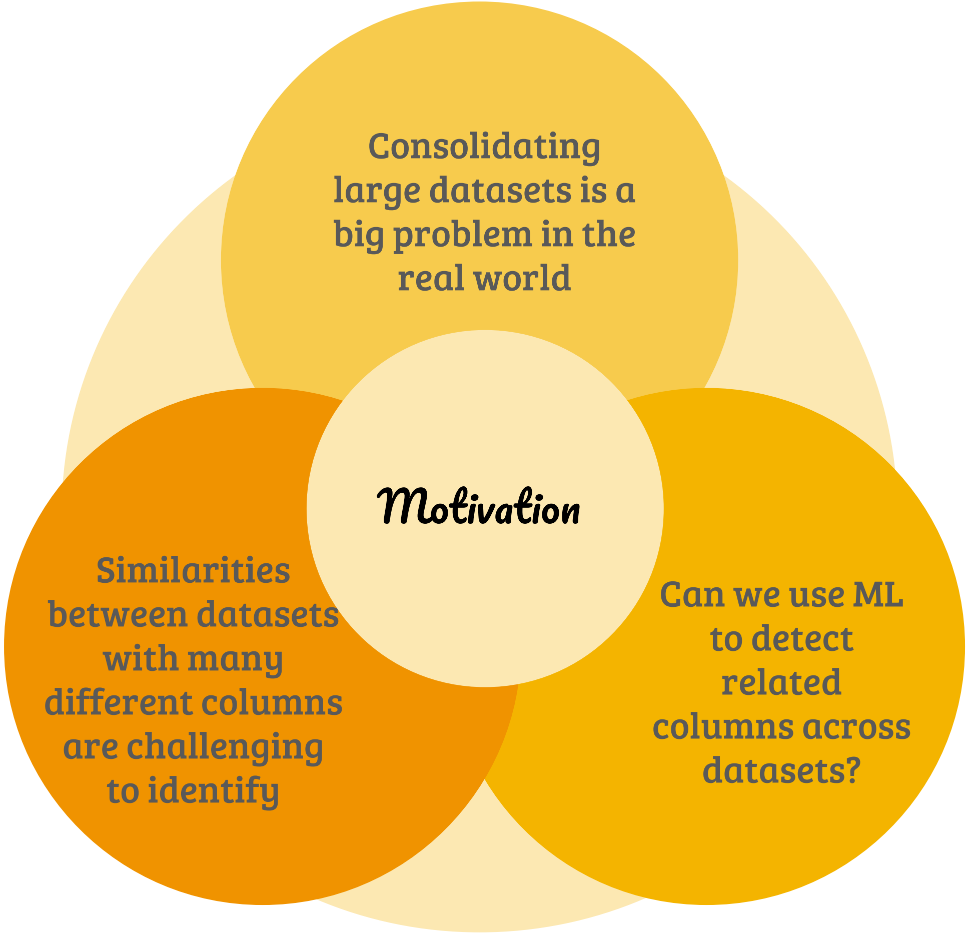
MnM

Match ‘n’ Merge!

IEOR 290: Data-X
“Fuzzy Joins” Group 1
December 4, 2017

Jordan Cox
Rob Holbrook
Ramon Lim
Stephanie Zhu



Motivation

Consolidating large datasets is a big problem in the real world

Similarities between datasets with many different columns are challenging to identify

Can we use ML to detect related columns across datasets?

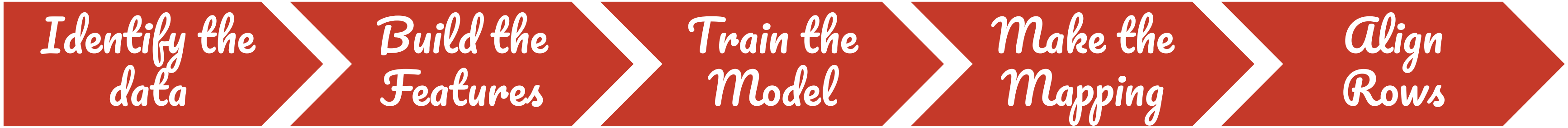
The Use Case

Facebook acquired Instagram in 2012. We can imagine Facebook may have had a difficult time integrating Instagram's years of data into their own system! Our tool can help.

Over 1,009 acquisitions occurred in 2016, worth $3.7 trillion dollars*

*Source: SeekingAlpha

The Approach



Identify the data

User uploads any two data sets to merge into the tool through a web portal. Data is loaded and processed.

Build the Features

Each column of both data set is analyzed and profiled based on a standard set of features

- Total length
- # letters
- # digits
- # whitespaces
- # punctuations
- % letters
- % digits
- % whitespaces
- % punctuations

Train the Model

Machine learning is deployed to train data set 1 column predictions based on features. Models applied to data set 2 to predict most likely similarities

Make the Mapping

Mappings are displayed and validated by the user to ensure proper matches

Align Rows

Rows from each dataset are mapped based on Ratcliff/Obershelp pattern recognition for string similarities. Rows are aligned for user review

Demo

Future improvements

- Improve code efficiencies for row mapping
- Create more features (feature engineering)
- Use machine learning to determine feature engineering
- Predict confidence % of prediction in addition to relative accuracy
- Fill in missing data values using inferential statistics (fuzzy joins)
- Expand the types of data set the model can merge

Our learning journey

Establishing the idea

How do we take the “Fuzzy Joins” idea and flip it on its head?

Learning Python, Pandas, Sklearn

Learn Python & all its dataframe and machine learning packages.

Training, testing, tweaking

Which datasets do we use? Which models should we choose? How do we implement the column match?

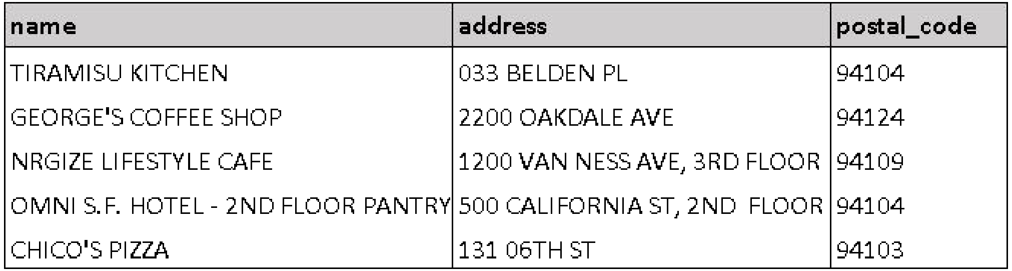
Presentation

Show what we’ve learned through the semester

Appendix

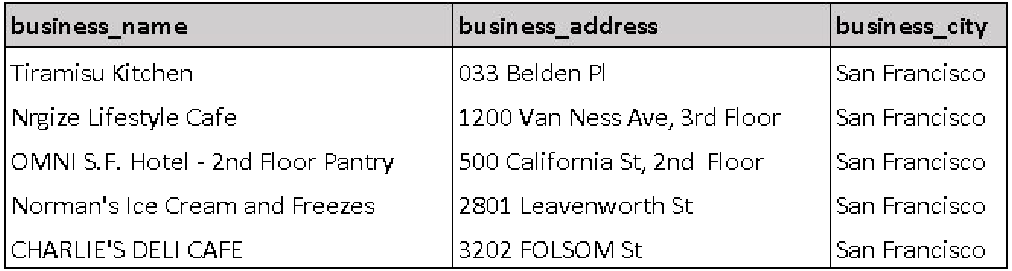
Choosing the datasets

Data Set A (Training Set)



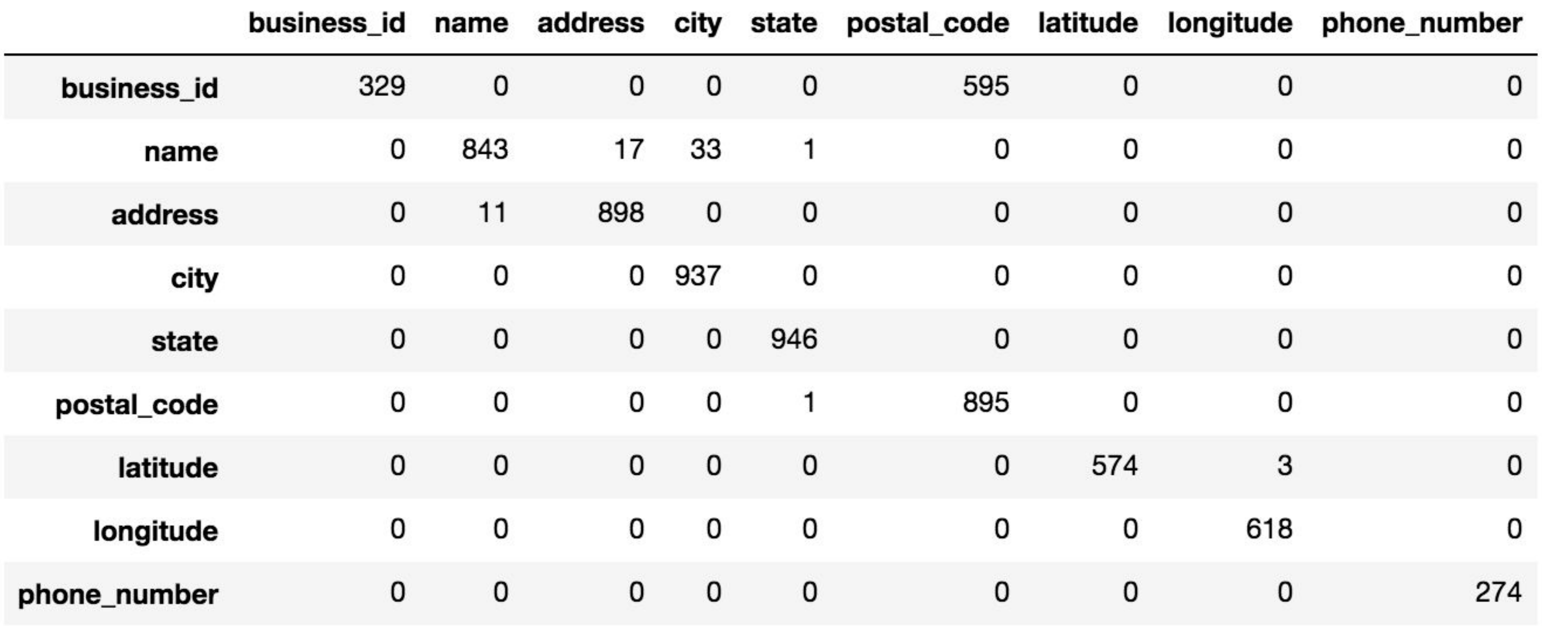
name	address	postal_code
TIRAMISU KITCHEN	033 BELDEN PL	94104
GEORGE'S COFFEE SHOP	2200 OAKDALE AVE	94124
NRGIZE LIFESTYLE CAFE	1200 VAN NESS AVE, 3RD FLOOR	94109
OMNI S.F. HOTEL - 2ND FLOOR PANTRY	500 CALIFORNIA ST, 2ND FLOOR	94104
CHICO'S PIZZA	131 06TH ST	94103

Data Set B (Validation Set)



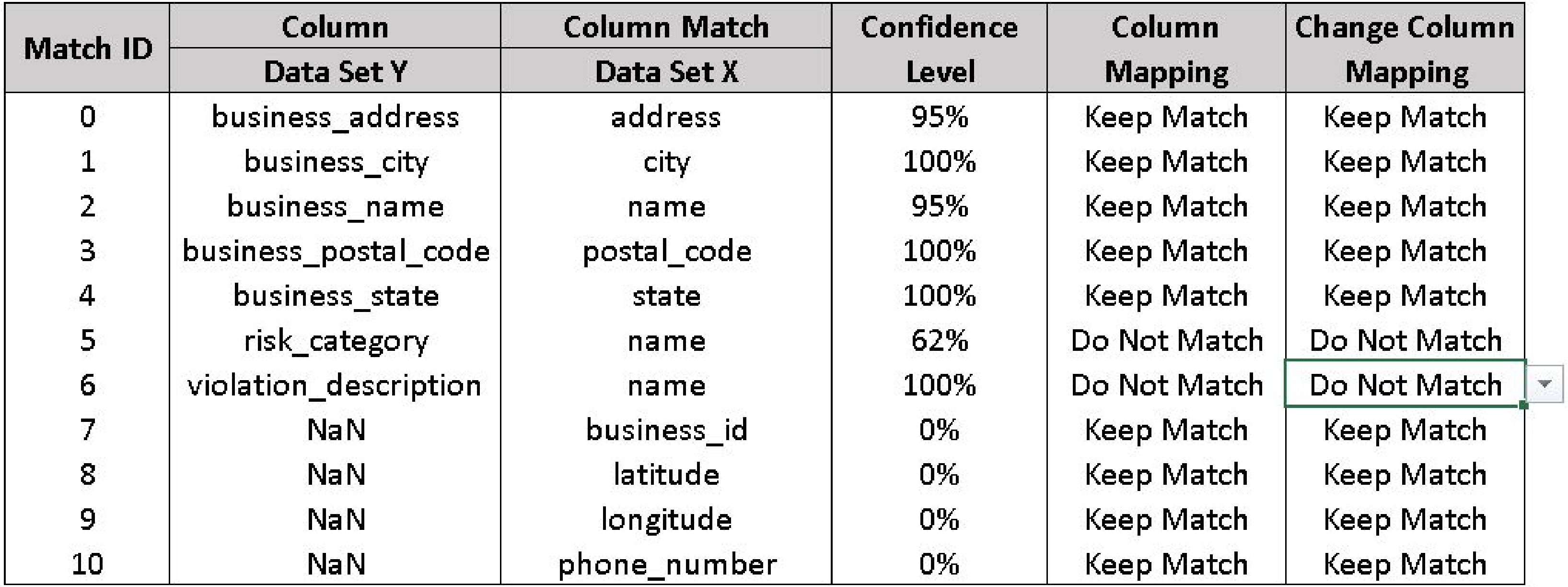
business_name	business_address	business_city
Tiramisu Kitchen	033 Belden Pl	San Francisco
Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	San Francisco
OMNI S.F. Hotel - 2nd Floor Pantry	500 California St, 2nd Floor	San Francisco
Norman's Ice Cream and Freezes	2801 Leavenworth St	San Francisco
CHARLIE'S DELI CAFE	3202 FOLSOM St	San Francisco

Training Dataset A



	business_id	name	address	city	state	postal_code	latitude	longitude	phone_number
business_id	329	0	0	0	0	595	0	0	0
name	0	843	17	33	1	0	0	0	0
address	0	11	898	0	0	0	0	0	0
city	0	0	0	937	0	0	0	0	0
state	0	0	0	0	946	0	0	0	0
postal_code	0	0	0	0	1	895	0	0	0
latitude	0	0	0	0	0	0	574	3	0
longitude	0	0	0	0	0	0	0	618	0
phone_number	0	0	0	0	0	0	0	0	274

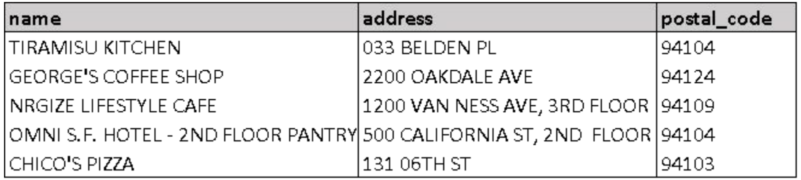
Testing Dataset B



Match ID	Column	Column Match	Confidence	Column	Change Column
	Data Set Y	Data Set X	Level	Mapping	Mapping
0	business_address	address	95%	Keep Match	Keep Match
1	business_city	city	100%	Keep Match	Keep Match
2	business_name	name	95%	Keep Match	Keep Match
3	business_postal_code	postal_code	100%	Keep Match	Keep Match
4	business_state	state	100%	Keep Match	Keep Match
5	risk_category	name	62%	Do Not Match	Do Not Match
6	violation_description	name	100%	Do Not Match	Do Not Match
7	NaN	business_id	0%	Keep Match	Keep Match
8	NaN	latitude	0%	Keep Match	Keep Match
9	NaN	longitude	0%	Keep Match	Keep Match
10	NaN	phone_number	0%	Keep Match	Keep Match

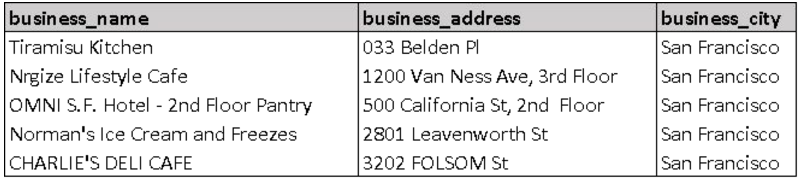
Merging the datasets

Data Set X



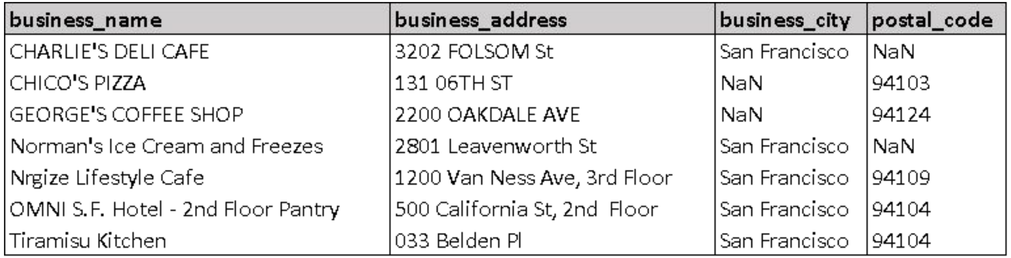
name	address	postal_code
TIRAMISU KITCHEN	033 BELDEN PL	94104
GEORGE'S COFFEE SHOP	2200 OAKDALE AVE	94124
NRGIZE LIFESTYLE CAFE	1200 VAN NESS AVE, 3RD FLOOR	94109
OMNI S.F. HOTEL - 2ND FLOOR PANTRY	500 CALIFORNIA ST, 2ND FLOOR	94104
CHICO'S PIZZA	131 06TH ST	94103

Data Set Y



business_name	business_address	business_city
Tiramisu Kitchen	033 Belden Pl	San Francisco
Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	San Francisco
OMNI S.F. Hotel - 2nd Floor Pantry	500 California St, 2nd Floor	San Francisco
Norman's Ice Cream and Freezes	2801 Leavenworth St	San Francisco
CHARLIE'S DELI CAFE	3202 FOLSOM St	San Francisco

Merged data set



business_name	business_address	business_city	postal_code
CHARLIE'S DELI CAFE	3202 FOLSOM St	San Francisco	NaN
CHICO'S PIZZA	131 06TH ST	NaN	94103
GEORGE'S COFFEE SHOP	2200 OAKDALE AVE	NaN	94124
Norman's Ice Cream and Freezes	2801 Leavenworth St	San Francisco	NaN
Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	San Francisco	94109
OMNI S.F. Hotel - 2nd Floor Pantry	500 California St, 2nd Floor	San Francisco	94104
Tiramisu Kitchen	033 Belden Pl	San Francisco	94104