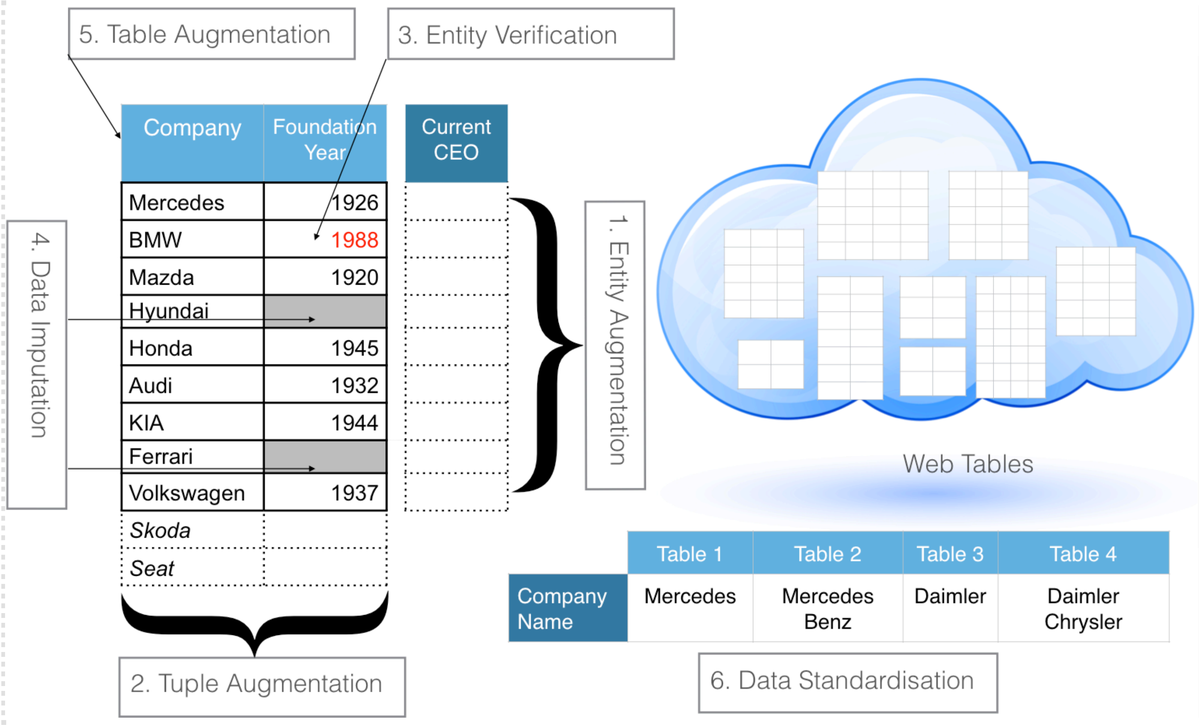
A Hybrid Imputation Approach Using Web Tables

Ahmad Ahmadov, Maik Thiele, Wolfgang Lehner, Robert Wrembel

What can be done to improve data quality?



Motivating Example Dataset

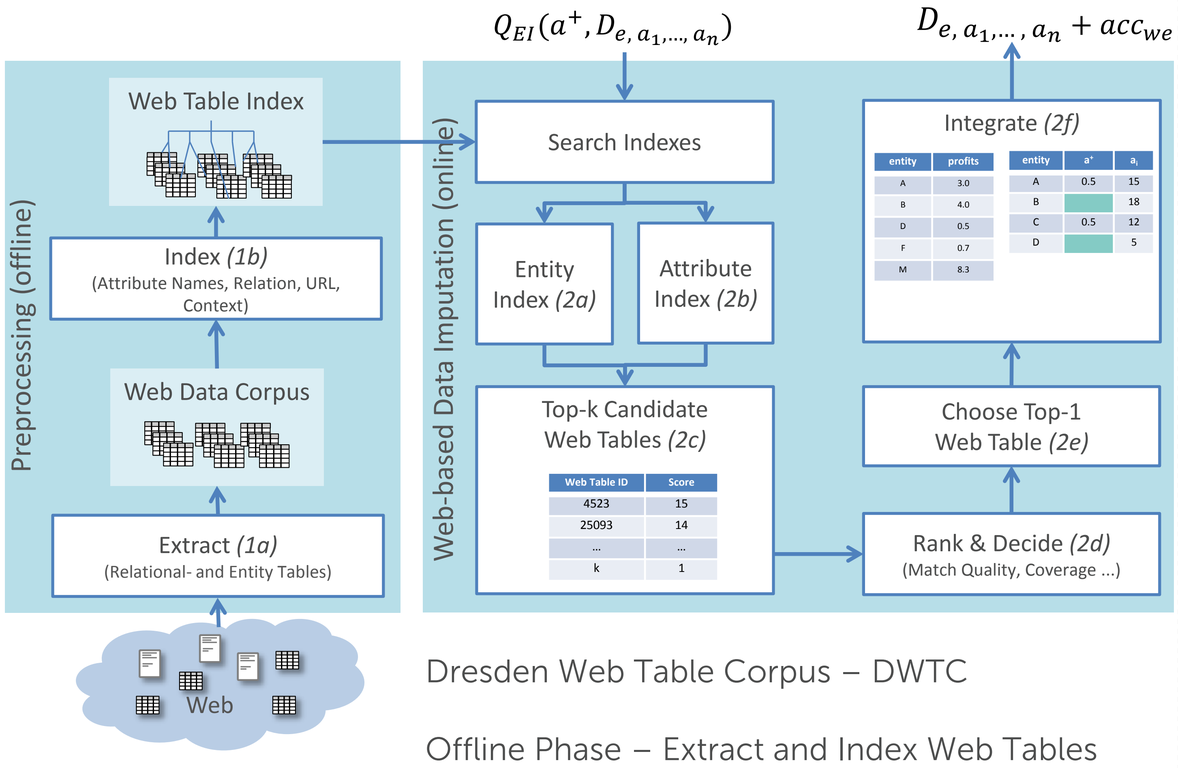
Missing data imputable: only by lookup, only by ML, by both ML and lookup

Company	Country	Industry	Sales	Profits	Assets	Marketvalue	Sales (in units)
Metro AG	Germany		54.12	0.47	22.94	14.38	23,000
AmerisourceBergen	USA	Health care	50.58		11.81	6.35	
Kroger	USA		53.23	1.04	20.61		30,000
Carrefour Group		Food markets	96.94		40.11	37.19	64,000
Tesco	UK	Food markets	41.48	1.49	25.9	33.99	
Costco Wholesale		Retailing	43.87		14.33	17.23	32,000
CVS	USA		26.59	0.85	10.54	14.81	8,000
McKesson	USA	Health care	66.45	0.61	15.95		43,000
Wal-Mart Stores	USA	Retailing	256.33	9.05	104.91	243.74	15,600

Data Lake

Web Tables

Web Based Imputation Approach

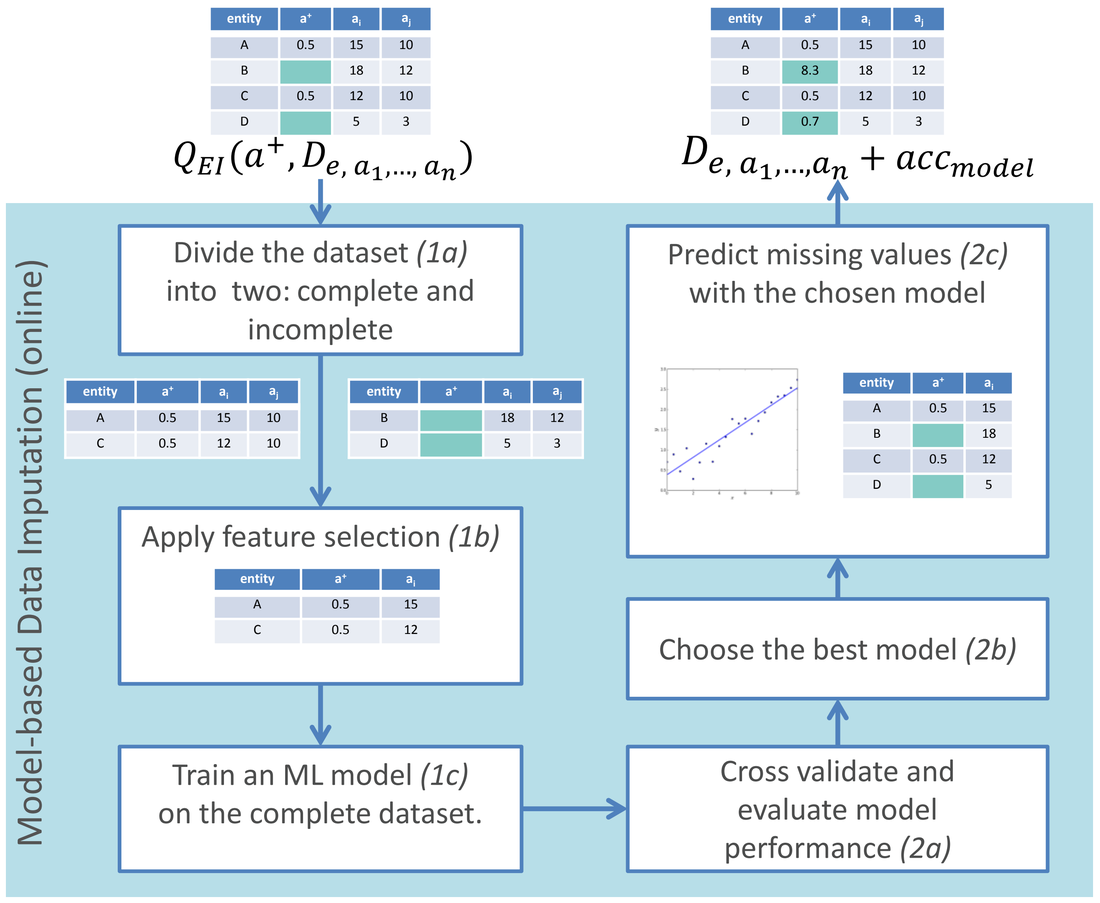


Dresden Web Table Corpus – DWTC

Offline Phase – Extract and Index Web Tables

Online Phase – Impute a dataset with missing values

Model Based Imputation Approach



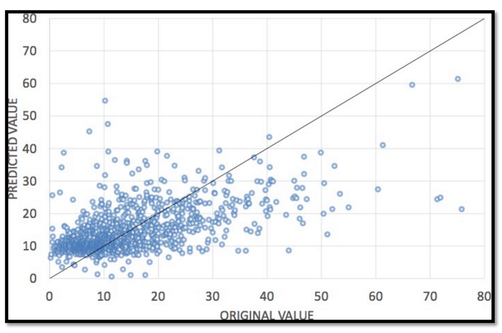
Hybrid Approach -Why Combine Machine Learning and External Lookup ?

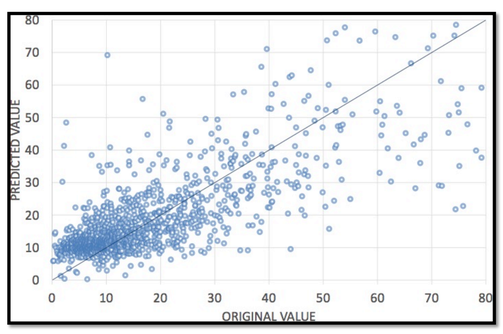
Using Hybrid Approach:

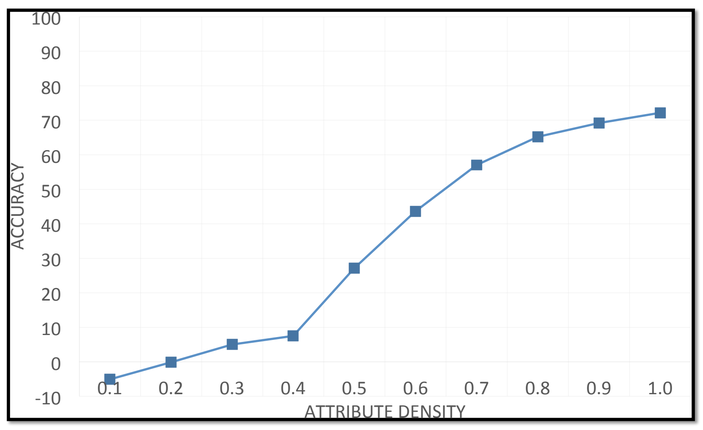
Companies Dataset: Significant increase of imputation accuracy by increasing coverage with the help of web data

Countries Dataset:

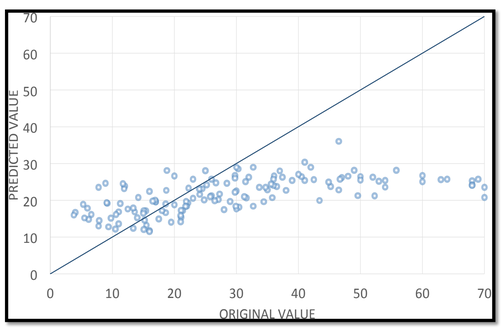
Slight increase in overall imputation accuracy using the hybrid approach.

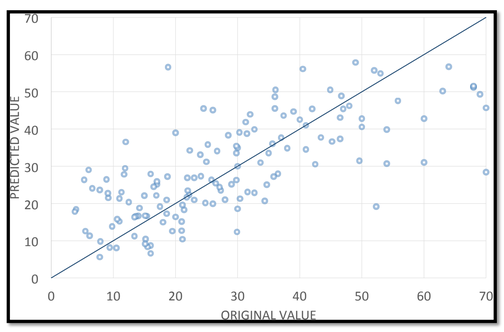


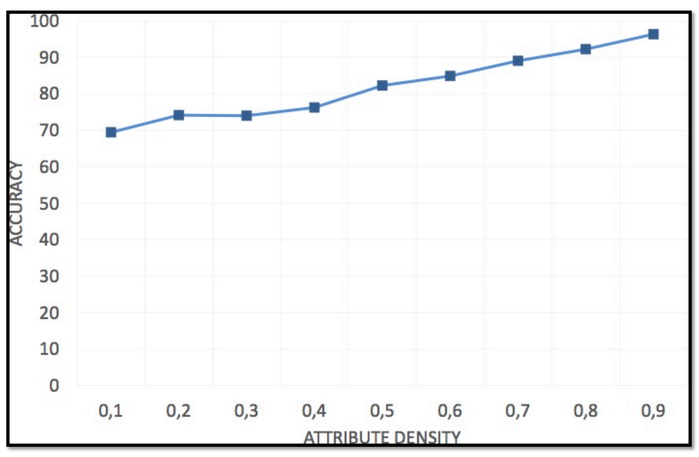




Companies Dataset – Impute Marketvalue







Countries Dataset – Impute Poverty