Algorithmic Effects on the Diversity of Consumption on Spotify

Ashton Anderson
University of Toronto
Spotify

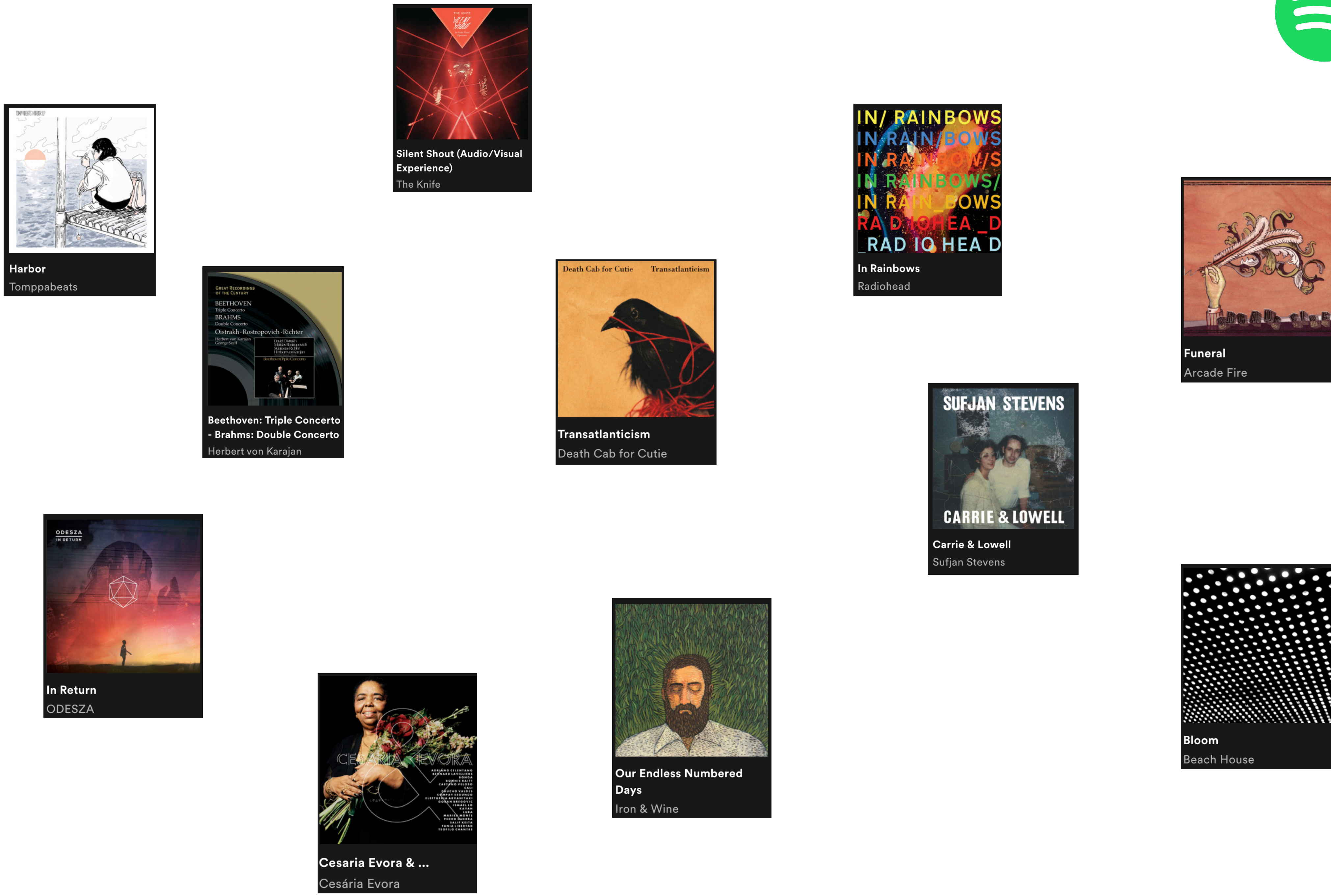
Ashton Anderson

Lucas Maystre

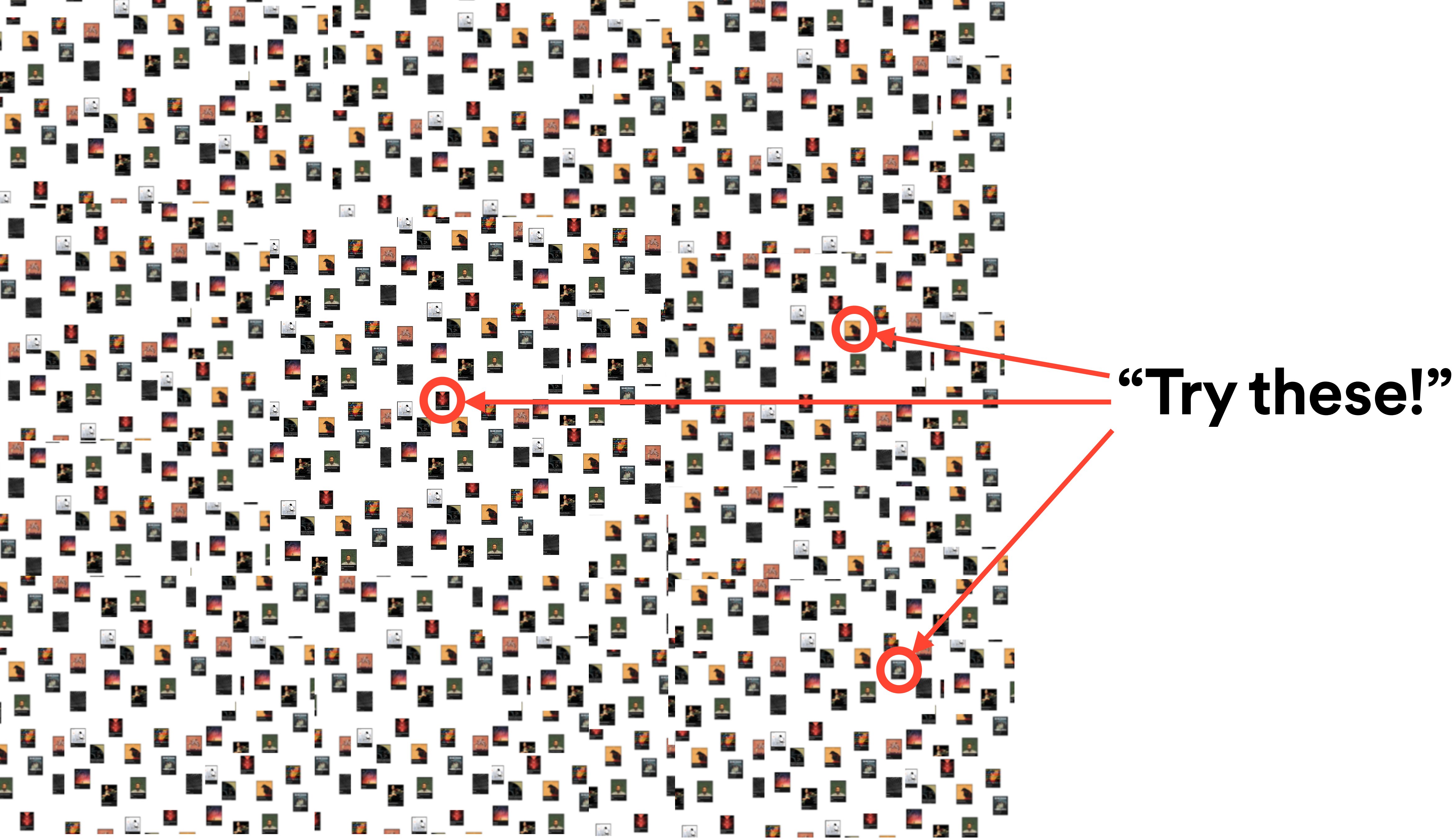
Ian Anderson

Rishabh Mehrotra

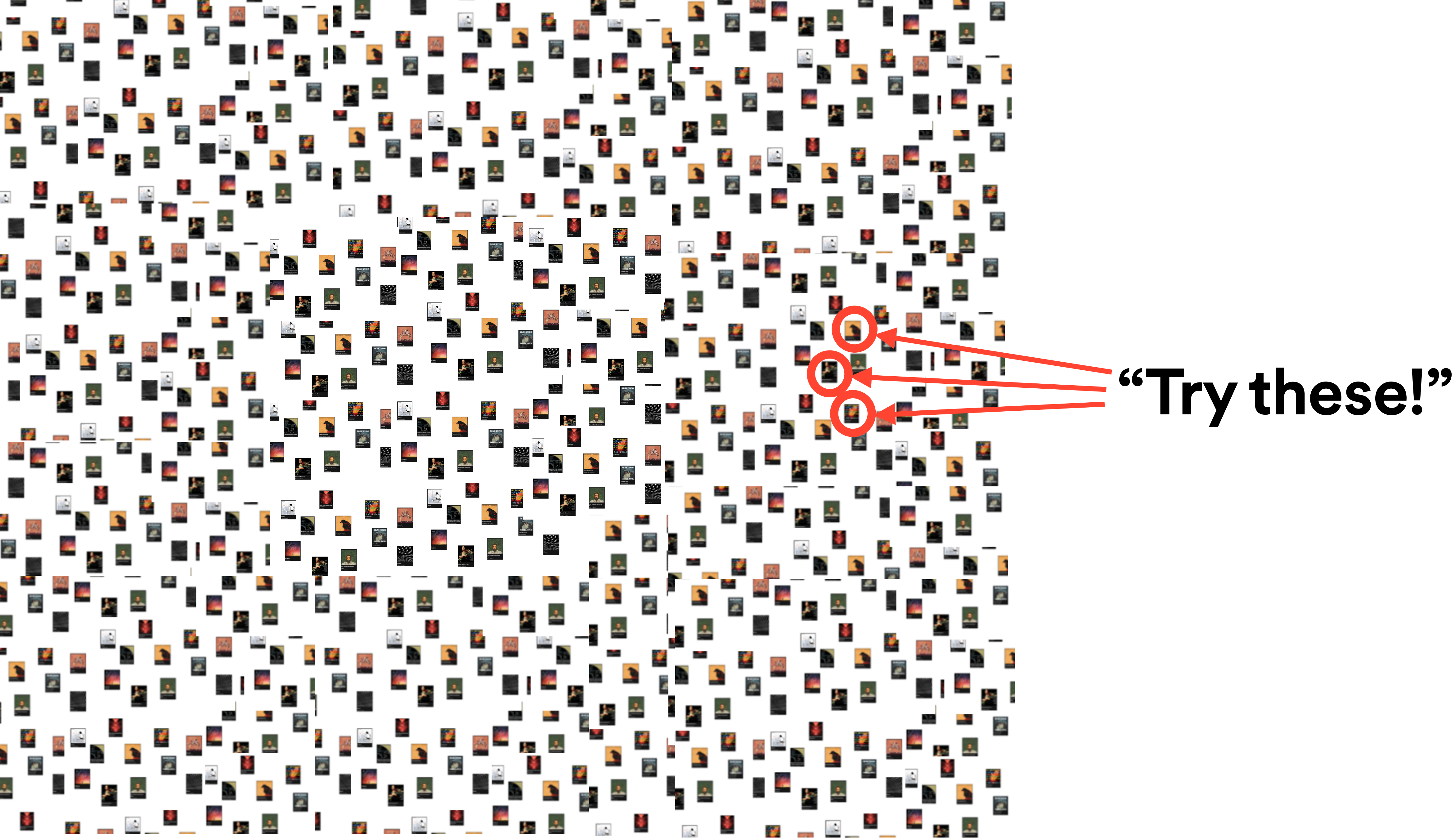
Mounia Lalmas



Online platforms present users with a universe of content to choose from

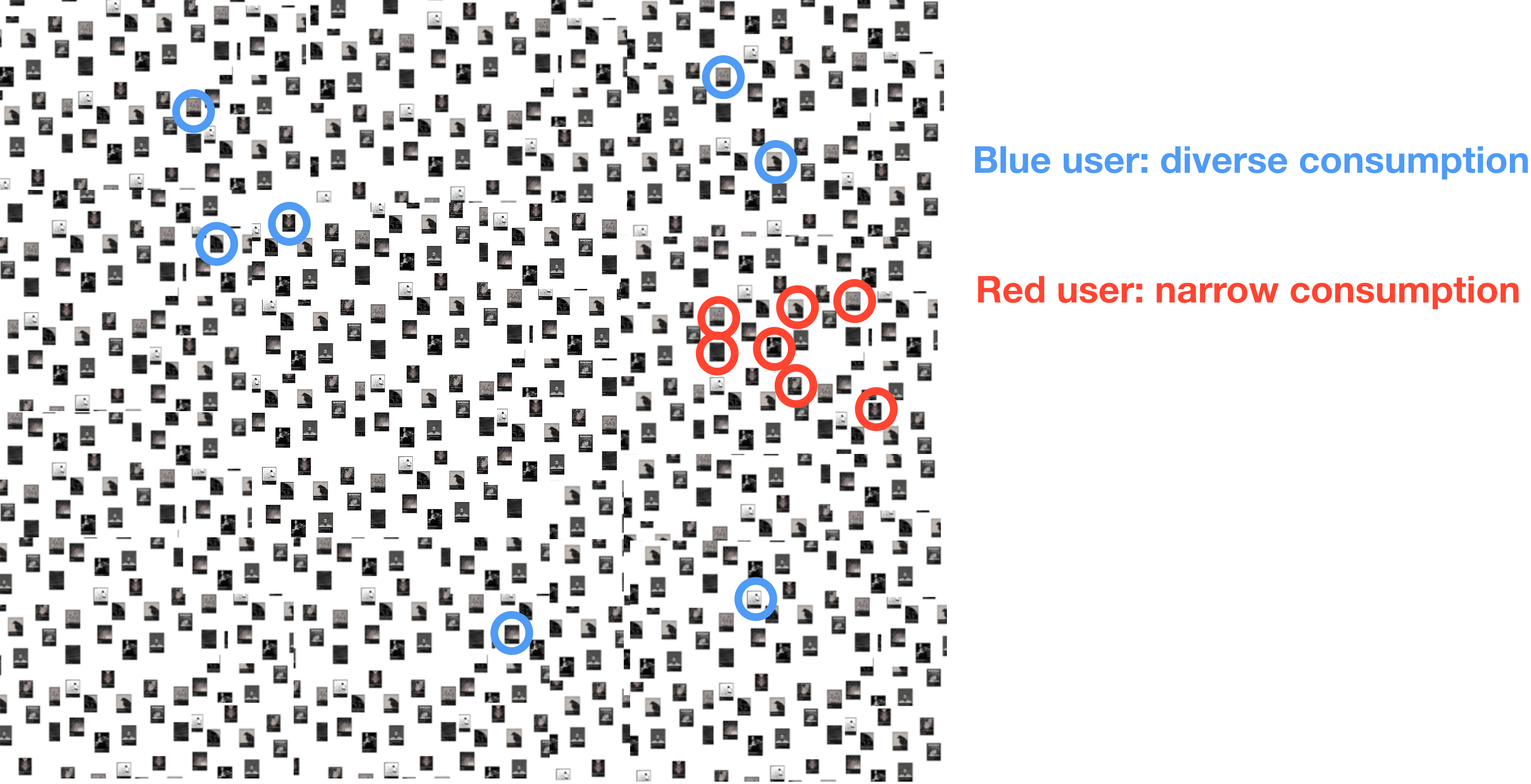


Recommendation algorithms influence which items users consume



But there is concern that recommendation algorithms concentrate on overly narrow sets of content

[1] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin, UK.
[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. *Exposure to ideologically diverse news and opinion on Facebook.* Science 348, 6239 (2015), 1130–1132.
[3] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. *Filter bubbles, echo chambers, and online news consumption.* Public opinion quarterly 80, S1 (2016), 298–320.



Consumption diversity is potentially beneficial for many reasons:

→ Exploring more of what the platform has to offer

→ Satisfying more needs on the platform

→ In many domains, diverse exposure is considered a virtue (news, information, etc.)

What is the association between algorithmic recommendation and the diversity of content users consume…

…and on the user experience in turn?

Data

We study consumption diversity on Spotify:

On Spotify, users can listen to over 50 million different songs on various digital devices.

There are free and premium versions of Spotify, with the free version being ad-supported and the premium version being subscription-based.

Main dataset:

→ Listening history of >100 million premium users during July 1–28, 2019
→ This comprises 70 billion "streams" (an instance of a user listening to a song)
→ For each user, we calculate their total activity and consumption diversity

How should we measure consumption diversity?

Measuring consumption diversity

What do we want from a measure of consumption diversity?

Captures breadth: diverse consumption means consuming from across the spectrum of content on the platform.
→ Diversity captures the extent to which consumed items are meaningfully different from each other.

Consistency: similarities and differences between items should be mutually consistent.
→ Measure of diversity should be comparable between users.

Scalability: should be efficiently computable for millions of items and millions of users.
→ Applicable to real-world online platforms.

Measuring consumption diversity

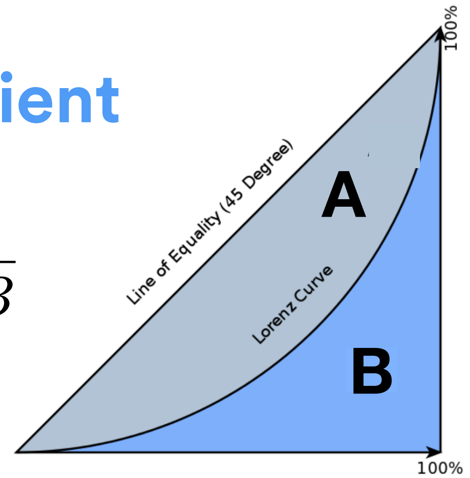
Existing approaches:

Entropy

$$H(X) = -\sum_i P_X(x_i) \log P_X(x_i)$$

Gini coefficient

$$Gini = \frac{A}{A+B}$$



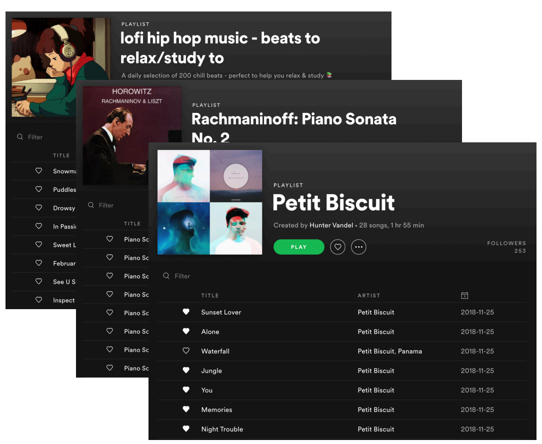
Problem: both entropy and Gini fail to capture breadth in a way that is sensitive to how **similar** the consumed items are

Compare:	10x "Let It Be" — The Beatles 10x "Blackbird" — The Beatles 10x "Imagine" — John Lennon	**vs.**	10x "Angel of Death" — Slayer 10x "Only Time" — Enya 10x "So What" — Miles Davis

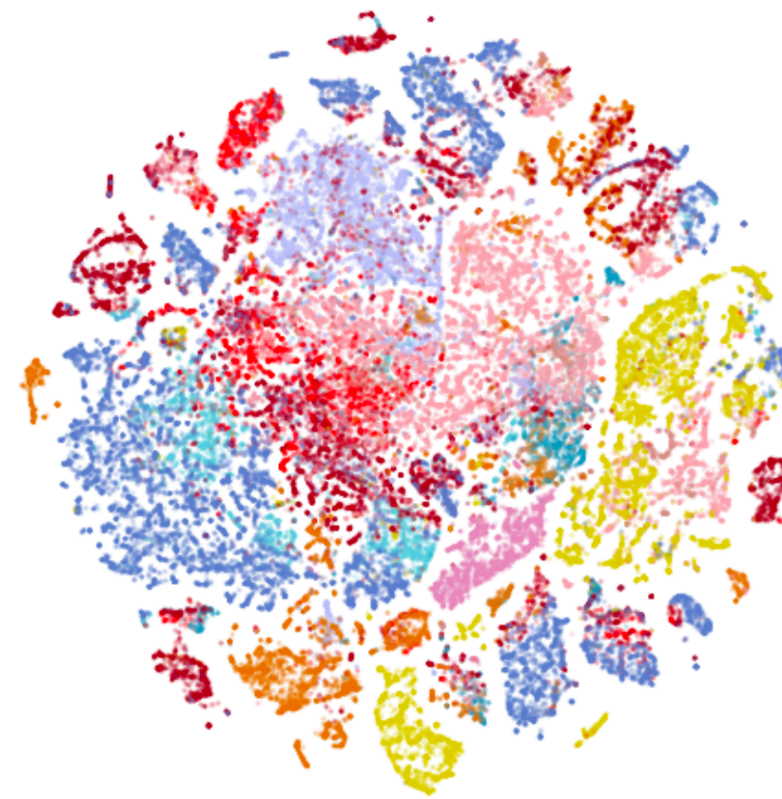
Entropy and Gini would consider these two users equally "diverse"!

Measuring consumption diversity

To consistently and scalably capture similarities between songs, we use song embeddings.



Treat user playlists as “documents” and songs as “words” and run word2vec



Songs are arranged in a high-dimensional space such that “similar” songs are close to each other in the space.

What does “similar” mean?

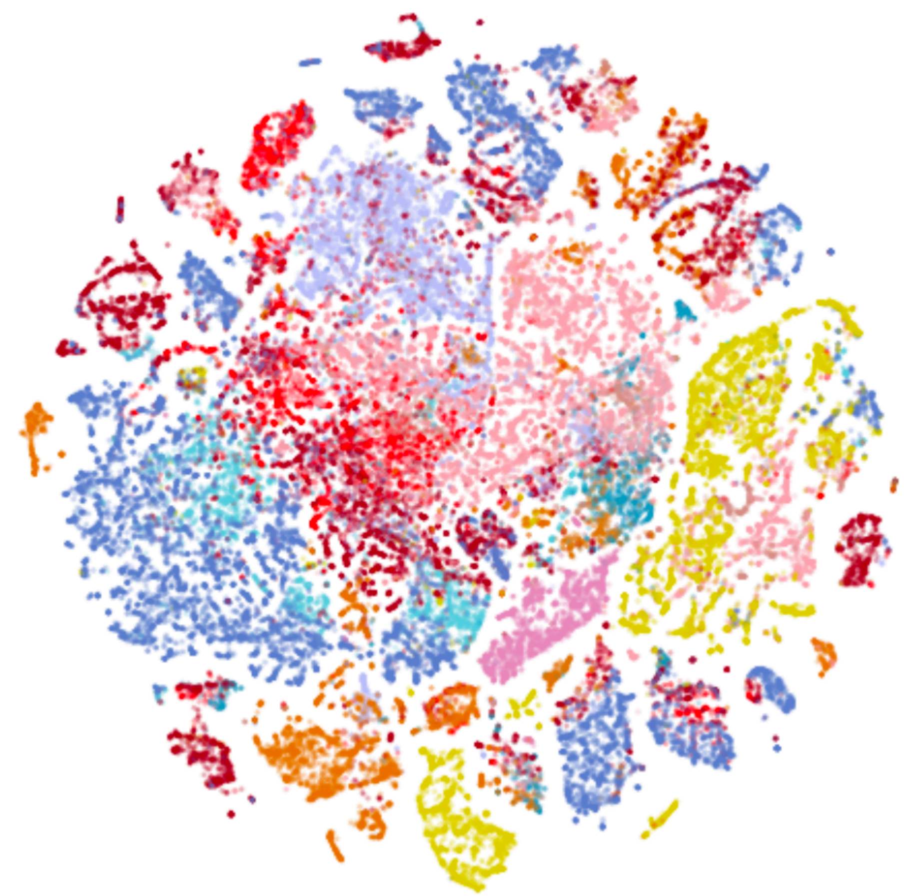
We define similarity empirically: the more often two songs appear together in user playlists, the more similar they are

Our song embeddings are:

→ 40-dimensional

→ trained on 850 million playlists

→ comprise millions of songs

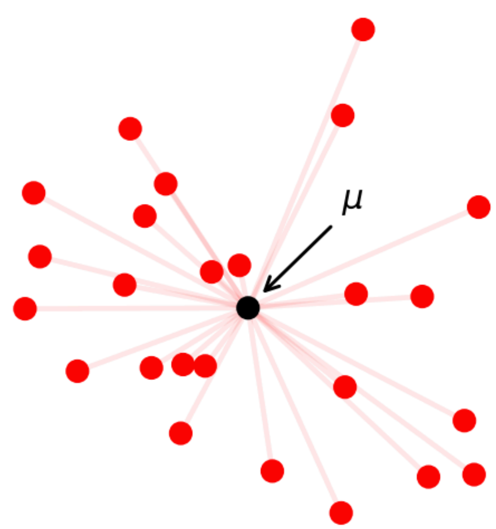


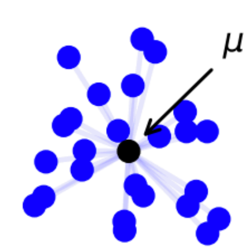
Song embeddings give us a way to consistently and scalably compute similarities between any pair of songs

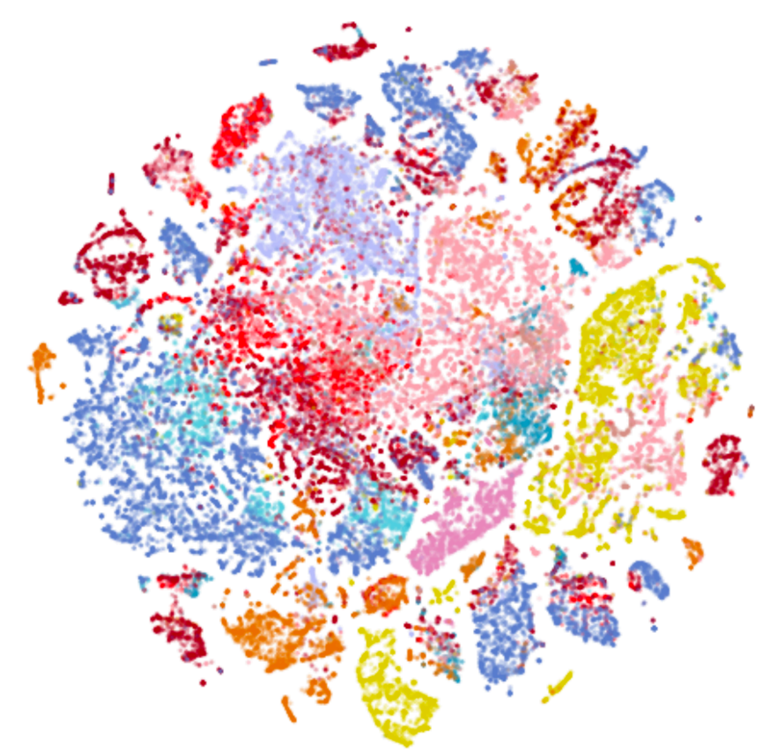
Measuring consumption diversity

With a proper way of measuring arbitrary song similarities in place, we can now measure the diversity of a user's consumption.

Intuition: the more "spread out" a user's songs are in the embedding space, the more diverse their consumption







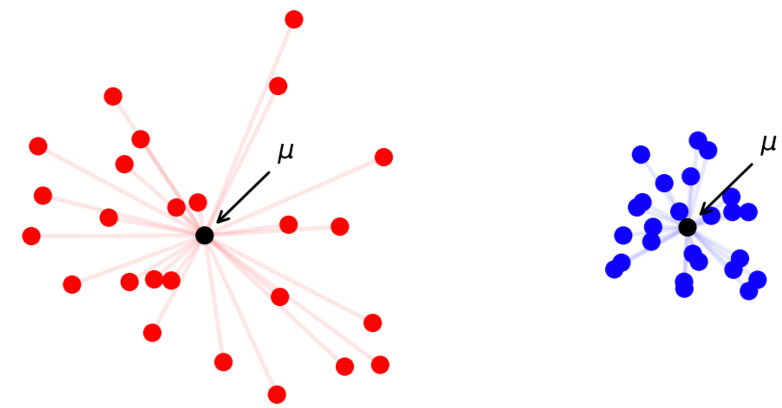
Diverse listening

Narrow listening

The GS-Score

We use a recently proposed definition of diversity based on embedding spaces

The **Generalist-Specialist (GS) score:** average cosine similarity between a consumed item and the user's center of mass



Diverse listening

Generalist

Narrow listening

Specialist

Say user i listens to song j w_j times, then:

Center of mass (simple mean):

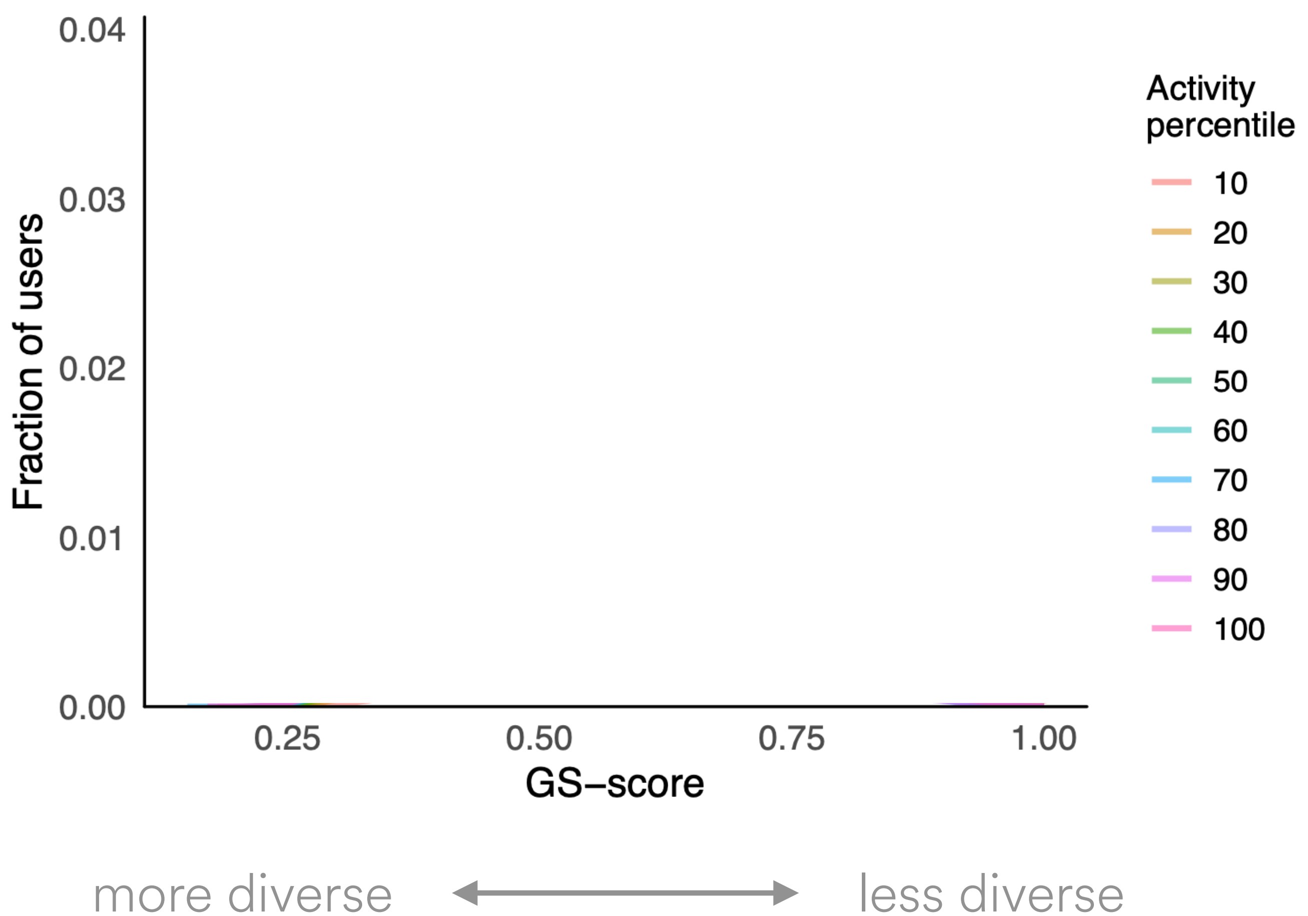
$$\vec{\mu}_i = \frac{1}{\sum w_j} \cdot \sum_j w_j \vec{s}_j$$

User consumption diversity measure:

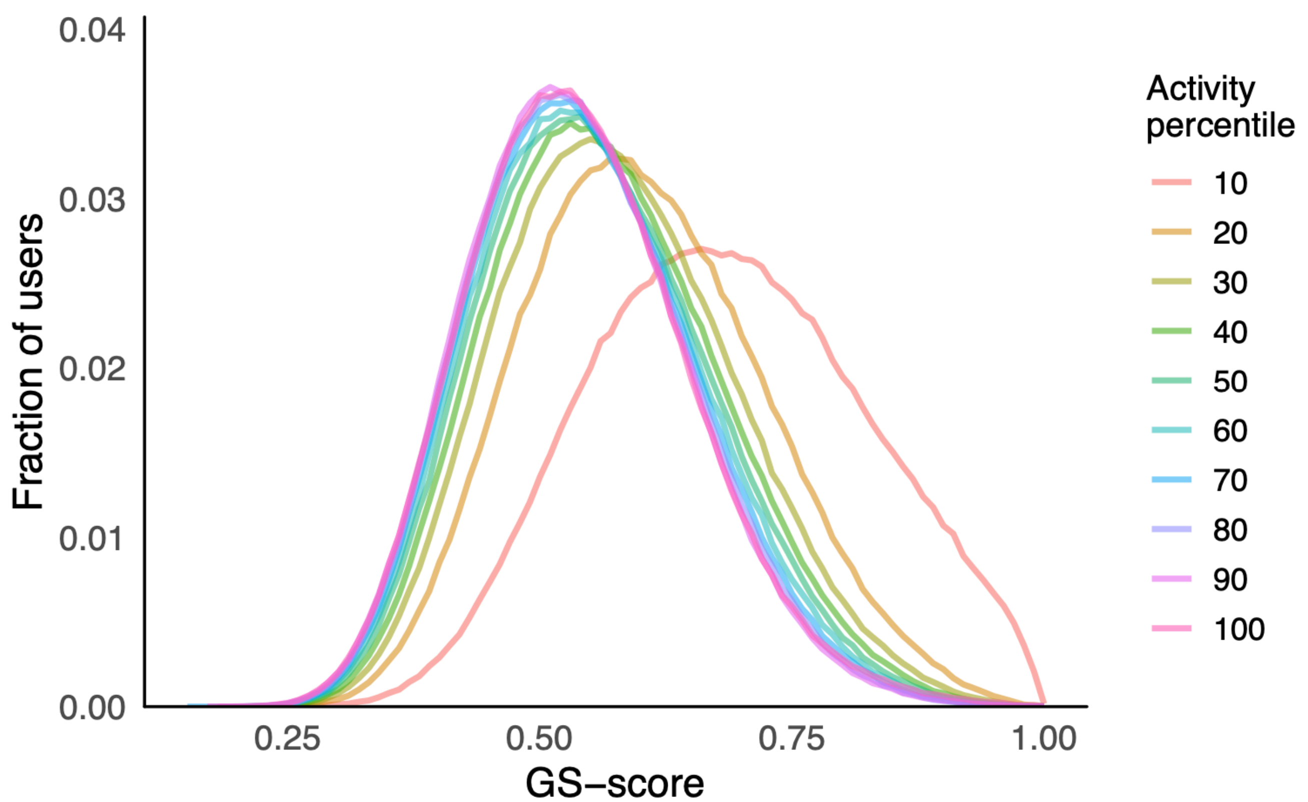
$$GS(u_i) = \underbrace{\frac{1}{\sum w_j} \sum_j w_j}_{\text{Weighted average over user's songs}} \underbrace{\frac{\vec{s_j} \cdot \vec{\mu_i}}{||\vec{s_j}|| \cdot ||\vec{\mu_i}||}}_{\text{Cosine similarity of song with center of mass}}$$

[1] Isaac Waller and Ashton Anderson. *Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms.* WWW 2019.

How is consumption diversity distributed?



How is consumption diversity distributed?

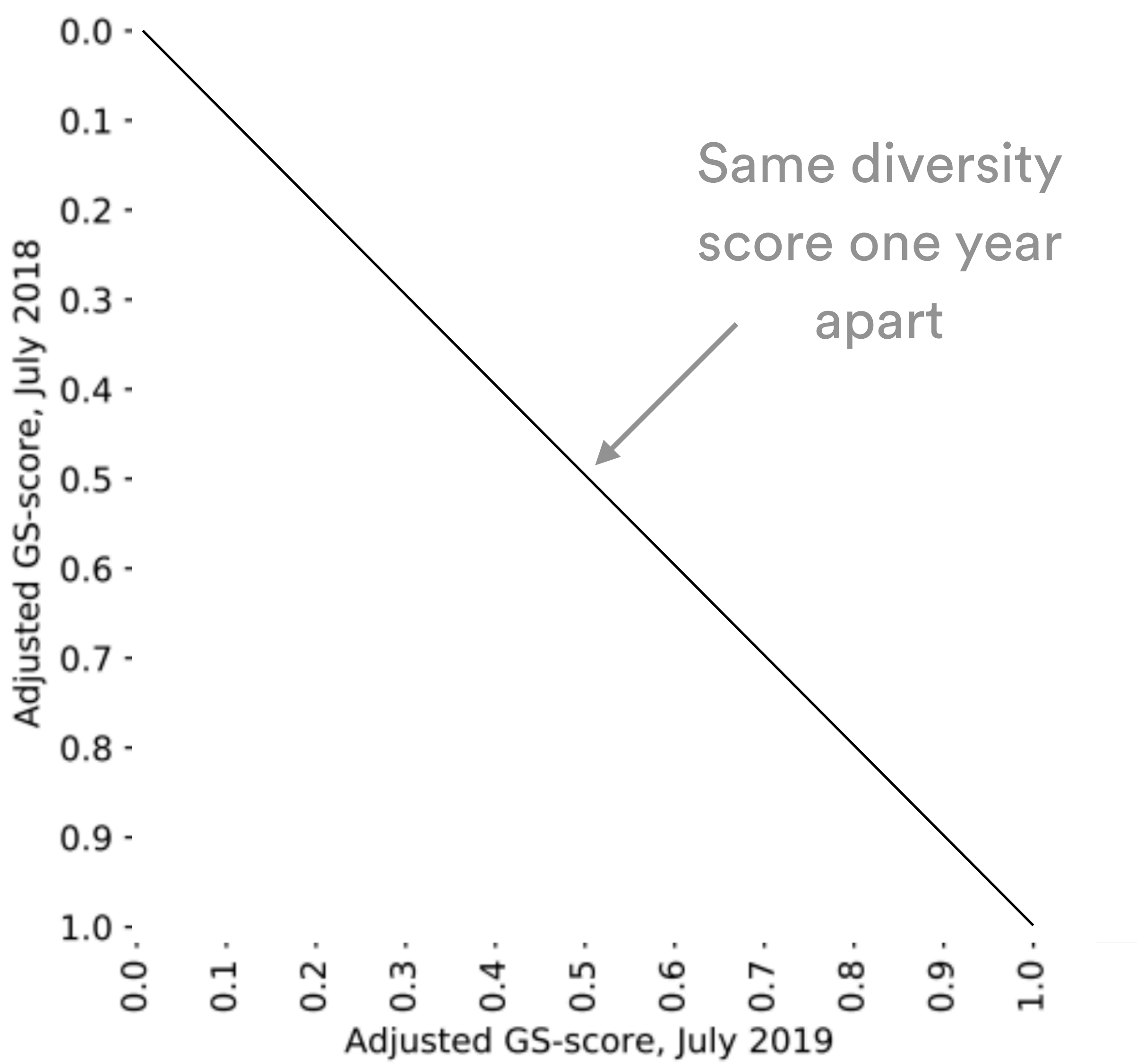


more diverse ⟷ less diverse

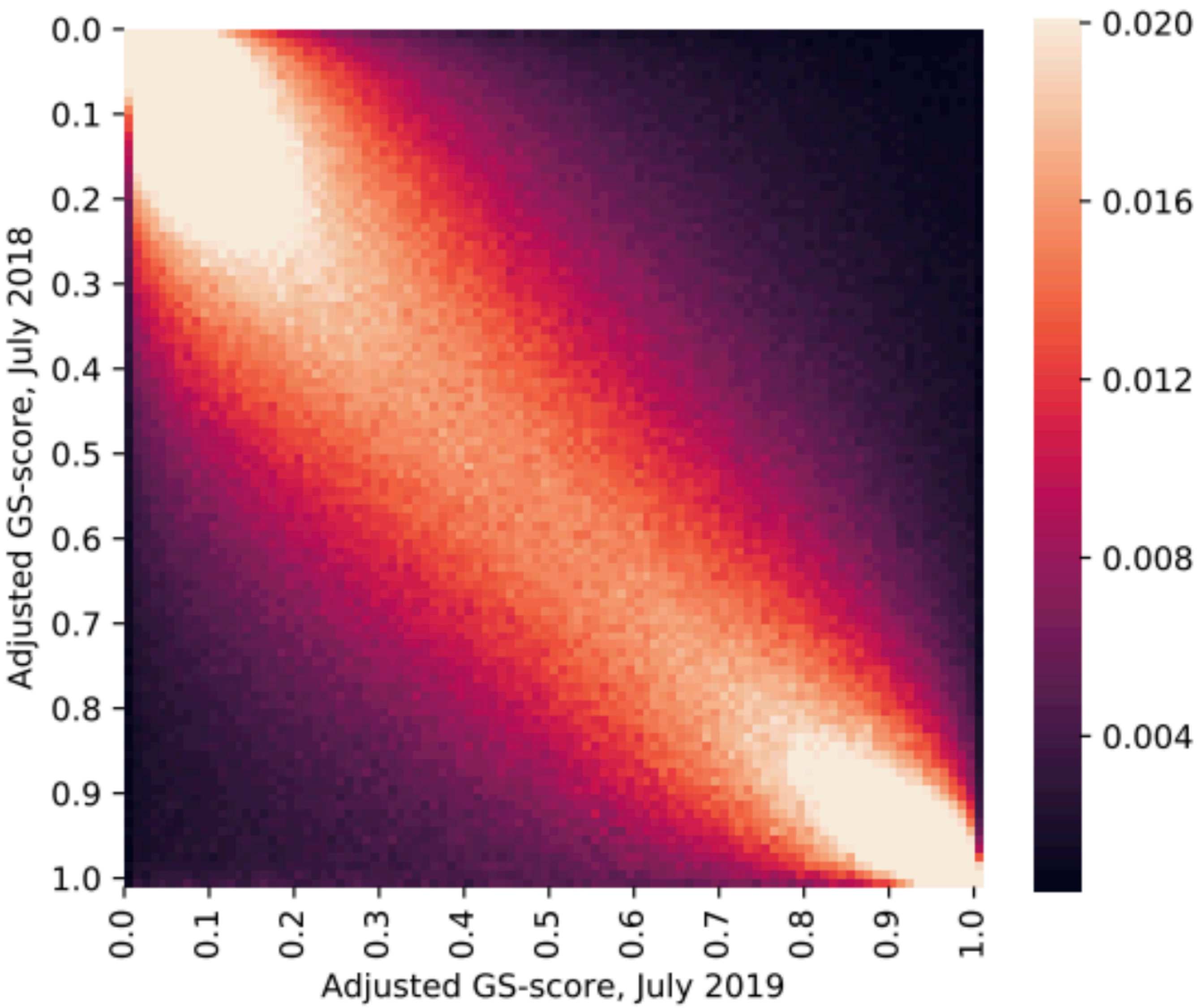
Wide range of consumption styles, from extremely narrow to extremely broad

Lowest activity users are more narrow, but activity and diversity are mostly uncorrelated

How stable is the diversity of a user's consumption over time?



How stable is the diversity of a user's consumption over time?

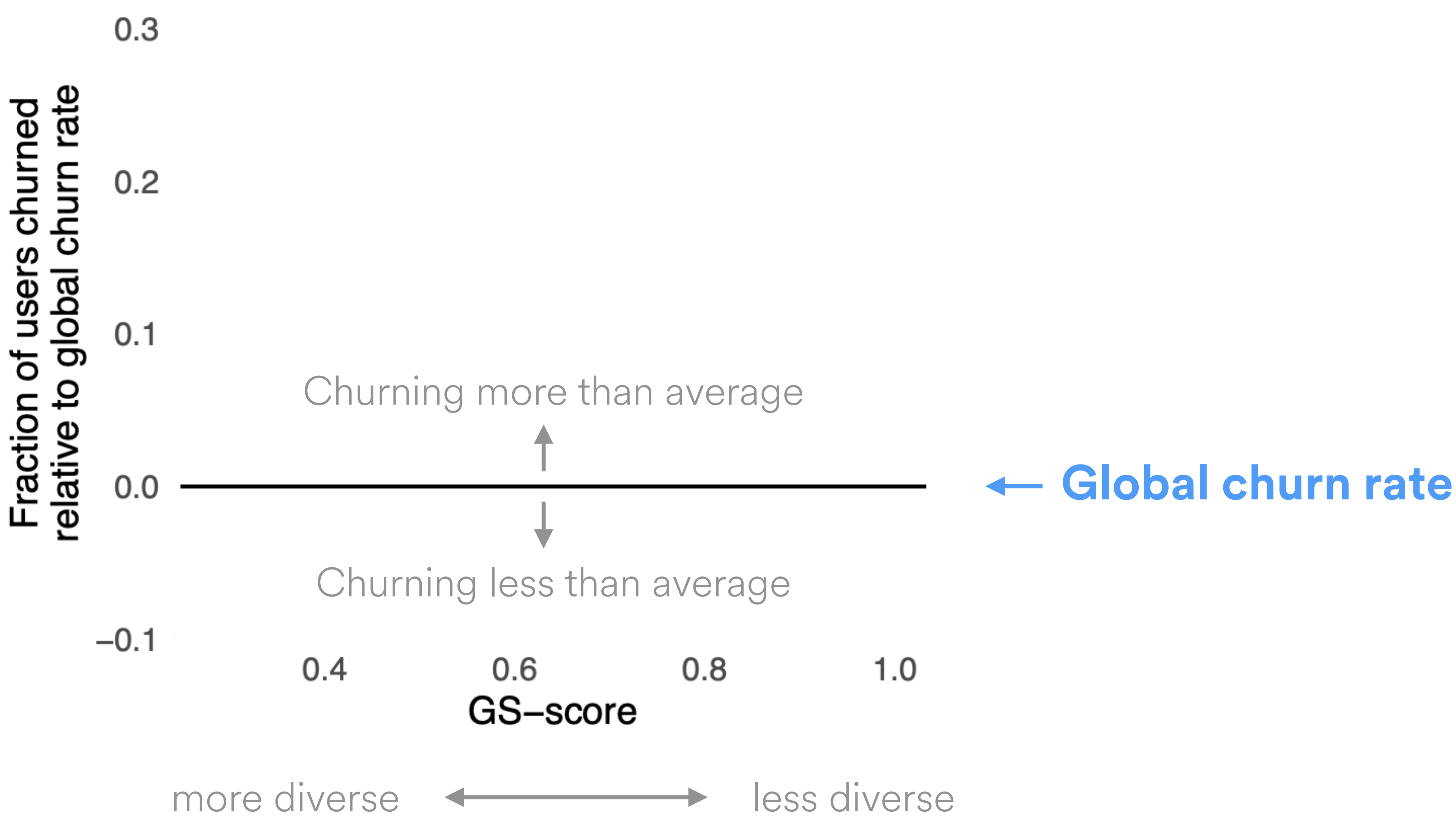


One year apart, user consumption diversity scores are similar, especially at the extremes

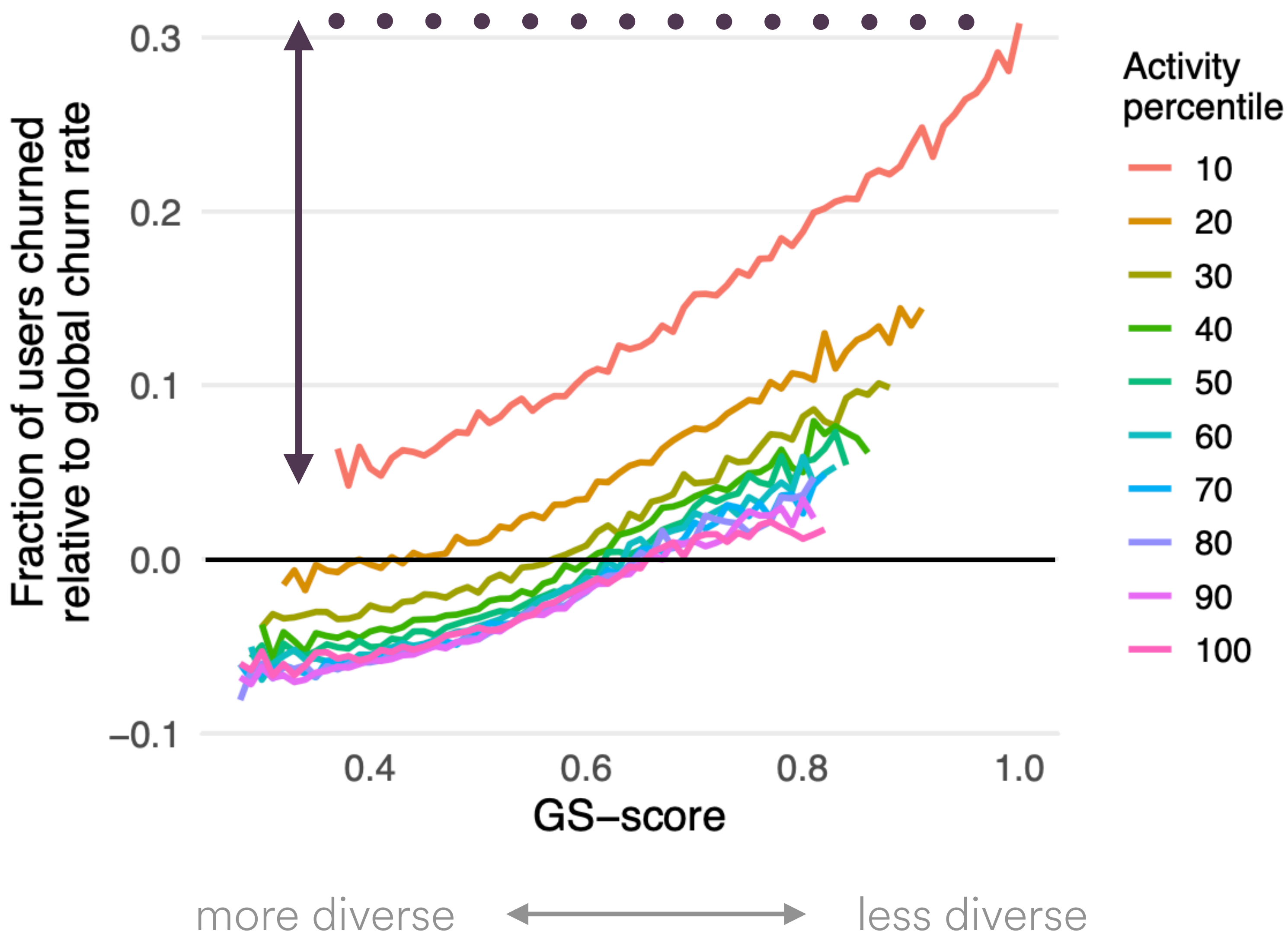
User consumption diversity is typically **stable over time**

What are the relationships between consumption diversity and important user outcomes?

How is diversity associated with churn?



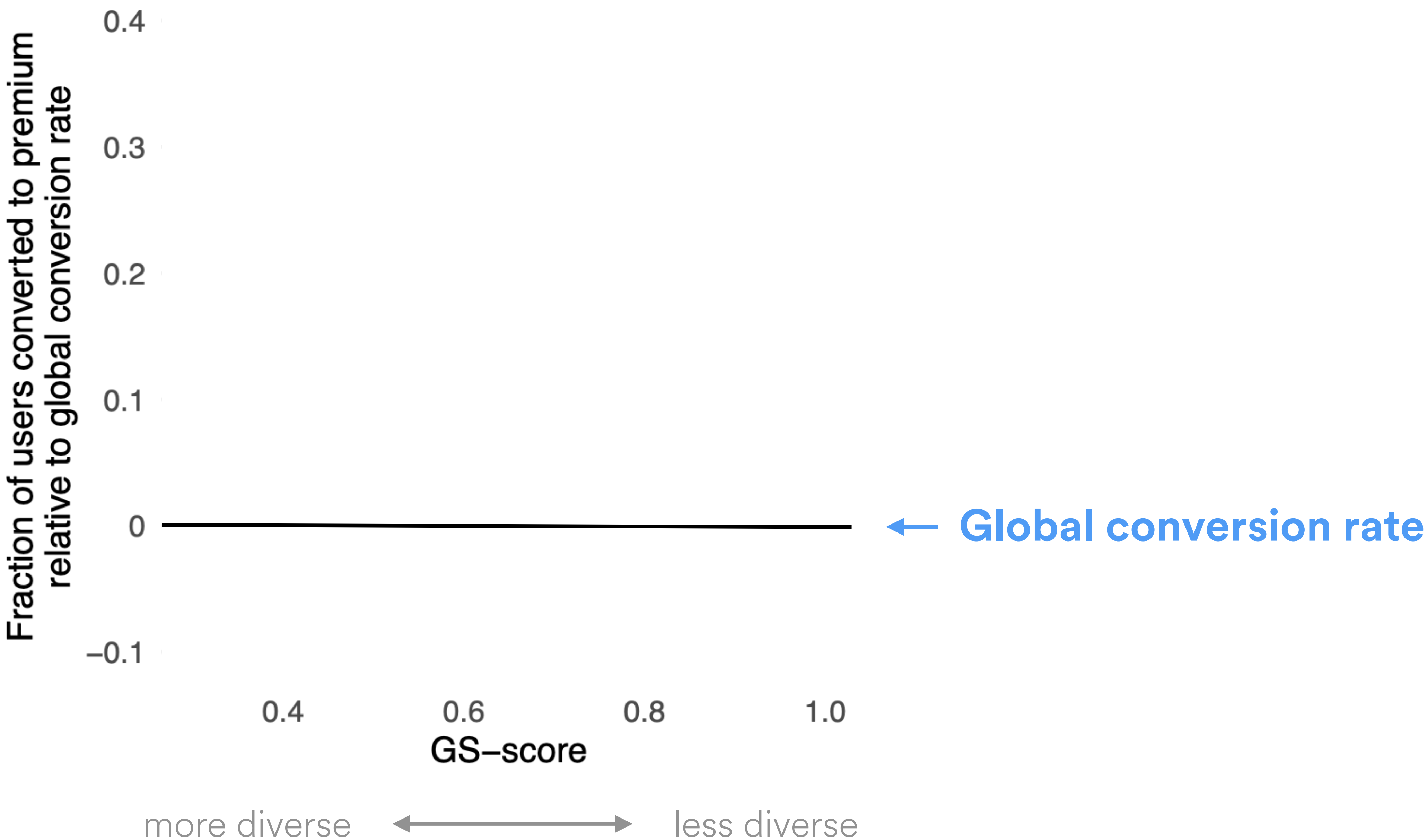
How is diversity associated with churn?



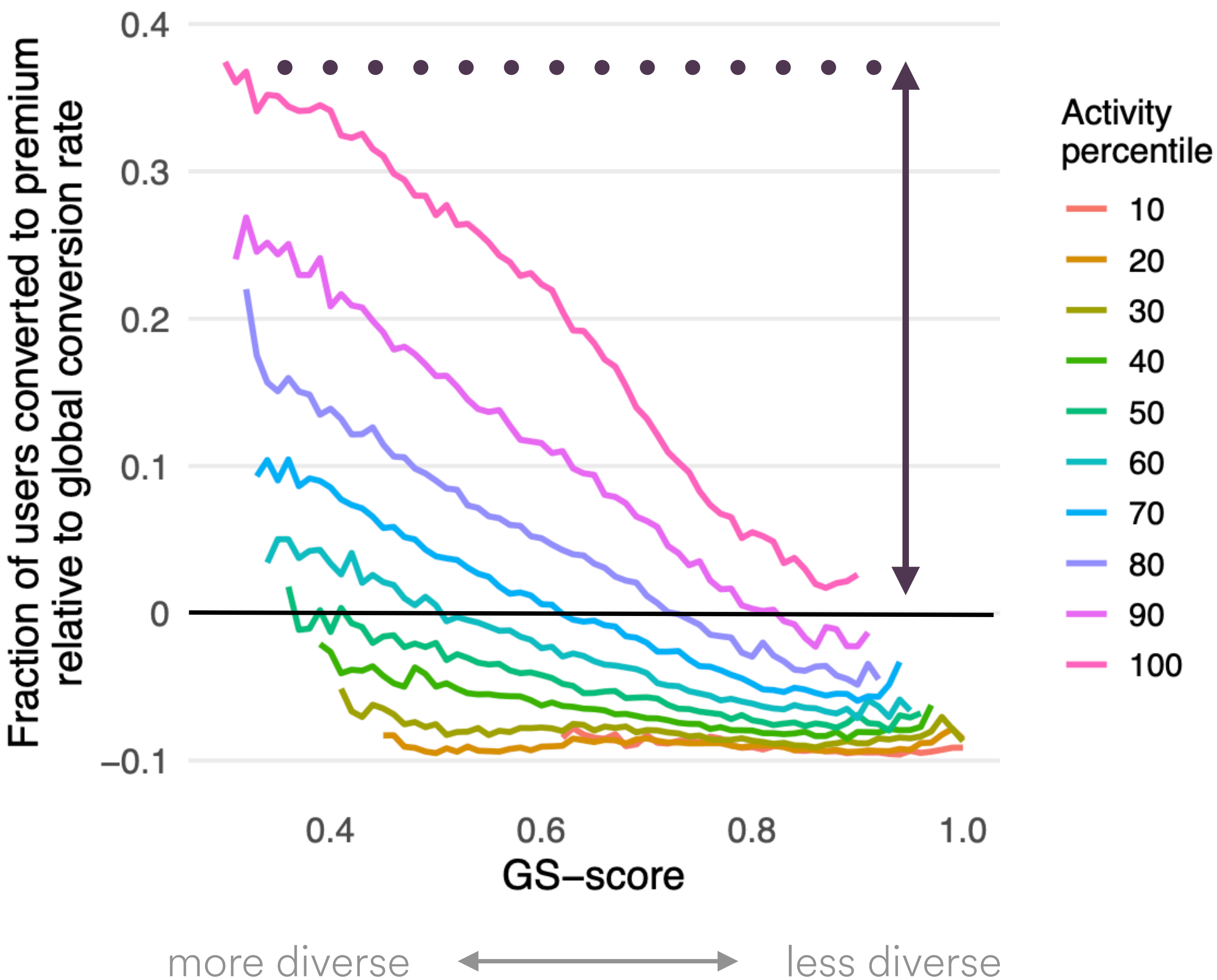
25 percentage points decrease in churn

Even controlling for activity, more diverse users are **far less likely** to churn

How is diversity associated with conversion?



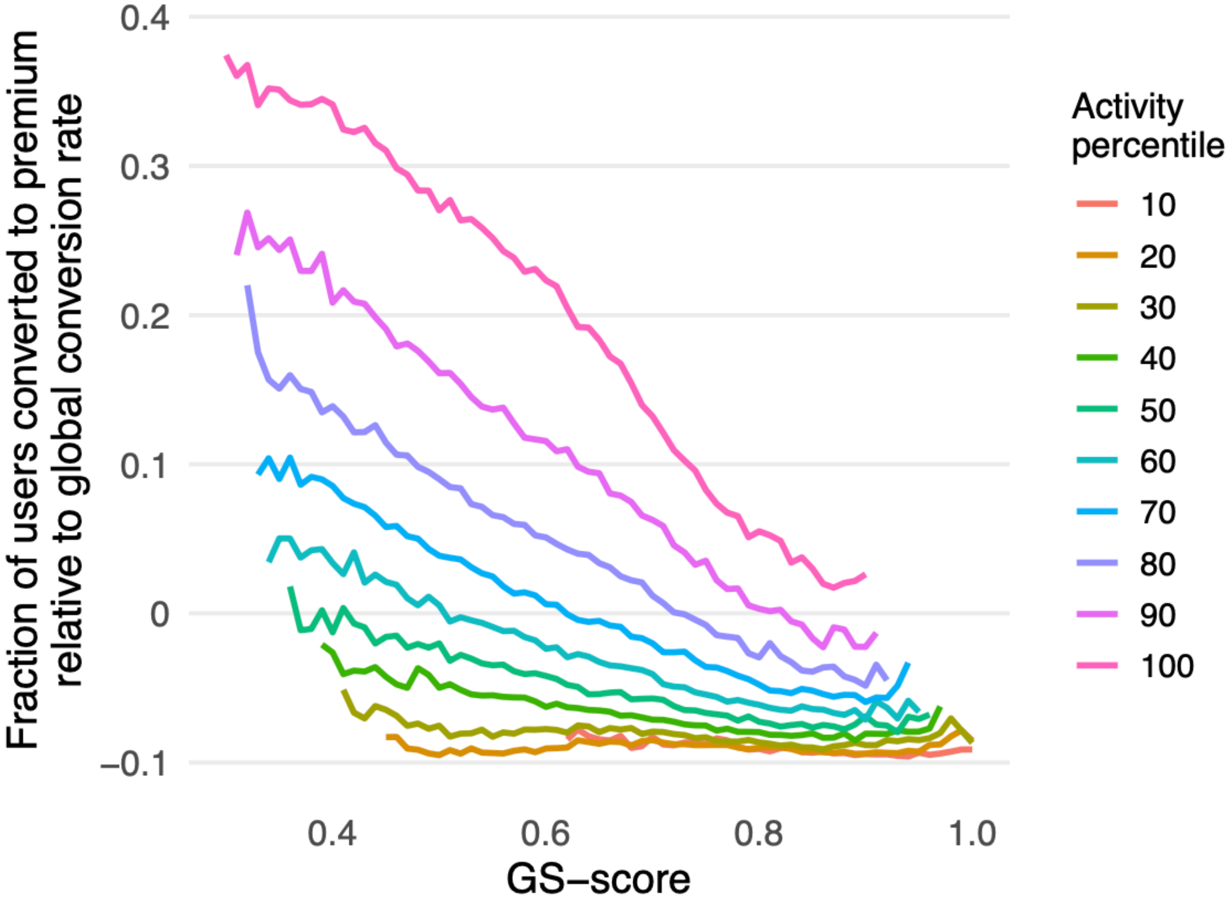
How is diversity associated with conversion?

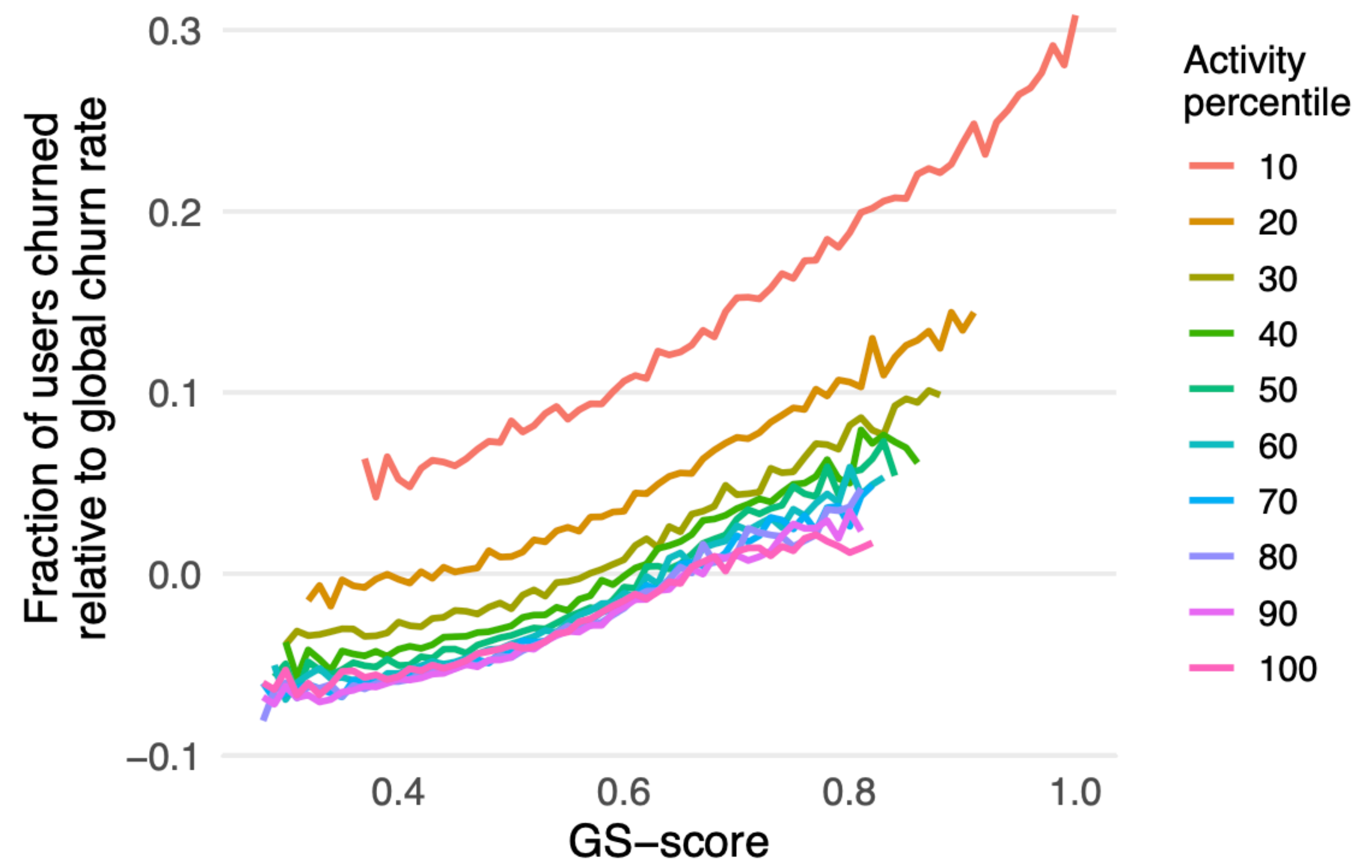


30 percentage points increase in conversion

Controlling for activity, more diverse users are **far more likely** to become premium members

Diverse listeners...





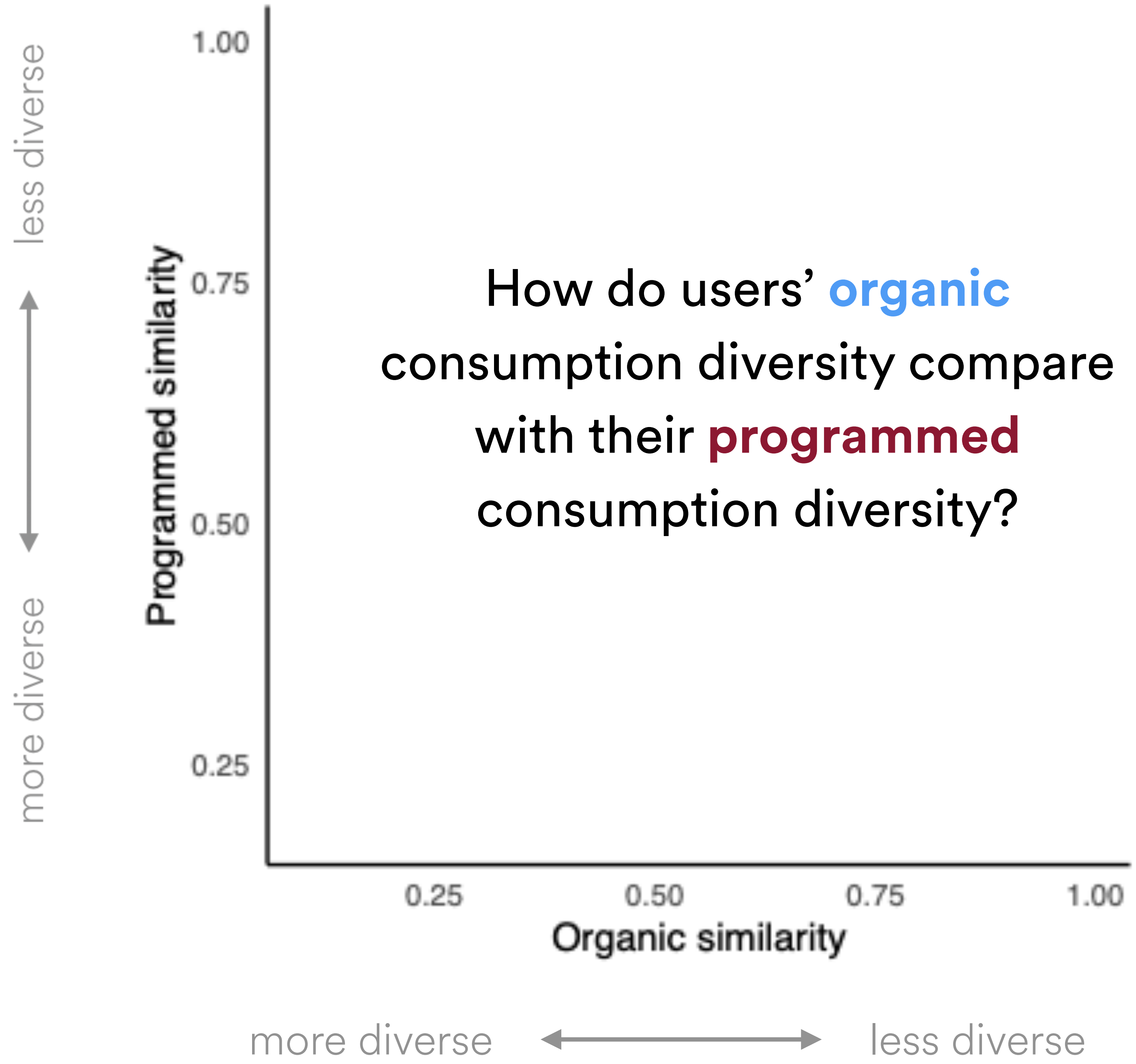
Convert more

and churn less

(controlling for activity!)

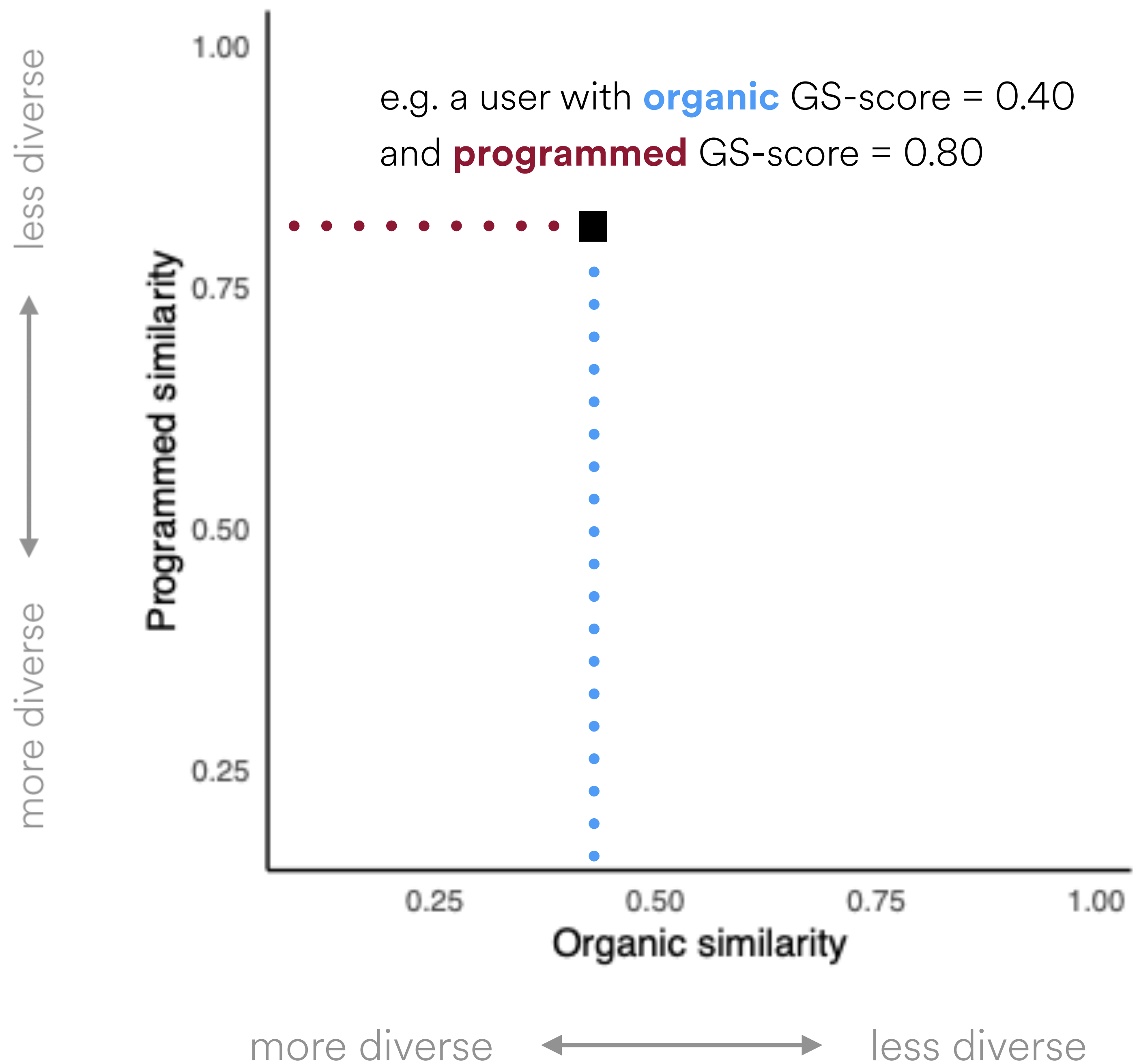
How are algorithmic recommendations related to the diversity of consumption?

Organic versus programmed consumption



How do users' **organic** consumption diversity compare with their **programmed** consumption diversity?

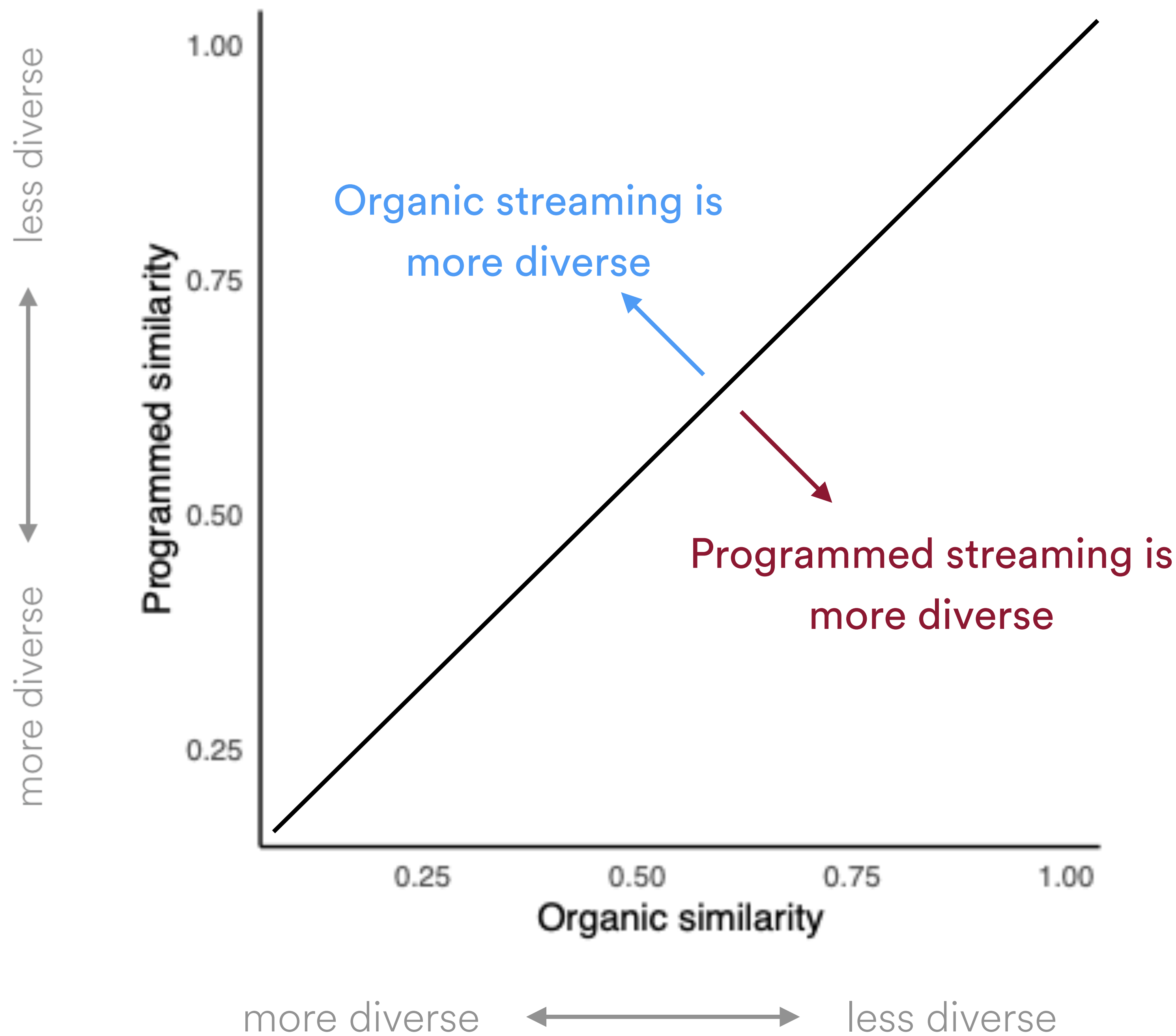
Organic versus programmed consumption



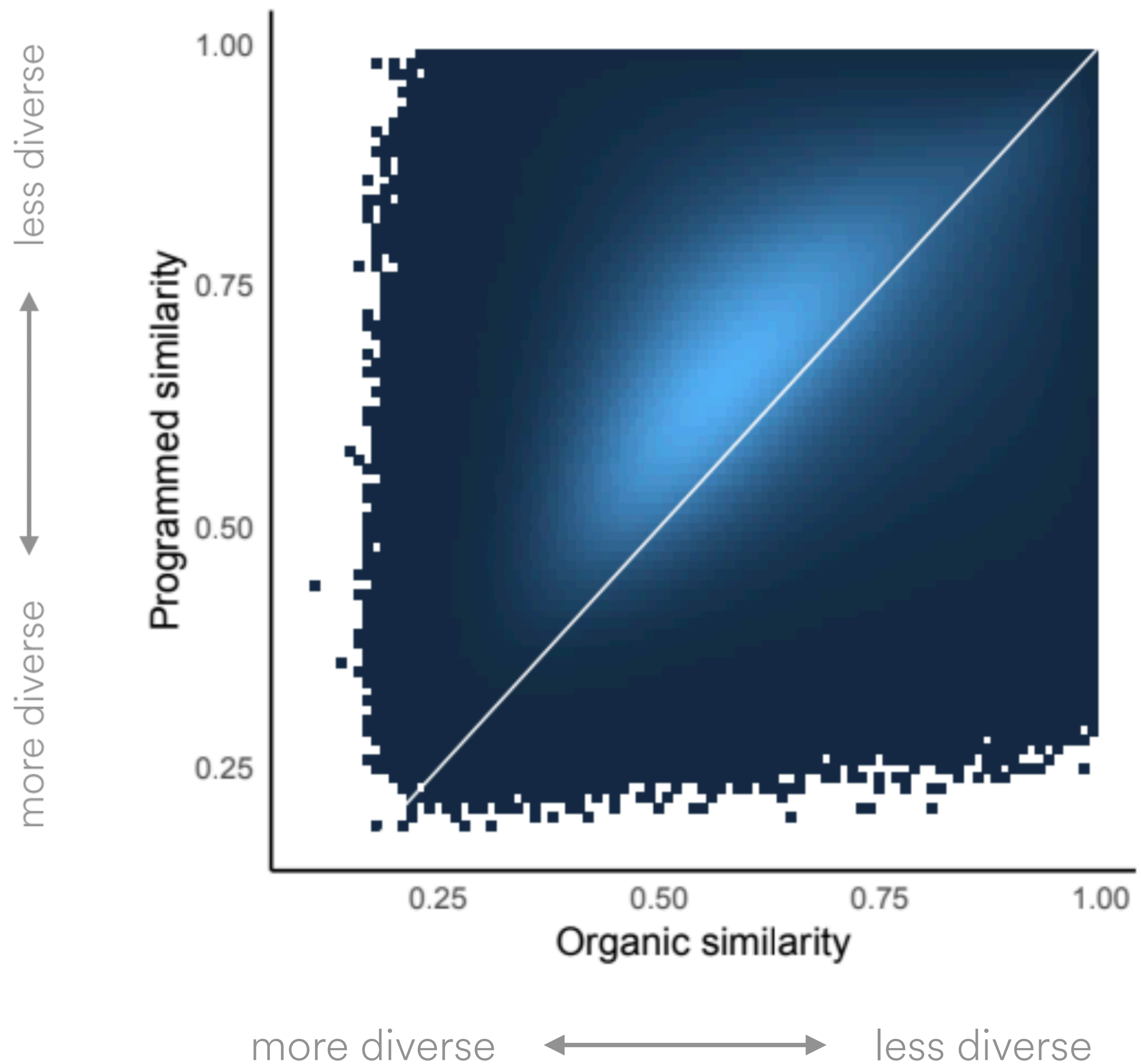
For each user, calculate **separate** diversity scores for their **organic** streaming only and their **programmed** streaming only

Examine joint distribution over all users to observe whether **organic** streaming tends to be **more or less diverse** than **programmed** streaming

Organic versus programmed consumption

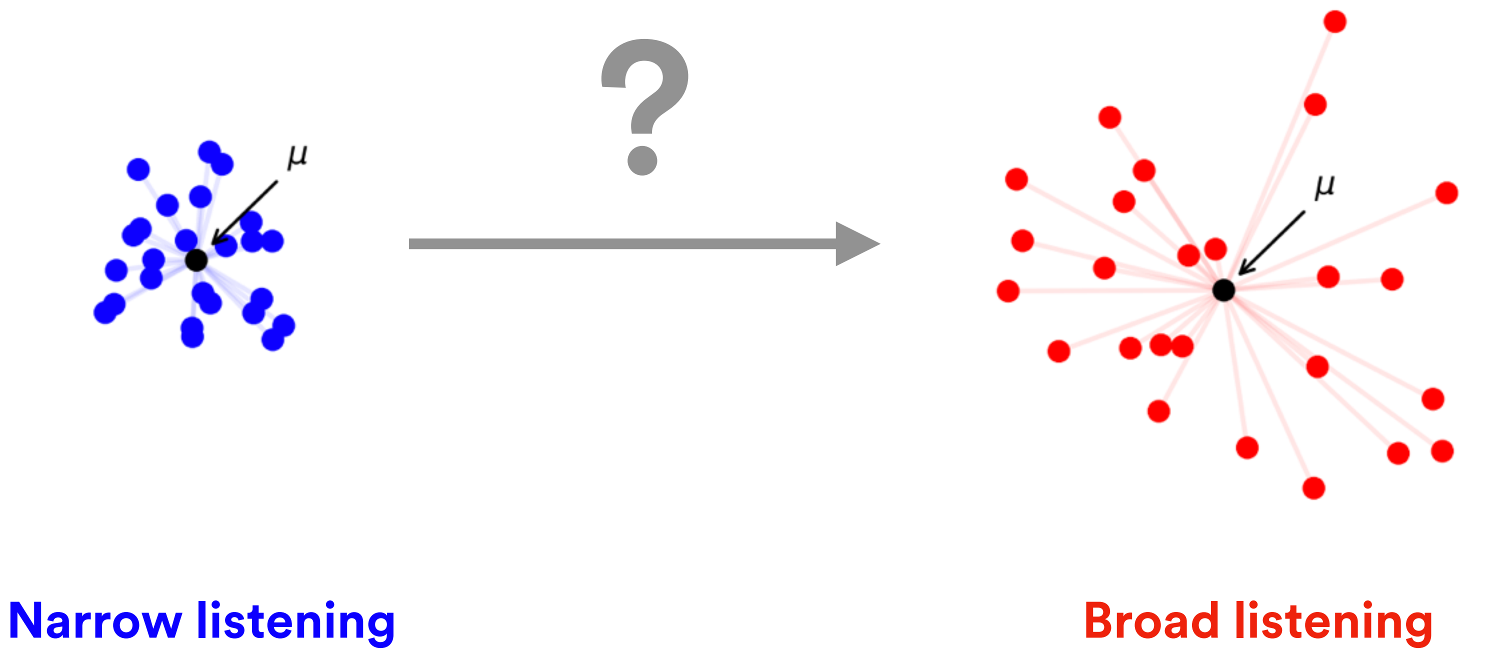


Organic versus programmed consumption



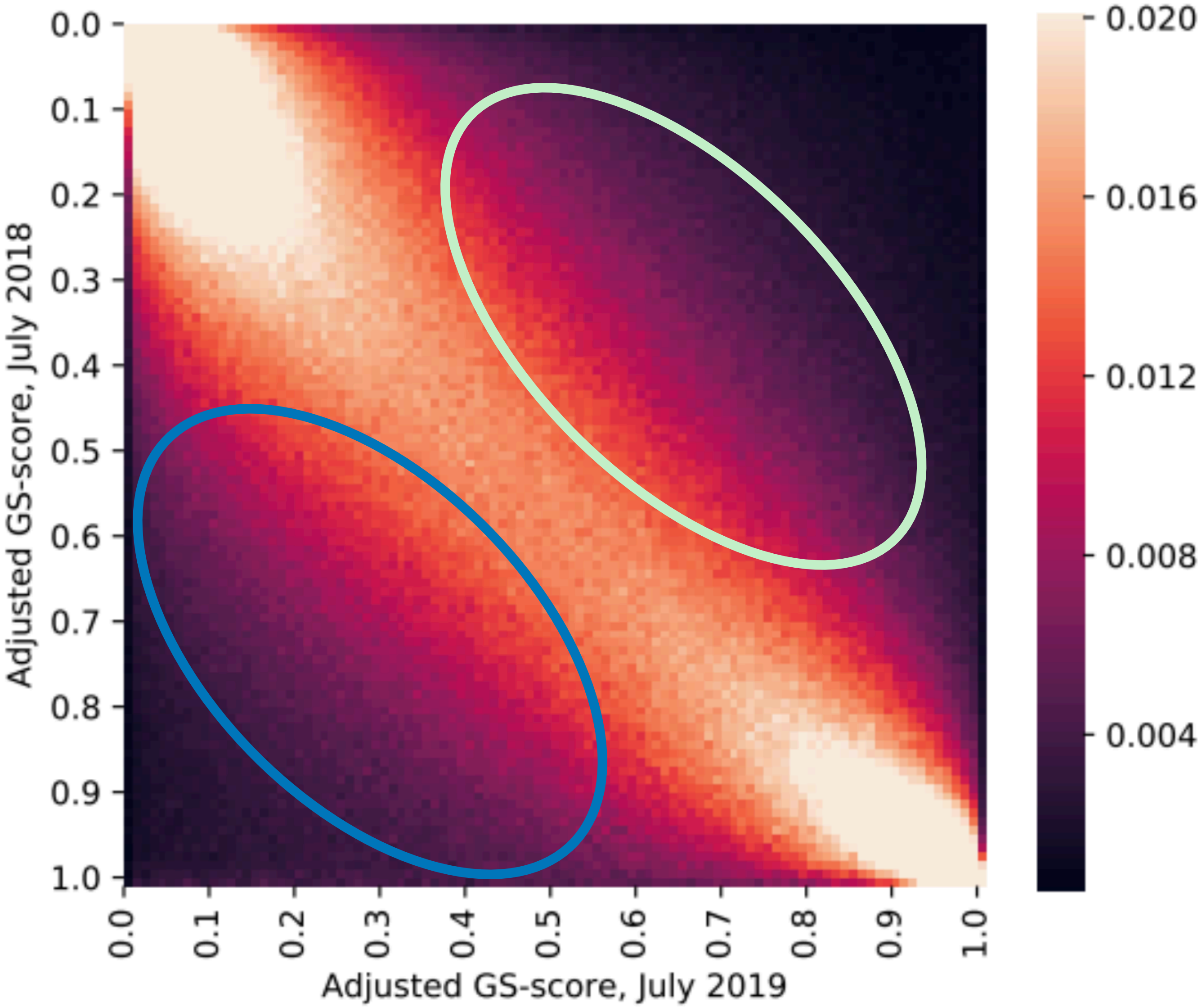
For most users, their **organic** consumption is more diverse than their **programmed** consumption

How does consumption become more diverse?



Dynamic view: when users change the diversity of their consumption over time, how do they do this?

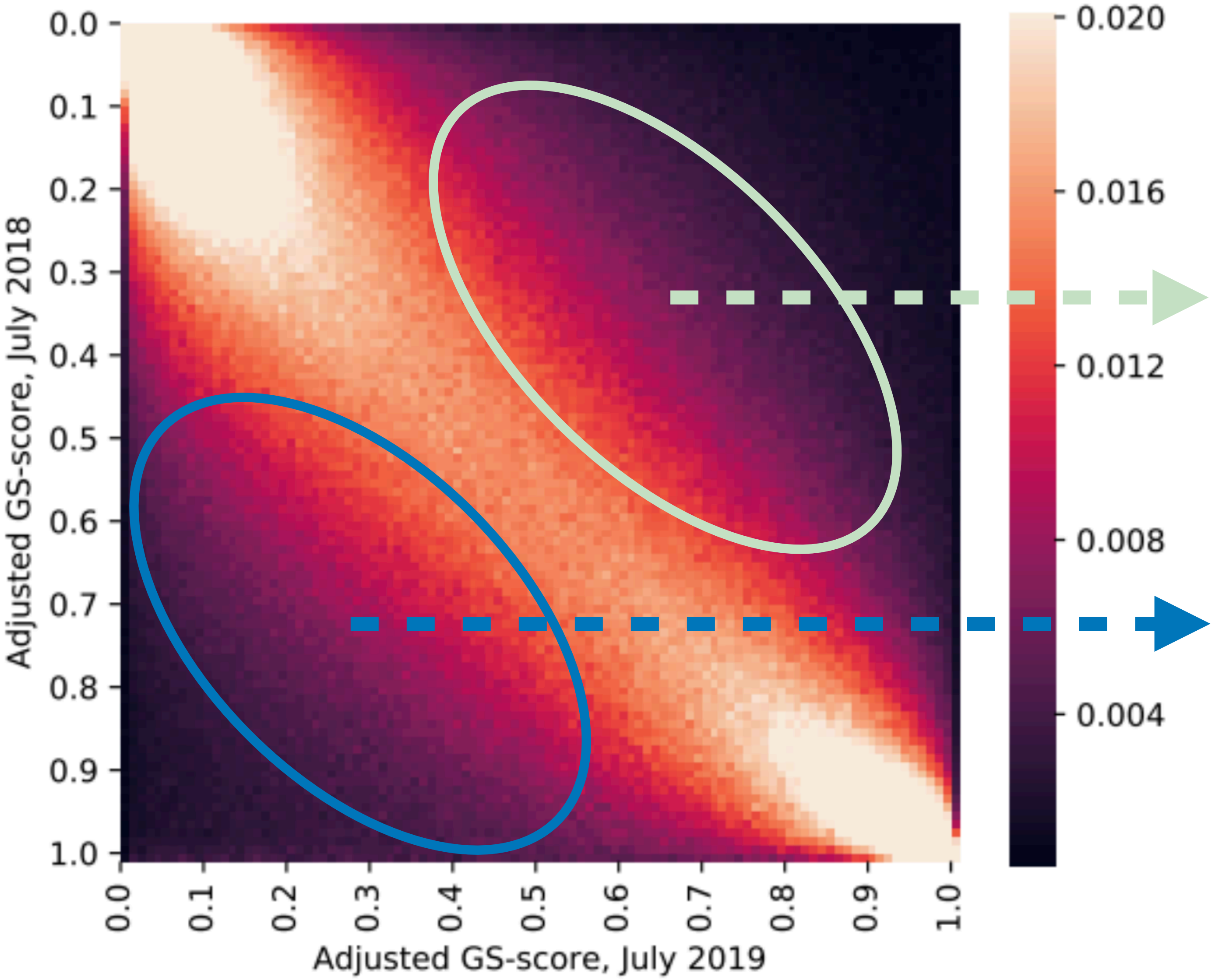
How does consumption become more diverse?



Although consumption diversity is typically stable, many users still **change the diversity of their listening substantially** from one year to the next

What changes in consumption patterns drive these changes in diversity?

How does consumption become more diverse?

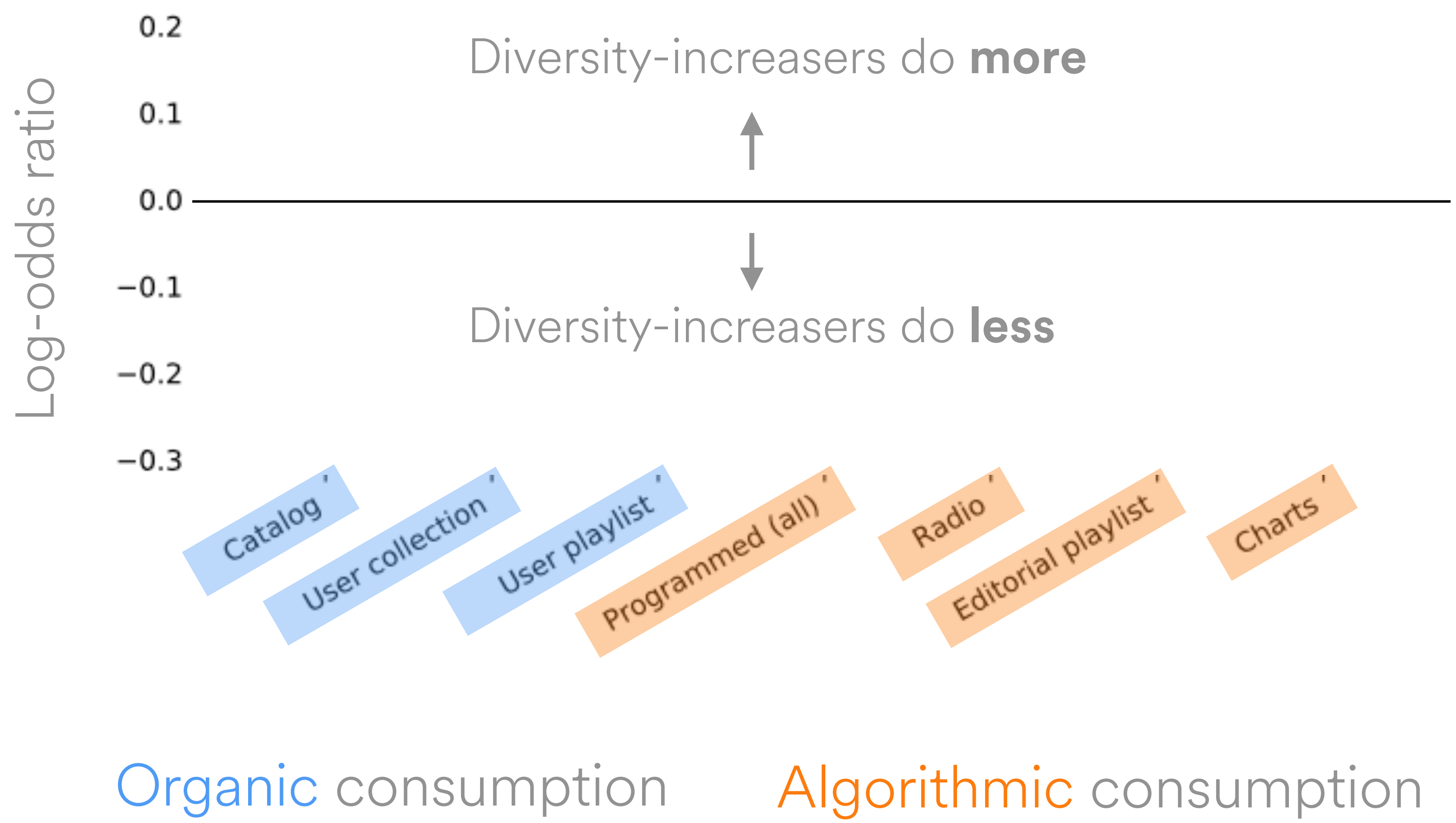


"Diversity-decreasers": those who went from broad to narrow consumption

"Diversity-increasers": those who went from narrow to broad consumption

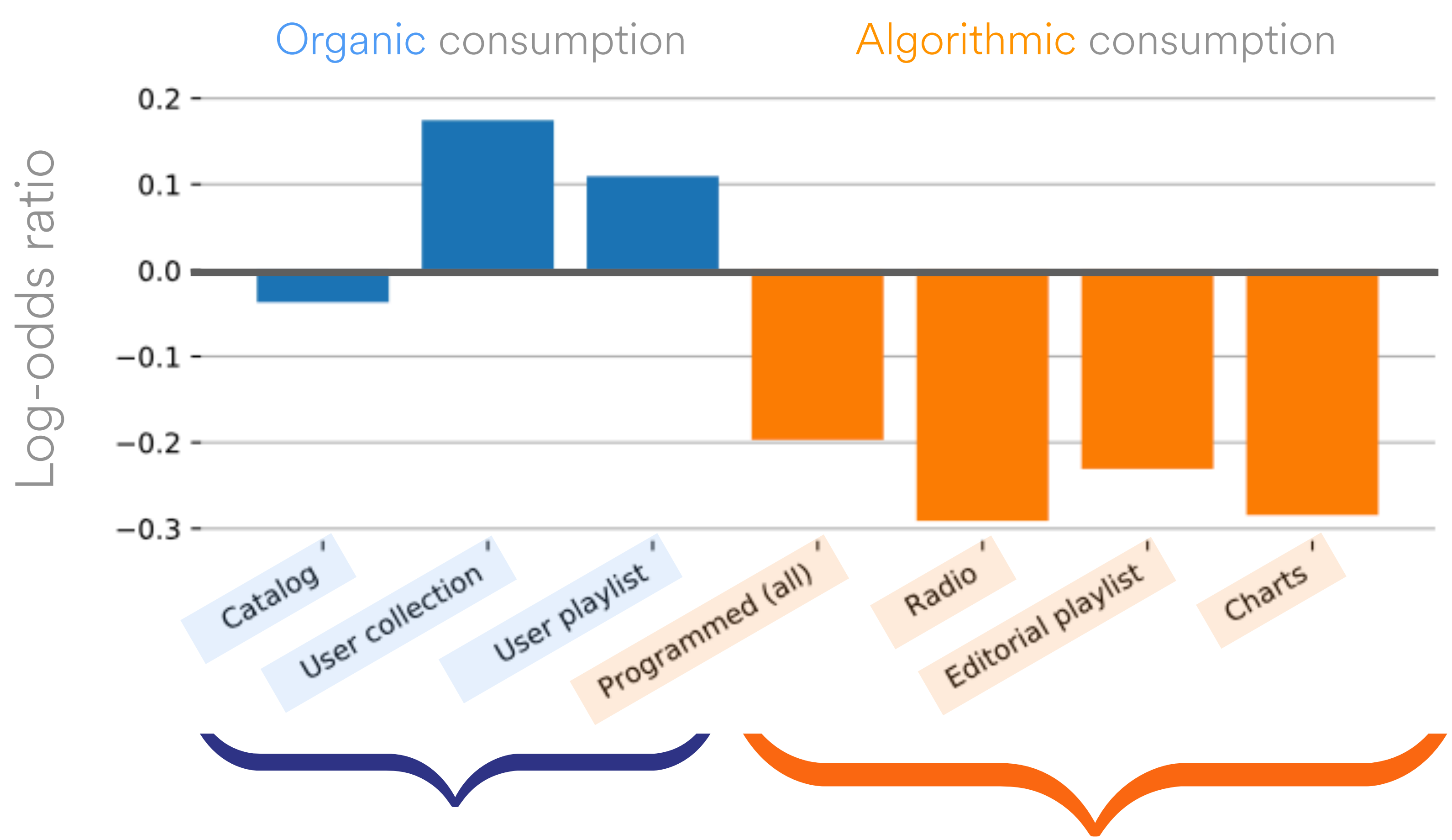
How does consumption become more diverse?

Analysis strategy: compare diversity-increasers to diversity-decreasers. How does listening between these two groups differ?



How does consumption become more diverse?

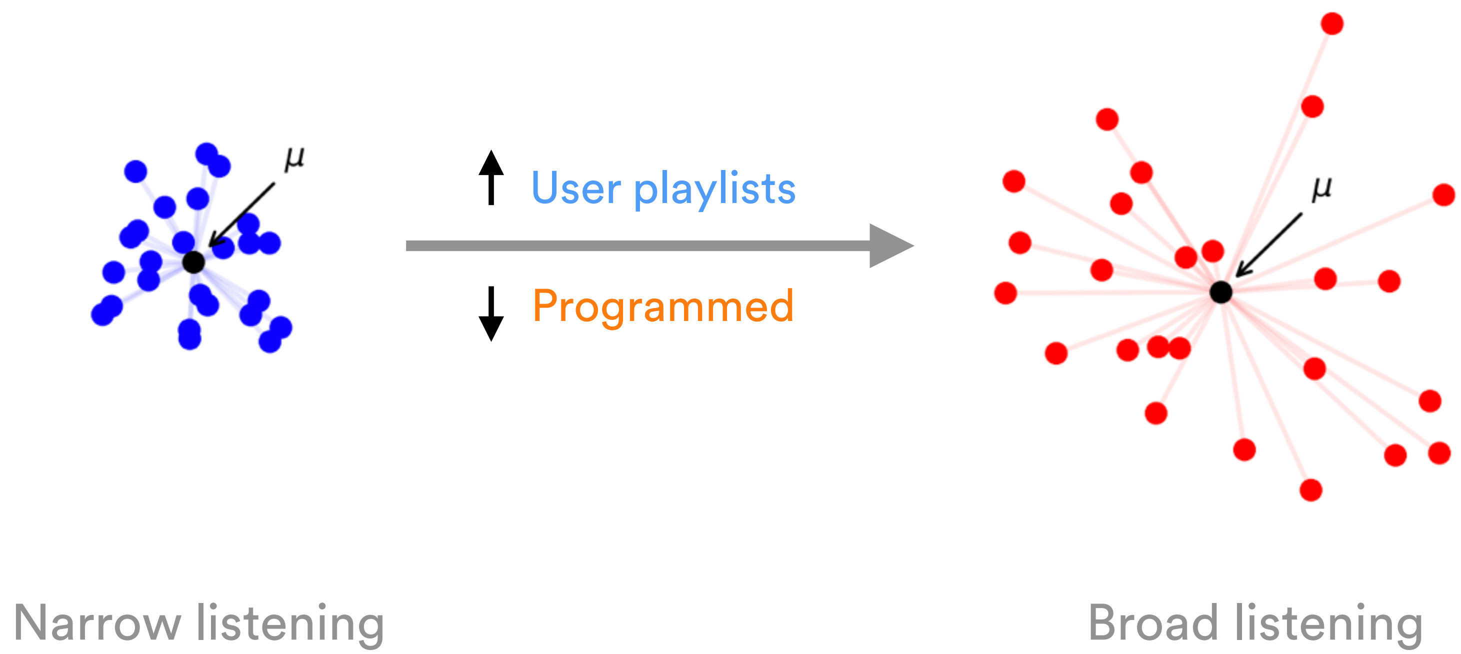
When users diversify their consumption over time, they:



Increase their organic listening

And decrease their algorithmic listening

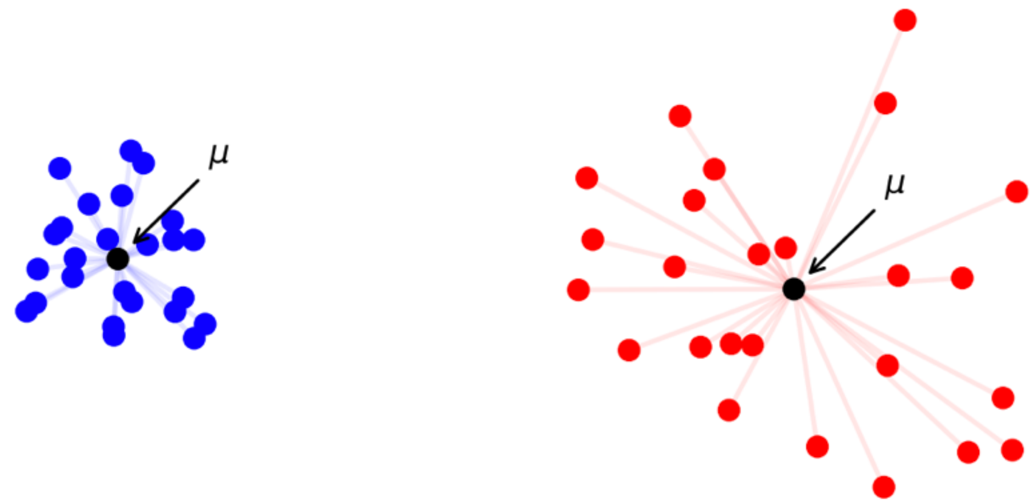
How does consumption become more diverse?



When users diversify their consumption, they choose more for themselves and rely less on algorithmic recommendations

How do users respond to algorithmic recommendations based on their consumption diversity?

Recommendation algorithms and consumption diversity



How do generalists and specialists respond to algorithmic recommendations?

Recommendation algorithms and consumption diversity

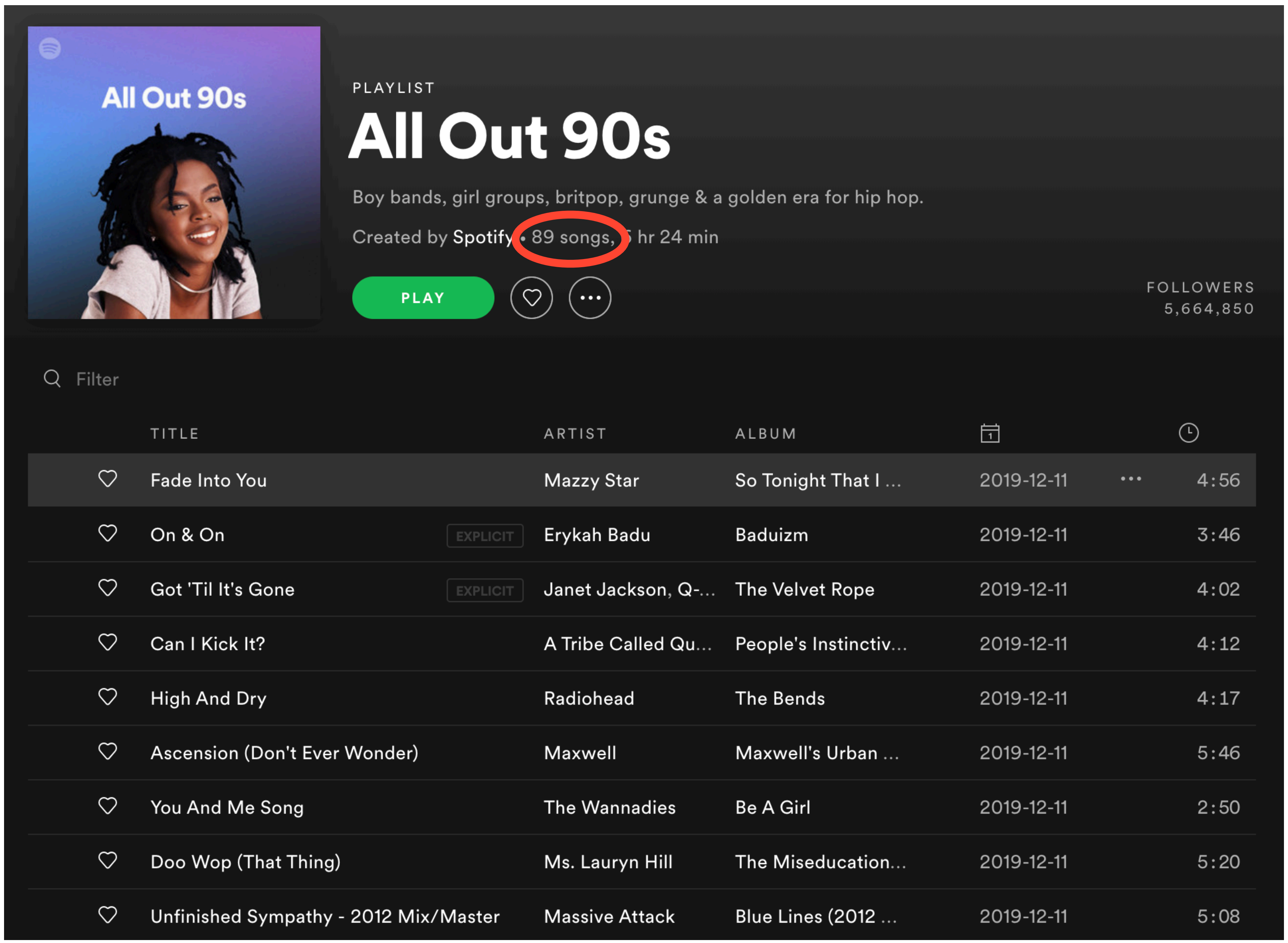
We ran a randomized experiment comparing how recommendation algorithms affect users based on their consumption diversity

Experimental context:

- Free users engaging with 7 popular playlists
- When they stream from one of them, a song is chosen at random from the top k songs
- Algorithms rank the songs to try to improve the user's experience

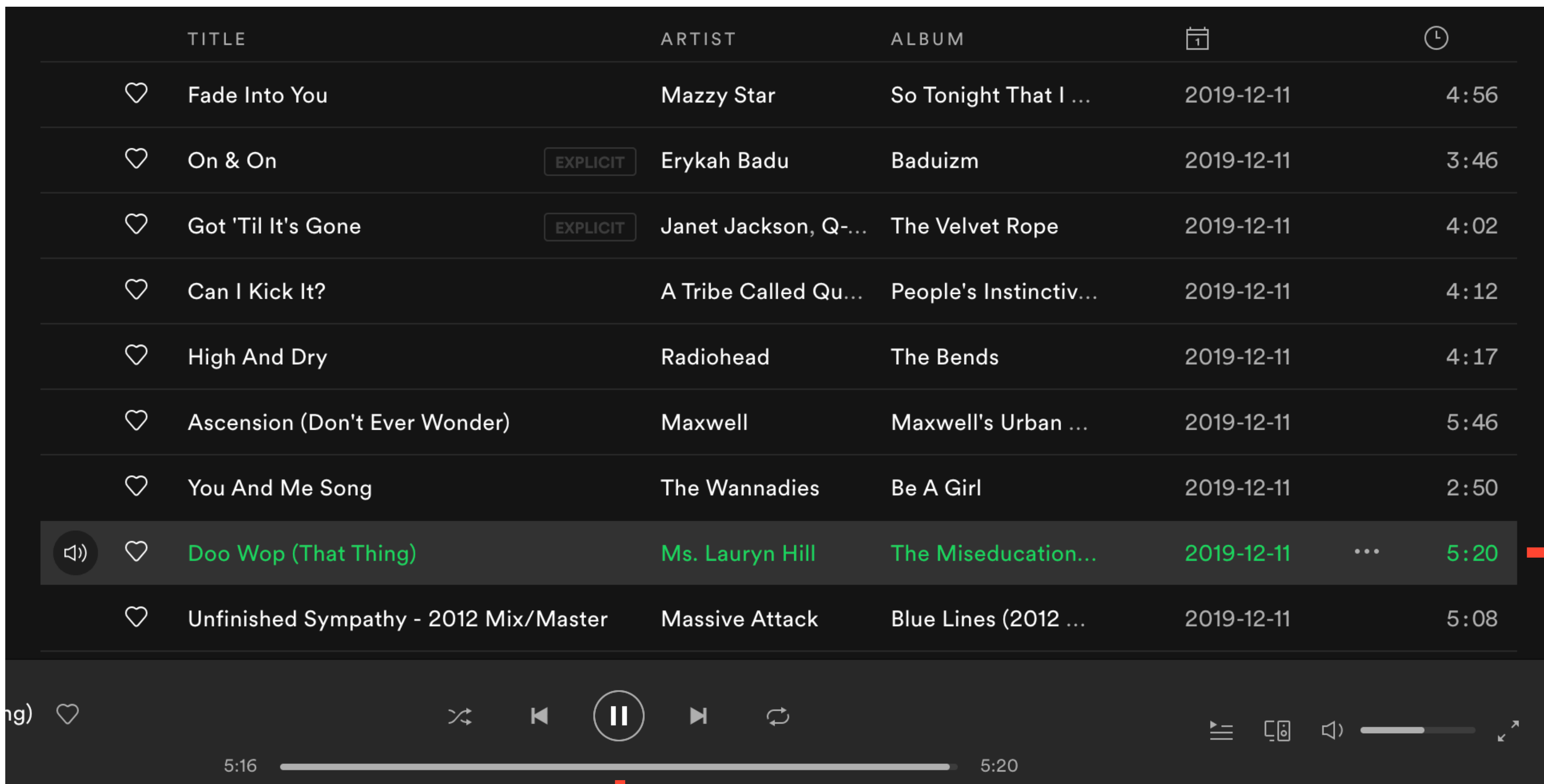
We randomly test 3 simple algorithms and measure their performance on generalists vs. specialists

Recommendation algorithms and consumption diversity



Rank the songs
in the playlist

Recommendation algorithms and consumption diversity



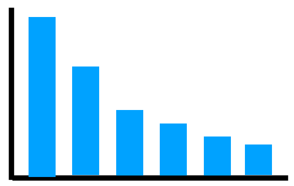
One of the top k songs plays at random

Measure number of plays (listening to the song) and skips (skipping the song)

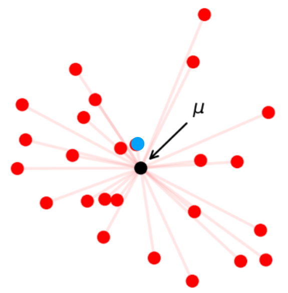
Recommendation algorithms and consumption diversity

3 ranking approaches:

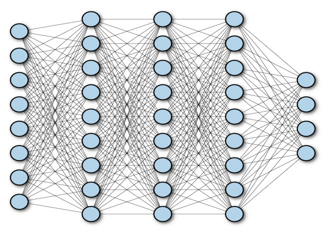
Popularity: rank songs by number of streams.
→ Un-personalized baseline.



Relevance: rank songs by proximity of track vector to user vector
→ Simple model of collaborative filtering algorithms.



Learned: neural regression model learned from user-level, song-level, and interaction-level features.
→ Learn from user's historical preferences.



User-level: country, affinity for various genres, user vector in embedding
Song-level: popularity, genres, song vector in embedding
Interaction-level: cosine similarity between user and song vectors (relevance), user's affinity for song's genres

Recommendation algorithms and consumption diversity

Experimental design:

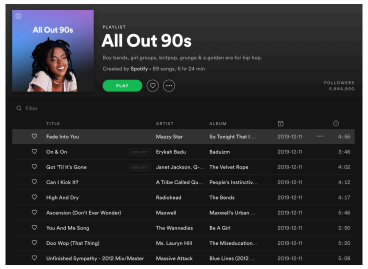
Randomly assign users to popularity, relevance, and learned conditions

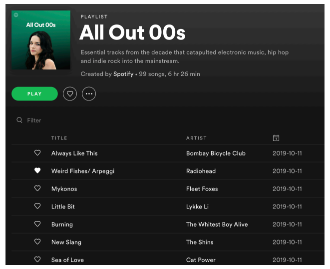
540,000 users during 1-week period

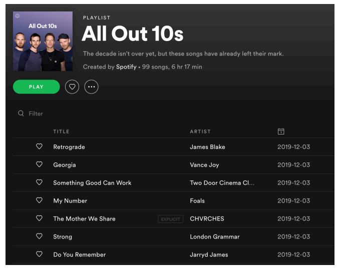
7 playlists: All Out 50s, 60s, 70s, 80s, 90s, 00s, 10s

Outcomes: streams and skips

Compare outcomes for generalists (high-diversity users) versus specialists (low-diversity users)







Recommendation algorithms and consumption diversity

Comparison	User Type	Song Streams	Song Skips
Relevance over Popularity	Generalists	+10.03%	+4.71%
	Specialists	+25.66%	+2.89%
Learned over Relevance	Generalists	+1.82%	+0.90%
	Specialists	+1.30%	−9.76%

1. Personalized algorithms deliver big short-term wins for all users, but **especially for specialists**.

2. Incorporating more information (learned) also benefits specialists more than generalists.

3. Algorithmic recommendations don't work as well for users with diverse consumption (generalists).

Diversity-aware recommendation

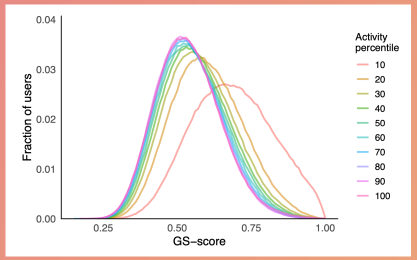
There is a clear need for diversity-aware recommendation algorithms.
→ Recommending content for specialists is different than recommending content for generalists.

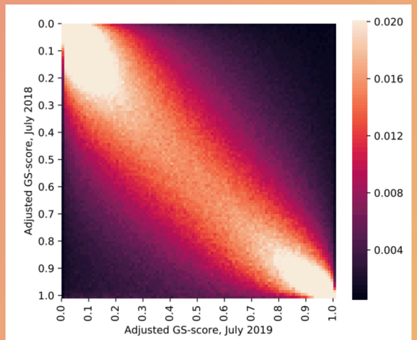
Recommendation algorithms may be over-optimizing for short-term goals at the expense of long-term goals.
→ A "grand challenge" for recommendation: How do you satisfy pressing user needs and keep the big picture in mind at the same time?

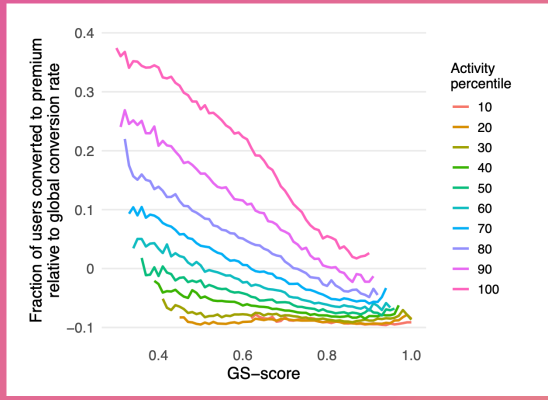
The causal effects of recommendation algorithms on consumption diversity and user outcomes are still unclear.
→ Our analyses are correlational.
→ A recent field experiment on Spotify suggests that personalized recommendation algorithms causally decrease diversity.

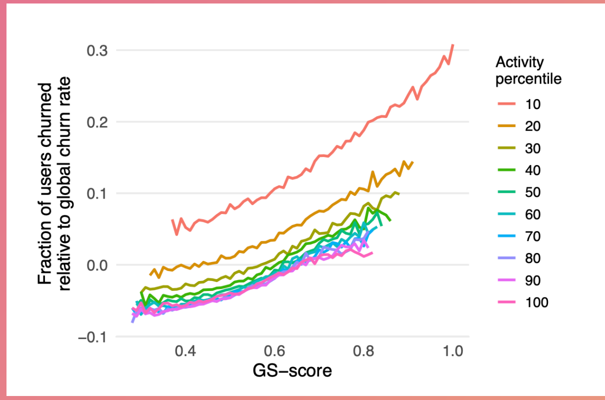
[1] David Holtz, Benjamin Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, Sinan Aral. *The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify.* Working paper.

Consumption diversity varies a lot between people but is typically stable within people.



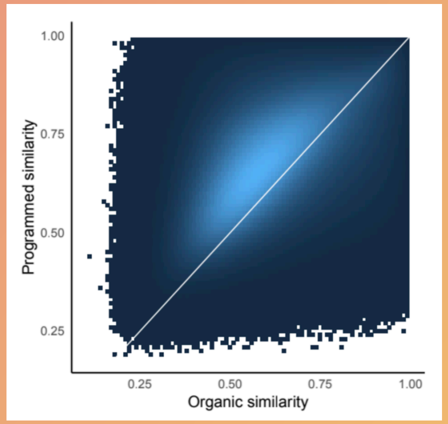


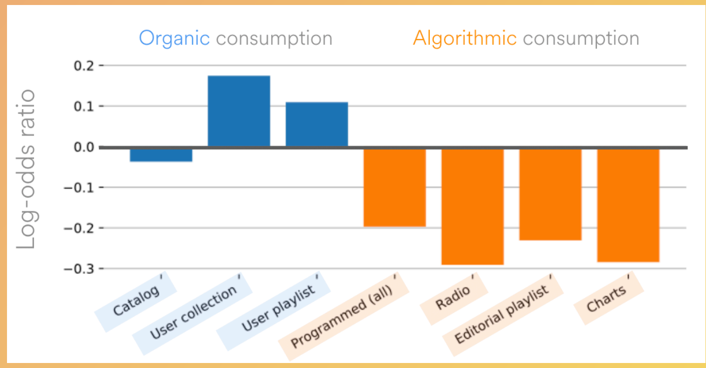




Diverse listeners convert more and churn less.

Recommendation algorithms are associated with reduced consumption diversity.





Comparison	User Type	Song Streams	Song Skips
Relevance over Popularity	Generalists	+10.03%	+4.71%
	Specialists	+25.66%	+2.89%
Learned over Relevance	Generalists	+1.82%	+0.90%
	Specialists	+1.30%	−9.76%

There is a need to develop diversity-aware recommendation algorithms.

Thank you!

ashtona@spotify.com