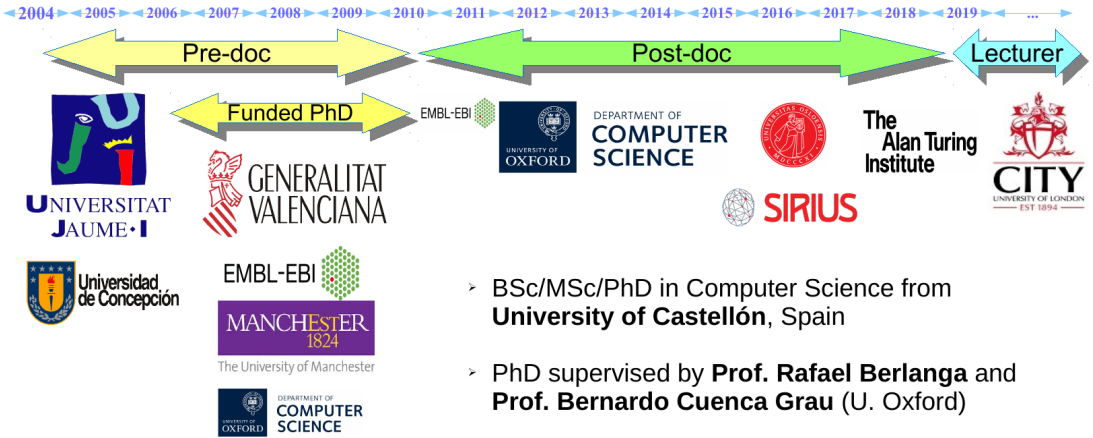
SIRIUS

Semantics for Data Access and Data Science

Ernesto Jiménez-Ruiz
Lecturer in Artificial Intelligence

Introduction

Research Journey



- BSc/MSc/PhD in Computer Science from **University of Castellón**, Spain
- PhD supervised by **Prof. Rafael Berlanga** and **Prof. Bernardo Cuenca Grau** (U. Oxford)

Research Overview

Main lines:

- Use of ontologies and thesauri for text mining (EBI)
- Application of logic-based methods in practice
- Reproducibility of research results
- Role of Semantics in Data Science
- Combination of Knowledge Representation and Machine Learning

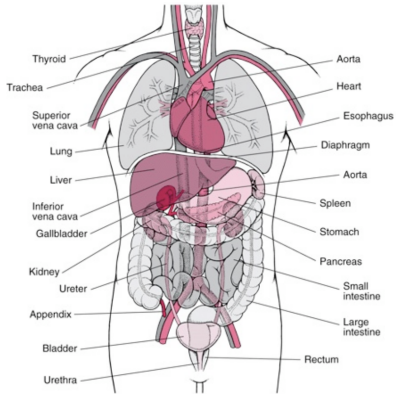
The Presentation in a Nutshell

- Brief introduction to Ontologies
- Overview of Optique and AIDA projects
- ColNet system
- A knowledge graph for ecotoxicological effect prediction

Ontologies

What is an Ontology?

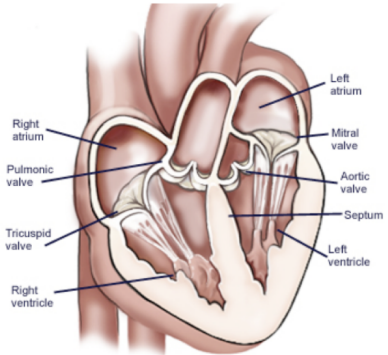
- Introduces **vocabulary** relevant to a domain
  - Anatomy



(*) Borrowed from Ian Horrocks' slides: **Ontologies and the Semantic Web: The Story So Far**. April 2010

What is an Ontology?

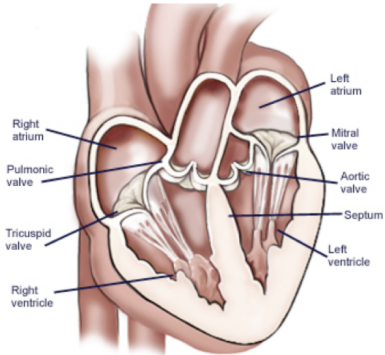
- Specifies meaning (**semantics**) of terms
  - Heart is a muscular organ that is part of the circulatory system



(*) Borrowed from Ian Horrocks' slides: **Ontologies and the Semantic Web: The Story So Far**. April 2010

What is an Ontology?

- Specifies meaning (**semantics**) of terms
  - Heart is a muscular organ that is part of the circulatory system
- **Formalised** using suitable logic
  - Heart SUBCLASSOF MuscularOrgan AND (isPartOf SOME CirculatorySystem)



(*) Borrowed from Ian Horrocks' slides: **Ontologies and the Semantic Web: The Story So Far**. April 2010

Why do we need Ontologies?

- Ontologies standardize, define and structure concepts in a knowledge domain
- They are essential for **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) data:
  - Using standard vocabularies to describe data is key for **Findability**, **Interoperability** and **Reusability**
  - A hierarchical structure improves **Findability** and enables **Interoperability**
  - Having a public knowledge model is key for **Accessibility**

What Ontologies are good for?

- Independence of logical/physical schema: **domain model**
- Vocabulary closer to domain experts: **more user-friendly**
- Incomplete and semi-structured data: **flexibility**
- Integration of heterogeneous sources: **unified view**

What is a Knowledge Graph (KG)?

– “Large network of entities, their semantic types, properties and relationships between entities.” (JWS Special Issue on Knowledge Graphs)

What is a Knowledge Graph (KG)?

- "Large network of entities, their semantic types, properties and relationships between entities." (JWS Special Issue on Knowledge Graphs)
- **(Light) Knowledge Base**: with a (light) terminology (ontology) and assertions (data)
- Nicer name than **RDF graph** (Resource Description Framework)
- Examples: Google Knowledge Graph, DBpedia (KG version of Wikipedia)

The EU Project Optique

EU Project Optique: Motivation

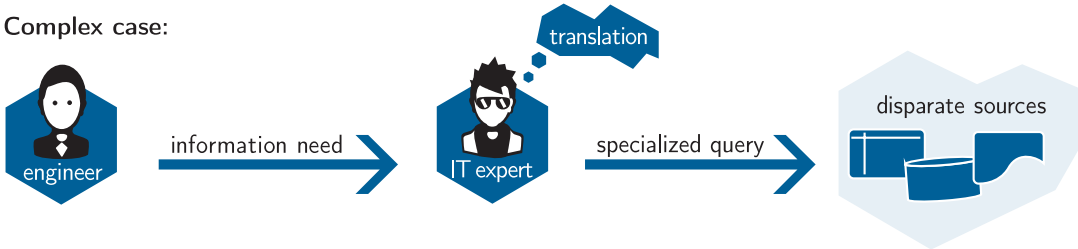
- **Statoil/Equinor**
  - Better quality information could add €1B/year net value to production
  - Poorer quality information and analysis costs €6M/weekend!
  - Data access is a challenge for engineers
- **IT industry:**
  - SAP deals with 80,000 queries/month at a cost of approx. €16M
  - SAP estimate 50% of support staff time spent searching for relevant information

(*) Data values may be obsolete, but may be even bigger.

EU Project Optique: Motivation

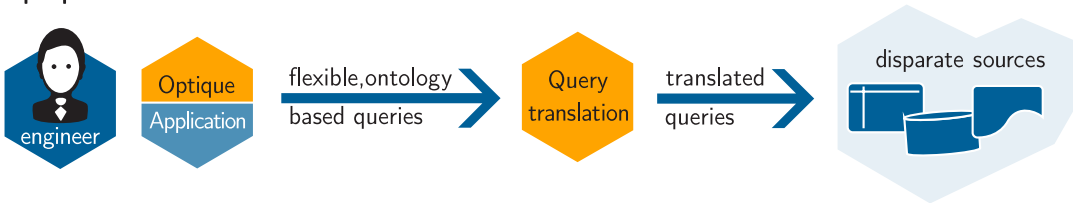
Simple case:

EU Project Optique: Motivation



EU Project Optique: Motivation

Optique solution



Optique in a Nutshell

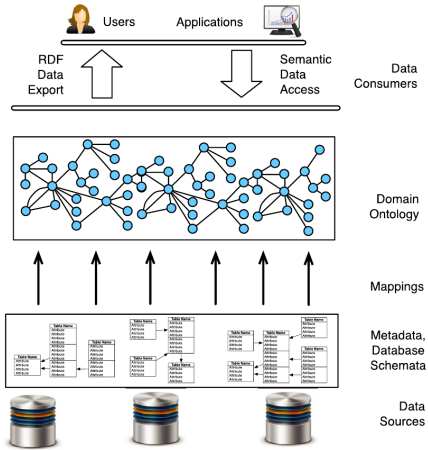
- Aims at facilitating **scalable end-user access to big data** in the oil and gas industry.
- Advocates for an **OBDA approach** (Ontology-Based Data Access)
  - ontology provides a virtual access to the data
  - mappings connect the ontology with the data source.
- Focuses around two demanding use cases provided by the industry partners **Siemens** and **Statoil**/**Equinor**

Ontology Based Data Access

- Three ingredients
  - **Onto vocabulary:** defines the terms from the RDB's domain
  - **Axioms:** describe the structure of the RDB's domain
  - **Mappings:** relate ontology terms to queries over the RDB

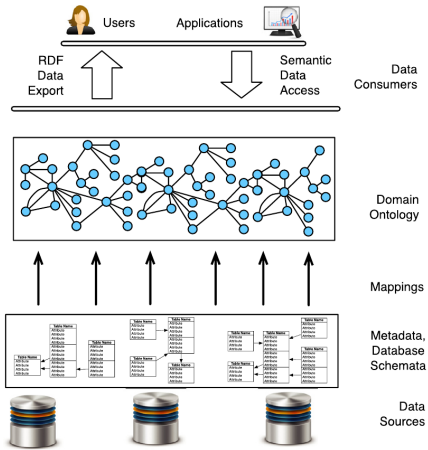
Mappings

$$Class(f_o(x)) \rightsquigarrow \texttt{SQL}(x)$$
$$objectProperty(f_o(x), f_o(y)) \rightsquigarrow \texttt{SQL}(x, y)$$
$$dataProperty(f_o(x), f_v(y)) \rightsquigarrow \texttt{SQL}(x, y)$$

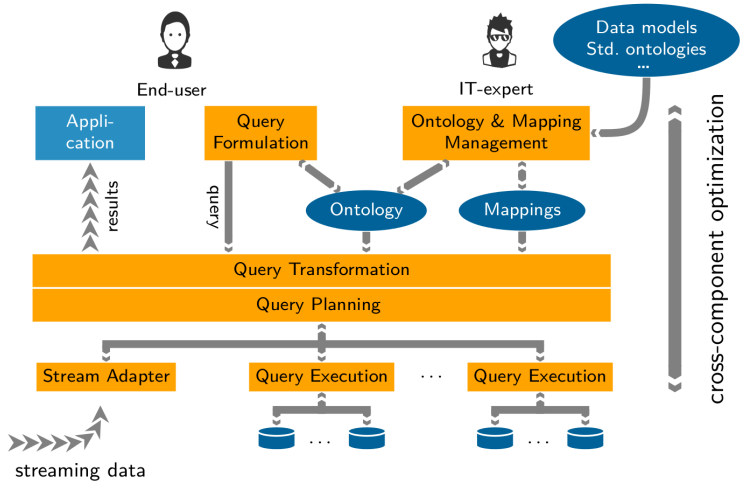


Exposing Relational Data via Ontologies

- **Virtual exposure of data** (OBDA) as in Optique
  - End-users' friendly access to "unfriendly" relational data
  - Pay as you go data integration
- **Data Export**
  - Transformation into a standard/clear schema
  - Easy to exchange data (over the Web)



Optique: Architecture



Optique: Lessons Learnt

- Requires **initial installation** and maintenance.
- **Modular** solution.
- **Evaluated** in Statoil/Equinor in a **controlled scenario**.
- Engineers were able to use the systems **without previous knowledge of semantic technologies**.
- The **visual query interface** was very important.
- **Much easier to formulate** complex and large SQL **queries**.
- **Great potential** for (large) enterprises, but not yet in production.
- Lead to the **SIRIUS centre** for research-based innovation.

AIDA Project
Artificial Intelligence for Data Analytics

Motivation

Data understanding and data preparation involves the 80% of work on a data mining project.

Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

AIDA project

The AIDA project seeks to reduce the time and effort needed for the data understanding and preparation.

- **Multidisciplinary work**: machine learning, semantic technologies and software engineering.
- **Modular system** based on AI assistants/experts designed to support individual steps.
- Integrated, **interactive system** that guides analysts through each step of the pipeline.
- **Open-source platform** to accelerate the task of practical data science.

There is not a public integrated toolchain to support the full data science process

The Role of Semantics in AIDA

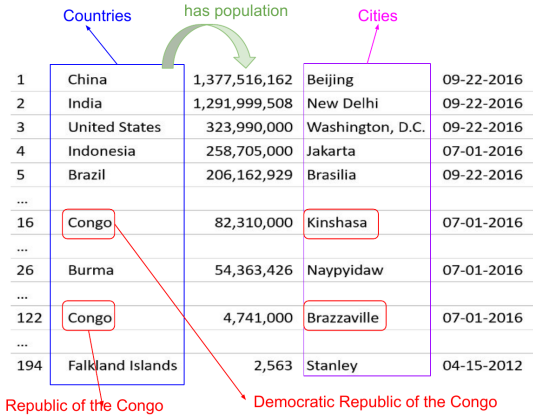
– The **lack of semantics and context in datasets** hinders the application of data analysis tools to, for example, identify errors like wrong values.

The Role of Semantics in AIDA

- The **lack of semantics and context in datasets** hinders the application of data analysis tools to, for example, identify errors like wrong values.
- **Adding semantics to Tabular Data**
  - Assigning a semantic type (e.g., a KG class) to an (entity) column
  - Matching a cell to a KG entity
  - Assigning a KG property to the relationship between two columns

(*) We assume the existence of a (possibly incomplete) ***Knowledge Graph (KG)*** *relevant to the domain.*

Adding semantics to Tabular Data



(*) Adapted from Efthymiou et al. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. ISWC 2017

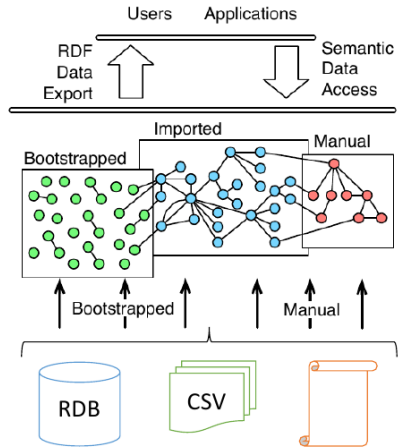
Contribution of Semantics in Data Wrangling Challenges

- *Data parsing*, e.g. converting csv's or tables.
- (+++)*Data dictionary*: basic types and semantic types.
- (++)*Data integration* from multiple sources (foreign key discovery).
- (++)*Entity resolution*: duplication and record linkage.
- *Format variability*: e.g. for dates and names.
- (+)*Structural variability* in the data.
- (++)Identifying and repairing *missing data*.
- (+)*Anomaly detection* and repair.
- (+++)**Metadata**/**contextual information**. (Semantic) data governance.

Exposing CSV files as Semantic Data

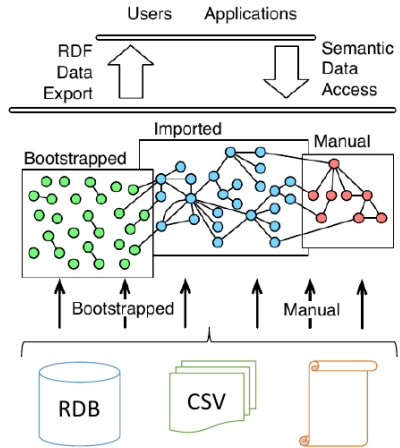
Virtual or Materialised

- **Virtual** (Knowledge Graph) exposure of the (CSV) data
- **Materialised** Data Export. Easy to exchange data.



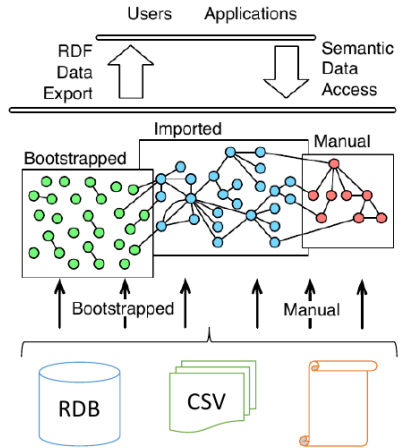
Motivation

- Enables the use of query languages like SPARQL
- Queries across CSV files
- Data described according to ontology vocabulary
- CSV files are not modified
- Queries as views to create "clean" CSV files
- Common language to integrate disparate resources



Ingredients

- **Ontology vocabulary**. Custom and/or given by a public KG.
- **Mappings**
  - RDB to RDF: relate ontology terms to queries over the RDB (W3C standard).
  - CSV to RDF: transformation functions.
- **Axioms** and **rules** (optional)



(Simple) Automatic Mappings

W3C directives:

- Table $\rightarrow$ class
- Row (primary key) $\rightarrow$ entity
- Foreign key $\rightarrow$ object property and relationship among entities
- Data column $\rightarrow$ data property
- Binary tables $\rightarrow$ fresh object properties

This transformation is very limited, especially in CSV files.

E. Jiménez-Ruiz et al. **BootOX: Practical Mapping of RDBs to OWL 2**, ISWC 2015

RDFox mappings/rules from CSV to RDF: `https://www.cs.ox.ac.uk/isg/tools/RDFox/`

Enhanced Transformation: Ongoing Work

- Assigning semantic column types: **ColNet**
- Cell to KG entities (canonicalization of entity mentions)
- Semantic relationships among columns
- Assigning column datatypes: **ptype**
- Canonicalization of values and units

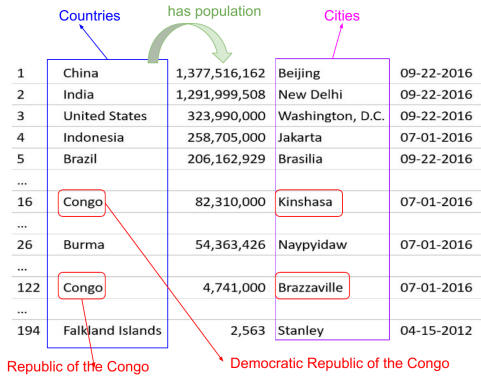
Enhanced Transformation (example)

- Simple (RDF) Triples:

  ⟨ ns:1, rdf:type, ? ⟩
  ⟨ ns:1, ns:col2, “China” ⟩

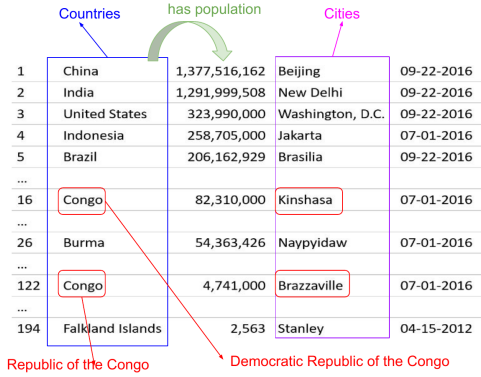
- Enhanced (RDF) Triples:

  ⟨ dbr:China, rdf:type, dbo:Country ⟩
  ⟨ dbr:Beijing, rdf:type, dbo:City ⟩
  ⟨ dbr:China, dbo:hasCapital, dbr:Beijing ⟩
  ⟨ dbr:China, dbo:hasPopulation, “. . . ” ⟩
  ⟨ dbo:hasPopulation, rdf:range, xsd:long ⟩



Enhanced Transformation (example)

- Enabled Query: “Countries with population greater than 10^6”
  - **SELECT ?c WHERE**
    ?c a dbo:Country .
    ?c dbo:hasPopulation ?p .
    ?p >= 1,000,000;
- Query results to be fed in subsequent analytical step.
- If data changes then only mappings may change



Adding Semantics to Tabular Data with ColNet

ColNet System

- **ColNet**: Embedding the Semantics of Web Tables for Column Type Prediction
  - Jiaoyan Chen, *Ernesto Jiménez-Ruiz*, Ian Horrocks, Charles Sutton
  - The Thirty-Third AAAI Conference on Artificial Intelligence (**AAAI 19**)
  - Extended version presented at **IJCAI 19**

DEPARTMENT OF
COMPUTER
SCIENCE

The
Alan Turing
Institute

THE UNIVERSITY
of EDINBURGH

Introduction

Challenges in Column Type Prediction

- **Multiple** and **hierarchical** classes
- Identifying a **fine-grained class** (dbo:BasketballPlayer VS dbo:Athlete VS dbo:Person)
- Column cells may have few or even empty KG entity correspondences, which is referred to as **knowledge gap**
- **Disambiguation**, e.g., “Virgin” as “Mary” or as “Virgin Media”

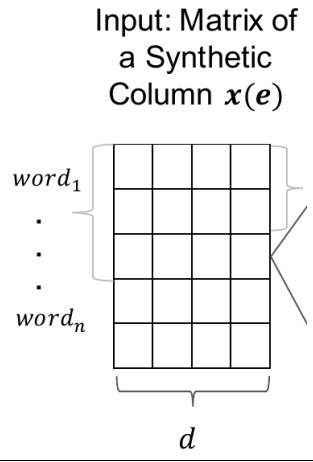
Methods

ColNet in a Nutshell

- utilizes **Convolutional Neural Networks** (**CNNs**), **semantic embeddings** and **Knowledge Graphs**.
- does **not** assume the existence of table **metadata**
- learns both **cell level** and **column level semantics**
- **automatically trains** prediction models relying on a Knowledge Graph
- uses **transfer learning** to address the knowledge gap
- **outperforms state-of-the-art** approaches when column entities are scarce

Samples and Embeddings

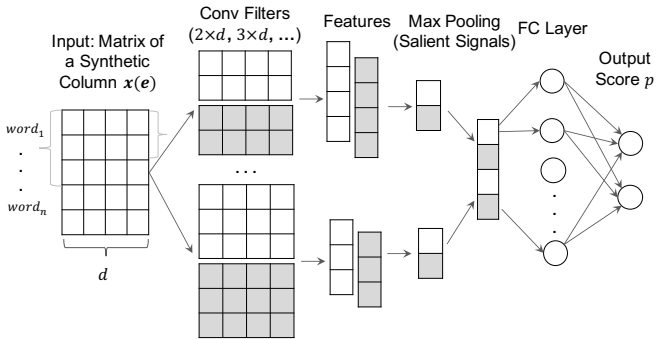
- The **CNNs** expect a matrix as input.
- **Semantic embeddings**: low-dimensional (vector space) representation of words.
- (Positive and negative) **samples** as a stack of word vectors
- In training, samples are **automatically labelled** with a KG class.



Training in ColNet

ColNet trains a CNN for each (candidate) KG class.

1. **pre-trains** the CNN with (general) samples from the KG,
2. **fine tunes** the CNN with (particular) samples from the table column.



Pre-training: (General) Samples from the Knowledge Graph

- We use **members/entities of the class in the KG**
  - For example (DBPedia): "dbr:Apple_Inc", "dbr:Microsoft", "dbr:Google", "dbr:Amazon.com" and "dbr:Alibaba Group" are members of the class "dbo:Company".

Pre-training: (General) Samples from the Knowledge Graph

- We use **members/entities of the class in the KG**
  - For example (DBPedia): "dbr:Apple_Inc", "dbr:Microsoft", "dbr:Google", "dbr:Amazon.com" and "dbr:Alibaba Group" are members of the class "dbo:Company".
- **A sample** (or synthetic column) is built by grouping a specific number of entities.
  - For example "dbr:Amazon" and "dbr:Alibaba Group" would form a sample of size 2 for the class "dbo:Company".
  - Matrix representation (stack of the word vectors): v(“Amazon”) $\oplus$ v(“Alibaba”) $\oplus$ v(“Group”)
- The **size of the sample** is one of our hyper-parameters.

Fine-Tuning: (Particular) Samples from Table Column

- **KG Look-up:** Lexical index based on entity labels
  * Cells $\rightarrow$ KG entity
  * e.g., Apple $\rightarrow$ "dbr:Apple", "dbr:Apple_Inc"

Column X
Apple
MS
Google

Fine-Tuning: (Particular) Samples from Table Column

- **KG Look-up:** Lexical index based on entity labels
  - Cells → KG entity
  - e.g., Apple → "dbr:Apple", "dbr:Apple_Inc"
- **KG Query:**
  - Entity → KG classes
  - e.g., "dbr:Apple" → "dbo:Fruit"
  - e.g., "dbr:Apple_Inc" → "dbo:Company"

Column X
Apple
MS
Google

Fine-Tuning: (Particular) Samples from Table Column

- **KG Look-up:** Lexical index based on entity labels
  - Cells → KG entity
  - e.g., Apple → "dbr:Apple", "dbr:Apple_Inc"
- **KG Query:**
  - Entity → KG classes
  - e.g., "dbr:Apple" → "dbo:Fruit"
  - e.g., "dbr:Apple_Inc" → "dbo:Company"

Column X
Apple
MS
Google

- **Sample generation:** segments of the column that are "dbo:Company".

Negative Samples for a KG class (training)

- **Balanced** positive and negative **samples**
- Source of (**general**) negative samples:
  * Exploits non members (especially from disjoint classes)

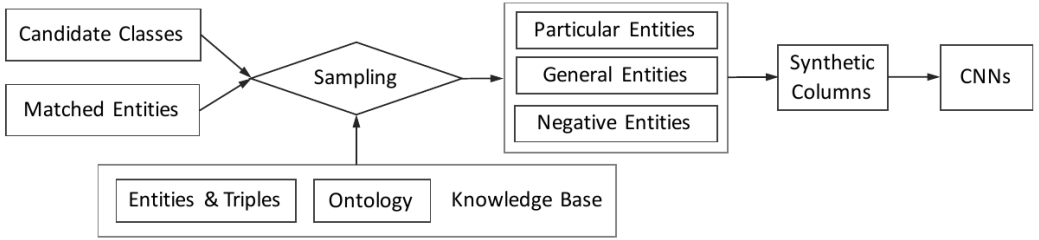
Negative Samples for a KG class (training)

- **Balanced** positive and negative **samples**
- Source of (**general**) negative samples:
  * Exploits non members (especially from disjoint classes)
- Source of (**particular**) negative samples:
  * Entities that are disjoint with the KG class and appear together in the column
  * e.g., members of "dbo:Fruit" like "dbr:Apple"

Training with Transfer Learning in ColNet

- Recall two steps CNN training: **pre-training** and **fine-tuning**
- **Benefits**:
  - **Pre-training**: deals with the shortage of particular samples: knowledge gap or short columns
  - **Fine-tuning**: bridges the data distribution gap between KG entities and table cells
- *[Impact analysis in the evaluation]*

Training and Sampling Example



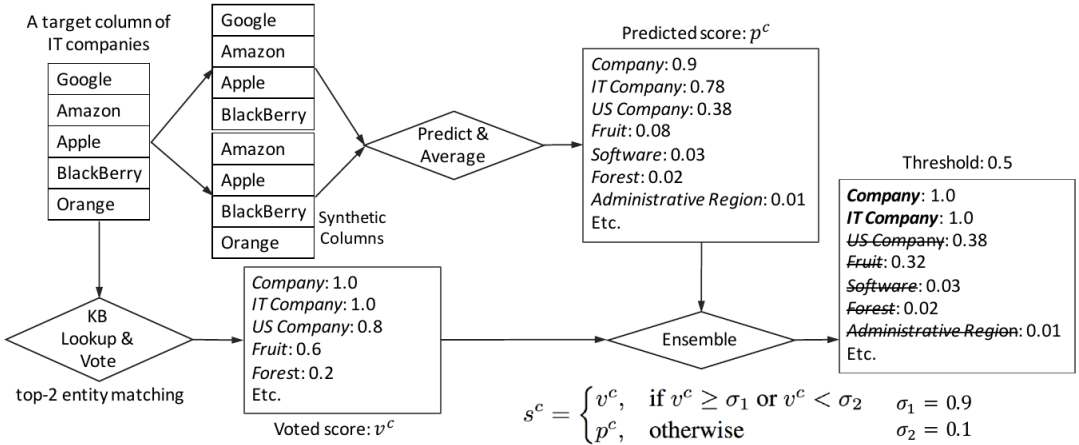
Prediction in ColNet

- **Prediction samples** are composed by segments of the column
- In our example: (“Apple", “MS", “Google") as column.
  - e.g. size 1: v(“Apple").
  - e.g. size 2: v(“Apple") $\oplus$ v(“Google").
  - e.g. size 3: v(“Apple") $\oplus$ v(“Google") $\oplus$ v(“MS").

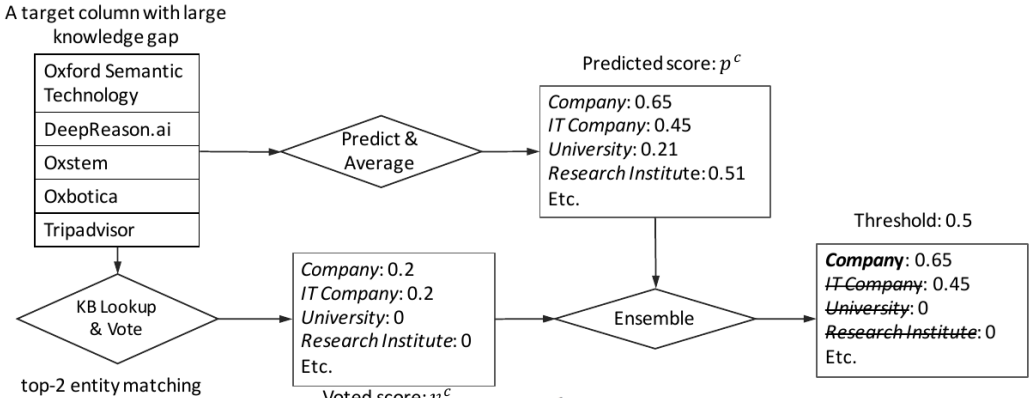
Prediction in ColNet

- **Prediction samples** are composed by segments of the column
- In our example: (“Apple", “MS", “Google") as column.
  - e.g. size 1: v(“Apple").
  - e.g. size 2: v(“Apple") $\oplus$ v(“Google").
  - e.g. size 3: v(“Apple") $\oplus$ v(“Google") $\oplus$ v(“MS").
- **Benefit of the sample size**: learn inter-cell correlations (locality features) by CNN
  - Expected prediction: “dbo:Company"
  - Prediction cell by cell: score from 0.33 to 0.66
  - Prediction score considering the correlation between cells $\approx$ 1.0
  - *[Impact analysis in the evaluation]*

Prediction in ColNet (Example 1)



Prediction in ColNet (Example 2)



Evaluation

Evaluation Setting: Data

- **DBPedia** as the KG
- Word embedding: **Word2vec** model trained with the latest dump of Wikipedia pages
- **T2Dv2** (tables from the Web) and **Limaye** (tables from Wikipedia pages) datasets
- Limaye dataset more challenging in terms of **knowledge gap**

Name	Columns	Avg. Cells	Different “Best” (“Okay”) Classes
T2Dv2	411	124	56 (35)
Limaye	428	23	21 (24)

Evaluation Setting: Ground Truth and Baselines

- **Evaluation models**
  - **Strict** (best hit only) and **tolerant** (compatible hits) evaluation models
- **Baselines**
  - DBPedia Lookup + Vote
  - T2K Match [Ritze et al. WIMS'15]
  - Efthymiou + Vote [Efthymiou et al. ISWC'17]
  - Other state of the art systems not available

Our methods: ColNet and $\text{ColNet}_{\text{Ensemble}}$

Overall Results on Limaye

Precision (P), Recall (R), F1 score (F)

Models	Methods	PK Columns
Tolerant	ColNet$_{Ensemble}$	**0.796**, 0.799, **0.798**
	ColNet	0.763, **0.820**, 0.791
	Lookup-Vote	0.732, 0.660, 0.694
	T2K Match	0.560, 0.408, 0.472
	Efthymiou17-Vote	0.759, 0.414, 0.536
Strict	ColNet$_{Ensemble}$	0.602, **0.639**, **0.620**
	ColNet	0.576, 0.619, 0.597
	Lookup-Vote	0.571, 0.447, 0.501
	T2K Match	0.453, 0.330, 0.382
	Efthymiou17-Vote	**0.626**, 0.357, 0.454

- **Prediction impact**
  - ColNet$_{Ensemble}$ and ColNet > Lookup-Vote
  - Improvement of recall
- **Ensemble impact**
  - ColNet$_{Ensemble}$ > ColNet
    - Improvement of precision
- **Comparison with the state-of-the-art**
  - ColNet$_{Ensemble}$ and ColNet > T2K Match
    - ColNet$_{Ensemble}$ and ColNet has competitive precision as Efthymiou17-Vote, but much higher recall and F1 score

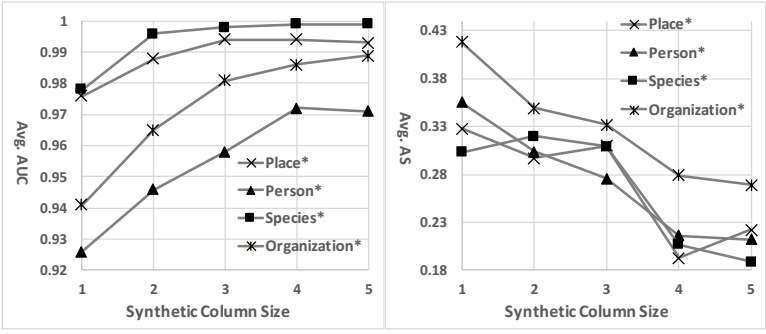
Overall Results on T2Dv2

Precision (P), Recall (R), F1 score (F)

Models	Methods	All Columns	PK Columns
Tolerant	ColNet$_{Ensemble}$	**0.917**, **0.909**, **0.913**	**0.967**, **0.985**, **0.976**
	ColNet	0.845, 0.896, 0.870	0.927, 0.960, 0.943
	Lookup-Vote	0.909, 0.865, 0.886	0.965, 0.960, 0.962
	T2K Match	0.664, 0.773, 0.715	0.738, 0.895, 0.809
Strict	ColNet$_{Ensemble}$	0.853, **0.846**, **0.849**	0.941, **0.958**, **0.949**
	ColNet	0.765, 0.811, 0.787	0.868, 0.898, 0.882
	Lookup-Vote	**0.862**, 0.821, 0.841	**0.946**, 0.941, 0.943
	T2K Match	0.624, 0.727, 0.671	0.729, 0.884, 0.799

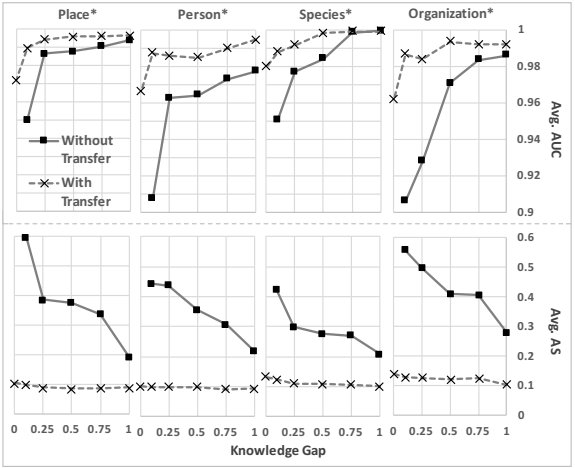
- **Prediction impact**
- **Ensemble impact**
- **Knowledge gap impact**
  - Limaye is harder than T2Dv2
  - Limaye has shorter columns in average, which causes larger knowledge gap
  - Improvements of ColNet$_{Ensemble}$ and ColNet on Limaye are more significant, since ColNet deals with the knowledge gap

Impact of Synthetic Column Size on CNNs



The testing **performance of CNNs on Truly Matched (TM)** classes [left] and **Falsely Matched (FM)** classes [right] for types of columns: Place, Person, Species & Organization. **AUC: area under ROC curve, AS: average score**

Impact of Transfer Learning and the Knowledge Gap on CNNs



- The testing performance of CNNs of TM classes [above] and FM classes [below]
  - under **different knowledge gaps**
  - **with and without transfer learning**
  - four types of columns: Place, Person, Species and Organization
- The knowledge gap is simulated by randomly selecting a ratio of particular entities for training. **The lower ratio, the larger gap.**

Future Work

Future work

- Learning stronger table locality features (contextual semantics) - IJCAI 19
- Use ColNet as the basis for other Web table to KG matching tasks

Future work

- Learning stronger table locality features (contextual semantics) - IJCAI 19
- Use ColNet as the basis for other Web table to KG matching tasks
- Application of ColNet in the data science pipeline as an AI assistant
  - Communication with *ptype* and *DataDiff*

Future work

- Learning stronger table locality features (contextual semantics) - IJCAI 19
- Use ColNet as the basis for other Web table to KG matching tasks
- Application of ColNet in the data science pipeline as an AI assistant
  - Communication with *ptype* and *DataDiff*
- Iterative creation of a shared KG: general knowledge, output from AI assistants, semantic data governance, etc.
  - KG shared among data analysis tools (**Semantics-aware AI assistants**)

Knowledge Graph Embedding for Ecotoxicological Effect Prediction

KG for Ecotoxicological Prediction

- **Knowledge Graph Embedding for Ecotoxicological Effect Prediction**
- Erik B. Myklebust, Ernesto Jimenez-Ruiz, Jiaoyan Chen, Raoul Wolf, and Knut Erik Tollefsen
- International Semantic Web Conference (**ISWC 2019**)
  - *Best Student In-Use Paper*

Ecological Risk Assessment

Exposure

Risk assessment is an estimation of cumulative risk on individuals, populations, communities, and ecosystems from chemical pollutants.

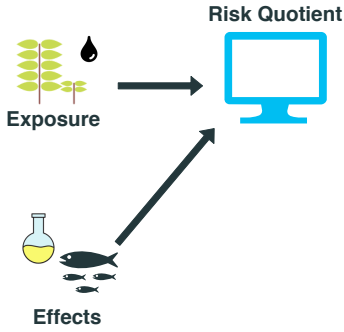
Ecological Risk Assessment

Exposure

Effects

Effect concentrations are found using organism experiments.

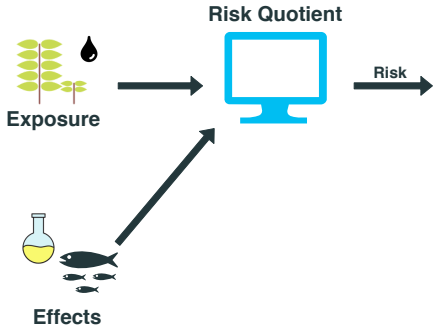
Ecological Risk Assessment



$$RQ = \frac{\text{environmental concentration}}{\text{effect concentration}}$$

RQs coverage is limited by effect concentration experiments.

Ecological Risk Assessment

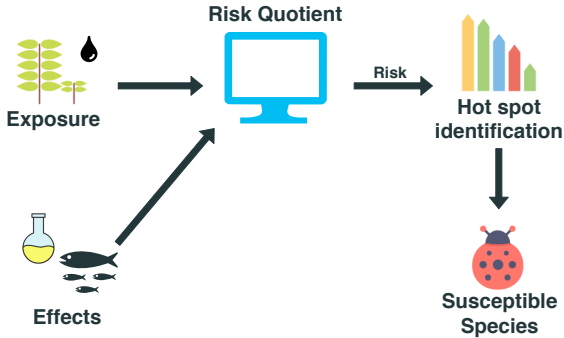


$$risk_{group} \approx \sum^{chemicals} RQ$$

Risk for a group of species.

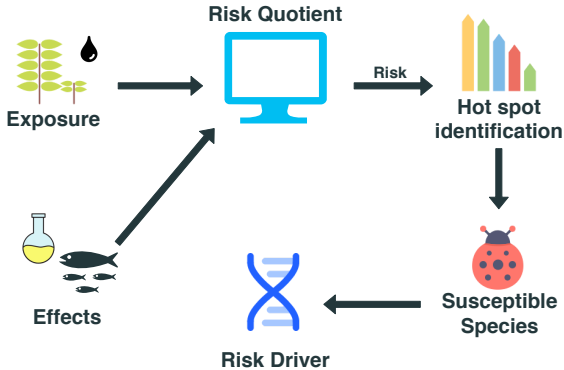
The group can contain all species in the ecosystem.

Ecological Risk Assessment



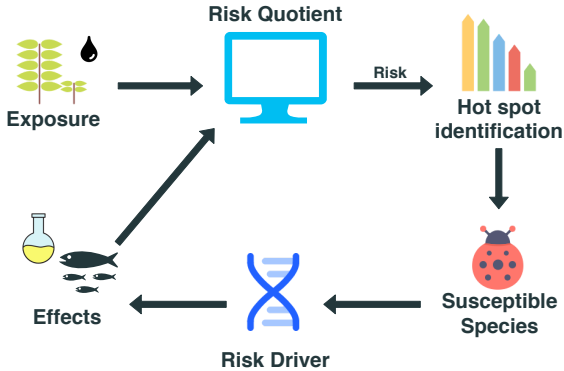
The risk is used to find further susceptible species.

Ecological Risk Assessment



Risk driver describes *how* the chemical affects an organism.

Ecological Risk Assessment



New effect hypotheses are then tested in the laboratory.

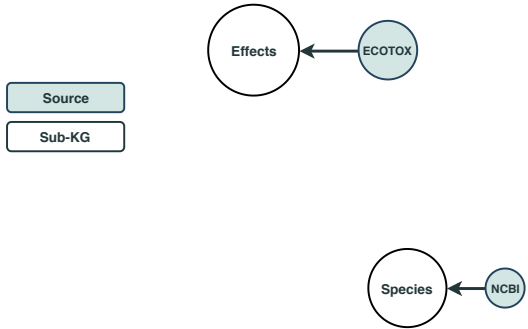
The TERA Knowledge Graph

The Toxicological and Risk Assessment (TERA) knowledge graph integrates data sources varying in format.

The TERA Knowledge Graph

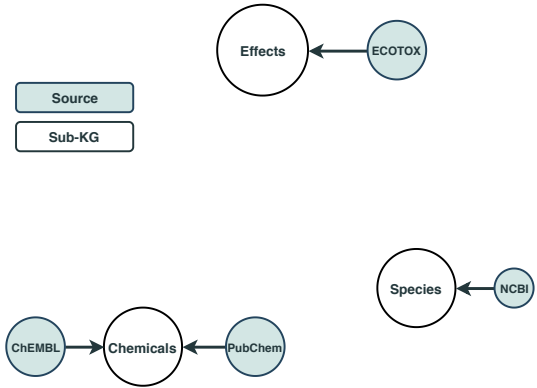
ECOTOX is the largest (public) source of effect data.

The TERA Knowledge Graph



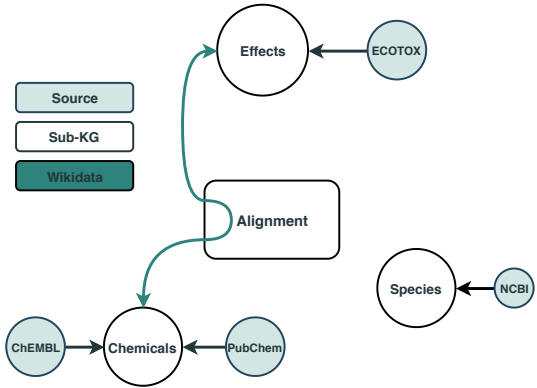
NCBI's tabular taxonomy is converted to a hierarchy.

The TERA Knowledge Graph



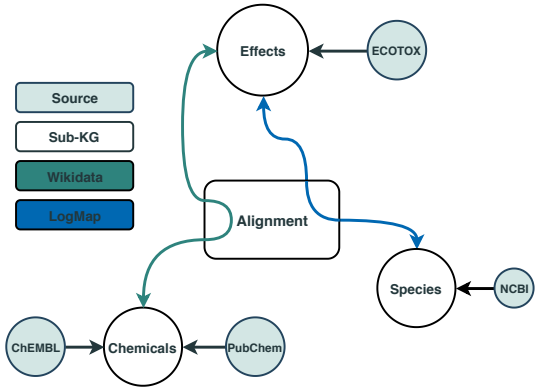
Importing the ChEMBL and PubChem knowledge graph.

The TERA Knowledge Graph



Aligning proprietary chemical identifiers in ECOTOX to open identifiers in PubChem.

The TERA Knowledge Graph



Aligning taxonomies using ontology alignment tool LogMap.

Effect Prediction Problem Definition

Effect Prediction Problem Definition

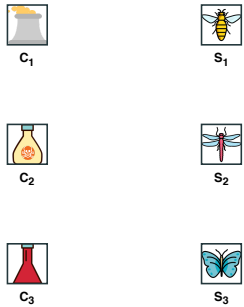
C_1

C_2

C_3

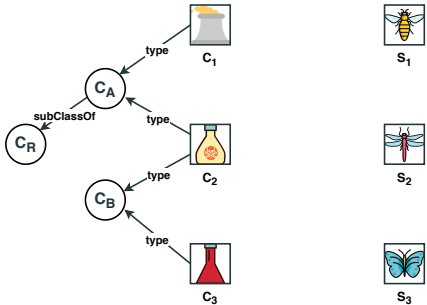
Chemicals

Effect Prediction Problem Definition



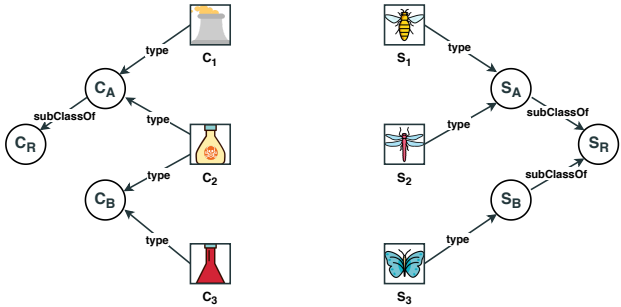
Species

Effect Prediction Problem Definition



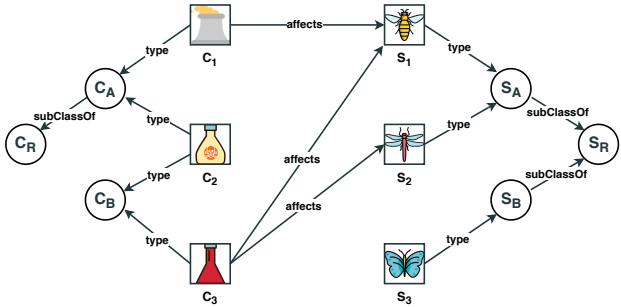
Chemical classification

Effect Prediction Problem Definition



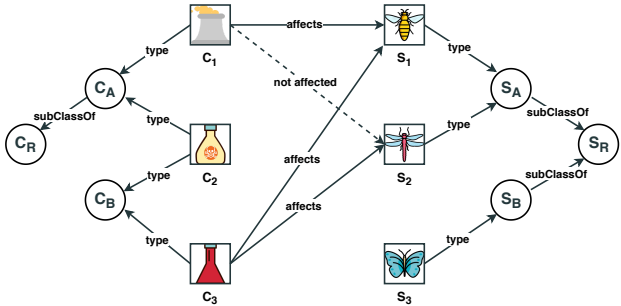
Taxonomy

Effect Prediction Problem Definition



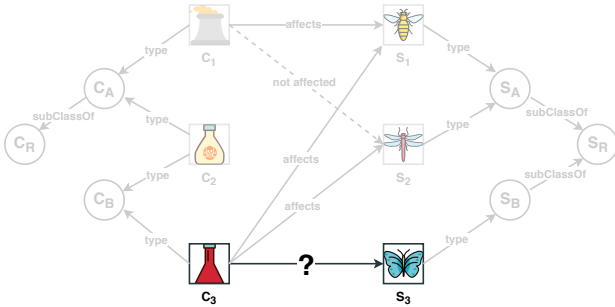
Positive samples

Effect Prediction Problem Definition



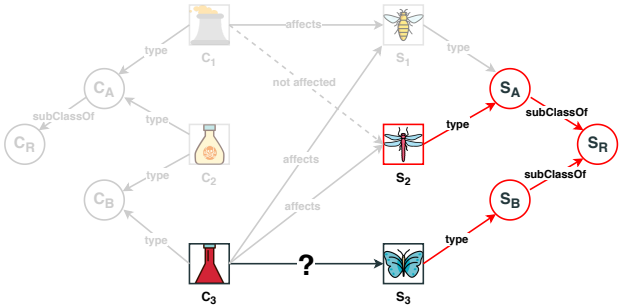
Negative samples

Taxonomic Distance Model - Baseline (BL)



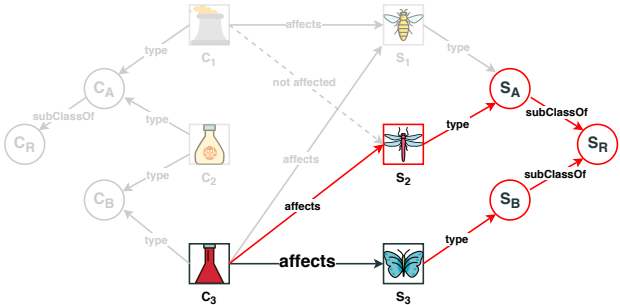
Does C_3 affect S_3?

Taxonomic Distance Model - Baseline (BL)



$$dist(S_3, S_2) = 4$$

Taxonomic Distance Model - Baseline (BL)

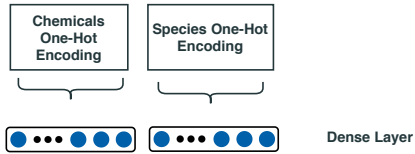


Yes, C_3 affects S_3

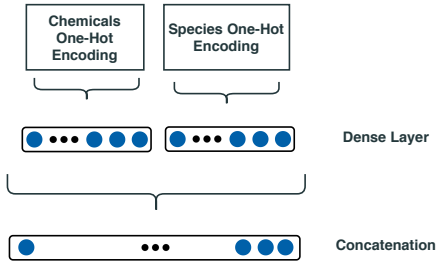
Multi-layer perceptron (MLP)

Multi-layer perceptron (MLP)

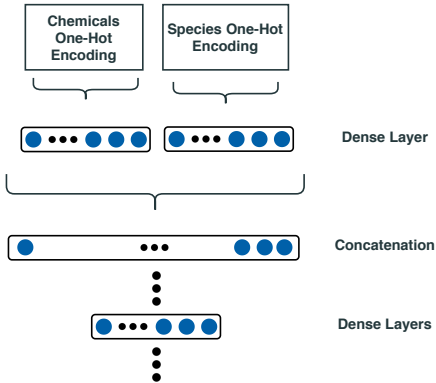
Multi-layer perceptron (MLP)



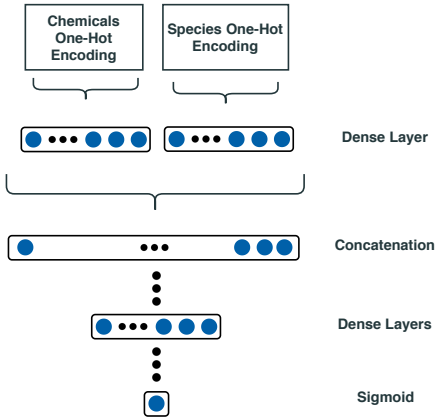
Multi-layer perceptron (MLP)



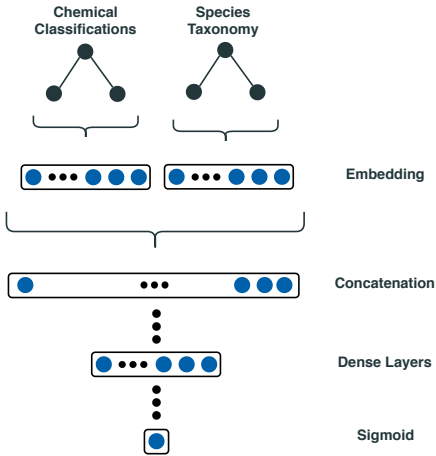
Multi-layer perceptron (MLP)



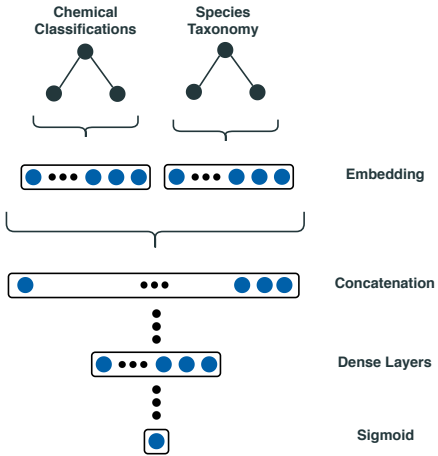
Multi-layer perceptron (MLP)



KG embedding + MLP



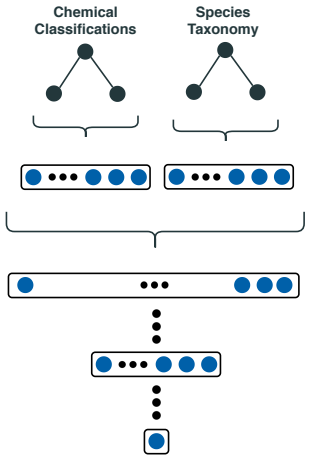
KG embedding + MLP



Three embedding models:

1. TransE
2. DistMult
3. HolE

KG embedding + MLP



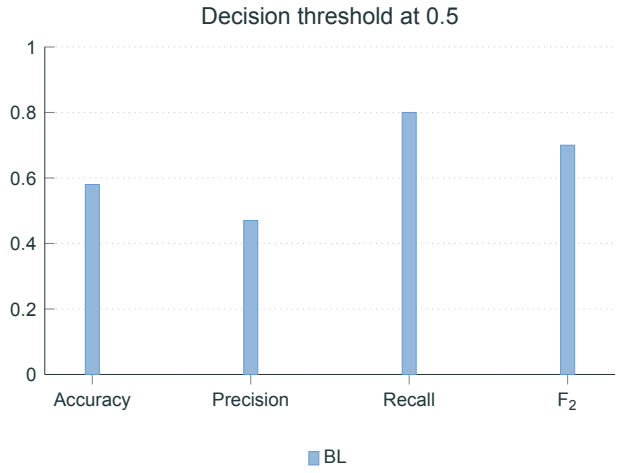
Three embedding models:

1. TransE
2. DistMult
3. HolE

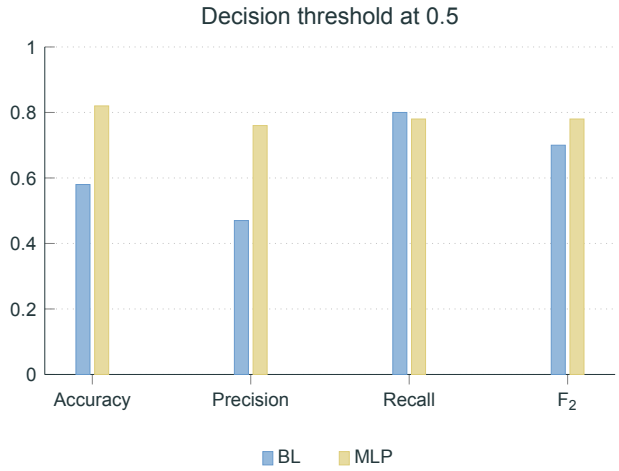
Optimization:

Simultaneous optimization of prediction and embedding models.

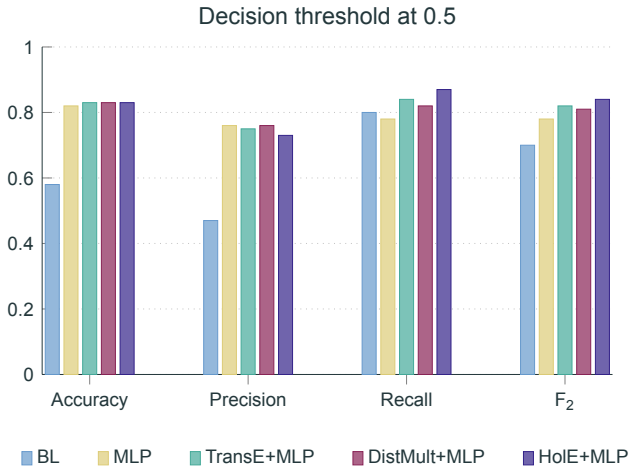
Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL

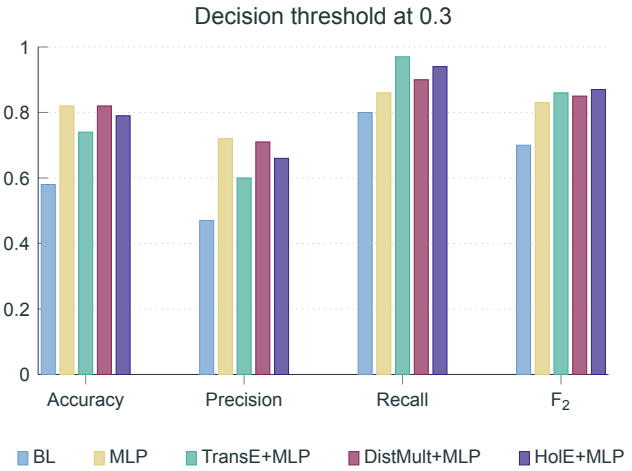
Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL
MLP

Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F2
BL
MLP
TransE+MLP
DistMult+MLP
HolE+MLP

Results


Decision threshold at 0.3
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F2
BL
MLP
TransE+MLP
DistMult+MLP
HolE+MLP

Summary and Future Work

☑ Improved data access using TERA KG.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

- ☐ Expand the TERA knowledge graph with other relevant data, *e.g.*, habitat.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

- ☐ Expand the TERA knowledge graph with other relevant data, *e.g.*, habitat.
- ☐ Explore the use of more sophisticated models

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

- ☐ Expand the TERA knowledge graph with other relevant data, *e.g.*, habitat.
- ☐ Explore the use of more sophisticated models
- ☐ Move from binary labels to chemical concentrations.

References and Acknowledgements

References and Acknowledgements

- Optique project: `http://optique-project.eu/`
- AIDA project: `https://www.turing.ac.uk/research/research-projects/artificial-intelligence-data-analytics`
- Semantics for AIDA: `https://github.com/alan-turing-institute/SemAIDA/`
- NIVA use case: `https://github.com/Erik-BM/NIVAUC`
- SIRIUS Centre for Scalable Data Access: `https://sirius-labs.no/`


SIRIUS


DEPARTMENT OF
COMPUTER
SCIENCE
The
Alan Turing
Institute
NIVA
Norwegian Institute for Water Research

Questions?

- Ernesto Jiménez Ruiz
- ernesto.jimenez-ruiz@city.ac.uk
- ernesto.jimenez.ruiz@gmail.com