Innovating Faster on Personalization Algorithms at Netflix Using Interleaving

Netflix Technology Blog
Nov 30, 2017 · 8 min read

By Joshua Parks, Juliette Aurisset, Michael Ramm

The Netflix experience is powered by a family of ranking algorithms, each optimized for a different purpose. For instance, the *Top Picks* row on the homepage makes recommendations based on a personalized ranking of videos, and the *Trending Now* row also incorporates recent popularity trends. These algorithms, along with many others, are used together to construct personalized homepages for over 100 million members.

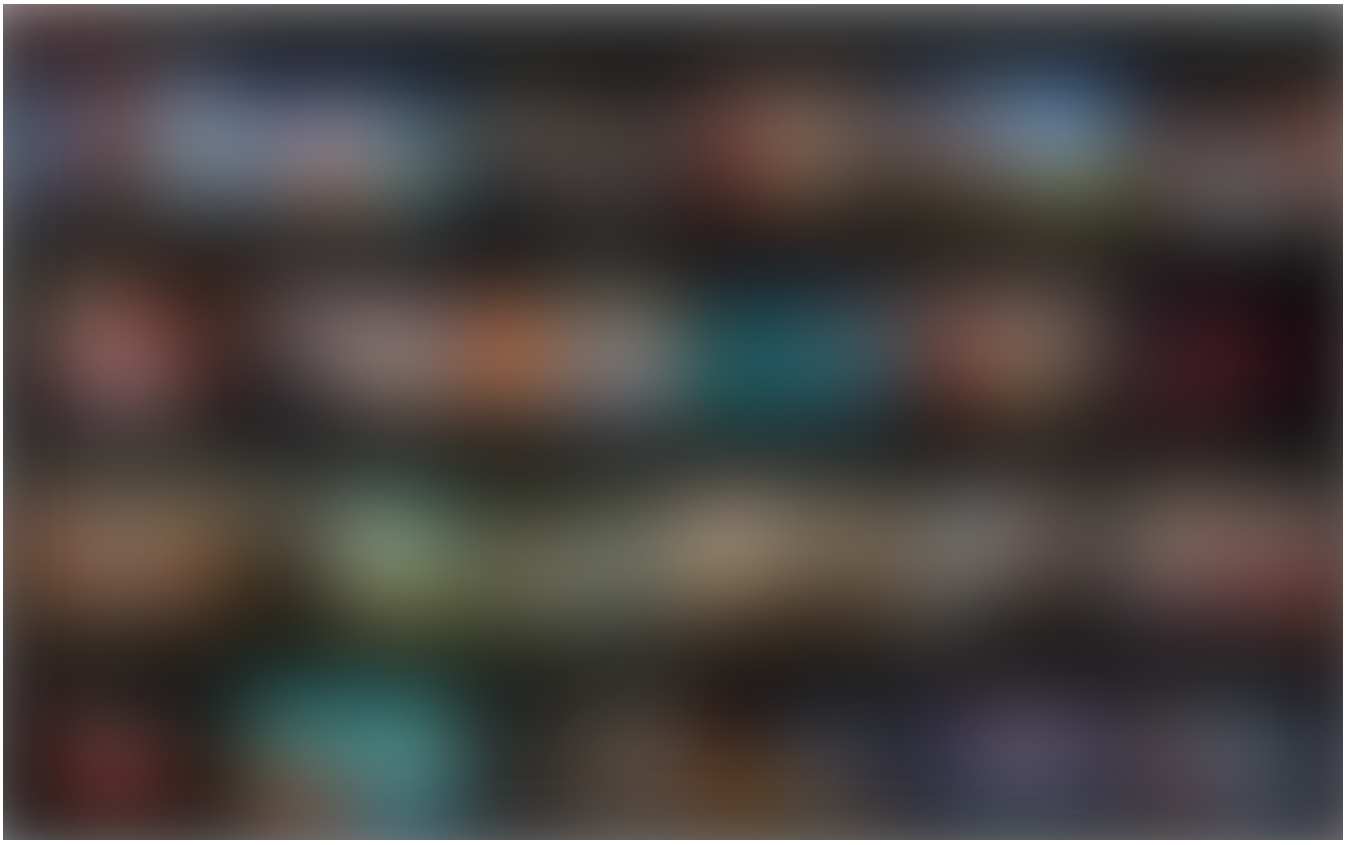


Fig. 1: An example of a personalized Netflix homepage. We use many ranking algorithms to provide personalized recommendations to our members. For a given row, the ordering of videos from left to right is

determined by a specific ranking algorithm.

At Netflix, we strive to continually improve our recommendations. The development process begins with the creation of new ranking algorithms and the evaluation of their performance offline. We then leverage A/B testing to conduct online measurements of core evaluation metrics that align closely with our business objective of maximizing member satisfaction. Such metrics include month-to-month subscription retention and member streaming hours. As the ranking algorithms and the overall Netflix product become optimized, discerning wins in these metrics requires increasingly large sample sizes and long experiment durations.

To accelerate the pace of algorithm innovation, we have devised a two-stage online experimentation process. The first stage is a fast pruning step in which we identify the most promising ranking algorithms from a large initial set of ideas. The second stage is a traditional A/B test on the pared-down set of algorithms to measure their impact on longer-term member behavior. In this blog post, we focus on our approach to the first stage: an *interleaving* technique that unlocks our ability to more precisely measure member preferences.

Faster algorithm innovation with interleaving

Increasing the rate of learning by testing a broad set of ideas quickly is a major driver of algorithm innovation. We have expanded the number of new algorithms that can be tested by introducing an initial pruning stage of online experimentation that satisfies two properties:

- It is highly sensitive to ranking algorithm quality, *i.e.*, it reliably identifies the best algorithms with considerably smaller sample size compared to traditional A/B testing.
- It is predictive of success in the second stage: the metrics measured in the first stage are aligned with our core A/B evaluation metrics.

We have achieved the above using an interleaving technique (cf. Chapelle *et al*.) that dramatically speeds up our experimentation process (see Fig. 2). The first stage finishes in a matter of days, leaving us with a small group of the most promising ranking algorithms. The second stage uses only these select algorithms, which allows us to assign fewer members to the overall experiment and to reduce the total experiment duration compared to a traditional A/B test.

Fig. 2: Faster algorithm innovation using interleaving. The world of new algorithms is represented by light bulbs. Among these, there is a winning idea (depicted in red). The interleaving approach allows us to quickly prune down the initial set of ranking algorithms to the most promising candidates, enabling us to conduct experiments a rate much faster than traditional A/B testing to identify winning ideas.

Using a repeated measures design to determine preferences

To develop intuition around the sensitivity gain that interleaving offers, let's consider an experiment to determine whether Coke or Pepsi is preferred within a population. If we use traditional A/B testing, we might randomly split the population into two groups and perform a blind trial. One group would be offered only Coke, and the second group would be offered only Pepsi (with neither drink having identifiable labels). At the conclusion of the experiment, we could determine whether there is a preference for Coke or Pepsi by measuring the difference in soda consumption between the two groups, along with the extent of uncertainty in this measurement, which can tell us if there is a statistically significant difference.

While this approach works, there may be opportunities for refining the measurement. First, there is a major source of measurement uncertainty: the wide variation in soda consumption habits within the population, ranging from those who hardly consume any soda to those who consume copious amounts. Second, heavy soda consumers may represent a small percentage of the population, but they could account for a large percentage of overall soda consumption. Therefore, even a small imbalance in heavy soda consumers between the two groups may have a disproportionate impact on our

conclusions. When running online experiments, consumer internet products often face similar issues related to their most active users, whether it is in measuring a change to a metric like *streaming hours* at Netflix or perhaps *messages sent* or *photos shared* in a social app.

As an alternative to traditional A/B testing, we can use a repeated measures design for measuring preference for Coke or Pepsi. In this approach, the population would not be randomly split. Rather, each person would have the option of either Coke or Pepsi (with neither brand having identifiable labels but yet still being visually distinguishable). At the conclusion of the experiment, we could compare, at the level of a person, the fraction of soda consumption for Coke or Pepsi. In this design, 1) we remove the uncertainty contributed by the wide range in population-level soda-consumption habits, and 2) by giving every person equal weight, we reduce the possibility that the measurement is materially affected by an imbalance in heavy soda consumers.

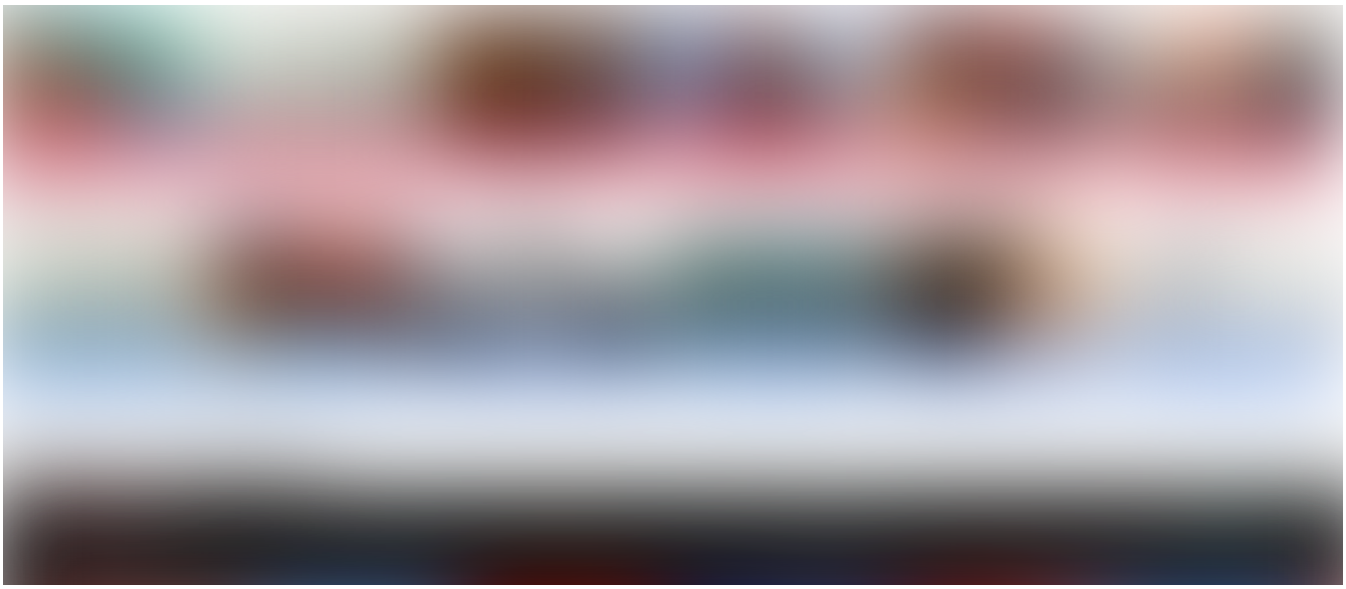
Interleaving at Netflix

At Netflix, we use interleaving in the first stage of experimentation to sensitively determine member preference between two ranking algorithms. The figure below depicts the differences between A/B testing and interleaving. In traditional A/B testing, we choose two groups of subscribers: one to be exposed to ranking algorithm *A* and another to *B*. In interleaving, we select a single set of subscribers who are exposed to an interleaved ranking generated by blending the rankings of algorithms *A* and *B*. This allows us to present choices side-by-side to the user to determine their preference of ranking algorithms. (Members are not able to distinguish between which algorithm recommended a particular video.) We calculate the relative preference for a ranking algorithm by comparing the share of hours viewed, with attribution based on which ranking algorithm recommended the video.

Fig. 3: A/B Testing vs. Interleaving. In traditional A/B testing, the population is split into two groups, one exposed to ranking algorithm A and another to B. Core evaluation metrics like retention and streaming are measured and compared between the two groups. In contrast, interleaving exposes one group of members to a blended ranking of rankers A and B. User preference for a ranking algorithm is determined by comparing the share of viewing hours coming from videos recommended by rankers A or B.

When generating an interleaved set of videos from two ranking algorithms *A* and *B* for a row on the Netflix homepage, we have to consider the presence of position bias: the probability of a member playing a video decreases as we go from left to right. For interleaving to yield valid measurements, we must ensure that at any given position in a row, a video is equally likely to have come from ranking algorithm *A* or *B*.

To address this, we have been using a variant of *team draft* interleaving, which mimics the process of how team selection occurs for a friendly sports match. In this process, two team captains toss a coin to determine who picks first. They then alternate picks, with each captain selecting the player who is highest on their preference list and is still available. This process continues until team selection is complete. Applying this analogy to interleaving for Netflix recommendations, the videos represent the available players and ranking algorithms *A* and *B* represent the ordered preferences of the two team captains. We randomly determine which ranking algorithm contributes the first video to the interleaved list. The ranking algorithms then alternate, with each algorithm contributing their highest-ranked video that is still available (see Fig. 4). The member preference for ranking algorithm *A* or *B* is determined by measuring which algorithm produced the greater share of hours viewed within the interleaved row, with views attributed to the ranker that contributed the video.



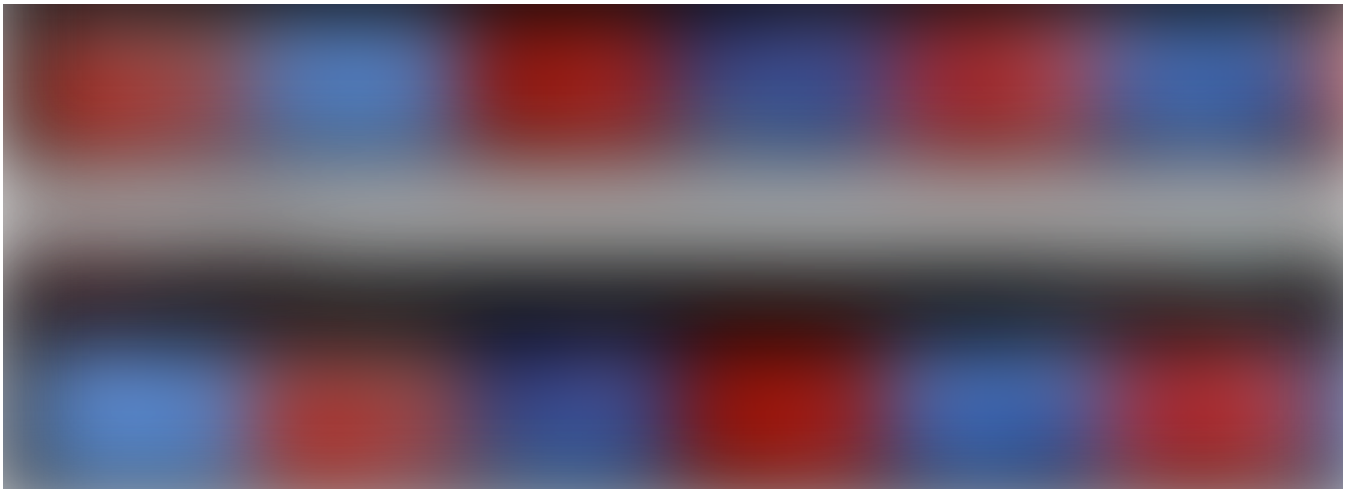


Fig. 4: Interleaving videos from two ranking algorithms using team draft. Ranking algorithms A and B will each have an ordered set of personalized videos. We start with a random coin toss that determines whether ranking algorithm A or B contributes the first video. Each algorithm then takes turns contributing the highest ranked video that is not yet in the interleaved list. Two possible outcomes for the interleaved list are shown depending on which ranker got to select first. We measure user preference by comparing the share of viewing hours attributed to each ranking algorithm.

Comparing the sensitivity of interleaving to traditional A/B testing

The first requirement that we laid out for using interleaving in a two-stage online experimentation process was that it needs to reliably identify the better ranking algorithm with a considerably smaller sample size. To evaluate how well interleaving satisfies this requirement, we turned to a case in which two ranking algorithms A and B were of known relative quality: ranker B is better than ranker A. We then ran an interleaving experiment in parallel with an A/B test using these 2 rankers.

To compare the sensitivity of interleaving vs. A/B testing, we computed both the interleaving preference and A/B metrics at various sample sizes using bootstrap subsampling. In performing the bootstrap analysis, we either simulated assigning N users to the interleaving cell or $N/2$ users to each cell of the traditional A/B experiment. If we were to randomly guess which ranker is better, the probability of disagreeing with the true preference would be 50%. When this probability is 5%, we are achieving 95% power to detect the difference in ranker quality. Therefore, a metric that crosses this threshold with a fewer number of subscribers is the more sensitive one.

Figure 5 shows the results from our analysis. We compare the interleaving preference with two metrics typically used in the A/B setting: overall streaming and an algo-specific engagement metric. The sensitivity of metrics used to evaluate A/B tests can vary over a wide range. We find that interleaving is *very* sensitive: it requires $>100\times$ fewer users than our most sensitive A/B metric to achieve 95% power.

Fig. 5: Sensitivity of interleaving vs. traditional A/B metrics for two rankers of known relative quality. Bootstrap subsampling was used to measure the sensitivity of interleaving compared to traditional engagement metrics. We find that interleaving can require >100× fewer subscribers to correctly determine ranker preference even compared to the most sensitive A/B metric.

Correlation of interleaving metrics with A/B metrics

Our second requirement was that the metrics measured in the interleaving stage need to be aligned with our traditional A/B test metrics. We now evaluate whether the interleaving preference is predictive of a ranker's performance in the subsequent A/B test.

The figure below shows the change in the interleaving preference metric versus the change in the A/B metric compared to control. Each data point represents a ranking algorithm that is evaluated against the production ranker, which serves as control. We find that there is a very strong correlation and alignment between the interleaving metric and our most sensitive A/B evaluation metric, giving us confidence that the interleaving preference is predictive of success in a traditional A/B experiment.

Fig 6: Correlation of the interleaving measurement with the most sensitive A/B metric. Each point represents measurements for a different ranking algorithm evaluated against the production algorithm. There is a strong correlation between the interleaving preference measurement and our most sensitive A/B metric

Conclusion

Interleaving is a powerful technique that has enabled us to accelerate ranking algorithm innovation at Netflix. It allows us to sensitively measure member preference for ranking algorithms and to identify the most promising candidates within days. This has enabled us to quickly test a broad set of new algorithms, and thus increase our rate of learning.

While interleaving provides an enormous boost in sensitivity and aligns well with A/B metrics, it does have limitations. First, implementing an interleaving framework can be fairly involved, which presents challenges from an engineering perspective. The presence of business logic can furthermore interfere, which requires building scalable solutions for consistency checks and automated detection of issues. Second, while interleaving enables quick identification of the best ranking algorithms, a limitation is that it is a relative measurement of user preference for a ranking algorithm. That is, it does not allow us to directly measure changes to metrics such as retention.

We address the latter limitation by running an A/B experiment in a second phase, where our initial set of ideas has been pruned to the best candidates. This gives us the option to power up the experiment by increasing the sample size per cell, which enables us to perform careful measurements of longer-term member behavior. Addressing these challenges and developing better measurements are aspects that we are continuing to explore.

If the work described here sounds exciting to you, please take a look at the jobs page. We are always looking for talented data scientists and researchers to join our team and

help innovate on experimentation methods at Netflix. Your work will help shape the product experience for the next 100M members worldwide!

Data Science Ab Testing Experimentation Algorithms Netflix