DataXFormer: A Robust Data Transformation System
Ziawasch Abedjan (MIT CSAIL), John Morcos (University of Waterloo), Ihab Ilyas (University of Waterloo), Mourad Ouzzani (QCRI), Paolo Papotti (QCRI), Michael Stonebraker (MIT CSAIL)

Example Use Case: Travel agencies
- Different date representations
- International date formats: DD/MM vs. MM/DD
- Location mapping:
- Price normalization

Data has to be transformed!

Many more transformations
- Syntactic: Can be computed by looking at the input values
- Semantic: Can be looked-up in third resource

Problem Statement
Find examples

Web tables
- Efficiently discovering transformation
  Examples:
  - Filter irrelevant tables
  - Overcome fragmentation
  - Dirty and Heterogeneous
  - Join tables for indirect transformations
  - Allow multiple columns as input values
  - Inductive EM model

DataXFormer
http://www.dataxformer.org

Knowledge Bases
- Infer types of examples
- Find patterns that connect examples
- Apply patterns on remaining input values

Web forms
- How to find Web forms?
- Use search engine
- How to use a Web form?
- Generate a wrapper
- Wrapping:
  - Parse the HTML and find request parameters

The crowd
- Evaluate transformations
- Solve conflicts
- Create transformations

Coverage of 120 queries:

Experimental setting:
- 120 queries
- 10 Input values per query
- Dataset: Dresden Web table Corpus
- 120 Million tables

Evaluation criteria:

Transformation quality:
Functional transformations:

Sampling strategies for supervision of non-functional transformations:

Contact author: abedian@csail.mit.edu