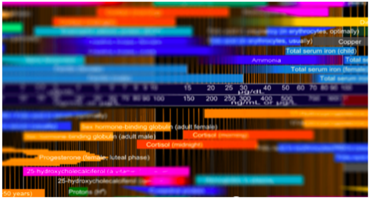
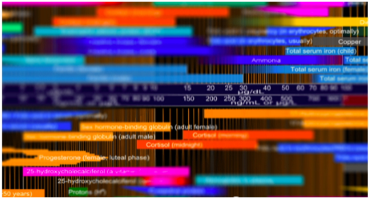
Preliminaries

Rebecca C. Steorts

August 28, 2024

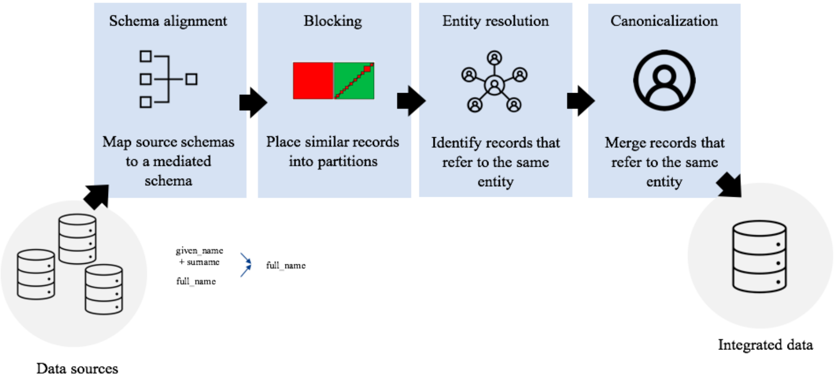

Copper
Total serum iron (child)
Ammonia
Sex hormone-binding globulin (adult female)
Cortisol (midnight)
Progesterone (female, luteal phase)
Protons (H+)



What do these datasets have in common?

- There is duplication in the data.
- The amount of duplication is typically small.
- Before we can apply inferential or prediction methods, any duplicate records must be removed.

Data Cleaning Pipeline



Entity resolution (ER) is the process of merging together noisy (structured) databases to remove duplicate entities, often in the absence of a unique identifier.

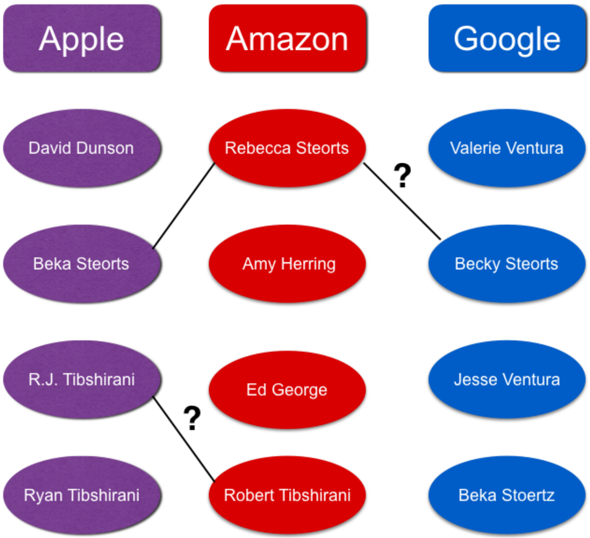
Other names for entity resolution:

record linkage, deduplication, duplicate detection, data matching, data integration, data cleansing.

Foundations and Terminology

A graph with no edges

The entity resolution graph



Entities are Real People (Objects, Businesses, Etc.)

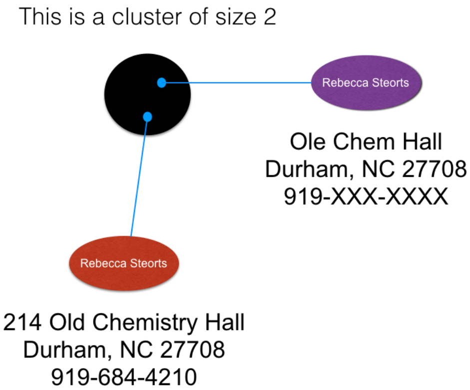
Rebecca Steorts

214 Old Chemistry Hall
Durham, NC 27708
919-684-4210

Becky Steorts

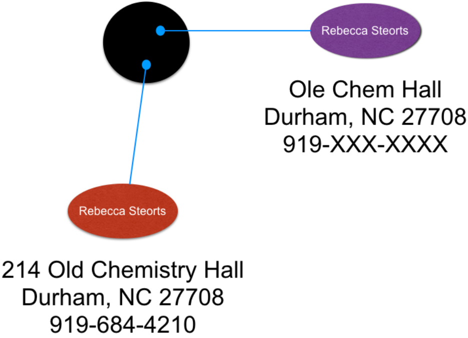
213 Main Street
Charleston, WV
304-XXX-XXXX

Goal of Entity Resolution



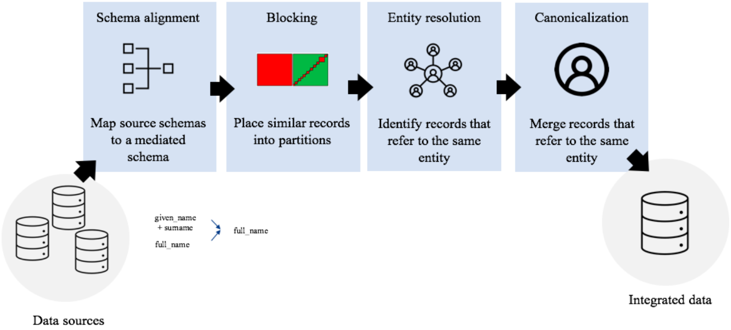
Goal of Entity Resolution

This is a cluster of size 2



To find the most representative records after ER, one must perform canonicalization (data fusion or merging).

In this talk, I will focus on the entity resolution task of the data cleaning pipeline.



[Christen (2012), Christophides+ (2021), Papadakis+ (2021), Binette and Steorts (2022)]

Challenges

Challenges of Entity Resolution

Costly manual labelling

Vast amounts of manually-labelled data are typically required for supervised learning and evaluation.

Scalability/computational efficiency

Approximations are required to avoid quadratic scaling. Need to ensure impact on accuracy is minimal.

Limited treatment of uncertainty

Given inherent uncertainties, it's important to output predictions with confidence regions.

Unreliable evaluation

Standard evaluation methods return imprecise estimates of performance.

Evaluation Metrics

How do we assess the effectiveness of entity resolution methods, where some ground truth is known?

Confusion Matrix

- No = Non-Match
- Yes = Match

Table: Confusion Matrix

N= Total Records		Actual Linkage	
		No	Yes
Predicted Linkage	No	true neg. (TN)	false neg. (FN)
	Yes	false pos. (FP)	true pos. (TP)

Confusion Matrix

Table: Confusion Matrix

		Predicted Linkage	
		Match	Non-Match
Actual Linkage	Match	true pos. (TP)	false pos. (FP)
	Non-Match	false neg. (FN)	true neg. (TN)

In the TP, FP, TN, FN terminology:

- "True" / "False" = prediction is correct/incorrect
- "Positive" / "Negative" = predicted class is positive/negative

Evaluation Metrics

$$\text{Accuracy (acc)} = \frac{TP + TN}{TP + FP + TN + FN}$$

- Commonly used in machine learning problems.
- Useful in situations where the data is balanced, i.e. matches and non-matches are roughly the same.
- The number of TN dominates, and leads to a class imbalance issue (and results that are misleading).

For an example, see page 167 of Christen (2012).

Evaluation Metrics

- False positive rate (FPR) $= \dfrac{FP}{FP + FN}$
  - Fraction of actual negatives that were predicted to be positive.
  - Specificity = Precision = 1 - FPR $= \dfrac{TP}{TP + FP}$
- True Positive Rate (TPR) $= \dfrac{TP}{TP + FN}$
  - Fraction of actual positives that were predicted to be positive.
  - Sensitivity = TPR.

- Useful in situations where the data is balanced, i.e. matches and non-matches are roughly the same.
- The number of TN dominates, and leads to a class imbalance issue (and results that are misleading).

Evaluation Metrics

$$\text{Precision} = \frac{TP}{TP + FP}$$

Measures how precise a method is in classifying true matches.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Measures how accurately the actual true matching pairs of records are correctly classified as matches.

Observe these metrics do not include TN. They do not suffer from a class imbalance issue.

Evaluation Metrics

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Harmonic mean of the precision and recall.
- Attempts to summarize all aspects of the effectiveness of an entity resolution method.