Interreg
ITALIA-SLOVENIJA
UNIONE EUROPEA
EVROPSKA UNIJA
TRAIN
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

Improving the reproducibility of experiments and reusability of research outputs in complex data analysis

Panče Panov
Jožef Stefan Institute, Ljubljana, Slovenia and
Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Institut
"Jožef Stefan"
Ljubljana, Slovenija

arrs
SLOVENIAN RESEARCH AGENCY

Outline

- Reproducibility issues in science
- Reusability of research outputs
- Adding semantics to research outputs
- Reproducibility and reusability in complex data analytics
- IMPERATRIX project
- Current work on reproducibility and reusability in the context of automated modeling dynamical systems

Reproducibility in Science

- Advances in science are heavily based on the premise of the concept of a trusted discovery
  - provided that the preformed research is done correctly, and
  - the research outputs can be reproducible by other scientists
- Reproducing of scientific experiment
  - one of the most important approaches that scientists use to gain confidence in their conclusions [Peng, 2011]
- Providing means for reproducible research
  - Defining minimal information about a performed experiment and the produced research outputs
  - Creating, storing and curating meta data about performed experiments and the research outputs
  - Enable search on collections on performed experiments and research outputs

Peng RD (2011). Reproducible research in computational science. *Science 334*: 1226–1227.

Reproducibility in Artificial Intelligence (AI)

- In computer science large number of experiments are fully conducted on computers
  - Makes the experiments more straightforward to document [Braun and Ong, 2014]
  - Most AI and machine learning research falls under this category
- AI must rely on reproducible experiments to validate results
- Reproducibility in AI is not easily accomplished [Gundersen et al., 2018]
  - AI publications fall short of providing enough documentation to facilitate reproducibility
  - Best practices for documenting research needs to be developed: Data, Source code, AI methods, Experiments

Braun, M. L; and Ong, C.S. 2014. Open science in machine learning. In Implementing Reproducible Research, page 343. CRC Press

Gundersen et al. (2018)On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI magazine 39(3)*

Reusable research outputs and FAIR

- Reusability of research outputs, such as data and produced models
  - very important aspect in performing scientific work
  - decreases duplication of efforts and ineffective use of resources
- Group of life scientists have proposed a set of guiding FAIR Data Principles [Wilkinson, 2016]
  - Findable, Accessible, Interoperable and Reusable (FAIR).
- Reusability of research data as one of the priorities of EC
  - endorsement by the Directorate General for Research and Innovation of the EC
  - rapid endorsement G20 forum participants
- The main point of FAIR is to ensure that research objects are reusable and thus becoming more valuable.

Wilkinson MD et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*: 160018

European Commission (Press release) - G20 Leaders' Communique Hangzhou Summit. URL: http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm

FAIR data principles

- **Findable:**
  - Easy to identify and find for both humans and computers
  - Provide metadata that facilitate searching for specific research objects
- **Accessible:**
  - stored so that they can easily be accessed and/or downloaded
  - well-defined access conditions at the level of metadata and level of the actual data
- **Interoperable:**
  - Ready to be combined with other research objects by humans or computers
  - No ambiguities in the meanings of terms and values
- **Reusable:**
  - Ready to be used for future research and to be further processed using computational methods
  - Information about how the objects were obtained and processed and an appropriate license (provenance)

Adding semantics to research outputs: Ontologies

- The research outputs that wish to fulfil the FAIR Data principles must be represented with a wide accepted machine-readable framework [Mons, 2017]
- Provide additional meaning or "semantics" to the data and knowledge contained in the research outputs.
- Allow computers to automatically interpret the represented information
  - Annotating this information with ontologies (generating meta-data) and creating a knowledge base
  - Ontologies are expressed in well-defined logical formalism (e.g., description logic)
  - The logical approach enables reasoners to perform automatic inference on top of the knowledge base

Mons B et al. (2017) Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf Serv Use.* 37: 49–56.

Semantic web technologies

- Semantic web technologies and formal ontologies
  - Currently a popular solution to data and knowledge sharing
  - Are in line also with the requirements of FAIR
- Set of technologies to support the knowledge-sharing process
  - Resource Description Framework (RDF) format: a simple data model that handles the representation of given facts
  - RDF Schema (RDFS) provides the basic constructs for expressing knowledge
  - Ontology Web Language (OWL) that provides semantic interpretation to RDF facts.
- Use of reasoners: inference of additional RDF statements from those given explicitly
- Querying: SPARQL query language
  - enables us to retrieve explicit facts from the RDF datasets
  - the RDF datasets are usually stored in so called RDF stores
  - SPARQL cannot perform any inference tasks by itself: it must interact with a reasoning component

Complex data analysis

- Complex data analysis methods
  - Originating from machine learning (ML) and data mining (DM)
  - Increasingly being used in applications from various domains of science, such as medicine, biology and ecology
- Examples of tasks in medicine, solved by using data analytics methods, include biomarker discovery, discovery of biological disease signatures, etc.
- All domain tasks can be mapped to a ML/DM task and adequate methods can be used to solve the task
  - Example: The task of biomarker discovery can be mapped to the task of feature ranking

Example task: Structured output prediction

- primitive output prediction task (classification and regression)
  - goal is to predict a single primitive value (numeric or discrete)
- structured output prediction task (SOP)
  - goal is to predict a structured object
- There is a significant body of research on specific SOP tasks that differ in terms of
  - the type of the structured output (tuple, sequence, set, tree, graph),
  - data availability (complete, partial), and
  - how the data examples arrive (offline or online setting)
- Examples of specific SOP tasks include: multi-label classification, hierarchical multi-label classification, multi-target prediction, time-series prediction and others.

How can we improve the reproducibility and reusability in complex data analysis? (1)

- Formally represent all the entities involved in the process of data analysis
  - Entites: datasets, algorithms, possible experimental scenarios, materialized experiments, experimental results, produced models etc.
- Efforts to identify and formally represent the knowledge about the process of data analysis and its core entities
  - State-of-the-art ontologies and vocabularies: OntoDM-core, OntoDT, OntoDM-KDD, DMOP, Exposè, KDDOnto, MEX, ML-schema and others
- Annotate all the participants in the data analysis process
- Store the annotations as digital objects in a database like structure
- Provide semantic querying on top of the created database

How can we improve the reproducibility and reusability in complex data analysis? (2)

- Experiment databases
  - "databases specifically designed to collect and organize all the details of large numbers of past experiments, performed by many different researchers, and make them immediately available to everyone" [Vanschoren, 2012].
- The experiment database stores all the details of the experimental procedures
  - It moving towards the scientific goal of a truly reproducible research.
  - Example: OpenML is web-based platform based on relational database that stores information about datasets, tasks, experimental settings and the experimental results
  - The database can be accessed with REST API or by writing direct queries using SQL query language.
- The experiments performed by running algorithms on sets of data represent a main source of objective information on how the algorithms behave [Vanschoren, 2014]
  - This is the base of meta-learning studies on different tasks.
- Experiment databases allows for previous experiments to be readily reused in further studies

Vanschoren J,et al. (2012) Experiment databases. *Machine Learning 87*: 127–158.

Vanschoren J, et al. (2014) OpenML. *ACM SIGKDD Explorations Newsletter 15*: 49–60

How can we improve the reproducibility and reusability in complex data analysis? (3)

- Inductive databases [Džeroski et al, 2010] is also highly relevant in the context of reusable research
  - it deals with storing and querying models and not only data
- Inductive databases regard patterns and models, generated from data, as "first class citizens", having the same status as the data
- This concept was first proposed by Imielinski and Mannila [1996]
  - with a viewpoint that a dataset contains patterns, just like it contains data.
  - That is, just like we can query a database for its content (database query), we should be able to query it for patterns and models (inductive query), and store them in the database.

Džeroski, et al. (2010) *Inductive Databases and Constraint-Based Data Mining*. Springer; 2010.

Imielinski and Mannila (1996) A database perspective on knowledge discovery. *Communications of ACM 39*: 58–64.

Storing experiments, research outputs and their annotations

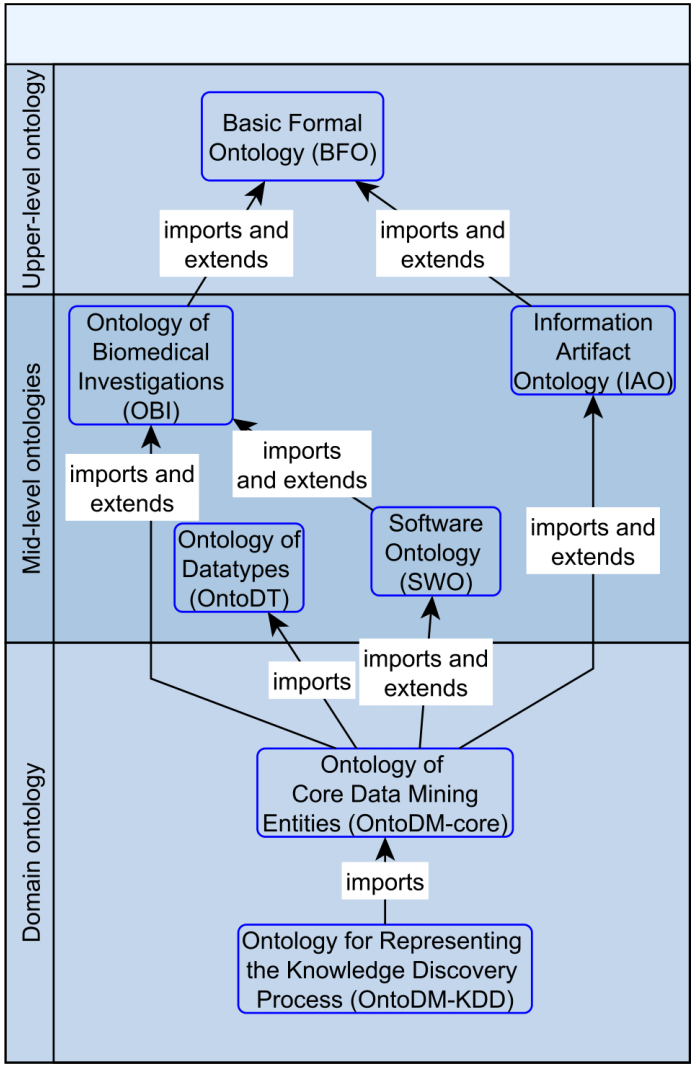
- Entities that appear in the process of data analysis (such as datasets, models, experimental settings) can have different nature
- Different database systems might be appropriate for different types of entities
  - Example: the same database system might not be appropriate for storing experimental settings/results and for storing predictive models.
  - Predictive models have dual nature, they can be seen as data objects and as functions
- A large number of database systems have been used as backend of existing RDF management systems
  - relational database management systems (e.g., MySQL, Oracle, etc)
  - NoSQL systems (e.g., MongoDB, Neo4J, Apache Jena TDB, RDF4J)
- Task of identifying the adequate database system to use (e.g., Relational or NoSQL)

Relevant ontological representations for machine learning and data mining

- Variety of ontologies, vocabularies and schemes
- OntoDM-core, OntoDT, OntoDM-KDD: a set of modular ontologies in the context of a general framework for data mining [Džeroski, 2006]
- Exposè [Vanschoren and Soldatova, 2010] and MEX vocabulary [Esteves et al., 2015]: representations of machine learning experiments
  - Exposè reuses parts of EXPO - ontology for representing general experiments [Soldatova and King, 2006]
- DMOP [Keet et al., 2015]: meta-mining of data mining workflows
- ML-schema [Publio et al., 2019]: an initiative to harmonize the developed ontologies and produce a lightweight schema, to be used for semantic annotation of ML experiments

Machine Learning Schema Community Group. URL: https://www.w3.org/community/ml-schema/

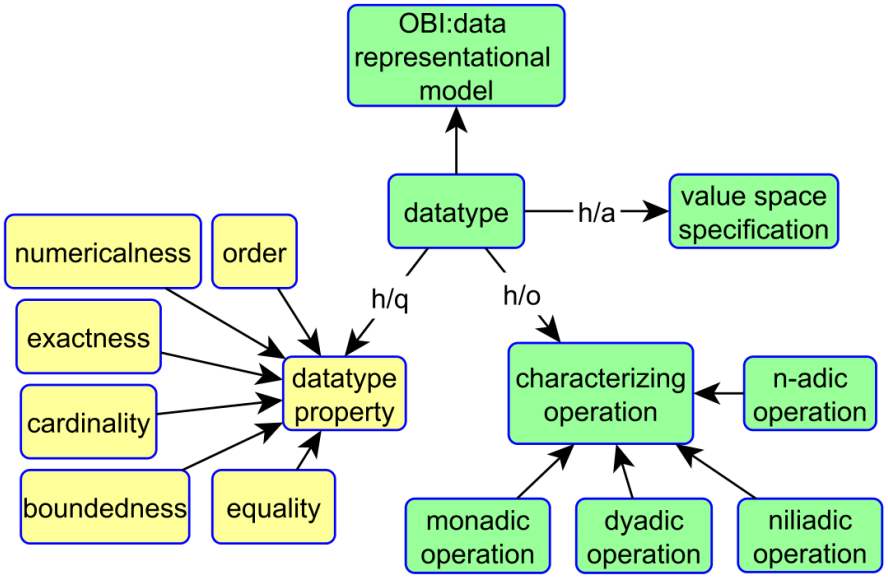
OntoDM:
a set of modular ontologies for data mining



- Built using best practices from biomedical domains [Smith et al., 2007]
- Complementary to and integrated with state-of-the-art ontologies for representing scientific knowledge
  - Interoperability with other resources
  - Allows for cross-domain reasoning
- Three modules: OntoDT, OntoDM-core, OntoDM-KDD

Smith et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology 25(11): 1251*

Generic ontology of datatypes - OntoDT



- Mid-level ontology
- Based on ISO/IEC 11404
- Can support a wide range of applications
- Datatypes are very important in data mining
  - Characterize the types of data contained in a dataset
  - Applicability of a data mining task on data from a given datatype
  - Applicability of a data mining algorithm on a dataset

Panov et al. (2016) "Generic ontology of datatypes",*Information sciences* 329:900-920

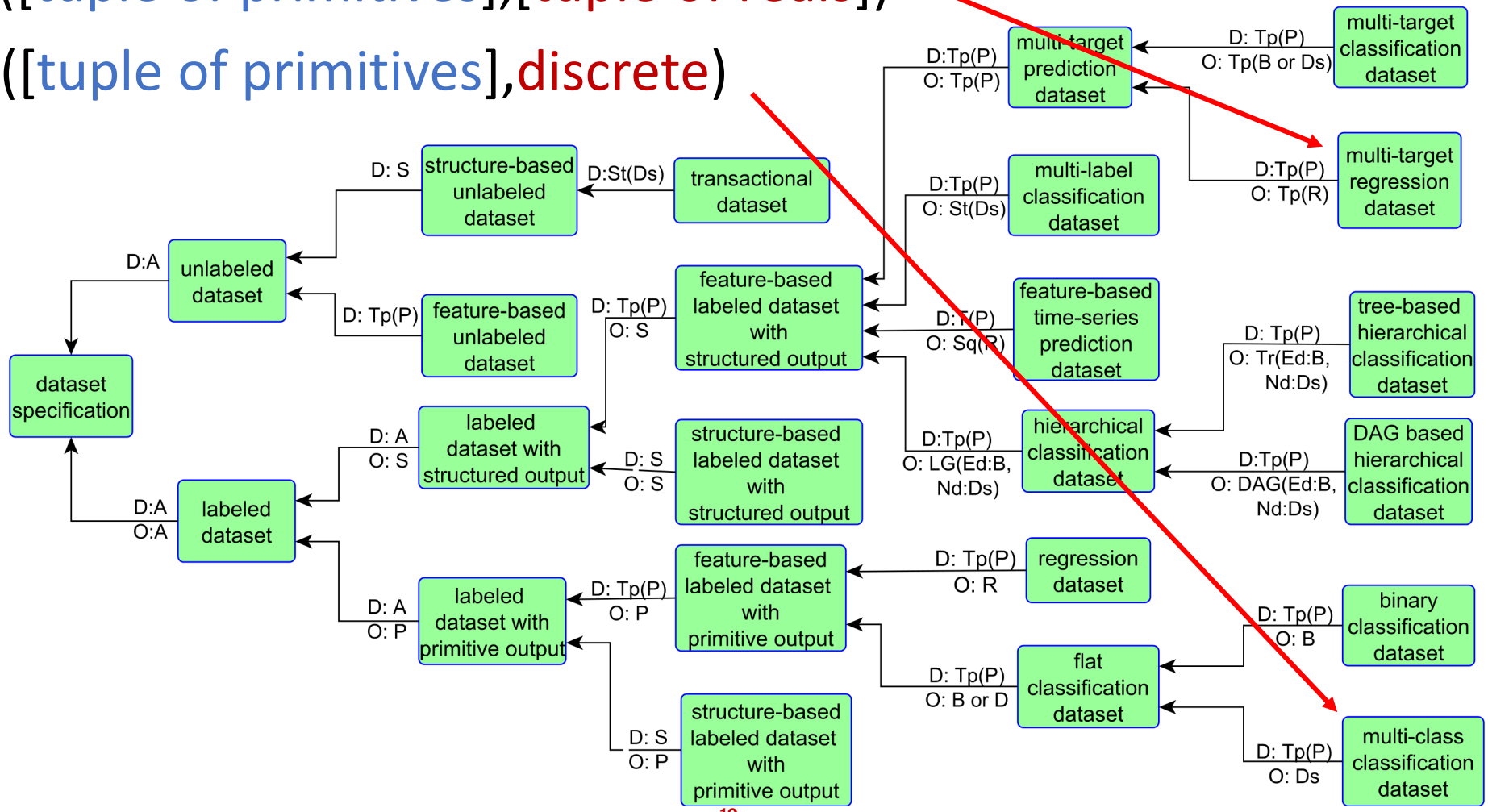
Examples of different datatypes in data mining

- Unlabeled data - only descriptive part
  - Feature-based data example ([tuple of primitives])
  - Transactional data example ({set of discrete})
- Labeled data - both descriptive and output parts
  - Feature-based data example with primitive output
    - ([tuple of primitives],real)
    - ([tuple of primitives],boolean)
    - ([tuple of primitives],discrete)
  - Feature-based data example with structured output
    - ([tuple of primitives],[tuple of reals])
    - ([tuple of primitives],[tuple of discrete])
    - ([tuple of primitives],{set of discrete})
    - ([tuple of primitives],(sequence of real))
    - ([tuple of primitives],tree with boolean edges and discrete nodes)
    - ([tuple of primitives],DAG with boolean edges and discrete nodes)

Example of dataset taxonomy constructed using OntoDT

([tuple of primitives],[tuple of reals])

([tuple of primitives],discrete)



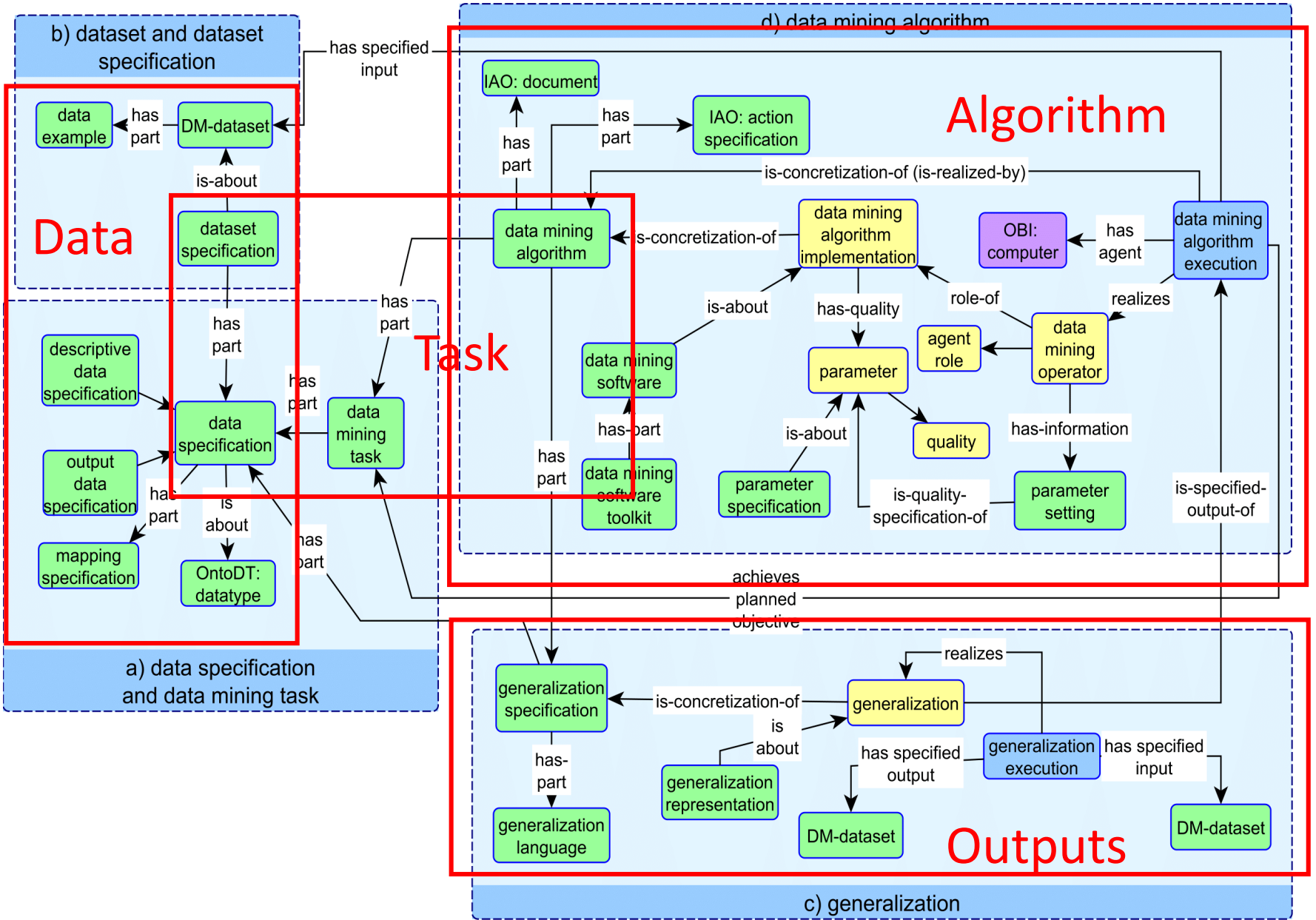
Ontology of core data mining entities - OntoDM-core

- Based on a proposal for a general framework for data mining [Džeroski, 2006]
- OntoDM-core describes the most essential data mining entities [Panov et al., 2014]
  - Data specification
  - Dataset
  - Data mining task
  - Data mining algorithm
  - Generalizations (patterns, models)
- Taxonomies of datasets, data mining tasks, generalizations, data mining algorithms based on the type of data.

Džeroski (2006) Towards a general framework for data mining. *Lecture Notes in Computer Science vol. 4747. pp. 259–300*

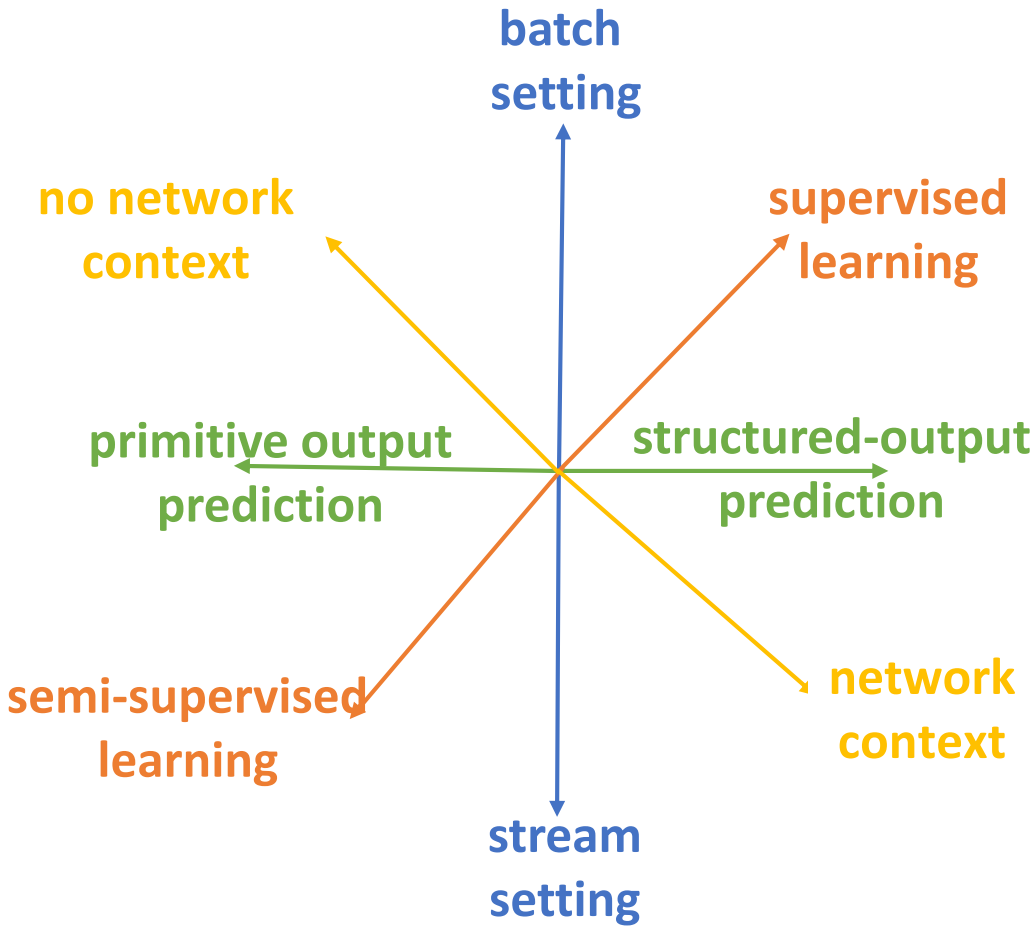
P. Panov et al. (2014) "Ontology of core data mining entities" *Data mining and knowledge discovery* 28(5-6):1222-1265

OntoDM-core structure



Representation of complex analysis tasks

- Extensions of OntoDM core in different dimensions
  1. Structured output prediction
  2. Streaming setting
  3. Semi-supervised learning setting
  4. Network context
- Examples of tasks
  - Semi-supervised online structured output prediction
  - Semi-supervised structured output prediction in a network context



How do we represent complex tasks?

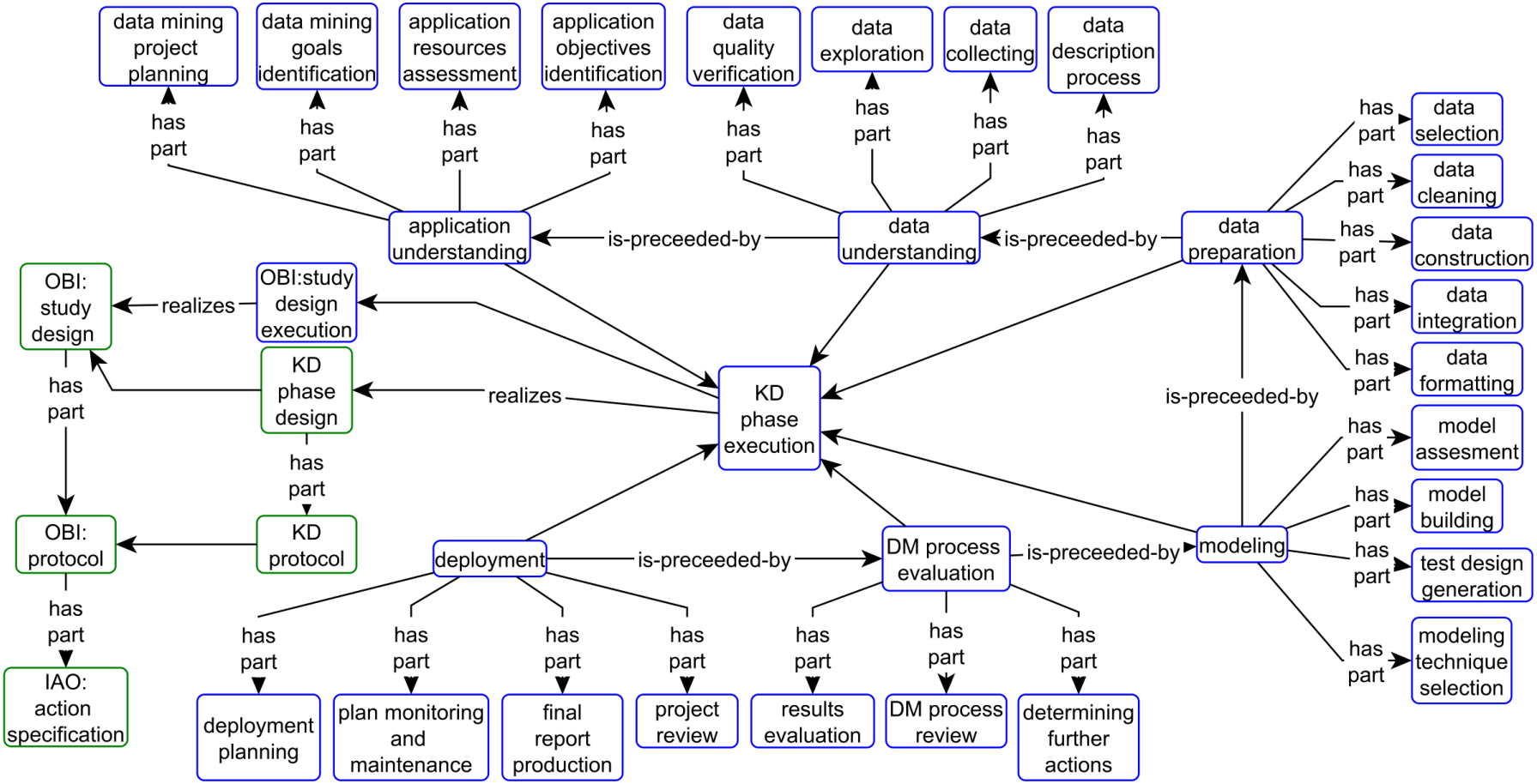
- Representation of the different dimensions at the level of datatype
- Here we used the ontology of datatypes OntoDT
  - Datatype descriptions are directly connected to the tasks
- Examples:
  - **semi-supervised multi-target regresion**
    ([tuple of primitives],choice([tuple of reals],void))
  - **semi-supervised multi-target regresion on data streams**
    sequence(([tuple of primitives],choice([tuple of reals],void)))
- In the same way we represented the tasks of learning with missing data in the context of all dimensions. For example:
  - **semi-supervised multi-label classification with missing data**
    ([tuple of choice(primitive,void)],choice(set{discrete},void))
- The task representations are further used to characterize algorithms that that solve the tasks

OntoDM-KDD

- Ontology for representing the knowledge discovery (KD) process [Panov et al., 2013]
- Based on the Cross Industry Standard Process for Data Mining (CRISP-DM)
- Most essential entities for describing data mining investigations
  - A taxonomy of KD specific actions and processes
  - Specifications of inputs and outputs of processes

Panov, et al. (2013) "OntoDM-KDD : ontology for representing the knowledge discovery process", 16th International Conference on Discovery science (DS 2013), Lecture notes in computer science vol. 8140, pg 126-140

OntoDM-KDD structure



IMPERATRIX project

- **IMPERATRIX project:** Improving Reproducibility of Experiments and Reusability of Research Outputs in Complex Data Analysis
- Slovenian national basic research project
  - 3 year project (2018-2021)
  - Financed by the Slovenian Research Agency
  - Value: 300K Eur

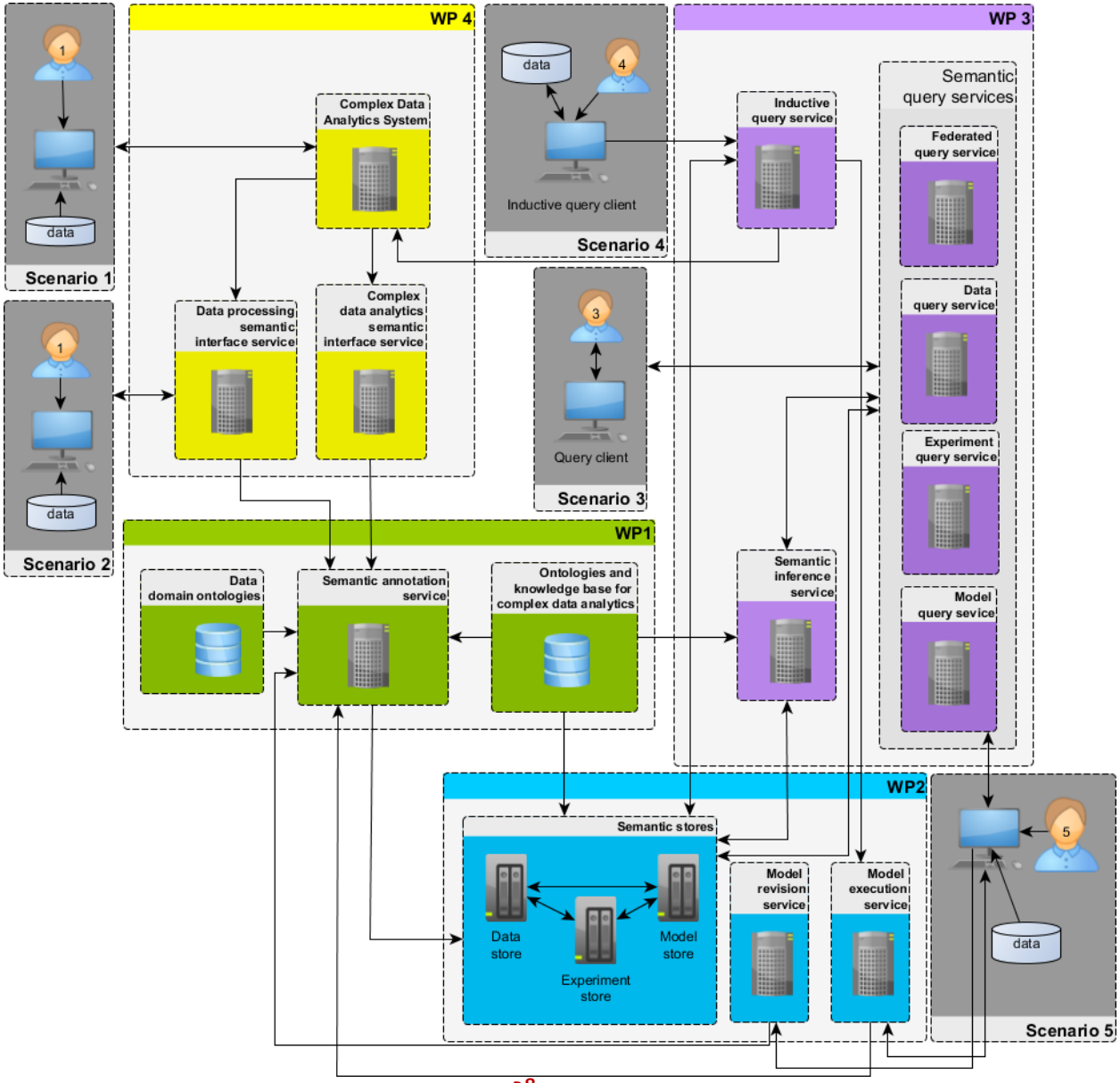
- **Main objective:** to improve the repeatability and reusability of data, experiments and outputs of experiments (data, models) in complex data analysis and moving towards a FAIR data analysis process.
  - Develop a prototype of a modular system for executing complex data analysis experiments, and semantically annotating, storing, querying and reusing results of experiments
- The project is combining approaches and ideas from the areas of complex data analytics, ontologies for science, semantic web and inductive databases

Project URL :http://kt.ijs.si/pance_panov/imperatrix/indexEN.html

Project objectives

- To design, implement and populate ontologies for complex data analysis to be used for semantic annotation
- To design and implement a prototype system for storing and querying semantically annotated data, experiments and models
- To test the querying capabilities of the prototype system in various scenarios
- To test the developed system in different use-case scenarios from various domains, such as machine learning, life sciences, space research and chemoinformatics

Vision for the IMPERATRIX prototype system



Current work: Reproducibility and reusability in the domain of dynamical systems modeling

- Storing metadata about models of dynamical systems in a machine readable form is one of the key steps towards their accessibility and reusability
- In the domain of process-based modeling of dynamical systems, the task is to construct an explanatory model of a dynamical system from domain knowledge and data expressed
- Ontology for Process-Based Modeling of Dynamical Systems (OntoPBM) [Kostovska et al., 2019]
  - Vocabulary of key terms about the process-based modeling paradigm
  - To capture the domain-specific characteristics, we extend OntoPBM with domain specific terms
  - The ontology is used for annotation of datasets, experiments and models in this context

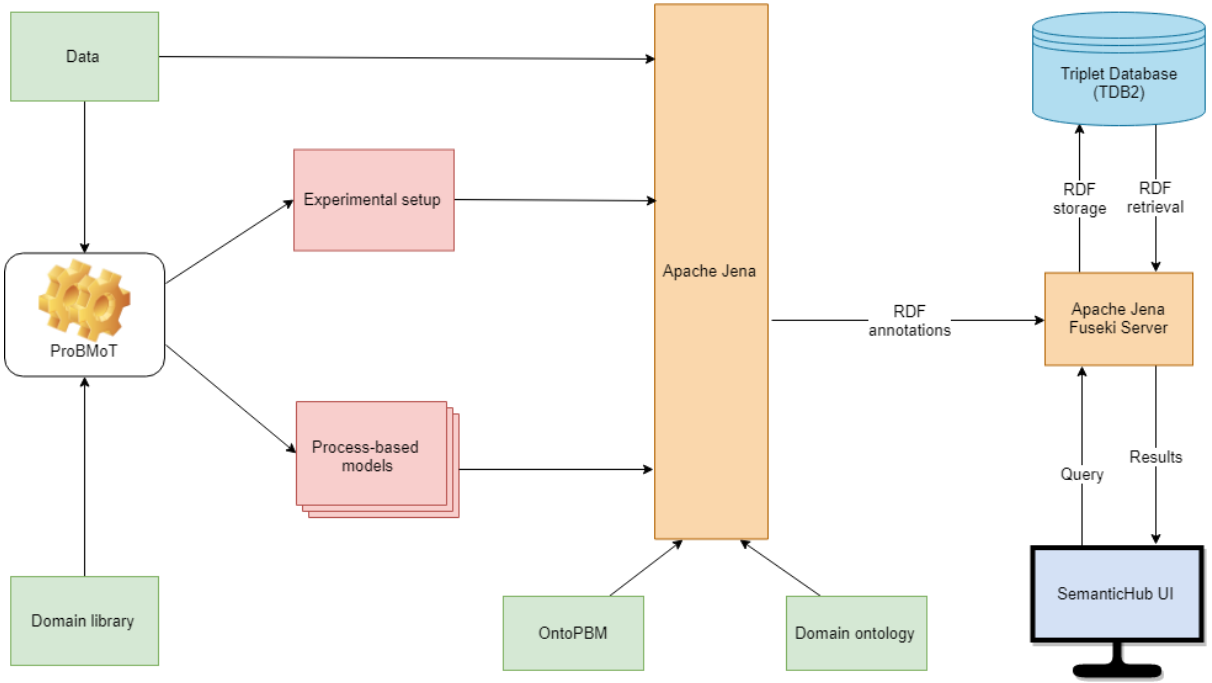
Kostovska et al. (2019) Process-based modeling of dynamical systems: An ontology of core entities (Tech. report)

Prototype system

- Ontology-based prototype system for annotation, storage and querying of process-based models, datasets and experiments [Tolovski, 2019]
  - The annotations for each experimental instance and produced model are stored in an RDF triple store.
  - We can execute SPARQL queries on facts asserted in the annotations, as well as facts inferred from the domain knowledge encoded in the ontology
- Applications
  - System identification and control engineering – Modeling of the two tanks system
  - Ecological modeling – Aquatic ecosystem modeling

Tolovski et al. (2019) Towards reusable process-based models of dynamical systems: A case study in the domain of aquatic ecosystems. In proceedings of MIPRO 2019, Opatija, Croatia (To appear)

Outline of the system prototype



Models	Train error	Test error	Validation error	Error function
BledIncomplete128_2019-01-28/18:31:438_127	0.644	0.664	0.934	RMSE
BledIncomplete128_2019-01-28/18:29:689_53	0.701	0.735	0.721	RMSE
BledIncomplete128_2019-01-28/18:29:115_28	0.705	0.967	0.689	RMSE
BledIncomplete128_2019-01-28/18:30:338_82	0.720	1.263	1.071	RMSE
BledIncomplete128_2019-01-28/18:29:769_29	0.724	0.982	0.747	RMSE
BledIncomplete128_2019-01-28/18:29:346_10	0.729	0.939	0.769	RMSE
BledIncomplete128_2019-01-28/18:29:641_46	0.733	0.918	0.955	RMSE
BledIncomplete128_2019-01-28/18:30:00_116	0.735	0.902	0.897	RMSE
BledIncomplete128_2019-01-28/18:30:760_98	0.737	0.871	0.963	RMSE
BledIncomplete128_2019-01-28/18:29:483_30	0.742	0.876	1.142	RMSE

Items per page: 10 1 - 10 of 128 Show query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX RO: <http://purl.obolibrary.org/obo/>
PREFIX DCAT: <http://www.w3.org/ns/dcat#>
PREFIX OntoPBM: <http://w3id.org/ontopbm#>
SELECT ?label ?downloadURL ?trainErrValue
       ?testErrValue ?validationErrValue
WHERE {
  ?PBMInstance rdfs:label ?label.
  ?PBMInstance OntoPBM:OntoPBM_00800 ?downloadURL.
  ?PBMInstance OntoPBM:OntoPBM_00727 ?trainErr.
  ?trainErr OntoPBM:OntoPBM_00496 ?trainErrValue.
  ?PBMInstance OntoPBM:OntoPBM_00726 ?testErr.
  ?testErr OntoPBM:OntoPBM_00496 ?testErrValue.
  ?PBMInstance OntoPBM:OntoPBM_00728 ?validationErr.
  ?validationErr OntoPBM:OntoPBM_00496 ?validationErrValue.
  ?PBMInstance DCAT:dataset 'BledData2000.data'.
  {
  select (count(?userRequestedProcesses)
                                as ?foundProcessesNum)
  where {
    values ?userRequestedProcesses
                        {"GrowthRate" "LightInfluence"}
    {
      select distinct ?foundProcesses
      where {
        ?PBMInstance RO:RO_0002351 ?processInstance.
        ?processInstance rdf:type ?processClass.
        ?processClass rdfs:subClassOf OntoPBM:OntoPBM_00453.
        ?processClass rdfs:subClassOf ?processSuperClass.
        ?processSuperClass rdfs:label ?foundProcesses.
      }
    }
    FILTER (?userRequestedProcesses = ?foundProcesses)
  }
  }
FILTER(?foundProcessesNum = 2)
}
```

Summary

- Reproducibility
  - Defining minimal information about a performed experiment and the produced research outputs
  - Development of best practices for documenting of research activities
- Reusable research outputs
  - FAIR data principles
  - Adding semantics to data objects: use of semantic web technologies
- Improving reproducibility and reusability in complex data analysis
  - Use of semantic representations in the domain of data analysis
  - IMPERATRIX project
  - Reproducible experiments and reusable models in the domain of dynamical systems modeling