Reinforcement Learning

나는 강화학습으로 축구한다

Google Research와 Manchester City F.C.의
인공지능 축구 프로젝트를 대한민국에서
재조명하고 직접 구현해보는 캠프

1

SESSION

한동대학교 빅데이터사업단
Winning Twelve / 전재영

Copyright Notice

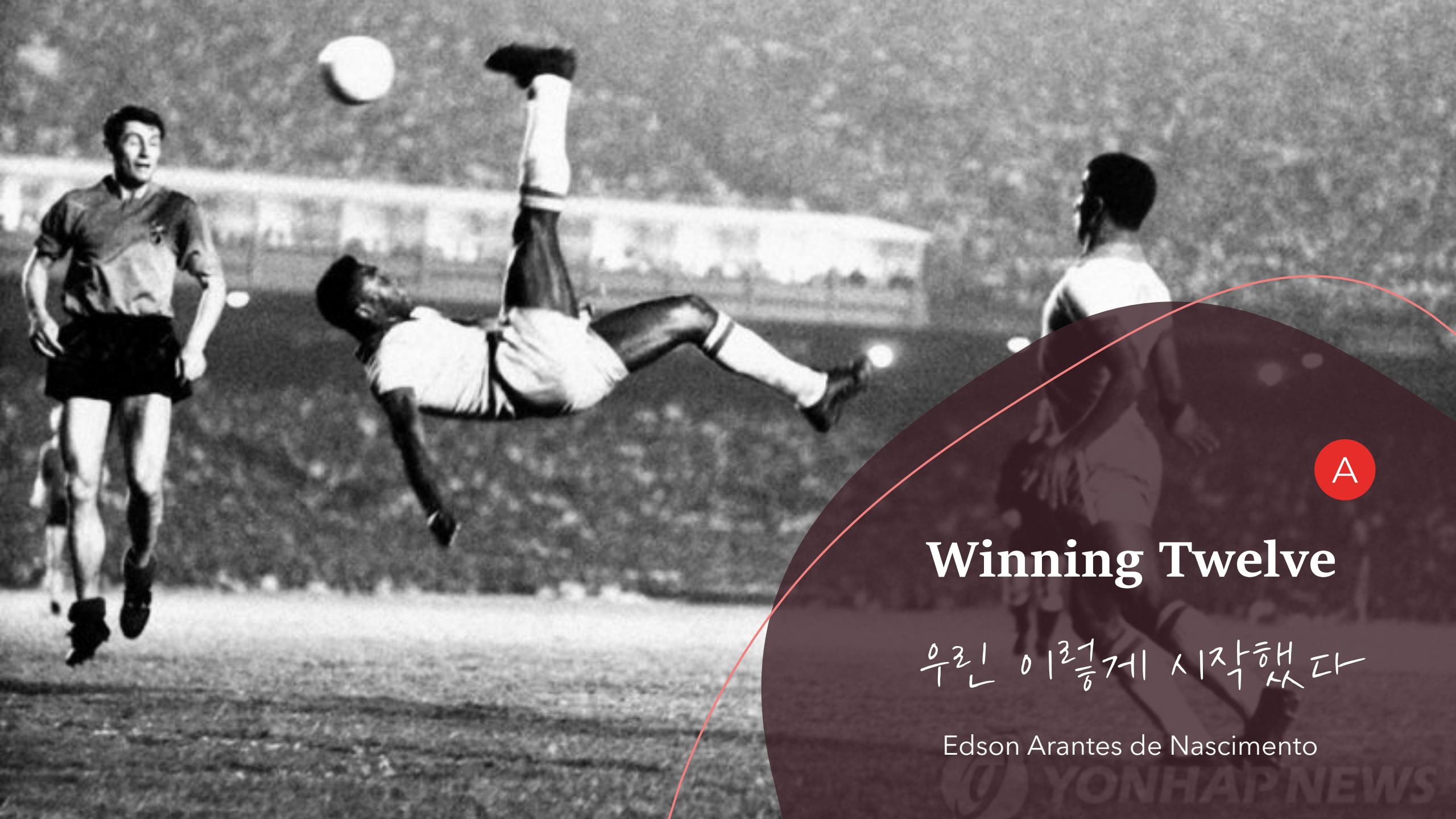
C

Google Research
Manchester City F.C.
재조명

B 강화학습

Winning Twelve

A



A

Winning Twelve

우린 이렇게 시작했다

Edson Arantes de Nascimento

A

Winning Twelve?

우린 이렇게 시작했다

- Edson Arantes de Nascimento
- Manchester United
- H-Milan

11 + 1

축구선수 데이터사이언티스트

Camp History

2023 Summer : 나는 데이터로 축구한다

2024 Summer : 나는 데이터로 축구한다

2025 Winter : 나는 데이터로 축구한다 (LLM/RAG 편)

2025 Summer : AI+X Ethiopia

2026 Winter : 나는 강화학습으로 축구한다

B

강화학습 / RL

Reinforcement Learning

- AlphaGo
- OpenAI Five (Dota 2)
- AlphaStar (Star Craft)
- OpenAI Gym
- …

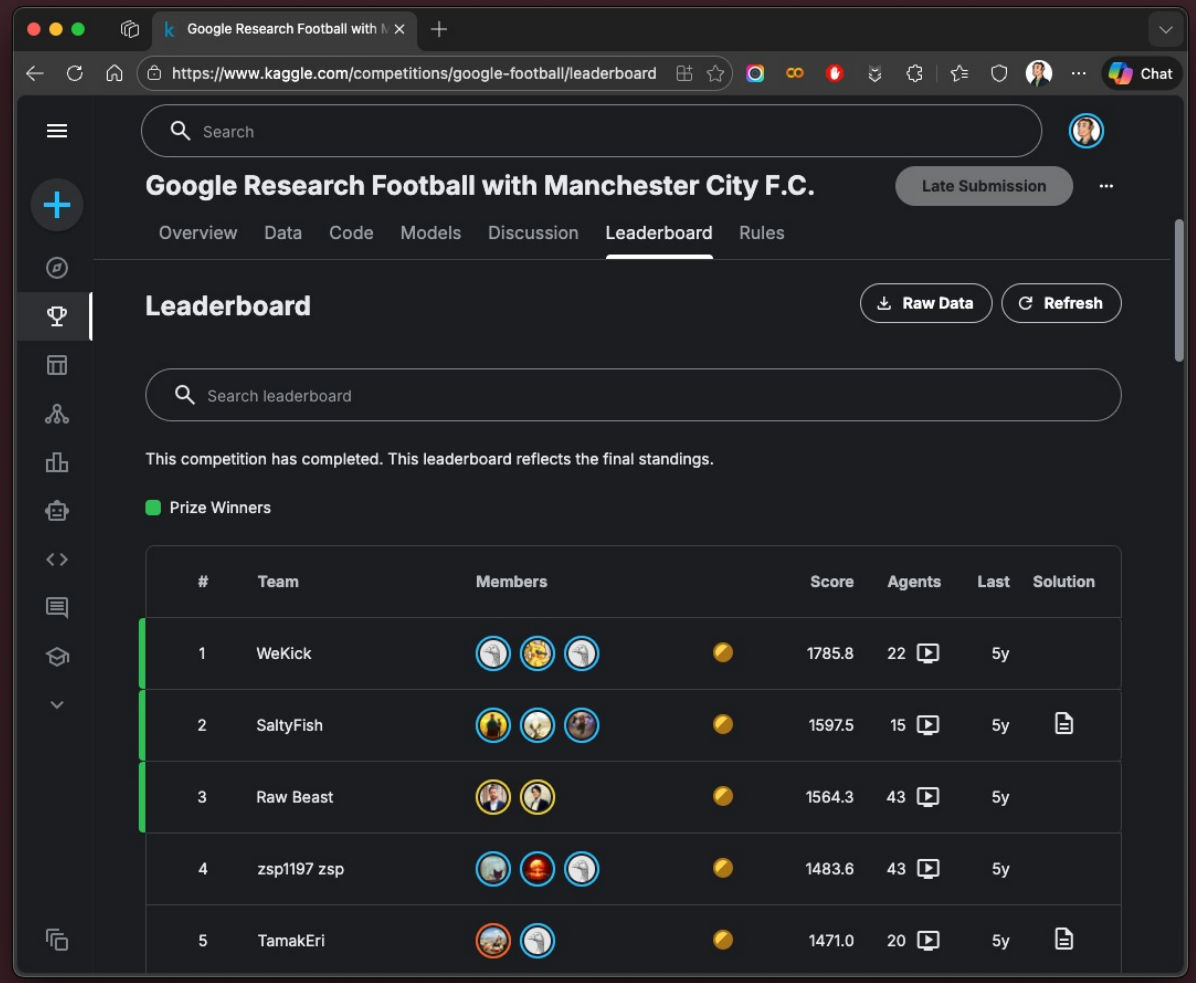
RL Applications B

- Games
- Robot, Drone, Self-Driving Car
- Traffic Signal Control
- Route Optimization
- Fleet Management
- Intrusion Detection
- Recommendation Systems
- Algorithmic Trading
- Dynamic Pricing
- Drug Dosage Optimization
- Human-in-the-loop Learning (RLHF)
- ...

C

2025/12/03 부터

결혼, 계절학기, 인턴, 탈영(?), ...



수학과 알고리즘

Maximize $J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[G_0]$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) G_t\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i) G_t^i$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_\theta(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i) G_t^i$$

$\boldsymbol{a} \sim \pi_\theta(\cdot\,|\boldsymbol{s}_t)$

- Markov Decision Process (MDP)
- Q-Learning
- Deep Q Network (DQN)
- Actor-Critic
- DDPG
- Proximal Policy Optimization (PPO)
- Impala CNN

Initialize policy $\boldsymbol{\pi_\theta}$ and value function $\boldsymbol{V_\phi}$
for iter = 1..K:
$\boldsymbol{\theta_{old}} \leftarrow \boldsymbol{\theta}$
$\boldsymbol{\phi_{old}} \leftarrow \boldsymbol{\phi}$

Collect rollouts with old policy
$\boldsymbol{D}$ = {($\boldsymbol{s}_{t,i}$, $\boldsymbol{a}_{t,i}$, $\boldsymbol{r}_{t+1,i}$, $\boldsymbol{s}_{t+1,i}$, $done_i$)}
collected by $\boldsymbol{\pi_{\theta,old}}$

Compute return and advantages
$R_{t,i} = r_{t+1,i} + \gamma\, \boldsymbol{V_{old}}(\boldsymbol{s}_{t+1,i}) \times$(not $done_i$)
$\widehat{\boldsymbol{A}}_{t,i} = R_{t,i} - \boldsymbol{V_{old}}(\boldsymbol{s}_{t,i})$

for epoch = 1..E:
for minibatch $\boldsymbol{B} \subset \boldsymbol{D}$:

Actor loss (clip objective)
$r_{t,i} = \exp(log\boldsymbol{\pi_\theta}(\boldsymbol{a}_{t,i}|\boldsymbol{s}_{t,i}) - log\boldsymbol{\pi_{\theta.old}}(\boldsymbol{a}_{t,i}|\boldsymbol{s}_{t,i}))$
$\boldsymbol{L_{clip}} = \text{mean}_{i\in B}(\min(r_{t,i}\times\widehat{\boldsymbol{A}}_{t,i}, \text{clip}(r_{t,i}, 1-\varepsilon, 1+\varepsilon)\times\widehat{\boldsymbol{A}}_{t,i}))$

Critic loss (MSE)
$\boldsymbol{L_v} = \text{mean}_{i\in B}(\boldsymbol{V_\phi}(\boldsymbol{s}_{t,i}) - R_{t,i})^2$

Update
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}\boldsymbol{L_{clip}}(\boldsymbol{\theta})$
$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta\nabla_{\boldsymbol{\phi}}\boldsymbol{L_v}(\boldsymbol{\phi})$

Cost & Time

> First, I tried a single V100 GPU, 12 vCPU cores, 32 GB of RAM, 128Gb regular disk. With this machine, I achieved about 17 SPS (Steps per second) and about 50–60 Game FPS (Frames per Second) for each environment. We'll run 16 environments in parallel, per environment, averaging ~280 FPS total. Estimated to take 50h and cost 108$ for 50M steps.

$108

50 HOURS

2019 년이지만...

시간이 없다.

돈도 없다.

쉽지 않다.

Agenda

Football Academy

Environment, Rules, ...

RL Study

Competition

academy_empty_goal_close


00:01
FRQ 0
RBA 0
TURING
academy_empty_goal


00:01
FRQ 0
RBA 0
ARCHIMEDINHO
TURING
academy_run_to_score


00:00
FRQ 0
RBA 0
academy_run_to_score_with_keeper

academy_pass_and_shoot_with_keeper

academy_run_pass_and_shoot_with_keeper

academy_3_vs_1_with_keeper


00:01
FRQ 0
RBA 0
BANNEKER
TURING
BlauwPrint
Indietopia
Indietopia
academy_corner


00:01
FRQ 0
RBA 0
NEWTON
academy_single_goal_versus_lazy


00:01
FRQ 0
RBA 0
1_vs_1_easy


00:01
FRQ 0
RBA 0
EINSTEIN
JOHNSON
5_vs_5

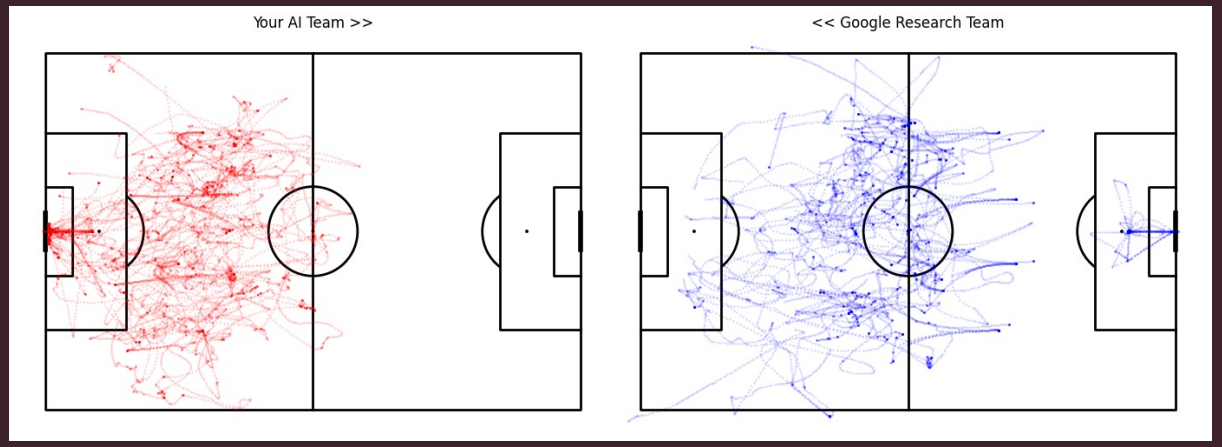

00:01
FRQ 0
RBA 0
EINSTEIN
TURING
11_vs_11_competition

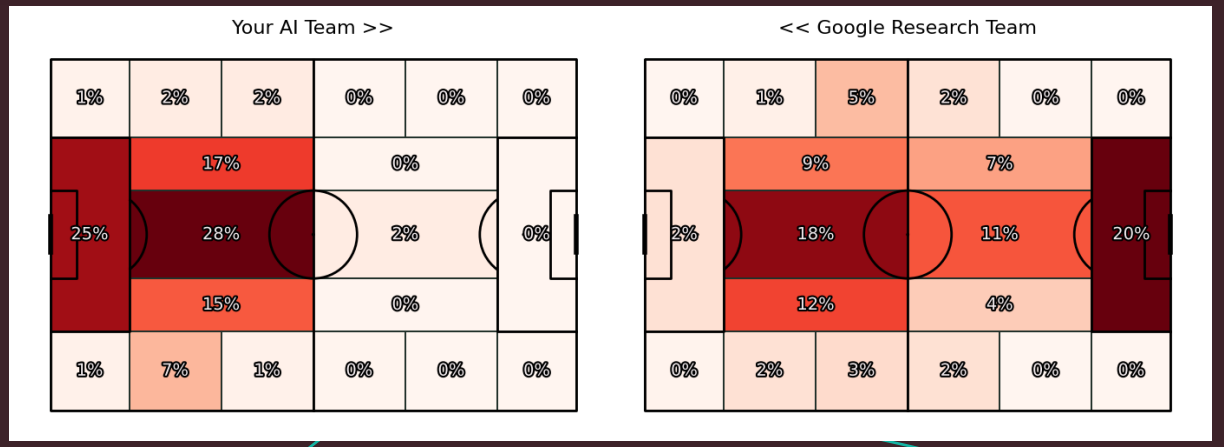
Five Scenarios

1. academy_empty_goal_close
2. academy_empty_goal
3. academy_run_to_score
4. academy_run_to_score_with_keeper
5. 5_vs_5

Winning Twelve World Cup

~~11_vs_11~~ **5_vs_5**

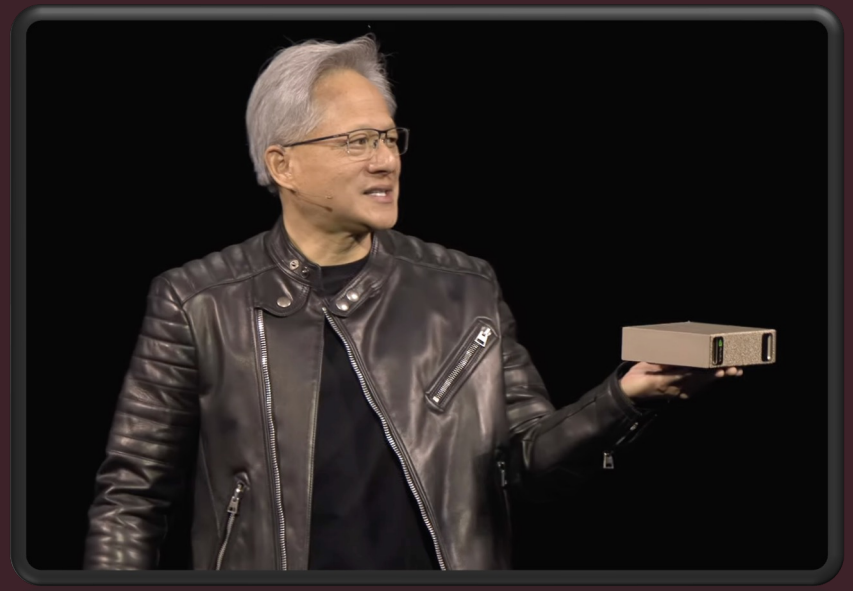


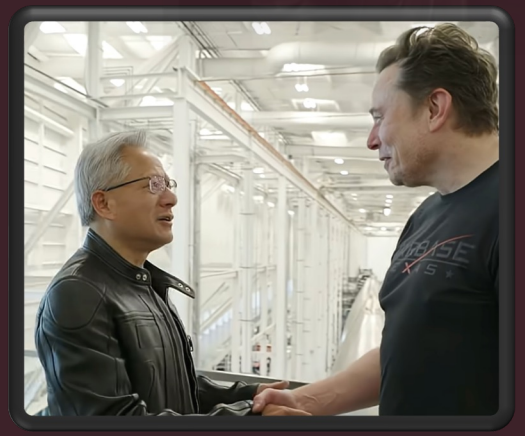


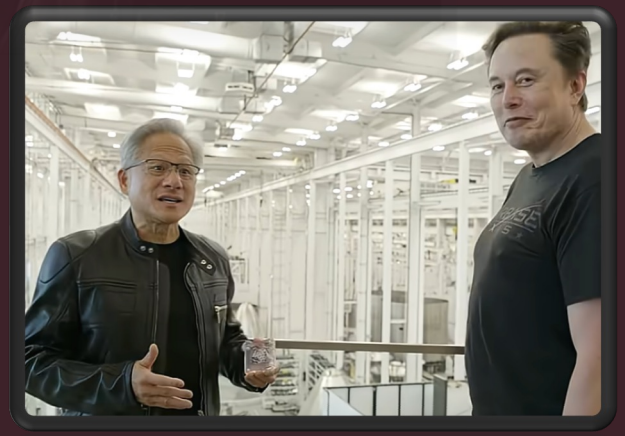
Leaderboard

- Ice breaking
- Run to Score with Keeper
- 시간엄수
- ...

NVIDIA DGX Spark @ GTC 2025 Keynote







Hand delivery to Elon Musk @ SpaceX