The Impact of Graph Embeddings on Enhancing Machine Learning Models

Student: Liam James Glennie England

Supervisor: Anna Queralt Calafat

Master in Data Science

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat d'Informàtica de Barcelona

FIB

Motivation

- Traditional ML models tend to struggle with poor quality datasets
  - Features may be scarce
  - Lack clear patterns
  - Exhibit weak relationships

- Graph embeddings could enrich datasets by
  - Capturing
    - Entity relationships
    - Structural information
    - Semantic information
  - Provide context to the ML model
  - Encode attributes from the KG related to the target to be predicted

- Main objective: Enhance ML models with graph embeddings

Table of Contents

Objectives

- **RO1:** Evaluate if graph embeddings can enhance machine learning models in classification and regression tasks on poor quality datasets
- **RO2:** Assess whether graph embedding algorithms can be improved for classification and regression tasks
- **RO3:** Analyse the behaviour of graph embedding algorithms depending on the quality and form of relevant data in the knowledge graph

Related Work

Graph Embedding Algorithms

- Translation Method
  - Treat link prediction as a goemetric challenge
  - TransE, TransH, TransR
- Matrix Factorisation
  - Preserve essential graph properties, like node similarity
  - Graph Laplacian Eigenmaps and Node Proximity Matrix Factorisation
- Deep Learning Methods
  - Random Walk
    - Suited for node classification, clustering, and node similarity
    - node2vec, RDF2Vec, OWL2Vec*
  - Non-random walk
    - Offer global representation of graph
    - AutoEncoders and GNN

Related Work

KG Embedding Evaluation

- Portisch et al. compared link prediction algorithms against deep learning algorithms
  - Used embedding algorithms like TransE and RDF2Vec
  - Evaluated embeddings on multiple ML tasks
    - TransE performed best in link prediction
    - RDF2Vec was best for data mining applications
    - TransE and RDF2Vec embeddings showed promise for classification and regression tasks
  - This work served as inspiration for main objective of this thesis

Related Work

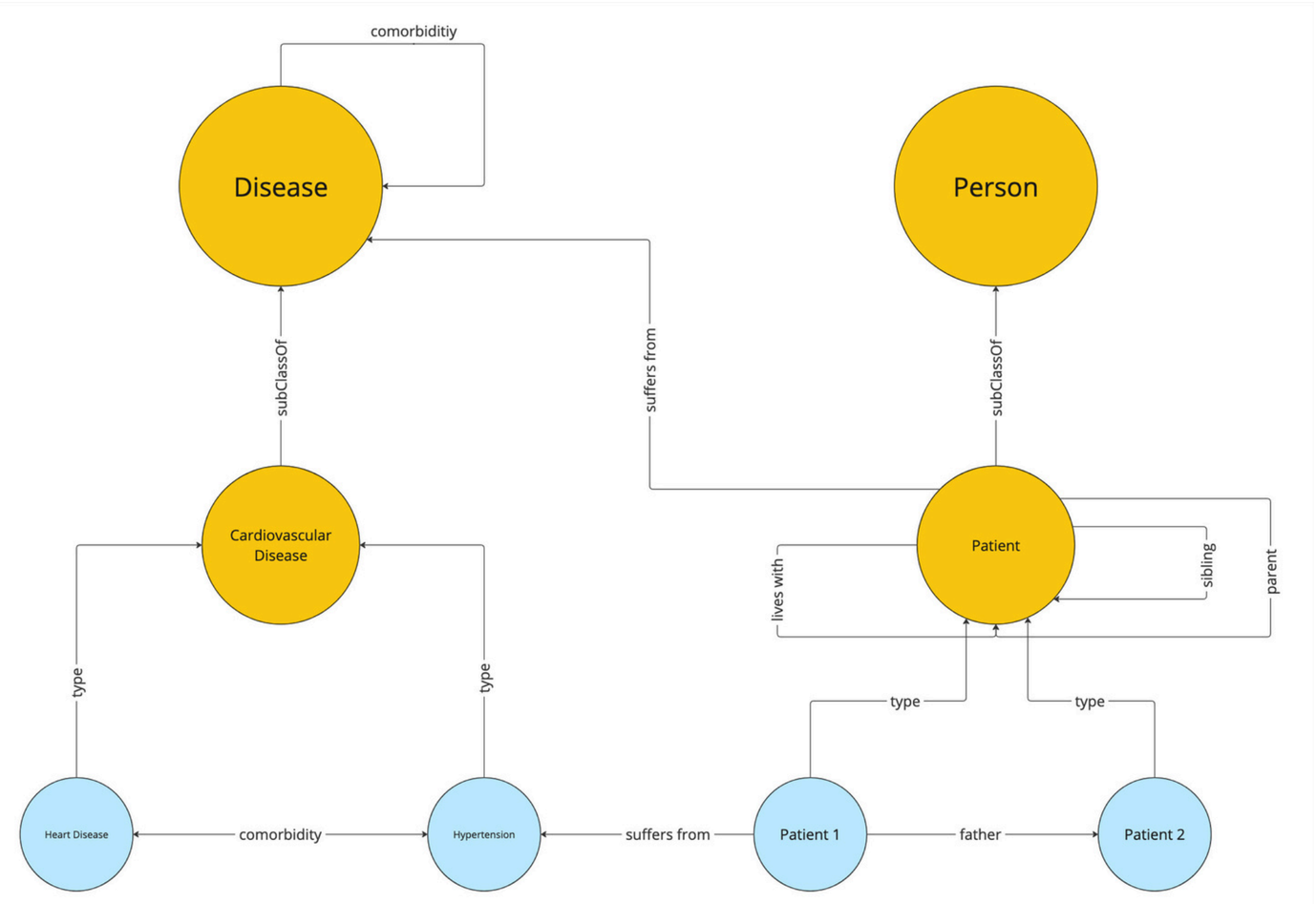
Dataset Enrichment with KGs

- Taveekarn and Vanderwiele both propose solutions to enrich datasets with KGs
  - Augment datasets with additional features derived from a KG
    - They implement feature selection over a KG
    - Integrate selected features into dataset as a column
  - Different to our solution
    - They enrich datasets with features as columns
    - We enrich with graph embeddings

Background

Knowledge Graph

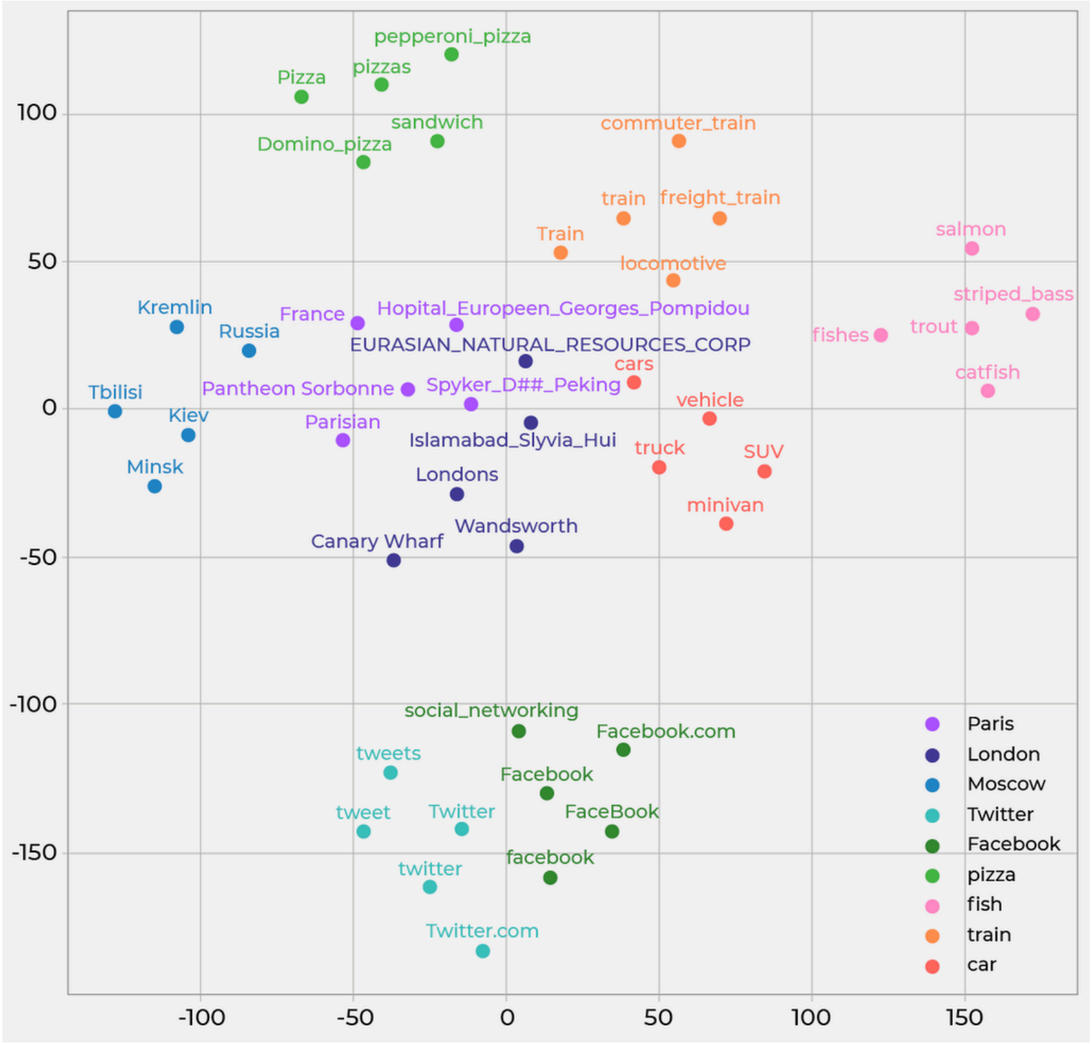
- Directed labelled graph
- Combine multiple sources to represent knowledge
- Two parts:
  - Schema
  - Instances
- Frameworks
  - RDF
  - RDF-S
  - OWL



Background

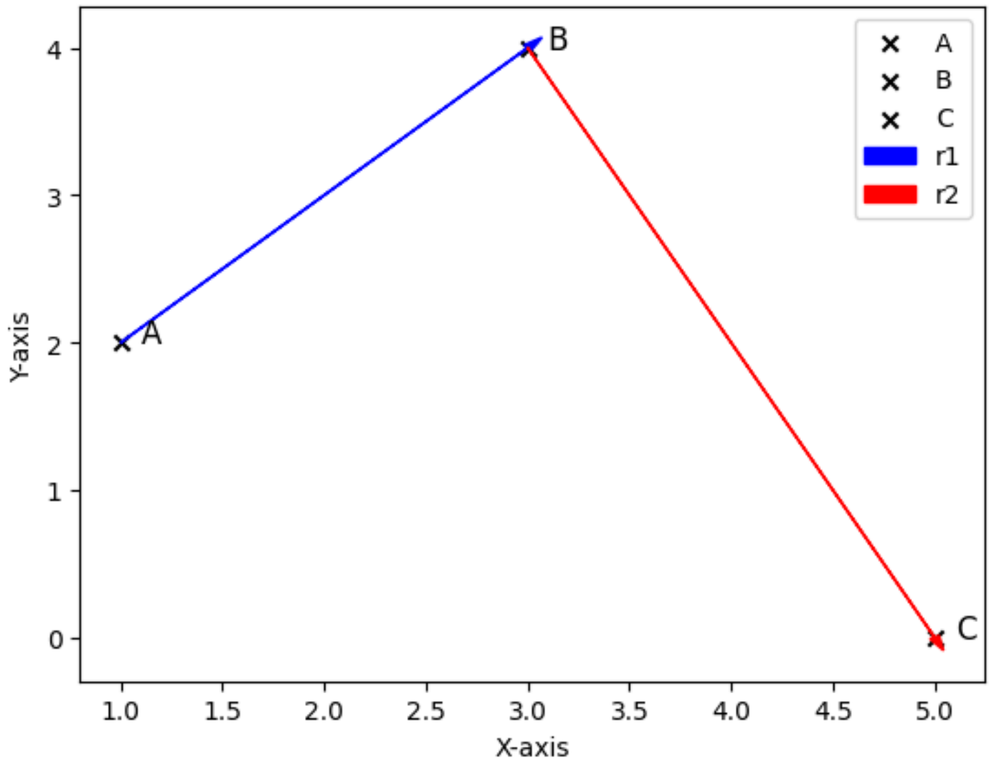
Embeddings

- Numerical representation of real world
- Map semi-structured or unstructured data to a vector space
- Similar entities found close together
- Graph Embeddings encode
  - Relationships
  - Semantics
  - Attributes



Background

TransE - Embedding Algorithm

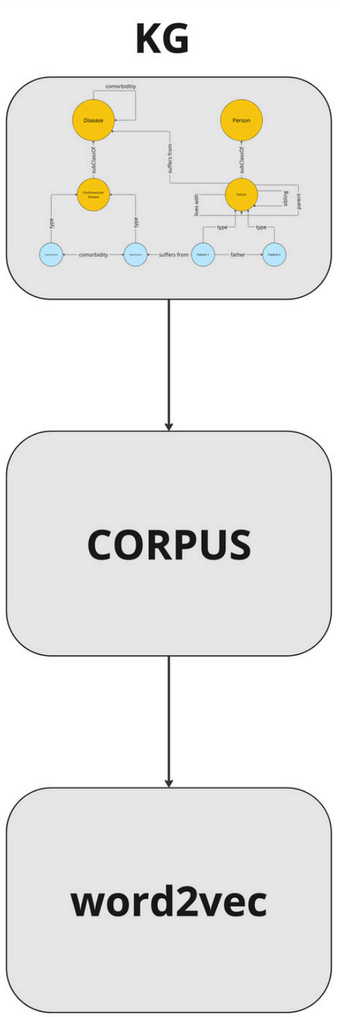


- Link Prediction Algorithm
- Focuses on relationship between entities
- Given:
  - 3 entities: A, B, C
  - Triples:
    - (A, r1, B)
    - (B, r2, C)
- Should hold:
  - A+r1≈B --> (1, 2) + r1 ≈ (3, 4)
  - B+r2≈C --> (3, 4) + r2 ≈ (5, 0)

Background

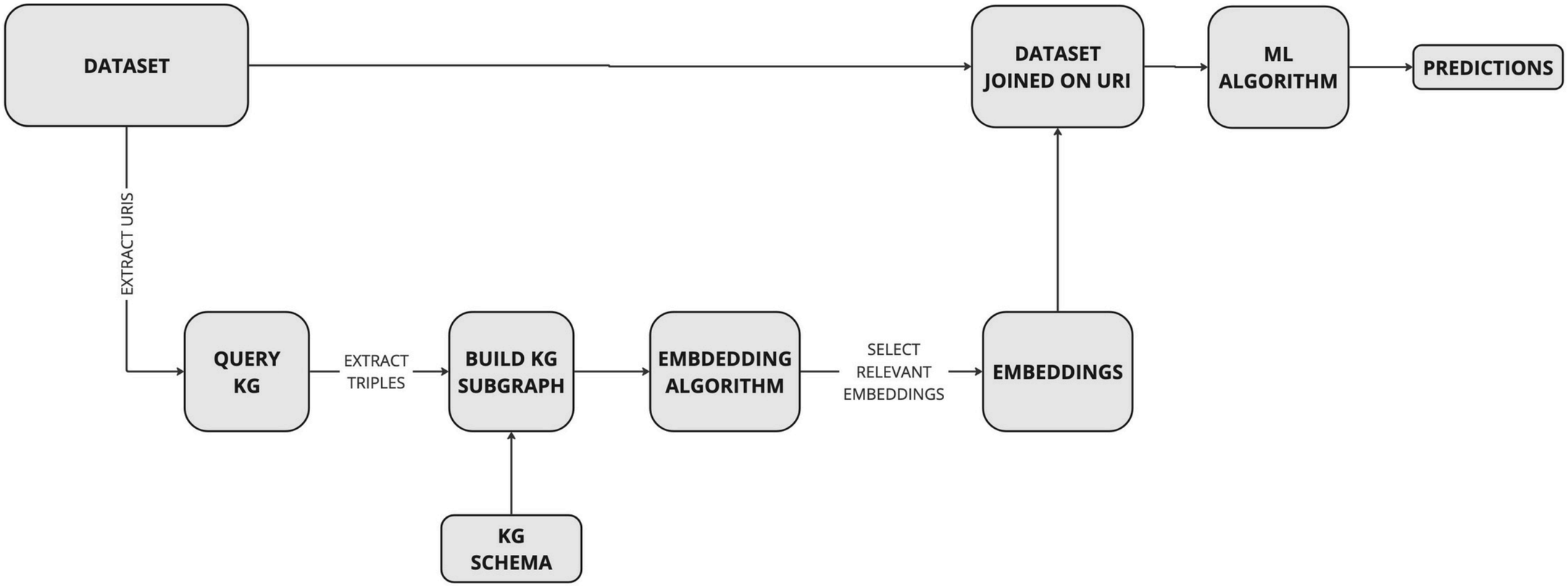
OWL2Vec* - Embedding Algorithm

- Random walk-based DL Algorithm
- Extends RDF2Vec
- Two components:
  - Corpus construction:
    - Random Walks on KG
    - Structure, lexical & combined documents
  - word2vec Training



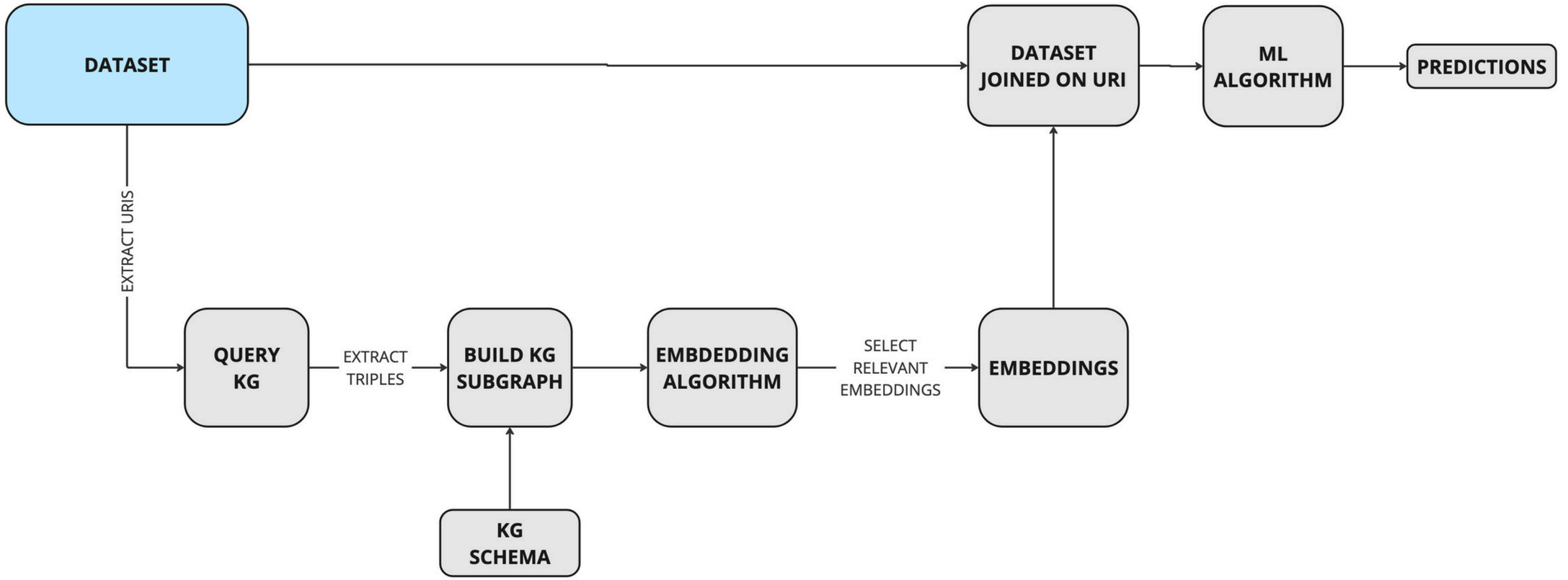
Methodology - Part 1

Process



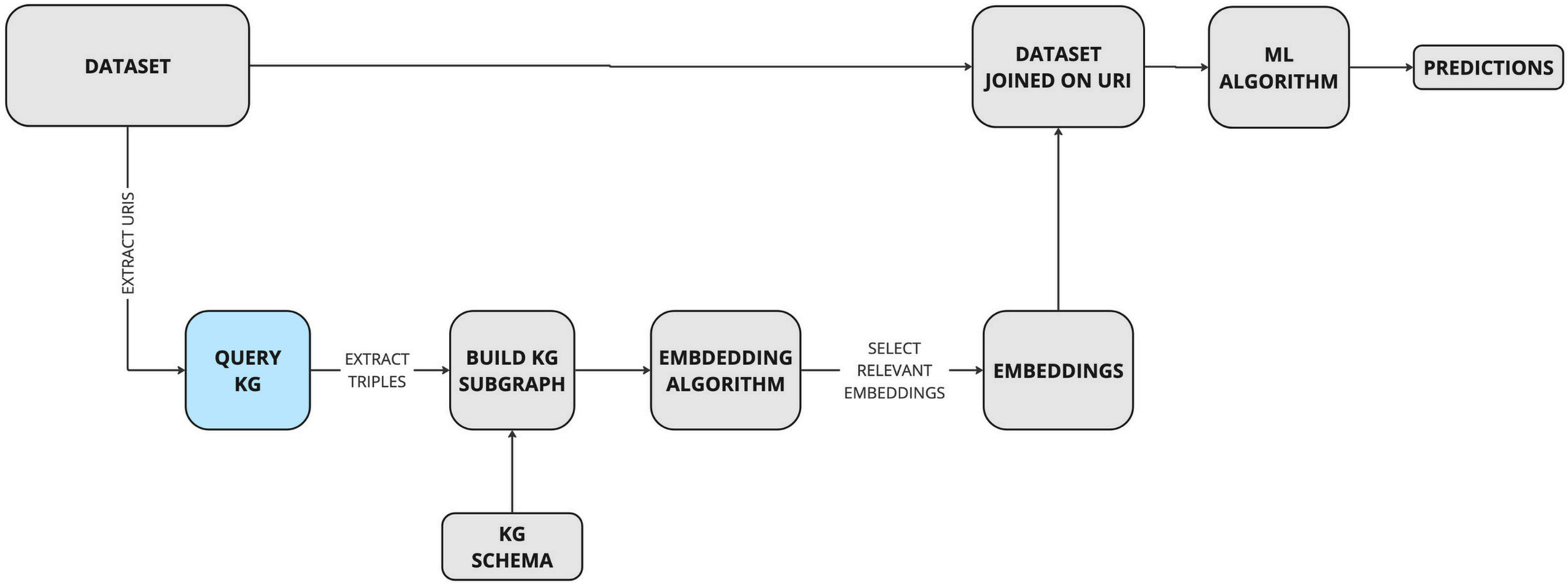
Methodology - Part 1

Process



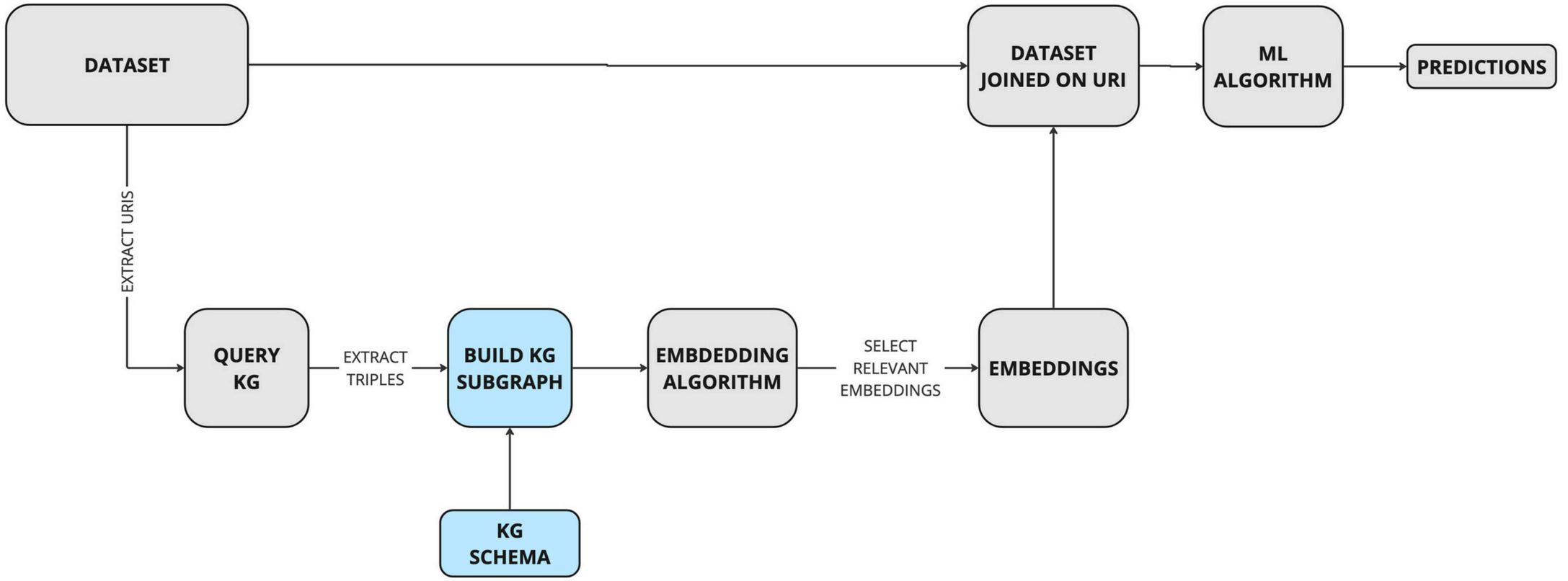
Methodology - Part 1

Process



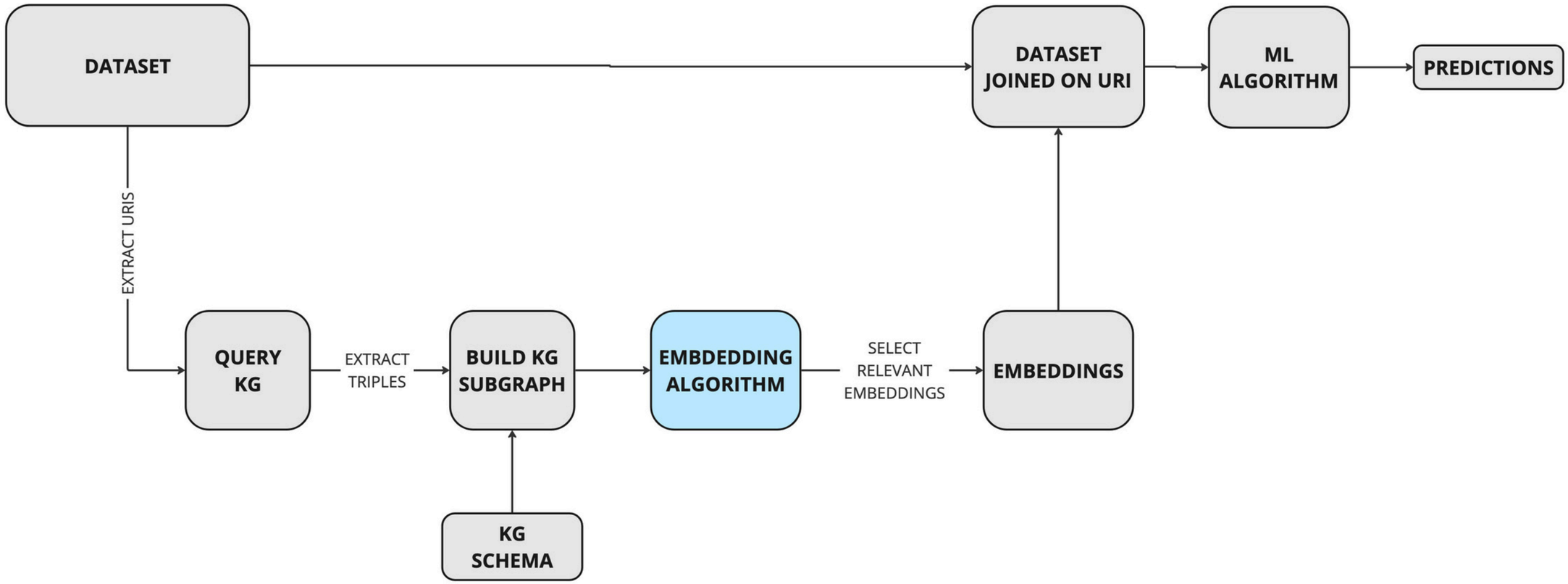
Methodology - Part 1

Process



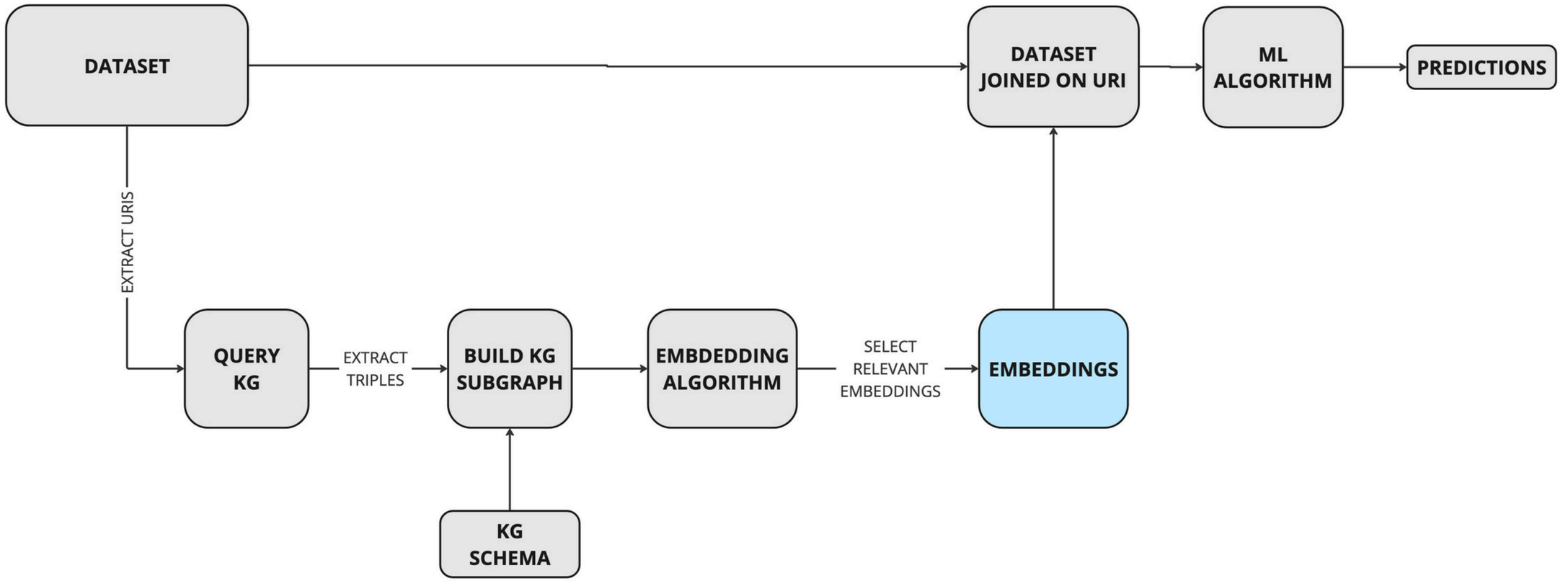
Methodology - Part 1

Process



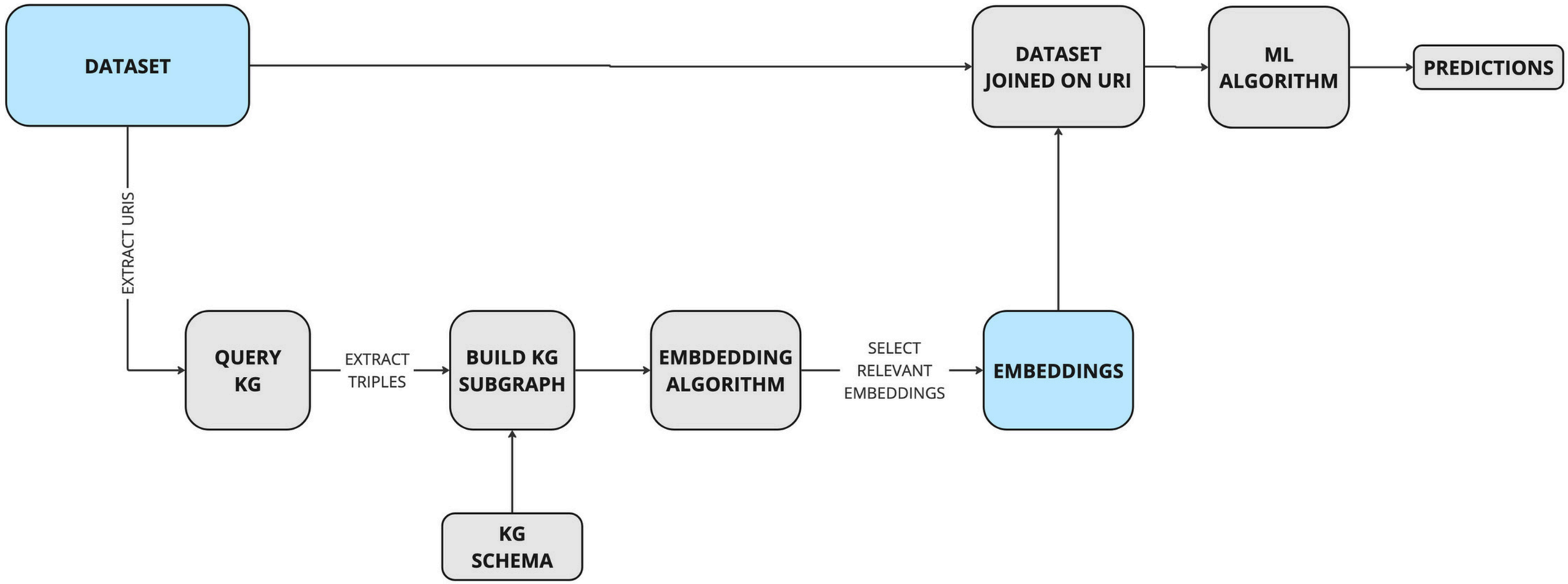
Methodology - Part 1

Process



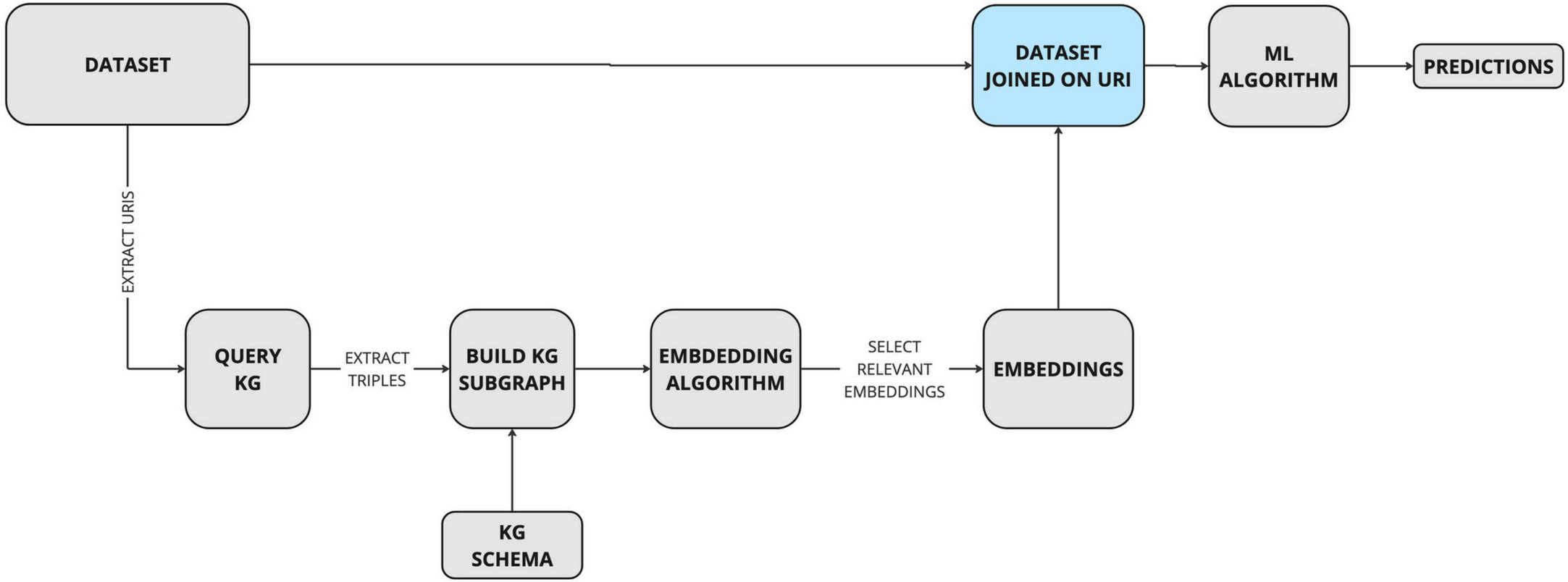
Methodology - Part 1

Process



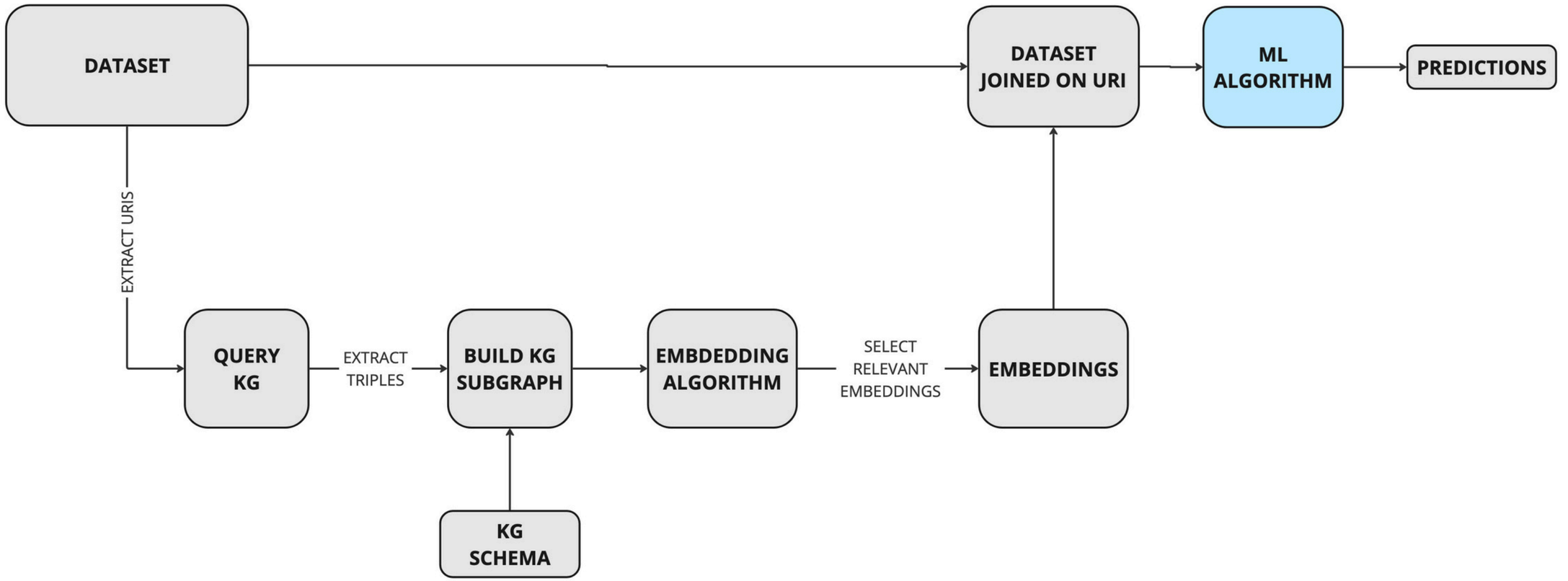
Methodology - Part 1

Process



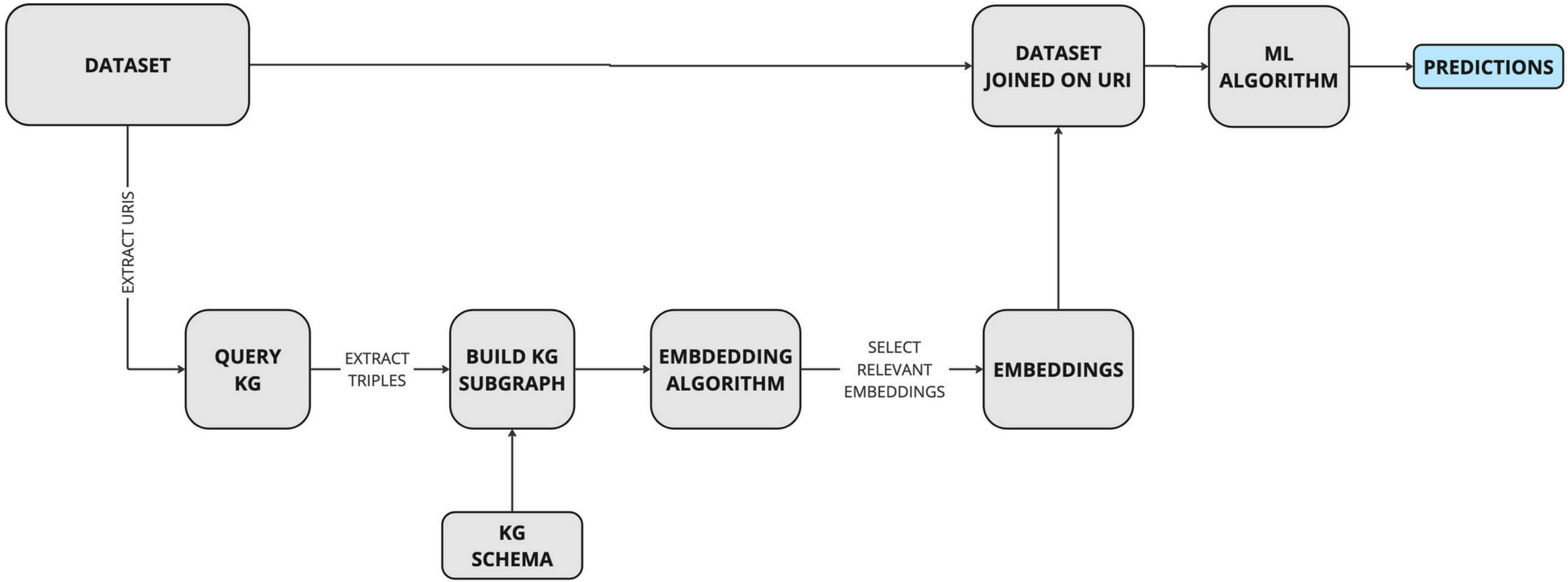
Methodology - Part 1

Process



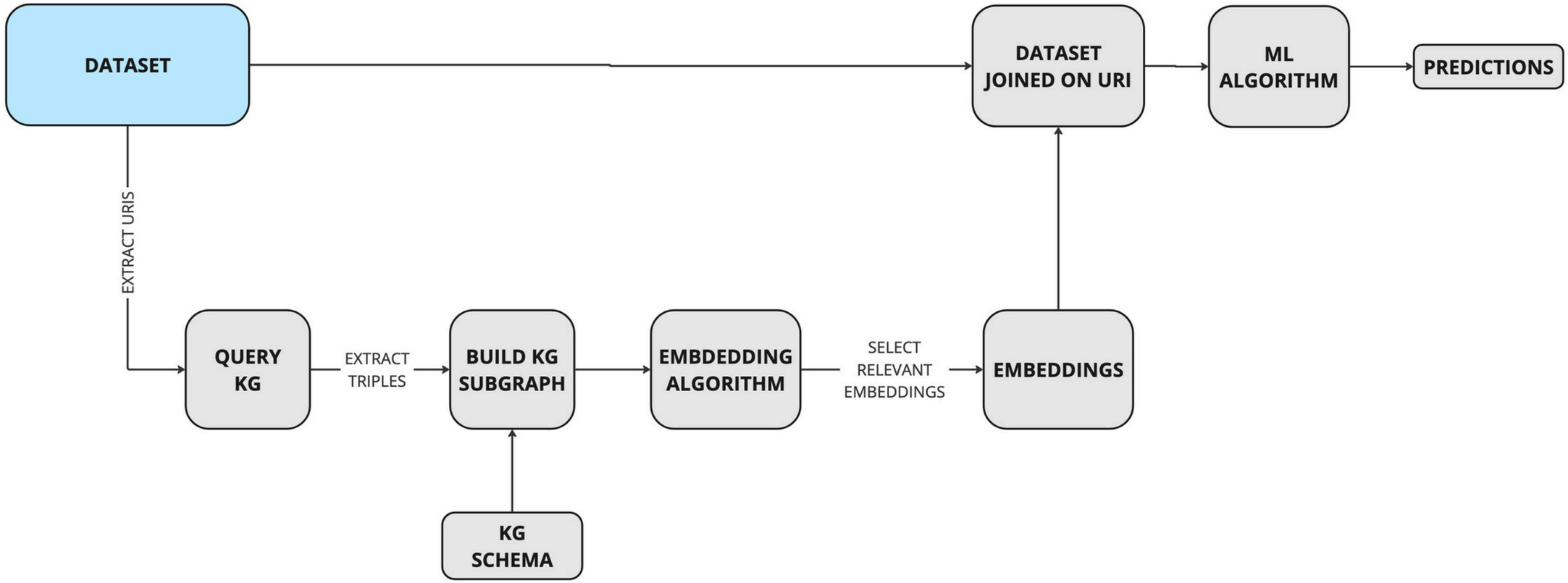
Methodology - Part 1

Process



Methodology - Part 1

Datasets



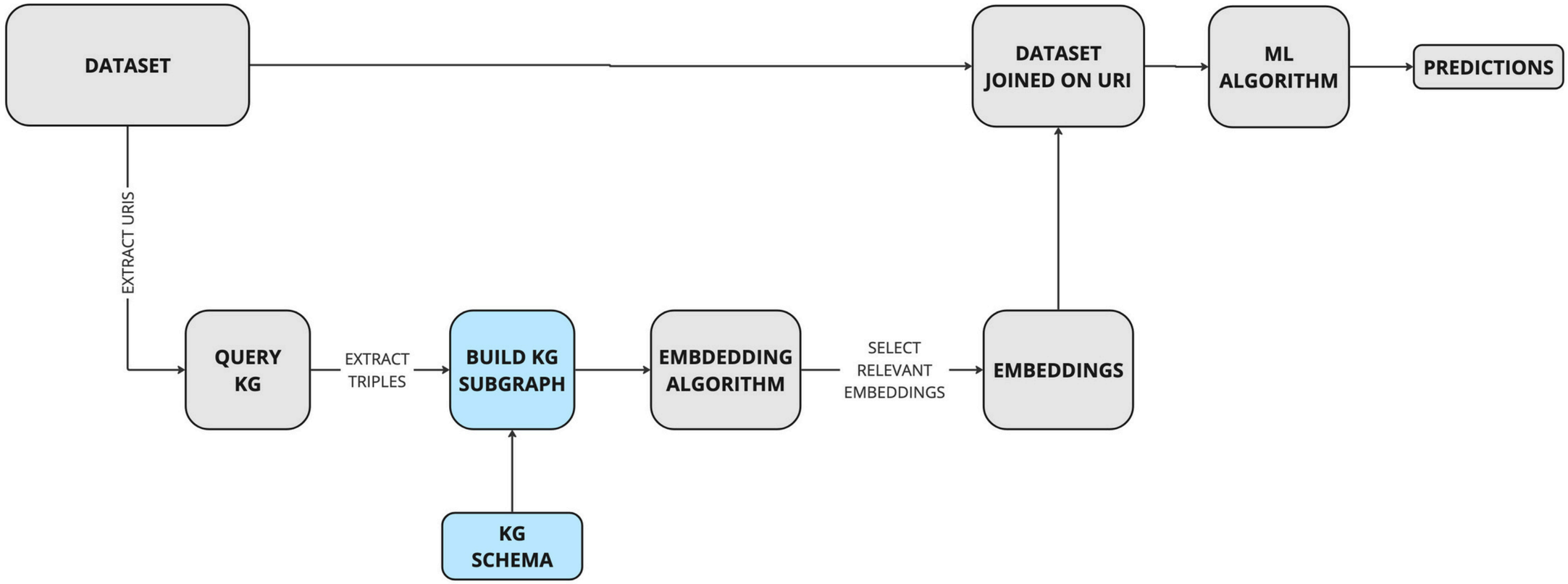
Methodology - Part 1

Datasets

- URI required to link instances with entities in KG
  - Difficult to find
  - Costly to build
- Ristoski et al. created repository with 22 datasets with DBpedia URIs
- Selection criteria:
  - Numeric features in orginal dataset
- 4 datasets selected:
  - Forbes
  - AAUP
  - Auto93
  - Auto MPG
- Highly correlated features removed
  - Emulate poor quality datasets
- Aggregation performed by URI
  - Multiple instances sharing same URI
  - Mean aggregation for numeric
  - Mode aggregation for categorical

Methodology - Part 1

KG Subgraphs



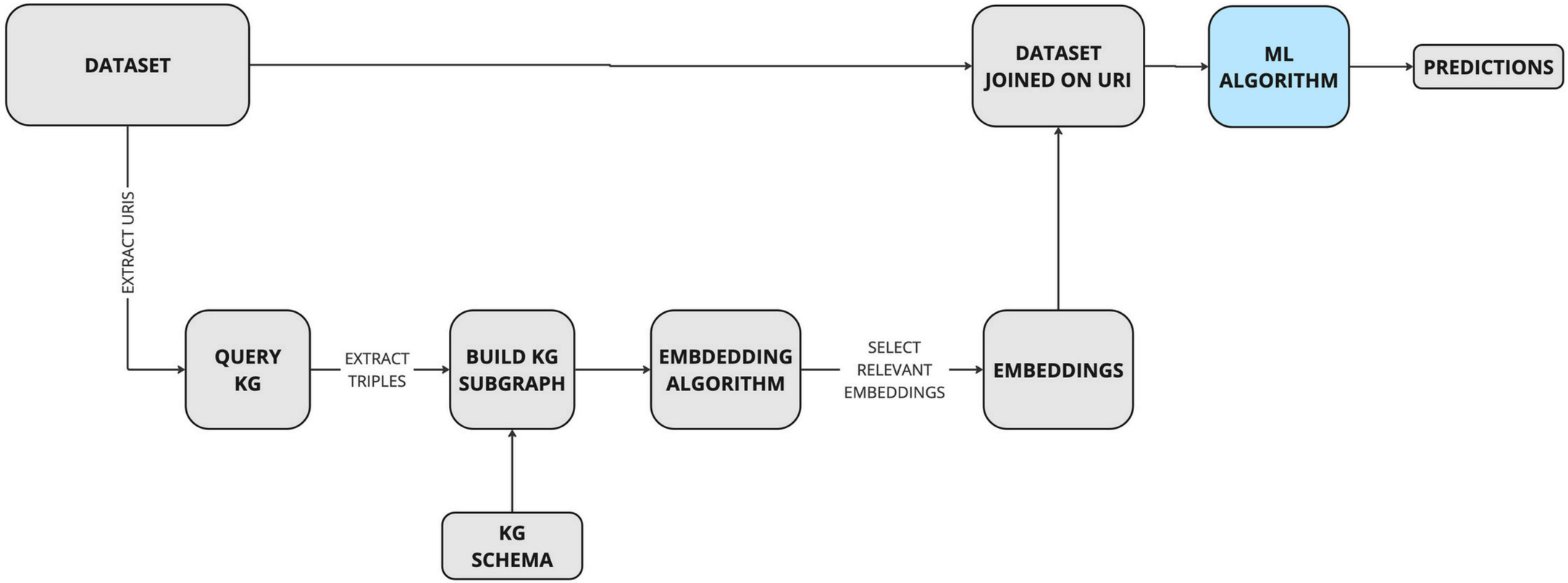
Methodology - Part 1

KG Subgraphs

- Subgraphs of DBpedia KG
- Subgraphs are created to improve quality of embeddings
  - Large KGs proabably have noise
  - Noise can confuse the embedding algorithms
- OWL syntax because it is more expressive and should produce higher quality embeddings
- Construction
  - Start with KG schema with no instances
  - Query KG to extract triples
    - Full query retrieves
      - Literals
      - Relationship properties
    - Simple query retrieves
      - Only rdf:type properties
  - Link triples from query to KG schema

Methodology - Part 1

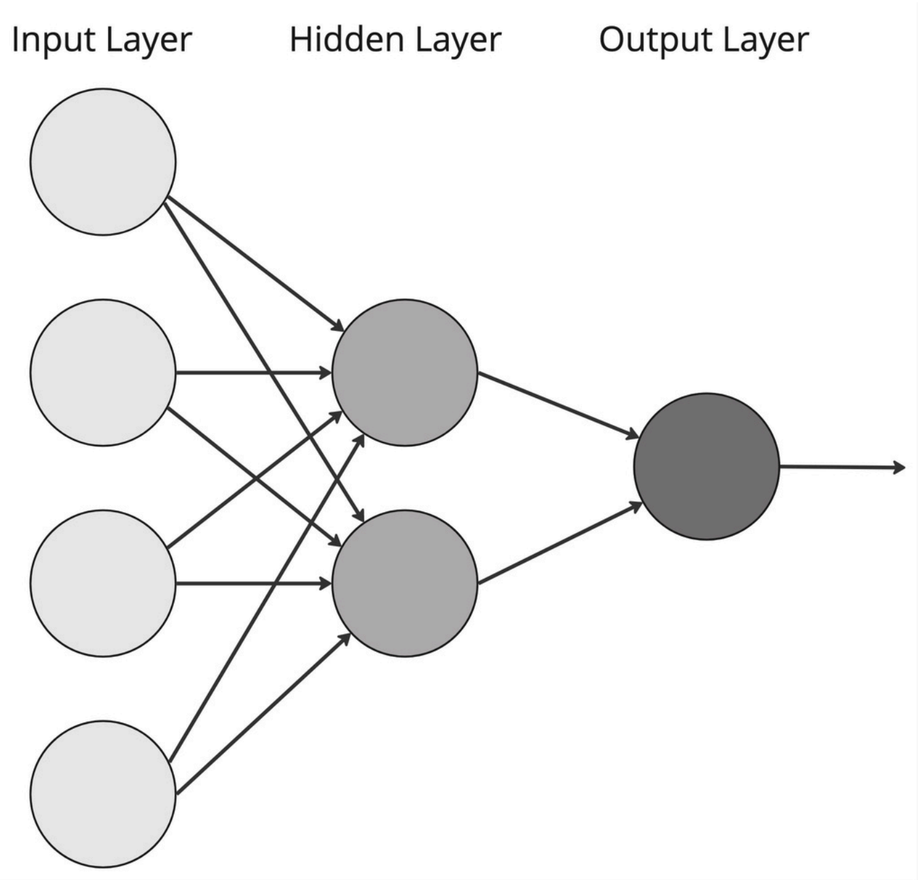
ML Algorithm



Methodology - Part 1

ML Algorithms

- Small Neural Networks
- Data augmentation on train set
- Experiments ran 10 times
- Mean of metric score calculated on validation set
- **Baseline Model**
  - Features from original datasets
  - Highly correlated features removed
  - No embeddings



Results - Part 1

Classification

Dataset	Algorithm	F1 Score	Baseline
Forbes	TransE OWL2Vec*	0.76 0.77	**0.79**
AAUP	TransE OWL2Vec*	0.76 0.78	**0.80**
Auto 93	TransE OWL2Vec*	0.69 0.69	**0.79**
Auto MPG	TransE **OWL2Vec***	0.67 **0.79**	0.76

Background

OWL2Vec* Extended

- After initial experiments showed that graphs embeddings did not beat the baseline model
  - We thoroughly investigated the OWL2Vec* algorithm
  - We found the Corpus did not inlcude any Literal values

- We decided to include Literals in the Corpus
  - Add additional features to embeddings
    - Income, number of employees, country...

  - Discretises Numeric Literals
    - word2vec struggles with numbers

    - Example
      - HSBC -- Income -- 1,000,000,000
      - HSBC -- Income -- very high

Results - Part 1

Classification

Dataset	Algorithm	F1 Score	Baseline
Forbes	TransE OWL2Vec* OWL2Vec* Ext	0.76 0.77 0.78	**0.79**
AAUP	TransE OWL2Vec* OWL2Vec* Ext	0.76 0.78 0.78	**0.80**
Auto 93	TransE OWL2Vec* OWL2Vec* Ext	0.69 0.69 0.77	**0.79**
Auto MPG	TransE OWL2Vec* **OWL2Vec* Ext**	0.67 0.79 **0.81**	0.76

Methodology - Part 2

Correlated Feature

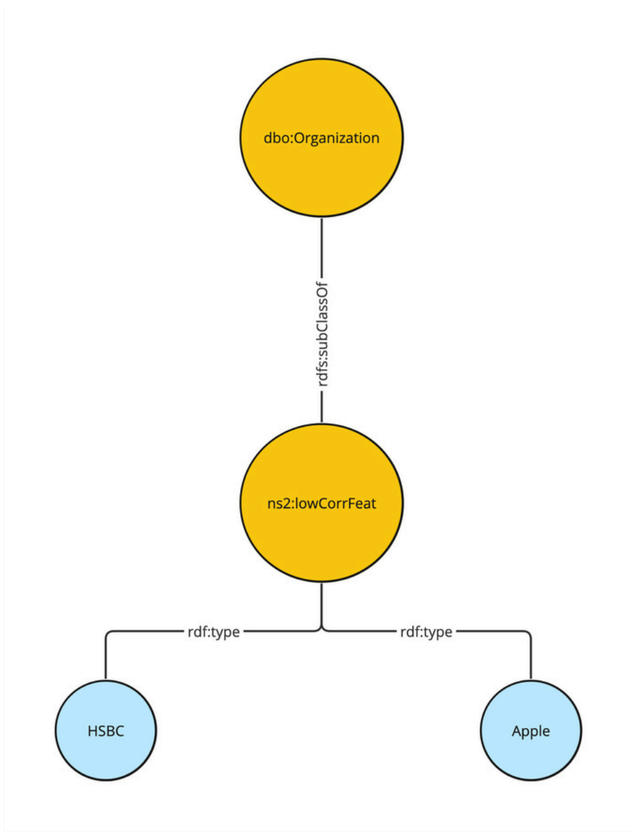
- A correlated feature to the Label to be predicted is added to the KGs

- The correlated feature should help understand
  - How the KG embedding algorithms work
  - What information is needed in KGs so embeddings can enhance ML models

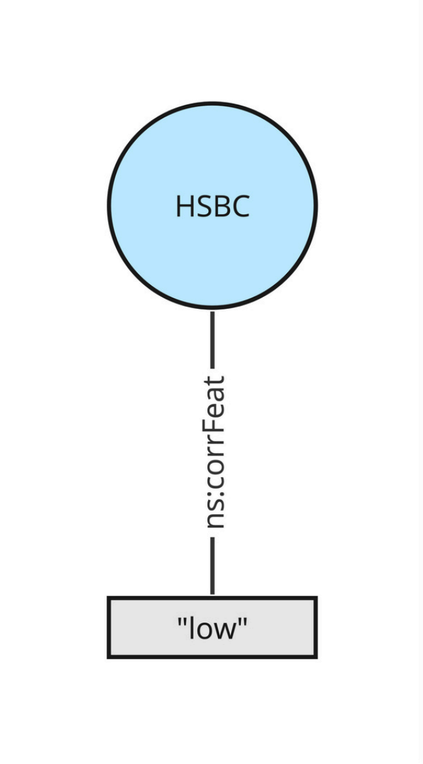
- Different degrees of correlation used
  - High
    - Correlated feature equal to Label being predicted in **90%** of instances
  - Medium
    - Correlated feature equal to Label being predicted in **50%** of instances
  - Low
    - Correlated feature equal to Label being predicted in **20%** of instances

Methodology - Part 2

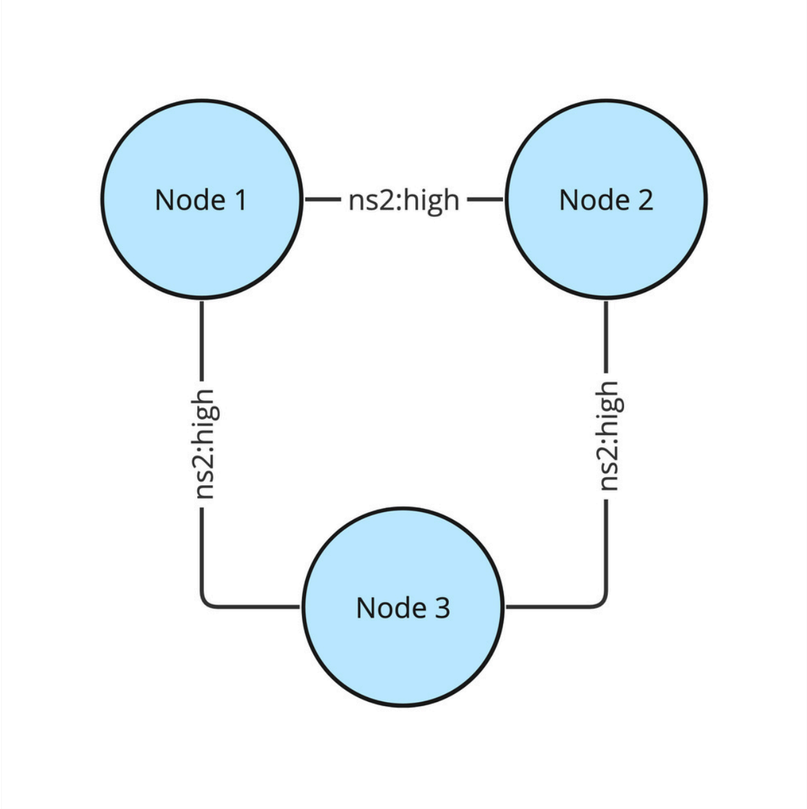
Class Modification



Attribute Modification



Property Modification



Results - Part 2

Classification - TransE with modified KGs

Dataset	Mod	F1 Score	Baseline
Forbes	Property Low Corr **Property High Corr** Attribute High Corr Class High Corr	0.80 **0.89** 0.75 0.77	0.79
AAUP	Property Low Corr **Property High Corr** Attribute High Corr Class High Corr	0.81 **0.91** 0.79 0.78	0.80
Auto 93	Property Low Corr **Property High Corr** Attribute High Corr Class High Corr	0.46 **0.81** 0.67 0.68	0.79
Auto MPG	Property Low Corr **Property High Corr** Attribute High Corr Class High Corr	0.75 **0.97** 0.74 0.80	0.76

Results - Part 2

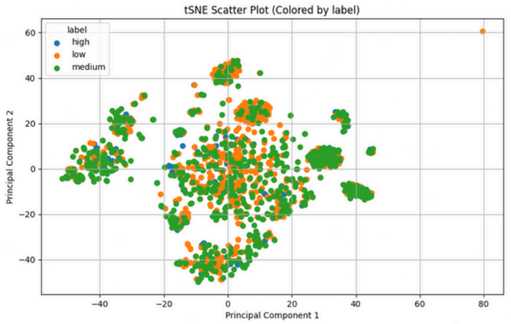
Classification - OWL2Vec* Ext with modified KGs

* Using simple query

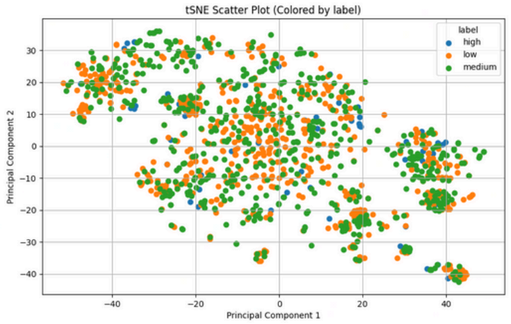
Dataset	Mod	F1 Score	Baseline
Forbes	Property High Corr Attribute High Corr* Class Low Corr **Class High Corr**	0.84 0.83 0.82 **0.92**	0.79
AAUP	Property High Corr Attribute High Corr* Class Low Corr **Class High Corr**	0.80 0.82 0.79 **0.88**	0.80
Auto 93	Property High Corr Attribute High Corr* Class Low Corr Class High Corr	0.52 0.71 0.40 0.52	**0.79**
Auto MPG	Property High Corr Attribute High Corr* Class Low Corr **Class High Corr**	0.75 0.70 0.76 **0.89**	0.76

Results - Part 2

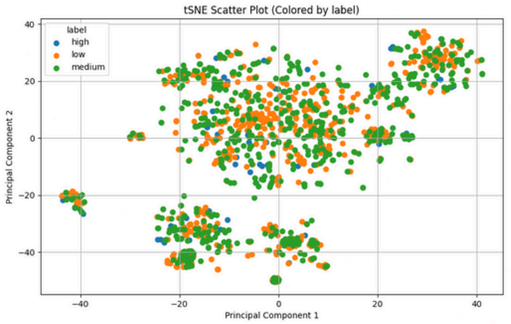
Clustering with modified KGs



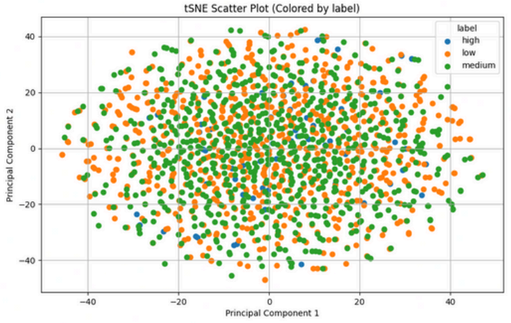
OWL2Vec* Class Mod



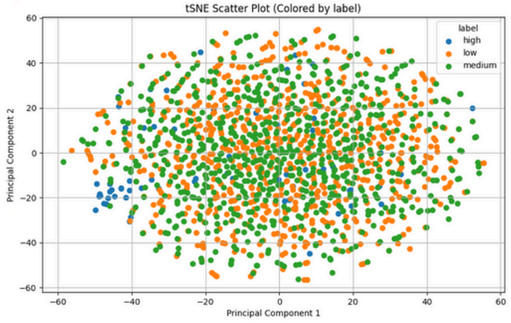
OWL2Vec* Attribute Mod



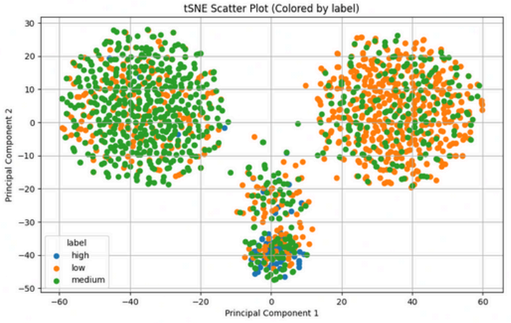
OWL2Vec* Property Mod



TransE Class Mod



TransE Attribute Mod



TransE Property Mod

Results - Part 2

Regression - TransE with Modified KGs

Dataset	Mod	RMSE	MAE	Baseline RMSE	Baseline MAE
Forbes	None	75.26	55.90	**57.23**	45.70
	Property Low Corr	83.03	59.73		
	Property High Corr	61.01	**44.60**		
	Attribute High Corr	70.23	52.40		
	Class High Corr	70.11	51.30		
Auto 93	None	12.29	9.55	29.68	24.84
	Property Low Corr	13.11	11.29		
	Property High Corr	**10.99**	**9.27**		
	Attribute High Corr	11.55	8.66		
	Class High Corr	12.53	10.42		
Auto MPG	None	13.55	9.97	34.79	28.47
	Property Low Corr	11.07	7.57		
	Property High Corr	11.08	9.06		
	Attribute High Corr	**9.33**	**7.03**		
	Class High Corr	13.30	10.54		

Results - Part 2

Regression - OWL2Vec* Ext with Modified KGs

* Using simple query

Dataset	Mod	RMSE	MAE	Baseline RMSE	Baseline MAE
Forbes	None Property High Corr Attribute High Corr* Class High Corr	62.19 63.70 60.36 58.21	46.33 45.47 44.79 **41.87**	**57.23**	45.70
Auto 93	None Property High Corr Attribute High Corr* Class High Corr	12.06 9.77 **8.29** 9.51	7.93 7.47 **6.33** 7.92	29.68	24.84
Auto MPG	None Property High Corr Attribute High Corr* Class High Corr	15.87 15.53 13.30 **12.93**	12.14 11.66 **9.93** 9.96	34.79	28.47

Conclusion

RO1: Evaluate if graph embeddings can enhance machine learning models in classification and regression tasks on poor quality datasets

Conclusion

RO1: Evaluate if graph embeddings can enhance machine learning models in classification and regression tasks on poor quality datasets

- Predict labels associated with datasets related to KG
- Results
  - Both embedding-enriched models failed to beat baseline

Conclusion

RO1: Evaluate if graph embeddings can enhance machine learning models in classification and regression tasks on poor quality datasets

- Predict labels associated with datasets related to KG
- Results
  - Both embedding-enriched models failed to beat baseline

RO2: Assess whether graph embedding algorithms can be improved for classification and regression tasks

Conclusion

RO1: Evaluate if graph embeddings can enhance machine learning models in classification and regression tasks on poor quality datasets

- Predict labels associated with datasets related to KG
- Results
  - Both embedding-enriched models failed to beat baseline

RO2: Assess whether graph embedding algorithms can be improved for classification and regression tasks

- Extended OWL2Vec* to include the values of Literals with discretisation
- Results
  - Extended OWL2Vec* performed better or equally than OWL2Vec*
  - To determine the better implementation more datasets and tasks need to evaluated

Conclusion

RO3: Analyse the behaviour of graph embedding algorithms depending on the quality and form of relevant data in the knowledge graph

Conclusion

RO3: Analyse the behaviour of graph embedding algorithms depending on the quality and form of relevant data in the knowledge graph

- Experiments included different degrees of correlated features in the KG in different forms

- Results
  - TransE better for properties

  - OWL2Vec* good classification and clustering performance
    - Class
    - Attribute

  - OWL2Vec* achieved better overall on all modifications than TransE

Conclusion

RO3: Analyse the behaviour of graph embedding algorithms depending on the quality and form of relevant data in the knowledge graph

- Experiments included different degrees of correlated features in the KG in different forms
- Results
  - TransE better for properties
  - OWL2Vec* good classification and clustering performance
    - Class
    - Attribute
  - OWL2Vec* achieved better overall on all modifications than TransE
- Drawbacks
  - Very specific design of KG
  - Unlikely case in real world
  - Feature selection over KG could be more beneficial

Conclusion

Future Work

- OWL2Vec* could be beneficial when KG schema is highly expressive
  - High class granularity

- OWL2Vec* & TransE embedding could be combined to exploit advantages from both
  - OWL2Vec* encodes semantics and entity attributes
  - TransE has a better understanding of relationships between entities

- Include LLMs in OWL2Vec*
  - LLMs have shown to outperform word2vec in multiple downstream tasks
  - Could produce superior embeddings
  - Some alternatives: BERT, DistilBERT, GPT

THANK YOU FOR LISTENING