Ministero dell'Università e della Ricerca

PRIN

S-PIC4CHU:
Semantics-Enriched Techniques for Data Preparation in Data Science

Gianvincenzo Alfano[1], Ilaria Bartolini[2], Diego Calvanese[3], Paolo Ciaccia[2], Sergio Greco[1], Davide Lanti[3], **Pasquale Leonardo Lazzaro**[5], Emilia Lenzi[4], Davide Martinenghi[4], Cristian Molinaro[1], Marco Patella[2], Letizia Tanca[4], Riccardo Torlone[5], and Irina Trubitsyna[1]

[1] University of Calabria
[2] University of Bologna
[3] Free University of Bozen-Bolzano
[4] Politecnico di Milano
[5] Roma Tre University

Motivation

- Real-world data is often incomplete, inconsistent, and biased
- Typically, fairness is neglected at the data level
- Data preparation = **bottleneck** in ML workflows.
- Data preparation techniques often lack domain knowledge, transparency, automation, and generalizability

Objectives

- Introduce a **semantic-based** **Data Preparation Pipeline** (DPP)
  - data cleaning
  - data integration
  - data transformation and reduction
  - deduplication
  - error detection and missing value imputation,
  - conflict reconciliation among different data quality dimensions
- Enable novel models and techniques for data preparation
- Embed provenance, integrity, and fairness in transformations
- Ensure interpretability and reusability

Proposed approach

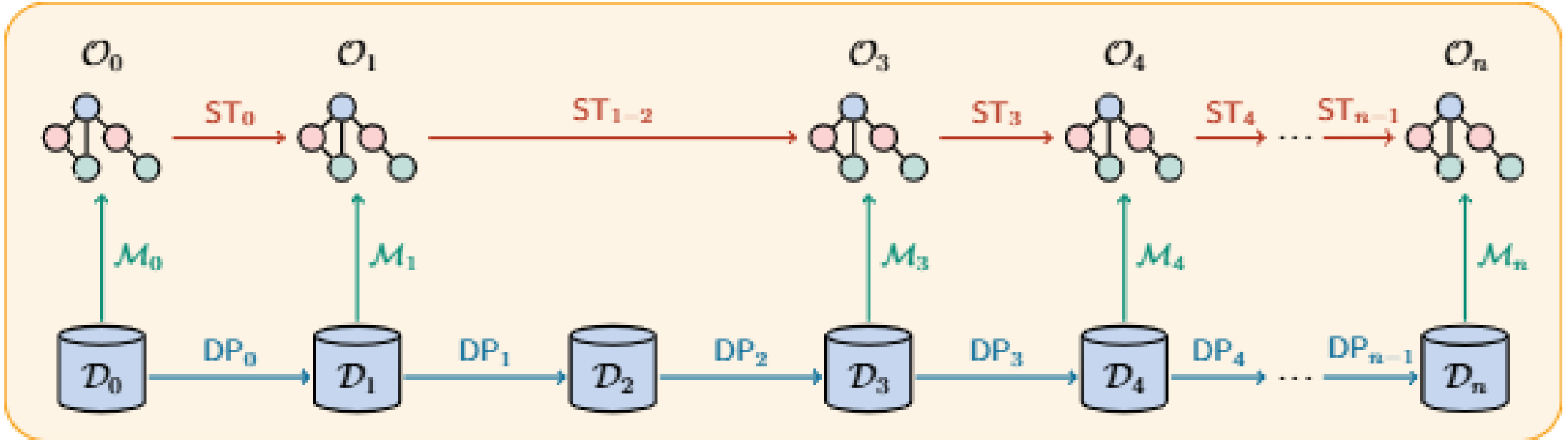
We enrich the traditional DPP with **semantic annotations**, enabling a deeper understanding of both the data and its transformations.

→ **Semantic Transformation Pipeline** (STP)

- The stages of the DPP are semantically described using ontologies
- They are exposed as (possibly virtual) knowledge graphs
- Key stages of the DPP (D_0, D_1, D_2, ...) are linked to domain ontologies (O_0, O_1, O_2, ...) via semantic mappings (M_0, M_1, M_2, ...).
- LLM4VKG: LLM-assisted framework for mapping and ontology enrichment (+17% F1 on RODI).
- These mappings allow us to reason about the transformations applied to the data (DP_0, DP_1, DP_2, ...) and reinterpret them at a higher semantic level as semantic transformation steps (ST_0, $ST_{1\text{-}2}$, ST_3, ...).

Semantic Transformation Pipeline

Semantic enrichment transforms a technical pipeline into a knowledge-aware, explainable workflow.



Semantic enrichment in Action: Data Quality

- Ontology-enriched imputation → missing values filled consistently.

- Preference-based inconsistency resolution → semantic rules choose most plausible data.

- Preferences can also be **context-dependent** (location, time, season, weather).

Semantic enrichment in Action: Ethics & Context

- **CDT** = Context Dimension Tree.
- **ERT** = Ethical Requirements Tree.
- Together: select preprocessing strategies based on context and ethical constraints.
- **Case study**: IMM Design Lab (Politecnico di Milano) to support urban policy makers applying IMM methodology for Sustainable Development Goals (SDGs).

Semantic enrichment in Action: Provenance

- Provenance captured at cell-level granularity (**PROLIT**).
- Implementation: **Pandas + Neo4j** with Observer pattern.
- Supports queries: Why, How, Why not.
- Evaluation datasets: German Credit, COMPAS, Census.
- Enables new queries: Item History, Impact on Feature Spread.

Semantic enrichment in Action: Ranking & Fairness

- **Partitioning strategies**: efficient skyline computation.
- **Directional queries**: outperform linear top-k in cumulative recall.
- **UF-skyline**: balance utility and fairness.

Conclusions & Future Work

- **Achievements:**
  - LLM4VKG for ontology/mapping construction.
  - Ontology-enriched imputation.
  - Preferences for conflict resolution.
  - CDT/ERT for ethics & context.
  - Fine-grained provenance.
  - Fairness-aware ranking.
- **Next steps**:
  - Extend to multimodal data (images, video, audio).
  - Validate in real-world domains.
  - Goal: adaptive, scalable, and responsible data preparation.

“Semantic enrichment makes data preparation transparent, explainable, and trustworthy.”