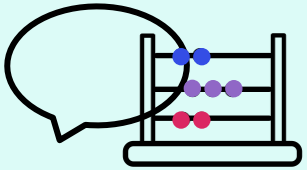


Mixed Methods Development of Evaluation Metrics

A Tutorial at KDD 2021

Praveen Chandar [Spotify]

Fernando Diaz [Google]

Christine Hosey [Spotify]

Brian St. Thomas [Spotify]

Introduction

What will you learn today and why it matters?

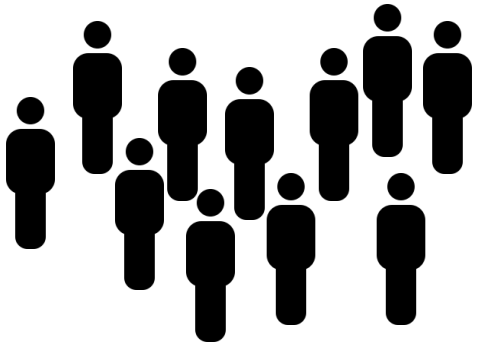
Why evaluate?

01 Ensure that the service is matching expectations.

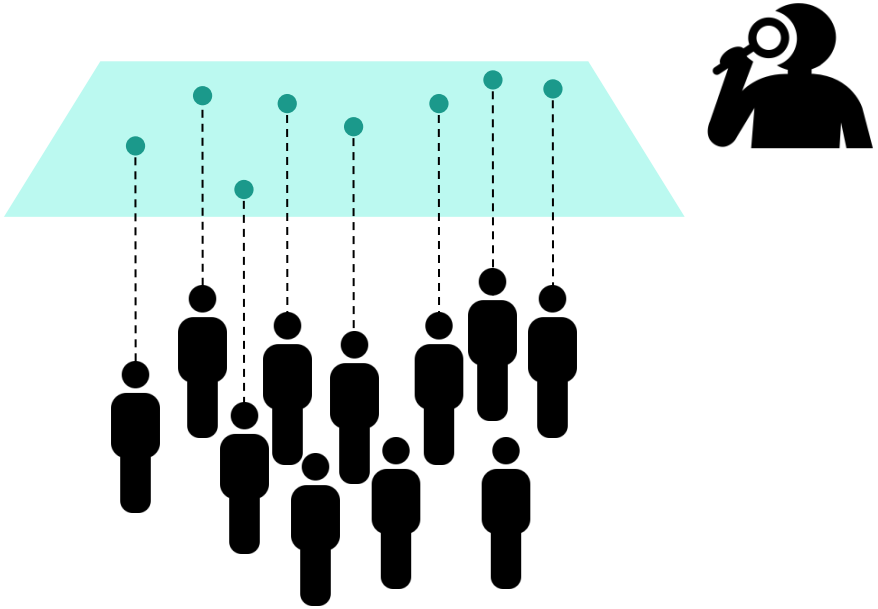
02 Improve products.

How do we evaluate?

users

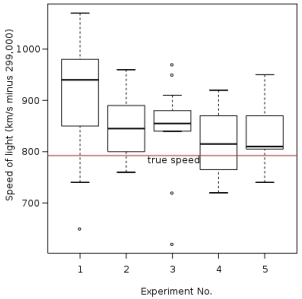


How do we evaluate?

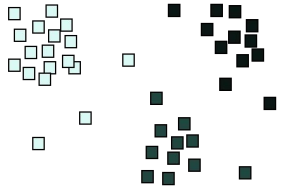


quantitative
evaluation

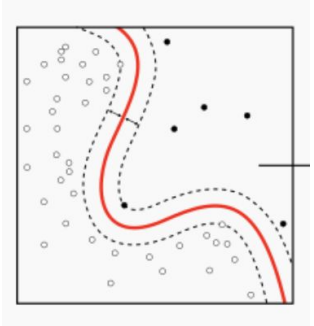
Quantitative methods



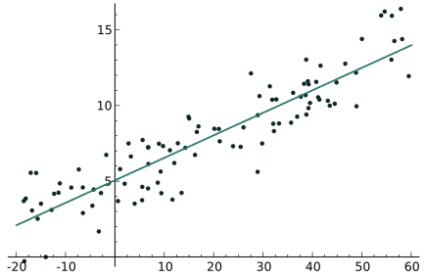
Descriptive statistics - summarize the data



Clustering - group together similar data points



Classification - place data points into categories



Regression - find relationships between variables

Why does the right metric matter?

Evaluation of recommender has long been divided between accuracy metrics (e.g., precision/recall) and error metrics (notably, RMSE and MAE). The mathematical convenience and fitness with formal optimization methods, have made error metrics like RMSE more popular, and they are indeed dominating the literature. However, it is well recognized that accuracy measures may be a more natural yardstick, as they directly assess the quality of top-N recommendations.

This work shows, through an extensive empirical study, that the convenient assumption that an error metric such as RMSE can serve as good proxy for top-N accuracy is questionable at best. There is no monotonic relation between error metrics and accuracy metrics. This may call for a re-evaluation of optimization goals for top-N systems. On the bright side we have presented simple and efficient variants of known algorithms, which are useless in RMSE terms, and yet deliver superior results when pursuing top-N accuracy.

Task: recommendation

Metric: minimize error of predicted ratings

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. RecSys, 2010.

Why does the right metric matter?

ASLIB CRANFIELD RESEARCH PROJECT

FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS VOLUME 1, DESIGN

by

Cyril Cleverdon, Jack Mills and Michael Keen

Part 1. Text

An investigation supported by a grant to Aslib by the National Science Foundation

Cranfield
1966

Optimizing Search Engines using Clickthrough Data

Thorsten Joachims
Cornell University
Department of Computer Science
Ithaca, NY 14853 USA
tj@cs.cornell.edu

ABSTRACT

This paper presents an approach to automatically optimizing the retrieval quality of search engines using clickthrough data. Intuitively, a good information retrieval system should present relevant documents high in the ranking, with less relevant documents following below. While previous approaches to learning retrieval functions from examples exist, they typically require training data generated from relevance judgments by experts. This makes them difficult and expensive to apply. The goal of this paper is to develop a method that utilizes clickthrough data for training, namely the query-log of the search engine in connection with the log of links the users clicked on in the presented ranking. Such clickthrough data is available in abundance and can be recorded at very low cost. Taking a Support Vector Machine (SVM) approach, this paper presents a method for learning retrieval functions. From a theoretical perspective, this method is shown to be well-founded in a risk minimization framework. Furthermore, it is shown to have an efficient implementation for linear ranking functions. The theoretical results are verified in a controlled experiment. It shows that the method can effectively adapt the retrieval function of a meta-search engine to a particular group of users, outperforming Google in terms of retrieval quality after only a couple of hundred training examples.

1. INTRODUCTION

Which WWW page(s) does a user actually want to retrieve when he types some keywords into a search engine? There are typically thousands of pages that contain these words, but the user usually has only a few pages in mind. One could simply ask the user for feedback. If we knew the set of pages actually relevant to the user's query, we could use this as training data for optimizing (and even personalizing) the retrieval function.

Unfortunately, experience shows that users are only rarely willing to give explicit feedback. However, this paper argues that sufficient information is already hidden in the logfiles of WWW search engines. Since major search engines receive millions of queries per day, such data is available in abundance. Compared to explicit feedback data, which is typically elicited in laborious user studies, any information that can be extracted from logfiles is virtually free and substantially more timely.

This paper presents an approach to learning retrieval functions by analyzing which links the users click on in the presented ranking. This leads to a problem of learning with preference examples like "for query q, document d_3 should be ranked higher than document d_7". More generally, I will formulate the problem of learning a ranking function over a finite domain in terms of empirical risk minimization. For this formulation, I will present a Support Vector Machine (SVM) algorithm that leads to a convex program and that can be extended to learn non-linear ranking functions. Experiments show that the method can successfully learn a highly effective retrieval function for a meta-search engine.

This paper is structured as follows. It starts with a definition of what clickthrough data is, how it can be recorded, and how it can be used to generate training examples in the form of preferences. Section 3 then introduces a general framework for learning retrieval functions, leading to an SVM algorithm for learning parameterized orderings in Section 4. Section 5 evaluates the method based on experimental results.

2. CLICKTHROUGH DATA IN SEARCH ENGINES

Clickthrough data in search engines can be thought of as triplets (q, r, c) consisting of the query q, the ranking r presented to the user, and the set c of links the user clicked on. Figure 1 illustrates this with an example: the user asked the query "support vector machine", received the ranking shown in Figure 1, and then clicked on the links ranked 1, 3, and 7. Since every query corresponds to one triplet, the amount of data that is potentially available is virtually unlimited.

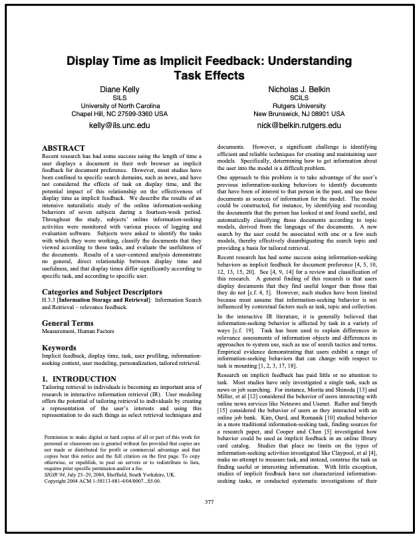
Clearly, users do not click on links at random, but make a (somewhat) informed choice. While clickthrough data is typically noisy and clicks are not "perfect" relevance judgments, the clicks are likely to convey some information. The key question is: how can this information be extracted? Below deriving a model of how clickthrough data can be analyzed, let's first consider how it can be recorded.

2.1 Recording Clickthrough Data

Clickthrough data can be recorded with little overhead and without compromising the functionality and usefulness

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'02 Edmonton, Alberta, Canada.
Copyright 2002 ACM 1-58113-561-X/02/0007 ...$5.00.

2002



Display Time as Implicit Feedback: Understanding Task Effects

Diane Kelly
SILS
University of North Carolina
Chapel Hill, NC 27599-3360 USA
kelly@ils.unc.edu

Nicholas J. Belkin
SCILS
Rutgers University
New Brunswick, NJ 08901 USA
nick@belkin.rutgers.edu

ABSTRACT

Recent research has had some success using the length of time a user displays a document in their web browser as implicit feedback for document preference. However, most studies have been confined to specific search domains, such as news, and have not considered the effects of task on display time, and the potential impact of this relationship on the effectiveness of display time as implicit feedback. We describe the results of an intensive naturalistic study of the online information-seeking behaviors of seven subjects during a fourteen-week period. Throughout the study, subjects' online information-seeking activities were monitored with various pieces of logging and evaluation software. Subjects were asked to identify the tasks with which they were working, classify the documents that they viewed according to these tasks, and evaluate the usefulness of the documents. Results of a user-centered analysis demonstrate no general, direct relationship between display time and usefulness, and that display times differ significantly according to specific task, and according to specific user.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – relevance feedback.

General Terms

Measurement, Human Factors

Keywords

Implicit feedback, display time, task, user profiling, information-seeking context, user modeling, personalization, tailored retrieval.

1. INTRODUCTION

Tailored retrieval of individuals is becoming an important area of research in interactive information retrieval (IR). User modeling offers the potential of tailoring retrieval to individuals by creating a representation of the user's interests and using this representation to do such things as select retrieval techniques and documents. However, a significant challenge is identifying efficient and reliable techniques for creating and maintaining user models. Specifically, determining how to get information about the user into the model is a difficult problem.

One approach to this problem is to take advantage of the user's previous information-seeking behaviors to identify documents that have been of interest to that person in the past, and use these documents as sources of information for the model. The model could be constructed, for instance, by identifying and recording the documents that the person has looked at and found useful, and automatically classifying those documents according to topic models, derived from the language of the documents. A new search by the user could be associated with one or a few such models, thereby effectively disambiguating the search topic and providing a basis for tailored retrieval.

Recent research has had some success using information-seeking behaviors as implicit feedback for document preference [4, 5, 10, 12, 13, 15, 20]. See [4, 9, 14] for a review and classification of this research. A general finding of this research is that users display documents that they find useful longer than those that they do not [c.f. 4, 5]. However, such studies have been limited because most assume that information-seeking behavior is not influenced by contextual factors such as task, topic and collection.

In the interactive IR literature, it is generally believed that information-seeking behavior is affected by task in a variety of ways [c.f. 19]. Task has been used to explain differences in relevance assessments of information objects and differences in approaches to system use, such as use of search tactics and terms. Empirical evidence demonstrating that users exhibit a range of information-seeking behaviors that can change with respect to task is mounting [1, 2, 3, 17, 18].

Research on implicit feedback has paid little or no attention to task. Most studies have only investigated a single task, such as news or job searching. For instance, Morita and Shinoda [11] and Miller, et al [12] considered the behavior of users interacting with online news services like Netnews and Usenet. Ruthe and Smyth [15] considered the behavior of users as they interacted with an online job bank. Kim, Oard, and Romanik [10] studied behavior in a more traditional information-seeking task, finding sources for a research paper, and Cooper and Chen [3] investigated how behavior could be used as implicit feedback in an online library card catalog. Studies that place no limits on the types of information-seeking activities investigated like Claypool, et al [4], make no attempt to measure task, and instead, construe the task as finding useful or interesting information. With little exception, studies of implicit feedback have not characterized information-seeking tasks, or conducted systematic investigations of their

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2004 ACM 1-58113-881-4/04/0007...$5.00.

377

2004

No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search

Jeff Huang
Information School
University of Washington
chi@jeffhuang.com

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Susan Dumais
Microsoft Research
Redmond, WA 98052
sdumais@microsoft.com

ABSTRACT

Understanding how people interact with Web sites is important in improving site design and the quality of services offered. The Web provides unprecedented opportunities to evaluate alternative design, interaction, and algorithmic methods at scale and *in situ* with actual customers doing their own tasks in their own environments [19]. Such studies typically involve measuring clicks which can be obtained easily at scale. However, they fail to capture behaviors that do not lead to clicks (e.g., which items are attended to, in what order, etc.) or subjective impressions. Gaze-tracking studies with participants present in the laboratory can provide more detailed insights but on a smaller scale. In this paper we consider how mouse movements, which can be collected remotely on a large scale, can be used to understand richer patterns of behavior.

We focus on understanding cursor activities in Web search behavior. People conduct Web searches to satisfy information needs. Their interaction with search engines begins by issuing a search query, then reviewing the search engine results page (SERP) to determine which, if any, results may satisfy their need. In doing so, they may move their mouse cursor around the page, hovering over and possibly clicking on hyperlinks. Small-scale laboratory studies have observed participants making many uses of the cursor on SERPs beyond hyperlink clicking [1,23,25]. These uses include moving the cursor as a reading aid, using it to mark interesting results, using it to interact with controls on the screen (e.g., buttons, scroll bars), or simply positioning the cursor so that it does not occlude Web page content. However, studying such behaviors in small-scale laboratory settings is limited in terms of what inferences can be made.

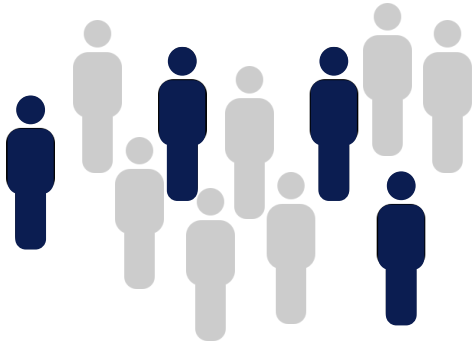
Tracking mouse cursor movements *at scale* can provide a rich new source of behavioral information to understand, model, and satisfy information needs. Recent research has shown that cursor movements correlate with eye gaze [6,13,25,26], and may therefore be an effective indicator of user attention. We believe that cursor data, like click data [18], can provide signals that reveal searcher intent and may be useful in improving the search experience. Cursor data can be used to complement click data in several ways. First, cursor data can be captured for uncommon queries where strong indicators of relevance such as result clicks may occur less frequently or not at all. For example, analyzing click logs for a query that has been issued several times but never clicked may provide limited relevance information, but cursor behavior on the SERP associated with the query may provide insight about relevance. Second, in cases of so-called *good abandonment* [20], where the content on the SERP satisfies the user's information need directly, a search result click may be unnecessary. Thus the lack of a click should not always be interpreted as a search failure. Cursor behavior may help in distinguishing between good and bad search abandonment.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05...$10.00.

2011

How do we evaluate?

qualitative
evaluation

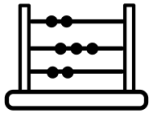


Qualitative methods

FOUNDATIONAL	GENERATIVE	EVALUATIVE
Needs, wants, existing habits	Building toward a solution	Assess usability, usefulness, delight
ETHNOGRAPHY	DESIGN SPRINT	USABILITY
CONTEXTUAL INQUIRY	PARTICIPATORY DESIGN	HEURISTIC EVALUATION
SEMI-STRUCTURED INTERVIEW	ANNOTATION STUDY	DIARY STUDY
	BEHAVIORAL EXPERIMENT	SURVEY
	CONCEPT TEST	COMPETITIVE BENCHMARKING

Qualitative

- Explore meaning and understanding of constructs
- Emphasize participant voices
- Small sample size



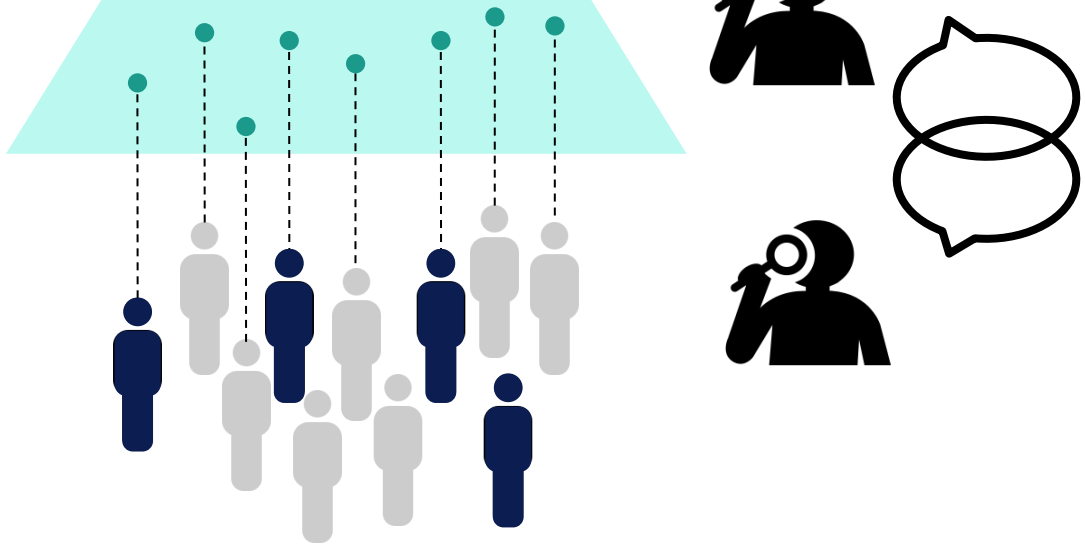
Quantitative

- Assess magnitude and frequency of constructs
- Generalize over a population
- Large sample size

Creswell, J.W., Klassen, A.C., Plano Clark, V.L. and Smith, K.C., 2011. Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health, 2013, pp.541-545.

How do we evaluate?

mixed methods
evaluation



When should we use mixed methods?

- To augment experiments by incorporating the **perspectives of individuals**
- To **explain** initial quantitative results
- To qualitatively **explore** questions, variables, and theories prior to a quantitative study
- To obtain more **complete and corroborated** results

Creswell, J.W. and Clark, V.L.P. (2017). 'The Nature of Mixed Methods Research.' In *Designing and conducting mixed methods research*. Sage publications.

Learning goals

01

How **quantitative** methods alone can build in assumptions that may lead you **astray**.

02

How to incorporate **qualitative** methods to generate hypotheses to **course correct**.

03

How to apply **quantitative** methods to **test qualitative** hypotheses.

Agenda

01 Introduction [10 min]

02 **General approach [10 min]**

03 **Case studies**

- **Discover Weekly [40 min]**

-------- Break [10 min] --------

- **Search [40 min]**
- **Home [30 min]**

04 **Lessons learned [10 min]**

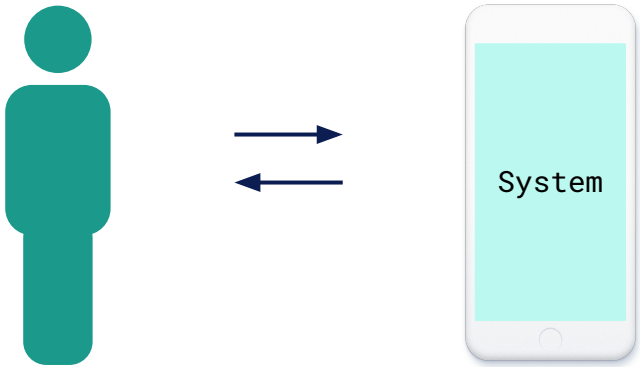
-------- Q & A [30 min] --------

General approach

How do we use mixed methods to improve metric development?

Metric development

Users interact with web and mobile applications.



We want to **measure** system performance and find ways to **improve** it through better **user understanding**.

01 Product background

Understanding the product hypotheses and assumptions about the user goals and needs.

02 Metric use

Identifying the intended use of the metric and the assumptions built into existing metrics.

03 Generating hypotheses

Conducting research to investigate existing assumptions directly and generate hypotheses about user-centric evaluation metrics.

04 Testing hypotheses

Final phase of mixed methods research to test hypotheses and report conclusions or recommend user-centric metrics.

05 Applications

Determining how to apply learnings to the product.

01 Product background

Understanding the product hypotheses and assumptions about the user goals and needs.

Existing user hypotheses

Assumptions about how users will interact with the product

Assumptions about what goals the product helps the user achieve

Solution hypotheses

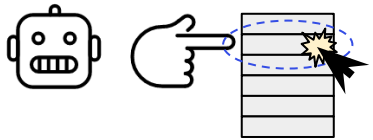
Assumptions about how the solution or system reacts to interactions and the environment

Assumptions about how different system responses will be perceived by the user

Overall product behavior

02 Metric use

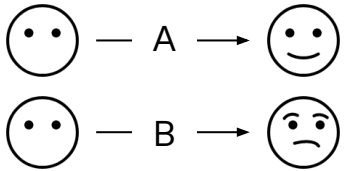
Identifying the intended use of the metric and the assumptions built into existing metrics.



Offline optimization

Methods need to be chosen with the granularity of the metric in mind.

Not all metrics are decomposable into item level



A/B test evaluation

User-centric evaluation of A/B tests is a common strategy to improve product decision making.

Sensitive, Trustworthy, Efficient, Debuggable, Interpretable (STEDI)*

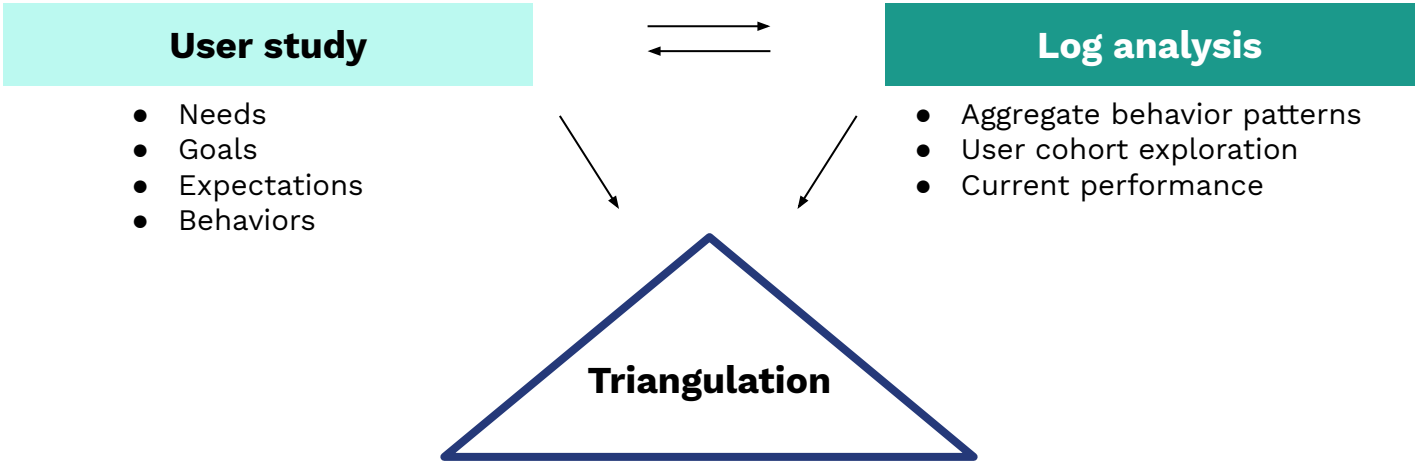
Monitoring

Long term metrics, strategic metrics, or metrics defined over data that is difficult to randomize.

*Somit Gupta et al. Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments. KDD 2019.

03 Generating hypotheses

Conducting research to investigate existing assumptions directly and generate hypotheses about user-centric evaluation metrics.



04 Testing hypotheses

Final phase of mixed methods research to test hypotheses and report conclusions or recommend user-centric metrics.

From our hypotheses, consider:

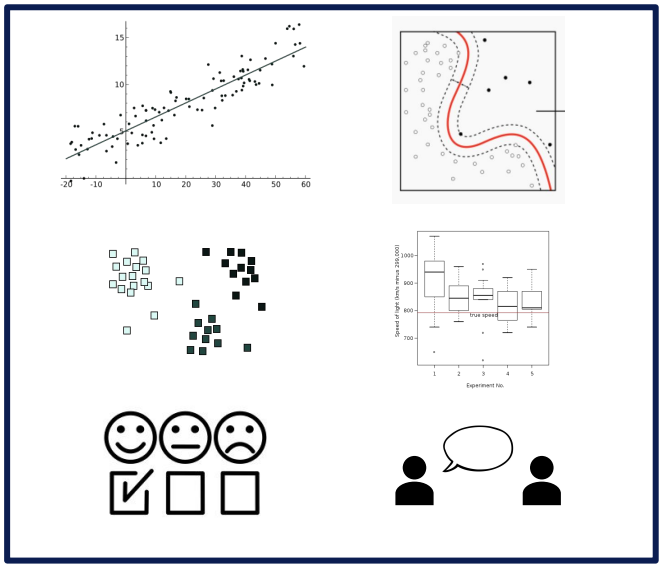
What would these hypotheses imply about the way a user generates data?

Will our data represent the user behavior we have hypotheses about?

How can we directly test the implications of these hypotheses?

Which methods are best suited for learning about our hypotheses at scale?

And then identify the most useful methods:



05 Applications

Determining how to apply learnings to the product.

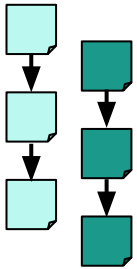
Insights from mixed methods can have implications for:

System Design

How to improve front-end, back-end, and user experience aspects of the system

Data Logging

Assumptions about navigation paths and task completion can lead to incomplete pictures



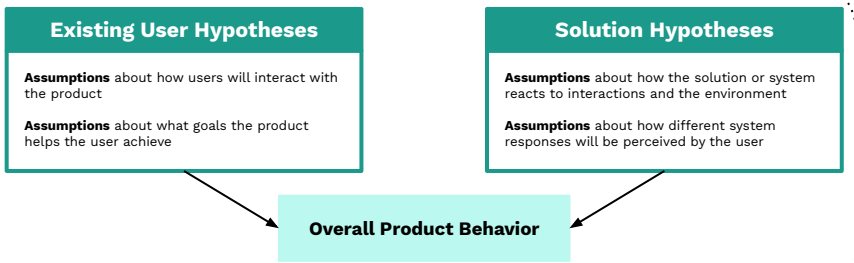
Evaluation Metrics

Developing metrics that better reflect user behavior and satisfaction

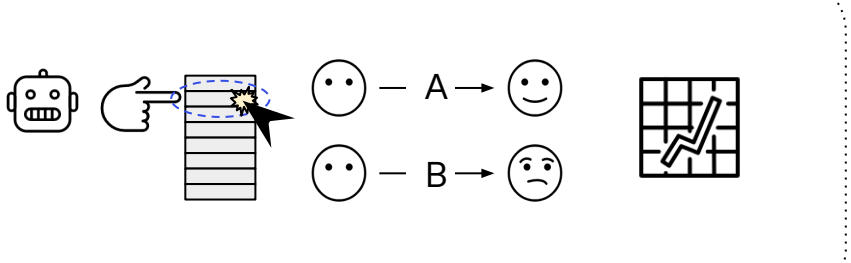
Identifying when to apply mixed methods research

Mixed methods are useful when...

... the **product hypotheses** are user-centric



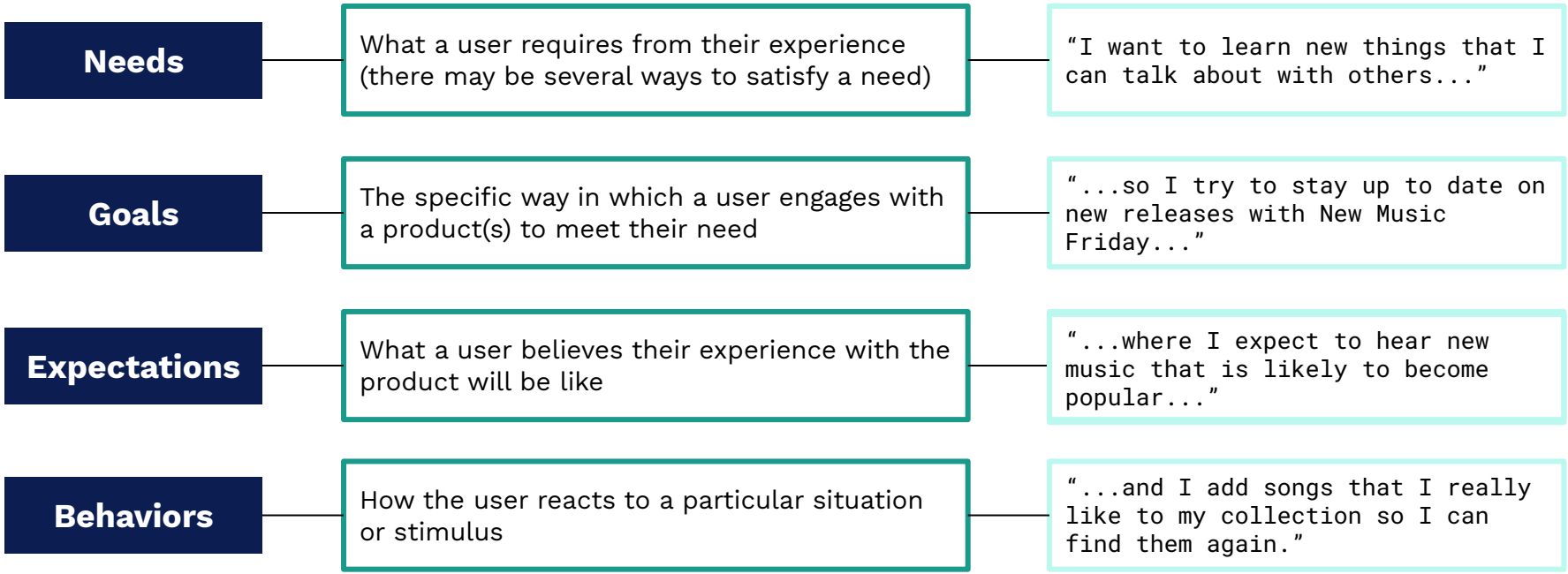
... the **intended evaluation** is user-centric



... you need to understand the extent to which business metrics **deviate** from the user experience

User-focused evaluation

The research questions in a mixed methods study can focus on understanding:



Mixed methods designs

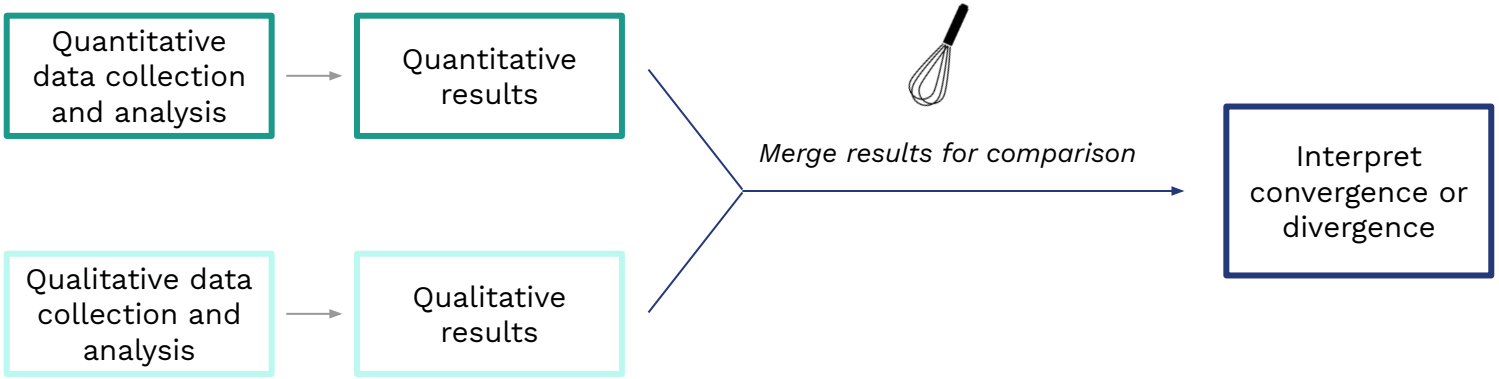
01 Convergent parallel design

02 Explanatory sequential design

03 Exploratory sequential design

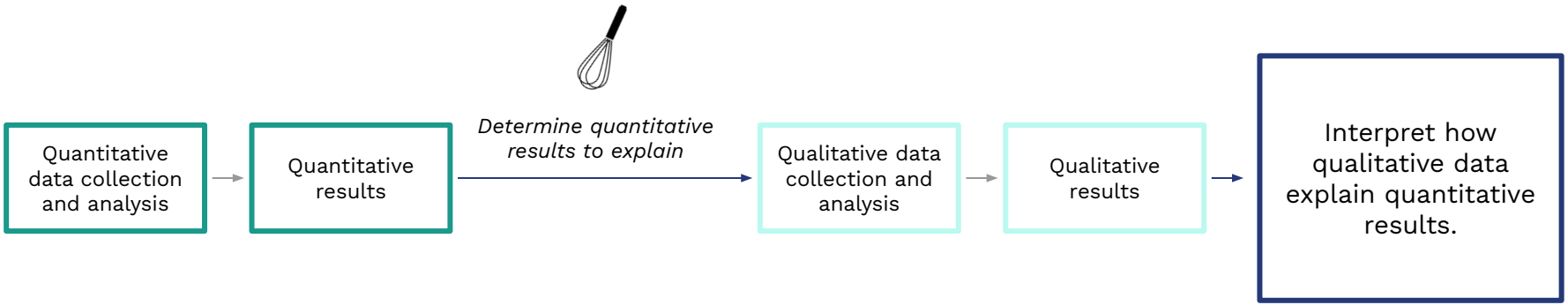
Creswell and Plano Clark, “Designing and Conducting Mixed Methods Research”, 2018.

01 Convergent parallel design



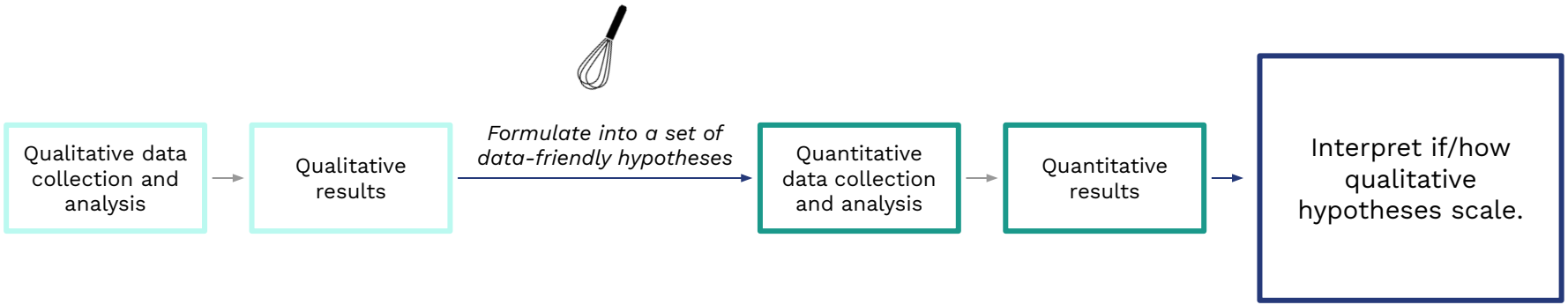
Creswell and Plano Clark, “Designing and Conducting Mixed Methods Research”, 2018.

02 Explanatory sequential design



Creswell and Plano Clark, “Designing and Conducting Mixed Methods Research”, 2018.

03 Exploratory sequential design



Creswell and Plano Clark, “Designing and Conducting Mixed Methods Research”, 2018.

03 Exploratory sequential design

Qualitative data collection and analysis → Qualitative results → *Formulate into a set of data-friendly hypotheses* → Quantitative data collection and analysis → Quantitative results → Interpret if/how qualitative hypotheses scale.

In metric development applications, we've found the exploratory sequential design useful because of how it enables and facilitates quantitative follow up.

This design is what we use throughout our case studies.

Creswell and Plano Clark, "Designing and Conducting Mixed Methods Research", 2018.

Summary of our general approach

01 Product background

Understanding the product hypotheses and assumptions about the user goals and needs.

02 Metric use

Identifying the intended use of the metric and the assumptions built into existing metrics.

03 Generating hypotheses

Conducting research to investigate existing assumptions directly and generate hypotheses about user-centric evaluation metrics.

04 Testing hypotheses

Final phase of mixed methods research to test hypotheses and report conclusions or recommend user-centric metrics.

05 Applications

Determining how to apply learnings to the product.

Case studies

How have we put these principles into practice?

For each case study, we will follow the general approach

01 Product background

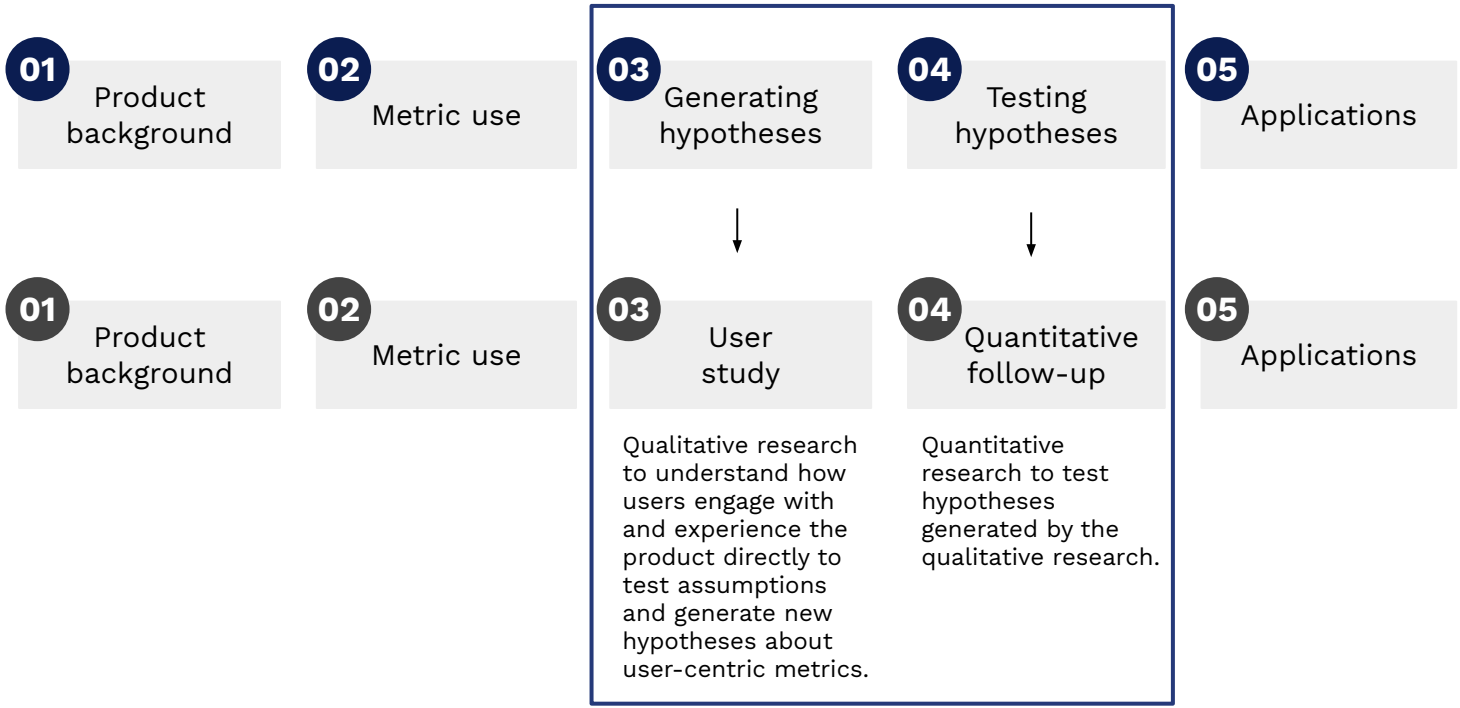
02 Metric use

03 Generating hypotheses

04 Testing hypotheses

05 Applications

But with a more specific take on steps 3 and 4 in line with the exploratory sequential design



We will cover three case studies

Within each case study, we highlight how qualitative findings shaped our approach to the subsequent quantitative analysis.

Discover Weekly

Overview

Goal: To develop metrics for Discover Weekly that align with how users engage with and experience the playlist.

01 Product background

What Discover Weekly is.

02 Metric use

How Discover Weekly is traditionally evaluated and the implicit assumptions that these metrics make.

03 User study

Research to understand how users engage with and experience Discover Weekly to test assumptions and generate new hypotheses about user-centric metrics.

04 Quantitative follow-up

The quantitative work to test hypotheses generated by the UR at scale.

05 Applications

Determining how to apply learnings to Discovery Weekly.

What is Discover Weekly?

01 Product background

02 Metric use

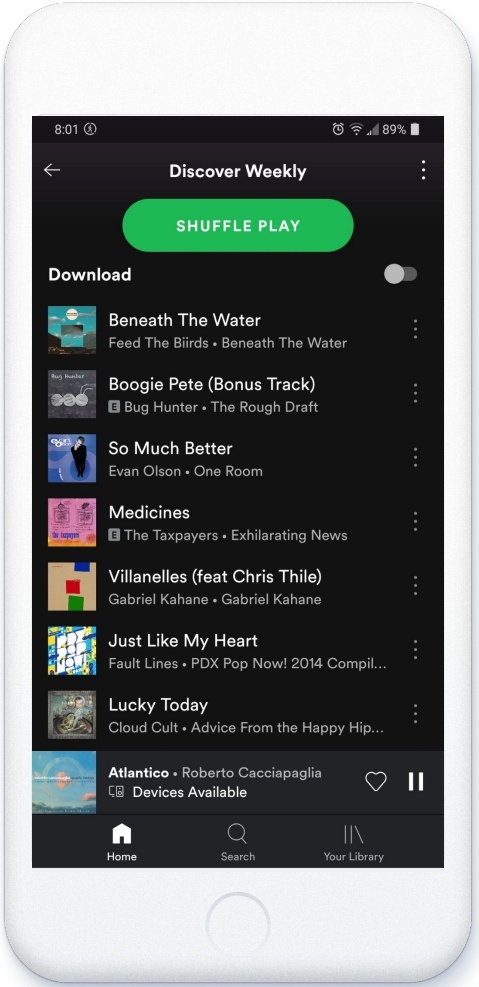
03 User study

04 Quantitative follow-up

05 Applications

What is Discover Weekly?

A playlist that is personalized with 30 tracks unfamiliar to the user every Monday.



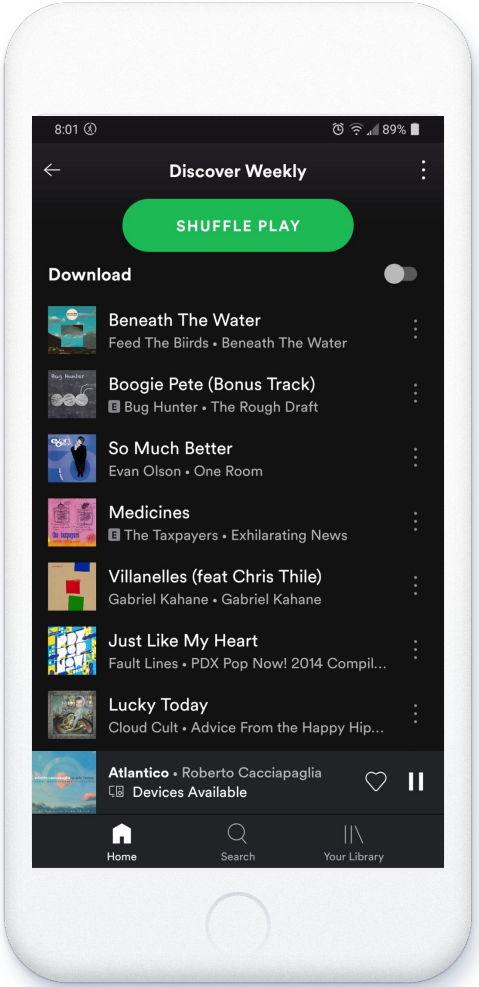
What is Discover Weekly?

From a system perspective:

Spotify populates the playlist with tracks that the user has never interacted with before. Tracks are meant to align with the user's tastes.

From a user perspective:

A playlist that provides "new" tracks with the goal of enabling effortless and delightful discovery.



How is Discover Weekly traditionally evaluated?

01 Product background

02 Metric use

03 User study

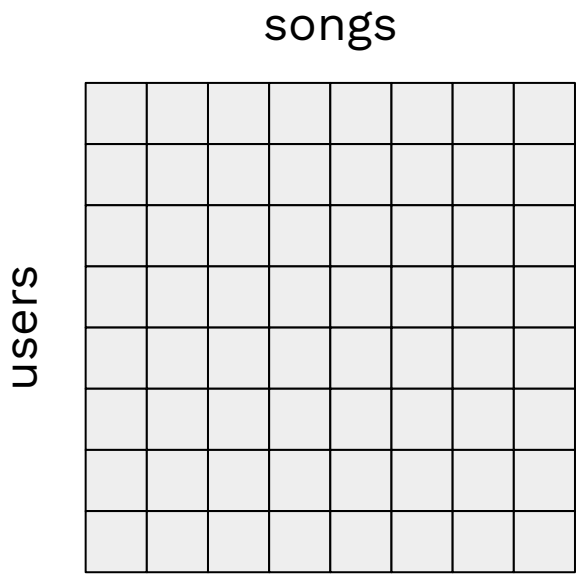
04 Quantitative follow-up

05 Applications

How is Discover Weekly traditionally evaluated?

Discovery is a recommendation problem.

- **Input**: incomplete user-song ratings matrix
- **Output**: imputed song ratings



Offline Evaluation Metrics: RMSE

Minimize error with held out ratings:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

Evaluation of recommender has long been divided between accuracy metrics (e.g., precision/recall) and error metrics (notably, RMSE and MAE). The mathematical convenience and fitness with formal optimization methods, have made error metrics like RMSE more popular, and they are indeed dominating the literature. However, it is well recognized that accuracy measures may be a more natural yardstick, as they directly assess the quality of top-N recommendations.

This work shows, through an extensive empirical study, that the convenient assumption that an error metric such as RMSE can serve as good proxy for top-N accuracy is questionable at best. There is no monotonic relation between error metrics and accuracy metrics. This may call for a re-evaluation of optimization goals for top-N systems. On the bright side we have presented simple and efficient variants of known algorithms, which are useless in RMSE terms, and yet deliver superior results when pursuing top-N accuracy.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. RecSys, 2010.

Offline Evaluation Metrics: Ranking Metrics

Assuming...

- browsing model for the user (e.g. sequential traversal of a ranked list)
- utility model for the user (e.g. relationship between rating and utility)

Simulate the user and compute expected utility

- normalized discounted cumulative gain (NDCG), expected reciprocal rank (ERR), rank-biased precision (RBP)

⚠️ Different ranking metrics make different assumptions about user behavior which may or may not be appropriate for your task.

More information: P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 2", 2020.

Production systems

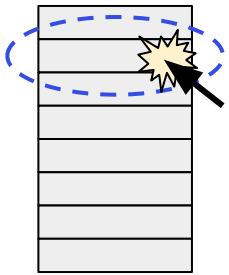
Item utility is contextual: utility can depend on the user's context

Partial feedback: not every item is rated

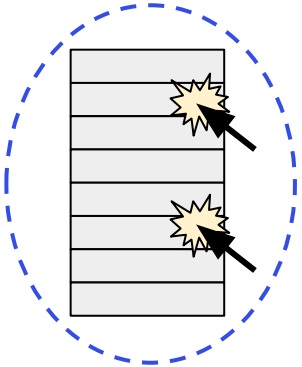
Implicit utility: users don't tell you their utility

Online evaluation uses logged data to determine system quality

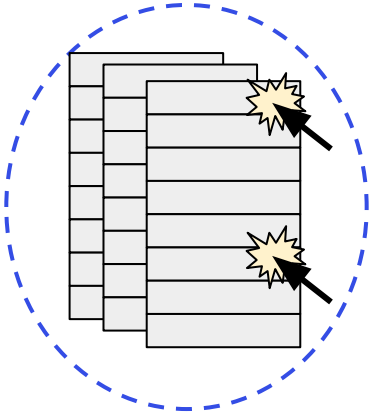
Utility Granularity



Item utility



Slate utility



Policy utility

Online Evaluation Metrics: Click-through Rate

Click on an item indicates the user found that system decision useful

Item utility is equal to the an item click (e.g. click: +1, skip: -1)

Slate utility is proportional to the number of clicks on items in slate

Policy utility is proportional to the number of clicks on items over multiple slates

Online Evaluation Metrics: Consumption Time

Consumption time is proportional to user utility

Item utility is proportional to the time the user spent consuming an item

Slate utility is proportional to the time the user spent consuming items in slate

Policy utility is proportional to the time the user spent consuming items over multiple slates

Online Evaluation Metrics: Retention

Cumulative user utility observed in monthly retention

Item utility is unobserved but contributes to retention

Slate utility is unobserved but contributes to retention

Policy utility is reflected in retention

Online Evaluation Metrics: Behavioral Metrics

Behavioral evaluation metrics are powerful, but to using them effectively is a modeling exercise on its own with short-but-rich history (e.g. new behavioral signals, position bias correction).

⚠️ Different online signals make different assumptions about user behavior which may or may not be appropriate for your system.

More information: P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 3", 2020.

Discover Weekly relied on key business metrics

Satisfaction proxy metrics

Reach
The share of Spotify users who use Discover Weekly in a given week

Depth
The share of this week's DW users who listened to a minimum threshold of time

WoW Retention
The share of last week's DW users who returned this week

These metrics have underlying assumptions

Satisfaction proxy metrics

Reach
The share of Spotify users who use Discover Weekly in a given week

Depth
The share of this week's DW users who listened to a minimum threshold of time

WoW Retention
The share of last week's DW users who returned this week

Assumption 1: the number of Spotify users who listen to DW is an indication of per user happiness.

Assumption 2: listening to DW for a longer amount of time is better than listening for a shorter amount of time.

Assumption 3: coming back next week depends on satisfaction with this week's experience.

What did we learn through user research?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

User study goals

01 Investigate underlying assumptions of existing Discovery Weekly evaluation.

02 Develop hypotheses around metrics that capture the user’s experience with Discover Weekly.

Investigating metric assumptions through user research

Reach
The share of Spotify users who use Discover Weekly in a given week

Depth
The share of this week's DW users who listened to a minimum threshold of time

WoW Retention
The share of last week's DW users who returned this week

How can we understand these metrics from the user's perspective?

We can't. This metric is meaningless on a per user basis.

What do longer versus shorter listening sessions signal?

How do users decide whether or not to return to Discover Weekly?

Study design

We conducted semi-structured interviews with 10 participants across 3 cohorts.

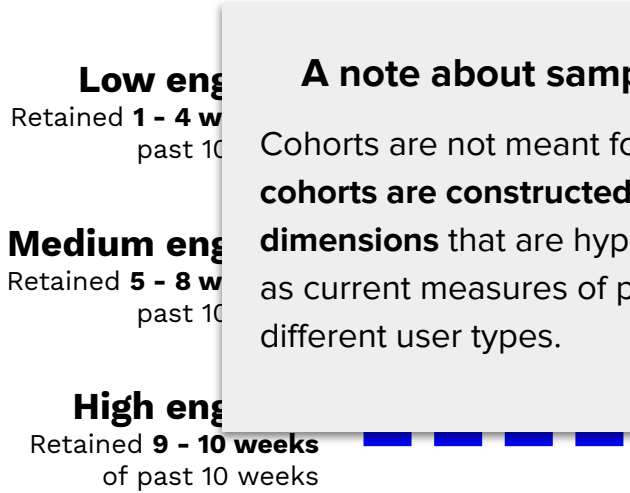
Low engaged
Retained **1 - 4 weeks** of past 10 weeks

Medium engaged
Retained **5 - 8 weeks** of past 10 weeks

High engaged
Retained **9 - 10 weeks** of past 10 weeks

Study design

We conducted semi-structured interviews with 10 participants across 3 cohorts.



A note about sampling for qualitative studies

Cohorts are not meant for direct comparison. Rather, **cohorts are constructed to be diverse along key dimensions** that are hypothesized to be important - such as current measures of performance and engagement or different user types.

Study design

We conducted semi-structured interviews with 10 participants across 3 cohorts.

Low engaged
Retained **1 - 4 weeks** of past 10 weeks

Medium engaged
Retained **5 - 8 weeks** of past 10 weeks

High engaged
Retained **9 - 10 weeks** of past 10 weeks

Interviews covered:

- Music preferences
- Spotify usage
- Discover Weekly
  - Attitudes
  - Good and bad experiences
  - Usage deep dive

Research questions

1. **Why** do users listen to Discover Weekly?
2. **How** do users listen to Discover Weekly?
3. How do users **evaluate** their Discover Weekly experience (e.g., what is a good vs. bad experience)?

Why do users listen to Discover Weekly?

All participants had discovery in mind but intended to experience it in different ways.

Goals in Discover Weekly

Play new **background** music	**Listen** to new music **now and later**	**Find** new music for **later**	**Engage** with new music
• Music that doesn't distract while working • Music to keep the user motivated while working out	• Music for when the user can half pay attention for passive discovery / curation	• Music to quickly sort through to find new music to add to their collection	• Music that will guide users to a new artist, album, or genre

Goals in Discover Weekly

Experience goals
Most listening occurs in DW

Play new **background** music	**Listen** to new music **now and later**	**Find** new music for **later**	**Engage** with new music
• Music that doesn't distract while working • Music to keep the user motivated while working out	• Music for when the user can half pay attention for passive discovery / curation	• Music to quickly sort through to find new music to add to their collection	• Music that will guide users to a new artist, album, or genre

Goals in Discover Weekly

Collection goals
Most listening occurs downstream

Experience goals
Most listening occurs in DW

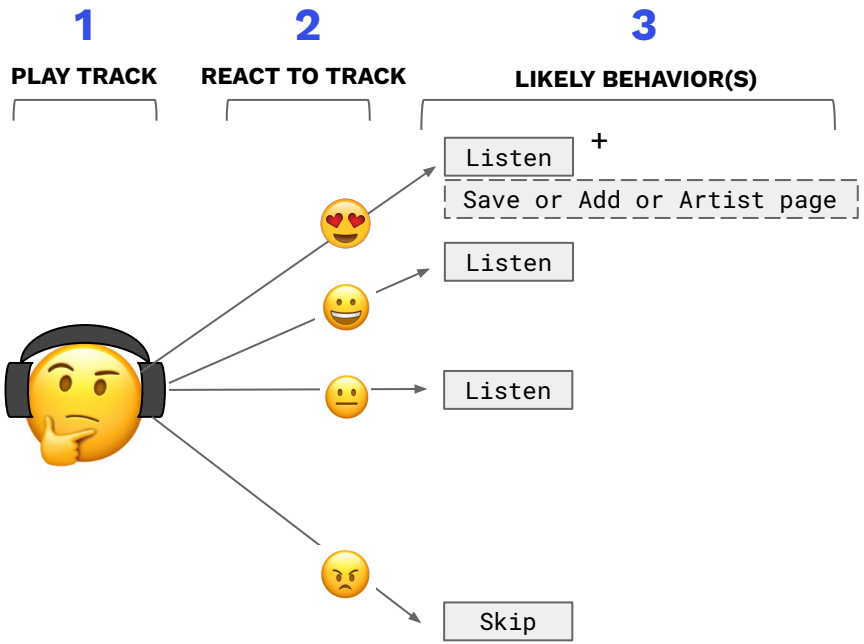
Play new **background** music	**Listen** to new music **now and later**	**Find** new music for **later**	**Engage** with new music
• Music that doesn't distract while working • Music to keep the user motivated while working out	• Music for when the user can half pay attention for passive discovery / curation	• Music to quickly sort through to find new music to add to their collection	• Music that will guide users to a new artist, album, or genre

How do users listen to Discover Weekly?

Behavioral patterns varied by the user’s goal.

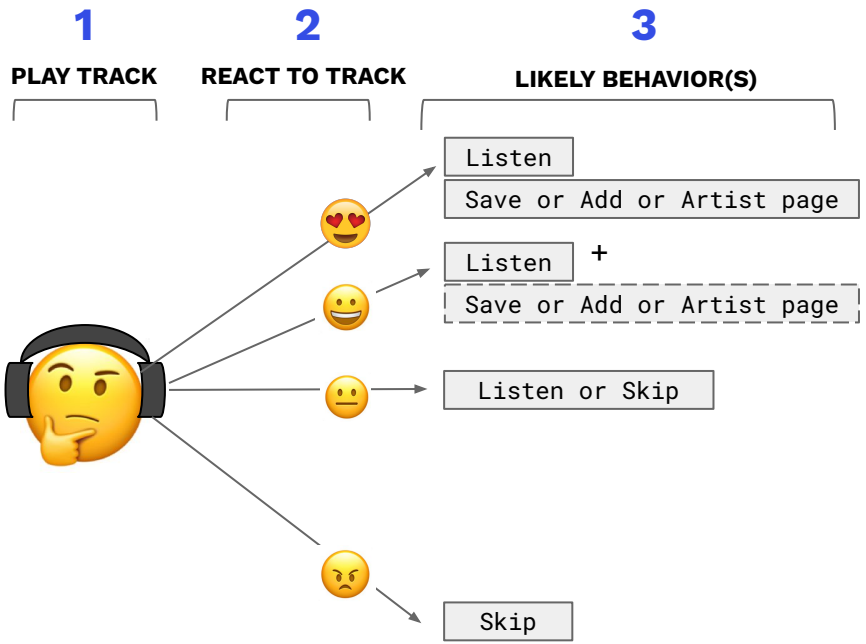
User behavior by goal

Play new
background music



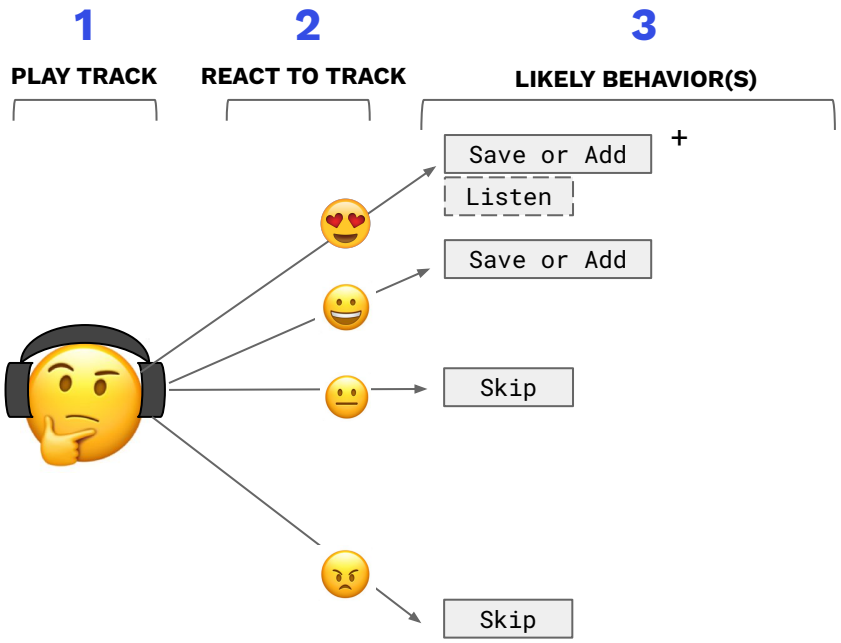
User behavior by goal

Listen to new music
now and later



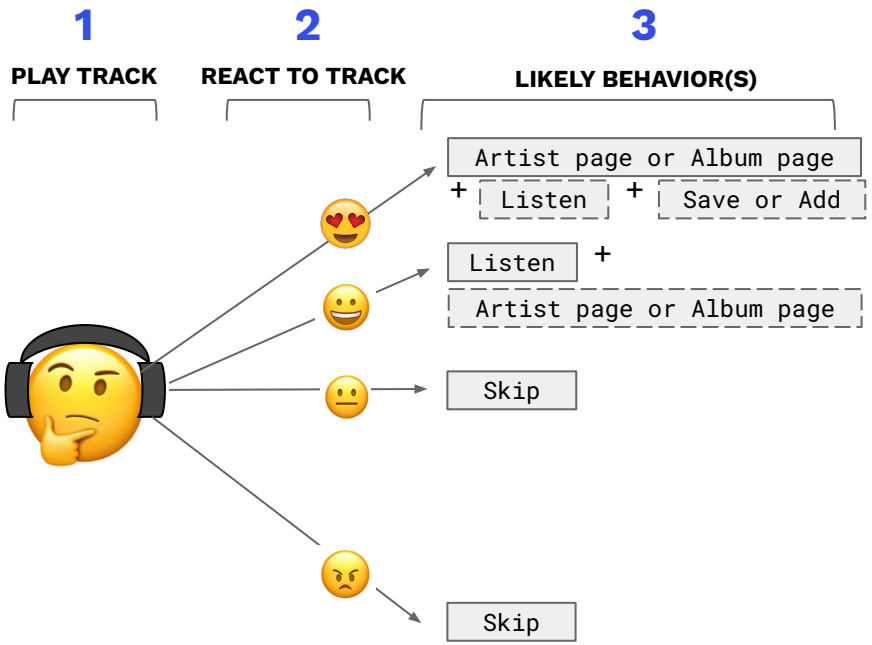
User behavior by goal

Find new music for **later**



User behavior by goal

Engage with new music



Listening patterns across time were user specific

Habitual users tend to return weekly

*"Usually **Monday** mornings I'll open it."*

A high-engaged user, 41

Some users need to be in the right mood

*"When there's **a stretch of time when I'm really in the mood to discover music and make playlists**, then I might be more invested and interested and I can't wait for Sunday and get my Discover Weekly."*

A medium-engaged user, 30

Some users return to check for improvement

*"If the first two songs on Discover Weekly aren't that great, I'll be like, 'This is a **tough week** for the algorithm or whatever.' … I **might stop and listen to normal stuff and then see next week.**"*

A low-engaged user, 21

Listening patterns across time were user specific

Habitual users tend to return weekly

*"Usually **Monday** mornings I'll open it."*

A high-engaged user, 41

Some users need to be in the right mood

Notice returning to Discover Weekly **often happens for reasons beyond the quality of recent experience**.

interested and I can't wait for Sunday and get my Discover Weekly."

A medium-engaged user, 30

Some users return to check for improvement

*"If the first two songs on Discover Weekly aren't that great, I'll be like, 'This is **a tough week** for the algorithm or whatever.' … I **might stop and listen to normal stuff and then see next week.**"*

A low-engaged user, 21

How do users evaluate Discover Weekly?

Participants relied on goal success and past experiences, often putting a lot of weight on a single track.

Satisfaction and dissatisfaction **differ by goal**

	Play new **background** music	**Listen** to new music **now and later**	**Find** new music for **later**	**Engage** with new music
😀	No skipping ↑ Saves or adds ↑ Listening time ↑ Sessions per week	↑ Saves or adds ↑ % tracks heard by EOW ↑ Streams over half the song ↑ Downstream listening	↑ Saves or adds ↑ Streams ↑ Downstream listening	↑ Artist page views ↑ Album page views ↑ Downstream listening
😠	↑ Skips (especially of the same track) ↑ Switch to different feature mid-session ↓ Listening time ↓ Sessions per week	Abandon for the week after one incomplete session ↓ Saves or adds ↑ Skips (without add/save) before halfway point ↓ Downstream listening	↓ Saves or adds ↓ Streams ↓ Downstream listening	↓ Artist page views ↓ Album page views ↓ Downstream listening

Satisfaction and dissatisfaction **depend on the past**

Play new **background** music	**Listen** to new music **now and later**	**Find** new music for **later**	**Engage** with new music
			
No skipping ↑ Saves or adds ↑ Listening time ↑ Sessions per week			↑ Artist page views ↑ Album page views ↑ Downstream listening
			
↑ Skips (especially of the same track) ↑ Switch to different feature mid-session ↓ Listening time ↓ Sessions per week	Abandon for the week after one incomplete session ↓ Saves or adds ↑ Skips (without add/save) before halfway point ↓ Downstream listening	↓ Saves or adds ↓ Streams ↓ Downstream listening	↓ Artist page views ↓ Album page views ↓ Downstream listening

Arrows are relative to past behavior
e.g., 5 saves is great for a user who typically saves 1 track and awful for a user who typically saves 15 tracks

Satisfaction and dissatisfaction can be **heavily swayed by a single track or artist**

😍 A great experience

"And every once in awhile I'll hear a song that is just my willing, or might be something that I've been holding onto, or that I've heard recently somewhere else, but ***I'm just like this particular song is amazing and I want to build an entire playlist around it.***"

A medium-engaged user, 30

😠 A terrible experience

"If there's ***one song and it really bothers me****, like maybe it has a really annoying melody or something like that, that* ***throws it off****, especially because I can't remove it, so it's just frustrating and that puts me off."*

A high-engaged user, 26

Let’s revisit our goals

01 Investigate underlying **assumptions** of existing Discovery Weekly evaluation.

02 Develop **hypotheses** around metrics that capture the user’s experience with Discover Weekly.

How do the metric **assumptions** hold up?

Satisfaction proxy metrics

Reach
The share of Spotify users who use Discover Weekly in a given week

Assumption 1: the number of Spotify users who listen to DW is an indication of per user happiness.

X This information is not even accessible to users.

Depth
The share of this week's DW users who listened to a minimum threshold of time

Assumption 2: listening to DW for a longer amount of time is better than listening for a shorter amount of time.

X Success with collection goals happens outside Discover Weekly.

WoW Retention
The share of last week's DW users who returned this week

Assumption 3: coming back next week depends on satisfaction with this week's experience.

X Users often return to DW out of habit or just being in the mood for discovery.

What **hypotheses** should we test?

Goals hypothesis

Behaviors provide clearer signals within the context of users' goals.

Past behavior hypothesis

Metrics should be normalized relative to each user's typical behavior.

Favorite hypothesis

One track recommendation can have a large effect on satisfaction.

For more information on how themes arise from qualitative research: J. Garcia-Gathright, C. Hosey, B. St. Thomas, B. Carterette, F. Diaz. "Mixed Methods for Evaluating User Satisfaction." Presented at: RecSys 2018.

How did we test UR hypotheses at scale?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Hypotheses to test

Goals hypothesis

Behaviors provide clearer signals within the context of users' goals.

↳ Can we find evidence of goals in the log data?

Do goals help us predict satisfaction?

Past behavior hypothesis

Metrics should be normalized relative to each user's typical behavior.

↳ Does knowing previous user behavior with Discover Weekly help us predict satisfaction?

Favorite hypothesis

One track recommendation can have a large effect on satisfaction.

↳ Do stand-out tracks within a single week help us predict satisfaction?

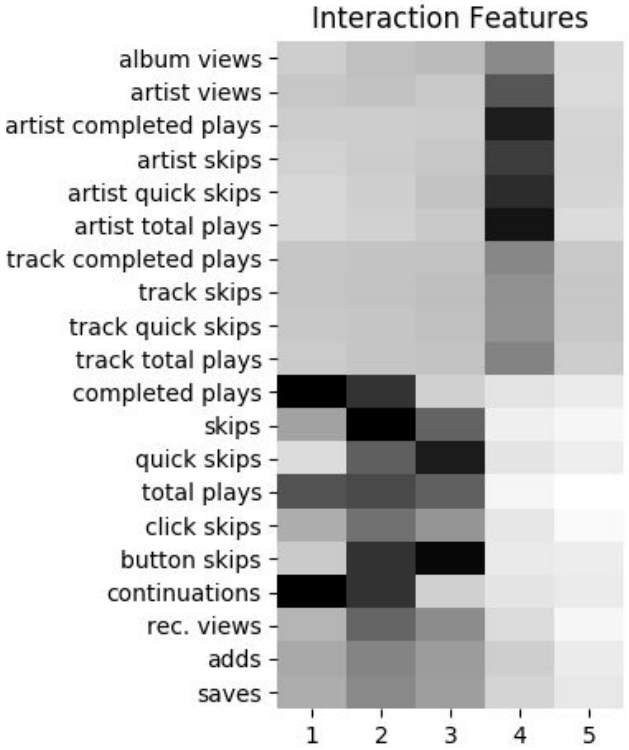
Goals hypothesis

Can we find evidence of goals in data?

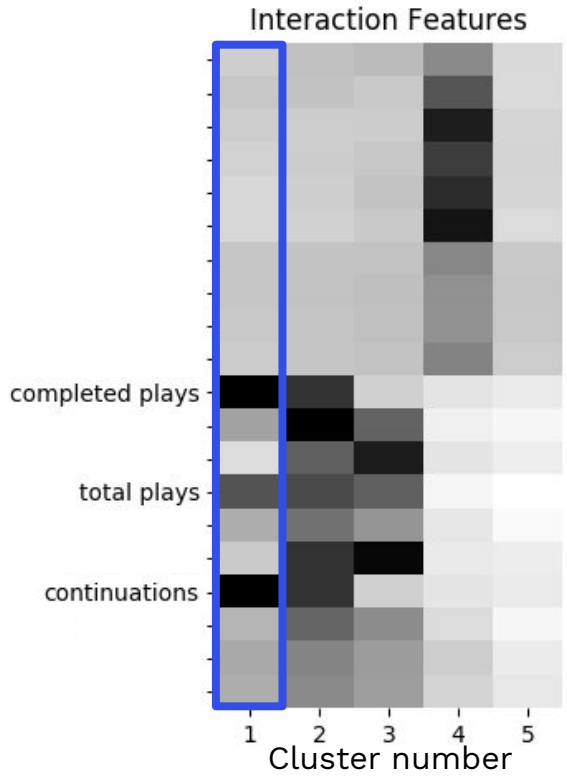
Method

We used **K-means clustering & elbow test** to look for patterns of behavior that aligned with the goals identified in the user study.

- Set of interactions was informed by user interviews
- Darker cells indicate stronger signal in that interaction
- Cluster interpretation was done manually



Cluster 1



Cluster behavior pattern

Plays, no skips

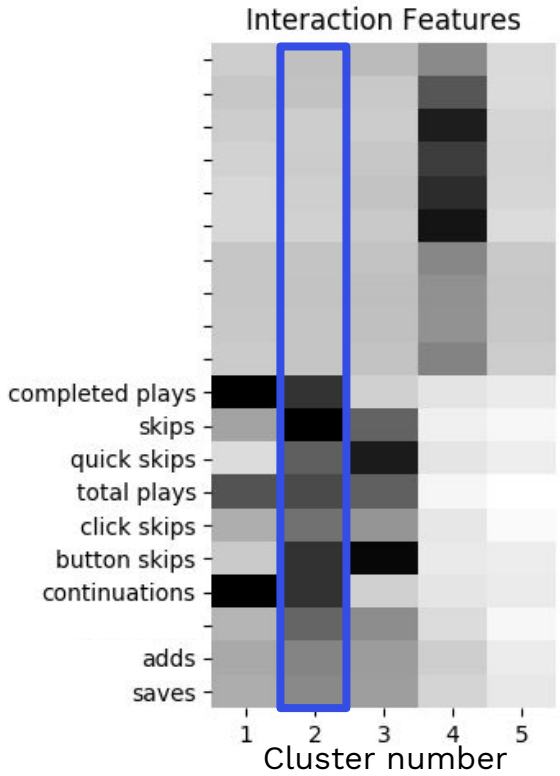
Goal alignment

Play new music in the background

Cluster size

16 %

Cluster 2



Cluster behavior pattern

Plays, skips, saves

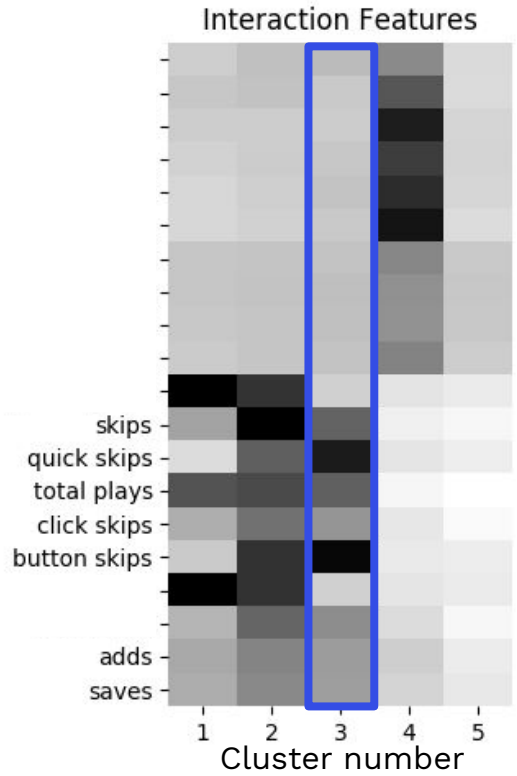
Goal alignment

Listen to new music now and later

Cluster size

15 %

Cluster 3



Cluster behavior pattern

Skips, saves

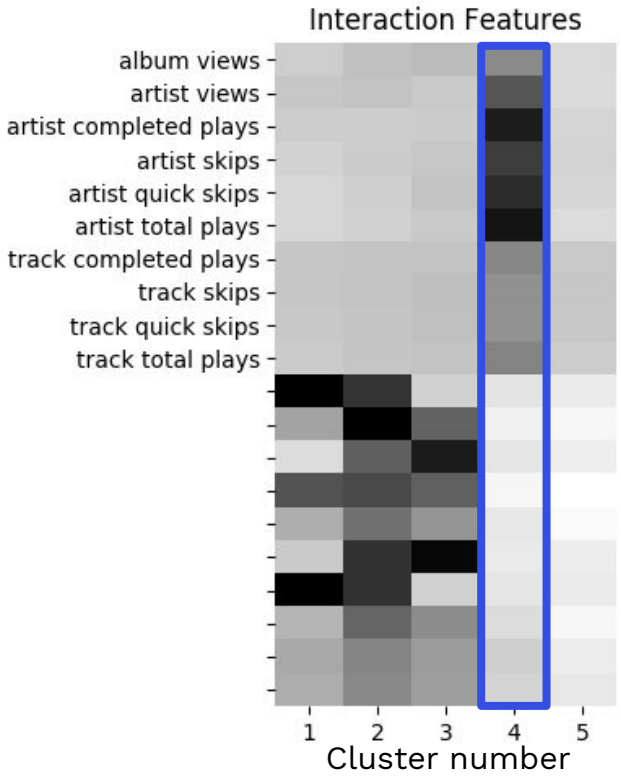
Goal alignment

Find new music for later

Cluster size

11 %

Cluster 4



Cluster behavior pattern

Play from artist page and beyond

Goal alignment

Engage with new music

Cluster size

2.5 %

Goals, experience, and favorite hypotheses

Does knowing a user’s goal, past behavior, and peak experience help us predict satisfaction?

Method

01 Collect ground truth satisfaction data

02 Include features that represent UR hypotheses

03 Train a supervised model to predict satisfaction

01 Collect ground truth satisfaction data

We sent an end-of-week **satisfaction survey** via email

- In interviews, users were highly expressive about their feelings toward their recommendations, so we judged the survey responses would be meaningful

OVERALL SATISFACTION
+ drivers:
- usability
- usefulness
- delight

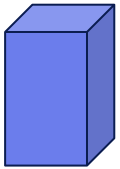
THIS WEEK SATISFACTION
+ drivers:
- goal achievement
- transparency
- fit
- track love
- track annoyance

USER GOAL RANK
- listen right now
- background music
- fits a specific activity
- save for later
- artist exploration
- genre exploration

PERSONAL PREFERENCES FOR DISCOVERY
- cohesion
- ‘ambitiousness’ of recommendations
- effort

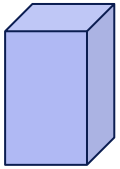
02 Include features that represent UR hypotheses

Favorite hypothesis →



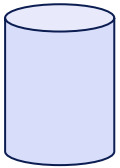
This Week's Data (User interactions with the playlist)

Past behavior hypothesis →



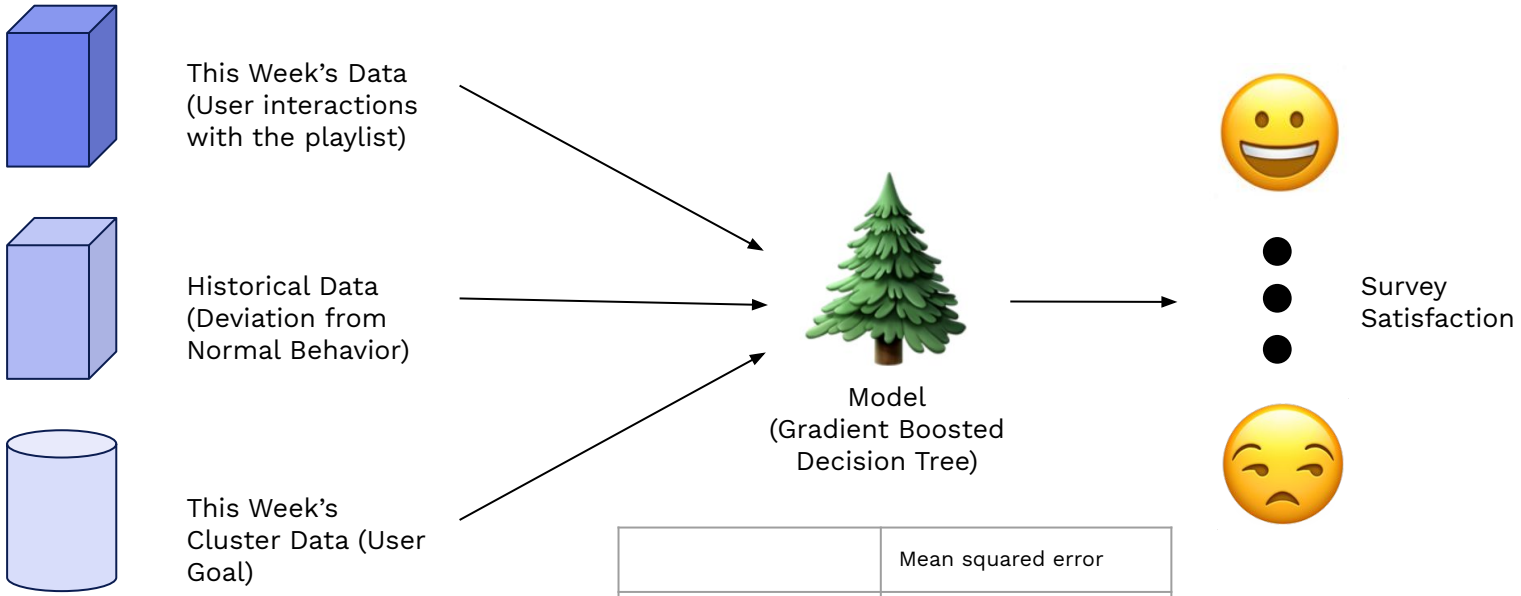
Historical Data (Deviation from Normal Behavior)

Goals hypothesis →



This Week's Cluster Data (User Goal)

03 Train a supervised model to estimate satisfaction



	Mean squared error
Intercept model	0.95
Final model	**0.28**

Findings

Model weights were consistent with hypotheses from user study

- Representing user **goals** (clusters) added the most information about satisfaction
- Taking **past patterns of behavior** into account (normalized) helped give context for the user experiencing success in their goal
- **Peak experiences** (max engagement over tracks) offered more information than total usage (sum engagement over tracks)



	Cluster	**Normalized**	**Max**	**Sum**
Gain (%)	70.6	15.4	5.6	2.3
Weight (%)	54.1	30.7	7.5	1.0

How did we apply the learnings to Discover Weekly?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Changes to Discover Weekly

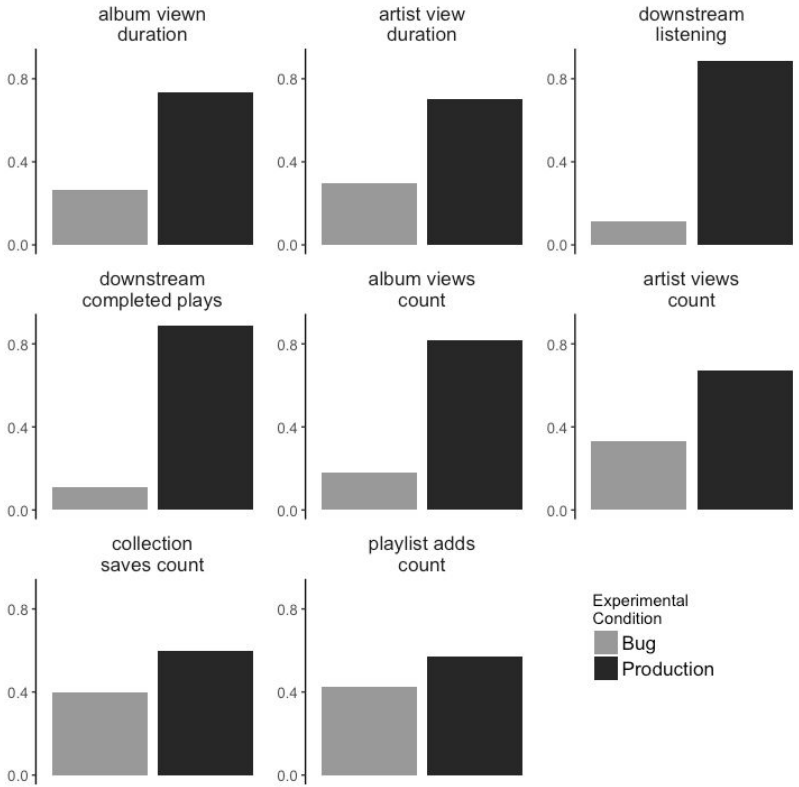
01 Model changes

Used cluster / normalized features to train new DW model

02 Metric changes

Incorporated downstream metrics into a/b test evaluation

Metric validation



We validated directionality and sensitivity of raw and normalized metrics in A/A and A/B tests

Break!

[10 Minutes]

Search

Overview

Goal: To develop metrics for Search that align with how users engage with and experience Search on Spotify.

01 Product background

What Search is.

02 Metric use

How Search is traditionally evaluated and the implicit assumptions that these metrics make.

03 User study

Research to understand how users engage with and experience Search directly to test assumptions and generate new hypotheses about user-centric metrics.

04 Quantitative follow-up

The quantitative work to test hypotheses generated by the UR at scale.

05 Applications

Determining how to apply learnings to Search.

What is Search?

01 Product background

02 Metric use

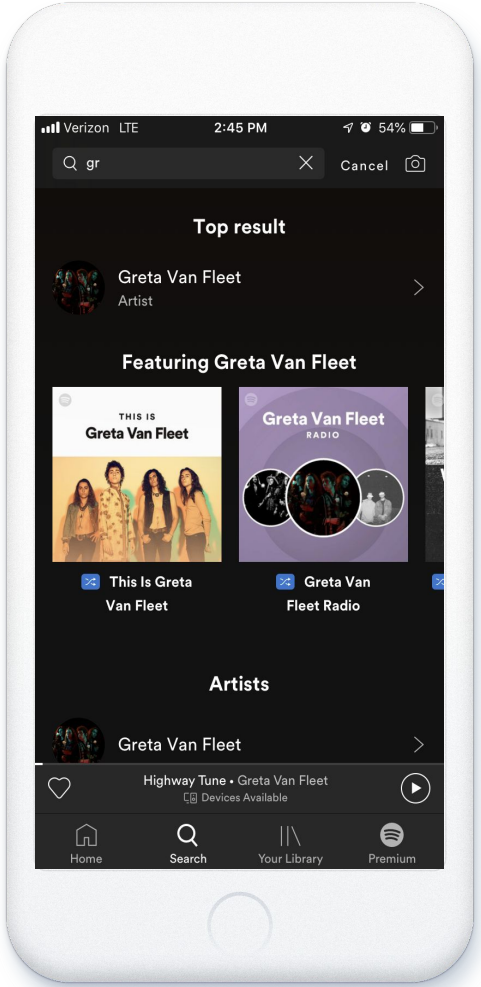
03 User study

04 Quantitative follow-up

05 Applications

What is Search?

A tool that enables users to sift through our catalog to find relevant audio content.



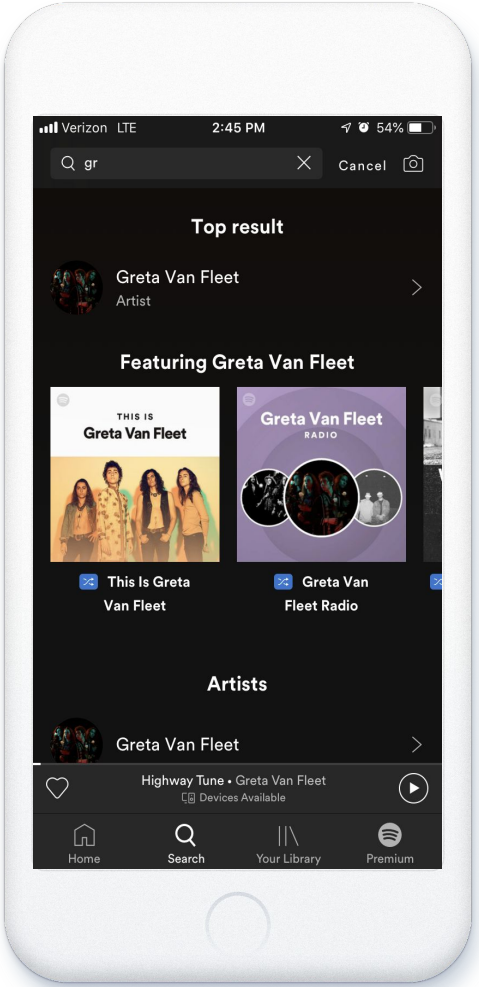
What is Search?

From a system perspective:

Spotify's search system uses *instant search* where results update with every keystroke.

From a user perspective:

A tool that requires *direct user input* to surface audio content that *meets a particular need* in that moment.



How is Search traditionally evaluated?

01 Product background

02 Metric use

03 User study

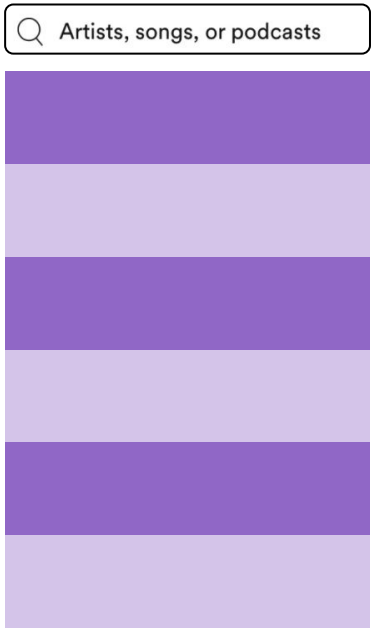
04 Quantitative follow-up

05 Applications

How is Search traditionally evaluated

User examines the list of items returned by the search engine for a given query.

They accumulate **utility** by traversing the list continuing until rank ***k***.

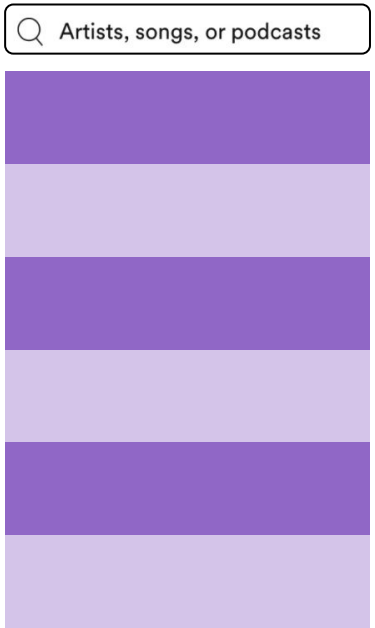


How is Search traditionally evaluated

User examines the list of items returned by the search engine for a given query.

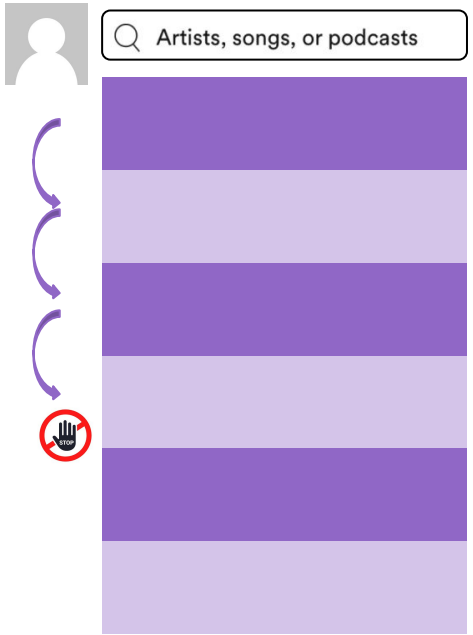
They accumulate **utility** by traversing the list continuing until rank ***k***.

- Browsing model
- Utility
- Utility accumulation



Browsing model

Common browsing model → user traverses ranked list one by one, stopping at rank k.



Utility model

Signals for estimating utility of an item in the ranked list

Implicit Signals

- Clicks
- Dwell time
- Eye-tracking
- Cursor movements
- Gestures on mobile
- Streams & skip behavior
- Bookmarks, saves, & shares

Explicit Signals

- Human annotations
- Ratings

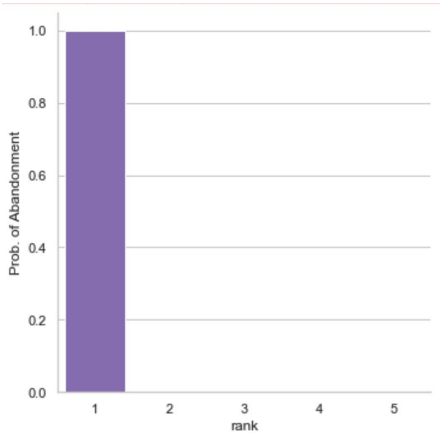
⚠️ Different signals make different assumptions about user behavior which may or may not be appropriate for your system.

More information: P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 3", 2020.

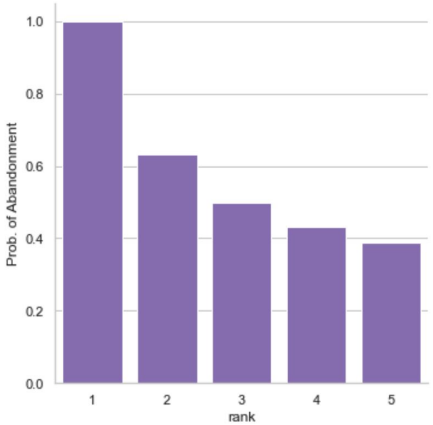
Utility accumulation model

Utility obtained from individual item can be aggregated in different ways:

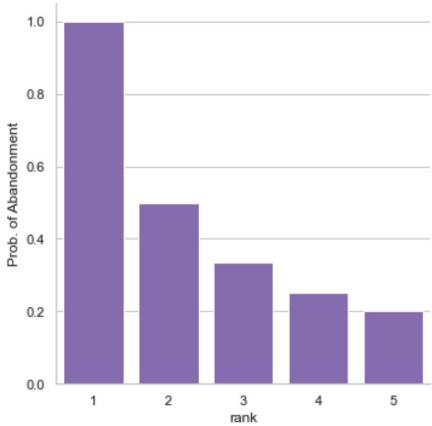
Top click



nDCG



Reciprocal rank



More information: *P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 2", 2020.*

Ranking metrics

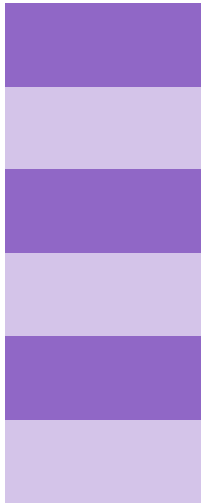
Browsing Model | **Utility Accumulation** | **Utility Model**

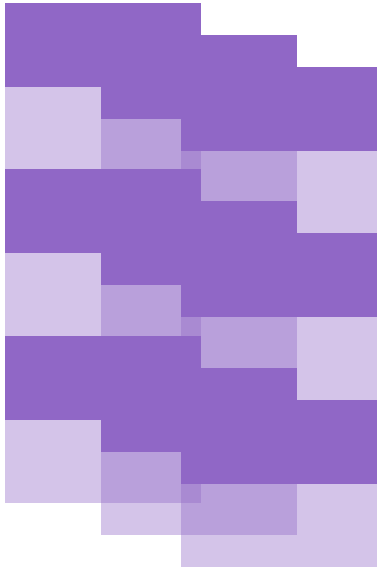
- Top click
- Precision @ K
- Reciprocal rank
- nDCG
- Expected search length

More information: *P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 2", 2020.*

Unit of measurement

Metrics can be applied at the ranked list or across a set of ranked lists.





Common Practice

Satisfaction proxy metrics

`Click-through rate`

`Top click (y/n)`

Which are applied to individual “search sessions”

`Start: first keystroke`

`End: click or 10 minutes`

These metrics have underlying assumptions

Satisfaction proxy metrics

`Click-through rate`

Assumption 1: seeing the search results page provides sufficient information for users to evaluate the results.

`Top click (y/n)`

Assumption 2: having the “correct” result at the top of the results page is important to users.

Which are applied to individual “search sessions”

`Start: first keystroke`

Assumption 3: users perceive their search to begin once they start typing.

`End: click or 10 minutes`

Assumption 4: users perceive their search to be complete once they have found a result worthy of clicking.

What did we learn through user research?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

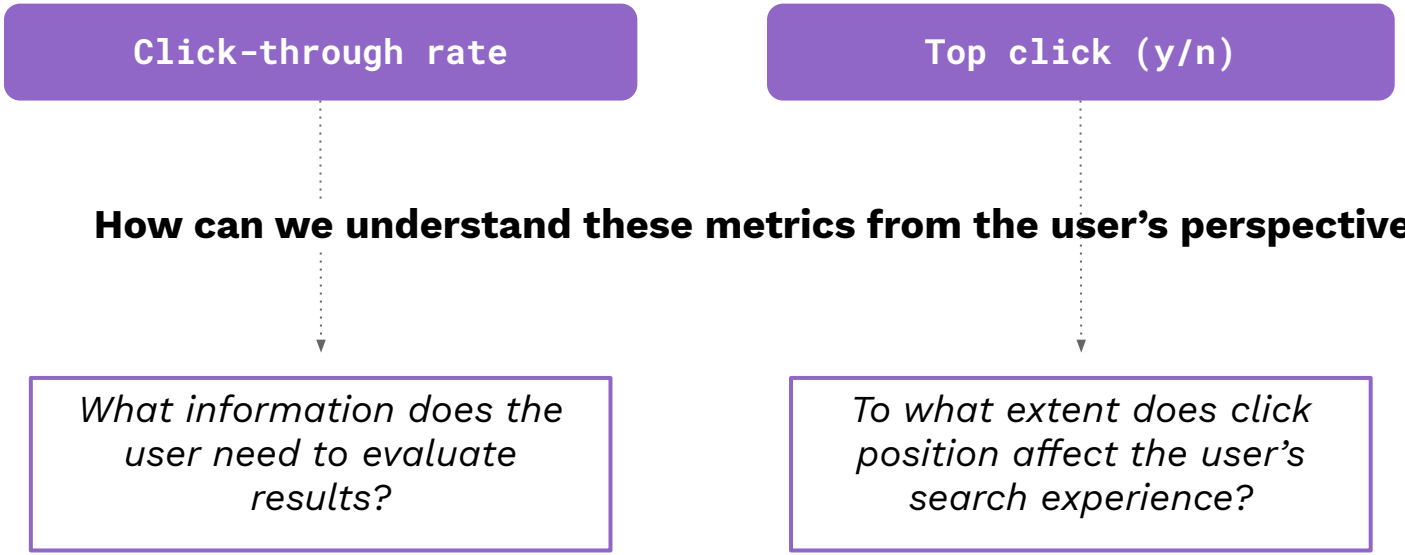
User study goals

01 Investigate underlying assumptions of existing Search evaluation.

02 Develop hypotheses around metrics that capture the user's experience with Search.

Investigating metric assumptions through user research

Satisfaction proxy metrics



Investigating metric assumptions through user research

Which are applied to individual “search sessions”

`Start: first keystroke`

`End: click or 10 minutes`

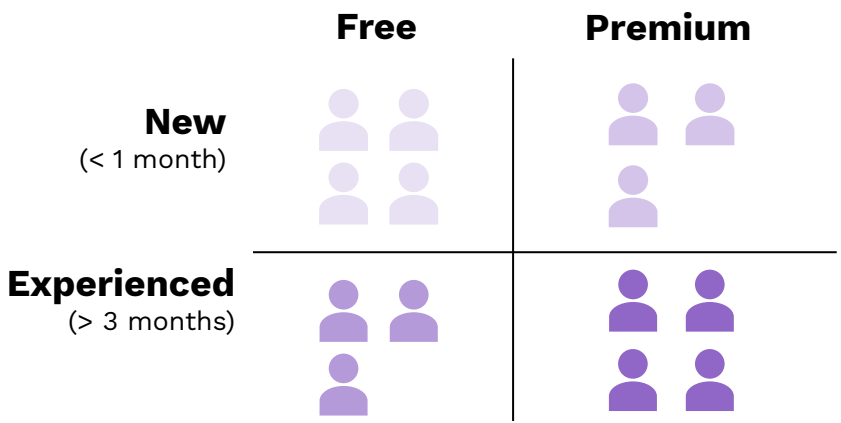
How can we understand these metrics from the user’s perspective?

When does the user perceive the search experience to begin?

Does a click end the search experience?

Study design

We conducted semi-structured interviews with 14 participants across 4 cohorts built around two dimensions: platform (free vs. premium) & account age (< 1 month vs. > 3 months).



Interviews covered:

- Music preferences
- Spotify usage
- Search
  - Attitudes
  - Good and bad experiences
  - Usage deep dive

Across cohorts, we ensured participants represented a range of search behaviors:
search frequency, average search session duration, average number of query reformulations

Research questions

1. **Why** do users Search?
2. **How** do users Search?
3. How do users **evaluate** their Search experience (e.g., what is a good vs. bad experience)?

Why do users search?

Participants had 4 goals: listen, organize, share, and fact check.

Goals in search

Participants goals were about Spotify generally, not Search specifically. Rather, Search was seen as a means to achieve more general goals.

LISTEN Have a listening session	**ORGANIZE** Curate for future listening	**SHARE** Connect with friends	**FACT CHECK** Find specific information
• Play background music • Hear a song stuck in your head	• Make a playlist • Build library	• Send music to a friend • Follow a friend	• Check own knowledge • Learn about concerts

*Most Common**

*Least Common**

**Based on the user interviews*

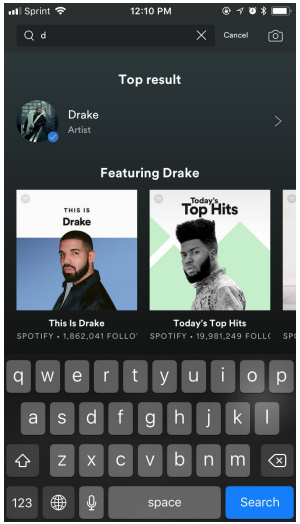
How do users search?

The user journey within Search includes 3 phases: type, consider, decide.

Overview of the user’s journey

TYPE

User communicates with us



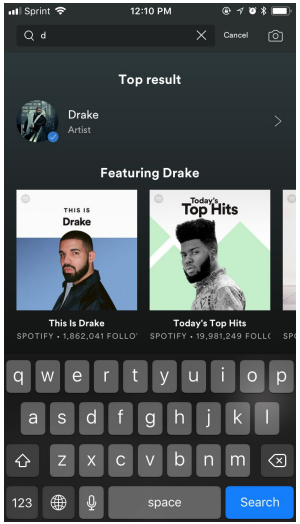
Overview of the user's journey

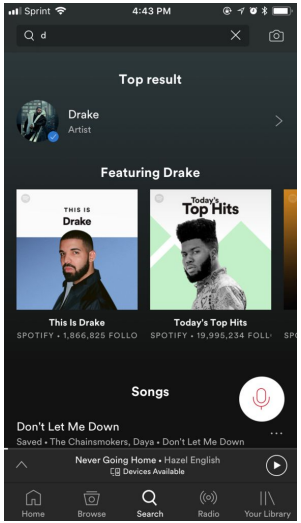
TYPE

User communicates with us

CONSIDER

User evaluates what we show them



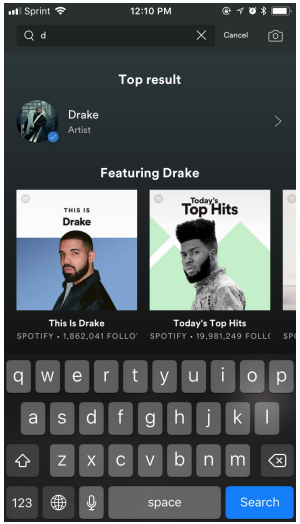


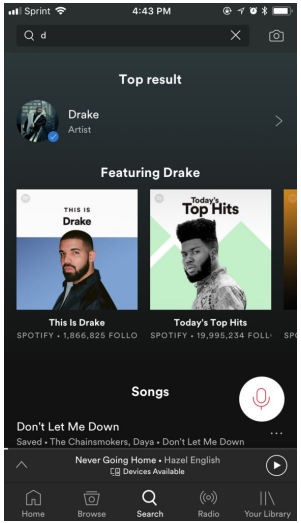
Overview of the user's journey

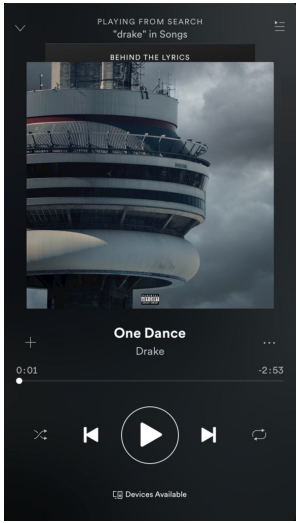
TYPE
User communicates with us

CONSIDER
User evaluates what we show them

DECIDE
User ends the search session







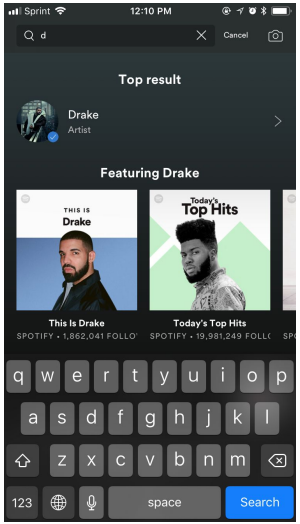
But users don’t come to Search as blank slates...

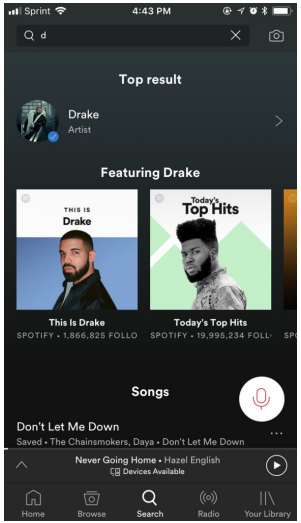
GOAL
What the user wants to do

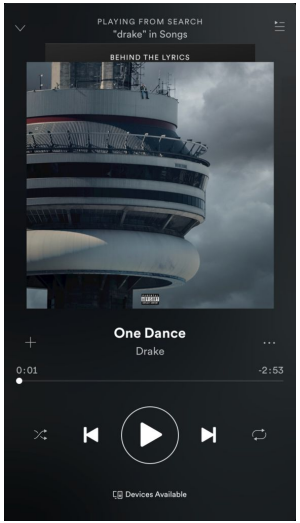
TYPE
User communicates with us

CONSIDER
User evaluates what we show them

DECIDE
User ends the search session







But users don’t come to Search as blank slates...

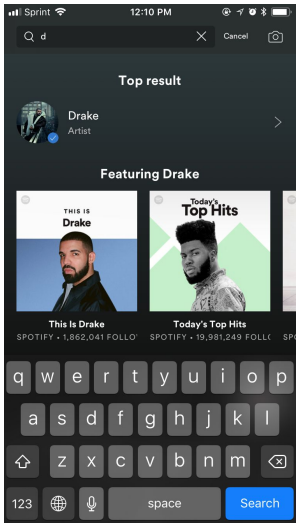
GOAL
What the user wants to do

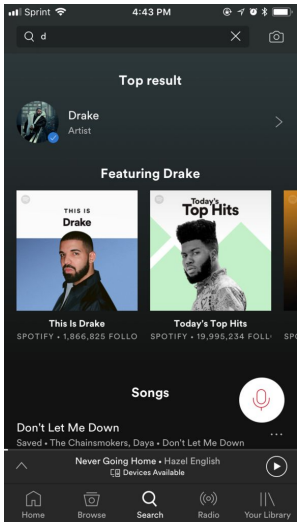
MINDSET
How the user thinks about results

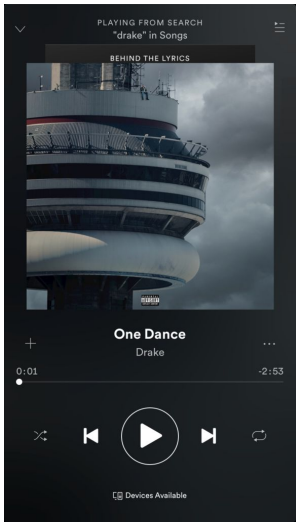
TYPE
User communicates with us

CONSIDER
User evaluates what we show them

DECIDE
User ends the search session







How do users evaluate Search?

By assessing success first and minimizing effort second.

Success

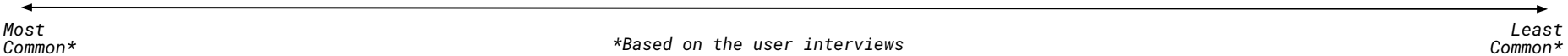
is finding what you were looking for

Effort

is how difficult the process of searching felt, whether or not is was successful

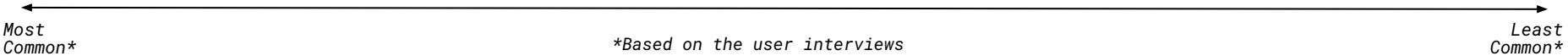
Goals help us understand success behaviors

LISTEN Have a listening session	**ORGANIZE** Curate for future listening	**SHARE** Connect with friends	**FACT CHECK** Find specific information
• Long stream • Add to queue	• Add to playlist • Save to library • Follow artists • Follow playlists • Download	• Share • Follow user	?



Goals help us understand success behaviors

LISTEN Have a listening session	**ORGANIZE** Curate for future listening	**SHARE** Connect with friends	**FACT CHECK** Find specific information
• Long stream • Add to queue	• Add to playlist • Save to library • Follow artists • Follow playlists • Download	• Share • Follow user	?



Success cannot typically be determined until after a click-through happens

Type and consider behaviors help us understand effort

TYPE

- Character entry
- Backspace
- Delete string
- Paste
- Toggle (drag cursor to fix typo)

CONSIDER

On results page

- Scrolling
- Clicking
  - Position

Beyond results page

- Quickback
- Click back
- Page view
- Short streams
- Previews (NFT)
- Google search (unobservable)

Type and consider behaviors help us understand effort

TYPE

- Character entry
- **Backspace**
- **Delete string**
- Paste
- **Toggle** (drag cursor to fix typo)

Going backward feels worse than going forward.

CONSIDER

On results page

- Scrolling
- Clicking
  - Position

Beyond results page

- **Quickback**
- **Click back**
- Page view
- Short streams
- Previews (NFT)
- Google search (unobservable)

Success and effort perceptions vary by mindset

There are 3 mindsets in Search:

FOCUSED
One specific thing in mind

OPEN
A seed of an idea in mind

EXPLORATORY
A path to explore

Mindsets in Search

FOCUSED
One specific thing in mind

*"Usually I'm pretty **focused** when I use Spotify. Generally, **I have an artist or band in mind that I'll search for.** I know right where to go."*

*Most Common**

**Based on the user interviews*

*Least Common**

Mindsets in Search

FOCUSED
One specific thing in mind

SUCCESS PERCEPTION

Binary
- Don’t find it
- Find it

EFFORT TOLERANCE

Low tolerance
- Quickest/easiest path to success is important

*Most Common**

**Based on the user interviews*

*Least Common**

Mindsets in Search

FOCUSED
One specific thing in mind

OPEN
A seed of an idea in mind

SUCCESS PERCEPTION

Binary
- Don’t find it
- Find it

EFFORT TOLERANCE

Low tolerance
- Quickest/easiest path to success is important

*“I guess my challenge is **sometimes I don't know what I wanna listen to**. But I **know one song or one artist** that I wanna listen to and then I kind of like I **think it flows naturally into something that's similar** and that's usually what a playlist will do.”*

Most Common*

*Based on the user interviews

Least Common*

Mindsets in Search

FOCUSED One specific thing in mind	**OPEN** A seed of an idea in mind

SUCCESS PERCEPTION

Binary
- Don’t find it
- Find it

Non-binary
- Nothing good enough
- Good enough
- Better than good enough

EFFORT TOLERANCE

Low tolerance
- Quickest/easiest path to success is important

Medium tolerance
- Willing to try some things out
- But still want to get to their goal efficiently

*Most Common**

**Based on the user interviews*

*Least Common**

Mindsets in Search

FOCUSED One specific thing in mind	**OPEN** A seed of an idea in mind	**EXPLORATORY** A path to explore

SUCCESS PERCEPTION

Binary
- Don’t find it
- Find it

Non-binary
- Nothing good enough
- Good enough
- Better than good enough

EFFORT TOLERANCE

Low tolerance
- Quickest/easiest path to success is important

Medium tolerance
- Willing to try some things out
- But still want to get to their goal efficiently

*“I've had **an interest in French rap**… so there's **one artist I know**, his name is Stromea. So, I'll find his song but then **I don't really know where to go from there**. This is all completely uncharted territory for me… **I can play these but I wouldn't know where to go from here or if I wanted to hear different French rappers.**"*

*Most Common**

**Based on the user interviews*

*Least Common**

Mindsets in Search

FOCUSED One specific thing in mind	**OPEN** A seed of an idea in mind	**EXPLORATORY** A path to explore

SUCCESS PERCEPTION

Binary
- Don’t find it
- Find it

Non-binary
- Nothing good enough
- Good enough
- Better than good enough

Unknown
- Difficult for users to assess success of exploration in the moment

EFFORT TOLERANCE

Low tolerance
- Quickest/easiest path to success is important

Medium tolerance
- Willing to try some things out
- But still want to get to their goal efficiently

High tolerance
- Users expect to be active in the discovery process
- Effort is expected

*Most Common**

**Based on the user interviews*

*Least Common**

Let’s revisit our goals

01 Investigate underlying **assumptions** of existing Search evaluation.

02 Develop **hypotheses** around metrics that capture the user’s experience with Search.

How do the metric **assumptions** hold up?

Satisfaction proxy metrics

Click-through rate

Assumption 1: seeing the search results page provides sufficient information for users to evaluate the results.

X Most goals are achieved beyond the click, so success cannot typically be determined until after a click-through happens.

Top click (y/n)

Assumption 2: having the “correct” result at the top of the results page is important to users.

~ Click position is a relevant proxy for effort, but it does not capture all effort.

How do the metric **assumptions** hold up?

Which are applied to individual “search sessions”

`Start: first keystroke`

Assumption 3: users perceive their search to begin once they start typing.

~ Users have a goal and a mindset before the first keystroke, but first keystroke is the cleanest starting point in data.

`End: click or 10 minutes`

Assumption 4: users perceive their search to be complete once they have found a result worthy of clicking.

X Users perceive their search to be complete when they have determined whether or not they have succeeded in their goal.

What **hypotheses** should we test?

Success and effort hypothesis

Users evaluate their experience with search through success and effort.

Mindset hypothesis

Mindsets shape the perception of success and effort.

For more information on how themes arise from qualitative research: J. Garcia-Gathright, C. Hosey, B. St. Thomas, B. Carterette, F. Diaz. “Mixed Methods for Evaluating User Satisfaction.” Presented at: RecSys 2018.

How did we test UR hypotheses at scale?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Hypotheses to test

Success and effort hypothesis

Users evaluate their experience with search through success and effort.

↳ `Can we distinguish user behavior in terms of success and effort?`

Mindset hypothesis

Mindsets shape the perception of success and effort.

↳ `Can we identify stable mindset constructs in data?`

`Do mindsets shape user behavior?`

Success and effort hypothesis

Can we distinguish success and effort behaviors in log data?

Method

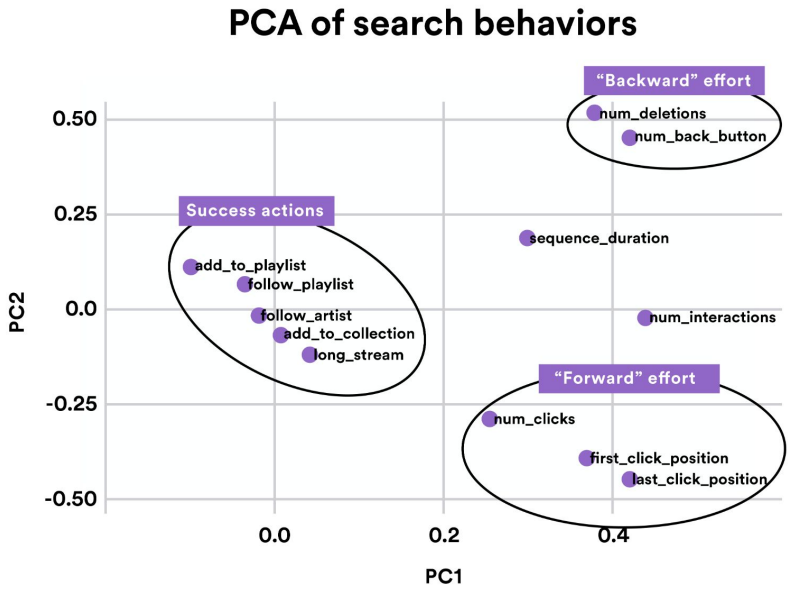
We used a principal component analysis (PCA) on the following logged search metrics to group related interactions:

- add_to_collection
- add_to_playlist
- first_click_position
- follow_artist
- follow_playlist
- last_click_position
- long_stream
- num_back_button
- num_clicks
- num_deletions
- num_interactions
- sequence_duration

Findings

We identified behavioral separation into success, backward effort, and forward effort components.

- Clicks and sequence duration are between “Success” and “Effort” dimensions
- Effort metrics organize into:
  - “forward” (clicking through)
  - “backwards” (deletions and back buttons)



Mindset hypothesis

Can we identify stable mindset constructs in data? And do these constructs shape search behavior?

Method

01 Collect ground truth mindset data

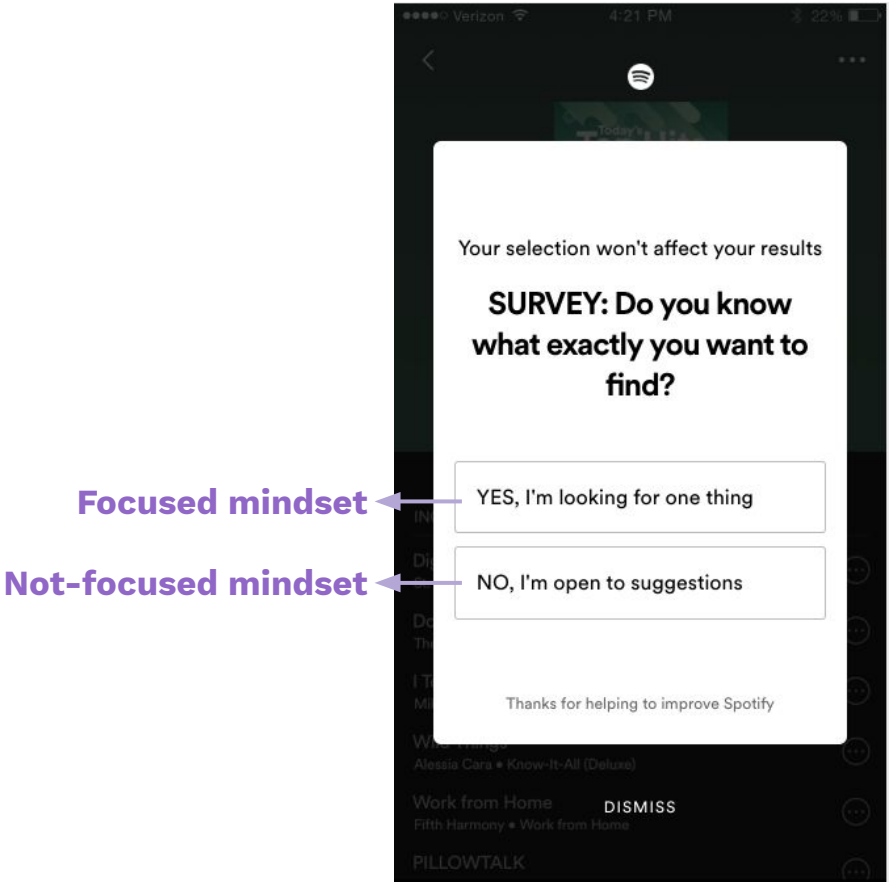
02 Compare mindset distribution across demographics and time

03 Linear regression to infer differences in behavior across mindsets

01 Collect ground truth mindset data

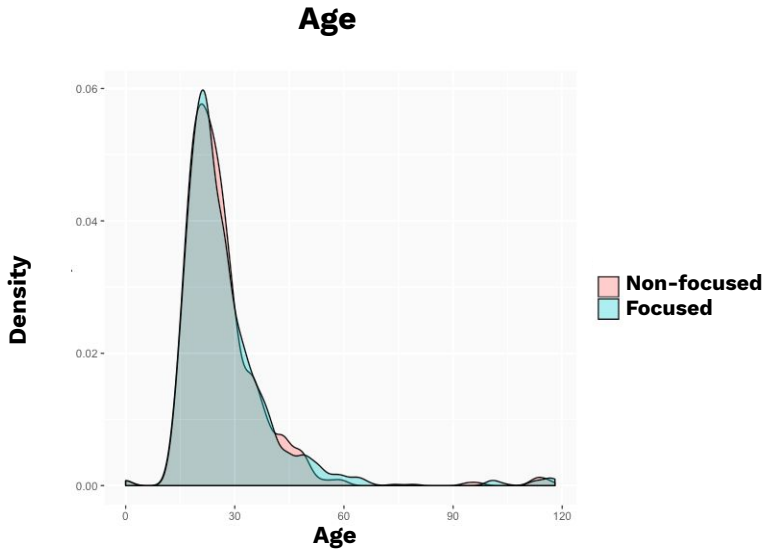
We sent an in-app **mindset survey**

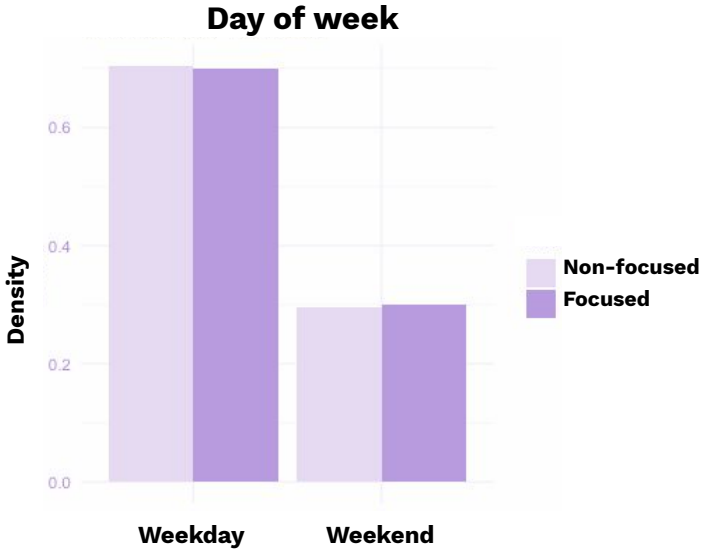
- Survey triggered on search page before user begins typing in the search bar
- Triggered ONCE per user lifetime
- Usability tested and iterated to ensure response options captured **focused** and **non-focused** mindsets



Findings

Mindsets **DID NOT** vary based on user age, gender, day of week, or time of day.





Findings

Mindsets **DID** vary by discovery rates in the past month, but **NOT** by account age or level of engagement.

- Users searching with a ***focused mindset*** had a **lower discovery rate** in the month leading up to that search session than users searching with a *non-focused mindset*.

  ($p < 0.05$, $\beta = -0.58$)

Findings

Mindsets **DID** influence user behavior during that session.

Focused mindset searches had:

Query / Search

- **Longer** search sessions ($p < 0.01$, $\beta = 0.055$)
- **More** complete queries with:
  - **More** total keystrokes
  - **Longer** queries (longer, more words)

Click

- **More** time before first click ($p < 0.05$, $\beta = 0.019$)
- **Lower** click position (**more** scrolling)
- Clicked items were:
  - **More** likely an album or track
  - **Less** likely playlist

Listen

- **Less** time streaming ($p < 0.1$, $\beta = -0.040$)
- **Fewer** long streams ($p < 0.1$, $\beta = -0.045$)
- **More** likely to save / add

How did we apply the learnings to Search?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Changes to Search

01 Metric changes

We re-defined success metrics to make them more user-centric and more sensitive to product changes.

02 Connecting session-level metrics to long-term metrics

We can understand how the new user-centric metrics relate to long-term business outcomes.

03 Data infrastructure and logging

Search updated how they logged their search data to better incorporate relevant interactions and meaningfully group them.

01 Re-defining success metrics

Click-through rate

Top click (y/n)

Success Signals

- Long Streams
- Add to playlist
- Save to library
- Follow artists/playlists
- Download

Effort Signals

- Character entry
- Backspace
- Delete string
- Scrolling
- Click Depth

01 Re-defining success metrics

Compared to CTR, the success signals provide a detailed view of search performance depending on the goals or mindset of the user.

In practice, it is often convenient to have a single metric for decision making. A simple way to create a composite metric that captures all the success signals is to take the union of them. We tested this composite metric for directionality and sensitivity using A/B tests and compared against CTR.

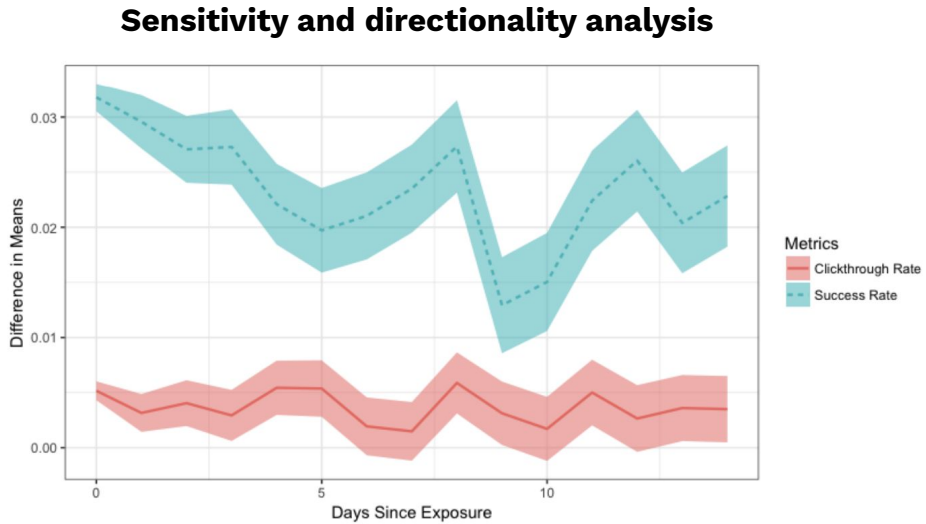
Success Signals

- Long Streams
- Add to playlist
- Save to library
- Follow artists/playlists
- Download

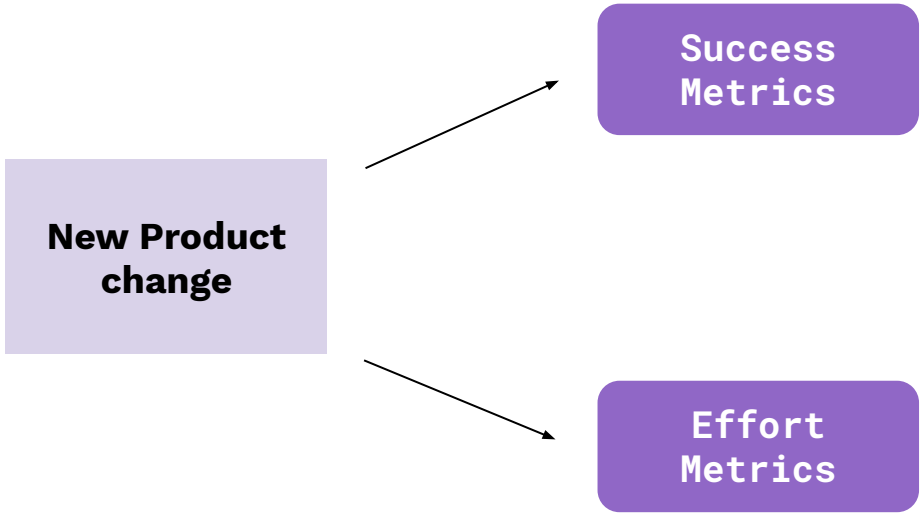
Composite Success Metric
(***Success Rate***)

01 Re-defining success metrics

New composite success metric is more sensitive to improvements than CTR (+~.3pp vs. +~.5pp)

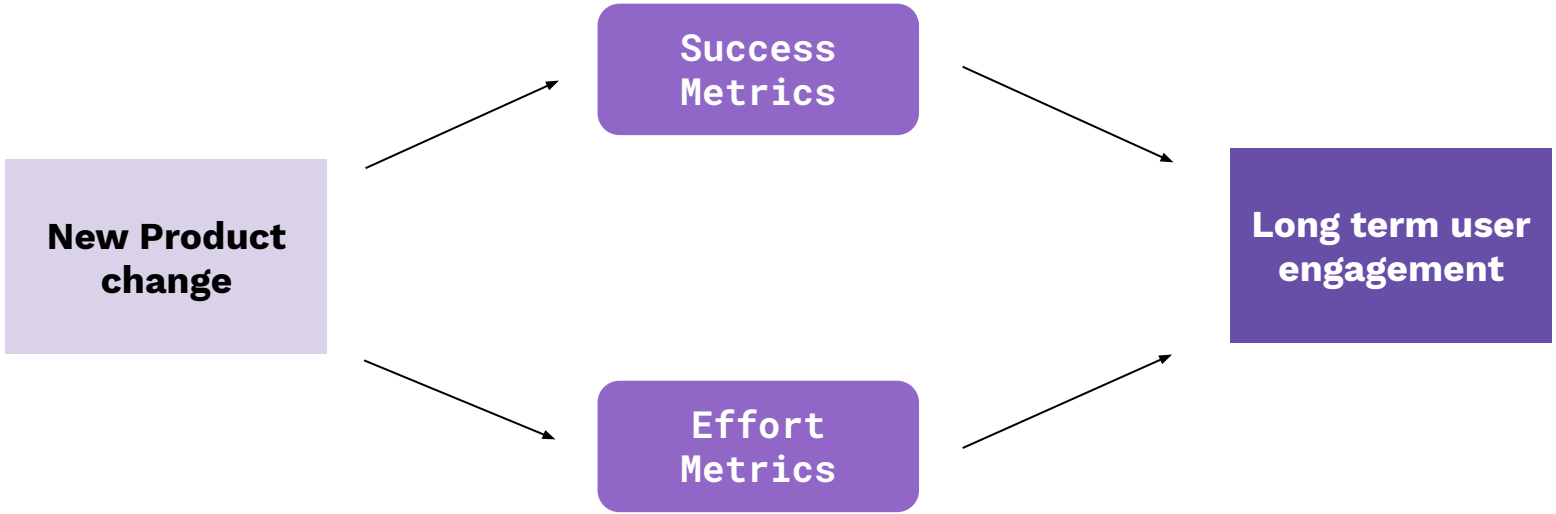


02 Informing A/B Test Decision Making



With a larger set of metrics to evaluate search quality, the metrics need to be prioritized or combined.

02 Using long term outcomes to learn metric combinations



⚠️ More information: *P. Chandar, F. Diaz, B. St. Thomas, "Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems: Part 3", 2020.*

02 Surrogate functions to weight metrics

Long term user engagement is not always observable within the A/B test, but it may be available *observationally* outside of A/B tests

Using our new search metrics as “surrogates”, we can construct a score function to use as a weights, so that we’d have a function like (imprecise notation):

Long Term Engagement = f(Short term metrics, Treatment Assignment)

Susan Athey, Raj Chetty, Guido W. Imbens, Hyunseung Kang. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. NBER 2019.

02 Learning Linear Surrogate Functions

Hohnhold et al 2015 used long-term experiments to fit a regression of long-term outcomes based on short-term search page metrics.

Learned weights from regression

$$\tilde{U}_{CTR} \approx k_1 \cdot \Delta\text{AdRelevance} + k_2 \cdot \Delta\text{LandingPageQuality} \quad (7)$$

Long term User Engagement

Search Metrics

Dmitriev et al, 2016 discuss challenges in collecting long-term decision criteria for this type of analysis

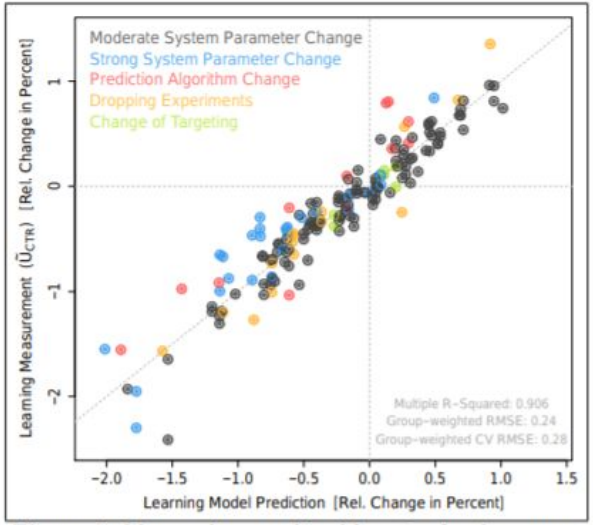


Figure 4: Measured vs. predicted learning for the current desktop macro-model.

Henning Hohnhold et al. Focusing on the Long-term: It's Good for Users and Business. KDD 2015; Pavel Dmitriev et al. Pitfalls of Long-Term Online Controlled Experiments. IEEE Big Data 2016.

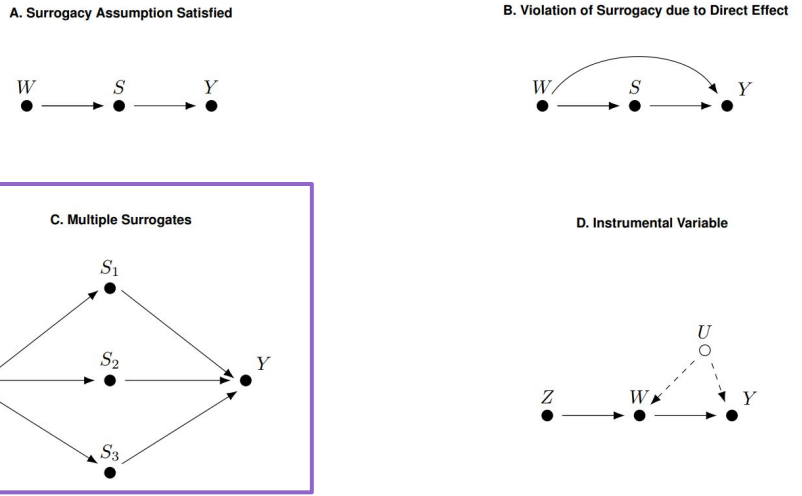
02 Statistical surrogacy without long-term experiments

Athey et al 2019 shows the conditions under which unbiased estimates of long-term effects can be constructed from ***observational data*** using the *surrogate score* and *surrogate index*.

FIGURE 1
Surrogacy Assumptions and Violations

Statistical Surrogacy:

P(*Y*, *W* | *S*) = P(*Y* | *S*) P(*W* | *S*)



With many surrogates measured between the treatment and the long term outcome, they are likely to lie near the true causal chain.

Susan Athey, Raj Chetty, Guido W. Imbens, Hyunseung Kang. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. NBER 2019.

02 Creating new A/B test criteria

New success and effort metrics
Introduces a need to combine and prioritize metrics

Long-term A/B tests available

Meta-analysis to understand which success and effort metrics predict long term outcomes best, for different types of experiments

Regression from meta-analysis on effects can be used to combine metrics.

Only observational data available for long-term outcomes

Surrogate score and index functions can be constructed from only observational data.

In A/B test analysis (with proper randomization) the surrogate score function can be used to combine metrics.

Home

Overview

Goal: To develop metrics for Home that align with how users engage with and experience the Spotify Home screen.

01 Product background

What Home is.

02 Metric use

How Home is traditionally evaluated and the implicit assumptions that these metrics make.

03 User study

Research to understand how users engage with and experience Home directly to test assumptions and generate new hypotheses about user-centric metrics.

04 Quantitative follow-up

The quantitative work to test hypotheses generated by the UR at scale.

05 Applications

Determining how to apply learnings to Home.

What is Home?

01 Product background

02 Metric use

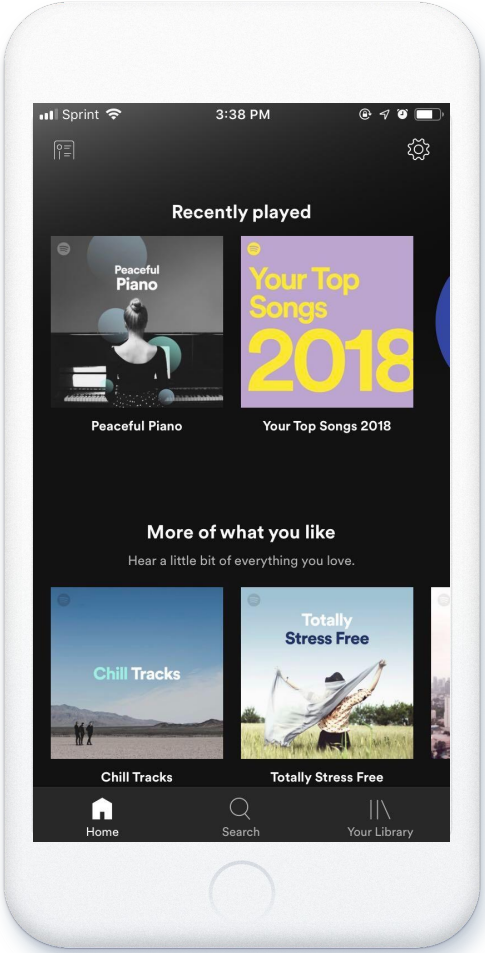
03 User study

04 Quantitative follow-up

05 Applications

What is Home?

The default screen on Spotify that offers personalized recommendations that users can scroll through.



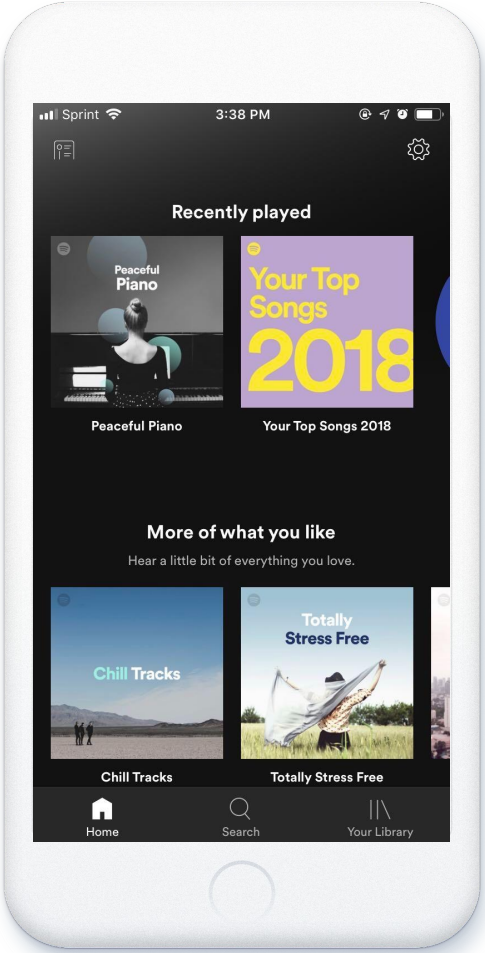
What is Home?

From a system perspective:

Home is made up of a personalized set “shelves” with personalized content on each shelf. Shelf content is accessed through horizontal scrolling and additional shelves are surfaced via vertical scrolling.

From a user perspective:

A tool that provide recommendation with passive interaction (scrolling) from the user, but the user has no direct input into what is surfaced.



How is Home traditionally evaluated?

01 Product background

02 Metric use

03 User study

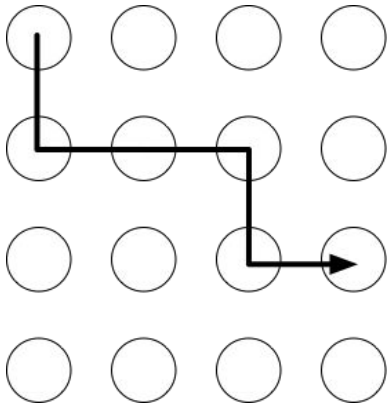
04 Quantitative follow-up

05 Applications

Offline evaluation metrics: grid-based metrics

User model: model of ordered user traversal of grid items; utility independent of stopping probability.

Metric: expected utility given simulated user behavior.



item utility is equal to editorial utility

slate utility is the expected cumulative utility

policy utility is the average slate utility

X Xie, J Mao, Y Liu, M de Rijke, Y Shao, Z Ye, M Zhang, S Ma. Grid-based evaluation metrics for web image search. WWW 2019.

Home relied on simple heuristics for evaluation

Satisfaction proxy metrics

`Click-through rate`

`Consumption time`

These metrics have underlying assumptions

Satisfaction proxy metrics

`Click-through rate`

Assumption 1: seeing the recommendations on the page provides sufficient information for users to evaluate the results.

`Consumption time`

Assumption 2: satisfaction is proportional to time streaming

What did we learn through user research?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

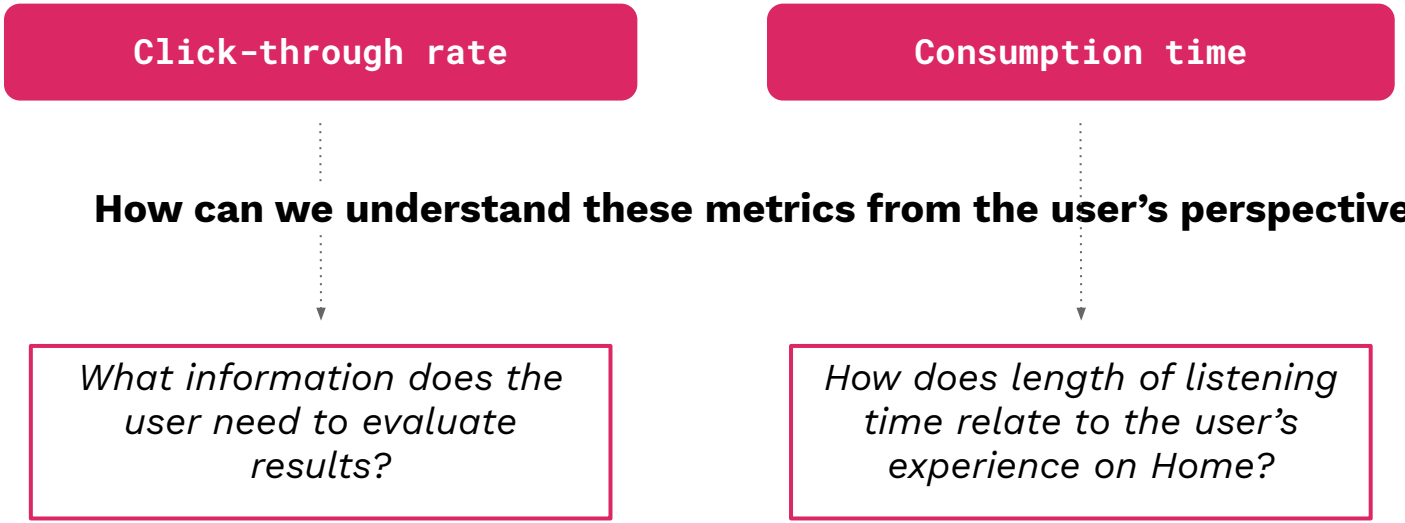
User study goals

01 Investigate underlying assumptions of existing Home evaluation.

02 Develop hypotheses around metrics that capture the user’s experience with Home.

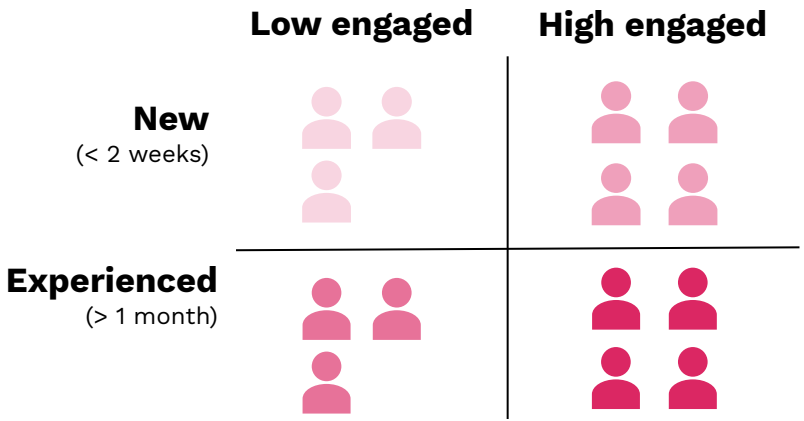
Investigating metric assumptions through user research

Satisfaction proxy metrics



Study design

We conducted semi-structured interviews with 14 participants across 4 cohorts built around two dimensions: Home engagement (low vs. high) & account age (< 1 month vs. > 3 months).



Interviews covered:

- Music preferences
- Spotify usage
- Home
  - Attitudes
  - Good and bad experiences
  - Usage deep dive

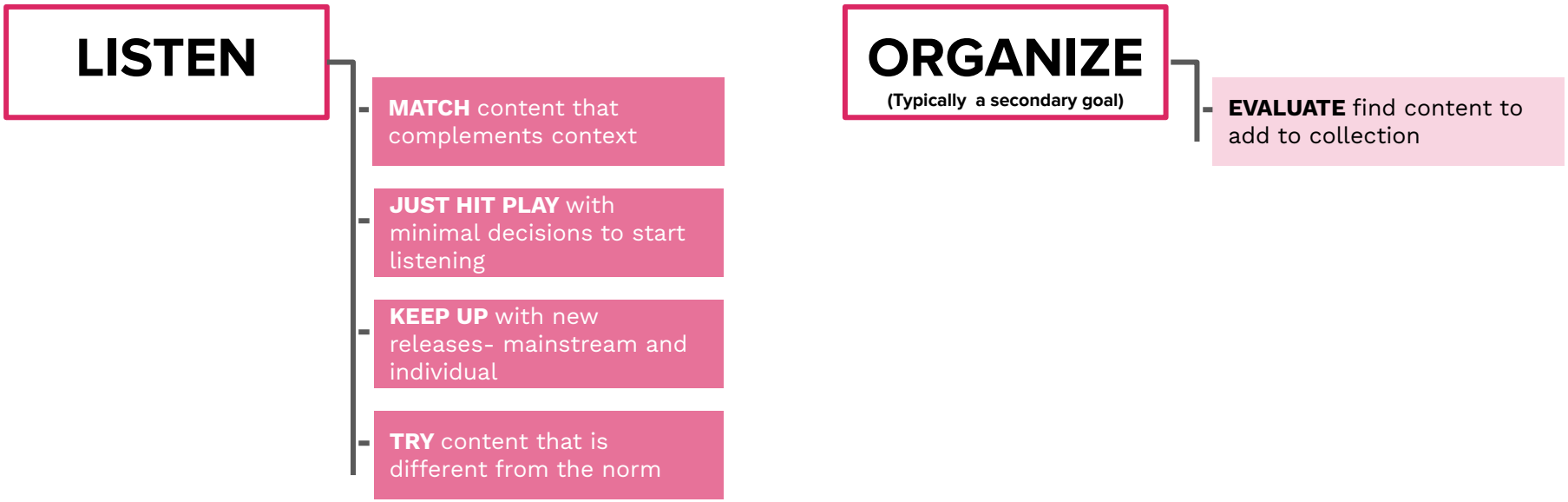
Research questions

1. **Why** do users engage with Home?
2. **How** do users engage with Home?
3. How do users **evaluate** their Home experience (e.g., what is a good vs. bad experience)?

Why do users engage with Home?

Participants typically used Home as an entry point for **listening**, occasionally with a secondary curation goal.

Goals can be broken down more granularly



Participants want Home to function as a shortcut to achieve their goals

Easier than other navigation

*“There was something on there that was what I wanted to listen to and **it was easier to just click on that than to search for it manually.**”*

An experienced, low-engaged user, 34

Relevant to their tastes

*“The subject on the main screen should be **what you like.**”*

A new, high-engaged user, 28

How do users engage with Home?

Participants gave Home only a sliver of attention.

Home typically receives only a glance

Participants demonstrated:

- A brief (and often subconscious) scan of Home
- Judgments that were made quickly, with an eye toward familiarity
- Most often without scrolling

It is important to optimize the “above the fold” space with scannable, relevant content.

But scrolling did happen sometimes

When scrolling does happen:

Effort

When there is a belief content *should* be there.

Curiosity

When there is time and an interest in seeing Spotify recommendations.

Trust

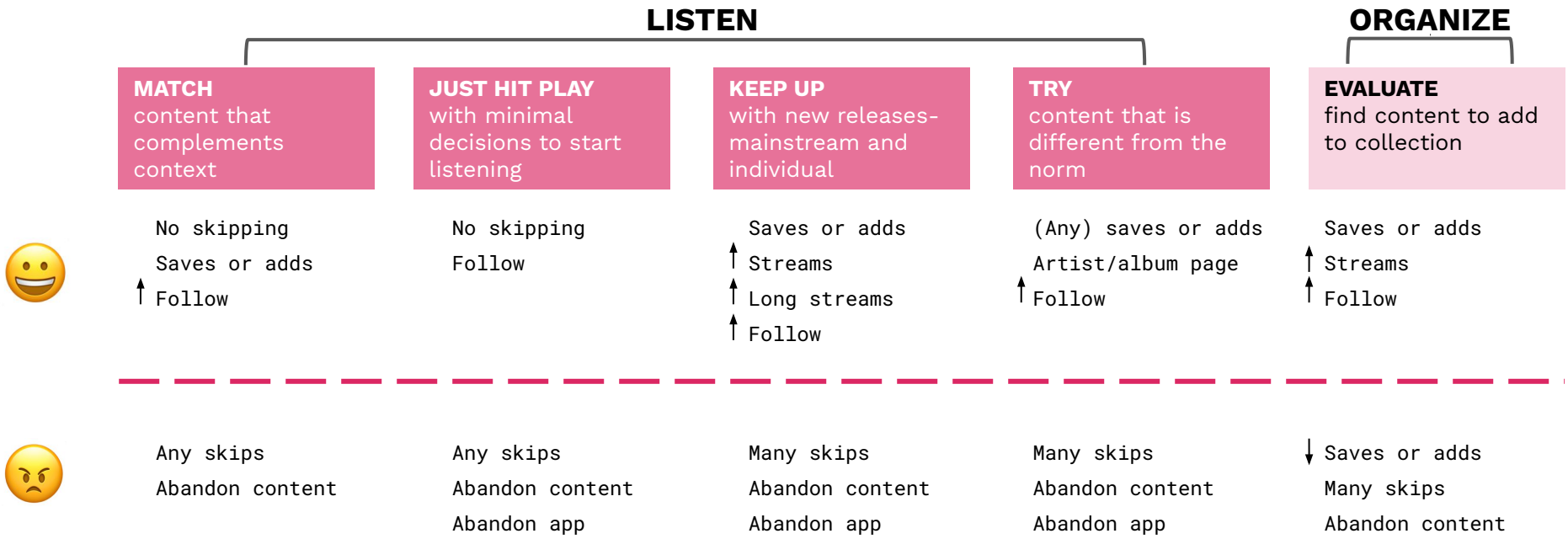
When there has been positive past Home experiences that make the user believe in quality recommendations.

Looking at scrolling both within and across session can tap into different scrolling motivations.

How do users evaluate Home?

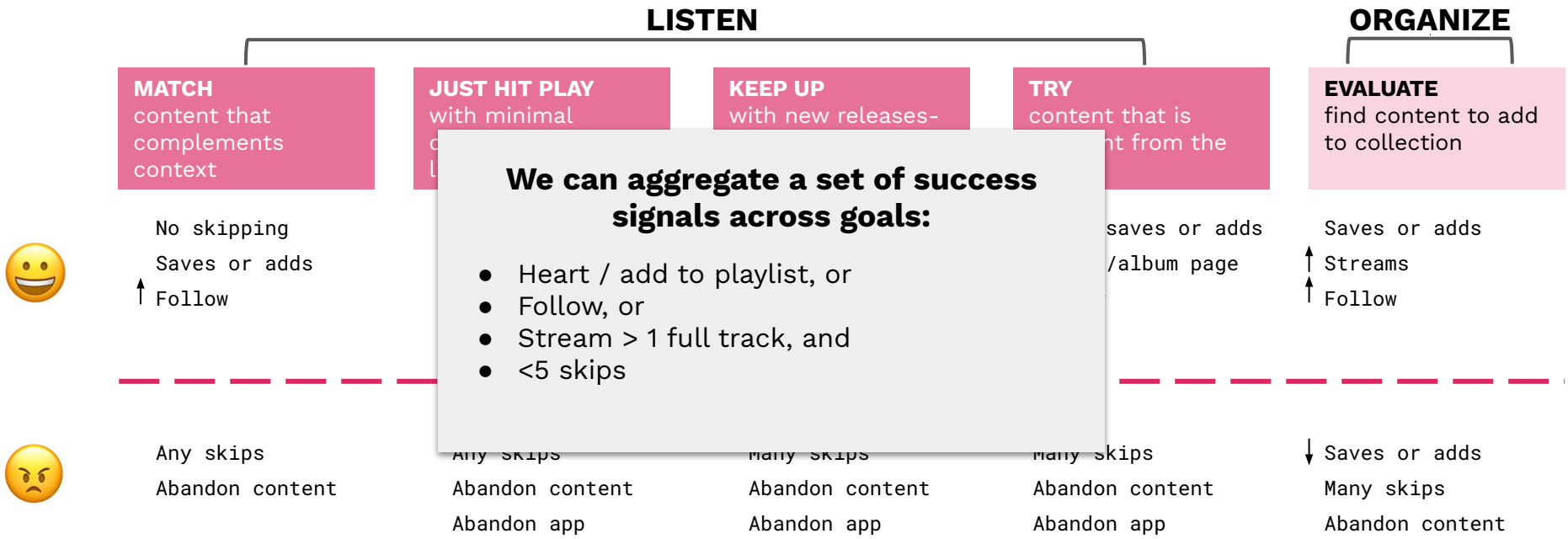
Participants evaluated Home primarily on goal success, secondarily on effort.

Success signals vary by goal



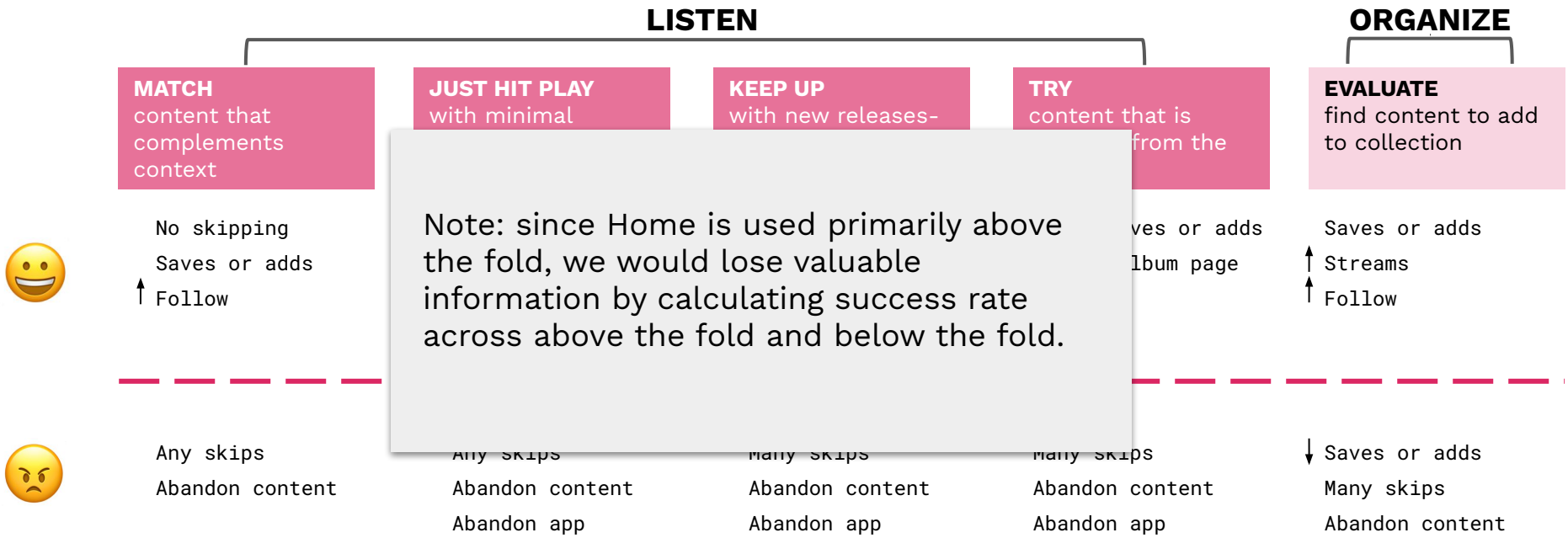
Note: signals with arrows are should be normalized by user

Success signals vary by goal



Note: signals with arrows are should be normalized by user

Success signals vary by goal



Note: signals with arrows are should be normalized by user

Let’s revisit our goals

01 Investigate underlying **assumptions** of existing Home evaluation.

02 Develop **hypotheses** around metrics that capture the user’s experience with Home.

How do the metric **assumptions** hold up?

Satisfaction proxy metrics

Click-through rate

Assumption 1: seeing the recommendations on the page provides sufficient information for users to evaluate the results.

X Most goals are achieved beyond the click, so success cannot typically be determined until after a click-through happens.

Consumption time

Assumption 2: satisfaction is proportional to time streaming

~ Streaming time is typically determined by constraints that are distinct from the user's goal.

What **hypotheses** should we test?

Shortcut hypothesis

We can represent most of the content users are currently consuming in a small space.

Scrolling hypothesis

Scrolling indicates effort, but a propensity to scroll indicates trust in the recommendations.

Data split hypothesis

Splitting data into “above the fold” and “below the fold” can better capture nuances in behavior.

For more information on how themes arise from qualitative research: J. Garcia-Gathright, C. Hosey, B. St. Thomas, B. Carterette, F. Diaz. “Mixed Methods for Evaluating User Satisfaction.” Presented at: RecSys 2018.

How did we test UR hypotheses at scale?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Hypotheses to test

Shortcut hypothesis

We can represent most of the content users are currently consuming in a small space.

↳ How top heavy is the distribution over a time period per user?

How fast do sources of content turn over?

Scrolling hypothesis

Scrolling indicates effort, but a propensity to scroll indicates trust in the recommendations.

↳ How does scrolling on Home relate to long-term outcomes?

Data split hypothesis

Splitting data into “above the fold” and “below the fold” can better capture nuances in behavior.

↳ Do we see differences in success metrics above versus below the fold?

Shortcuts hypothesis

How top heavy is the distribution over a time period per user?

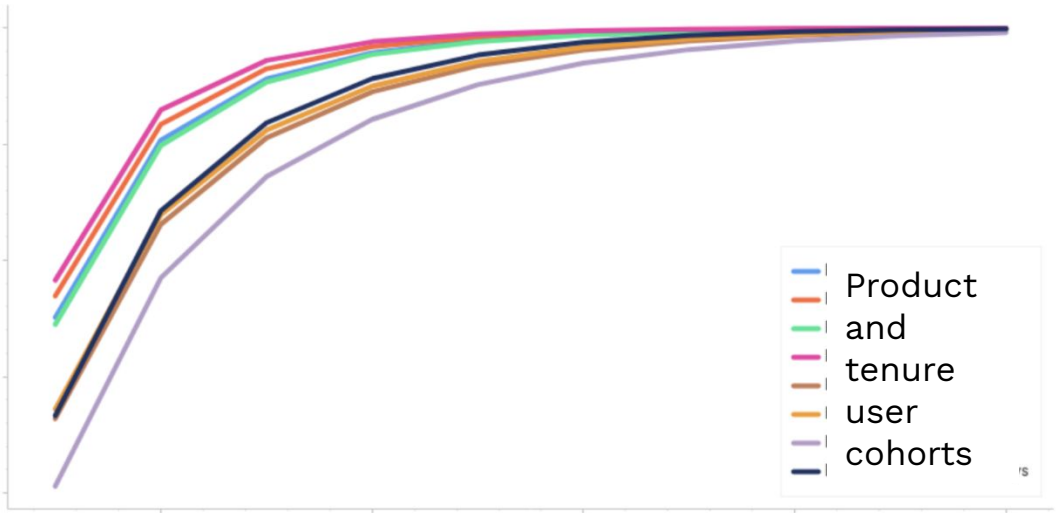
Most consumption comes from a handful of sources

Method

We looked at simple exploratory plots.

For each user, we measured how many content sources we would need to hit some threshold of consumption (e.g. 100%) over a period of time (e.g. 1 month).

Cumulative normalized consumption over sources of content for a time period



How fast do the sources of content turn over?

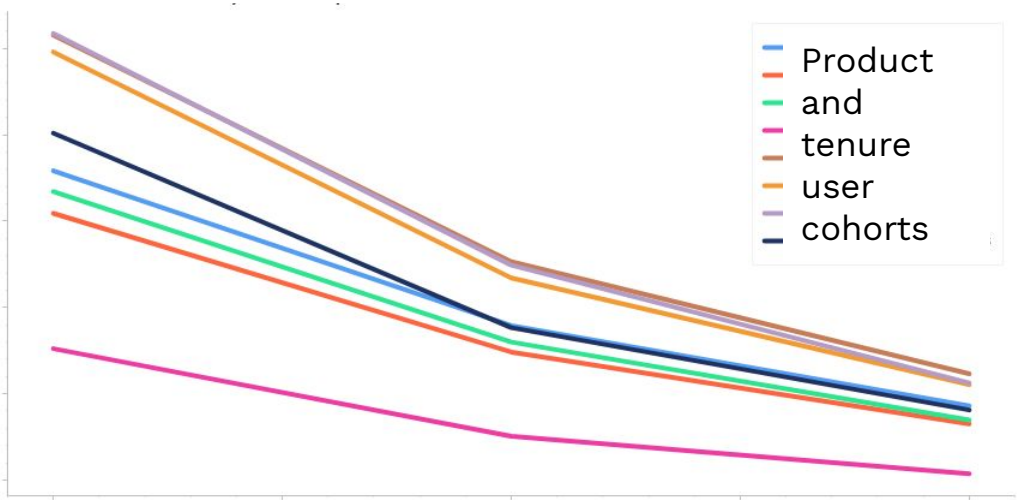
The top sources of content were fairly stable

Method

We used simple exploratory plots to understand what percent of top content sources were remained top content sources in the next time period.

Given a small number of top content sources that were stable over time, creating personalized content shortcuts should be feasible.

Conservation of top content sources between time periods



Scrolling hypothesis

How does scrolling on Home relate to long-term outcomes?

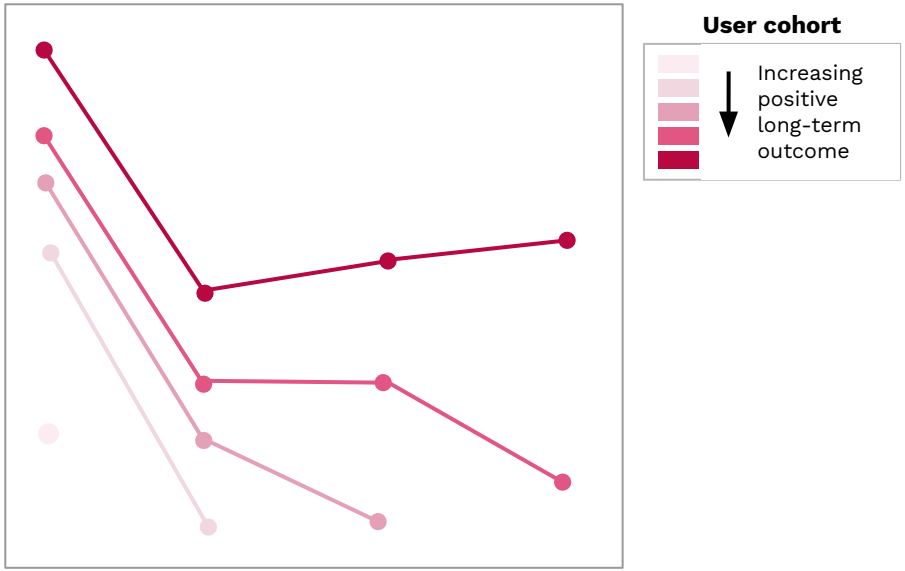
Scrolling on Home is more common with users who have better outcomes on Spotify

Method

We constructed simple exploratory plots showing how likely users are to scroll at different points in the user life-cycle.

This aligns with the finding that scrolling may mean more than just an effortful experience.

Impressions from scrolling on Home over time



Data split hypothesis

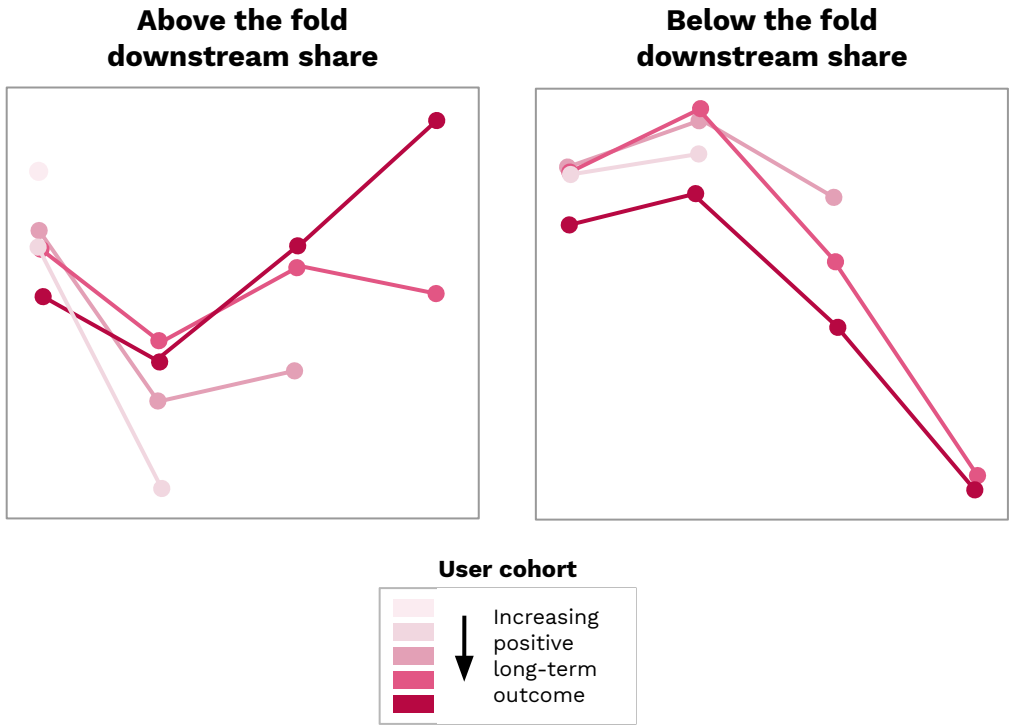
Do we see differences in success metrics above versus below the fold?

Differences in downstream engagement

Method

We plotted downstream engagement behaviors (e.g., saving tracks to a playlist) over time and split by above versus below the fold.

We saw different patterns of engagement in the “stickiness” of content based on where the content was discovered on Home.



How did we apply the learnings to Home?

01 Product background

02 Metric use

03 User study

04 Quantitative follow-up

05 Applications

Changes to Home

01 UI changes

Home was re-designed to optimize the “above the fold” space in line with the shortcuts hypothesis.

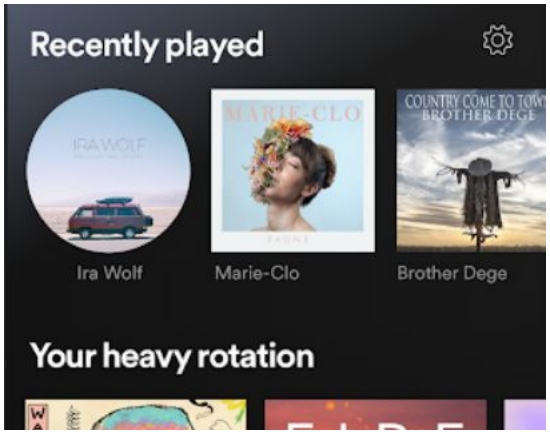
02 Metric changes

We introduced success and effort metrics and created a framework to split the data in line with the data split hypothesis.

01 Re-designed Home around the shortcuts hypothesis

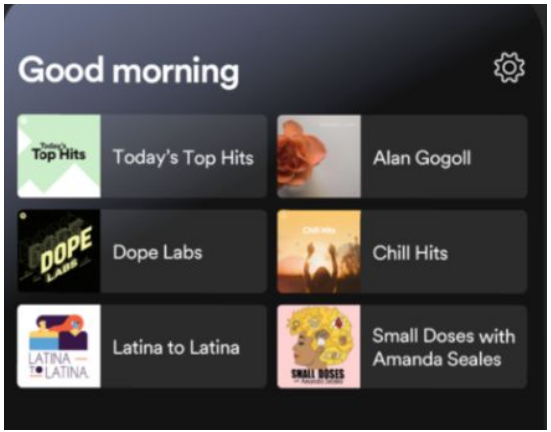
From:

3 options optimized using recency alone



To:

6 options optimized using additional information like consumption time



https://engineering.atspotify.com/2020/04/15/reach-for-the-top-how-spotify-built-shortcuts-in-just-six-months/

02 Updated Metric Framework for A/B test evaluation

Overall A/B test criteria

Long-term indicators of overall Home page quality

Secondary metrics

"Above the fold" metrics
- Success
- Effort
- Downstream engagement

"Below the fold" metrics
- Success
- Effort
- Downstream engagement

Lessons learned

What are the top **five** pro tips we picked up along the way?

01 Users don’t care about your product vision.

There is strong heterogeneity in how users use and evaluate a system.

Understanding that heterogeneity can add a lot of value to how to evaluate system performance.

Common assumption	Missed opportunity
“People who interact with this product we built to do [X] want to do [X]”	Understanding different user goals and the success criteria attached to them
“Habit formation means the product is successful”	Identifying positive long term outcomes and their short term surrogates
“People who measure a higher [Y] are having a better experience than people who measure a lower [Y]”	Understanding the measurable contexts relevant to a user’s evaluation of their experience.

02 Behavior metrics and system metrics do not always get along.

Optimizing for behavior-centric metrics (e.g., CTR, consumption time) can lead to **degraded system-centric metrics** (e.g. long-term retention, fairness, algorithmic responsibility, satisfaction).

Exploratory data analysis alone will likely miss biases inherent to behavior-centric metrics, as there is no way to understand why a behavior occurred.

→

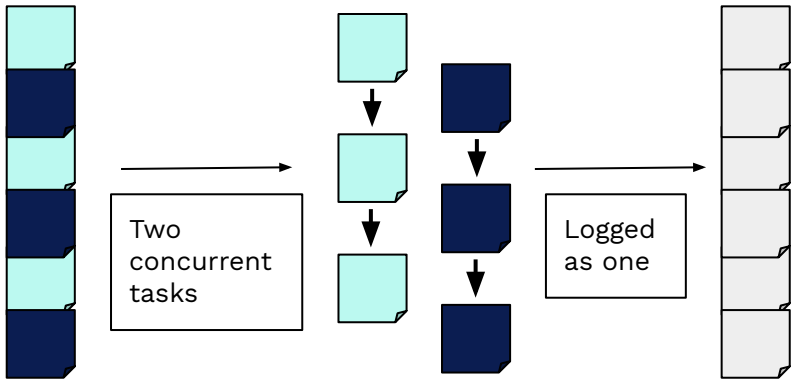
Qualitative analysis can uncover and address potential discrepancies between short-term user behavior and overall system metrics.

03 You may need to restructure your data to test qualitative hypotheses.

It is **annoying**, but often **worth the effort**.

The structure of logged data may be out of sync with the mental models and behavioral patterns the qualitative research uncovered.

For example:



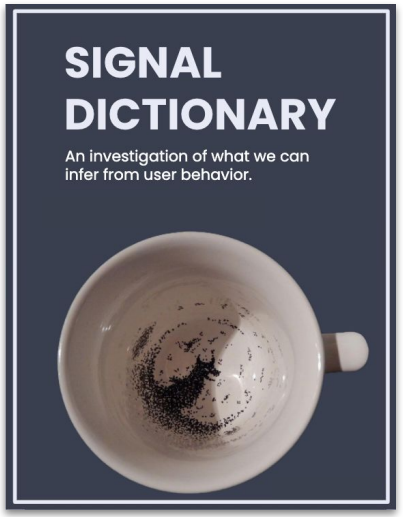
04 Start specific, then generalize.

It’s tempting to want to start with finding metrics that work across products, but **don’t**.

Focus on one product at a time for a deep and nuanced understanding. Once you do this for a variety of products, you can look across studies to:

- Synthesize a holistic understanding of how to interpret metrics in different settings
- Build and adapt mental models from previous research
- Enable meta-analysis of both qual and quant research

Ex:

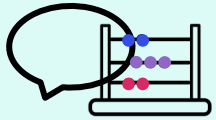


05 ~~Mix~~ Emulsify your methods.

High bandwidth communication throughout qualitative and quantitative phases can spark **positive, creative** discussion.

Examples that have worked well for us:

- Joint literature review across disciplines
- Questioning methods and assumptions, not just reporting results
- Discussing incremental progress, low-level details
- Bouncing ideas off each other
- Asking (and answering) lots of questions
- Encouraging more questions



Thanks!

Questions / thoughts?