

LaBiD: Automating Weak Supervision to Find Missing Labels for Big Data

Mona Nashaat¹, Aindrila Basak¹, James Miller¹, Shaikh Quader²

¹University of Alberta, Edmonton, Canada, ²IBM Canada

Motivation

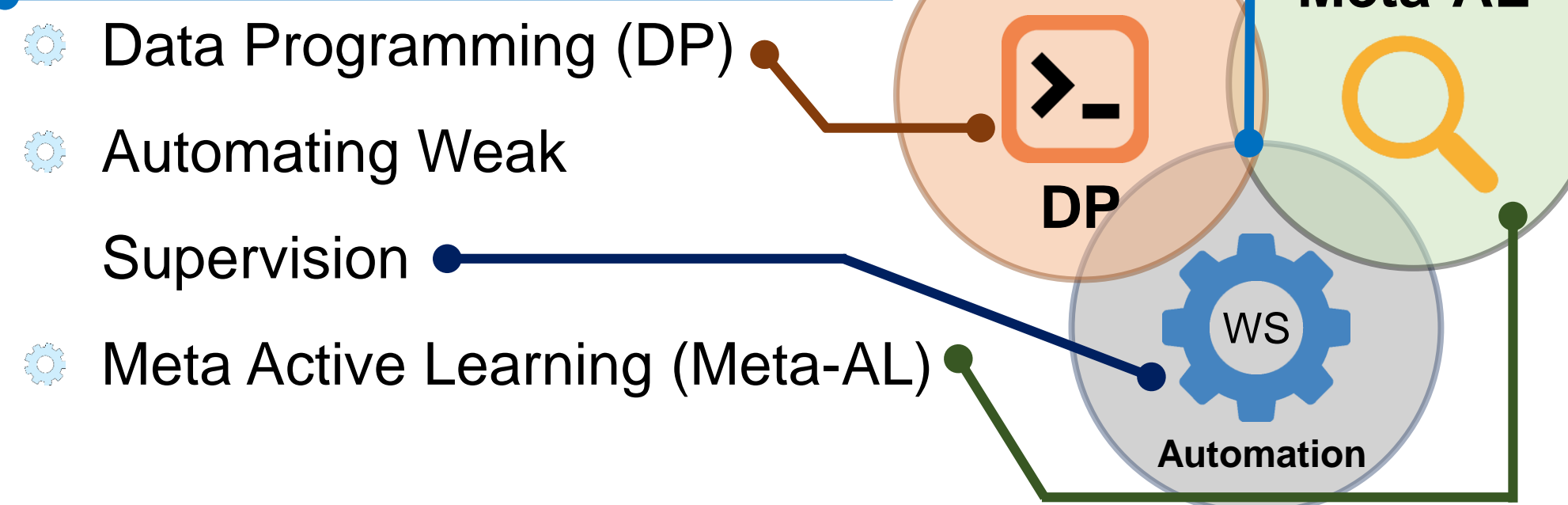
Obtaining Accurate Labels is Expensive!

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

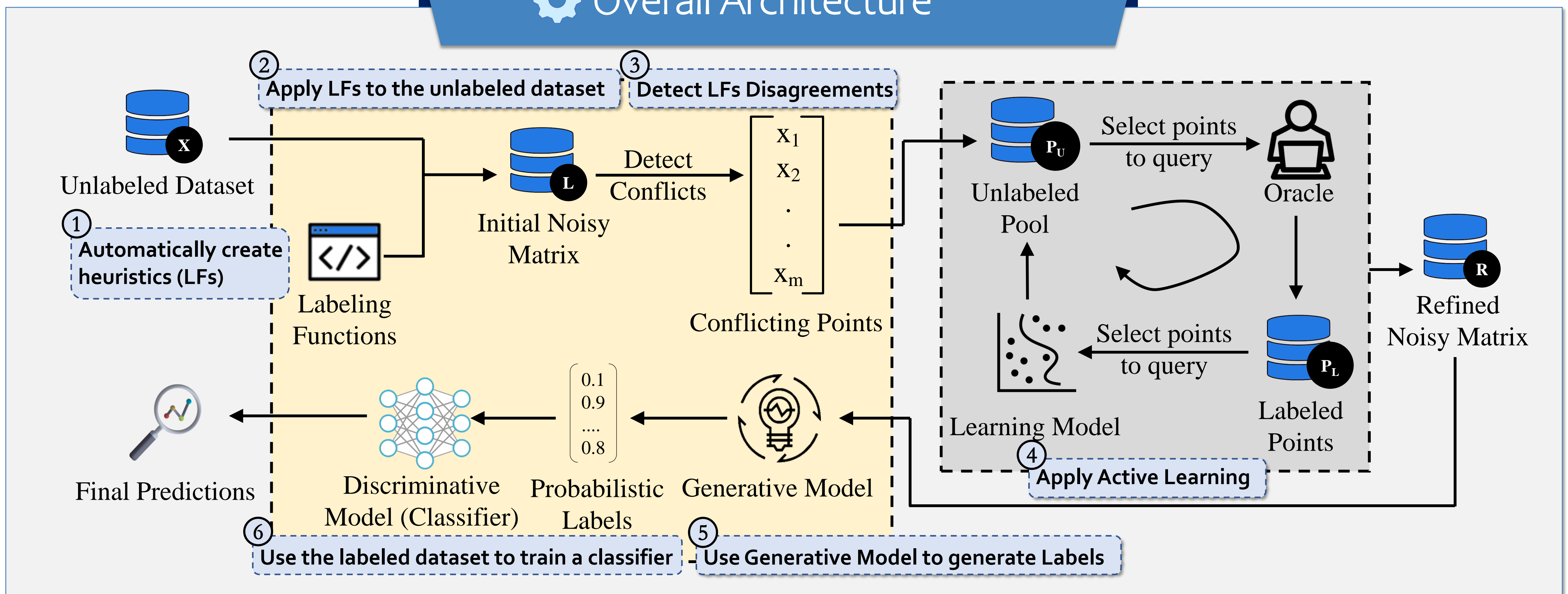
Guru Banavar, IBM data scientist

Proposed Solution: LaBiD

LaBiD: Labeler for Big Data



Overall Architecture

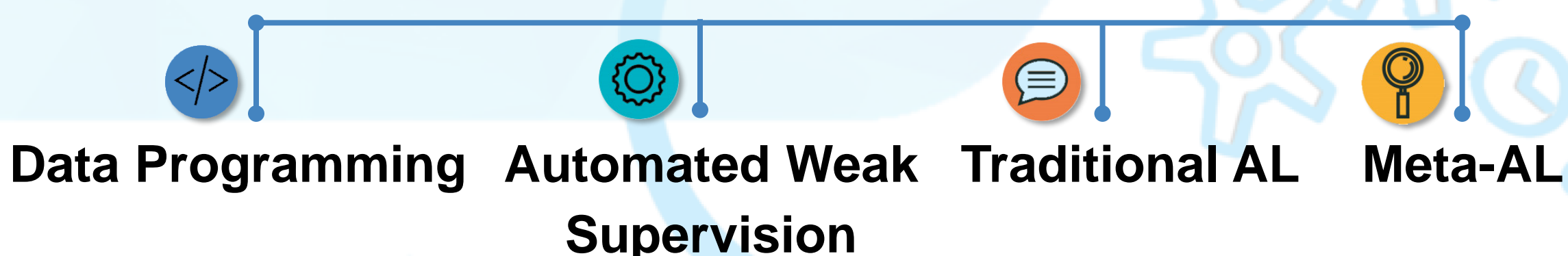


Experimental Results

1 Datasets

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

2 Existing Approaches



3 Results

