

# Masterclass: RAG mit InterSystems IRIS



Andreas Schuetz  
Sales Engineer



Felix Vetter  
Sales Engineer Intern



GitHub Repo mit Code:

<https://github.com/intersystems-dach/RAG-Demo>



**Artificial Intelligence**

**Machine Learning**

Image Recognition

Survival Analysis

...

**Tabular ML**

Classification

Regression

Clustering

...

**Deep Learning**

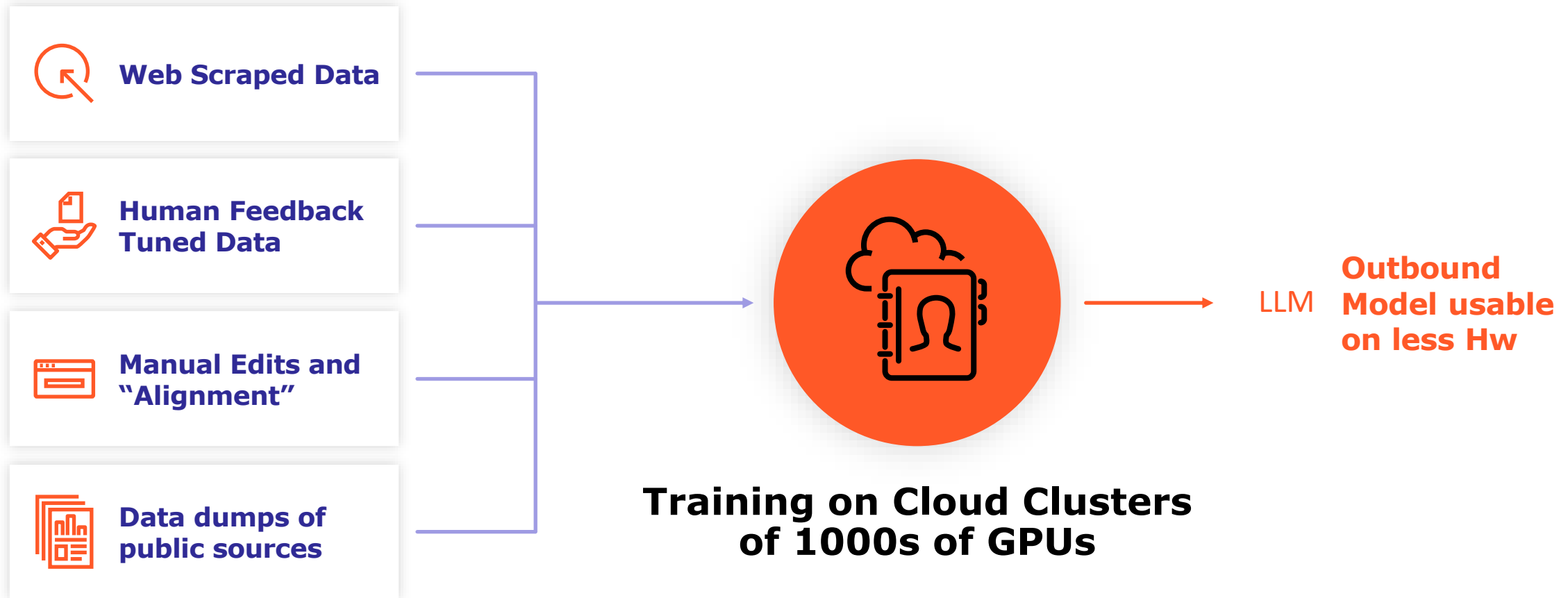
**Generative AI**

Large Language Models

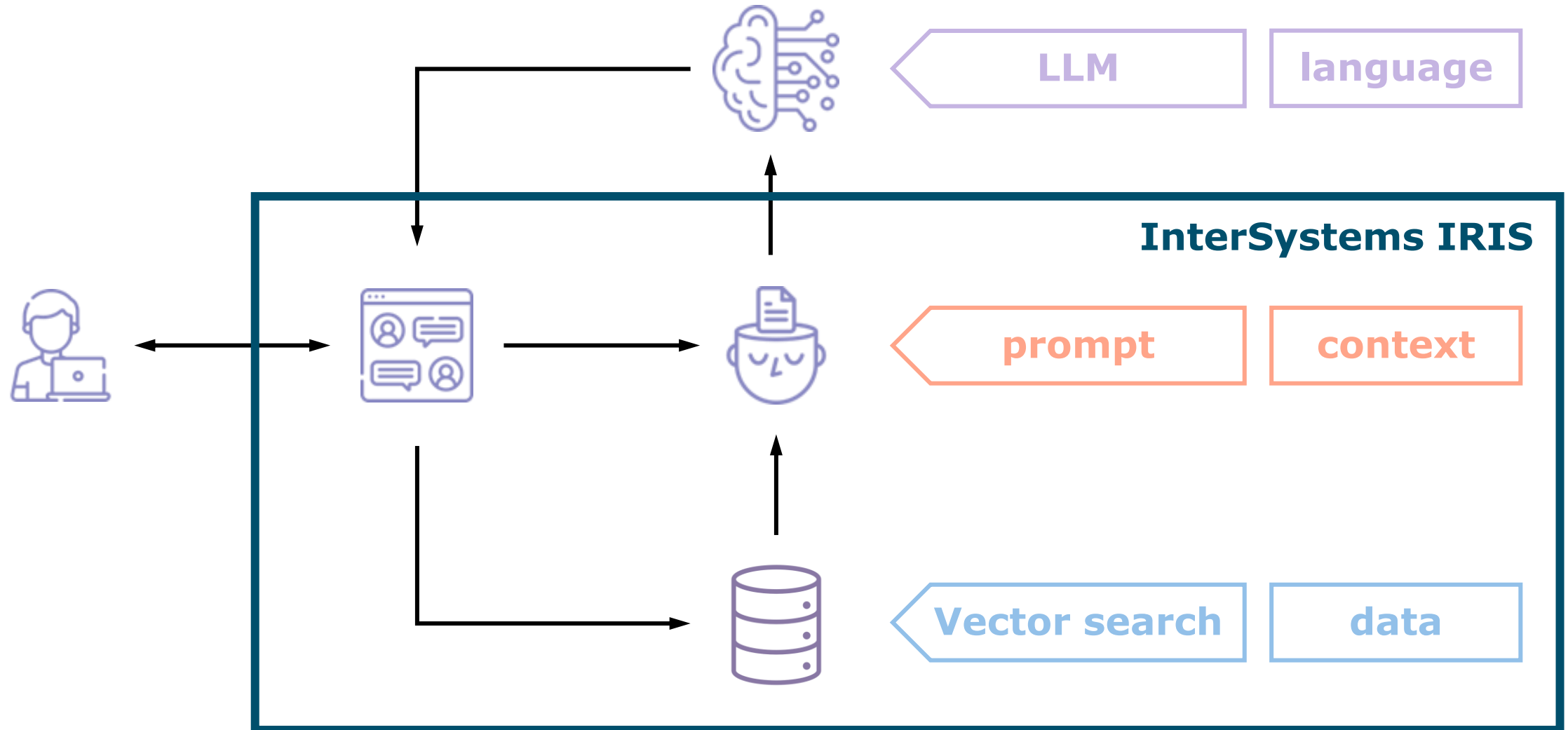
Vector Search

...

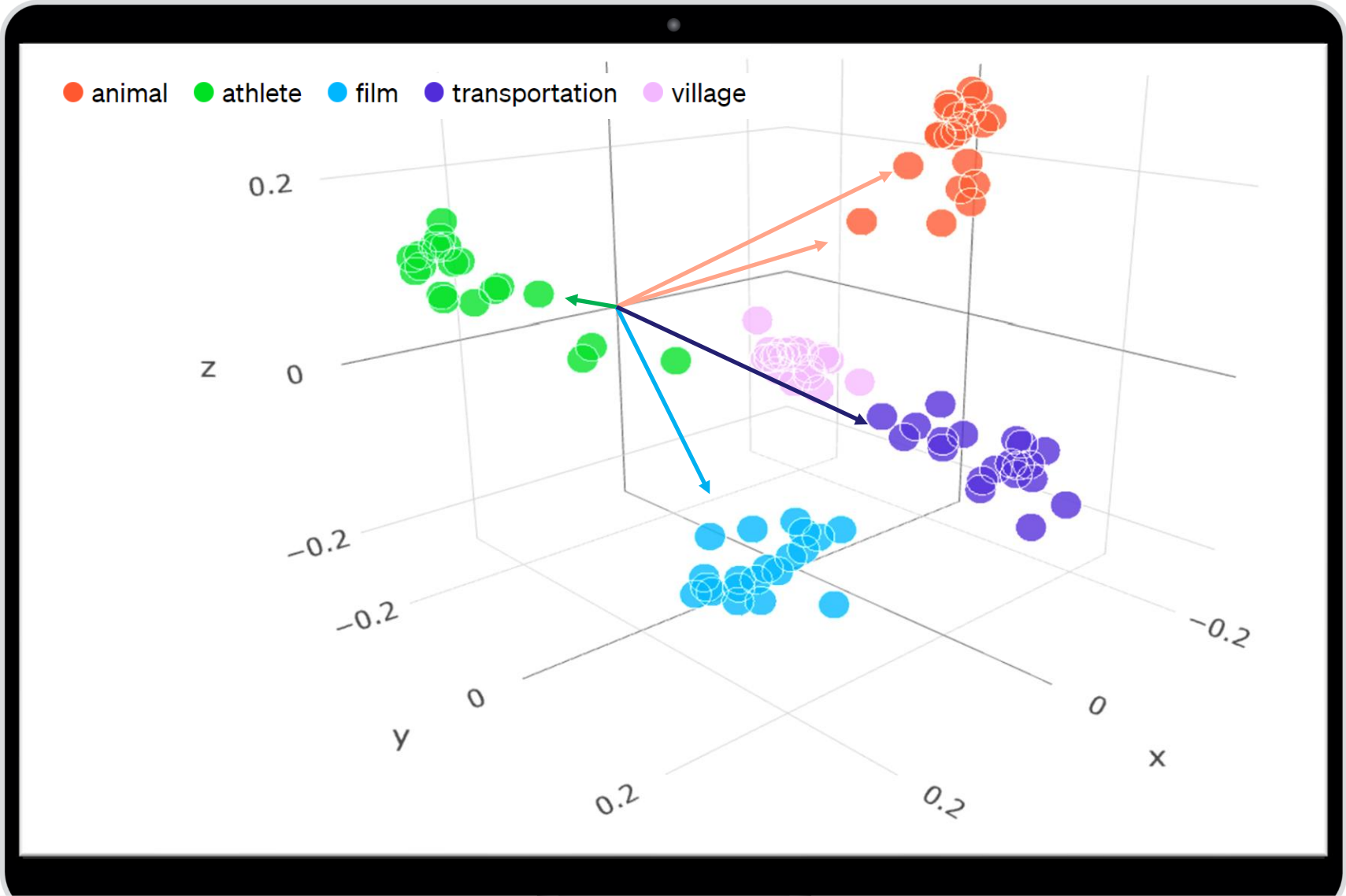
# LLMs are trained on a WIDE variety of sources

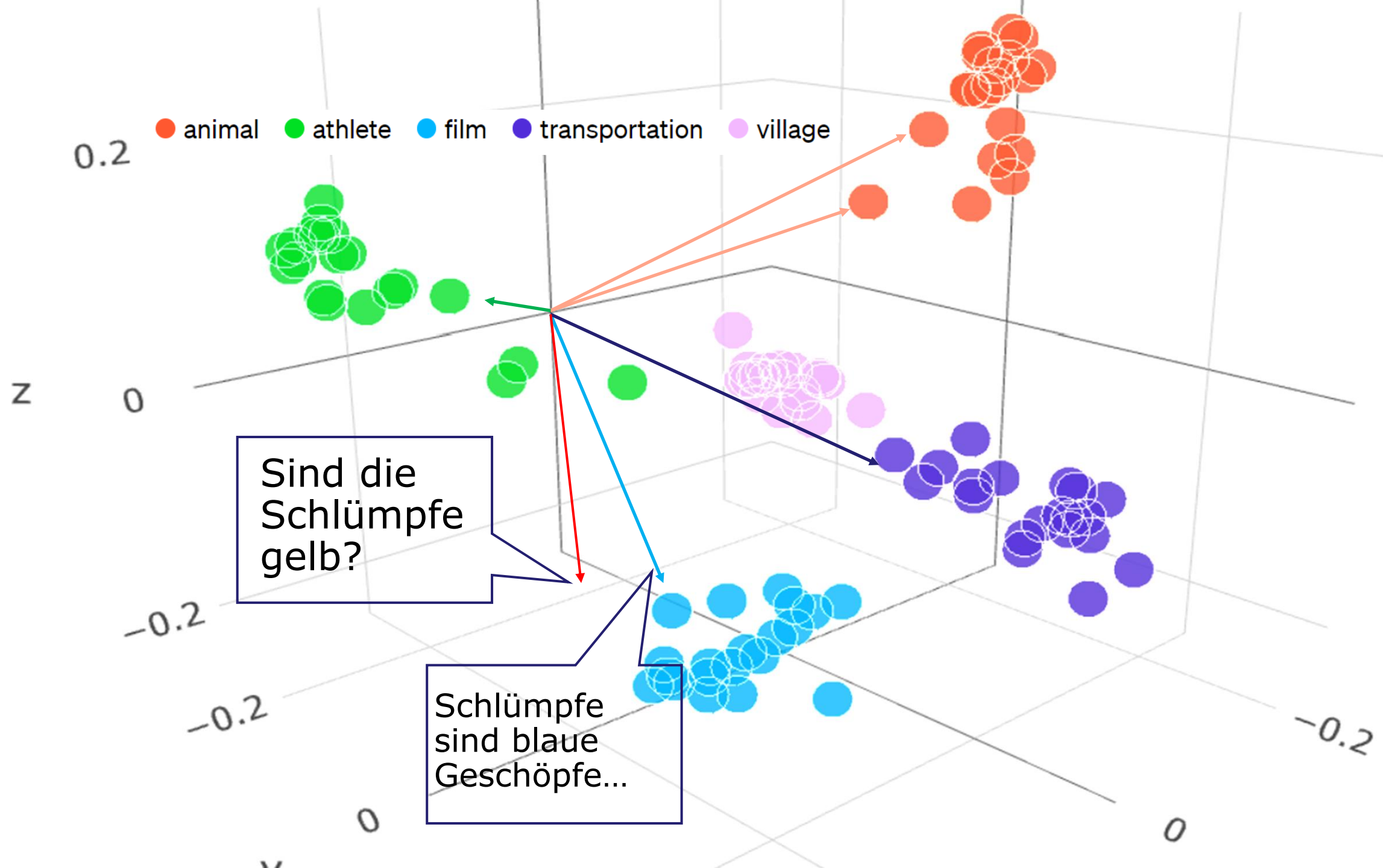


# Retrieval Augmented Generation



# Sentences & context as vectors

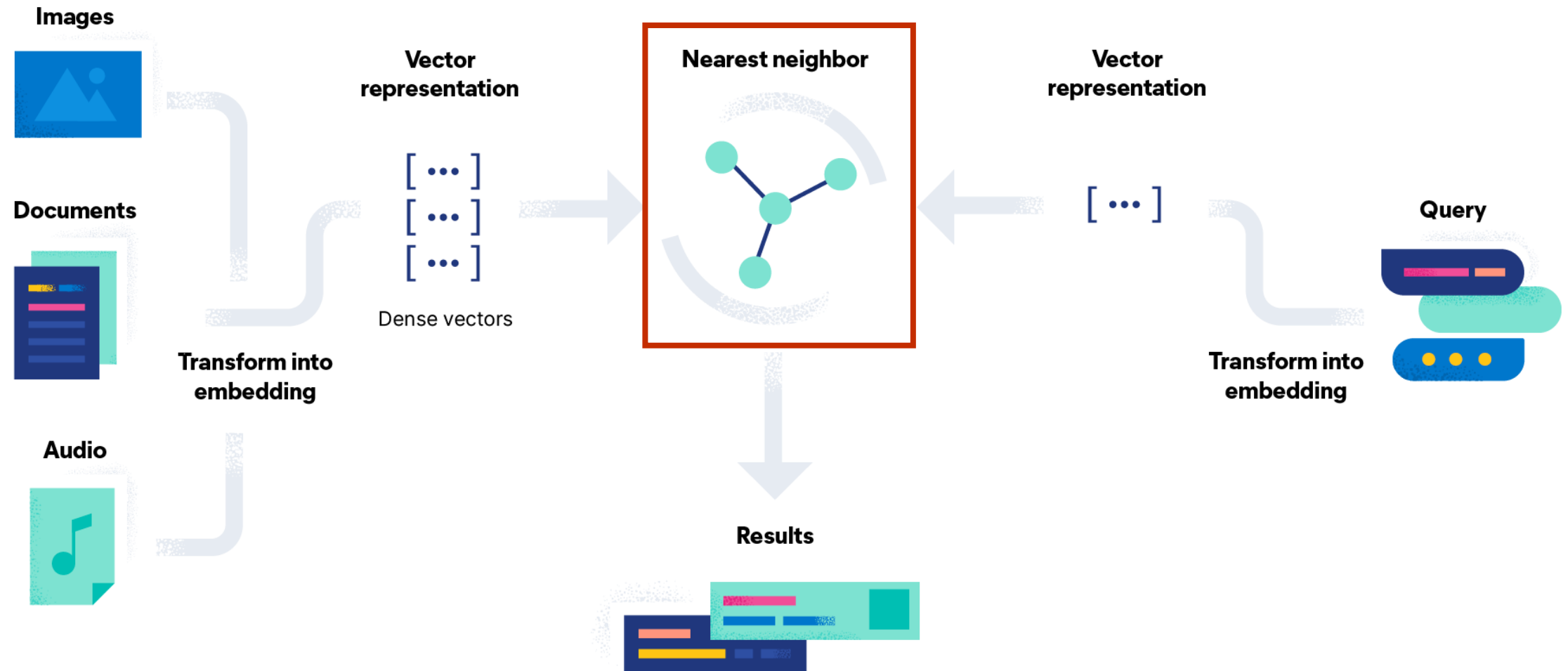




# Vector Search

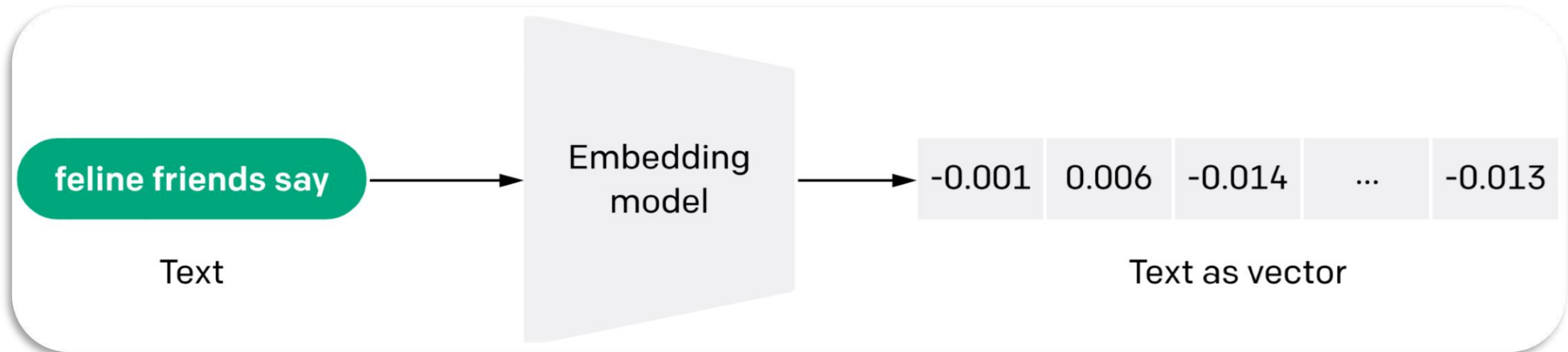
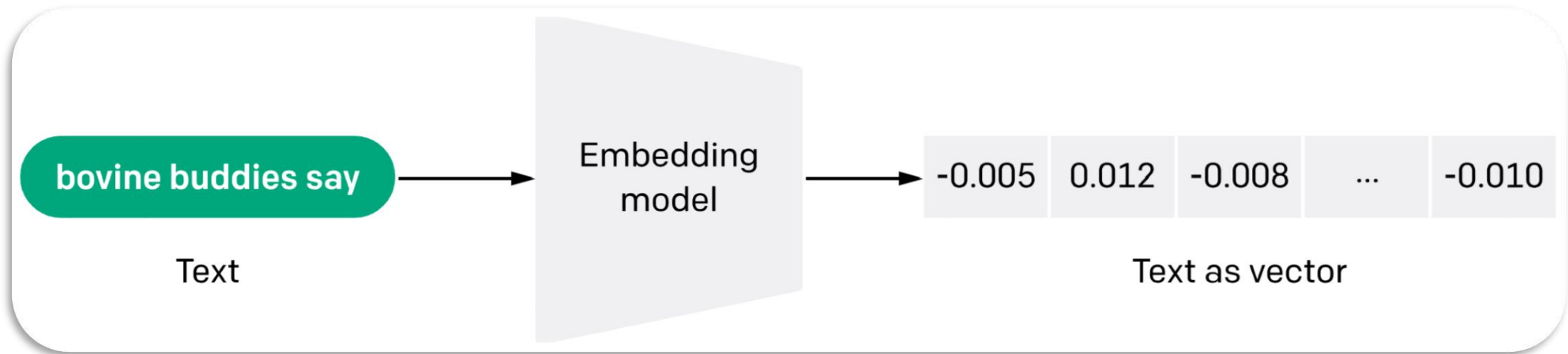


- **Vector search** stellt Datenelemente durch numerische Vektoren dar, was die Indizierung und Ausführung von Abfragen über numerische Inhalte ermöglicht





# Word Embeddings



# Chunking & Overlapping



Upload .txt

Splitter: Character Splitter 

Chunk Size: 29 

Chunk Overlap: 0 

Total Characters: 352

Number of chunks: 13

Average chunk size: 27.1

Logystic's logistics fleet: The company has several aircraft. Air transport cargo aircraft:  
A fleet of 20 Boeing 767 freighters and 10 Airbus A330-200Fs serving international routes.  
Containerised mobile warehouses: Innovative mobile storage units that can be rapidly deployed in  
disaster areas or used to increase storage capacity during peak periods.

Upload .txt

Splitter: Character Splitter 

Chunk Size: 252 

Chunk Overlap: 55 

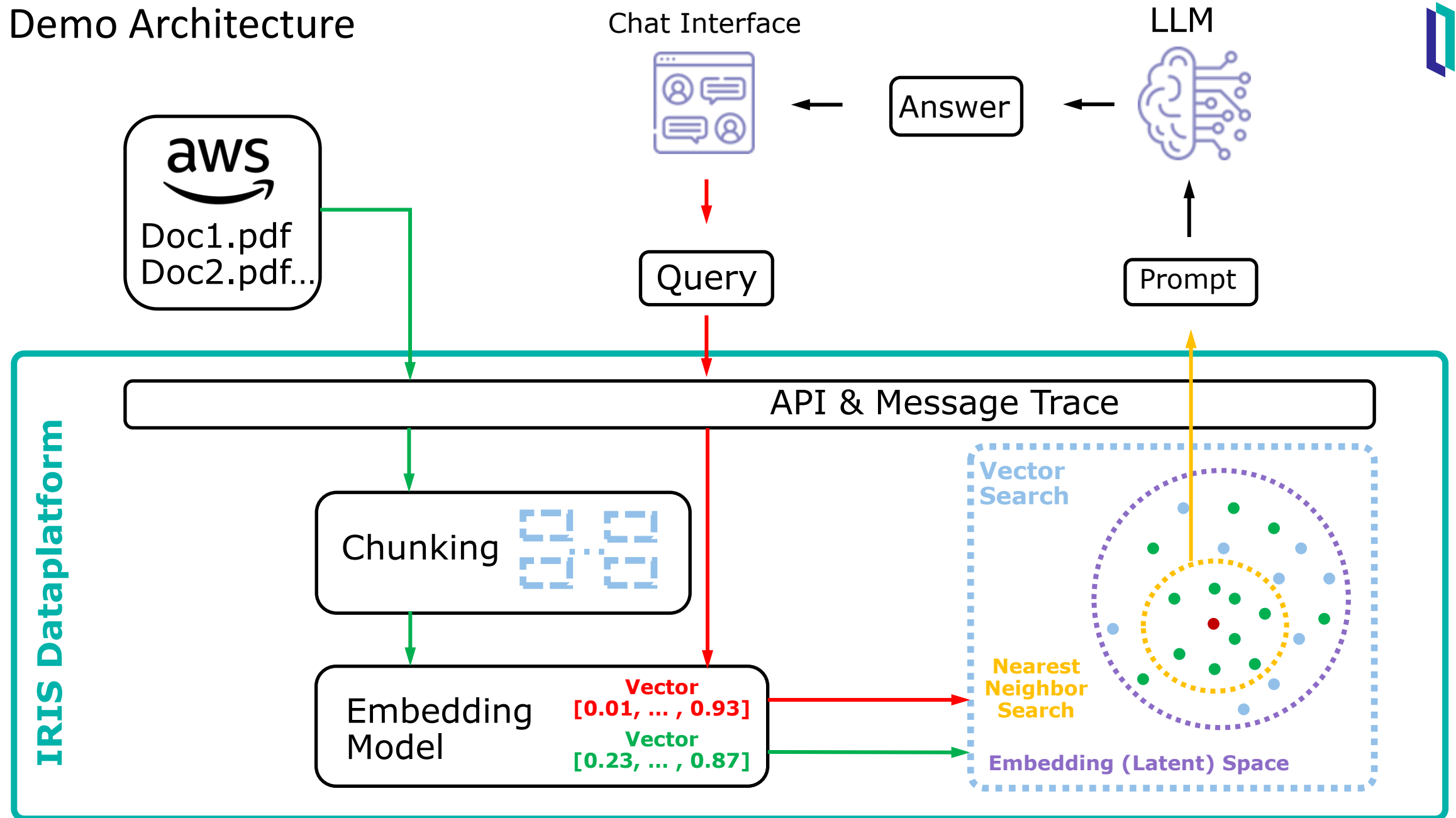
Total Characters: 407

Number of chunks: 2

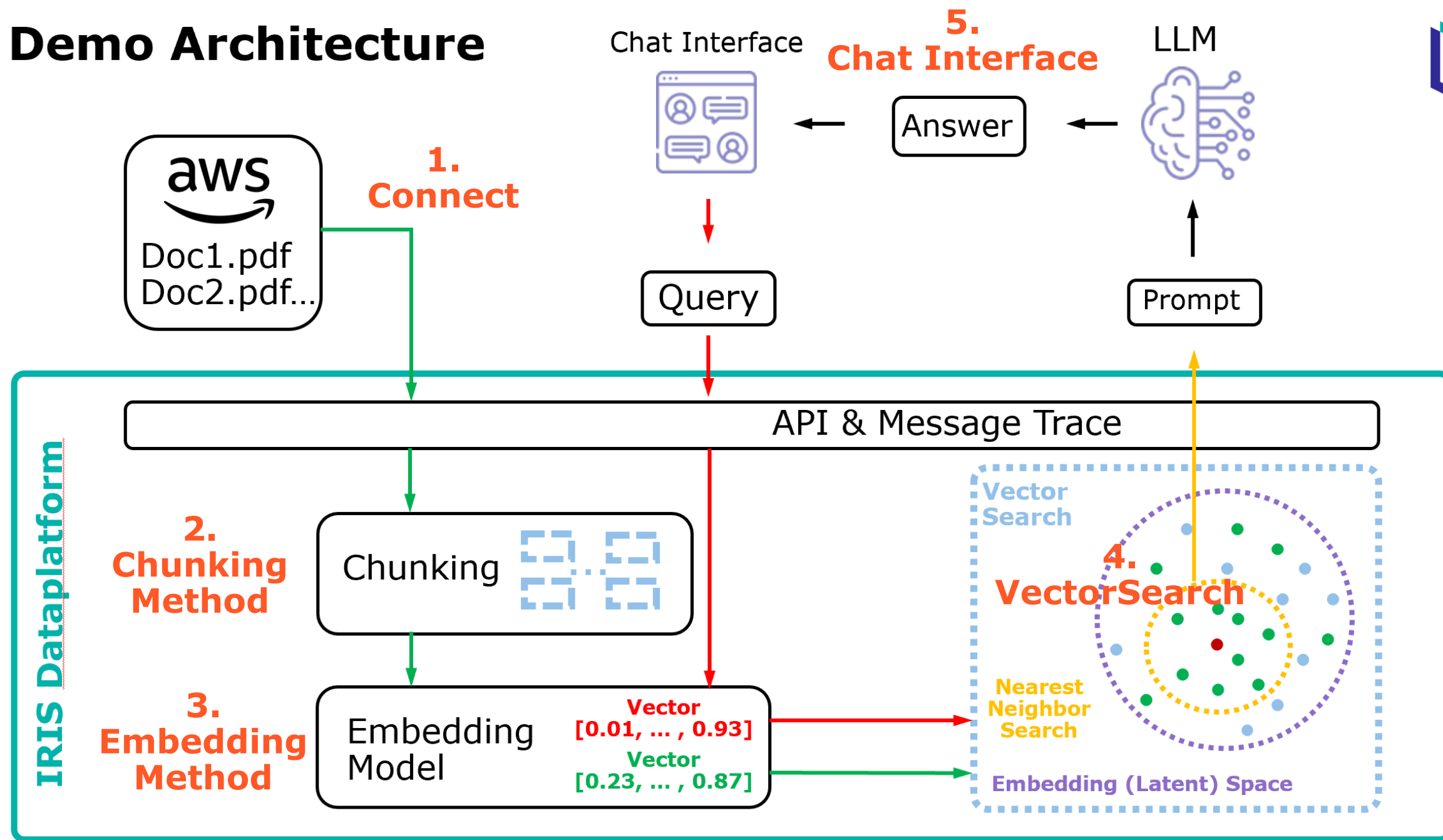
Average chunk size: 203.5

Logystic's logistics fleet: The company has several aircraft. Air transport cargo aircraft:  
A fleet of 20 Boeing 767 freighters and 10 Airbus A330-200Fs serving international routes.  
Containerised mobile warehouses: Innovative mobile storage units that can be rapidly deployed in  
disaster areas or used to increase storage capacity during peak periods.

# Demo Architecture



# Demo Architecture





## Aufgabe 0: Verbindung von S3 zu IRIS



```
if(self.Bucket != "" and self.Key != ""):

    from io import BytesIO
    from PyPDF2 import PdfReader
    import boto3
    from botocore import UNSIGNED
    from botocore.client import Config
    from botocore.exceptions import ClientError

    # Initialize a session using Amazon S3 with unsigned configuration
    s3 = boto3.client('s3', config=Config(signature_version=UNSIGNED))

    response = s3.get_object(Bucket=self.Bucket,Key=self.Key)

    reader = PdfReader(BytesIO(response['Body'].read()))

    text = ""

    for page in reader.pages:
        text += page.extract_text()

    return text
else:
    return ""
```



## Aufgabe 1: Die GetChunks() Methode



```
## Utils.cls Method GetChunks
```

```
from langchain.text_splitter import RecursiveCharacterTextSplitter
```

```
text_splitter = RecursiveCharacterTextSplitter(  
    chunk_size = tChunkSize,  
    chunk_overlap = tChunkOverlap  
)
```

```
docs = text_splitter.split_text(tText)
```

```
return docs
```

```
}
```



## Aufgabe 2: Die GetEmbeddingPy() Methode

- Gehe zu [Docs.Intersystems.com](https://docs.intersystems.com) und suche nach GetEmbeddings() Methode
- Kopiere es in die Demo.Utills Klasse



```
## Utils.cls Method GetEmbeddingPy
```

```
import json
```

```
# import the package
```

```
import sentence_transformers
```

```
# create the model and form the embeddings
```

```
model = sentence_transformers.SentenceTransformer('all-MiniLM-L6-v2')
```

```
embeddings = model.encode(sentences)
```

```
# convert the embeddings to a string
```

```
embeddings_list = [str(embedding.tolist()) for embedding in embeddings]
```

```
# print(embeddings_list[0])
```

```
return embeddings_list
```



## Aufgabe 3: Definiere das Property

- Gehe zu Demo.RecordEmbeddings.cls
- Erstelle ein Property names "Embedding" als Type %Vector.
- Die Länge sollte 384 sein.



```
Property Embedding As %Vector(LEN = 384);
```





## Aufgabe 4: Definier das SQL Statement

Baue das SQL Statement:

- Insert into Demo.RecordEmbeddings
- Wir wollen in DataSourceId,SourceId,Text,Embedding abspeichern
- Die Werte sind (tId, tRecordId, tChunk, tVector) **Beachte Speicherung von tVector**



```
&sql(insert into Demo.RecordEmbeddings (DataSourceId,SourceId,Text,Embedding)  
      values (:tId,:tRecordId,:tChunk,T0_VECTOR(:tVector)))
```

# Aufgabe 5: Das Chat Interface




Developer Roadshow

localhost:8501

Message Viewer SQL Production Configu... InterSystems IRIS D...

Deploy

# Welcome to Developer Roadshow

**InterSystems®**  
Creative data technology

User: How many boeings does logistics have? Only write the number

LLM: 20

You:



## Aufgabe 6: Die Vector Search implementieren

Baue das SQL Statement:

- Select the TOP 5 ID from the Demo.RecordEmbeddings
- Ordne absteigend nach dem Ergebnis von VectorSearch(Embedding, tVector)



```
SELECT TOP 5 ID FROM Demo.RecordEmbeddings  
ORDER BY VECTOR_DOT_PRODUCT(Embedding, TO_VECTOR(?)) DESC
```



## Aufgabe 7: Die Prompt Operation verstehen

## Aufgabe 8: Die OpenAI Operation verstehen

## Aufgabe 9: Starte die Production



## Aufgabe 10: Starten des StreamlitUI Servers

- Öffne den RAG-Masterclass auf dem Desktop
- Rechts Klick "Im Terminal öffnen"
- Führe "python -m streamlit run app.py" aus

Das UI sollte sich im Browser öffnen.

**Frage an den Chatbot: "Who is the venture capitalist and how much was invested?"**

# Aufgabe 11: Wir brauchen Kontext



- Starte den "Start PDF Import" Service durch Doppelklick auf den Service

**Production Configuration** Start Stop

**Production Stopped** Category: All Legend Production Settings

Services +	Processes +	Operations +
<input type="radio"/> Start PDF Import	<input type="radio"/> Injection Process	<input type="radio"/> Embedding Operation
<input type="radio"/> Streamlit Service	<input type="radio"/> LLM Router	<input type="radio"/> OpenAI Operation
		<input type="radio"/> PDF Operation
		<input type="radio"/> Prompt Operation
		<input type="radio"/> RAG Operation

## Aufgabe 12: Let's RAG



**Frage an den Chatbot erneut:**

- **“Who is the venture capitalist and how much was invested?”**



Thank you