Entity Resolution Framework for Big Data

- Sakshi Rai (1806059)
- Samdarshi jha (1806060)
- Shashikant Singh (1806068)
- Ananya (1806458)
- Manisha Godara (1806486)

Under The Guidance Of Prof. Bhaswati Sahoo

ENTITY RESOLUTION

Problem of identifying and linking/grouping different manifestations of the same real world object.

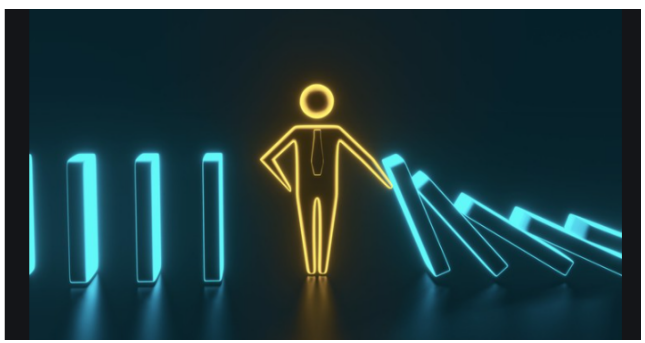
Examples of manifestations and objects:

- Different ways of addressing (names, Email addresses, FaceBook accounts) the same person in text.
- Web pages with differing descriptions of the same business.
- Different photos of the same object.

This clearly has many applications, particularly Linking Census Records, public health data, web search, comparison shopping, law enforcement, and more.

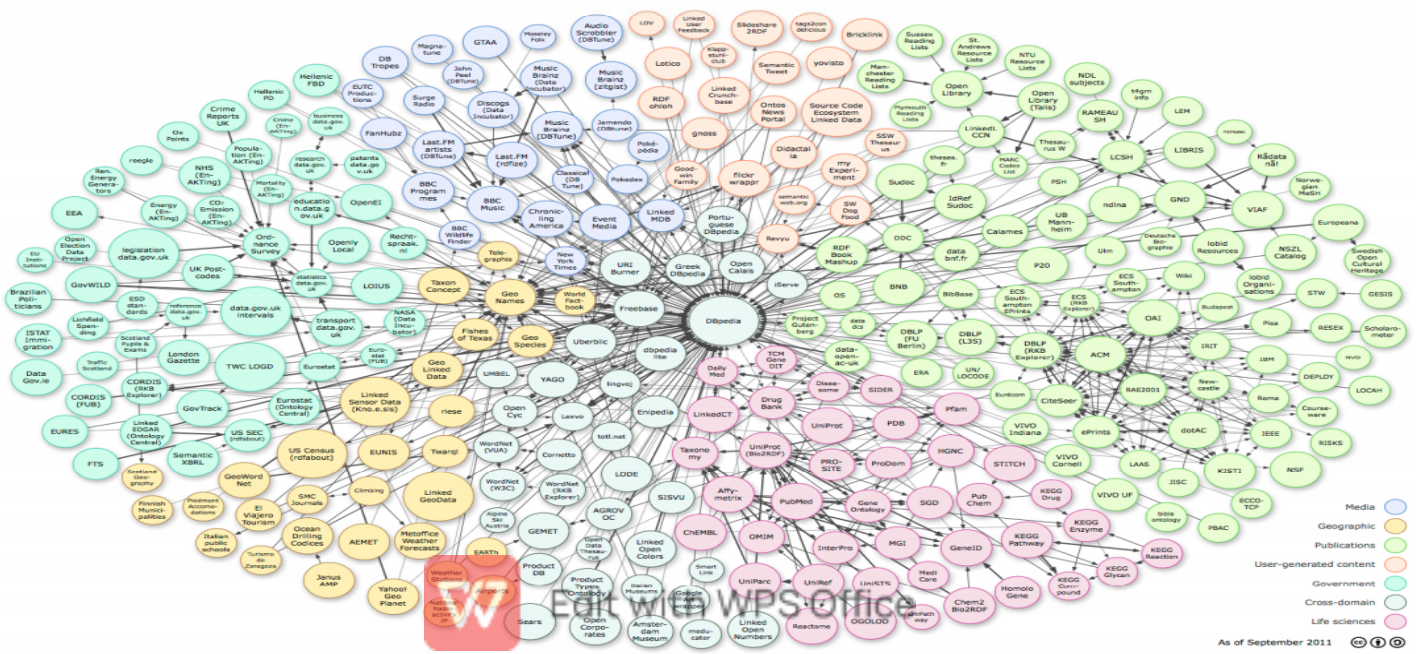
TRADITIONAL CHALLENGES IN ER

- Name/Attribute ambiguity
- Errors due to data entry
- Missing Values
- Changing Attributes
- Data formatting
- Abbreviations / Data Truncation



BIG DATA CHALLENGES IN ER

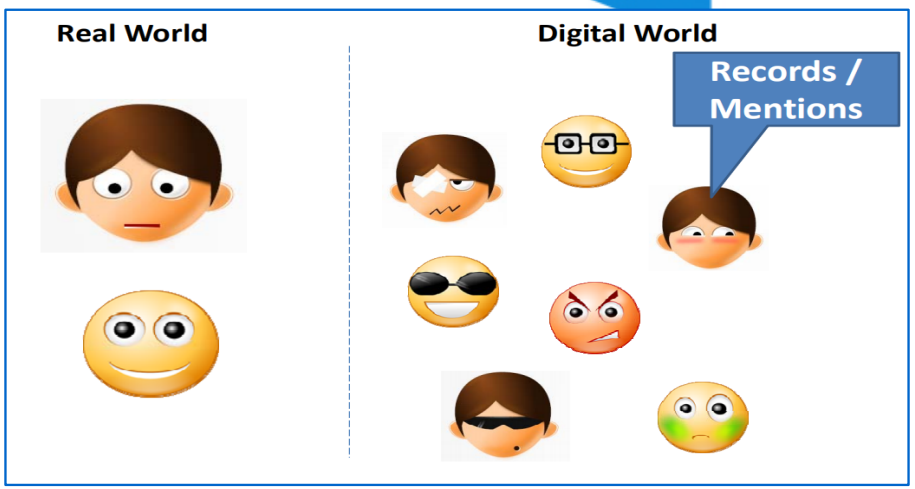
- Larger and more Datasets Need efficient parallel techniques
- Unstructured, Unclean and Incomplete data. Diverse data types.
- Need to infer relationships in addition to “equality”
- Deal with structure of entities (Are Walmart and Walmart Pharmacy the same?)

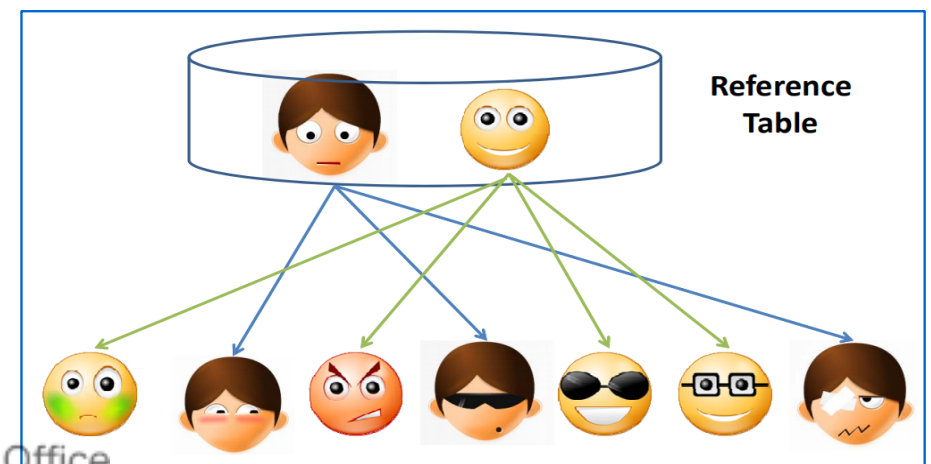


TASKS IN ENTITY RESOLUTION

There exists in the real world entities, and in the digital world, records and mentions of those entities. The records and mentions may take many forms, but they all refer to only a single real world entity.

We can therefore discuss the ER problem as one involving matching record pairs corresponding to the same entity, and as a graph of related records/mentions to related entities.





THE GOAL

The goal is to construct, for a pair of records, a “Comparison Matrix” of similarity scores of each component attribute.

Similarity scores can simply be Boolean (match or non-match) or they can be real values with distance functions.

After we have constructed a vector of component-wise similarities for a pair of records, we must compute the probability that the pair of records is a match.

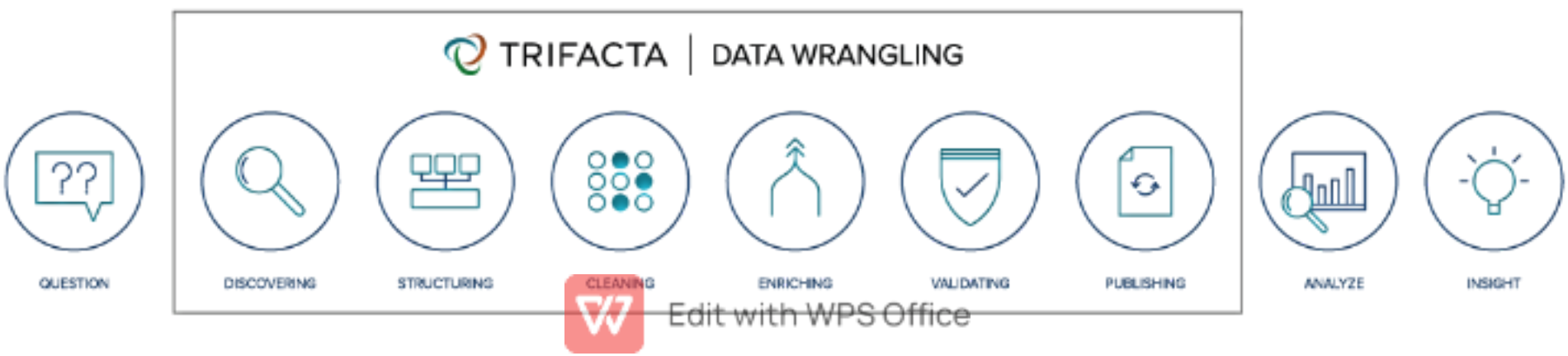
j / i	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
0.0	1.0	0.382571	0.532699	0.439488	0.681190	0.533750	0.563029	0.598446	0.413979	0.410752
1.0	NaN	1.000000	0.467336	0.664375	0.306781	0.241162	0.388742	0.228590	0.428832	0.660477
2.0	NaN	NaN	1.000000	0.551375	0.539465	0.396098	0.444757	0.421963	0.668599	0.480889
3.0	NaN	NaN	NaN	1.000000	0.418856	0.245248	0.504123	0.319967	0.510964	0.570572
4.0	NaN	NaN	NaN	NaN	1.000000	0.621446	0.655645	0.613013	0.481490	0.329167
5.0	NaN	NaN	NaN	NaN	NaN	1.000000	0.497234	0.444127	0.336801	0.268338
6.0	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.497442	0.379787	0.332879
7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.442287	0.256362
8.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.448167
9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000

WORKFLOW


Get Data
Data Preparation
Check Column Types in Data
Binning
Blocking
Pairwise Matching
Entity Resolution Framework
By Similarty Score Matrix
Edit with WPS Office

DATA PREPARATION

The first tasks of schema and data normalization are required preparation for any Entity Resolution Algorithms. In this task, schema attributes are matched (e.g. contact number vs phone number), and compound attributes like addresses are normalized. Data normalization involves converting all strings to upper or lower case and removing whitespace. Data cleaning and dictionary lookups are also important.



ANALYSING ATTRIBUTES

It is important to determine the types of columns present in a database. It give a clear view to perform any sort of analytical operations. In our Module we determine following types:

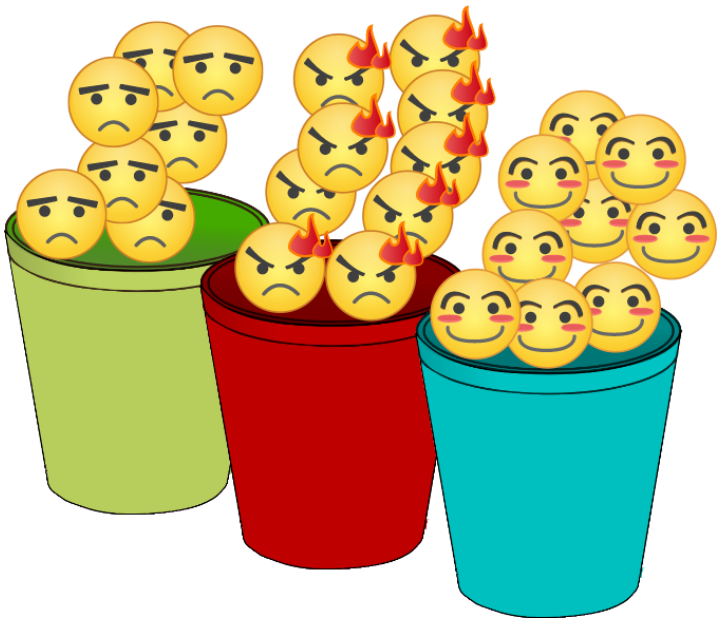
- Constant
- Boolean
- Numeric With Length=1
- Numeric Continuous
- Uniformly Spaced
- Non-Uniformly Spaced
- Uniformly Distributed
- Date
- Only Text
- Text with Number & Special Character

TITANIC DATASET EXAMPLE

	Name	Type_col
0	PassengerId	Numeric Continuous
1	Pclass	Numeric With Length=1
2	Name	Text with Number & Special Character
3	Sex	Boolean
4	Age	Uniformly Distributed
5	SibSp	Numeric With Length=1
6	Parch	Numeric With Length=1
7	Ticket	Text with Number & Special Character
8	Fare	Numeric Continuous
9	Cabin	Text with Number & Special Character
10	Embarked	Only Text

BINNING

Sometimes numerical values make more sense if clustered together. Data binning, which is also known as bucketing or discretization, is a technique used in data processing and statistics. Binning can be used for example, if there are more possible data points than observed data points.

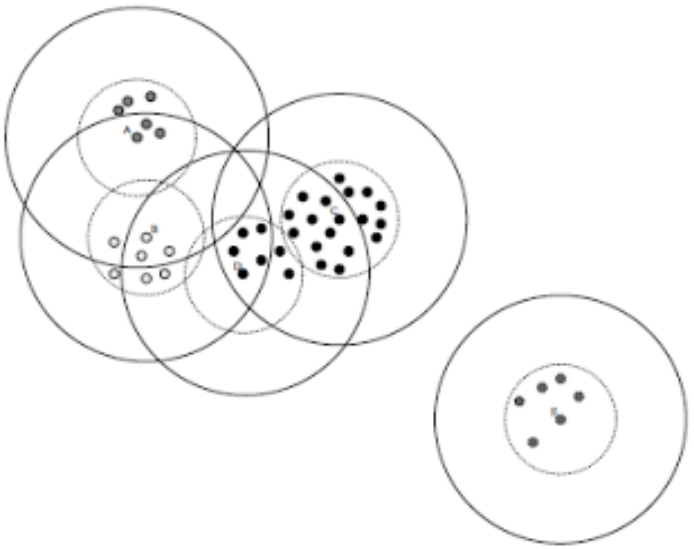


pandas.cut() Bin values into discrete intervals. Use cut when you need to segment and sort data values into bins.

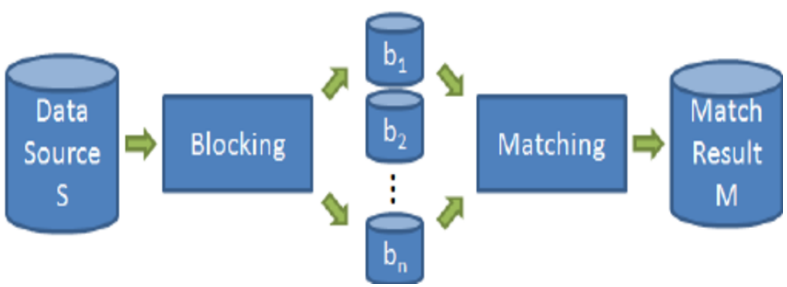
pandas.cut

BLOCKING

The first approach to scaling to big data is to use a blocking approach. Consider a Naïve pairwise comparison with 1, 000 business mentions from 1,000 cities. This is 1 trillion comparisons, which is 11.6 days with ms comparisons! However, we know that business mentions are probably city-specific, therefore only comparing mentions within cities reduces the comparison space to the more manageable 1 billion comparisons, which takes only 16 minutes at the same rate.



It cluster similar entities into blocks so that pair-wise comparisons are executed only between co-occurring entities.



PAIRWISE MATCHING

It finds the similarity score across all records in a dataframe. We must compute the probability that the pair of records is a match. There are several methods for discovering the probability of a match. Two simple proposals are to use a weighted sum or average of component similarity scores, and to use thresholds.

CONCLUSION

- Entity resolution is becoming an increasingly important task as linked data grows, and the requirement for graph based reasoning extends beyond theoretical applications. With the advent of big data computations, this need has become even more prevalent.

- As data, noise, and knowledge grows , greater the needs & opportunities for intelligent reasoning about entity resolution.

FIRST EVER COMPANY

Senzing

Senzing is the First Blue Chip Company to Introduce Real-Time AI For Entity Resolution. Jeff Jonas and team developed a vision to revolutionize entity resolution.

DATA

REVIEW

THANK YOU!