

Better NAICS Classifications with Text Mining of Public Business Data

Nathaniel Burbank

Civic Digital Fellow - Summer 2017

Big Data Center (CBDRA)

The views presented are those of the author and not necessarily the views of the U.S. Census Bureau. All data presented have been review to ensure privacy and confidentiality.

N. American Industry Classification System (NAICS)

- A hierarchical 6-digit industry coding system for business entities in North America.
- The first two digits represent one of 20 industry categories, while the remaining four digits specify one of 1000+ more detailed industry classifications.
- Since 1999, the Social Security Administration (SSA) has worked with the Census Bureau to assign NAICS codes to incoming Employer Identification Number (EIN) applications (which serve as a proxy for new business entity creation.)
- Beginning in 2004 the Census Bureau utilized the first version of an automated classifier to assign NAICS codes to a portion of incoming EIN applications based primarily on the business name and description fields from the submitted tax forms
- By 2007, more than 70% of new business entities were being automatically classified by this “auto-coder”
- By 2015, after some additional questions were added to the online version of the EIN application form, the auto-coder was being used for nearly 80% of EIN applications, with the **SSA classifying the remaining 20% by hand.**

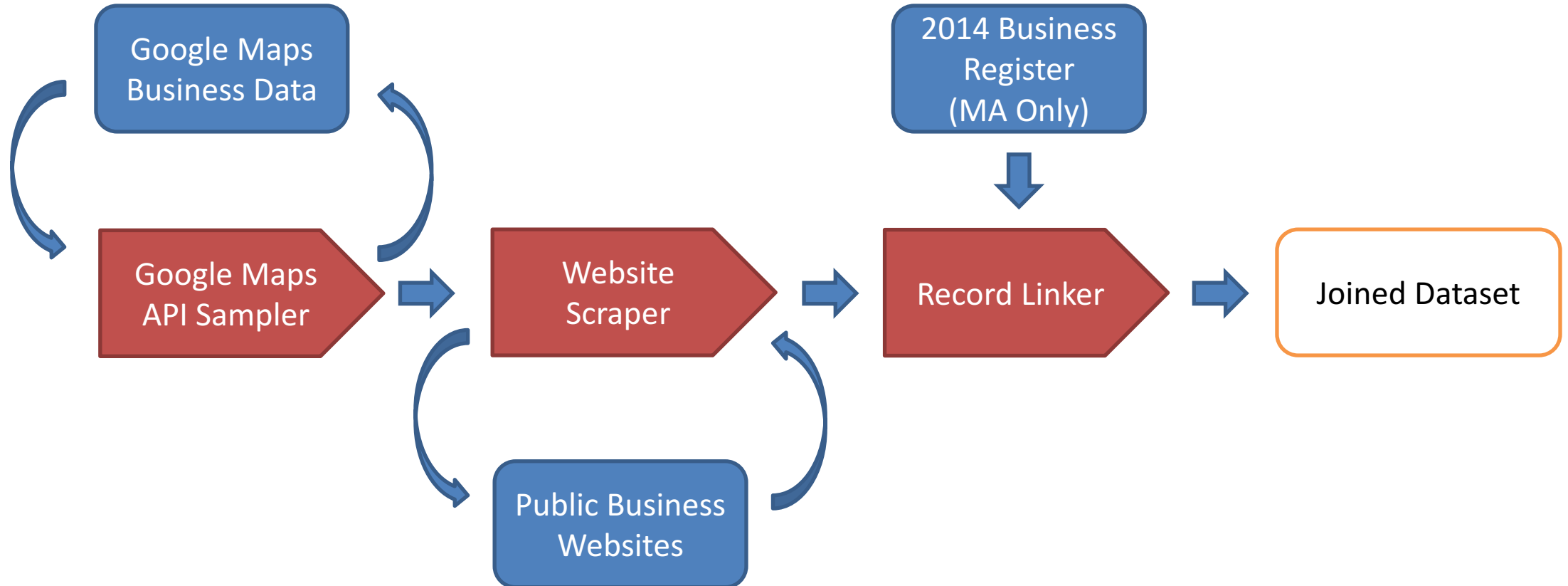
Challenges with NAICS classification

- Ground truth is nebulous – classifications from BLS, current auto coder and economic census only match ~84% of the time
- Trained human classifiers don't agree on which category a business entity should be classified as 10% to 15% of the time
- Existing SS4 derived prediction features are extremely limited

Central Question:

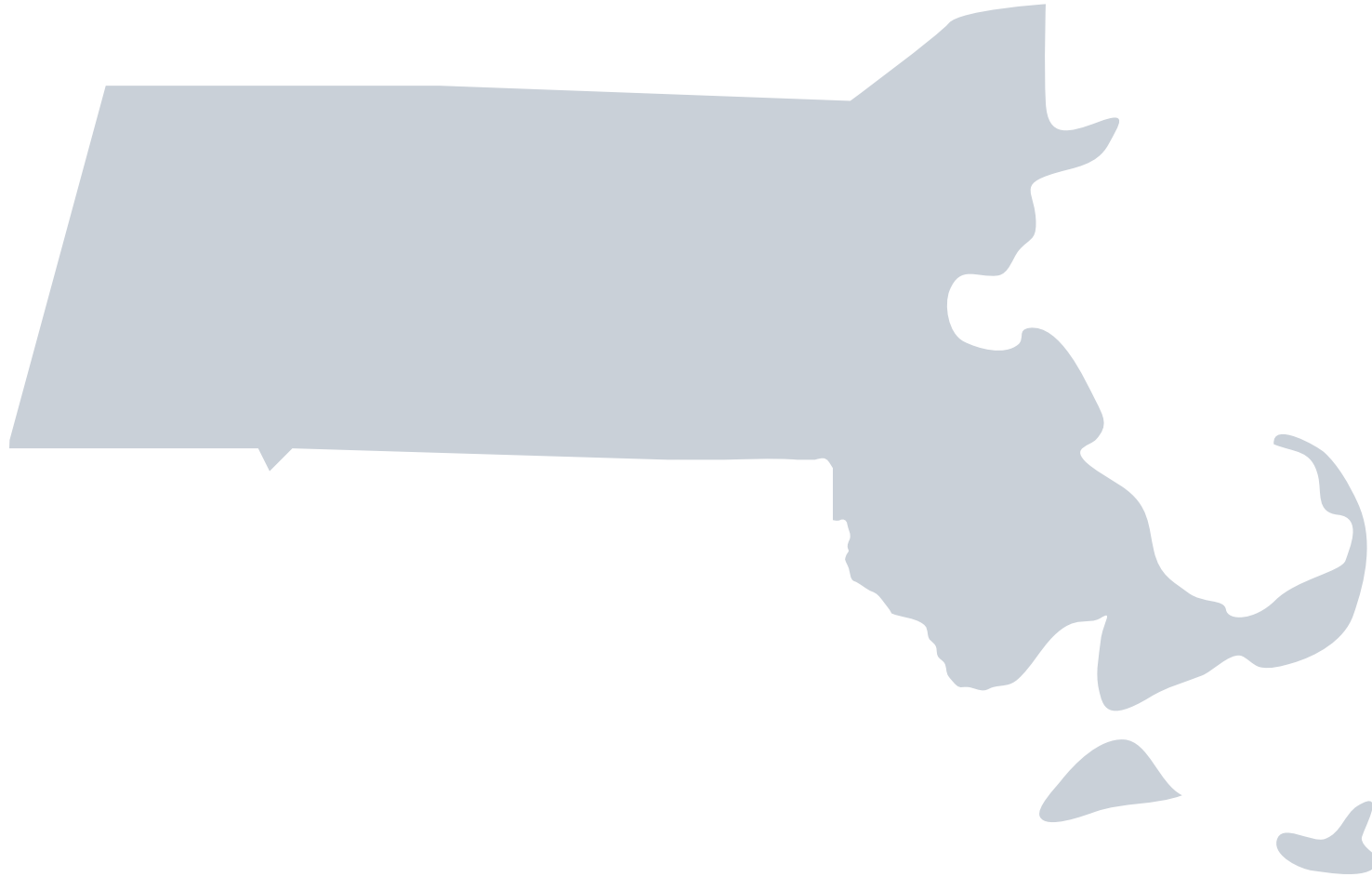
Can NAICS classification challenges be solved with better data?

Data Collection Overview

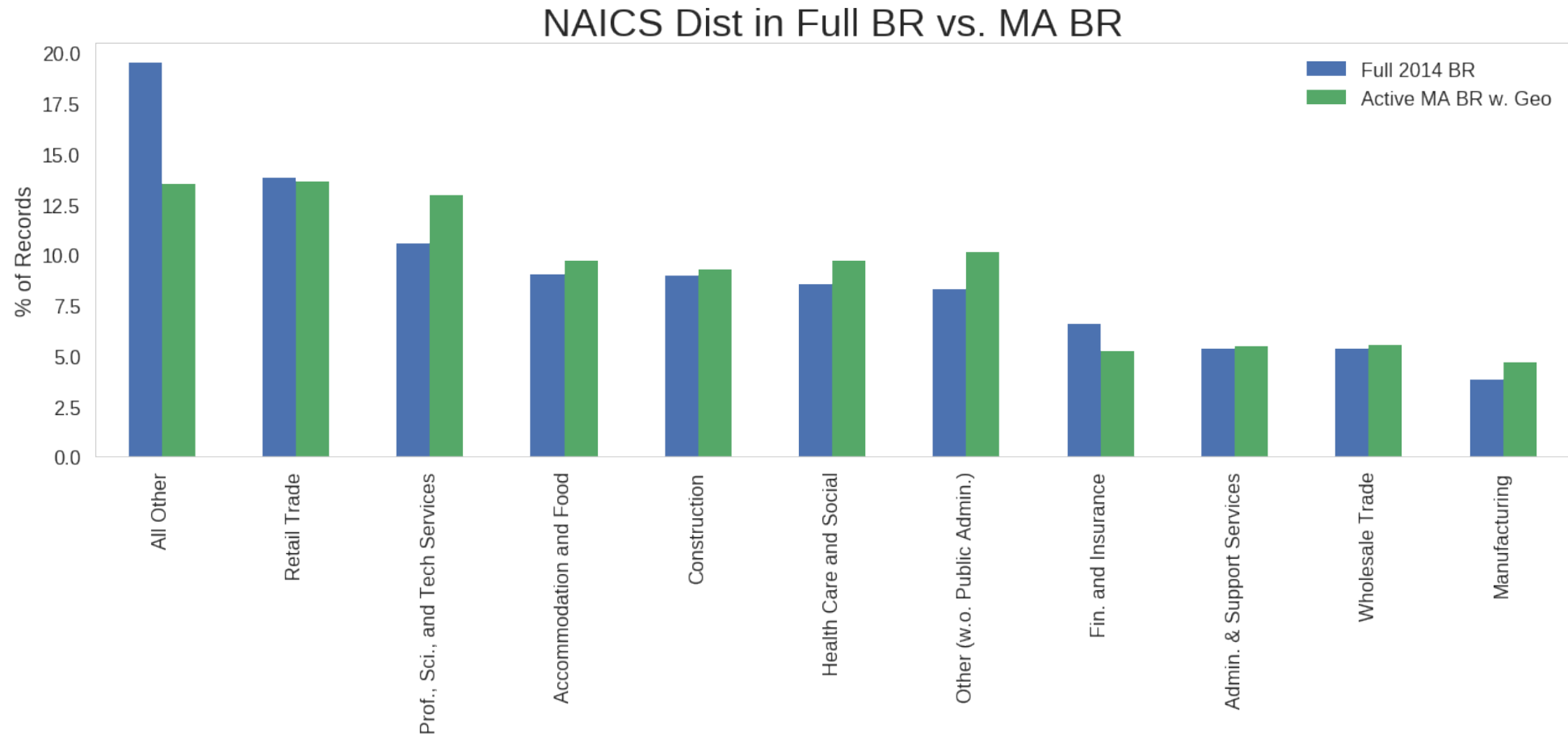


To avoid disclosing Title 13/Title 26 protected data, we build an entirely external-sourced dataset of business listings and then link those records with Business Register establishments

To make problem tractable within time constraints,
focused on Massachusetts exclusively



Distribution of Business entities by 2-digit NAICS category in United States Vs. Massachusetts



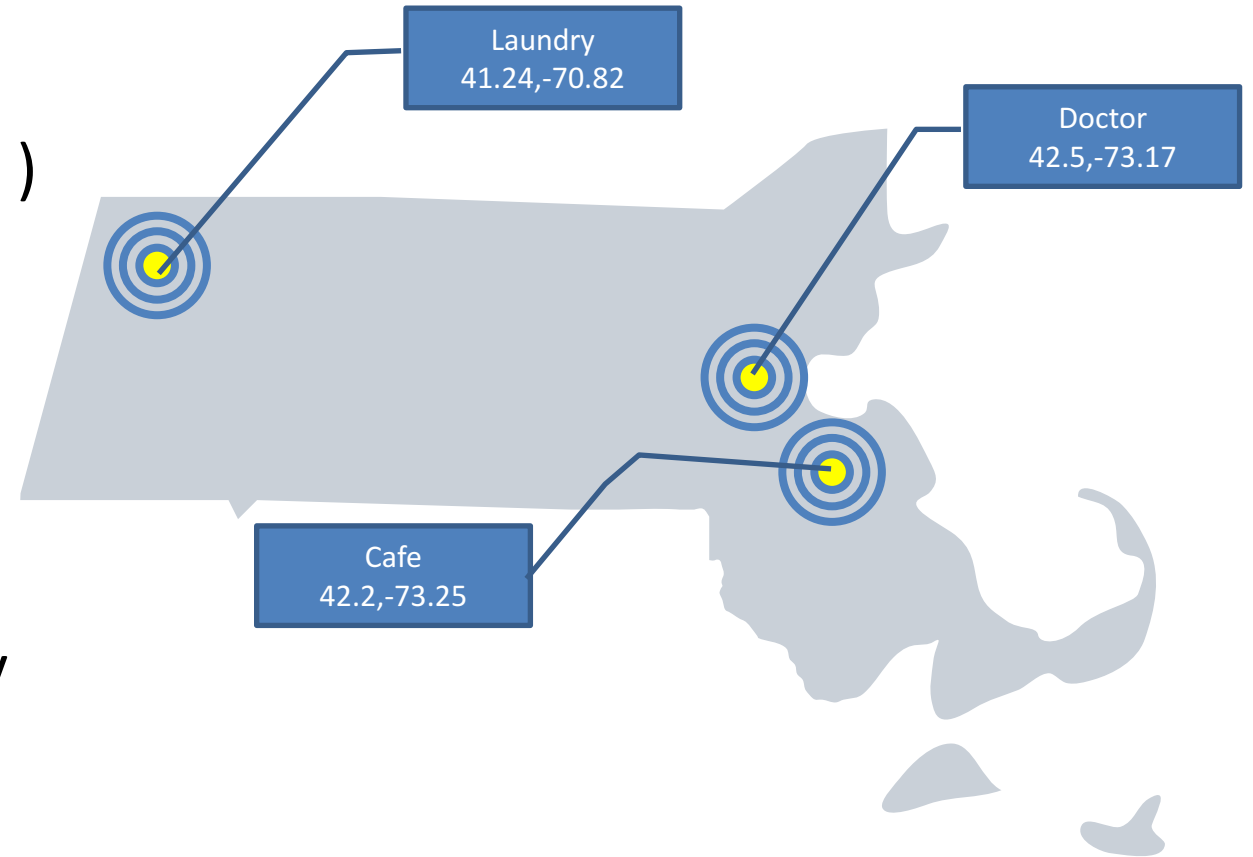
The distribution of establishments by 2-digit NAICS category is not particularly skewed in Massachusetts vs. the United States overall.

Why is Google Places API attractive for business classification?

- Comprehensive up-to-date dataset with uniform standards for business name, address, and geo coordinates
- JSON is easy to process and store
- Includes type tags (“Café”, “Doctor”, “Grocery”) and text of user reviews that could be useful for classification
- Provides lots of additional business metadata, such as phone numbers and website urls
- Free for non-profit use (up to a point)

Google Places API Limitations

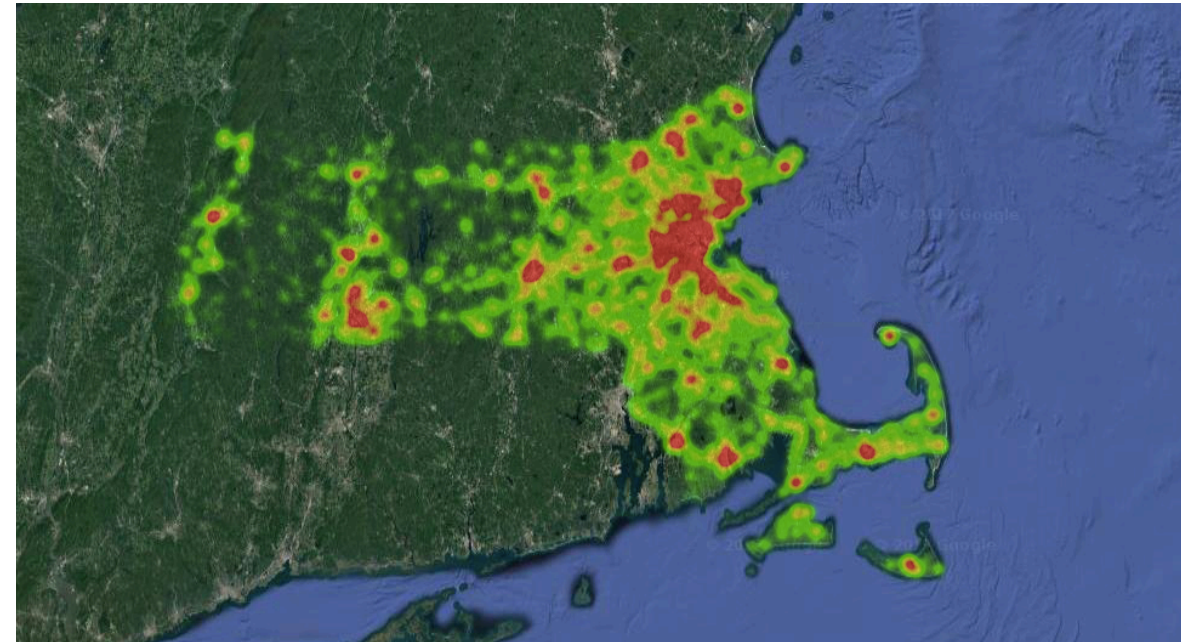
- Must supply a latitude/longitude coordinate pair and keyword
- Returns a maximum of 60 (nearest) results per query
- Not possible to specify already found entities
- Not possible to know when all places have been identified
- Limited to 150,000 queries per day (free tier)
- Takes up to 6 seconds per query



Google Places API Sampling

Google API Sampling Results

- Built a parallelized downloader
- Uniformly sampled lat/long locations within Massachusetts, with randomly selected keywords from list derived from NAICS classification guide
- Made 1.37 million queries between July 14th and July 31st
- Identified 224,000 unique place listings in Massachusetts




Website Scraper

224,000 Google Place entities identified

- URLs identified from Google maps
- HTML downloaded with parallelized scraper
- Visible text from home page extracted with BeautifulSoup Library (no links followed)
- Images, keywords, links, and other features discarded

With URLs	73.8%
With working website	66.5%
With (at least 1) user review	42.2%
With working website & reviews	31.24%
With working website or reviews	77.5%

Now we have two datasets...



224,000 Google
Maps entities in
Mass. w. text from
159,000 websites



202,000 active
entities in Mass.
with geo
coordinates from
2014 Business
Register

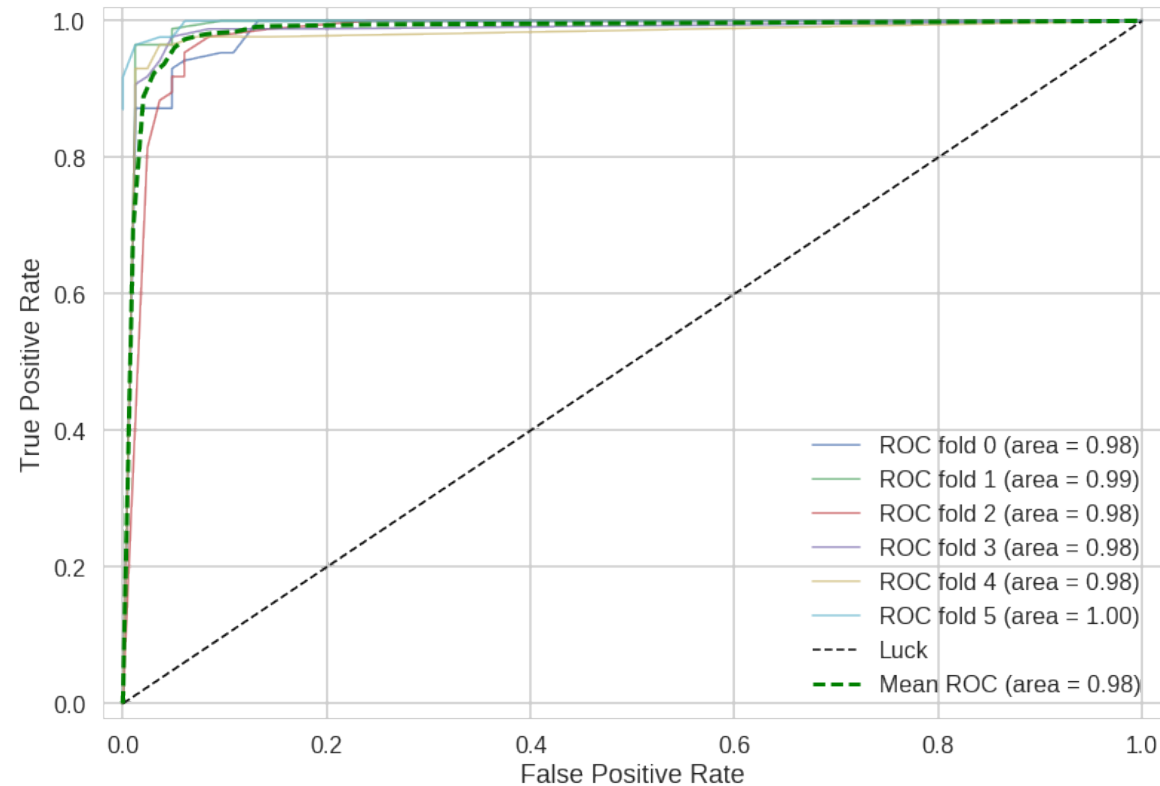
Challenges of matching names & addresses between Google Dataset and Business Register

- Cross product of possible matches between datasets is large
- No single unique identifier between datasets
- Matches are not necessarily 1-to-1 in either direction
- Name fields in Business Register are populated inconsistently

Challenges of matching names & addresses

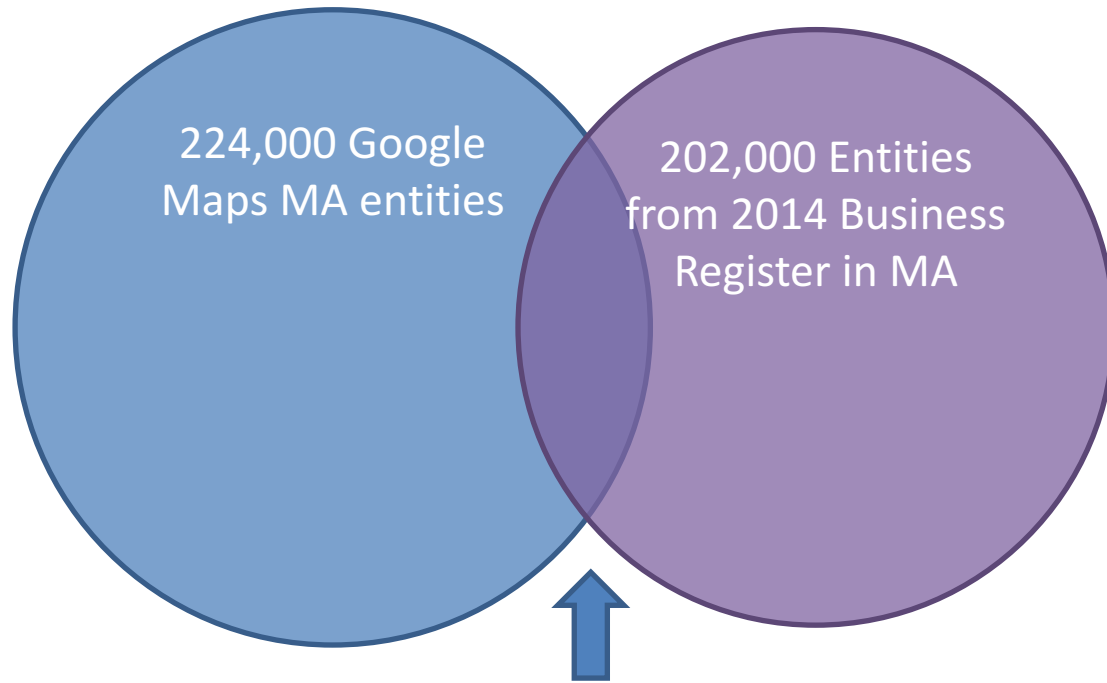
- Preprocess Business Register fields to correct name field difficulties and extract street number, street, and suite numbers into separate fields
 - (e.g., “24 Winterfell Street. Tower 3” -> “24”, “Winterfell St”, “3”)
- For each BR row, “block” on geo coordinates, to filter down to 15 closest Google results that might be a match
- Score string similarity of name, street number, street, and suite fields between datasets using Jaro Winkler and qgram distance algorithms)
- Hand-code 1000+ record pairs as matches or non-matches
- Train a Random Forest Classifier to identify true matches

Random Forrest Match Classifier ROC Curve



Because we would be making predictions off of the joined dataset, we wanted to avoid false-positive matches as much as possible. After training on human-coded data, we only marked rows as “matched” if the Random Forrest classifier predicted an 80% chance (or higher) of being a true match.

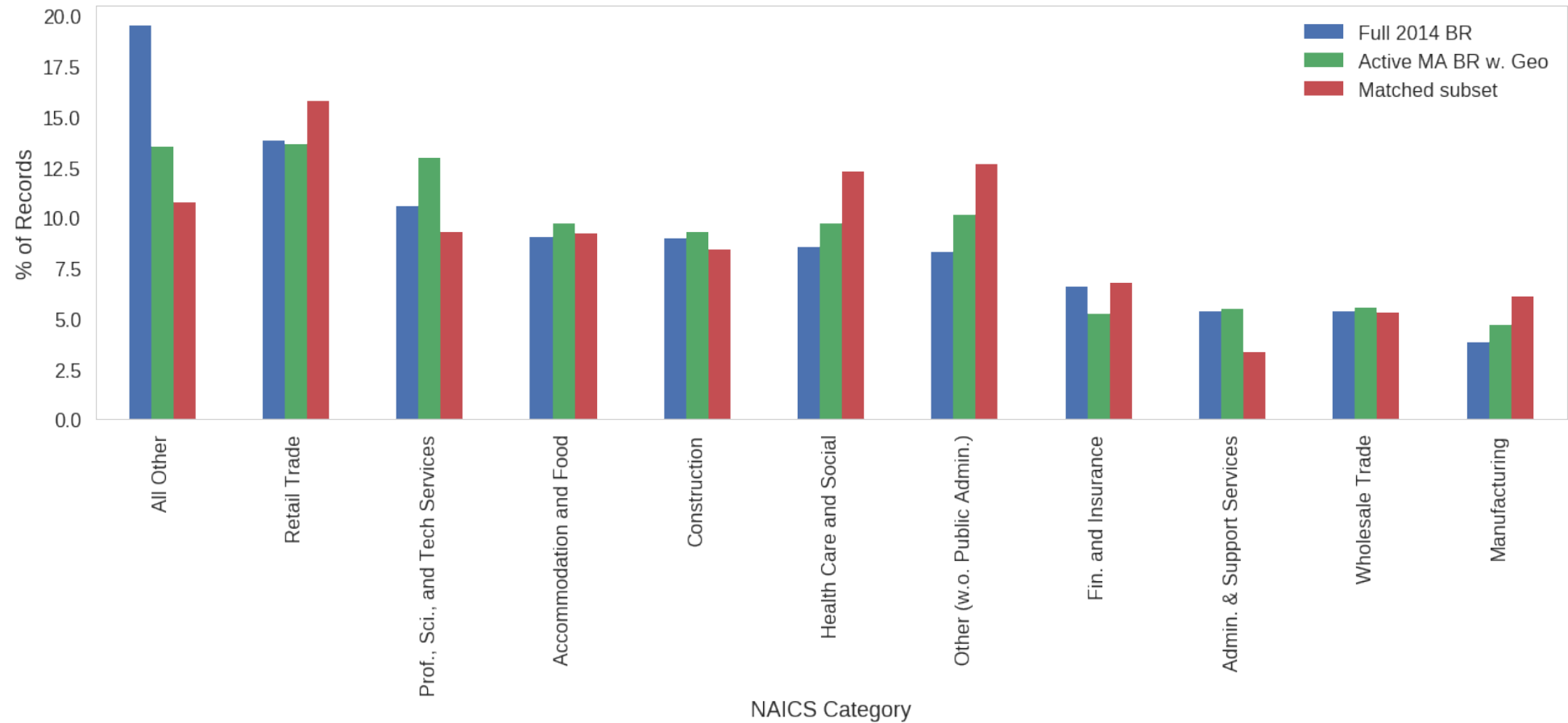
Matched Dataset



27,000 Matched Records

~27,000 Matched Records	
With user reviews	51.4%
With websites	71.6%
With websites & reviews	39.1%
With websites or reviews	83.9%

Distribution of Business entities by 2-digit NAICS category in U.S. vs. MA vs. Matched Subset



Potential Predictive Features

- Search Keywords
- Google Type Tags
- Business Name
- Text of user reviews (51.4%)
- Text of front page of business website (71.6%)
- Other metadata

Model 1: Linear SVC

- Features:
 - Name Keywords
 - Remove stop words (“inc.” “co.” etc.), and extract out 750 most common words in business name and flag if present)
 - Search Keywords
 - Google Type Tags
 - “Has website” flag

Model 1: Linear SVC

(Means from 5-fold Cross Validation)

	Obs.	Recall	Precision	F1 Score
Accommodation and Food	2500	0.906	0.895	0.900
Admin. & Support Services	900	0.540	0.655	0.590
Construction	2300	0.874	0.798	0.834
Fin. and Insurance	1800	0.932	0.926	0.929
Health Care and Social	3300	0.922	0.831	0.874
Manufacturing	1700	0.660	0.630	0.645
Other (w.o. Public Admin.)	3400	0.812	0.831	0.822
Prof., Sci., and Tech Services	2500	0.823	0.797	0.810
Retail Trade	4300	0.834	0.805	0.819
Wholesale Trade	1400	0.403	0.552	0.466
All Other	2900	0.614	0.883	0.711
Overall	27000	0.791	0.783	0.785

Based on the name and search keyword features alone, the first model actually does pretty well, classifying 79.1% of the records into the correct 2-digit NAICS category overall.

Model 2: Multinomial Naïve Bayes

- Features:
 - Concatenated user reviews and website text into single string
 - (83.9% of subset has at least one of these)
- Pre Processing:
 - Remove stop words
 - Tf-idF Vectorizer
 - No stemming

Model 2: Multinomial Naïve Bayes

(Means from 5-fold Cross Validation)

	Obs.	Recall	Precision	F1 Score
Accommodation and Food	2400	0.909	0.794	0.848
Admin. & Support Services	750	0.384	0.762	0.510
Construction	1400	0.789	0.625	0.697
Fin. and Insurance	1700	0.865	0.954	0.907
Health Care and Social	2700	0.901	0.830	0.864
Manufacturing	1400	0.582	0.649	0.613
Other (w.o. Public Admin.)	3000	0.739	0.719	0.729
Prof., Sci., and Tech Services	2000	0.769	0.675	0.719
Retail Trade	3800	0.726	0.755	0.741
Wholesale Trade	1100	0.455	0.544	0.496
All Other	2500	0.555	0.822	0.648
Overall	23000	0.734	0.731	0.726

On the subset of data for which we have free-form text features, the Naïve Bayes classifier correctly predicts the 2-digit NAICS category of 73.4% of the time.

Model 2: Multinomial Naïve Bayes

Top 10 most important words per NAICS category

NAICS Cat.	Top 10 Words
Accommodation and Food	coffee chicken order restaurant menu place good great pizza food
Admin. & Support Services	security maintenance business contact travel job company cleaning service services
Construction	contact heating company commercial home project work projects services construction
Fin. and Insurance	savings personal checking credit financial loans business bank banking insurance
Health Care and Social	office center children patients patient services dr health dental care
Manufacturing	design metal product contact company printing quality manufacturing custom products
Other (w.o. Public Admin.)	place work services funeral hair great auto repair service car
Prof., Sci., and Tech Services	estate attorney business legal contact firm clients tax services law
Retail Trade	place products new prices good selection shop great service store
Wholesale Trade	services sales product company new contact service parts equipment products

Model 3: Ensemble Random Forrest

Goal with Ensemble model is the combine the predictive power of Models 1 and 2.

- Features:
 - All Bernoulli features from Model 1 (search keywords etc.)
 - Probabilities from Naïve Bayes model trained on features from Model 1
 - Probabilities from Model 2 (zeroed for observations without text features)

Model 3: Ensemble Random Forrest

(Means from 5-fold Cross Validation)

	Obs.	Recall	Precision	F1 Score
Accommodation and Food	2500	0.896	0.901	0.899
Admin. & Support Services	900	0.611	0.660	0.633
Construction	2300	0.860	0.821	0.840
Fin. and Insurance	1800	0.911	0.934	0.922
Health Care and Social	3300	0.909	0.879	0.894
Manufacturing	1700	0.676	0.647	0.661
Other (w.o. Public Admin.)	3400	0.810	0.834	0.821
Prof., Sci., and Tech Services	2500	0.825	0.830	0.827
Retail Trade	4300	0.822	0.836	0.829
Wholesale Trade	1400	0.524	0.536	0.530
All Other	2900	0.641	0.874	0.731
Overall	27000	0.796	0.797	0.796

The Ensemble Model achieves an overall accuracy of 79.6%. This is only slightly improved overall over the Model 1, but it makes some key gains in hard to predict categories such as Wholesale Trade.

Model Comparison -- Overall Accuracy (Recall) when Predicting 2-Digit NAICS Categories

	Obs.	LinearSVC	MultinomialNB	Ensemble RF
Accommodation and Food	2500	0.906	0.909	0.896
Admin. & Support Services	900	0.540	0.384	0.611
Construction	2300	0.874	0.789	0.860
Fin. and Insurance	1800	0.932	0.865	0.911
Health Care and Social	3300	0.922	0.901	0.909
Manufacturing	1700	0.660	0.582	0.676
Other (w.o. Public Admin.)	3400	0.812	0.739	0.810
Prof., Sci., and Tech Services	2500	0.823	0.769	0.825
Retail Trade	4300	0.834	0.726	0.822
Wholesale Trade	1400	0.403	0.455	0.524
All Other	2900	0.614	0.555	0.641
Overall	27000	0.791	0.734	0.796

Potential Next steps

- Link with SS4 predictive features to investigate joined predictive performance
- Extracting out keywords associated with different NAICS categories
- More sophisticated text mining and stemming
- More sophisticated web crawling
- Investigate model accuracy at more detailed levels of NAICS classification
- Build larger dataset with including other states

Thank you!