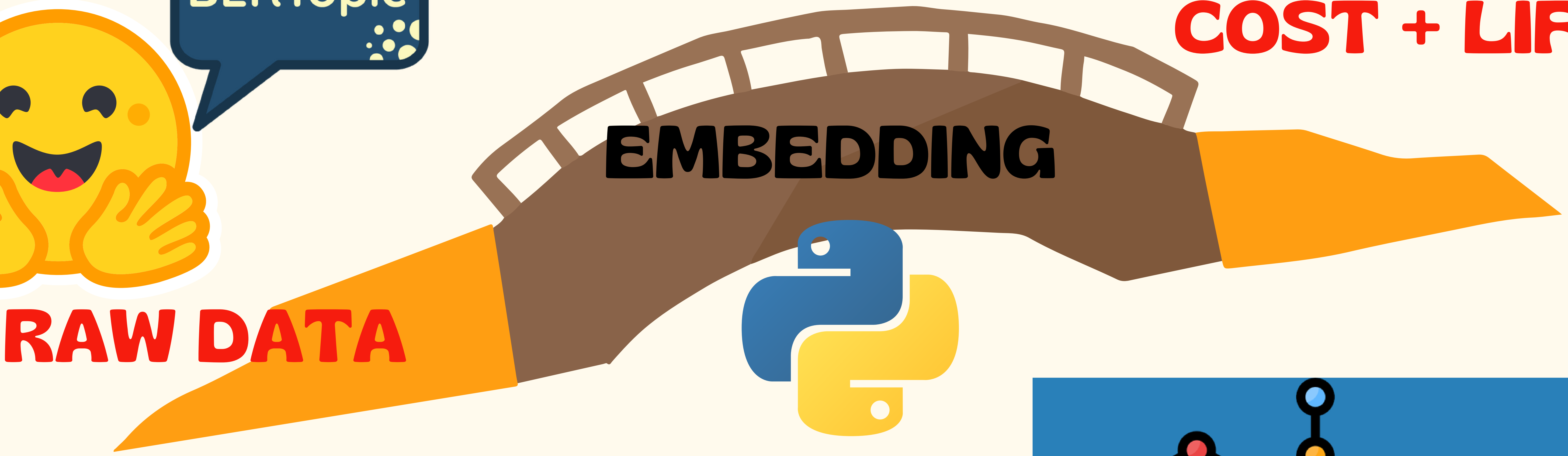
EXPLORING THE POWER OF EMBEDDINGS WITH CLUSTERING

DECISION INSIGHTS SAVING COST + LIFE

GROOTENDORST, MAARTEN}

BERTopic



BUILDING BRIDGES BETWEEN DATA & INSIGHTS IN THE AI ERA

SBERT.net

HTTPS://GITHUB.COM/INSIGHTBUILDER

CHALLENGE SOLVED: WHERE & HOW EMBEDDINGS ARE USED

- REAL LIFE APPLICATION:
  - MARKET SEGMENTATION
  - IMAGE SEGMENTATION (CANCER CELL DETECTION)
  - ANAMOLY DETECTION (CREDIT CARD / NETWORK ANALYSIS)
  - LAND / NETWORK USAGE ANALYSIS
  - SEARCH ENGINES
  - CROSS ENCODERS
  - IMAGE SEARCH
- ANY APPLICATION THAT WILL REQUIRE CLUSTERING IN VOICE, VIDEO ALSO CAN WORK

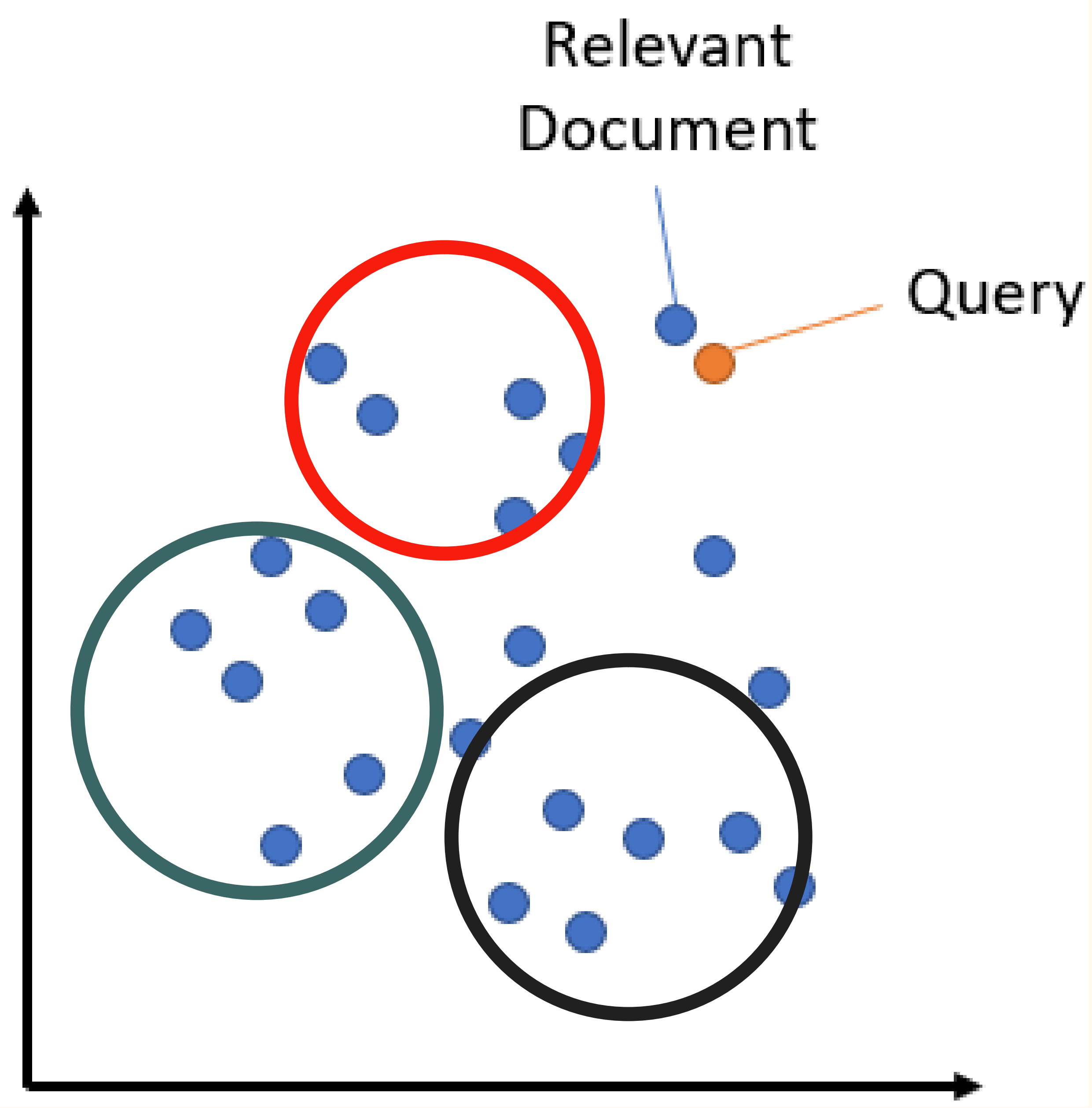
- CLUSTERING ALGORITHMS / PROCESSES:
  - PARTITION CLUSTERING
    - K-MEANS
  - DENSITY BASED CLUSTERING
    - MEAN-SHIFT ALGORITHM
  - DISTRIBUTION MODEL-BASED CLUSTERING
    - DENSITY BASED SPATIAL CLUSTERING & NOISE
  - HIERARCHICAL CLUSTERING
    - AGGLOMERATIVE CLUSTERING
    - AFFINITY PROPOGATION
  - FUZZY CLUSTERING

OPEN SOURCE LIBARIES : CHALLENGES THEY SOLVE

- SENTENCE-TRANSFORMERS : PROVIDE EMBEDDING (HTTPS://WWW.SBERT.NET)
- BERTOPIC: TOPIC MODELING + VISUALISATION (HTTPS://MAARTENGR.GITHUB.IO/BERTOPIC)
- PICKLE : SAVE THE EMBEDDING DATA AS FILE
- SAFETENSOR : SAFER ALTERNATIVE OF SAVING EMBEDDING DATA
- KEYBERT: EXTRACTING KEYWORDS FROM CORPUS
- SKLEARN: PROVIDE ML ALGORITHMS FOR CLUSTERING
- HDBSCAN: LIBRARY FOR DOING DBSCAN CLUSTERING + LOT MORE (HTTPS://HDBSCAN.READTHEDOCS.IO/)
- TRANSFORMERS : LOAD NEURAL NETWORK MODELS, TRAIN & PREDICT OUTPUT
- PYTORCH: CREATE NEURAL NETWORK MODEL AND TRAIN + PREDICT OUTPUT
- HUGGINGFACE_HUB : SAVE AND LOAD NEURAL NETWORK MODELS IN THE HUB
- RAPIDS : MOVE THE ML OPERATIONS TO GPU (RAPIDS.AI)

CLUSTERING: DOES NATURE CREATE CLUSTERS

- NATURE JUST CREATES, MATH ALGORITHMS PLACE THE CIRCLES OVER THE CREATIONS TO MAKE LIFE OF THE OBSERVER EASIER
- WHAT TO DO WITH THE OUTLIERS? WHY DO THE EXIST
- MODEL THAT CAN CHOOSE WHICH BAG THE DATA POINT WILL GO IS EASY TO CREATE.
- WHEN THE NUMBER OF POINTS INCREASES THEN THE QUESTION OF WHETHER TO INCREASE THE BAGS ARISES
- CAN THERE BE CLUSTERS WITHIN CLUSTERS? HIERARCHY.
- WHAT IF I DON'T KNOW ANYTHING ABOUT NUMBER OF CLUSTERS AVAILABLE



AND WHAT ABOUT THE TOPICS OF THESE CLUSTERS?

TOPIC MODELS: TYPES & METHODS

- **APPROXIMATE TOPIC DISTRIBUTION WITH SLIDING WINDOW ON DOCS**
- **ONLINE TOPIC MODELING IS USED WHEN DATA IS FLOWING INCREMENTALLY**
- **SEMI-SUPERVISED CAN HELP IF YOU HAVE SOME CATEGORIES AVAILABLE.**
- **SUPERVISED MODELING INVOLVES REGRESSION TO TRAIN**
- **MANUAL MODE SKIPS DIM REDUCTION & CLUSTERING. HEADS TO TOPIC**
- **GUIDING THE TOPICS WITH SIMILARITY SEARCH**
- **USING C-TF-IDF TO CREATE HIERARCHICAL CLUSTERING**
- **LOOKING AT THE TOPIC CHANGING WITH DYNAMIC TOPIC MODELING**

Method	Code
Topic Distribution Approximation	.approximate_distribution(docs)
Online Topic Modeling	.partial_fit(doc)
Semi-supervised Topic Modeling	.fit(docs, y=y)
Supervised Topic Modeling	.fit(docs, y=y)
Manual Topic Modeling	.fit(docs, y=y)
Multimodal Topic Modeling	.fit(docs, images=images)
Topic Modeling per Class	.topics_per_class(docs, classes)
Dynamic Topic Modeling	.topics_over_time(docs, timestamps)
Hierarchical Topic Modeling	.hierarchical_topics(docs)
Guided Topic Modeling	BERTopic(seed_topic_list=seed_topic_list)

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER