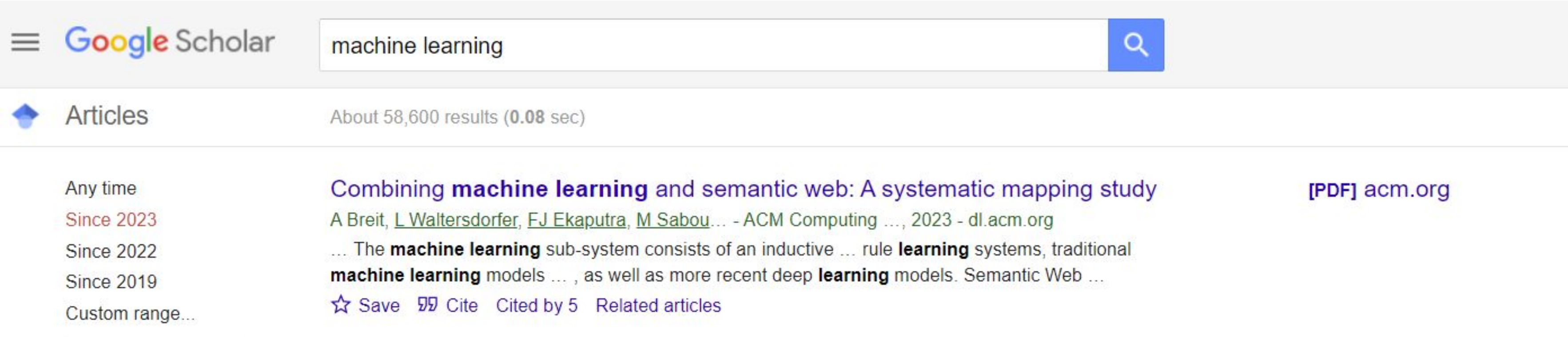
# Research Paper Recommender

Topic modeling using Bert

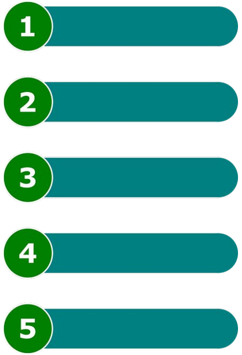
# Problem



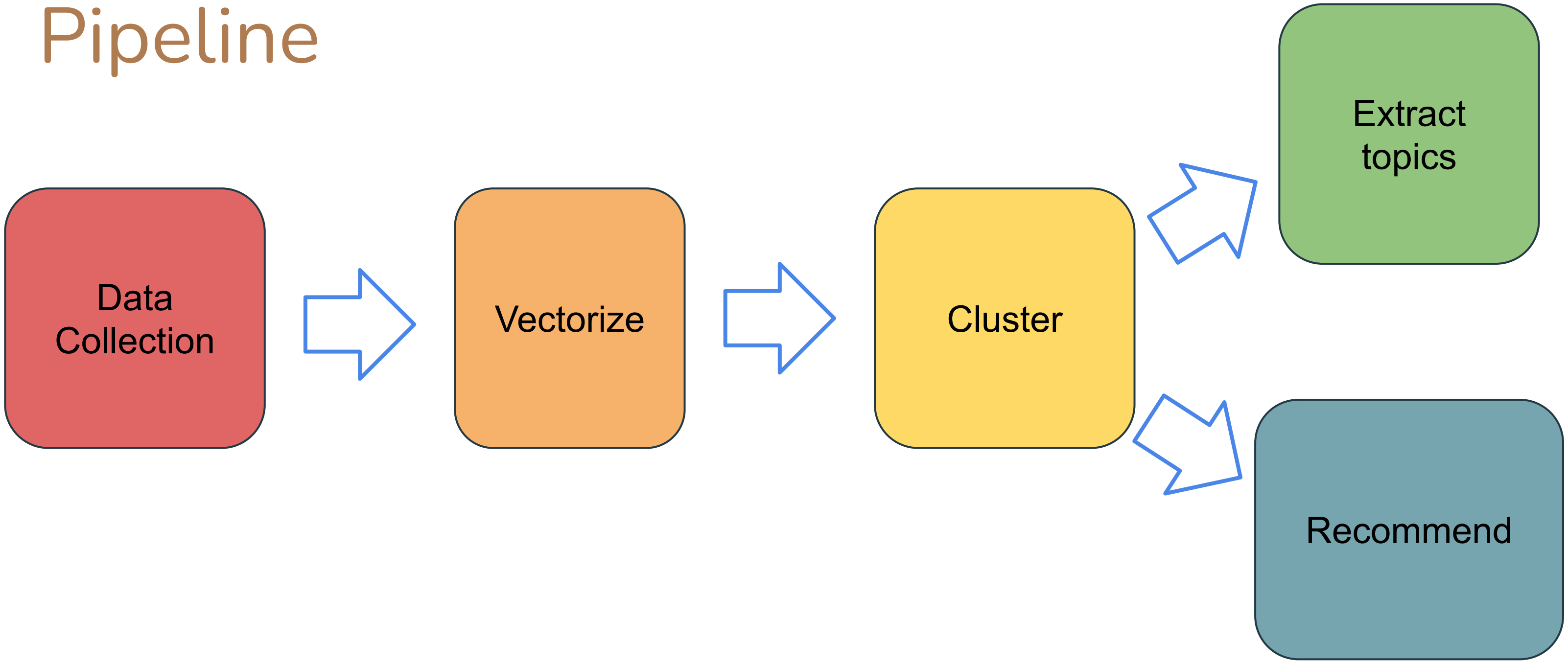

There are too many papers published to keep up with, even in a specific field.

# Project Goal

- To help researchers find relevant research
- Provide a high level overview of conferences
- Starting point for downstream AI augmented tasks



# Pipeline


Data Collection
Vectorize
Cluster
Extract topics
Recommend

# Data Collection

- Scrape NeurIPS | 2018, 2019, 2022
  - Title
  - Author
  - Abstract
  - Year

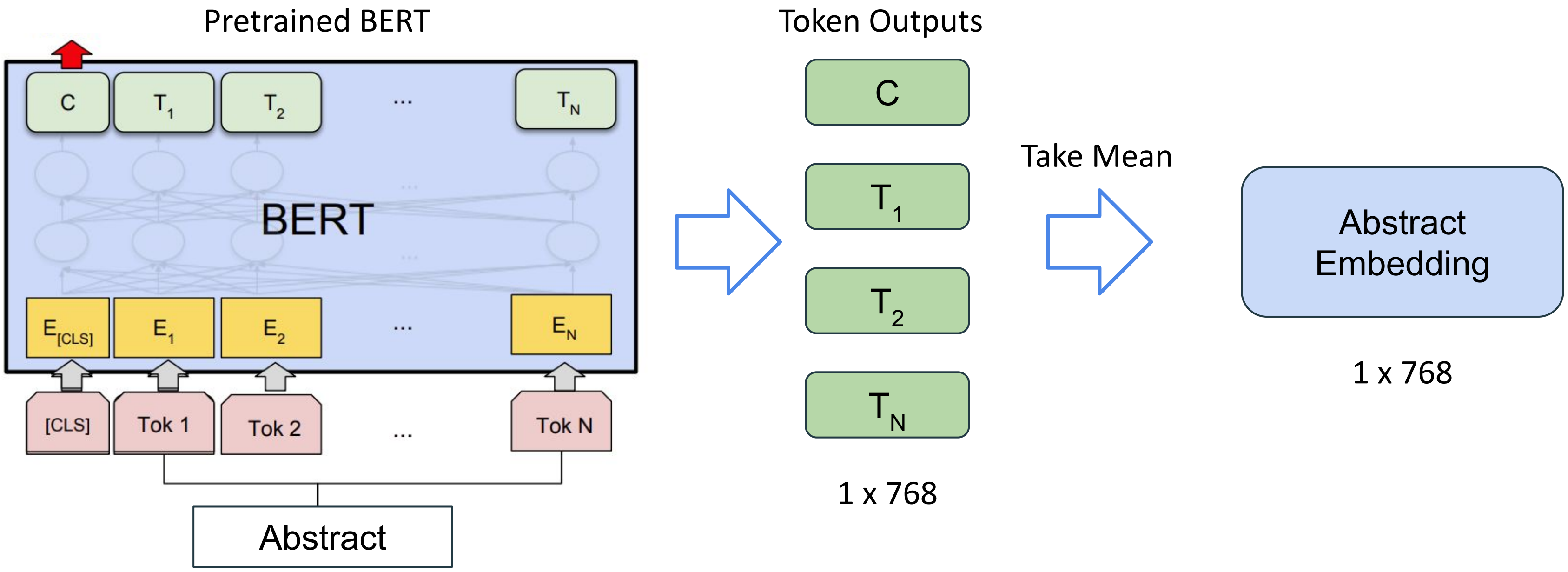
❌ https... and other website links

❌ Latex math symbols

| | Unnamed: 0 | title | authors | abstract | year |
|---|---|---|---|---|---|
| 0 | 0 | Synthesized Policies for Transfer and Adaptati... | Hexiang Hu, Liyu Chen, Boqing Gong, Fei Sha | The ability to transfer in reinforcement learn... | 2018 |
| 1 | 1 | Self-Supervised Generation of Spatial Audio fo... | Pedro Morgado, Nuno Nvasconcelos, Timothy Lang... | We introduce an approach to convert mono audio... | 2018 |
| 2 | 2 | On GANs and GMMs | Eitan Richardson, Yair Weiss | A longstanding problem in machine learning is ... | 2018 |
| 3 | 3 | Batch-Instance Normalization for Adaptively St... | Hyeonseob Nam, Hyo-Eun Kim | Real-world image recognition is often challeng... | 2018 |

4507 rows × 5 columns

# Vectorize each Abstract



# Clustering

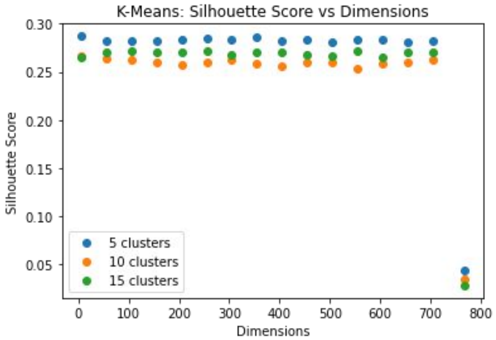
- K-means
  - minimize the euclidean distance between data points and their cluster centers.
- Mixture of Gaussians
  - models each cluster as a Gaussian distribution using maximum likelihood estimation
- DBSCAN
  - identifies clusters as dense regions of points separated by areas of lower density

# Dimensional Reduction

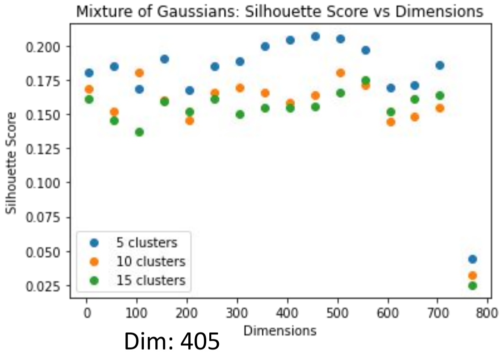
- Each abstract embedding is a vector in 768 dimensional space!
  - All three clustering techniques suffer from the **curse of dimensionality**
- U-map
  - Nonlinear
  - Preserves and highlights global structure of the data
  - Conducive to clustering

# Experiment

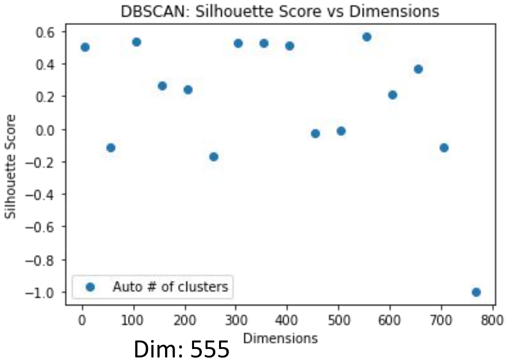
- 16 tested dimensions: 5, 55, ...,, 655, 705, 768 (original dimension)
- 3 specified cluster number: 5, 10, 15 (DBScan doesn't need cluster)




Dim: 5
Cluster: 5 0.287

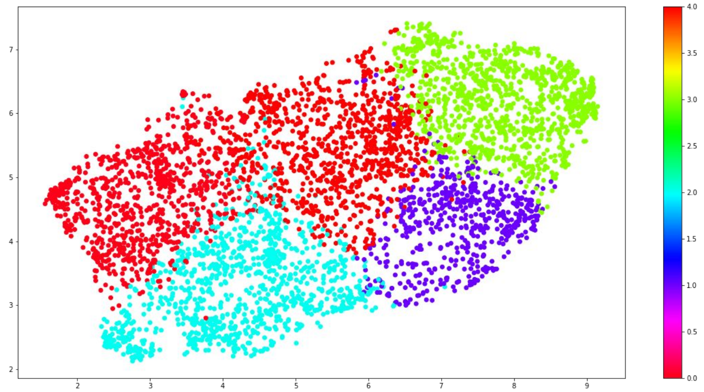



Dim: 405
Cluster: 5 0.207



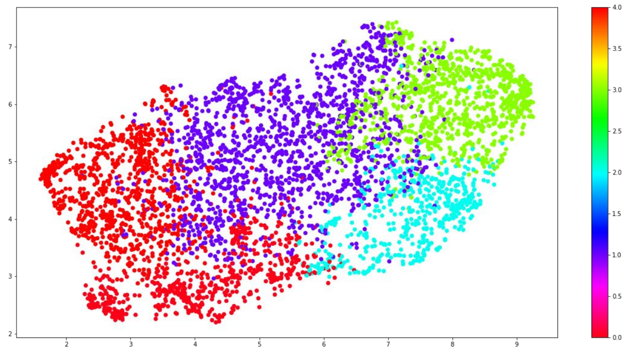

Dim: 555
Cluster: 22 0.565

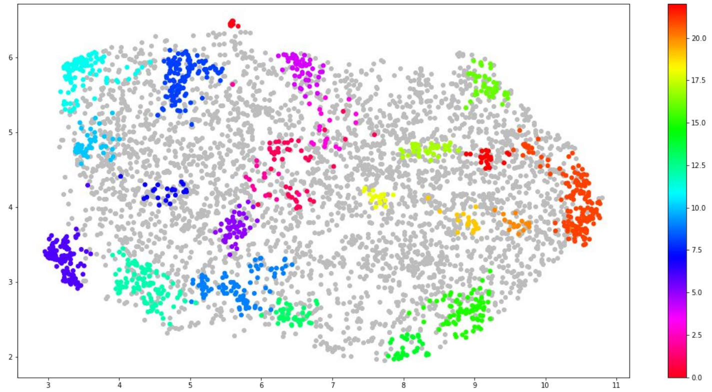
# Clusters Results



K-Means

GMM





DBSCAN

# Topic Extraction

- Combine all abstracts from each cluster
- Calculate a cluster specific TF-IDF (Term Frequency-Inverse Document Frequency)

$$c - TF - IDF_i = \underbrace{\boxed{\frac{t_i}{w_i}}}_{\text{TF}} \times \underbrace{\boxed{\log \frac{m}{\sum_j^n t_j}}}_{\text{IDF}}$$

Source: https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6

# Results of Topic Extraction (K-Means)

Cluster 1

```
[('image', 0.010455527349121162),
 ('3d', 0.009338199797847122),
 ('images', 0.007872895766817491),
 ('object', 0.0078358751067602),
 ('semantic', 0.006623888327007003),
 ('video', 0.0063063096932340895),
 ('art', 0.006305684280094541),
 ('attention', 0.006046105110822836),
 ('text', 0.005817588456637443),
 ('segmentation', 0.005568266949247685)]
```

Computer Vision

Cluster 2

```
[('regret', 0.01676209786682996),
 ('setting', 0.009170052708645252),
 ('optimal', 0.009098231216397733),
 ('policy', 0.008930715619463148),
 ('bandit', 0.008713460566124896),
 ('online', 0.008014194945893777),
 ('algorithm', 0.007686849370323906),
 ('reward', 0.007635184692618447),
 ('algorithms', 0.007570931240640078),
 ('games', 0.006998430221836676)]
```

Reinforcement Learning?

# Results of Topic Extraction (K-Means)

Cluster 3

```
[('agent', 0.006604908974015055),
 ('policy', 0.006145742945052291),
 ('tasks', 0.006114656026612 5),
 ('reinforcement', 0.006026089386762524),
 ('rl', 0.005944230960590833 4),
 ('task', 0.005813555170400954),
 ('human', 0.005574303698401548 5),
 ('adversarial', 0.005350507283604001),
 ('agents', 0.005322553960396518),
 ('reward', 0.005033949524843321)]
```

Reinforcement Learning

Cluster 4

```
[('convex', 0.009203464569474 54),
 ('convergence', 0.008829418795822337),
 ('gradient', 0.008741732051570611),
 ('stochastic', 0.008159357598137748),
 ('bounds', 0.007857653597210 75),
 ('descent', 0.007210433150682484),
 ('optimization', 0.007191138534732545),
 ('bound', 0.007012929832571141),
 ('linear', 0.006873387174704888),
 ('non', 0.006808549815941425)]
```

Optimization/ learning

# Results of Topic Extraction (K-Means)

Cluster 5

```
[('graph', 0.00841625274867288),
 ('inference', 0.006089938083782041),
 ('networks', 0.005505296119215530 5),
 ('deep', 0.005215817628757047),
 ('graphs', 0.005207559518619444),
 ('variational', 0.005163280208286707),
 ('neural', 0.004954506349573331),
 ('latent', 0.004850544412235576 4),
 ('bayesian', 0.004781959467815738),
 ('distribution', 0.004696655844755842)]
```

Graph Neural Network

Overall, It seems like the topics are easily interpretable

# Recommendation

1. Use BERT to generate an embedding of the user input
2. Find which cluster the user input belongs to
3. Rank papers using Cosine similarity

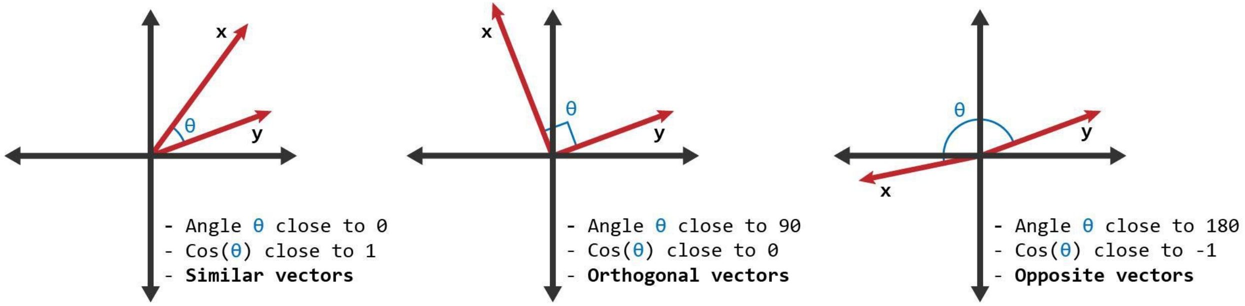



Image Credit: https://datascience-enthusiast.com/DL/Operations_on_word_vectors.html

# Show Recommendation Model Results

I want to learn more about “Image Generation”

**K-Means**

1. title: QC-StyleGAN - Quality Controllable **Image Generation** and Manipulation
2. title: Multi-View Silhouette and Depth Decomposition for High Resolution **3D Object** Representation
3. title: FNeVR: Neural Volume **Rendering** for Face Animation

# Some Challenges

I want to learn more about “Image Generation”

**DBSCAN**

```
The predicted cluster is: [-1]

[('algorithms', 0.0038961218755742847),
 ('approach', 0.003816919525161166),
 ('optimization', 0.0037525583066266937),
 ('paper', 0.0037502642373708687),
 ('performance', 0.0037464070714998829),
 ('deep', 0.003743298631327031),
 ('new', 0.0037160996486909886),
 ('state', 0.003710392352912112),
 ('work', 0.0037038899917861195),
 ('proposed', 0.0037011288298504823)]
```

# Show Recommendation Model Results

I want to learn more about "Text to Image"

1. Context-aware Synthesis and Placement of Object Instances
2. PatchComplete: Learning Multi-Resolution Patch Priors for 3D Shape Completion on Unseen Categories
3. Language Conditioned Spatial Relation Reasoning for 3D Object Grounding
4. Learning Dense Object Descriptors from Multiple Views for Low-shot Category Generalization

# Some Challenges

I want to learn more about "Reinforcement Learning"

**K-Means**

```
The predicted cluster is: [4]

[('graph', 0.00841625274867288),
 ('inference', 0.006089938083782041),
 ('networks', 0.005505296119215305),
 ('deep', 0.005215817628757047),
 ('graphs', 0.005207559518619444),
 ('variational', 0.005163280208286707),
```

# Some Challenges

I want to learn more about "Value-function-based methods have long played an important role in reinforcement learning. …

K-Means

1. Exponentially Weighted Imitation Learning for Batched Historical Data
2. Robust exploration in **linear quadratic reinforcement learning**
3. Near-Optimal **Multi-Agent Learning** for Safe Coverage Control

Input abstract from: Arthur Delarue et. al, Reinforcement Learning with Combinatorial Actions: An Application to Vehicle Routing

# Key Takeaways

- We were able to generate semantically meaningful clusters.
- BERT is good at vectorizing text while capturing its meaning
  - Length based bias
  - Generally reasonable recommendations
- Machine learning research is very diverse with a couple hot pockets
  - Many outliers according to DBSCAN