Hands on GraphRAG Workshop

Erste Bank Graph Day 2025

Erik Bijl | Neo4j
Jesus Barrasa | Neo4j

Who are we?

Erik Bijl

```
(Erik) - [:LIVES_IN] → (`the Netherlands`)
(Erik) - [:WORKS_FOR] → (`Neo4j`)
(Erik) - [:HAS_ROLE] → (`Solution Engineer`)
(Erik) - [:IS_PART_OF] → (`EMEA Field Team`)
```

Jesus Barrasa

```
(Jesus) - [:LIVES_IN] → (`England`)
(Jesus) - [:WORKS_FOR] → (`Neo4j`)
(Jesus) - [:HAS_ROLE] → (`Field CTO - GenAI`)
(Jesus) - [:Loves] → (`Real Madrid`)
```

Side note: You can read Cypher queries now!

Agenda

1. **Workshop objectives**
   Setting the stage

2. **Groups and setup**
   Let’s get set up

3. **Modules**
   Run the Notebooks

4. **Wrap up**
   Resources and Q&A

Workshop Rules

- Ask questions straight away, this is an interactive session
- Raise your hand if you are stuck
- Slides & notebooks will be shared
- Have fun!

Connect to the notebooks:

- https://erstebank-jp.graphdatabase.ninja/
- Attendeexxx (If you got number 105 => attendee105)

Neo4j Browser : https://browser.neo4j.io

Quick Poll (by show of hands)

Modules

1 **Explore the Graph**
Run some queries

2 **Vector Index**
Set up the Vector Search

3 **Graph Analytics**
Run Graph Algorithms

4 **GraphRAG Chatbot**
Run a Chatbot on the Graph

5 **GraphRAG Agent**
Create Agents with Tools

6 **Wrap up**
Resources and Q&A

Module 1

Explore the Graph: Let’s run some queries

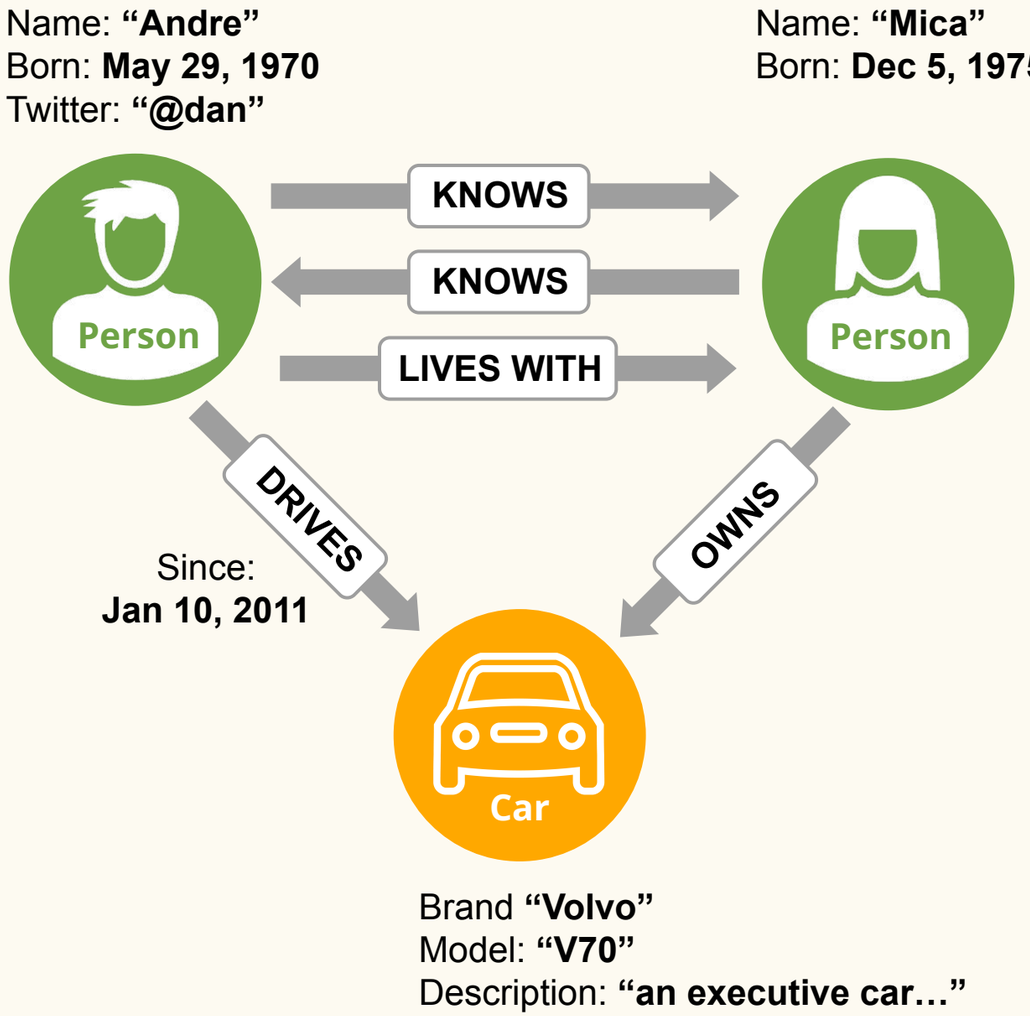
Knowledge Graph = design patterns to organize & access interrelated data

Property Graph Data Model

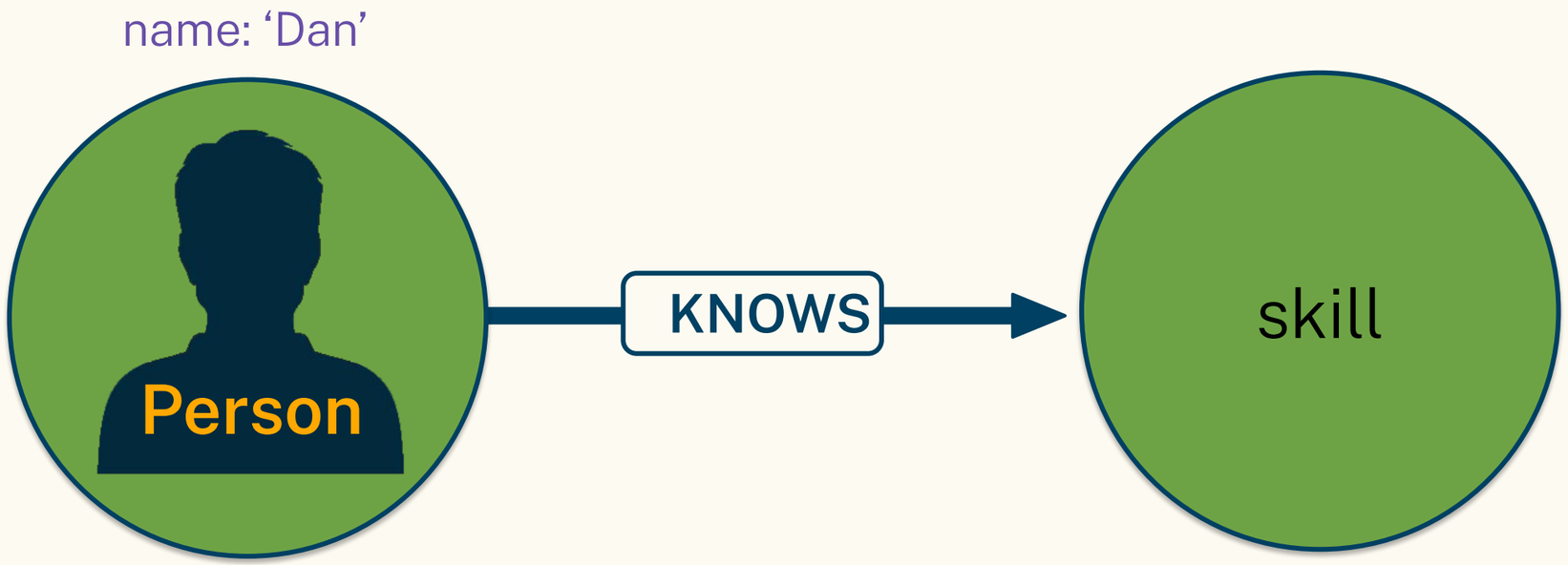
Nodes represent entities in the graph

Relationships represent associations or interactions between nodes

Properties represent attributes of nodes or relationships



Cypher: A Powerful & Expressive Query Language



NODE — RELATIONSHIP TYPE — NODE

```
MATCH (p:Person {name:“Dan”} ) -[:KNOWS]-> (skill)
```

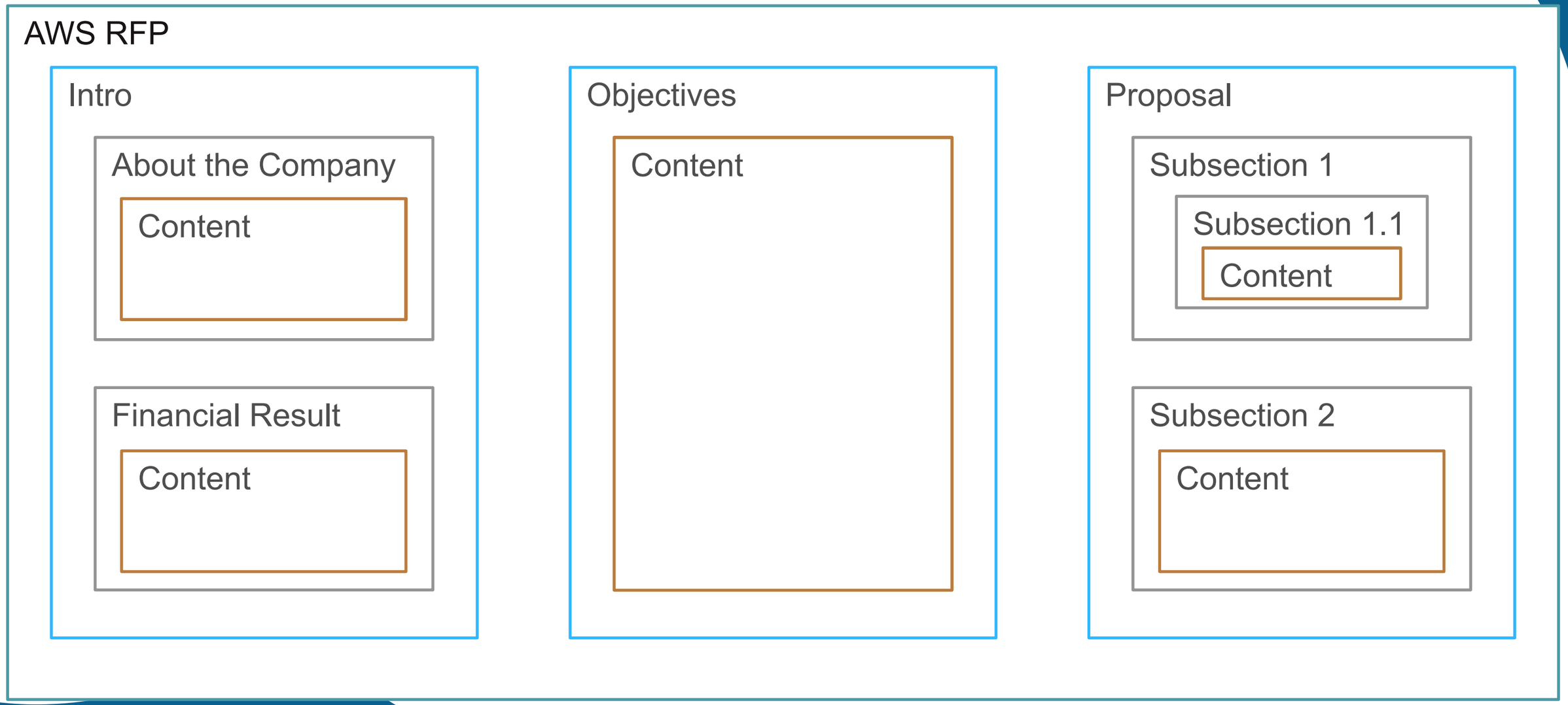
VARIABLE — LABEL — PROPERTY — VARIABLE

```
RETURN p.name as person, skill
```

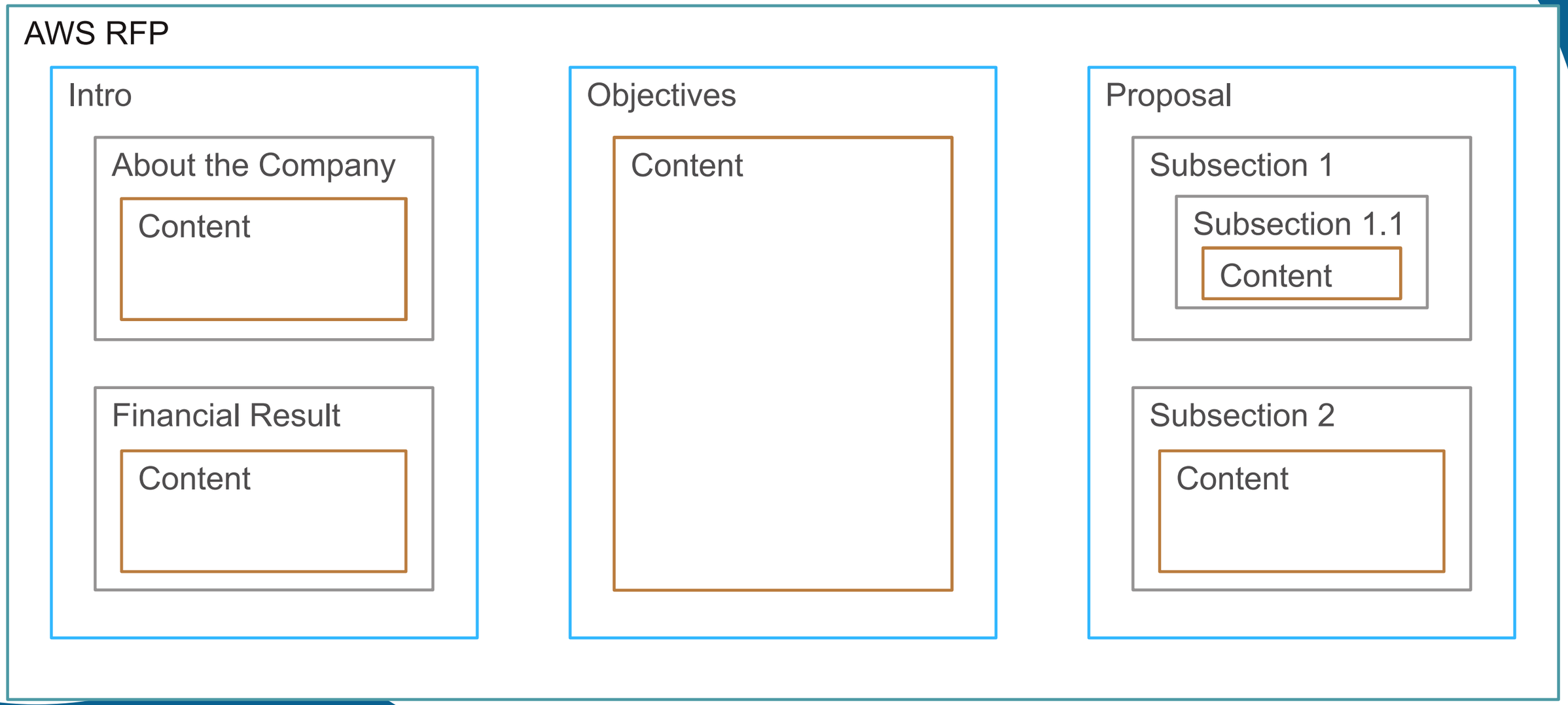
From Unstructured to Structured: An example...

RFP Generation GenAI App

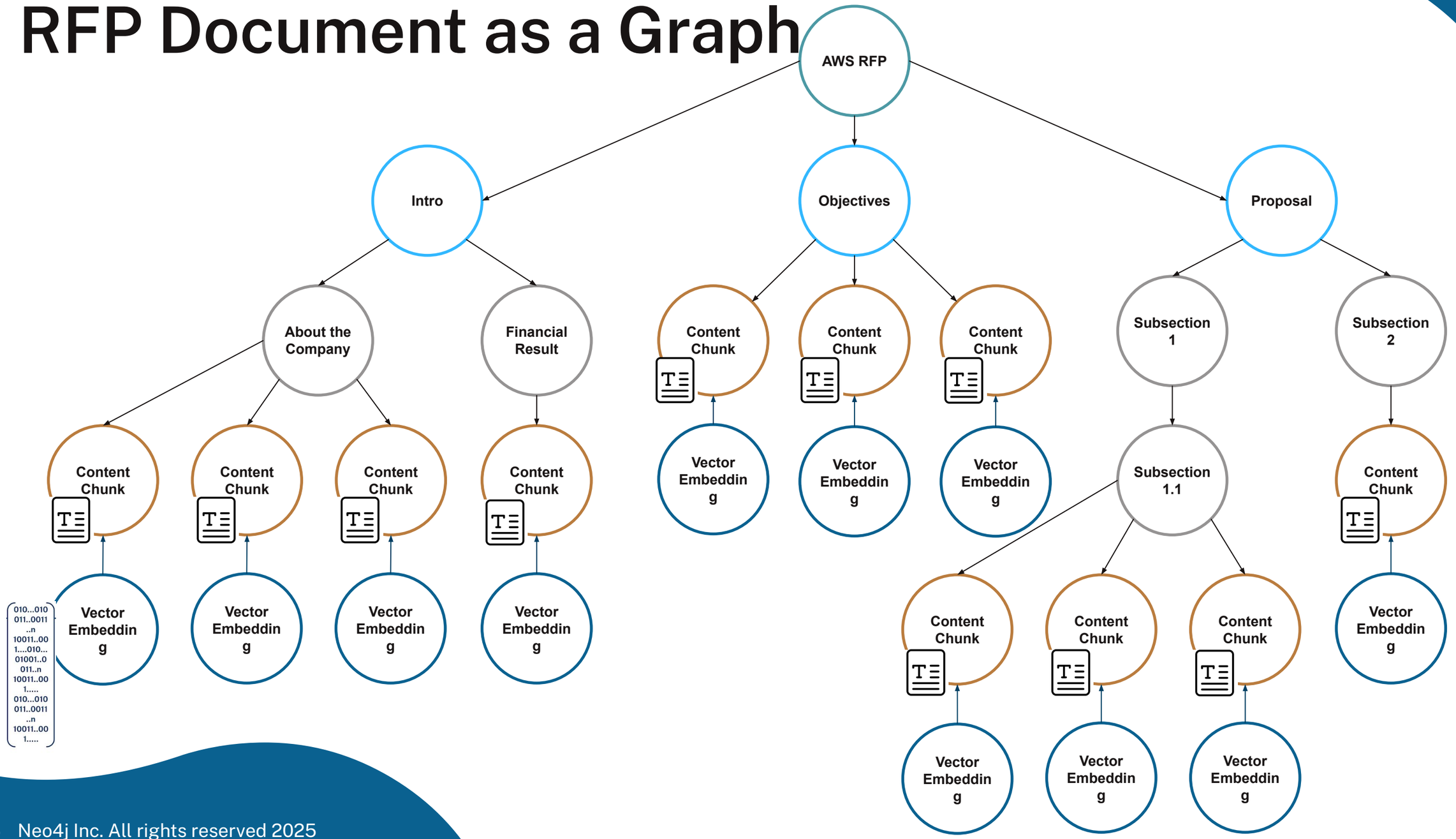
Anatomy of an RFP Document



Anatomy of a Document

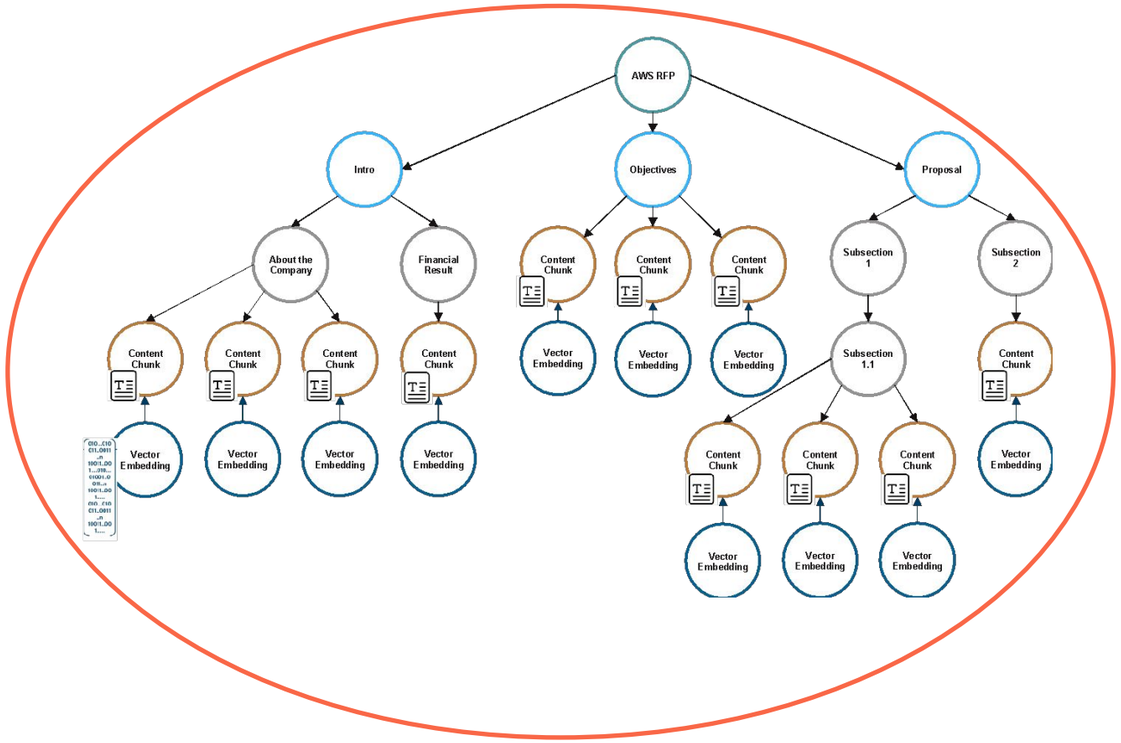


RFP Document as a Graph

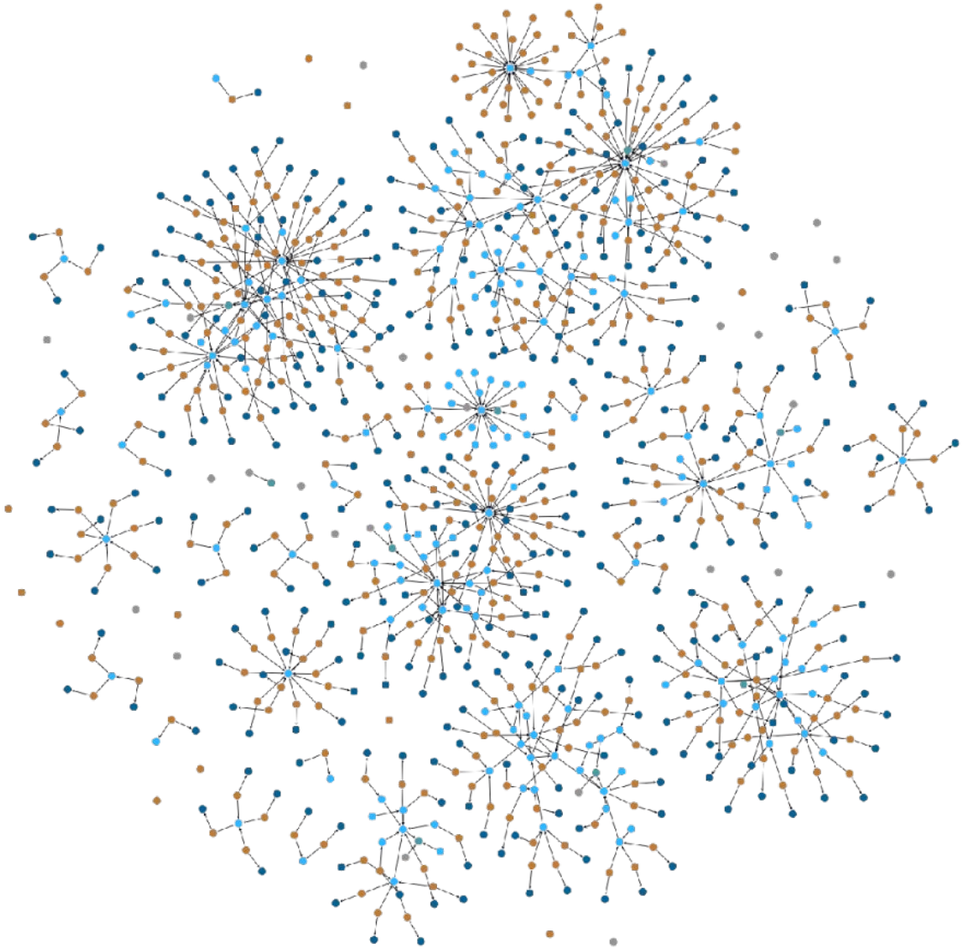


Knowledge Graph as the Knowledge Base

Document in a KG

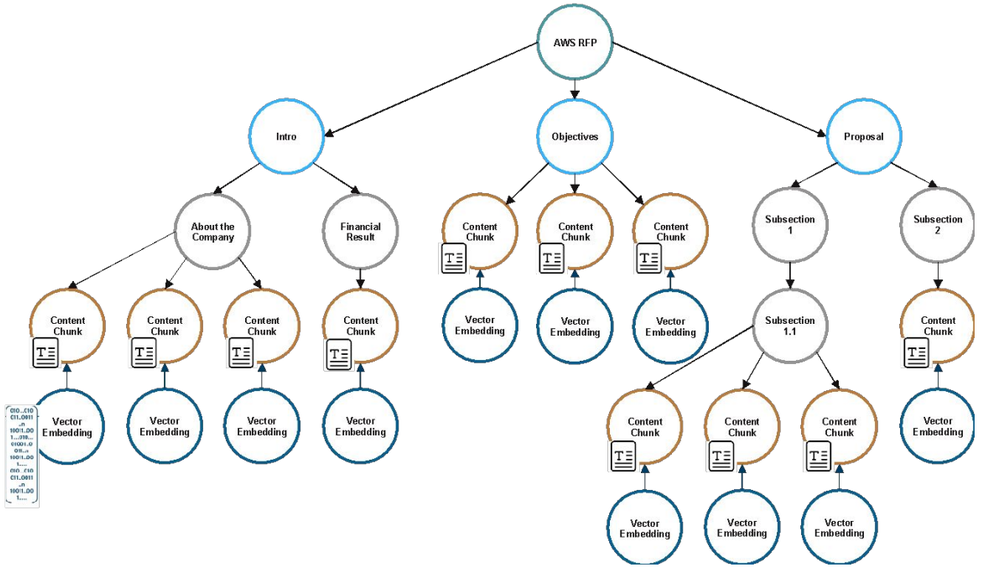


Knowledge Graph

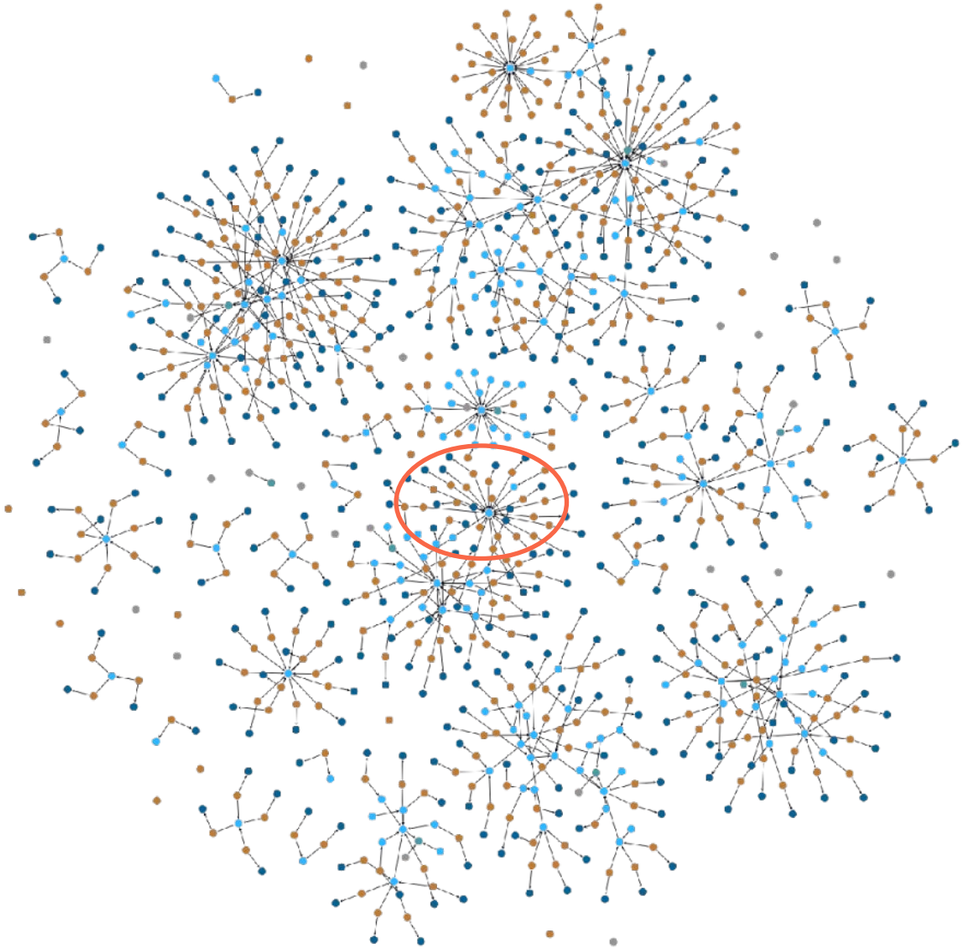


Knowledge Graph as the Knowledge Base

Document in a KG



Knowledge Graph

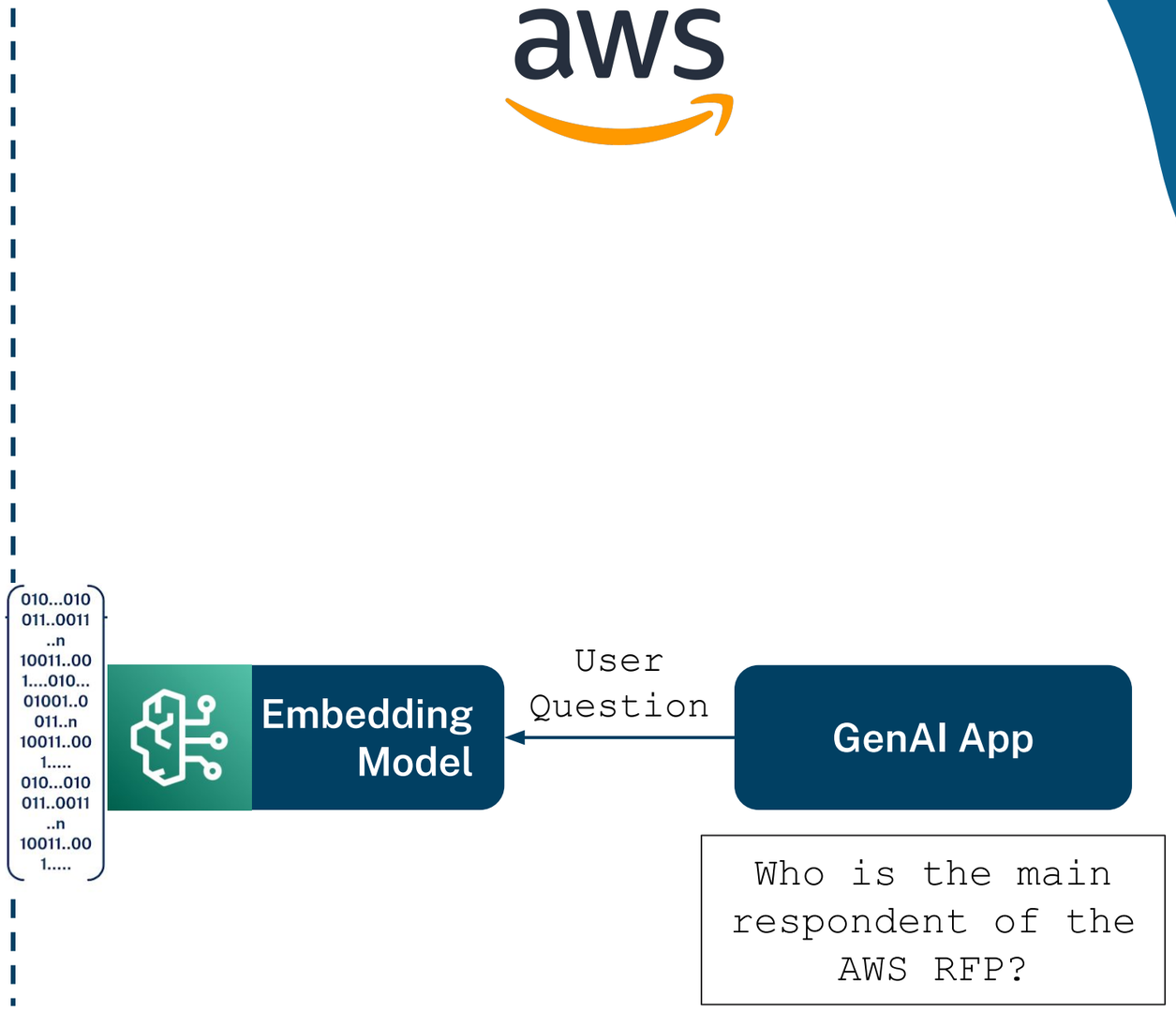


Accurate, Contextual and Explainable

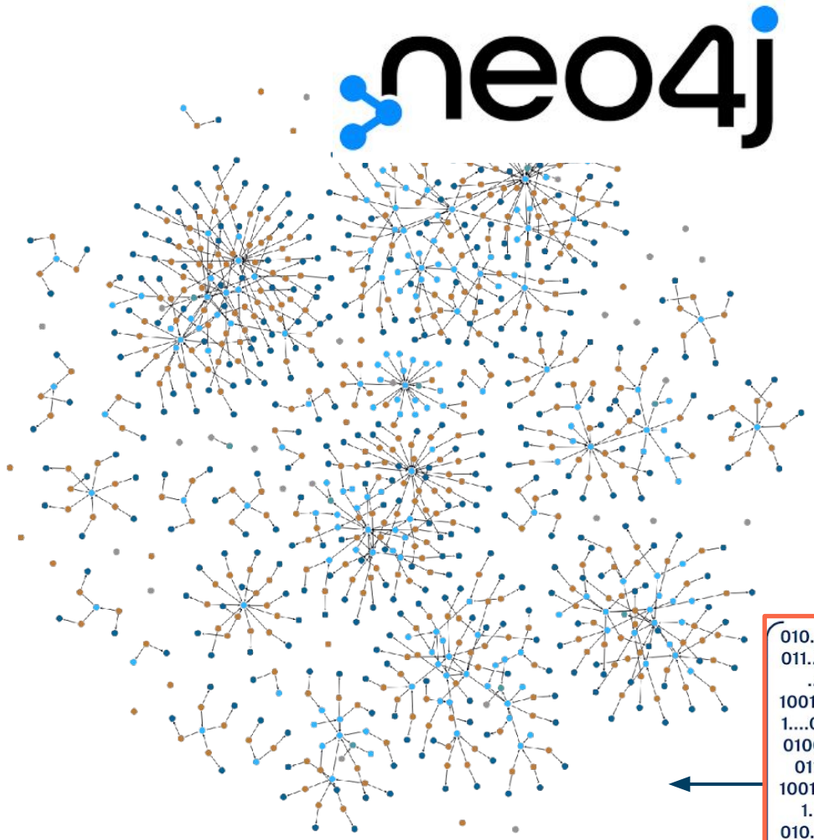
Who is the main respondent of the AWS RFP?

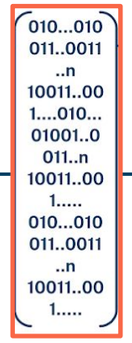
GenAI App

Accurate, Contextual and Explainable



Accurate, Contextual and Explainable





Vector Embedding

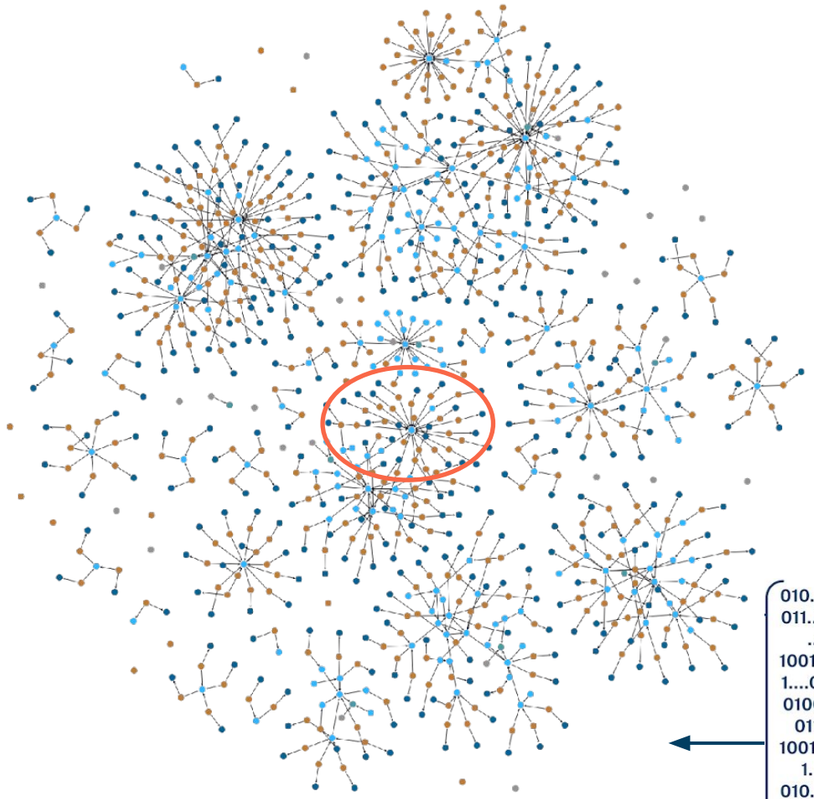
Embedding Model

User Question

GenAI App

Who is the main respondent of the AWS RFP?

Accurate, Contextual and Explainable



Similarity Search using
Neo4j Vector Index

010...010
011..0011
..n
10011..00
1....010...
01001..0
011..n
10011..00
1.....
010...010
011..0011
..n
10011..00
1.....

Vector Embedding

Embedding Model

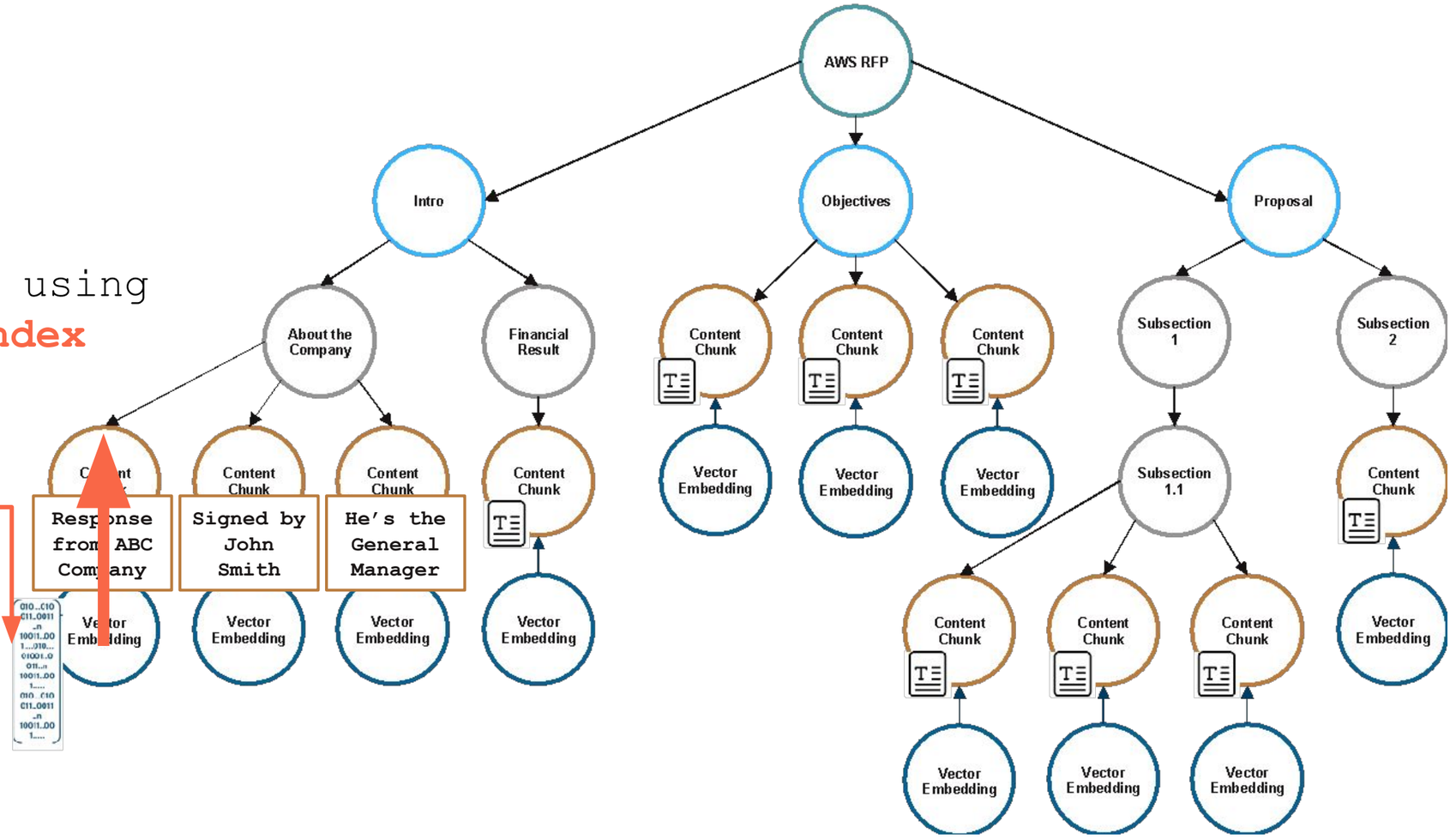
User Question

GenAI App

Who is the main respondent of the AWS RFP?

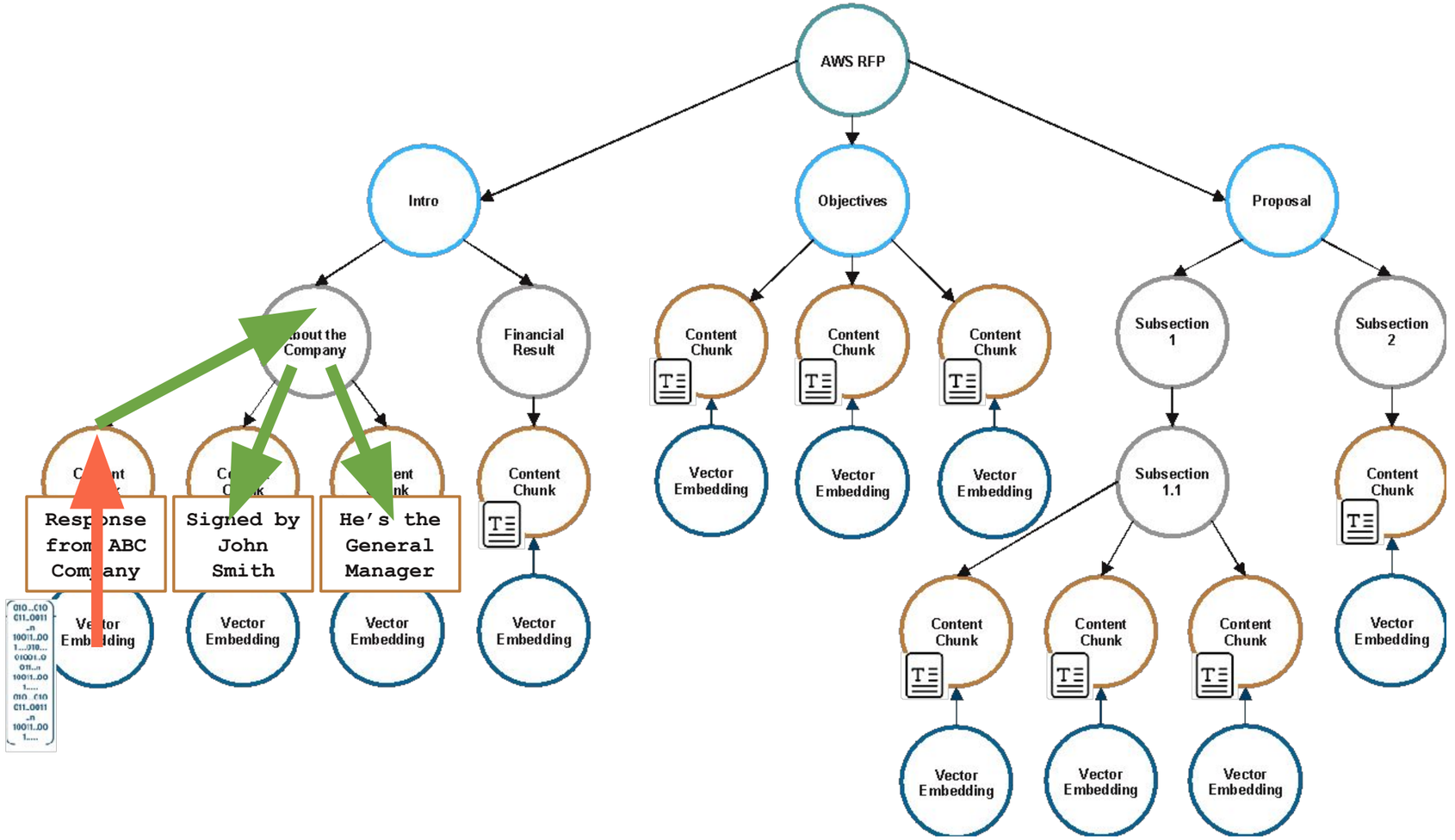
Accurate, Contextual and Explainable

Similarity Search using
Neo4j Vector Index



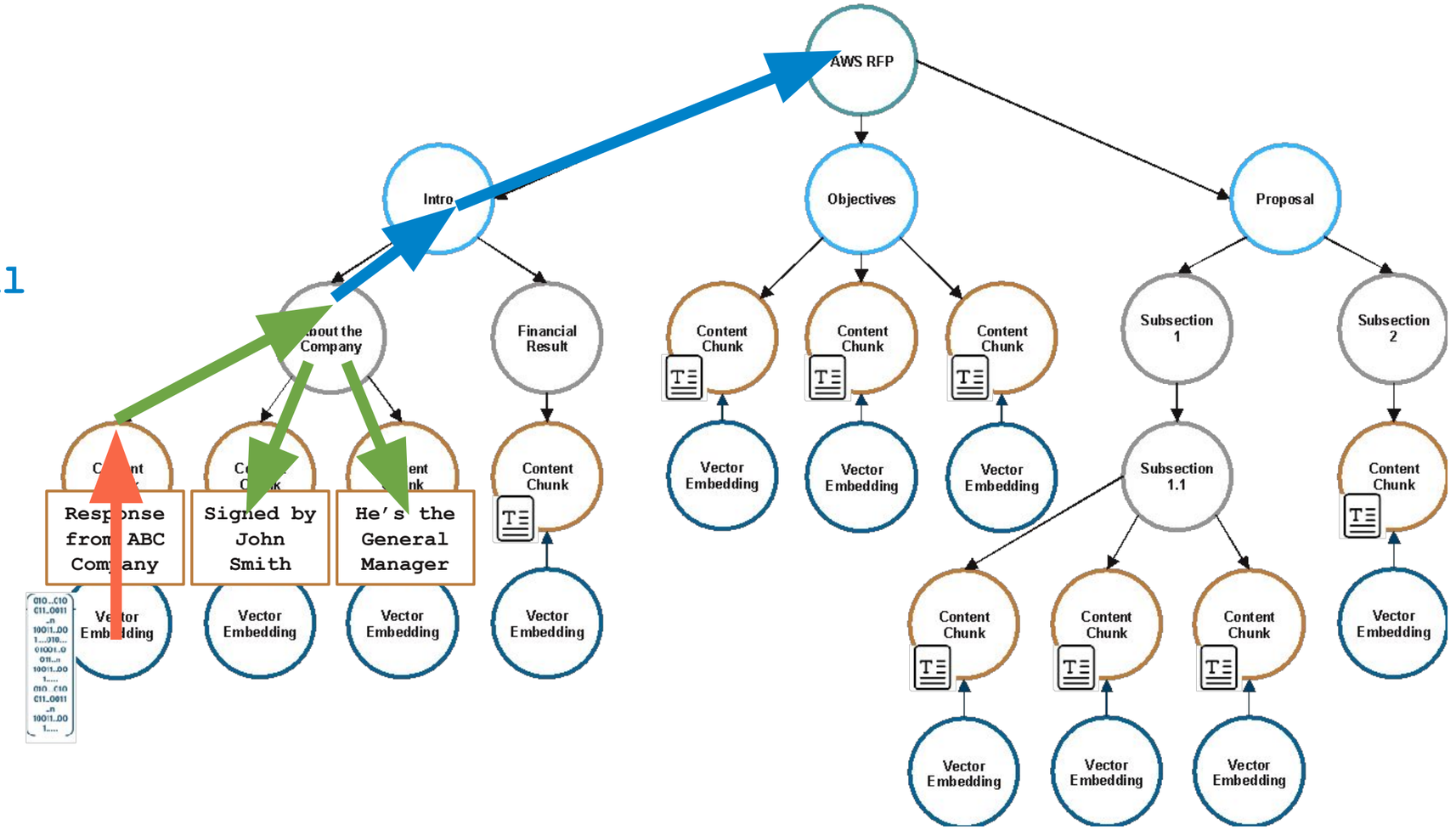
Accurate, Contextual and Explainable

Contextual Knowledge Retrieval within Neo4j KG



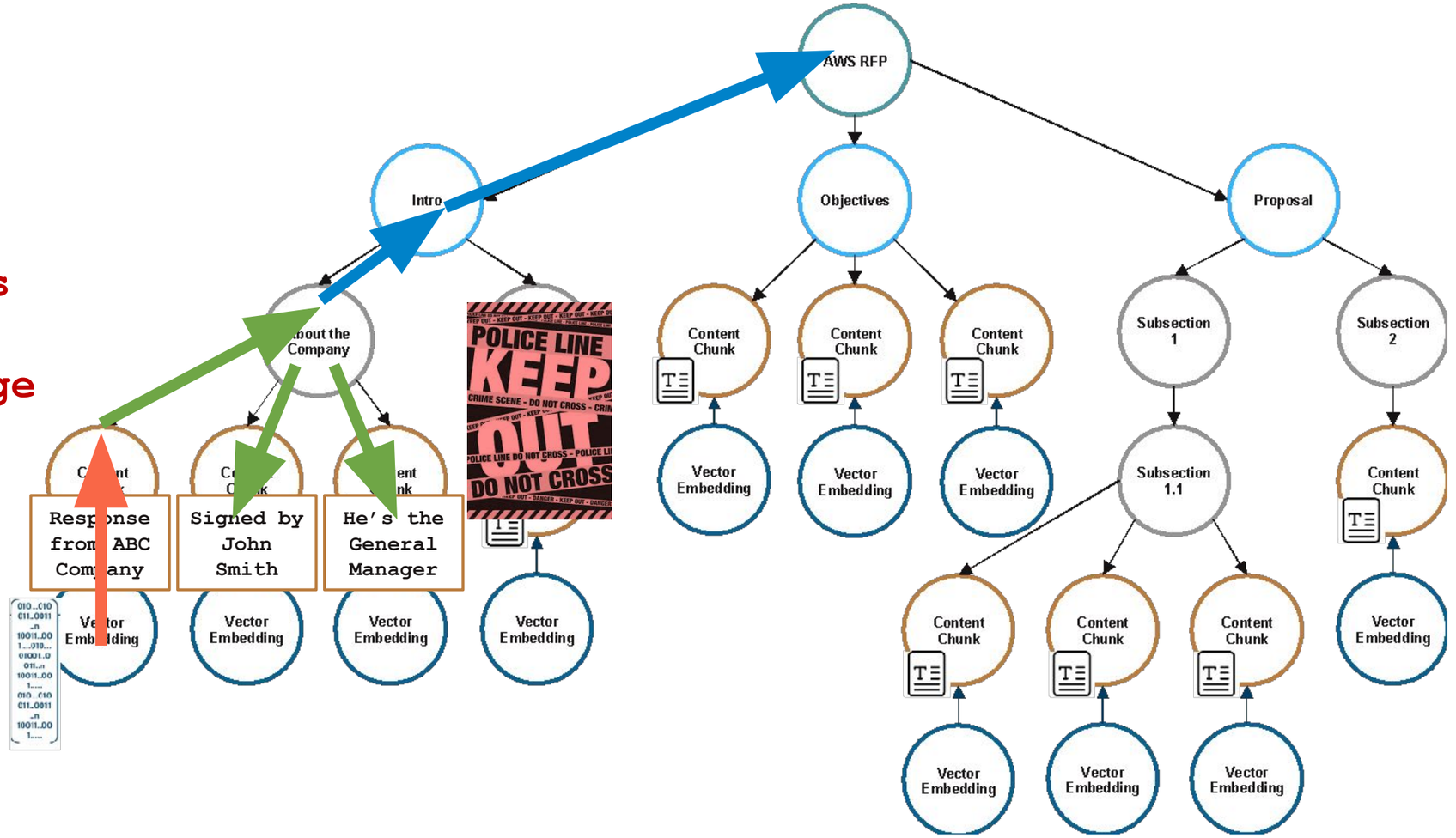
Accurate, Contextual and Explainable

Knowledge Retrieval
to aid in
Explainability

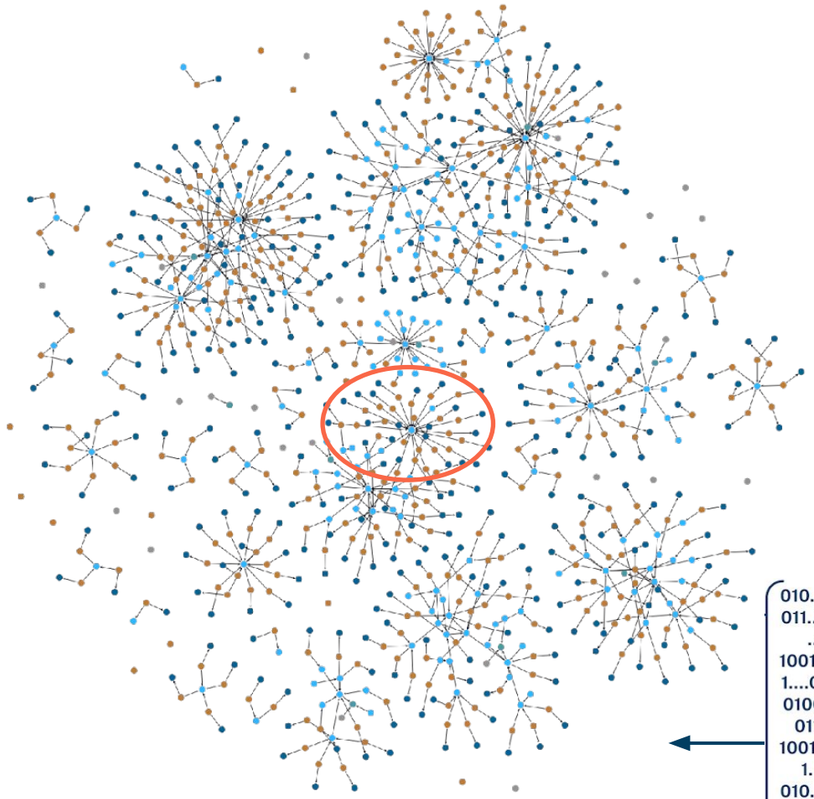


Accurate, Contextual and Explainable

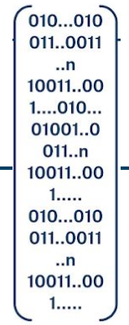
Fine Grained Access Control to prevent **unwarranted Knowledge Retrieval**



Accurate, Contextual and Explainable



Similarity Search using
Neo4j Vector Index



Vector Embedding

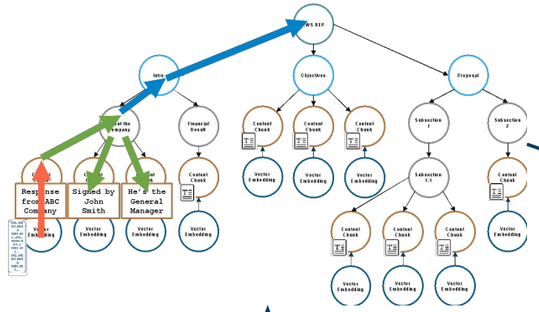
Embedding Model

User Question

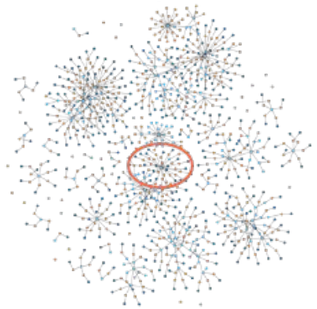
GenAI App

Who is the main respondent of the AWS RFP?

Accurate, Contextual and Explainable



Similarity Result



Similarity Search using Neo4j Vector Index

010...010
011..0011
..n
10011..00
1....010...
01001..0
011..n
10011..00
1.....
010...010
011..0011
..n
10011..00
1.....

Vector Embedding

Embedding Model

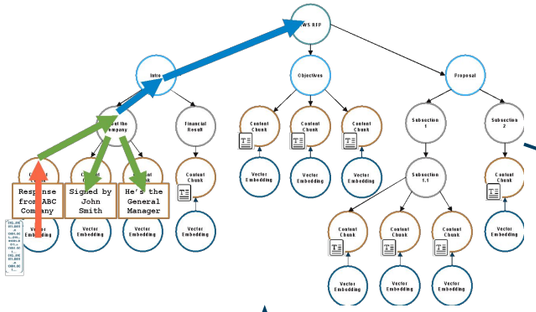
User Question

GenAI App

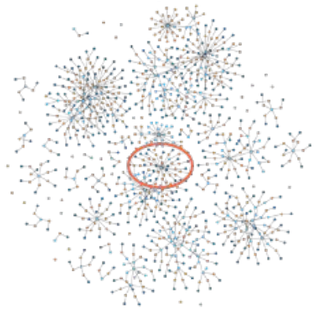
LLM

Who is the main respondent of the AWS RFP?

Accurate, Contextual and Explainable



Similarity Result



Similarity Search using Neo4j Vector Index

010...010
011..0011
..n
10011..00
1....010...
01001..0
011..n
10011..00
1.....
010...010
011..0011
..n
10011..00
1.....

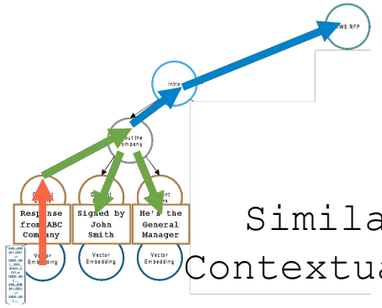
Vector Embedding

Embedding Model

User Question

GenAI App

Who is the main respondent of the AWS RFP?



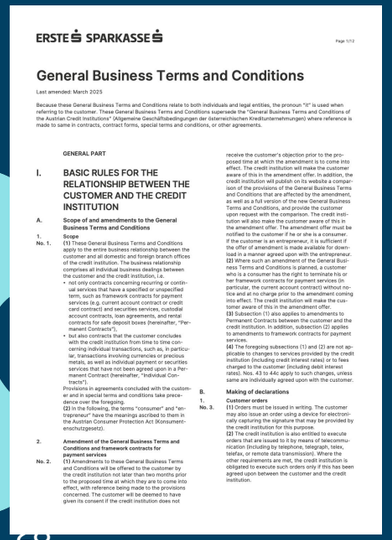
Similarity + Contextual Result

LLM

The main respondent of the AWS RFP is John Smith, General Manager XYZ Group representing ABC Company (source: RFP_AWS.pdf, page: 38)

Erste Bank Documents

- Allgemeine Reiseversicherungsbedingungen
- Reiseversicherung der Kreditkarten von Erste Bank und Sparkasse
- Krankenversicherung
- General information on consumer payment services
- General Business Terms and Conditions
- Informationen über uns und unsere Wertpapierdienstleistungen



ERSTE SPARKASSE

Page 1/12

General Business Terms and Conditions

Last amended: March 2023

Because these General Business Terms and Conditions relate to both individuals and legal entities, the pronoun "it" is used when referring to the customer. These General Business Terms and Conditions supersede the "General Business Terms and Conditions of the Austrian Credit Institutions" (Allgemeine Geschäftsbedingungen der österreichischen Kreditunternehmungen) where reference is made to same in contracts, contract forms, special terms and conditions, or other agreements.

GENERAL PART

I. BASIC RULES FOR THE RELATIONSHIP BETWEEN THE CUSTOMER AND THE CREDIT INSTITUTION

A. Scope of and amendments to the General Business Terms and Conditions

Scope

No. 1. (1) These General Business Terms and Conditions apply to the entire business relationship between the customer and all domestic and foreign branch offices of the credit institution. The business relationship comprises all individual business dealings between the customer and the credit institution, i.e.

- not only contracts concerning recurring or continual services that have a specified or unspecified term, such as framework contracts for payment services (e.g. current account contract or credit card contract) and securities services, custodial account contracts, loan agreements, and rental agreements for deposit boxes (hereinafter, "Permanent Contracts"),
- but also contracts that the customer concludes with the credit institution from time to time concerning individual transactions, such as, in particular, transactions involving currencies or precious metals, as well as individual payment or securities services that have not been agreed upon in a Permanent Contract (hereinafter, "Individual Contracts").

Provisions in agreements concluded with the customer and in special terms and conditions take precedence over the foregoing.

(2) In the following, the terms "consumer" and "entrepreneur" have the meanings ascribed to them in the Austrian Consumer Protection Act (Konsumentenschutzgesetz).

Amendment of the General Business Terms and Conditions and framework contracts for payment services

No. 2. (1) Amendments to these General Business Terms and Conditions will be offered to the customer by the credit institution not later than two months prior to the proposed time at which they are to come into effect, with reference being made to the provisions concerned. The customer will be deemed to have given its consent if the credit institution does not receive the customer's objection prior to the proposed time at which the amendment is to come into effect. The credit institution will make the customer aware of this in the amendment offer. In addition, the credit institution will publish on its website a comparison of the provisions of the General Business Terms and Conditions that are affected by the amendment, as well as a full version of the new General Business Terms and Conditions, and provide the customer upon request with the comparison. The credit institution will also make the customer aware of this in the amendment offer. The amendment offer must be notified to the customer if he or she is a consumer. If the customer is an entrepreneur, it is sufficient if the offer of amendment is made available for download in a manner agreed upon with the entrepreneur.

(2) Where such an amendment of the General Business Terms and Conditions is planned, a customer who is a consumer has the right to terminate his or her framework contract for payment services, in particular, the current account contract) without notice and at no charge prior to the amendment coming into effect. The credit institution will make the customer aware of this in the amendment offer.

(3) Subsection (1) also applies to amendments to Permanent Contracts between the customer and the credit institution. In addition, subsection (2) applies to amendments to framework contracts for payment services.

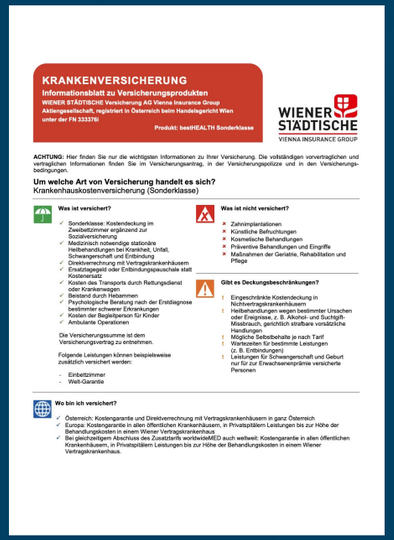
(4) The foregoing subsections (1) and (2) are not applicable to changes to services provided by the credit institution (including credit interest rates) or to fees charged to the customer (including debit interest rates). Nos. 43 to 44c apply to such changes, unless same are individually agreed upon with the customer.

B. Making of declarations

Customer orders

No. 3. (1) Orders must be issued in writing. The customer may also issue an order using a device for electronic data transmission if such possibility has been provided by the credit institution for this purpose.

(2) The credit institution is also entitled to execute orders that are issued to it by means of telecommunication (including by telephone, telegraph, telex, telefax, or remote data transmission). Where the other requirements are met, the credit institution is obligated to execute such orders only if this has been agreed upon between the customer and the credit institution.



KRANKENVERSICHERUNG
Informationsblatt zu Versicherungsprodukten
WIENER STÄDTISCHE Versicherung AG Vienna Insurance Group
Aktiengesellschaft, registriert in Österreich beim Handelsgericht Wien unter der FN 33337fi
Produkt: bestHEALTH Sonderklasse

WIENER STÄDTISCHE
VIENNA INSURANCE GROUP

ACHTUNG: Hier finden Sie nur die wichtigsten Informationen zu Ihrer Versicherung. Die vollständigen vorvertraglichen und vertraglichen Informationen finden Sie im Versicherungsantrag, in der Versicherungspolizze und in den Versicherungsbedingungen.

Um welche Art von Versicherung handelt es sich?
Krankenhauskostenversicherung (Sonderklasse)

Was ist versichert?
- ✓ Sonderklasse: Kostendeckung im Zweibettzimmer ergänzend zur Sozialversicherung
- ✓ Medizinisch notwendige stationäre Heilbehandlungen bei Krankheit, Unfall, Schwangerschaft und Entbindung
- ✓ Direktverrechnung mit Vertragskrankenhäusern
- ✓ Ersatztagegeld oder Entbindungspauschale statt Kostenersatz
- ✓ Kosten des Transports durch Rettungsdienst oder Krankenwagen
- ✓ Bestand durch Hebammen
- ✓ Psychologische Beratung nach der Entbindung bestimmter schwerer Erkrankungen
- ✓ Kosten der Begleitperson für Kinder
- ✓ Ambulante Operationen

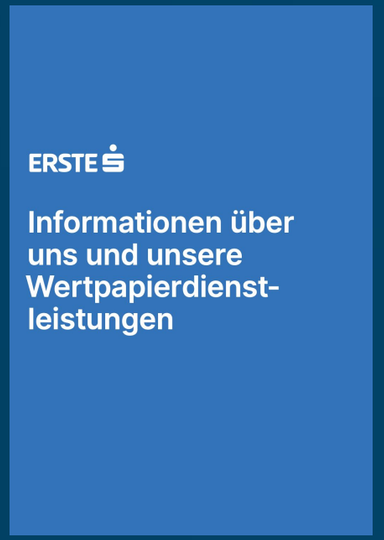
Die Versicherungssumme ist dem Versicherungsvertrag zu entnehmen.

Folgende Leistungen können beispielsweise zusätzlich versichert werden:
- Einbettzimmer
- Welt-Garantie

Was ist nicht versichert?
- ✗ Zahnimplantationen
- ✗ Künstliche Befruchtungen
- ✗ Kosmetische Behandlungen
- ✗ Präventive Behandlungen und Eingriffe
- ✗ Maßnahmen der Geriatrie, Rehabilitation und Pflege

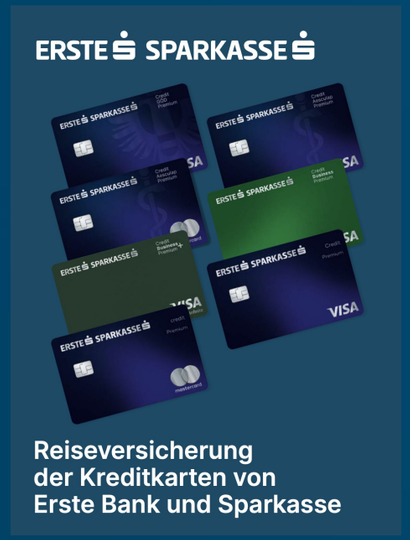
Gibt es Deckungsbeschränkungen?
- ! Eingeschränkte Kostendeckung in Nichtvertragskrankenhäusern
- ! Heilbehandlungen wegen bestimmter Ursachen oder Ereignisse, z. B. Alkohol- und Suchtgift-Missbrauch, gerichtlich strafbare vorsätzliche Handlungen
- ! Wartezeiten für bestehende je nach Tarif
- ! Wartezeiten für bestimmte Leistungen (z. B. Entbindungen)
- ! Leistungen für Schwangerschaft und Geburt nur für zur Erwachsenenprämie versicherte Personen

Wo bin ich versichert?
- ✓ Österreich: Kostengarantie und Direktverrechnung mit Vertragskrankenhäusern in ganz Österreich
- ✓ Europa: Kostengarantie in allen öffentlichen Krankenhäusern, in Privatspitälern Leistungen bis zur Höhe der Behandlungskosten in einem Wiener Vertragskrankenhaus
- ✓ Bei gleichzeitigem Abschluss des Zusatztarifs worldwideMED auch weltweit: Kostengarantie in allen öffentlichen Krankenhäusern, in Privatspitälern Leistungen bis zur Höhe der Behandlungskosten in einem Wiener Vertragskrankenhaus.



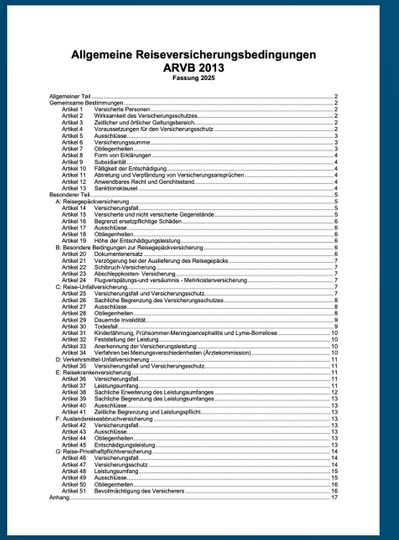
ERSTE

Informationen über uns und unsere Wertpapierdienst-leistungen

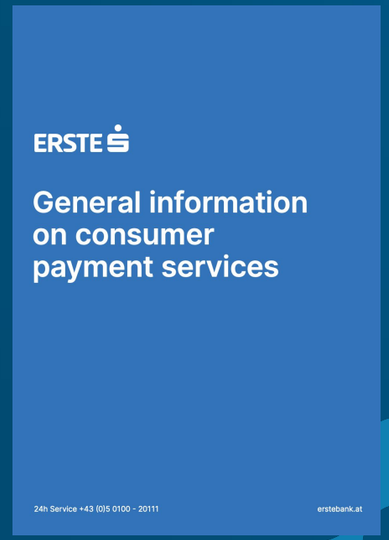


ERSTE SPARKASSE

Reiseversicherung der Kreditkarten von Erste Bank und Sparkasse



Allgemeine Reiseversicherungsbedingungen ARVB 2013
Fassung 2025



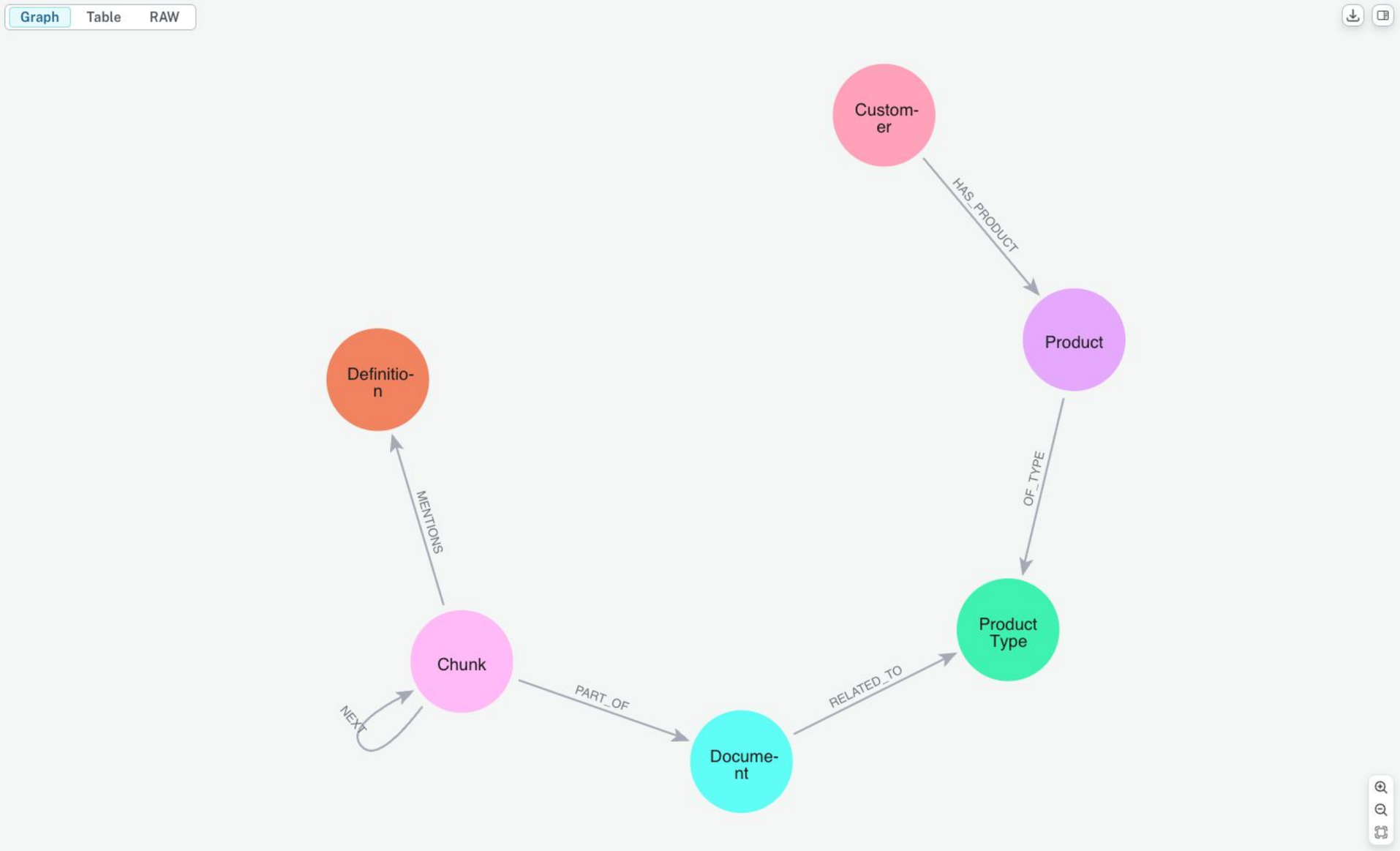
ERSTE

General information on consumer payment services

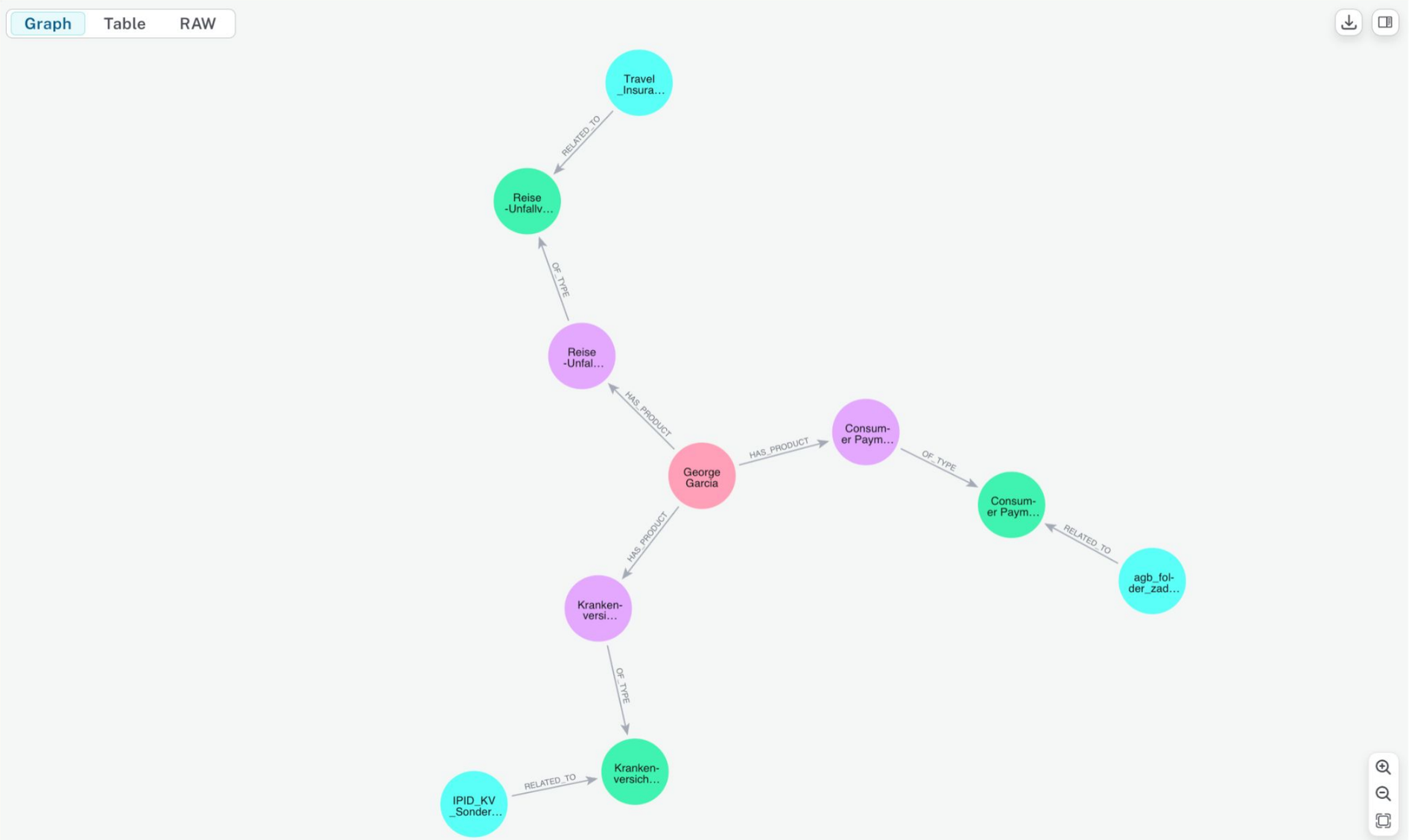
24h Service +43 (0)5 0100 - 20111

erstebank.at

Our Data Model Today

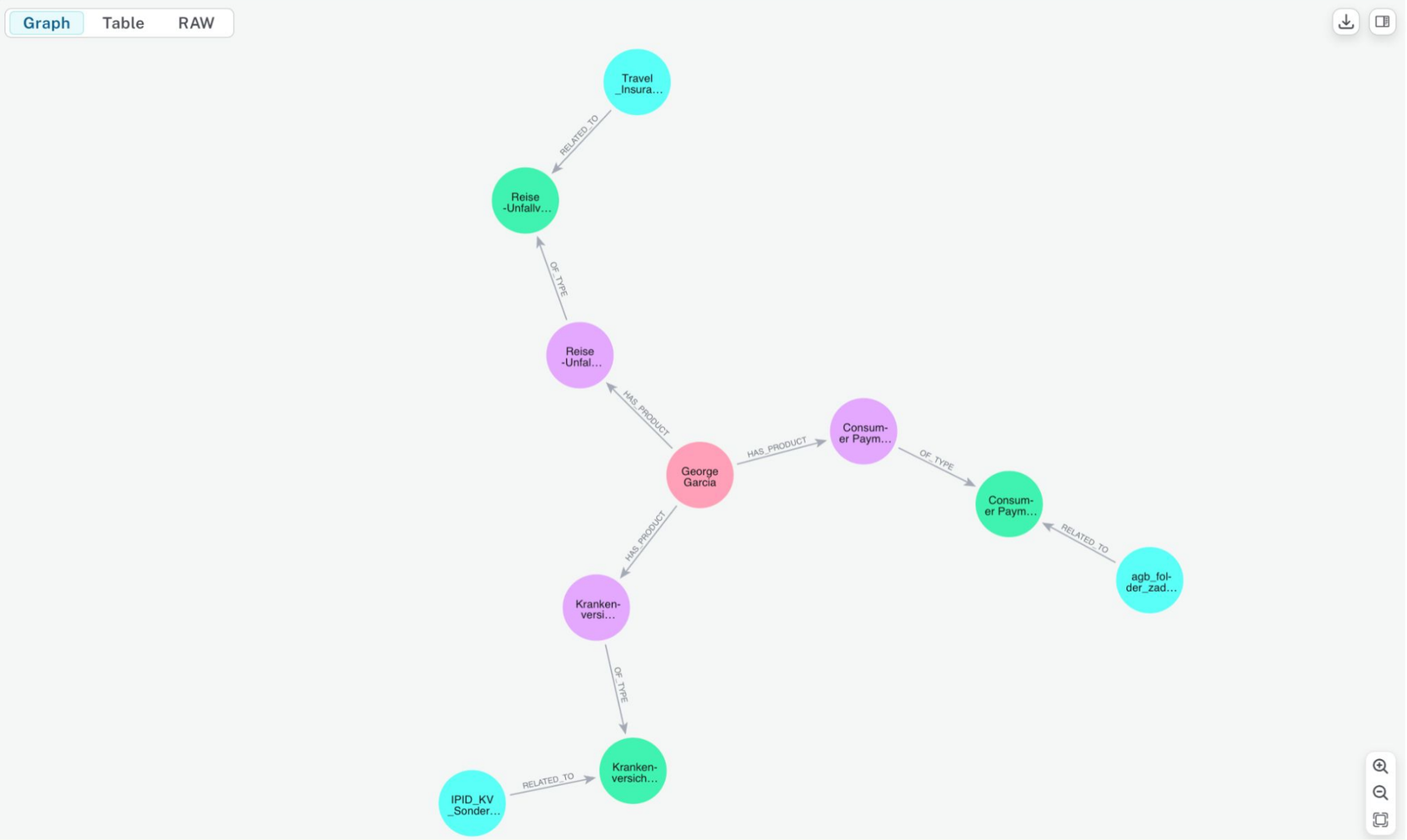


Our Data Model Today



Module 1
Go to Notebook

Our Data Model Today




```
1 MATCH p=(:Customer {name:"George Garcia"})-[]->(:Product)-[]->(:ProductType)-[]-(:Document)
2 RETURN p
```

Module 2

Vector Index: Set up the vector Search

Rabobank
neo4j

Knowledge Graphs – New & Improved!

NOW WITH VECTORS!

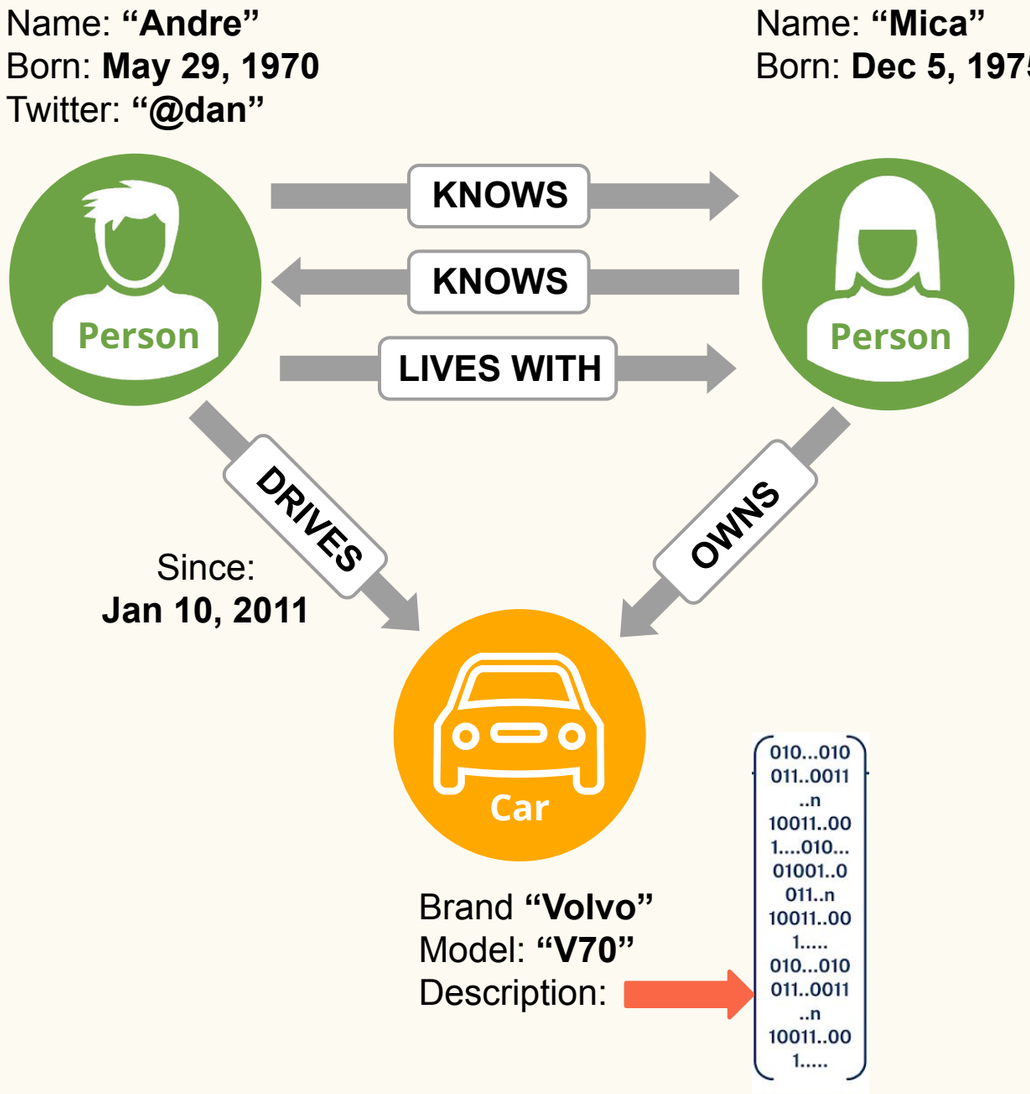
Knowledge Graph = design patterns to organize & access interrelated data

Property Graph Data Model

Nodes represent entities in the graph

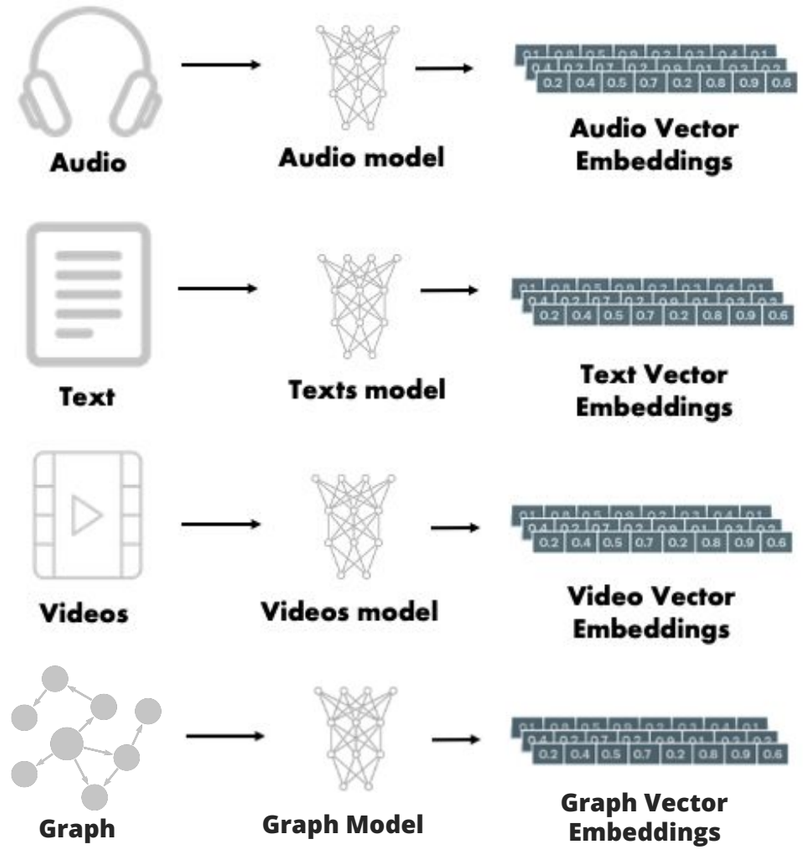
Relationships represent associations or interactions between nodes

Properties represent attributes of nodes or relationships



What are Embeddings?

- A type of data compression
- Transform messy data (text, images, audio, etc.) into a compact format for ML algorithms
- Most often numeric **vectors** (a.k.a arrays) with 100s or 1000s of elements
- Preserve information such that ***"similar" items have proportionally "similar" embedding vectors***
- Similarity is measured with vector algorithms (cosine, euclidean, etc.)



Search & Vectors in Neo4j

Neo4j makes search efficient through a variety of indexes including

- **Range:** General index for predicates based on equality and range.
- **Point:** Predicts on geospatial points like distance bounding boxes, etc.
- **Text:** Predicates on strings like contains, ends with, etc.
- **FullText:** Text search based on tokenization and analyzers
- ***Vector: ANN (Approximate Nearest Neighbor) search on vectors***

How is similarity measured

What would you use for openai embedding models?

For OpenAI embedding models, **cosine similarity** is the preferred choice. Here's why:

Why Cosine Similarity?

1. **OpenAI Embeddings Are Normalized**
   - OpenAI embedding models (like `text-embedding-ada-002`) generate **normalized vectors**, meaning each vector has a unit length (i.e., a magnitude of 1).
   - In such cases, **cosine similarity and Euclidean distance are proportional**, but cosine similarity is computationally more efficient.
2. **Focus on Semantic Meaning**
   - Since cosine similarity measures the **angle between vectors**, it captures semantic similarity between words, sentences, or documents effectively.
   - Euclidean distance, on the other hand, is sensitive to magnitude differences, which aren't meaningful in normalized embeddings.
3. **Industry Best Practice**
   - Most applications using OpenAI embeddings (e.g., search ranking, recommendation systems, and semantic clustering) rely on cosine similarity.
   - OpenAI's own documentation recommends using **cosine similarity or dot product** over Euclidean distance.

Module 2
Go to Notebook

Module 3

Graph Analytics: Run some Graph Algorithms

50+ Graph Algorithms in Neo4j

Pathfinding & Search

- Shortest Path
- Single-Source Shortest Path
- All Pairs Shortest Path
- A* Shortest Path
- Yen's K Shortest Path
- Minimum Weight Spanning Tree
- K-Spanning Tree (MST)
- Random Walk
- Breadth & Depth First Search

Centrality / Importance

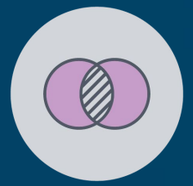
- Degree Centrality
- Closeness Centrality
- Harmonic Centrality
- Betweenness Centrality & Approx.
- PageRank
- Personalized PageRank
- ArticleRank
- Eigenvector Centrality

Community Detection

- Triangle Count
- Local Clustering Coefficient
- Connected Components (Union Find)
- Strongly Connected Components
- Label Propagation
- Louvain Modularity
- K-1 Coloring
- Modularity Optimization

Link Prediction

- Adamic Adar
- Common Neighbors
- Preferential Attachment
- Resource Allocations
- Same Community
- Total Neighbors



Similarity

- Euclidean Distance
- Cosine Similarity
- Node Similarity (Jaccard)
- Overlap Similarity
- Pearson Similarity
- Approximate KNN

Embeddings

- Node2Vec
- Random Projections
- GraphSAGE

... Auxiliary Functions:

- Random graph generation
- Graph export
- One hot encoding
- Distributions & metrics

Today we will explore

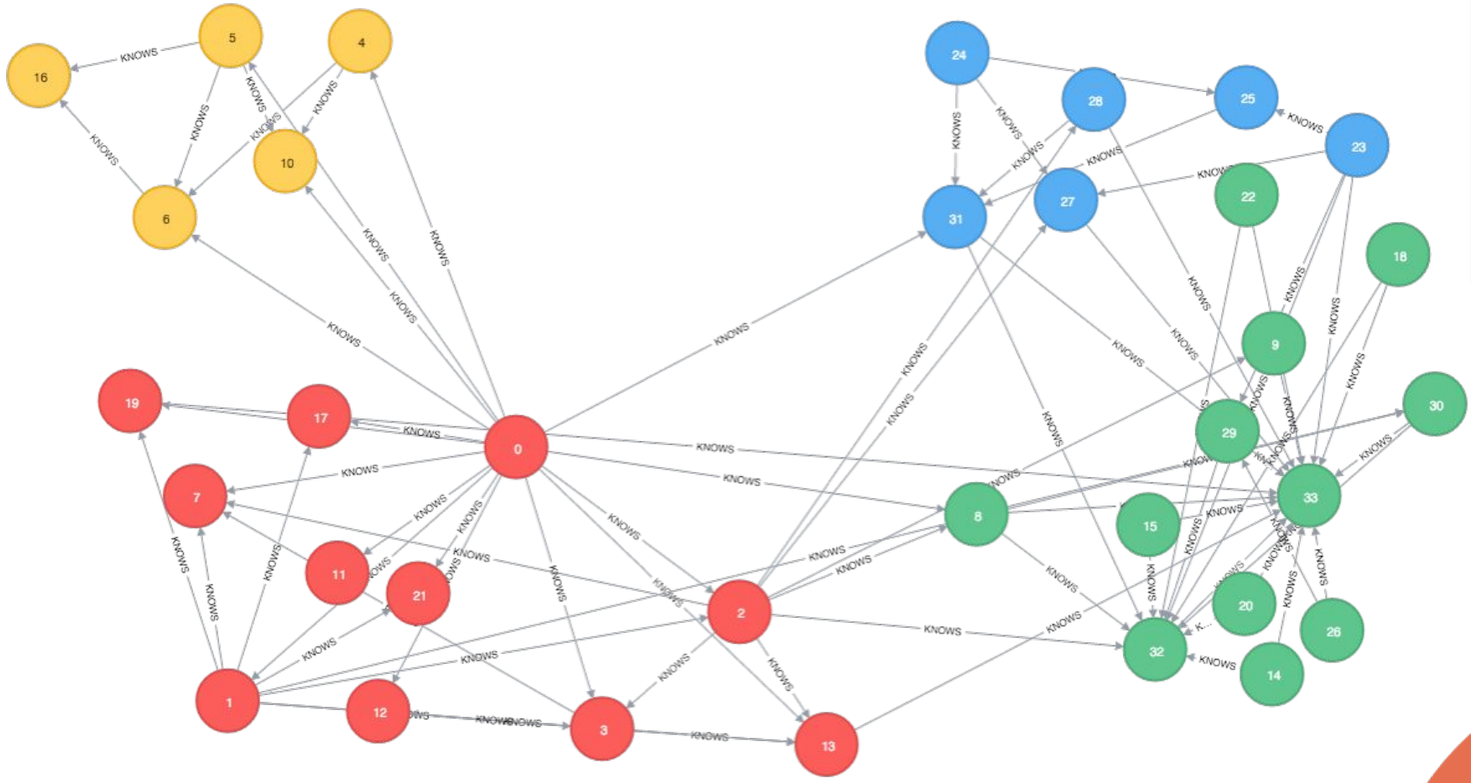
Community Detection

Evaluate how groups of nodes may be clustered or partitioned in a graph.

Community id properties assigned to node based on relationship structure.

Useful for:

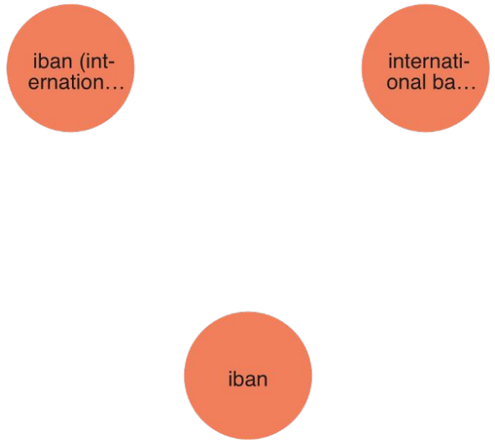
- Segmentation
- Clustering
- Entity resolution
- Summarization (for AI)



Entity Resolution

! Note !

- **The LLM is very good at extracting entities, but the end product can contain very similar entities**
- **This will make it more difficult to obtain insight from the graph**
- **Solution: perform Entity Resolution first**




```
1 MATCH (d:Definition)
2 WHERE LOWER(d.term) CONTAINS "iban"
3 RETURN d.term
```

Table RAW

	d.term
1	"iban"
2	"international bank account number (iban)"
3	"iban (international bank account number)"

Started streaming 3 records in less than 1ms and completed after 1 ms.

Module 3
Go to Notebook

Module 4

GraphRAG Chatbot: Run a Chatbot on the Graph

Knowledge Graphs + LLM: Bringing it Together

Facts
Explicit
Explainable

KG + LLMs

Words
Implicit
Opaque

BETTER TOGETHER

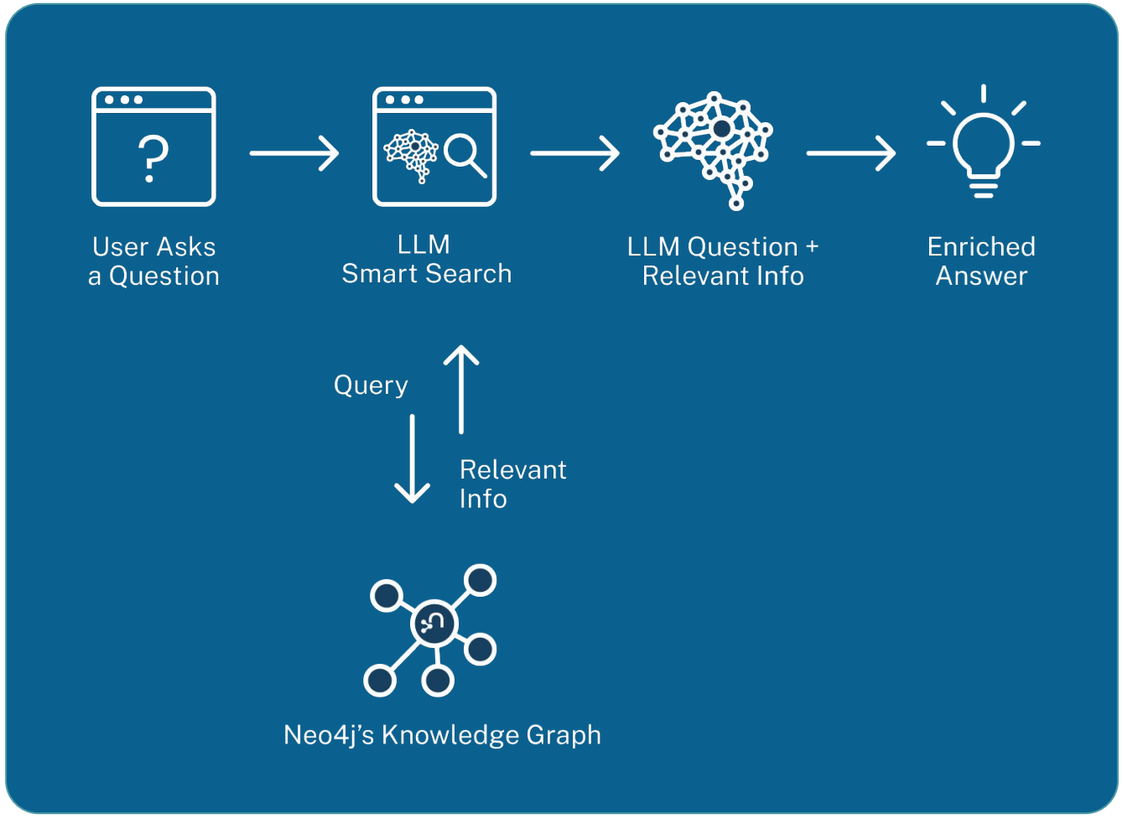
Building a RAG-application in Neo4j

Preparation Steps:

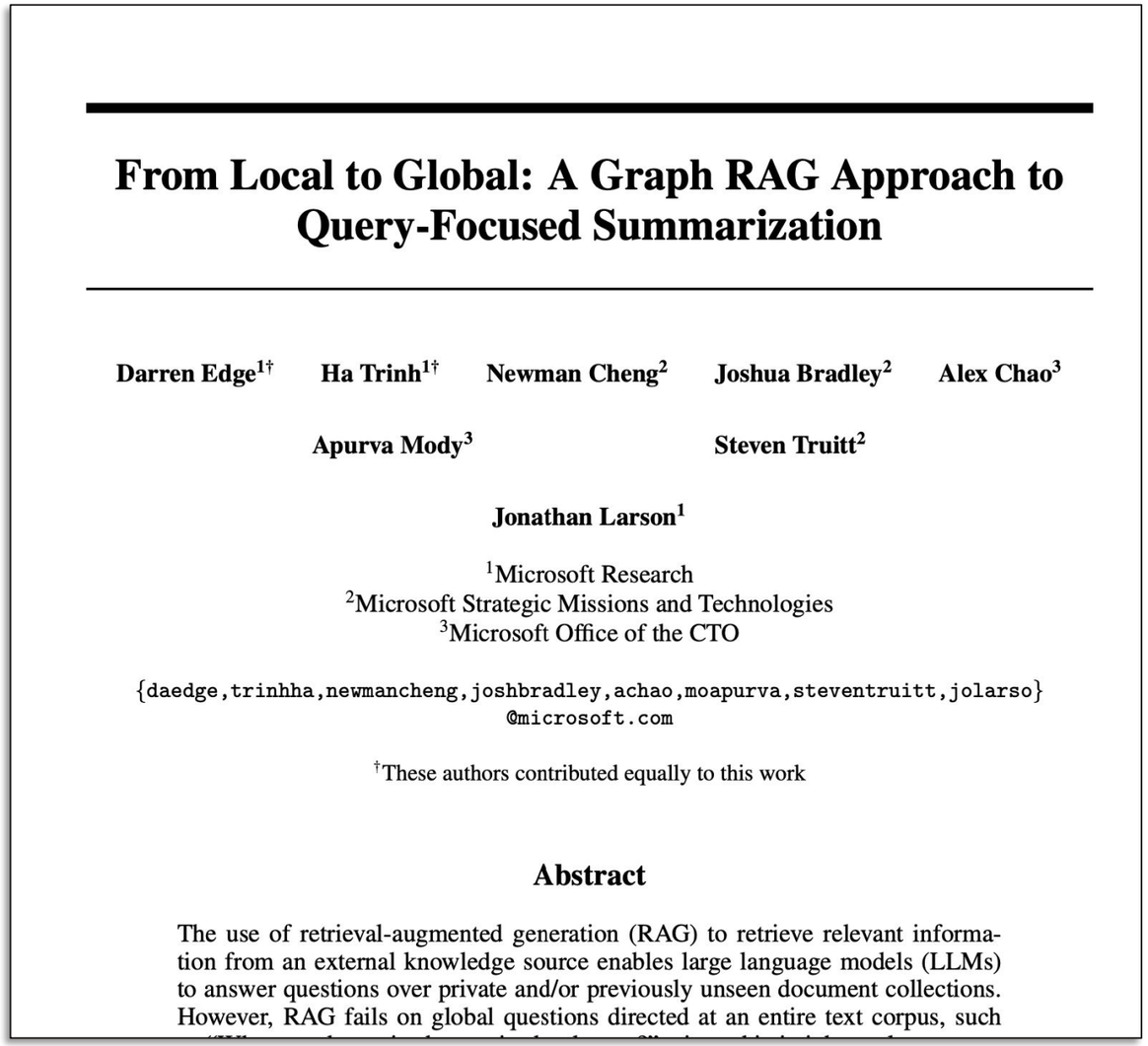
1. Chunk Documents
2. Create Embeddings
3. Load Chunks & Embeddings to the Database
4. Create a Vector Index

RAG Steps:

1. User Provides the User Query
2. Embed the User Query
3. Retrieve Relevant Documents
4. Prompt the LLM with User Query and Relevant Documents
5. Provide Answer to User



GraphRAG



From Local to Global: A Graph RAG Approach to Query-Focused Summarization

Darren Edge[1†] **Ha Trinh**[1†] **Newman Cheng**[2] **Joshua Bradley**[2] **Alex Chao**[3]

Apurva Mody[3] **Steven Truitt**[2]

Jonathan Larson[1]

[1]Microsoft Research
[2]Microsoft Strategic Missions and Technologies
[3]Microsoft Office of the CTO

{daedge,trinhha,newmancheng,joshbradley,achao,moapurva,steventruitt,jolarso}
@microsoft.com

[†]These authors contributed equally to this work

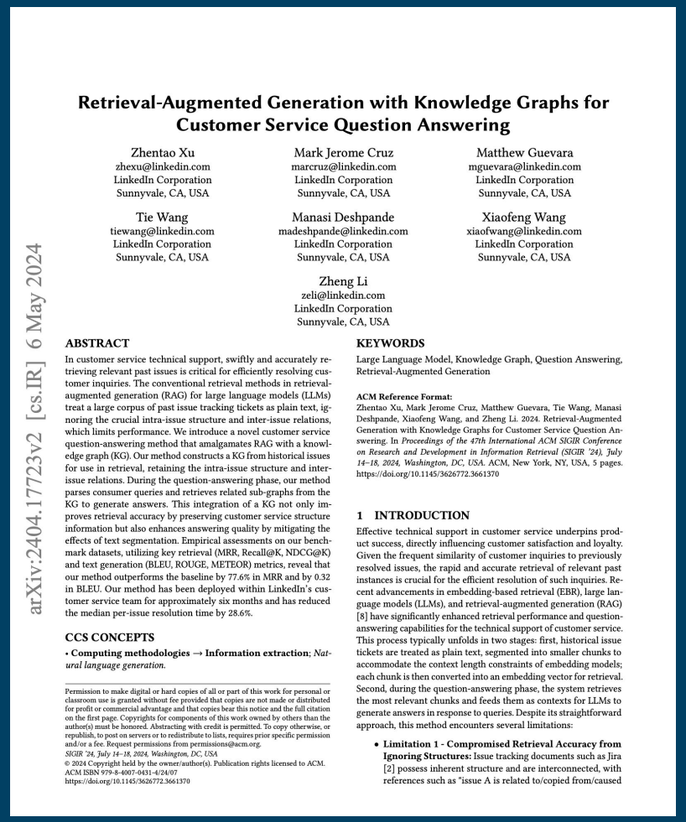
Abstract

The use of retrieval-augmented generation (RAG) to retrieve relevant information from an external knowledge source enables large language models (LLMs) to answer questions over private and/or previously unseen document collections. However, RAG fails on global questions directed at an entire text corpus, such

GraphRAG

Advanced RAG Patterns that use Graph Data Structures for Retrieval for relevant context and higher explainability.

Patterns



arXiv:2404.17723v2 [cs.IR] 6 May 2024

Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering

Zhentao Xu
zhexu@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Mark Jerome Cruz
marcruz@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Matthew Guevara
mguevara@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Tie Wang
tiewang@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Manasi Deshpande
madeshpande@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Xiaofeng Wang
xiaofwang@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

Zheng Li
zeli@linkedin.com
LinkedIn Corporation
Sunnyvale, CA, USA

ABSTRACT

In customer service technical support, swiftly and accurately retrieving relevant past issues is critical for efficiently resolving customer inquiries. The conventional retrieval methods in retrieval-augmented generation (RAG) for large language models (LLMs) treat a large corpus of past issue tracking tickets as plain text, ignoring the crucial intra-issue structure and inter-issue relations, which limits performance. We introduce a novel customer service question-answering method that amalgamates RAG with a knowledge graph (KG). Our method constructs a KG from historical issues for use in retrieval, retaining the intra-issue structure and inter-issue relations. During the question-answering phase, our method parses consumer queries and retrieves related sub-graphs from the KG to generate answers. This integration of a KG not only improves retrieval accuracy by preserving customer service structure information but also enhances answering quality by mitigating the effects of text segmentation. Empirical assessments on our benchmark datasets, utilizing key retrieval (MRR, Recall@K, NDCG@K) and text generation (BLEU, ROUGE, METEOR) metrics, reveal that our method outperforms the baseline by 77.6% in MRR and by 0.32 in BLEU. Our method has been deployed within LinkedIn's customer service team for approximately six months and has reduced the median per-issue resolution time by 28.6%.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Natural language generation.*

KEYWORDS

Large Language Model, Knowledge Graph, Question Answering, Retrieval-Augmented Generation

ACM Reference Format:

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3661370

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3661370

1 INTRODUCTION

Effective technical support in customer service underpins product success, directly influencing customer satisfaction and loyalty. Given the frequent similarity of customer inquiries to previously resolved issues, the rapid and accurate retrieval of relevant past instances is crucial for the efficient resolution of such inquiries. Recent advancements in embedding-based retrieval (EBR), large language models (LLMs), and retrieval-augmented generation (RAG) [8] have significantly enhanced retrieval performance and question-answering capabilities for the technical support of customer service. This process typically unfolds in two stages: first, historical issue tickets are treated as plain text, segmented into smaller chunks to accommodate the context length constraints of embedding models; each chunk is then converted into an embedding vector for retrieval. Second, during the question-answering phase, the system retrieves the most relevant chunks and feeds them as contexts for LLMs to generate answers in response to queries. Despite its straightforward approach, this method encounters several limitations:

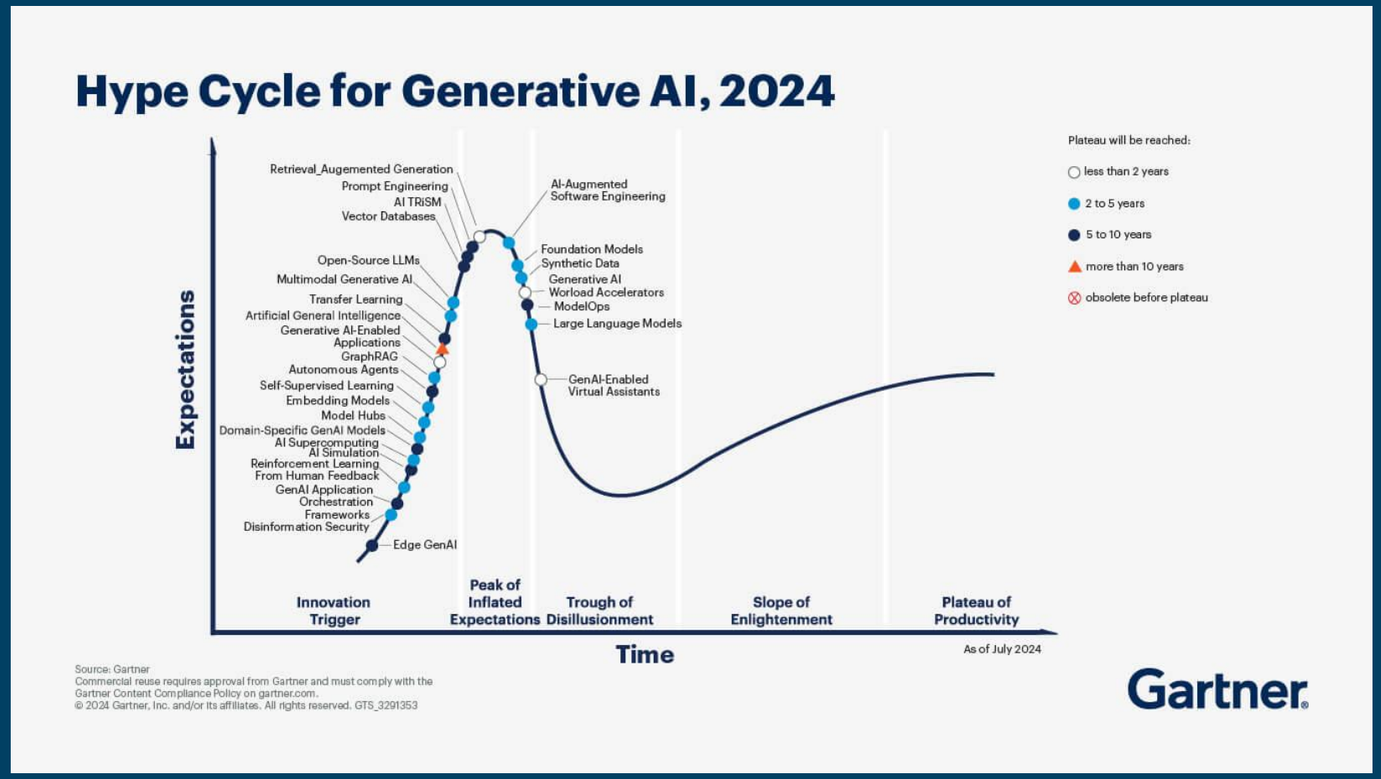
- **Limitation 1 - Compromised Retrieval Accuracy from Ignoring Structures:** Issue tracking documents such as Jira [2] possess inherent structure and are interconnected, with references such as "issue A is related to/copied from/caused

*We introduce a novel customer service question-answering method that **amalgamates RAG with a knowledge graph (KG)**. (...)*

*Empirical assessments on our benchmark datasets, utilizing key retrieval and text generation metrics, **reveal that our method outperforms the baseline by 77.6% in MRR and by 0.32 in BLEU**. Our method has been deployed within LinkedIn's customer service team for approximately six months and has reduced the median per-issue resolution time by 28.6%.*

Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering - LinkedIn

Gartner®



Gartner Hype Cycle Generative AI, November 2024

***GraphRAG**: This technique improves the accuracy, reliability and explainability of retrieval-augmented generation (RAG) systems. The approach uses knowledge graphs (KGs) to improve the recall and precision of retrieval, either directly by pulling facts from a KG or indirectly by optimizing other retrieval methods.*

Example Pattern

Name: Graph Enhanced Vector Search

Description: The user question is embedded using the same embedder used to create chunk embeddings. A vector similarity search is executed on the chunk embeddings to find k (number previously configured by developer/user) most similar chunks. A traversal of the Domain Graph starting at the found chunks is executed to retrieve more context.

Context: The biggest problem with basic GraphRAG patterns is finding all relevant context necessary to answer a question. The context can be spread across many chunks not being found by the search. Relating the real-world entities from the chunks to each other and retrieving these relationships together with a vector search provides additional context about these entities that the chunks refer to. They can also be used to relate chunks to each other through the entity network.

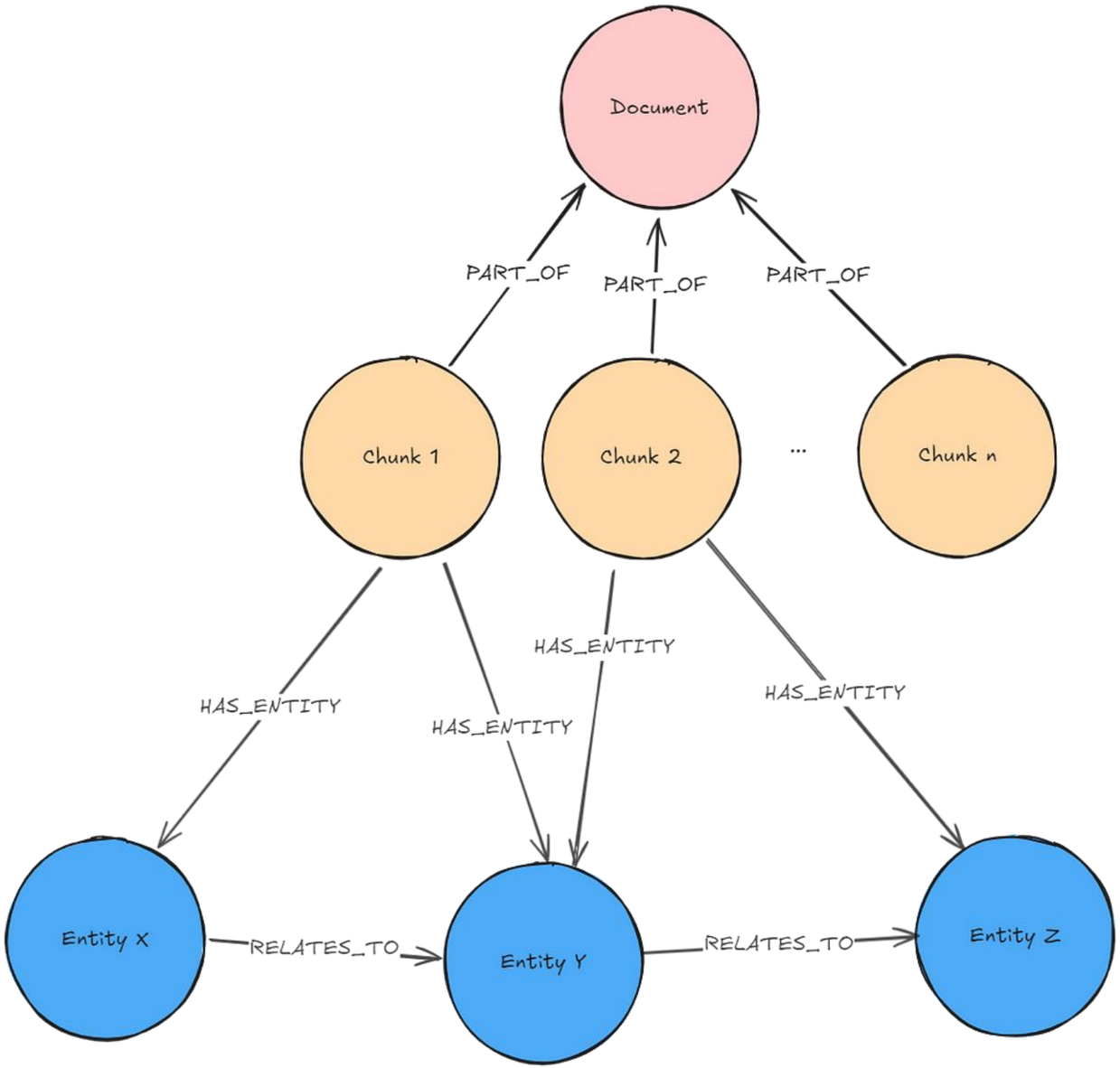
Required pre-processing: Use an LLM to execute entity and relationship extraction on the chunks. Import the retrieved triples into the graph.

Variations: Entity disambiguation, Question-guided/Schema-defined extraction, Entity embeddings, Ontology-driven traversal

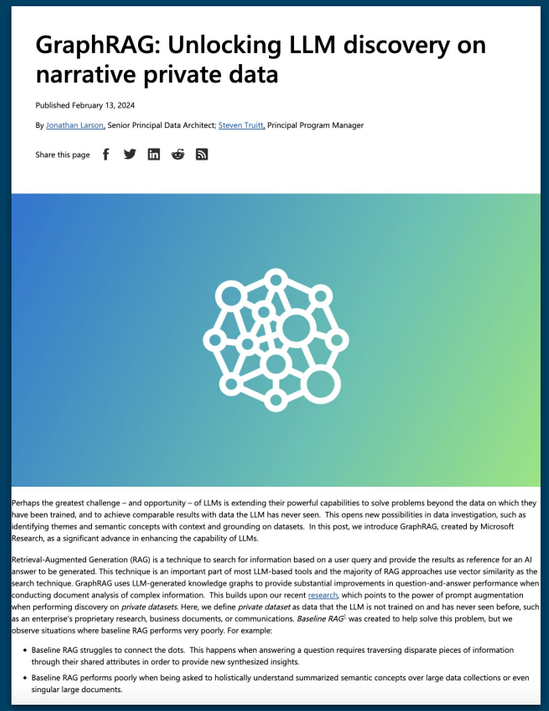
```
MATCH (node)-[:PART_OF]->(d:Document)
MATCH (node)-[:__HAS_ENTITY__]->(e)
MATCH path=(e)(()-[rels:!HAS_ENTITY&!PART_OF]-()){0,2}(:!Chunk&!Document)
...
RETURN ...
```

AKA: Graph + Vector, Augmented Vector Search

Required graph pattern: Lexical Graph with Extracted Entities



Graph Enhanced Vector Search



GraphRAG: Unlocking LLM discovery on narrative private data

Published February 13, 2024

By Jonathan Larson, Senior Principal Data Architect; Steven Truitt, Principal Program Manager

Share this page

Perhaps the greatest challenge – and opportunity – of LLMs is extending their powerful capabilities to solve problems beyond the data on which they have been trained, and to achieve comparable results with data the LLM has never seen. This opens new possibilities in data investigation, such as identifying themes and semantic concepts with context and grounding on datasets. In this post, we introduce GraphRAG, created by Microsoft Research, as a significant advance in enhancing the capability of LLMs.

Retrieval-Augmented Generation (RAG) is a technique to search for information based on a user query and provide the results as reference for an AI answer to be generated. This technique is an important part of most LLM-based tools and the majority of RAG approaches use vector similarity as the search technique. GraphRAG uses LLM-generated knowledge graphs to provide substantial improvements in question-and-answer performance when conducting document analysis of complex information. This builds upon our recent research, which points to the power of prompt augmentation when performing discovery on *private datasets*. Here, we define *private dataset* as data that the LLM is not trained on and has never seen before, such as an enterprise's proprietary research, business documents, or communications. *Baseline RAG*[2] was created to help solve this problem, but we observe situations where baseline RAG performs very poorly. For example:

- Baseline RAG struggles to connect the dots. This happens when answering a question requires traversing disparate pieces of information through their shared attributes in order to provide new synthesized insights.
- Baseline RAG performs poorly when being asked to holistically understand summarized semantic concepts over large data collections or even singular large documents.

Microsoft - GraphRAG

*Initial results show that GraphRAG **consistently outperforms** baseline RAG.*

Steps:

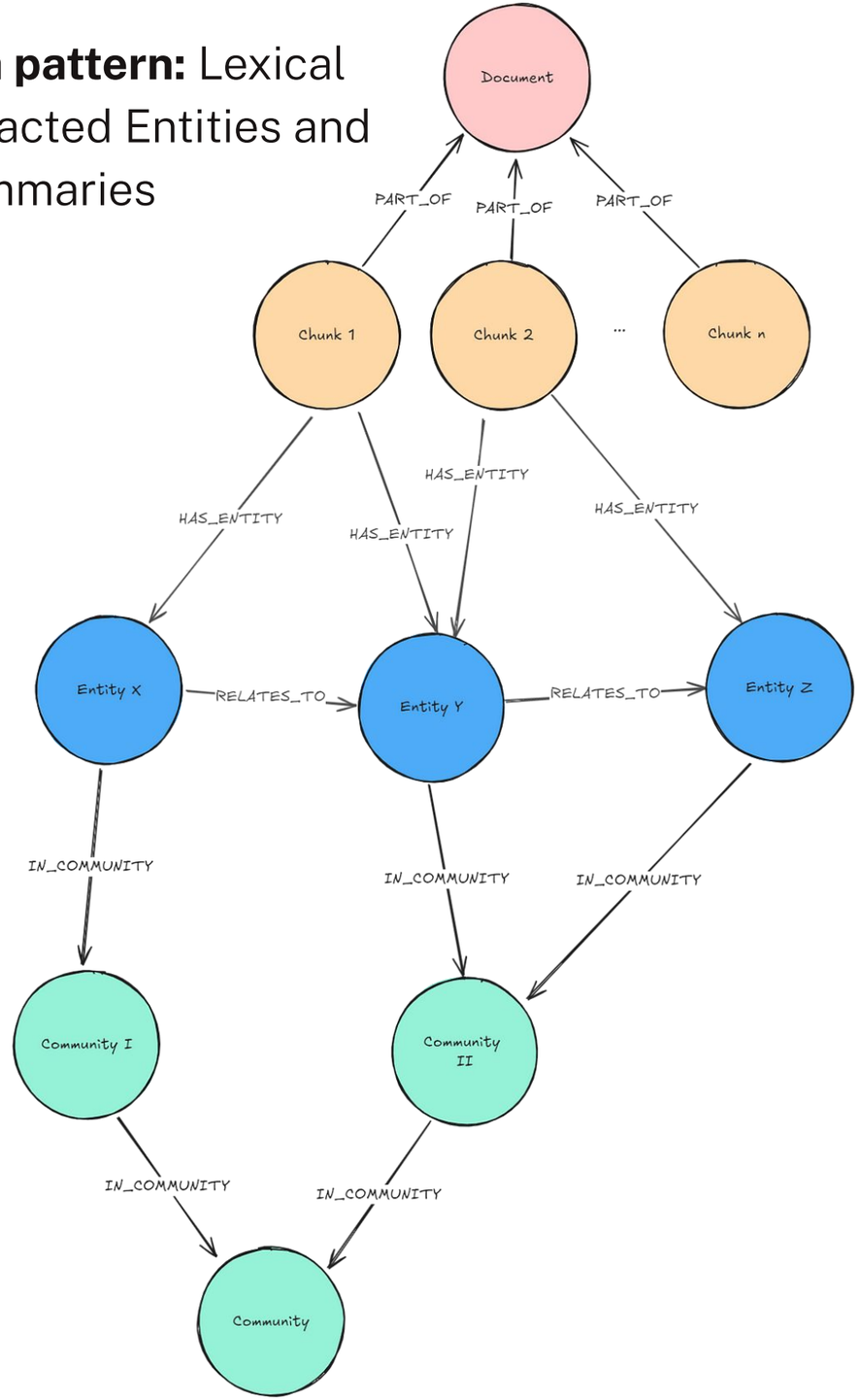
1. *Ingest Text Data*
2. *Generate Knowledge Graph*
3. *Import Into Graph Database*
4. *Create Semantic Hierarchies*
5. *Augment Retrieval*
6. *Deeper Understanding*

Global Community Summary Retriever

AKA: *Microsoft GraphRAG*, Global Retriever
Context: Certain ***questions that can be asked on a whole dataset*** do not just relate to things present in some chunks but rather search for an overall message that is overarching in the dataset.
Required pre-processing: In addition to extracting entities and their relationships, we need to form hierarchical communities within the Domain Graph. For every community, an LLM summarizes the entity and relationship information into Community Summaries.

```
MATCH (c:__Community__)
WHERE c.level = $level
RETURN c.full_content AS output
```

Required graph pattern: Lexical Graph with Extracted Entities and Community Summaries

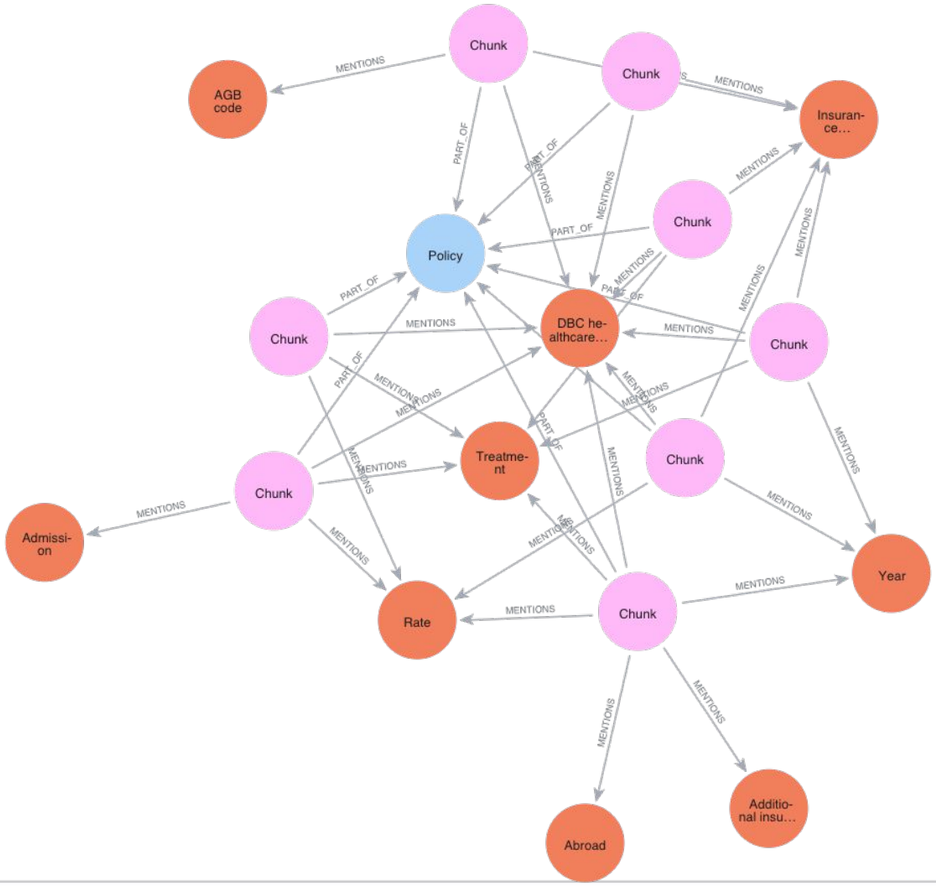


Add Domain Knowledge

- There are references in the document that give more context
- These add more context to the chunks

See also:

- Abroad (definitions)
- DBC healthcare product (definitions)
- Rate (definitions)
- Treatment (definitions)
- Year (definitions)



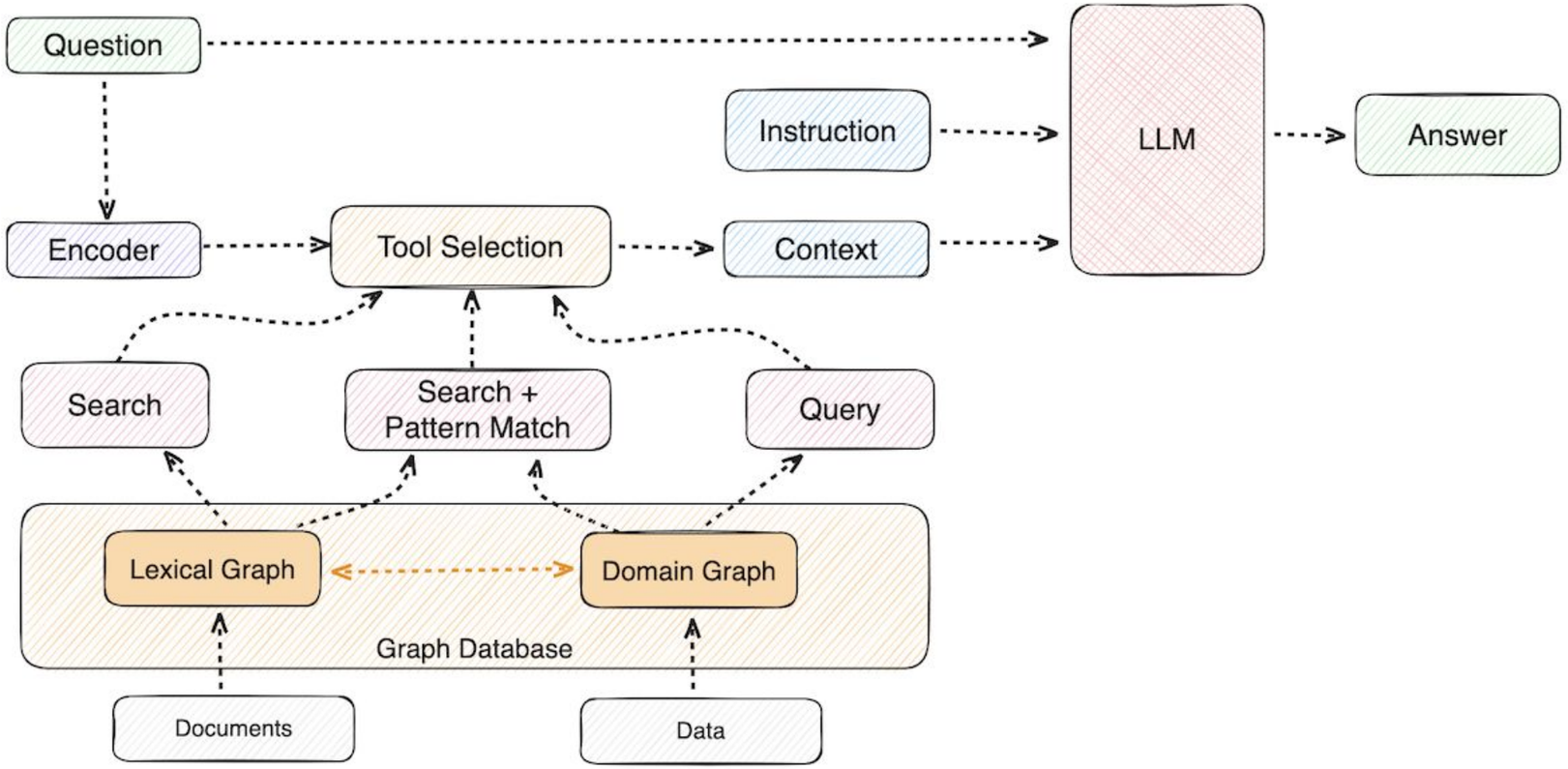

```
1 MATCH (c:Chunk)-[:MENTIONS]-(d:Definition)
2 WHERE c.id in $chunk_ids
3 WITH DISTINCT d as d
4 RETURN d.definition as definition, d.description as description
```


Module 4
Go to Notebook

Module 5

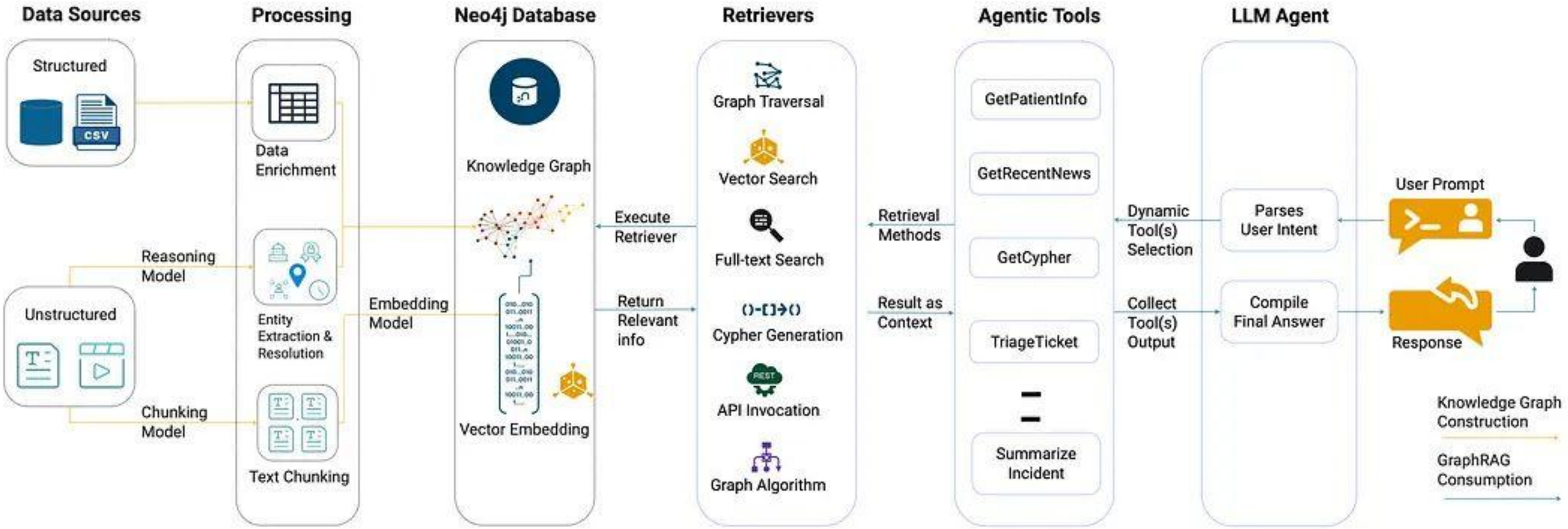
GraphRAG Agent: Create Agents with Tools

What is GraphRAG?



[Blog - What is GraphRAG?]

Agentic GraphRAG Architecture



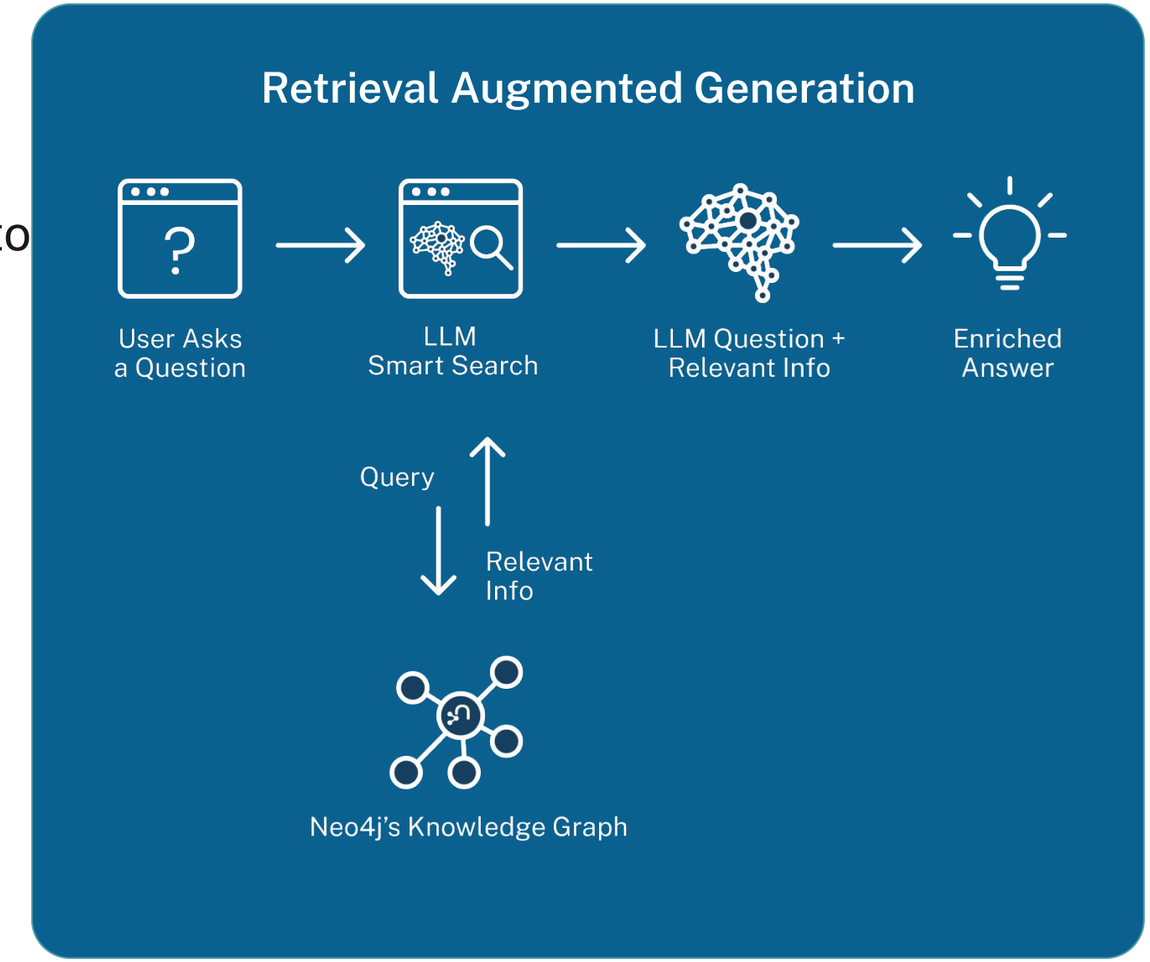
Module 5
Go to Notebook

Wrap Up!

Neo4j Knowledge Graph as Database of Truth

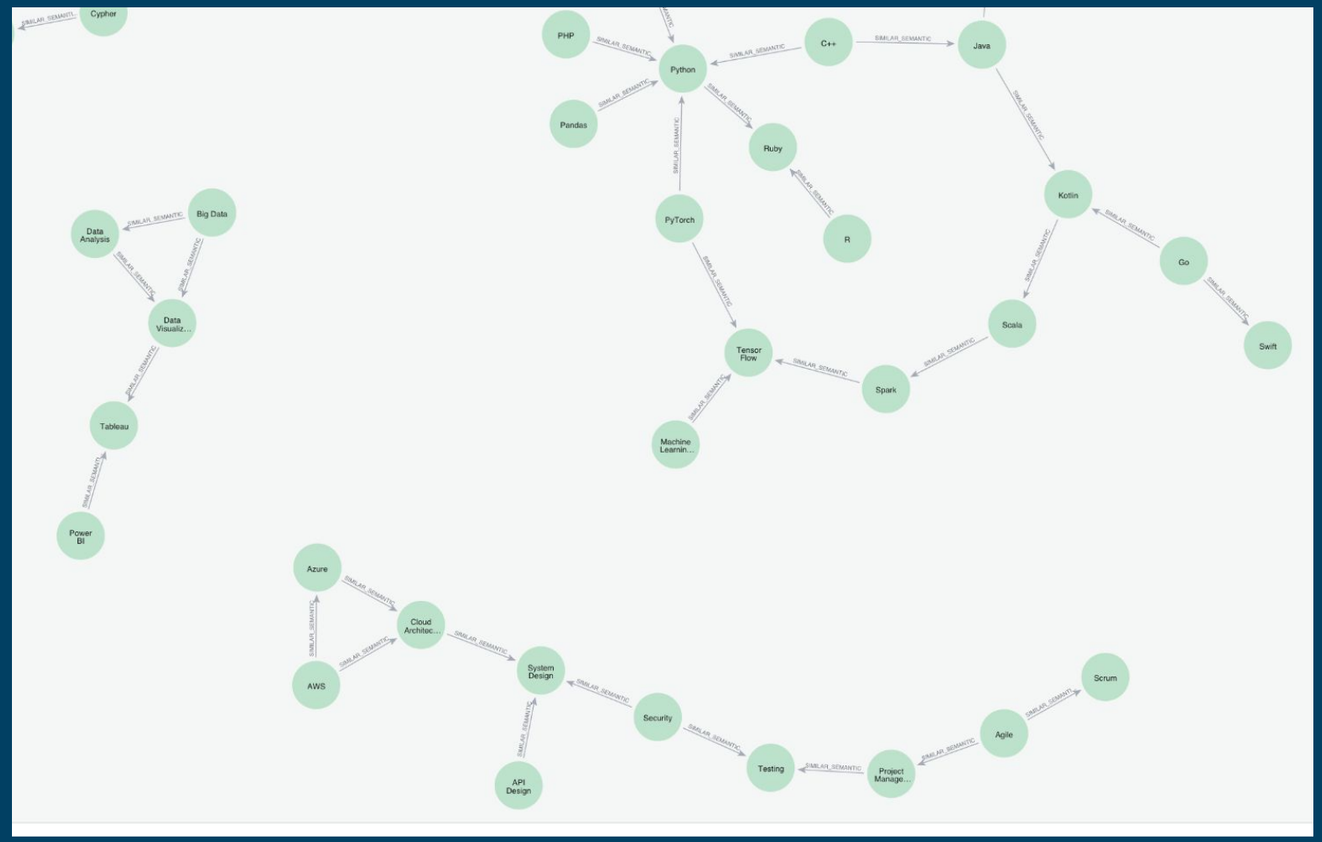
A Neo4j Knowledge Graph combined with LLM's obtain improvements:

- **Accuracy** - Obtain better answers compared to plain vector searches
- **Explainability**: Provide the user with more reasoning on how the results were obtained.
- **Acceleration**: Having all the capabilities in one platform increases understanding and decreases time to value.



Rabobank
neo4j

Easier Development



Feedback from an AI Engineer

Here is the PR with the changes

I kinda replicated the same action-based cache already in place for Pinecone but thanks to the graph nature of Neo4J most of the operations yield better results:

- thanks to the `Neo4j graph data science` plugin we can store embeddings and calculate cosine similarity at the database level
- getting related actions is as simple as following the relationships between nodes
- **the cache can be visualised**. This is an extremely valuable debugging tool for us to understand if/when and how the cache might be broken/misbehaving (I actually already fixed a couple of bugs just thanks to this)

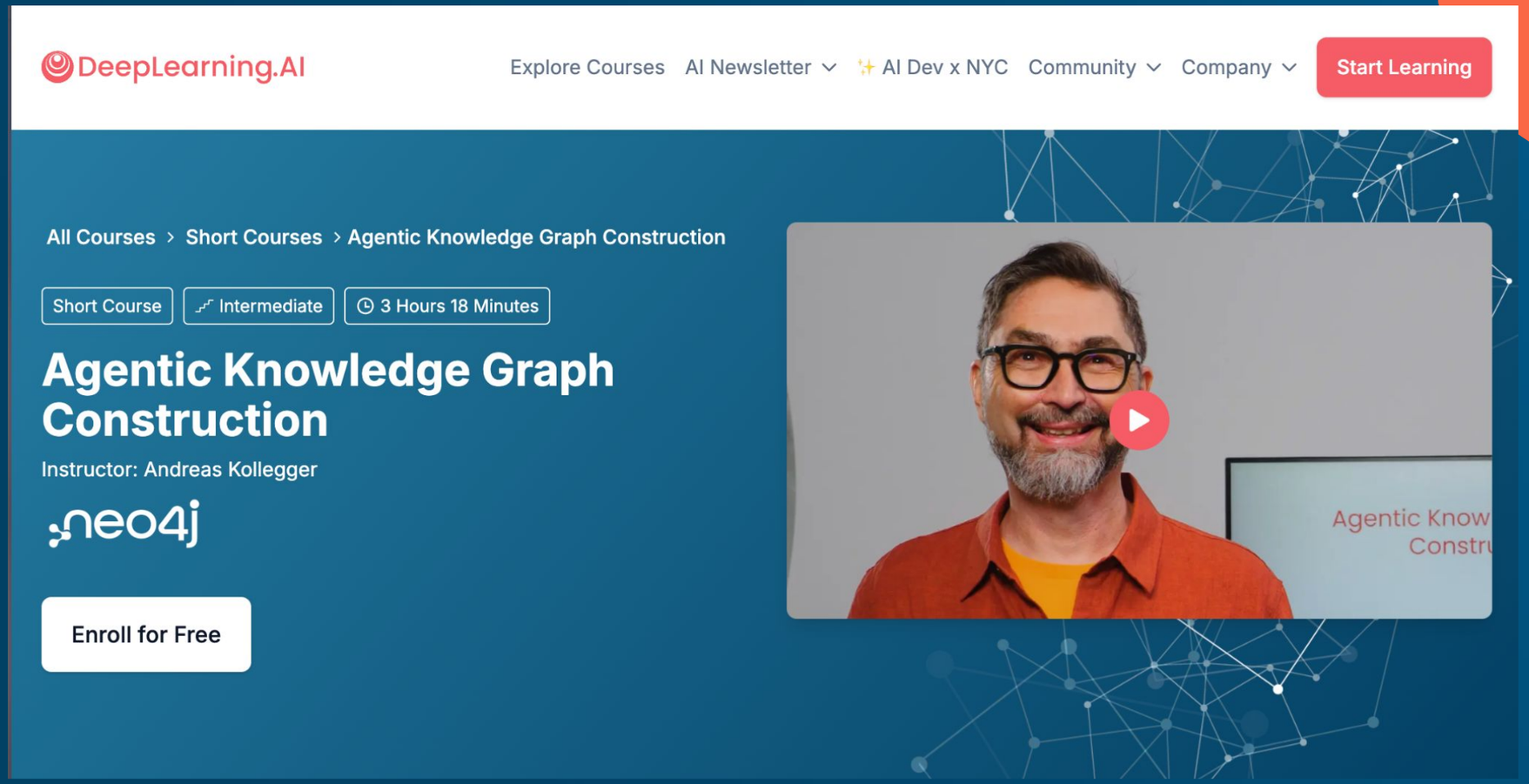
GraphRAG Resources

Github Repository with this Workshop

Free Online Graph Academy Courses,
Videos & Webinars

Developer Guides and Coded Examples

GraphRAG Resources




Thank you!