The Problem

- Processing knowledge from arbitrary web tables proves difficult when the information is frequently incomplete, ambiguous or references external information.
- Approaches to handling table ambiguity often rely on mapping table data to an external reference; however, this limits discovery to known concepts and requires taxonomy engineering.
- Assigning labels to table concepts without an external reference requires concept inference from the data itself.

The Goal

Develop a simple knowledge base from arbitrary tabular data that allows for discovery and question answering without requiring ontology engineering, external references, or manual manipulation.

Why Web Tables?

- Contain open domain knowledge, not limited to a specified schema—allows for discovery
- Can be parsed more reliably than unstructured text
- Frequently contain information of interest that is not available in surrounding text
- Table structure may provide insight into relations between entities
- The challenge: A table describing “smartphones” may never explicitly contain the label “smartphone”

Methods

- Treat each table structure as its own taxonomy
  - Table Structure | Taxonomy Feature
    | Table | Concept/Class
    | Headers | Class Attributes
    | Non-Header Rows | Instance of Class
    | Non-Header Cells | Instance of Class Attributes
    | Index Column (if present) | Label for Instance of Class
- Map taxonomy relations to graph structure

Results

- Table concepts can be inferred through analysis of is-A and has-A relationships implicit in the structure of relational tables, even when the concept is not explicitly referenced locally in the table
- Non-descriptive header labels are frequent in web tables (e.g. “Name: Giza” vs “City: Giza”)
  - When table information overlaps, descriptive labels can be inferred from other tables
- Relations extracted across tables lend to clustering of similar concepts (e.g. “City”, “Country”, “Location”)
- External is-A lookup resource is not needed for concept labeling if descriptive labels occur across multiple tables

Graph query for “Cairo” indicating that it is-A Name, Place, Capital, Metropolitan area, and an concept embodied by table 317.csv.

317.csv is inferred to be a Metropolitan area, a City, or a Location based on is-A relations extracted from index columns of other tables.

Graph data model manifests numerous concept label scoring methodologies for future evaluation