Reward innovation for long-term member satisfaction

Gary Tang, Jiangwei Pan, Henry Wang and Justin Basilico

NETFLIX

Goal

Create a **personalized homepage** to help members find content to watch and enjoy

that

maximizes **long-term** member **satisfaction**.

Long-term satisfaction for Netflix

Member enjoys watching Netflix

So continues the subscription

and

tells their friends about it

Batch learning from bandit feedback

Production policy

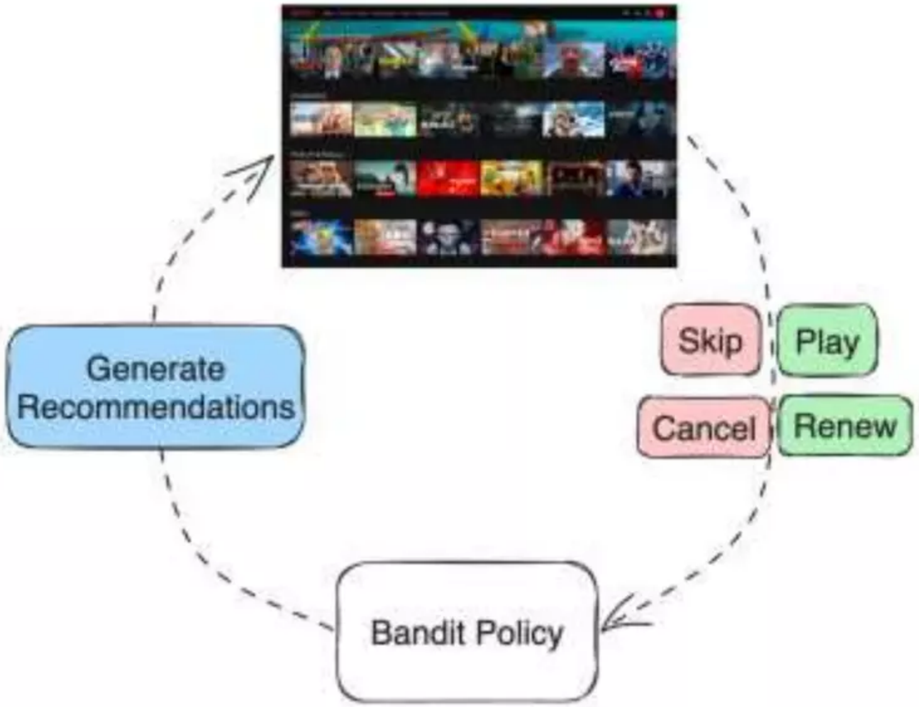
- show recommendations to the member

Member gives feedbacks on recommendations

- immediate: skip/play a show
- long-term: cancel/renew subscription

Goal

- train a policy to maximize the long-term reward



Batch learning from bandit feedback

Production policy

- show recommendations to the me

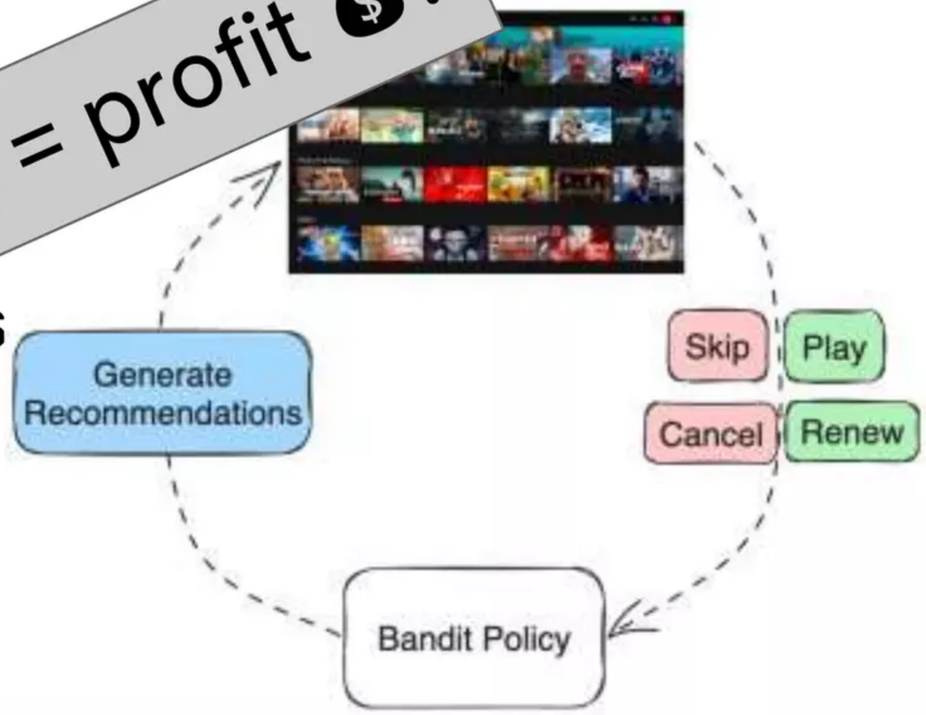
Member gives feedbacks on r ations

- immediate: skip/
- long-term: ew subscription

Goal

- train policy to maximize the long-term reward

Easy: (1) reward = retention (2) = profit 💰!



Challenges with long-term retention

Noisy signal

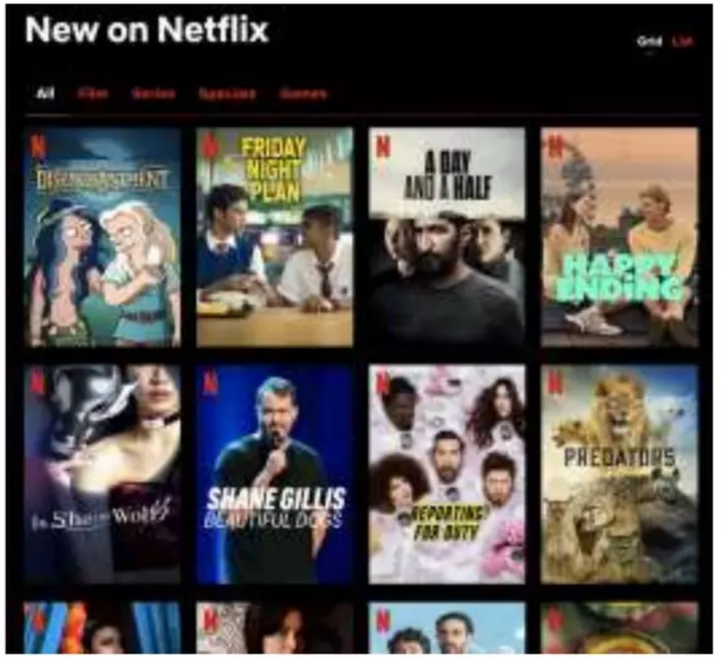
Influenced by external factors

Nonsensitive signal

Only sensitive for “borderline” members

Delayed signal

Need to wait a long time & hard to attribute



Proxy reward

Train policy to optimize *proxy* reward

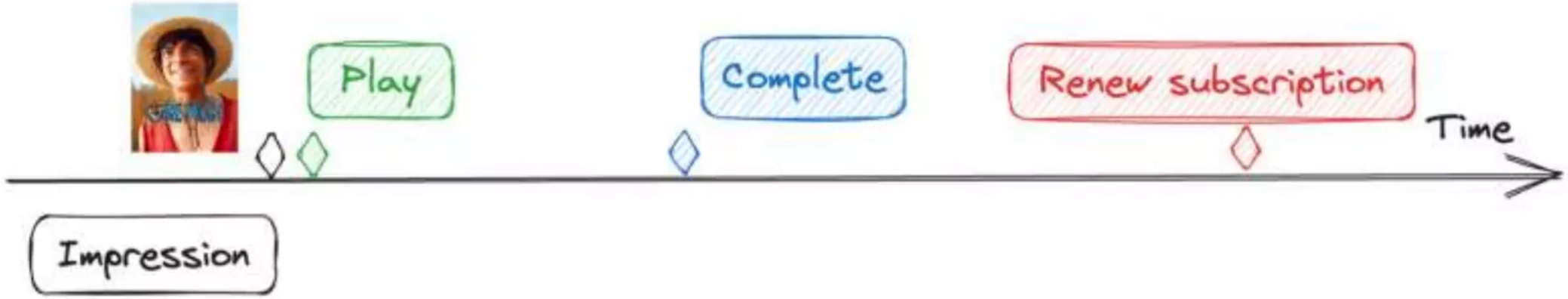
- highly correlated with long-term satisfaction
- sensitive to individual recommendations

Immediate feedback as proxy

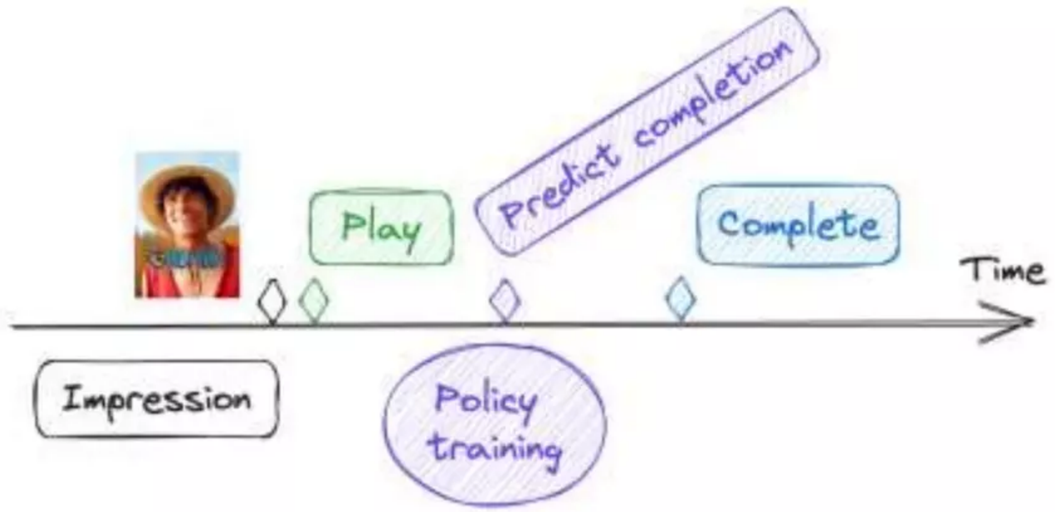
- e.g. start to play a show

Delayed feedback could be more aligned

- e.g. completing a show



Delayed proxy reward



Need to wait a **long time** to observe the delayed reward

- can not be used in training immediately
- can hurt **coldstarting** of model

Don’t want to wait? predict delayed reward

- use all user actions up to policy training time

Train policy to maximize **predicted reward**

Note: predicted reward can not be used online as it uses post-recommendation user actions

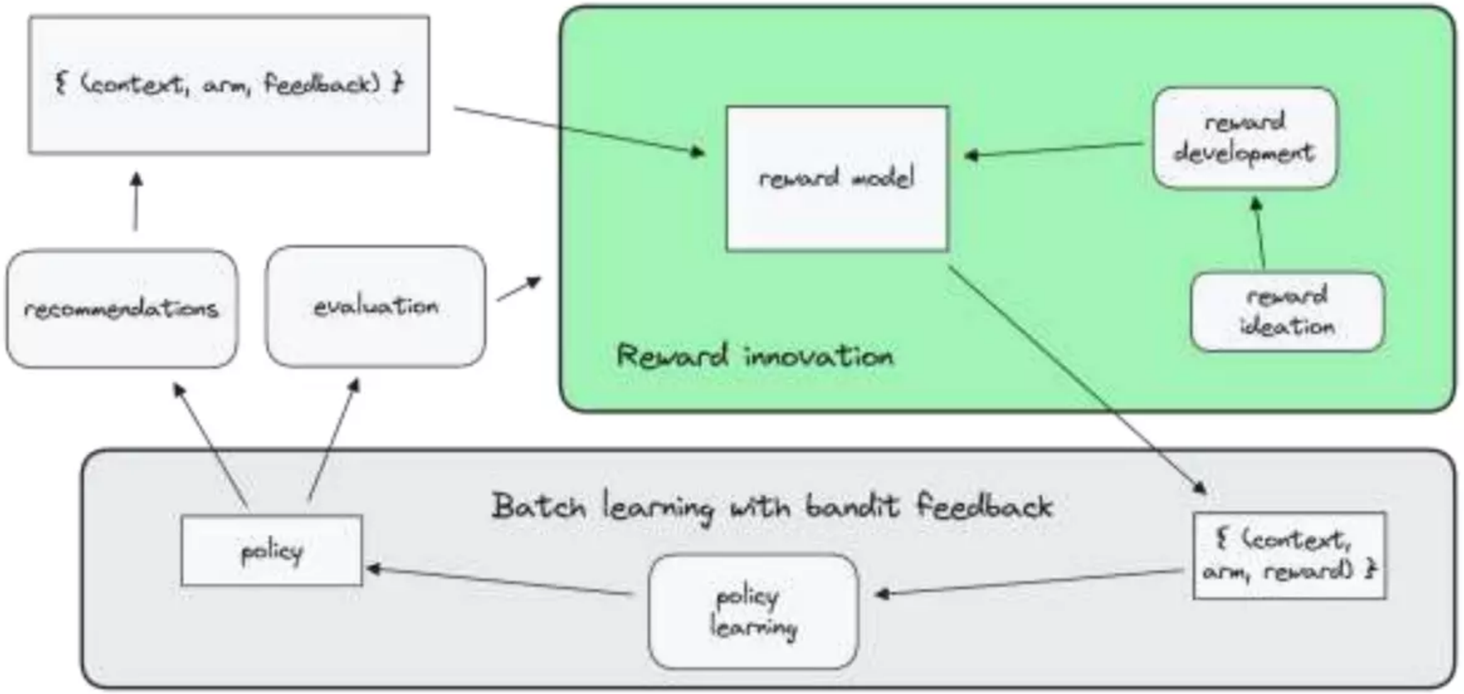
The ideas are not new

Related work:

- Long-term optimization in recommenders
- Reward shaping in reinforcement learning
- Online reward optimization

We focus on reward innovation as an important product development workstream at Netflix

Integrating reward in bandit

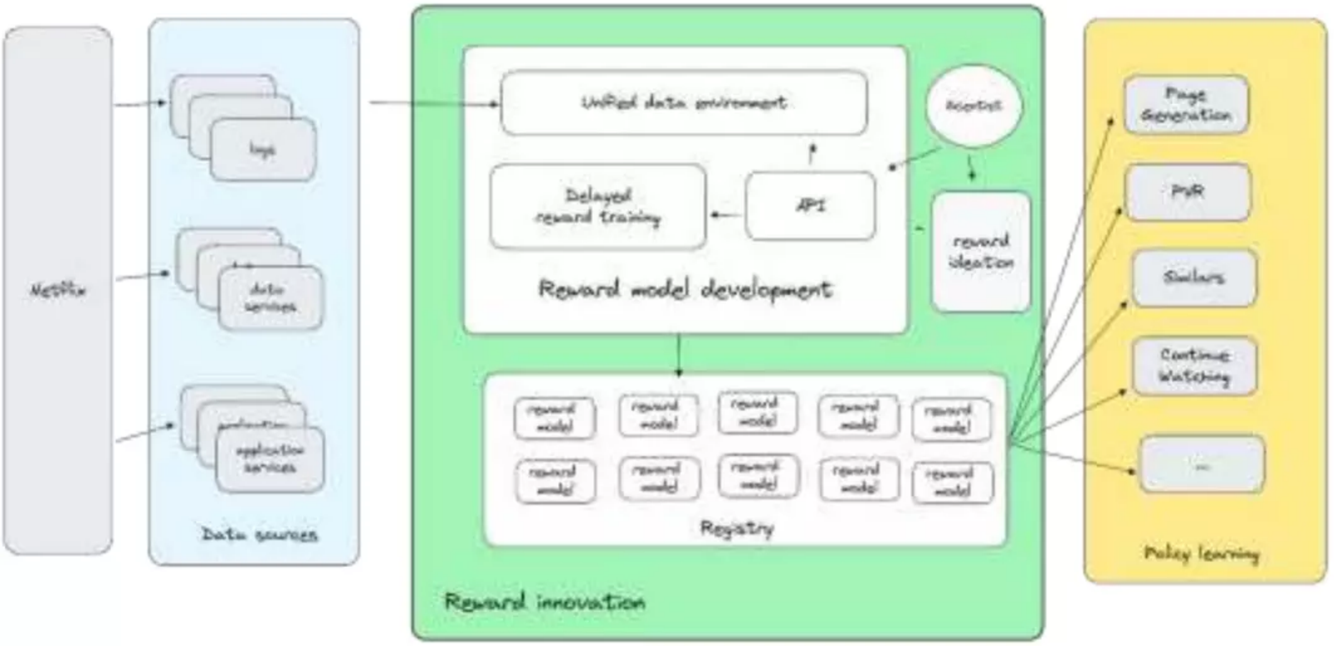


Reward component provides the **objective** of bandit policy training data

Reward innovation

- **Ideation**: what aspect of long-term satisfaction hasn't been captured as a reward?
  - Requires balancing perspectives of ML, business, and psychology. Not easy!
- **Development**: how do we compute this new reward for every recommended item?
  - Collecting immediate feedbacks, predicting delayed feedbacks
- **Evaluation**: whether this is a good reward for the bandit policy to maximize?

Reward infrastructure



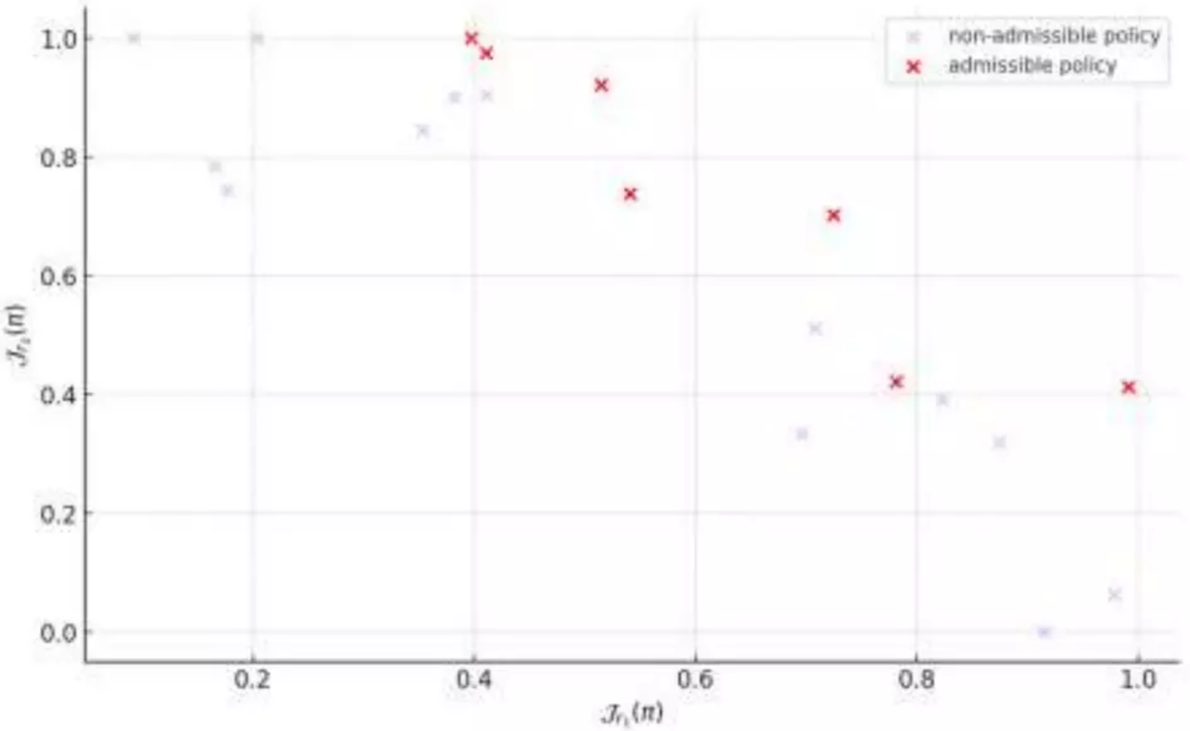
Common development patterns

- Computing immediate proxy rewards
- Predicting delayed proxy rewards
- Combining multiple rewards
- Sharing rewards across multiple recommender components

Reward evaluation

- Online testing is expensive (time, resources)
- Use offline evaluation to determine promising reward candidates for online testing
- **Challenge**: hard to compare policies trained with different rewards

Offline reward evaluation



- compare policies along multiple reward axes using OPE
- choose small number of candidate policies on pareto front
- compare candidate policies online using long-term user satisfaction metrics

Practical learnings

- **Reward normalization**: dynamic range of reward can affect SGD training dynamics
- **Reward features**: pairing a reward with a correlated feature tends to improve the model's ability to optimize that reward
- **Reward alignment**: make different parts of the overall recommender system "point in the same direction"

Summary

- **Proxy rewards**: we can train bandit policies using proxy rewards to optimize long-term member satisfaction
- **Art and science**: to come up with good reward hypotheses
- **Supporting infrastructure**: we develop infrastructure to help iterate on new hypotheses quickly

Open challenges

- **Proxy rewards:** how can we identify proxy rewards that are aligned with long-term satisfaction in a more principle way?
- **Reinforcement learning**: can we use reinforcement learning to optimize long-term reward directly for recommendations?

Thank you!

Questions?

Jiangwei Pan
jpan@netflix.com