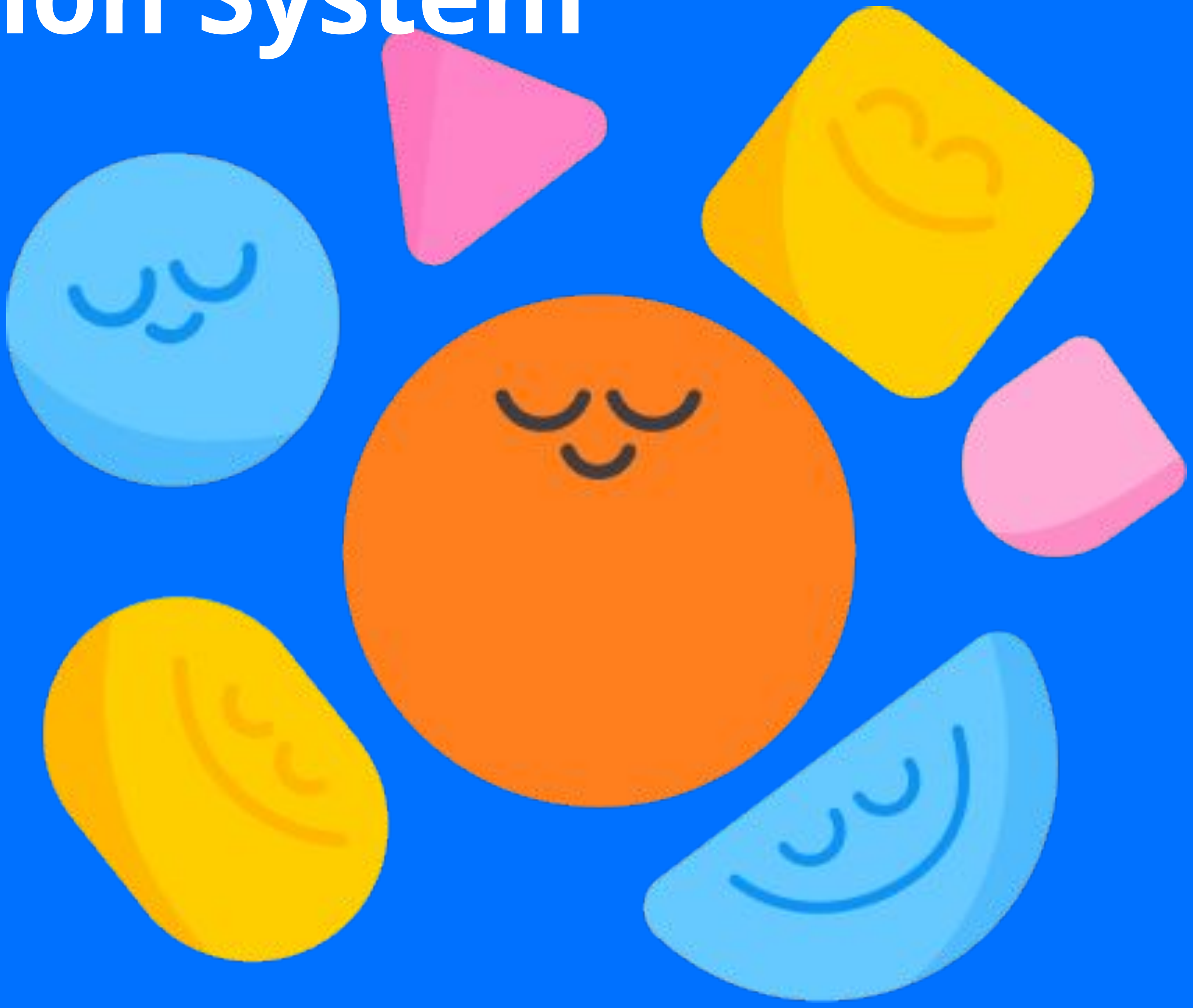
Lessons Learned from Building a Large-Scale Recommendation System at Headspace



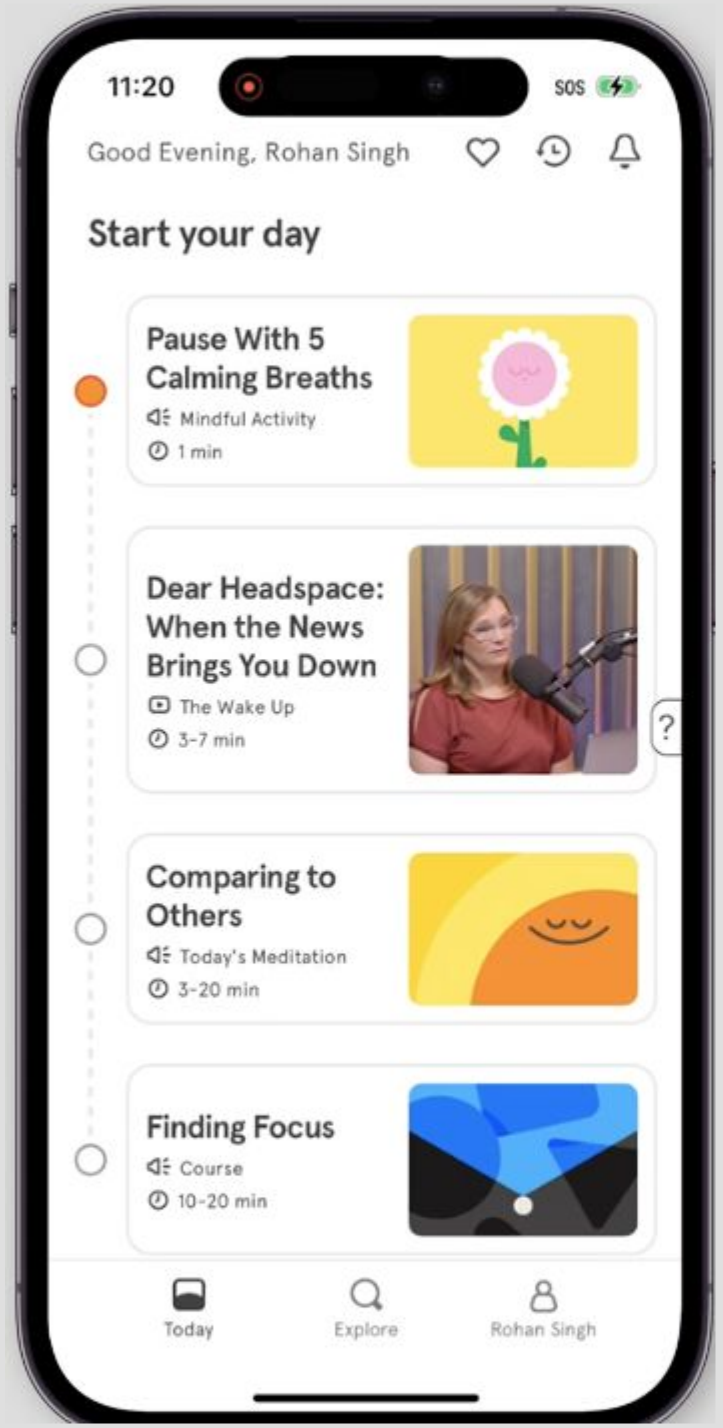
Rohan Singh Rajput
Senior Data Scientist
Headspace

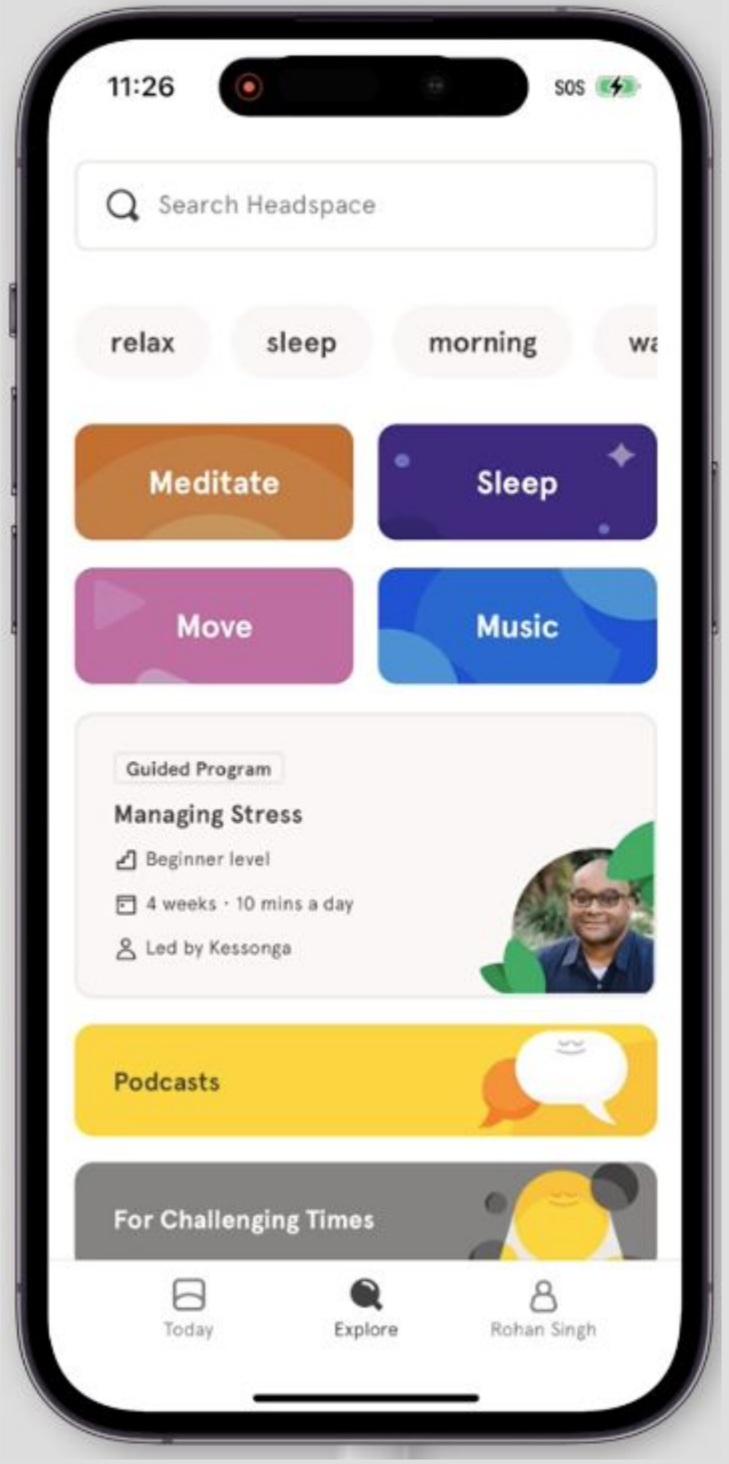
Agenda

- Introduction
- History
- Models
- Current State
- Challenges
- Opportunities

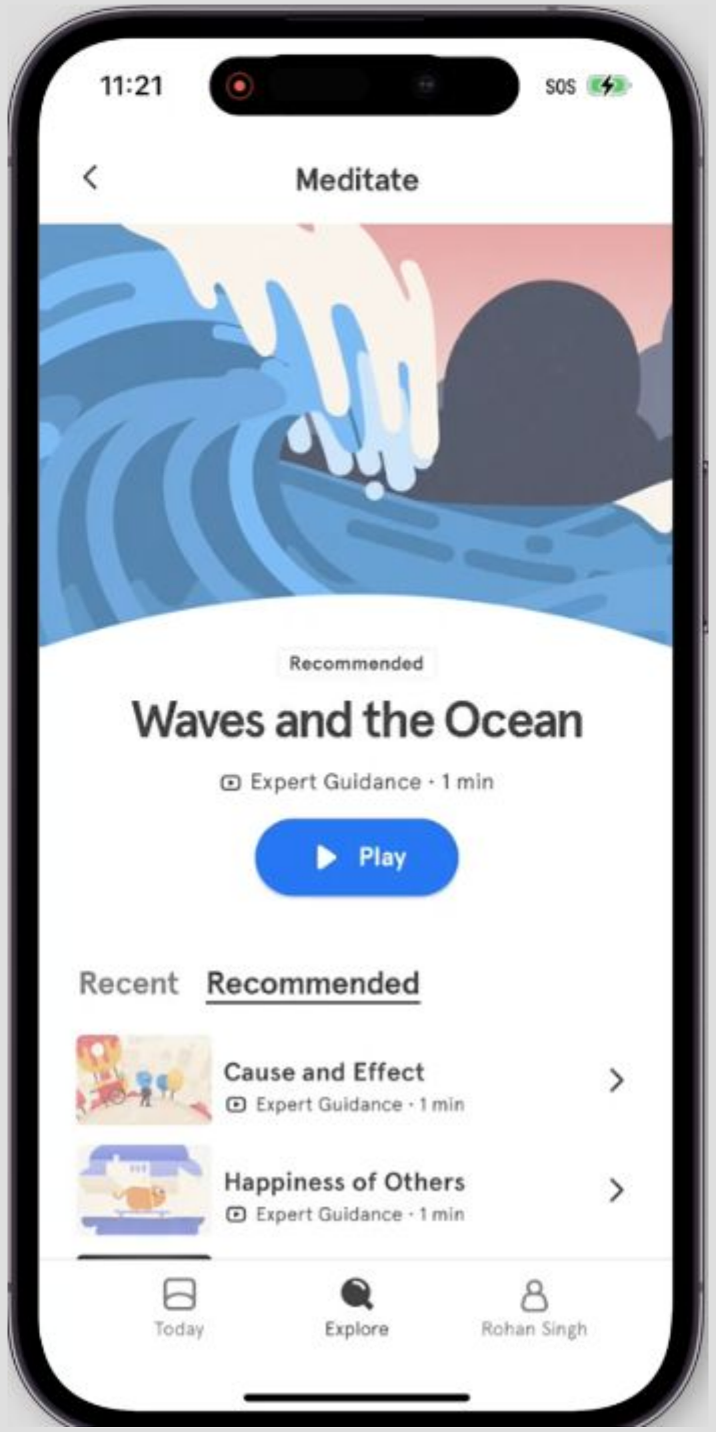
headspace

Headspace is an online meditation and mindfulness company


11:20
SOS
Good Evening, Rohan Singh
Start your day
Pause With 5 Calming Breaths
Mindful Activity
1 min
Dear Headspace: When the News Brings You Down
The Wake Up
3-7 min
Comparing to Others
Today's Meditation
3-20 min
Finding Focus
Course
10-20 min
Today
Explore
Rohan Singh


11:26
SOS
Search Headspace
relax
sleep
morning
Meditate
Sleep
Move
Music
Guided Program
Managing Stress
Beginner level
4 weeks · 10 mins a day
Led by Kessonga
Podcasts
For Challenging Times
Today
Explore
Rohan Singh


11:22
SOS
Sleep
Recommended
Goodnight
Wind Down · 1-20 min
Play
Recent
Recommended
Sleeping
Wind Down · 3-10 min
Goodnight Body With Elmo
Expert Guidance · 3 min
Today
Explore
Rohan Singh


11:21
SOS
Meditate
Recommended
Waves and the Ocean
Expert Guidance · 1 min
Play
Recent
Recommended
Cause and Effect
Expert Guidance · 1 min
Happiness of Others
Expert Guidance · 1 min
Today
Explore
Rohan Singh

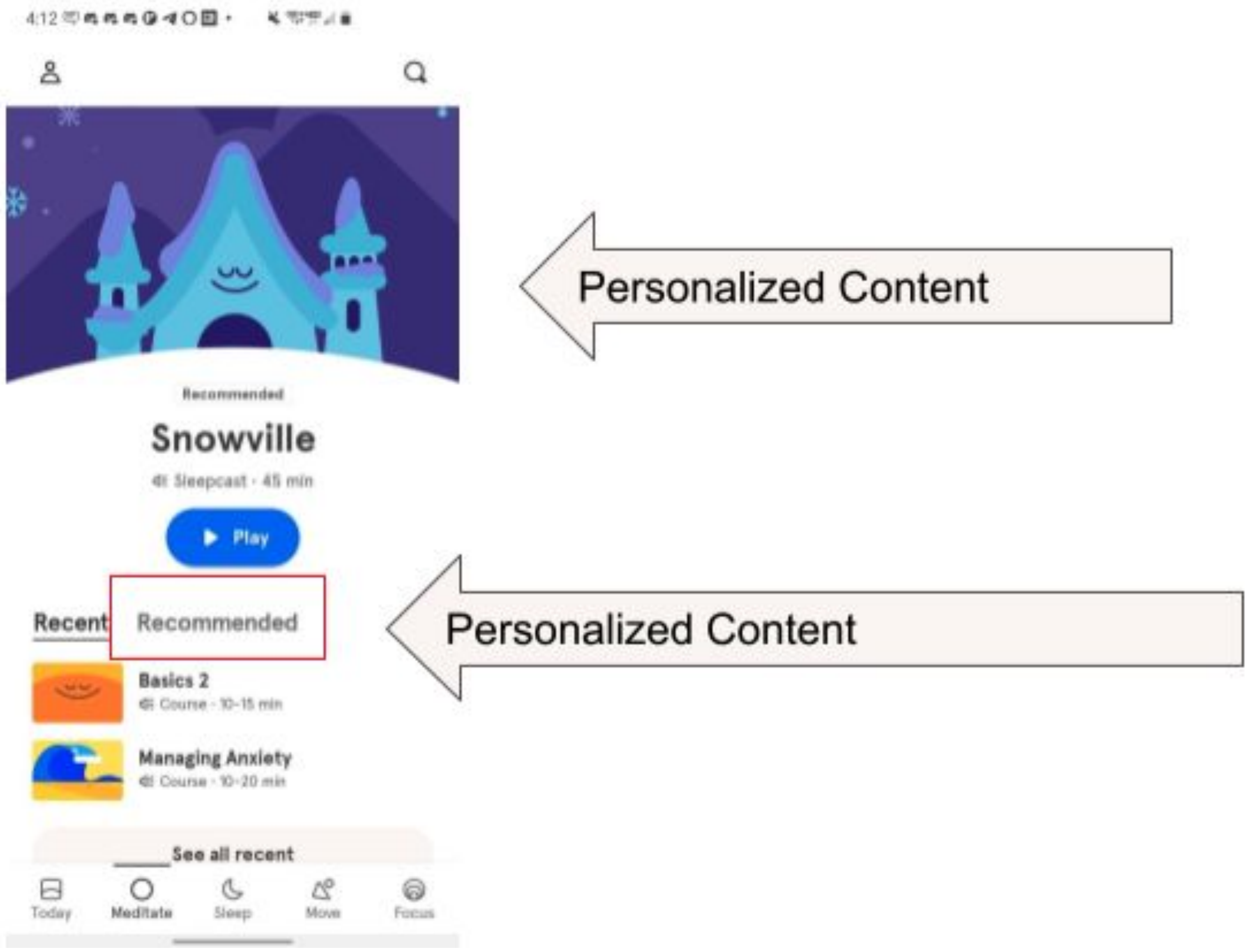
History

- Launched in 2021, the content customizer started to countering the declining retention rate.
- Every headspace user was getting the same editorial content that used to refresh everyday.
- Headspace had over 1000+ unique content and more than 2 million B2C subscribers

Hypothesis

Personalized recommendations will help members find more relevant content, thereby increasing their engagement with Headspace

- Content Customizer (a.k.a CC) is our recommendation system that serves personalized recommendations
- It uses historical user-item interaction data to generate recommendation everyday

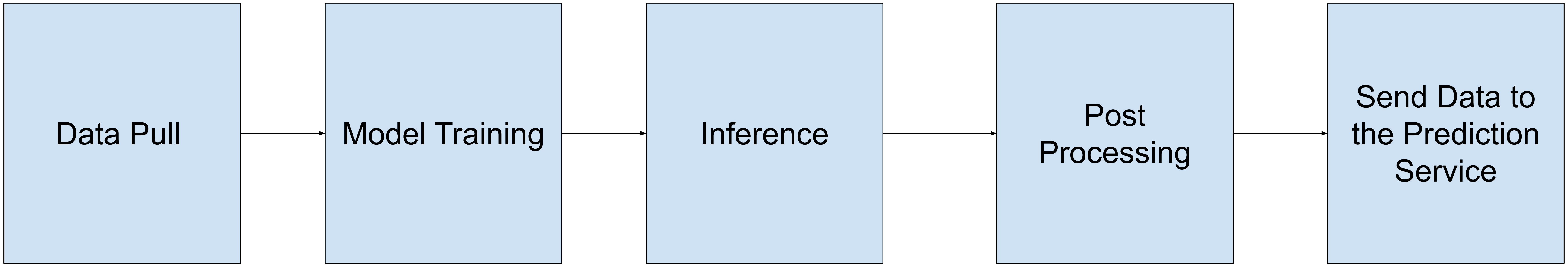


Components of Modeling

- **Data**
  - B2C and B2B Subscribers
  - Feature Store
    - User tenure, demographics, language, time of the day etc.
    - Item type, playtime duration, content creation date etc.
  - IOS and Android
- **Evaluation**
  - **Offline:** NDGC@K , HITRate@K, Precision@K, etc.
  - **Online:** A/B Testing
- **Inference and Post processing**
  - Nightly batched training
  - Payload to prediction service
  - Fix sequence, remove popular and recently completed

Modeling Pipeline

- Deployed for both B2C and B2B members
- Deployed to both iOS and Android platform



LightFM Model

- Reasons
  - Matrix Factorization
  - Implicit and Explicit Feedback
  - GPU Optimization
  - Highly scalable

- Result
  - **Offline:** Precision@k and Recall@k
  - **Online:** No statistically significant lift (content start/complete) in A/B test

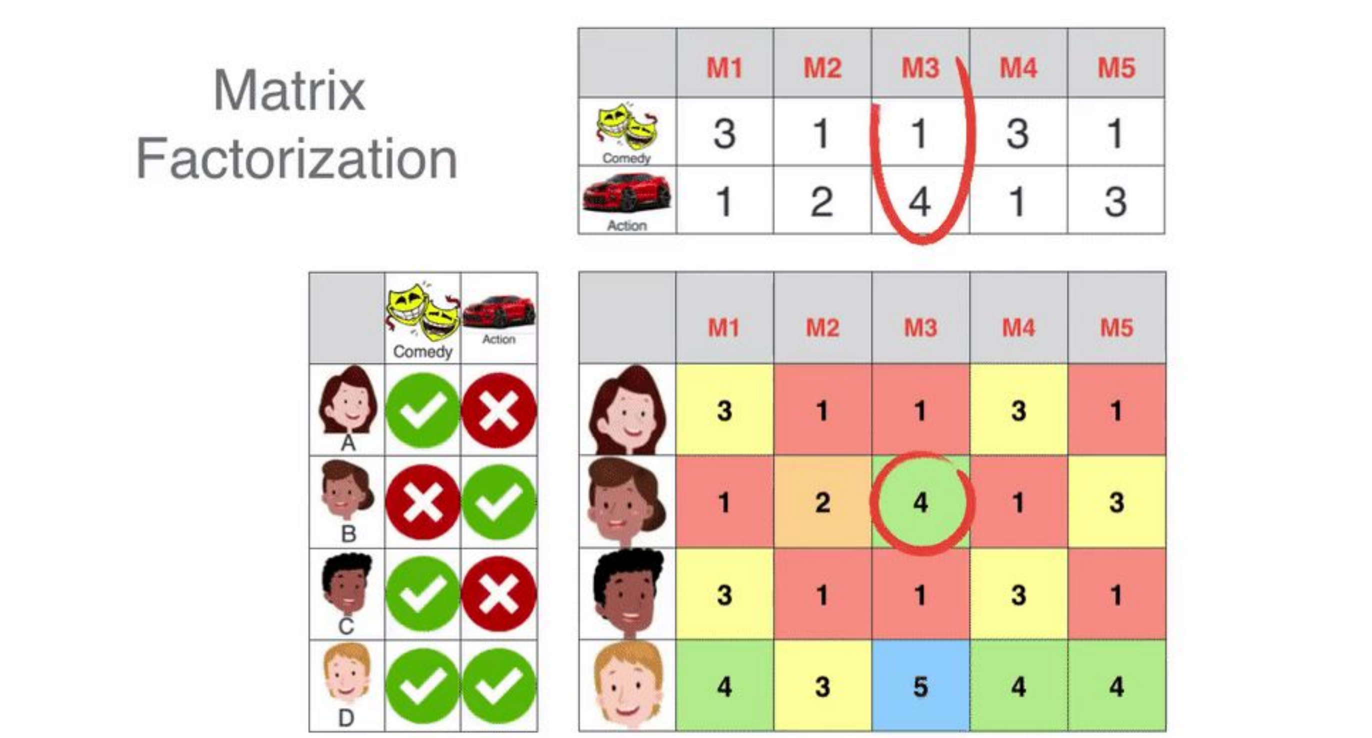
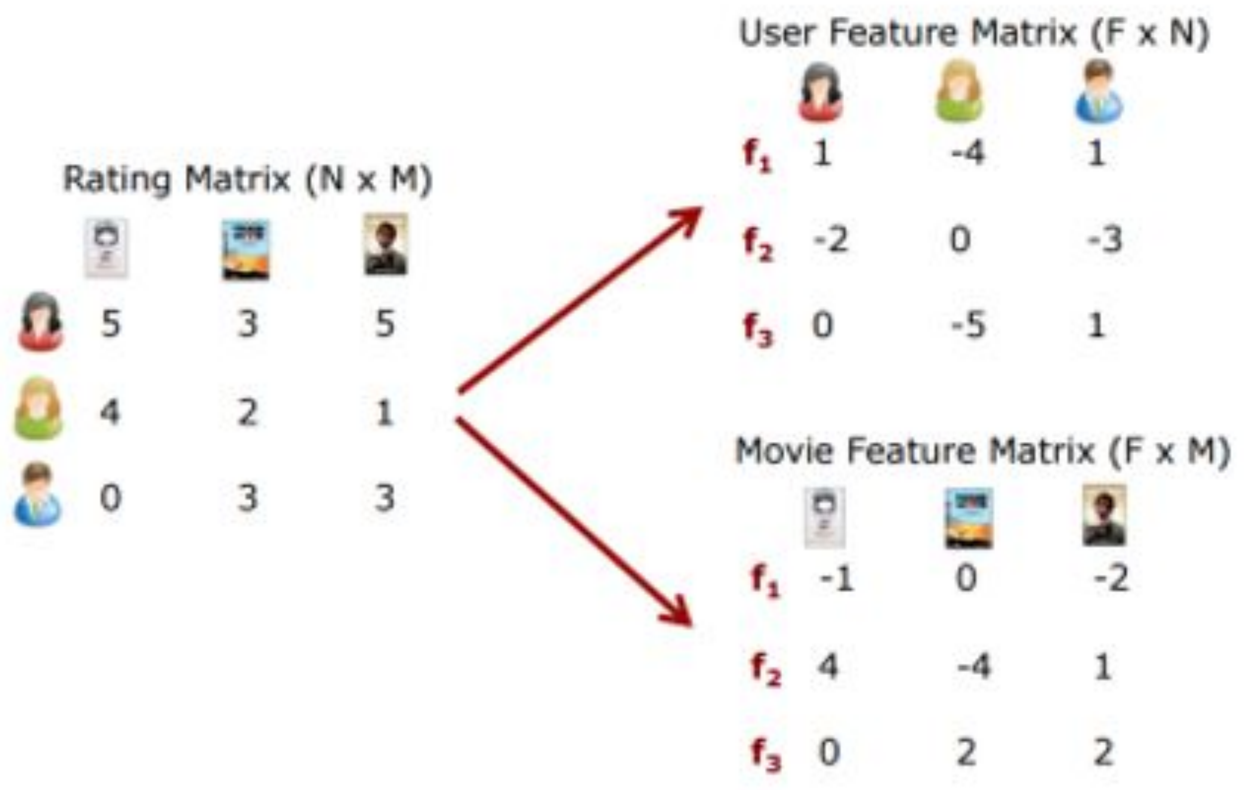


Image Credit: Google Developer

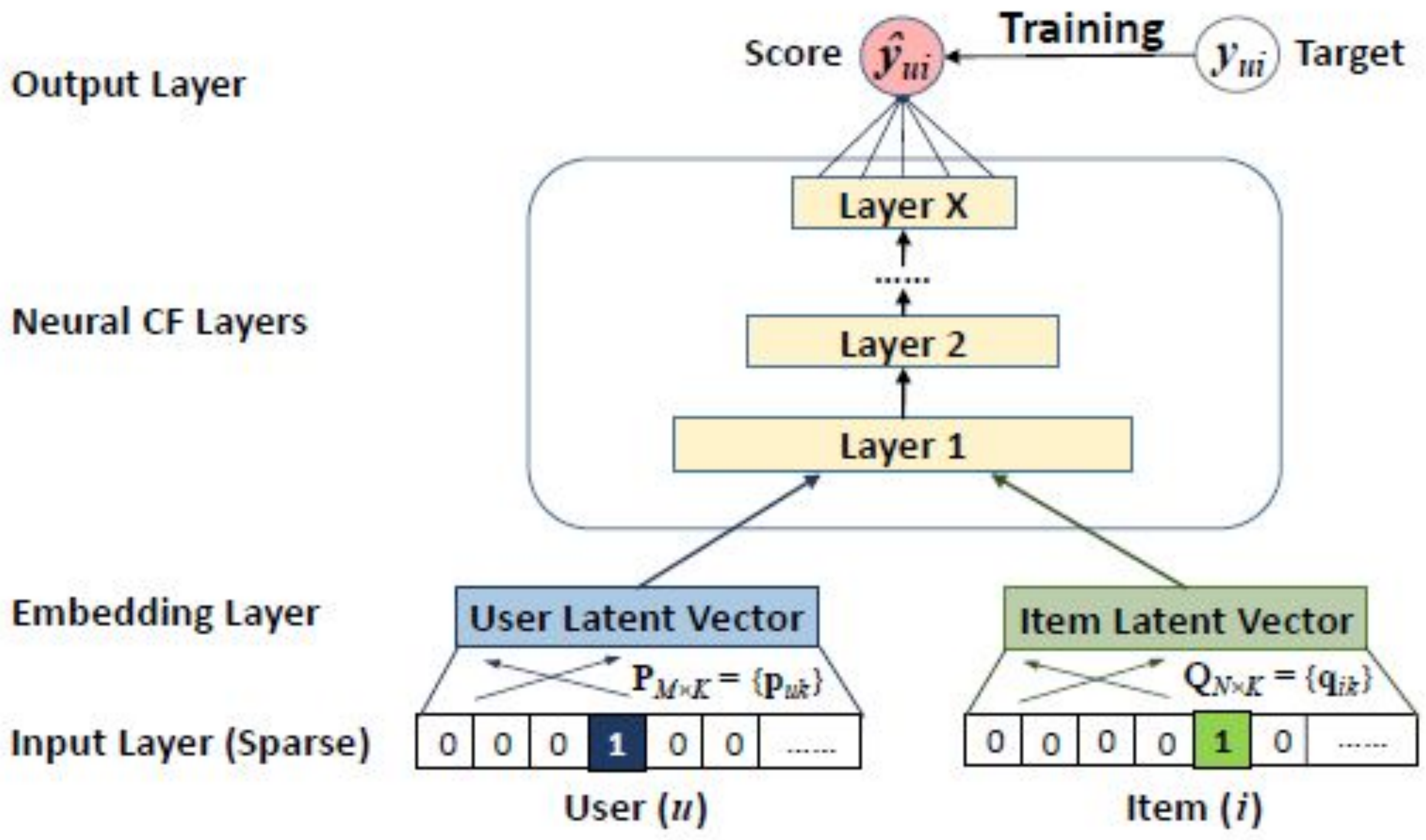


Kula, Maciej. "Metadata embeddings for user and item cold-start recommendations." arXiv preprint arXiv:1507.08439 (2015)

Neural Collaborative Filtering Model

- Reasons
  - Non-linearity
  - User-item features
  - Sparsity

- Result
  - **Offline:** Precision@k and Recall@k
  - **Online:** Over **2%** statistically significant lift (content start/complete) in A/B test

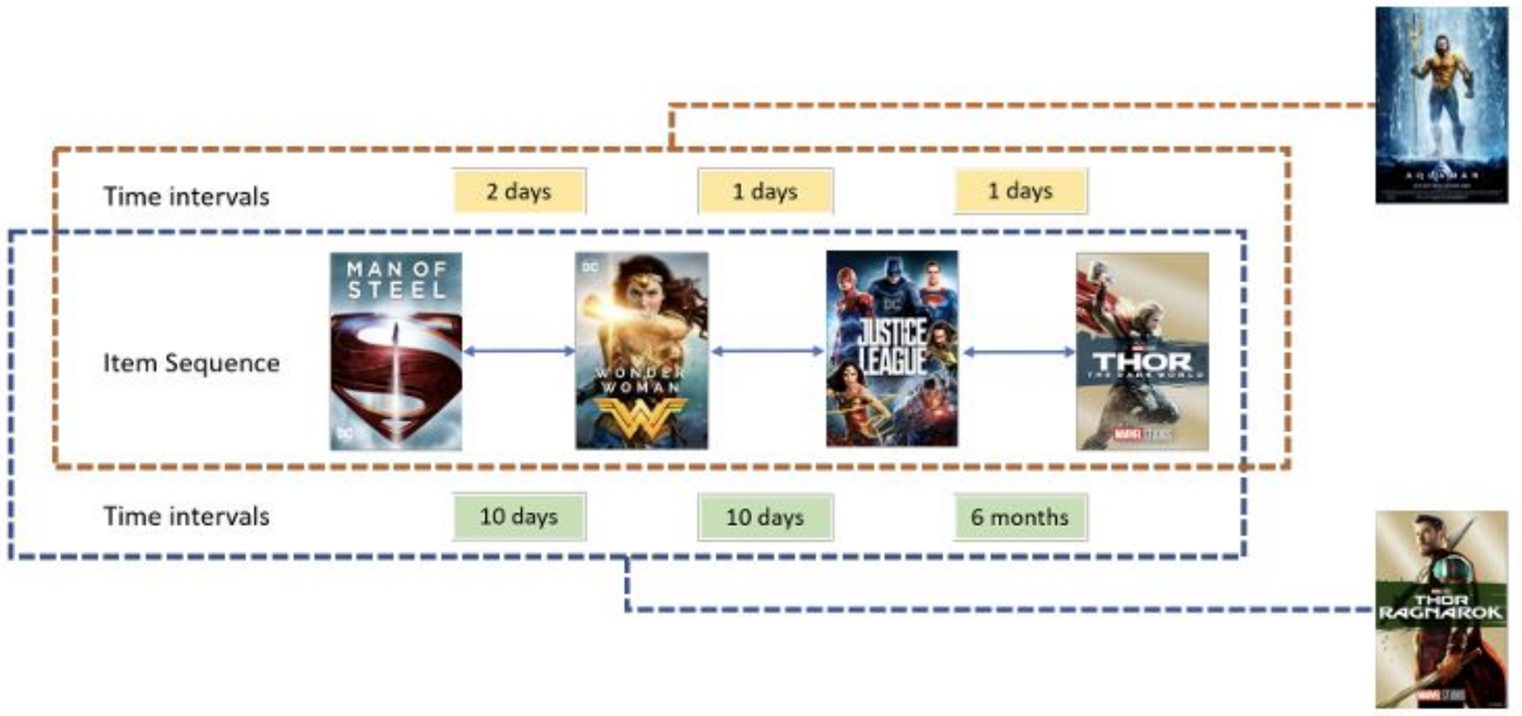


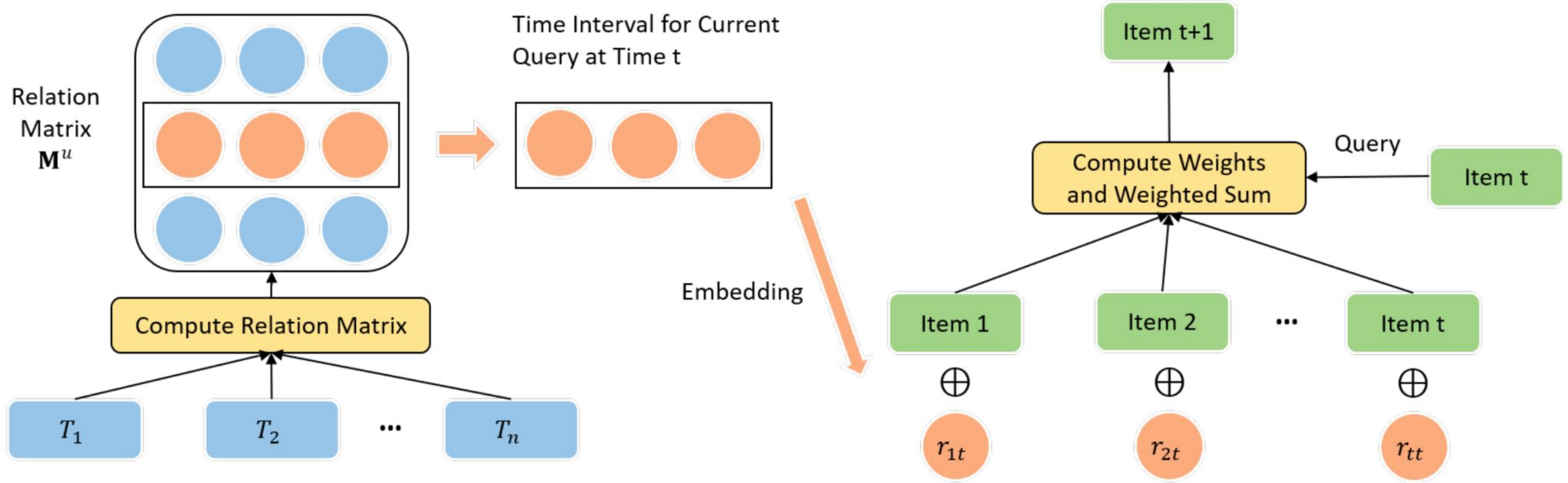
He, Xiangnan, et al. "Neural collaborative filtering." Proceedings of the 26th international conference on world wide web. 2017.

Time Interval Aware Self-Attention for Sequential Recommendation (TiSASRec)

- Reasons
  - Sequential Data
  - Temporal Dynamics
  - Transformer Architecture

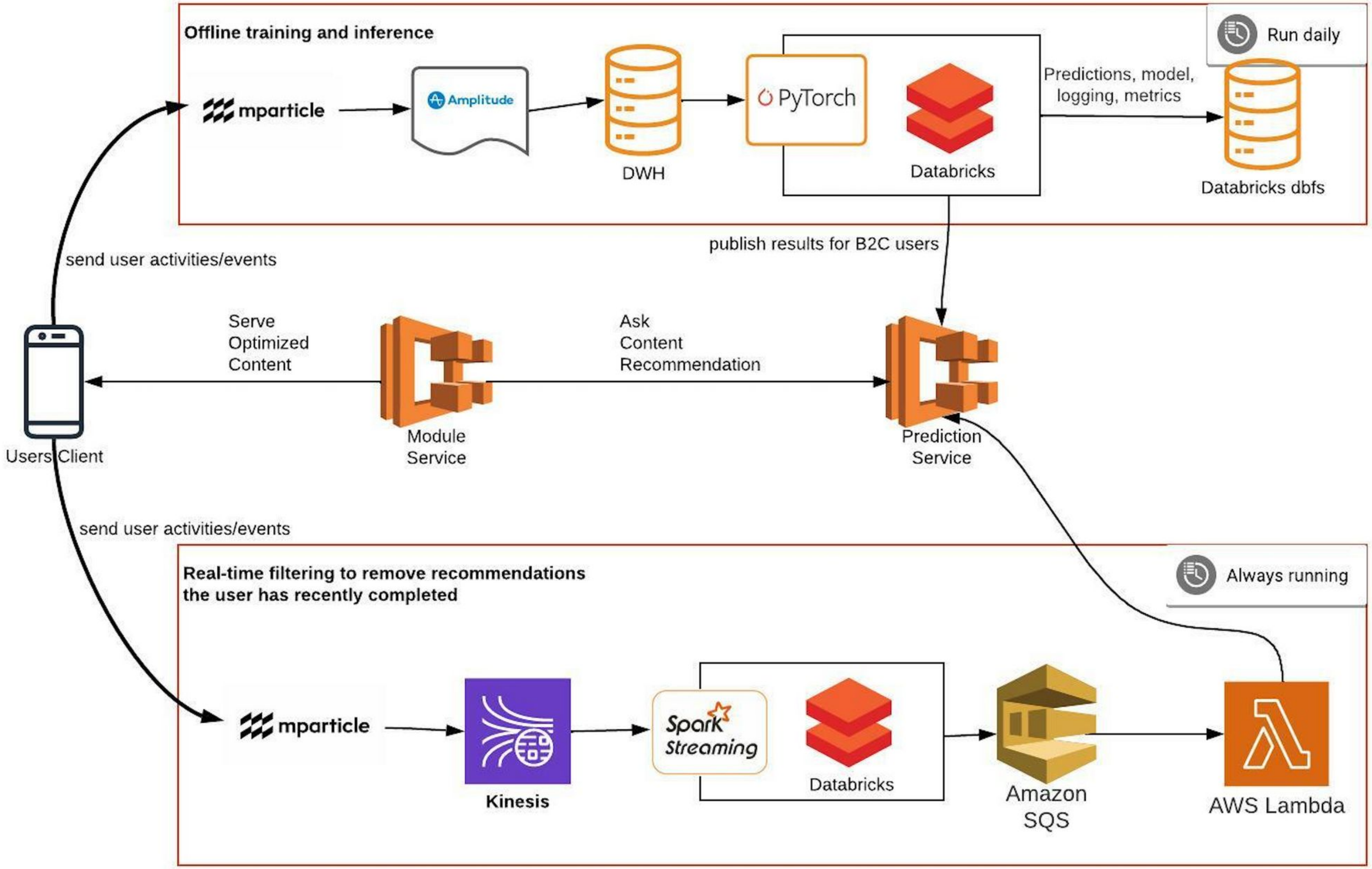
- Result
  - **Offline:** NDCG@k and HitRate@k
  - **Online:** Over **4.68%** statistically significant lift (content play) in A/B test



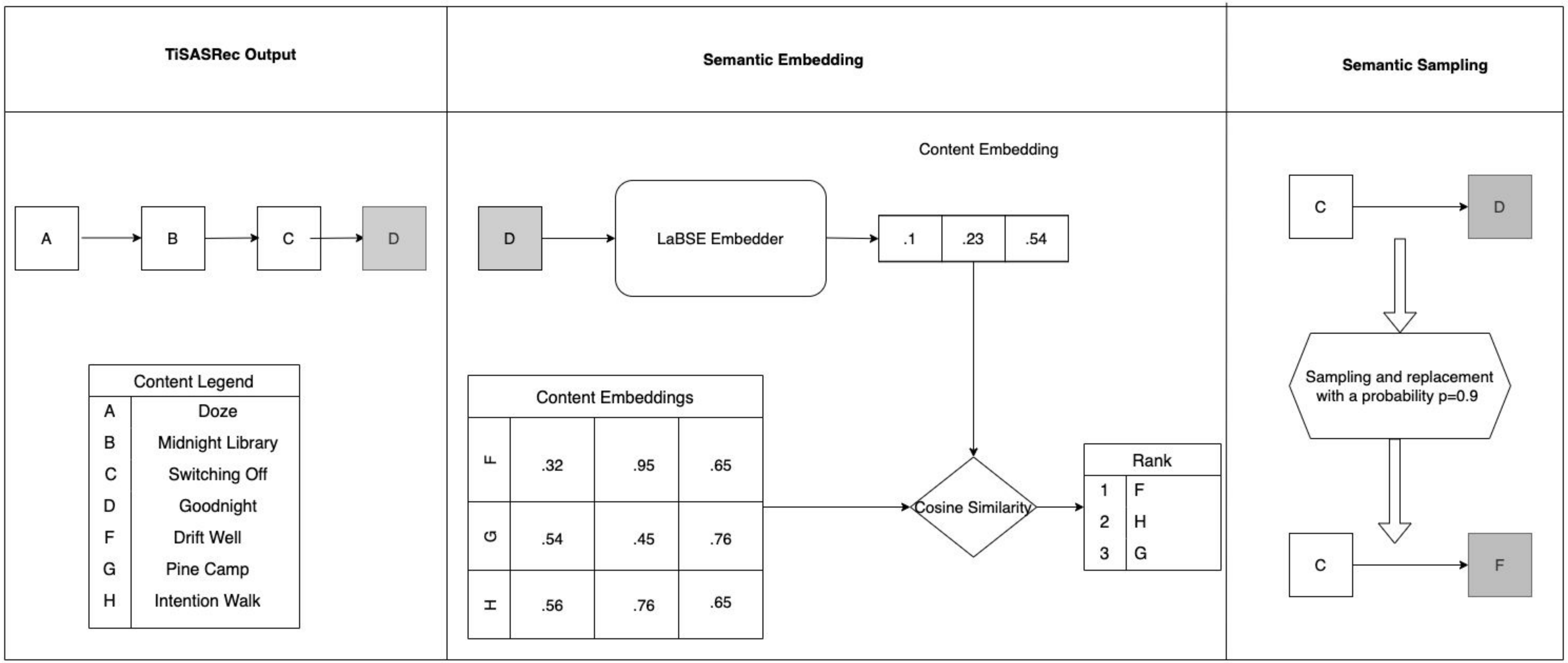


Li, Jiacheng, Yujie Wang, and Julian McAuley. "Time interval aware self-attention for sequential recommendation." *Proceedings of the 13th international conference on web search and data mining*. 2020.

System Architecture



Semantic Sampling



Language-agnostic BERT Sentence Encoder (LaBSE) : https://huggingface.co/setu4993/LaBSE

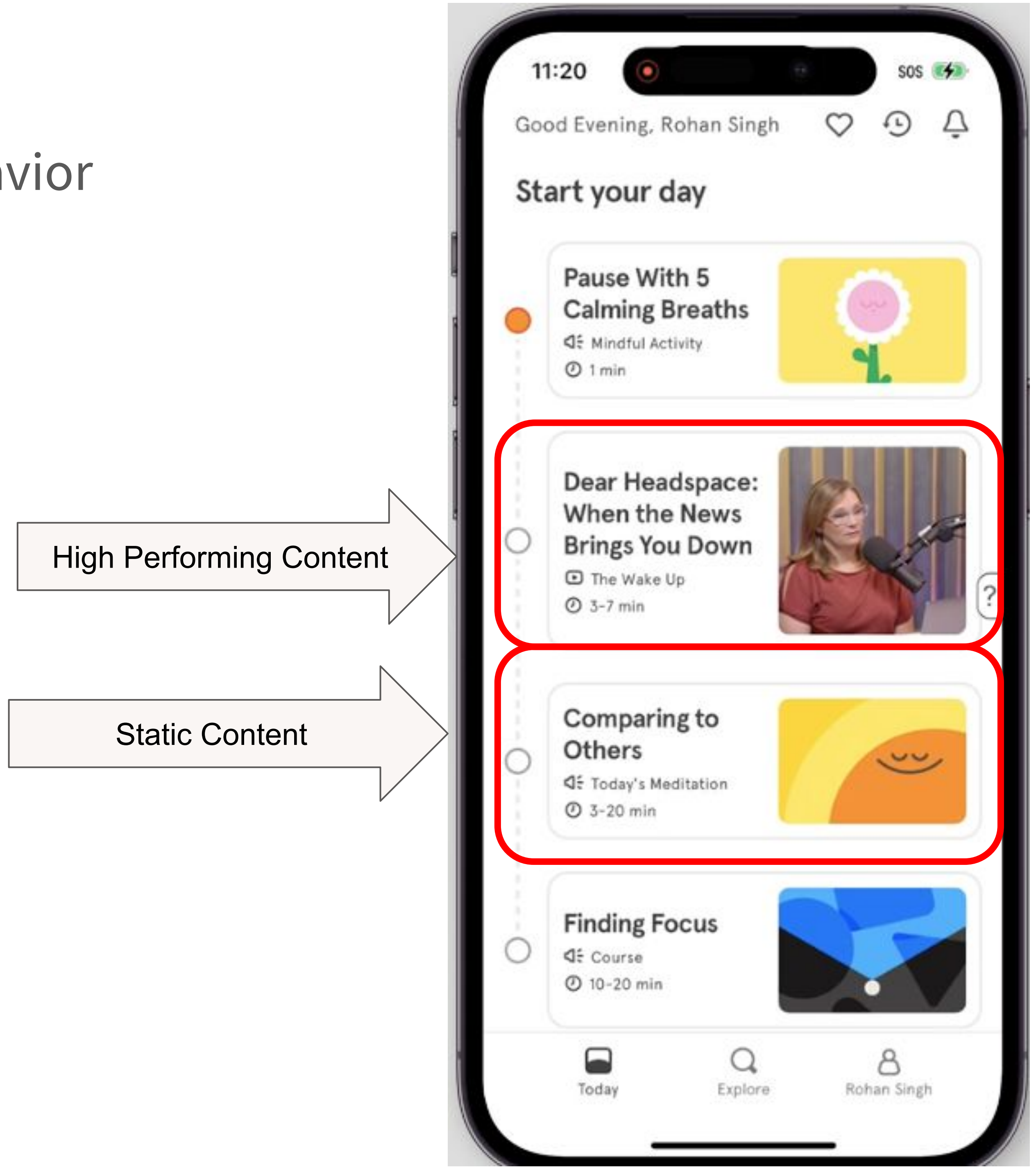
Experimented Models

Version Name	Outcome
RL Batch-Constrained Q-learning algorithm to a discrete-action setting	**Model didn't win in the experiment**
Transformer4Rec implement with NVIDIA team	**Didn't beat the existing model** **Used user-item features**
Onboarding Question Cold start model	**Dropped the engagement** Utilized onboarding question with static NLP embedding

Challenges

- **Feature Space**
  - User features are not rich enough to understand the user behavior
    - Subscription mismatch
    - Sparse interaction
    - Privacy and Compliance
  - Item features are fragmented
    - Many features are stored in multiple system
    - Overlapping content type

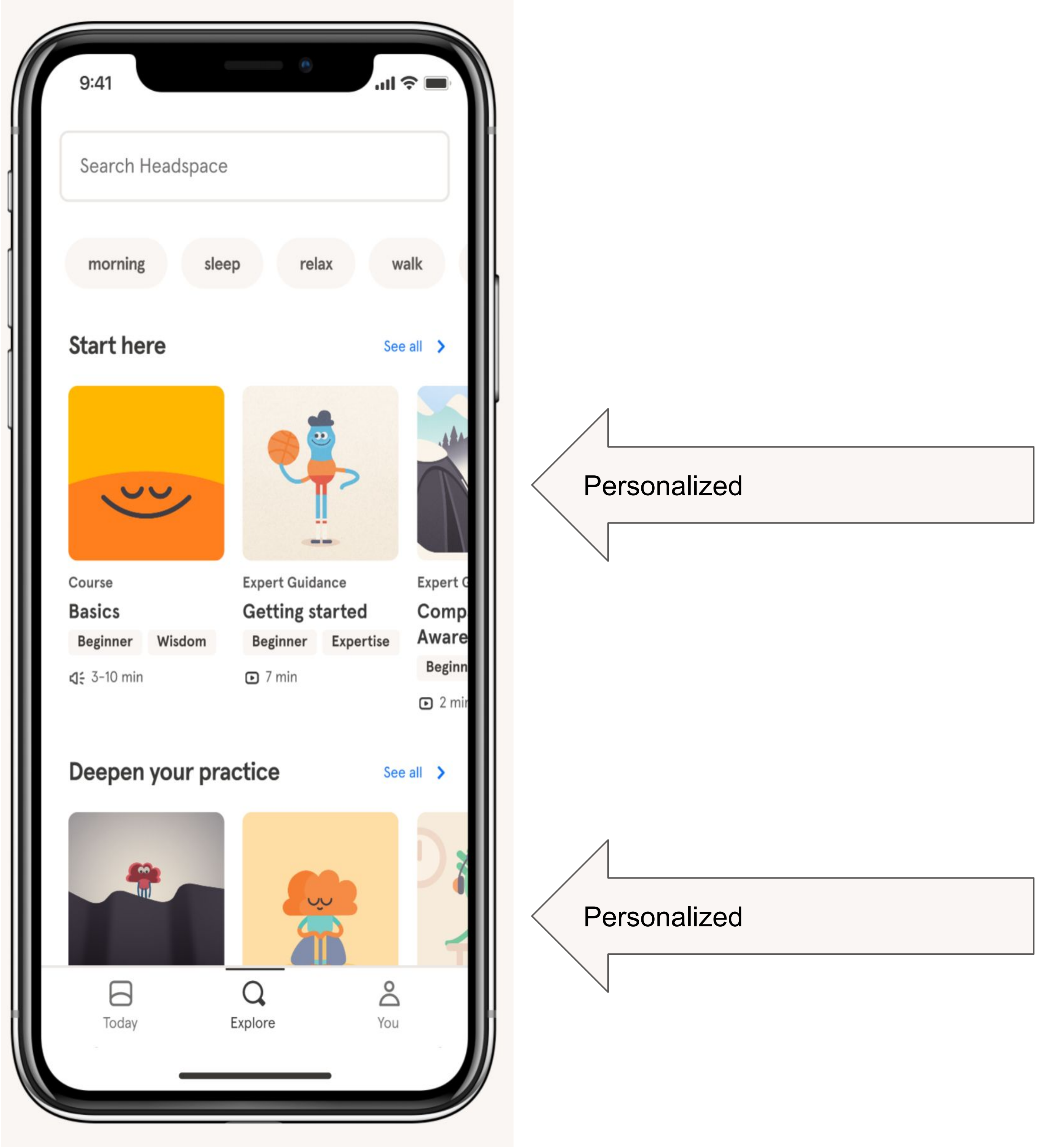
- **UI Rigidity**
  - Fluid design for multiple screens
  - High performing static content
  - Single model everywhere
  - Explicit Feedback



Opportunities

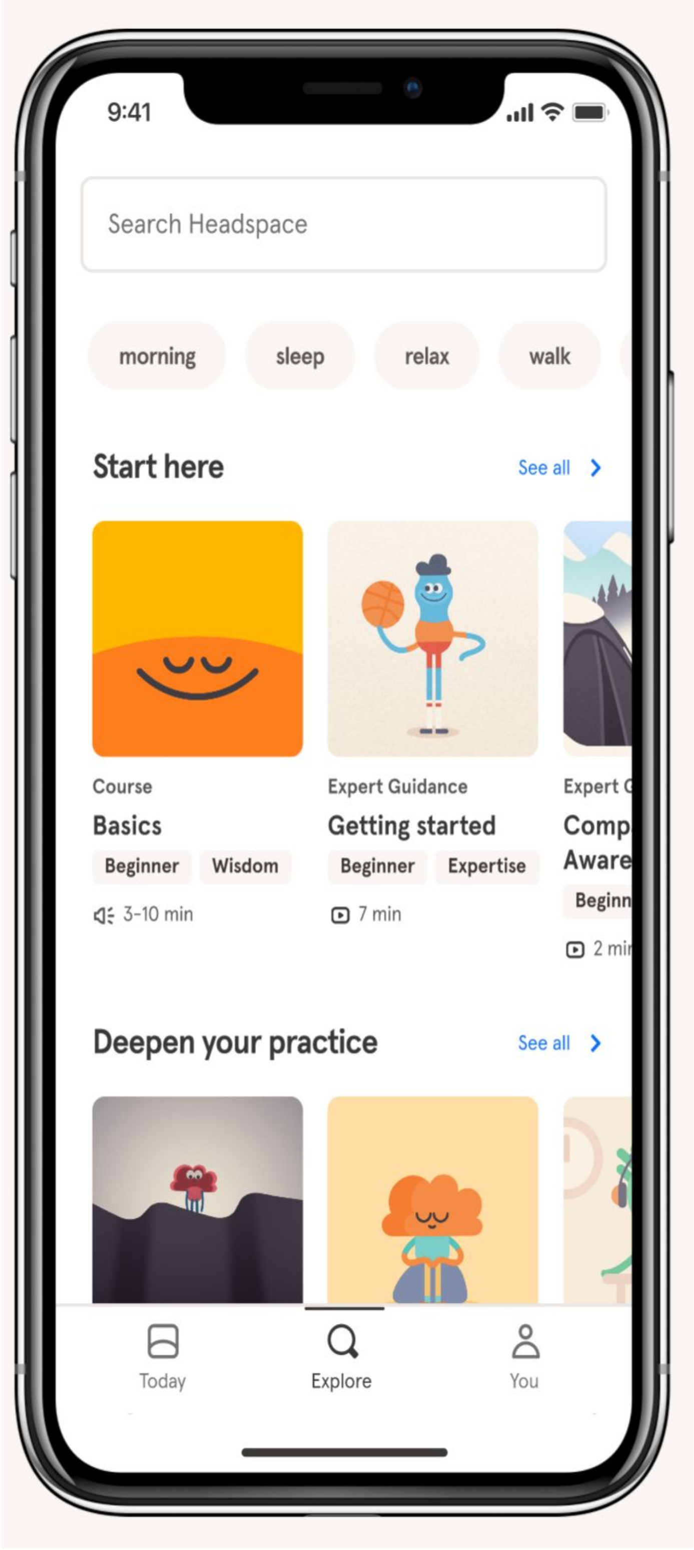
- **Client Reform**
  - The new shelves based design helps to surface context based recommendations
  - Provide separate screen to present notified recommendations
  - Search based recommendations

- **Modeling**
  - **Long term value** (Align with the product's vision)
  - Multi-modal recommendation (text, image, audio)
  - Personalized shelves ranking
  - Causal AI to get explainability
  - Optimize novelty, diversity and relevance



Opportunities

- **Robust E2E System**
  - Impression data
  - Analyze data drift, model drift and concept drift
  - Smart caching
  - Real time inference with continuous learning
  - Interconnected recommendation (email, push, notification)



Personalized

Personalized

Thank you

https://www.linkedin.com/in/rohan3/