Open Data for Tobacco Retail Mapping

Felicia Chen
felicia.chen@duke.edu
Nikhil Pulimood
nikhil.pulimood@duke.edu
James Wang
chenyang.wang@duke.edu

Project Manager: **Mike Dolan Fliss**
mike.dolan.fliss@gmail.com

Introduction

There is no national database of tobacco retailers.

- Only 37 states require licenses to sell tobacco.
- Tobacco products consist of 36% of sales revenue in convenience stores.
- There are weak incentives to obtain proper licensing

But having the knowledge of tobacco retailers' location is important.

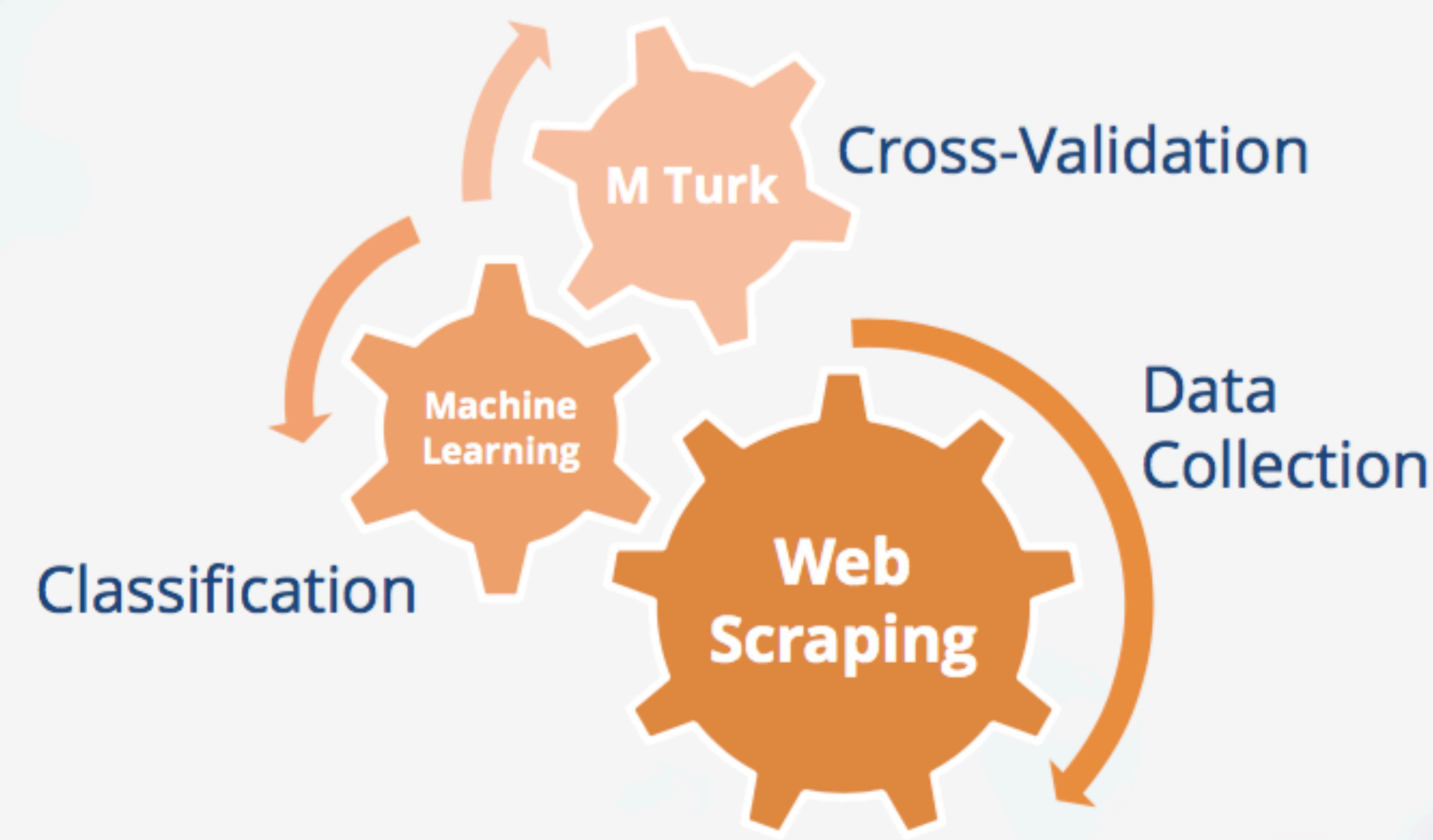
- Youth are more likely to begin smoking in areas with lots of tobacco retailers.
- The density of tobacco retailers correlates with many indicators of social disadvantage, including lack of healthcare.
- Regulations are often under enforced.

Objective

Evaluate novel techniques for building a tobacco retailer dataset.

- Web-scraping tobacco retailer locations.
- Machine learning to predict characteristics of retailers.
- Amazon Mechanical Turk as an inexpensive and accurate method to cross-validate data.

Method Overview

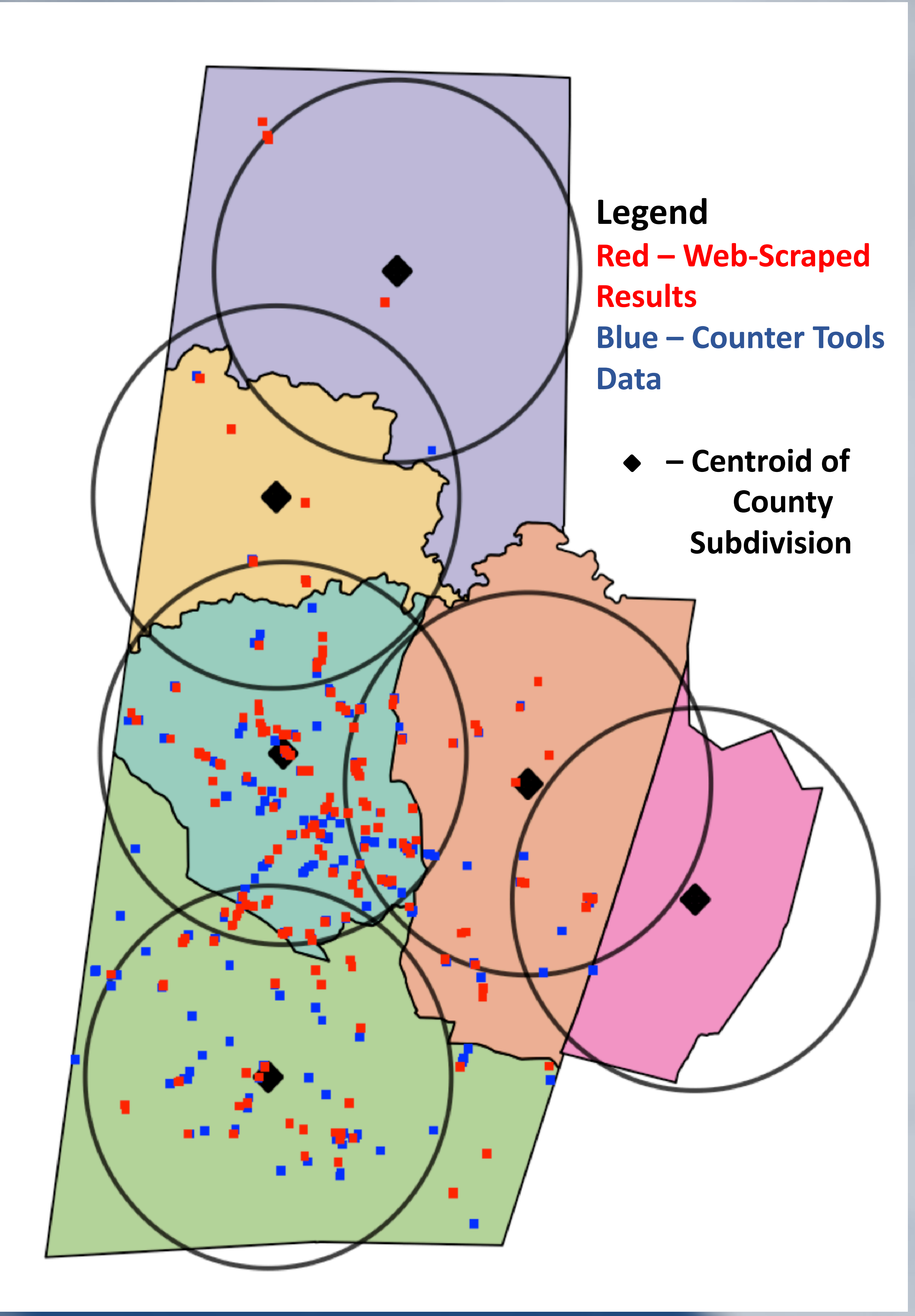


Web Scraping

In order to efficiently obtain a list of tobacco retailers, we looked to scrape data from webpages.

Used R to code an automated web crawler that parses HTML script

- Collected basic store information from Yellow Pages such as the store name, address, and phone number

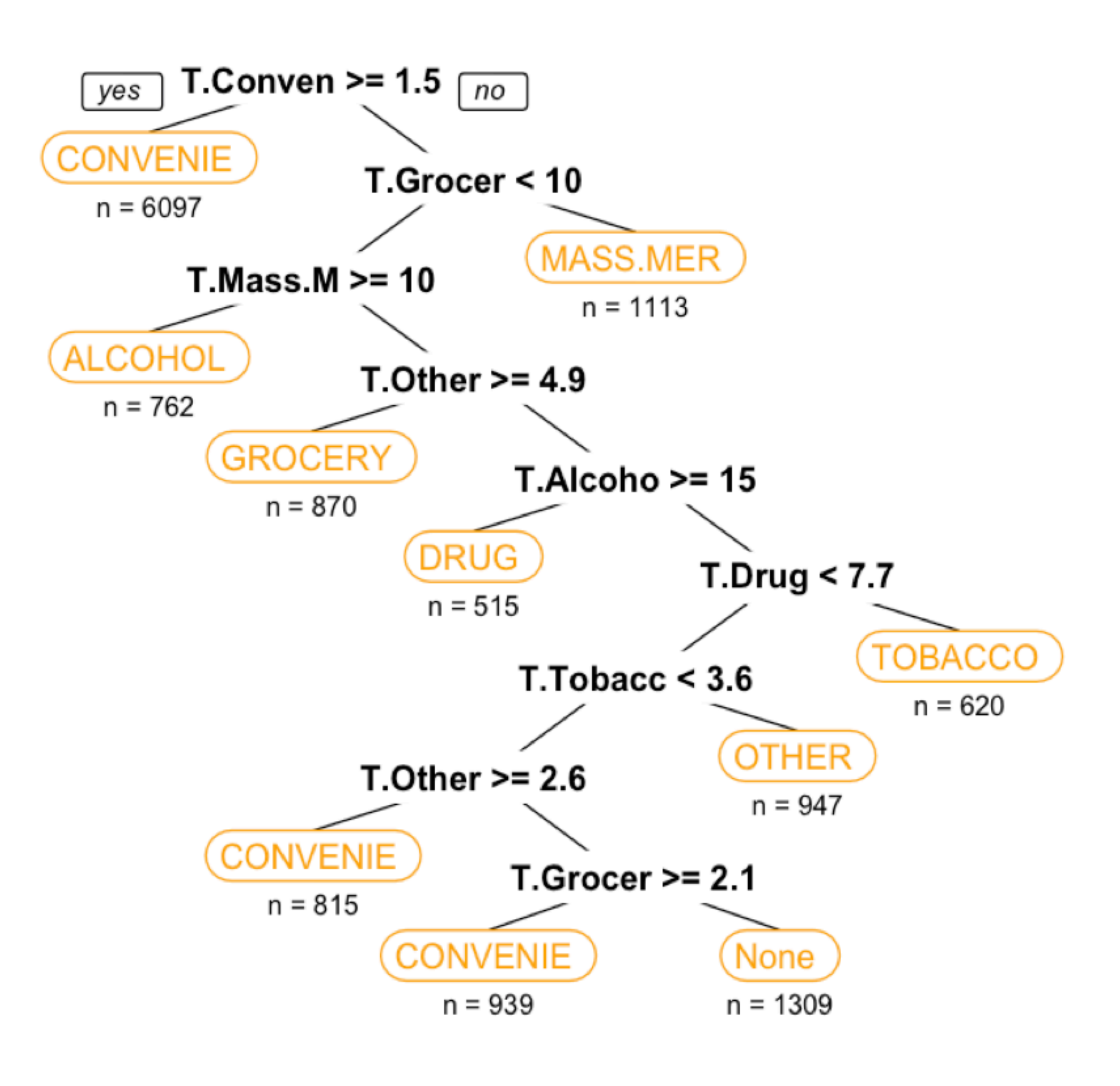


Web-Scraped Stores vs. Counter Tools Dataset in Durham County

Machine Learning

Our aggregated dataset contains many retailers.
But not all may actually sell tobacco products. The next step was predicting such characteristics of a store.

- Tokenized store names by breaking them down into n-grams. Calculated a modified version of the term frequency–inverse document frequency (tf-idf) score for each n-gram within each category.
- Used Jenks Natural Breaks to cluster tokens with similar scores together, and to determine which tokens were the best predictors for a store being in each category.
- Modeled a decision tree through R, where are training set was 70% of our data and our test set the other 30%.



Decision Tree for Store Type Classification

Results

- Aggregated 15,502 unique retailers in North Carolina, and 266 unique retailers in Durham County through web-scraping.
- Found that all 266 retailers matched the dataset of a community partner.
- Created and trained a decision tree using 19,619 retailers that were not in North Carolina, to predict the store types of 363 North Carolina retailers with an accuracy of 85.15%.

% accurately coded by text-mining machine learning methods

Original Coding on Store Visit	n stores	Alcohol	Convenience	Drug	Grocery	Hookah	Mass Merch.	None	Other	Tobacco	Vape
Alcohol	384	50%	30%		4%			8%	7%		
Convenience	2,818	0%	93%		1%		1%	2%	2%		
Drug	239		4%	82%				14%			
Grocery	600	1%	40%		49%		3%	3%	4%		
Hookah	2		50%					50%			
Mass Merch.	395		10%		1%		88%	1%			
None	1,020	3%	57%	2%	4%		3%	26%	6%		
Other	384	2%	47%		3%		3%	28%	18%		
Tobacco	138	1%	43%					50%	6%		
Vape	15		27%					73%			

Conclusion

- **Web-scraping** is the most effective method of data collection
- **Machine learning** with text mining is a relatively precise method for classification.
- **M Turk** is cost-effective for human cross-validation. It only costs $1.25 to validate a retailer.

Other Applications

Item	Web Scraping	Machine Learning	M Turk
Tobacco	All relevant stores	Classify store types using store names via text analysis	Cross-validate if a store sells tobacco
Produce	Stores that sell organic produce/ accept SNAP	Classify farmer markets, co-ops, grocery stores	Validating SNAP availability and food freshness
Overdoses	Surrounding retailers and establishments	Classify to predict areas that may be prone to incidents	