# okcupid

Ron Butbul
Rotem Mustacchi
Yuval Horesh

# PROBLEM STATEMENT

## Motivation:

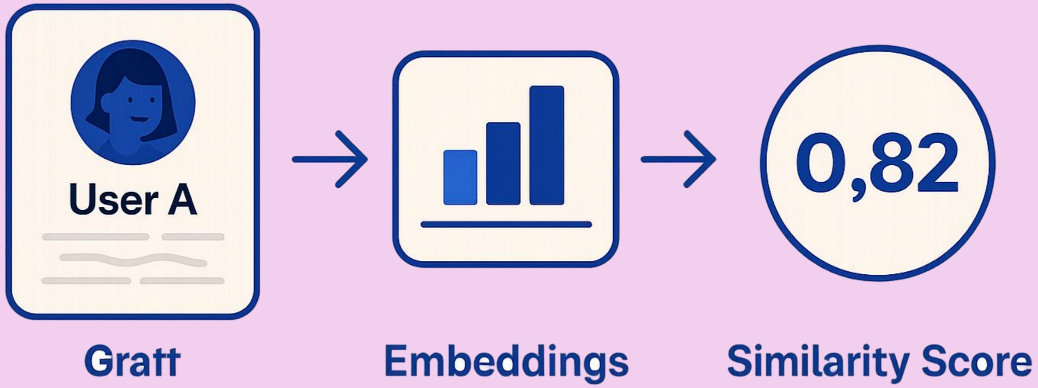
- Online dating relies heavily on manual browsing. Matching based on NLP-derived insights can improve the quality and efficiency of matches.

## Challenges:

- Free-form, noisy, and subjective text. Profiles may differ in length, tone, style, and vocabulary.

## Problem Definition:

- Input: Textual profiles (essays, self-summaries)
- Output: Similarity score or ranking between profiles
- Task: Text Similarity / Recommendation



# DATA

## Data Type:

- Unlabeled, real-world profile text. No ground-truth labels for "match“.

## Problem Definition:

- OkCupid Profiles – Kaggle. Contains multiple essay fields per profile.

## Example:

- Input: two users "Self-summary" or "What I’m doing with my life“.
- Output: similarity score between profiles.

I'm a huge fan of live music and checking out new bands I also enjpy hik-

I love sci-fi and coffee shops, I'm a big fan of deep conversations.

I like getting outside as much as possible, biking, I also love going to concerts.

# EVALUATION

## Metrics:

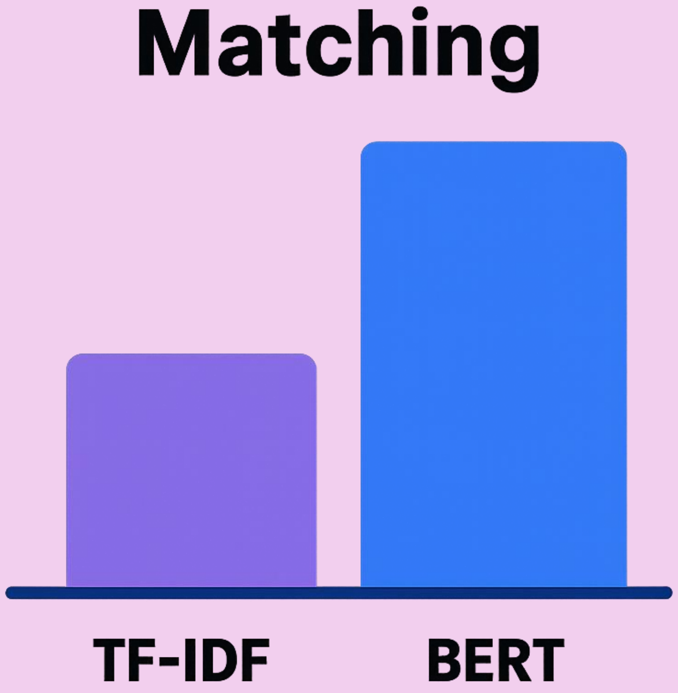
- Cosine similarity,
- Precision at K
- Jaccard similarity

## Evaluation Plan:

- Hold-out validation – train on 80%, test on 20% randomly selected profiles.

## Baseline:

- TF-IDF + Cosine Similarity.
- Compare with BERT / SBERT / LLM embeddings


Matching
TF-IDF
BERT