UNIVERSITY OF ALBERTA

LaBD

Automating Weak Supervision to Find Missing Labels for Big Data

Presented by: Mona Nashaat

Slides are inspired from our paper:

M. Nashaat, A. Ghosh, J. Miller, S. Quader, C. Marston, and J.-F. Puget, “Hybridization of Active Learning and Data Programming for Labeling Large Industrial Datasets,” in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 46–55, 2018.

① Motivation: Why Labels?

Derive Value from Business Data

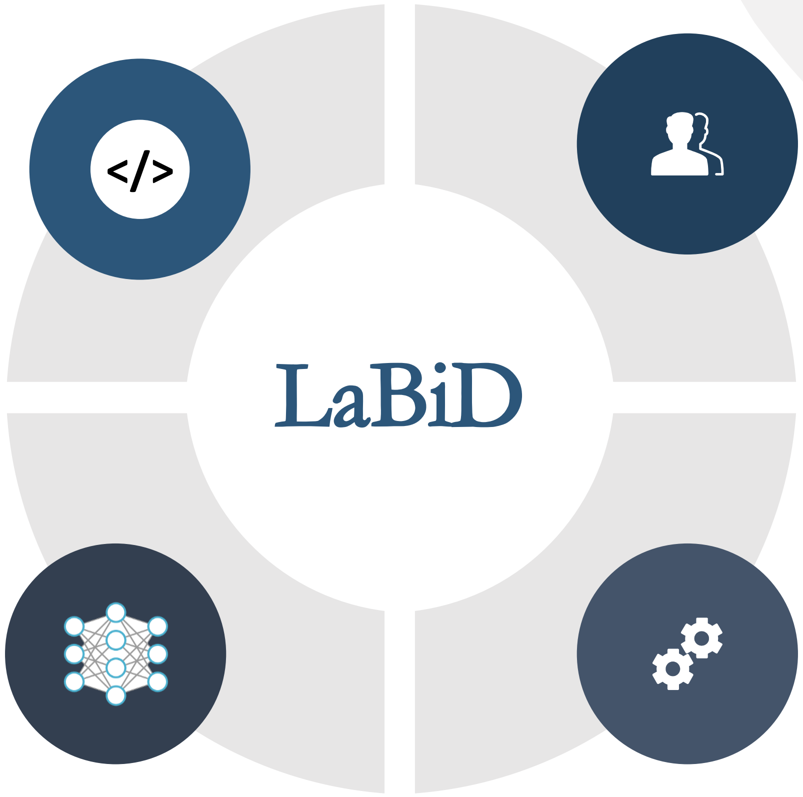
- Essential to build supervised machine learning models.
- The quality and the size of training data limits the performance of predictive systems.
- Labeled training datasets do not exist.

> “About 70% of complex analytical tasks today are related to data preparation. There have to be people who are preparing and labeling data for machines to understand. Here’s a situation in which human labor automation driven by ML creates new job opportunities.”

Guru Banavar, IBM data scientist

Obtaining Accurate Labels is Expensive

② Proposed Solution: LaBiD



LaBiD

Data Programming

Learns a model of the training set that includes labeling functions.

Meta Active Learning

Treats active learning algorithm design as a meta-learning problem and learn the best criterion from data

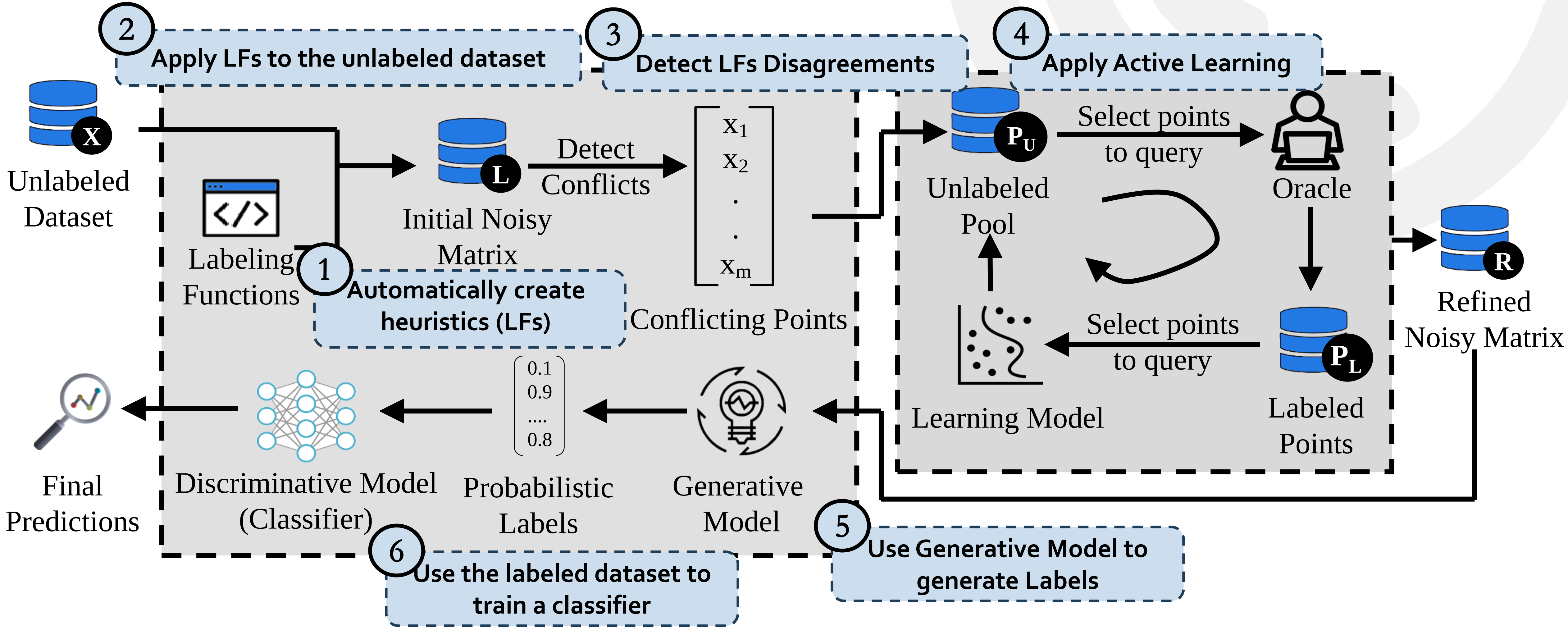
Gets a lower-quality labels more efficiently and/or at a higher abstraction level

Weak Supervision

Automating the process of generating heuristics that assign training labels to unlabeled data

Automating Weak Supervision

3 Overall Architecture



4 Experimental Results

DATASETS USED IN THE EXPERIMENTS

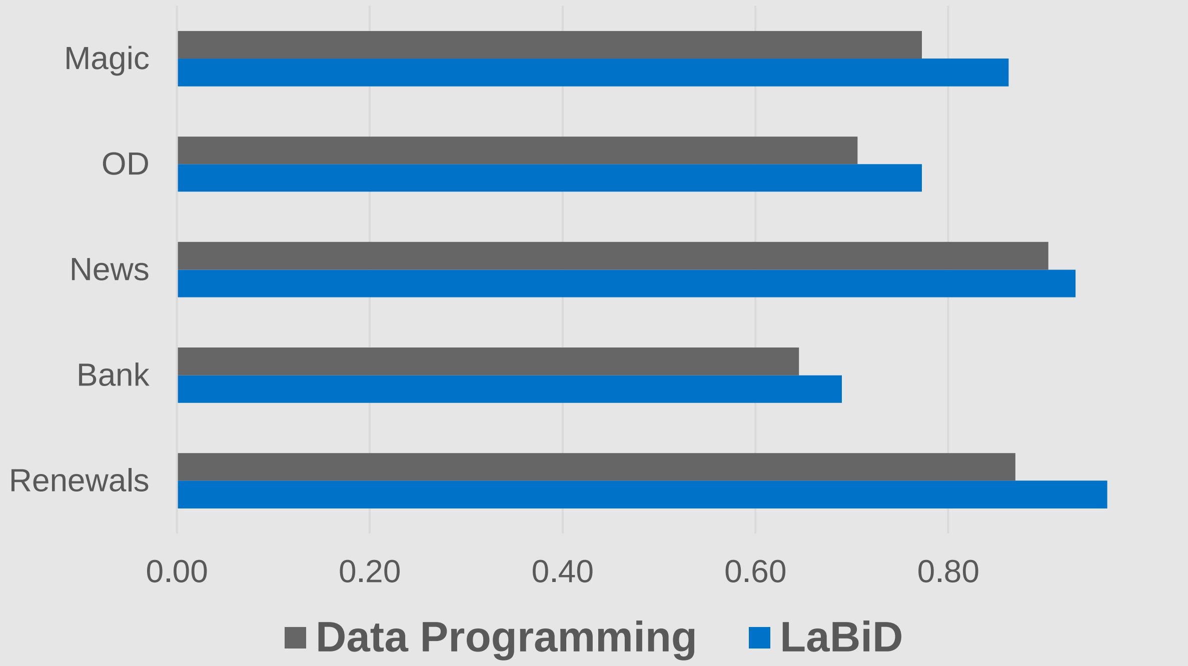
Dataset	# of records	# of attributes
Higgs	11,000,000	28
Renewal Sales	1,354,704	11
Rain Prediction	142,000	24
Travel Insurance	63,300	11
Bank	45,211	17
News	39,797	61
Credit Card	30,000	24
Tenancy Detection	20,560	7
Magic	19,020	12

- **Data Programming**
- **Active Learning**
- **Meta Active Learning**
- **Automated Weak Supervision**

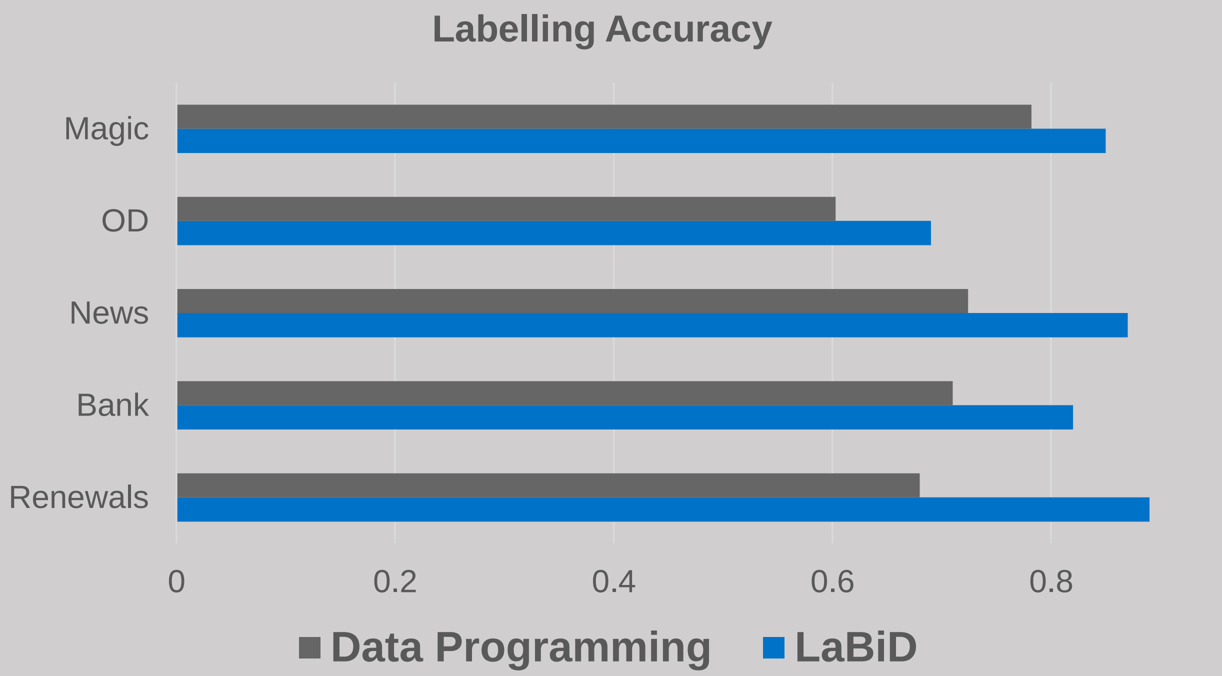
EXISTING APPROACHES

4 Experimental Results

PREDICTION PERFORMANCE



LABELS ACCURACY



⑤ Application to Database

Analytics challenges

- Never assume the data is clean.
- Automatically create heuristics.
- Apply the LaBiD flow and compare the results with ground truth.
- Double check with the user to detect outliers and missing values.

> Bad data is bad for business. Poor quality data is costing businesses at least 30% of revenues.

Reported by Ovum Research

Next steps…

UNIVERSITY OF ALBERTA

THANK YOU

✉ nashaata@ualberta.ca