SEMINAR

AI AND DUPLICATE DETECTION TO LEVERAGE EXTERNAL DATA SOURCE

April 21st

ABOUT MYSELF

DAVID BEAUCHEMIN
Ph. D. candidate in ML
Laval University

- B. Sc. in actuarial science
- M. Sc. in computer science, ML and NLP
- 5 years in academic ML and NLP researc project
- +6 years of various applied business projects
- Founder and creator of OpenLayer* a bilingual AI podcast
- Founder of .Layer* a data science non-profit organisation

MENU

1. What is our problem?
2. How can we develop a solution for that?
3. How can we compute the similarity between two entities?
4. Why develop our own solutions?

What is our problem?

WHAT IS OUR PROBLEM?

Data is not cheap

WHAT IS OUR PROBLEM ?

A lot of open data now (data.gouv*)

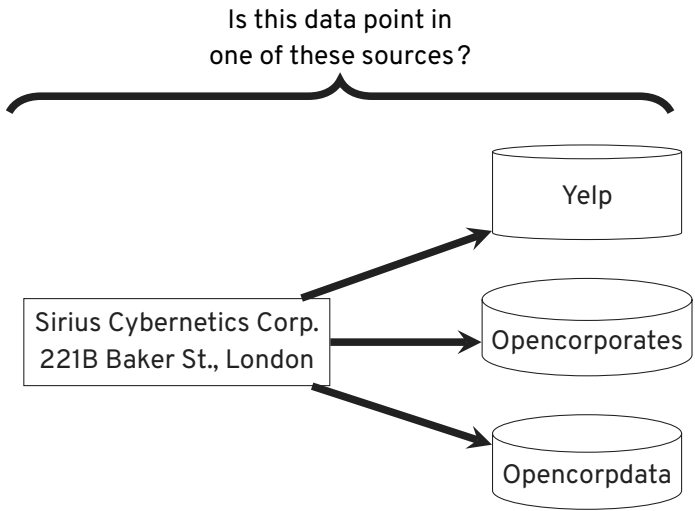
WHAT IS OUR PROBLEM?

Mapping between datasets

AN EXAMPLE

Only using the name and the address of a business client is it possible to match these with external sources to leverage their content ?

AN EXAMPLE

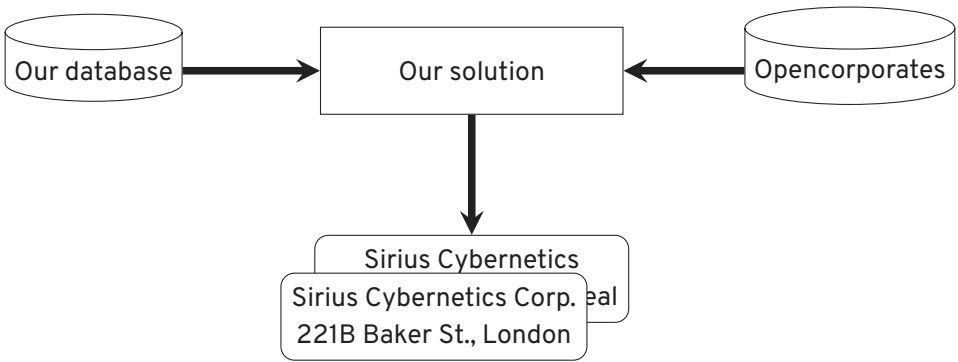

Is this data point in
one of these sources?
Yelp
Sirius Cybernetics Corp.
221B Baker St., London
Opencorporates
Opencorpdata

AN EXAMPLE

How can we match our information with external data sources if we do not control it and our information is not unique?

How can we develop a solution for that?

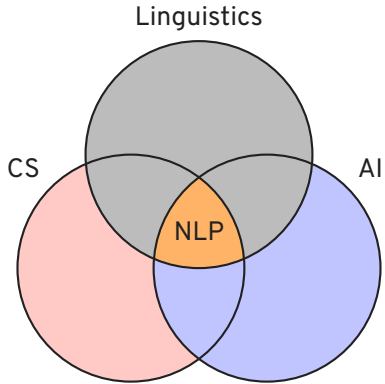
WHAT OUR SOLUTION NEEDS TO DO?



WHAT DO WE NEED TO ACHIEVE THAT?

- Natural language processing (NLP)
- A duplicate detection mechanism

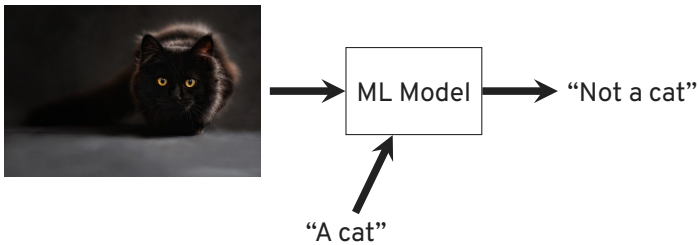
WHAT IS NLP?



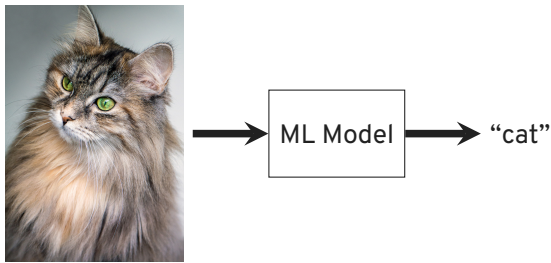
WHAT IS AI?

A cat or not?

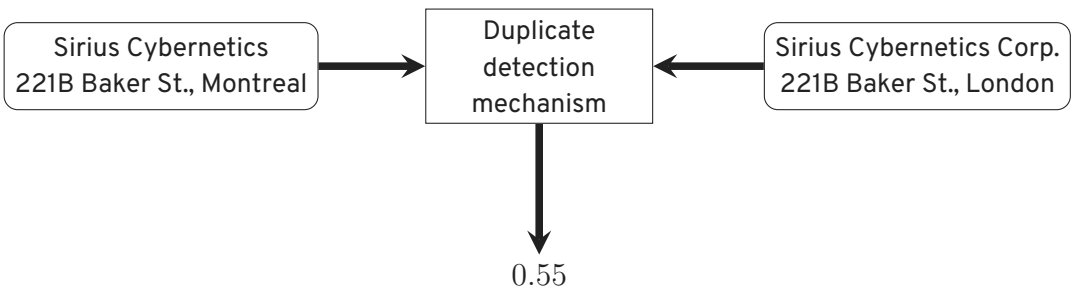
WHAT IS AI?



WHAT IS AI?



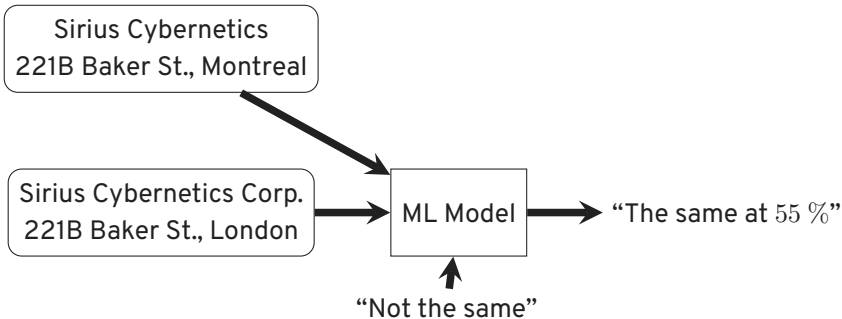
DUPLICATE DETECTION MECHANISM



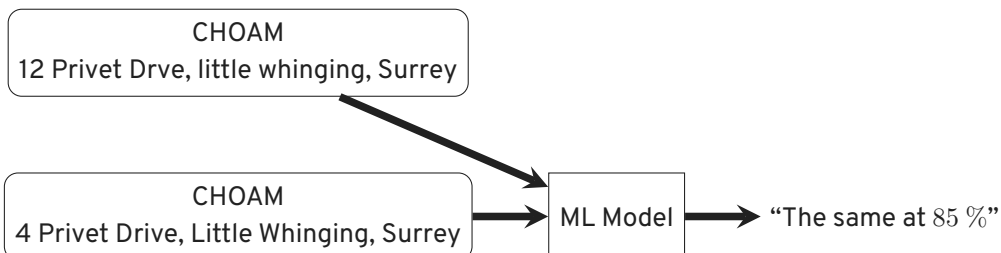
DUPLICATE DETECTION MECHANISM

- Rules-based
- Probabilistic
- ML

ML DUPLICATE DETECTION MECHANISM

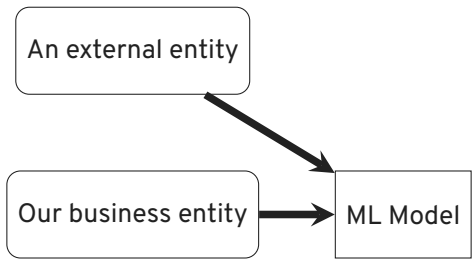


ML DUPLICATE DETECTION MECHANISM

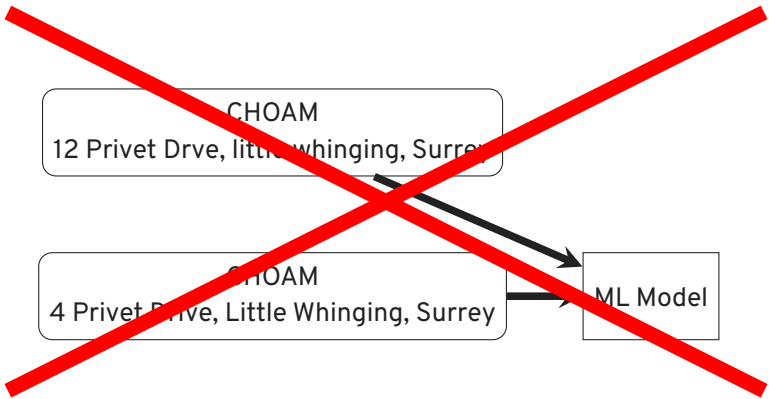


How can we compute the similarity between two entities?

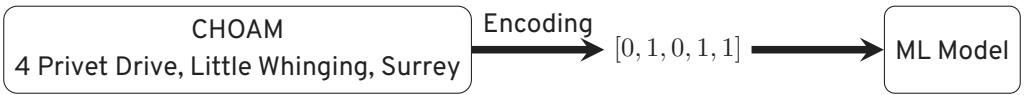
SIMILARITY ALGORITHM



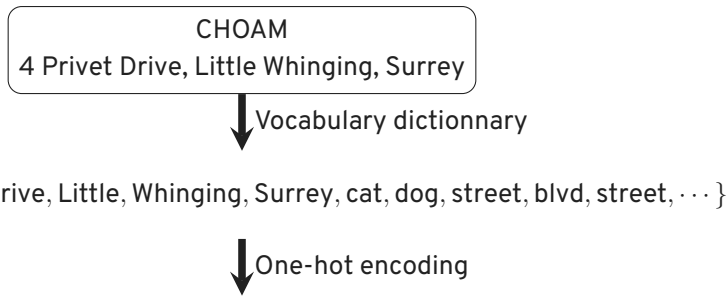
SIMILARITY ALGORITHM



NLP ML ALGORITHM



ONE-HOT ENCODING



ONE-HOT ENCODING

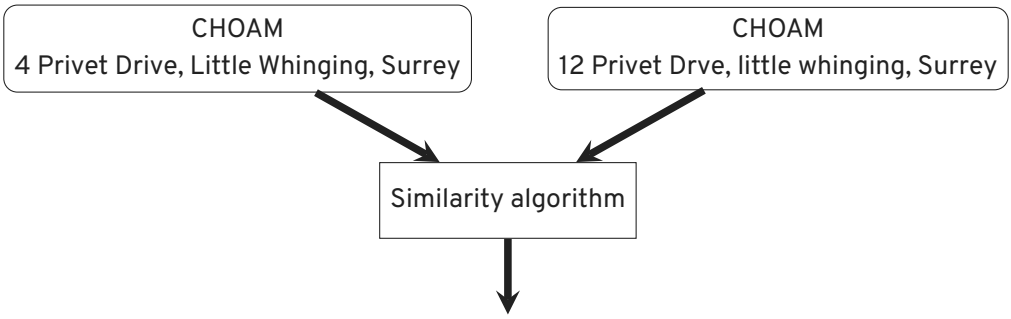
Advantages

- Easy to implement
- Not complex
- Easy to encode

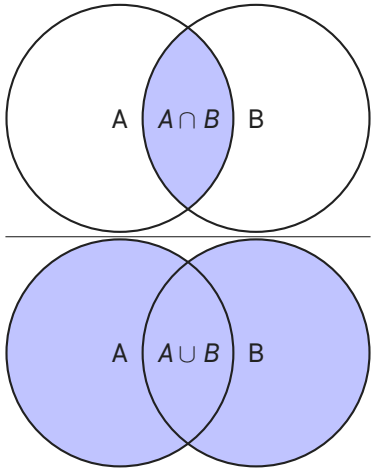
Disadvantages

- Sparse
- Out-of-vocabulary

SIMILARITY ALGORITHM



JACCARD



JACCARD

$$\frac{\{\text{CHOAM}, \text{Privet}, \text{Surrey}\}}{\{\text{CHOAM}, 4, 12, \text{Privet}, \text{Drive}, \text{Drve}, \text{Little}, \text{little}, \text{Whinging}, \text{whinging}, \text{Surrey}\}} = \frac{3}{11} = 0.2727$$

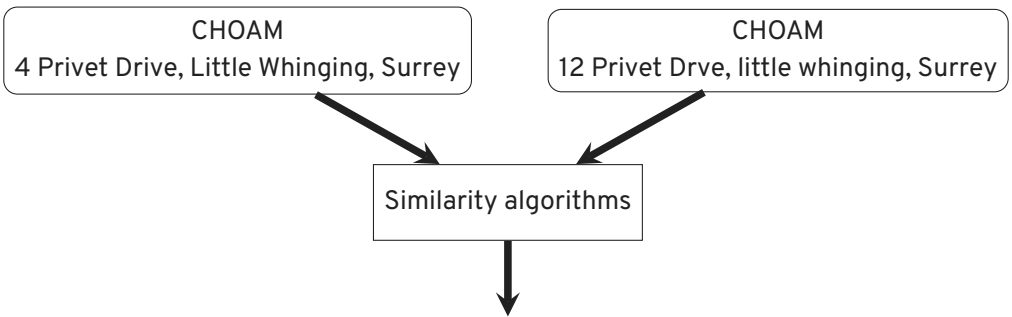
JACCARD

$$\frac{\{\text{choam}, \text{privet}, \text{little}, \text{whinging}, \text{surrey}\}}{\{\text{choam}, 4, 12, \text{privet}, \text{drive}, \text{drve}, \text{little}, \text{whinging}, \text{surrey}\}} = \frac{5}{9} = 0.5556$$

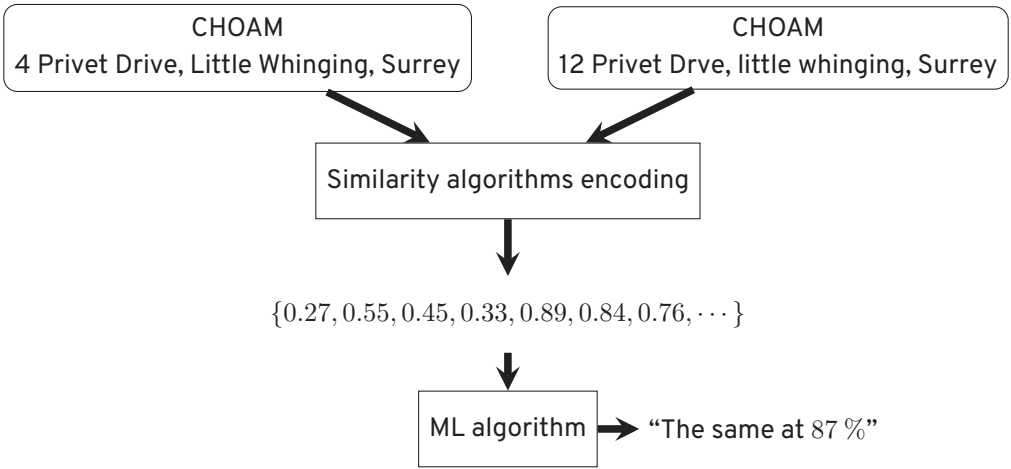
SIMILARITY ALGORITHM

- Jaro
- Jaro-Winkler
- Levenshtein
- Longest common subsequence
- Cosinus
- MASI
- Monge Elkan
- Overlap

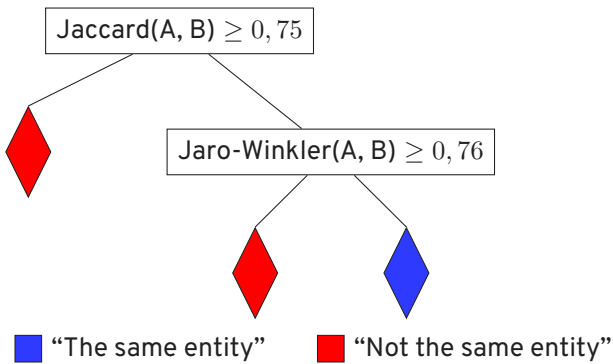
SIMILARITY ALGORITHM



FINAL APPROACH



ML ALGORITHM



NAME-ADDRESS INFORMATION VECTOR GENERATOR

Example of an information vector

StoS	Levenshtein	Jaro-Winkler	LCSP	Jaccard	Cosinus	-
0.00	0.15	0.25	0.35	0.15	0.15	-

StoS	Levenshtein	Jaro	LCSP	Jaccard	Cosinus	CSS
0.00	0.16	0.55	0.15	0.45	0.37	0.48

RESULTS

	Logistic Regression	Random Forest	Multilayer perceptron	Jaccard
Recall (%)	66.67	73.54	**79.73**	72.51
Precision	**89.77**	81.06	87.55	81.78

Detection of Duplicates Among Non-structured Data From Different Data Sources *

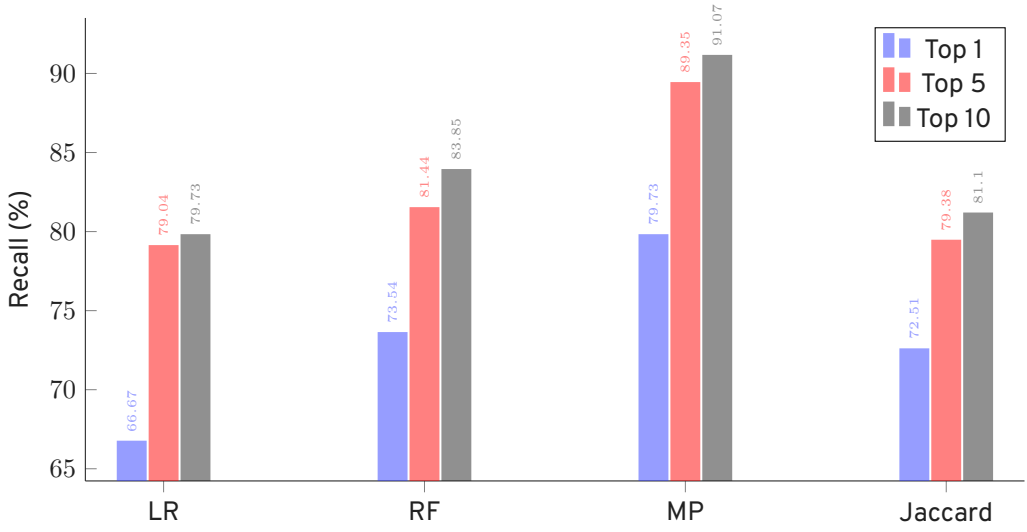
INFERENCE TIMES

(second)	Logistic Regression	Random Forest	Multilayer Perceptron	Jaccard
Time	1,32	1,74	1,34	0,25

IMPROVING THE RESULTS - *N* MOST SIMILAR

We consider a matching is good when the pair *(Our database entity, external data source entity)* is included in the *N* most similar rather than the top-1.

IMPROVING THE RESULTS - *N* MOST SIMILAR



Why develop our own solutions?

WHY DEVELOP OUR OWN SOLUTIONS?

There is commercial solutions out there, why not buying it?

WHY DEVELOP OUR OWN SOLUTIONS ?

- Help empower your teams on important NLP steps,
- It is a straightforward classification problem.

CONCLUSION

- NLP can help you bring value to your business.
- I have presented a solution that one can develop to detect duplicates using NLP and ML.
- Without deep learning, you can achieve interesting results in duplicate detection with external data sources.

SEMINAR

THANK YOU FOR LISTENING!