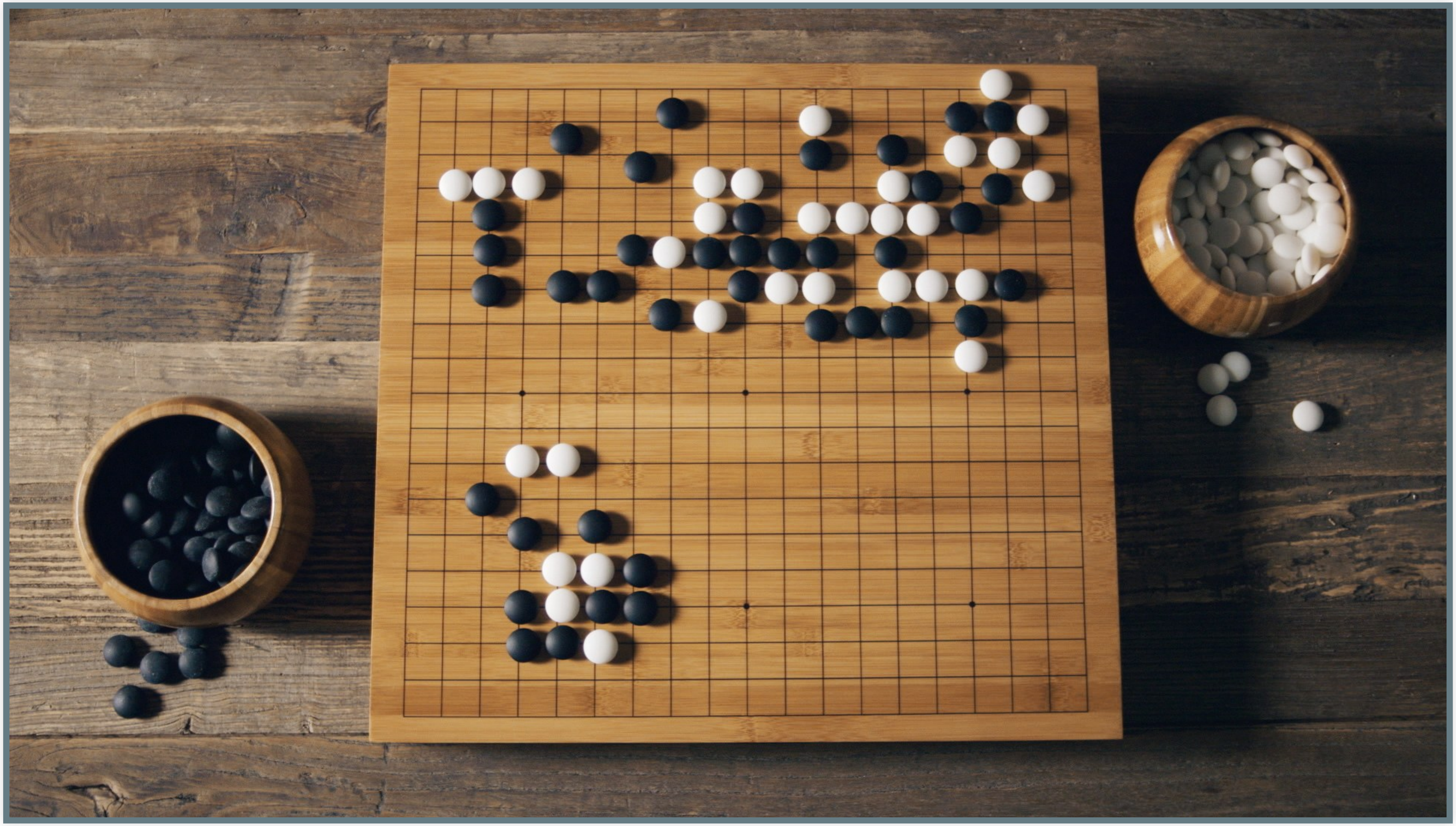
$[\alpha_0]$ t0 3I0

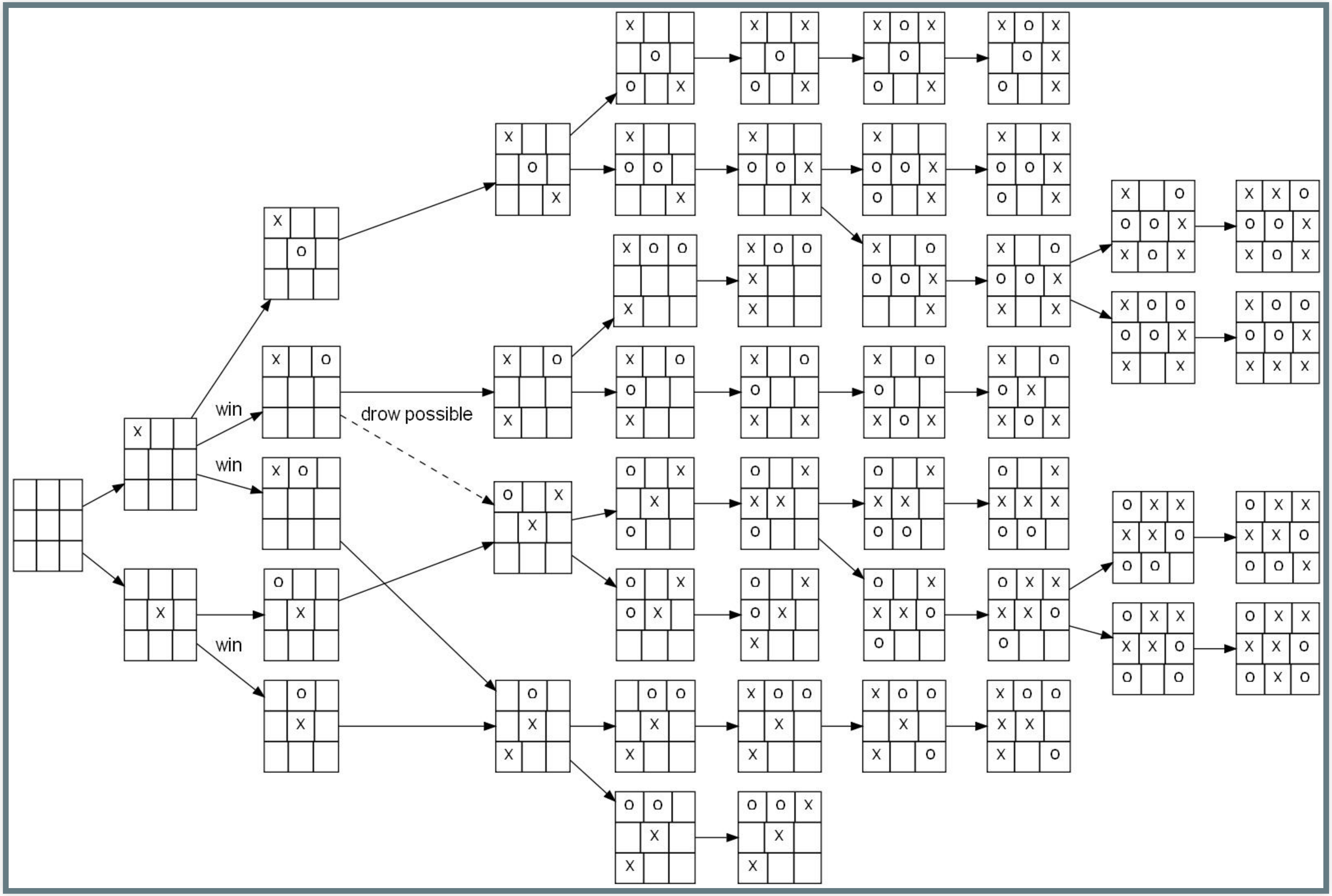
With are only 3 equations ...

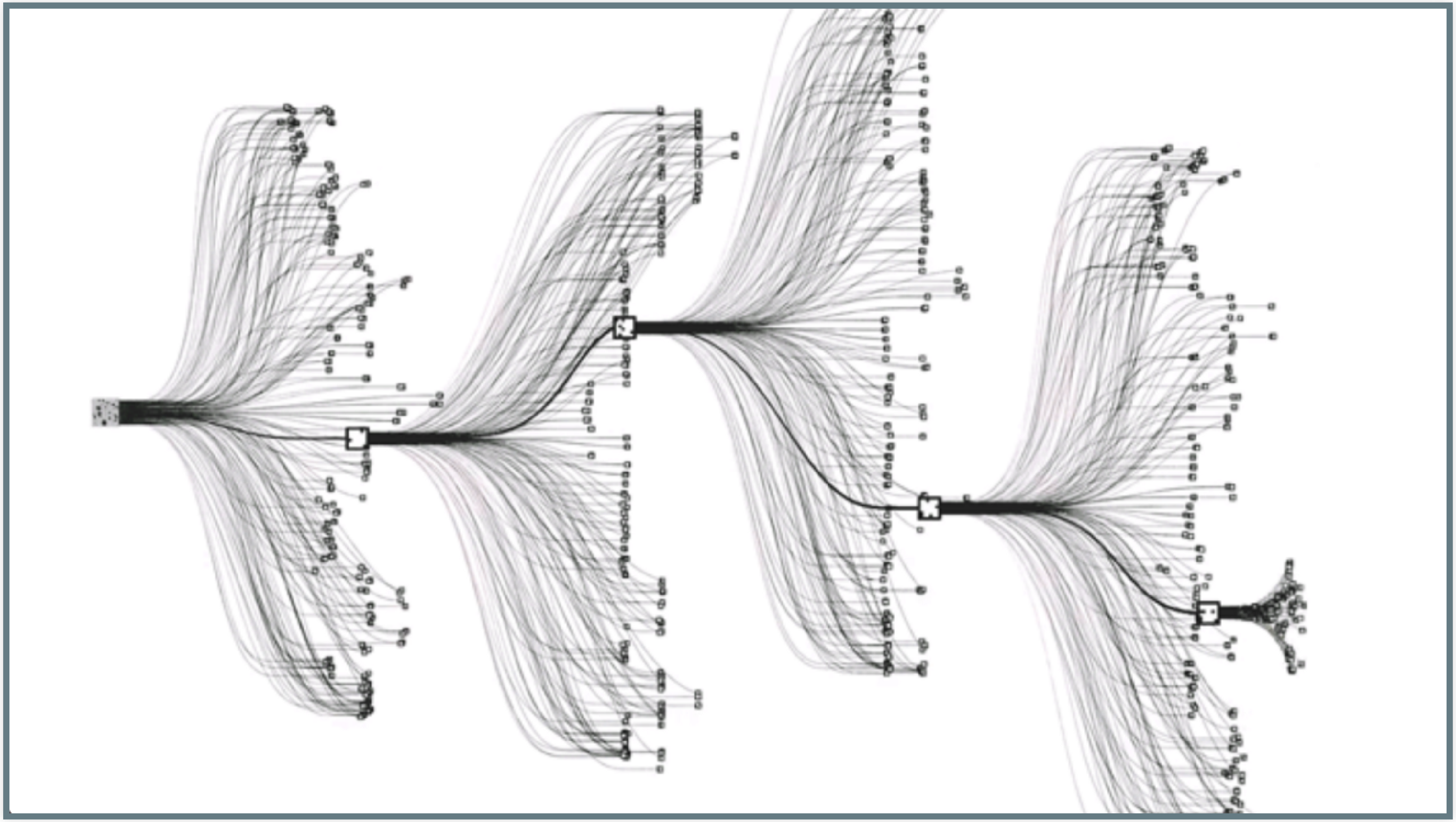
Outline

1. The math behind Go
2. From Crazy Stone -> AlphaGO
3. AlphaGo vs AlphaZero
4. Policy Iteration
5. Policy Improvement (Math alert!)
6. Policy Evaluation
7. Code and demo



- 10^{170} possible states
- 10^{360} possible games for each starting state
- 250 legal moves from each state
- 150 moves for each match


win
win
win
drow possible



AI in Go

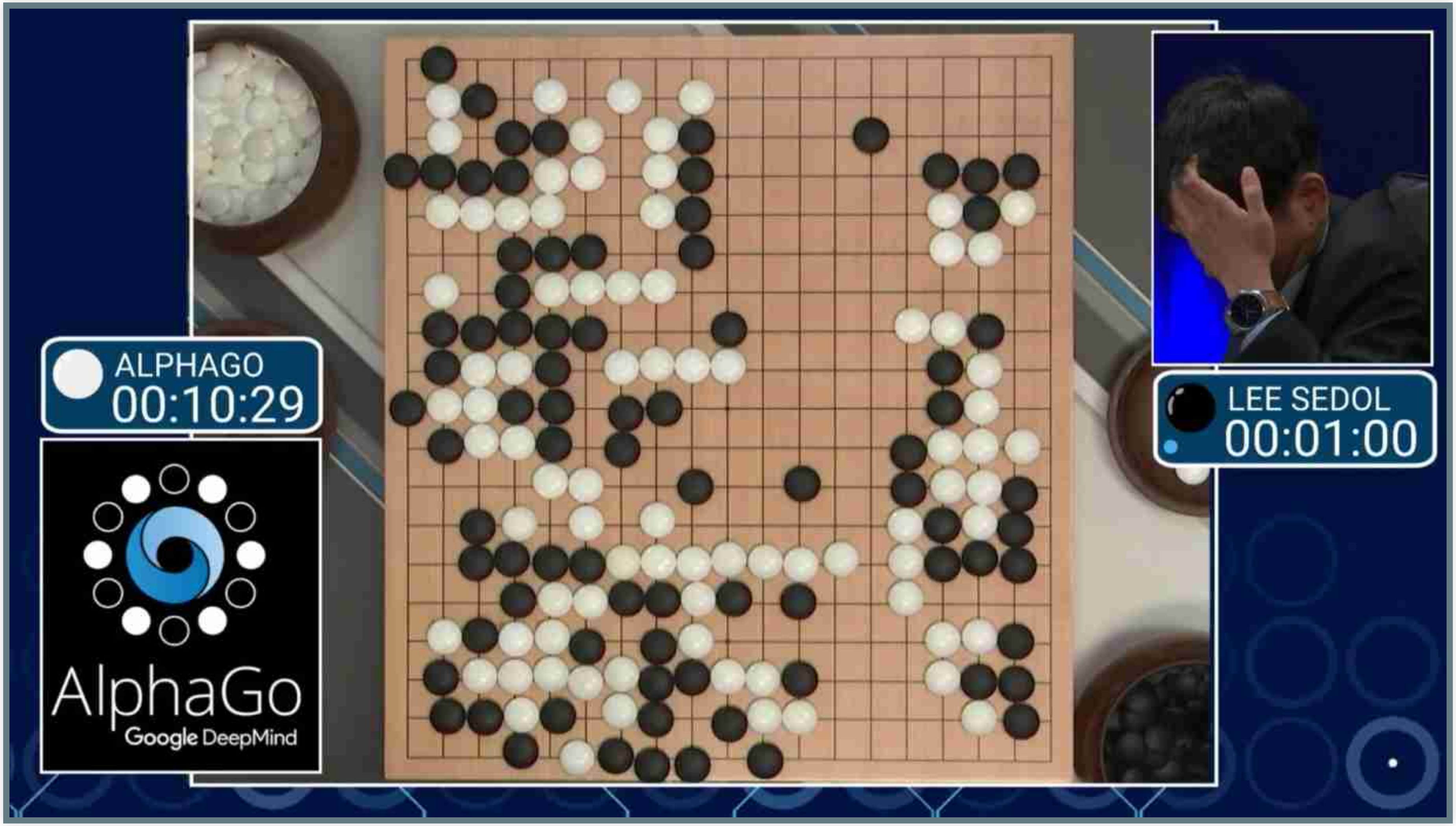
"The mystery of Go, the ancient game that computers still can't win" - Wired 2014

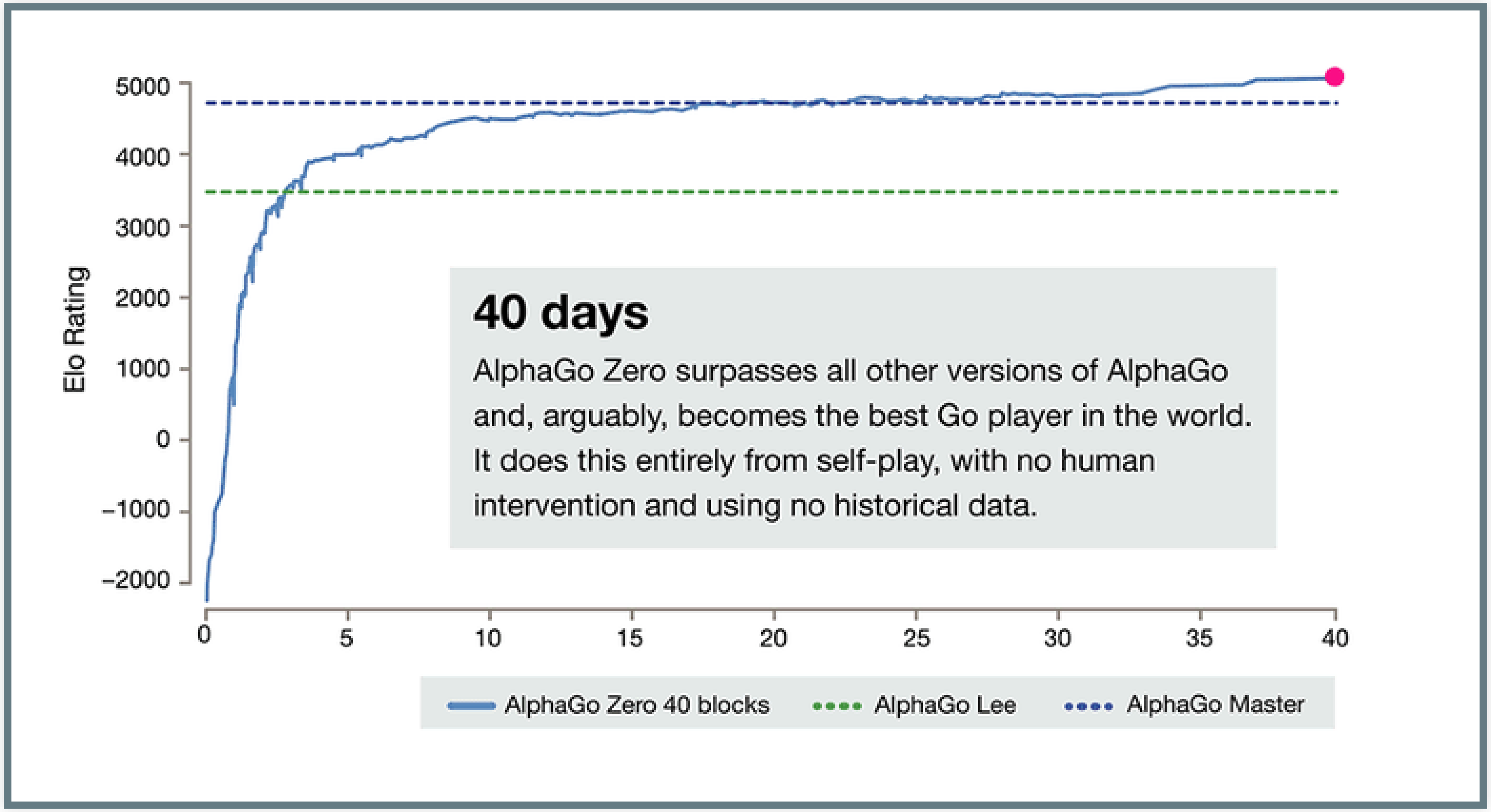
- Go is constructive
- Difficult to build an evaluation functional
- Humans describe more as intuititive game

- Adversarial
- Deterministic
- Fully observable

CrazyStone

AlphaGo


ALPHAGO
00:10:29
LEE SEDOL
00:01:00
AlphaGo
Google DeepMind

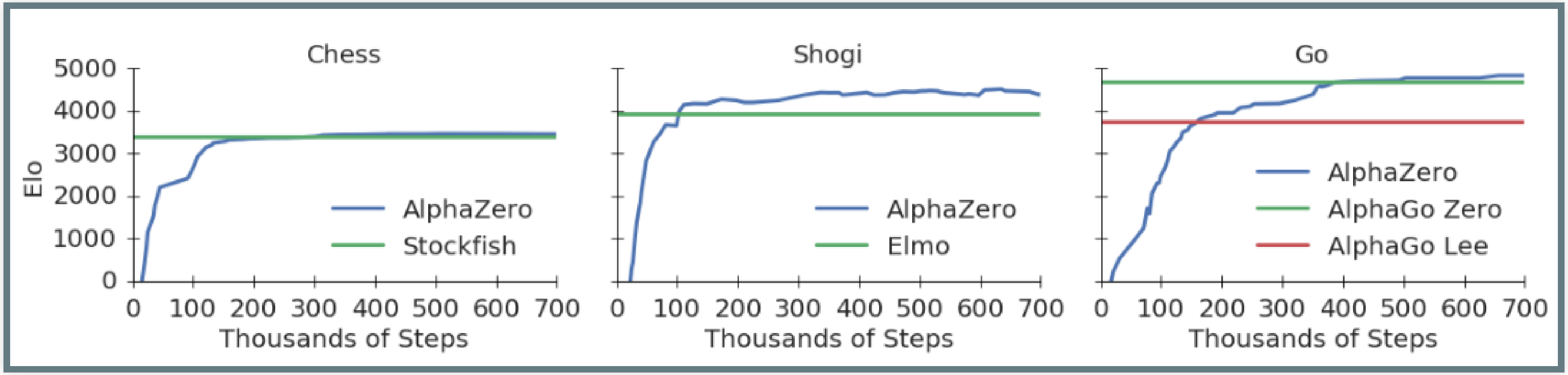


40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.

AlphaGo Zero vs AlphaZero

- No data augmentation
- No threshold update
- Diff. exploration noise for each game


Chess
Shogi
Go
Elo
5000
4000
3000
2000
1000
0
0 100 200 300 400 500 600 700
Thousands of Steps
AlphaZero
Stockfish
Elmo
AlphaGo Zero
AlphaGo Lee

Reinforcement Learning

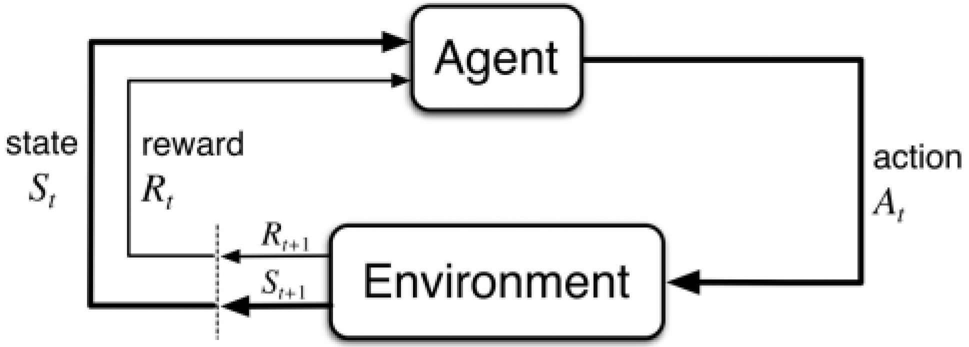


Figure 3.1: The agent–environment interaction in a Markov decision process.

$$v_\pi(s) = E_\pi\left[\sum_t \gamma^t R_t \mid S_t\right]$$

where:

$$\pi(a \mid s) = P(a \mid s) \;\; \forall s \in S$$

Policy Iteration

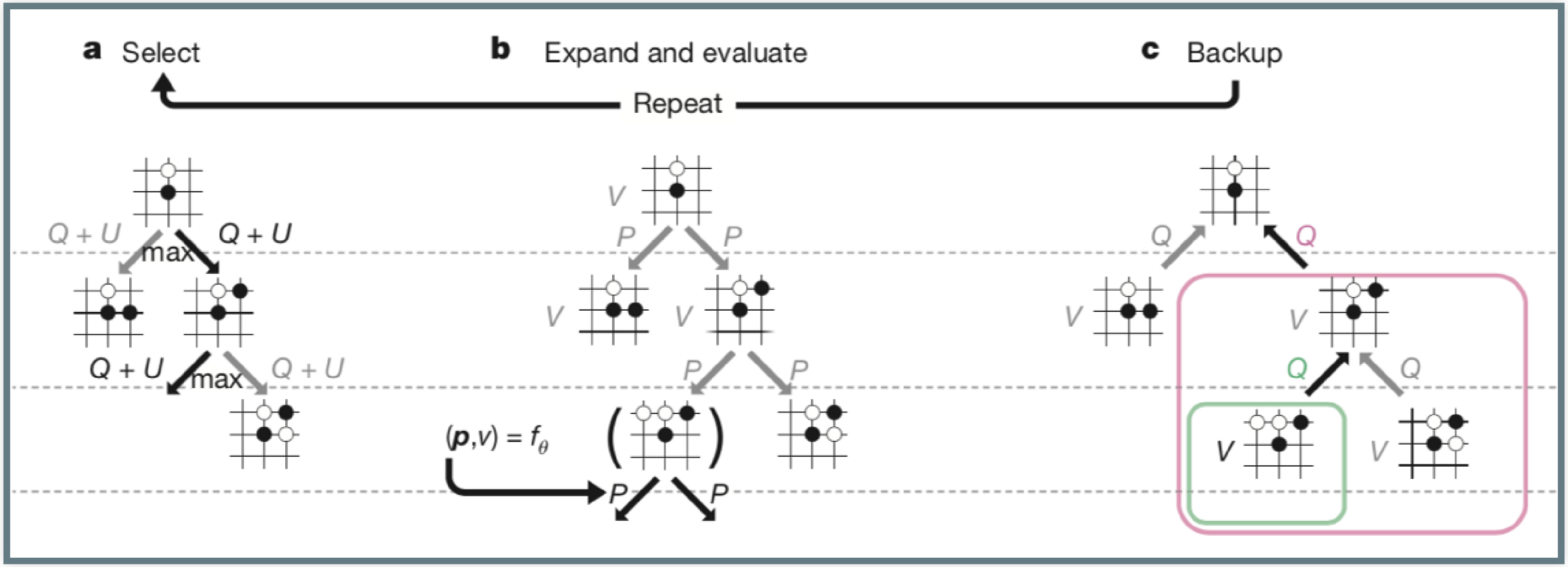
```
- Add sudo code policy iterat
```

Policy Improvement

Monte-Carlo Tree Search

MCTS is an algorithm to perform sampling based lookahead search.

```
- Add sudo code MCTS
```

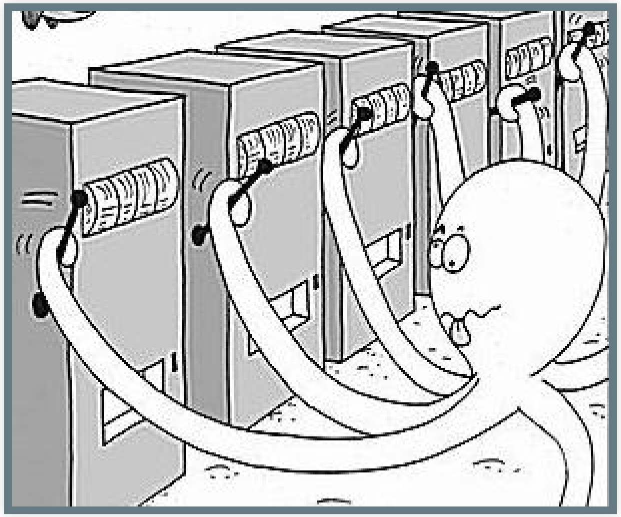

a Select
b Expand and evaluate
c Backup
Repeat
Q + U
Q + U
max
Q + U
max
Q + U
V
P
P
V
V
P
P
$(\boldsymbol{p}, v) = f_\theta$
P
P
Q
Q
V
V
Q
Q
V
V

With the backup operation we keep track of:

- N(s,a) visit count
- W(s,a) total action value
- Q(s,a) mean action value
- P(s,a) prior probability

Exploration

- $\epsilon - greedy$
- Bandits



$$cP(s, a)\frac{\sqrt{\sum_b N(s,b)}}{1+N(s,a)}$$

Policy Evaluation

Training

Architecture

Demo

Thank you!

github/mosc