Meta-X: Metadata Knowledge Discovery for Context Aware Business Intelligence

Hiba Khalid
Supervisor: Esteban Zimanyi
Co-Supervisor: Robert Wrembel
CPC Chair: Oscar Romero

Presentation Outline

- Introduction
- Literature Review (State of the Art)
- Problem Statement
- Methodology
- Progress Review
  - Project Planning
  - Research Dissemination Planning
- Conclusion

Introduction

Dataset Collection Explosion

Many organizations face the problem of integrating **multiple data sources** to attain business intelligence (BI)

Data Lakes & Freedom

The problem is typically addressed using the no **uniformity** and freedom provided by **data lakes.** Data lakes have no uniformity and that’s a problem.

Lack of Integration Principle

The BI teams do not have a generalized framework for dataset integration or processing

Data Fishing From Lakes

The next challenge is to fish or **retrieve the right dataset** for the corresponding problem, process or operation.

Problem Statement

The integration of independent data sources using metadata as knowledge base of meta-learning in order to obtain enhanced business intelligence.

Defined Goals

G1: Integrate independent data sources

G2: Construct knowledge base using metadata

G3: Minimize number of fetch requests in data lakes

G4: Increase accuracy of business analytics using deep learned metadata

Metadata Redefined-I

Metadata Classes/ Groups

The traditional metadata does not provide conformity to increasing data integration and analytic processes. In order to attain power over data itself scientists have concluded a redefined class of metadata based on context and domain of application operation.

Metadata Groups	Metadata
Basic	Size, formats, aliases, last modified time, access, control lists
Content Based	Schema, number of records, data fingerprints, key fields, frequent data tokens, similar datasets
Provenance	Reading jobs, writing jobs, downstream datasets, upstream datasets
User Supplied	Descriptions, annotations
Team and Project	Project description, owner, team name
Temporal	Change history

1. *Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang. GOODS: Organizing Google's Datasets. SIGMOD 2016.*
2. *Philip W. Lee. Metadata Representation and Management for Context Mediation. Composite Information Systems Laboratory (CISL) Sloan School of Management Massachusetts Institute of Technology Cambridge, MA 02142.*
3. *Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion.*

Metadata Redefined-II

Metadata at Use

Data Table

Table ID 102: Credit Risk Members

CRM-ID	CR-ID	Name	Age	Region
101030	343	John	43	Brussels
201829	345	Alice	23	Antwerp
123478	567	Karen	29	Dallas
245678	589	Siena	35	New York
111999	456	Ariana	67	Rome

Descriptive Metadata

Created By: Lee
Date of Creation: 13 JUNE 2012
Version: 2.0
Subject: Member listing

Used for

-For discovery of data
-For displaying data such as transactional data (OLTP)
-For interoperability

Structural Metadata

Table Index: 102
P-key: Yes
F-Key: Yes
Interdependencies: 104, 106, 109

Used for

-Navigation and presentation
-Internal structure + relationship description
-Foreign Key CR-ID was included in table 102)

Administrative Metadata

Data Type: Integer, real, Text
Access Rights: Admin Only
Data Migration: Yes
Last Migration: 10 January 2012

Used for

-Short-term and long-term management processing
-Technical data: creation, quality control, rights management, access control

1. Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang. GOODS: Organizing Google's Datasets. SIGMOD 2016.
2. Philip W. Lee. Metadata Representation and Management for Context Mediation. Composite Information Systems Laboratory (CISL) Sloan School of Management Massachusetts Institute of Technology Cambridge, MA 02142.
3. Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion.

Literature Review

The research is based on aggregated literature in field of heterogeneous data sources, data lakes, integration of data sources, metadata extraction, enrichment, profiling.

Paper Title	Details	Drawback
GOODS: Organizing Google's Datasets	Metadata classification, enrichment, logging, cataloging, Provenance enrichment, enhancement of metadata	Not designed for common business analytics in companies. Costly Requires domain altering-Not generic
Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion	Automatic Knowledge base construction. Structured data repository. Probabilistic modelling.	Only considers structured representations
Evaluation of Metadata Representations in RDF stores	Model for storing metadata along RDF data for big data at implementation and conceptual level.	Does not handle the cost of querying disintegrated data.
Web Tables: Exploring the Power of Tables on the Web	The identification and extraction of labelled schema as structured datasets. Analyzes and answers the query of traversing structured web tables in search engines. Schema auto complete. Attribute co-relations etc.	Only works for web tables and structured data. Auto-Complete schemas but no correlation between disintegrated schema representations.

Research Scenario: Credit Risk Modelling

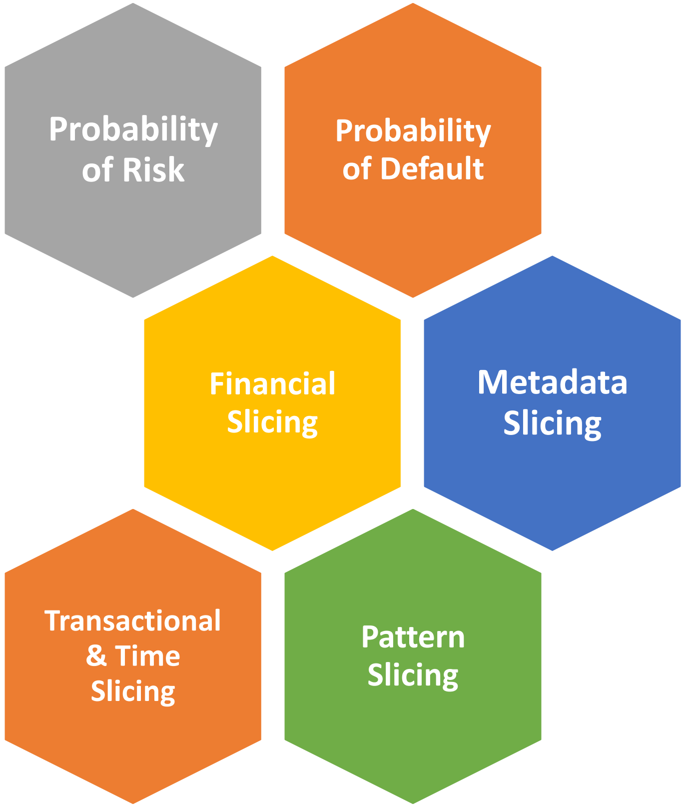
Information

- Stock purchase: 2013; Name: ‘John Smith’
- Tennis Match Ticket: 2014 ; Purchase Type: Online; Payment Status: Paid
- Online Shopping: 2015; Purchase Mode: Credit: Online; Payment Status: Unpaid; Name: John

Possible Queries

Q1: Will John smith pay her credit card bill by the end of 2015?

Q2: Is John smith and John the same entities?



Research Explication

1

Metadata Collection

To gather metadata of existing data, extract it and clean it for further processing

2

Deep Metadata Layer

To run the metadata through deep learning network and create a ‘Deep self actualized metadata layer’

3

Process Answering

To use this MD library for answering business and customer queries

4

Reinforcement

To Reinforce the Membrane using Feedback Loop in the form of rewards and failures

Research Methodology

Process Details: Types of metadata formats for collection

```
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<Description about="http://www.w3.org/Press/99Folio.pdf">
<dc:title>The W3C Folio 1999</dc:title>
<dc:creator>W3C Communications Team</dc:creator>
<dc:date>1999-03-10</dc:date>
<dc:subject>Web development, World Wide Web
Consortium, Interoperability of the Web</dc:subject>
</Description>
</RDF>
```

Example Metadata to be extracted includes: Title, author, date, creator, about, subject etc.

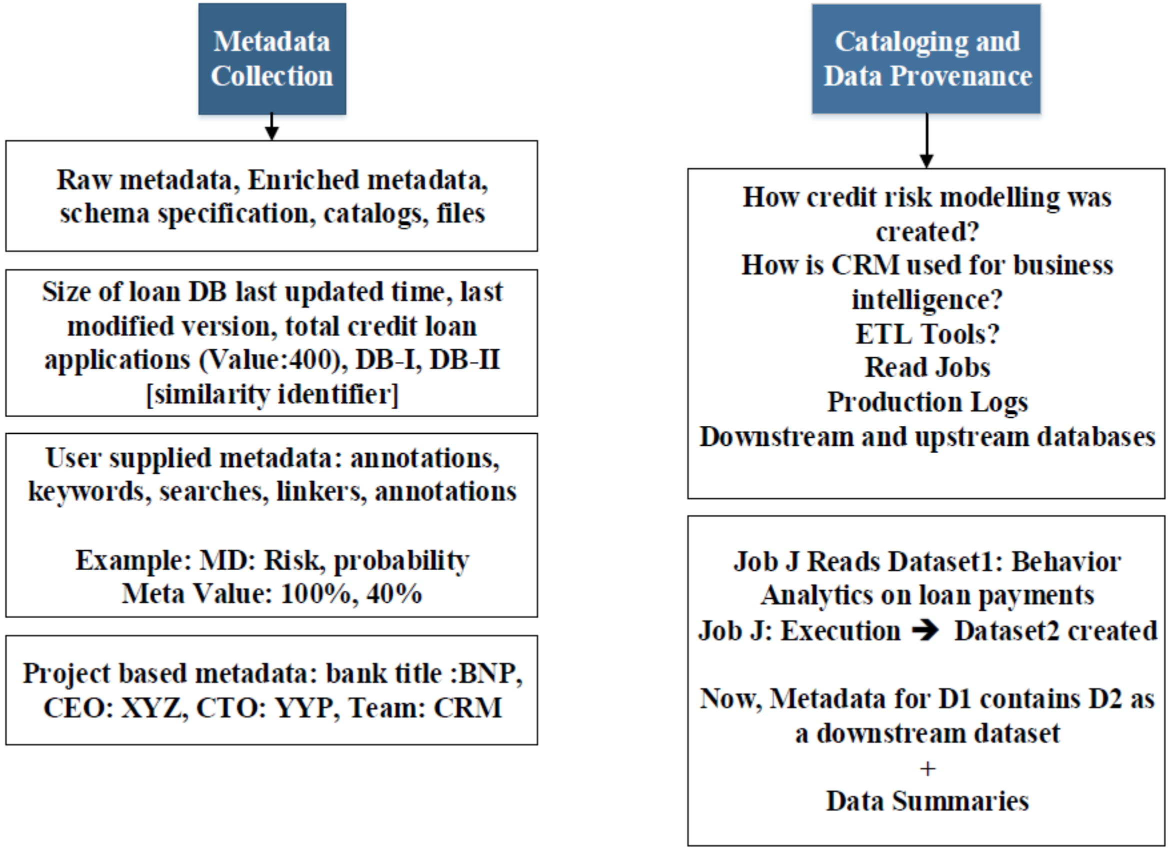
+

RDF Schema Specifications

1: The metadata can be collected in different forms such as xml, RDF representation, even log files, schema etc.

Research Methodology

Process Details: Metadata Collection & Enhancement

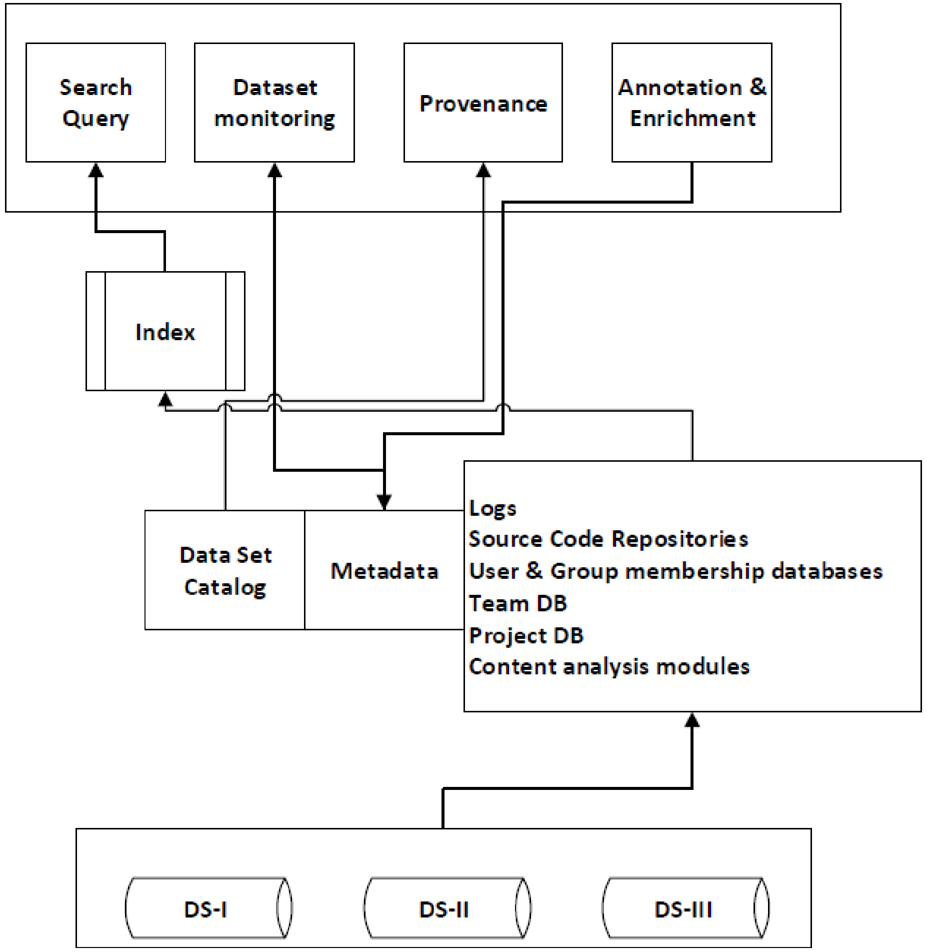


2: The collection leads into metadata classes and categories for assignment

3: The third step focuses on enhancement of collected and categorized metadata

Metadata Visualization Process

Metadata			
Size	Provenance	Schema	…
100	Written by: Task A	Credit.nlu.BNP	
400	Read by: Reader A	Behaviour.anl.Schema	
800	Corrected by: Task A	…..	
600	Updated by: Task B	……	
300	Trained: Task D	……	



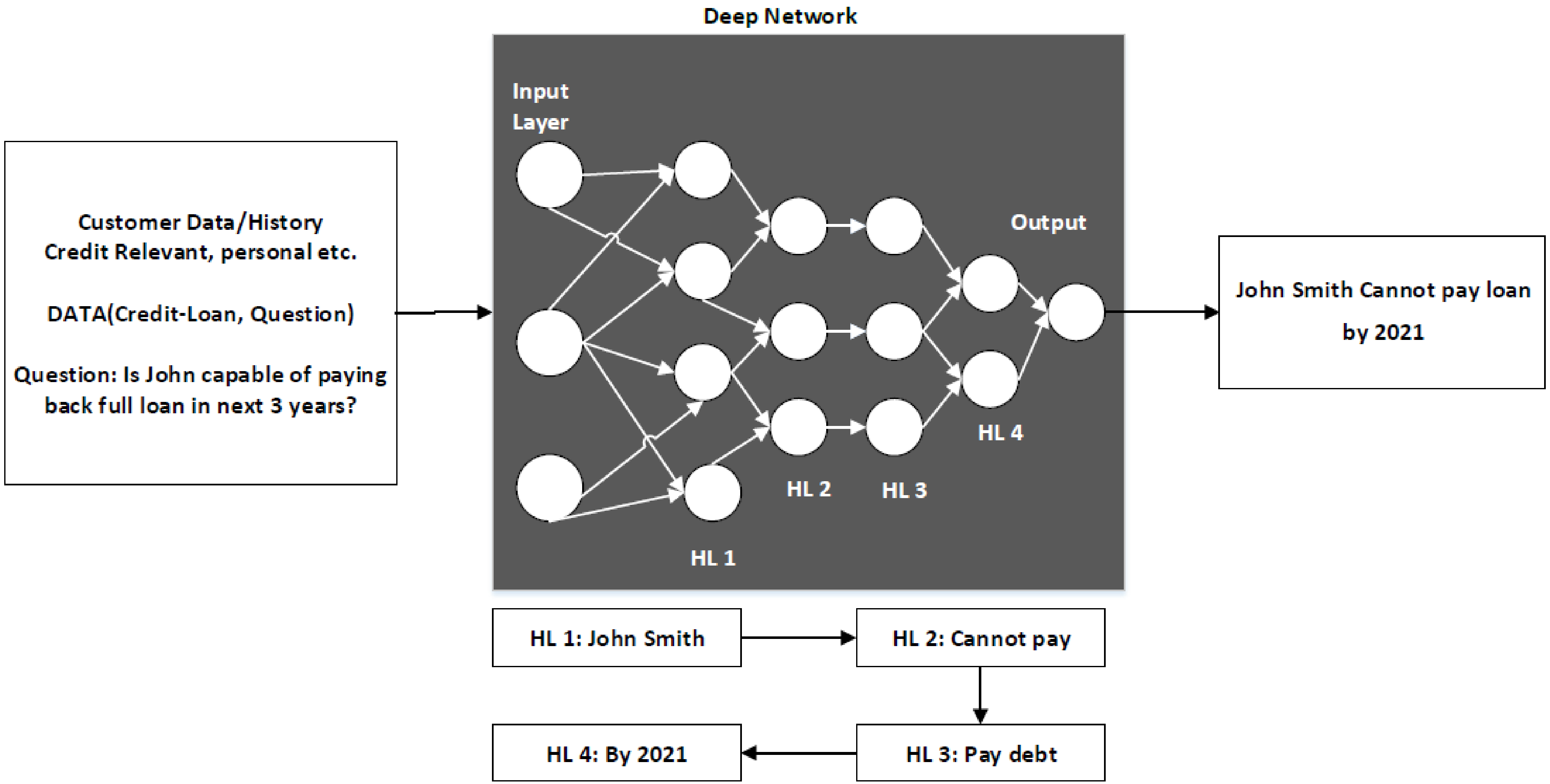
Metadata Collection Layer

Deep Learning

- ANN’s but better !
- Multiple Hidden Layers
- Supervised, Unsupervised or Semi-supervised
- Learning Data Representations

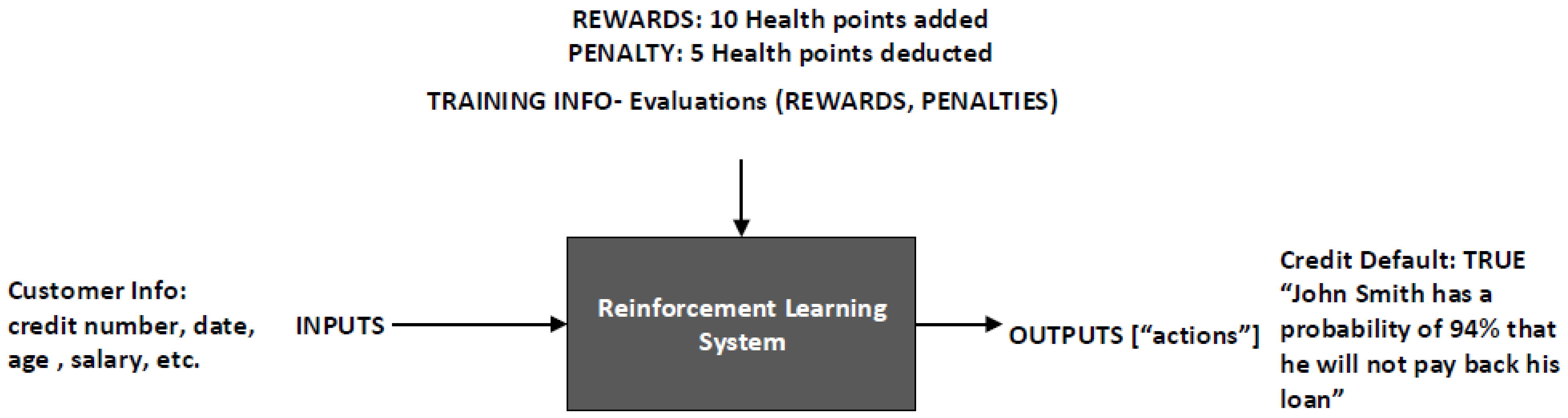
- Automatic feature detection
- Hierarchical feature learning
- Multiple level representation

Learning Visualization: Learning to Predict Customer Behavior



Learning Visualization: Feedback & Reinforcement

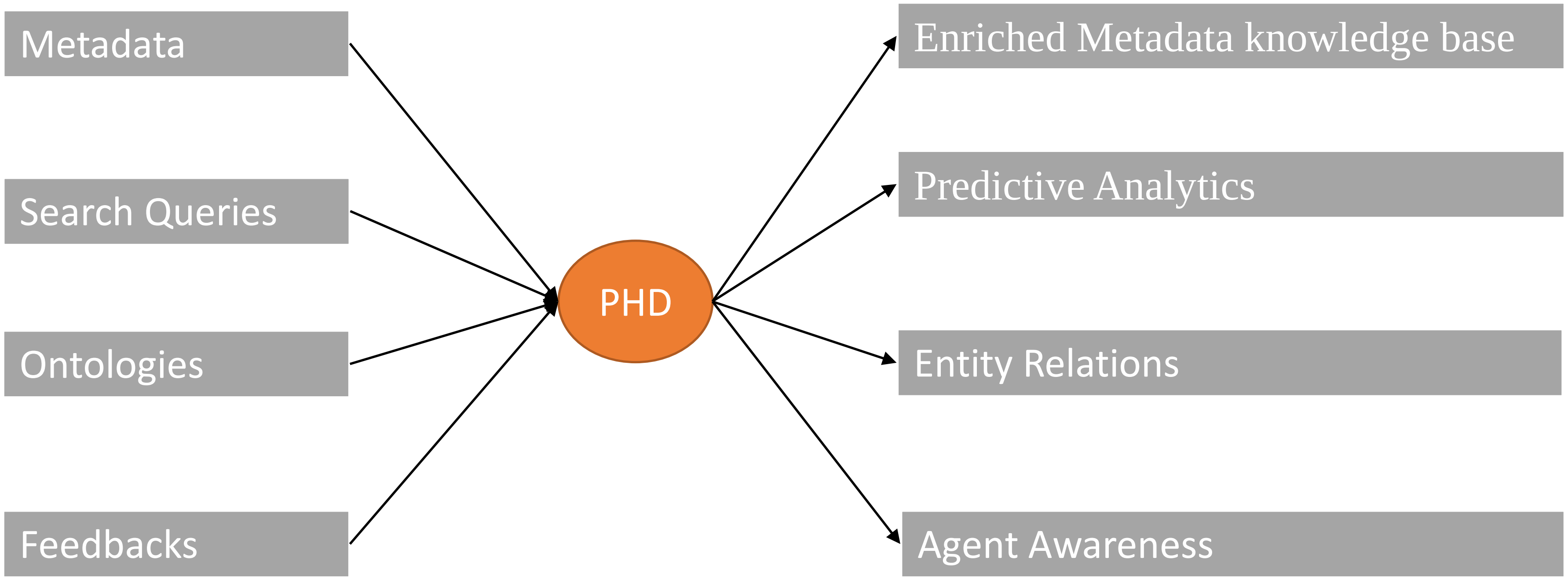
RL offers the algorithms to learn from its own actions of classifications, predictions and decisions.



Research Inputs to Outputs Mapping

Inputs:

Outputs



Research Limitations & Constraints

Restrictions & Constraints

- Metadata Enrichment
- Profiling
- Limited independent data sources (4)

Limitations of Project

- Does not handle ontology alignment
- Does not handle ontology evolution
- Does not contain aggregated datasets

Summary & Conclusion

Tasks	Description
Milestones Achieved	Literature Review Client Side Query Analysis (Text Based) Research Article-I based on scientific study variables and associated dependencies. Research Article-II is based on the construction of data composites for metadata discovery
Milestones Planned	Metadata Extraction Metadata Enrichment Metadata Classification & Learning

ECTS Planned

Activity	Place	ECTS	G/I/PC	Status
January 2017 to December 2017				
Summer School	Brussels	2	PC	In Progress
French Language Course	Brussels	1.5	G	Planned
Research Publishing	Brussels	2	I	Planned
OPEN HPI Semantic Web Course	Online	1	PC	In Progress
January to December 2018				
Software Development Studio	Poznan	3	PC	Planned
Technical & Scientific Writing	Poznan	3	G	Planned
Data Mining and Analysis	Poznan	5	PC	Planned
Polish Language Course	Poznan	1.5	G	Planned
January 2019 to January 2020				
Research Seminar	Brussels	1	PC	Planned
Internship	Brussels	2	PC	Planned

PHD Timeline

Sessions	Details	Status
Spring 2017 [Jan-July]	Research methodology Construction (Done) Literature Review (Done) Submission of Research Paper-I in ADBIS (Rejected). Corrected & Ready for Re-submission. Submission of DPP (Done) **Planned Publications:** **a. Deep Metadata: A Scientific Study into the needs of Intelligent Business Semantics. (Completed, Reviewed, ready for Submission)** b. **Deep metadata Knowledge graphs for Semantic Web & Business Intelligence.**	Completed
Fall 2017-2018[Sept-Feb]	Moving to host university (PUT) Submission of Conference Paper-II (In Progress) **Planned Publications:** **a. Understanding the metadata modelling in semantic business;** **b. Conceptual meta-data to automated metadata.**	Planned
Spring 2018 [Mar-July]	Submission of TPR **Submission of Journal-I: SEMX: A learning model of metadata for high speed semantic knowledge**	Planned
Fall 2018-2019[Jan-July]	Proof of concept-I **Submission of Journal II: Learning as a reinforced activity: Meta Composites in Business Intelligence**	Planned
Spring 2019 [Sept-Feb]	Thesis Evaluation cycle	Planned
Fall 2019-2020 [Jan-July]	Thesis write up and Defense	Planned

References

- Punit Pandey, Deepshika Pandey, and Dr. Shirshir Kumar, “Reinforcement Learning by Comparing Immediate Reward,” (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 5, August 2010.
- Volodymyr Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou Daan Wierstra Martin Riedmiller, “Playing Atari with Deep Reinforcement Learning”, DeepMind Technologies.
- Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In Proceedings of the 12th International Conference on Machine Learning (ICML 1995), pages30–37.
- M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. PVLDB, 1(1):538–549, 2008.
- M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. SIGMOD Rec., 34(4):27–33, Dec. 2005.
- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279, 2013.
- Marc G Bellemare, Joel Veness, and Michael Bowling. Investigating contingency awareness using atari 2600 games. In AAAI, 2012.
- Marc G. Bellemare, Joel Veness, and Michael Bowling. Bayesian learning of recursively factored environments. In Proceedings of the Thirtieth International Conference on Machine Learning (ICML 2013), pages 1211–1219, 2013
- Nicolas Heess, David Silver, and Yee Whye Teh. Actor-critic reinforcement learning with energy-based policies. In European Workshop on Reinforcement Learning, page 43, 2012.
- I. Konstantinou, E. Angelou, D. Tsoumakos, and N. Koziris. Distributed indexing of web scale datasets for the cloud. In Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud, MDAC ’10, pages 1:1–1:6, 2010.
- Vanden Berghen Frank, Q-Learning, IRIDIA, Universit Libre de Bruxelles.
- Acosta, M, Simperl, E, Flöck, F, Vidal, M.-E, and Studer, R. (2015). RDF-Hunter: Automatically Crowdsourcing the Execution of Queries Against RDF Data Sets. Arxiv preprint arXiv:1503.02911 (2015).