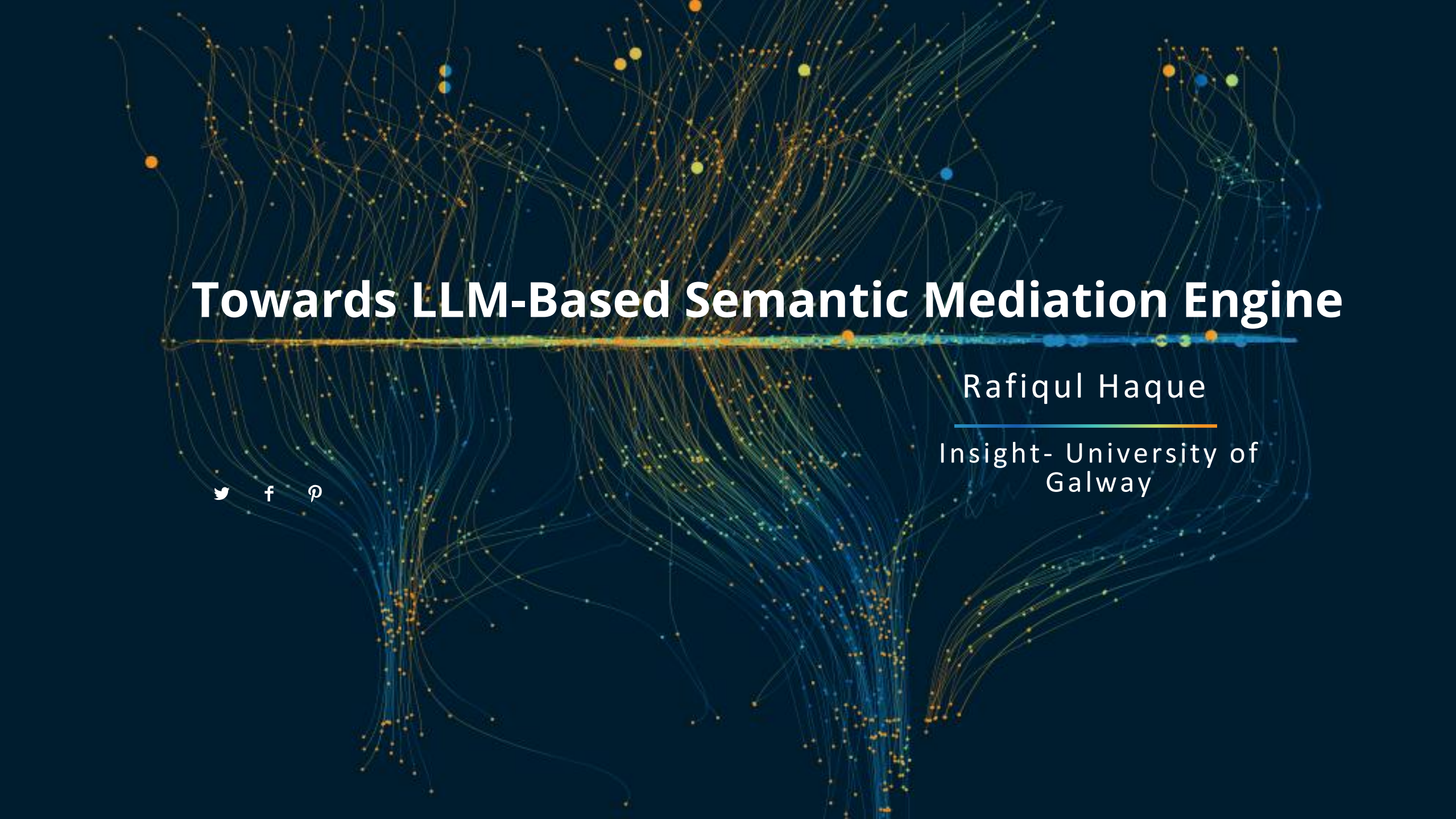


Towards LLM-Based Semantic Mediation Engine

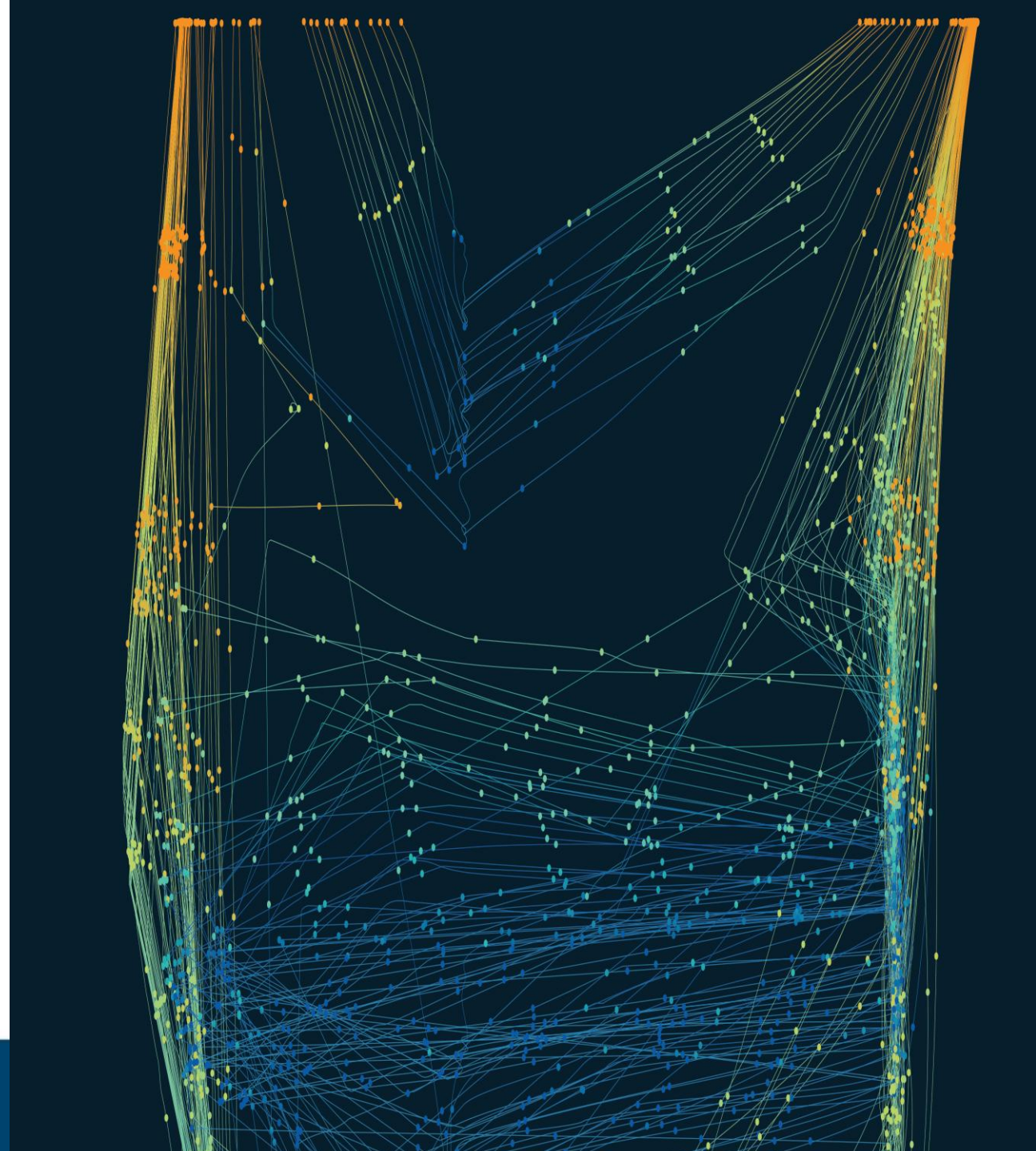
Rafiqul Haque

Insight- University of Galway

1

Challenges

Outline the challenges related to semantic interoperability in data spaces



2

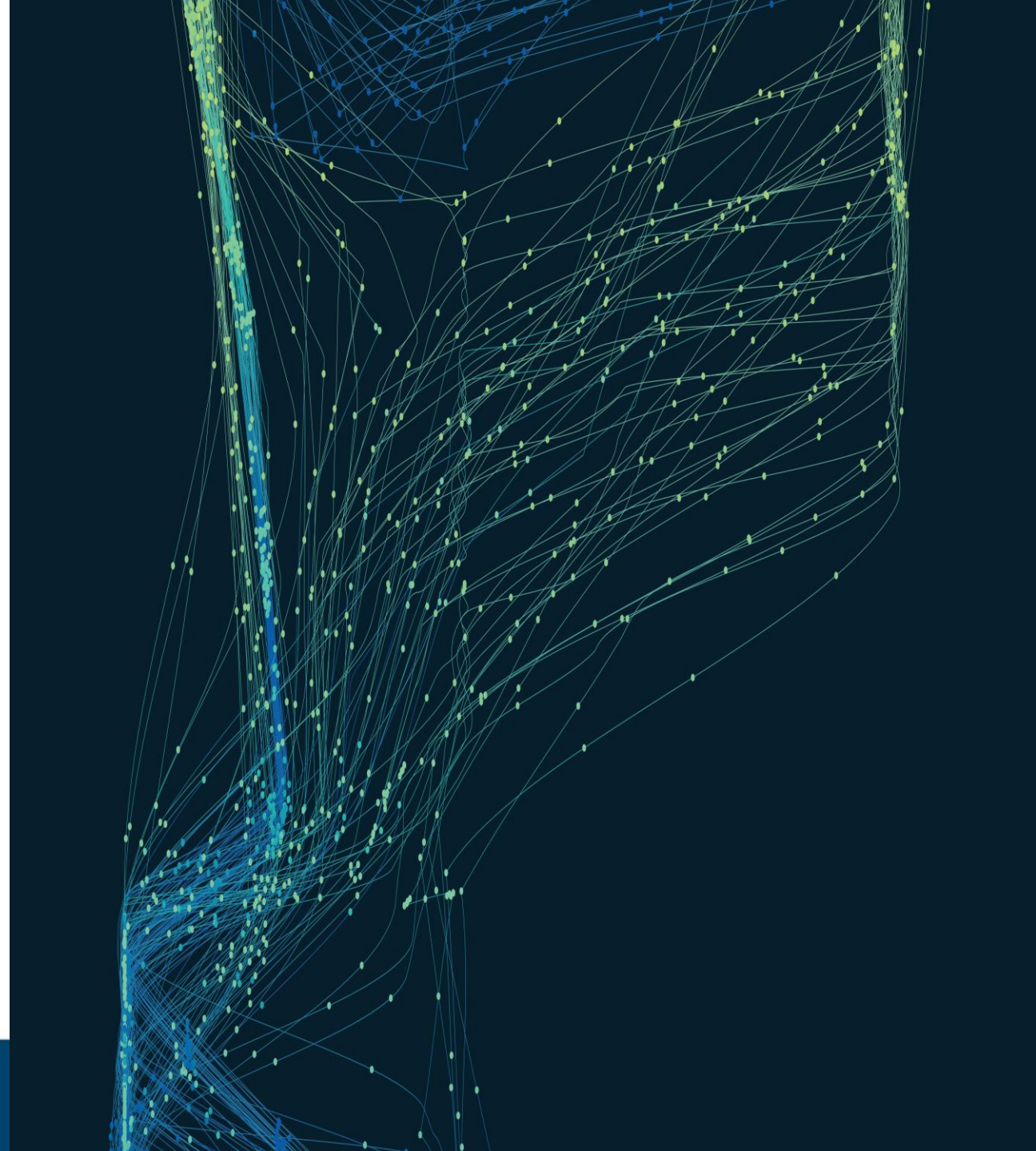
LLM-Based Semantic Mediation Engine

Introduce Large Language-based Mediation Engine for achieving semantic interoperability in data spaces

3

Core Functions of LSME

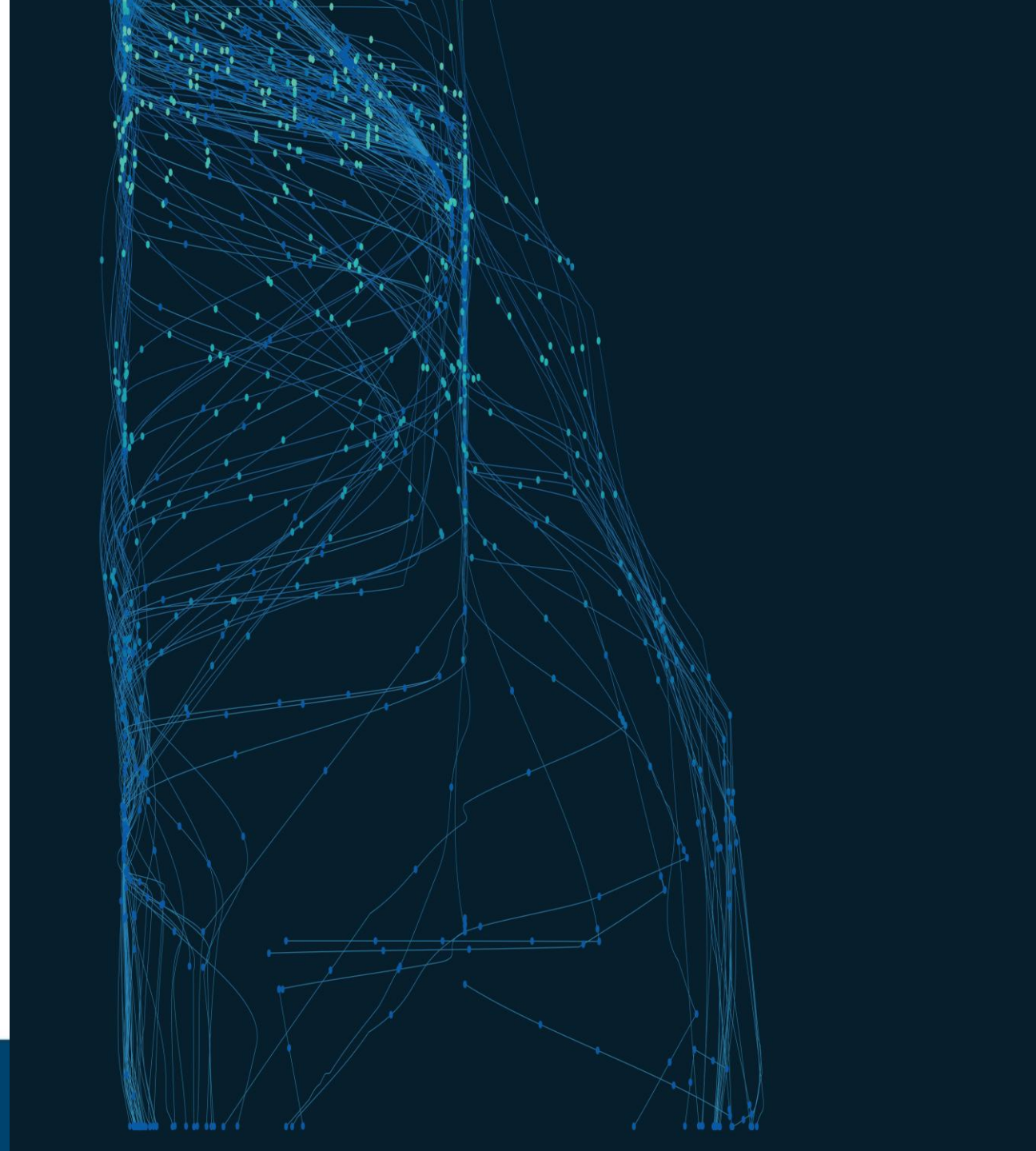
Provides the details of major core functions of LSME



4

High-level Architecture of LSME

Presents a high-level architecture of Large Language Model Semantic Mediation Engine

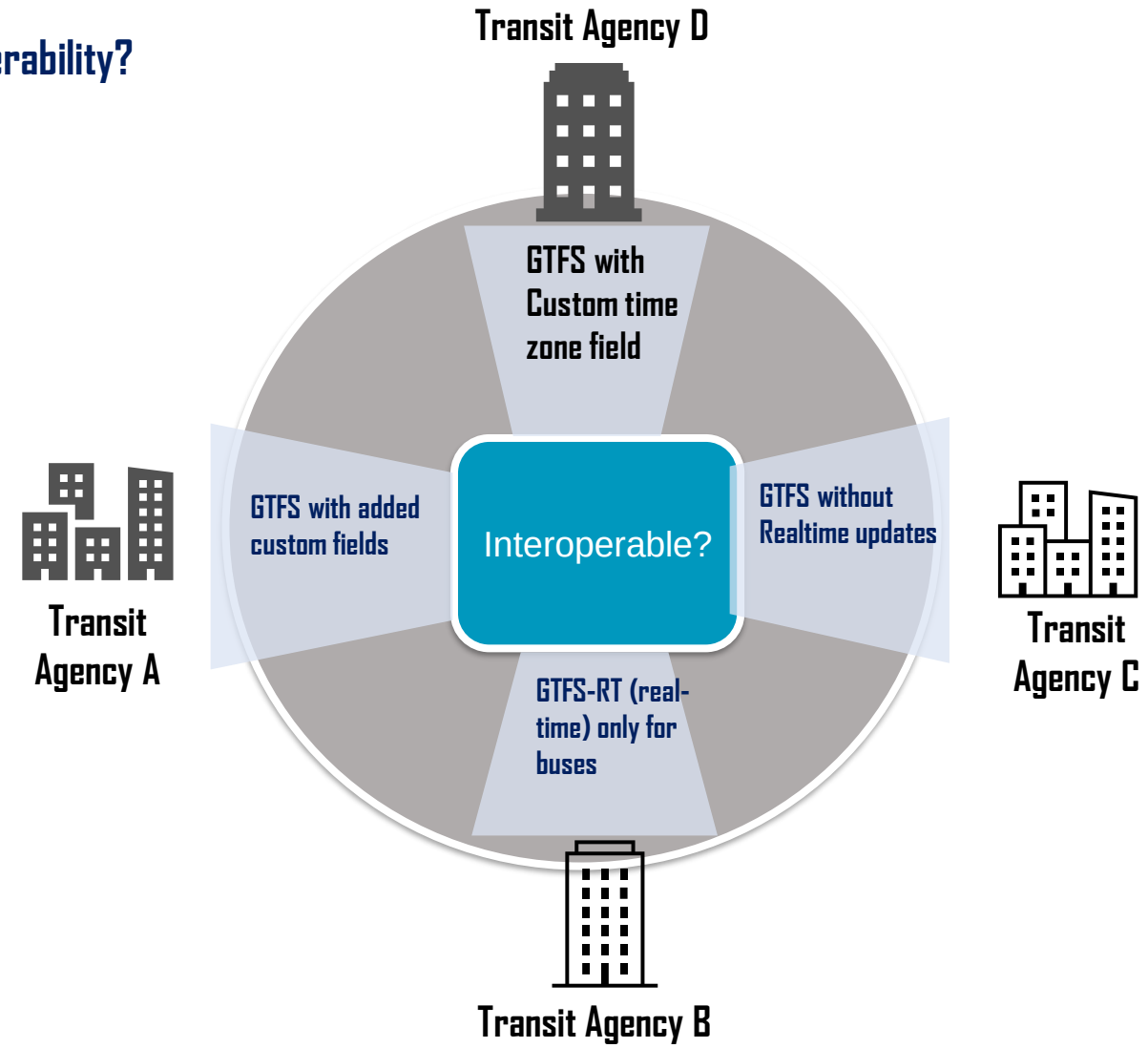


Semantic Interoperability Challenges

Can standardization alone ensure semantic interoperability?

- **Fragmented Adoption**: Different organizations/data spaces adopt different standards or versions of the same standard.

- **Impact on Mobility Sector**:
  - Application developers struggle to build a unified journey planner
  - Data Integration for MaaS platforms is slowed down.



Semantic Interoperability Challenges

Can standardization alone ensure semantic interoperability?

- **Semantic Difference**:
  - Two datasets may comply with the same format but interpret terms differently (e.g., what constitutes a "trip" may vary).
  - Semantic interoperability is often missing even with syntactic alignment

Trip data using a shared format like JSON

```json
{
  "trip_id": "12345",
  "start_time": "2025-04-07T08:15:00Z",
  "end_time": "2025-04-07T08:45:00Z",
  "mode": "bus"
}
```

$\neg$(PTA_trip = RSC_trip = MDS_trip)

Public Transit Authority
A *trip* is defined as a scheduled transit service instance:
PTA_trip=$\{s_1,s_2,...,s_n\}$ s_i is a scheduled stop. It is independent of passenger activity and focuses on the operational execution of the service route.

Ride-Sharing Company
A *trip* is defined as a completed user-requested ride:
RSC_trip = (o, d, t_u, t*c*, f) where o is the origin, d is the destination, t_u is the user request time, t*c* is the trip completion time, and f is the fare.

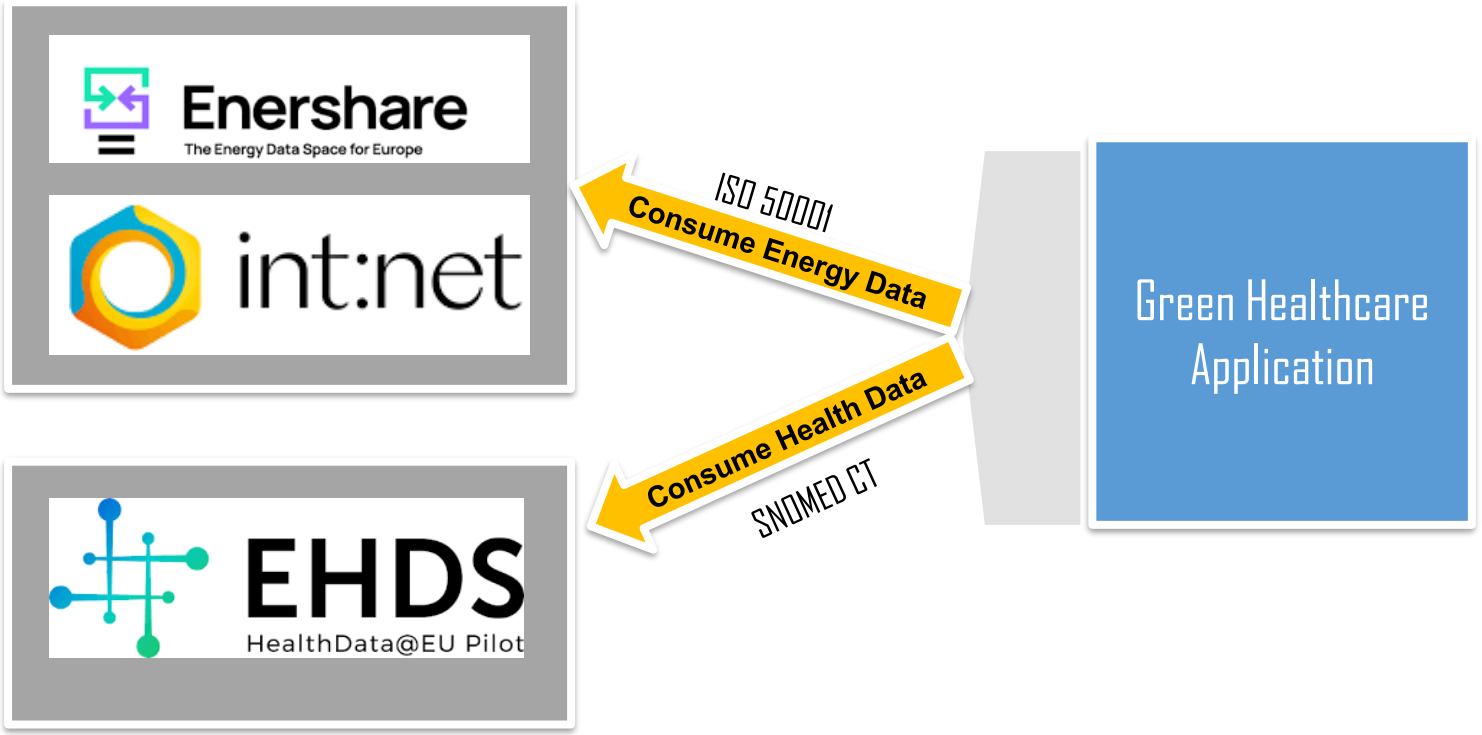
Mobility-as-a-Service
A *trip* is defined as a mobility segment within a multimodal journey:
MaaS_trip = (m, o, d, t_s, t_e) where m is the mode of transport (e.g., walk, bike, train), o and d are segment-level origin and destination, and t_s and t_e are the segment's start and end times.

Semantic Interoperability Challenges

Can standardization alone ensure semantic interoperability?

Green Healthcare Use Case

An application that tracks energy consumption in hospitals and integrates this data with patient care to ensure that energy usage aligns with specific healthcare practices (e.g., reducing energy usage during high-demand periods while maintaining patient care quality).



Challenges

- Different schemas making cross-domain data interpretation difficult without **mapping mechanisms**
- **Contextual alignment** is essential to make these datasets semantically interoperable. **Standards define structure, not meaning-in-context**

Large Language Model Based Semantic Mediation Engine (LSME)

- **What is LLM-powered Semantic Mediator?**
  - It acts as an intelligent semantic bridge between desperate datasets and vocabularies.
  - It relies on Large Language Model.
  - It complements traditional standards by dynamically bridging semantic gaps, enabling broader participation, and agile data integration at scale.

- **Purpose:** Intended for use in dataspaces to the following support:
  - **Understanding context and terminology**: Bridges gaps in meaning across systems that use different vocabularies.
  - **Extracting structured concepts from unstructured notes**: Unlocks meaning from free-text records, making them interoperable with structured data sources.
  - **Mapping between datasets**: Enables consistent understanding across datasets.
  - **Cross-domain knowledge fusion**: Reconciling knowledge and data from distinct and heterogeneous domains.

Large Language Model Based Semantic Mediation Engine (LSME)

- **Key Values**
  - **Semantic Reasoning and Inference**: Enables reasoning over implicit meanings and relationships in data, helping infer missing links or derive higher-level concepts.
  - **Fine-Tunable for Domain-Specific Interoperability**: The mediation engine can be fine-tuned using in-domain data and ensuring alignment with sector-specific interoperability requirements.
  - **Multilingual and Cross-Cultural Semantics**: Supports interpretation of semantics across languages and cultural contexts, crucial for data exchange.
  - **Less rework when standards change**: Spend less time reengineering mappings every time a schema is updated.
  - **One engine replaces multiple brittle pipelines:** Instead of maintaining custom scripts or point-to-point transformations for each integration, it enables to use a single adaptable mediation layer.

LSME Core Functions: Schema Mapping

- **Overview**
  - LLMs use **few-shot or zero-shot prompting** to identify **semantically equivalent fields** across datasets.
  - LLMs can understand both **contextual meaning** and **structural intent**, generating schema alignment suggestions.
  - The LLM processes the prompt and applies learned relationships between similar data structures, identifying common patterns and semantic links.
  - Schema Mapping Generation: Based on the inferred relationships, the model suggests schema mappings

- **Example**
  - **Prompt**: "Map the following source fields to the target schema using natural language understanding"

Source Schema

```
Source:
{'device_id',
'reading_time',
'CO2_ppm'}.
```

Target Schema

```
Target: {'sensorID',
'timestamp',
'carbonDioxideLevel'}."
```

Output: Schema Mapping

```
device_id → sensorID, reading_time → timestamp,
CO2_ppm → carbonDioxideLevel.
```

Use Case Scenario

- Different participants in a data space may represent similar concepts with different labels, column names, or schema structures

LSME Core Functions: Cross-Standard Terminology Translation

- **Overview**
  - LLMs translates terms between controlled vocabularies or classification systems
  - Utilize context-aware translation to disambiguate meaning.
  - **Example**: Differentiating between Type 1 and Type 2 diabetes in patients with ambiguous diagnosis history

- **Key Techniques**
  - Contextual Embedding
  - Cross-vocabulary embedding alignment
  - Cosine similarity
  - **Retrieval Augmented Generation** to provide domain-specific documents to the LLM during inference.

Terminology Translation Example

`Input: Prompt`

- `Translate the following ICD-10 code to its SNOMED CT equivalent:ICD-10: E11.9`

Output: ICD-10 ↔ SNOMED

- E11.9 → Type 2 diabetes mellitus without complications

Context Used: Adult patient, prescribed metformin, no documented complications.

LSME Core Functions: Cross-Domain Mediation and Fusion

- **Overview**
  - Facilitate cross-domain knowledge fusion by using Large Language Models (LLMs) to unify diverse datasets into a single, semantically coherent view.
  - Infers equivalence even with domain-specific terminology or aliasing.
  - Provides explanations for mappings to support trust and review
  - Proposes schemas that unify disparate datasets under a common structure

- **Key Techniques**
  - Semantic Field Alignment
  - Terminology Bridging
  - Schema Harmonization and Fusion Schema Generation
  - Natural Language-Based Mapping Interface
  - Contextual Reasoning Over Metadata and Sample Data

Example

Semantic Field Alignment

Mobility domain: `trip_start_time`
Healthcare domain: `admission_time`
Unified concept: `event_time`

Terminology Bridging

`ICD-10: J45` (Asthma)
`Air Quality Index` threshold breach
`Fusion Insight:` LLM connects poor air quality with likely respiratory distress events.

Schema Harmonization and Fusion Schema Generation

```
{
  "entity_id": "Derived from: patient_id, user_id, commuter_id",
  "event_type": "Derived from: hospital_admission, travel_start, service_usage",
  "location": "Derived from: hospital_unit, station_code, sensor_region",
  "timestamp": "Standardized from: admission_time, boarding_time, detection_time"
}
```

LSME Core Functions: Semantic Normalization Across Contexts

- **Overview**
  - LLMs will detect and normalize synonyms or concept variants using contextual embeddings.
  - LLM will use retrieval augmentation with domain-specific corpora.

- **Techniques**
  - Embed input text using the LLM's encoder.
  - Compare against a reference vocabulary or controlled terminology using cosine similarity.
  - Assign canonical form to harmonize data inputs.

Example

1. Embed raw text (e.g., vehicle type field: "BEV", "EV") using the LLM's encoder

2.
```
{
  "electric vehicle": ["EV", "BEV", "electric car", "zero-emission vehicle", "e-vehicle"]
}
```

3. Harmonize data inputs by assigning the **canonical label** "electric vehicle" to all recognized variants

LSME Core Functions: Structured Concept Extraction (1/2)

- **Overview**
  - The engine scans unstructured free-text records to identify key entities and concepts.
  - Leveraging the LLM's language understanding, the engine decodes meaning from the unstructured text, grasping context, relationships, and nuances that might be missed in a standard keyword-based approach.
  - The engine maps these concepts to predefined terminologies, or schemas in structured datasets, ensuring consistency across different systems.

- **Techniques**
  - Text Pre-processing
  - Named Entity Recognition
  - Context embedding and linking
  - Concept mapping

LSME Core Functions: Structured Concept Extraction (2/2)

Use Case

1 Input: A data provider uploads anonymized clinical documents

"Patient reports worsening fatigue and episodes of dizziness. MRI revealed left ventricular hypertrophy. Prescribed ACE inhibitor. Follow-up in 2 weeks."

2 **Using LLMs trained on biomedical corpora, the engine:**

- Identifies symptoms: fatigue, dizziness
- Detects diagnosis indicators left ventricular hypertrophy
- Extracts interventions: ACE inhibitor therapy
- Captures temporal reference: follow-up in 2 weeks

4 **These entities are mapped to structured codes:**

- Fatigue → SNOMED CT: 84229001
- Dizziness → SNOMED CT: 404640003
- Left Ventricular Hypertrophy → ICD-10: I51.7
- ACE Inhibitor → ATC: C09AA
- Planned follow-up → FHIR Appointment Resource

5

```
{
  "condition": [
    { "code": "I51.7", "display": "Left Ventricular Hypertrophy" }
  ],
  "symptoms": [
    { "code": "84229001", "display": "Fatigue" },
    { "code": "404640003", "display": "Dizziness" }
  ],
  "medication": [
    { "code": "C09AA", "display": "ACE Inhibitor" }
  ],
  "followUp": "2 weeks"
}
```

Core Functions: Semantic Search and Discovery (1/2)

- **Overview**
  - Enables intelligent, context-aware search capabilities across heterogeneous data sources by leveraging the semantic understanding of Large Language Models.
  - Surpasses traditional keyword-based search by interpreting the intent, context, and conceptual relationships in user queries and target data

- **Key Techniques**
  - Natural Language Query Understanding
  - Semantic Embedding Generation
  - Synonym and Concept Expansion
  - Contextual Disambiguation

Core Functions: Semantic Search and Discovery (2/2)

Use Case

Input:

A clinical researcher queries the Health Data Space with:

"Show me patient data related to heart failure treatment outcomes.**"**

Matching Semantics:

"Heart failure" with clinical equivalents like "CHF," "Herzinsuffizienz," "insuffisance cardiaque."

Data Discovery: Semantic mediation engine uses the LLM to match the **intent** of the query with the **semantically indexed metadata**

In Italy, a dataset tagged "**studio su esiti terapeutici dello scompenso cardiaco negli anziani" (study on therapeutic outcomes of heart failure in the elderly)** is matched.

In Germany, a dataset labeled "**Klinische Ergebnisse bei Patienten mit Herzinsuffizienz über 65 Jahren**" (clinical outcomes for patients with heart failure over 65) is retrieved.

Core Functions: Semantic Indexing (1/2)

- **Metadata Extraction Overview**
  - It leverages LLM to extract structured metadata from unstructured, semi-structured, or inconsistently structured data sources.
  - Identifies and classifies key entities,
  - Identifies and classifies key entities, attributes, relationships, and events.
  - Understands contextual meaning of terms to assign the right labels (e.g., knowing that "Paris" in one context is a person, in another it's a city).
- **Semantic Indexing Overview**
  - Assigns meaning-aware tags and identifiers to content based on its semantic content.
  - Builds a knowledge-rich index that links concepts, topics, and entities across documents, systems, and data sources.
  - Enables semantic search and retrieval that goes beyond keyword matching

Core Functions: Semantic Indexing (2/2)

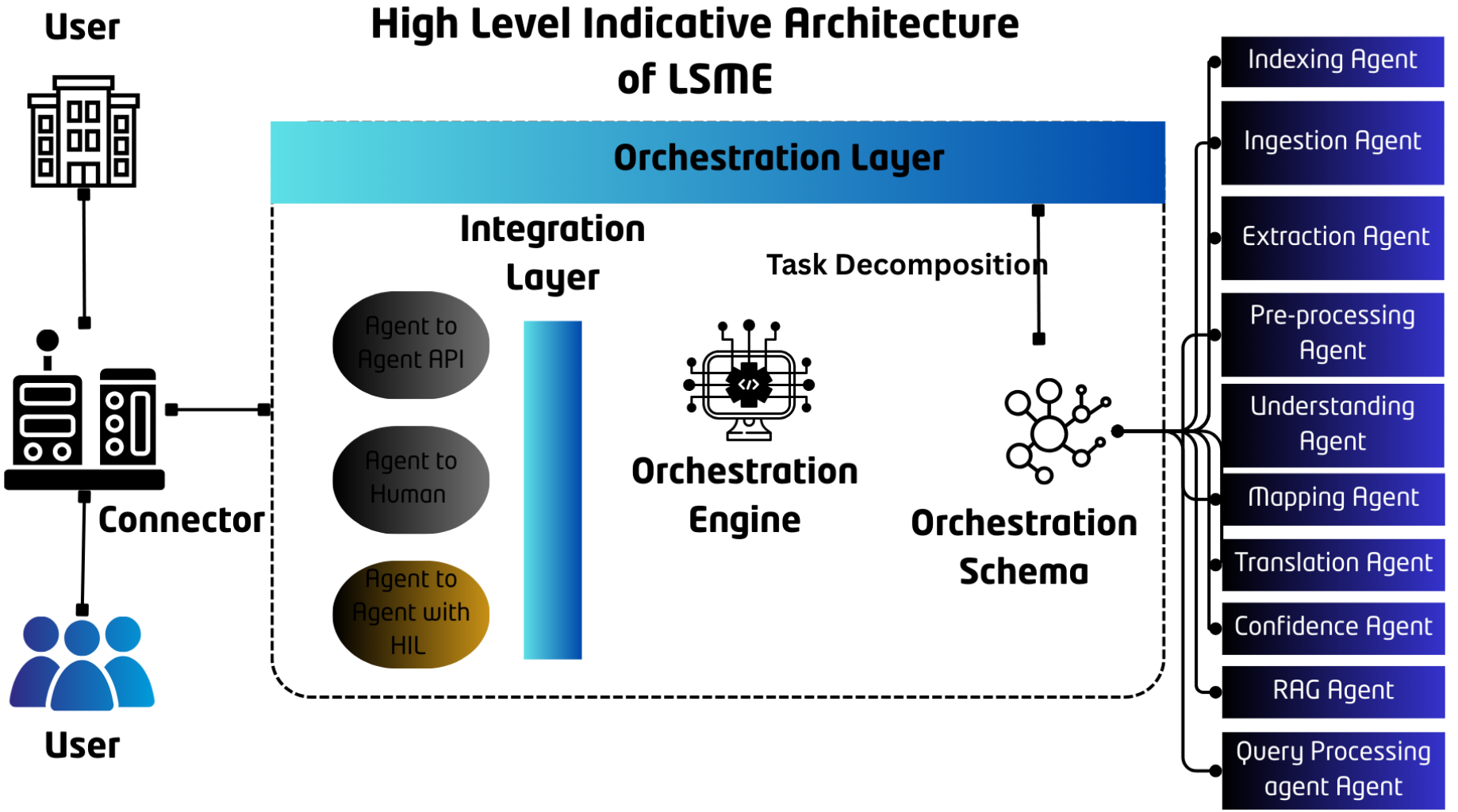
Key Techniques

- **Concept extraction**: Extracts key topics, terms, and themes from datasets and documentation
- **Semantics Embeddings**: Converts content into high-dimensional vectors using LLM embeddings
- **Semantic Tag Generation**: Tags datasets using keywords, and inferred categories.
- **Vector Indexing**: Builds vector index (e.g., FAISS, Weaviate, Qdrant) to enable semantic similarity search
- **Cross Lingual Normalization**: translates and aligns terms across languages using multilingual embeddings or translation models
- **Temporal and Contextual Dimensions**: Associates semantic tags with time, location, and context for multidimensional querying

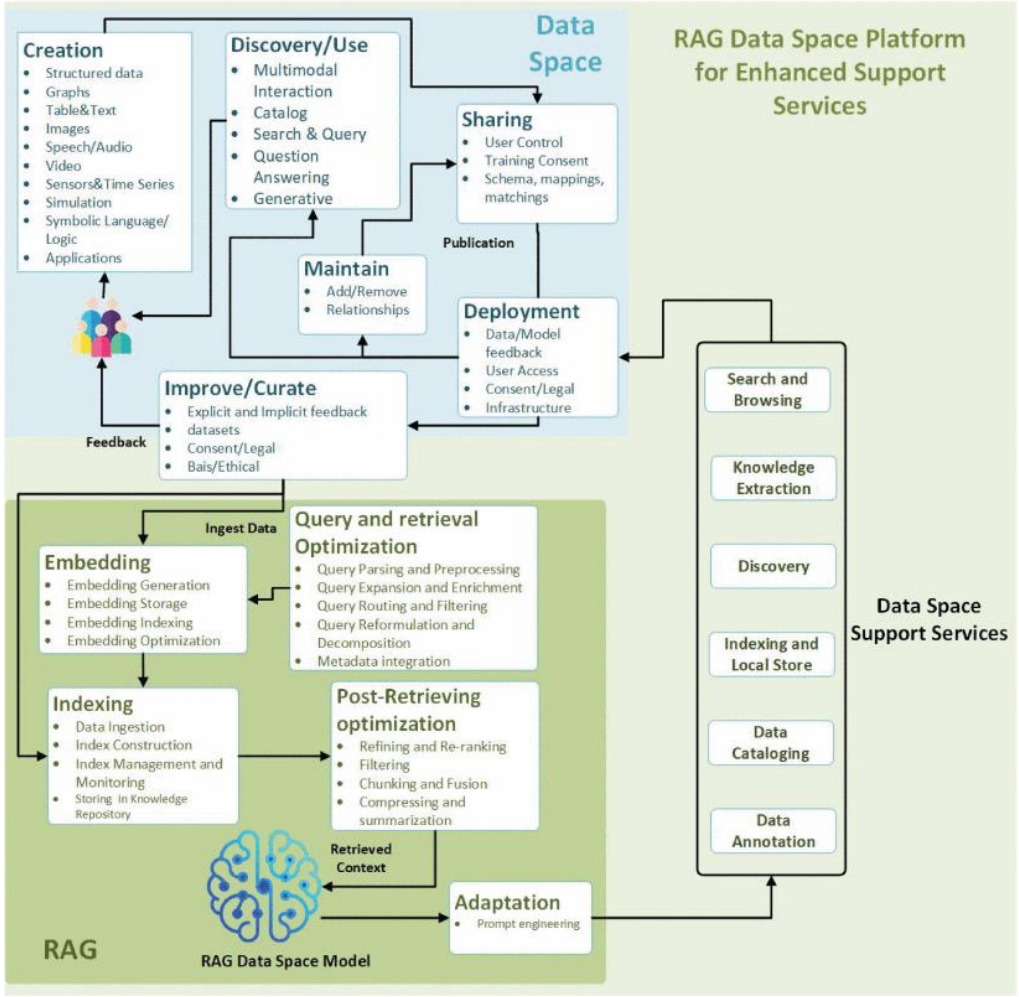
Use case scenario

Sector	Metadata Extraction using LSME	Metadata Extraction using LSME
Healthcare	Extracts patient demographics, medications, and conditions from clinical notes	Indexes records semantically to support symptom-based or case-based search
Mobility	Extracts service coverage areas, route patterns, operator details from datasets	Indexes MaaS services by function, region, or regulatory compliance

A High-Level Architecture of LSME



Current Progress - RAG4DS



Support Service	Subtasks
Knowledge Extraction	1. Entity Recognition 2. Relationship Extraction 3. Concept Extraction 4. Semantic Role Labeling
Data Labeling	1. Data Annotation
Discovery	1. Semantic Search 2. Cross-Modal Discovery 3. Relationship Identification
Indexing and Local Store	1. Caching and Monitoring 2. Data Integration 3. Schema Mapping 4. Data Storage Management and Index Building
Data Cataloging	1. Metadata Management 2. Data Classification
Search and Browsing	1. Approximate Query Processing 2. Keyword Search and Ranking

Al-Qatf, Majjed, et al. "RAG4DS: Retrieval-Augmented Generation for Data Spaces—A Unified Lifecycle, Challenges, and Opportunities." *IEEE Access* (2025).