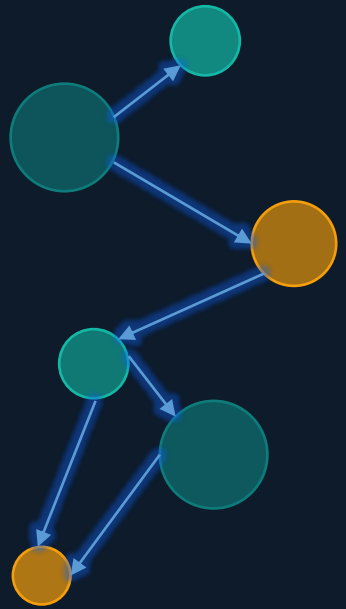
A Knowledge Graph Driven Approach to Semantic SDTM Validation

Bridging the Gap in Clinical Data Integrity Through Semantic Reasoning

GITIKA KISHOR
Junior Associate, Statistical Data Scientist
Pfizer, India



DISCLAIMER

- The views and opinions expressed in this presentation are my own and do not necessarily reflect those of my organization. All information is presented for educational and informational purposes only.

AGENDA

01

The Validation Challenge

Limitations of traditional rule-based SDTM validation approaches

02

Knowledge Graph Fundamentals

Nodes, edges, semantic relationships

03

Framework & Architecture

End-to-end pipeline: Python → rdflib → networkx → Neo4j

04

Building the SDTM Knowledge Graph

Domain modelling, CT mapping and cross-domain link creation

05

Case Scenarios

Cypher queries, anomaly detection and traceable outputs

06

Results, Benefits & Future Directions

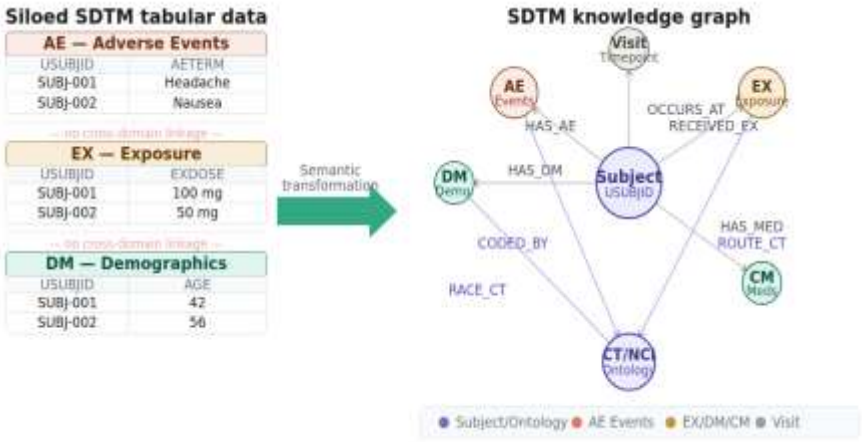
Scalability, regulatory alignment, and open-source roadmap

THE VALIDATION CHALLENGE

Why Traditional Approaches Fall Short

- Regulatory submissions require SDTM datasets that are technically compliant and semantically consistent across all domains.
- Traditional SDTM validation handles domains independently, often missing cross-domain context.

A Knowledge Graph-driven approach connects all domains semantically, enabling contextual, traceable, and standards-aware validation in a single query.



Static Rule Sets

Rule-based programs check one domain at a time, missing critical inter-domain inconsistencies and timing violations

Cross-Domain Awareness

Missing USUBJID inconsistencies, timing gaps, and inter-domain linkage failures

Traceability

Issues logged without linkage to metadata definitions for regulatory review

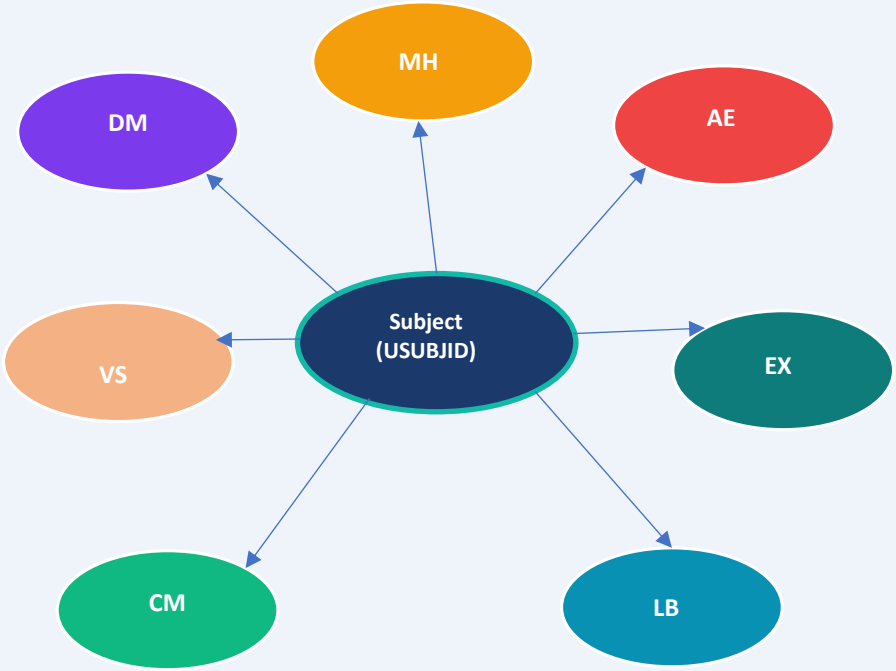
Regulatory Compliance

CDISC standards evolve continuously

KNOWLEDGE GRAPH FUNDAMENTALS

What is a Knowledge Graph?

- A structured representation of entities (nodes) and their semantic relationships (edges)
- Encodes domain knowledge with meaning, not just raw data values
- Supports end-to-end queries across linked domains in a single step
- Enables inference, anomaly detection, and contextual reasoning
- Widely used in drug discovery, biomedical research, and regulatory science
- Every SDTM concepts, subjects, domains, variables, visits, and CT codes are all represented as structured, interconnected nodes



Every SDTM domain, variable, subject and controlled term becomes a node, relationships become edges enabling semantic queries across the entire study.

BUILDING THE SDTM KNOWLEDGE GRAPH

Graph Node Types

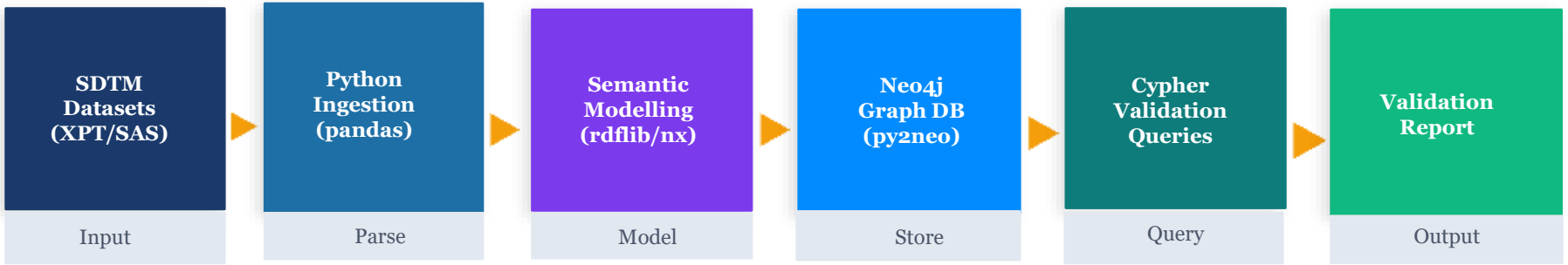
- **Subject** — USUBJID — unique subject identifier node
- **Domain** — DM, AE, LB, CM, EX, VS, MH ...
- **Variable** — Domain-specific dataset variables and values
- **CT Value** — NCI codelist values
- **Visit** — Planned/unplanned study visits and timepoints

Relationship (Edge) Types

- **HAS_AE** → Subject → Adverse Event record
- **RECEIVED_EXPOSURE** → Subject → Exposure/Treatment
- **HAS_LAB_RESULT** → Subject → Laboratory finding
- **CODED/CT** → AE/MH → Controlled terminology
- **OCCURS_AT_VISIT** → Assessment → Study visit node

```
graph.add_node(usubjid, type="Subject")   |   graph.add_edge(usubjid, ae_seq, relation="HAS_AE", aestdtc=onset_dt)
```

FRAMEWORK & ARCHITECTURE



👉 *Click here to explore the SDTM Knowledge Graph*

Data Ingestion

Load XPT/CSV SDTM datasets; extract domain metadata, variable attributes and controlled terminology values using pandas

Graph Construction

Create nodes for subjects, domains, variables, CT values and visits; define edges encoding timing, causation and coding relationships

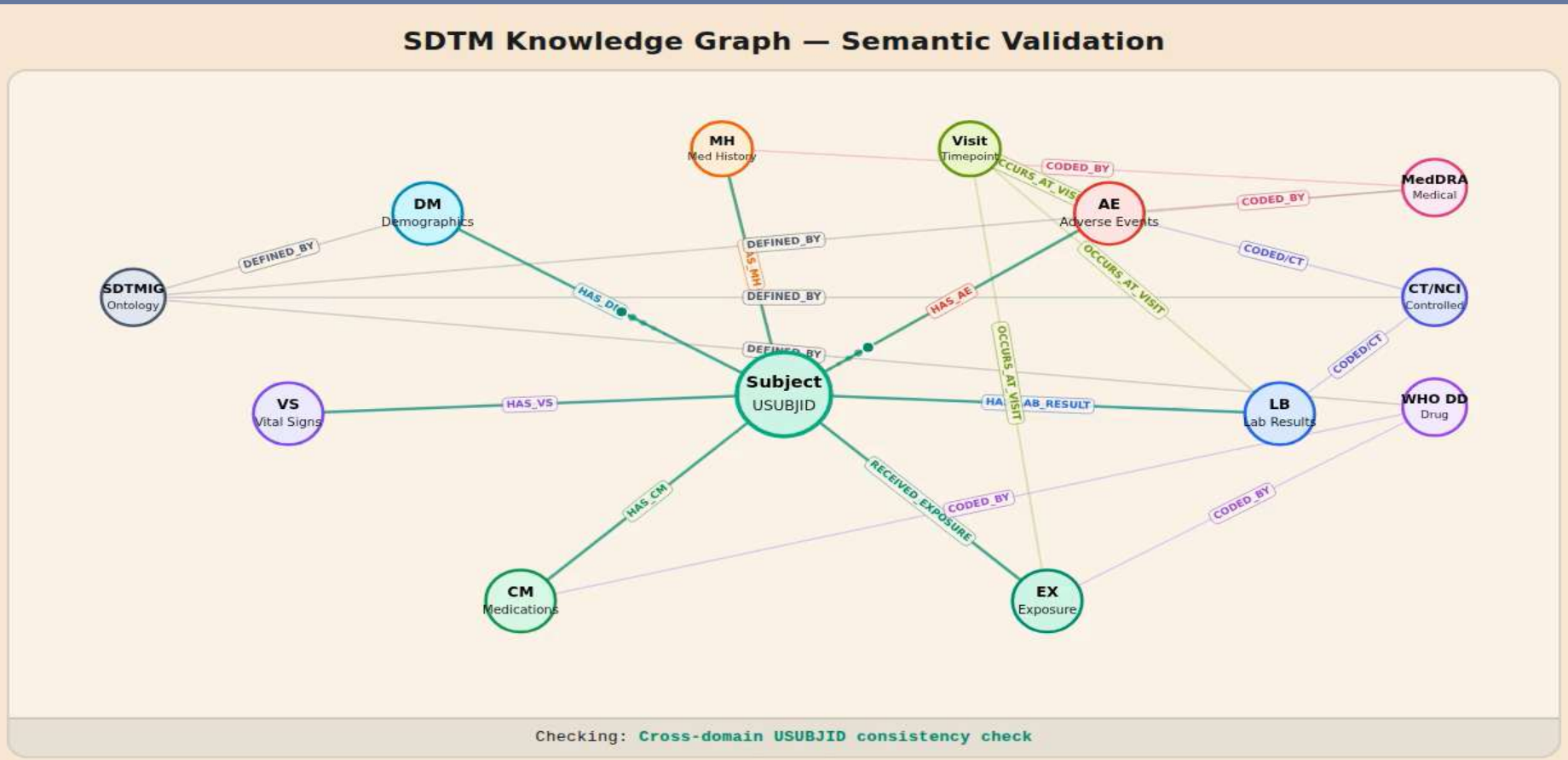
CDISC Alignment

Map graph schema to CDISC SDTM Implementation Guide, WHO, MedDra and NCI controlled terminology; use rdflib for RDF/OWL compliant semantic modelling

Semantic Queries

Execute Cypher pattern-matching queries to detect anomalies: missing domain links, CT violations, identifier inconsistencies, temporal errors

SDTM KNOWLEDGE GRAPH — VISUAL SCHEMA



CASE SCENARIOS

Scenario 01: Cross-domain USUBJID inconsistency

AE domain			
USUBJID	**AETERM**	**AESTDTC**	**AESEV**
SUBJ-105	Headache	2024-01-10	MILD
SUBJ-107	Nausea	2024-01-12	MODERATE
SUBJ-107	Fatigue	2024-01-15	MILD
DM domain			
USUBJID	**AGE**	**SEX**	**RFSTDTC**
SUBJ-105	42	M	2024-01-05
SUBJ-106	58	F	2024-01-06
---missing---	-----	-----	-----

```
MATCH (s:Subject)
WHERE NOT (s)-[:ENROLLED_IN]->(:DM)
RETURN s.usubjid AS MissingInDM
```

Detects subjects in AE/LB/EX with no DM record; a critical submission error

Scenario 02: Controlled terminology violation

AE domain			
USUBJID	**AETERM**	**AESEV**	**AEREL**
SUBJ-042	Rash	MILD	POSSIBLE
SUBJ-055	Anaphylaxis	LIFE-THREATENING	PROBABLE
SUBJ-061	Anaphylaxis	LIFE-THREATENING	DEFINITE

NCI CT		
Code	**Submission Value**	**Status**
C49494	MILD	Current
C49495	MODERATE	Current
C49496	SEVERE	Current
-----	LIFE-THREATENING	Removed

```
MATCH (ae:AE)-[:CODED_BY]->(ct:CT)
WHERE NOT ct.code IN $validNCI
RETURN ae.aeterm, ct.code AS Invalid
```

Flags AEOUT/AESEV/AESOC values absent from NCI codelist

Scenario 03: Temporal dependency violation

AE domain			
USUBJID	**AETERM**	**AESTDTC**	**AEENDTC**
SUBJ-214	Headache	2024-01-16	2024-01-20
SUBJ-215	Anaphylaxis	2024-01-10	2024-01-16
SUBJ-215	Rash	2024-01-18	2024-01-22
EX domain			
USUBJID	**EXTRT**	**EXSTDTC**	**EXDOSE**
SUBJ-214	Drug A	2024-01-14	100mg
SUBJ-215	Drug A	2024-01-15	100mg
SUBJ-215	Drug A	2024-01-22	100mg

```
MATCH (ae:AE)-[:OCCURS_AT]->(v2:Visit)
MATCH (ex:EX)-[:OCCURS_AT]->(v1:Visit)
WHERE v2.date < v1.date RETURN ae, ex
```

Identifies AE onset before first study drug exposure; impossible sequence

Scenario 04: Missing domain link — no CM for on-treatment subject

AE domain (on-treatment)			
USUBJID	**AETERM**	**AECONTRT**	**AESTDTC**
SUBJ-329	Fatigue	Y	2024-02-10
SUBJ-331	Nausea	Y	2024-02-12
SUBJ-333	Headache	Y	2024-02-14
CM domain			
USUBJID	**CMTRT**	**CMSTDTC**	**CMENDTC**
SUBJ-329	Ibuprofen	2024-02-11	2024-02-15
— SUBJ-331 missing —	—	—	—
SUBJ-333	Ondansetron	2024-02-14	2024-02-14

```
MATCH (s:Subject)
WHERE NOT (s)-[:HAS_CM_RECORD]->()
AND s.aecontrt = 'Y'
RETURN s.usubjid AS NoCM
```

Finds on-treatment subjects with no CM records; potential data omission

VALIDATION OUTPUT & TRACEABILITY

Each validation finding is fully traceable to its source metadata definition

Sample Validation Report Output

Check ID	Domain	Subject	Issue Description
KG-001	AE	SUBJ-0042	AESEV value 'LIFE-THREATENING' not in NCI C66768
KG-002	LB	SUBJ-0107	No LB records found for on-treatment subject
KG-003	EX	SUBJ-0215	AE onset date precedes first drug exposure date
KG-004	DM	SUBJ-0331	USUBJID missing in CM and VS domains
KG-005	AE	SUBJ-0089	AEOUT 'RECOVERING' used — non-standard term

```
Running date consistency validation...
✗ Found 1 date consistency issues

Running cross-domain reference validation...
✓ All cross-domain references valid

Running required variables validation...
✓ All required variables present

Running controlled terminology validation...
✓ All controlled terminology valid
```

BENEFITS & IMPACT

Scalability

faster cross-domain checks

Graph DB indexing and Cypher handle millions of nodes with sub-second responses, scaling across large study datasets.

Transparency

100%

traceable findings

Links every validation issue to source metadata, SDTMIG section, and NCI CT code for a complete regulatory audit trail.

Adaptability

Auto

version-aware validation

Graph schema dynamically reflects current CDISC standards; CT and SDTMIG updates propagate instantly, no rule reprogramming required.

Consistency

Zero

inter-domain blind spots

A unified graph model connects all SDTM domains seamlessly ensuring consistent relationships, eliminating data silos, and giving teams full cross-domain visibility without writing complex queries.

FUTURE EXPANSIONS

- Initial exploratory phase, requires structured feasibility checks before full-scale development.
- Conduct scenario-based testing across multiple study types, therapeutic areas, and data complexities to validate robustness.
- Identify edge cases and failure modes early to refine design assumptions before scaling.
- Validate the KG's ability to handle incomplete, inconsistent, or evolving metadata through pilot studies.
- Real-time validation monitoring dashboard
- Extend the KG's to cover ADaM datasets and derived records — enabling end-to-end study data traceability
- By amalgamating data from the life sciences domain, processes, and technology, KGs provide a robust data foundation for Gen AI models, ensuring transparency and evidence-backed responses.
- KGs can be continuously enriched with new data, keeping AI systems aligned with the latest scientific discoveries and regulatory updates
- Combined with large language models, this enables powerful, interactive analytics without the risk of AI hallucinations.

REFERENCES

- https://www.wisecube.ai/blog/revolutionizing-the-biopharma-industry-the-role-of-knowledge-graphs-in-shifting-to-a-data-centric-paradigm/
- https://altair.com/docs/default-source/resource library/da_print_eventbooklet_knowledgegraphs_a5_web.pdf?sfvrsn=2f665e71_1
- https://pubmed.ncbi.nlm.nih.gov/33843398/
- https://www.lexjansen.com/phuse-us/2025/pp/PAP_PP23.pdf
- https://pharmasug.org/proceedings/2025/SS/PharmaSUG-2025-SS-344.pdf
- https://www.ltm.com/industries/life-sciences/connecting-the-dots-knowledge-graphs-transforming-pharma-production

Thank You!

Gitika.Kishor@pfizer.com

www.linkedin.com/in/gitika-kishor-1a1499280