How AI/ML is changing information retrieval

Jon Handler
Director, Solutions Architecture, AWS
handler@amazon.com

Accessing Information

As soon as we started writing stuff down, we needed to find it

Ancient Sumer

Callimachus—*Pinakes* ca. 245 BCE

Card catalogs

- Organization
- Static
- Not interactive

But you needed to talk to a Librarian

OpenSearch

The boom in AI/ML

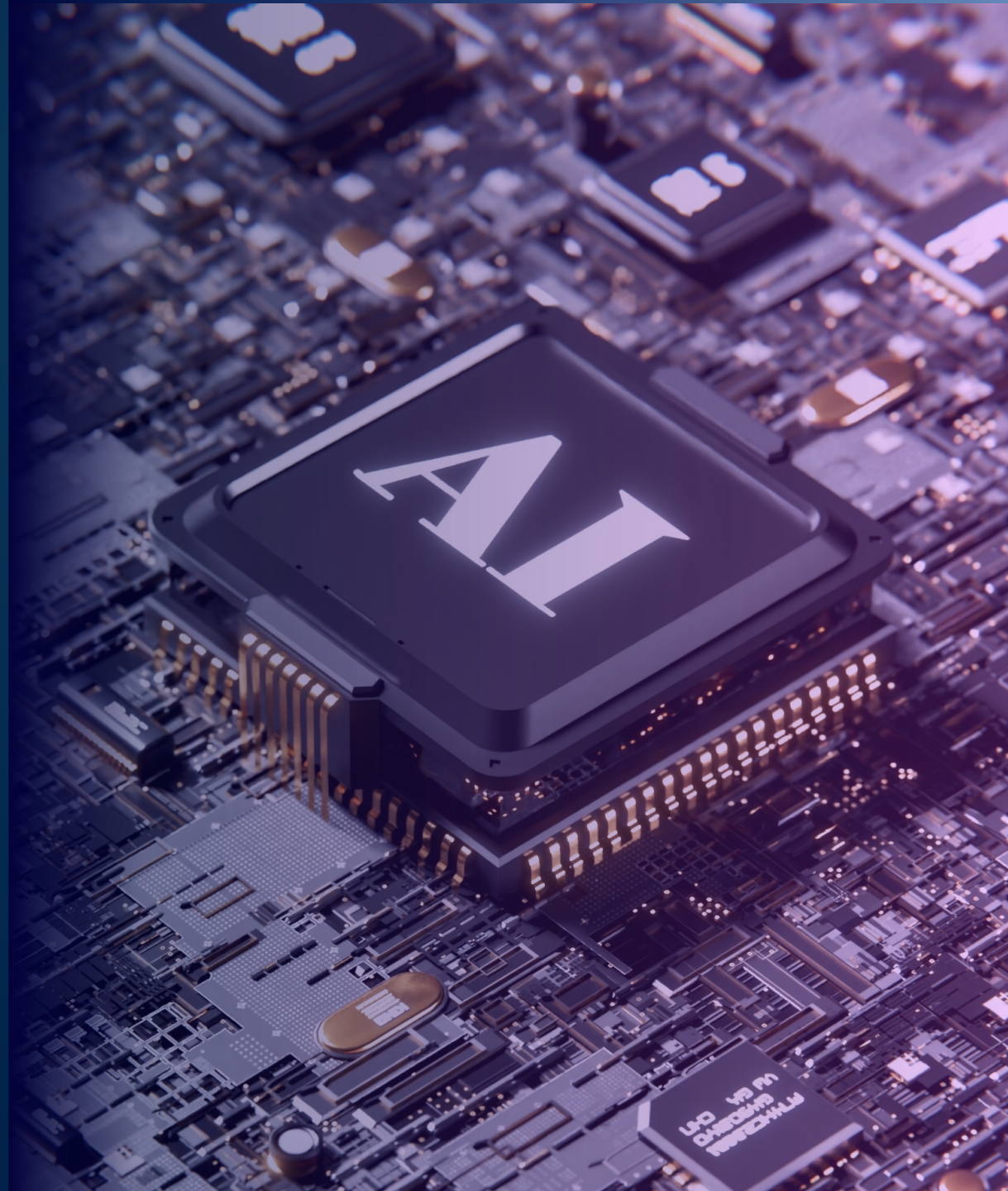
Large language models (LLM) are powering a new wave of generative AI

Embedding generating LLMs capture meaning from language

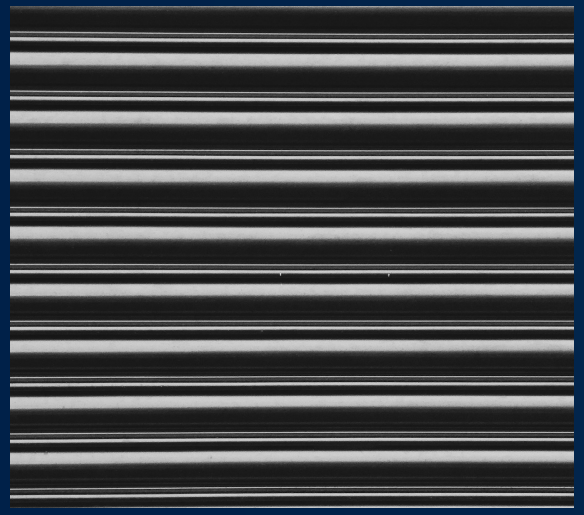
Text generating LLMs enable human-like chats in AI assistants and chatbots

Natural language is becoming the currency and yardstick for search

OpenSearch



Where we're headed



10 blue links will still be useful for some cases like parts search, where you know the exact name of the thing you want

Finding and interacting with information will become more long-term, and more based on conversation between human and computer

Lexical Search

OpenSearch

The basic idea

Words are units of meaning that code the concepts behind them, from concrete to abstract

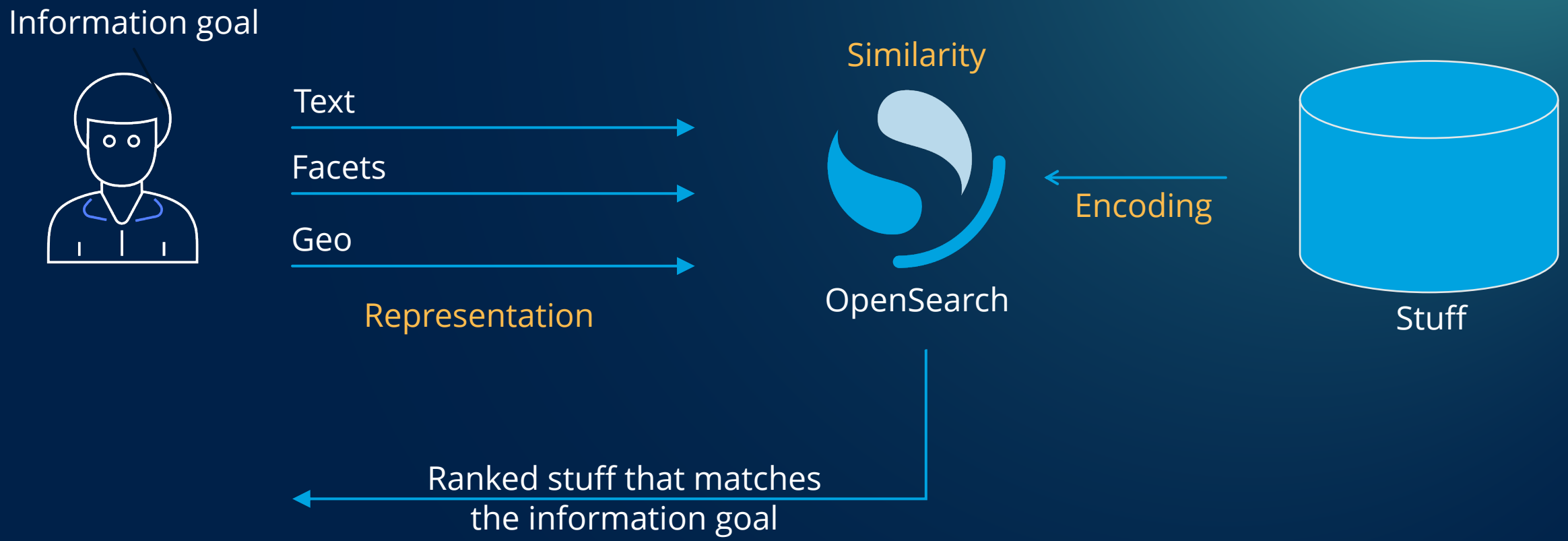
Words created for a purpose express that purpose through the assembly of concrete and abstract meaning

Search queries express user purpose through words (and some UI elements)

Match words in order to match meaning

The more matches the better, and the more information-specific to the purpose, the better

Search engines match words and score



OpenSearch

Encoding

Search documents represent search targets

Fields provide scoped units of meaning

Field names carry meaning as well, but only for the builders


```
{
  "question_text": "Im 5'8  and pretty thin.
                    Should i go for a small or a medium?\"",
  "category_name": "t-shirts",
  "question_type": "WH",
  "answer_aggregated": "N/A",
  "answers": [
    {
      "answer_text": "I'm 5 8 also , usually wear women's M.
                      I ordered a small and it's fitting
                      but good. I would say a small for you.\"",
      "gender": "other",
      "user_lon": -63.723771,
      "name": "Michelle Davenport",
      "user_lat": 6.134147,
      "age": 9,
      "product_rating": 3
    }
  ],
  "chunk": "This Parks & Recreation T-shirt features
            a I Met Li'l Sebastian design on an
            adult-sized cotton tee. QINOL Parks &
            Recreation - Lil' Sebastian T-Shirt Heather
            (S) Grey 100% Cotton Fully Machine
            Washable Fast And Free Shipping
            Printed in the U.S Brand New",
  "brand_name": "",
  "item_name": "QINOL Parks & Recreation -
                Lil' Sebastian T-Shirt
                Heather (S) Grey",
  "asin": "B005TGLE64",
  "product_description": "This Parks & Recreation
                         T-shirt features a I Met
                         Li'l Sebastian design on
                         an adult-sized cotton tee.",
  "question_id": "Tx125VNPD96ZLI6",
  "bullets": "100% Cotton Fully Machine Washable
              Fast And Free Shipping Printed in
              the U.S Brand New"
}
```

Text analysis: prep for matching

Source	Analyzed
Miss Kobayashi's Dragon Maid Tohru Cosplay Dress Outfit Package:dress+tie+belt+gloves Fabric:Cotton 90% & Polyester blend. Washable! Practical cotton cosplay! Christmas, Halloween, Birthday parties, Barbecue Party AND Daily Kitchen.	miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen

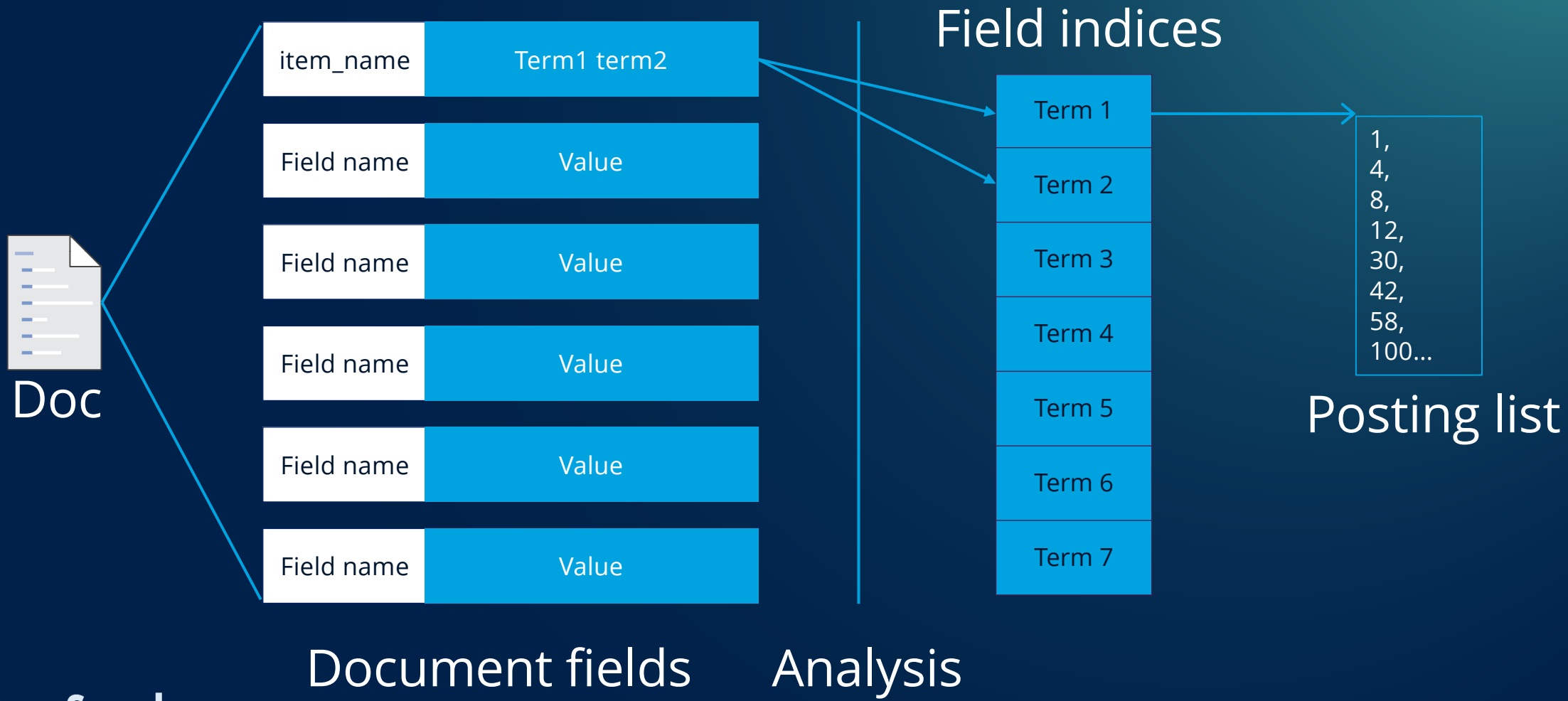
Stemming: brings words to a common form

Stop words: remove common terms that don't provide matching or discriminatory value

Synonyms: Add to increase matchability

All of these transformations work at the token level to improve the *meaning* of a token-token matching

Search indices map terms to posting lists



OpenSearch

Searching: query

```
GET amazon_pqa/_search
{
  "query": {
    "multi_match": {
      "query": "cotton washable
        clothes",
      "fields": ["bullets",
        "product_description",
        "item_name"],
      "operator": "and"
    }
  }
}
```

22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washable cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown

Searching: analysis

1. Analyze query: “cotton washable clothes” becomes “cotton washabl cloth”

22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washable cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

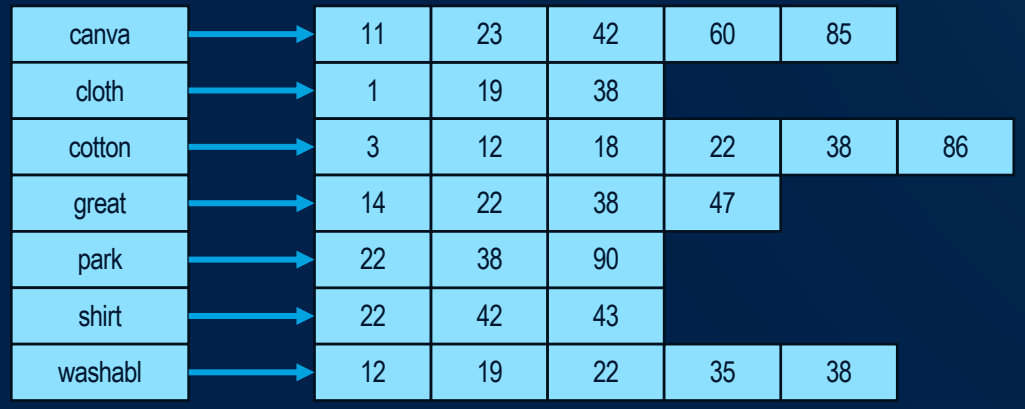
38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown

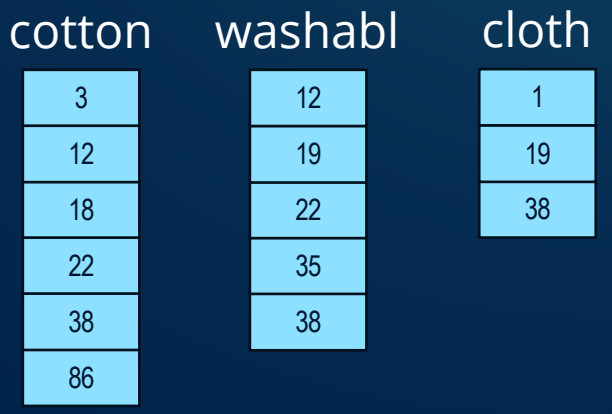
Searching: map

1. Analyze query: “cotton washable clothes” becomes “cotton washabl cloth”
2. Match terms and retrieve posting lists

22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washable cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown





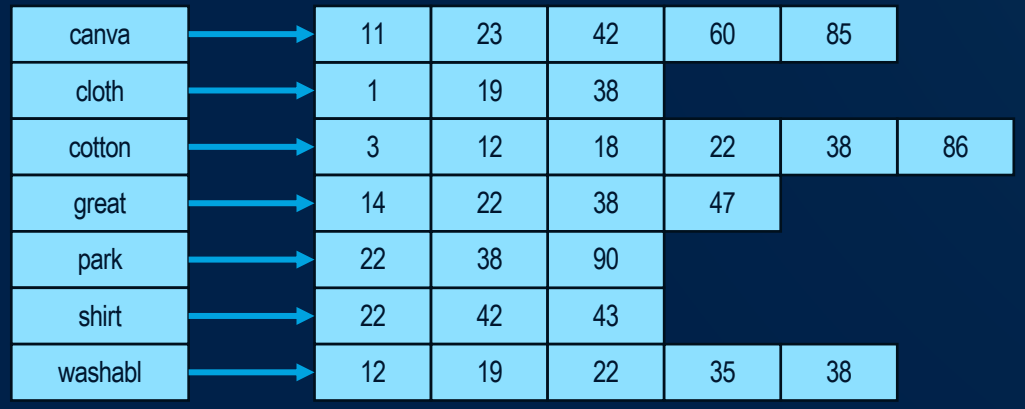
OpenSearch

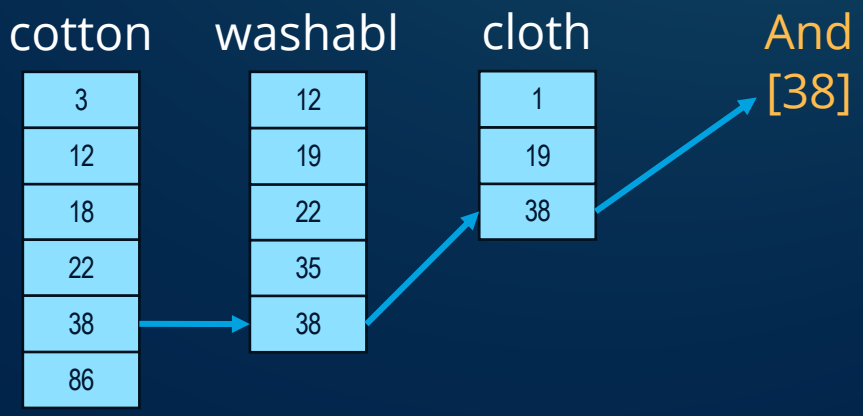
Searching: reduce

1. Analyze query: “cotton washable clothes” becomes “cotton washabl cloth”
2. Match terms and retrieve posting lists
3. Merge posting lists

22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washable cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown





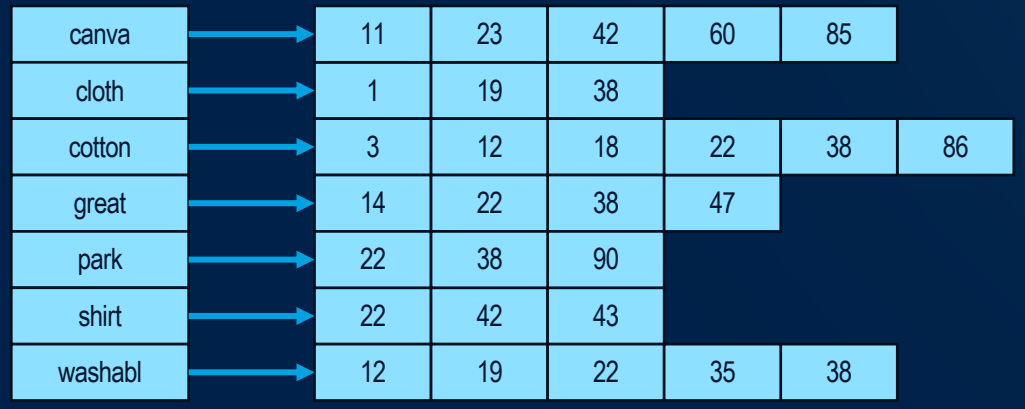
OpenSearch

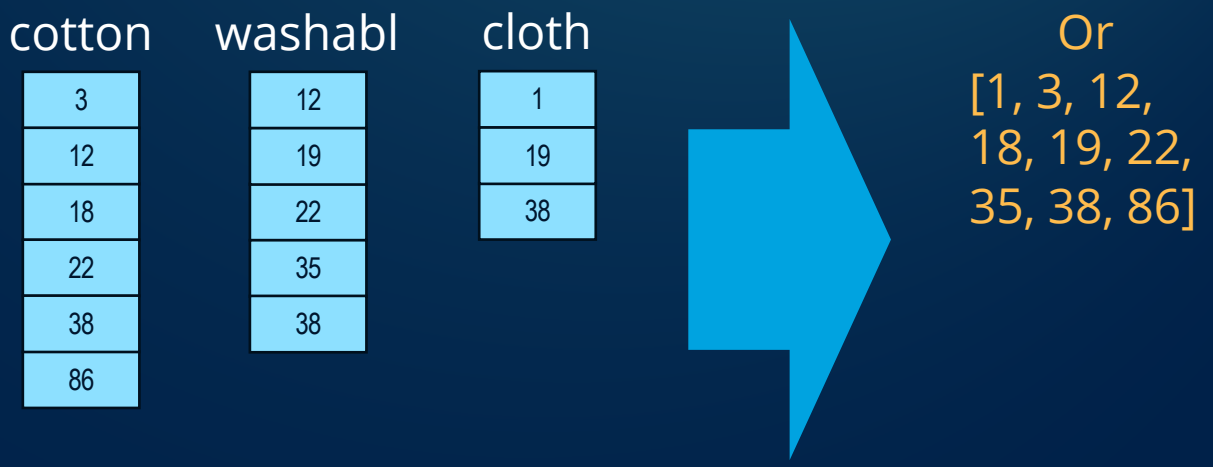
Searching: reduce

1. Analyze query: “cotton washable clothes” becomes “cotton washabl cloth”
2. Match terms and retrieve posting lists
3. Merge posting lists

22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washable cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown





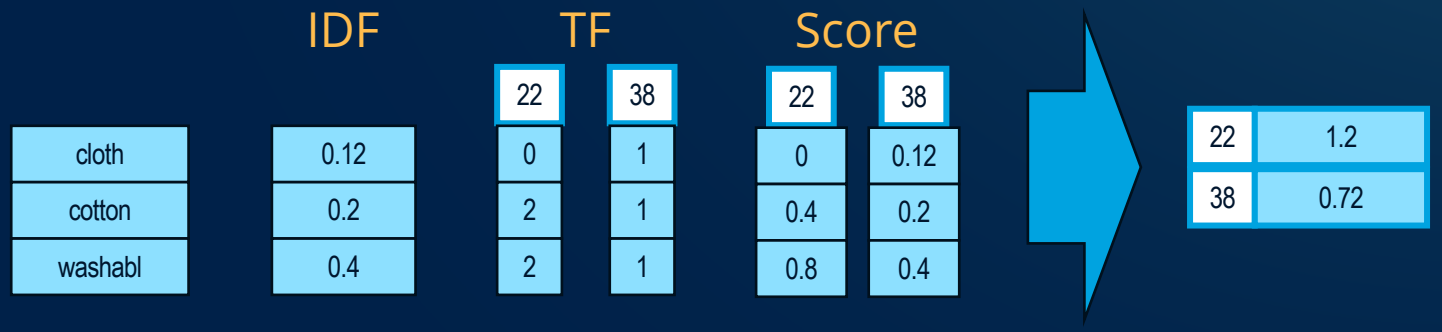
OpenSearch

Searching: score

1. Analyze query: “cotton washable clothes” becomes “cotton washabl cloth”
2. Match terms and retrieve posting lists
3. Merge posting lists
4. Score

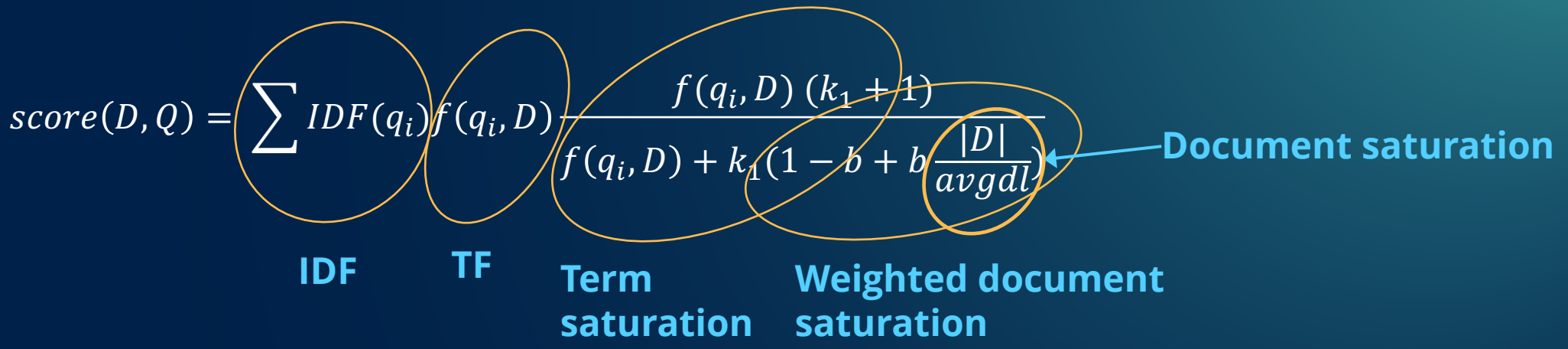
22: miss kobayashi dragon maid tohru cosplai dress outfit package:dress tie belt glove fabric:cotton 90 polyest blend washabl practic cotton cosplai christma halloween birthdai parti barbecu parti daili kitchen washabl cotton kitchen apron cosplai woman m 155 160cm miss kobayashi dragon maid tohru

38: veri soft silicon vinyl stuf pp cotton bodi babi 100 handmad craftwork hand print dimens 55cm 22 inch real babi size deviat exist due differ measur wai weight about 1.2 kg type silicon vinyl doll doll can put water ey close ey hair mohair implant hand high fidel washabl can comb cloth just pictur shown



OpenSearch

Okapi BM25

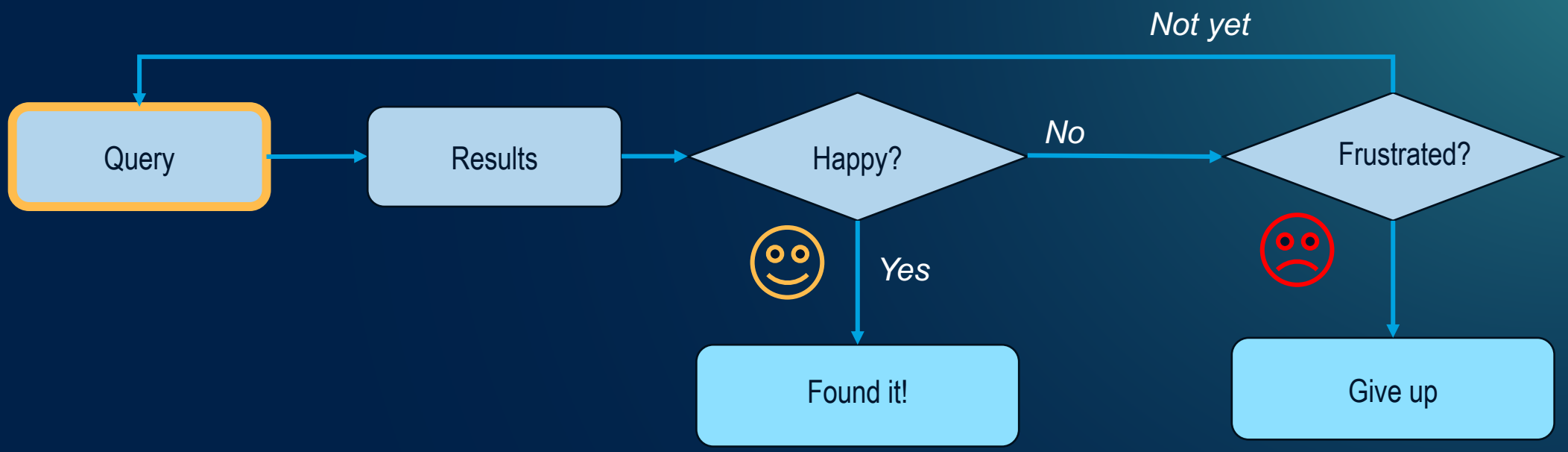


Based on probabilistic retrieval model

Considers term saturation & document length

Reward short documents, while penalizing matches in long documents

Did you get what you came for?



Lexical search leads to an iterative pattern of one-shot searches

To go beyond requires matching beyond single terms

OpenSearch

Vectors

OpenSearch

Why vectors?

Some ML Models (Large Language Models—LLMs) capture the meaning of words or blocks of text

They can emit vectors with values in many dimensions to represent this text

Related words and concepts cluster together

OpenSearch

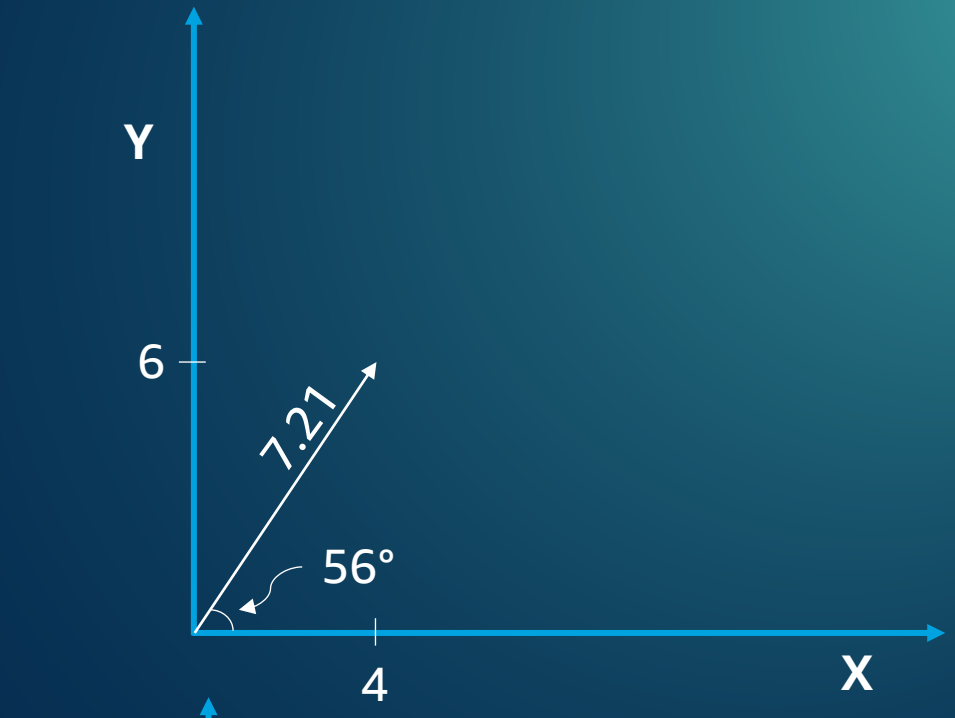
Vectors

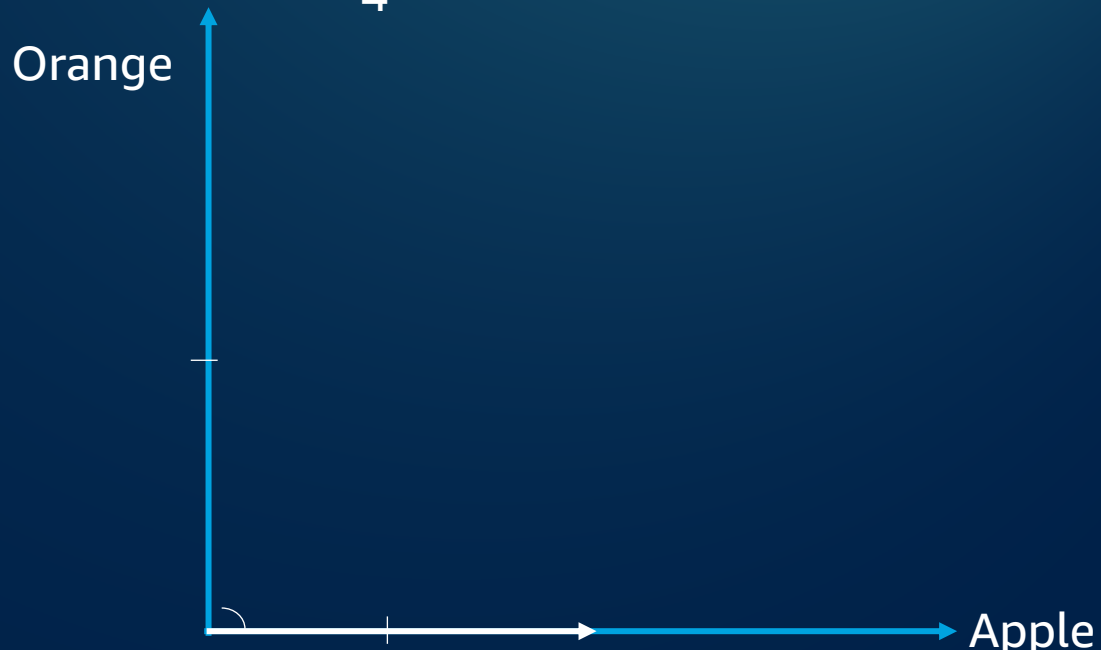
A vector is a quantity with a magnitude and a direction

In two dimensions, you can represent it with **X** and **Y** coordinates: (4, 6), on perpendicular axes

But these two directions could mean anything!

E.g.: Y axis corresponds to “Orange”, X axis corresponds to “Apple”, we could represent “Apple” as a vector





OpenSearch

One-hot coding

Apple	[1, 0, 0, ...]
Aardvark	[0, 1, 0, ...]
Banana	[0, 0, 1, ...]
Fruit	⋮
Red	
Vegetable	
White	[..., 0, 0, 1]

Dimensions are the same as the number of words in the corpus

Still useful – some ML techniques use 1-hot

Generalization is limited: Apple + Banana = ?

Sparse coding

TOMS Charcoal Felt Men's Searcher Boot. The Searcher Boot was constructed with the world traveler or urban explorer in mind. Featuring a combination of felt and leather on the upper, a ventilated footbed for breathability and TPR outsole for comfortable support and durability. Tongue gussets keeps out unwanted elements.

Dimensions are << the number of words in the corpus
Some generalization from collapse of tokens
Relationship maintained between tokens and values


```
"##ed": 0.0890511348435745,
"##r": 0.3354767560958623,
"##er": 0.934960901737213 1,
"man": 0.4592981934547424 3,
"found": 0.3969778120517730 7,
"best": 0.3757555782794952 4,
"men": 0.922002553939819 3,
"black": 0.1964904963970184 3,
"top": 0.0152543811127543 45,
"felt": 0.969756841659545 9,
"near": 0.074368007481098 18,
"##man": 0.028174642473459244,
"find": 0.721498548984527 6,
"research": 0.290300071239471 44,
"support": 0.31443119049072266,
"feel": 0.462964713573455 8,
```


OpenSearch

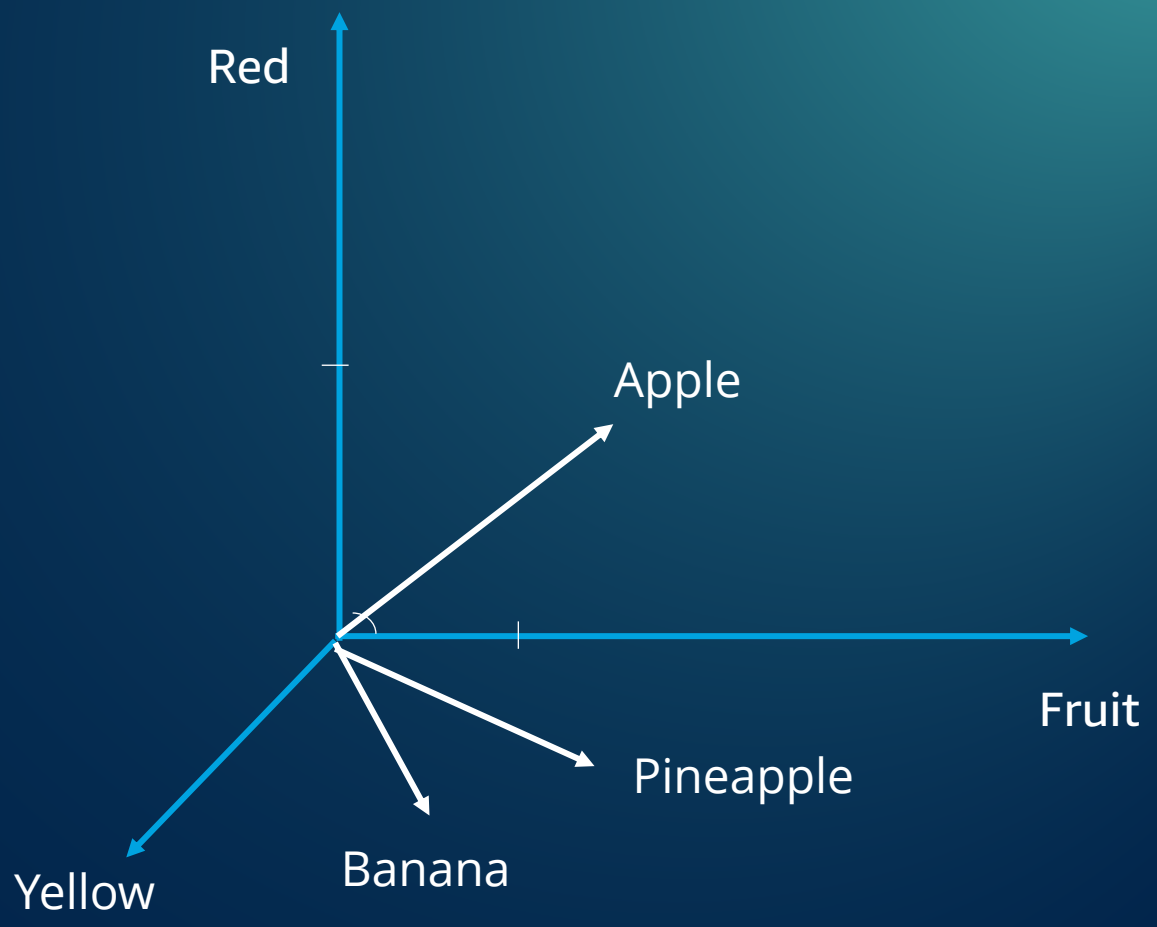
Dense coding

Select a few concepts to represent everything else

E.g. Y axis “Red”, X axis “Fruit” – “Red” + ”Fruit” = ”Apple”

Each dimension adds additional capacity

How to pick? Neural nets + backpropagation. Via backprop, the neural network learns the important concepts from your corpus



OpenSearch

Dense coding

Dimensions << # of words

Generalization is high… but hallucinations

Language is complex, non-linear!

The general equation for a line in n dimensions is

$$a_1x_1 + a_2x_2 + \cdots + a_n\,x_n = c$$

To represent language you need a complicated function to represent the space, billions of parameters

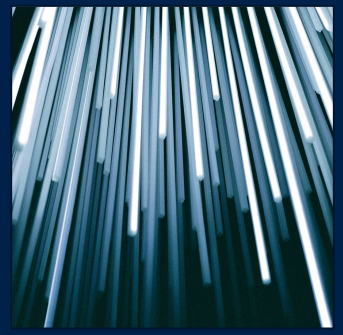
[0.46289062, -0.39257812, 0.25195312, 0.016845703, 0.20996094, 0.13867188, -0.20410156, 0.00041770935, 0.002029419, -0.15332031, 0.046142578, 0.4765625, 0.08642578, -0.07763672, -0.009399414, -0.123535156, 0.23730469, 0.15234375, 0.115234375, 0.079589844, 0.15527344, 0.078125, -0.24023438, 0.38867188, 0.078125, -0.3203125, -0.0058898926, 0.061035156, 0.33007812, -0.29101562, 0.46875, -0.078125, -0.11816406, -0.31640625, 0.20898438, -0.13183594, 0.07714844, 0.1640625, 0.34765625, -0.19824219, 0.35351562, -0.26171875, -0.12988281, -0.27148438, 0.0007209778, -0.16992188, -0.22460938, 0.14746094, 0.16894531, 0.28515625, -0.12109375, -0.07421875, 0.609375, -0.15039062, 0.55859375, 0.47265625, 0.20996094, 0.029541016, -0.15625, 0.064941406, 0.296875, 0.100097656, -0.31640625, 0.4375, 0.063964844, -0.011413574, 0.06201172, -0.14453125, 0.111328125, -0.038330078, -0.4296875, 0.09423828, -0.27734375, -0.5703125, 0.17285156, 0.0020141602, -0.048095703, 0.18066406, 0.30078125, -0.38085938, -0.107421875, 0.34765625, -0.045654297, 0.29101562, -0.0061035156, 0.31640625, -0.26953125, -0.2421875, 0.00013828278, -0.32226562, 0.19921875, 0.38476562, 0.62890625, -0.035888672, 0.025634766, 0.12792969, 0.33789062, -0.076660156, -0.91015625, -0.17382812, -0.15527344, -0.16503906, 0.06225586, 0.08300781, -0.010559082, 0.24121094, 0.46484375, 0.24121094, 0.16015625, 0.11621094, 0.328125, 0.026855469, -0.41210938, 0.24804688, -0.22753906, -0.34765625, -0.23730469, -0.20996094, -0.39453125, 0.36132812, 0.5625, 0.0065612793, 0.034179688, 0.018554688, -0.05834961, -0.24804688, -0.03173828, 0.19140625, 0.06640625, 0.18261719, 0.09667969, -0.265625, 0.006652832, 0.29101562, 0.203125, -0.19042969, -0.25, -0.09863281, 0.22363281, -0.0022125244, -0.40234375, -0.20996094, -0.1796875, -0.28320312, 0.011108398, -0.42773438, -0.10595703, 0.1015625, -0.24023438, -0.18359375, -0.087890625, -0.20898438, -0.3359375, 0.6875, 0.4453125, 0.025512695, -0.4921875, 0.50390625, -0.16210938, -0.3671875, 0.15234375, -0.16894531, 0.23339844, 0.03515625, -0.22070312, -0.15625, 0.19921875, 0.32421875, 0.46679688, 0.025268555, -0.032226562, 0.19726562, -0.16601562, -0.13085938, 0.08544922, 0.56640625, 0.40625, -0.13476562, -0.74609375, 0.33007812, 0.12988281, -0.5234375, 0.036132812, -0.049072266, 0.5390625, 0.16308594, 0.41601562, 0.009460449, -0.16503906, 0.14941406, -0.026733398, -0.118652344, 0.045410156, 0.036376953, 0.15625, -0.2578125, 0.28320312, -0.06738281, -0.24511719, -0.110839844, -0.037597656, -0.08300781, -0.39453125, -0.22753906, -0.30273438, -0.11376953, -0.21386719, -0.056396484, 0.067871094, 0.20117188, -0.32421875, -0.46289062, -0.14941406, -0.2578125, 0.26953125, 0.080566406, 0.1015625, 0.22851562, -0.33398438, 0.46484375, -0.06738281, -0.30078125, -0.06933594, -0.44921875, -0.24511719, -0.099121094, 0.28515625, -0.13671875, 0.23046875, -0.07910156, 0.52734375, -0.048583984, 0.19628906, -0.66015625, 0.08203125, -0.16699219, -0.27148438,…

OpenSearch

Natural Language and Search

OpenSearch

New paradigms for searching



Sparse

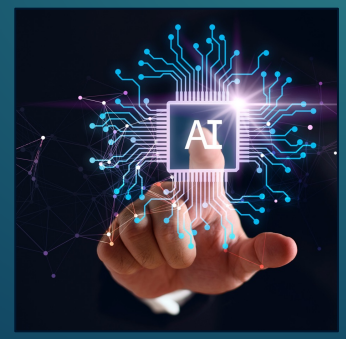
Sparse models more closely encode source tokens and retain better relevance for exact matching

Multi-modal

LLMs and other technologies can create embeddings for text, images, audio, etc. People are using these additional media to improve search relevance

Hybrid

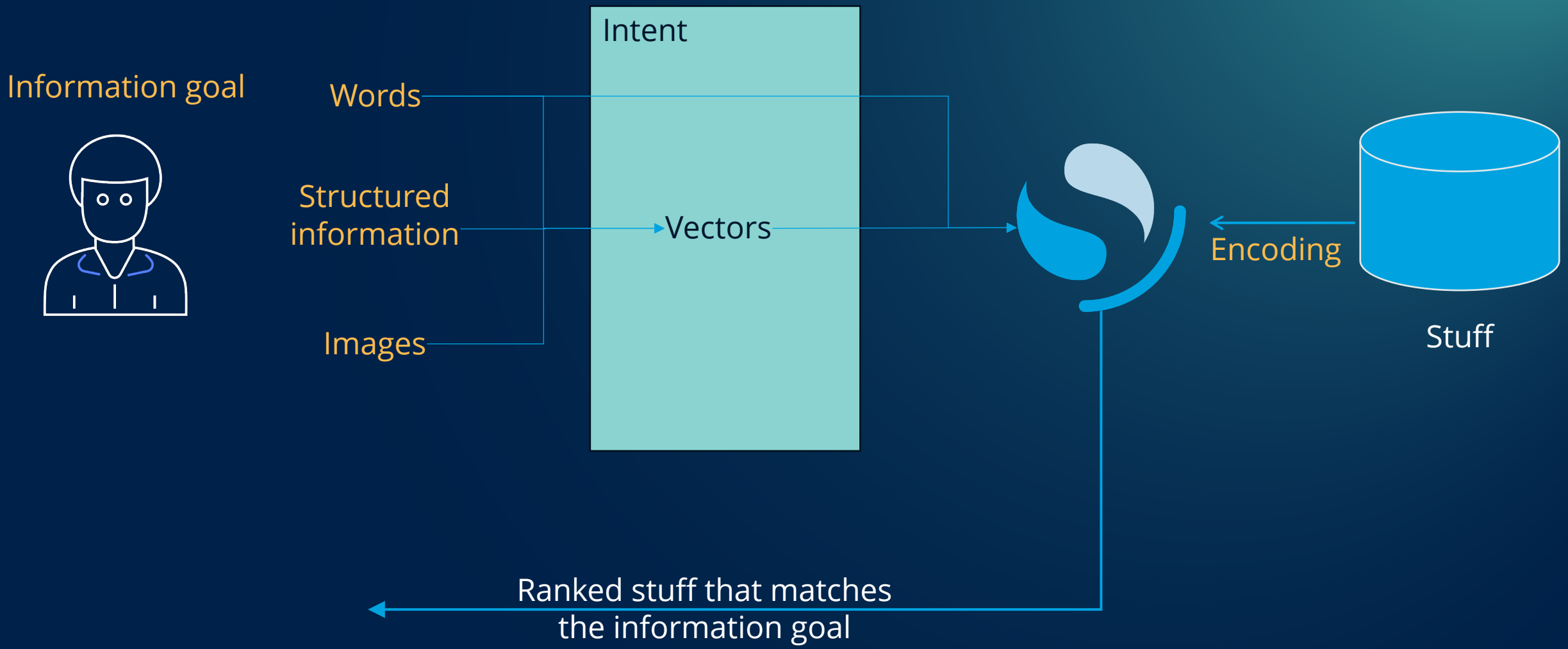
Lexical search and semantic search both have their benefits and right applications. Hybrid blends scores from vector and lexical search to improve overall relevance



Conversational

As NL capabilities improve and chatbots are commonplace, people are turning to these bots for simple searching.

The basic idea



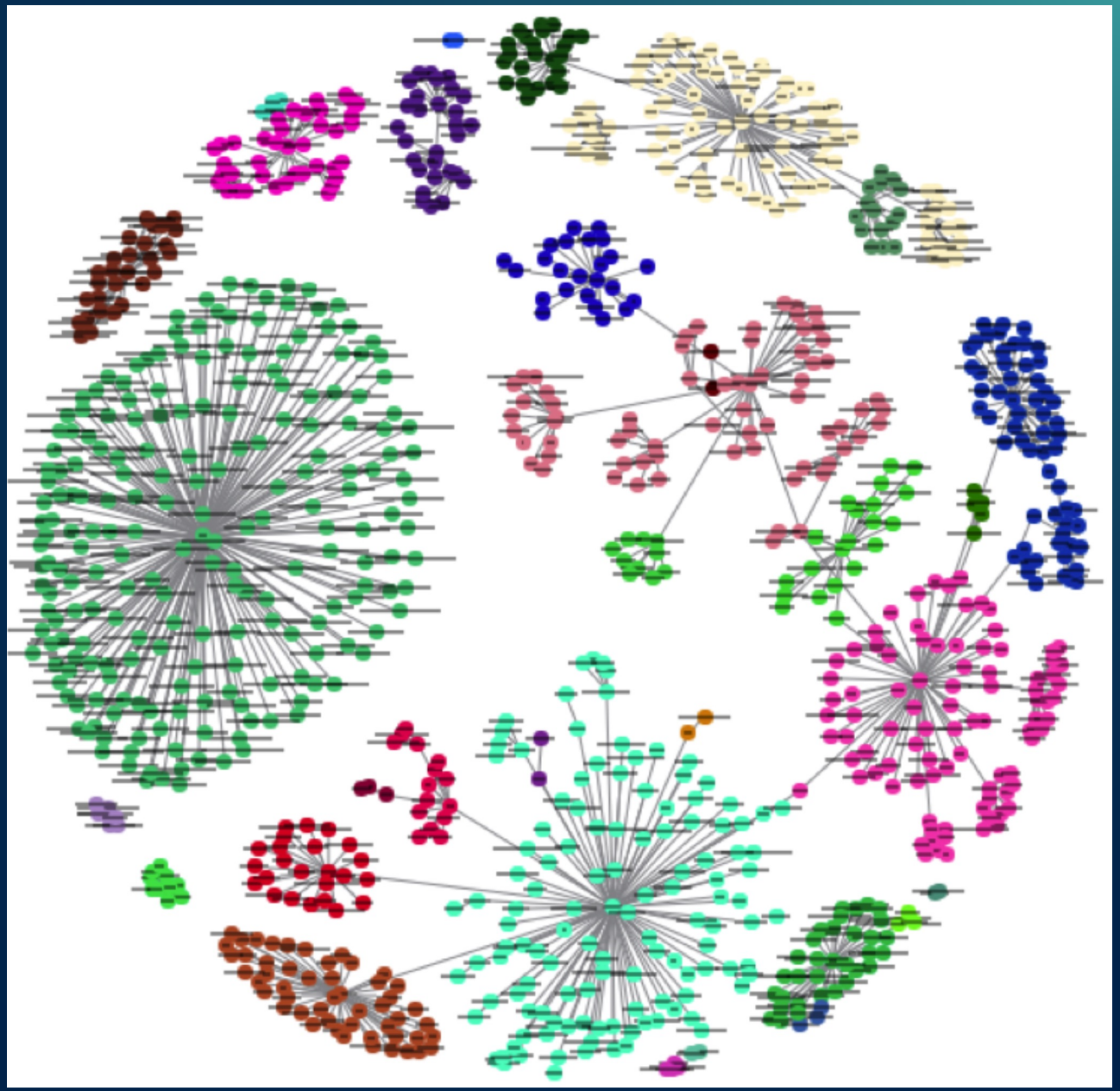
OpenSearch

Vector similarity

Documents with similar “meaning” are encoded “near” one another

The score for a query-document pair is measured by distance

Different distance measures include L1, L2 (Euclidean), LInf, cosine, inner product, and Manhattan,



OpenSearch

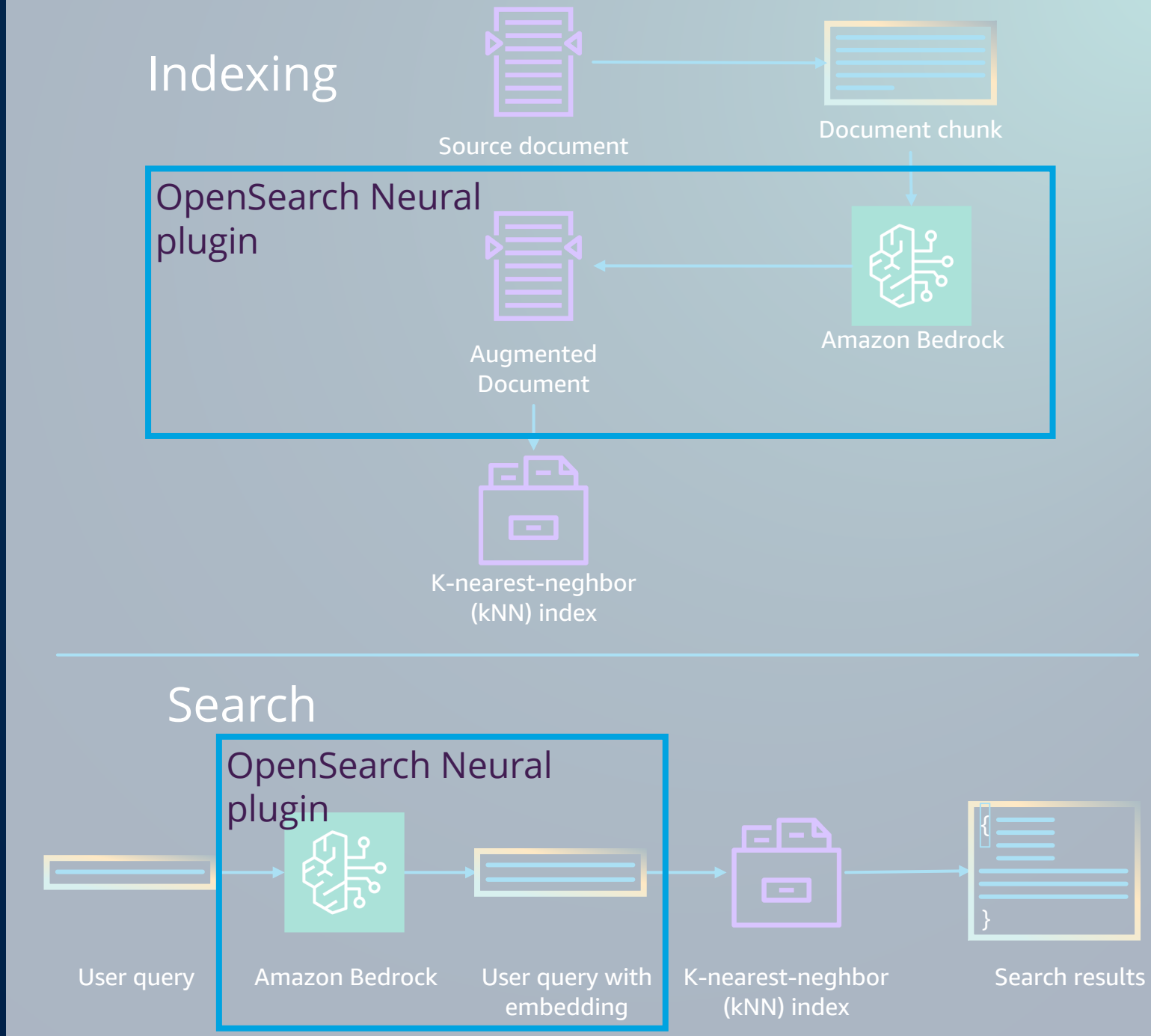
Neural Plugin

OpenSearch provides connectors to 3P model-hosting systems – e.g., Amazon Bedrock

Indexing: Select chunks from the source document, send to the 3P system for embeddings

Search: Create embedding for the query then find nearest neighbors

OpenSearch



Ingest, search pipelines for Neural

```
PUT /_ingest/pipeline/nlp-ingest-pipeline
{
  "description": ”Text embedding pipeline",
  "processors": [ {
    "text_embedding": {
      "model_id": "XXXX",
      "field_map": {
        ”chunk": ”chunk_embedding"
      }
    }
  } ]
}
```

```
PUT /_search/pipeline/nlp-search-pipeline
{
  "description": "Hybrid search",
  "phase_results_processors": [{
    "normalization-processor": {
      "normalization": {
        "technique": "min_max"
      },
      "combination": {
        "technique": "arithmetic_mean",
        "parameters": {
          "weights": [ 0.3, 0.7 ]
        }
      }
    }
  }]
}
```

OpenSearch

Hybrid

bullets: Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.
chunk_embedding: [0.078125, -0.24023438, ...]

OpenSearch

Query: Styling Kicks

Amazon Bedrock

```
PUT /_search/pipeline/nlp-search-pipeline
{
  "description": "Hybrid search",
  "phase_results_processors": [
    {
      "normalization-processor": {
        "normalization": {
          "technique": "min_max"
        },
        "combination": {
          "technique": "arithmetic_mean",
          "parameters": {
            "weights": [
              0.3,
              0.7
            ]
          }
        }
      }
    }
  ]
}
```

```
GET amazon_pqa/_search?search_pipeline=nlp-search-pipeline
{
  "query": {
    "hybrid": {
      "queries": [
        {
          "match": {
            "bullets": {
              "query": "stylin kicks"
            }
          }
        },
        {
          "neural": {
            "chunk_embedding": {
              "query_text": "stylin kicks",
              "model_id": "7OfGJY8BoFwiDML8j-zu",
              "k": 5
            }
          }
        }
      ]
    }
  }
}
```

OpenSearch

Sparse

Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.

OpenSearch

Query: Styling Kicks

Amazon Sagemaker sparse model

```
PUT /amazon_pqa_sparse
{
  "settings": {
    "default_pipeline": "sparse"
  },
  "mappings": {
    "properties": {
      "chunk_sparse": {
        "type": "rank_features"
      },
      "chunk": {
        "type": "text"
      }
    }
  }
}
```

OpenSearch

Sparse

Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.

OpenSearch

Query: Styling Kicks

Amazon Sagemaker sparse model

```
PUT /_ingest/pipeline/sparse
{
  "description": "Sparse encoding",
  "processors": [
    {
      "sparse_encoding": {
        "model_id":
          "7OfGJY8BoFwiDML8j-zu",
        "field_map": {
          "chunk": "chunk_sparse"
        }
      }
    }
  ]
}
```

OpenSearch

Sparse

Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.

OpenSearch

Query: Styling Kicks

Amazon Sagemaker sparse model

```
PUT /amazon_pqa_sparse
{
  "settings": {
    "default_pipeline": "sparse"
  },
  "mappings": {
    "properties": {
      "chunk_sparse": {
        "type": "rank_features"
      },
      "chunk": {
        "type": "text"
      }
    }
  }
}
```

```
GET amazon_pqa/_search
{
  "query": {
    "neural": {
      "chunk_embedding": {
        "query_text": "stylin kicks",
        "model_id":
          "7OfGJY8BoFwiDML8j-zu",
        "k": 100
      }
    }
  }
}
```

OpenSearch

Multimodal

Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.

OpenSearch

Query: Styling Kicks

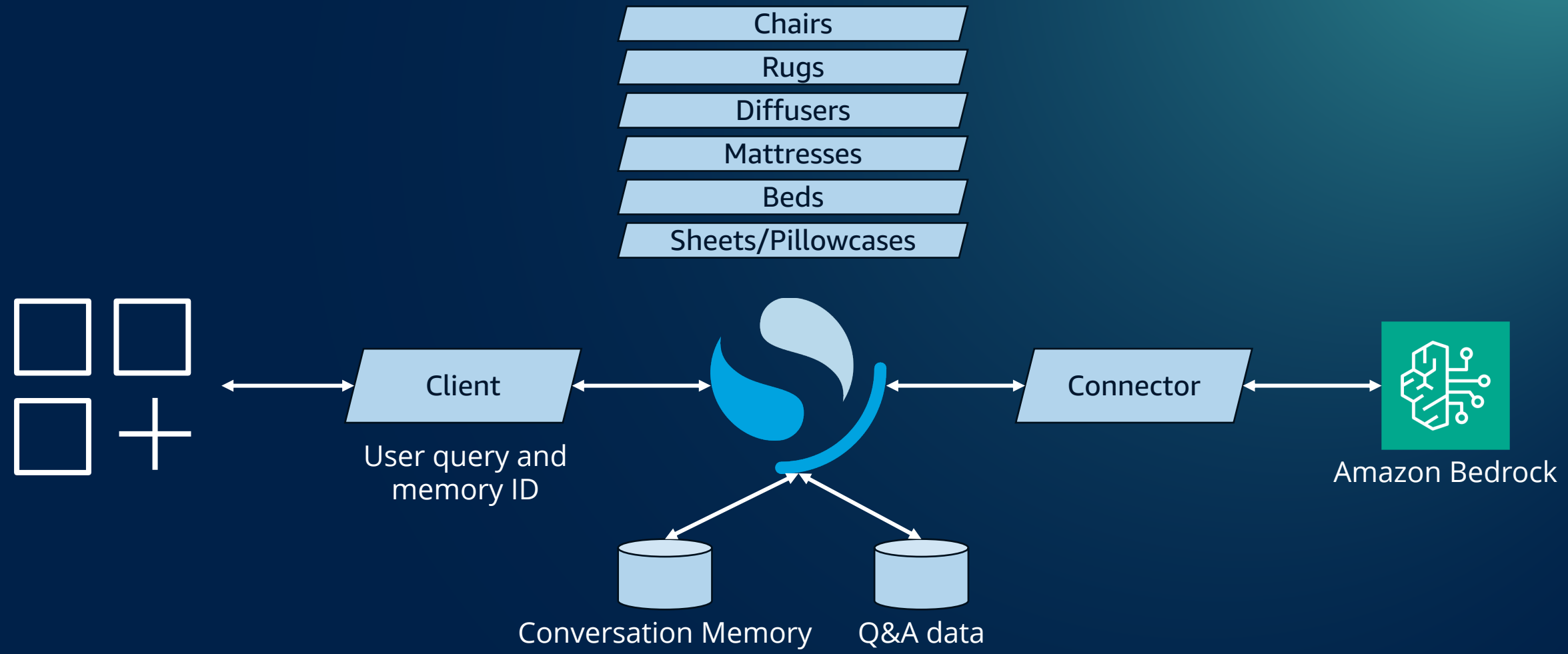
Amazon Bedrock

```
PUT /_ingest/pipeline/multimodal
{
  "description": "multimodal
    pipeline",
  "processors": [
    {
      "text_image_embedding": {
        "model_id":
          "afYQAosBQkdnhhBsK593",
        "embedding":
          "chunk_embedding",
        "field_map": {
          "text": "chunk",
          "image": "image_binary"
        }
      }
    }
  ]
}
```

```
GET /my-nlp-index/_search
{
  "size": 10,
  "query": {
    "neural": {
      "chunk_embedding": {
        "query_text": "stylin kicks",
        "query_image":
          "iVBORw0KGgoAAAANSUI...",
        "model_id": "
          -fYQAosBQkdnhhBsK593",
        "k": 5
      }
    }
  }
}
```

OpenSearch

Conversational search



OpenSearch

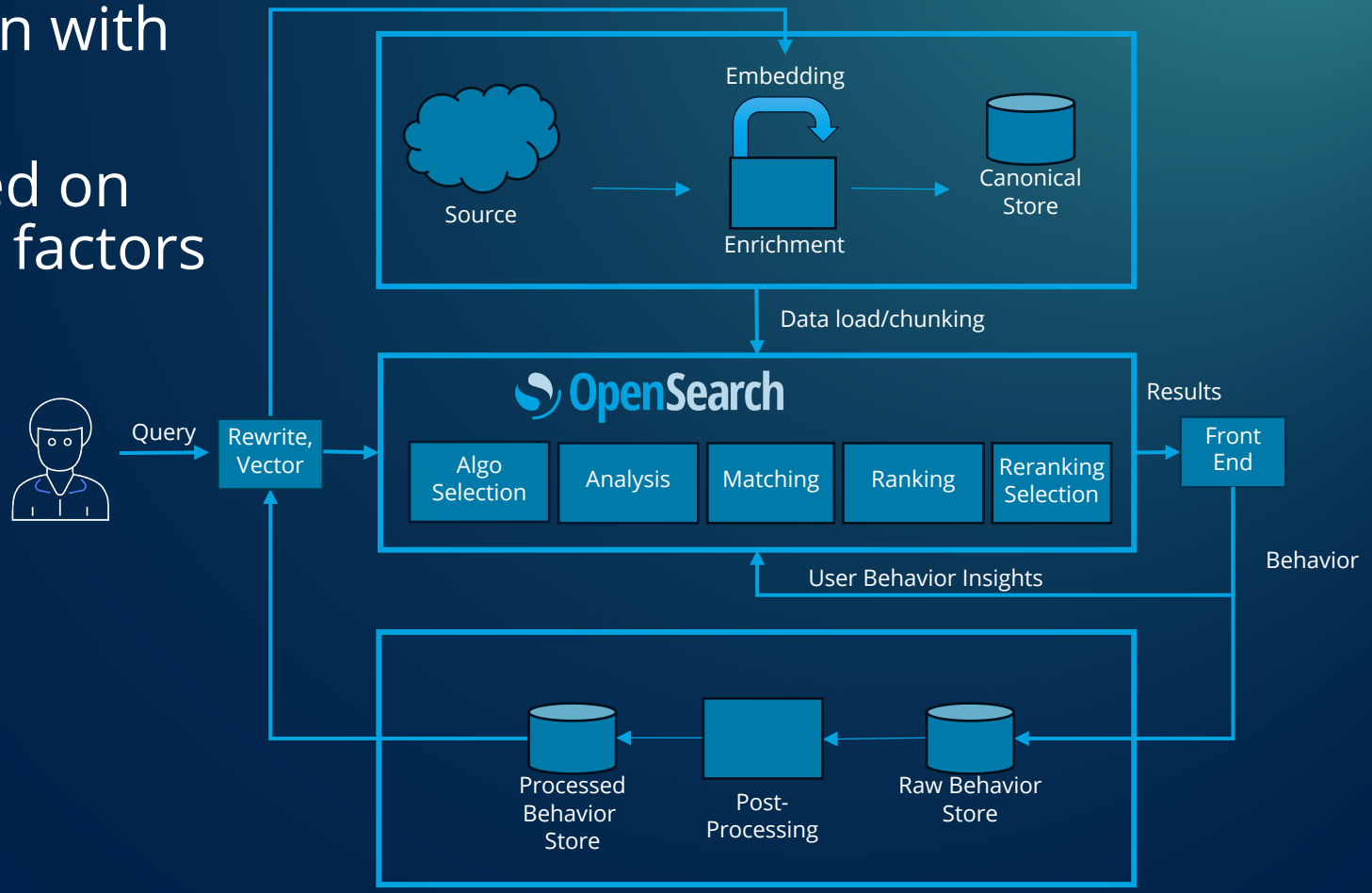
Where to from here?

PREPARING, ANALYZING, VECTORIZING, RANKING, AND RERANKING DOCUMENTS

Employing behavior information with User Behavior Insights tracking

Query algorithm selection based on concrete vs. abstract and other factors

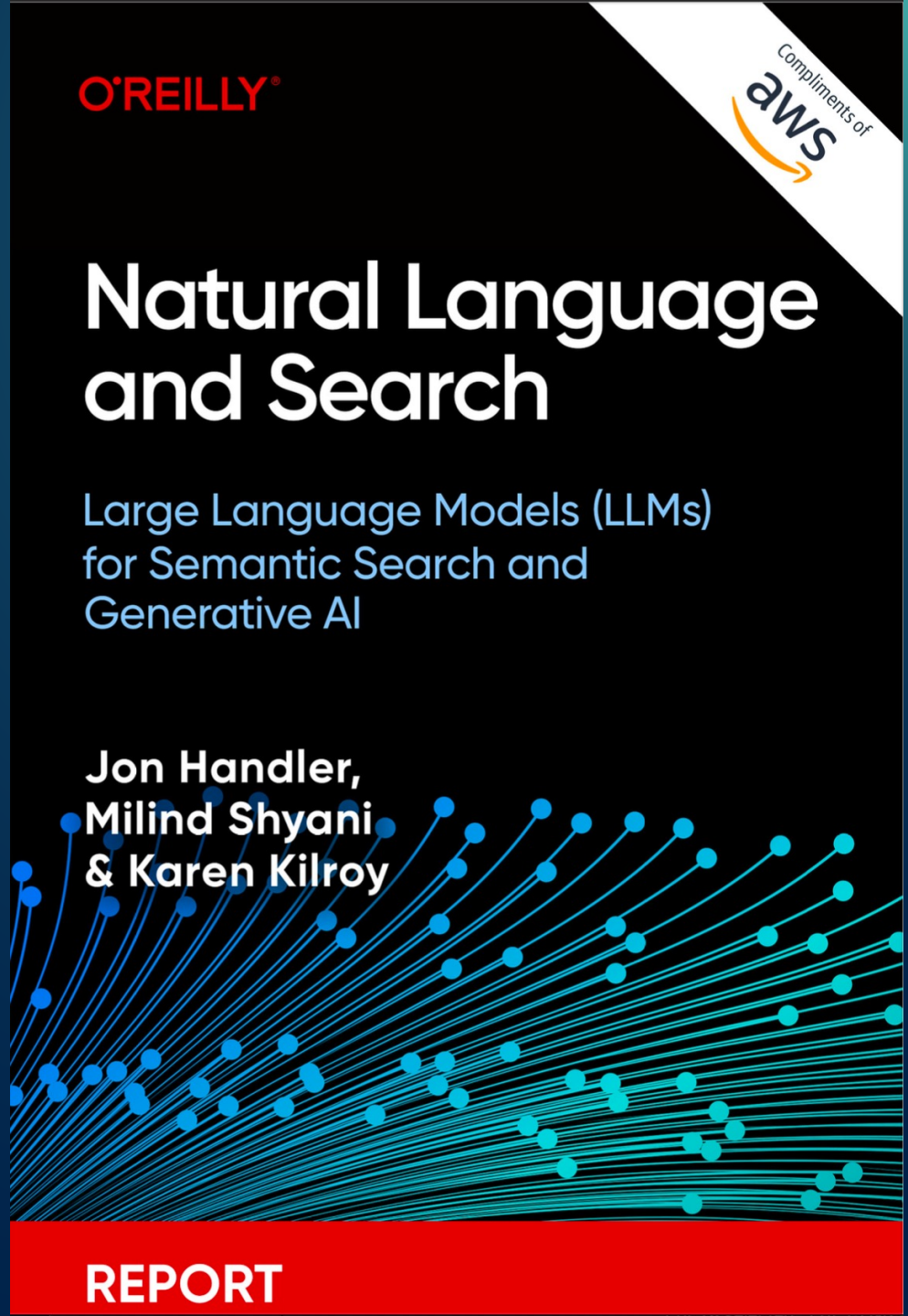
Reranking selection based on past user behavior and other factors



OpenSearch

https://d1.awsstatic.com/aws-analytics-content/OReilly_book_Natural-Language-and-Search_web.pdf

OpenSearch



Wrap

It was always about language – language to record information and language to find information

Improvements in language processing have enabled the capture and generation of language

How you interact with information in your enterprise, and in the world, mediated by technology is changing