Kahoona

Y-DATA

BEHAVIORAL CROSS-SESSION USER IDENTIFICATION

Evgenia Amineva
Rima Rohana

THE TEAM

Y-Data Students

Evgenia Amineva, background in Project Management

Rima Rohana, background in Software Engineer

Kahoona

Noy Nissim, Data scientist, Y-Data graduated

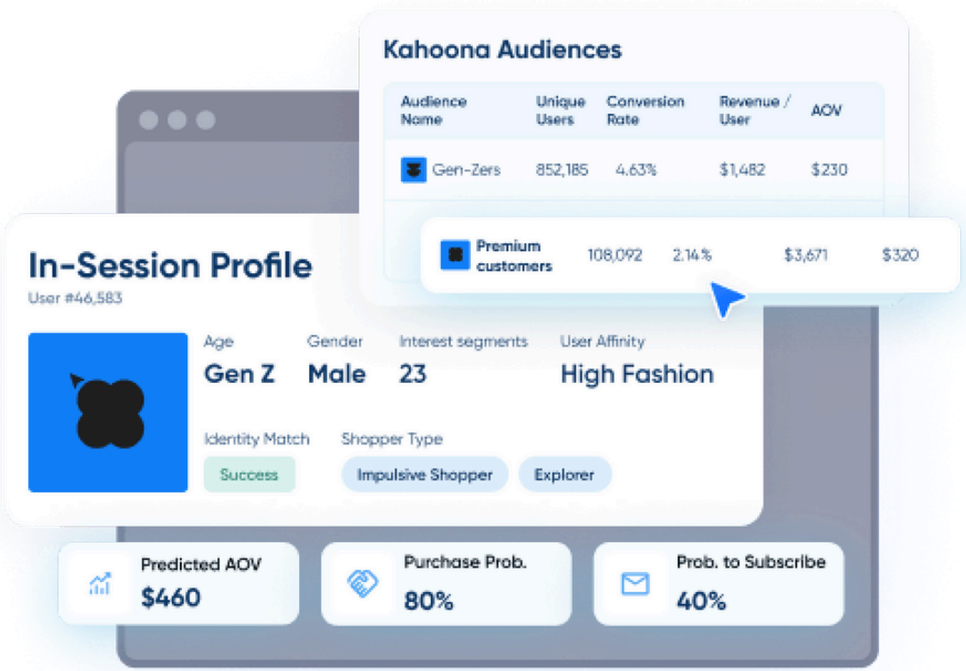
Nadav Kallenberg, Data Science Team lead

The Mentor

Oren Elisha, Data Science Manager, Forter

THE COMPANY

Kahoona



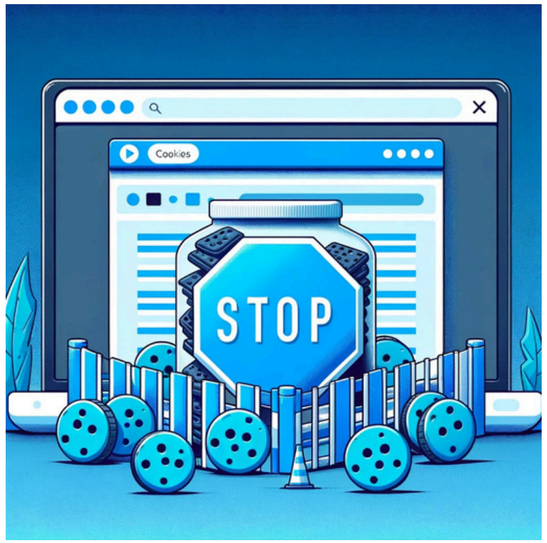
A first-party data activation platform that provides an essential solution for a cookie-free and identity-free digital ecosystem

FOUNDATIONS

95%

of E-commerce websites visitors are unidentified

Third-party cookie restriction

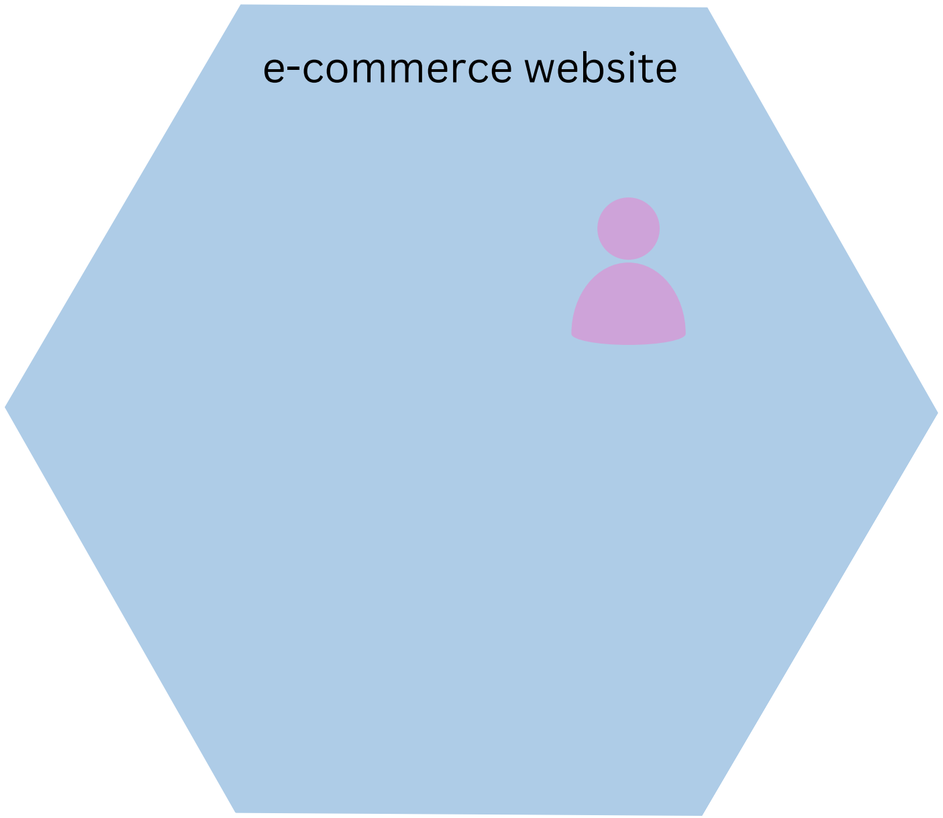


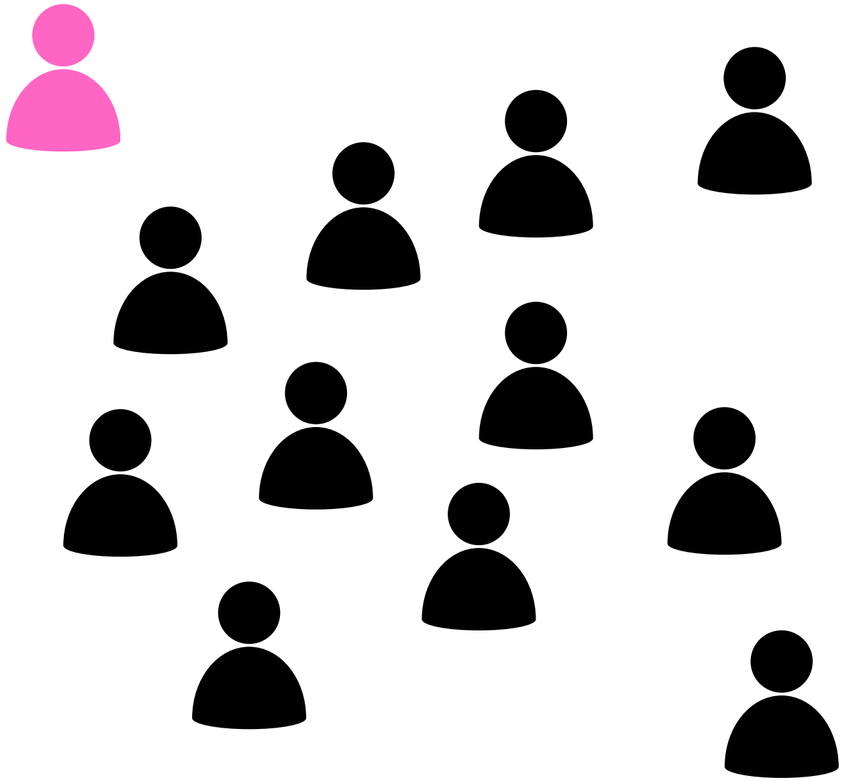
User behavior

New perspective on user behavior in E-commerce websites that focuses on the ‘how’ of user interaction rather than ‘what’.

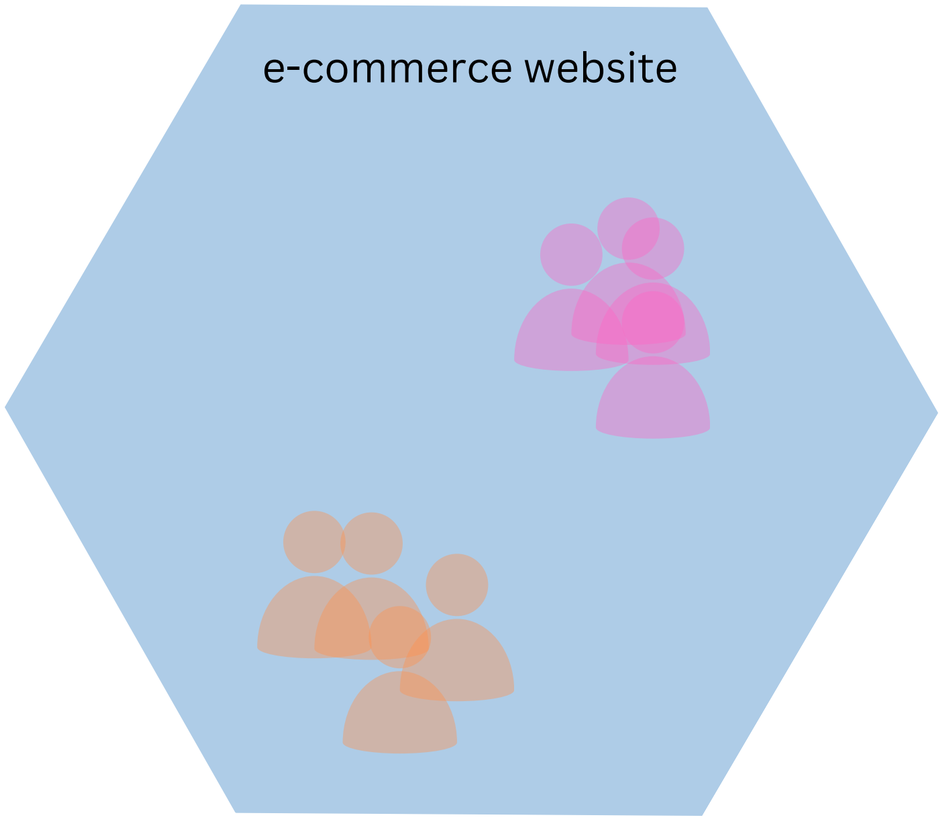
PROBLEM

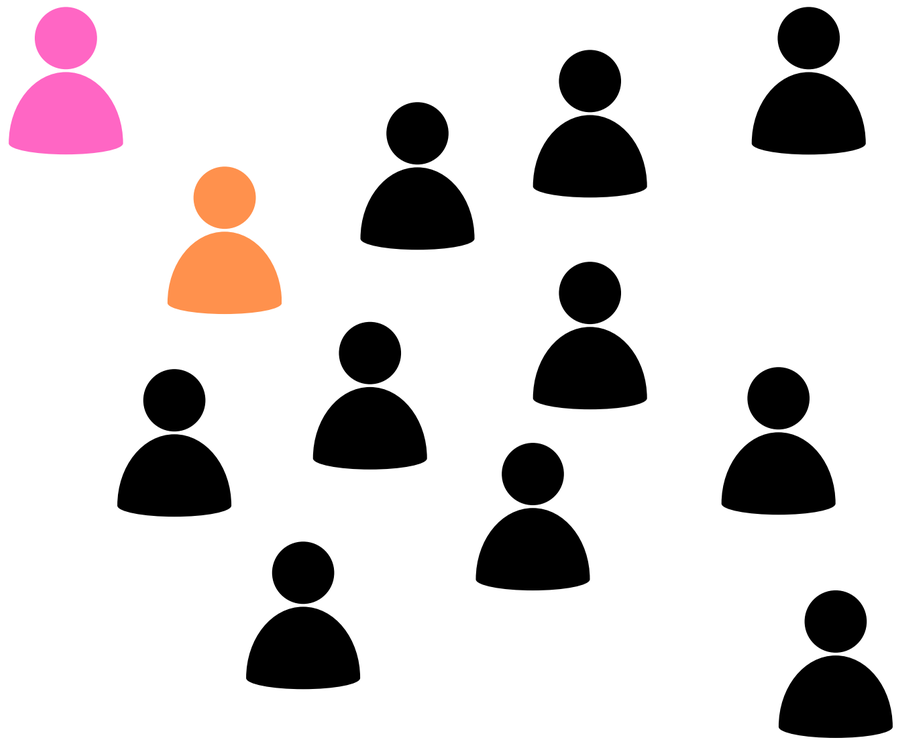
Identification of not logged-in returned users





SIMILARITY PROBLEM





INITIAL GOALS

Main goal:

Find 3 most likely matching users and beat the company's accuracy 90%

Advances goal:

Develop a mechanism that picks the most likely matched user

THE DATA

Hashed IDs: session, pageview ...	1db1bb7c-242e-4bb8-bb63-c3b56e4459a3	2f274e9a-5e37-45fa-aec4-a71d5d190150	
Timestamps: session, pageview ...	1677352958273	1677353614099	1677352958249
Environmental data: user_agent, referrer	Mozilla/5.0 (iPhone; CPU iPhone OS 15_4_1 like...		
External IDs:	None	None	abFBKzTsO0FBSP8LhM2ACMzwel
Behavioral data: business_data ...	[{'reloads': 0, 'eventtype': 'khn_loaded', 'ev...	[{"events":"[{\"et\":\"load\",\"ets\":16773536...	
Label:			**user_id** 7eb1033072cc795e9340dc7f09e57a29524dfc89652c70...

THE DATA

33 features
2 339 008 entries

Period

January 21, 2023 – March 12, 2024

Parquet

file format

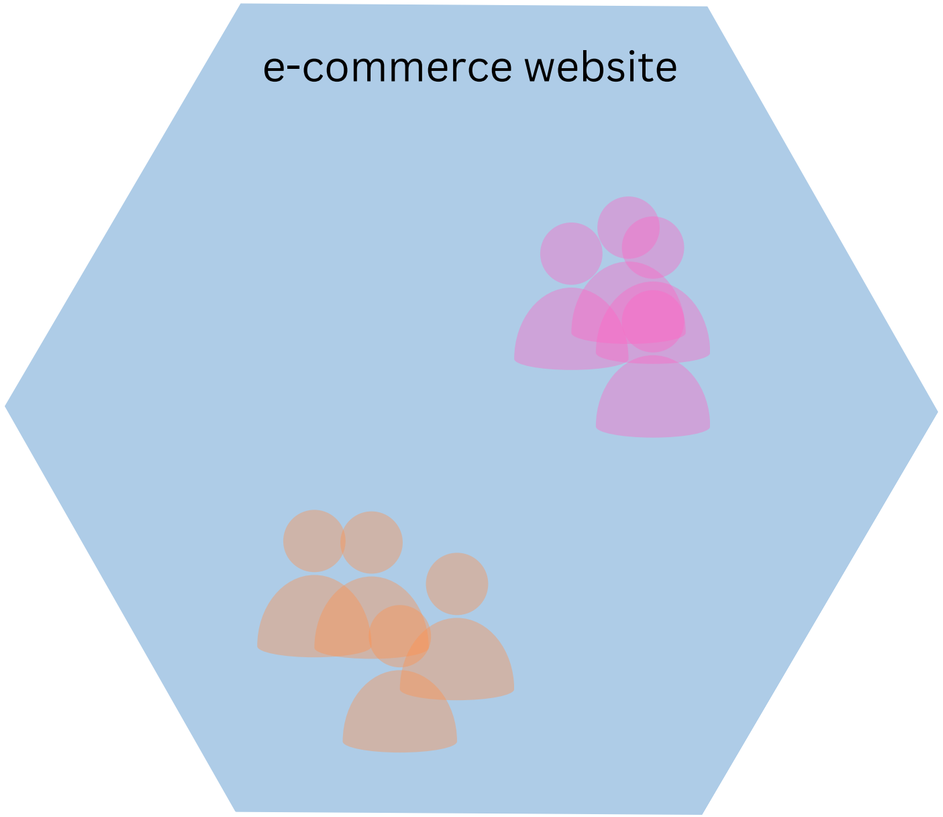
8.32 GB

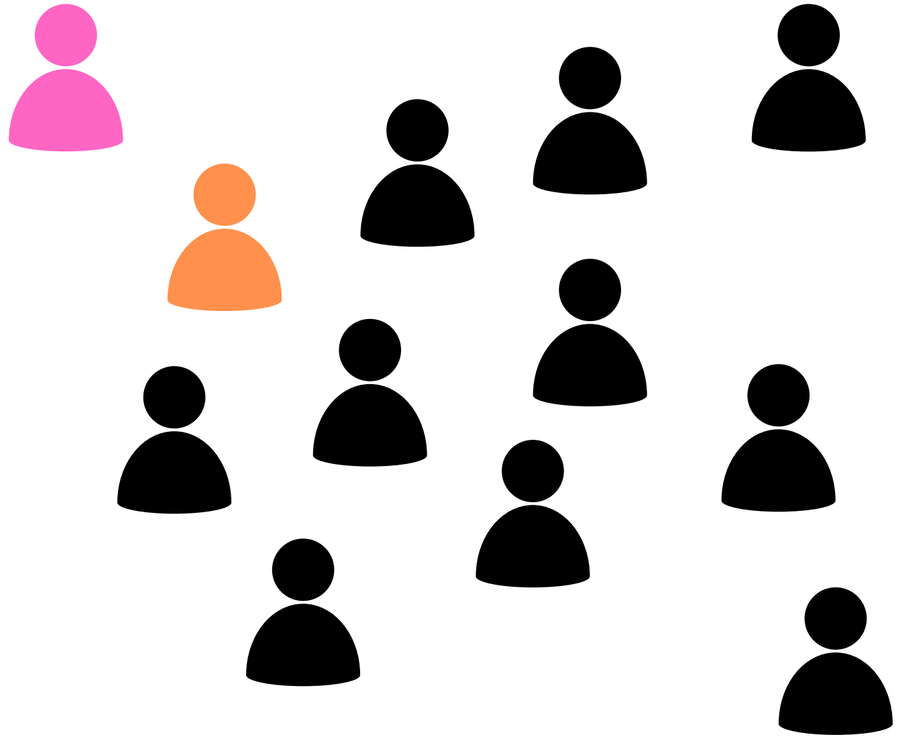
Source

e-commerce
website

USER TRACES

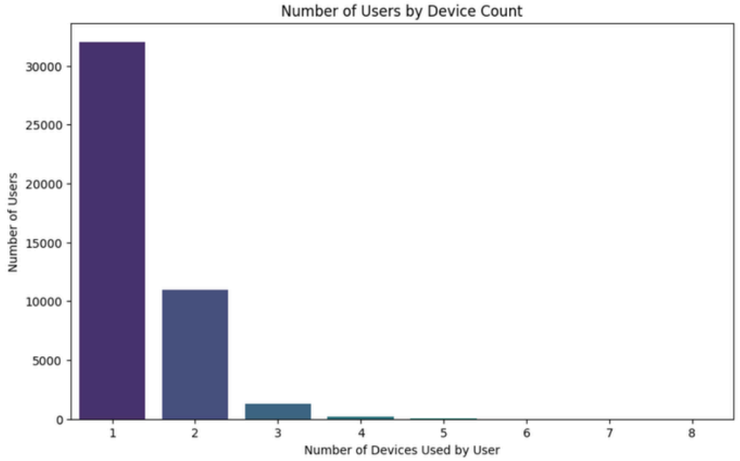
why we assume that users leave traces?

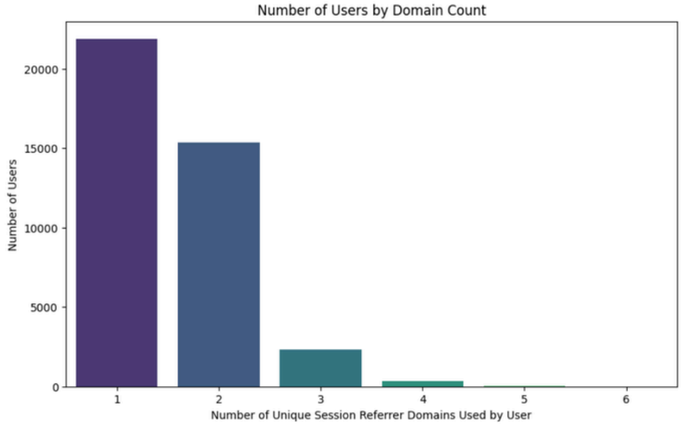


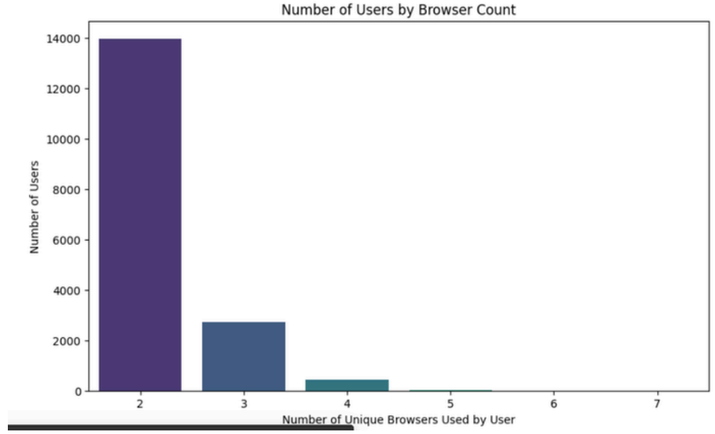


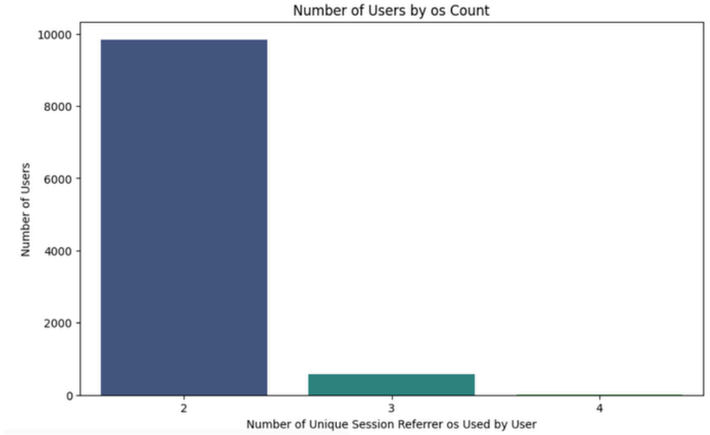
USER TRACES

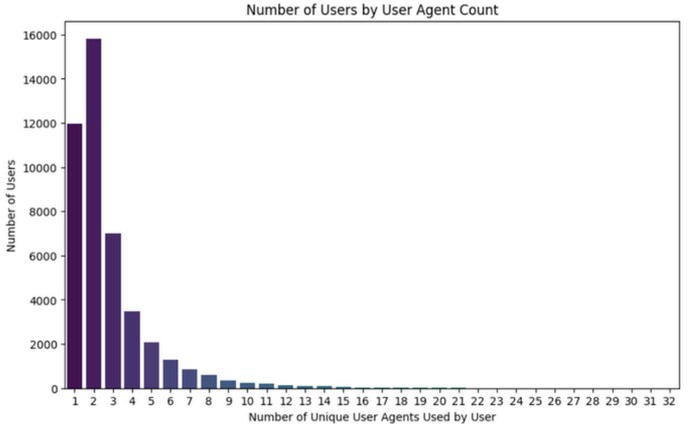
Environmental Data EDA

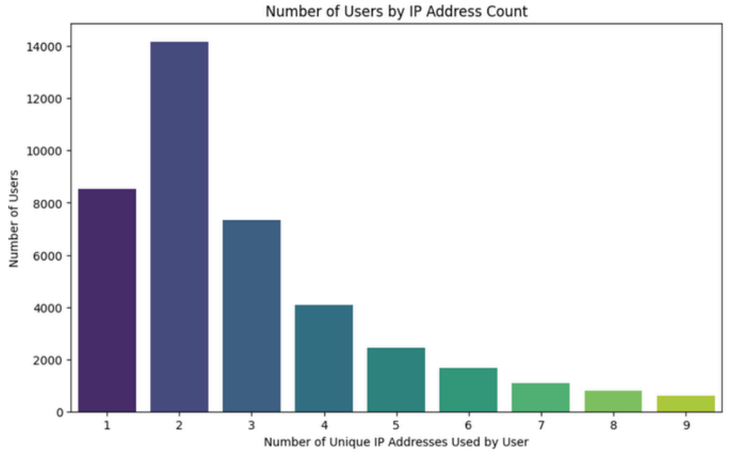






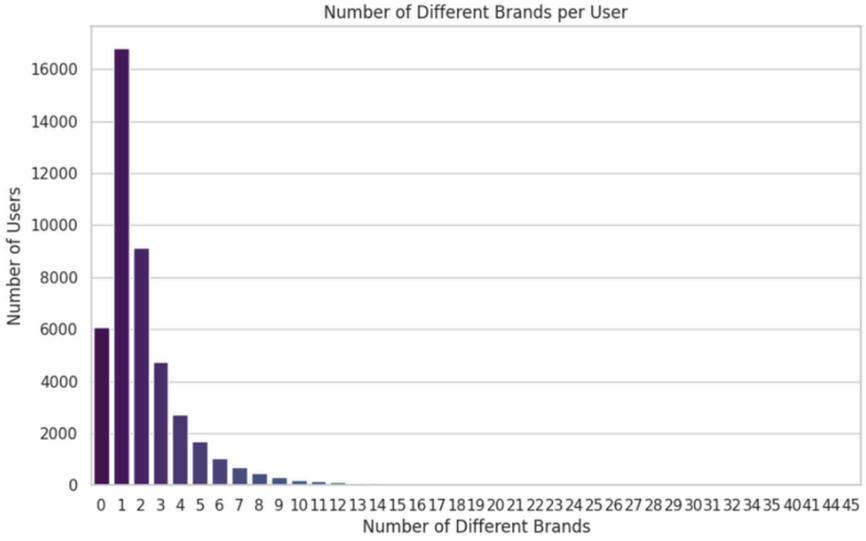


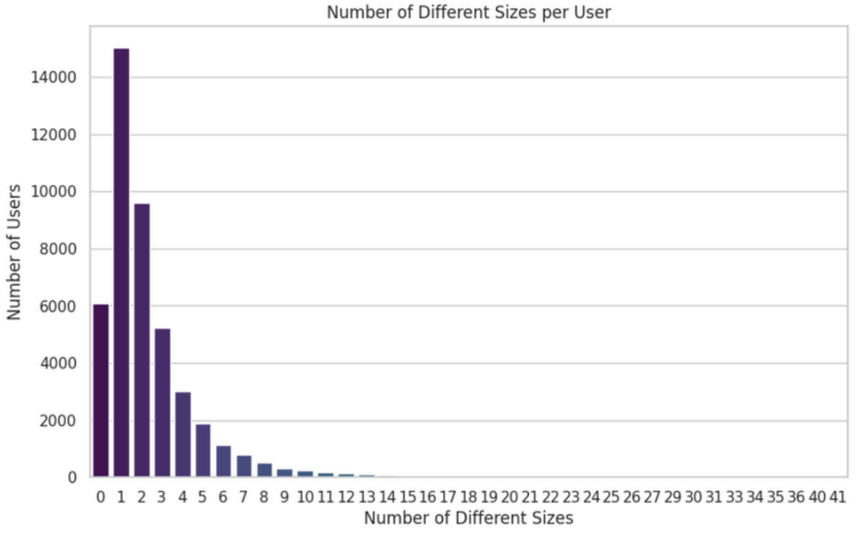


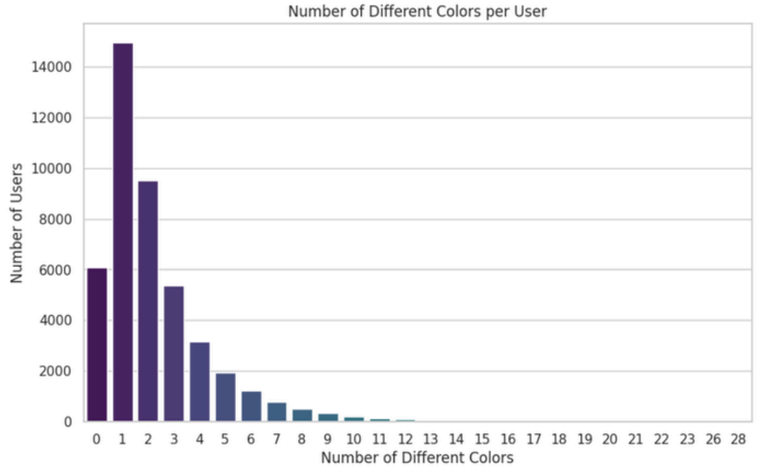


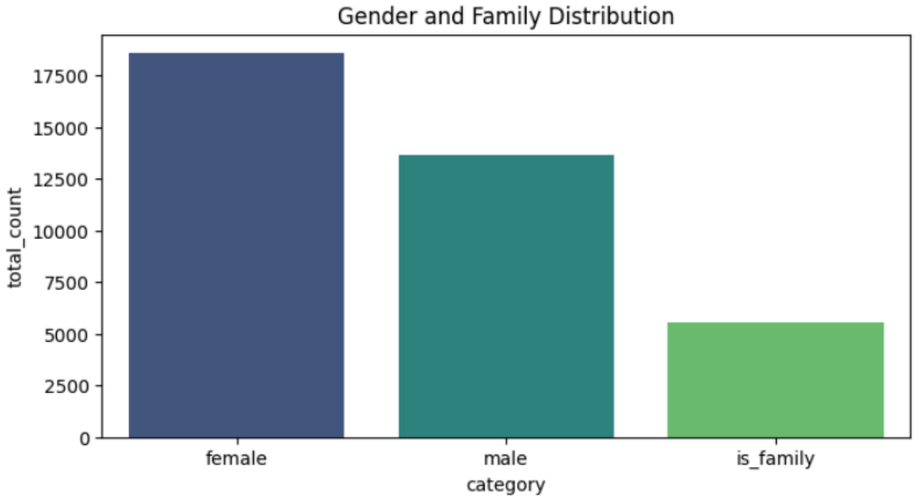
USER TRACES

Behavioral Data EDA

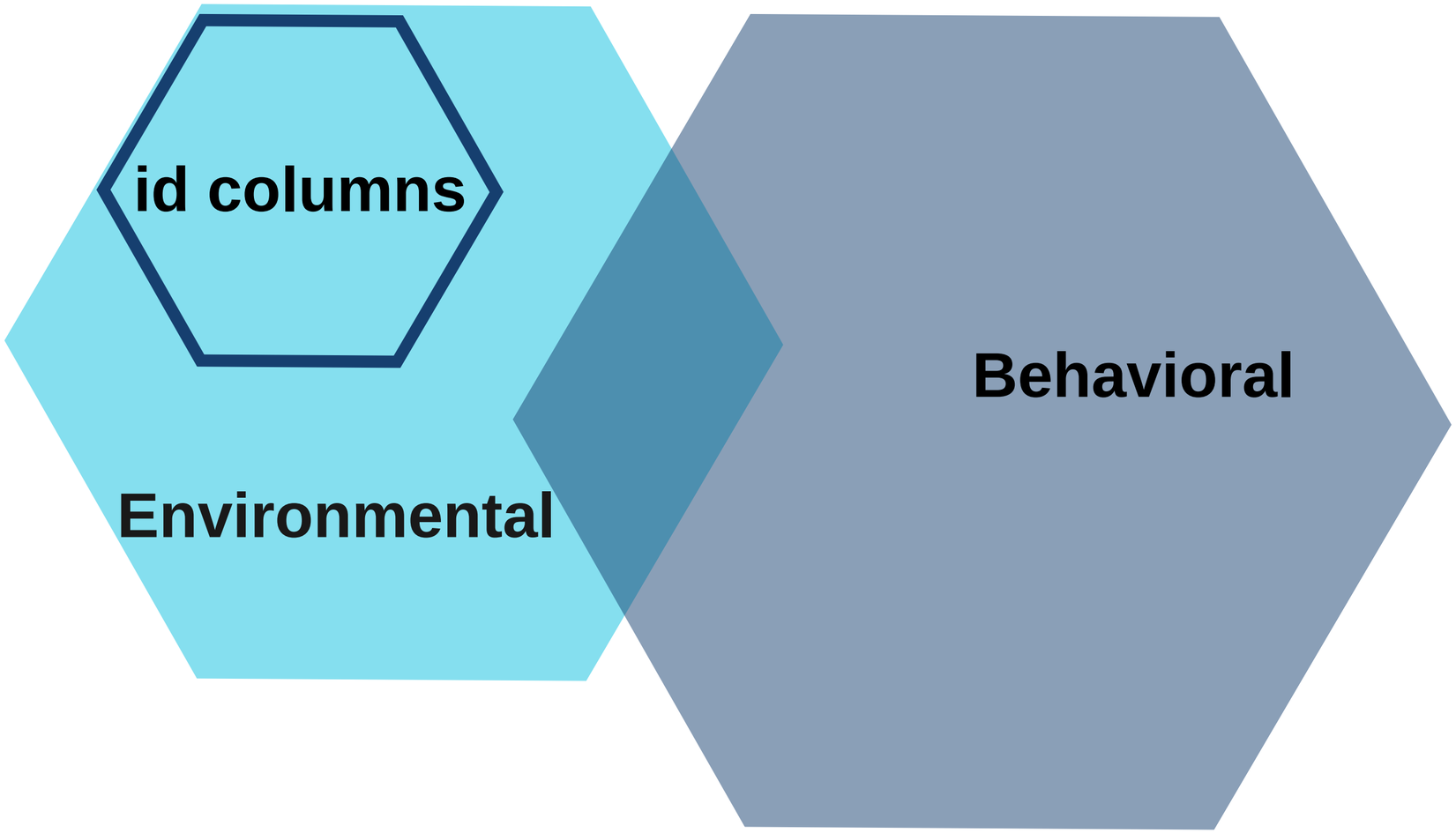




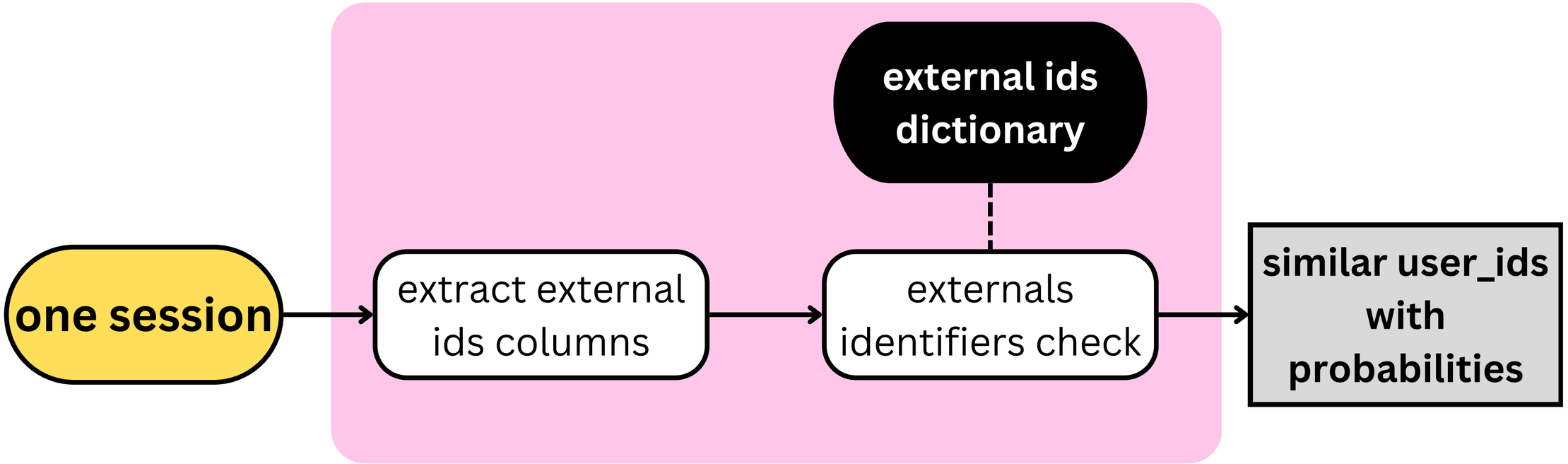




THE DATA



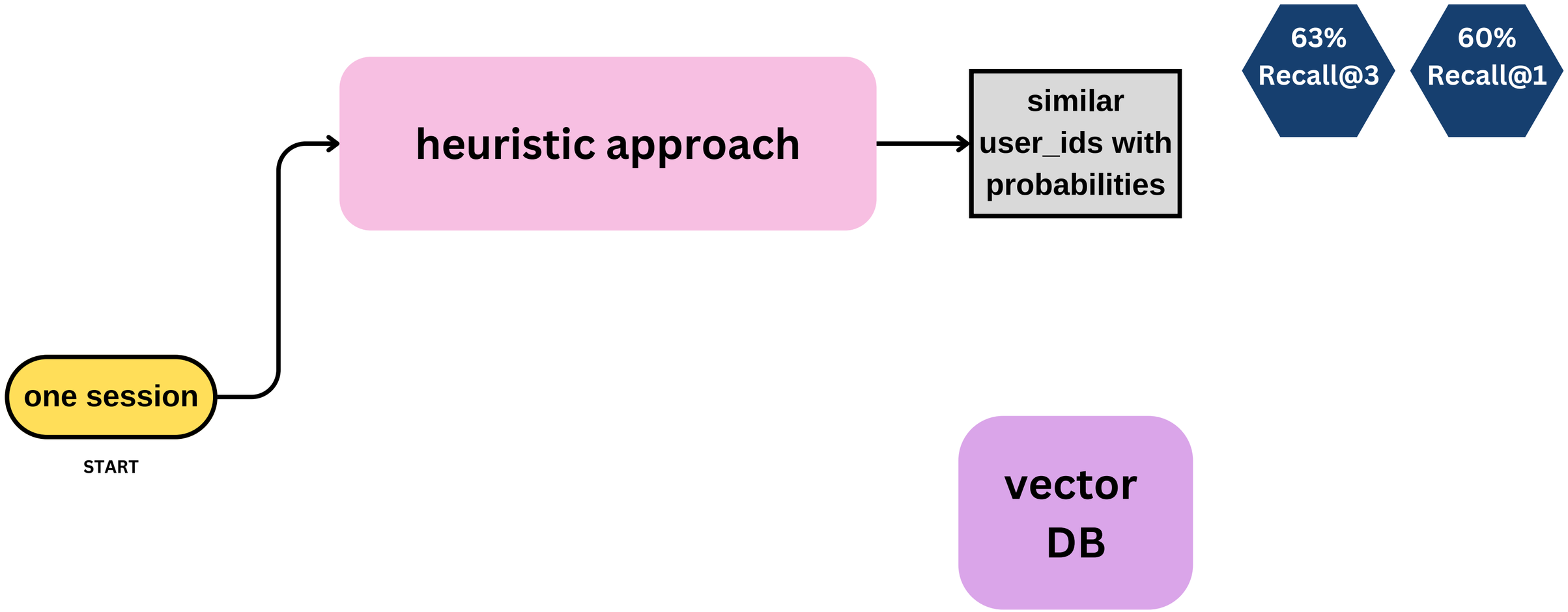
WORKFLOW



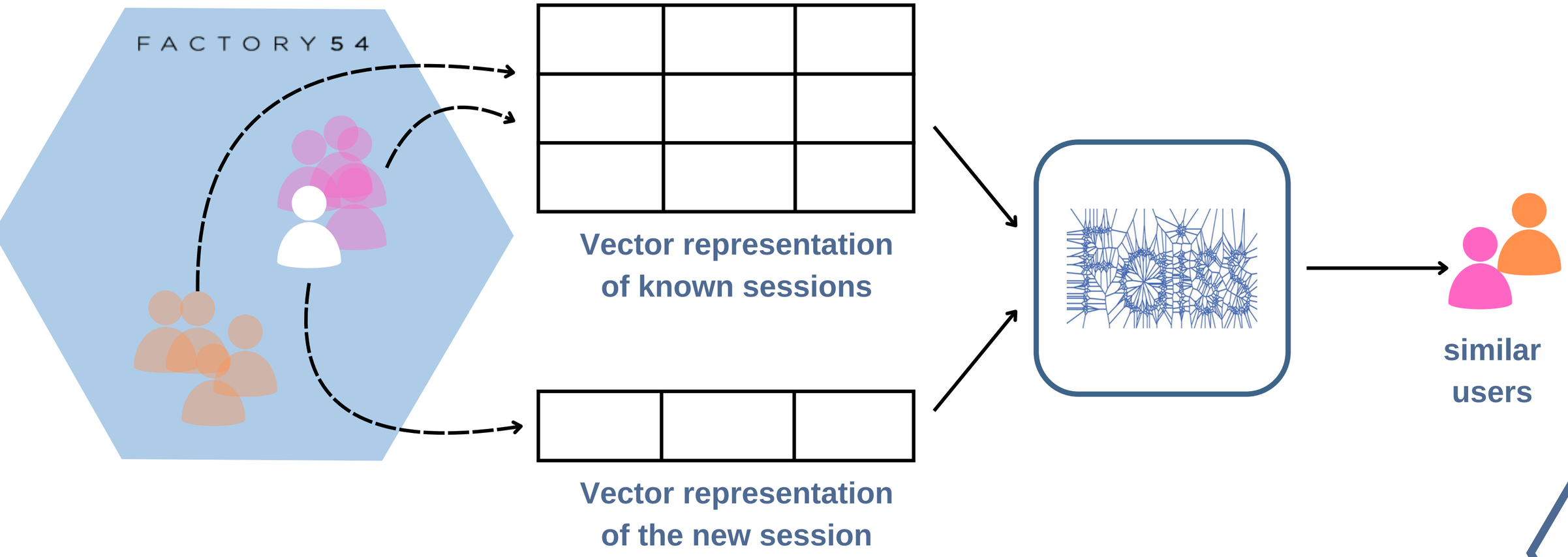
63% Recall@3

60% Recall@1

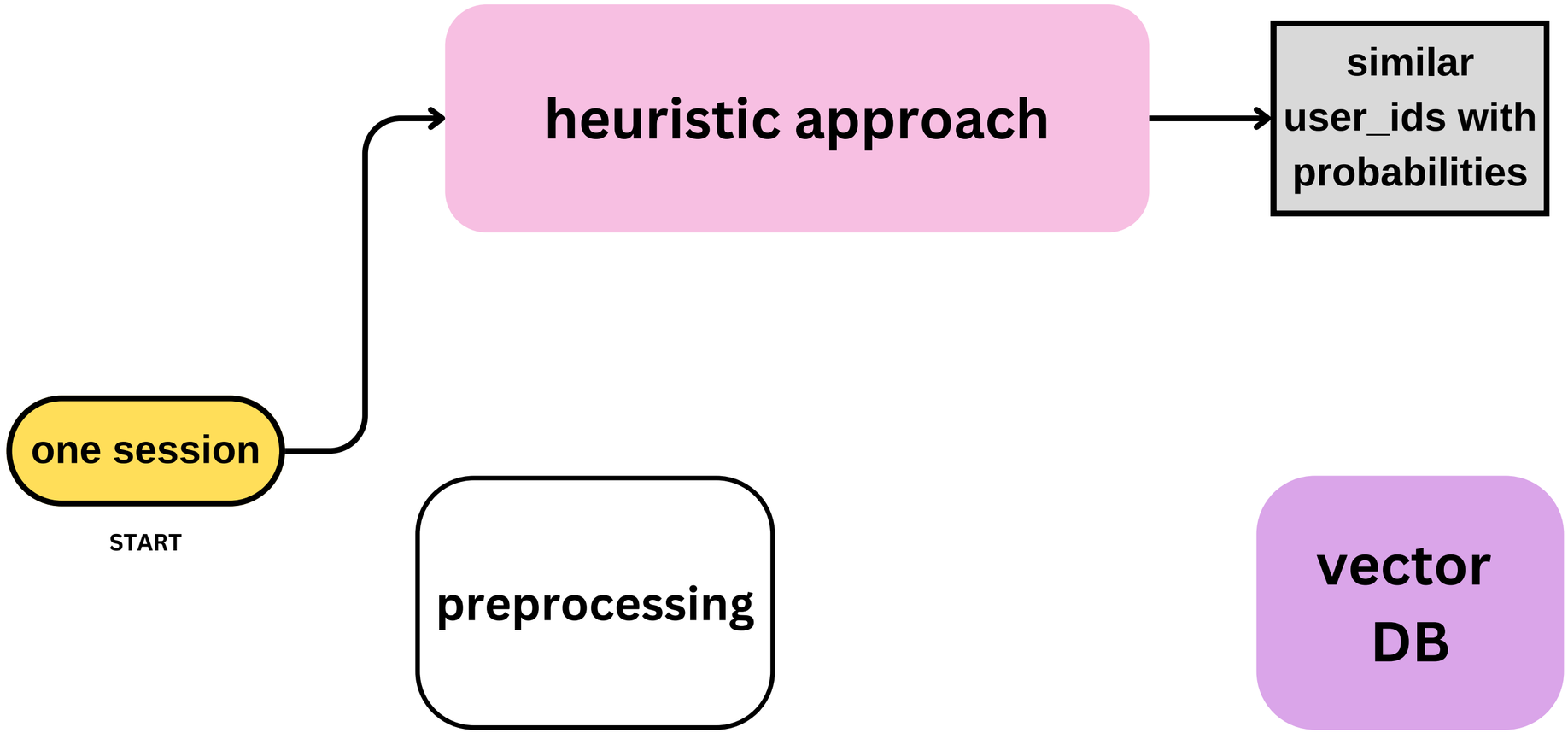
WORKFLOW


63%
Recall@3
60%
Recall@1
heuristic approach
similar
user_ids with
probabilities
one session
START
vector
DB

FAISS DB


FACTORY 54
Vector representation
of known sessions
Vector representation
of the new session
similar
users

WORKFLOW


heuristic approach
similar user_ids with probabilities
one session
START
preprocessing
vector DB

DATA PREPROCESSING

business event

[{"pageType":"productPage"},
{“events":
“[
{\\"et\\":\\"load\\",
\\”images\\”:
[\\"/on/demandware.static/-/Sites-master-catalog/default/dw46ad99d3/images/models/893624238_L_2.JPG\\",
\\"/on/demandware.static/-/Sites-master-catalog/default/dw778a6840/images/large/893624238_P_1.png\\",
\\"/on/demandware.static/-/Sites-master-catalog/default/dwfcbe4a2c/images/models/893624645 893623745
893624534 893624683 893624238_L T.JPG\\",\\"/on/demandware.static/-/Sites-master-
catalog/default/dw4f3e7a7d/images/large/893624238_P_2.png\\",
\\”/on/demandware.static/-/Sites-master-catalog/default/dw73f6e730/images/large/893624238_P_3.png\\",\
\”/on/demandware.static/-/Sites-master-catalog/default/dwf89e499a/images/large/893624238_P_4.png\\”
],
\\”productId\\":\\"893624238\\",
\\"brand\\":\\"CALVIN KLEIN\\”,
\\"price\\":\\"₪ 409.00\\”,
\\”ets\\":1674580753264}]“,
”pageType":"productPage"},
{“events":
“[{\\"et\\":\\"changeColor\\",
\\”productId\\":\\"893624238\\",
\\"brand\\":\\"CALVIN KLEIN\\”,
\\"price\\":\\"₪ 409.00\\”,
\\”previousColor\\":\\"BEIGE\\",
\\”pickedColor\\":\\"BLACK\\",
\\”ets\\":1674580756472}]",
“pageType":"productPage"},

{"pageType":"productPage"}]'

- size
- top brand
- gender
- language

DATA PREPROCESSING

Grouping by session

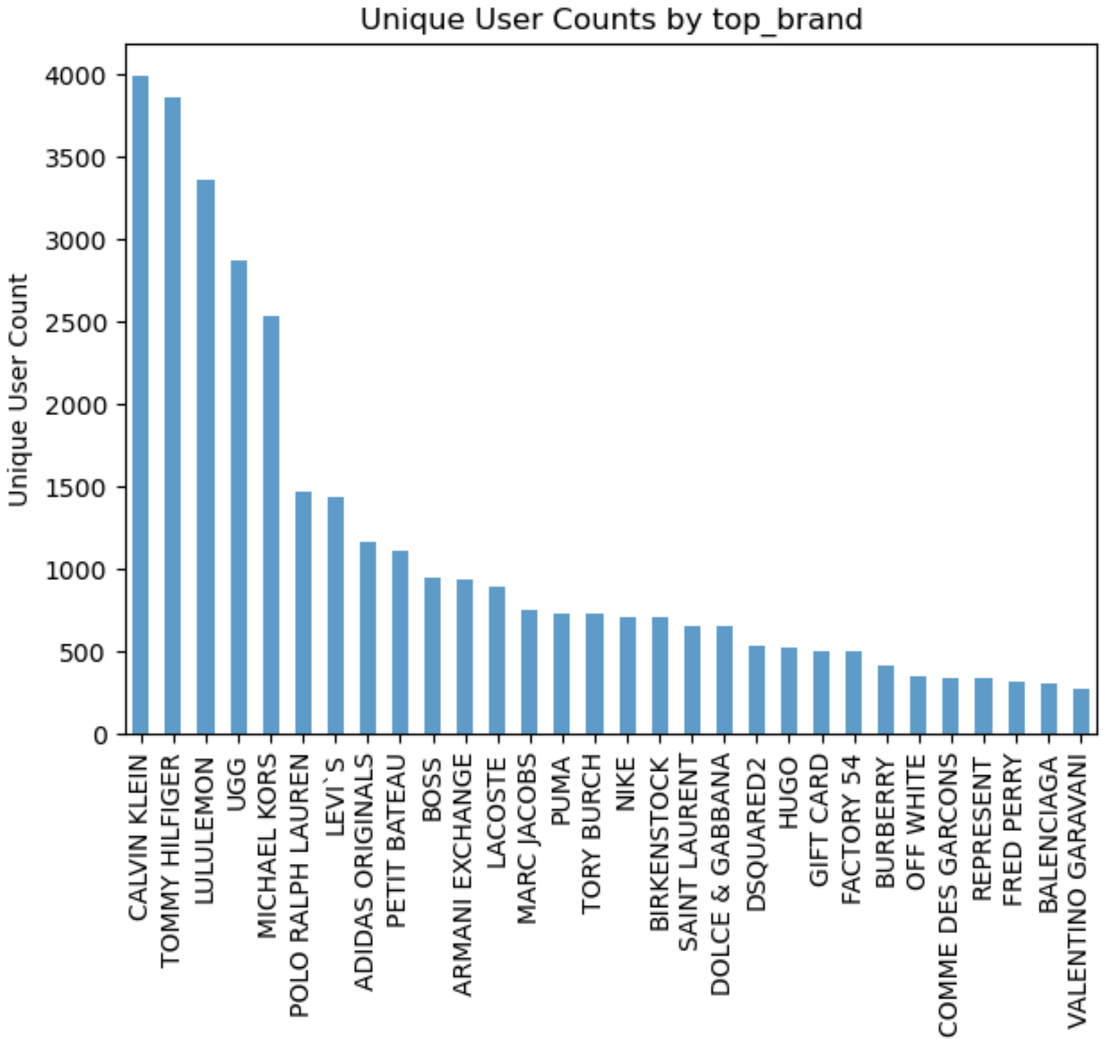
sum(): document_scroll_height

count(): pageview, tabs

mode(): in case of ambiguity during the session

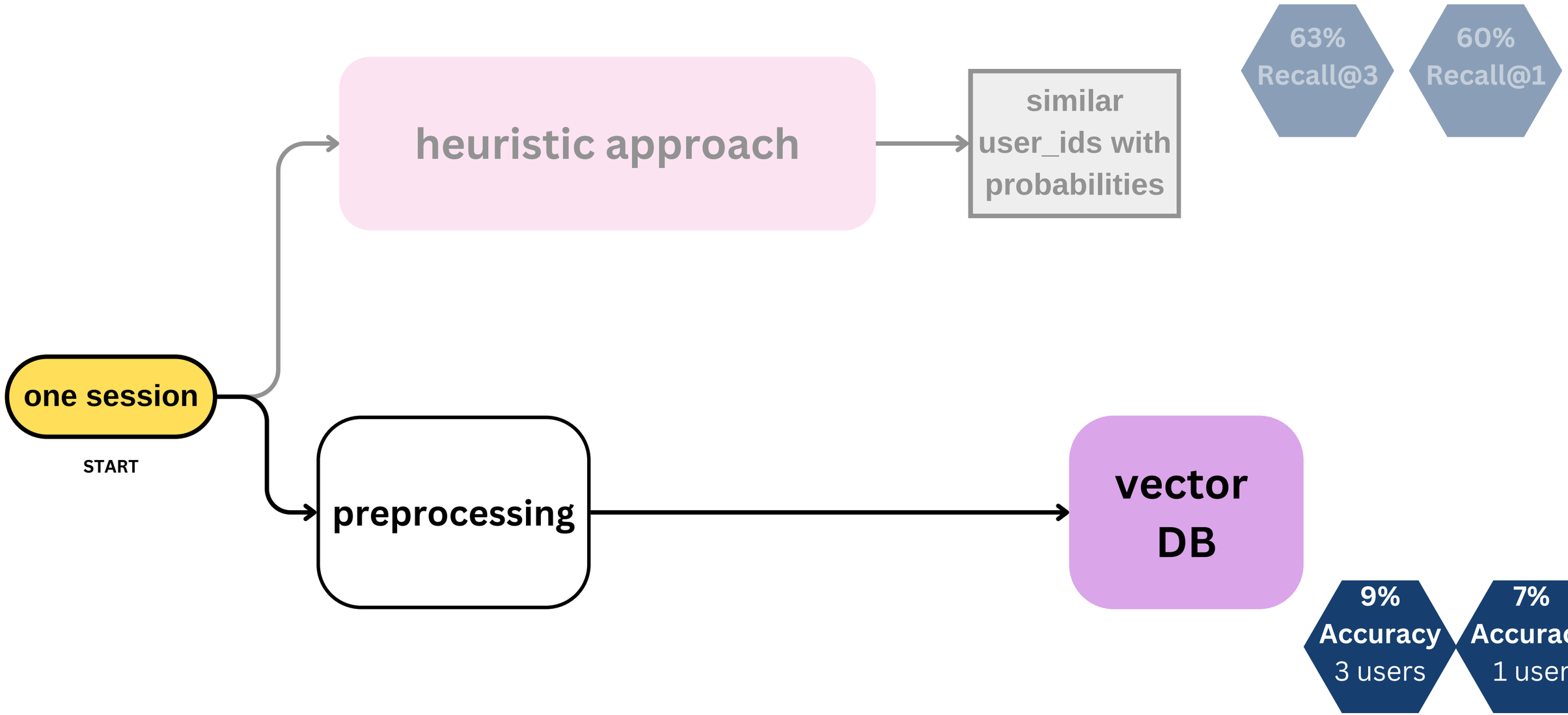
DATA PREPROCESSING

Handling categorical features

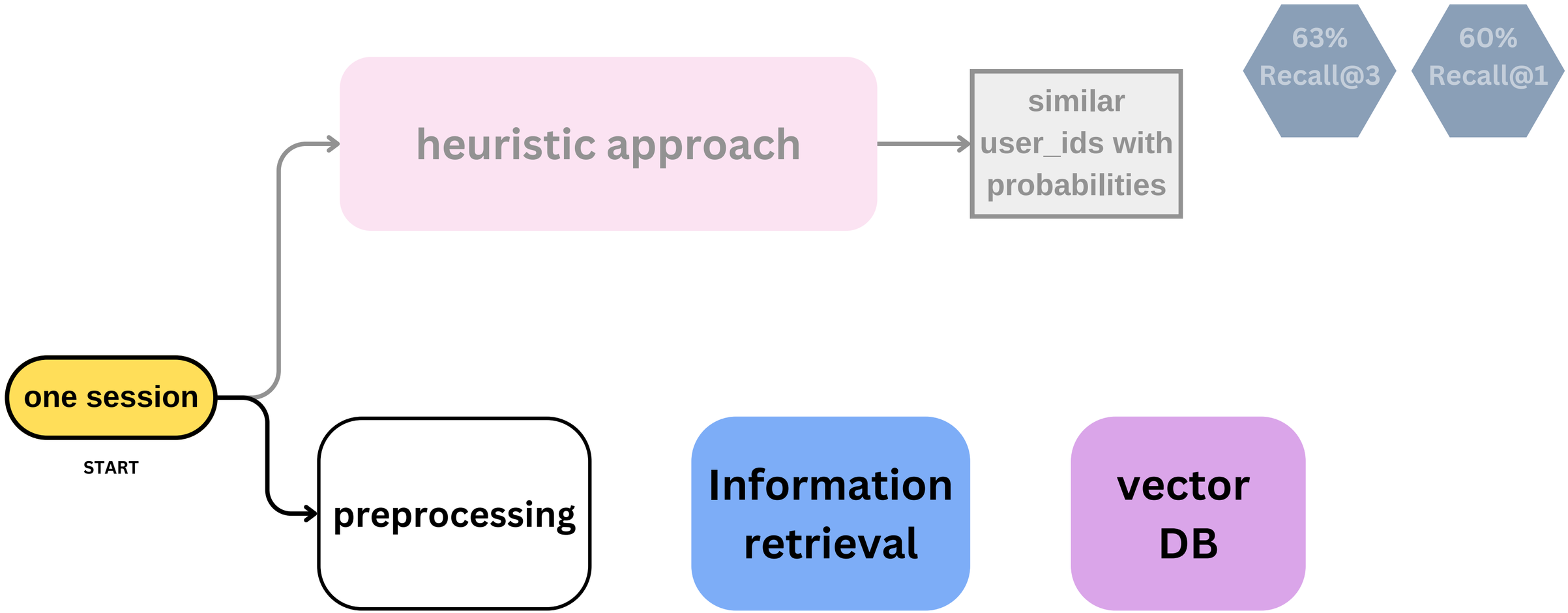


For nominal categories:
find the top ~10 values that distinguish users

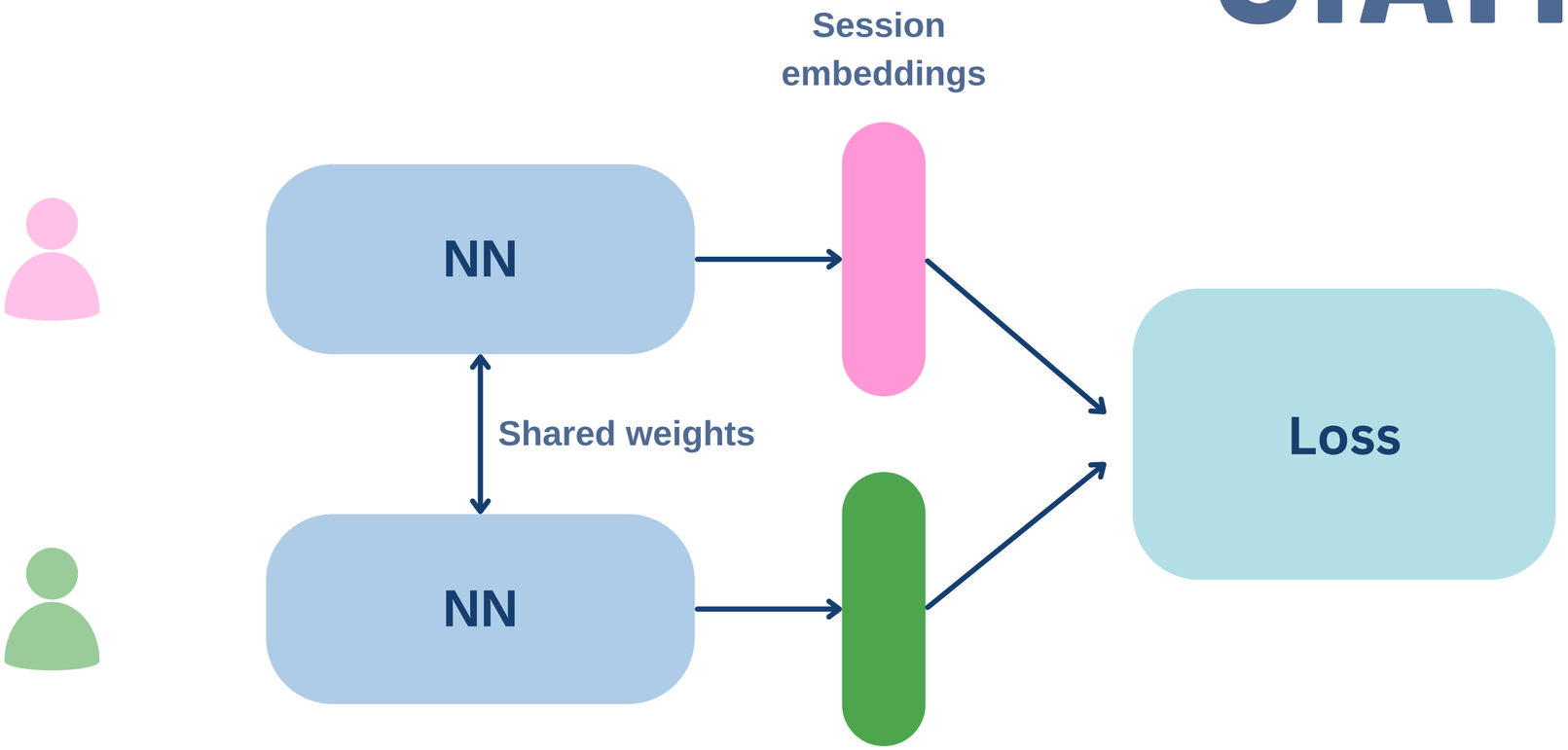
WORKFLOW


63%
Recall@3
60%
Recall@1
heuristic approach
similar
user_ids with
probabilities
one session
START
preprocessing
vector
DB
9%
Accuracy
3 users
7%
Accuracy
1 user

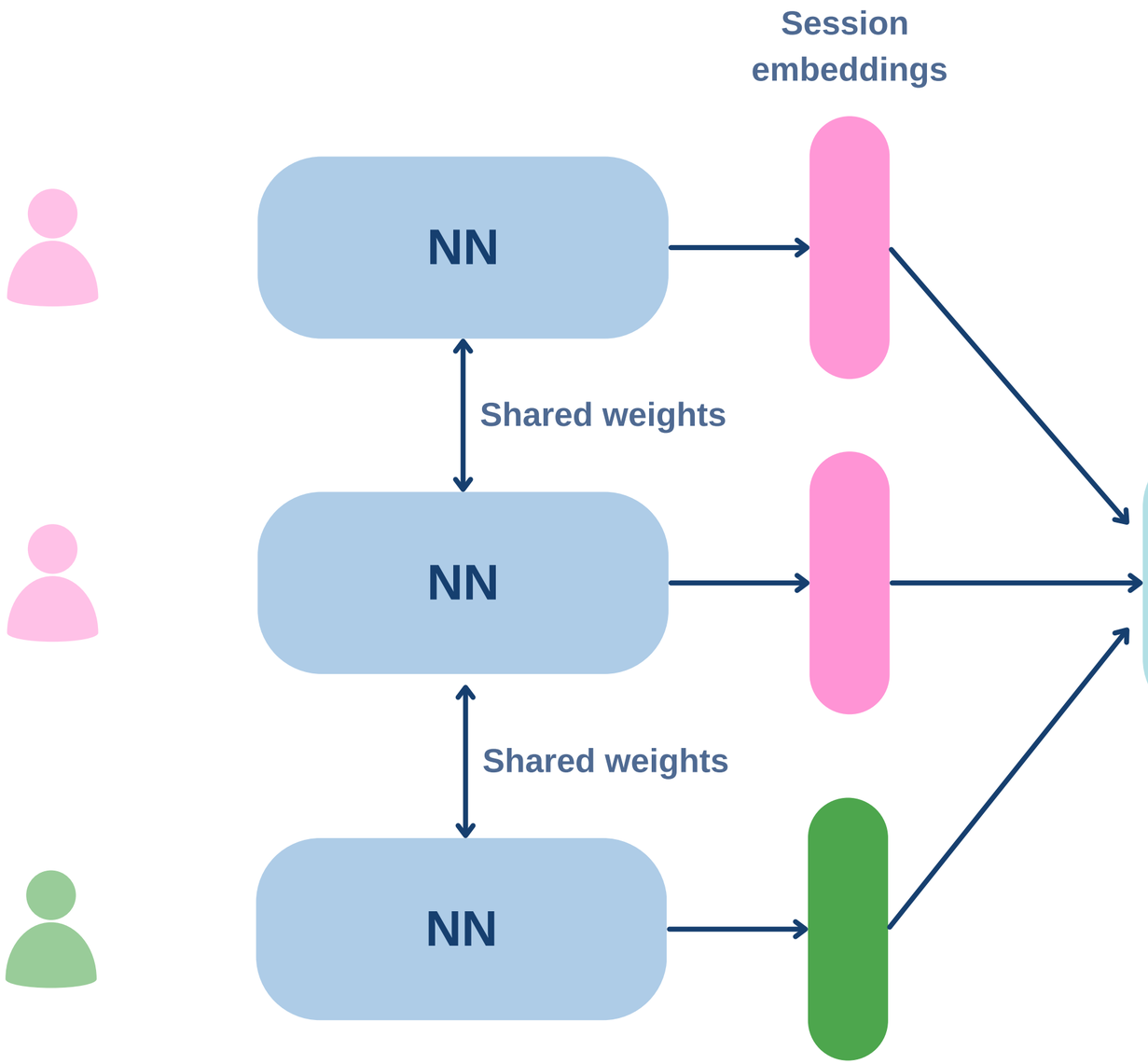
WORKFLOW


63%
Recall@3
60%
Recall@1
heuristic approach
similar
user_ids with
probabilities
one session
START
preprocessing
Information
retrieval
vector
DB

SIAMESE NN



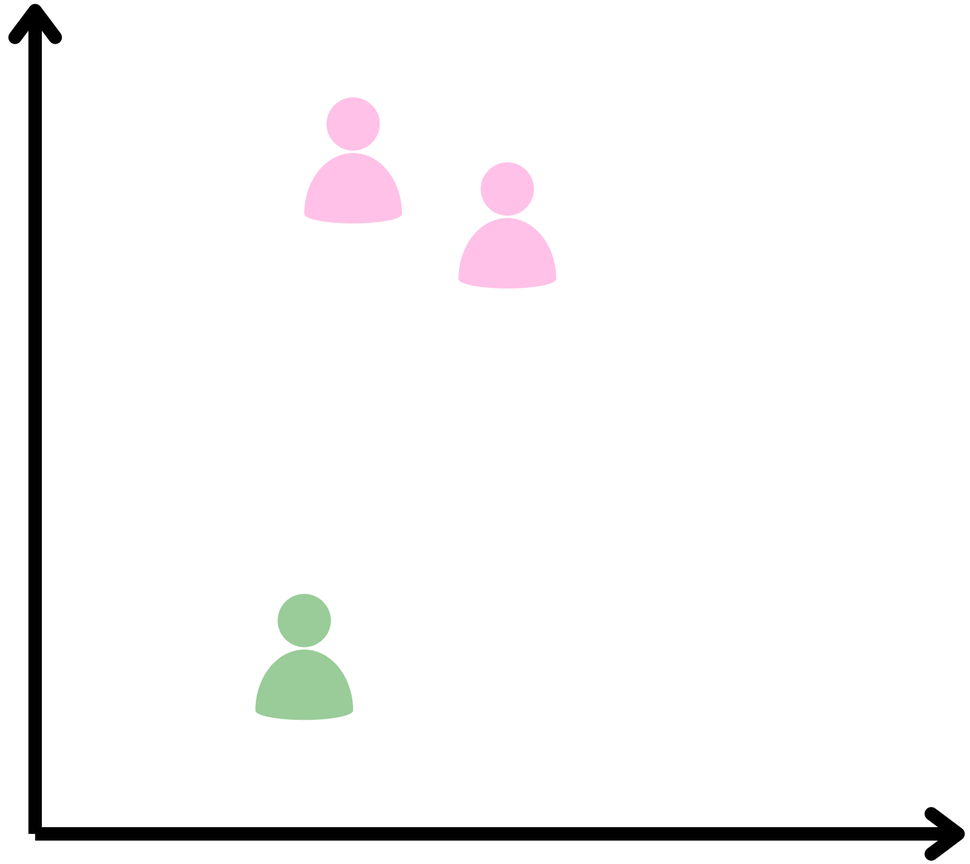
SIAMESE NN



$$L(a, p, n) = \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\}$$

$$d(x_i, y_i) = \|\mathbf{x}_i - \mathbf{y}_i\|_p$$

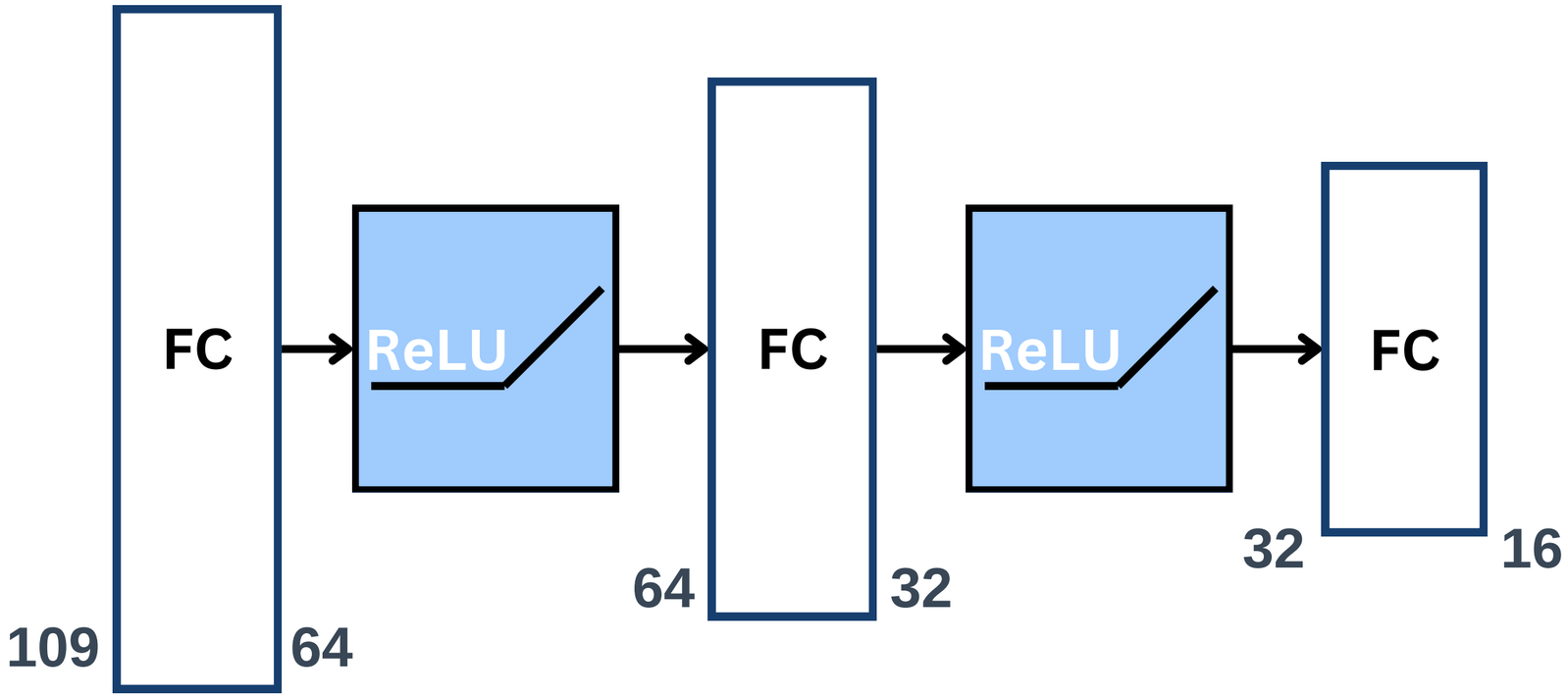
SIAMESE NN



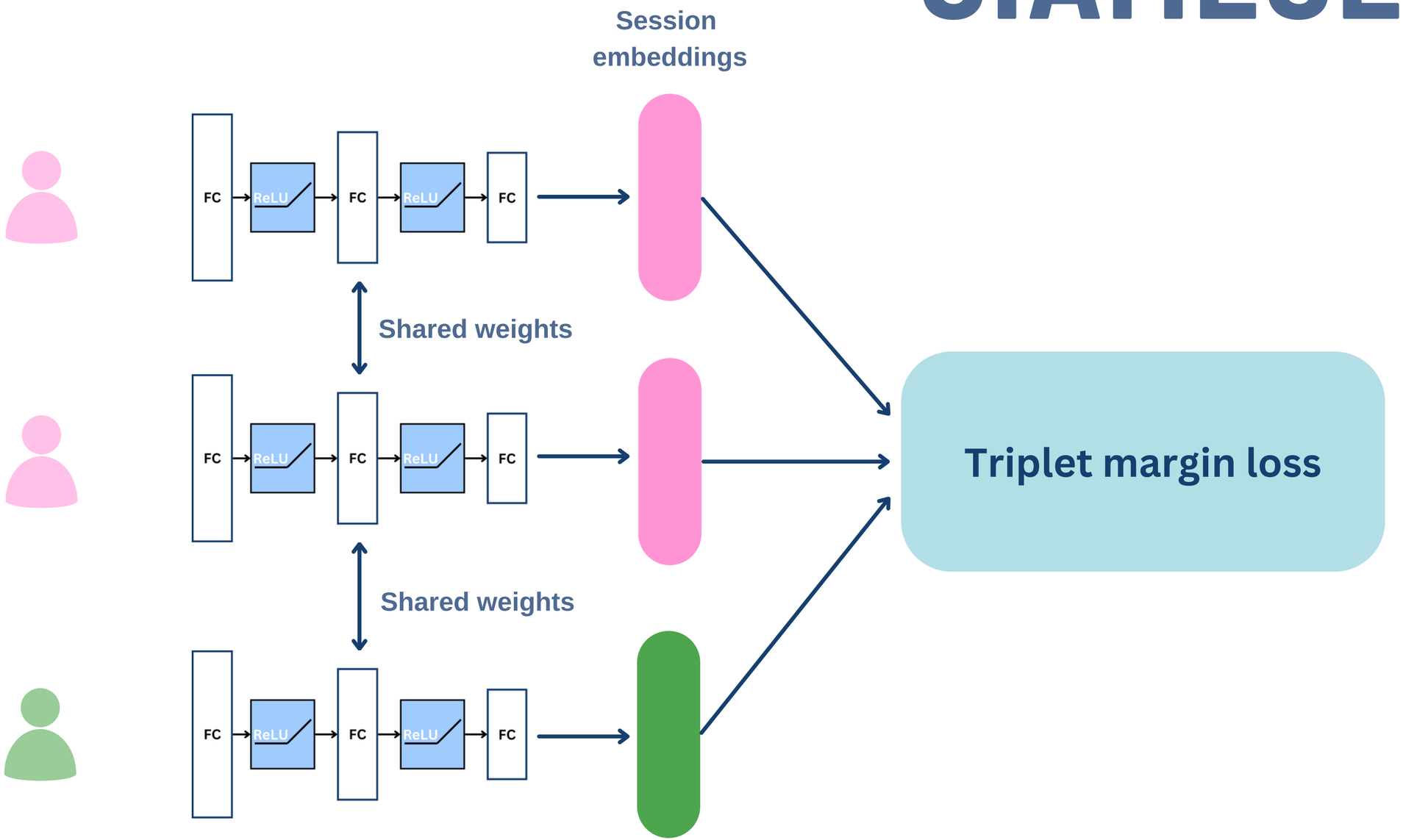
MODEL ARCHITECTURE

Custom architecture

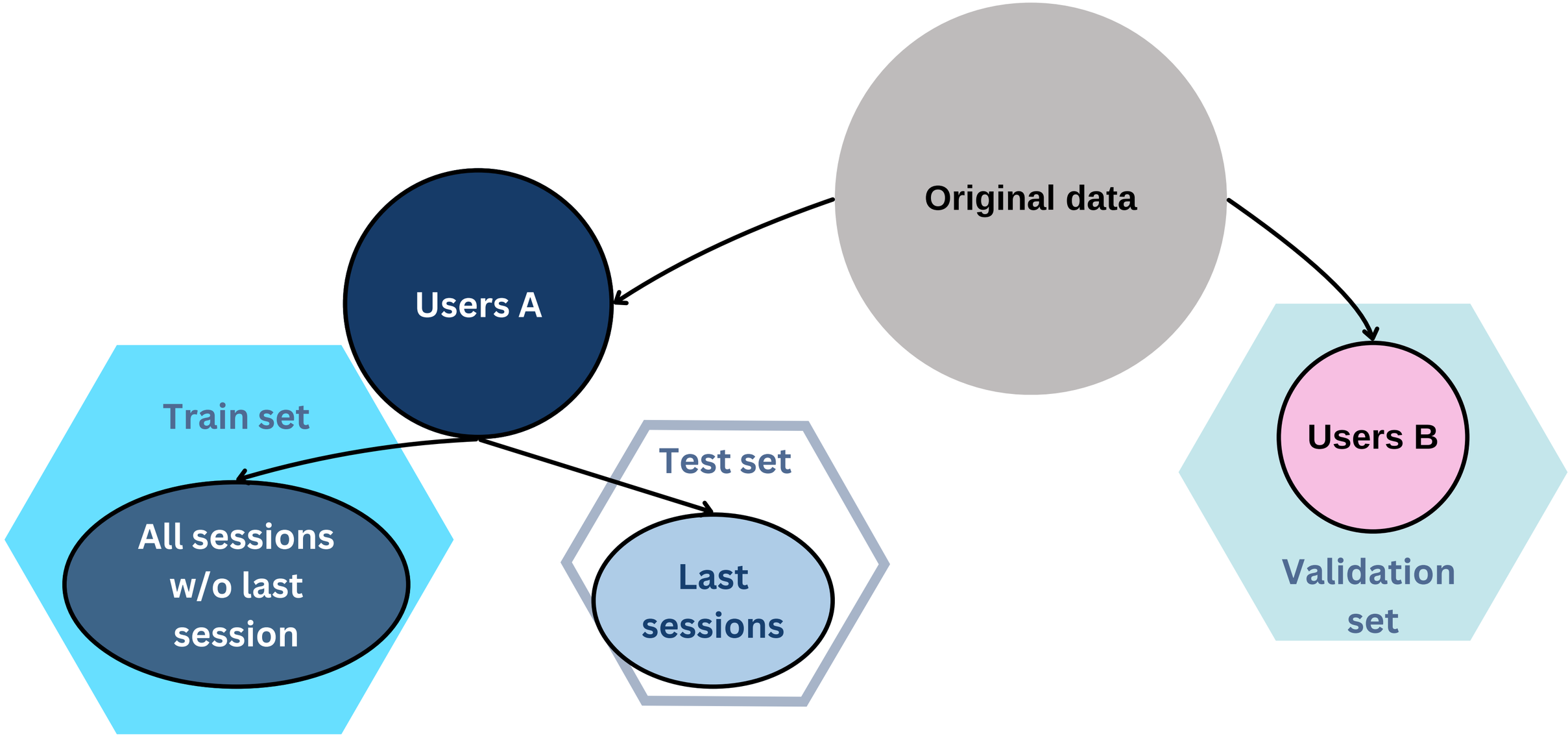
Total params: 8,432



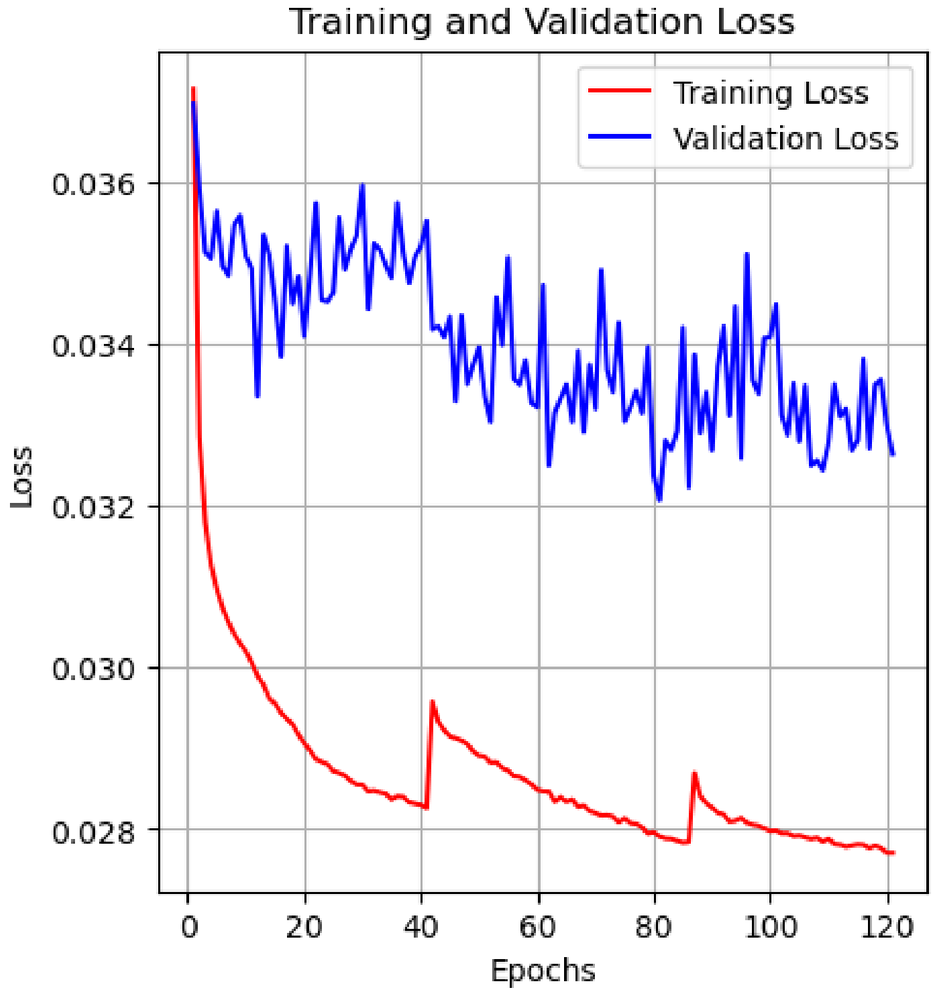
SIAMESE NN

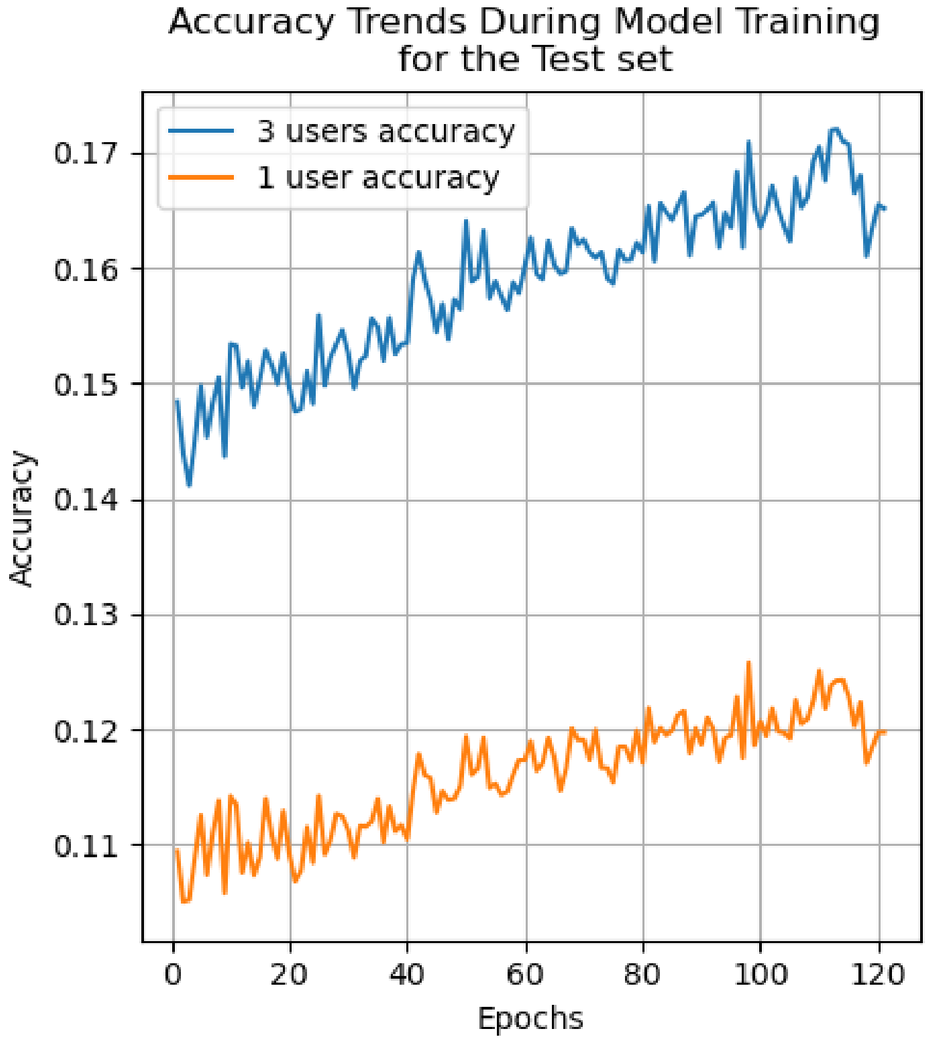

Session embeddings
FC
ReLU
FC
ReLU
FC
Shared weights
FC
ReLU
FC
ReLU
FC
Shared weights
FC
ReLU
FC
ReLU
FC
Triplet margin loss

DATASETS


Original data
Users A
Users B
Train set
All sessions w/o last session
Test set
Last sessions
Validation set

MODEL TRAINING. PROCESS





WORKFLOW

63% Recall@3

60% Recall@1

heuristic approach

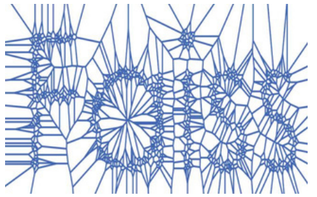
similar user_ids with probabilities

one session

START

preprocessing

Information retrieval



WORKFLOW

one session

START

heuristic approach

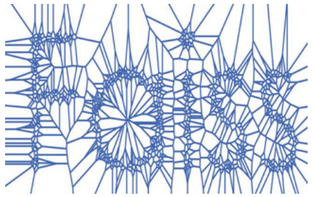
similar user_ids with probabilities

63% Recall@3

60% Recall@1

preprocessing

Information retrieval



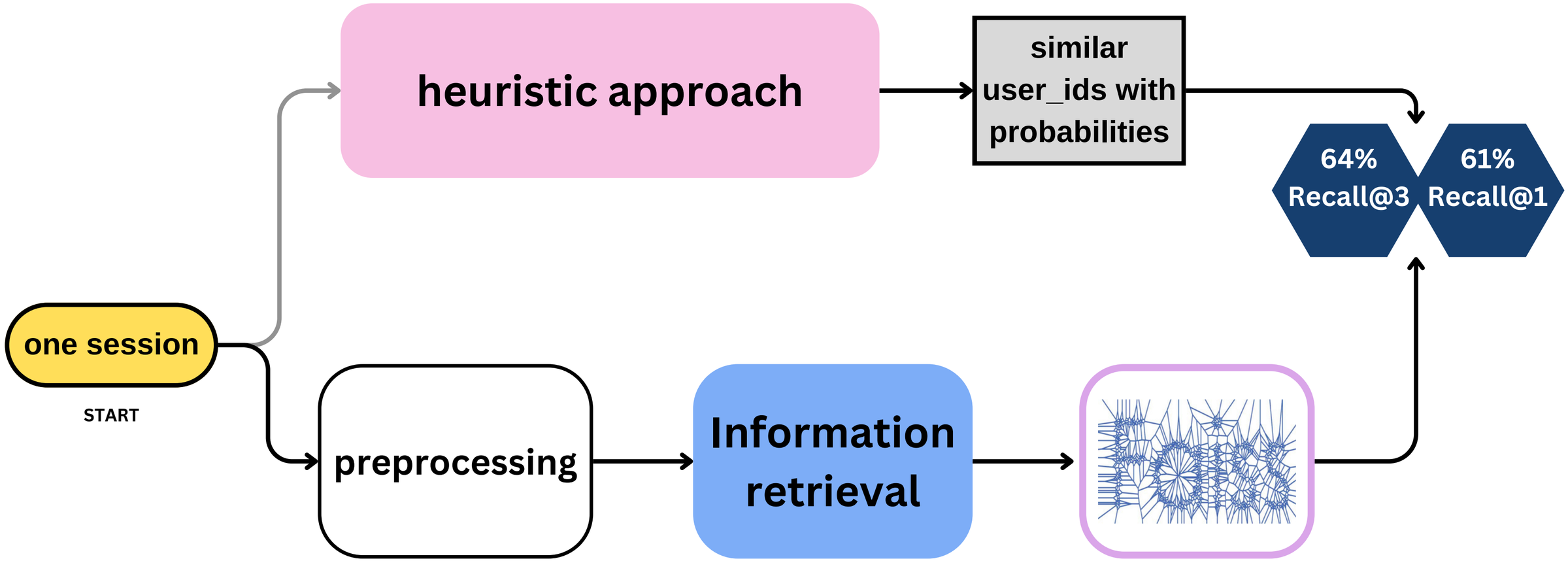
17% Recall@3

13% Recall@1

9% Recall@3

7% Recall@1

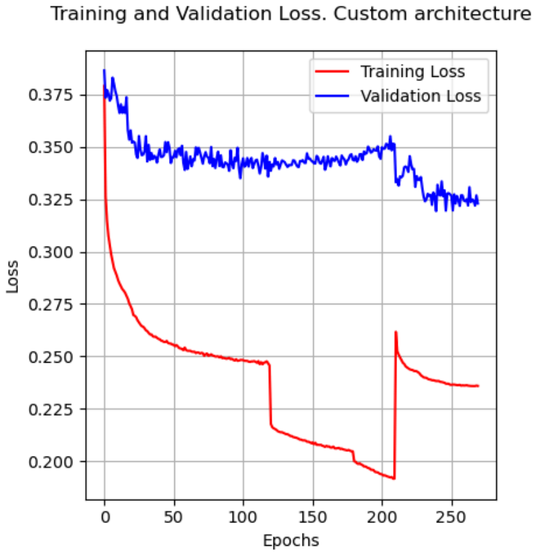
WORKFLOW

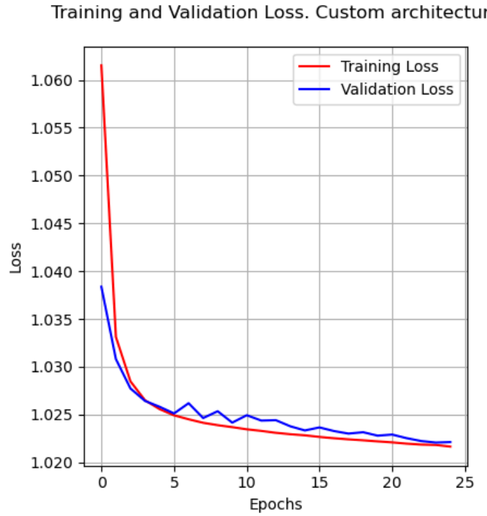

heuristic approach
similar user_ids with probabilities
64% Recall@3
61% Recall@1
one session
START
preprocessing
Information retrieval

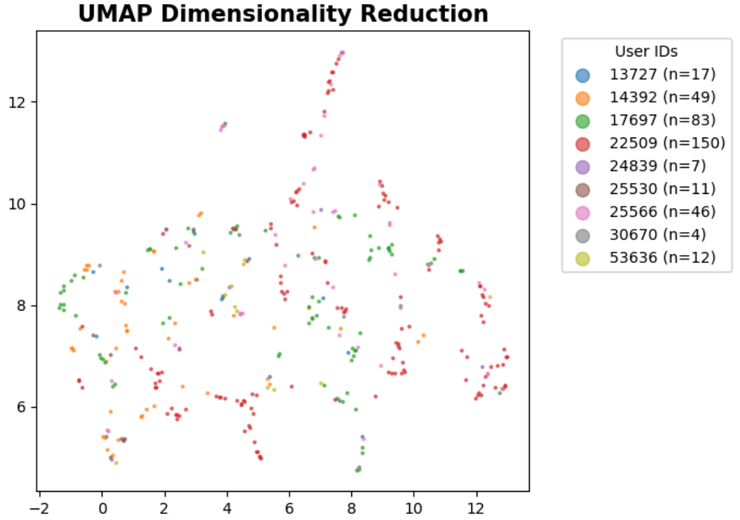
CHALLENGES

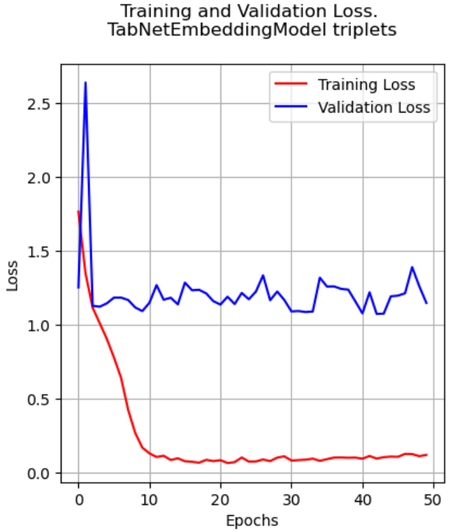
- Data preprocessing
- Handling categorical features
- Training NN models from scratch
  - different model architectures
  - high variation in parameters
  - time consuming

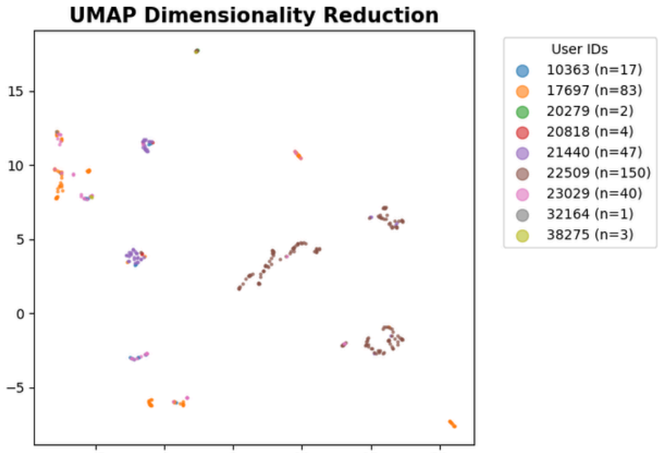
APP. 1. EXPERIMENTS



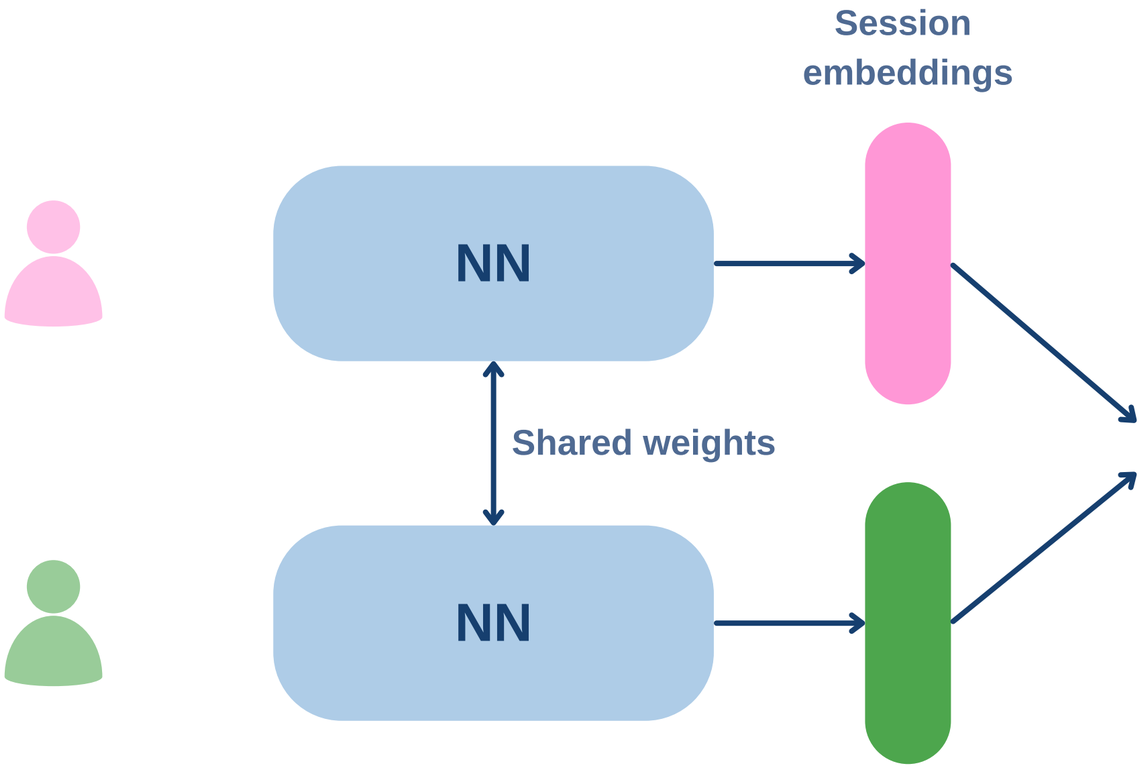








APP. 2. SIAMESE NN



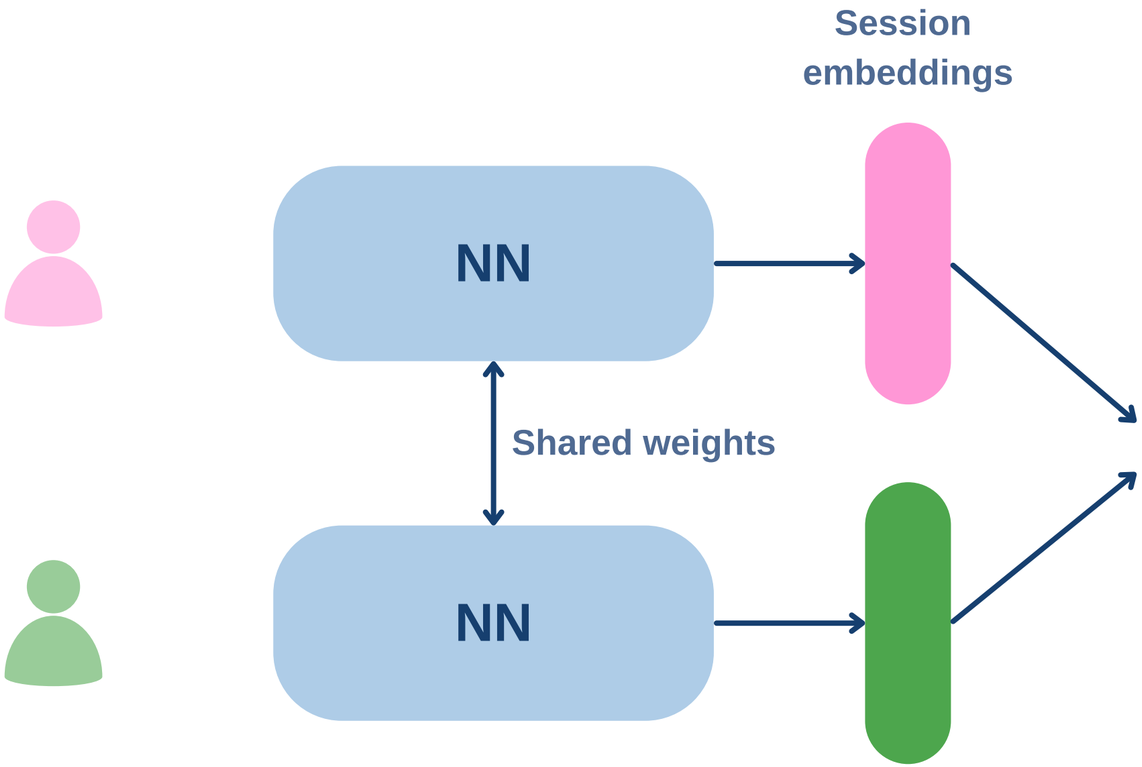
Contrastive loss

$$L = y * D^2 + (1 - y) * \max(0, m - D)^2$$

D = Euclidean distance

y = 1 for similar pairs
y = 0 for dissimilar pairs

APP. 2. SIAMESE NN



Contrastive loss

$$L = y * CD^2 + (1 - y) * \max(0, m - CD)^2$$

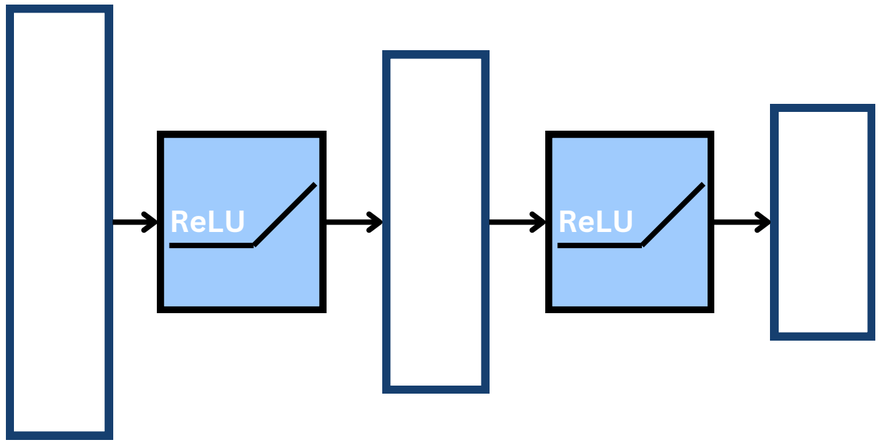
$$CD = 1 - \text{cosine\_similarity}$$

y = 1 for similar pairs
y = 0 for dissimilar pairs

APP. 3. MODEL ARCHITECTURES

Custom architecture

Total params: 8,432



TabNet architecture

Total params: 15,092