

A decorative graphic in the top-left corner consisting of a network of interconnected nodes and lines. Some nodes are solid blue circles, while others are white circles with blue outlines. The lines are thin and grey.

# **A Practical Introduction to Vector Search**

**Andy Yun**

sqlbek@gmail.com

*Welcome! I'm  
Sebastian!*





# Andy Yun

*Field Solution Architect*

- SQL Server DBA & DB Developer
- SQL Server 2014



Group (n)  
– Director-at-Large  
– Organizer



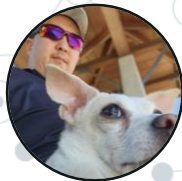
[sqlbek@gmail.com](mailto:sqlbek@gmail.com)

<https://sqlbek.wordpress.com/>

<https://www.github.com/sqlbek/>



What do you want for dinner?



Uhh, how about a stew or noodle soup?

Sure. We have chicken thighs & carrots we need to use.



Don't forget, Sebastian has a doggie play date at 7PM. So we need to eat before then.

# What's for D

USE RecipesDemoDB  
GO

SELECT TOP 100 \*  
FROM dbo.recipes

Don't y'all  
work with

	recipe_id	name	description	ingredients	ingredients_raw	steps	servings	serving_size	tags
1	38	Low-Fat Berry Blue Frozen Dessert	This is yummy and low-fat, it always turns out pe...	["blueberries", "granulated sugar", "vanilla yogurt", ...	["4 cups blueberries, fresh or frozen ","1/4 c...	["Toss 2 cups berries with sugar.", "Let stand for 45 ...	4.0	1 (225 g)	["weeknight", "time-to-make", "course", "preparat...
2	39	Biryani	Delhi, India	["safron", "milk", "hot green chili peppers serranos...	["1 tablespoon safron","4 teaspoons milk...	["Soak safron in warm milk for 5 minutes and pur...	6.0	1 (799 g)	["weeknight", "time-to-make", "course", "main-in...
3	40	Best Lemonade	This is from one of my first Good House Keeping...	["lemons lemon white rind", "fresh water", "fresh le...	["1 1/2 cups sugar","1 tablespoon lemons, ...	["Into a 1 quart Jar with tight fitting lid, put sugar a...	4.0	1 (212 g)	["60-minutes-or-less", "time-to-make", "course", "...
4	41	Carina's Tofu-Vegetable Kebabs	This dish is best prepared a day in advance to allo...	["firm tofu", "eggplant", "small zucchini", "red pepp...	["12 ounces extra firm tofu, water-packed ","...	["Drain the tofu, carefully squeezing out excess wat...	2.0	1 (932 g)	["weeknight", "time-to-make", "course", "main-in...
5	43	Best Blackbottom Pie	Sweet, chocolatey, yummy	["graham cracker crumbs", "butter", "milk", "egg yo...	["1 1/4 cups graham cracker crumbs","1/4 cu...	["Graham Cracker Crust: In small bowl, combine gr...	8.0	1 (171 g)	["weeknight", "time-to-make", "course", "cuisine", ...
6	44	Chicken a la King	I copied this one out of a friend's book so many ...	["chicken", "flour", "celery", "button mushrooms", "...	["12 ounces chicken, cooked (in fairly large c...	["Melt 1 1/2 ozs butter, add the flour and cook for ...	2.0	1 (634 g)	["60-minutes-or-less", "time-to-make", "course", "...
7	45	Buttermilk Pie With Gingersnap Crumb Crust	Yum	["margarine", "egg whites", "flour", "buttermilk", "g...	["3/4 cup sugar","1 tablespoon margarine"]...	["Preheat oven to 350°F.", "Make pie crust, using 8 ...	8.0	1 (91 g)	["weeknight", "time-to-make", "course", "main-in...
8	46	A Jad - Cucumber Pickle	It is a traditional accompaniment to snacks such a...	["graham cracker crumbs", "butter", "milk", "egg yo...	["1/2 cup rice vinegar","5 thangkwa (cucu...	["Slice the cucumber in four lengthwise, then slice t...	1.0	1 (59 g)	["30-minutes-or-less", "time-to-make", "course", "...
9	49	Chicken Breasts Lombardi	Cheese, Chicken, and Mushrooms with Marsala.	["fresh mushrooms", "butter", "boneless skinless ch...	["2 cups fresh mushrooms, sliced ","2 tables...	["Cook mushrooms in 2 tbsp butter in a large skill...	6.0	1 (477 g)	["weeknight", "time-to-make", "course", "main-in...
10	52	Cafe Cappuccino	NULL	[]	["1/2 cup instant coffee","3/4 cup sugar","1 ...	["Stir ingredients together.", "Process in a blender u...	18.0	1 (16 g)	["15-minutes-or-less", "time-to-make", "course", "...
11	54	Carrot Cake	This is one of the few recipes my husband every ...	["carrots", "vegetable oil", "white sugar", "all - purp...	["3 cups carrots, grated ","4 eggs","1 1/4 ...	["Beat together the eggs, oil, and white sugar. Blen...	12.0	1 (161 g)	["30-minutes-or-less", "time-to-make", "course", "...
12	56	Buttermilk Pie	This recipe was originally noted by my wife on a c...	["butter margarine", "flour", "salt", "vanilla", "butter...	["1/2 cup butter or 1/2 cup margarine, melt...	["Preheat oven to 400°F.", "Beat the butter and sug...	8.0	1 (123 g)	["time-to-make", "course", "main-ingredient", "cul...
13	58	Low-Fat Burgundy Beef & Vegetable Stew	(high fiber)	["vegetable oil", "dried thyme leaves", "beef broth", ...	["1 1/2 lbs beef eye round","1 tablespoon ...	["Trim fat from beef, cut into 1-inch pieces.", "In Du...	6.0	1 (368 g)	["weeknight", "time-to-make", "course", "main-in...
14	59	Lou's Fabulous Bruschetta	this one is different!	["French baguette", "butter", "garlic powder", "ricott...	["1 French baguette, sliced 1/4-1/2 inch thic...	["Cut baguette into slices.", "Butter and then sprinkl...	8.0	1 (94 g)	["60-minutes-or-less", "time-to-make", "course", "...
15	62	Black Bean, Corn, and Tomato Salad	This is easy, delicious, colorful, delicious, uses cur...	["fresh lemon juice", "olive oil", "black beans", "fres...	["3 tablespoons fresh lemon juice","2 tables...	["In a bowl whisk together lemon juice, oil, and salt...	2.0	1 (319 g)	["30-minutes-or-less", "time-to-make", "course", "...
16	64	Almond Pound Cake	--Adopted Recipe--	["butter", "almond paste", "cornstarch", "baking po...	["2/3 cup butter, softened ","3 1/2 ounces a...	["Preheat oven to 350 degrees Fahrenheit", "Crea...	10.0	1 (115 g)	["weeknight", "time-to-make", "course", "main-in...
17	66	Black Coffee Barbecue Sauce	It's great to know folks like this sauce so much! I ...	["brewed coffee espresso", "ketchup", "red wine vin...	["1/2 cup brewed coffee, espresso preferred "]...	["Combine all ingredients in a saucepan and simm...	1.0	1 (1401 g)	["lactose", "30-minutes-or-less", "time-to-make", ...
18	67	Bourbon Pecan Pound Cake	NULL	[]	["1/2 lb butter","2 1/2 cups sugar","6 eg...	["Combine butter and sugar in bowl of electric mix...	12.0	1 (186 g)	["weeknight", "time-to-make", "course", "main-in...
19	68	Chicago Style Pizza	NULL	[]	["1 (1/4 ounce) package dry yeast","1 1/4 cup...	["For crust, dissolve yeast in water.", "Add sugar, s...	8.0	1 (174 g)	["weeknight", "time-to-make", "course", "cuisine", ...
20	69	Chicha Peruana	Chicha (corn beer). Chicha is made in South and, t...	["jora malted corn", "piloncillo cone brown sugar", ...	["1 1/2 lbs jora (malted corn)","1 lb piloncill...	["Procedure: Mash for 90 minutes at 160°F.", "We d...	1.0	1 (7 g)	["weeknight", "time-to-make", "course", "cuisine", ...
21	71	Chicken and Dumplings	(Welk)	["carrot", "celery", "chicken bouillon cubes", "flour", ...	["4 lbs chicken","1 carrot","2 stalks cele...	["Place chickens in large saucepan, cover with wat...	8.0	1 (333 g)	["weeknight", "time-to-make", "course", "main-in...
22	73	Bratwurst	Although they can be made with a fine grind, we ...	["veal", "pork shoulder", "milk", "white pepper", "gi...	["3 lbs veal, trimmed ","7 lbs pork shoulde...	["Grind the veal and pork with a 3/8 in. (0.95 cm) ...	1.0	1 (6047 g)	["weeknight", "time-to-make", "course", "preparat...
23	74	Brownie Cheesecake Torte	From the Zaar Adopt-a-Recipe Program, have not...	["low - fat fudge brownie mix", "instant coffee gran...	["15 1/4 ounces low-fat fudge brownie mix (1...	["Preheat oven to 425 degrees F.", "Combine first 4...	12.0	1 (87 g)	["weeknight", "time-to-make", "course", "preparat...
24	75	California Chilled Salsa	NULL	[]	["2 cups tomatoes, peeled and chopped ","1 ...	["Also delicious made with red sweet peppers or ...	10.0	1 (71 g)	["30-minutes-or-less", "time-to-make", "course", "...
25	76	Alfred's Sauce	This is my son's favorite meal. I make it with chri...	["cousin butter butter", "heavy cream light cream", ...	["1/4 lb. cousin butter (1/2 cup), light butter, ...	["Place butter in microwave safe pot and heat on hi...	4.0	1 (107 g)	["15-minutes-or-less", "time-to-make", "course", "...

# Querying for a Recipe

```
SELECT recipe_id,  
       name, ingredients, tags  
FROM dbo.recipes  
WHERE (
```

```
EXEC dbo.sp_search_recipes  
    N'Give me some chicken stew or  
    noodle recipes with carrots that  
    I can cook in less than an hour';
```

```
)  
AND (  
    tags LIKE '%60-minutes-or-less%'  
    OR tags LIKE '%main-dish%'  
    OR tags LIKE '%weeknight%'  
)
```



# Agenda

- ◎ Fundamentals of Vector Search
- ◎ SQL Server as a Vector Store (CTP 2.1)
- ◎ Demo!!!
- ◎ ***A Note from Today's Sponsor: Pure Storage***
- ◎ Practical Possibilities
- ◎ *Warning: Fast Moving Crash Course*

*I'm fast  
too!*



Chapter 1:

# Fundamentals of Vector Search

*Are you going  
to mention  
that buzzword  
yet?*



# What is AI Today?

- ◎ Tools that enable...
  - Data Insights
  - Enhanced Decision-Making
  - Forecasting
  - Human-like Interaction

*What does  
AI really  
mean?!*





# Unraveling Components of AI

- ◎ Large Language Model (LLM)
- ◎ Embedding Model
- ◎ Vector Embedding
- ◎ KNN and ANN Search
- ◎ RAG Architecture

# Large Language Models

- ◎ OpenAI - *aka “ChatGPT”*
  - GPT-3, GPT-3.5, ...
- ◎ Anthropic
  - Claude 1 / 2 / 3, ...
- ◎ Google
  - Gemini 2.0 Flash, 2.5 Pro, ...
- ◎ Meta (Facebook) – LLaMA
- ◎ DeepSeek

# What are Large Language Models

- ◎ A glorified pattern recognition system to understand “language”
- ◎ Predict next word based on context
- ◎ Pre-trained on text data
- ◎ Good for summarizing information

# LLM Limitations

- ◎ Finite knowledge
  - Training cut-off
- ◎ No real-time data access
- ◎ Does NOT create new ideas
- ◎ Context Window Challenge...
  - Hallucination?

# Context Window

- ◎ ... includes:
  - Prompt
  - Dialogue history
  - Retrieved information
- ◎ Akin to available working memory

# Context Window Measurement

## ⦿ Measured in Tokens:

- 1 English word = ~1.5 tokens
- Page of text = ~500-750 tokens

## ⦿ Model Capacity

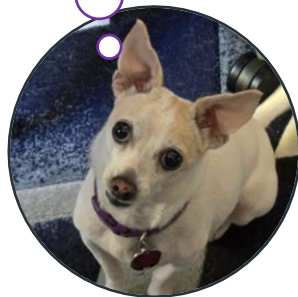
- Early models = ~4k - ~32k tokens
- Latest models = ~128k - 200k tokens
- Cutting edge models = ~1m - ~100m tokens!!!



# Context Window Challenge

- ◎ Fill up the bucket?
- ◎ Truncation - First In, First Out
  - I forgot...
- ◎ LLM Priority: Coherent & contextually appropriate response
  - “I don’t know...?”
  - Fills in gaps with nonsense
  - Hallucinations

*No, you  
didn't give  
me a treat 5  
minutes ago*



# Vector Search

- ◎ We can't scan and process EVERYTHING
- ◎ Intelligent pre-filtering
  1. Vector Embedding
  2. Embedding Models

# Embedding Models

## ◎ Semantic Similarity:

- “How do I reset my password?”
- “Steps to recover account access”
- “I forgot my password, what do I do?”

## ◎ Contextual Understanding:

- “That’s the **right** answer”
- “Turn **right** at the next corner”

# Vector Embedding

- ◎ Embedding = sequence of floating point numbers
- ◎ Vector = a position in space
  - 2D Vector  $(x, y) = [3, 4]$
  - 3D Vector  $(x, y, z) = [3, 5, -1]$
- ◎ ex: 5 dimensional vector embedding  
 $[93.124, 395.0423, 0.123, -2.3404, 53.23498]$

# Vector Store

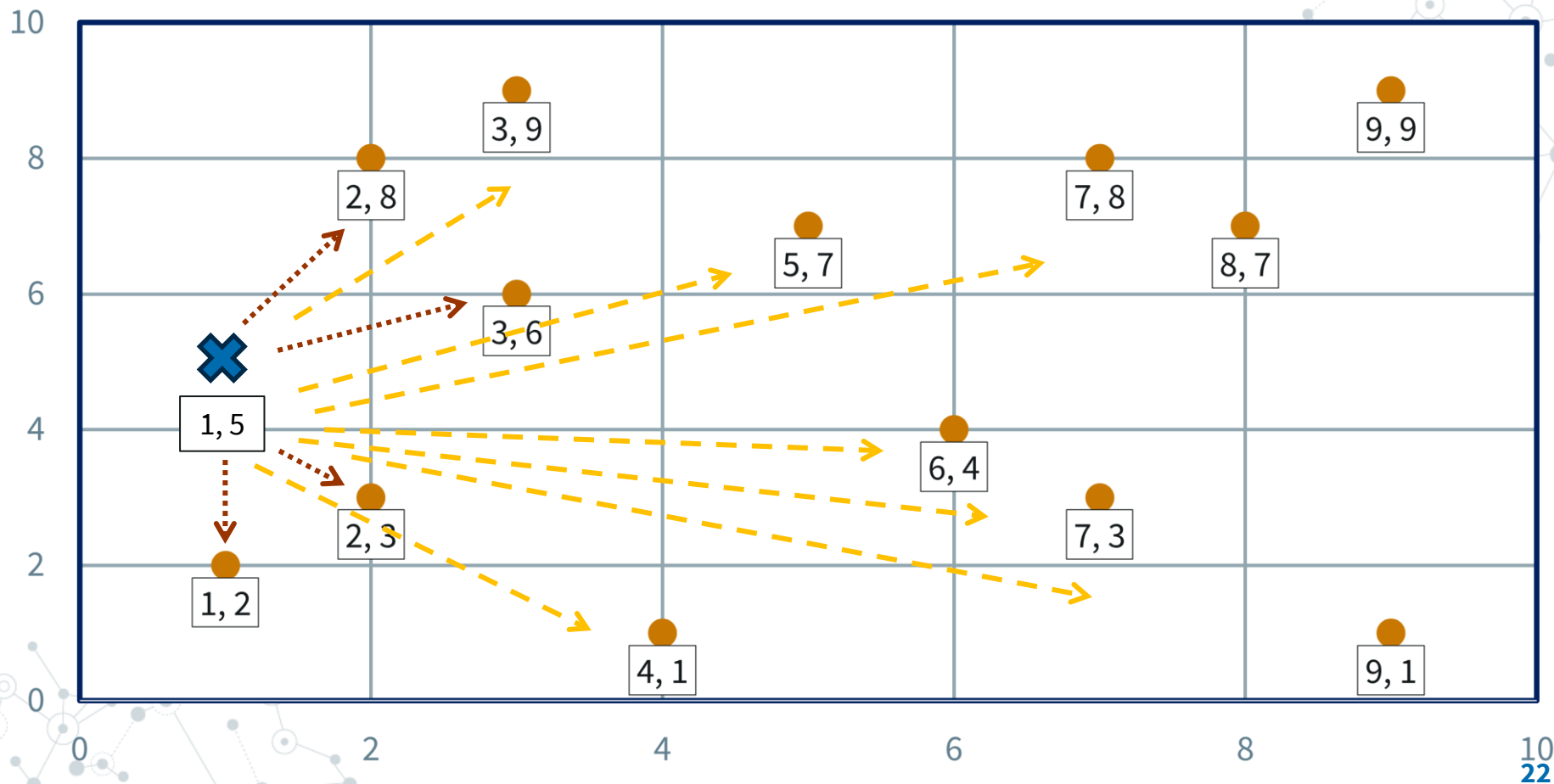
- ◎ Vector Search Embeddings:
  - 768 & 1536 dimensions (typical)
- ◎ Stored in a database
- ◎ Was a dedicated Vector DB...

# KNN & ANN Searching

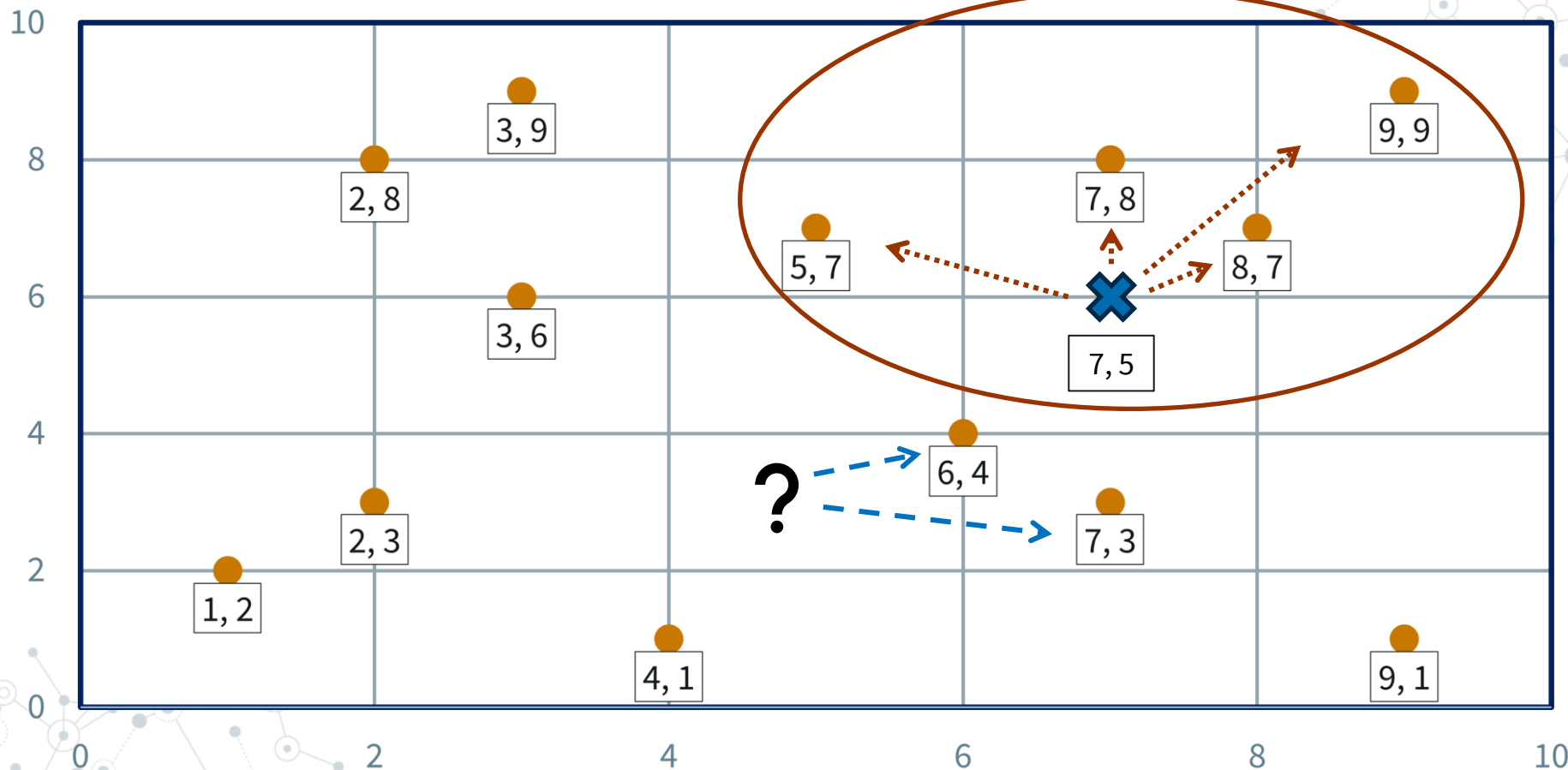
- ◎ Nearest Neighbor search
    - Top N
  - ◎ KNN = k-Nearest Neighbor
    - “exact”
    - Calculates distance against EVERY OTHER data point
  - ◎ ANN = Approximate Nearest Neighbor
    - “good enough”
- Multiple algorithms – ex: DiskANN



## 2D Vector Space – KNN Search

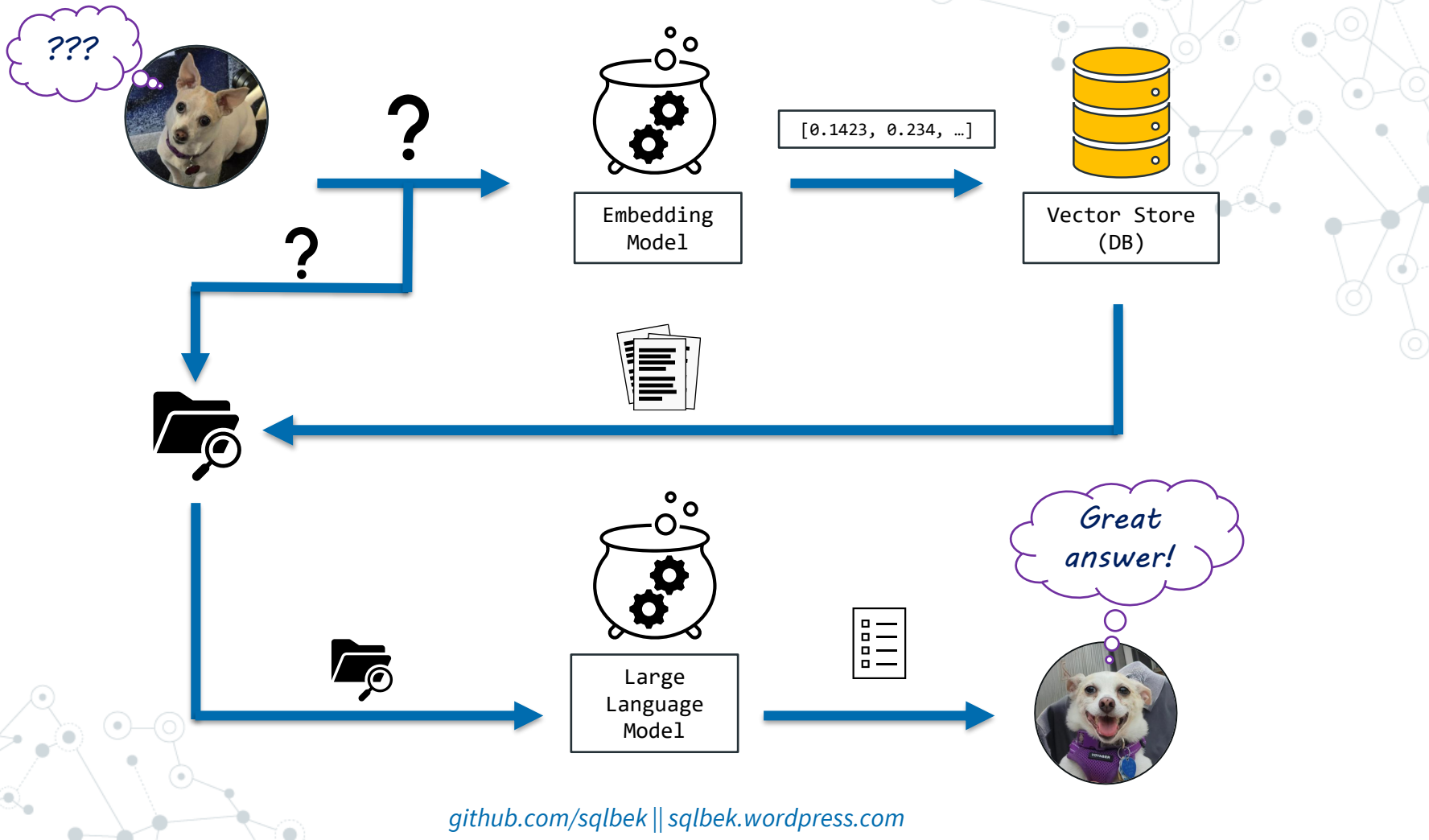


## 2D Vector Space – ANN Search



# RAG Architecture

- ◎ Retrieval Augmented Generation
  - “*Workflow*” instead of “*Architecture*”
  - Groups all of these components together
- ◎ Combines pre-trained knowledge of a LLM with external data



# Recap

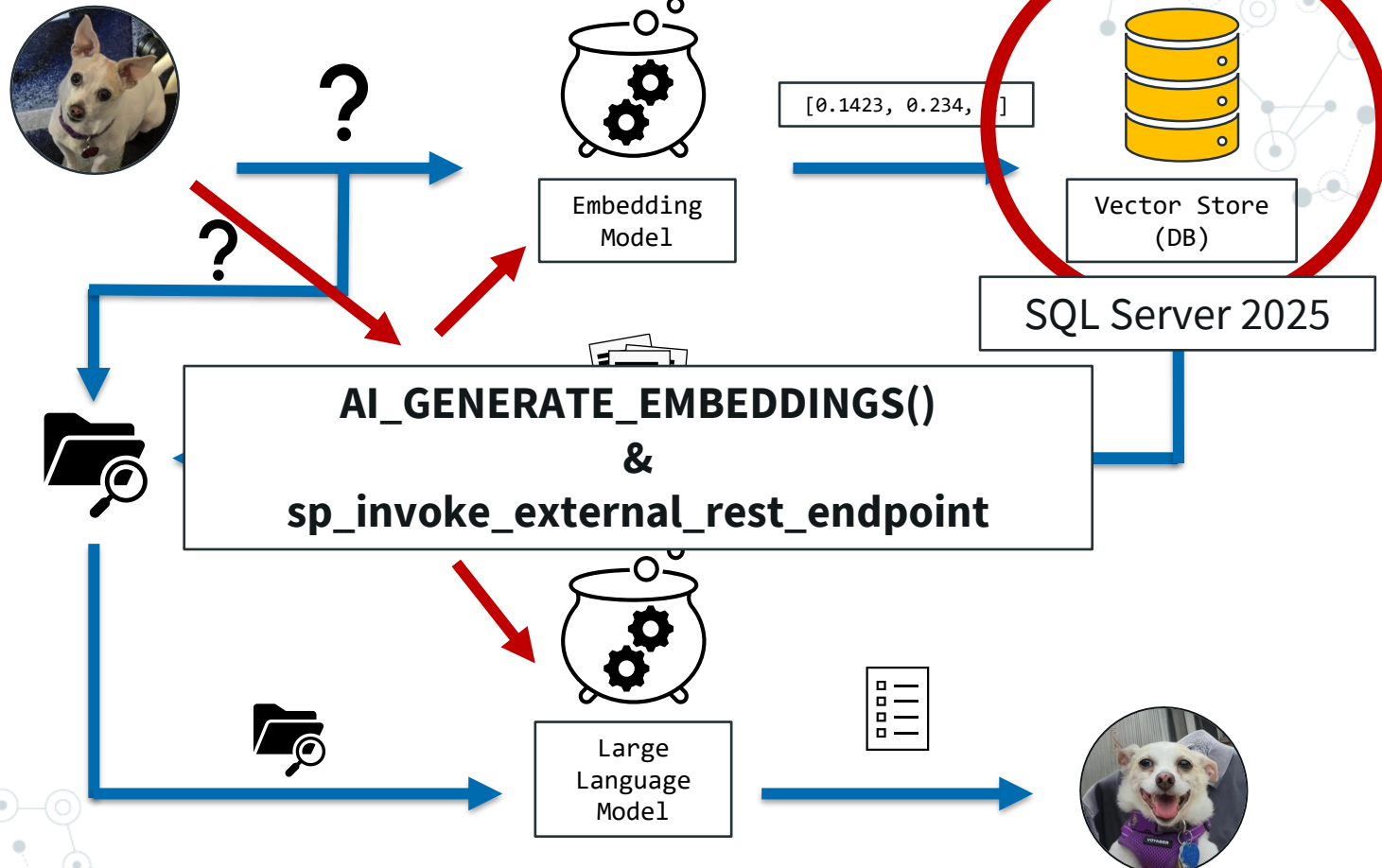
- ◎ Large Language Model (LLM)
- ◎ Embedding Model
- ◎ Vector Embedding
- ◎ KNN and ANN Search
- ◎ RAG Architecture

# Chapter 2: SQL Server as a Vector Store

*SSaaVS?*







# SQL Server 2025 as a Vector Store

## ◎ New Vector datatype

- Without datatype, would have to store vector data elsewhere –another Vector DB

## ◎ Vector Index - DiskANN algorithm

- VECTOR\_SEARCH() – for ANN
- VECTOR\_DISTANCE() – for KNN

# What does SQL Server NOT Provide?

- ◎ No embedding model or LLM supplied
  - YOU Choose!
  - Azure AI Foundry
  - Ollama – on-premise

# Why Build This Into SQL Server?

- ◎ Converged Data Platforms
- ◎ Transactional Engine vs Analytical Engine
  - Data Movement / Copy Data Management
- ◎ Reduce complexity

# Demo!!!

*Wait, share  
your setup  
first!*



# My Setup

- ◎ SQL Server 2025 CTP 2.1 + Ollama + nginx
  - Embedding Model: nomic\_embed\_text
  - LLM: llama3.2
- ◎ Quick Start Guide available on my blog
  - Search “sqlbek sql server 2025 ollama”

*NOW it's  
demo time!*





# Current State in SQL Server 2025

## ◎ Public Preview – CTP 2.1

- Remember how columnstore was on V1?

## ◎ Data must be “transformed” twice

- Create vector embeddings
- Create vector index

## ◎ Davide Mauri (*Principal Product Mgr*)

<https://devblogs.microsoft.com/azure-sql/database-and-ai-solutions-for-keeping-embeddings-updated/>  
<https://devblogs.microsoft.com/azure-sql/storing-querying-and-keeping-embeddings-updated-options-and-best-practices/>

# Price of Vector Search

- ◎ VECTOR(768) or VECTOR(1536)...
  - 768 or 1536 floating point numbers... EACH
- ◎ Regular Float(n) datatype = 4 or 8 bytes
- ◎ GPU intensive? Storage intensive!

# A Brief Word From Today's Sponsor



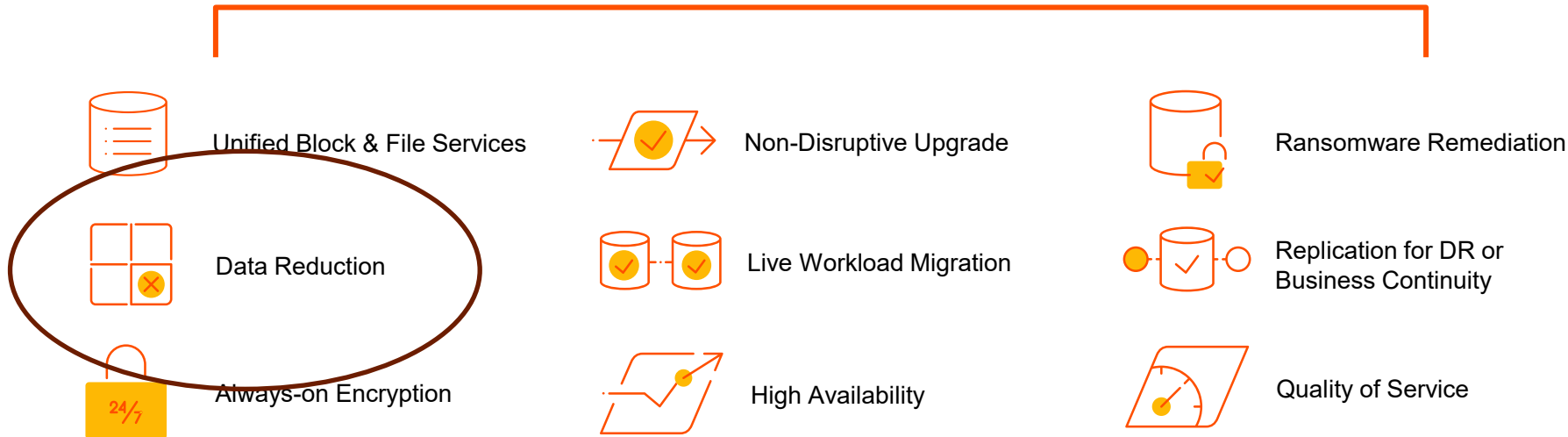
**PURESTORAGE®**

Uncomplicate Data Storage, Forever

# Pure Storage's Enterprise Data Services

Data Services That Transform And Accelerate Your Organization

Purity **FA**



# FlashArray Data Reduction (DRR)


- ◎ Compression + Deduplication
- ◎ 512 byte dedupe block size\*
- ◎ Globally across array – not just single volume


# RecipesDemoDB

	recipe_count		
1	492630		
	filegroup_name	used_space_GB	allocated_space_GB
1	PRIMARY	2.13916015625	2.83203125000
2	VECTOR_FG	22.96386718750	43.00341796875

Storage					
2	VECTOR_FG	22.96386718750	43.00341796875		

Array   Presets   Workloads   Hosts   Servers

 > Volumes > vol-ayun-sql25-02-b19e5b0-vg1 Vector Embeddings

Size	Virtual	Data Reduction	Unique	Snapshots	Total
5.00 T 	43.05 G	4.5 to 1	8.92 G	0.00	8.92 G

Data Reduction:  
2.57:1

# FlashArray 170XL-R5

## Results and Analysis

High-dimensional vector embeddings can occupy substantial storage space, making it difficult to manage capacity requirements and maintain operational efficiency. The table below shows the results of FlashArray 170XL-R5 data reduction on a 536GB vector data set.

Metric	Footprint
Logical Data Footprint Post Indexing	1.5TB
Physical Data Footprint	670GB
Data Reduction Ratio	2.4:1
Footprint Reduction	56.77%

WHITE PAPER

## Performance and Scale for

Data Reduction for Vector Embedding Dataset

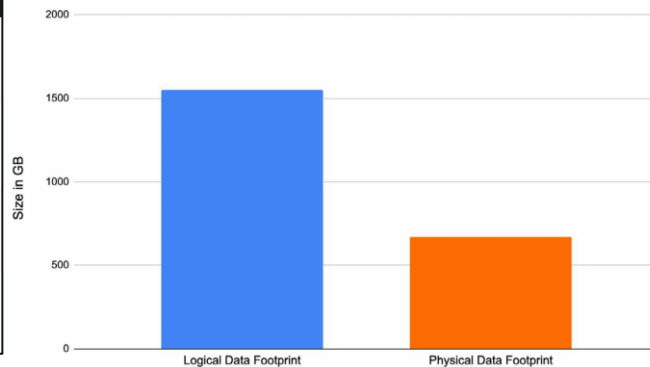


FIGURE 7 Vector embedding storage footprint before and after data reduction

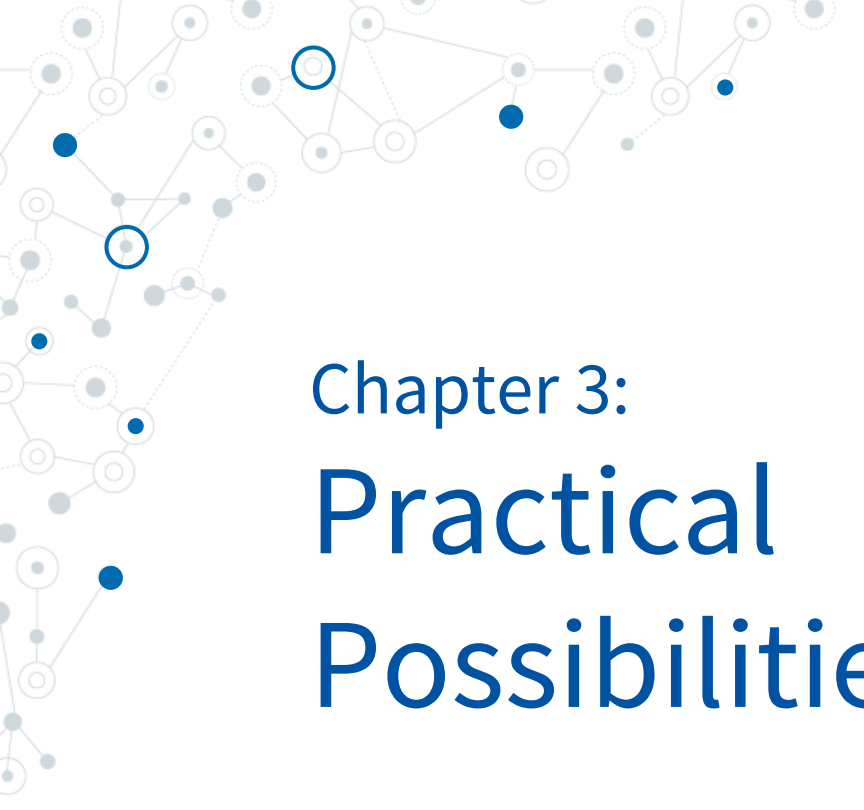
FlashArray//XL170-R5 applies inline compression and deduplication to reduce the data set's physical footprint by over half. A 2.4 to 1 data reduction ratio means that 2.4GB of logical data requires only 1GB of physical storage. This reduction results in a 56.77% decrease in capacity consumption.

# Learn More...


- 🕒 Webinar:  
Get Ready for SQL Server 2025:  
Accelerating AI, Performance, and Resilience  
—From Query to Infrastructure  
*w. Anthony Nocentino (Pure) & Dhananjay Mahajan (Microsoft)*

<https://www.purestorage.com/video/webinars/accelerating-ai-performance-and-resilience/6372781959112.html>



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with blue dots. The lines are thin and grey, creating a subtle background pattern.

# Chapter 3: Practical Possibilities

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a cluster of nodes connected by lines, with several nodes highlighted in blue. The overall style is clean and modern, with a focus on connectivity and structure.

# What Problem Does This Solve?

- ◎ Advanced Search in Applications
  - Semantic vs Keyword searching
- ◎ Anomaly Detection
- ◎ Data Summarization

# Static Datasets

- ◎ Product Manuals, Schematics, & FAQs
- ◎ Research Data & Reports
- ◎ Customer Service Guides & Knowledge Bases
- ◎ Historical Trade Data
- ◎ Structured + Unstructured Data

# Food for Thought

*What data in YOUR  
applications could  
benefit your business with  
RAG powered search  
capabilities?*



*Treat?*

# Conclusion

*Is it almost  
time for  
my walk?!*



# Recap

- ◎ Crash Course in RAG, LLMs, & Vector Search
- ◎ SQL Server as a Vector Store
- ◎ Practical Possibilities

# Parting Thought

*We search data  
with T-SQL every day...*

*consider...*

*how Vector Search  
adds a new dimension*



*Seriously?*

# Learn More: Resources

SQL Server 2025 + Ollama Quick Start Guide: Andy Yun

<https://sqlbek.wordpress.com/2025/05/19/ollama-quick-start/>

Ollama Fast Start w. Docker Compose: Anthony Nocentino

<https://www.nocentino.com/posts/2025-05-19-ollama-sql-faststart>

#learnwithmz Series: Muazma Zahid

[https://www.linkedin.com/posts/muazmazahid\\_learnwithmz-ai-machinelearning-activity-7349463742461464579-BEsj/](https://www.linkedin.com/posts/muazmazahid_learnwithmz-ai-machinelearning-activity-7349463742461464579-BEsj/)

Blog & AI for Everyday IT (book): Chrissy LeMaire

<https://blog.netnerds.net/>

Microsoft Developer Blog Series: Davide Mauri

<https://devblogs.microsoft.com/azure-sql/author/damauri-2/>

Using local Large Language Model OLLAMA with SQL Server: Sebastiao Pereira

<https://www.mssqltips.com/sqlservertip/8245/large-language-model-ollama-with-sql-server/>



# Learn More: Resources

Get Ready for SQL Server 2025: Accelerating AI, Performance, and Resilience—  
From Query to Infrastructure: Anthony Nocentino (Pure) & Dhananjay Mahajan (Microsoft)

<https://www.purestorage.com/video/webinars/accelerating-ai-performance-and-resilience/6372781959112.html>

Performance and Scale for Modern Database Workloads: Whitepaper

<https://www.purestorage.com/docs.html?item=/type/pdf/subtype/doc/path/content/dam/pdf/en/white-papers/wp-performance-scale-for-modern-database-workloads.pdf>

SQL Server 2025: Enterprise AI without the Learning Curve

<https://blog.purestorage.com/purely-technical/sql-server-2025-enterprise-ai-without-the-learning-curve/>

*See you  
again soon!*

# Thank You!

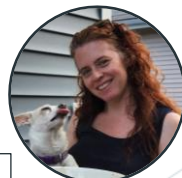
## Any Questions?



**Andy Yun**

ayun@purestorage.com  
sqlbek@gmail.com

<https://sqlbek.wordpress.com>  
<https://github.com/sqlbek>



Did you ever decide  
what's for dinner?

*Keep in Touch:*  
SQL Slack Community  
<https://dbatools.io/slack/>  
Reddit  
r/SQLServer

Special thanks to all the people who made and released these awesome resources for free:  
Presentation template by [SlidesCarnival](#)  
CC0 images sourced from [Unsplash](#), [pixabay.com](#), [wannahpik.com](#) & [pexels.com](#)