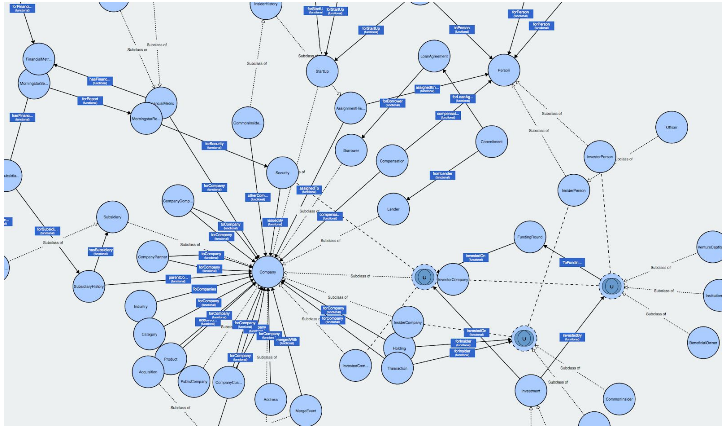
Semantic Enrichment of Data for AI Applications

Fatma Özcan (Google)

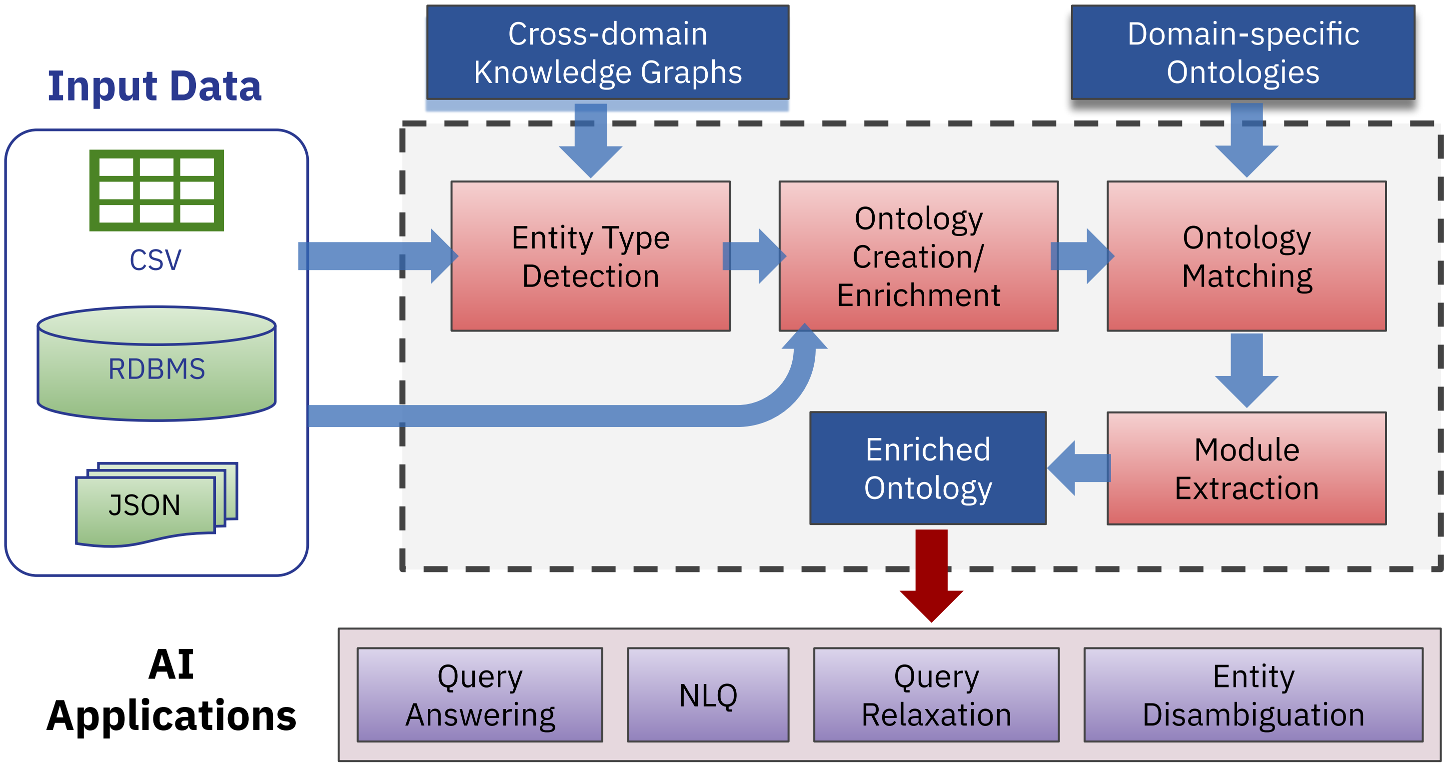
Joint work with
Chuan Lei (IBM Research)
Abdul Quamar (IBM Research)
Vasilis Efthymiou (ICS-FORTH)

More Context and Semantics

ID	First Name	Last Name	Gender	Email	DOB
1	John	Smith	M	john@abc.com	--
2	Tim	Cook	M	tim@apple.com	--



- Rich external knowledge sources
    - KGs and ontologies
- Context and semantics give better search results
- Formal semantics for reasoning
- **Ontologies** describe domains in terms of, concepts (aka classes) and roles (aka binary relationships)
- **Knowledge Graph :** Graph data model connecting real-world entities and events


Input Data
CSV
RDBMS
JSON
Cross-domain Knowledge Graphs
Domain-specific Ontologies
Entity Type Detection
Ontology Creation/ Enrichment
Ontology Matching
Module Extraction
Enriched Ontology
AI Applications
Query Answering
NLQ
Query Relaxation
Entity Disambiguation

Identify Identification: Detecting Column-level Entities

ID	First Name	Last Name	Gender	Email	DOB
1	John	Smith	M	john@abc.com	--
2	Tim	Cook	M	tim@apple.com	--

Similarity measures: lexical, word embedding

Metadata-level matching using the type hierarchy of Schema.org

Open Type Hierarchy describing schemas for structured data

Instance-level matching using a search index with Wikidata

Open, cross-domain KG

Entity Identification: Detecting Table-level Entities

ID	First Name	Last Name	Gender	Email	DOB
1	John	Smith	M	john@abc.com	--
2	Tim	Cook	M	tim@apple.com	--

Step 1: Detect column-level entities

First Name, Family Name, Gender, Email, Date

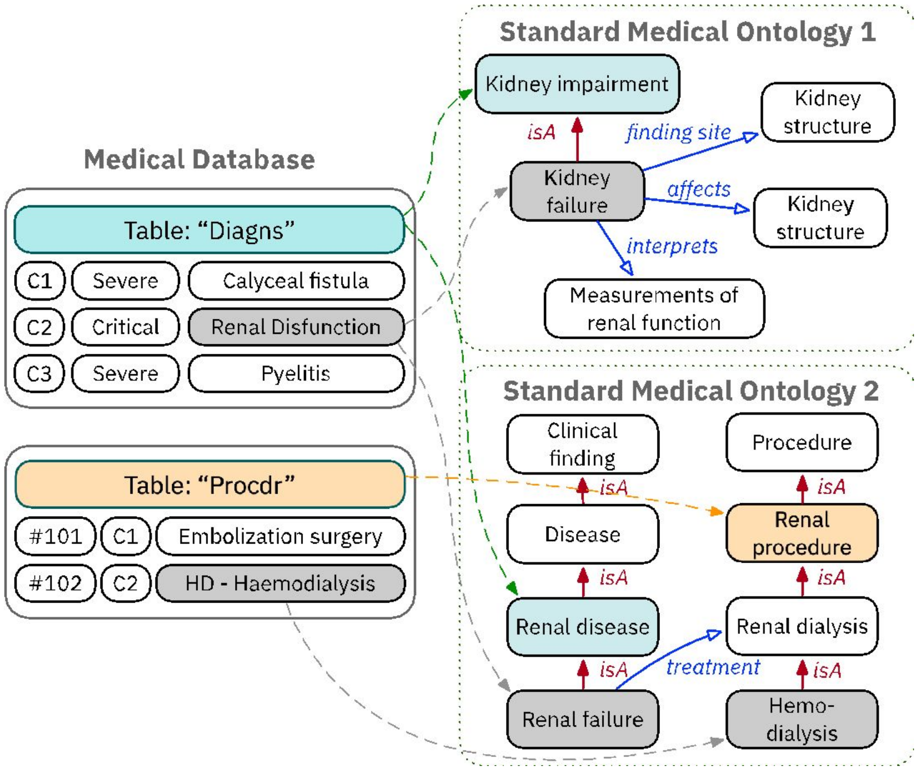
Step 2: Detect table-level entities

schema.org

Person

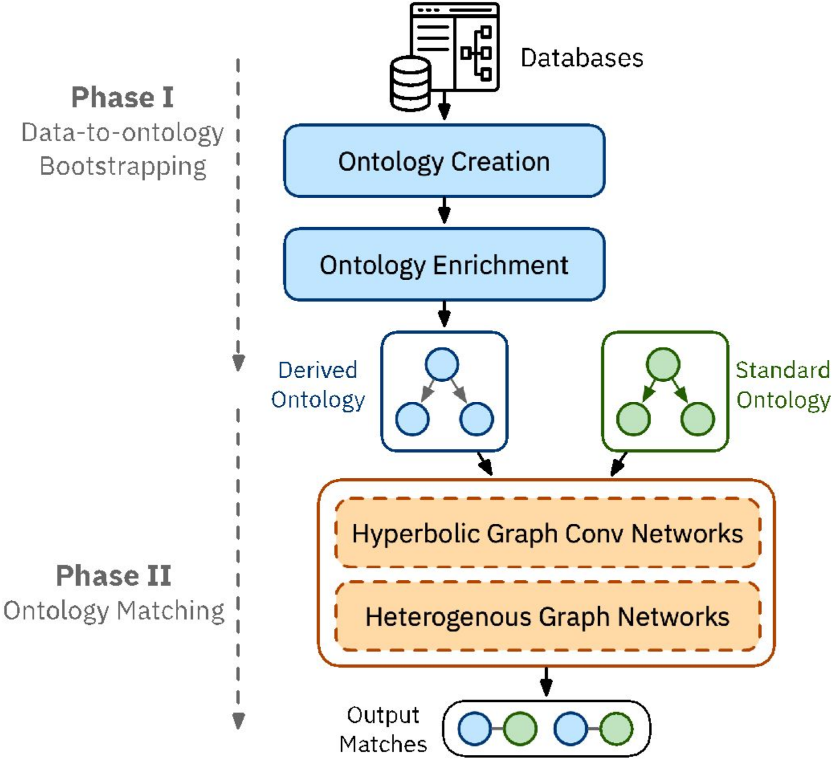
Using top-3 column-level matches, identify table-level entities based on voting

Data to Ontology Matching



- Data matches with an ontology
  - Standardize terminologies
  - Enrich semantic information for downstream applications
- Ontologies often have rich hierarchical structures, different levels from general to specific

Data to Ontology Matching: MEDTO Approach



- There is not always an ontology for a data set
- Data to ontology matching in two steps
  - Data to ontology bootstrapping
    - Derive an ontology from its schema and data instances
    - Bootstrap seed matches between the derived and standard ontologies for training
  - Ontology matching
    - Inputs – derived and standard ontologies as well as seed matches
    - Ontologies are captured via GNNs for both semantical and structural representation learning

J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, and W. Wang, "MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks", KDD 2021

Ontology Creation from Data

Relational data to Ontology

- Tables → Concepts
- PK/FK to infer relationships
- ISA hierarchies are not straightforward

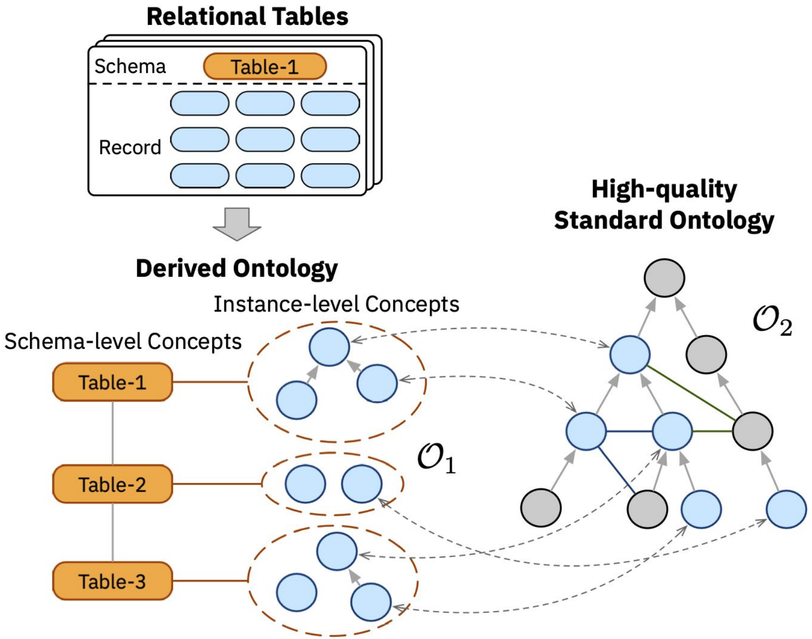
JSON data to Ontology

1. Infer data guides FIRST
2. Data guides to ontology using two simple rules
   Path A.b ⇒ Concept A, Property b of A
   Path A.B.c ⇒ Concept A, Concept B, Relation A to B, property c of B

C. Lei, F. Ozcan, A. Quamar, A. Mittal, J. Sen, D. Saha, K. Sankaranarayanan, "Ontology-Based Natural Language Query Interfaces for Data Exploration", IEEE Data Engineering Bulletin 41 (3), 52-63

Data to Ontology Bootstrapping

- Bootstrap an ontology (graph) for matching using both metadata and instances from relational databases
- Ontology creation
  - Create schema-level concepts from the metadata of the relational database
- Ontology enrichment
  - Concept augmentation
  - Neighborhood augmentation

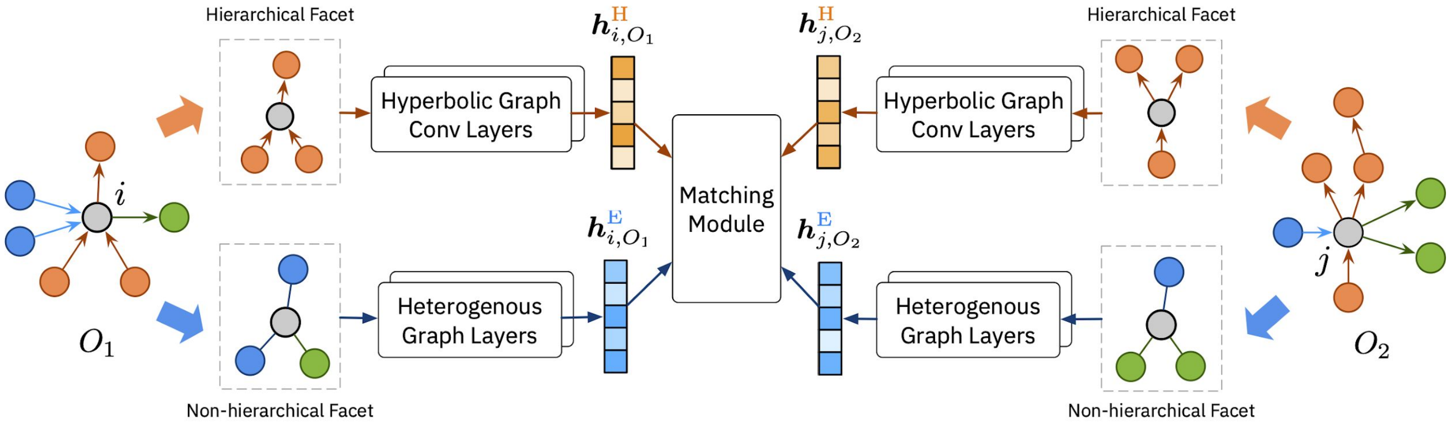


J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, and W. Wang, "MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks", KDD 2021

Ontology Matching: MEDTO Architecture

Two graph encoders and one matching module

- Hyperbolic graph conv encoder: hierarchical facet in ontology
- Heterogeneous graph encoder: local relational structure and global context



J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, and W. Wang, "MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks", KDD 2021

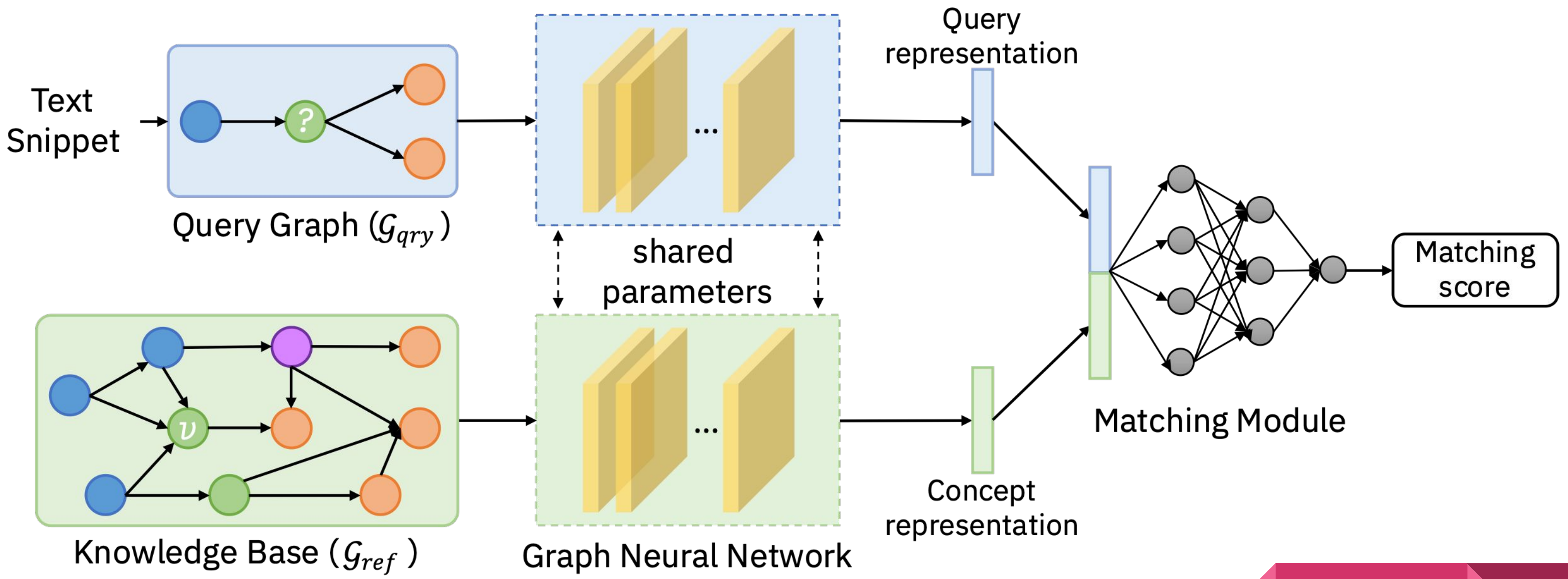
Applications

Medical Entity Disambiguation

- Medical knowledge graph (ontology) curation/maintenance
- Editorial staff often refer to a medical entity (a.k.a concept/class) in a knowledge graph with acronyms, typos and colloquial terms
- Example
  - An entity in KG: *acute renal failure*
  - Text snippet from an editorial staff: *Aspirin can cause nausea, indicating a potential ARF, nephrotoxicity, or proteinuria.*
  - *ARF* is the ambiguous term
- Collectively learn contextual and structural information of entities in a text snippet
- Capture discriminative contextual information of entities in a medical KB

A. Vretinaris et. al, "Medical Entity Disambiguation Using Graph Neural Networks" in SIGMOD 2021: Data Curation and Integration 6/24/2021: 18:30-2PM EST

ED-GNN Architecture

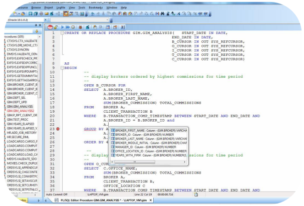


A. Vretinaris, C. Lei, V. Efthymiou, X. Qin, and F. Özcan, “Medical Entity Disambiguation Using Graph Neural Networks”, SIGMOD 2021

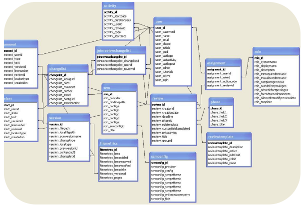
Ontology-based Natural Language Querying

Motivation

Easy Access for Business Users



User does need to know SQL or any other complex lang !!



Exact knowledge of underlying data is not required

Conversational interfaces

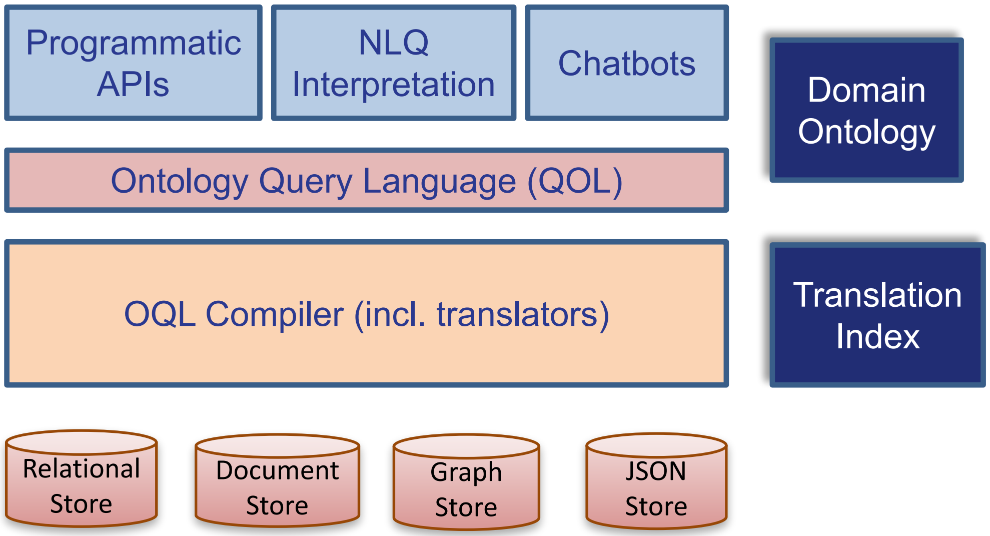
Challenges

- Understanding user intent (disambiguation)
- Converting the intent to target language (e.g., SQL)

ATHENA (PVLDB 2016), ATHENA++ (PVLDB 2020),
SIGMOD 2019 demo, SIGMOD 2020 industrial, Data Engineering Bulletin 2018

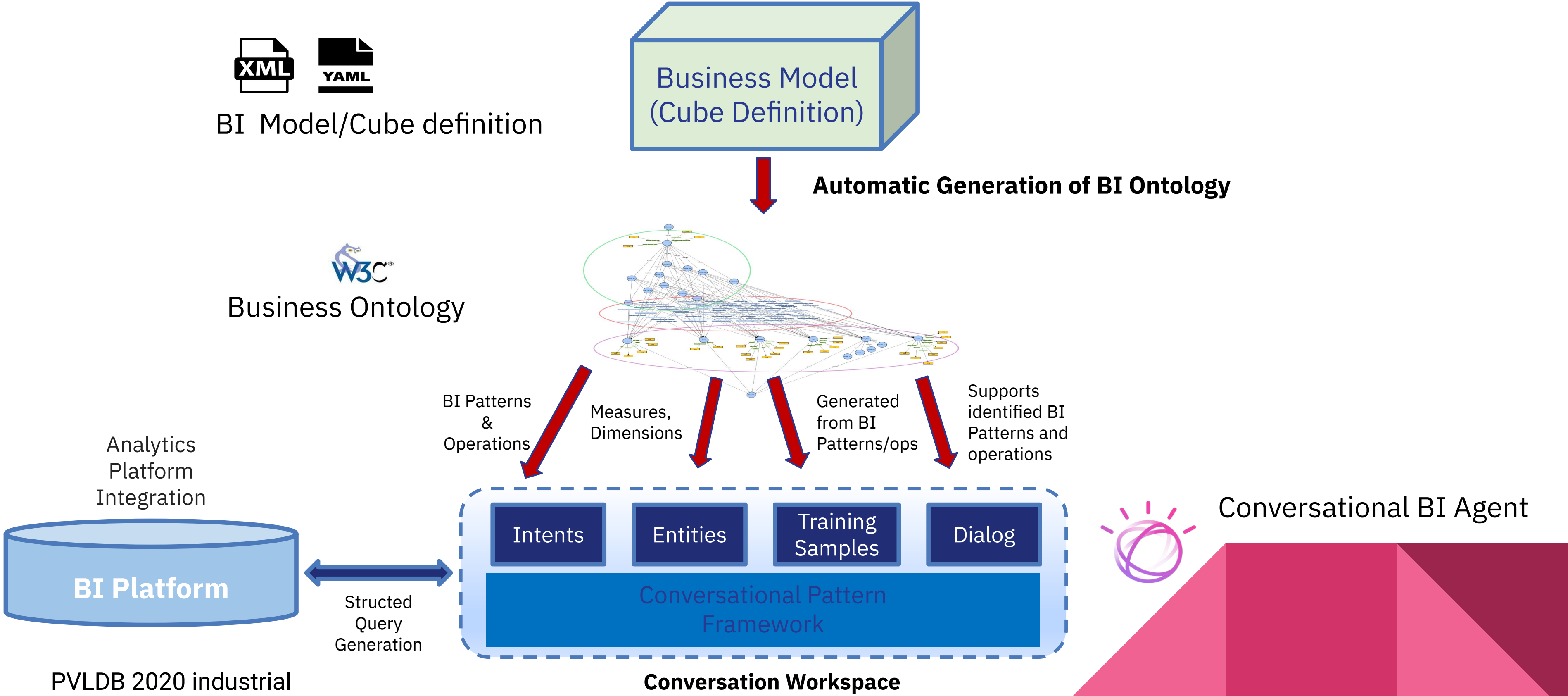
Ontology-Based NL Overview

- **Ontology: Entity-based interpretation**
  - Rich semantic modelling of the domain schema
- **General purpose NLQ interpretation engine with a 2-phase approach**
  - NLQ->OQL, capturing use intent, reasoning over domain schema
  - OQL to one or more target data stores
- **Ontology Query Language (OQL): A query language over the domain schema agnostic to physical data model**
  - Supports multiple data models and backends
  - Allows querying at a higher semantic level
- **Translation Index**
  - Captures domain vocabulary
  - Incorporates external domain knowledge



ATHENA (PVLDB 2016), ATHENA++ (PVLDB 2020),
SIGMOD 2019 demo, SIGMOD 2020 industrial, Data Engineering Bulletin 2018

An Ontology-driven Approach for Conversational BI



PVLDB 2020 industrial

References

1. J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, and W. Wang, "MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks", SIGKDD 2021
2. A. Vretinaris, C. Lei, V. Efthymiou, X. Qin, and F. Özcan, "Medical Entity Disambiguation Using Graph Neural Networks", SIGMOD 2021
3. S. Ahmetaj, V. Efthymiou, R. Fagin, P. G. Kolaitis, C. Lei, F. Özcan, and L. Popa, " Ontology-Enriched Query Answering on Relational Databases", IAAI 2021
4. A. Quamar, C. Lei, D. Miller, F. Özcan, J. Kreulen, R. J. Moore, and V. Efthymiou, "An Ontology-Based Conversation System for Knowledge Bases", SIGMOD 2020 (industrial)
5. C. Lei, V. Efthymiou, R.Geis, and F. Özcan, "Expanding Query Answers on Medical Knowledge Bases", EDBT 2020 (industrial)
6. A. Quamar, F. Özcan, D. Miller, R. Moore, R. Niegus, and J. Kreulen, "Conversational BI: An Ontology-Driven Conversation System for Business Intelligence Applications", PVLDB 2020 (industrial)
7. J. Sen, C. Lei, A. Quamar, F. Özcan, V. Efthymiou, A. Dalmia, G. Stager, A. Mittal, D. Saha, K. Sankaranarayanan, "ATHENA++: Natural Language Querying for Complex Nested SQL Queries", PVLDB 2020
8. J. Sen, F. Özcan, A. Quamar, G. Stager, A. R. Mittal, M. Jammi, C. Lei, D. Saha, and K. Sankaranarayan, "Natural Language Querying of Complex Business Intelligence Queries", SIGMOD 2019 (demo)
9. C. Lei, F. Özcan, A. Quamar, A. R. Mittal, J. Sen, D. Saha, and K. Sankaranarayanan, "Ontology-Based Natural Language Query Interfaces for Data Exploration", IEEE Data Engineering Bulletin, 41(3), September 2018
10. D. Saha, A. Floratou, K. Sankaranarayanan, U. F. Minhas, A. Mittal, and F. Özcan, "ATHENA: An Ontology-Driven System for Natural Language Querying over Relational Data Stores", PVLDB 2016

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
рахмат
Баярлалаа
спасибо
gracias
merci
obrigado
bedankt
dziękuje
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
arigatô
dakujem
takk
grazie
tapadh leat
hvala
mauruuru
dank je
mersi
kia ora
barka
welalin
tack
spas
faafetai lava
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
kiitos
dankie
nanni
nandri
dhanyavad
bayarlalaa
gràçie
köszönöm
enkosi
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
akun
dankon
aciu
sobodi
dėkuji
sagolun
mèsi
didi madloba
najis tuke
kam sah hamnida
তোমাকে ধন্যবাদ
rahmat
ありがとう
감사합니다
tanemirt
rahmet
xièxie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
trugarez