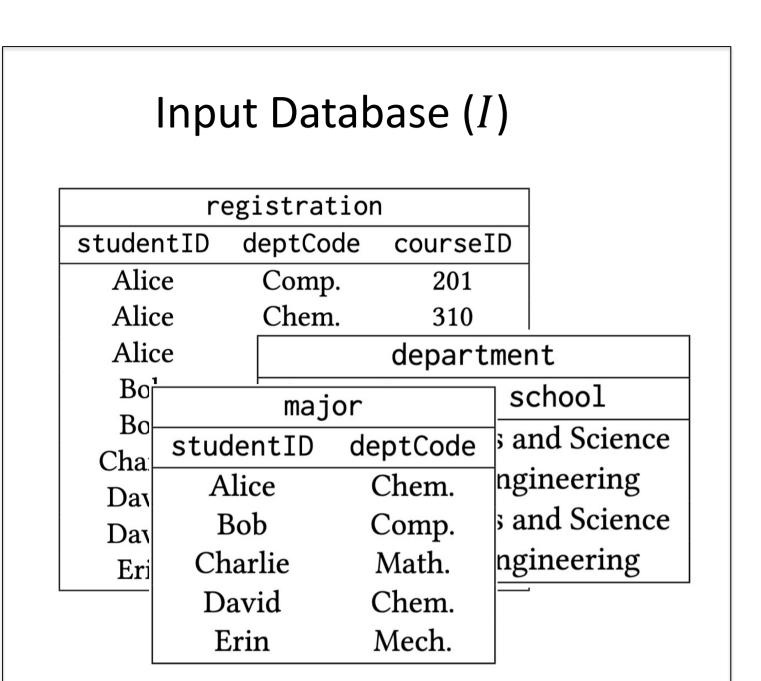


Relational Query Synthesis > Decision Tree Learning

Aaditya Naik, Aalok Thakkar, Adam Stein, Mayur Naik, Rajeev Alur asnaik@seas.upenn.edu



Synthesizing Select-Project-Join Programs



Output Labels Negative Labels (O⁻) Positive Labels (O^+) Alice Charlie Bob David

Problem: Given I, O^+ , and O^- , synthesize SPJ query $oldsymbol{Q}$ such that

$$O^+ \subseteq [[Q]](I)$$
, and $O^- \cap [[Q]](I) = \phi$

Target Query

SELECT registration.studentID FROM registration JOIN department ON

registration.deptCode = department.deptCode WHERE registration.courseID < 500 AND

department.school = "Engineering"

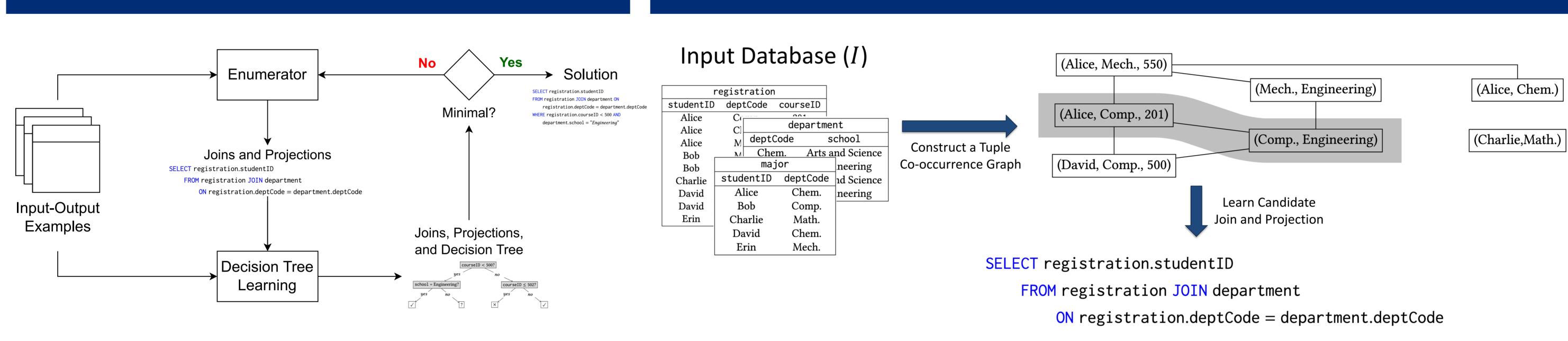
Challenge 2: Learning Selections **Existing Solutions: Decision Trees**

Challenge 1: Learning Joins and Projections **Existing Solutions: Inductive** Logic Programming

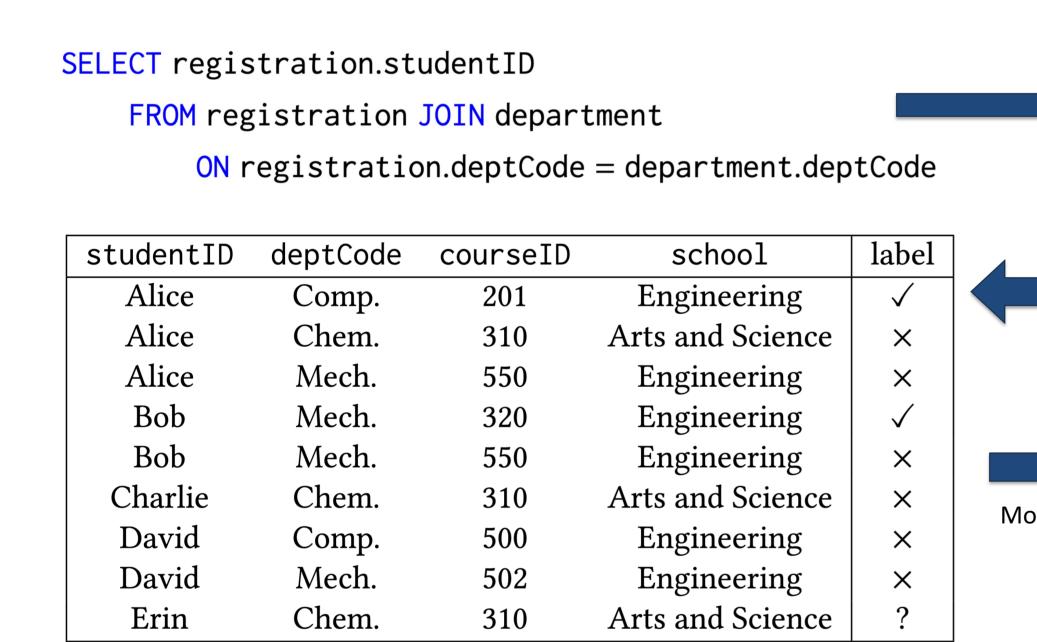
How to simultaneously address both challenges?



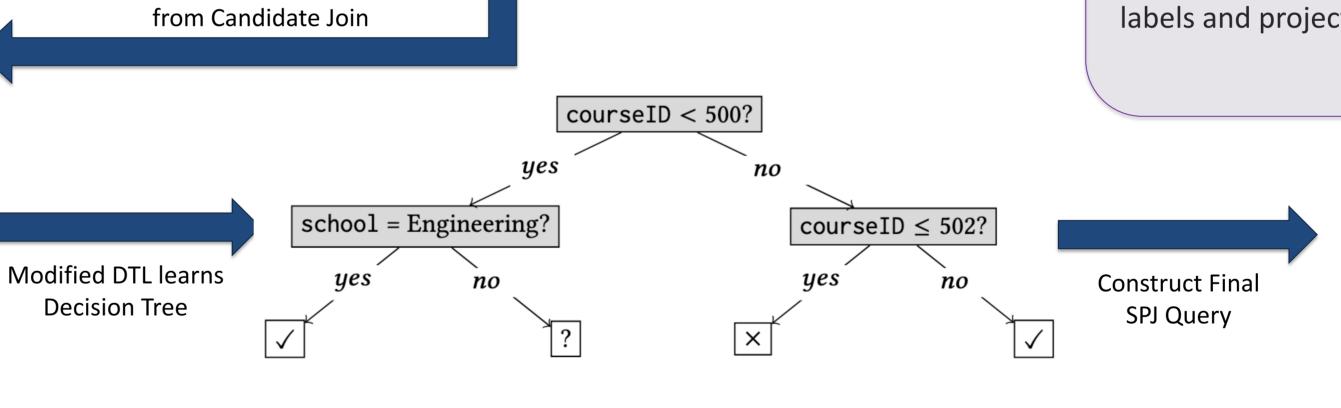
Learning Joins and Projections



Learning Selections







Modified Decision Tree Learning

 $p = P(O^+ | \pi_Y(T_C), O^+ \cup O^-)$ **Modified Entropy equation** $n = P(O^- | \pi_{\gamma}(T_C), O^+ \cup O^-)$ $S(N) = -\left(p \log_2 p + n \log_2 n\right)$ allows support for partial labels and projected columns

SELECT registration.studentID FROM registration JOIN department ON registration.deptCode = department.deptCode

WHERE registration.courseID < 500 AND department.school = "Engineering"

Evaluation Setup

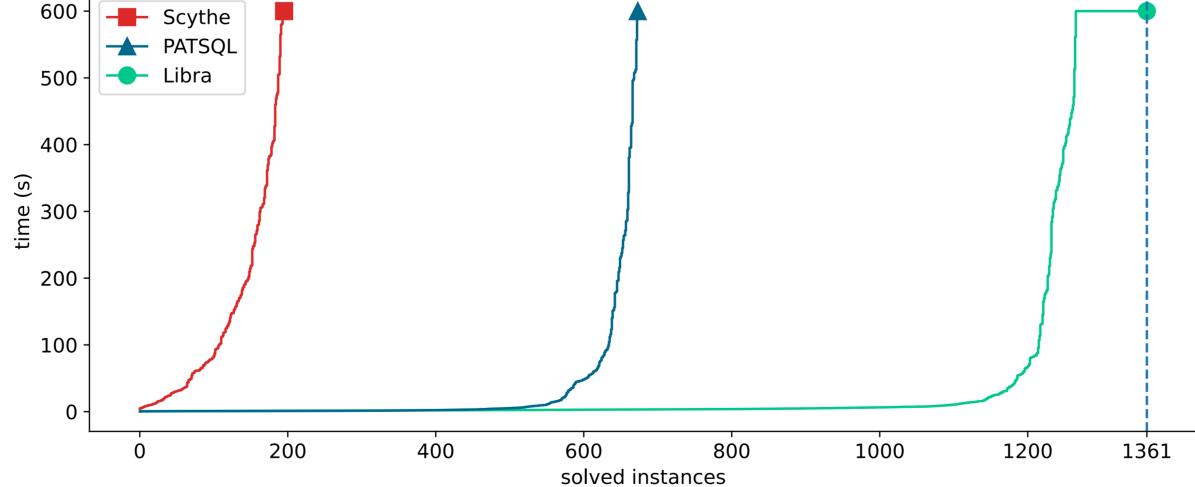
Baselines Scythe **PATSQL**

Benchmarks

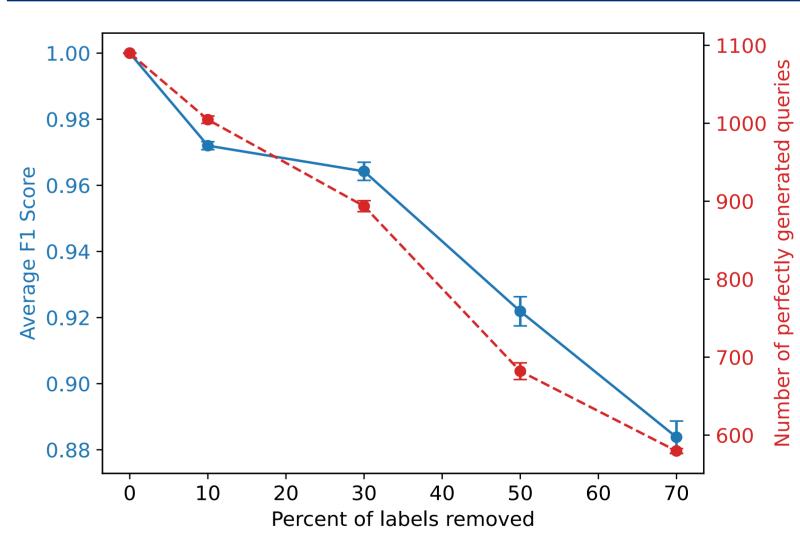
1496 SPJ queries over 160 databases from the Spider and Geography benchmarks

For each query we provide the full database along with positive and negative labels.

Performance



Sensitivity to Partial Labels



Succinctness 600

Scythe **PatSQL** Libra Reference of benchi 100 Size of the target query

Conclusion

- We present a novel technique to synthesize SPJ queries by interleaving the synthesis of joins and selections
- We synthesize joins using an example-guided enumerator and modify the classical decision tree learning algorithm to synthesize selections
- We implement the algorithm in a tool named Libra and evaluate it over around 1500 queries from 160 databases
- Libra solves **1,361 benchmarks**, outperforming PATSQL (673) and Scythe (195).
- 99% of Libra's solutions are minimal.
 - Libra is robust to partial labels, producing solutions with an F1 score of **0.88** given only **30%** of the labels.