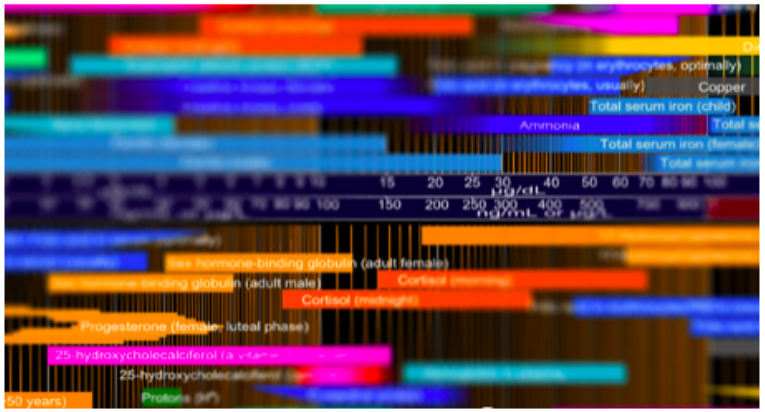
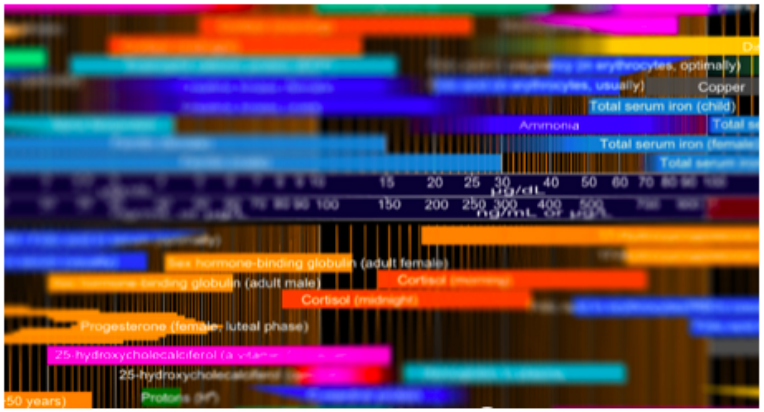
Preliminaries

Rebecca C. Steorts

September 6, 2024

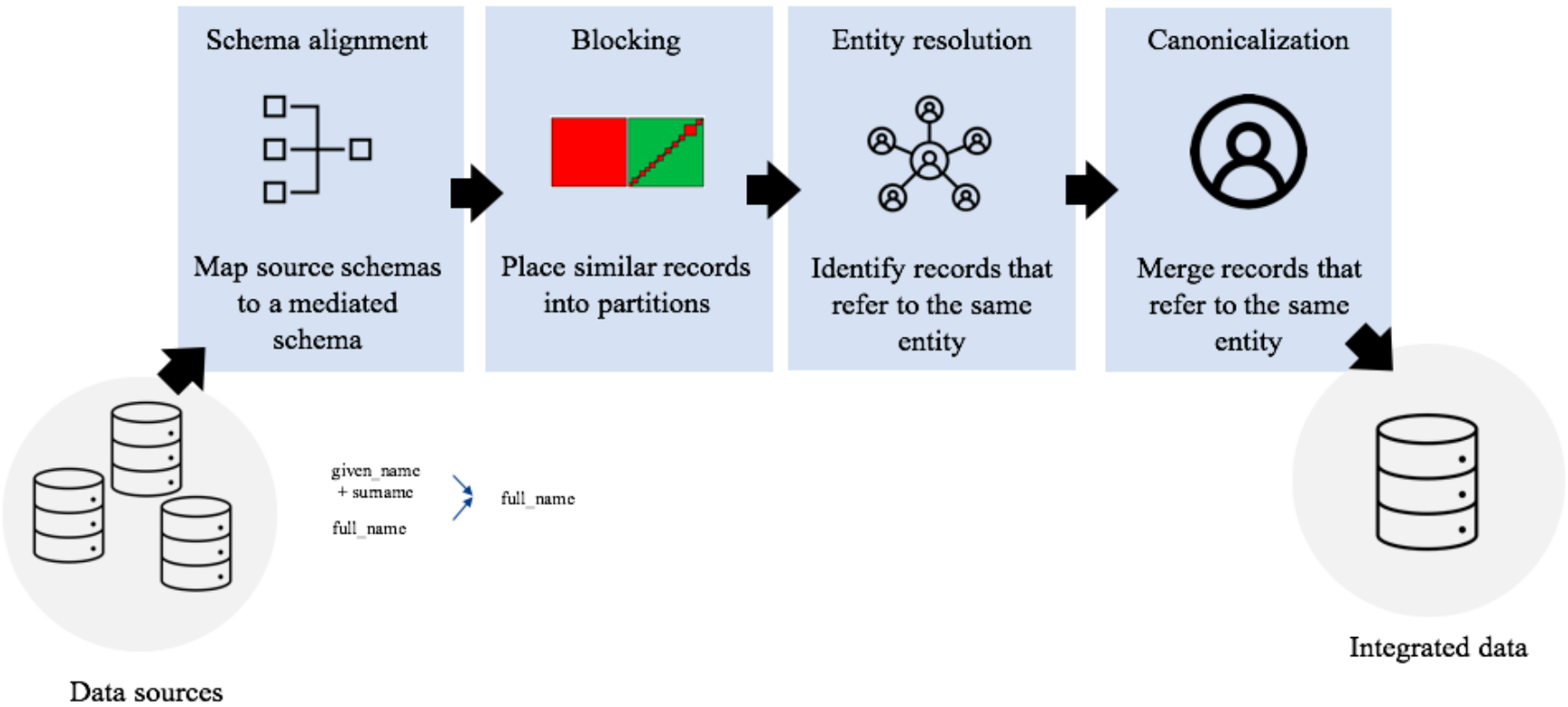

Total serum iron (child)
Ammonia
Total serum iron (female
Cortisol (midnight)
Progesterone (female, luteal phase)
50 years)



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.

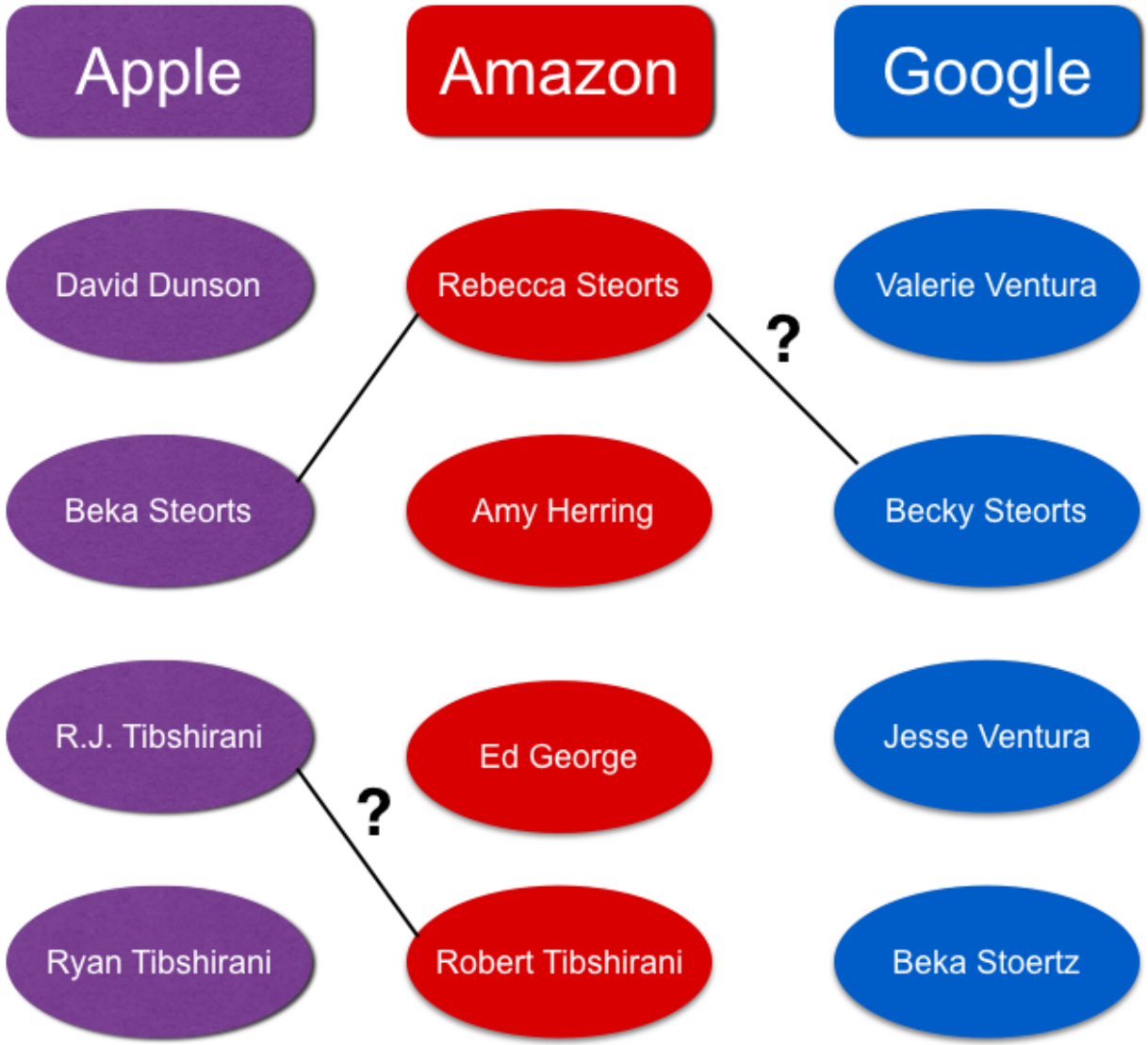
Other names for entity resolution:

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

Foundations and Terminology

A graph with no edges

The entity resolution graph

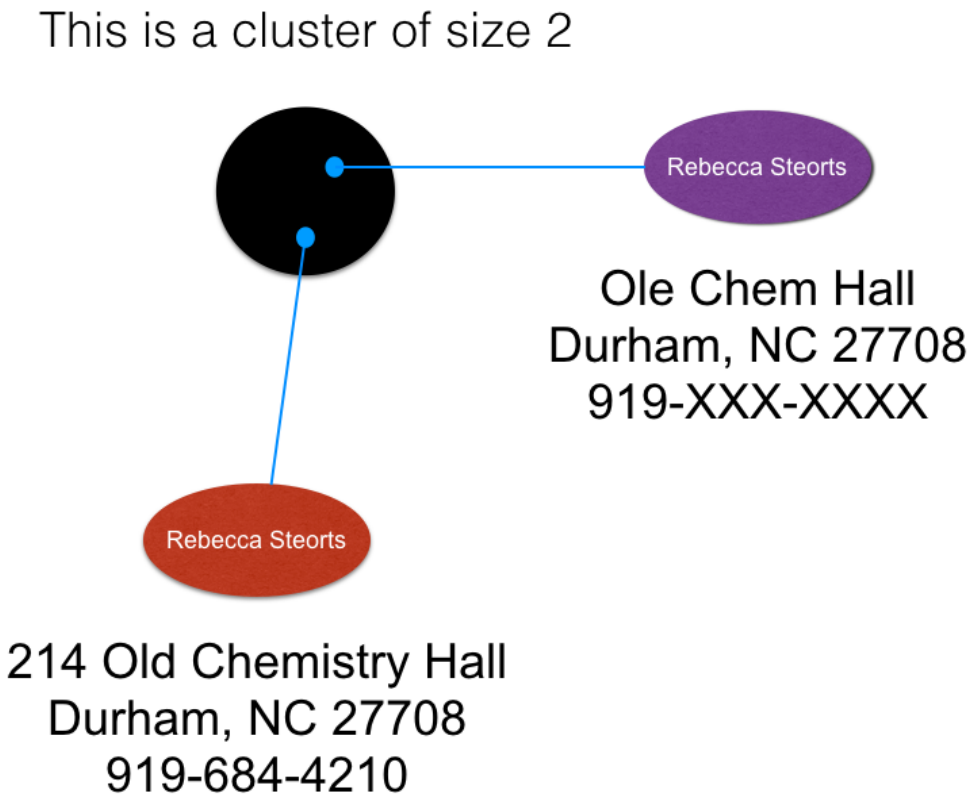


Entities are Real People (Objects, Businesses, Etc.)

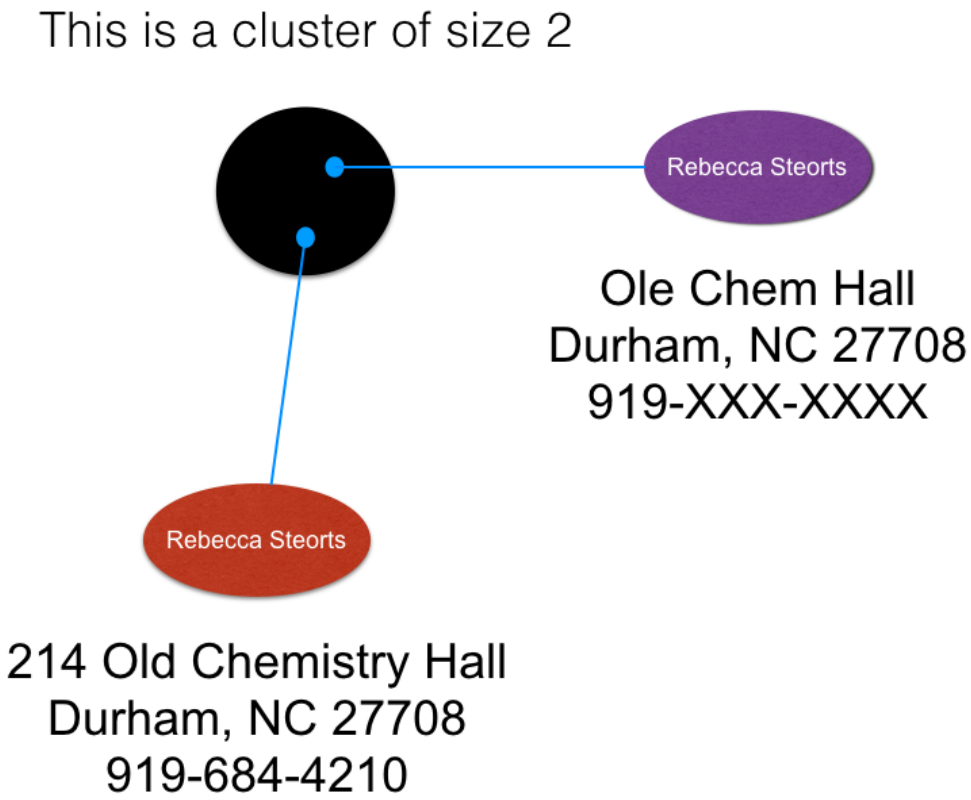
214 Old Chemistry Hall
Durham, NC 27708
919-684-4210

213 Main Street
Charleston, WV
304-XXX-XXXX

Goal of Entity Resolution

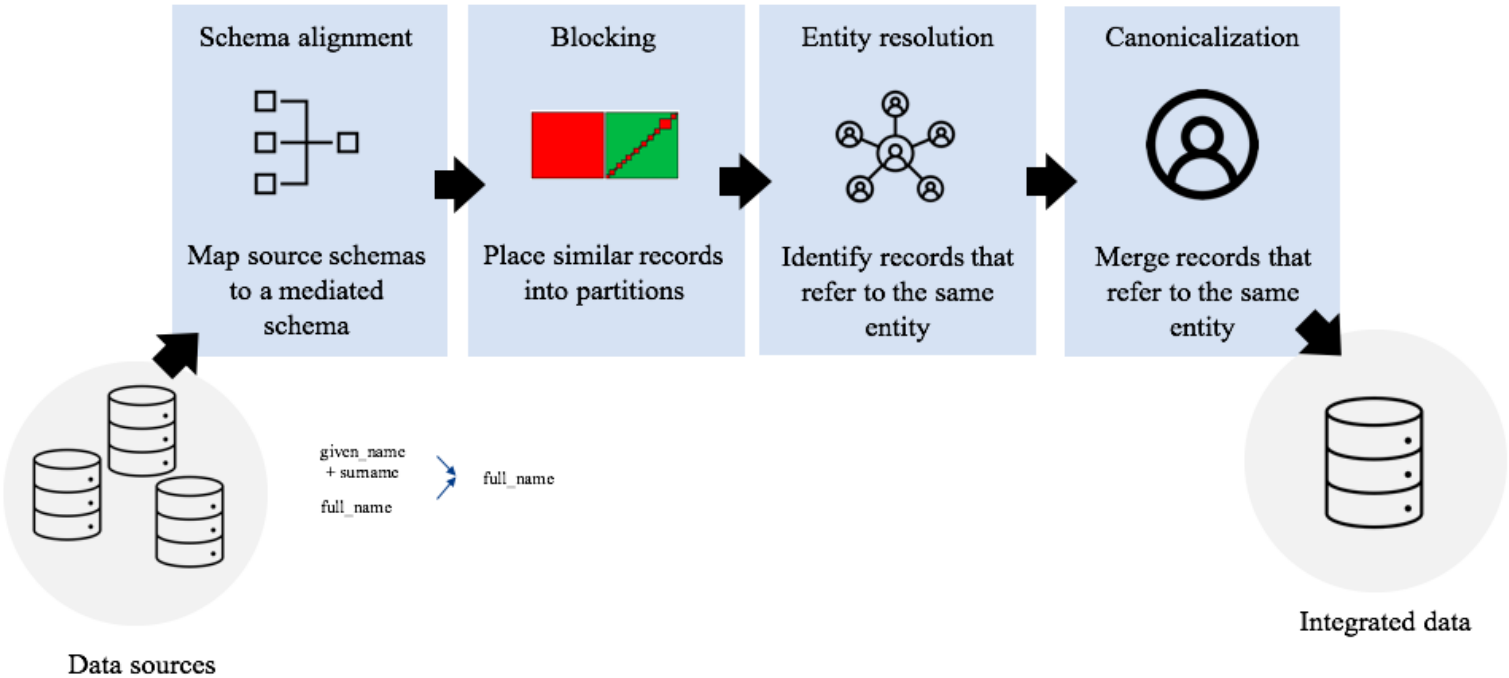


Goal of Entity Resolution



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?

True Positive (True Matches)

True Positive (TP): These are records that are classified as matches and are true matches.

These are pairs of records that refer to the same person (entity).

True Negatives (True Non-Matches)

True Negative (TN): These record pairs are classified as non-matches, they are true non-matches.

The two records refer to two different entities.

False Positive (False Matches)

False Positive (FP): These are record pairs that have been classified as matches, but they are not true matches.

The model made a wrong decision with the record pairs and falsely declared them to be matches.

False Negatives (False Non-Matches)

These are record pairs that have been classified as non-matches, but they are actually true matches.

The two record pairs refer to the same entity, but the method made a mistake.

Confusion Matrix

- Match (M)
- Non-Match (NM)
- N = total records

N		Predicted	
		M	NM
Actual	M	TP (true matches)	
	NM		TN (true non-matches)

- True Positive (TP): The model correctly predicted a positive match (the actual outcome was positive).
- True Negative (TN): The model correctly predicted a negative outcome (the actual outcome was negative).

Confusion Matrix

- Match (M)
- Non-Match (NM)
- N = total records

N		Predicted	
		M	NM
Actual	M		FN
	NM	FP	

- False Positive (FP): The model incorrectly predicted that the record pairs are a true match.
- False Negative (FN): The model incorrectly predicted that the record pairs are a non-match.

Confusion Matrix

- Match (M)
- Non-Match (NM)
- N = total records

N		Predicted	
		M	NM
Actual	M	TP (true matches)	FN
	NM	FP	TN (true non-matches)

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) $= \dfrac{FP}{FP + TN}$
  - Fraction of actual negatives that were predicted to be positive.
- True Positive Rate (TPR) $= \dfrac{TP}{TP + FN}$
  - Fraction of actual positives that were predicted to be positive.
  - Sensitivity $=$ TPR.
- True negative rate (TNR) $=$ specificity $= \dfrac{TN}{TN + FP}$

Which metrics suffer from a class imbalance issue and which ones do not?

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.

Summary

- What is entity resolution (and other names for it)?
- What are challenges of entity resolution?
- Know the components of a confusion matrix.
- What are the evaluation metrics used for entity resolution (and ones that we do not consider)? Be sure to know why!