Motivation

Why ASIN?

📄 <u>ProductQA Traffic Investigation</u>

- Current retrieval system
  - ○ Some answers found from top-10 ASINs
- Rewriting questions to keyword queries
  - ○ More can be answered by ASINs

Motivation

ASIN is a good resource to answer product question

- What is the name of the boat curved cutting knife from Alaska?
  - Rewrite to: **Alaska Knife**
    - **ALASKAN ULU KNIFE:** Title of the ASIN (www.amazon.com/dp/B0006Z877C)

- What's the score available habanero pepper
  - Directly search using the proposed system
    - **100k-350k on the Scoville scale:** In the bullets of the ASIN (www.amazon.com/dp/B07SNGQHZ7)

Motivation

- **High precision**
  - Provide only frequently purchased/clicked ASINs

- **Naturally decide whether to answer based on ASIN**
  - When no result is similar enough to the PQ
    - → ASIN is not suitable to the provided question
    - → ASIN is not suitable to the provided question

- **Support Multimodality**
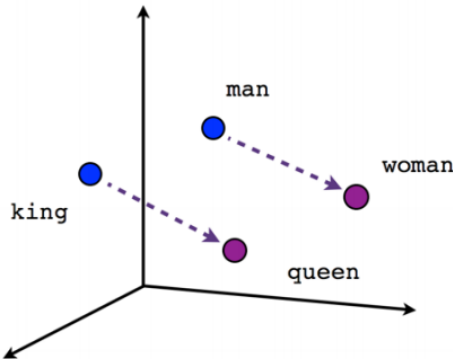 - Retrieving ASINs beyond the use of text fields

Motivation

Token Matching IR v.s. Semantic Search System
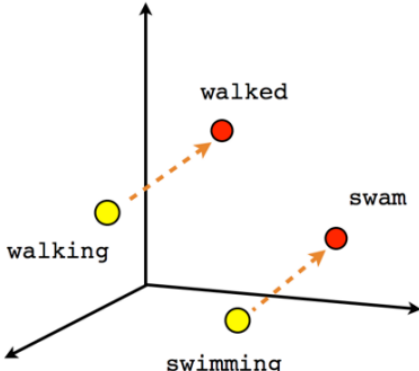
- **Token-based is unideal in retrieving ASINs**
  - A gap between ASINs and Alexa questions
    - Terminology:
      - Specific product (ASINs)
      - General questions (PQs)
    - Format:
      - Include lots of info (ASINs)
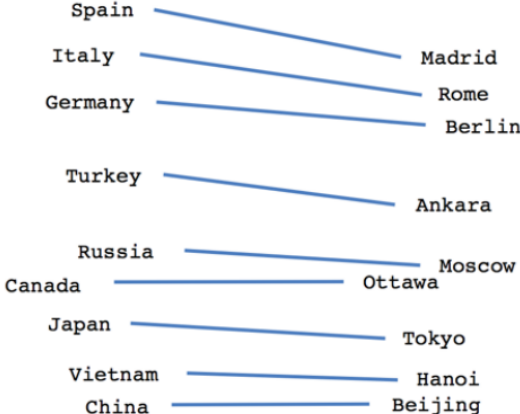   - Natural Language Questions (PQs)

Background —
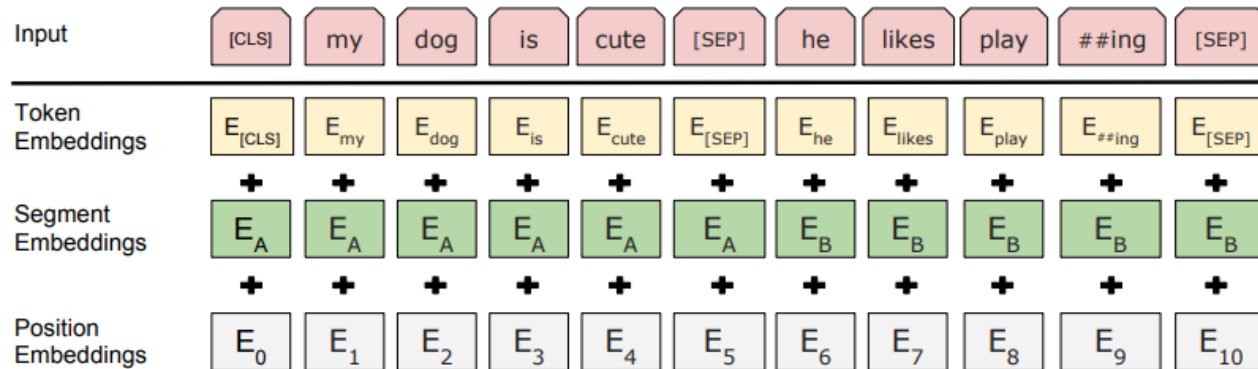# Dense Vectors for Semantic Search

Male-Female

Verb tense

Country-Capital

Background — Bert
Bidirectional Encoder Representations from Transformer

- BERT considers contextual meaning

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

- Experiment on [CLS] and average

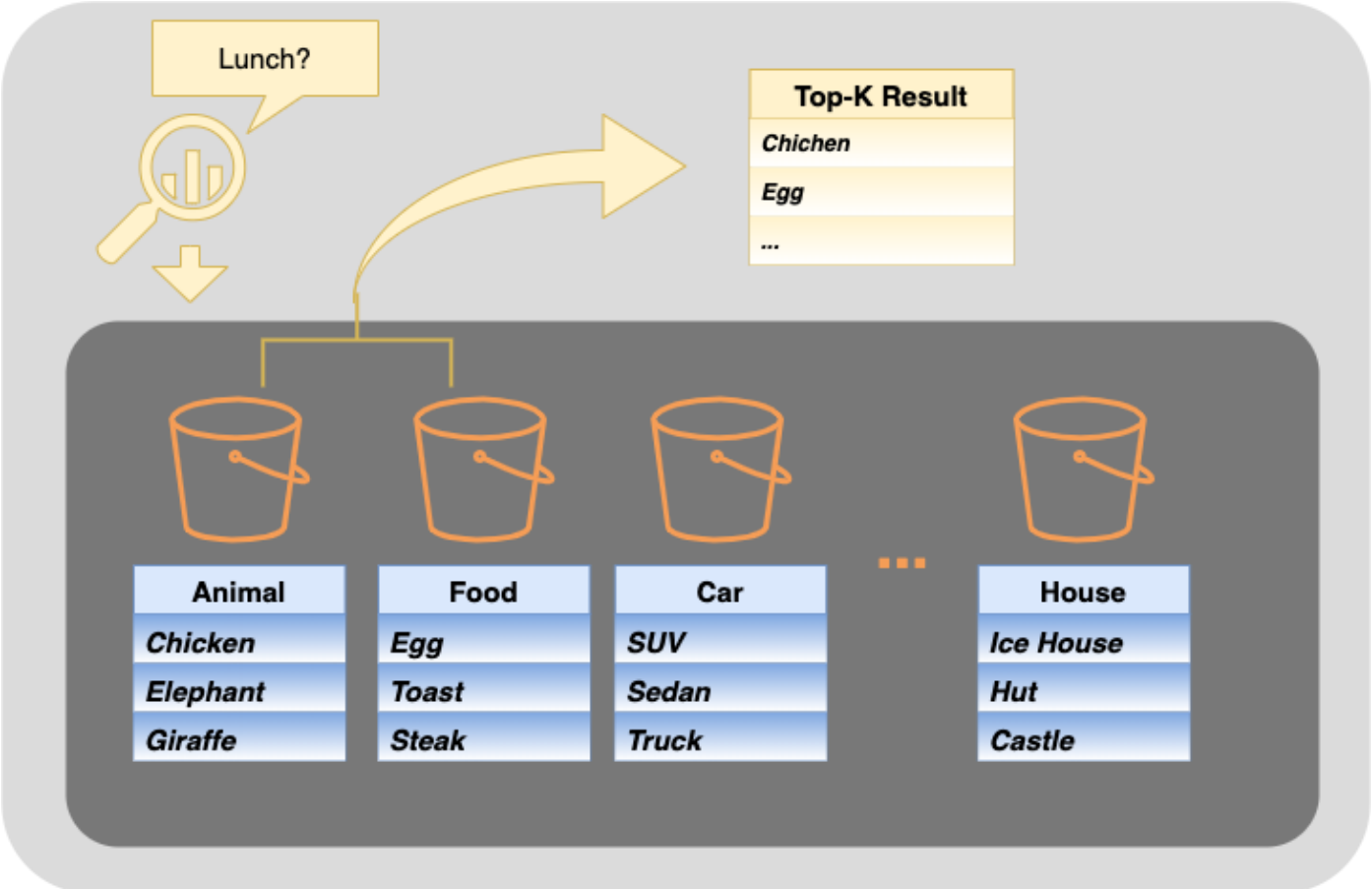**Use average representation from pre-trained BERT**

Background —
Locality-Sensitive Hashing (LSH)

- Similar input items in the same "buckets"

- Search visits some buckets among all buckets
  - **Reduce computation**

Background — Locality-Sensitive Hashing (LSH)

Background —
Facebook AI Similarity Search (Faiss)

- A package for efficient similarity search

- One of the implementations is LSH

- **"Bucket"** is called **"Voronoi cell"** in Faiss
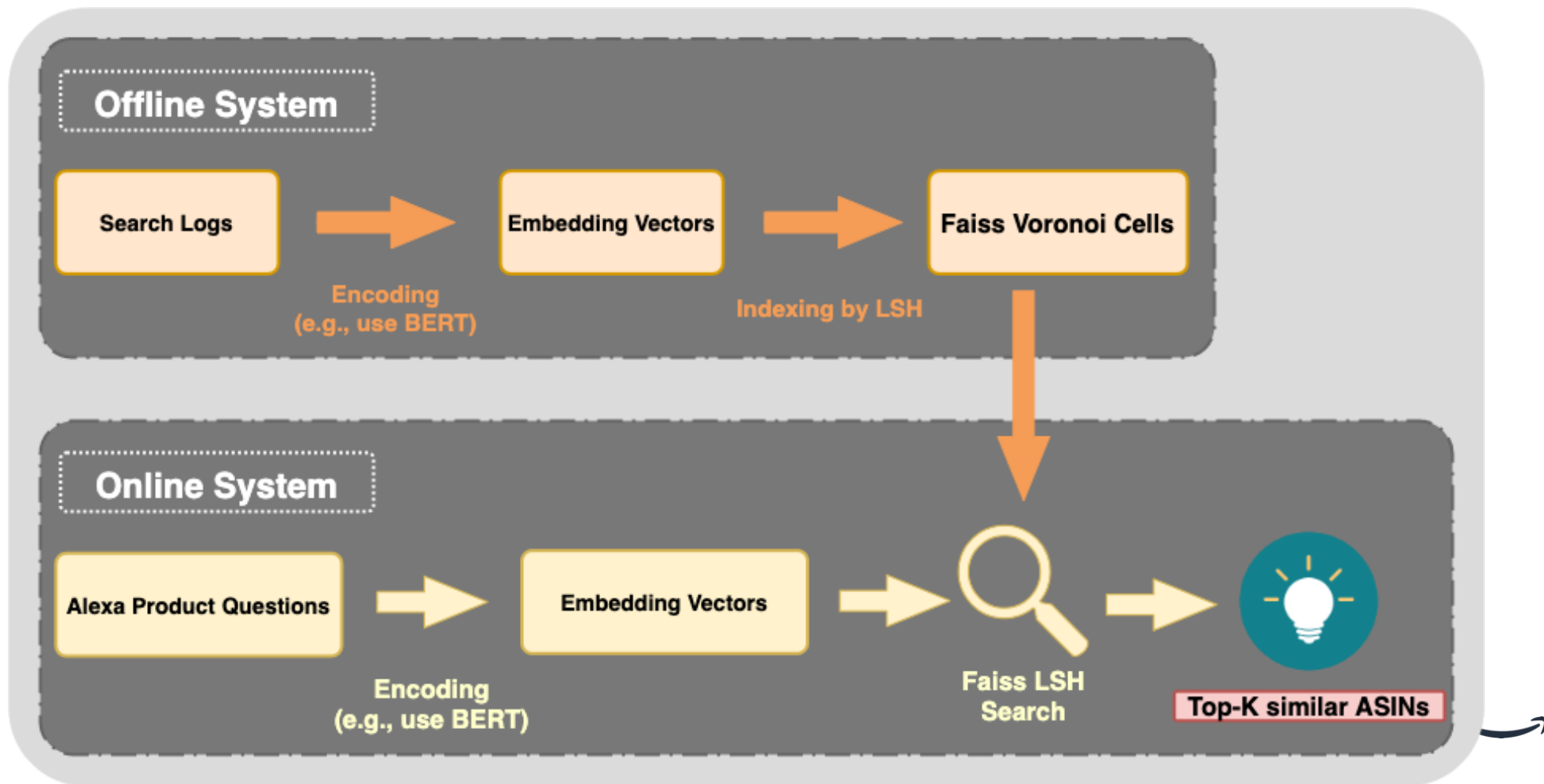
→ **We use Faiss to retrieve ASINs**

Overview

- **Two components in the ASIN based PQA system**
  - **ASIN Retrieval System**
  - ASIN Question Answering System

Method

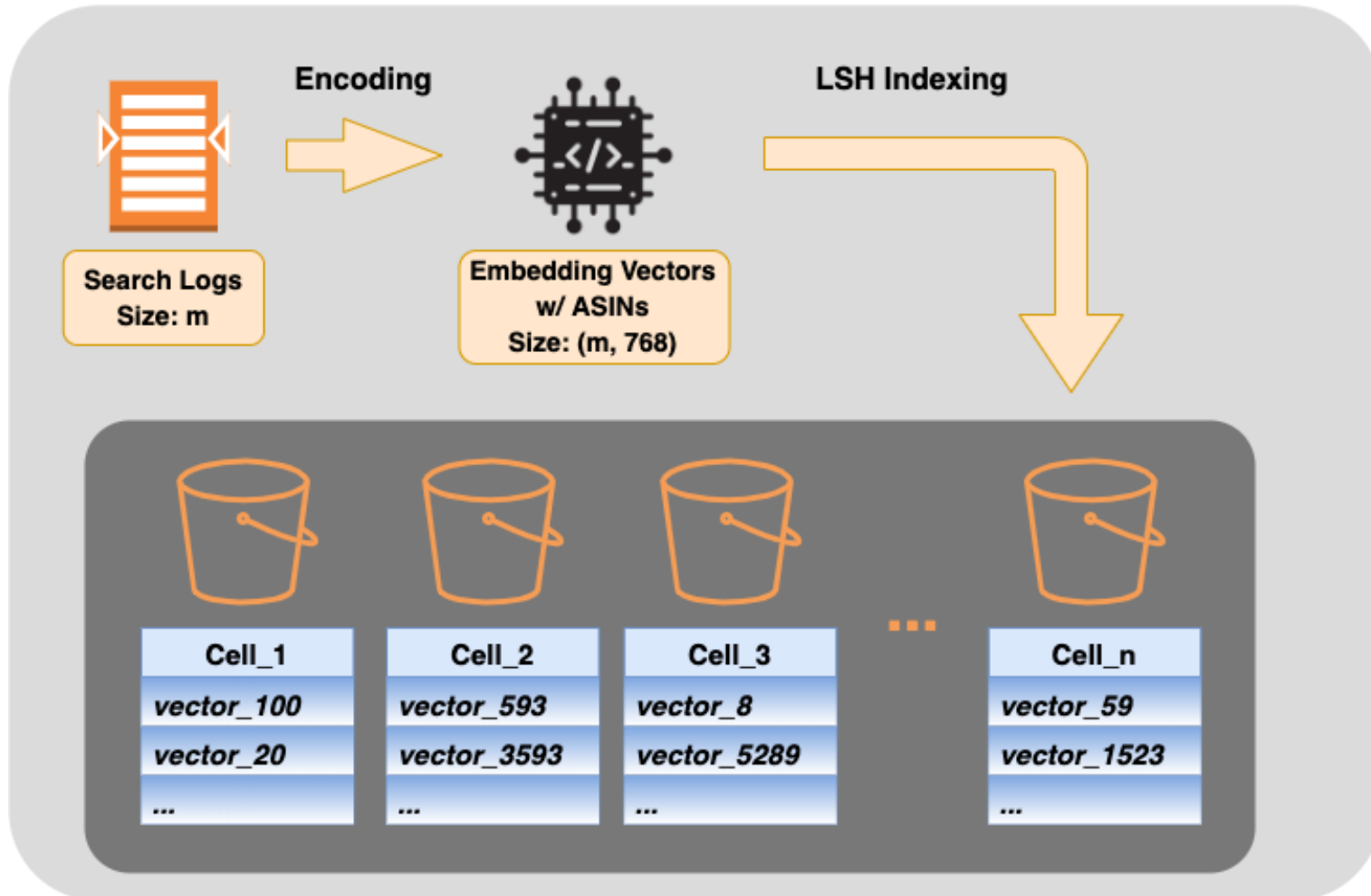**Offline System**

Search Logs → Embedding Vectors → Faiss Voronoi Cells

Encoding (e.g., use BERT)

Indexing by LSH

Online System

Alexa Product Questions → Embedding Vectors → Faiss LSH Search → Top-K similar ASINs

Encoding (e.g., use BERT)

Method

- **Two Implementations on Buckets**

Implementation	Index Loading	Inference	RAM Requirement
On Disk	Fast (mins)	Slow	Small (300 MB)
On RAM	Slow (hrs)	Fast	Huge (1.1TB)

Example — Offline System

Encoding

LSH Indexing

Search Logs
Size: m

Embedding Vectors
w/ ASINs
Size: (m, 768)

Cell_1
vector_100
vector_20
...

Cell_2
vector_593
vector_3593
...

Cell_3
vector_8
vector_5289
...

Cell_n
vector_59
vector_1523
...

Example — Online System

- Product Question:
  - What is the most up-to-date apple watch

- Embedding Vector
  - [0.286, -0.426, ..., -0.368, -0.978]
  - Dimension: 768

- Top-k similar ASIN(s)
  - B07XR5T85R
  - Apple Watch Series 5 (GPS, 44mm) - with Silver Aluminum Case with White Sport Band

Evaluation — Hardware

- **Offline Stage that indexes the (keyword, ASINs):**
  - Four p2.16xlarge
    - GPUs: 16
    - vCPUs: 64
    - RAM (GiB): 732

- **Online Stage that retrieves ASINs given a PQ:**
  - One x1e.32xlarge
    - vCPUs: 128
    - RAM (GiB): 3904

Evaluation — Setup

- **Pre-trained Models Selection**
  - Triplets: **<PQ, True Title, Random Title>**

  - Precision: $\dfrac{\sum \delta[L(E_{PQ},\ E_{true}) < L(E_{PQ},\ E_{random})]}{\sum \delta[PQ]}$, $\begin{cases} E: Embedding\ Vector \\ L: Distance \\ \delta: 1\ if\ true,\ else\ 0 \end{cases}$

Model	Version	Precision
RoBERTa	roberta-base	80%
ELMo	small	82%
BERT	**bert-base-uncased**	**90%**
BERT	bert-base-multilingual-uncased	85%

Evaluation — Setup

- **Efficiency:**
  - Avg. time to retrieve top-5 ASINs for 923K questions

    (AmazonQA)

- **Effectiveness:**
  - Manual annotations (explained in the following slides)
  - 492 non-critical PQs
  - Only consider text fields in an ASIN
  - Relevance/answerability of top-1 retrieved ASIN per PQ

Evaluation — Annotation, Relevance

Query: Are sun dried tomatoes black?

Retrieved ASIN: B074MG6VGH

365 Everyday Value, Sundried Tomatoes, 4 oz

by Brand: 365 by Whole Foods Market

★★★★☆ ∨   69 ratings

Currently unavailable.

We don't know when or if this item will be back in stock.

- Brought to you by Whole Foods Market. The packaging for this product has a fresh new look. During this transition, you may get the original packaging or the new packaging in your order, but the product and quality is staying exactly the same. Enjoy!
- Made without Sulfites.
- Fat Free, No Sugar Added.

Evaluation — Annotation, Answerable

Query: Are Triscuits crackers?

Retrieved ASIN: B074XMX17J

Triscuit Dill, Sea Salt & Olive Oil Whole Grain Wheat Crackers, 8.5 oz

by Visit the Triscuit Store

⭐⭐⭐⭐⯪ ⌄   324 ratings

Price: **$2.99** ($0.35 / Ounce) Pantry orders are delivered for $5.99 and do not qualify for FREE Delivery. Learn more

pantry

Pantry orders of $35 and up ship for free- a $5.99 shipping fee applies to orders under $35. Due to high demand, delivery may take longer than usual. Learn more

In Stock.

Ships from and sold by Amazon Pantry.

Want it Thursday, Sep 10 ? Order within the next **19 hrs and 2 mins** Details

- One 8.5 oz box of Triscuit Dill, Sea Salt & Olive Oil Whole Grain Wheat Crackers
- Dill, sea salt and olive oil flavors complement the whole grain wheat base
- Square woven crackers offer a classic, refined snacking option
- Baked, wholesome sea salt crackers with a hearty woven texture
- Savor the taste by itself or top with cheese or spreads

Results — Index Construction

- **Time and Space Consumption for Index Construction**
  - Build the index per quarter before merging

	Q1	Q2	Q3	Q4	On Disk	On RAM
#Q	78M	80M	104M	115M	378M	
Time (hr)	21.5	23.9	31.2	34.36	**34.36**	**110.96**
Size(GB)	224.5	231.5	299.2	330.1	1085.3	

Results — Efficiency under Two Configurations

- **Percent of Overall Data Touched / Search**
  - 16 Buckets / 65536 Buckets → **0.02%**

- **Runtime (underestimated**)**

Type	Include Index Loading	Top-K	ms per question
On Disk	No		**68**
	Yes		69
On RAM	No	5	**57**
	Yes		65

Results — Efficiency under Two Configurations

- **Why is it underestimated**
  - Over 923K ASINs accumulated
  - Retrieved ASINs are stored on a list
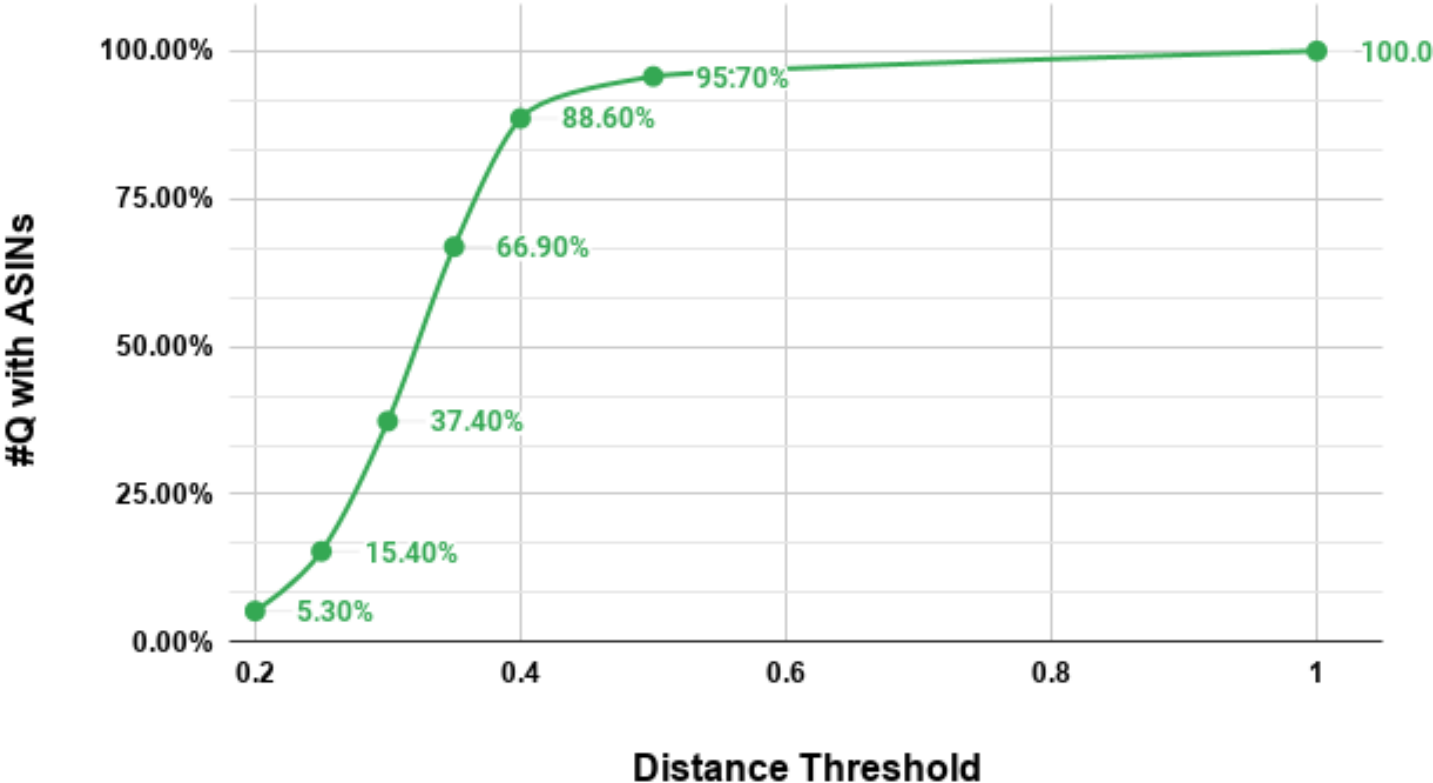 - Time consuming to maintain the list

Results — Effectiveness

- Experimented on different cut-off thresholds for distance

- The number of PQs have relevant retrieved ASINs @top-1

- The number of PQs have answerable retrieved ASINs @top-1

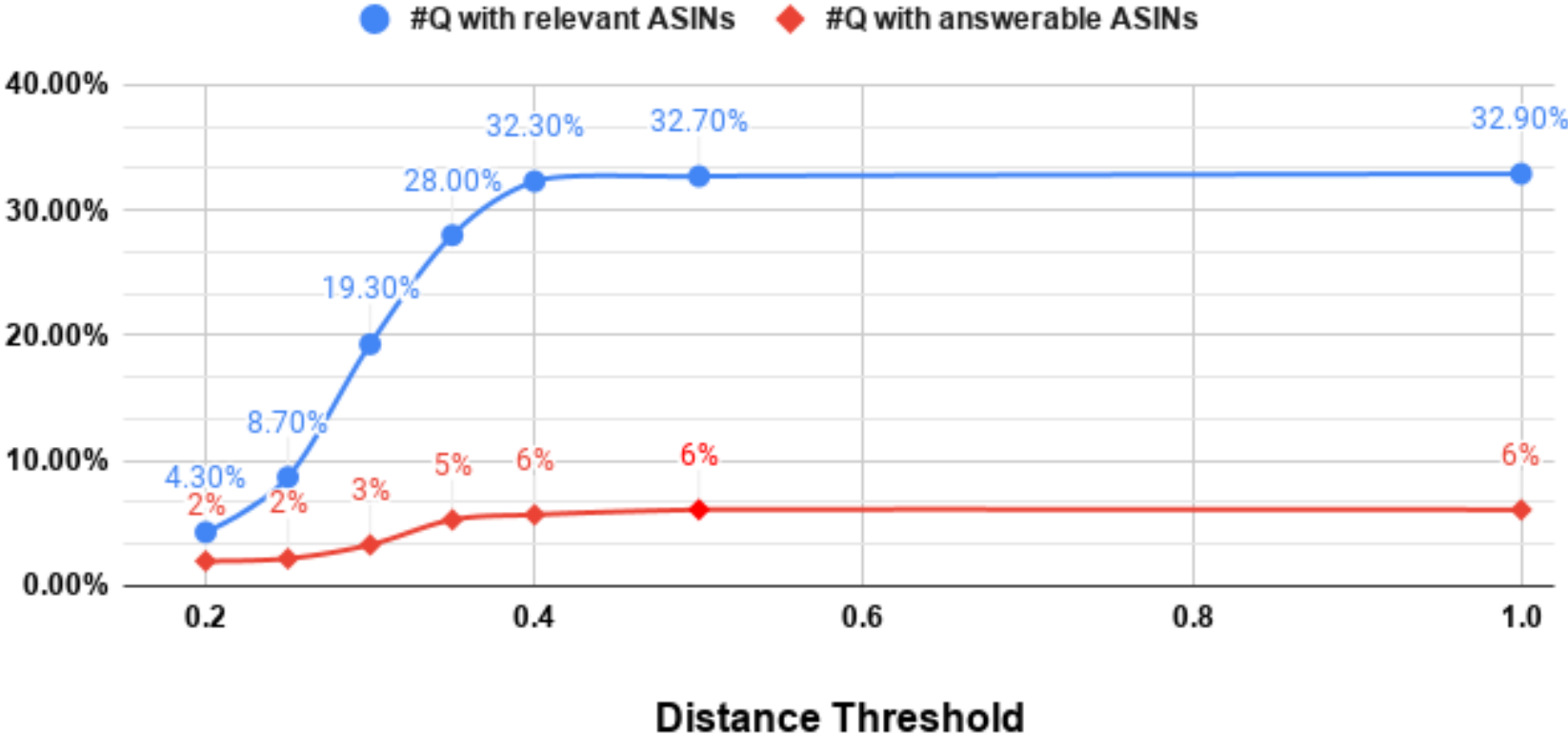| Query	Retrieved ASINs
Query 1	B007ZEMGS8 **(0.12)**, B078GZDJMH **(0.16)**, ~~B00UW06EVW~~ **~~(0.27)~~**
Query 2	~~B07RYB9G7Z~~ **~~(0.25)~~**
Query 3	B00LB4REF8 **(0.03)**, B0000EXU0S **(0.11)**

Results — Effectiveness

#Q with ASINs vs. Distance Threshold

Results — Effectiveness

#Q with Relevant/Answerable ASINs vs. Distance Threshold

Results – Effectiveness

	Distance Threshold	#Q with ASINs	#Q with relevant ASINs	#Q with ASINs including an answer	#Q including answer/#Q with ASIN
1	< 0.2	26 (5.3%)	21 (4.3%)	10 (2.0%)	38.5%
2	< 0.25	76 (15.4%)	43 (8.7%)	11 (2.2%)	14.5%
3	< 0.3	184 (37.4%)	95 (19.3%)	16 (3.3%)	8.7%
4	< 0.35	329 (66.9%)	138 (28%)	26 (5.3%)	7.9%
5	< 0.4	436 (88.6%)	159 (32.3%)	28 (5.7%)	6.4%
6	< 0.5	471 (95.7%)	161 (32.7%)	30 (6.1%)	6.4%
7	N/A	492 (100%)	162 (32.9%)	30 (6.1%)	6.1%

Result — Remaining Issues

- **QA system works only if ASINs include answers**
  - Some questions are not suitable to use ASIN as answer source
    - E.g. "Can my dog eat Cheetos"

- **Issue with our index on Amazom.com 2019**
  - Some ASINs may no longer exist

Future Work

- **Apply big data system** (e.g. Spark)

  ○ We are using 1-year data, there might be 10-year data
- **Collect product info related product questions**
  ○ Required to determine threshold
- **Some optimisations to further speed up**
  ○ Current runtime is ok, but preferably to be faster
  ○ Re-estimate the runtime
- **Encode PQs & keywords with supervised learning**
  ○ Current training dataset:

  ▪ BookCorpus
  English Wikipedia