APPLIED DATA SCIENCE

ONTOLOGY ALIGNMENT CLASSIFIER

BASED ON NLP

Era Alcani
Francesco Dal Cero
Francesco Mastrosimone

Politecnico di Torino
1859

Repertorio
UNLOCKING PATIENT PREFERENCES

TABLE OF CONTENTS

PROJECT OVERVIEW

01

PROJECT CONTEXT AND SOLUTION

THE CHALLENGE: SEMANTIC INCONSISTENCY

Context : Healthcare evidence is scattered across isolated "data silos".
Problem : Different terms (e.g., Fatigue vs. Asthenia) prevent meaningful integration of this data.

THE SOLUTION: AI SEMANTIC ENGINE

Goal : Develop an AI-powered engine to automatically harmonize fragmented data into structured evidence.
Impact : Replaces manual "brittle scripts" with a scalable, automated pipeline.

OBJECTIVES AND RESEARCH PERSPECTIVE

THE GOAL: AUTOMATED HARMONIZATION

To replace manual, brittle scripts with a robust framework that transforms raw attributes into high-confidence clinical evidence.

THE RESEARCH OBJECTIVES

SYSTEM LEVEL: Build a reusable framework for interpretable ontology alignment.

RESEARCH LEVEL: Evolve from simple string matching to capturing actual clinical meaning and context.

USER PERSONA

Name: Elias Vance
Role: Data Engineer
Company: Repertorio

"I need a pipeline that cleans messy data automatically so i can focus on structure"

GOALS

Automation: Reduce manual review of 1000s of attributes.
Scalability: Pipeline must handle daily data spikes.
Accuracy: Needs "Scientific" context, not just keywords.

PAIN POINTS

Synonym Trap: Keyword search misses obvious matches.
Maintenance Difficulty: Can't retune every single week.
Noisy Data: Raw inputs are full of typos and errors.

THE REPERTORIO SCENARIO

"Elias receives 50k raw attributes daily. He needs a 'Filter-then-Rank' system to act as a gatekeeper- automating the easy 60% and flagging the hard 40%"

RESEARCH HYPOTHESIS

02

RESEARCH HYPOTHESIS

Efficiency and Ambiguity (H1)

Question: Do we always need complex deep learning models to align clinical terms?
Hypothesis: No. Model complexity should match semantic ambiguity: deterministic rules handle clear cases, while deep models are needed only for ambiguous or synonymous terms.

Architecture (H2)

Question: How do we balance high accuracy with scalability?
Hypothesis: Efficiency demands a "Filter-then-Rank" architecture. Scalability is achieved by filtering candidates with a fast retriever before applying expensive deep semantic scoring.

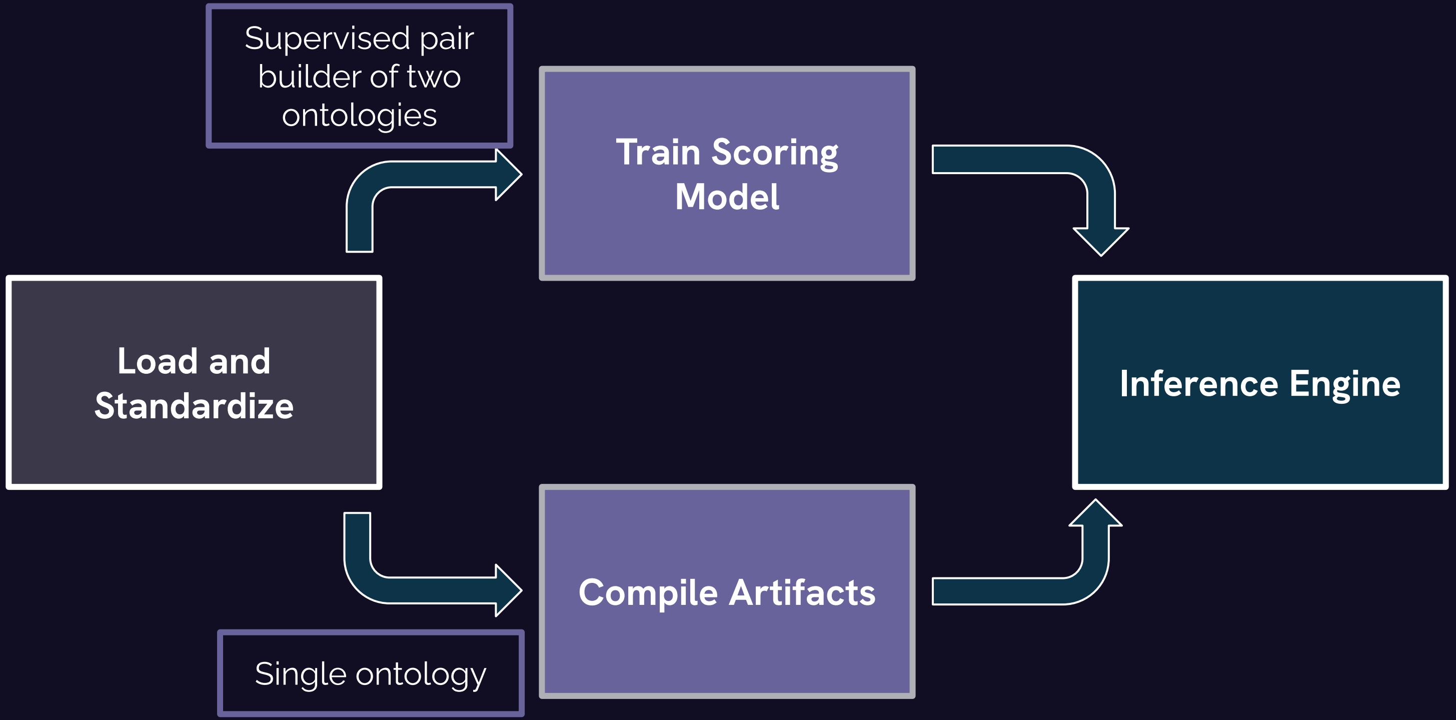
Data Strategy (H3)

Question: Is random training data sufficient for clinical accuracy?
Hypothesis: No. To generalize well, the model must be explicitly trained on "Hard Negatives"—tricky, similar-looking concepts—rather than just random non-matches.

SYSTEM DESIGN

03

SYSTEM DESIGN - MODULAR OVERVIEW



- Reusable across studies – build once, run many times

SYSTEM WORKFLOWS & ENTRY POINTS

1

Training

- Ontology loading & standardization
- Reference alignment loading
- Training dataset construction
- Model fine-tuning

From standardized ontology →

Offline Preprocessing

- Unified ontology representation
- Exact-match lookup structures
- Candidate retrieval indices
- Semantic representations (for hybrid retrieval)

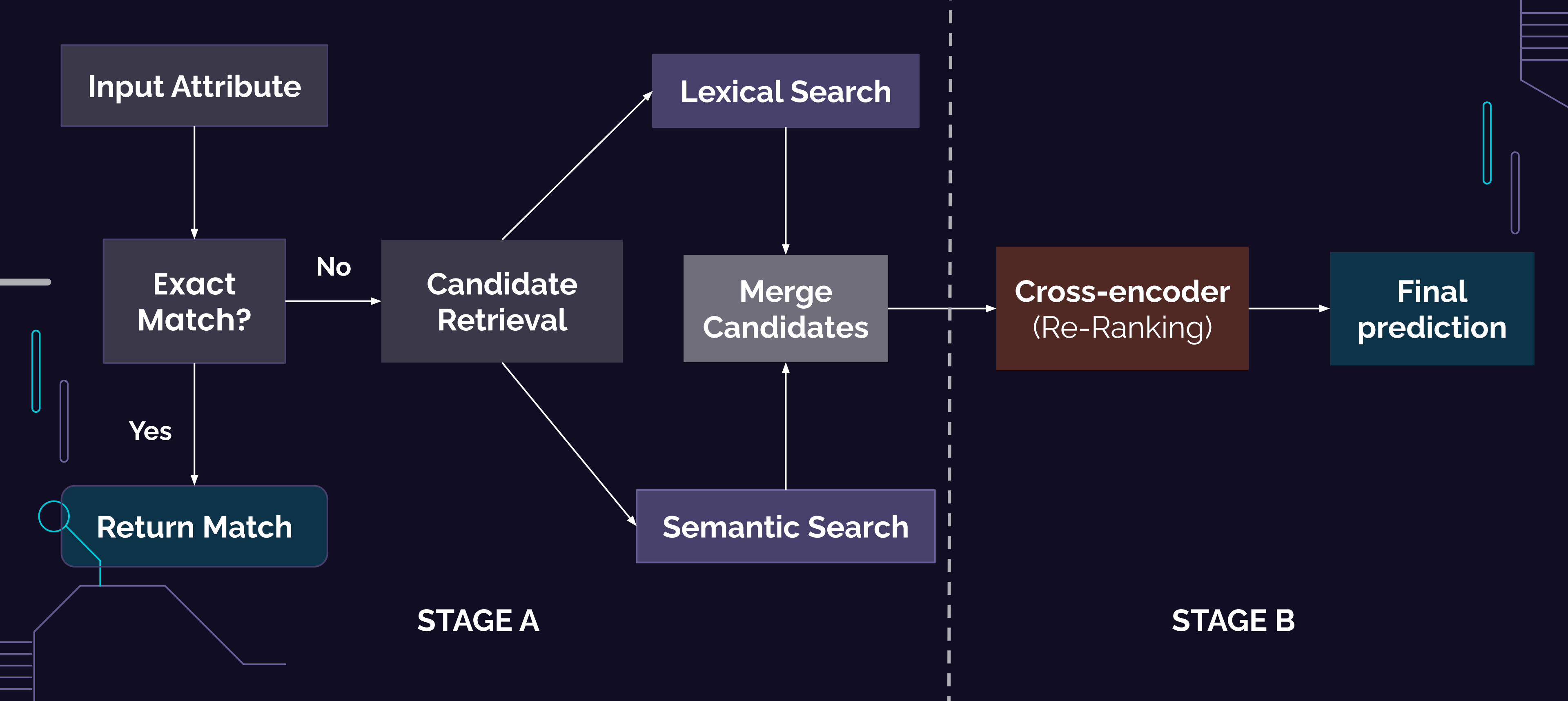
2

3

Inference

Exact match shortcut → Candidate selection → Neural scoring

INFERENCE PIPELINE



DATA ENGINEERING

04

DATA PIPELINE: From Ontology to Supervised Training

ONTOLOGY STANDARDIZATION

Process: Parses heterogeneous OWL/RDF files into a Unified View.

Normalization: Maps diverse metadata (IRI, Labels, Synonyms, Definitions) into a standardized, ontology-agnostic schema.

Goal: Ensures the model sees consistent input regardless of the source format.

SUPERVISED DATASET CONSTRUCTION

Structure: Converts data into pairwise samples ***(Attribute, Concept, Label).***

Positives: Drawn from Gold Standard alignments.

Negatives:

- ***Random***: For broad coverage.
- ***Hard Negatives***: Mined from semantically close "near-miss" candidates.

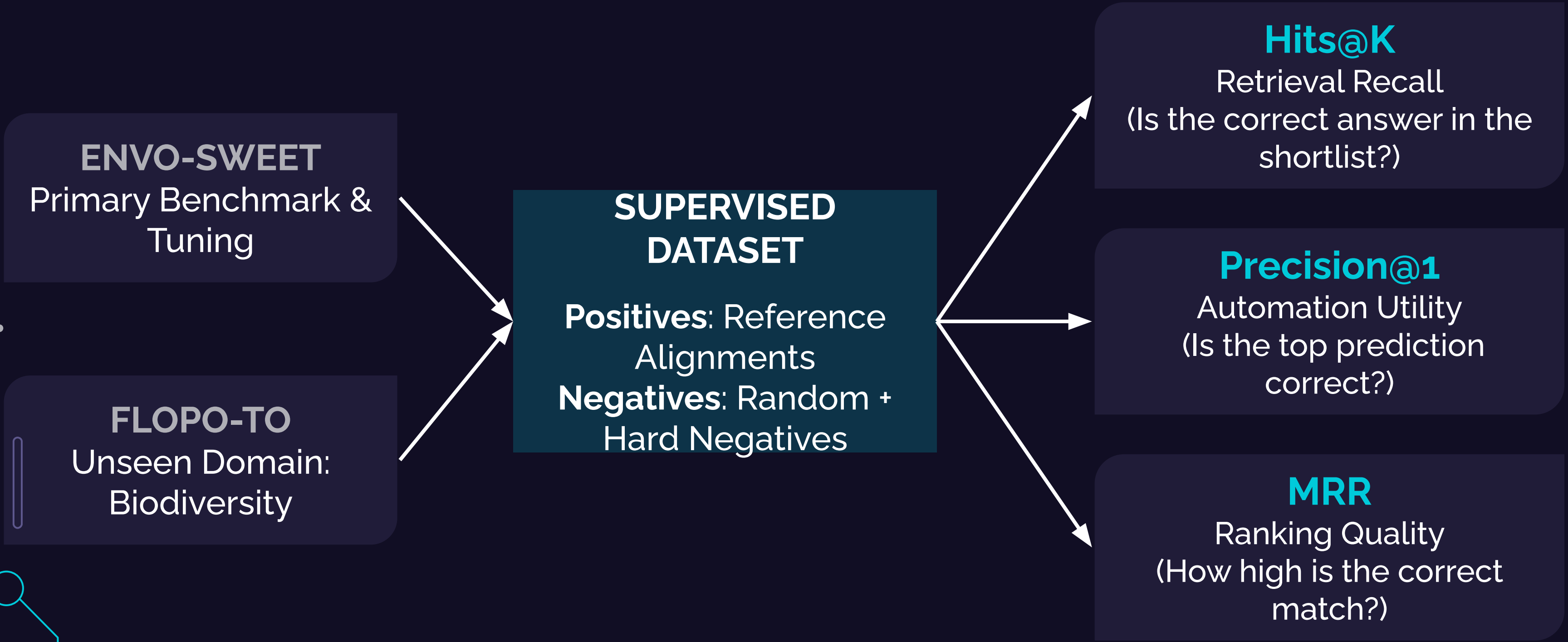
KEY OUTCOME

Balanced Training Set: The pipeline ensures a 1:1 ratio of positives to negatives, preventing class imbalance and forcing the model to learn fine-grained semantic distinctions.

EVALUATION & RESULTS

05

EVALUATION SETUP & METRICS



MODEL SCREENING & SELECTION

The Insight: All models were equally good at *finding* candidates (Recall ~ 70%), but SciBERT was the only one capable of consistently *ranking* the correct one at the top.

The Reason: While both models are trained on scientific papers, SciBERT covers a broader range of domains, whereas PubMedBERT is strictly biomedical. This wider scope gave SciBERT a slight edge in handling the mixed terminology of our dataset.

BASELINES & EVALUATION CONVENTION

EXACT MATCH (Reference)	LEXICAL MODE (Baseline)
Definition: Strict resolution via label or synonym matching only.	**Definition**: Integrates Exact Match with Subword-based Lexical Retrieval.
Constraint: Attributes with no direct correspondence are treated as incorrect.	**Ranking**: Uses SciBERT Cross-Encoder for re-ranking.
Role: Captures the proportion of cases resolvable deterministically.	**Constraint**: Semantic Retrieval is DISABLED.
Metric: **Precision@1: 0.6119**	**Metrics**: Precision@1: 0.6716

KEY INSIGHTS

Structured lexical retrieval combined with discriminative re-ranking provides a ~6% gain over pure exact matching. This isolates the contribution of ranking *before* we introduce Dense Semantic Search.

H1

H2

ABLATION STUDY: HARD NEGATIVES

H3

Training Configuration	Precision@1
SciBERT (Standard)	0.6866
SciBERT (NO HN)	0.64 (-5% Drop)

Training Configuration	Hits@20
SciBERT (Standard)	0.70
SciBERT (NO HN)	0.70 (Unchanged)

Why Precision Drops?

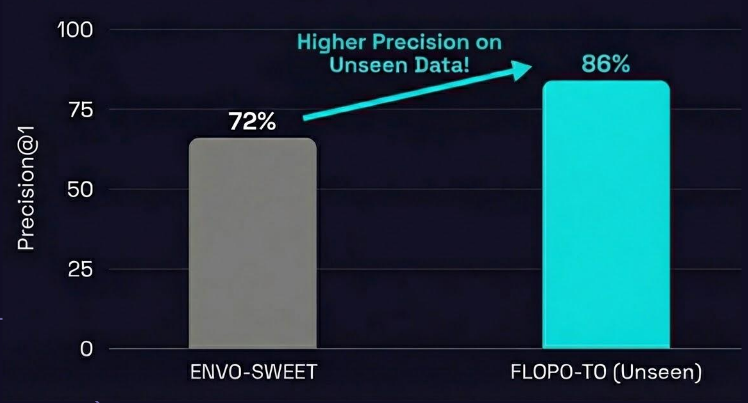
Without Hard Negatives, the model gets confused by "Near-Misses". It treats related concepts as correct, lowering the automation score.

Why Recall Stays Same?

Hard Negatives don't help the model *find* candidates. They only teach the model how to *rank* the shortlist correctly. Thus, the search ability (Hits@20) remains identical.

OPTIMIZATION & ZERO-SHOT GENERALIZATION

OPTIMIZATION	ZERO-SHOT ON FLOPO-TO
Config: Hybrid Retrieval (Lexical + Semantic) + Hard Negative Training **Result (ENVO-SWEET)** Achieved best overall performance: *Precision@1: 0.72 * Hits@20: 0.78	**Dataset** FLOPO-TO (Unseen Domain: Biodiversity). **Metrics** - Precision@1: 0.86. - Hits@20: 0.92.



Key Insight

The system achieves even higher precision (0.86) on an unseen domain than on the training set. This proves the model learned general alignment logic, rather than just memorizing the specific training data .

CONCLUSIONS

06

CONCLUSIONS

Scalability via "Filter-then-Rank" Design

High accuracy is achieved without high latency by combining Deterministic Shortcuts with targeted SciBERT scoring. This ensures the system is computationally viable for real-time production use.

The Critical Role of Hard Negatives

Standard training is insufficient for fine-grained ontology alignment. Mining "Hard Negatives" is essential to prevent the model from confusing near-miss concepts.

Strong Zero-Shot Generalization

The system is not limited to its training data. It achieved superior performance on the unseen FLOPO-TO dataset, proving that it learns general alignment logic rather than just memorizing specific terminologies

Alignment quality emerges from the interaction of System Design (Filtering), Data Strategy (Hard Negatives), and Model Choice (SciBERT)—not from any single component in isolation.

THANK YOU

ANY QUESTIONS?