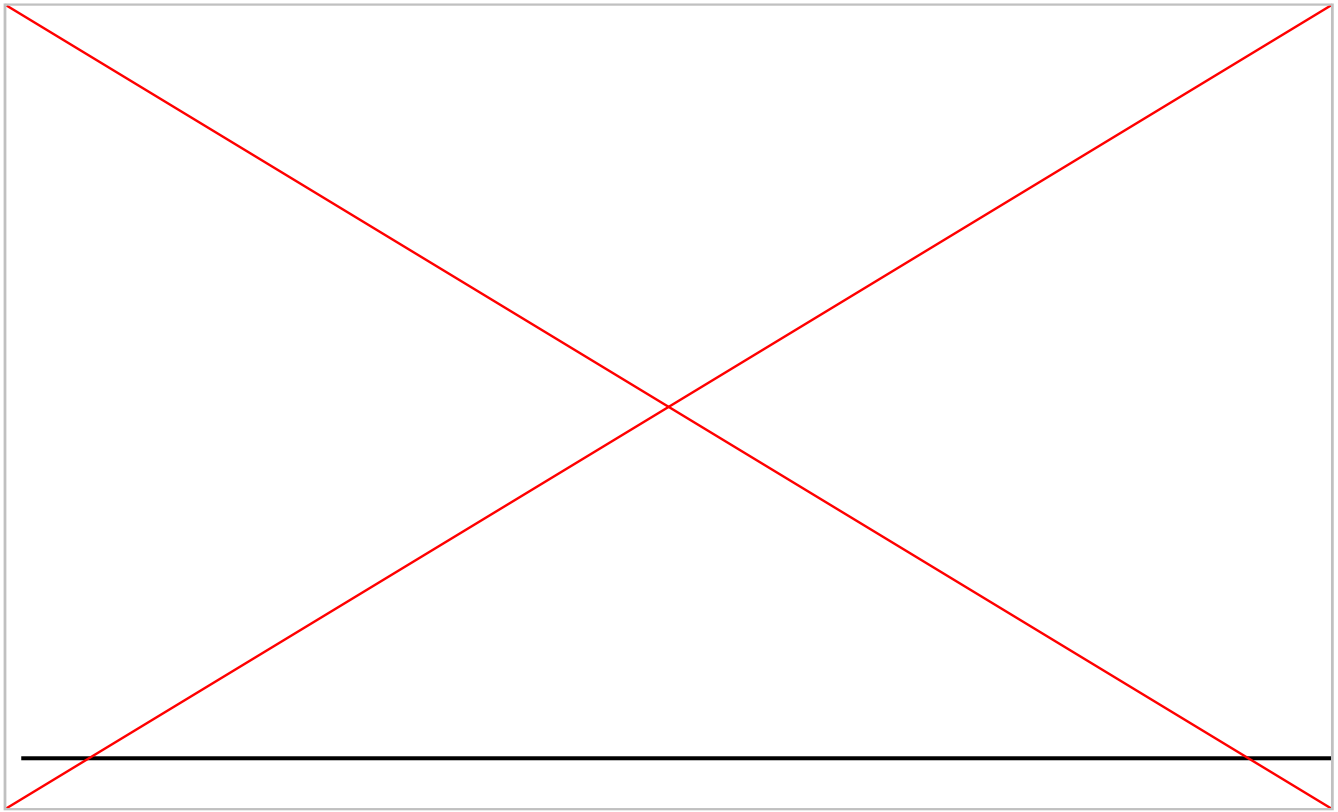
allegro

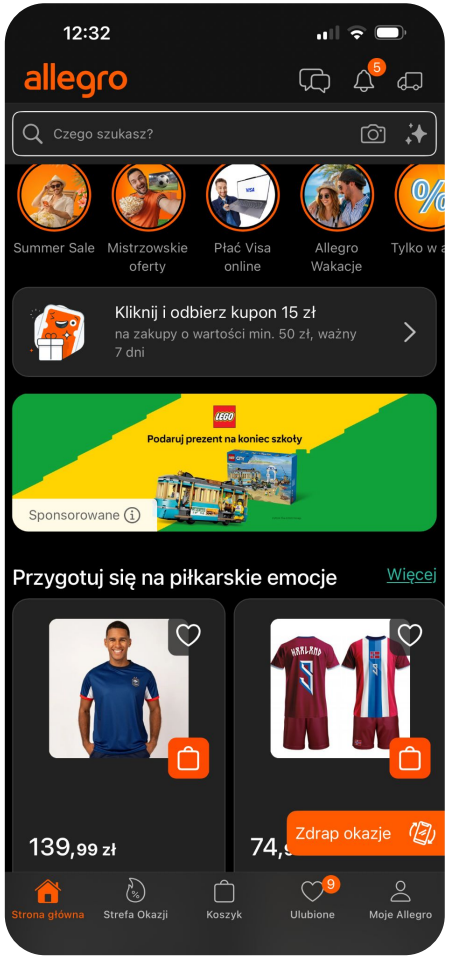
Deep Learning at Scale Tailoring the Home Page for Millions

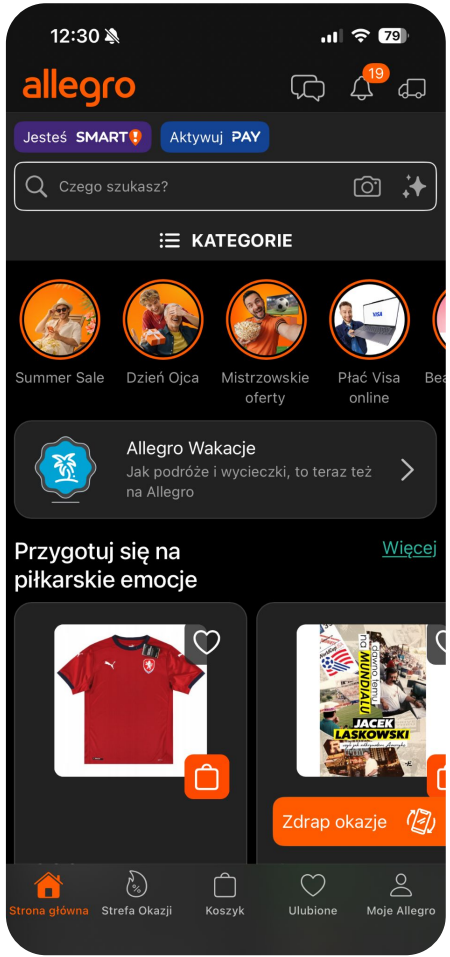
Marcin Szpunar, Jakub Żmujdzin

19/06/2026

What is Home Page?



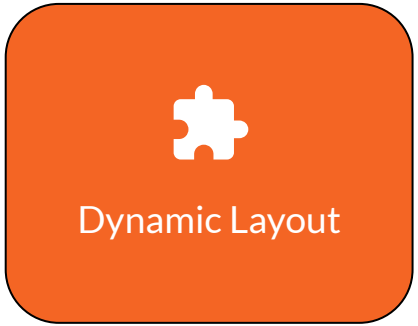




The "Same App" Problem

- **Same look, different users:** We are all different, but the app looks similar for everyone.
- **Hard to find:** Users waste time searching for features they need right now.
- **One size fits none:** Static apps are boring and don't help the user effectively.

Our Goal: A “Living” App



The initial strategy

Naive Approach

- Just use some carousel embeddings of previously viewed items and then serve more carousels that are the closest in terms of cosine distance
- What embeddings?
- Trained under what objective?
- Do we even have this kind of embeddings?

It’s not as easy as it looks

The real challenges:

- Which user actions are actually worth tracking?
- Which contextual types should be prioritized initially?
- How to transform raw context into ML-ready features?

Redefinition

User inspiration using real-time user actions and 30-day transaction history.

The Step-by-Step Plan

User Profile

Long-term preferences and historical behavior.

Real-Time Interactions

Immediate signals (live clicks).

Final Ranking

Scoring and ordering the top items based on retrieval stage and business KPIs.

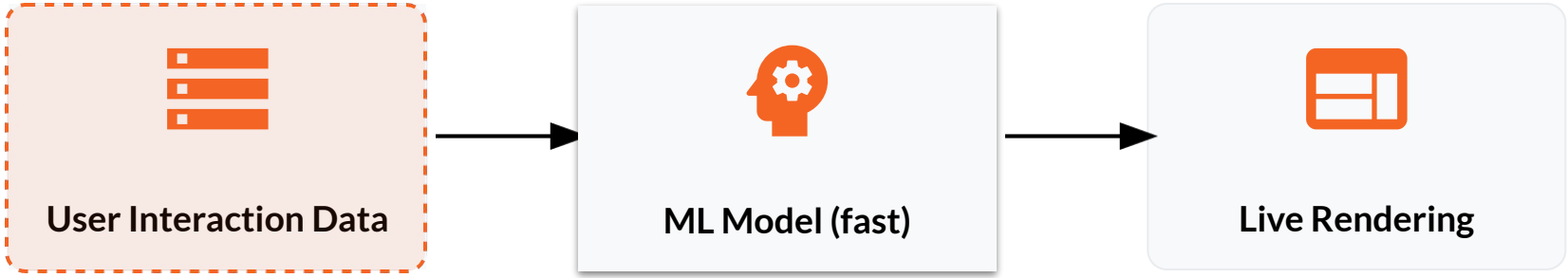
Offer Space

Structuring and grouping offers for home page carousels.

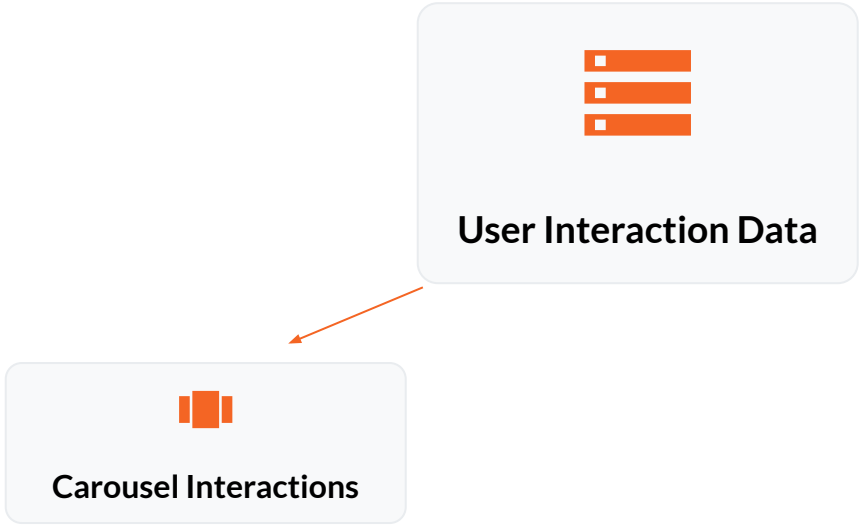
Modelling

Filtering down a set of options to the best candidates.

Personalization Engine Concept



User Interaction Data

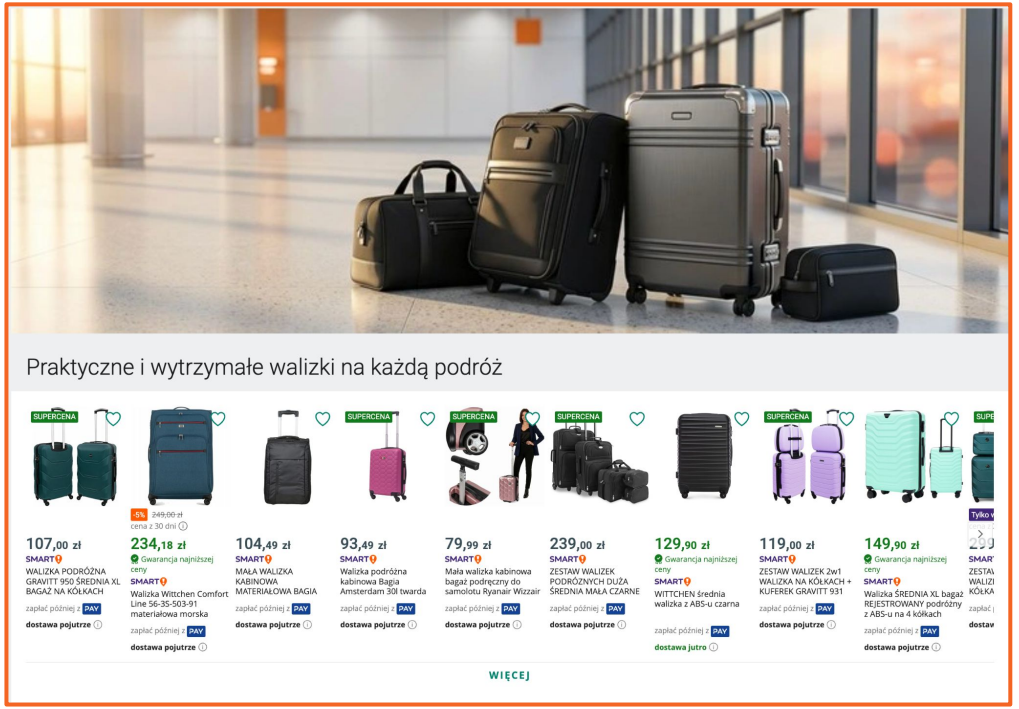


Single carousel composition

(optional) image →

title →

offers →

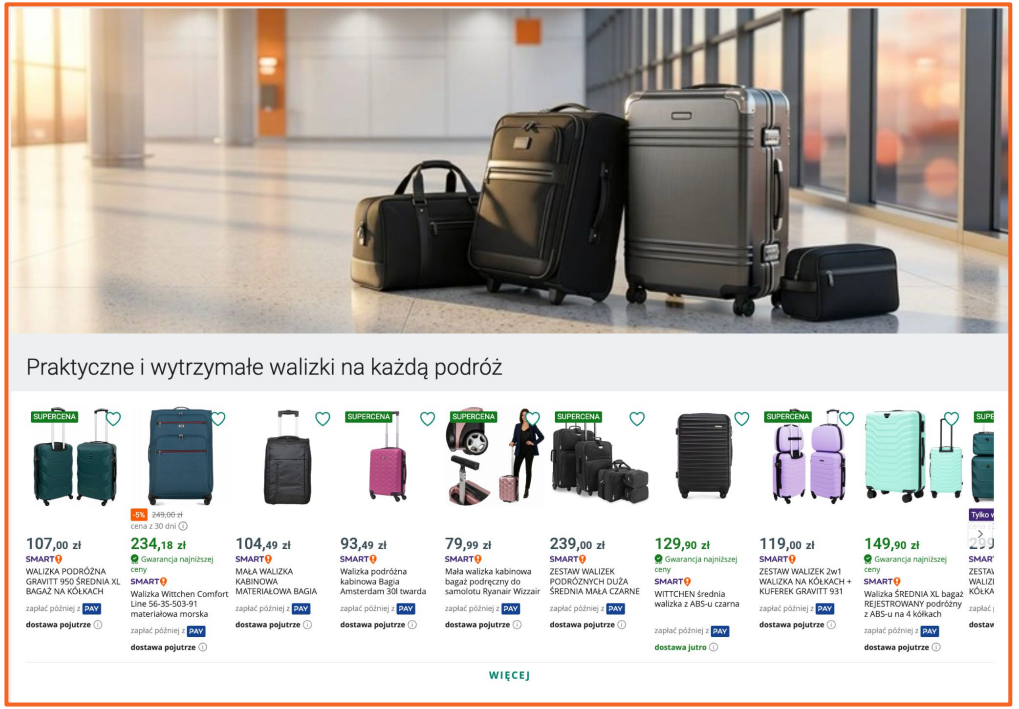


Single carousel composition

(optional) image

title

offers



Capturing the Live Session

Granular Engagement Metrics

- Tracking immediate user actions like clicks, scrolls etc.

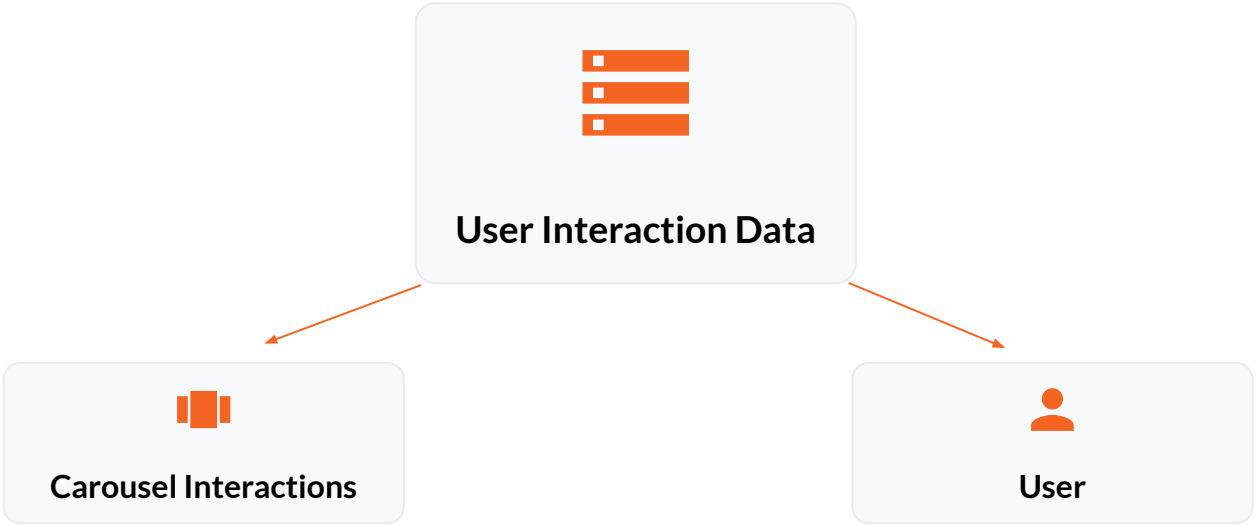
Sequence-Based Aggregation

- Grouping individual interactions into n-element sequences to capture immediate browser momentum.

Strict Time-Window Constraints

- Filtering data through a short-term time window to focus only on active, current intent

User Interaction Data



User Embeddings

Dedicated Model

A specialized encoder that turns raw buying patterns into dense vectors.

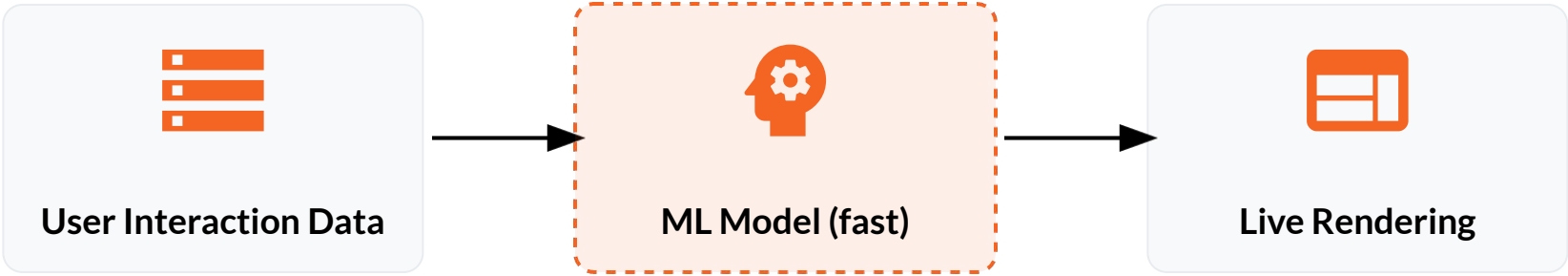
Daily Scoring Lifecycle

Continuous updates to reflect the latest user state in our feature store.

30-Day Transactional Window

Focusing on active shoppers with confirmed purchase history.

Personalization Engine Concept



What is the actual ML task type?

- Binary Classification?
- Multilabel Classification?
- Contrastive Learning!

Contrastive Learning

What is it and how does it work?

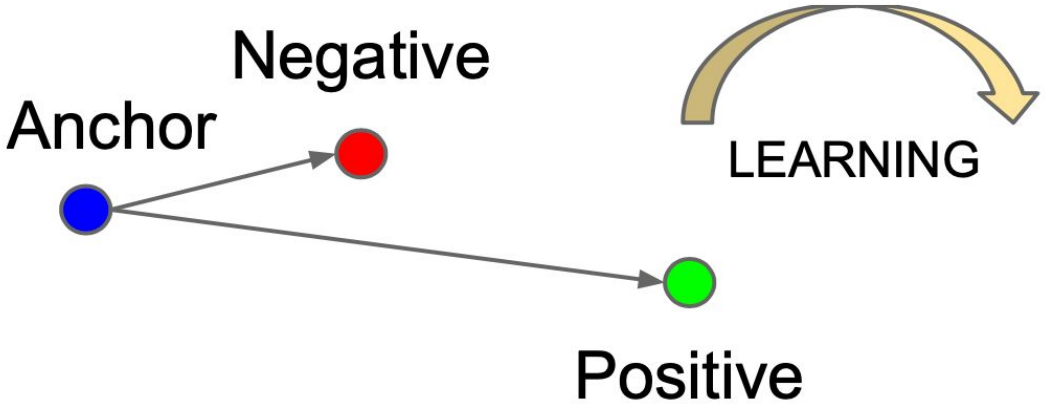
- Objective: make samples from same class have similar embeddings
- … and samples from different class - different embeddings

Contrastive Learning

Measuring embedding similarity

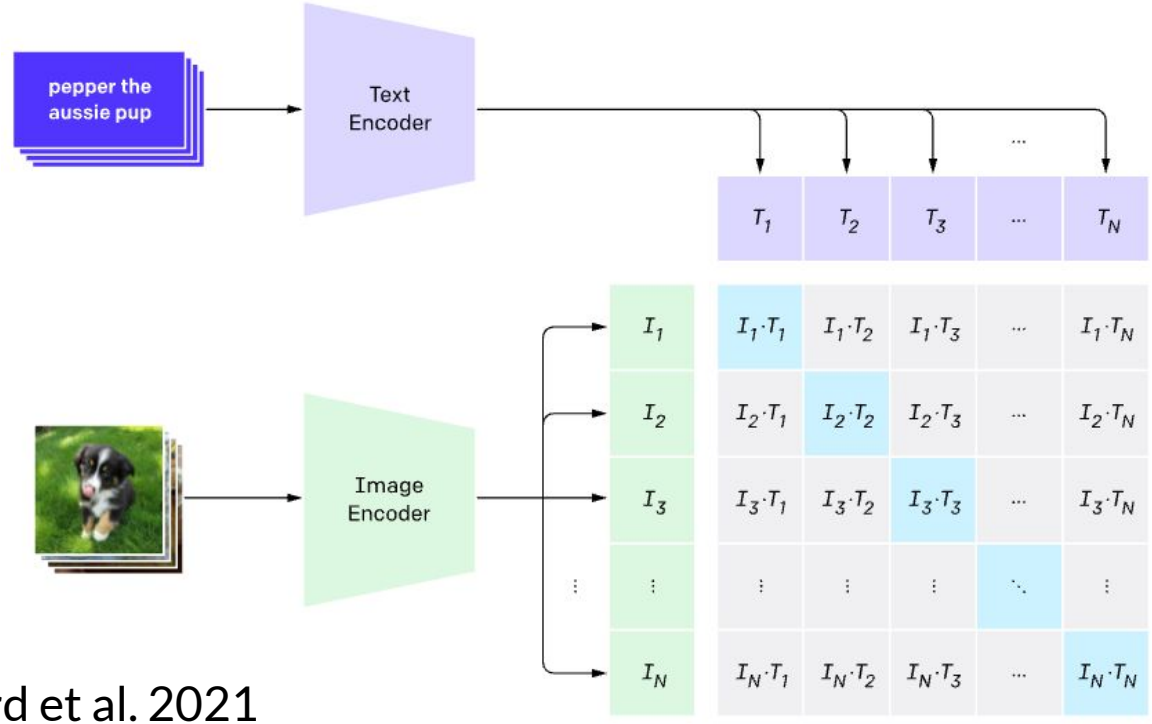
- Any distance metric (e.g. Euclidean) (lower = better)
- Cosine of angle between vectors (higher = better)
  - Linear transform to Cosine distance (lower = better)
- Dot product - vector magnitude times cosine
  - With L2-normed vectors, Dot product = Cosine

Contrastive Learning - Triplet Loss



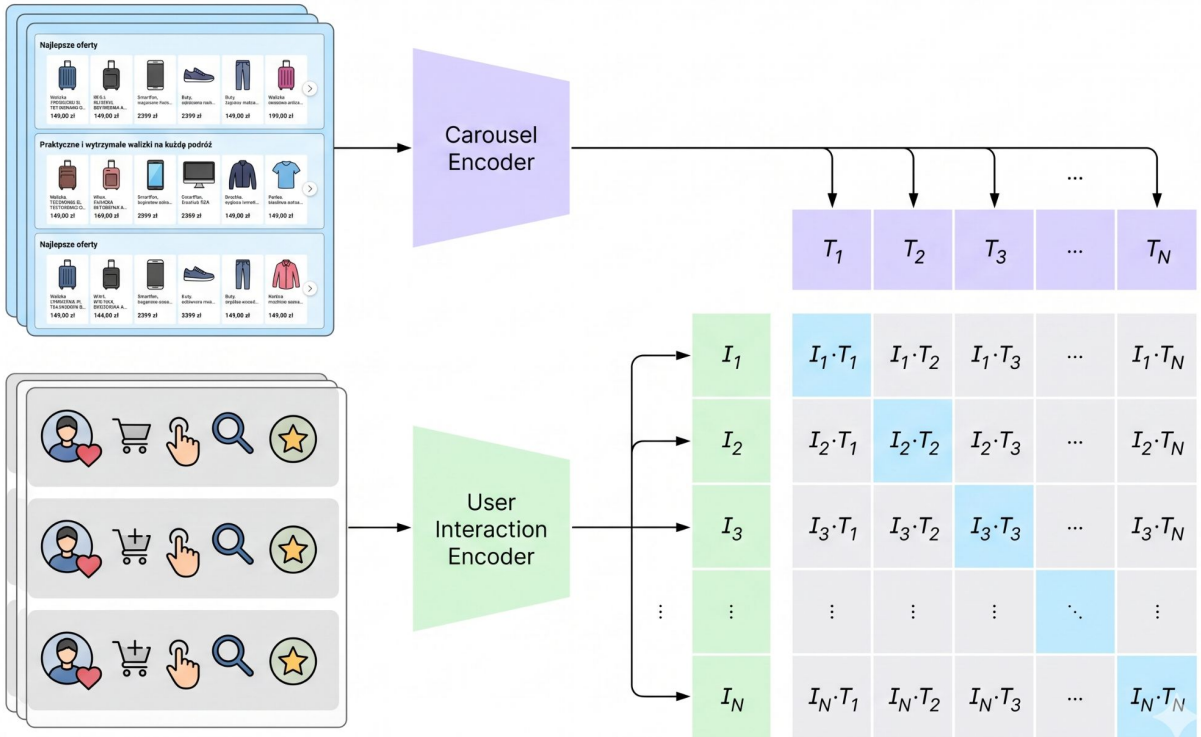
Source: Schroff et al. 2015

Contrastive Learning - InfoNCE



Source: Radford et al. 2021

Contrastive Learning - InfoNCE



Model Evaluation

How to construct a fair eval?

- Constantly changing carousel inventory
  - New carousels pop in
  - Within carousel item ordering can change
- Can we have users from different days in the same batch?

Model Evaluation

How to construct a fair eval?

- Constantly changing carousel inventory
  - New carousels pop in
  - Within carousel item ordering can change
- Can we have users from different days in the same batch?
  - Yes, but not too many
    - We can try to learn high-level patterns in user interactions

Model Evaluation

How to construct a fair eval?

- The order of carousels was already optimized using some algorithm
  - How do we make sure that the model can learn unbiased patterns?
  - Does appearing at the top = more clicks?

Model Evaluation

How to construct a fair eval?

- The order of carousels was already optimized using some algorithm
  - How do we make sure that the model can learn unbiased patterns?
    - Randomization test
  - Does appearing at the top = more clicks?
    - Yes, but lower Click-Through Rate

Model Evaluation

How to construct a fair eval?

- The upstream embedding models have to be retrained
  - How do we ensure that our input data doesn't drift?

Model Evaluation

How to construct a fair eval?

- The upstream embedding models have to be retrained
  - How do we ensure that our input data doesn't drift?
    - Let's construct an adapting layer
    - Zielnicki, Hsiao 2025. Orthogonal Low Rank Embedding Stabilization

Inference

Choosing what to display

- To compute similarity between user and a single carousel, we need to process each of these vectors through a dedicated encoder
  - Recompute needed whenever input changes (e.g. a new click)
- How to make this feasible, given limited time for inference and potentially 1000+ carousels?

Inference

Inventory Assumptions

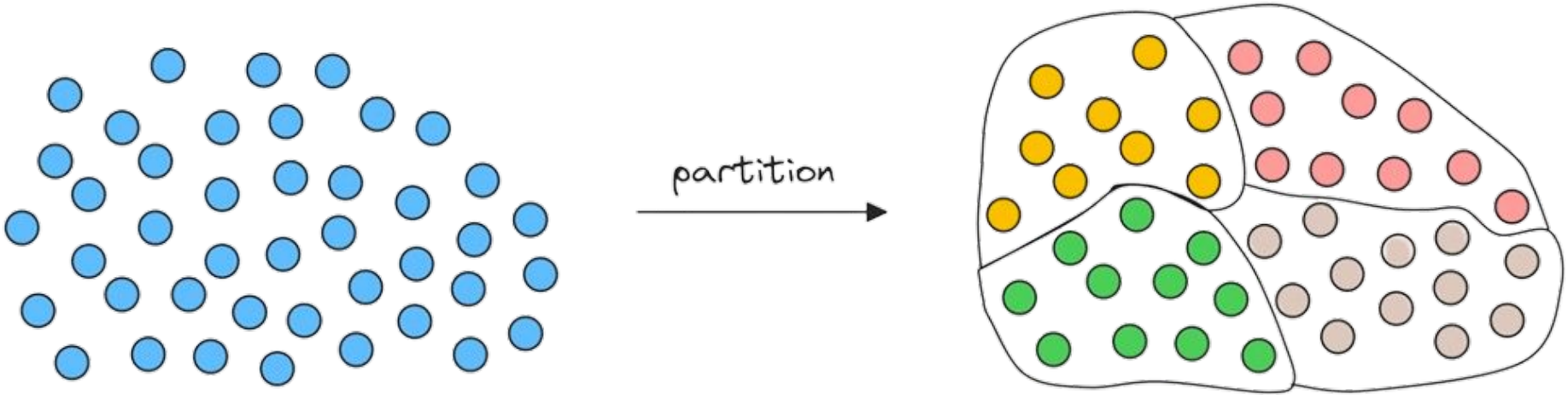
- Carousel Embeddings change less frequently
- We'll just assume that the Carousel Inventory is fixed at a given point in time, and compare User Embedding to the Carousels

Inference

Feasibility

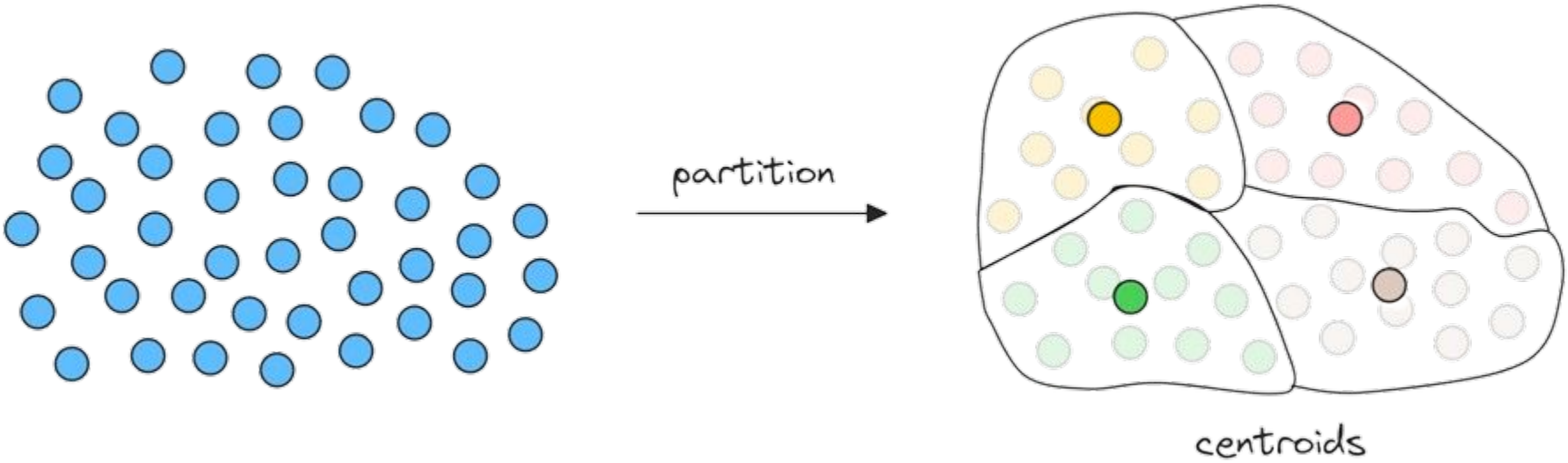
- Computing the distances for all carousels is too slow
  - Maybe we actually don’t need the exact NN, given large enough inventory?

Contrastive Learning - ANN



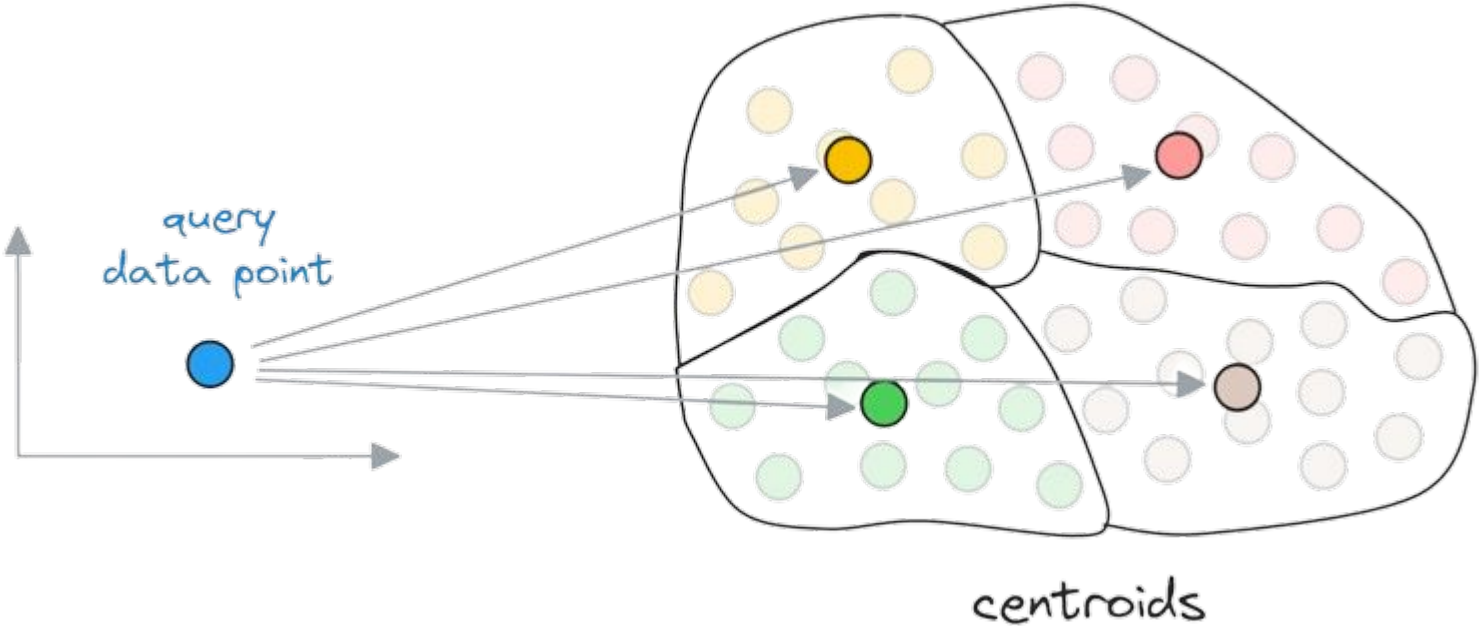
Source: Daily Dose of Data Science

Contrastive Learning - ANN



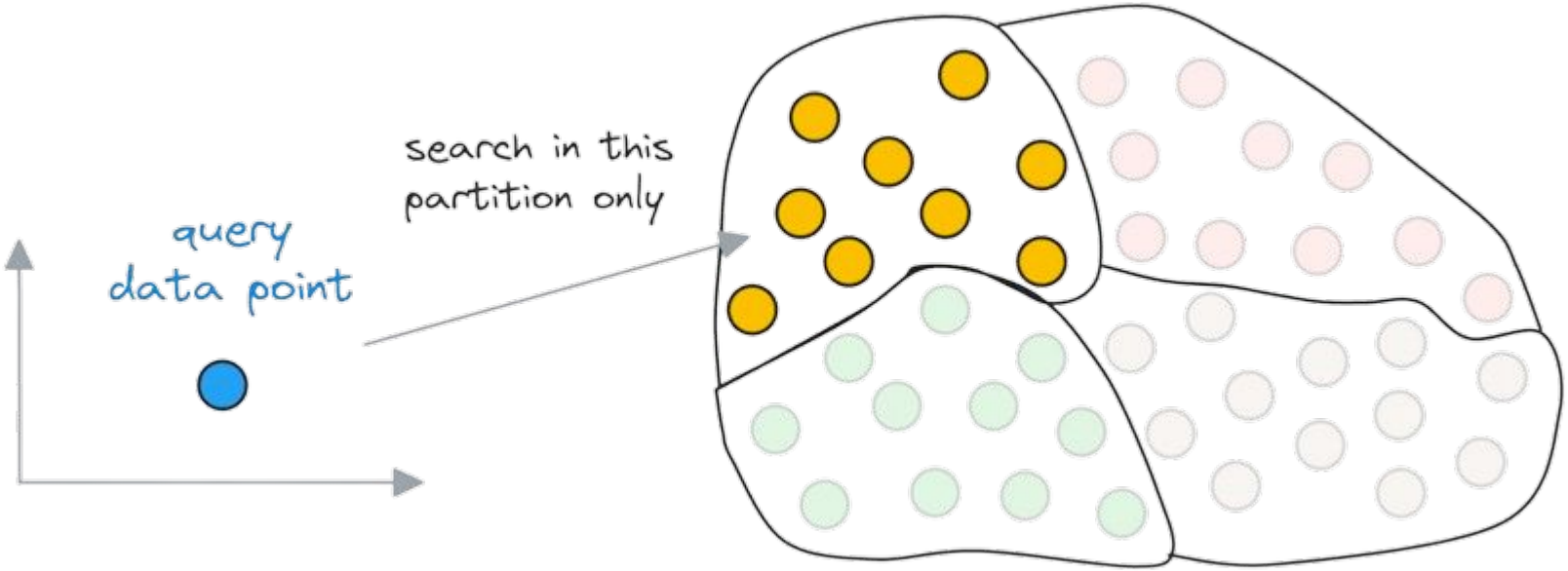
Source: Daily Dose of Data Science

Contrastive Learning - ANN



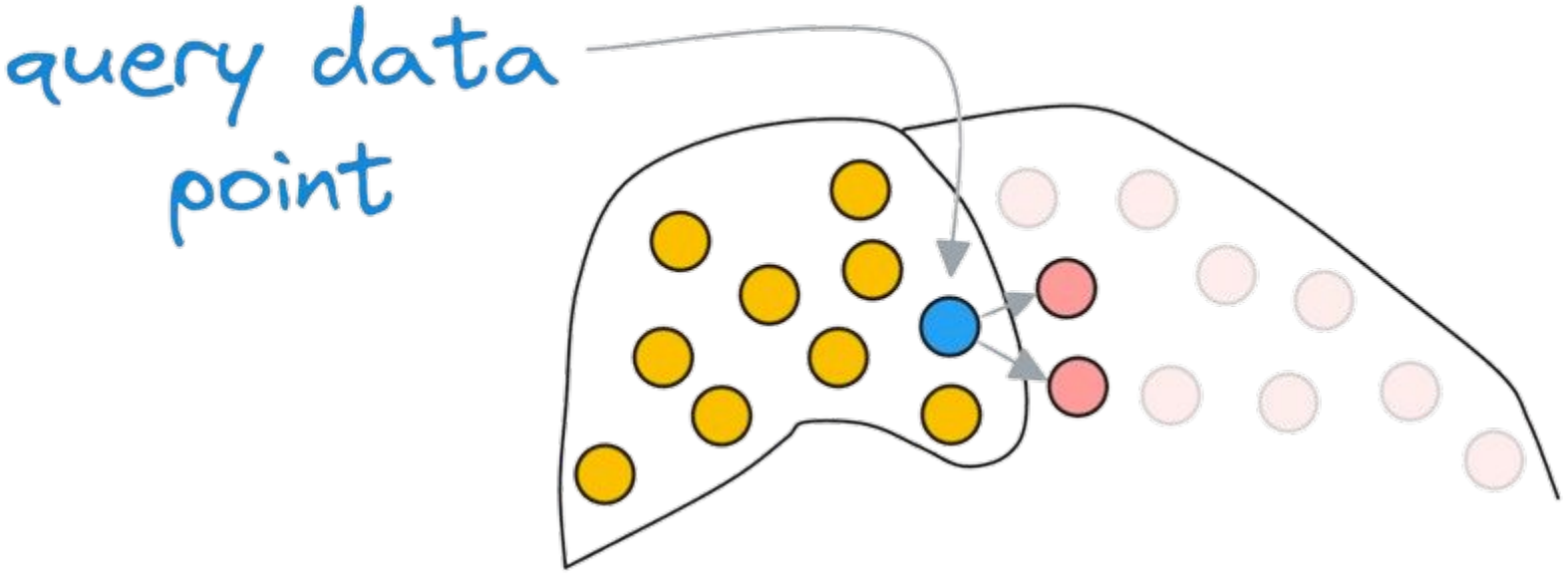
Source: Daily Dose of Data Science

Contrastive Learning - ANN



Source: Daily Dose of Data Science

Contrastive Learning - ANN



Source: Daily Dose of Data Science

Retrieval

Choosing what to display

- ANN result
  - K nearest neighbors (approx.) and the distance metrics
- Do we just display these carousels to the user?

Retrieval

Choosing what to display

- Do we just display these carousels to the user?
  - No, these were rough estimates of user interests
  - How do we optimize the display order?

Retrieval + Ranking

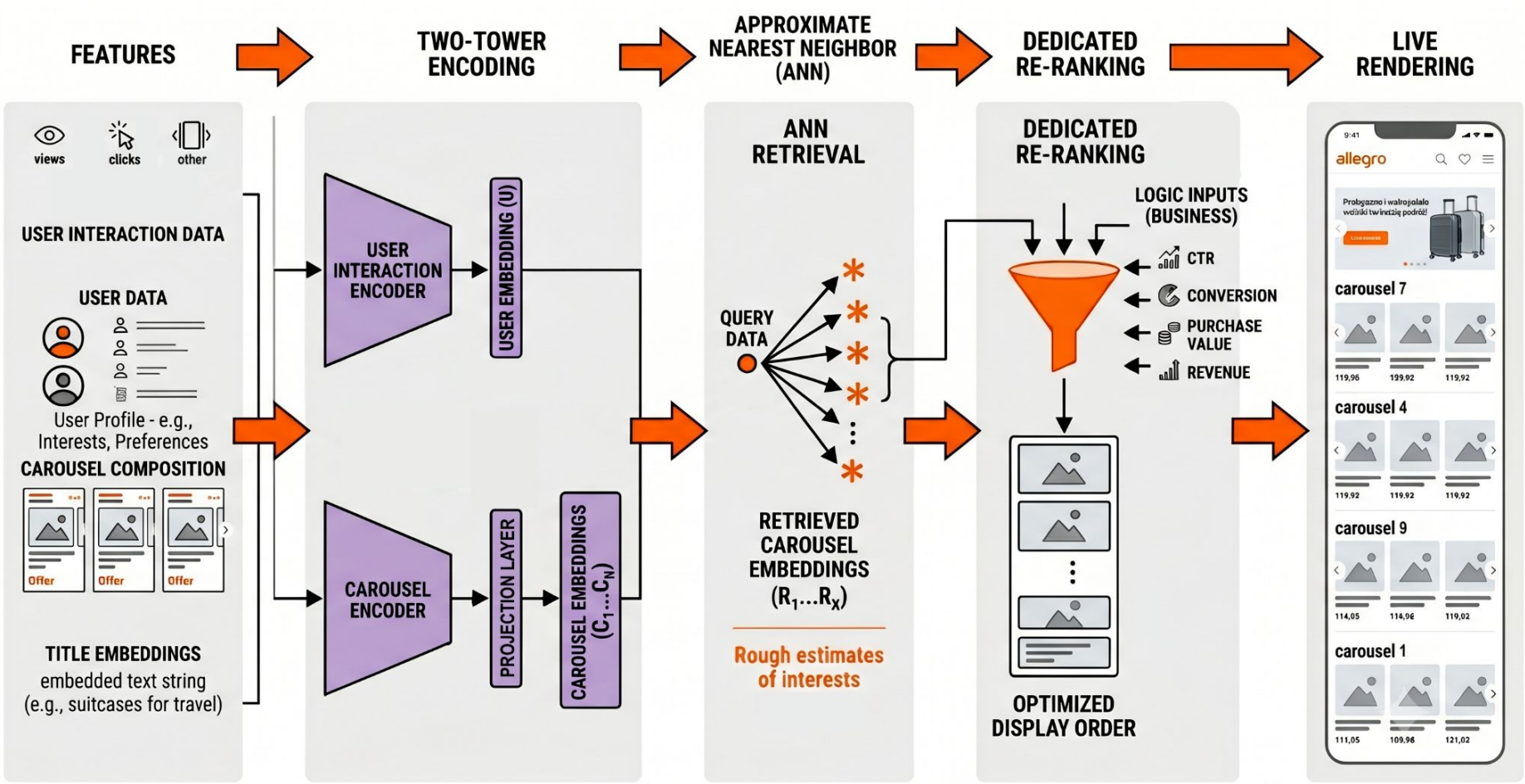
Choosing what to display

- How do we optimize for display order?
  - Dedicated Ranking step
  - We'll also take into account business-relevant metrics

Retrieval + Ranking

Ranking

- Complexity vs control trade-off
  - Black box model that could rank very well
  - Simpler, controllable model, to adjust, depending on needs


FEATURES
TWO-TOWER ENCODING
APPROXIMATE NEAREST NEIGHBOR (ANN)
DEDICATED RE-RANKING
LIVE RENDERING
views
clicks
other
USER INTERACTION DATA
USER DATA
User Profile - e.g., Interests, Preferences
CAROUSEL COMPOSITION
Offer
Offer
Offer
TITLE EMBEDDINGS
embedded text string (e.g., suitcases for travel)
USER INTERACTION ENCODER
USER EMBEDDING (U)
CAROUSEL ENCODER
PROJECTION LAYER
CAROUSEL EMBEDDINGS ($C_1 \ldots C_N$)
ANN RETRIEVAL
QUERY DATA
RETRIEVED CAROUSEL EMBEDDINGS ($R_1 \ldots R_X$)
Rough estimates of interests
DEDICATED RE-RANKING
LOGIC INPUTS (BUSINESS)
CTR
CONVERSION
PURCHASE VALUE
REVENUE
OPTIMIZED DISPLAY ORDER
9:41
allegro
carousel 7
119,96
199,92
119,92
carousel 4
119,92
119,92
119,92
carousel 9
114,05
114,96
119,02
carousel 1
111,05
109,98
121,02

Feedback

Zeskanuj kod i zostaw swoją opinię

DATA SCIENCE SUMMIT AI EDITION

Deep Learning at Scale: Tailoring the Home Page for Millions.

Jakub Żmujdzin, Marcin Szpunar

https://ml.dssconf.pl/user.html#!/lecture/DSSML26-8638/rate

Visit:

jobs.allegro.eu

ai.allegro.tech