# Reddit Comments and I: A Love Story

# Data

**91,558,594 - all**

That's a lot of comments

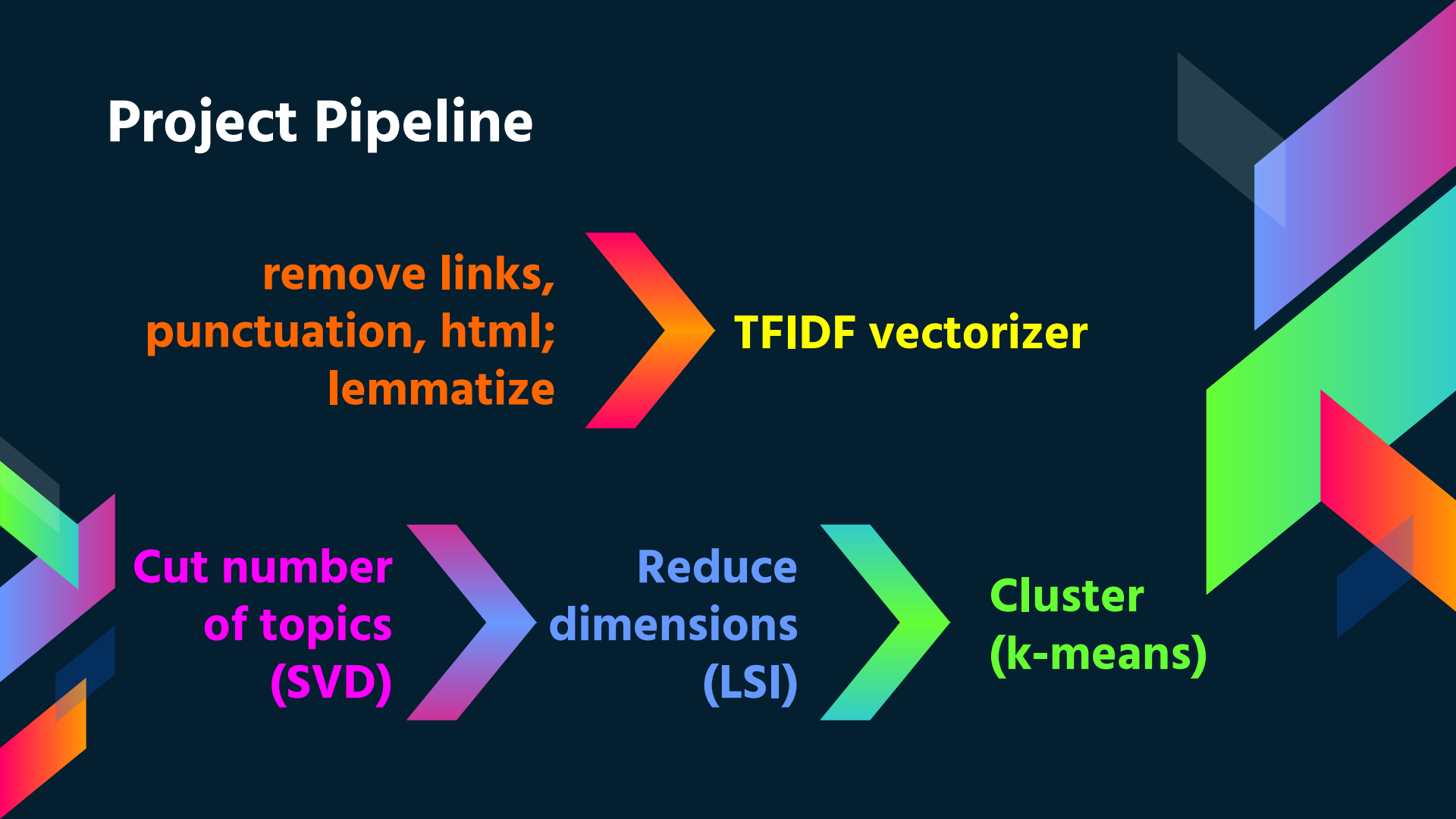
**10k, 100k, 1m**

Sampling tests

**494,610 - r/gaming**

But a subset works, as well

# Project Pipeline



# Results: Overview

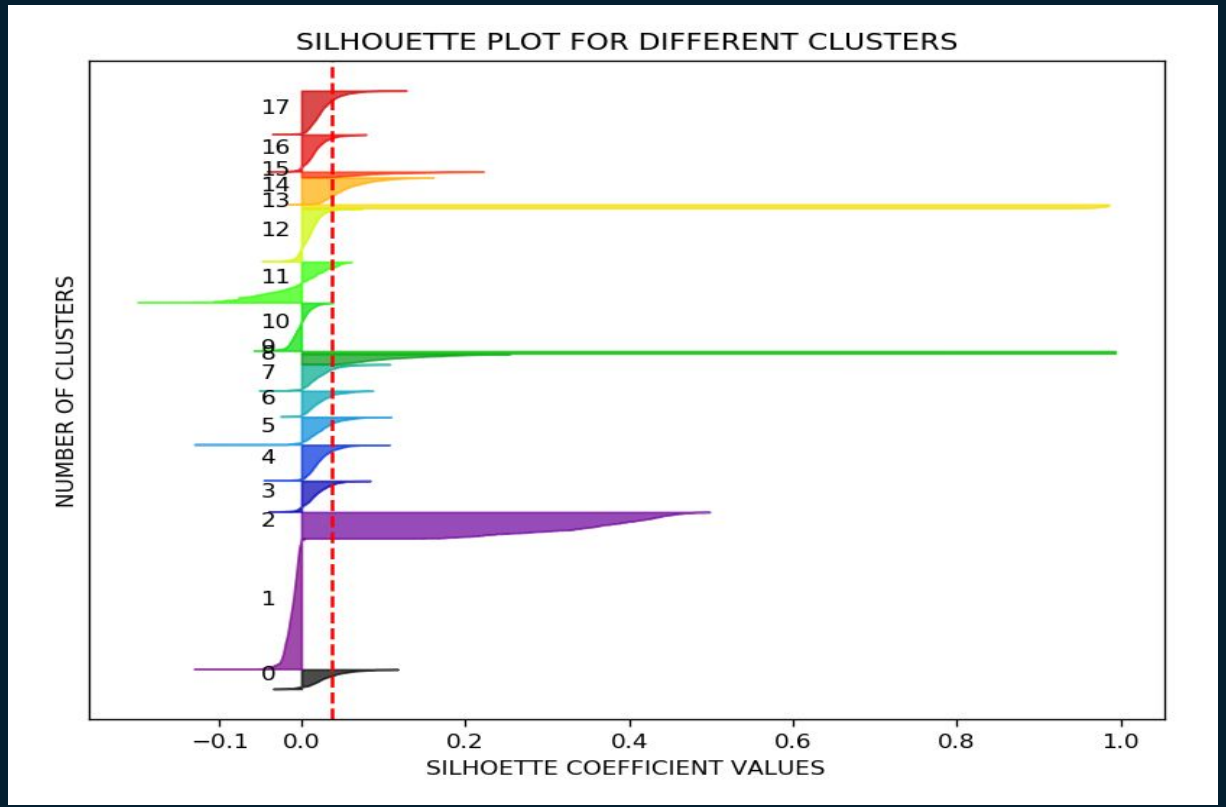
## Clustering on the entire data set:

- Only 49% of variance explained by features
- Wild silhouette plots (clusters too spread out from centroids)
- Hard to make sense out of clusters upon inspection

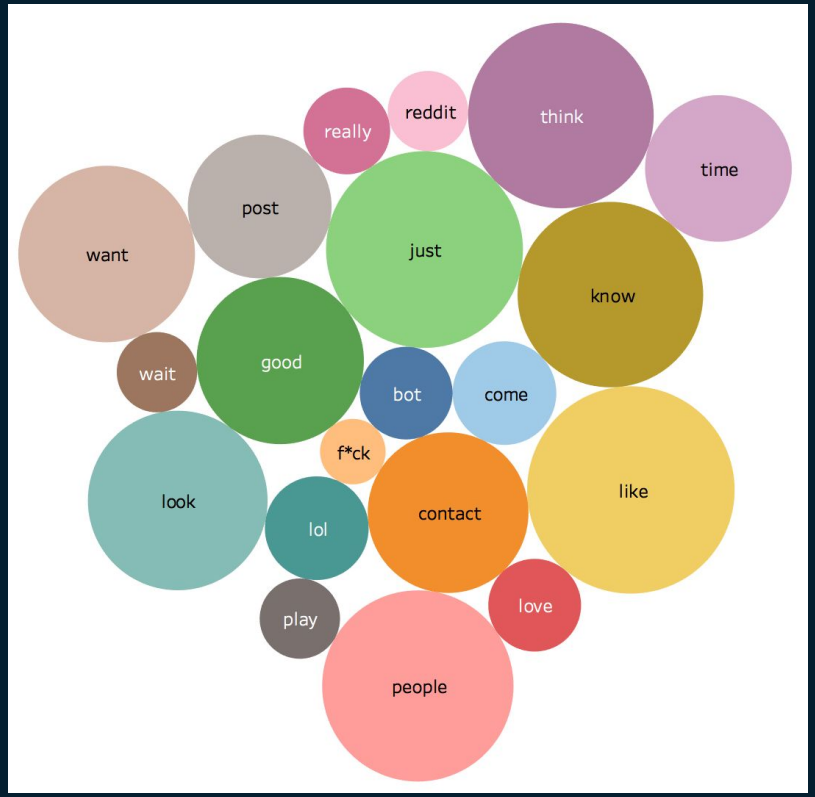
## Clustering on r/gaming:

- 5.2% of variance explained by features
- Same issues as the entire data set

# Results: Silhouette Scores (subset)



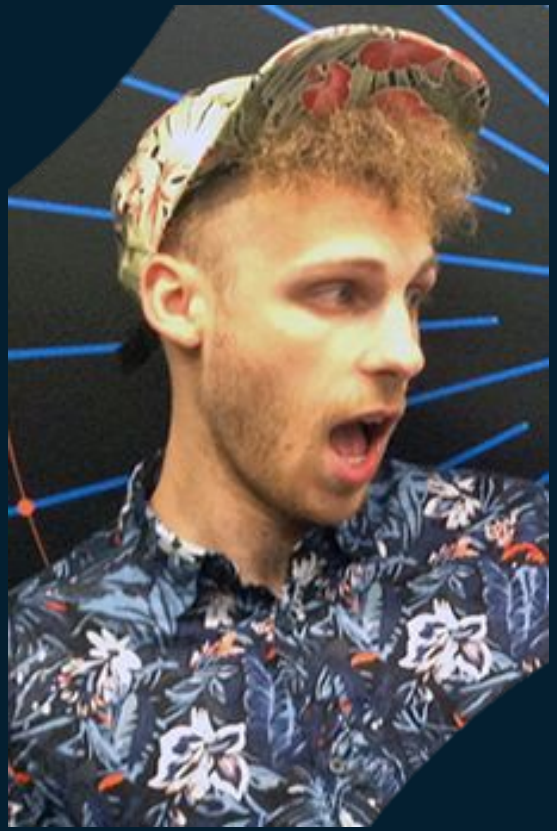
# Results: 20 Most Common Words




- Cannot be discarded
- Multiple meanings
- ***Like***: verb, noun, preposition, adverb

# Conclusion

- Words which carry meaning also add noise
- Data set too diverse for clustering
- Data may not work well for analysis
- May need a different approach altogether



# Big thanks to:

- Emy 😅
- Michael and Mike 🤔
- Perry 😍
- Varun 😌
- Ivan 😝
- Debbie 🤩