# Deep RL and Beyond

2019 Fall

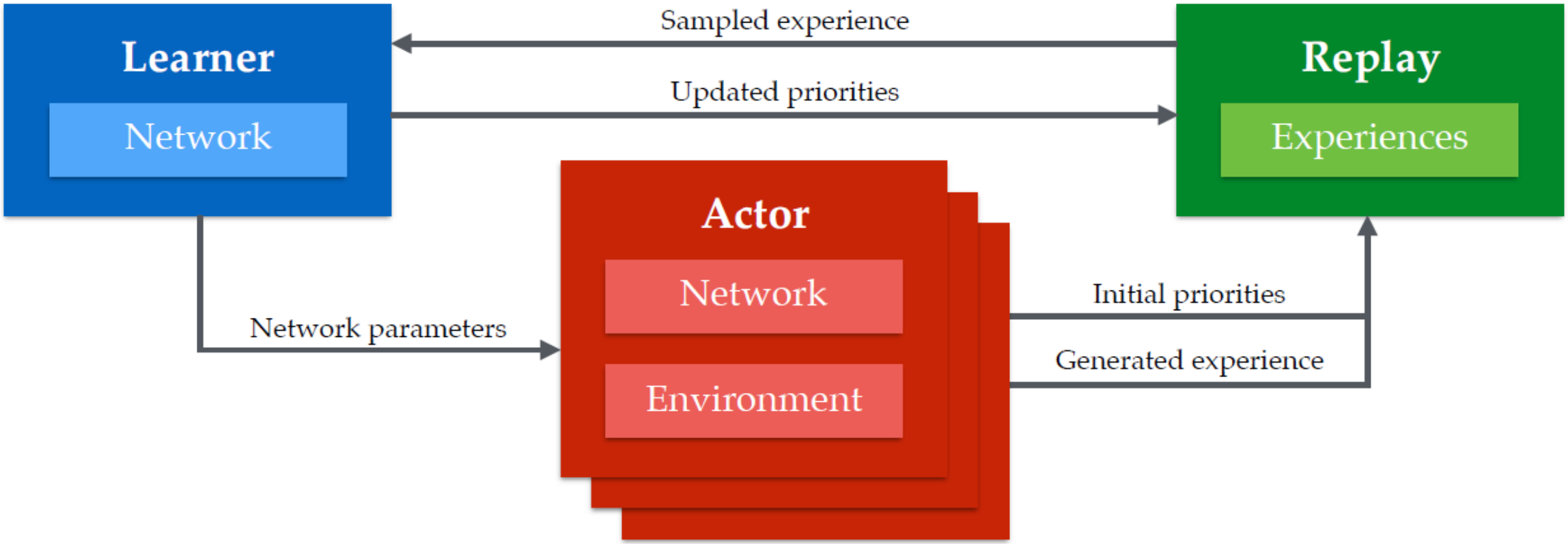
Yusung Kim

yskim525@skku.edu

# Distributed Prioritized Experience Replay [ICLR 2018]

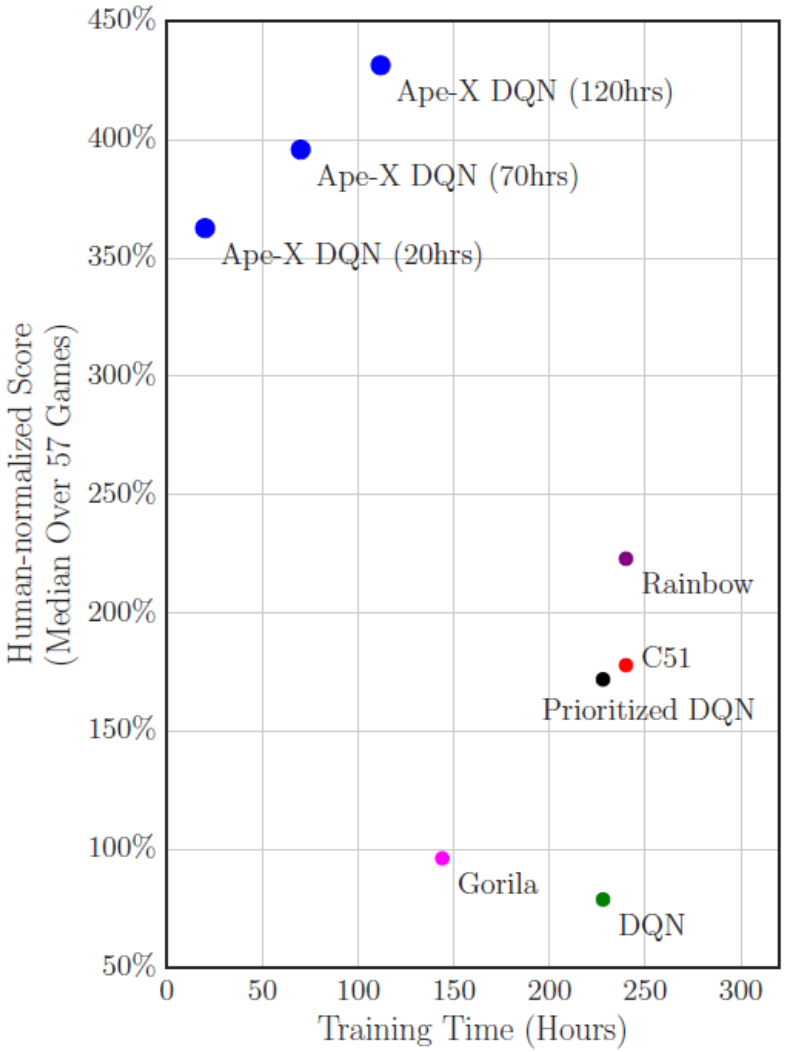
- A distributed architecture based on Prioritized Experience Replay
- Actors paly and store Experience in shared memory
  - each actor uses a different epsilon value for $\epsilon$-greedy
  - periodically copy $\theta$ from Learner
- A single Learner with GPU learns from shared Experience Replay
  - Dueling Double Q-learning with PER & Multi-Step

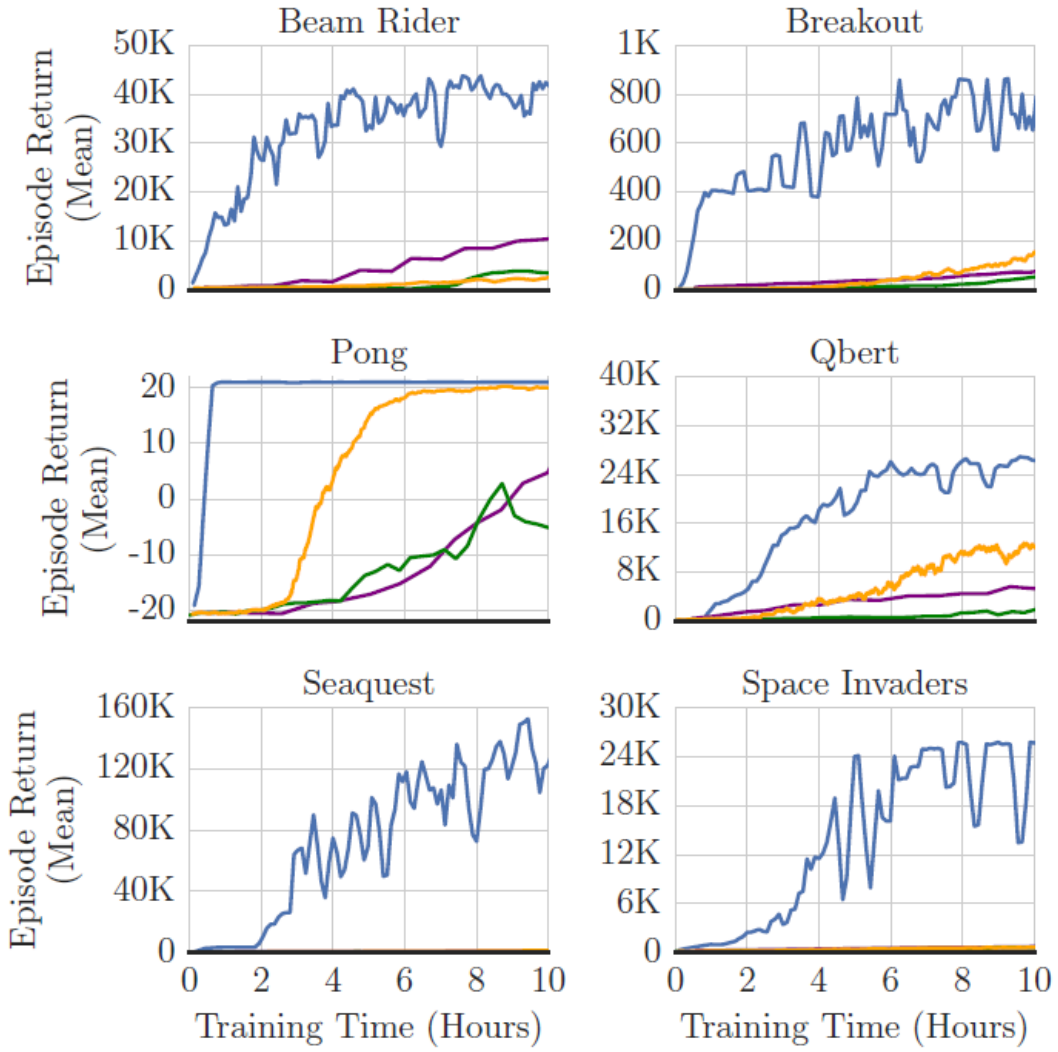
# Ape-X Architecture




360 actors, 139 FPS per actor ( 360 * 139 = 50K FPS )

# Performance Results

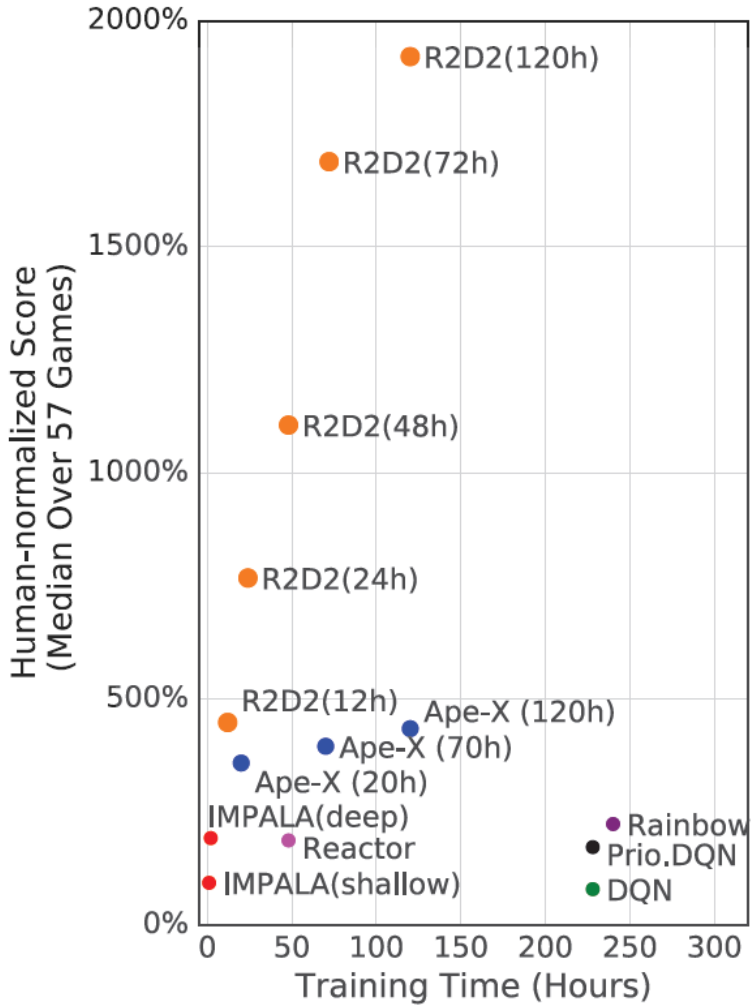


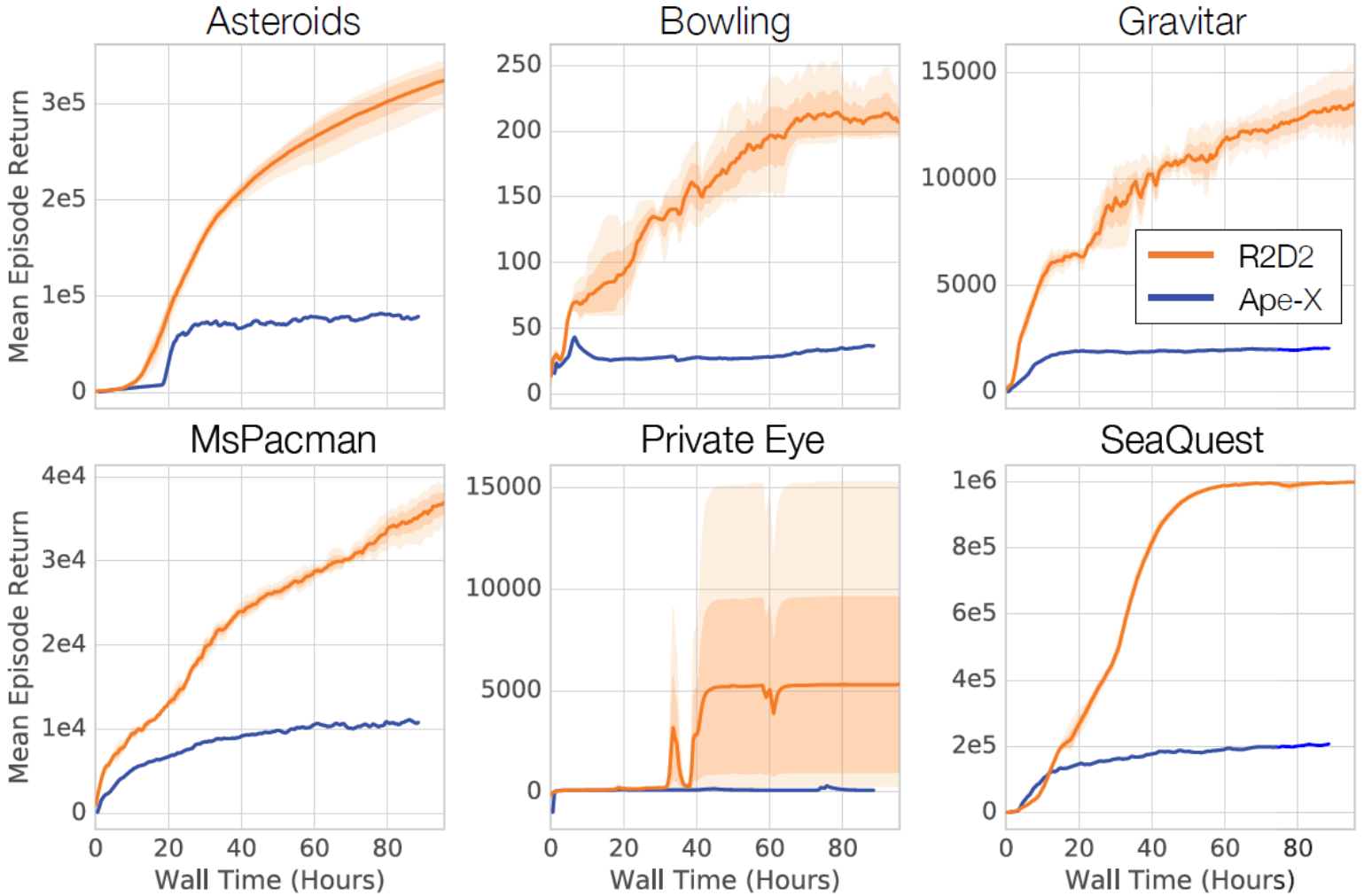



# Recurrent Experience Replay in Distributed Reinforcement Learning [ICLR 2019]

- Investigate the training of RNN-based RL agents from distributed prioritized experience replay
- Study the effects of parameter lag resulting in representational drift and recurrent state staleness
- The first record to exceed human-level performance in 52 of the 57 Atari games !

# R2D2 Performance Results



# Montezuma Revenge ...

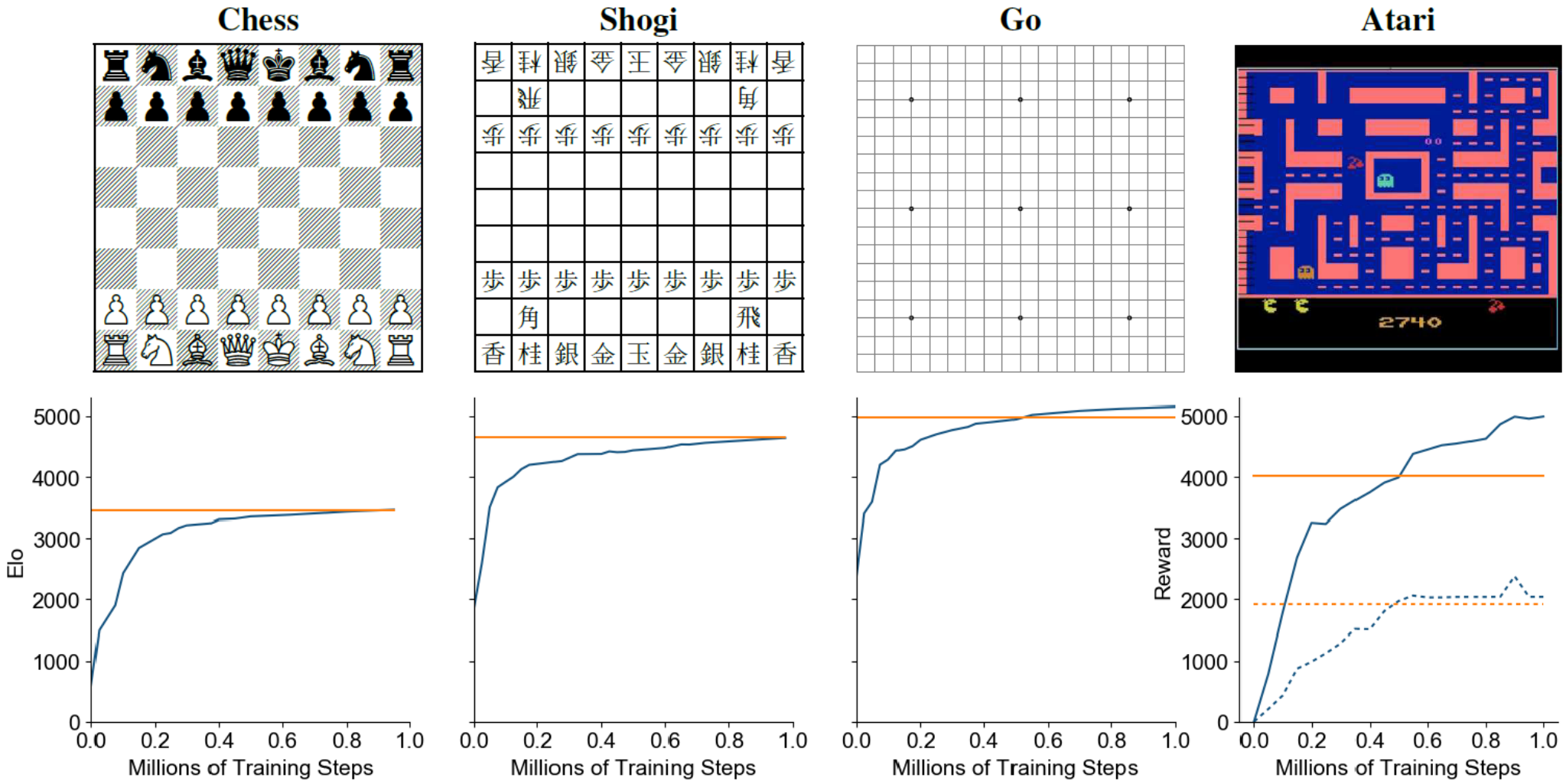
| GAMES | HUMAN | REACTOR | IMPALA (S/D) | APE-X | R2D2 |
|---|---|---|---|---|---|
| frostbite | 4334.7 | 8042.1 | 269.6/317.8 | 9328.6 | **315456.4** |
| gopher | 2412.5 | 69135.1 | 1002.4/66782.3 | 120500.9 | **124776.3** |
| gravitar | 3351.4 | 1073.8 | 211.5/359.5 | 1598.5 | **15680.7** |
| hero | 30826.4 | 35542.2 | 33853.2/33730.6 | 31655.9 | **39537.1** |
| ice_hockey | 0.9 | 3.4 | -5.3/3.5 | 33.0 | **79.3** |
| jamesbond | 302.8 | 7869.3 | 440.0/601.5 | 21322.5 | **25354.0** |
| kangaroo | 3035.0 | 10484.5 | 47.0/1632.0 | 1416.0 | **14130.7** |
| krull | 2665.5 | 9930.9 | 9247.6/8147.4 | 11741.4 | **218448.1** |
| kung_fu_master | 22736.3 | 59799.5 | 42259.0/43375.5 | 97829.5 | **233413.3** |
| montezuma_revenge | **4753.3** | 2643.5 | 0.0/0.0 | 2500.0 | 2061.3 |
| ms_pacman | 6951.6 | 2724.3 | 6501.7/7342.3 | 11255.2 | **42281.7** |
| name_this_game | 8049.0 | 9907.1 | 6049.6/21537.2 | 25783.3 | **58182.7** |
| phoenix | 7242.6 | 40092.3 | 33068.2/210996.5 | 224491.1 | **864020.0** |

# Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model [DeepMind 2019]

- A tree-based search with a learned model
- Without any knowledge of the underlying dynamics
  - No knowledge of the game rules

    ( Go, Chess, Shogi, and 57 Atari games )

# MuZero Performance Results



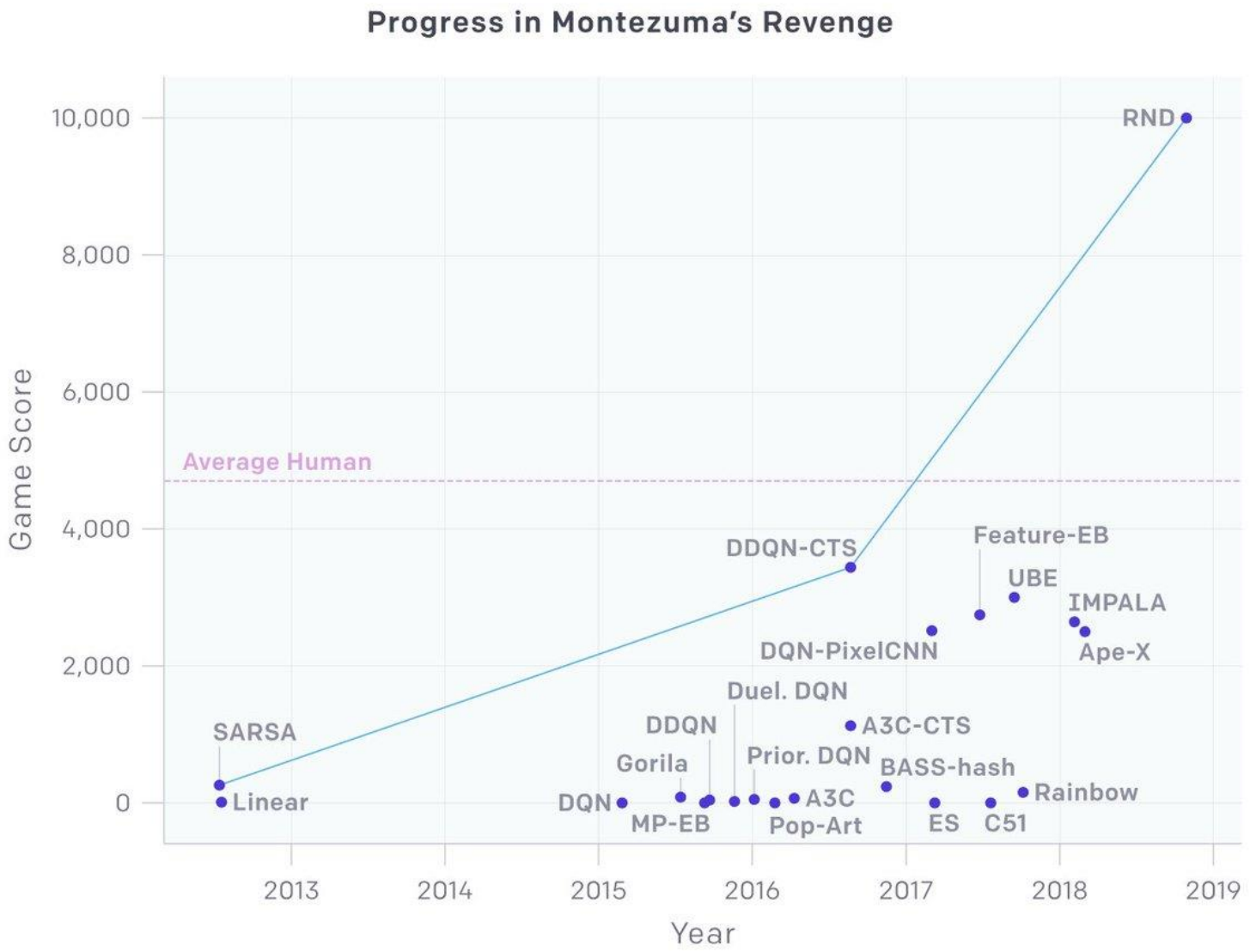
# Montezuma Revenge ...

| Game | Random | Human | SimPLe [20] | Ape-X [18] | R2D2 [21] | *MuZero* | *MuZero* normalized |
|---|---|---|---|---|---|---|---|
| centipede | 2,090.87 | 12,017.04 | - | 12,974.00 | 599,140.30 | **1,159,049.27** | 11,655.6 % |
| chopper command | 811.00 | 7,387.80 | 979.40 | 721,851.00 | 986,652.00 | **991,039.70** | 15,056.4 % |
| crazy climber | 10,780.50 | 35,829.41 | 62,583.60 | 320,426.00 | 366,690.70 | **458,315.40** | 1,786.6 % |
| defender | 2,874.50 | 18,688.89 | - | 411,944.00 | 665,792.00 | **839,642.95** | 5,291.2 % |
| demon attack | 152.07 | 1,971.00 | 208.10 | 133,086.00 | 140,002.30 | **143,964.26** | 7,906.4 % |
| double dunk | -18.55 | -16.40 | - | **24.00** | 23.70 | 23.94 | 1,976.3 % |
| enduro | 0.00 | 860.53 | - | 2,177.00 | 2,372.70 | **2,382.44** | 276.9 % |
| fishing derby | -91.71 | -38.80 | -90.70 | 44.00 | 85.80 | **91.16** | 345.6 % |
| freeway | 0.01 | 29.60 | 16.70 | **34.00** | 32.50 | 33.03 | 111.6 % |
| frostbite | 65.20 | 4,334.67 | 236.90 | 9,329.00 | 315,456.40 | **631,378.53** | 14,786.7 % |
| gopher | 257.60 | 2,412.50 | 596.80 | 120,501.00 | 124,776.30 | **130,345.58** | 6,036.8 % |
| gravitar | 173.00 | 3,351.43 | 173.40 | 1,599.00 | **15,680.70** | 6,682.70 | 204.8 % |
| hero | 1,026.97 | 30,826.38 | 2,656.60 | 31,656.00 | 39,537.10 | **49,244.11** | 161.8 % |
| ice hockey | -11.15 | 0.88 | -11.60 | 33.00 | **79.30** | 67.04 | 650.0 % |
| jamesbond | 29.00 | 302.80 | 100.50 | 21,323.00 | 25,354.00 | **41,063.25** | 14,986.9 % |
| kangaroo | 52.00 | 3,035.00 | 51.20 | 1,416.00 | 14,130.70 | **16,763.60** | 560.2 % |
| krull | 1,598.05 | 2,665.53 | 2,204.80 | 11,741.00 | 218,448.10 | **269,358.27** | 25,083.4 % |
| kung fu master | 258.50 | 22,736.25 | 14,862.50 | 97,830.00 | **233,413.30** | 204,824.00 | 910.1 % |
| montezuma revenge | 0.00 | **4,753.33** | - | 2,500.00 | 2,061.30 | 0.00 | 0.0 % |
| ms pacman | 307.30 | 6,951.60 | 1,480.00 | 11,255.00 | 42,281.70 | **243,401.10** | 3,658.7 % |
| name this game | 2,292.35 | 8,049.00 | 2,420.70 | 25,783.00 | 58,182.70 | **157,177.85** | 2,690.5 % |
| phoenix | 761.40 | 7,242.60 | - | 224,491.00 | 864,020.00 | **955,137.84** | 14,725.3 % |

# Other Interesting Topics in RL

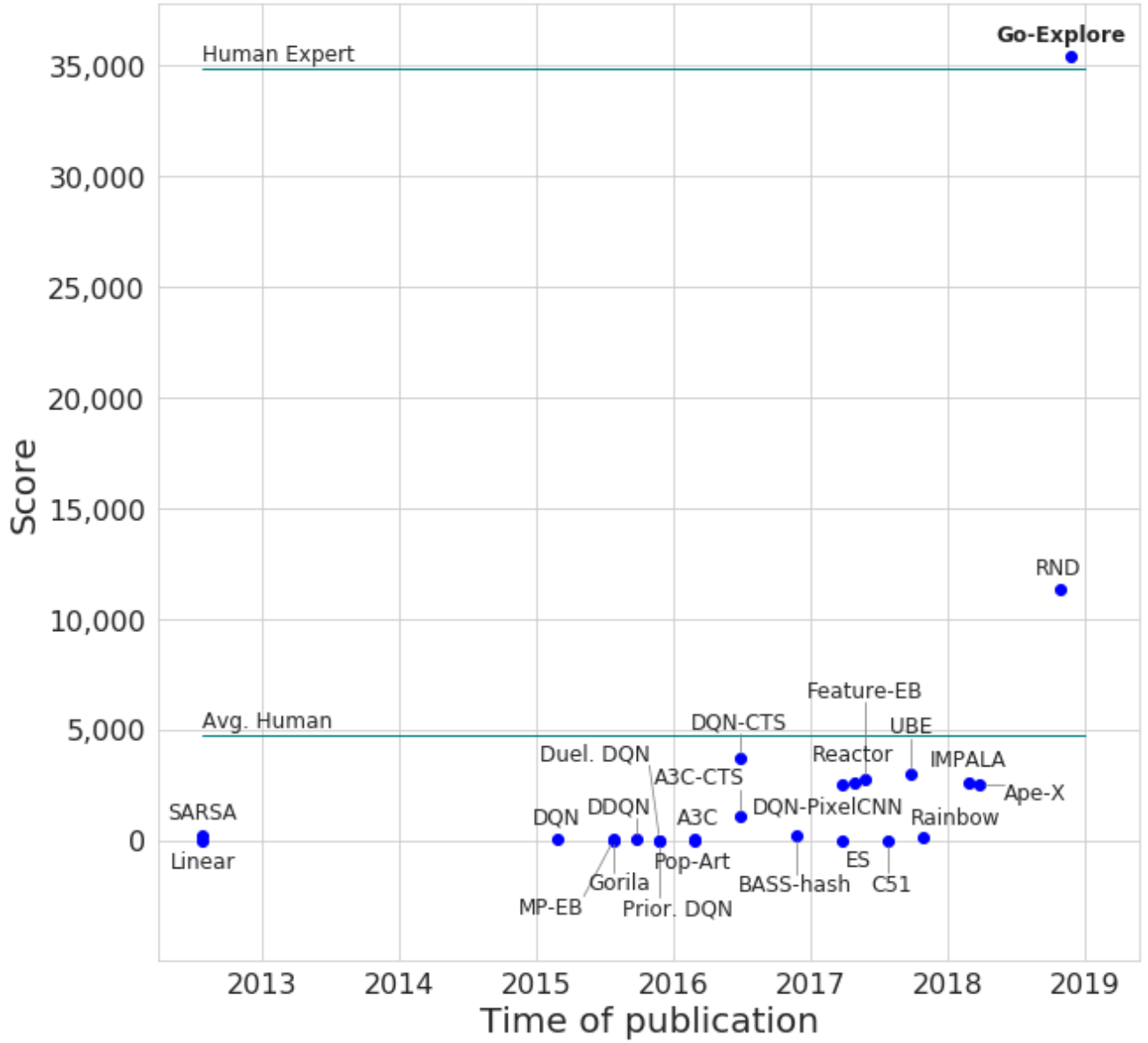
- Transfer, Multitask, and/or Meta Reinforcement Learning
- Imitation Learning and Inverse Reinforcement Learning
- Multi-Agent Reinforcement Learning
- Safe Reinforcement Learning
- Exploration / Intrinsic Motivation

# Exploration by Random Network Distillation [ICLR 2019]



# Go-Explore: a New Approach for Hard-Exploration Problems [Uber-AI 2019]

The mean score is 35,410 which beats human experts for the first time !




https://eng.uber.com/go-explore/

# Concluding Remarks

- Have studied from MDP to Deep Reinforcement Learning.
- RL researches are getting general, stable, and scalable.
- Many open source codes & various RL environments
- Interests about RL have been growing rapidly.
- Above all, RL is really fun !

Emergence of Locomotion Behaviours in Rich Environments
https://www.youtube.com/watch?v=hx_bgoTF7bs&t=98s