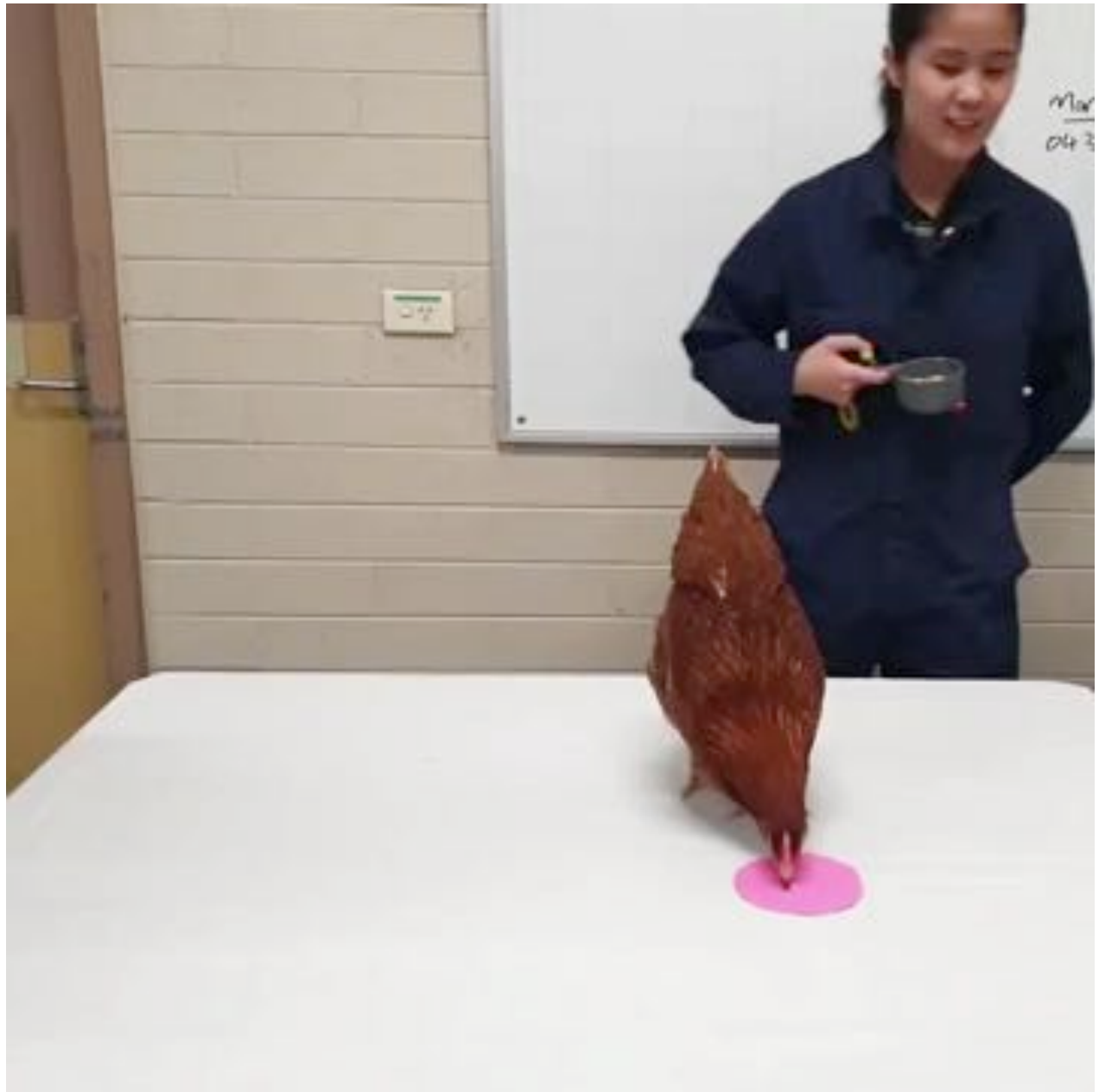
Week 11: Real-World RL

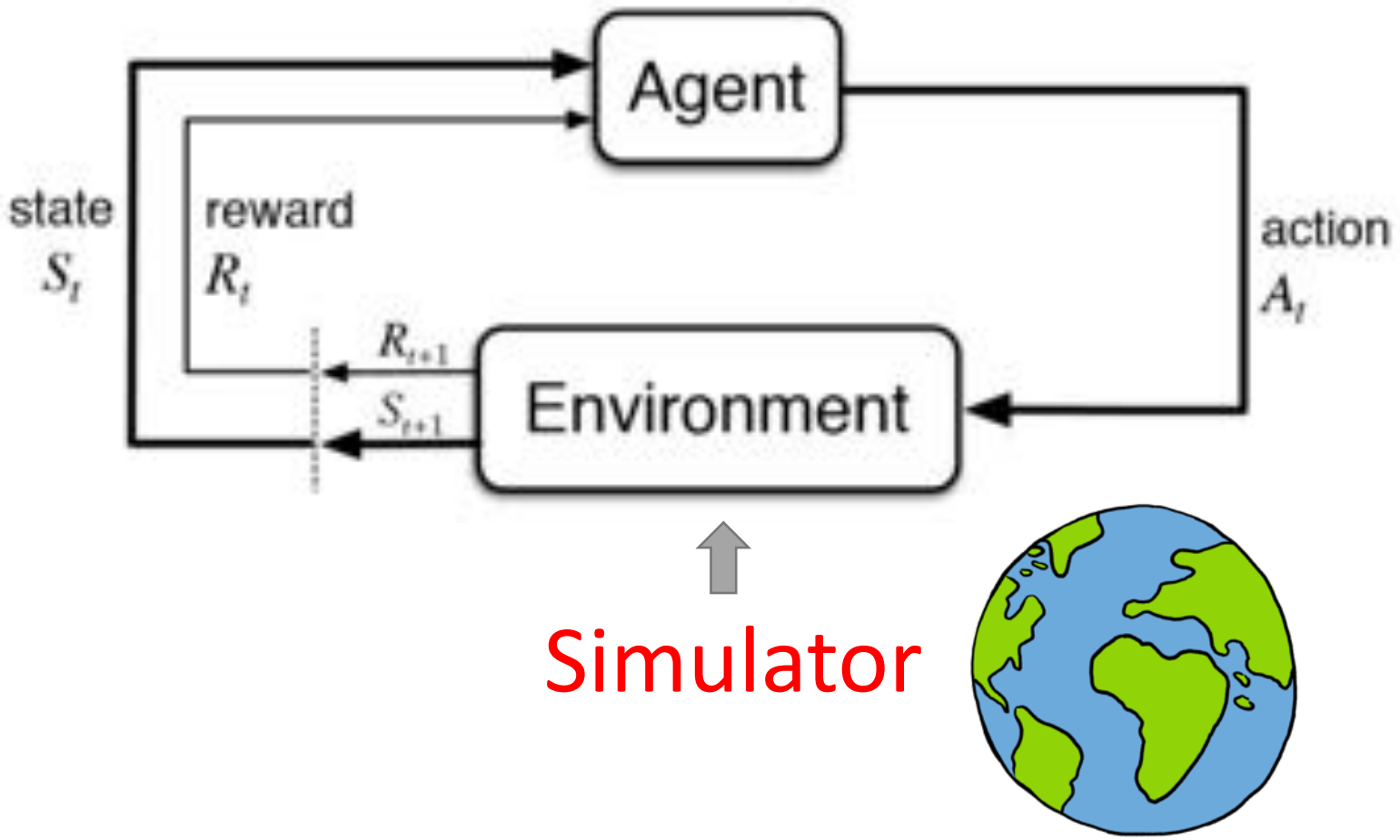
Bolei Zhou
The Chinese University of Hong Kong



Outline

- Real-world RL
- Offline RL
- RL applications

Simulator RL versus Real-World RL



Priorities in Real-World RL

- Policy Gradient ↓
- Complex representations ↓
- Computational efficiency ↓
- Control environment ↓
- Learning ↓
- Last policy ↓

Guided Policy Learning ↑

Generalization ↑

Sample efficiency ↑

Environment controls ↑

Evaluation ↑

Every policy ↑

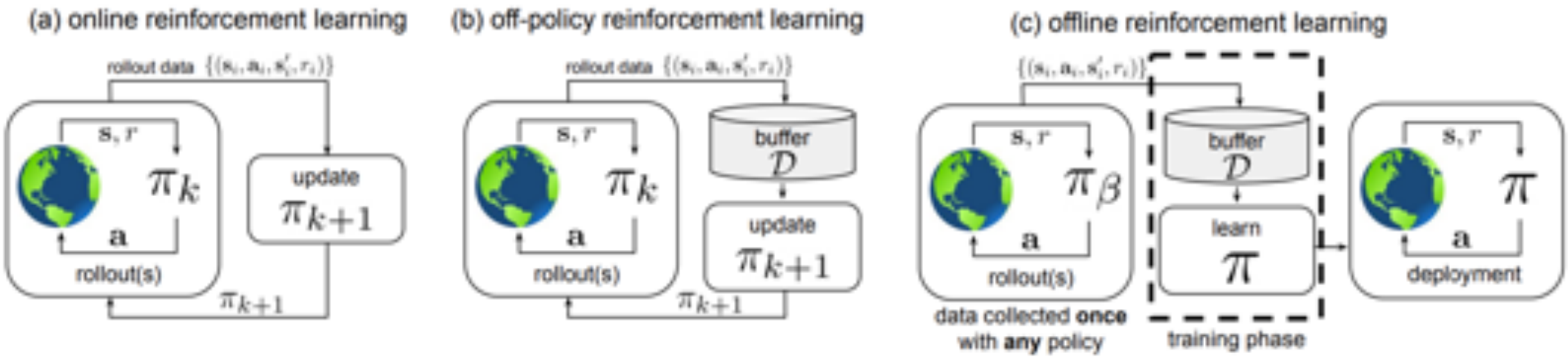
Offline Reinforcement Learning

- Power plants, robots, healthcare systems, or self-driving cars are expensive to run
- Goal of offline RL to learn a policy directly from that logged data without interacting with the environment.

https://deepmind.com/research/publications/RL-Unplugged-Benchmarks-for-Offline-Reinforcement-Learning

Offline Reinforcement Learning

- Problem Setup: data-driven formulation of RL problem
- Training set: static dataset of transitions $\mathcal{D} = \{(\mathbf{s}_t^i, \mathbf{a}_t^i, \mathbf{s}_{t+1}^i, r_t^i)\}$



A very recent survey on offline RL: https://arxiv.org/pdf/2005.01643.pdf

Offline Reinforcement Learning

- Fundamental challenge: Distributional shift
- Our function approximator (policy, value function, or model) might be trained under one distribution on the static dataset, it will be evaluated on a different distribution

Off-policy evaluation via importance sampling

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\sum_{t=0}^{H}\left(\prod_{t'=0}^{t}\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\beta(\mathbf{a}_t|\mathbf{s}_t)}\right)\gamma^t r(\mathbf{s},\mathbf{a})\right] \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H} w_t^i \gamma^t r_t^i.$$

Off-policy policy gradient

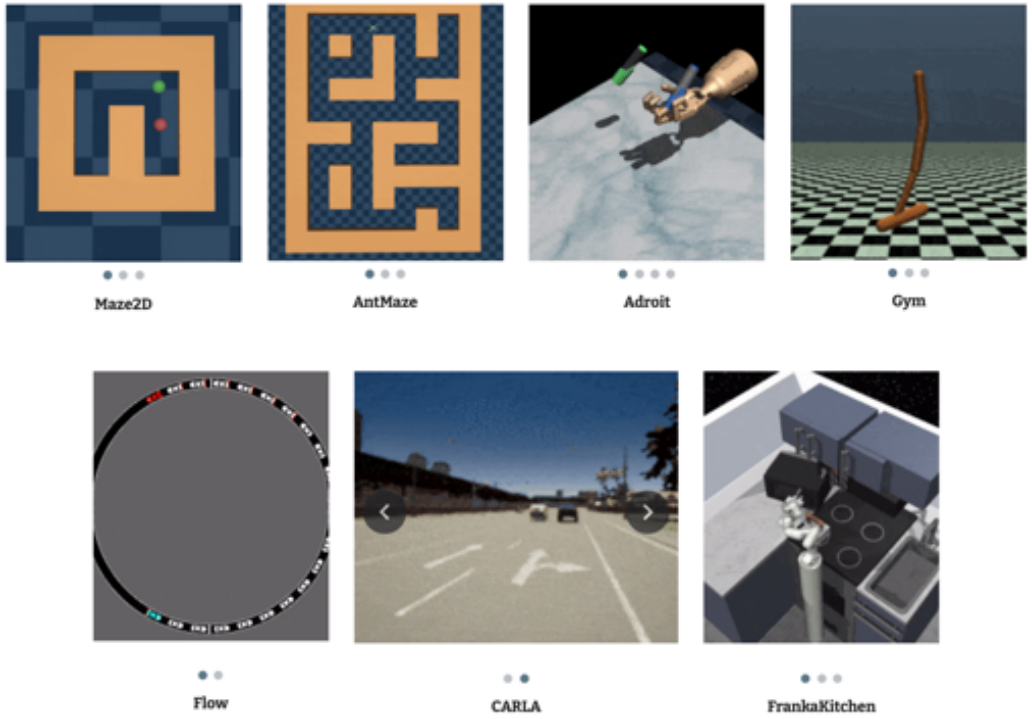
$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\frac{\pi_\theta(\tau)}{\pi_\beta(\tau)}\sum_{t=0}^{H}\gamma^t\nabla_\theta\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\hat{A}(\mathbf{s}_t,\mathbf{a}_t)\right]\\ &= \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\left(\prod_{t=0}^{H}\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\beta(\mathbf{a}_t|\mathbf{s}_t)}\right)\sum_{t=0}^{H}\gamma^t\nabla_\theta\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\hat{A}(\mathbf{s}_t,\mathbf{a}_t)\right]\\ &\approx \sum_{i=1}^{n} w_H^i\sum_{t=0}^{H}\gamma^t\nabla_\theta\log\pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)\hat{A}(\mathbf{s}_t^i,\mathbf{a}_t^i),\end{aligned}$$

A very recent survey on offline RL: https://arxiv.org/pdf/2005.01643.pdf

Offline Reinforcement Learning

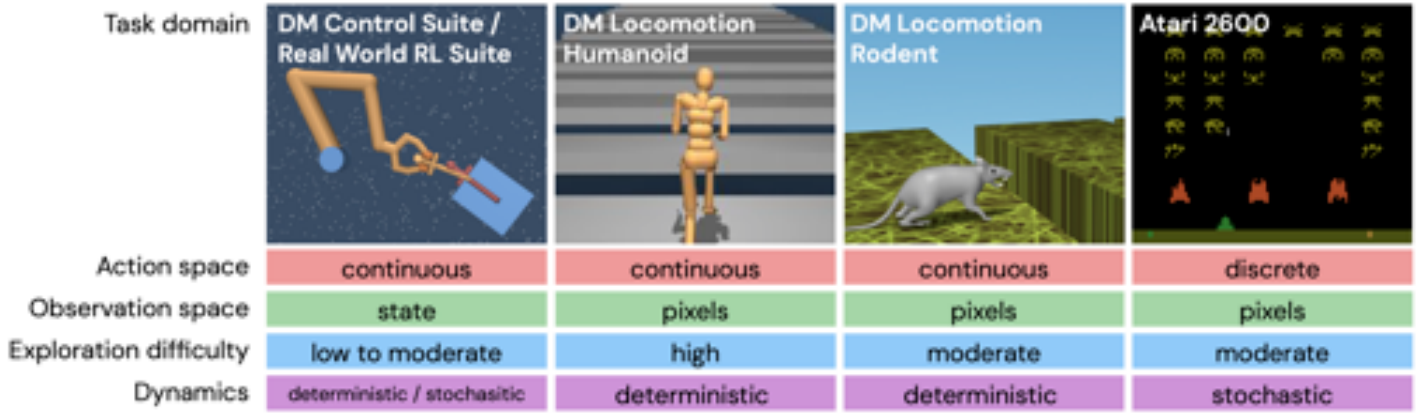
- Many offline RL datasets proposed in 2020

D4RL



https://sites.google.com/view/d4rl/home

RL unplugged



https://github.com/deepmind/deepmind-research/tree/master/rl_unplugged

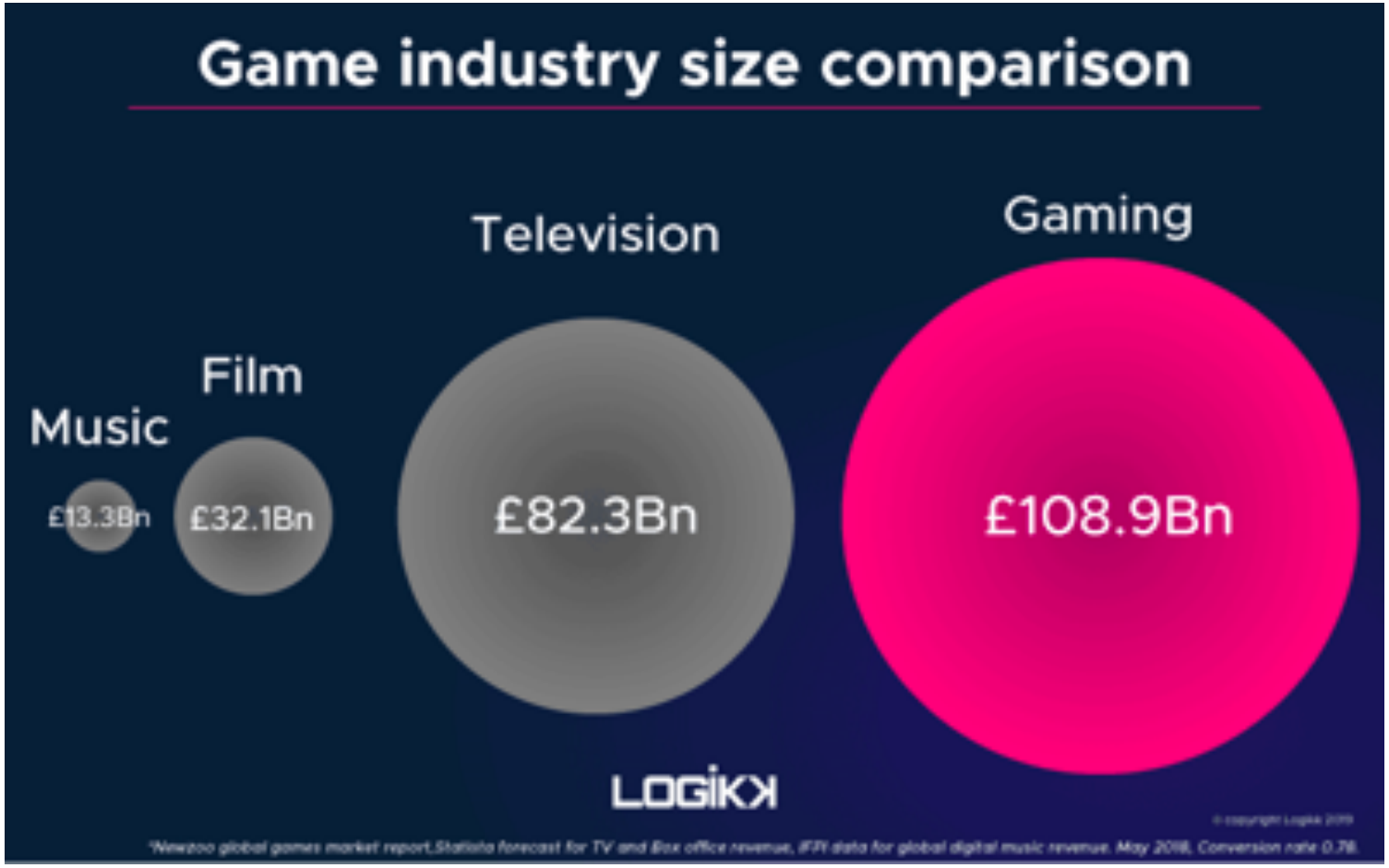
Difference between Imitation Learning and offline RL

- In imitation learning you are trying to match a behavior policy, thus achieving the same performance as the policy that generated the data.
- In offline/batch reinforcement learning you are not trying to imitate the data generating policy, instead you are trying to learn the best policy possible given the fixed dataset.

RL Applications

Only game playing??? How can I make a living???

Machine Learning in Gaming Industry



Machine Learning in Game Development

Algorithms Playing as NPCs
NPCs will respond to your actions in unique, unexpected ways.

Modelling Complex Systems
The game could predict and alter downstream effects.

Making Games more Beautiful
Textures and objects will render dynamically as you get closer

More Realistic Interactions
NLP will create more realistic conversations and responses.

Universe Creation on the Fly
Open world games have the potential to be unlimited in size.

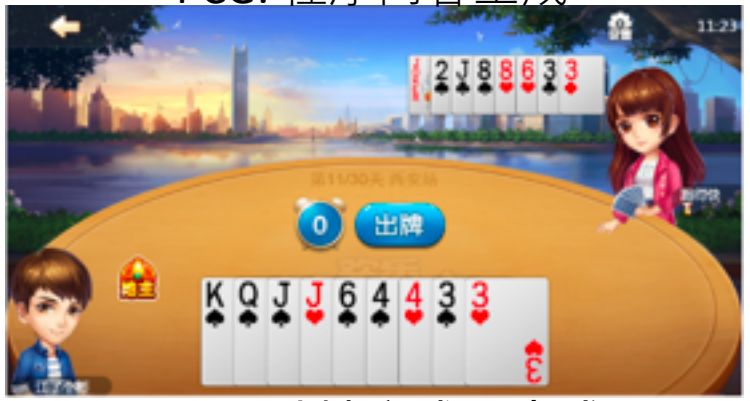
More Engaging Mobile Games
AI chips in phones will bring the power of ML to phones.

www.logikk.com LOGIKK

Training gaming AI bots

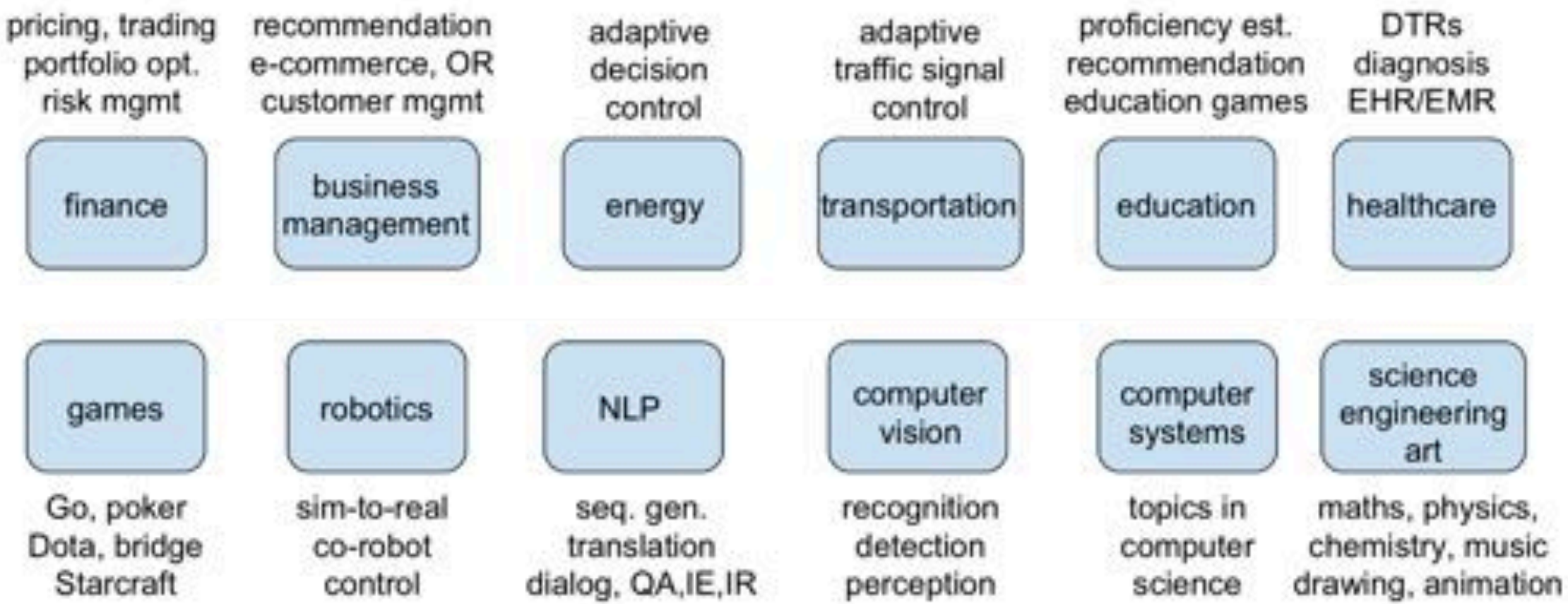
腾讯AI在QQ飞车手游的应用

PCG: 程序内容生成



QQ斗地主残局生成

RL Applications

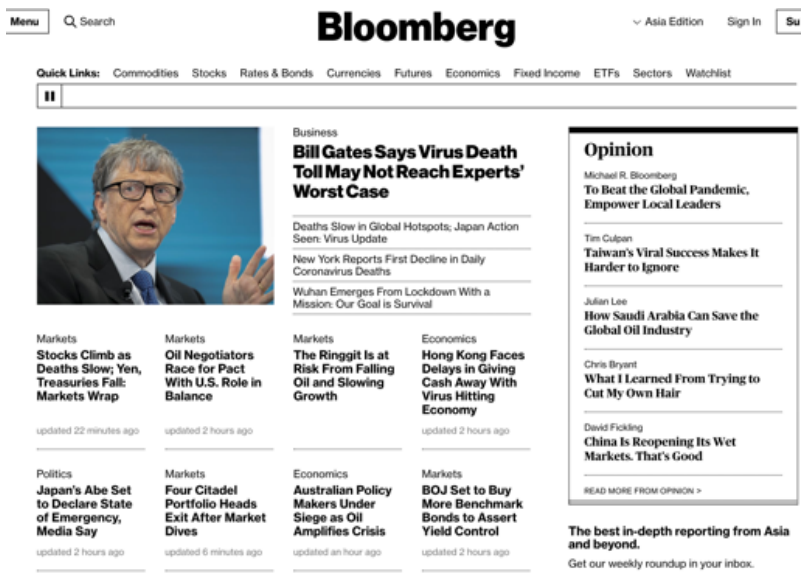


https://arxiv.org/pdf/1908.06973.pdf

Application to e-commence

Contextual Bandits

- In real-world, there is usually some context that help you make a decision
- For example:
  - Patient data for clinical trials
  - Consumer data for news/movie recommendation



Contextual Bandits

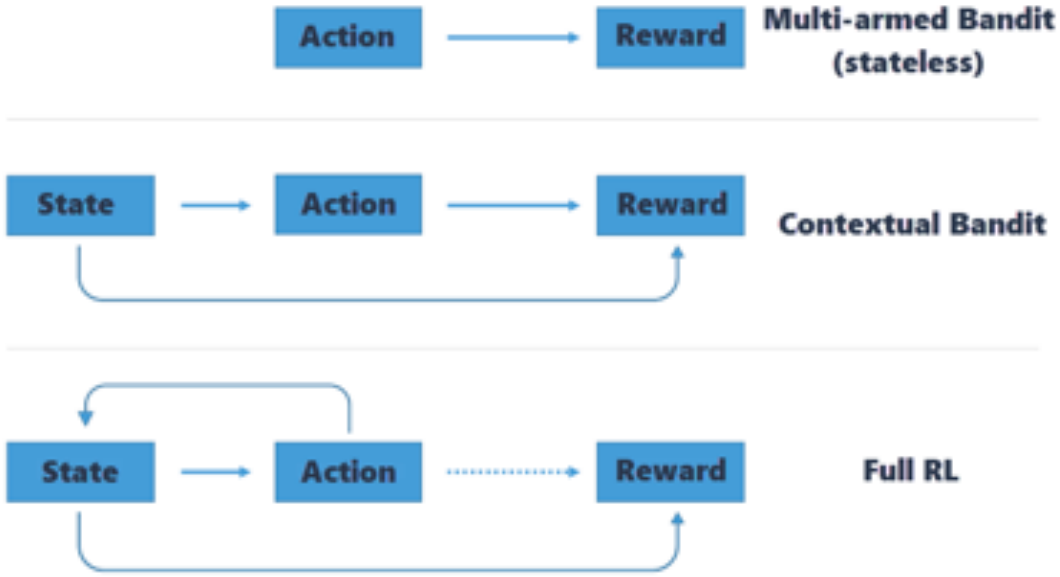
- We are running a sports news website. Today, there are K big sports related news stories.
- Every time a user visits our set, we must decide then and there which headlines to display to him/her on the front page.
- The goal is to maximize the number of clicks.

Contextual Bandits

Repeatedly:

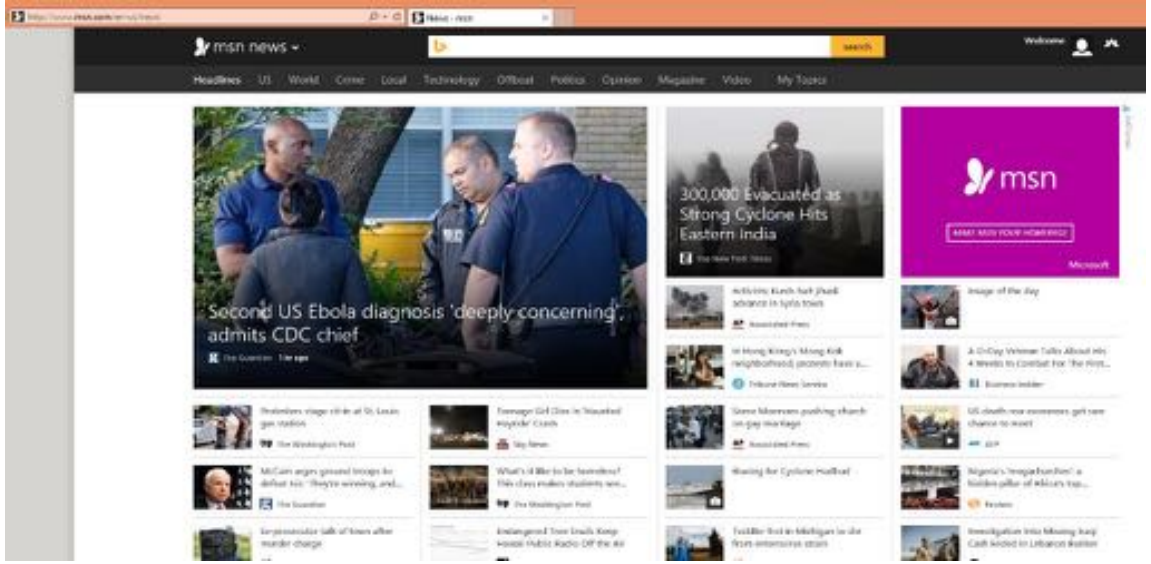
1. Observe features x
2. Choose action $a \in A$
3. Observe reward r

Goal: Maximize expected reward



Contextual bandit can be considered as one-step RL

News Recommendation



1. Use contextual bandit to learn best action for top slot
   - with a score-based policy, i.e. $\pi(x) = \underset{a}{\operatorname{argmax}}\, f(x, a)$
2. Use the ordering from f for actions in other slots

SIGAI Industry Award to Real World Reinforcement Learning Team from **Microsoft**

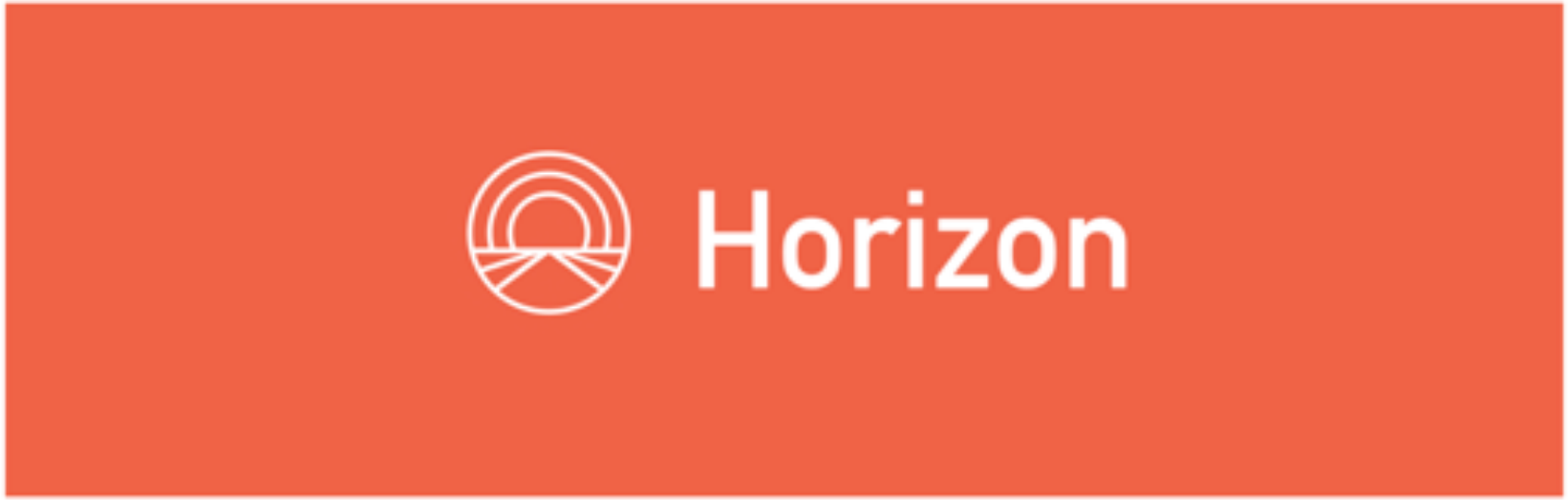
- Decision Service created by the Real World Reinforcement Learning Team from Microsoft, has been chosen as the winner of the inaugural 2019 award.
- Identification and development of cutting-edge research on contextual-bandit learning throughout the broad range of Microsoft products

https://www.microsoft.com/en-us/research/project/real-world-reinforcement-learning/

Facebook RL system in production

POSTED ON NOV 1, 2018 TO AI RESEARCH, ML APPLICATIONS

Horizon: The first open source reinforcement learning platform for large-scale products and services



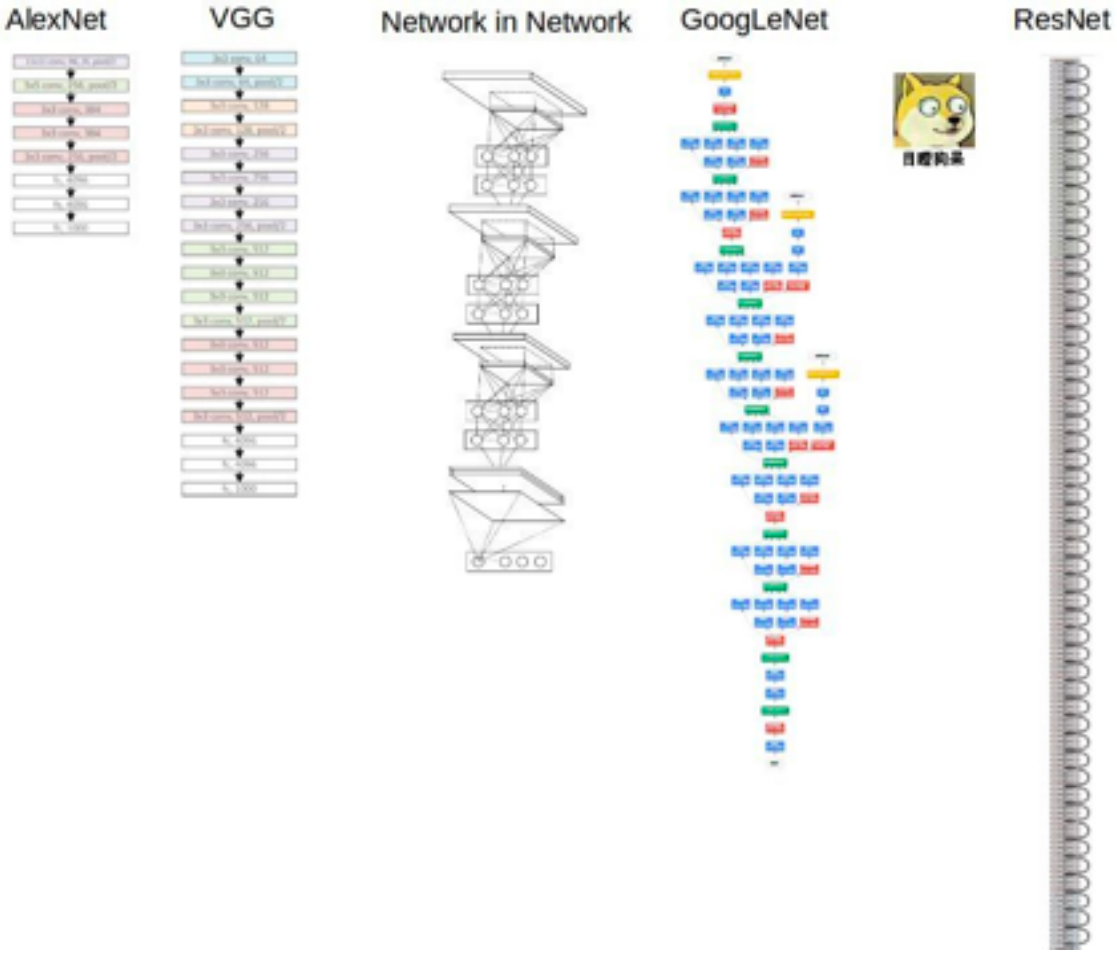
By Jason Gauci, Edoardo Conti, Kittipat Virochsiri

https://github.com/facebookresearch/ReAgent
https://research.fb.com/wp-content/uploads/2018/10/Horizon-Facebooks-Open-Source-Applied-Reinforcement-Learning-Platform.pdf?

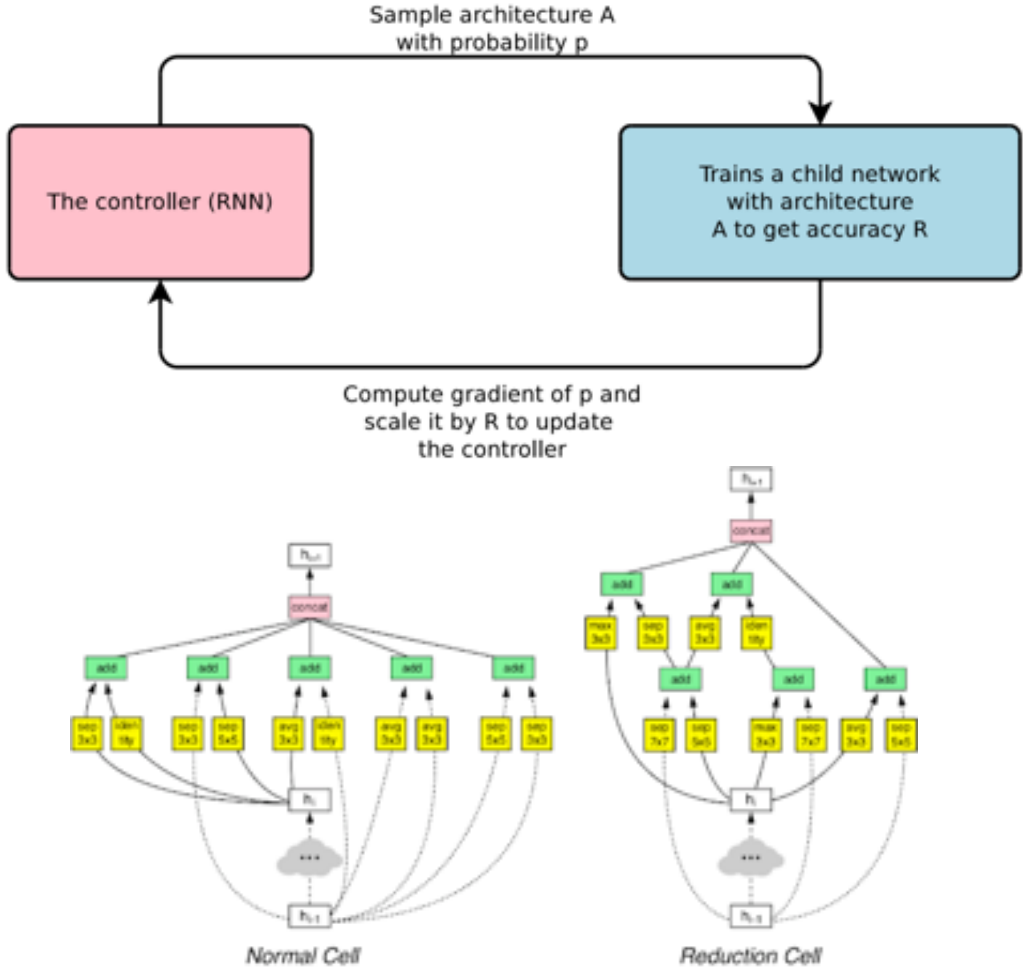
Application to Deep Learning

AutoML: Neural Architecture Search

Manually designed networks

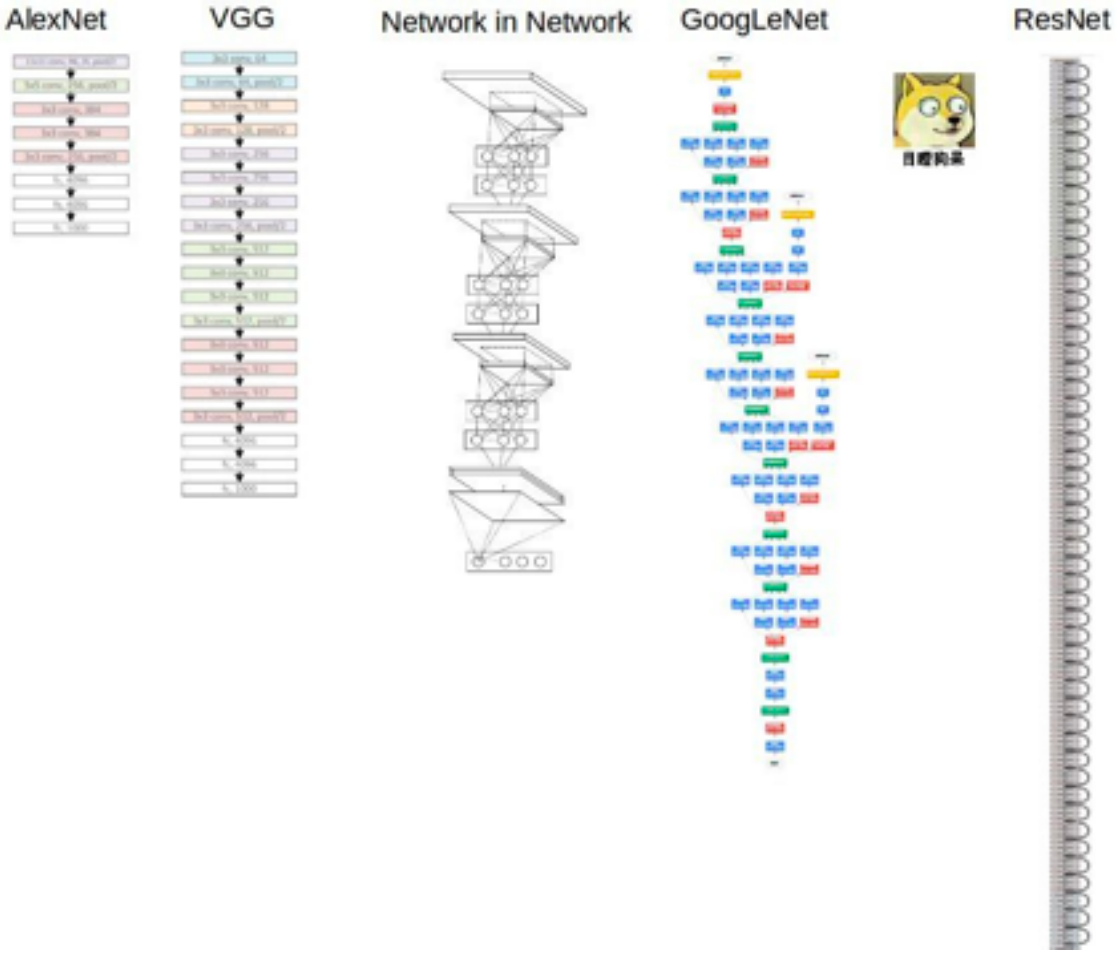


RL for network architecture search

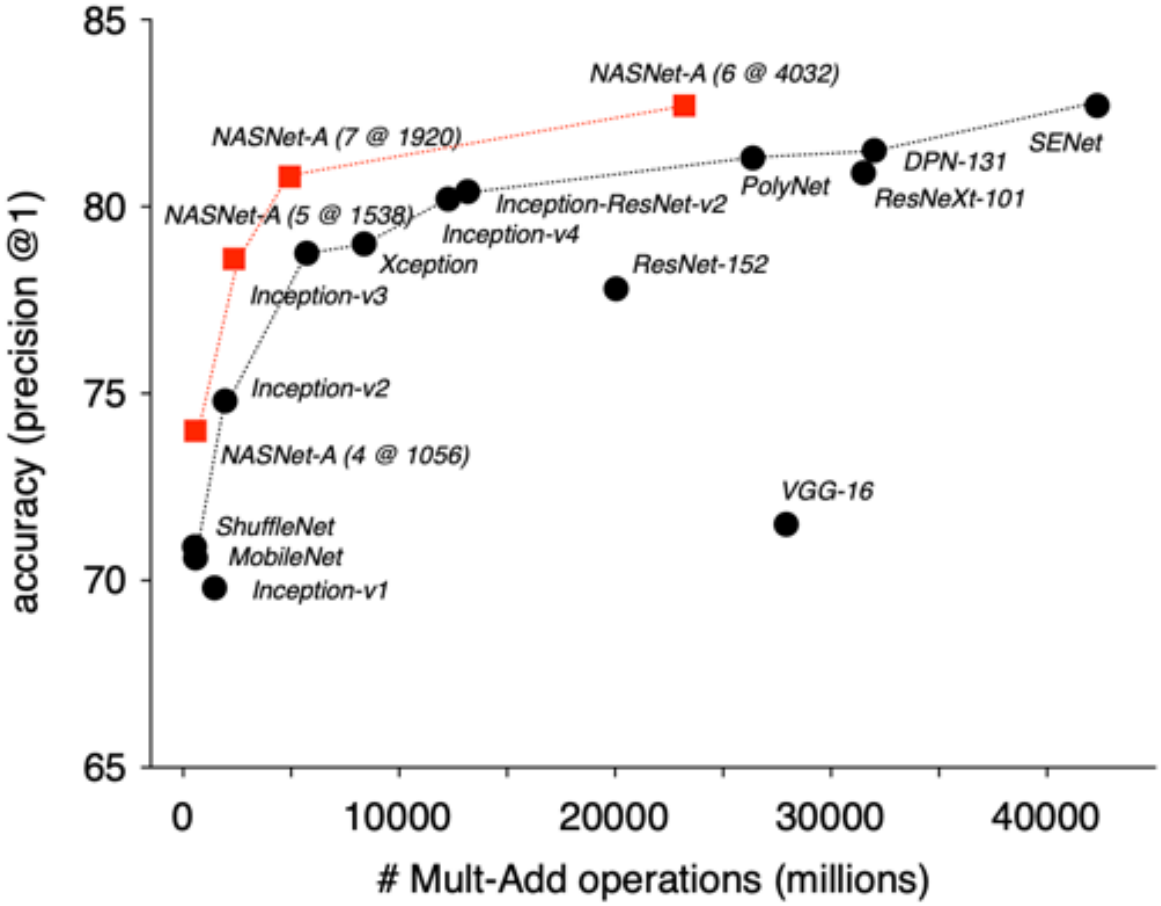


AutoML: Neural Architecture Search

Manually designed networks



RL for network architecture search



https://ai.googleblog.com/2017/11/automl-for-large-scale-image.html

AutoML

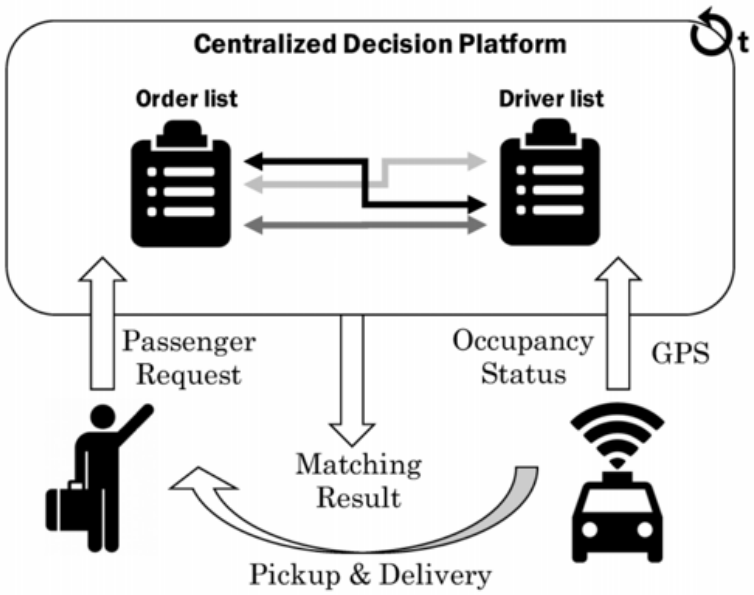
Winter is coming for some of the ML researchers/engineers

Jeff Dean's talk at ICML'19
https://slideslive.com/38917526/an-overview-of-googles-work-on-automl-and-future-directions

Application to Transportation

Large-scale Order Dispatch for Taxis

- NP-hard/combinatorial optimization



KDD 2018 paper from DiDi Research Institute

https://zhuanlan.zhihu.com/p/47193506 https://drive.google.com/file/d/17BoHSK-js0NPwOWJQzEFATINfbYj4OKR/view

Large-scale Order Dispatch for Taxis

Offline learning + Online planning

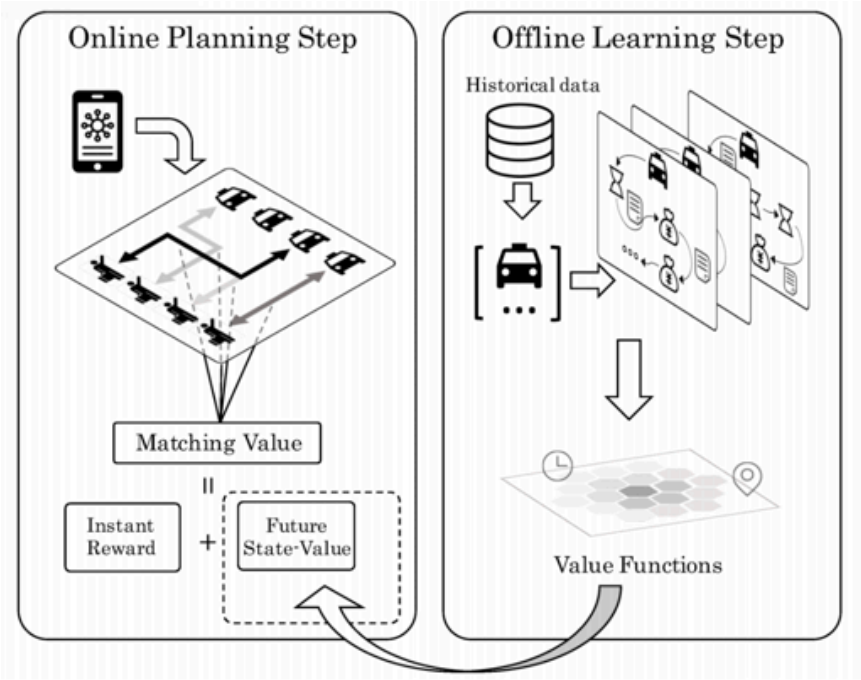


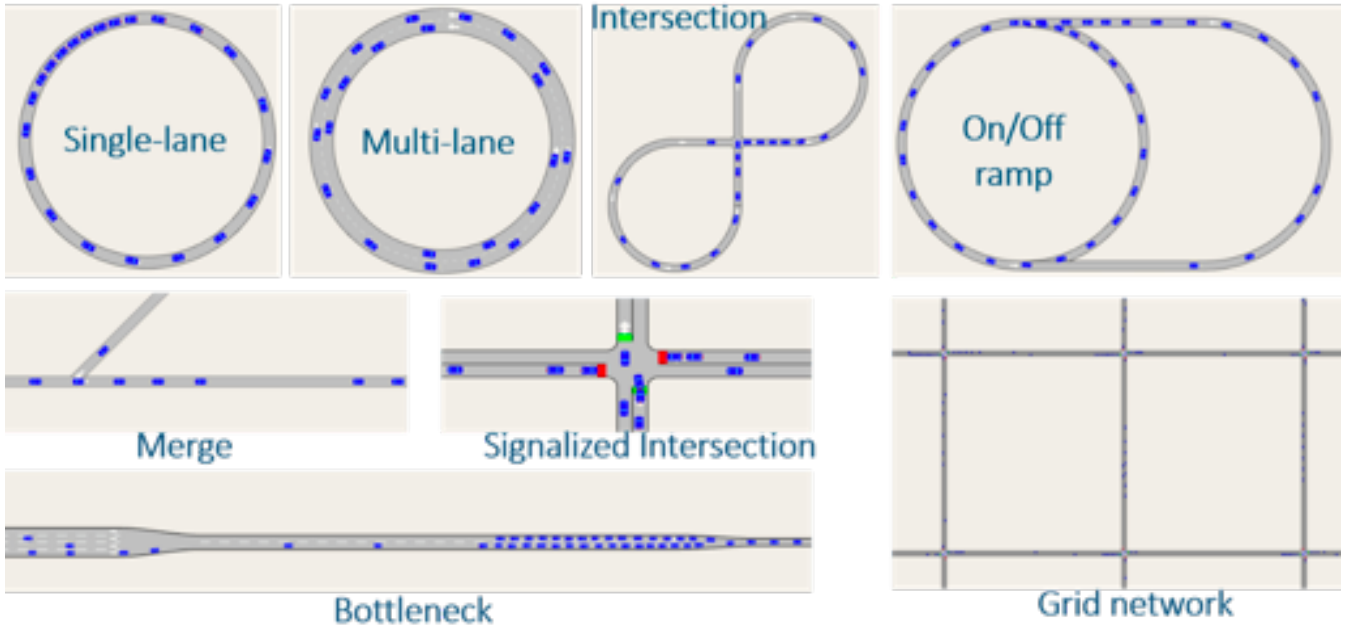
Figure 1: Illustration of the proposed algorithm.

Take-home message:

- How to model real-world tasks into RL framework
- How to simplify the problem so that it is solvable

https://drive.google.com/file/d/17BoHSK-js0NPwOWJQzEFATINfbYj4OKR/view

Multi-agent system for traffic simulation



https://flow-project.github.io/index.html

Multi-agent system for traffic simulation

https://flow-project.github.io/index.html

Other RL Applications

Application to Robot Learning

Dexterous Manipulation from OpenAI

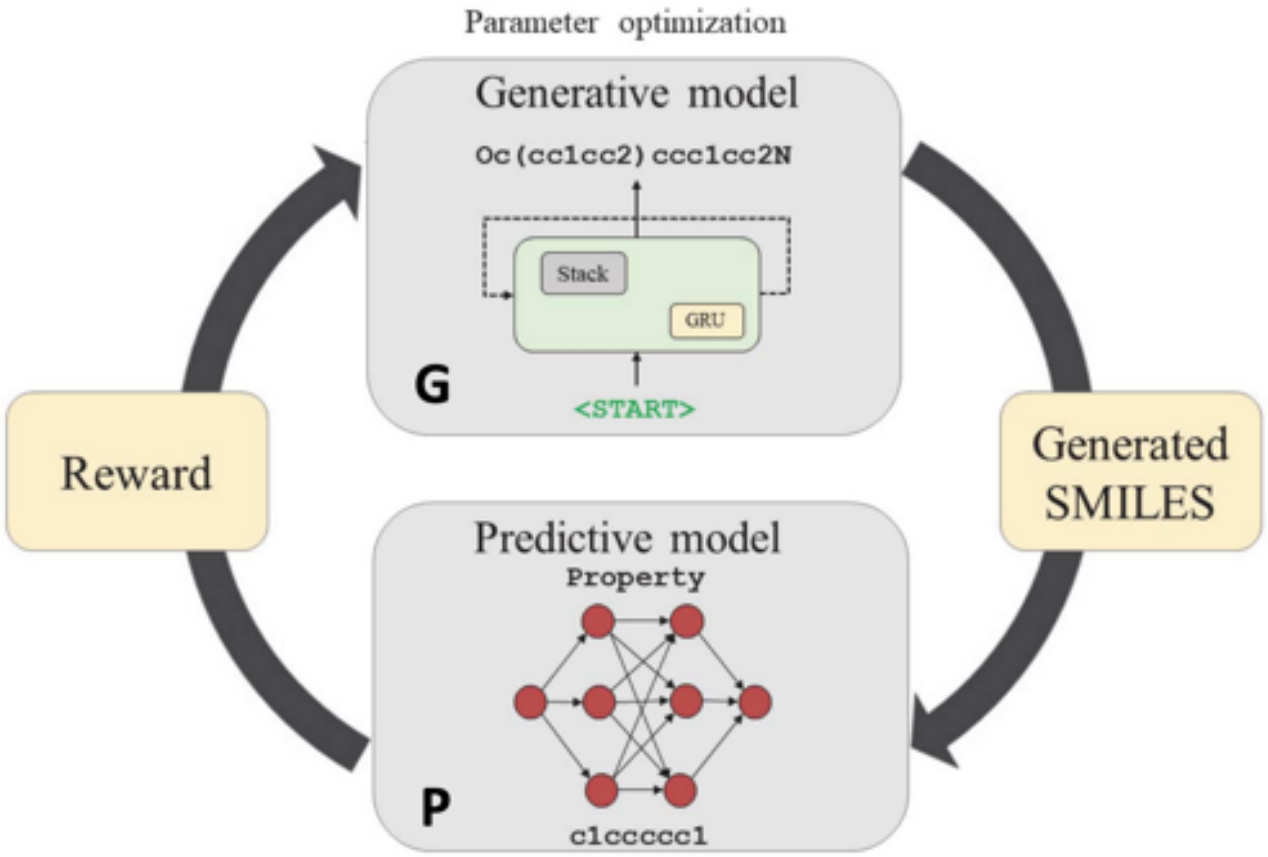
https://openai.com/blog/learning-dexterity/

CoRL (new annual conference on robot learnings ince 2017)

https://sites.google.com/robot-learning.org/corl2019

CoRL 2019

Application to Drug Design



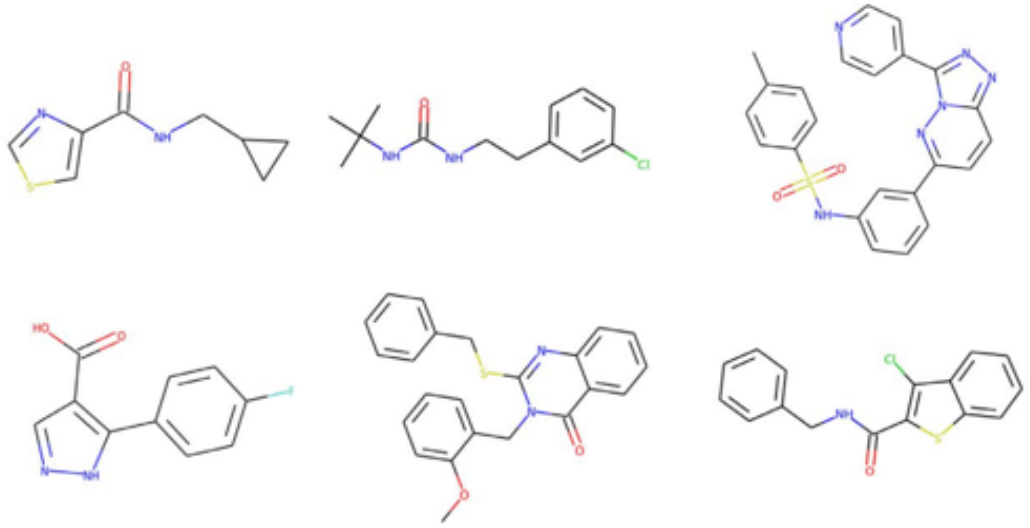


Fig. 2. A sample of molecules produced by the generative model.

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Science Advances, 4(7).

Application to Drug Design

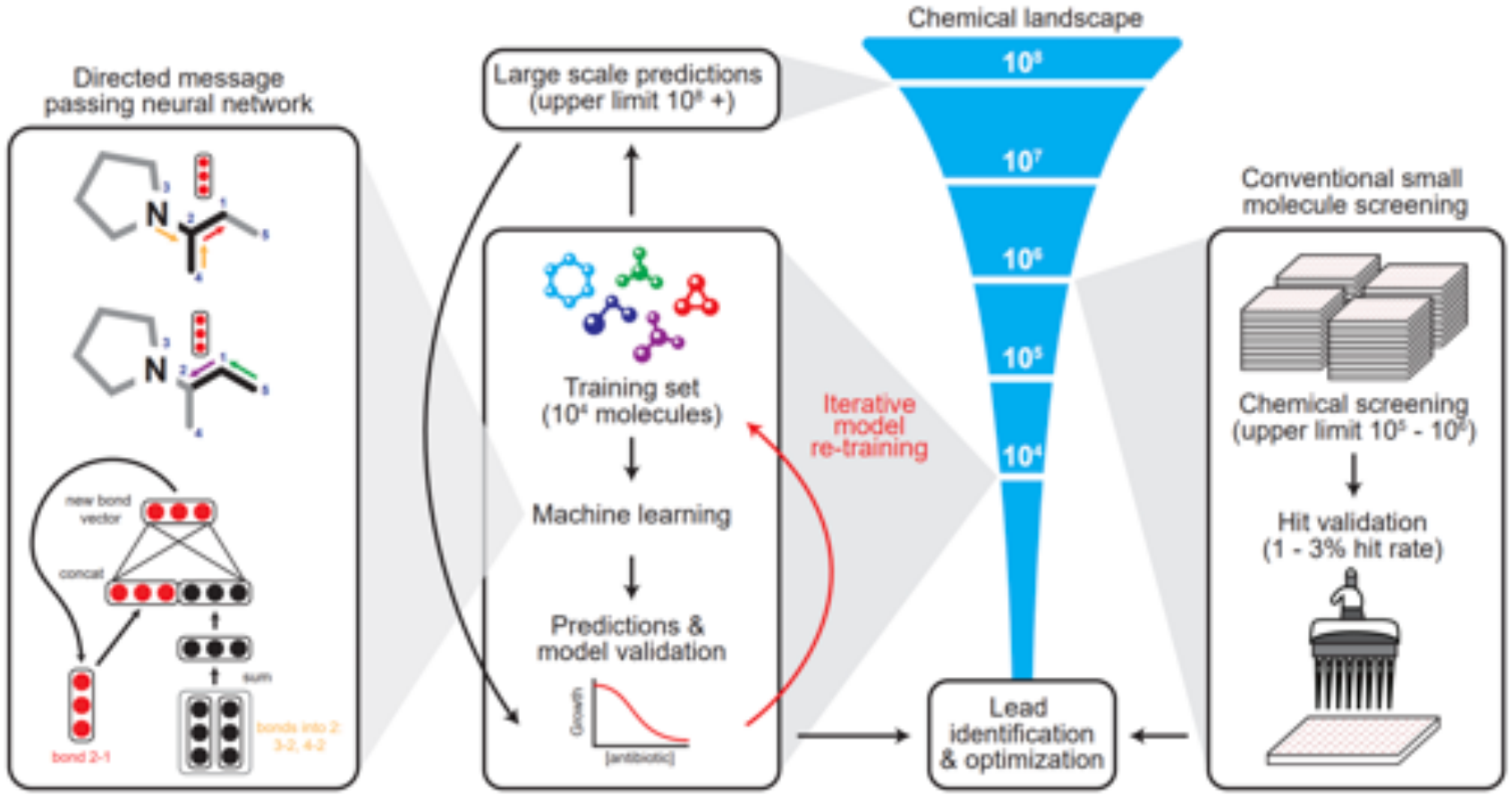


Figure 1. Machine Learning in Antibiotic Discovery

https://www.cell.com/cell/pdf/S0092-8674(20)30102-1.pdf

A Deep Learning Approach to Antibiotic Discovery. Stokes, et al. Cell 2020.

Application to Finance

- TensorTrade is an open-source Python framework for building, training, evaluating, and deploying robust trading algorithms using reinforcement learning
- https://towardsdatascience.com/trade-smarter-w-reinforcement-learning-a5e91163f315
- https://github.com/tensortrade-org/tensortrade

Other Resources on Real-World RL

- RL for Real Life ICML'19 Workshop: https://sites.google.com/view/RL4RealLife
- Recent survey on RL application:
  - https://arxiv.org/pdf/1908.06973.pdf
  - https://medium.com/@yuxili/rl-applications-73ef685c07eb