DataPrism: Exposing Disconnect between Data and Systems

Sainyam Galhotra, Anna Fariha, Raoni Lourenço, Juliana Freire, Alexandra Meliou, and Divesh Srivastava
sainyam@uchicago.edu

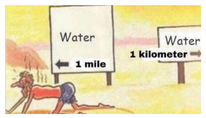
Motivation

Mismatch of pipeline's assumptions/expectations of input and the actual input data can cause malfunction

Input Dataset → Pipeline Implementation

90% Accuracy

30% Accuracy



Metric vs Imperial
KG vs lbs

Example Scenario

Age	Sex	Race	Marital Status	zip	Income
32	Female	White	Unmarried	14552	<50K
37	Female	Black	Married	94560	>50K
24	Male	Black	Married	01002	<50K
24	Male	White	Unmarried	64533	>50K

ML Model → Unfairness score 0.13

Age	Sex	Race	Marital Status	zip	Income
45	Female	Black	Unmarried	01004	<50K
33	Male	Black	Married	01014	>50K
23	Male	Black	Married	?	>50K
?	Male	White	?	94523	>50K

ML Model → Unfairness score 0.95

- The correlation of Male married individuals with high income causes unfairness in the ML model
- Fewer married females in second dataset causes ML model to ignore those tuples

Profile-Violation-Transformation (PVT)

- **Profile:** A property or a set of constraints that are satisfied by the tuples
  - Domain (Marital Status) = {Married, Unmarried}
  - Correlation(Sex, Income) = 0.63
- **Violation Function:** Quantifies violation of a dataset wrt a profile
- **Transformation Function:** Alter tuples to reduce violation

Interventional Approach

- Forcefully transform the dataset and check its effect on system malfunction
- Validates causality of a PVT

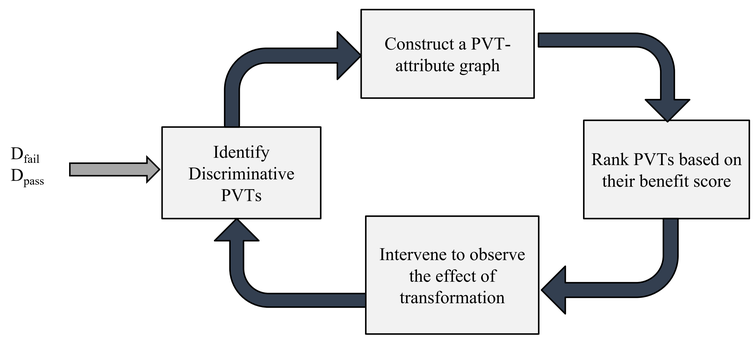
Age	Sex	Race	Marital Status	zip	Income
45	Female	Black	Unmarried	01004	0
33	Male	Black	Married	01014	1
23	Male	Black	Married	?	1
33	Male	White	?	94523	1

ML Model → Malfunction score: 0.73

Problem Statement

What are the minimal set of PVTs such that the transformed failing dataset $T(D_{fail})$ has malfunction score less than 0.2?

Our Approach



- Insights

  - Attributes with mistakes have multiple discriminative PVTs

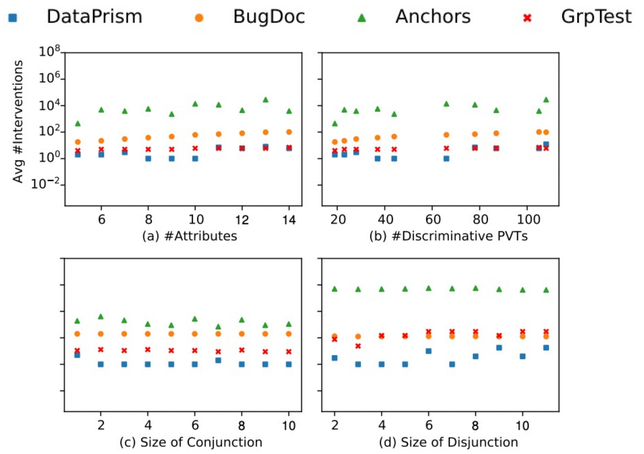
  - PVT for which the failing dataset incurs higher violation score is more likely to be a potential explanation of malfunction

  - A transformation function that affects a large number of data tuples is likely to result in a higher change in the malfunction score, after the transformation is applied.

Experiments

1. Runtime error in **entity linking** application
2. System crash of a **data visualization** tool
3. Functional dependency violation in a **data integration** pipeline
4. Data representation mismatch in **sentiment analysis** pipeline
5. Data unit mismatch in an **ML model**
6. Correlation of attributes in a **fairness-based** application

Case Study	Data Prism	BugDoc	Anchors	Group Testing
Sentiment	4	4	303	4
Income	**2**	20	103	10
Cardio vascular	**5**	100	3503	-
Flights	**5**	78	601	-
Amazon	**2**	8	303	4
Open data	**8**	14	102	-
Physicians	**30**	46	104	47



DataPrism grows sub-linearly with increase in number of discriminative PVTs