DRN: A Deep Reinforcement Learning Framework for News Recommendation

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang,

Nicholas Jing Yuan, Xing Xie, Zhenhui (Jessie) Li

PennState

Microsoft

Introduction: Why reinforcement recommendation

First round

Second round

Equal rewarding recommendation for current round

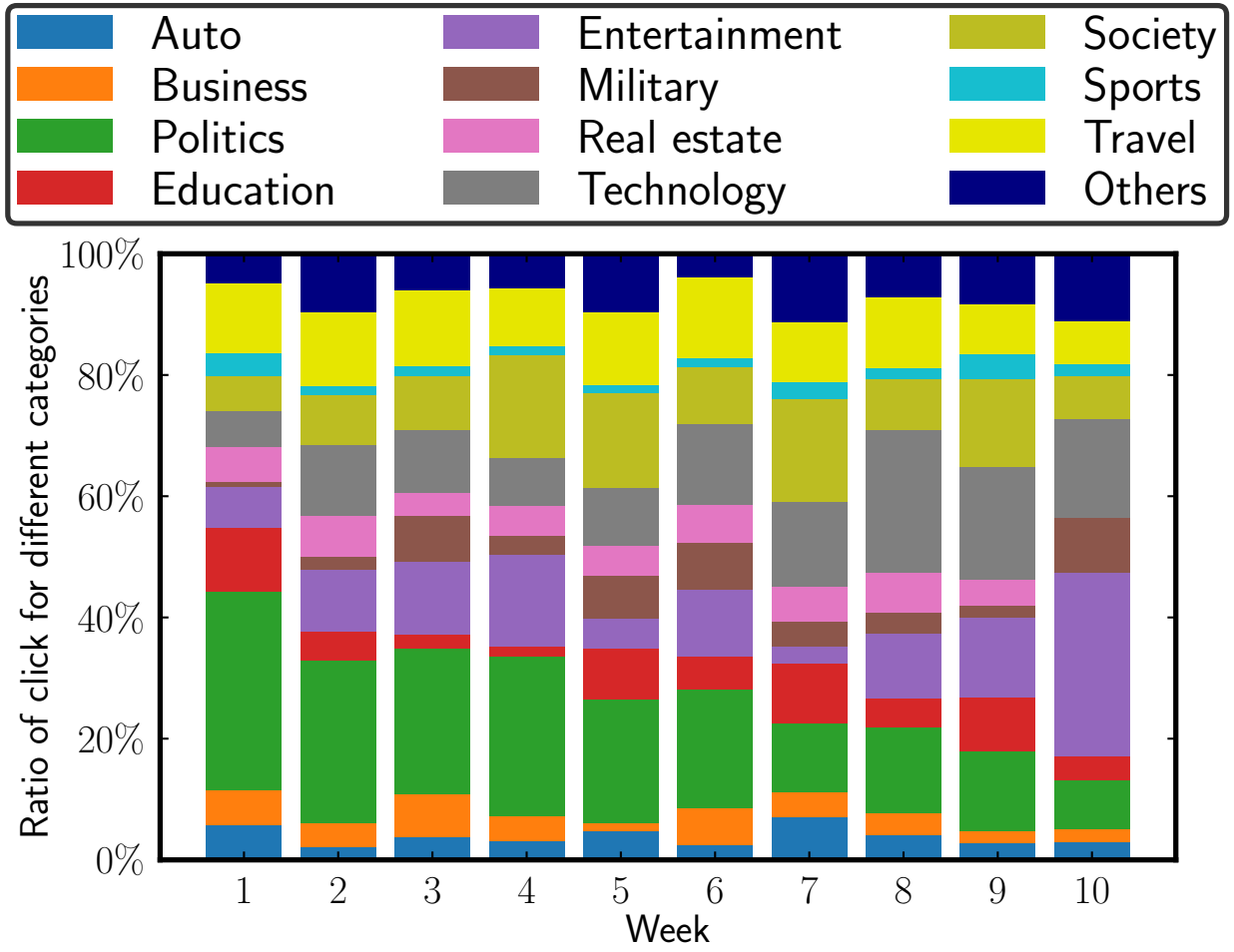
Images from:

1. https://www.ohio.com/akron/sports/cavs/cavaliers-j-r-smith-expects-lebron-james-to-keep-opening-night-streak-alive-despite-sprained-ankle
2. https://www.cnbc.com/2018/03/05/kobe-bryant-has-won-an-oscar-heres-what-he-says-it-takes-to-succeed.html
3. https://twitter.com/NBA
4. https://www.stgeorgeutah.com/news/archive/2016/06/11/djg-alert-significant-weather-strong-thunderstorm-rolls-through-washington-county/#.WsqHUdPwb6Y
5. http://www.clipartpanda.com/clipart_images/question-mark-36565633

Introduction: News recommendation is dynamic

The life period for news is usually very short.

User's interest may change during time.

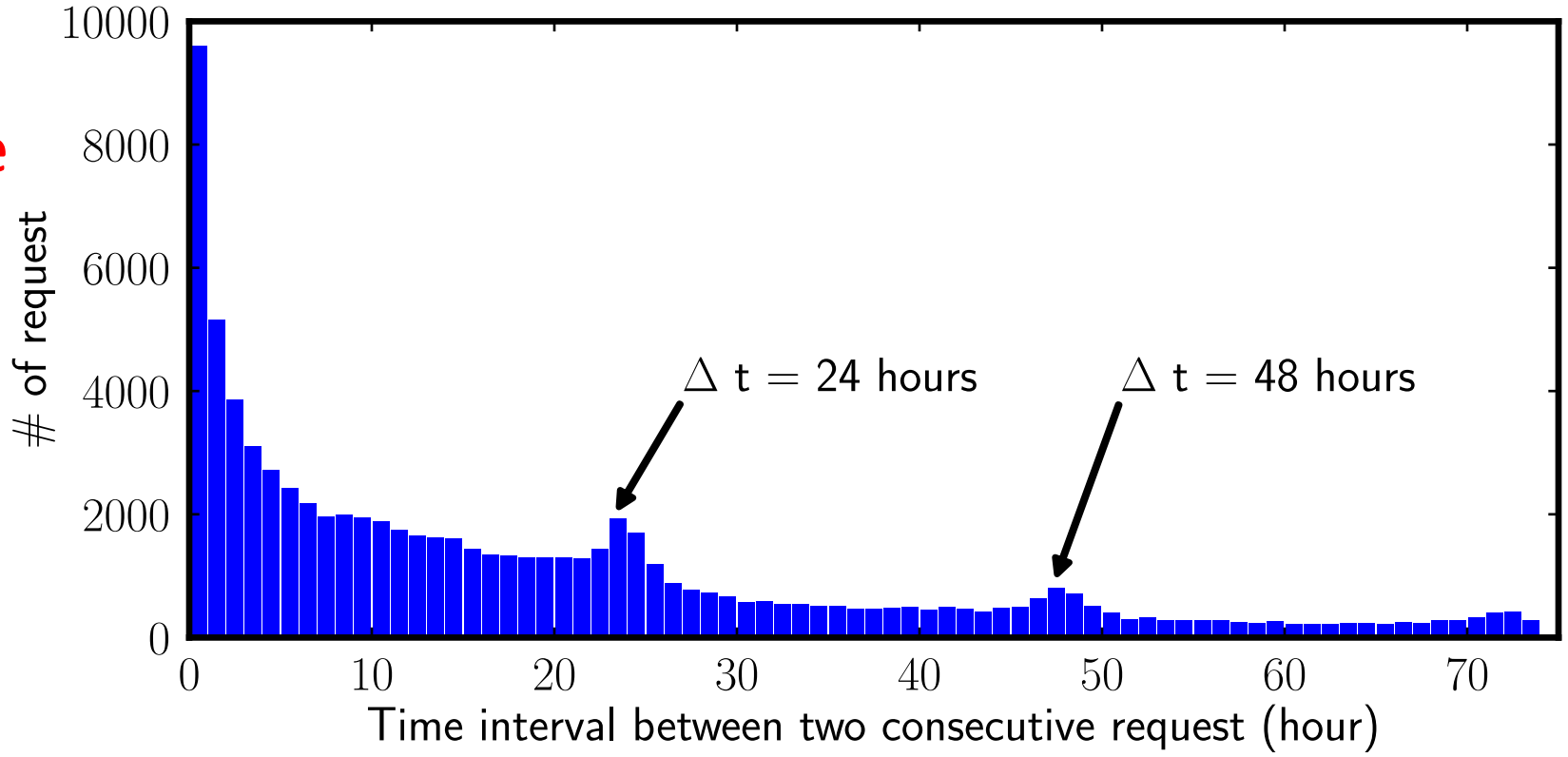


Introduction: Is there more than click/no-click?

User's clicks on news are usually very dense in a short period.

Then, user usually leave the app!

User may return everyday!



Introduction: Should we keep recommending similar items?

Lebron James will be the MVP!

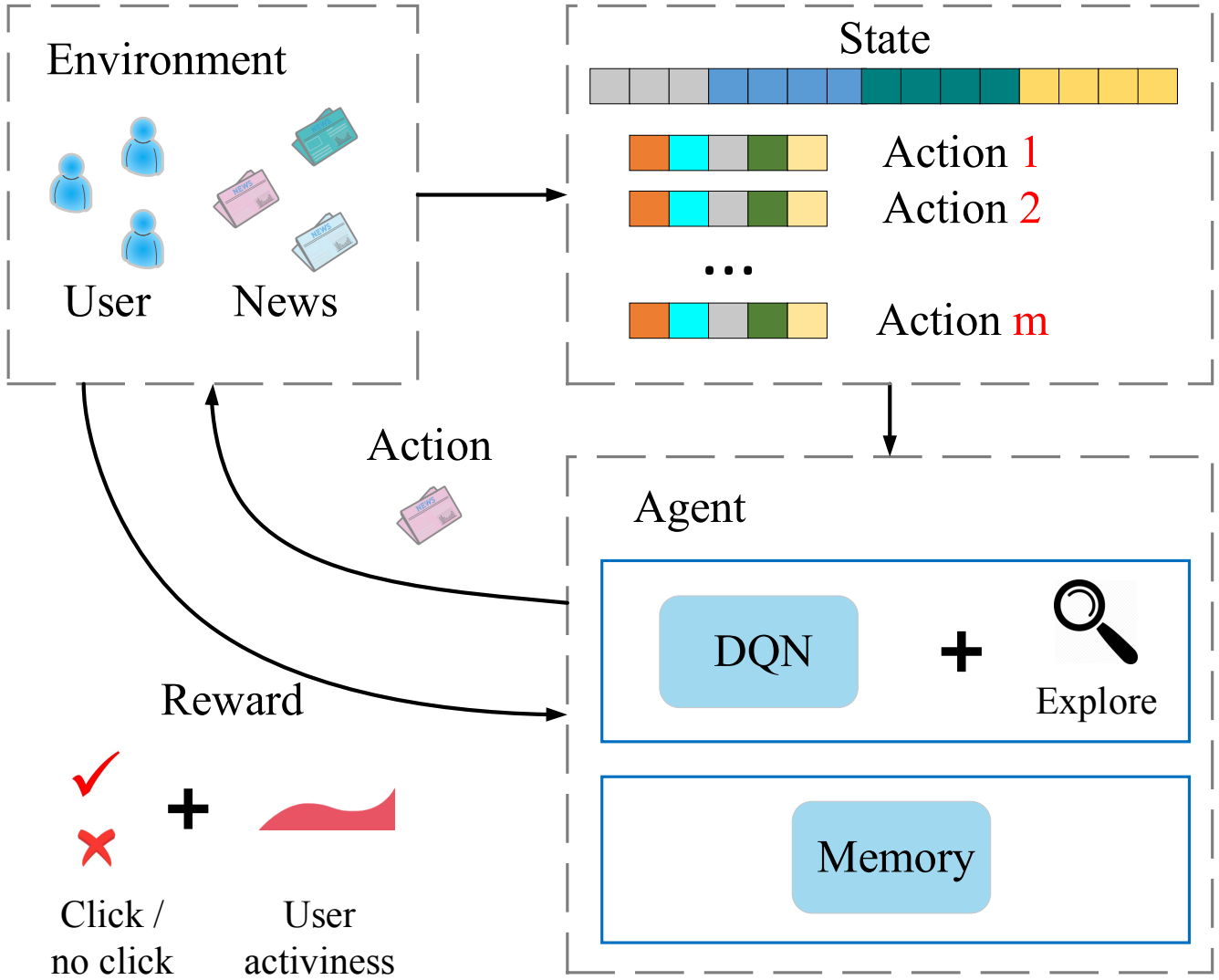
Tony Parker has come back from injury!

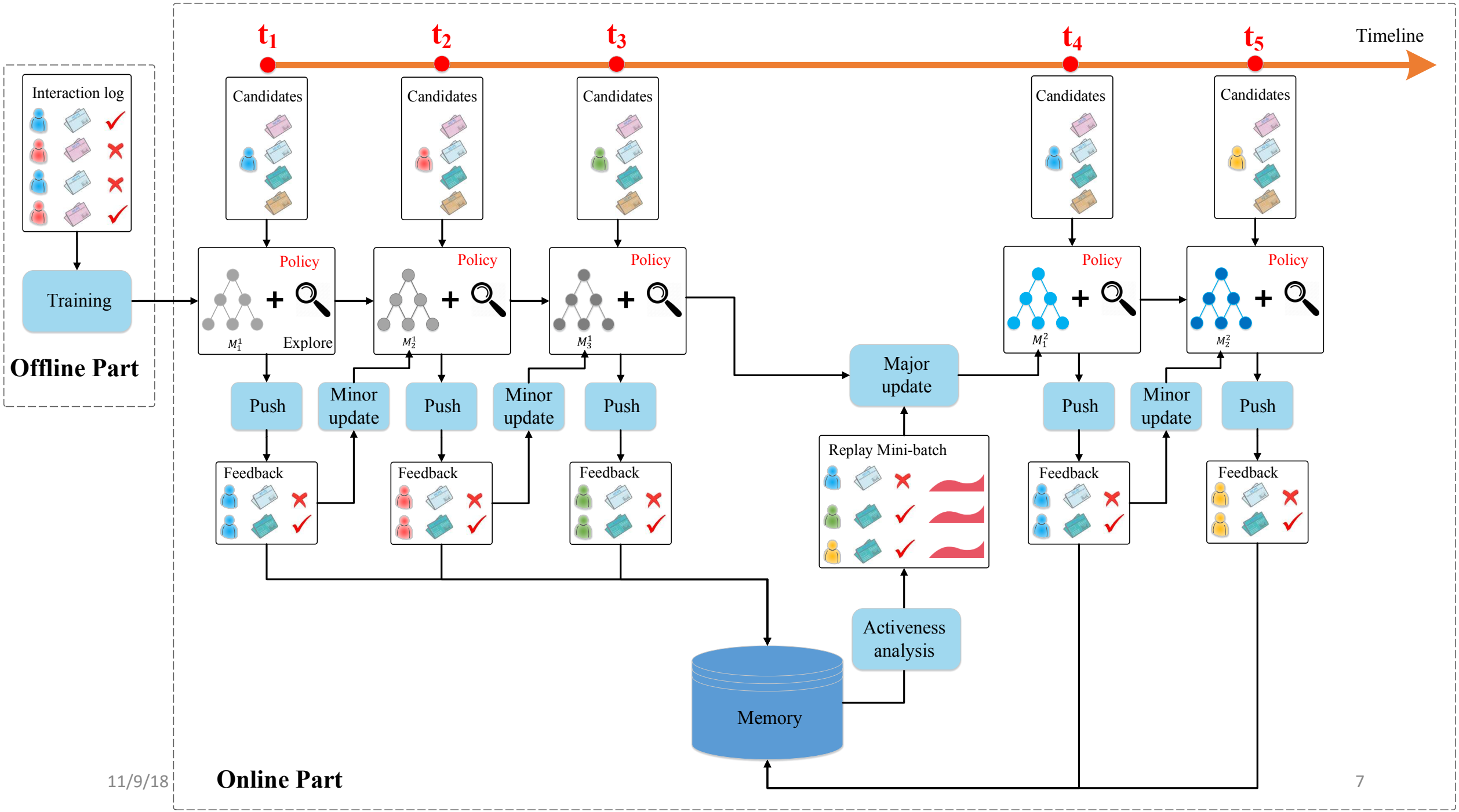
Paul Gasol promises to help the Spurs in the playoff.

Will you **get bored** if all the recommended news are from NBA when you are browsing the sports news?

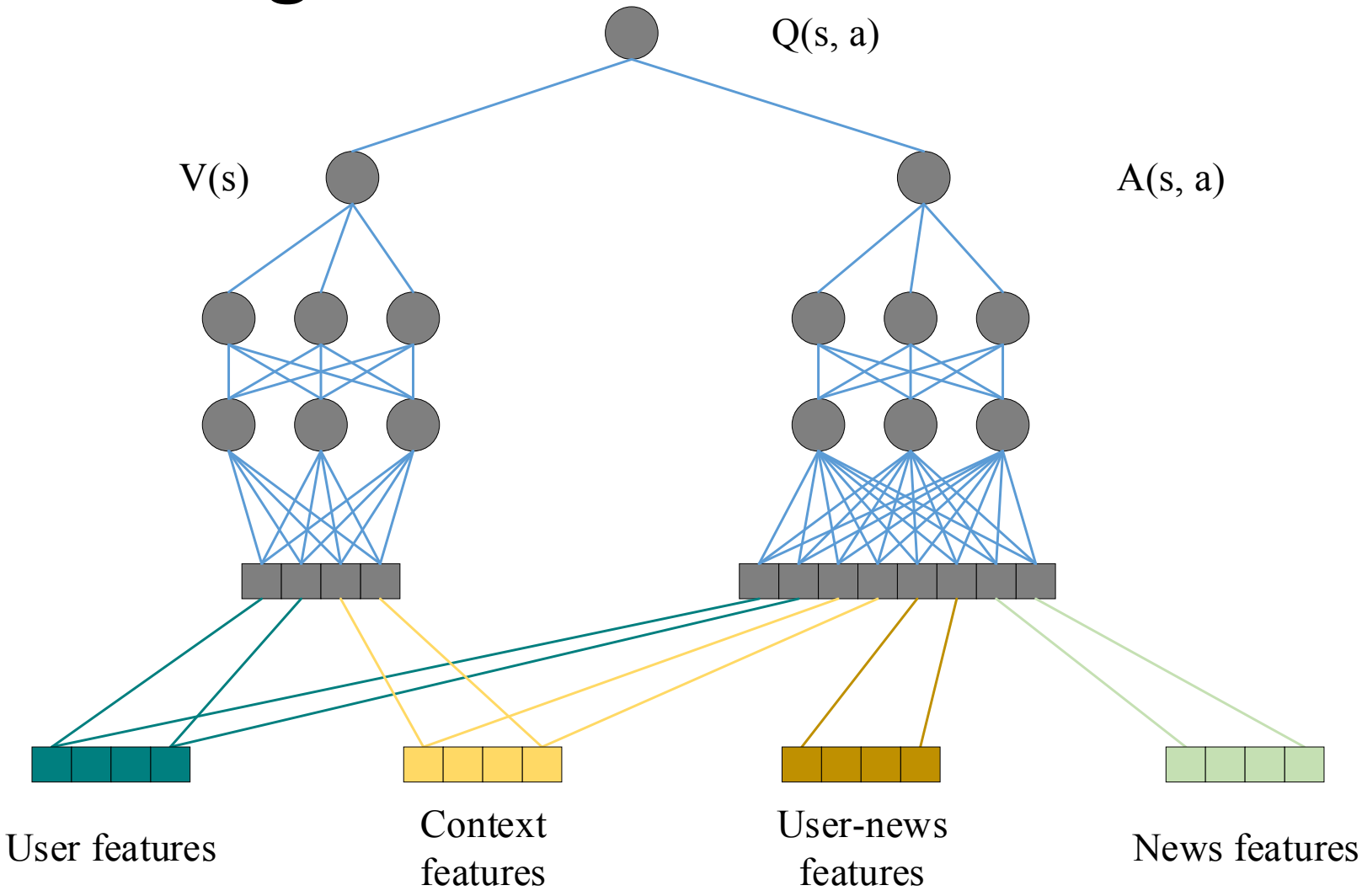
Images from: https://twitter.com/nba

Method: Using reinforcement learning to do recommendation




Interaction log
Training
Offline Part
t1
t2
t3
t4
t5
Timeline
Candidates
Policy
Explore
Push
Minor update
Feedback
Major update
Replay Mini-batch
Activeness analysis
Memory
Online Part

Method: Dueling network structure – value and advantage function



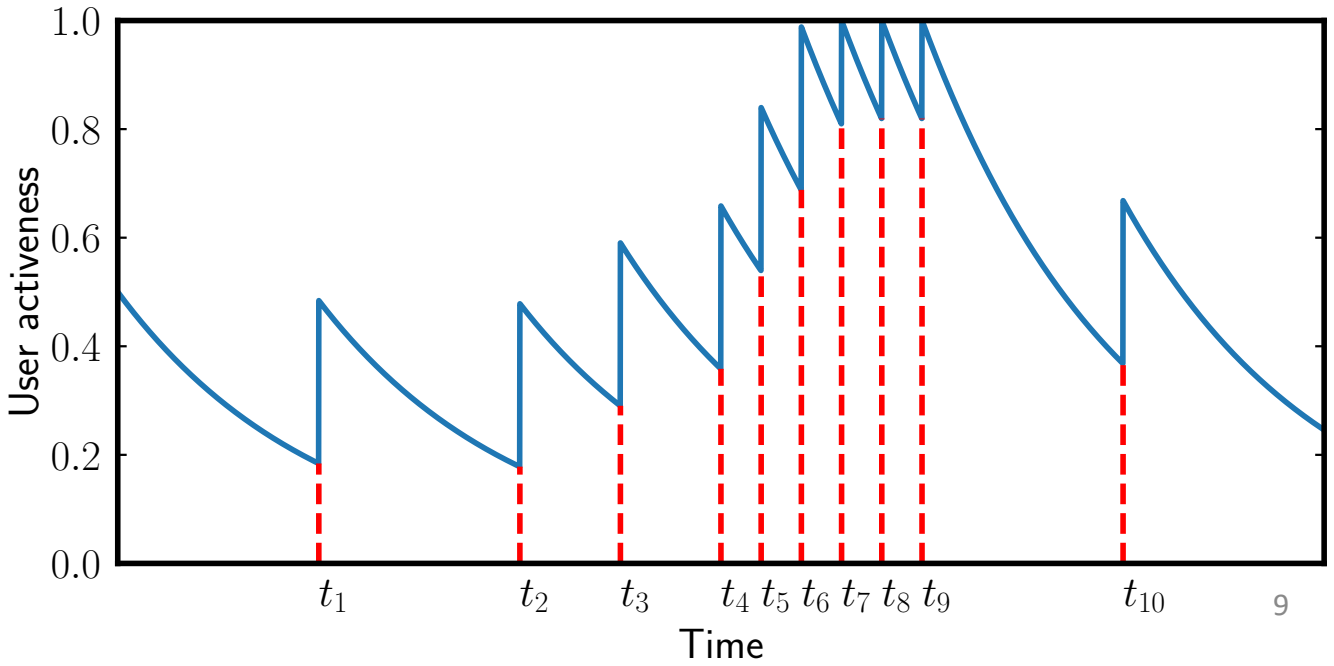
Method: user activeness modeling -- survival analysis

User activeness decay function

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr\{t \leq T < t + dt | T \geq t\}}{dt}$$

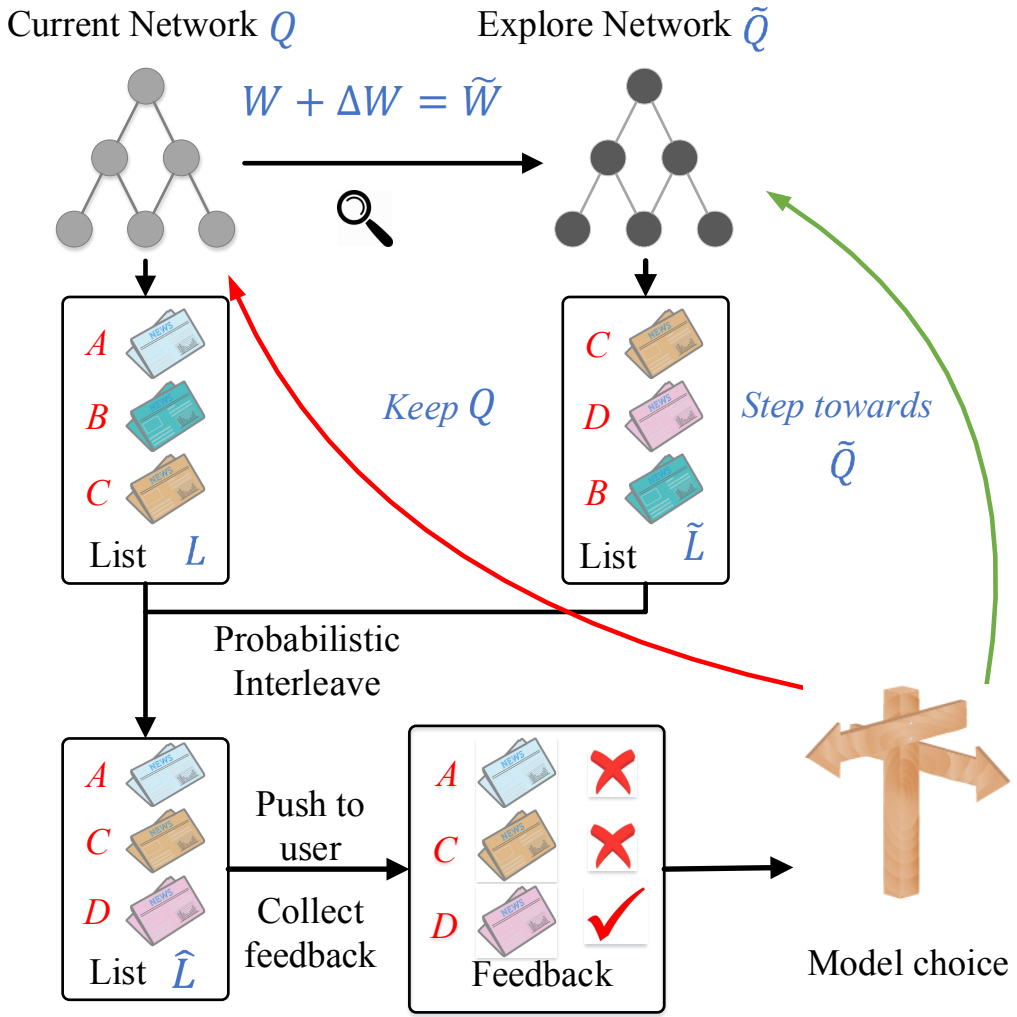
User activeness

$$S(t) = e^{-\int_0^t \lambda(x)dx}$$



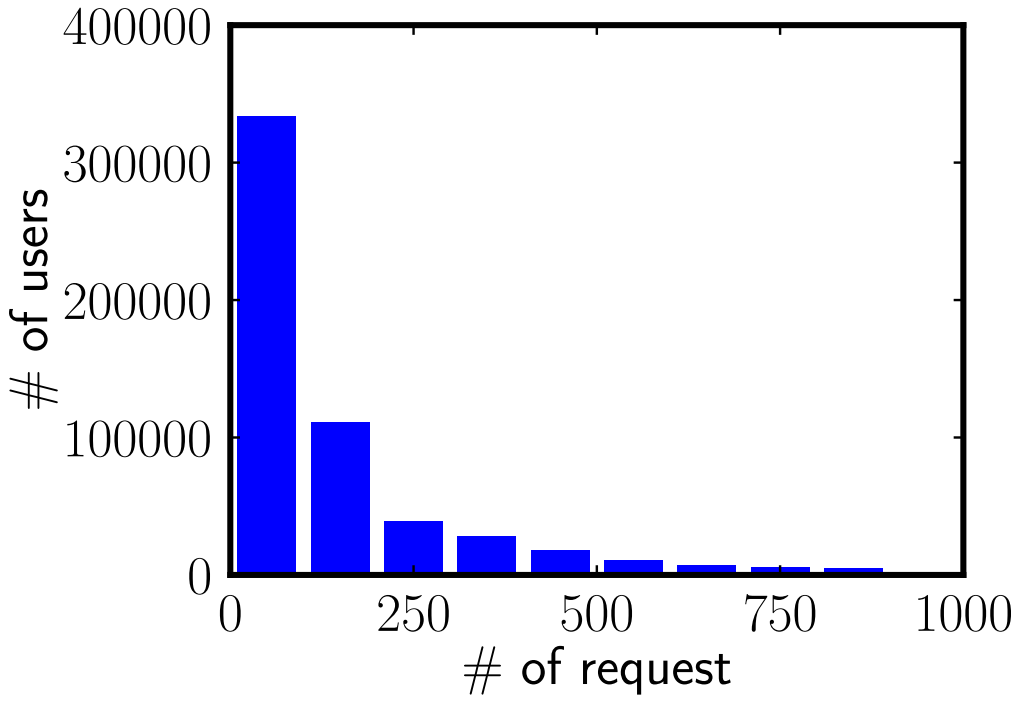
Method: Effective exploration

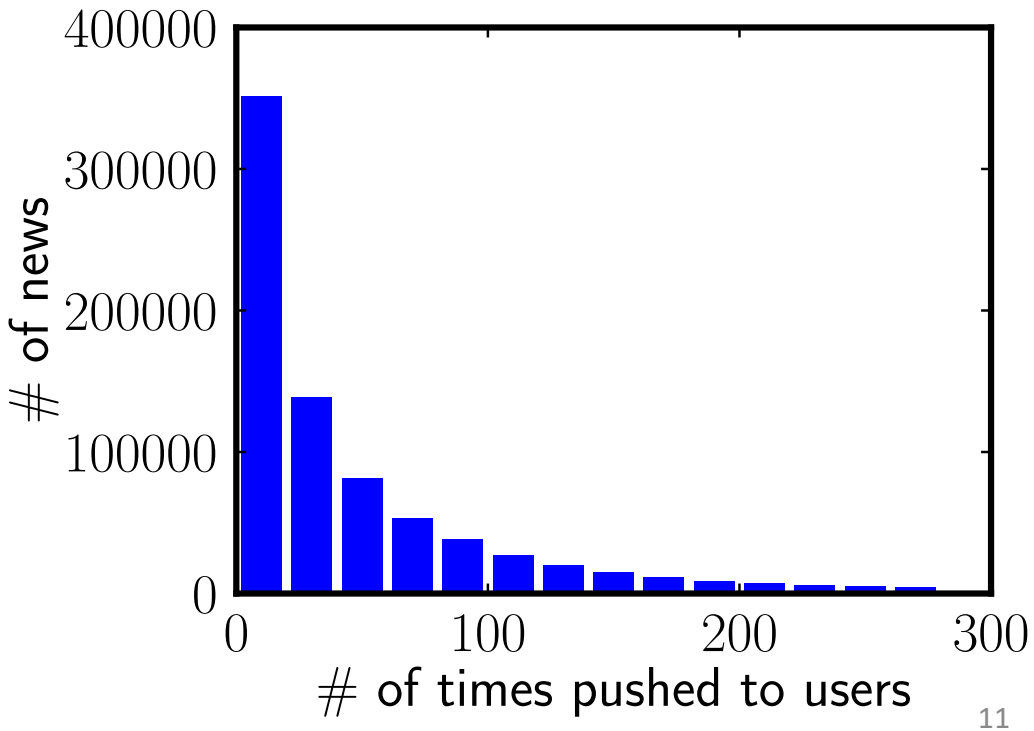
- Step1: get recommendation from Q and $\tilde{Q}$
- Step2: probabilistic interleave these two lists
- Step3: get feedback from user and compare the performance of two network
- Step4: if $\tilde{Q}$ performs better, update Q towards it



Dataset

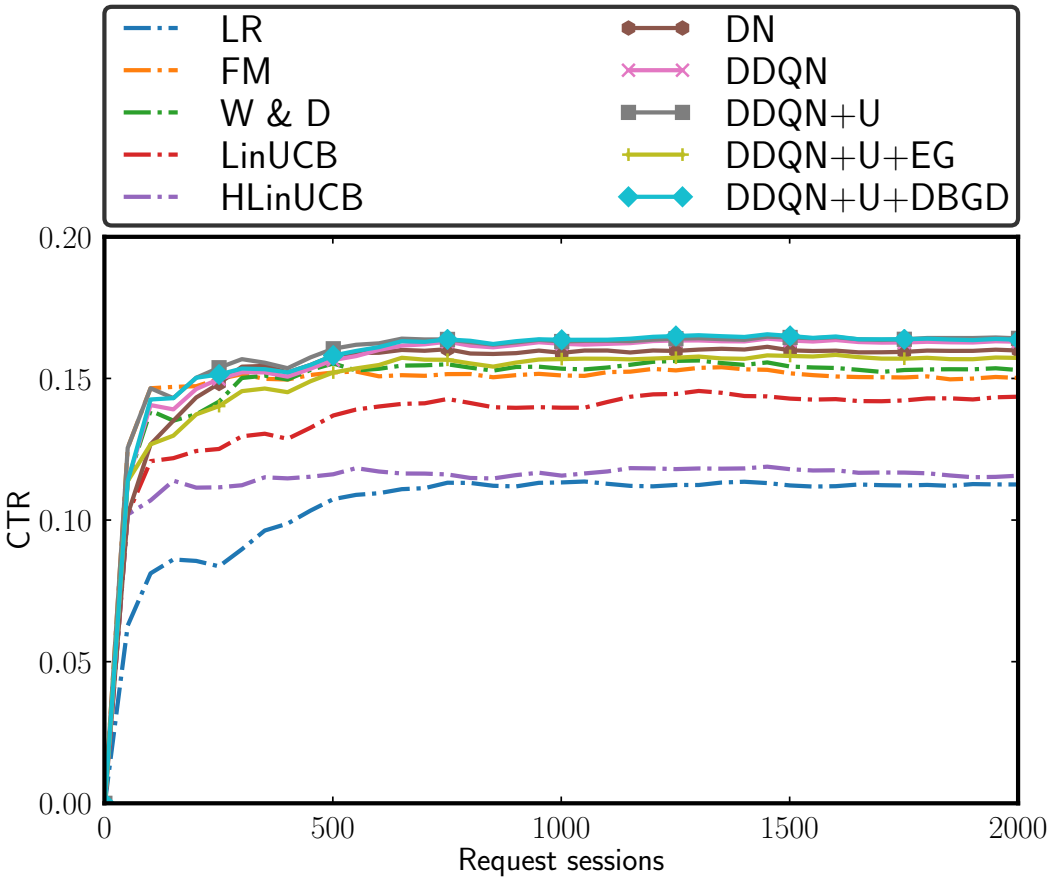
Stage	Duration	# of users	# of news
Offline stage	6 months	541,337	1,355,344
Online stage	1 month	64,610	157,088





Results: Offline

Method	CTR	nDCG
LR	0.1262	0.3659
FM	0.1489	0.4338
W&D	0.1554	0.4534
LinUCB	0.1447	0.4173
HLinUCB	0.1194	0.3491
DN	0.1587	0.4671
DDQN	**0.1662**	**0.4877**
DDQN + U	**0.1662**	**0.4878**
DDQN + U + EG	0.1609	0.4723
DDQN + U + DBGD	**0.1663**	**0.4854**



Accuracy

Results: Online

Method	CTR	Precision@5	nDCG
LR	0.0059	0.0082	0.0326
FM	0.0072	0.0078	0.0353
W&D	0.0052	0.0067	0.0258
LinUCB	0.0075	0.0091	0.0383
HLinUCB	0.0085	0.0128	0.0449
DN	0.0100	0.0135	0.0474
DDQN	0.0111	0.0139	0.0477
DDQN + U	0.0089	0.0110	0.0425
DDQN + U + EG	0.0083	0.0100	0.03391
DDQN + U + DBGD	**0.0113**	**0.0149**	**0.0492**

Accuracy

Method	ILS
LR	0.1833
FM	0.2014
W&D	0.1647
LinUCB	0.2636
HLinUCB	0.1323
DN	0.1546
DDQN	0.1935
DDQN + U	0.1713
DDQN + U + EG	0.1907
DDQN + U + DBGD	**0.1216**

Diversity

Summary of motivation and solution

Motivation	Solution
Long term effect in recommendation	**Deep reinforcement learning (DRL)**
• Dynamic nature of news recommendation • Consider more measures for long term effect • Recommendation diversity	• Online learning feature of DRL • Reward function design of DRL • Explore in DRL

Conclusion

- We propose a **reinforcement learning framework** to do **online personalized news recommendation**, taking care of both immediate and future reward. Our framework can be generalized to many other recommendation problems.
- We consider **user activeness** to help improve recommendation accuracy, which can provide extra information than simply using user click labels.
- Our system has been **deployed online in a commercial news recommendation application**. Extensive offline and online experiments have shown the superior performance of our methods.