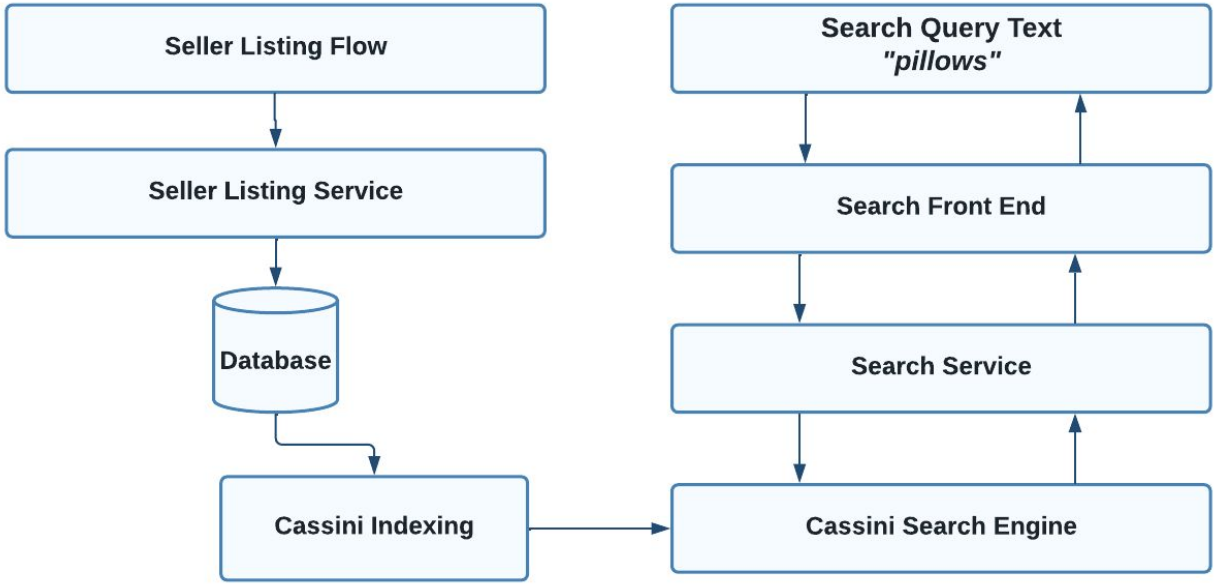
Scaling embedding models to serve a billion queries

Senthilkumar Gopal
@sengopal

ebay

Journey of a Query @ eBay



Search @ eBay

How can we discover items without describing them?

This is a problem across many domains where search is a core functionality.

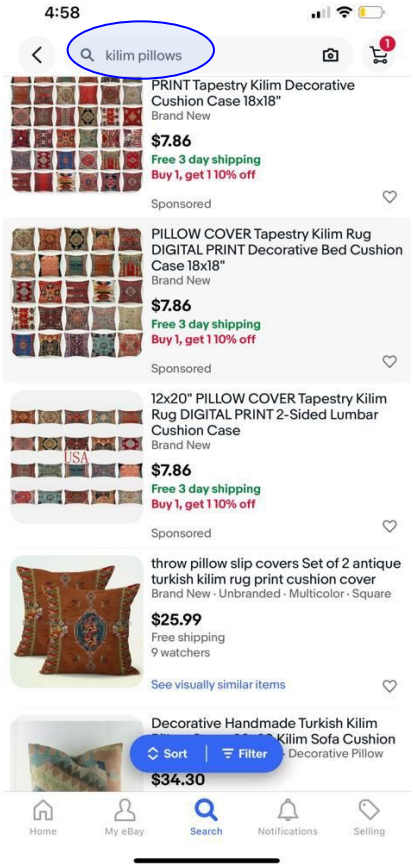
Question to ponder

can we provide users with the ability to “discover” through visual cues instead?

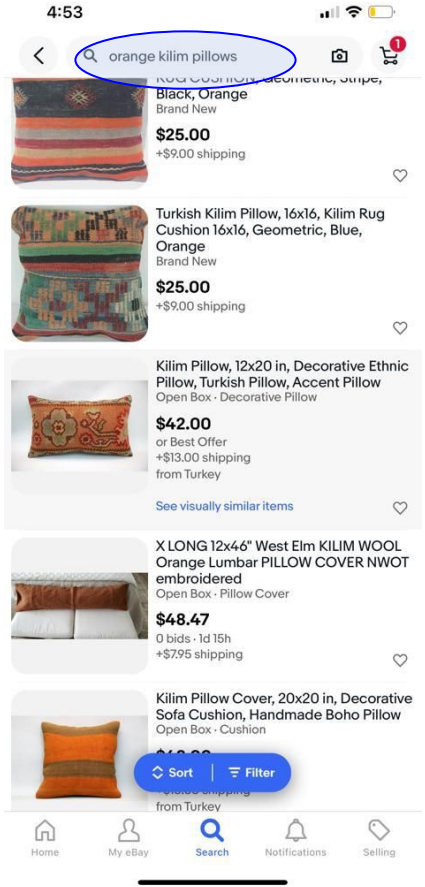
https://unsplash.com/photos/2oUiUu5QAys

Current Search Experience

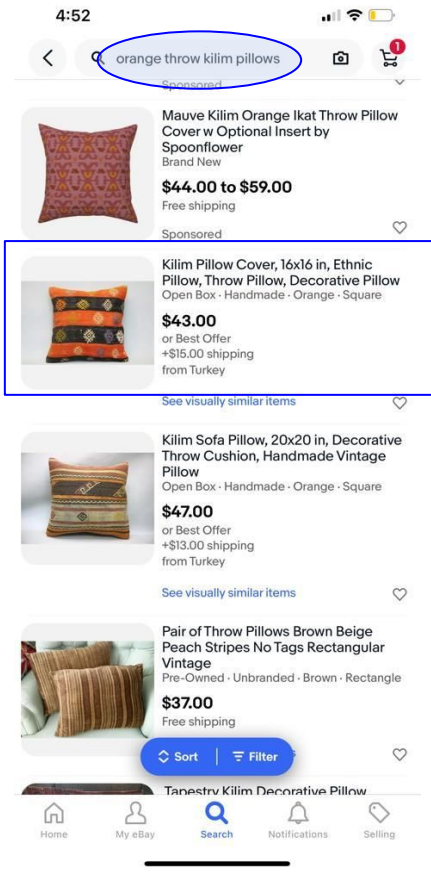
Nice Kilim Pillow for my couch!



Is this a ***kilim pillow***?



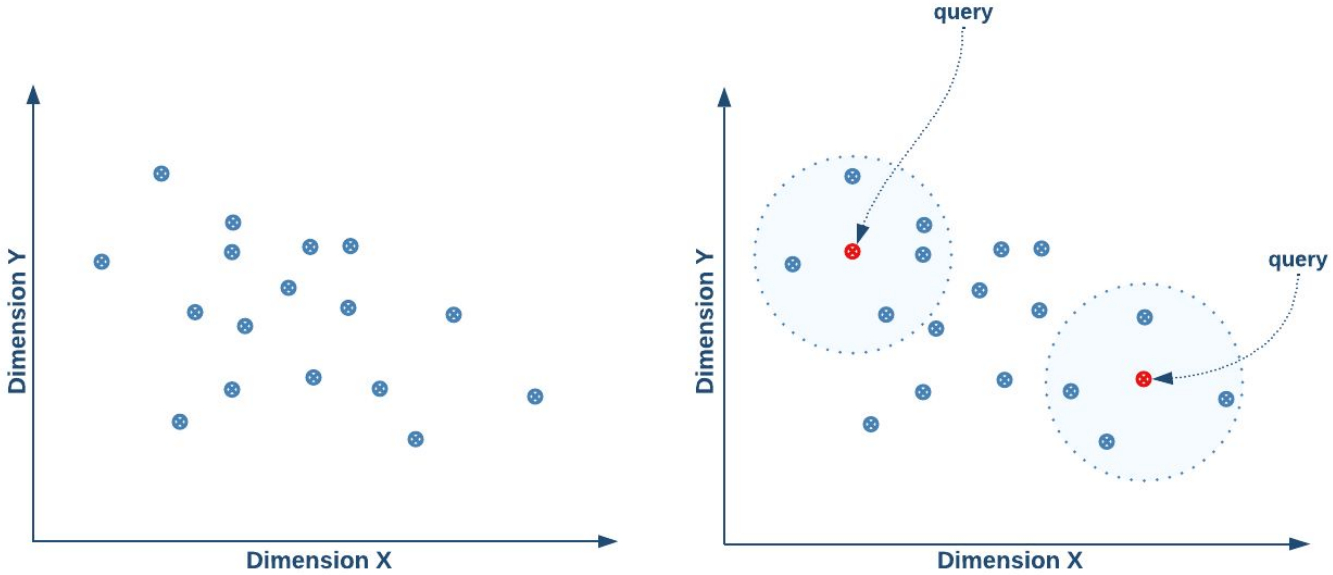
Or a ***orange kilim pillows***?



Perhaps a ***orange throw kilim pillow***?

k nearest neighbours search - A thought experiment

Let’s represent an item **TITLE** as a 2-dimensional vector



So what is an embedding then?

Represents Semantic Similarity

$$q_{couch} = [2.5, 9.1, 6.4, ...]$$
$$q_{sofa} = [2.3, 9.4, -5.5, ...]$$

Similarity (sofa, couch)

$$\frac{q_{sofa} \cdot q_{couch}}{||q_{sofa}|| . ||q_{couch}||} = cos(\phi)$$

A real word example [R_{768}]

```
[[  4.3323   2.5935   3.2519 ...  60.3621 -62.5823 -26.8413]
 [ 16.1435 -46.3839 -13.0966 ...  44.3534  -8.0482  12.7218]
 [ 51.5475  15.9534  14.3011 ...  21.5839 -38.7423   9.219 ]
 ...
 [ 34.8775  60.488   39.4437 ...   2.802   55.0218 -57.1433]
 [ 16.3728 -13.69    17.4932 ...  41.0666  46.8029  44.1613]
```

https://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html

So what is an embedding then?

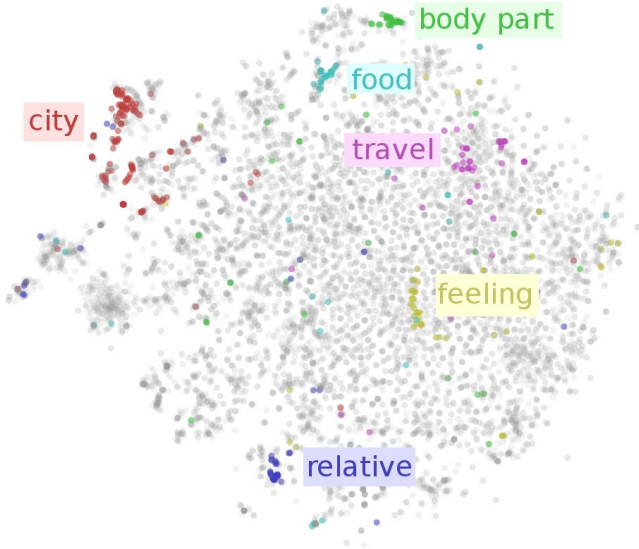
> You shall know a word by the company it keeps
>
> - *(Firth, J. R. 1957:11)*

Large Language Models - GPT 3 [175 B]

- 45 TB text data - Wikipedia and books

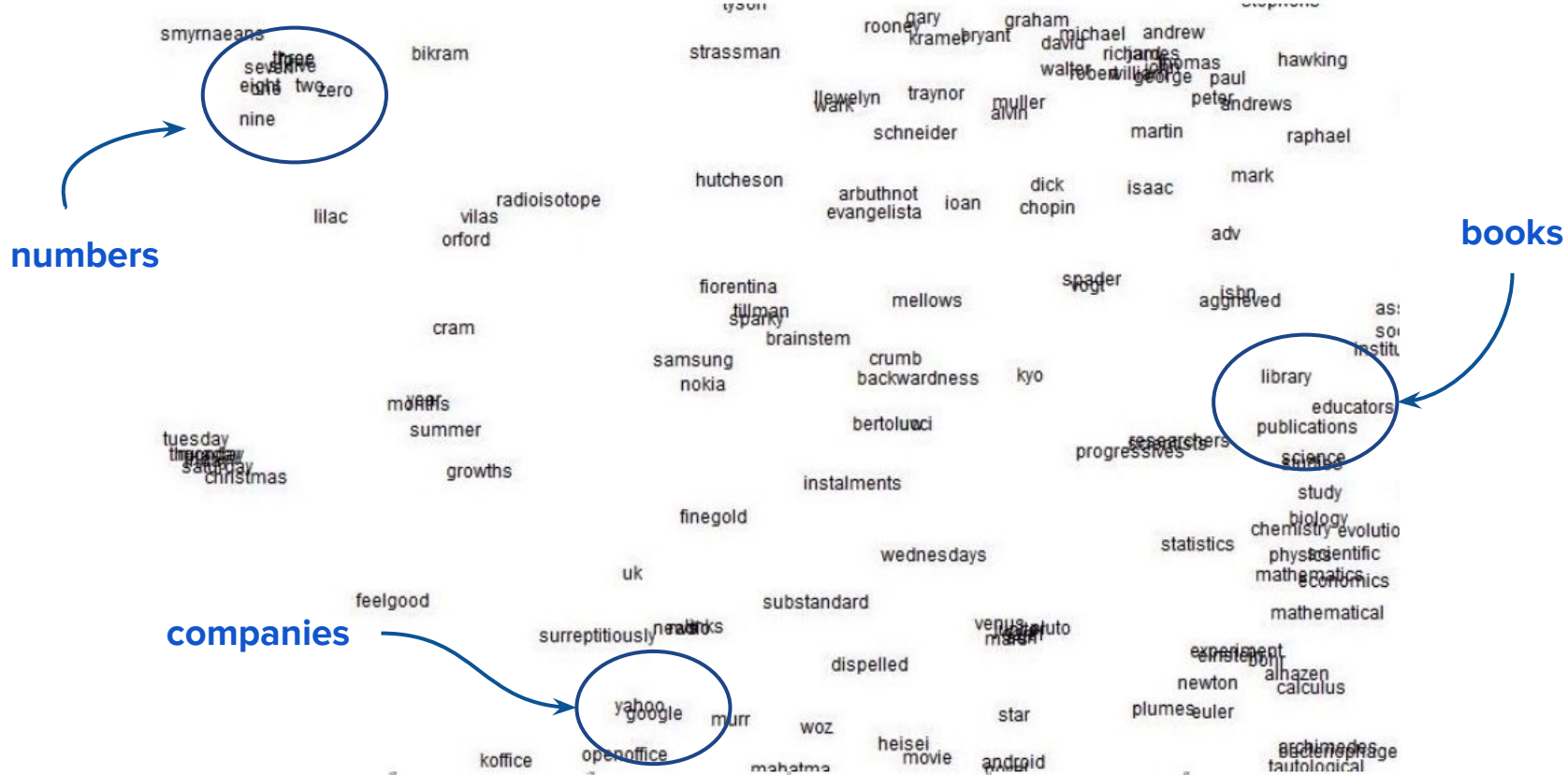
Neural network learns word associations from a large corpus.

- Detects synonymous words.
- Suggests words for a partial sentence.



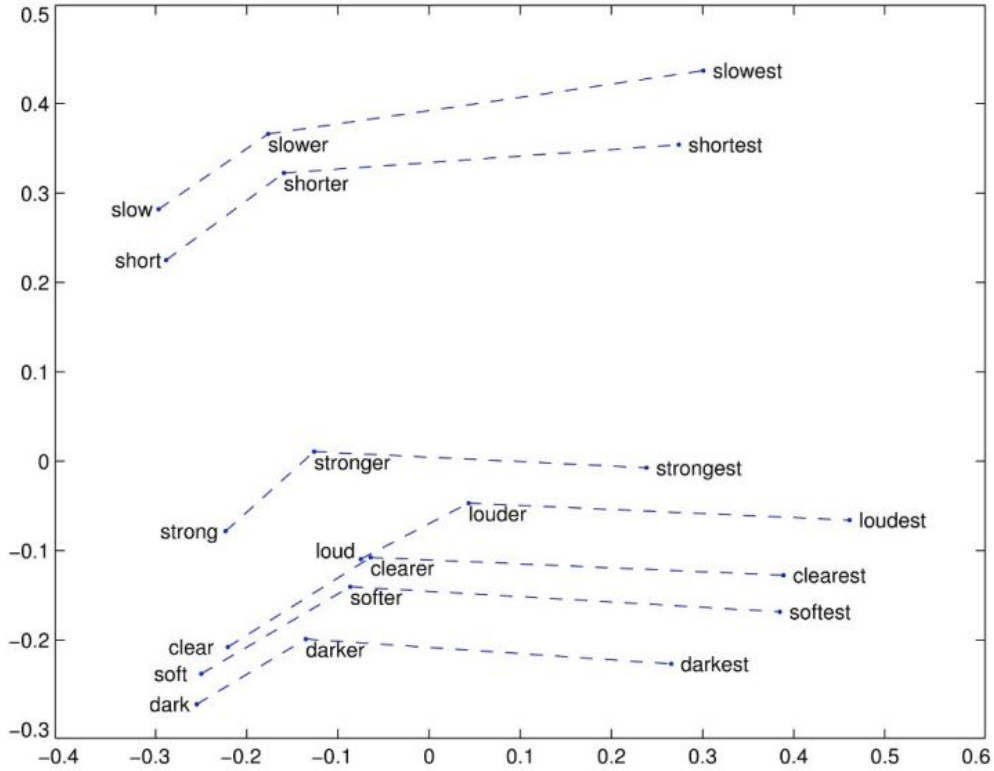
https://en.wikipedia.org/wiki/Word2vec
https://ruder.io/word-embeddings-1/

So what is an embedding then?



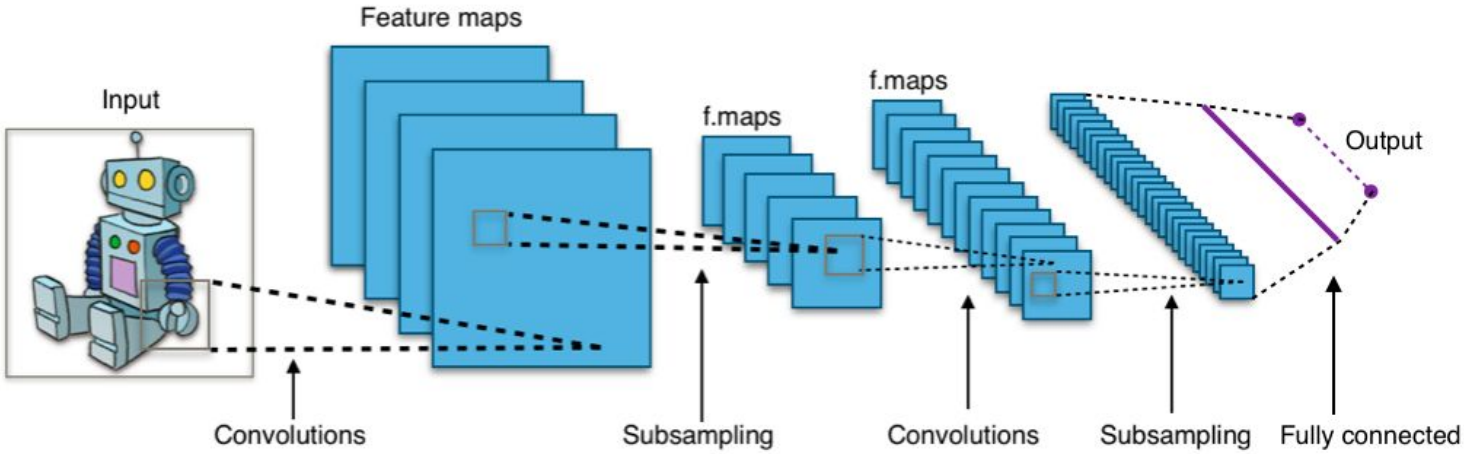
http://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture02-wordvecs2.pdf

So what is an embedding then?



http://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture02-wordvecs2.pdf

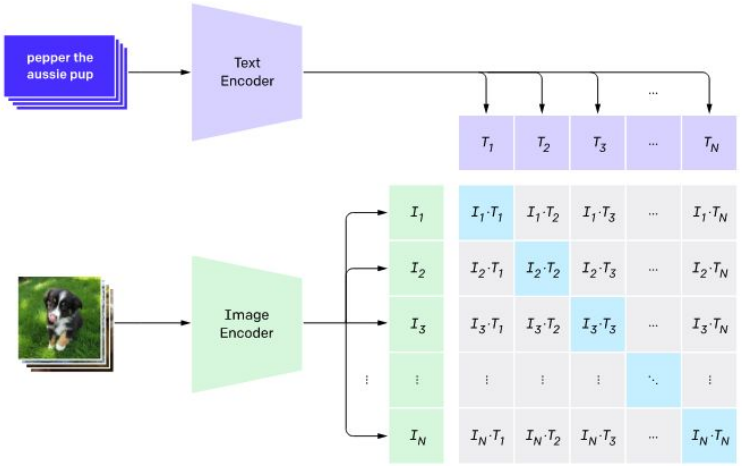
What about an image?

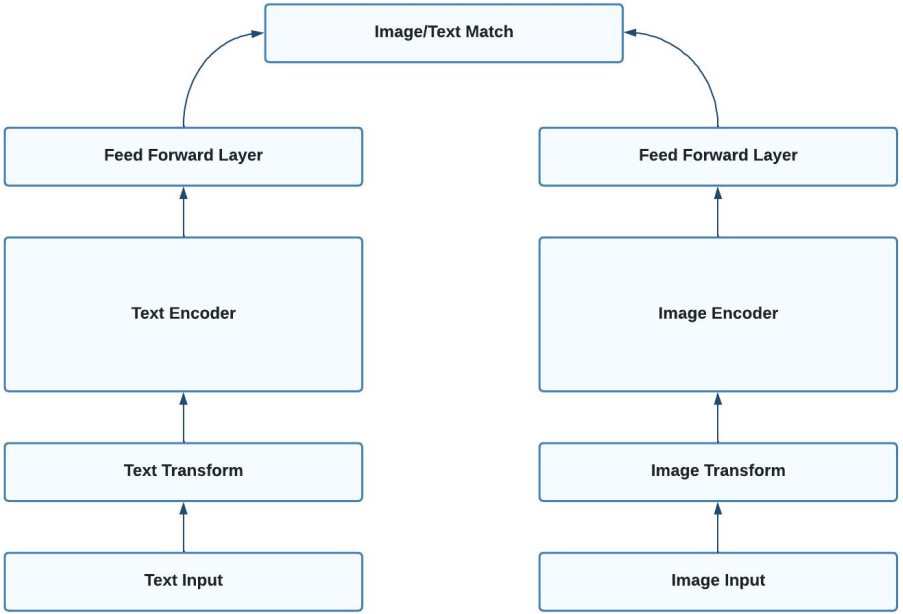


https://en.wikipedia.org/wiki/Convolutional_neural_network

Model Architecture

Multiple Modalities - Inspired by CLIP *



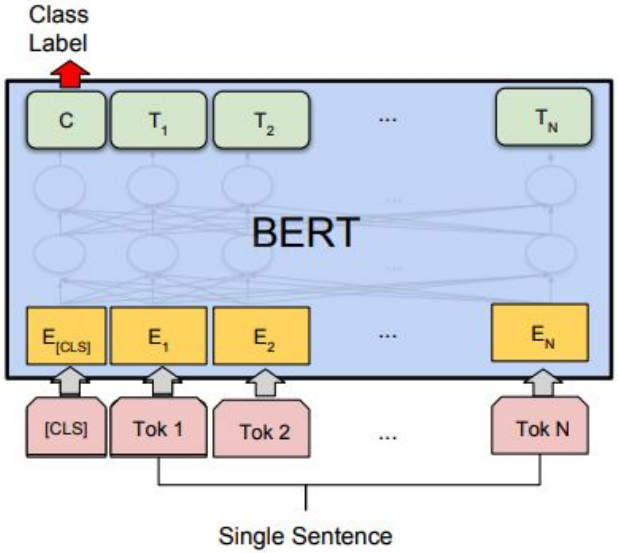


https://openai.com/blog/clip/

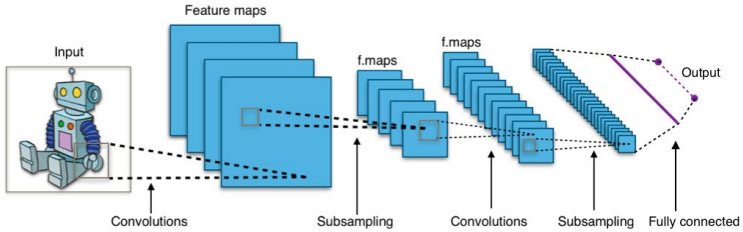
How do we “learn” an embedding?

R_{768}

That's how an embedding looks like!!!




```
[[  4.3323   2.5935   3.2519 ...  60.3621 -62.5823 -26.8413]
 [ 16.1435 -46.3839 -13.0966 ...  44.3534  -8.0482  12.7218]
 [ 51.5475  15.9534  14.3011 ...  21.5839 -38.7423   9.219 ]
 ...
 [ 34.8775  60.488   39.4437 ...   2.802   55.0218 -57.1433]
 [ 16.3728 -13.69    17.4932 ...  41.0666  46.8029  44.1613]]
```



Text Encoder

Image Encoder

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
https://en.wikipedia.org/wiki/Convolutional_neural_network

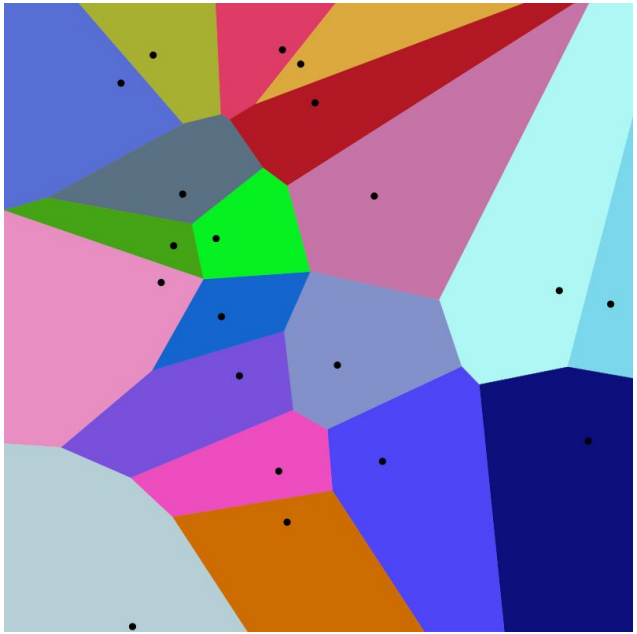
Why do we need ANN?

All problems start with **SCALE**

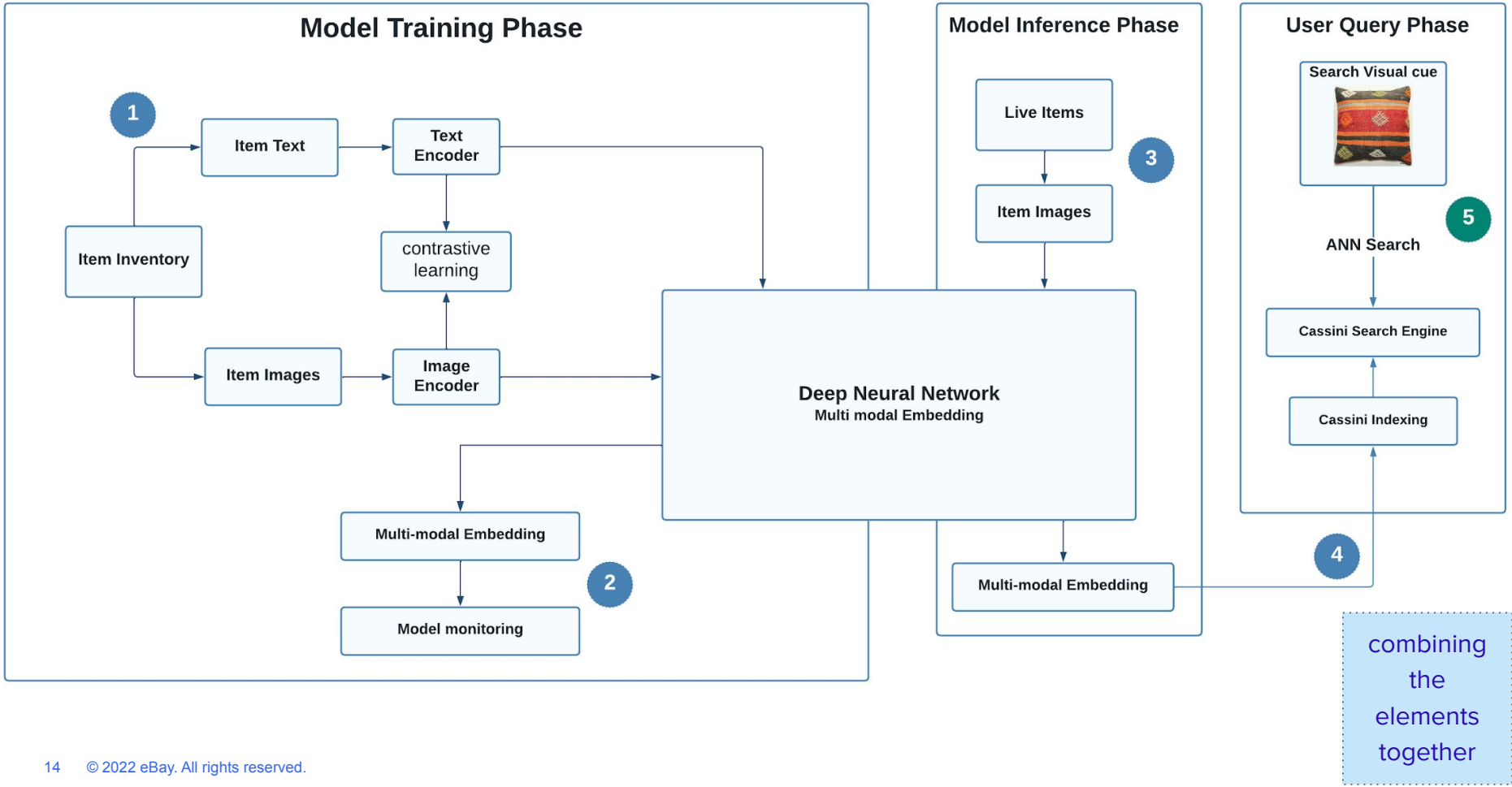
Exhaustive search
curse of dimensionality

ANN
Approximate Nearest Neighbours

rpforest BallTree(nmslib) SW-graph(nmslib)
vamana-pq(diskann) vamana(diskann) faiss-ivf
hnswlib pynndescent faiss-ivfpqfs flann
hnsw(vespa) hnsw(faiss) n2 milvus scann
hnsw(nmslib) annoy mrpt puffinn vald(NGT-panng)
NGT-panng bruteforce-blas elastiknn-l2lsh kgraph
NGT-qg opensearchknn NGT-onng kd sptag
ckdtree

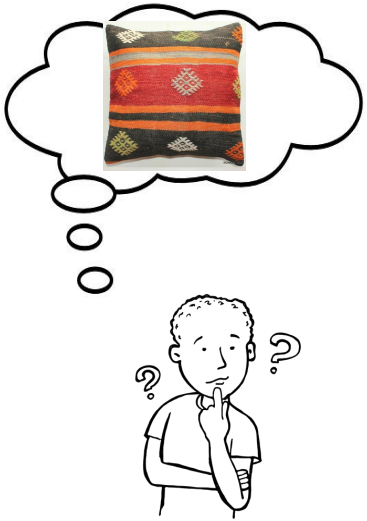


https://en.wikipedia.org/wiki/Proximity_analysis#/media/File:Euclidean_Voronoi_diagram.svg
http://ann-benchmarks.com/index.html#algorithms

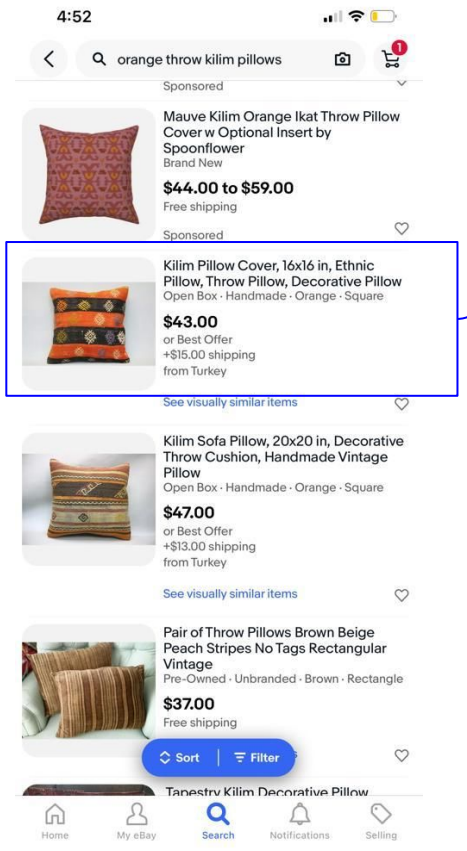


combining the elements together

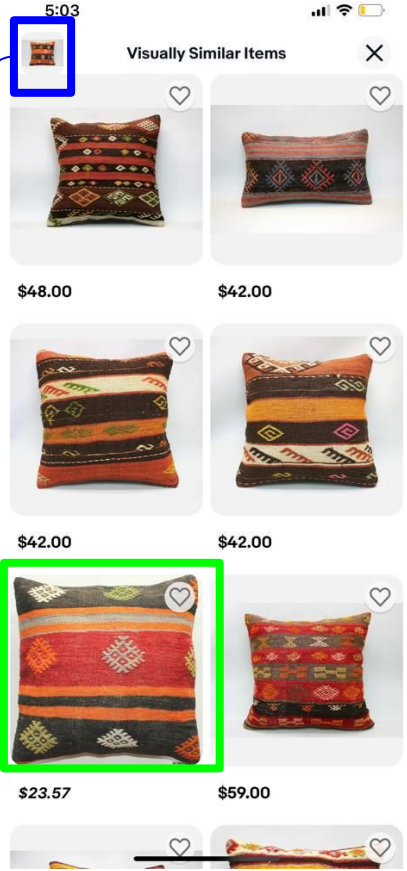
How does this function?



Display all inventory matching my visual appeal



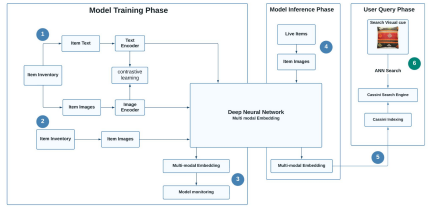
Is this a ***orange throw kilim pillow?***

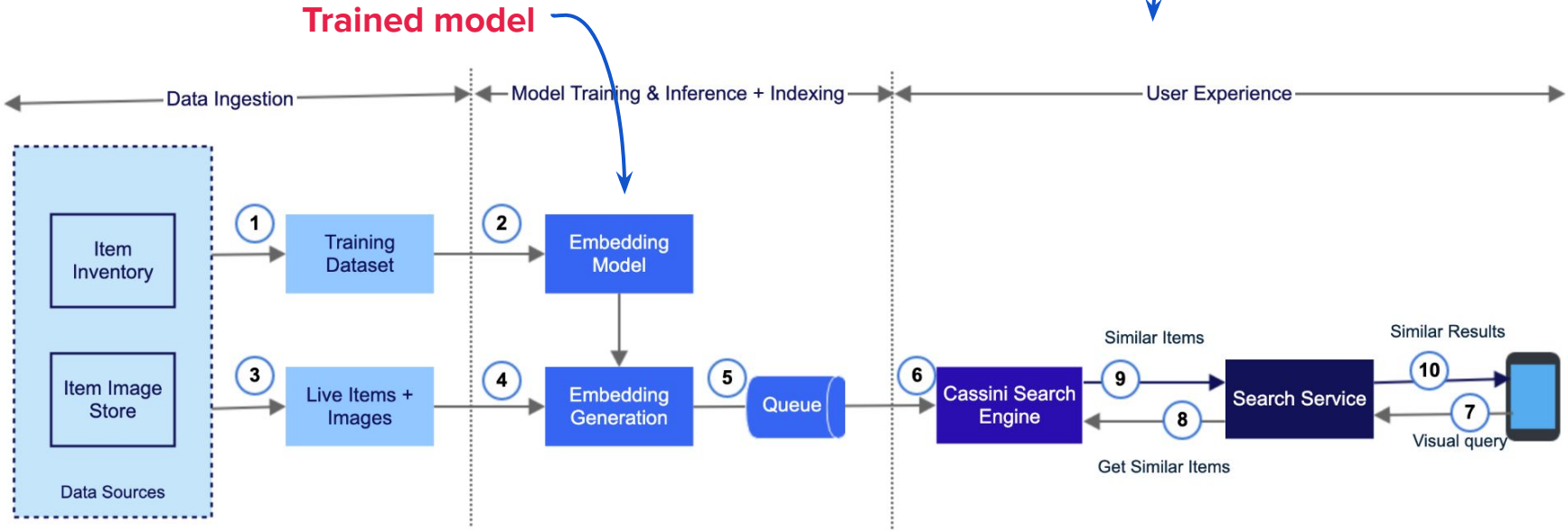


Quickly pivot to entire inventory using a visual first cue

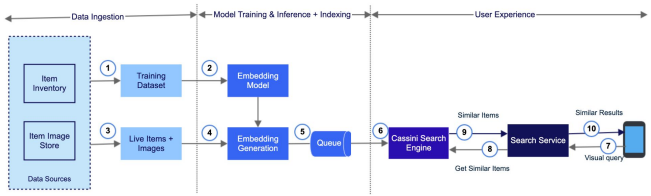
Data Engineering

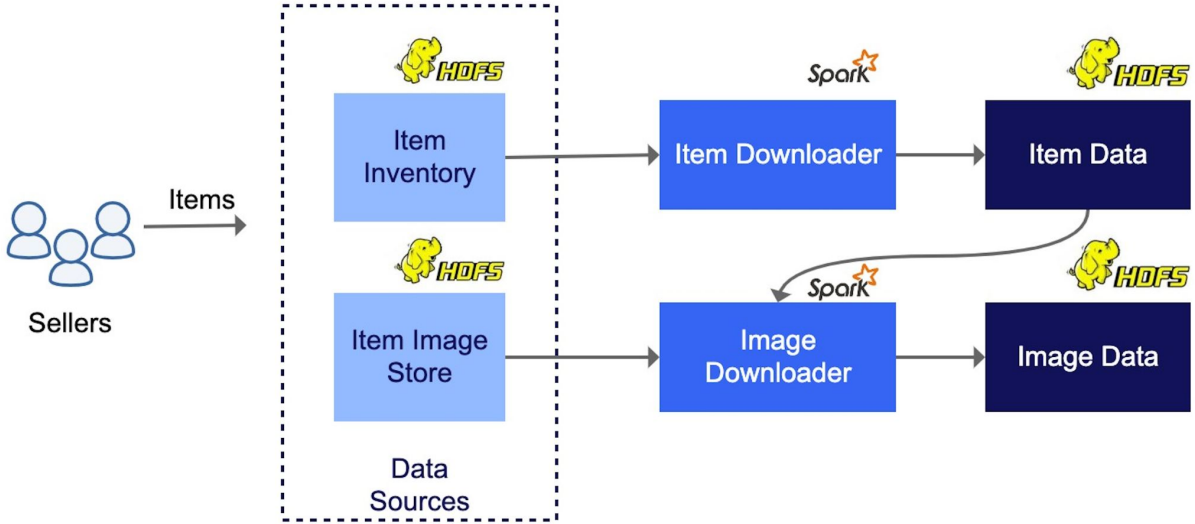
Design





Data Ingestion

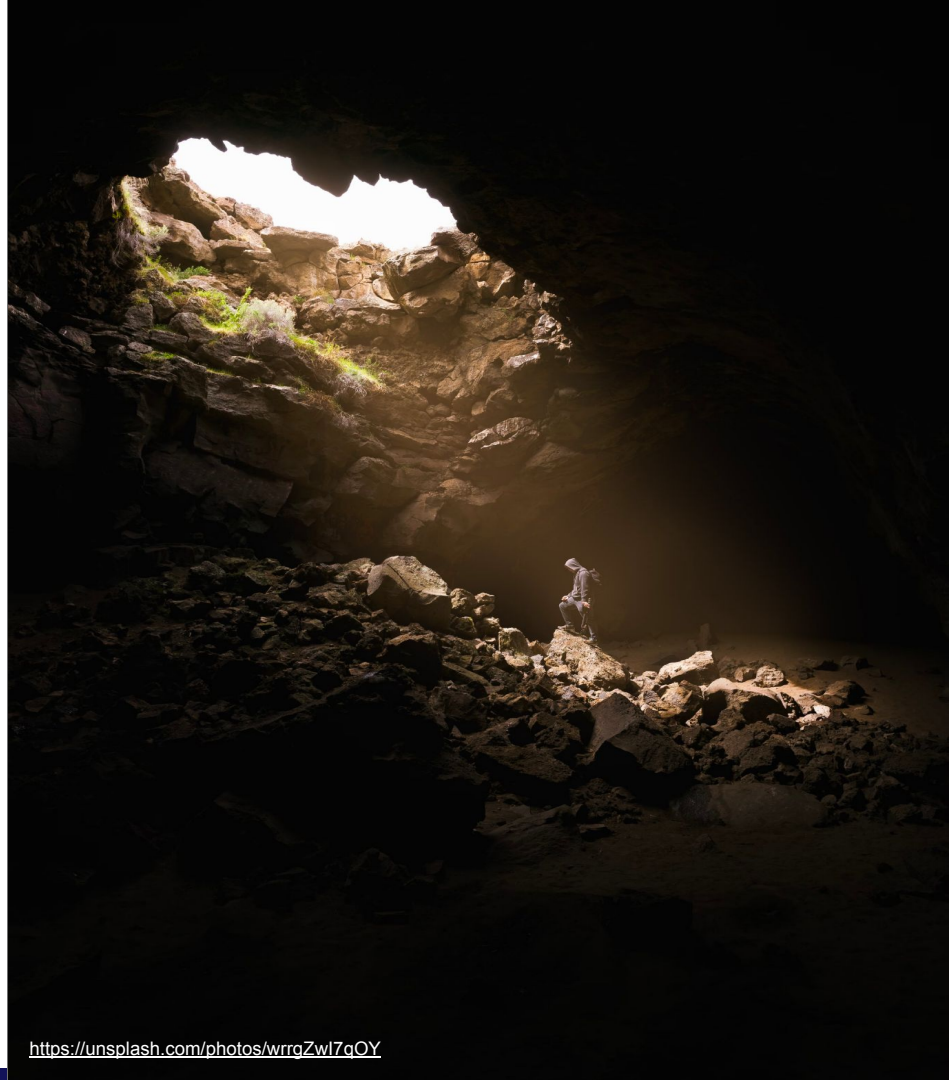




Data Ingestion

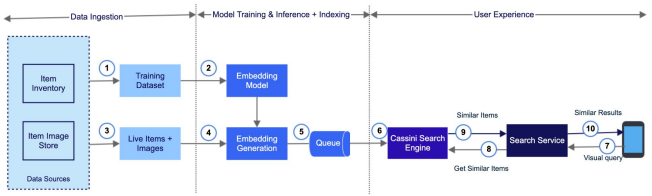
Challenges

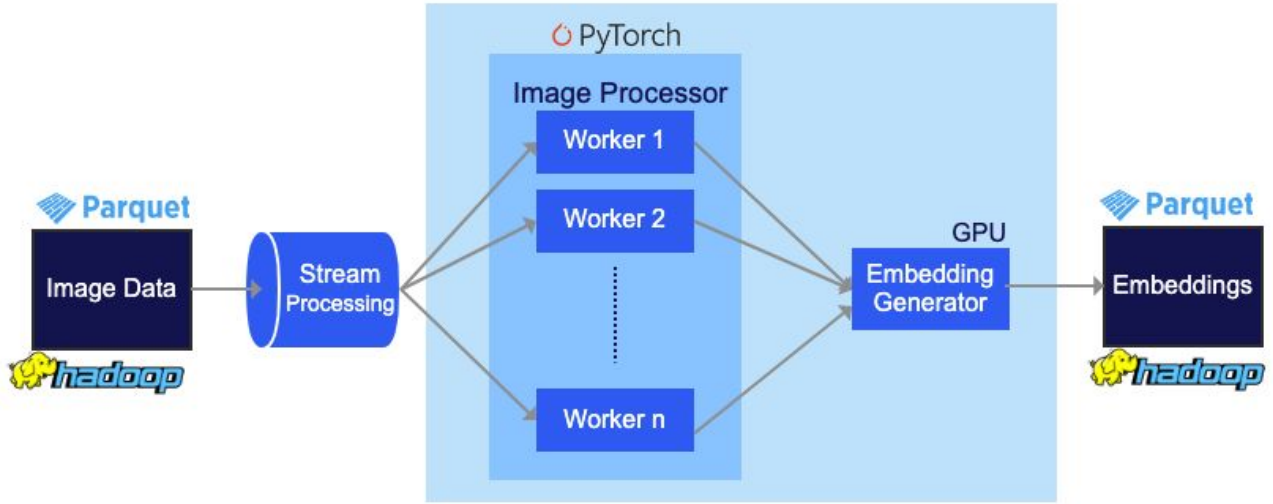
- Speed vs. resource trade off
- Storage
- Download errors
- Downstream dependencies



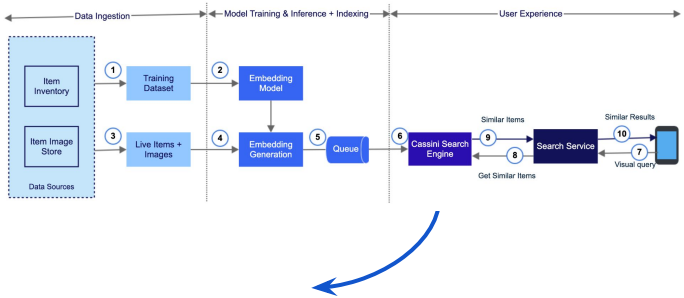
https://unsplash.com/photos/wrrgZwI7qOY

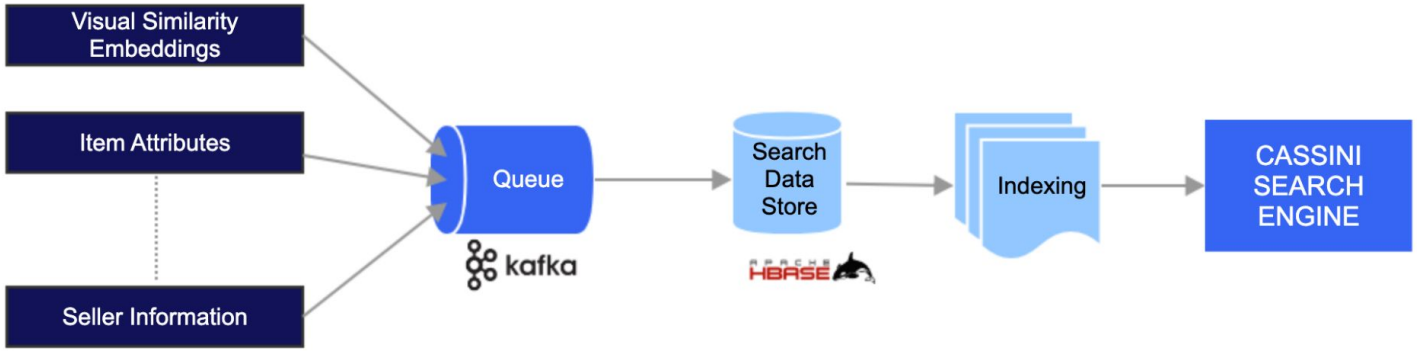
ML Platform and Inference





Cassini Indexing





The Architecture of eBay Search Trotman, Andrew, Jon Degenhardt, and Surya Kallumadi - eCOM@ SIGIR. 2017.

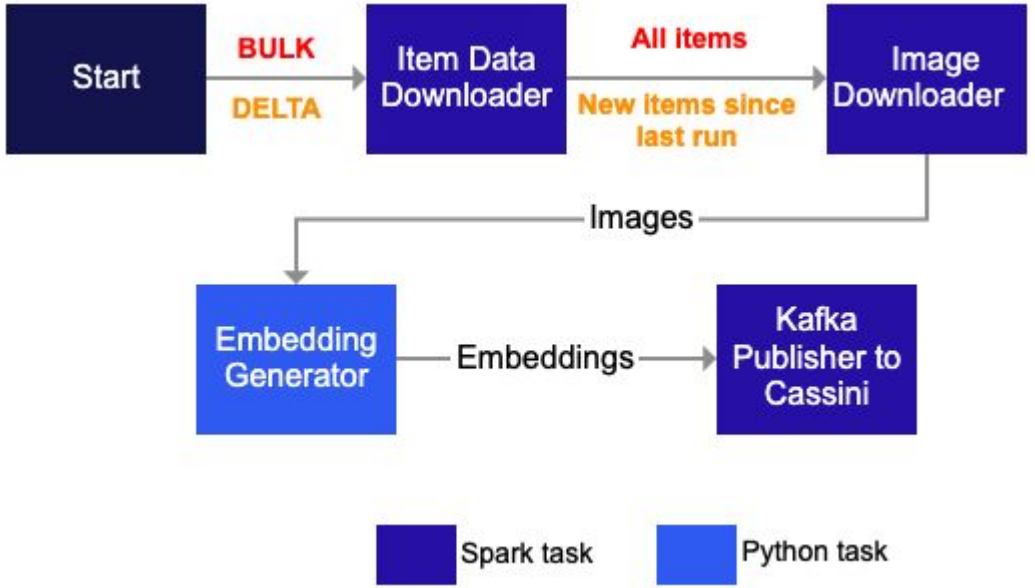
Orchestration

https://unsplash.com/photos/yUJVHiYZCGQ

Workflow Orchestration using Apache Airflow

Processing modes

- BULK
- DELTA



Apache Airflow Logo

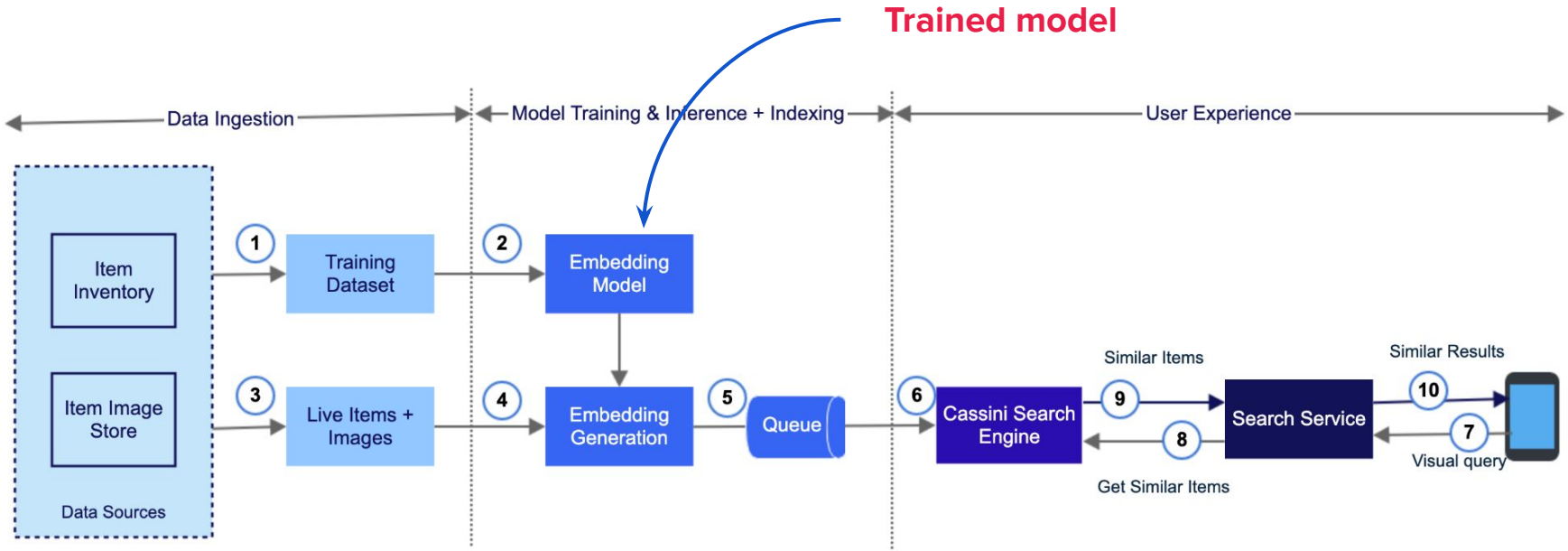
Challenges with Apache Airflow

Challenge	Solution
Multiple Spark versions	Define task level parameters
Multiple Docker image versions	Python virtual environment packages
Different platforms, zones, and network flakiness	Retries, system monitoring

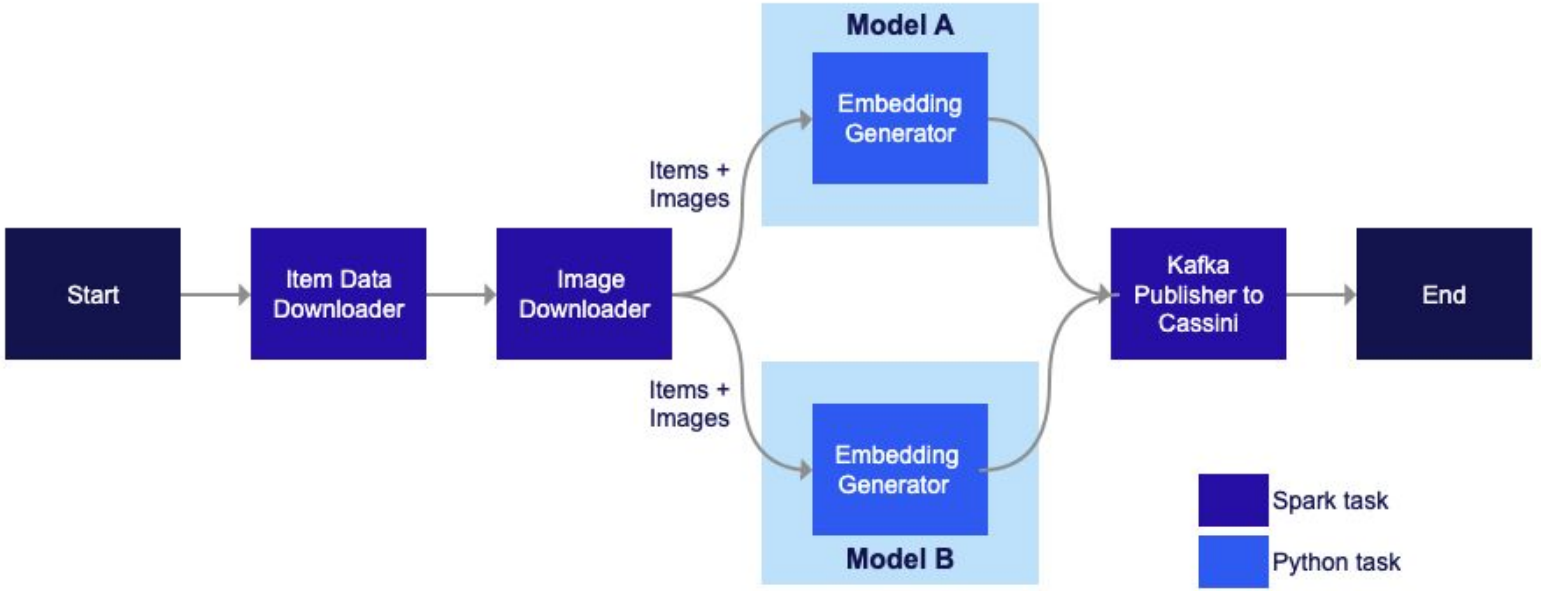
Apache Airflow Logo

A/B Testing

How do we test different models in production?



Data Publishing for A/B Tests using Airflow



Evolution

Model Drift

- Seasonality
- Aging of the models

Actions

- Metrics monitoring
- Downstream evaluation
- Retraining

Data Drift

- Data Integrity
- Data pipelines

Actions

- Fault tolerance
- Monitoring of time, cpu, memory, disk

Key takeaways

Similarity

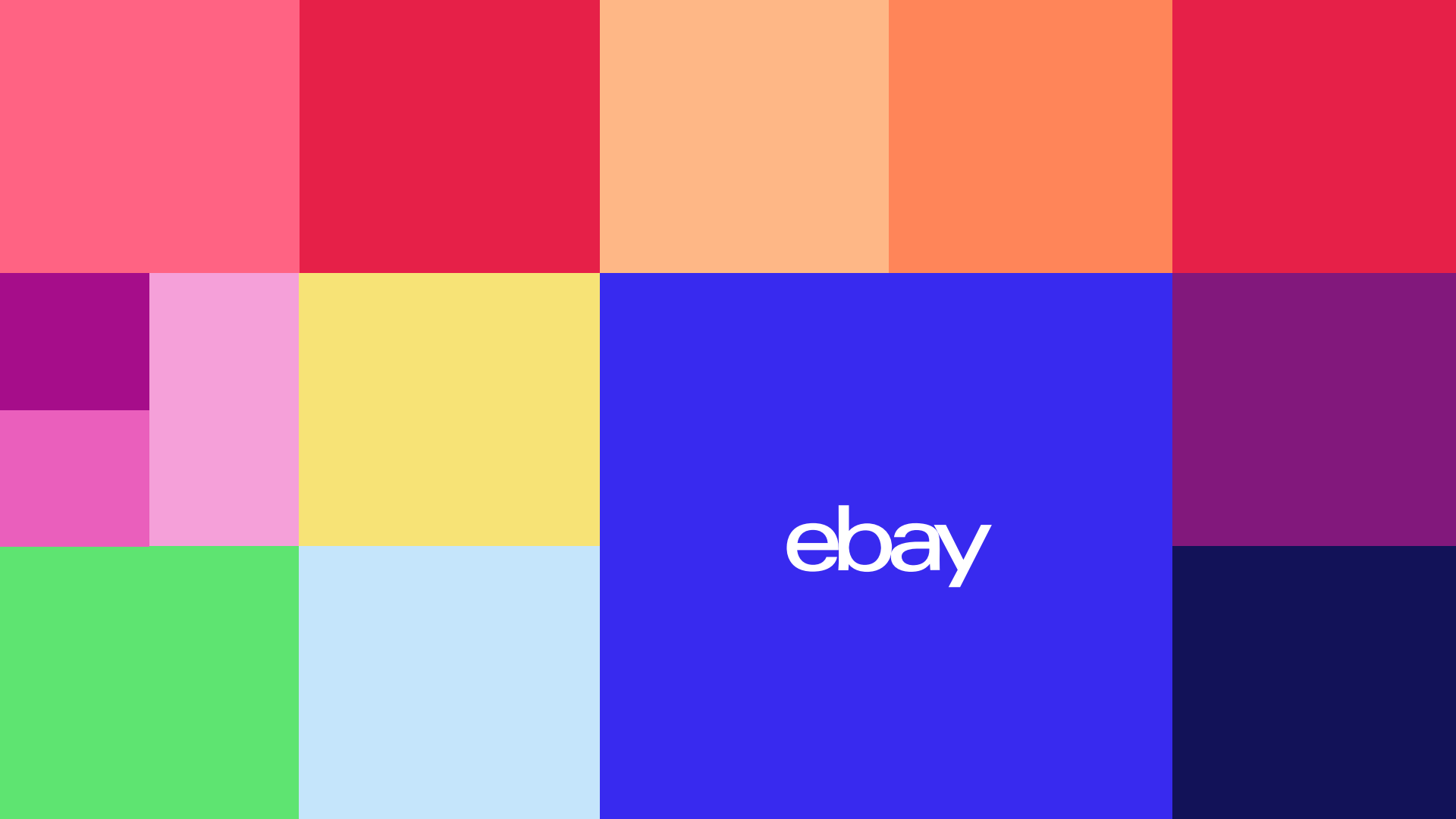
Scalability

Monitoring

Questions?

slides are available at **https://bit.ly/ebay-ml**

https://unsplash.com/photos/4V1dC_eoCwg


ebay