User-Centered Data Population of Knowledge Graphs

Fayaz Shaik, Justin Lubin, Sarah E. Chasins
Ohlone College, Fremont | PLAIT Lab, UC Berkeley

2021 Transfer-to-Excellence Research Experiences for Undergraduates Program (TTE REU Program)

Abstract: A knowledge graph is a data structure that describes structured relationships between entities. Knowledge graphs are widely used in artificial intelligence systems, but sometimes require data that can only be found on the internet in a semi-structured format such as a bulleted list. Unfortunately, semi-structured data can be difficult to parse, requiring users to write custom web-scraping programs for each web page to extract the necessary data. After presenting several ideas to domain experts for their opinions in a mock formative study, we built a prototype that prompts the user to highlight relevant information in a webpage, then uses these highlights as input-output examples for a custom program synthesizer that automatically generates web scraping scripts. These scripts are customized to each website and parse the semi-structured data into a tabular format easily transferable to a knowledge graph. Such a tool heightens the level of automation accessible to domain experts so that they may better leverage the vast amount of data available online for use in artificial intelligence systems.

AI can leverage data stored in **knowledge graphs**…
but writing code to scrape data from the web is **hard**.

What if it could be **automated**?

User-Centered Design

Live feedback

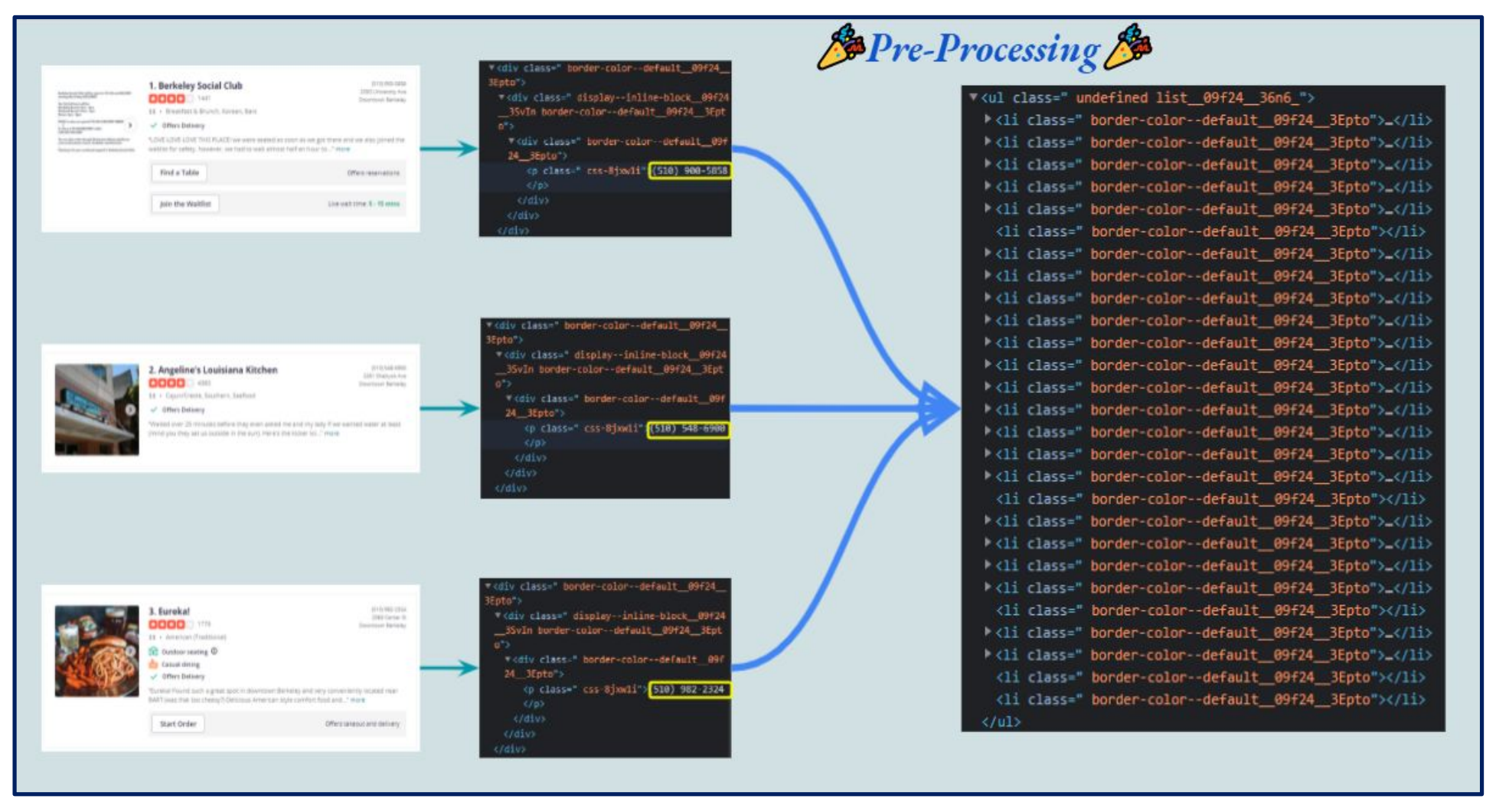
Modifiable table

Column approach

Program Synthesis Algorithm

```
Procedure Scrape(userHighlights):

  ancestor ←
    SmallestCommonAncestor(userHighlights)

  selector ←
    ExtractPath(userHighlights)

  Return Evaluate(selector, Children(ancestor))

End
```

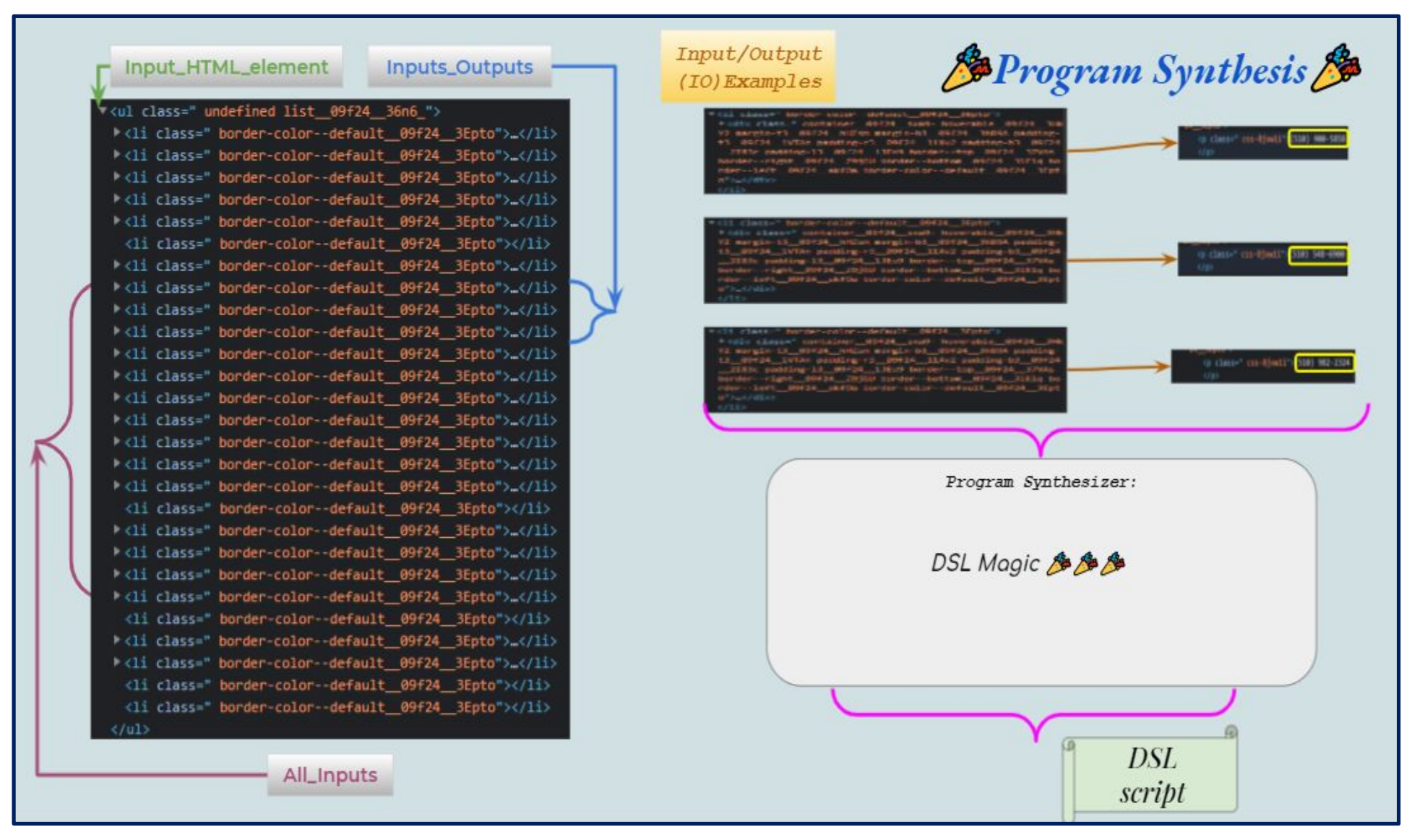
System Overview



Take websites

& HTML base

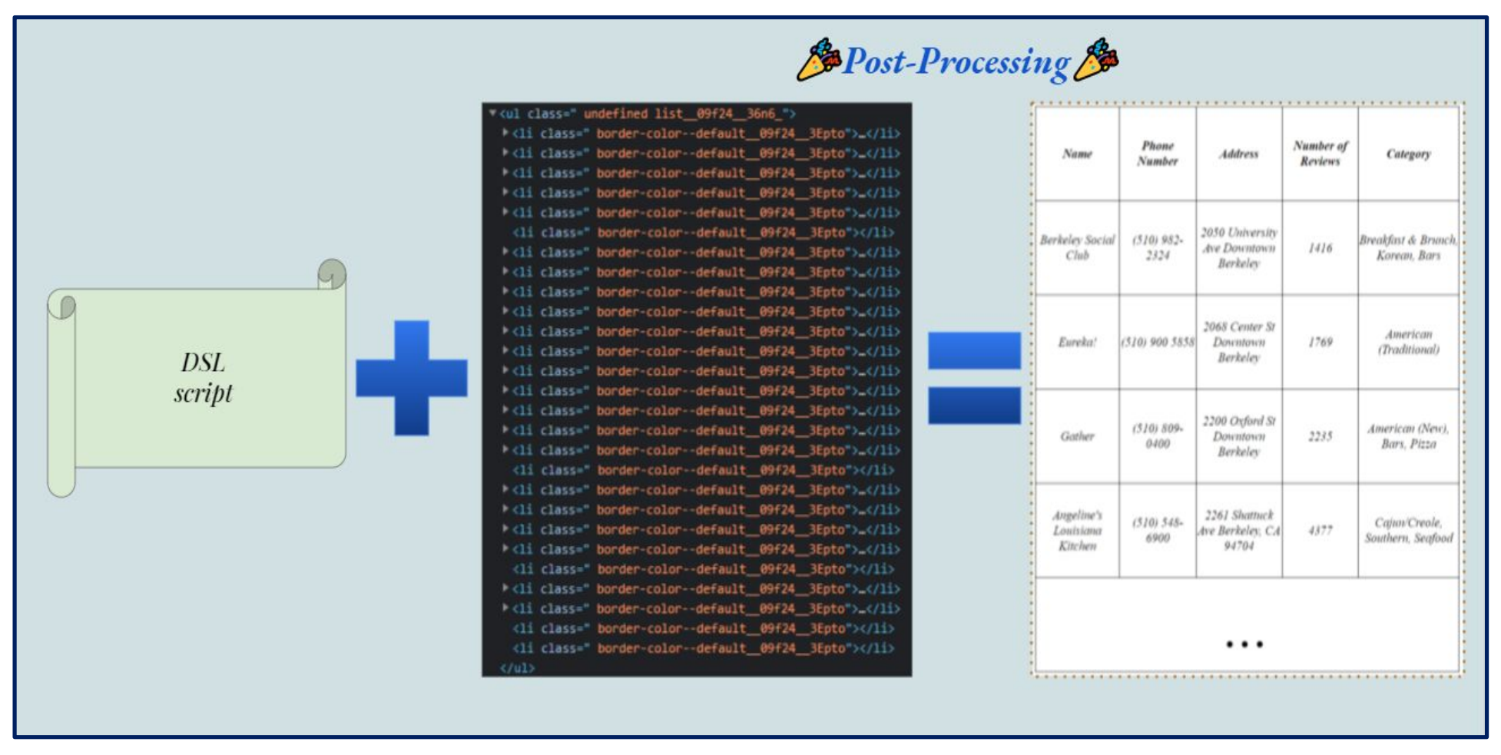
***GET* Ancestral list**



Input: HTML code

Program Synthesis

Output: DSL Script



DSL Script

+ HTML base

= Knowledge Graph

Support Information
This work was funded by National Science Foundation Award #1757690 and the Hopper Dean Foundation.

Contact Information
Email: fshaik3@student.ohlone.edu