A/B Testing and Beyond: Improving the Netflix Streaming Experience with Experimentation and Data Science

Published on June 14, 2017

Nirmal Govind + Follow
Director, Studio Production & Streaming Data Science at Netflix
3 articles

225 5 34

[Originally posted on the Netflix Tech Blog]

Golden Globes for The Crown. An Oscar win for The White Helmets. It's an exciting time to be a Netflix member, with the power to stream such incredible content on a Smart TV or a mobile phone, from the comfort of one's home or on a daily commute.

With a global launch in January 2016 that brought Netflix to over 130 new countries, Netflix is now a truly global Internet TV network, available in almost every part of the world. Our member base is over 100 million strong, and approximately half of our members live outside the US. Since more than 95% of the world's population is located outside the US, it is inevitable that in the near future, a significant majority of Netflix members will be located overseas. With our global presence, we have the opportunity to watch, learn, and improve the service in every part of the world.

A key component of having a great Internet TV service is the streaming quality of experience (QoE). Our goal is to ensure that you can sit back and enjoy your favorite movie or show on Netflix with a picture quality that delights you and a seamless experience without interruptions or errors. While streaming video over the Internet is in itself no small feat, doing it well at scale is challenging (Netflix accounts for more than a third of Internet traffic at peak in North America). It gets even more complex when we're serving members around the world with not only varying tastes, network infrastructure, and devices, but also different expectations on how they'd like to consume content over the Internet.

Mumbai, and Bangkok as it is in San Francisco, London, or Paris? The scientists at Netflix continuously innovate to ensure that we can provide possible. To enable this, Netflix has a culture of experimentation and da making that allows new ideas to be tested in production so we get data our members. In this post, I'll focus on the experimentation that we do QoE, including the types of experiments we run, the key role that data s also how the Netflix culture enables us to innovate via continuous expe will not delve into the statistics behind experimentation but I will outlin statistical challenges we're working on in this space.

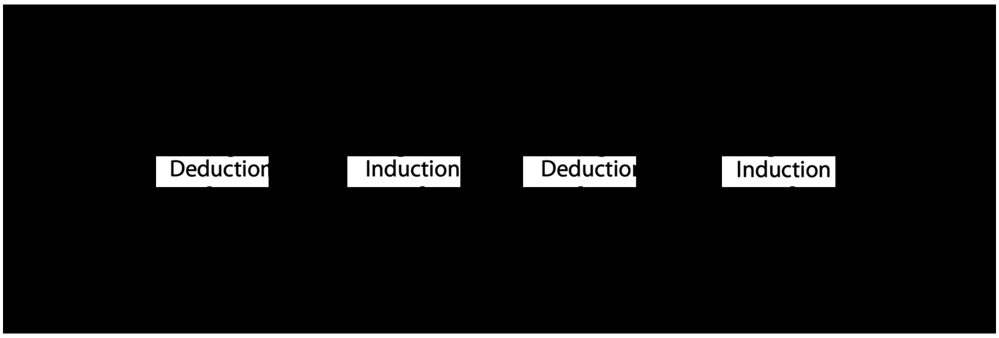
We also use data to build Machine Learning and other statistical models to improve QoE; I will not focus on the modeling aspect here but refer to this blog post for an overview and this post highlights one of our modeling projects. It's also worth noting that while the focus here is on QoE, we use experimentation broadly across Netflix to improve many aspects of the service, including user interface design, personalized recommendations, original content promotion, marketing, and even the selection of video artwork.

Need for Experimentation

Before getting to the experiments that we run, it's useful to develop some intuition around why a structured approach to experimentation is not just a nice-to-have but a necessary part of innovation.

Enabling the Scientific Method

Experiments and empirical observation are the most important part of the scientific method, which allows engineers and scientists to innovate by formulating hypotheses, gathering data from experiments, and making conclusions or formulating new hypotheses. The scientific method emphasizes an iterative learning process, alternating between deduction and induction (see figure below, courtesy of the famous statistician George Box).



Deduction is the process of going from an idea or theory to a hypothesis to actual observations/data that can be used to test the hypothesis, while induction is the process of generalizing from specific observations/data to new hypotheses or ideas. Experimentation plays a critical role in collecting data to test hypotheses and enabling the deduction-induction iterations as part of the scientific method.

Establishing Causation

Experimentation is a data-based approach to ensure we understand the i
we make to our service. In our case, we’d like to understand the impact
related algorithm or a change to an existing algorithm. Usually, we’re i
answering two questions: 1) How does the change (“treatment”) affect (
What effect does the change have on member behavior: do our member
experience or the old one?

In general, an experiment allows us to obtain a causal read on the impa
allows us to make a claim, with some degree of confidence, that the res
caused by the change we made. In controlled experiments such as A/B tests, proper randomization ensures that the control and treatment groups in a test differ only in the experience or “treatment” they receive, and other factors (that may or may not affect the experiment’s results) are present in equal proportions in both groups. This makes A/B testing a popular approach for running experiments and determining if experience “A” or experience “B” works better.

It’s worth noting that experiments help establish causation as opposed to relying on correlation in observed data. In this regard, experimentation may be thought of as being superior to most ML approaches that are based on observational data. We do spend a significant amount of effort in researching and building ML models and algorithms. Carefully exploiting patterns in observed data is powerful for making predictions and also in reaffirming hypotheses, but it’s even more powerful to run experiments to get at causation.

Data-driven Judgment

Last but not least, experiments are a powerful way to let data guide decision-making. Making decisions based on data from experiments helps avoid the HiPPO (Highest Paid Person’s Opinion) problem, and also ensures that intuition alone does not drive decisions. When combined with human judgment, experiments are a powerful tool to ensure that the best ideas win. Culture plays an important role here, more on that later in the post.

Streaming Quality of Experience

There are several aspects that determine QoE, and I’ll provide a brief overview of three key components before getting into the types of experiments we run at Netflix to improve QoE.

For each movie or episode of a show we stream, the encoding process creates files at different video quality levels (bitrates), which are then cached on our servers distributed around the world. When a member initiates play, client-side adaptive streaming algorithms select the best bitrate to stream based on network and other considerations, and server-side algorithms determine how best to send packets of data over to the client.

Let’s take a closer look at these components, starting with the algorithms that run on a member’s device.

Adaptive Streaming

Messaging

for streaming. Netflix is available on thousands of devices ranging from tablets to game consoles, computers, and Smart TVs. Most of these dev streaming algorithms developed by Netflix that decide what bitrate shou various times during a streaming session. These bitrate selection decisic quality of the video on the screen and also directly influence how quick the device is depleted. When the buffer runs out, playback is interrupted occurs.

We obsess over great playback experiences. We want playback to start i quality, and we never want playback to stop unexpectedly. But in reality last mile connectivity issues may make this impossible to achieve. What we can do is design algorithms that can quickly detect changes in network throughput and make adjustments in real-time to provide the best experience possible.

Given the large number of networks, network conditions, and device-level limitations as we serve content to millions of members around the world, it's necessary to rely on the scientific method to tune existing algorithms and develop new algorithms that can adapt to a variety of scenarios. The adaptive streaming engineers use experimentation to develop and continuously improve the algorithms and configurations that provide the best experience for each streaming session on Netflix.

Content Delivery

Open Connect is Netflix's Content Delivery Network (CDN), and it's responsible for serving the video and audio files needed to play content during a streaming session. At a high level, Open Connect allows us to locate content as close as possible to our members in order to maximize delivery efficiency and QoE. The Open Connect team does this by partnering with Internet Service Providers (ISPs) to localize their Netflix traffic by embedding servers with Netflix content inside the ISP network. Open Connect also peers with ISPs at interconnect locations such as Internet Exchanges around the world. For more on how Open Connect works, check out this blog post.

The engineers in Open Connect optimize both the hardware and the software on the servers used to serve Netflix content. This allows us to tune the server configuration, software, and algorithms for the specific purpose of video streaming. For example, caching algorithms determine what content should be stored on servers distributed around the world based on what content is likely to be watched by members served by those servers. Engineers also develop network transport algorithms that determine how packets of data are sent across the internet from the server to a member's device. For more on some of the problems in this space, refer to this blog post.

Similar to adaptive streaming on the client-side, experimentation enables rapid iteration and innovation in Open Connect as we develop new architectures and algorithms for content delivery. There is additional complexity in this area due to nature of the system; in some scenarios, it's impractical to do a controlled randomized experiment, so we need to adapt experimental techniques to get a causal read. More on this further below.

Encoding

The perceptual quality of content is also an important aspect of
impact on what's seen on the screen. Perceptual quality is tied to a proc
which compresses the original "source" files corresponding to a movie
files or "encodes" at different bitrates. The encoding algorithms are an
innovation at Netflix, and our encoding team has made some significant
provide better perceptual quality at a given network bandwidth or use le
quality level. More recently, the engineers have been working on more
low-bandwidth streaming.

Encoding changes pose a different challenge for experimentation as the
usually specific to the content in each movie or show. For example, the
change may be different for animated content versus an action-packed thriller. In addition, it's also important to ensure that encoding changes are compatible with the client application and the decoders on the devices used to stream Netflix.

Before we roll out a new encoding algorithm, which also means re-encoding the entire Netflix catalog, the encoding team runs experiments to validate changes and measure the impact on QoE. The experimental design for such experiments can be challenging due to content-specific interactions and the need to validate on sufficiently diverse content and devices.

Experiments to Improve QoE

Let's take a look at the types of experiments we run at Netflix across the areas outlined above to improve QoE. Broadly, there are three classes of experiments we run to improve QoE and understand the impact of QoE on member behavior.

System Experiments

The goal of system experiments is to establish whether a new algorithm, change to an existing algorithm, or a configuration parameter change has the intended effect on QoE metrics. For example, we have metrics related to video quality, rebuffers, play delay (time between initiating playback and playback start), playback errors, etc. The hypotheses for these experiments are typically related to an improvement in one or more of these metrics.

System experiments are usually run as randomized A/B experiments. A system test may last a few hours or may take a few days depending on the type of change being made, and to account for daily or weekly patterns in usage and traffic. Our 100 million strong member base allows us to obtain millions of "samples" relatively quickly, and this allows for rapid iteration and multiple system experiments to be run sequentially to optimize the system.

From an experimenter's standpoint, these fast-paced system experiments allow for exploration of new experimentation methodologies. For example, we can test new strategies for allocation to control and treatment groups that allow us to learn quickly. We're also adapting Response Surface Methodology techniques to build statistical models from experimental data that can reduce the number of iterations needed to achieve a set goal.

Testing in this area poses a number of challenges that also motivate our research; below are a couple examples.

Messaging

testing methods that account for such distributions. For this reason, we nonparametric statistical methods in our analysis to establish statistical Nonparametric methods on really large datasets can be rather slow, so t of research we're exploring.

Furthermore, in these experiments, we typically measure several QoE n correlated, across multiple treatment cells, and need to account for the r problem.

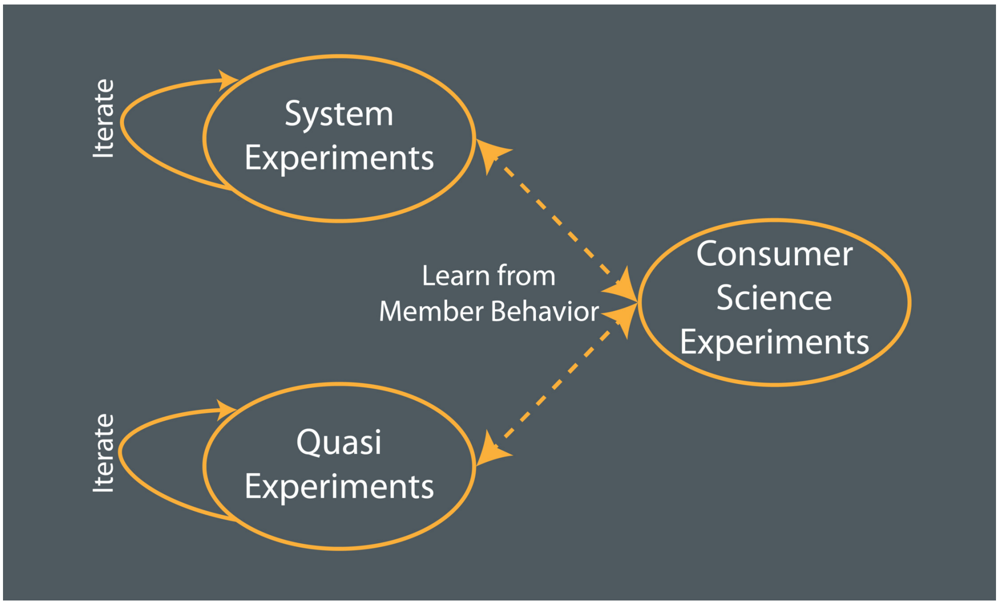
Quasi-Experiments and Causal Inference

Most of our system experiments are controlled randomized A/B experiments. However, in certain situations where randomization isn't feasible, we resort to other approaches such as quasi-experiments and causal inference.

One area where we're exploring quasi-experiments is to test changes to algorithms in Open Connect. Consider an Internet Exchange with two identical server (or cache) clusters where one cluster serves member traffic from ISP #1 and the other cluster serves traffic from ISP #2. If we're interested in testing a new algorithm for filling content on caches, ideally we would run an A/B experiment with one cache cluster being control and the other being the treatment. However, since the traffic to these clusters cannot be randomized (peering relationships are difficult to modify), an A/B experiment is not possible.

In such situations, we run a quasi-experiment and apply causal inference techniques to determine the impact of the change. Several challenges abound in this space such as finding a matching control cluster, determining the appropriate functional relationship between treatment and control, and accounting for network effects.

Consumer Science Experiments



Experiments designed to understand the impact of changes on Netflix member behavior are called Consumer Science experiments. Typically, these experiments are run after several iterations of system experiments or quasi-experiments are completed to

Messaging

algorithm or configuration change has the intended effect on QoE me
study the impact of QoE changes on member behavior: 1) Do members
if they have better video quality or lower rebuffers or faster playback st
retain better after the free trial month ends and in subsequent months?

We can also study the effect on member behavior of making tradeoffs a
do members prefer faster playback start (lower play delay) with lower v
they prefer to wait a bit longer but start at a higher quality?

Consumer Science experiments typically run for at least one month so v
member retention after the free month for new members. An interesting
experiments is to identify segments of the member base that may differ in their expectations around QoE. For example, a change that drastically reduces play delay at the expense of lower initial video quality may be preferred by members in parts of the world with poorer internet connectivity, but the same experience may be disliked by members on stable high-speed internet connections. The problem is made more difficult due to the fact that changes in QoE may be subtle to the member, and it may take a while for behavior changes to manifest as a result of QoE shifts.

A Culture of Experimentation

Last but not least, I'd like to discuss how company culture plays an important role in experimentation. The Netflix culture is based on the core concept of “freedom and responsibility”, combined with having stunning colleagues who are passionate and innovative (there's more to our culture, check out the Netflix culture deck). When you have highly talented individuals who have lots of great ideas, it's important to have a framework where any new idea can be developed and tested, and data, not opinion, is used to make decisions. Experimentation provides this framework.

Enabling a culture of experimentation requires upfront commitment at the highest level. At Netflix, we look for ways to experiment in as many areas of the business as possible, and try to bring scientific rigor into our decision-making.

Data Science has a huge role to play here in ensuring that appropriate statistical rigor is applied as we run experiments that determine the kind of product and service that our members experience. Data Science is also necessary to come up with new ideas and constantly improve how we run experiments at Netflix, i.e. to experiment with our approach to experimentation. Our data scientists are heavily involved in the design, execution, analysis, and decision making for experiments we run, and they also work on advancing experimentation methodology.

In addition to the science, it's also important to have the infrastructure in place to run experiments and analyze them, and we have engineering teams that are focused on improving our experimentation platform. This platform enables automation of the steps needed to kick off an experiment as well as automated generation of analysis reports and visualizations during various phases of the experiment.

Netflix is leading the Internet TV revolution and we're changing how people around the
world watch movies and TV shows. Our data scientists and engineers w

all walks of life with stories from different cultures, and it is both inspi

satisfying. We're hiring so reach out if you'd like to join us in this amaz

225 Likes

5 Comments

Show previous comments

Apurva Kansara 1y ···
Senior Software Engineer at Netflix

Well written post.

Like Reply | 1 Like

Kashif Haidery 1y ···
Manager-Software Engineering; Lead an agile program| OTT | Connected/Video Broadcast|Cloud Vid...

Nicely written article, linking QOE to company culture and approach to experimentation.

Like Reply

Add a comment...

Nirmal Govind

Director, Studio Production & Streaming Data Science at Netflix

+ Follow

More from Nirmal Govind

Optimizing Content Quality Control at Netflix with Predictive Modeling

Nirmal Govind on LinkedIn

Streaming Science at Netflix: Turning Data into a Better Viewin...

Nirmal Govind on LinkedIn

Messaging