What?

We study the relevance of the common RecSys practice consisting in:

1. Learning **item embeddings** to summarize similarities between some recommendable items.
2. **Averaging** them to represent **users** or **other recommendable concepts** in the same space.

*Averaging embeddings of songs from a playlist (**Ex 1**) or user's listening history (**Ex 2**), to obtain a playlist or user embedding.*

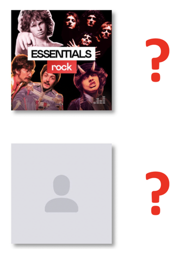
Why?

Averaging embeddings is simple and scalable. But this practice is often adopted without a clear **theoretical justification** from a RecSys standpoint:

Ex 1: Would songs similar* to the playlist be similar to songs in this playlist?

Ex 2: Would songs similar* to the user be relevant recommendations for them?

**Similar in the embedding space, according to some similarity metric, e.g., the inner product.*



How?

1. We propose a **consistency score**, measuring the faithfulness of average embeddings relative to the recommendable items they should summarize:

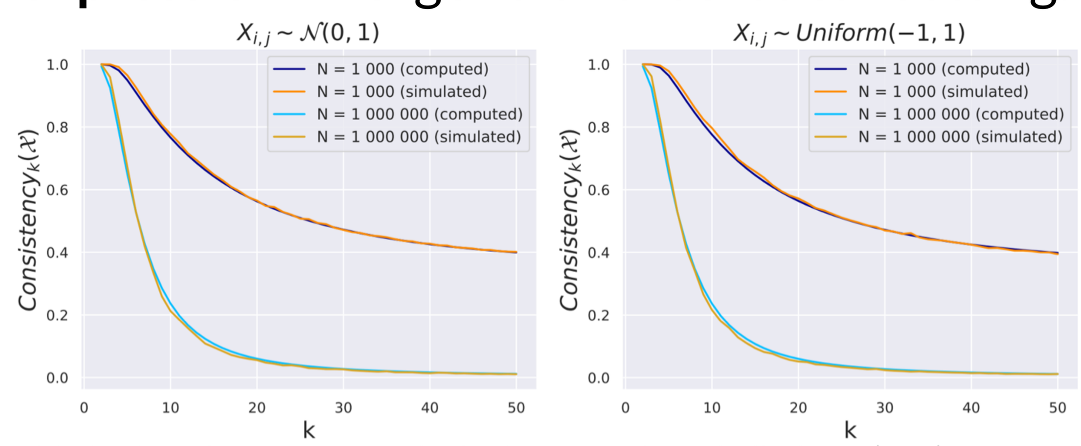
$$\text{Consistency}_k(\mathcal{X}) = \mathbb{E}_{\mathcal{U}\in\mathcal{X}_k}\Big[\text{Precision}_k(\mathcal{U})\Big], \text{ where } \text{Precision}_k(\mathcal{U}) = \frac{|\mathcal{X}_k(\mu_{\mathcal{U}})\cap\mathcal{U}|}{k} \text{ and } k \in \{1,\dots,N\}.$$

Notation: *$\mathcal{X}$: set of $N \in \mathbb{N}^*$ d-dim. item embeddings. $\mathcal{X}_k$: set of subsets of $\mathcal{X}$ of cardinality k. For any $\mathcal{U} = \{u_i\}_{1\leq i\leq k} \in \mathcal{X}_k$, its average embedding is $\mu_{\mathcal{U}} = \frac{1}{k}\sum_{u_i\in\mathcal{U}} u_i$. $\mathcal{X}_k(\mu_{\mathcal{U}})$: set of the k nearest neighbors of $\mu_{\mathcal{U}}$ in $\mathcal{X}$ according to a similarity metric s.*

Interpretation: *Higher values = on expectation, $\mu_{\mathcal{U}}$ averages comprise more items from $\mathcal{U}$ in their top-k neighborhood.*

2. We prove its **mathematical expression** in a general theoretical setting:

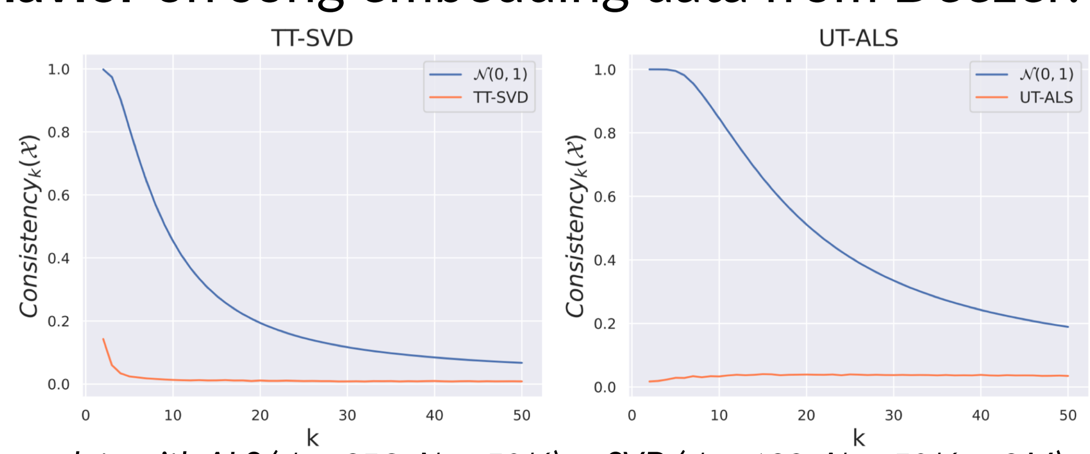
- $\text{Consistency}_k(\mathcal{X})$: $\nearrow$ with the **dimension** d.
- $\searrow$ with the **catalog size** N and **cardinality** k.
- $\searrow$ with the items' **kurtosis** ($\sim$ more outliers).
- Scores are **close to 1** for a small k.
- $u \in \mathcal{U}, v \in \overline{\mathcal{U}} \Rightarrow \mathbb{P}(s(u,\mu_{\mathcal{U}}) > s(v,\mu_{\mathcal{U}})) > 0.5$.



Theoretical setting*: $\mathcal{X}$: i.i.d. r.v., with i.i.d. elements and finite first 4 moments – s is the inner product similarity: $s(x,y) = x^\intercal y$.*

3. We analyze its **empirical behavior** on song embedding data from Deezer:

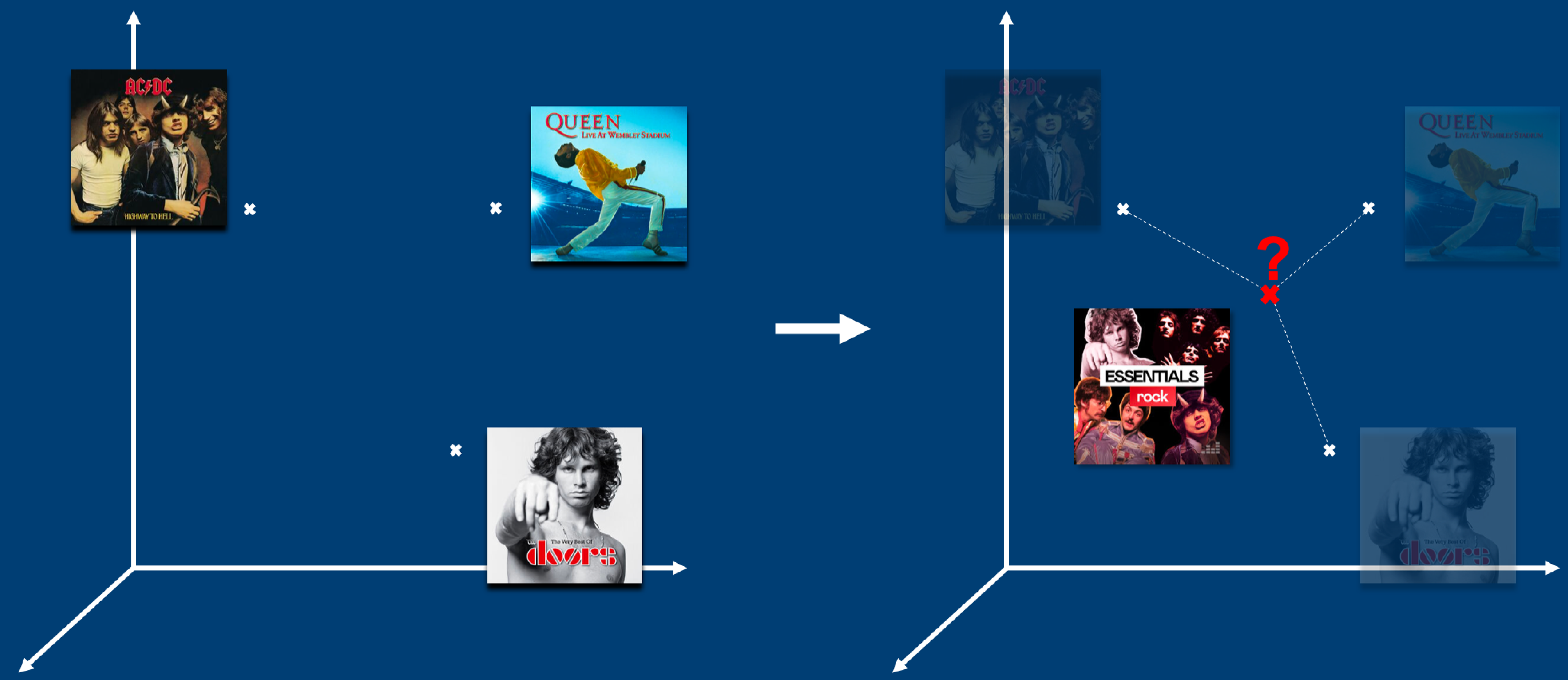
- **"Real-world" averages are less consistent**!
- Even for a small k, averages do **not** always remain similar to the items they summarize.
- **ALS:** steady with k. **SVD**: declining with k.
- Future research: **align** embeddings with our theoretical setting, e.g., via a regularization.



Data*: 3 song embedding variants, computed from usage data with ALS ($d = 256$, $N = 50K$) or SVD ($d = 128$, $N = 50K$ or $2M$).*

Be careful! In an embedding-based recommender system:

Averaging item embeddings does not always consistently **summarize** them.



Paper

Code

On the Consistency of Average Embeddings for Item Recommendation

Walid Bendada[1,2], Guillaume Salha-Galvan[1], Romain Hennequin[1], Thomas Bouabça[1], Tristan Cazenave[2]

[1]Deezer Research [2]LAMSADE, Université Paris Dauphine, PSL