

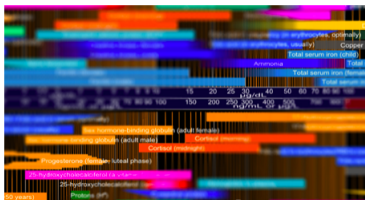
# An Overview of Entity Resolution

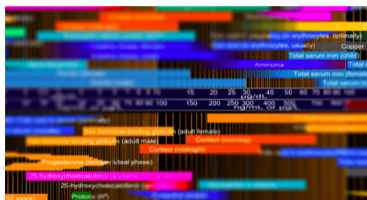
Rebecca C. Steorts

Department of Statistical Science, affiliated faculty in  
Computer Science, Biostatistics and Bioinformatics, the  
information initiative at Duke (iiD) and  
the Social Science Research Institute (SSRI)  
Duke University and U.S. Census Bureau

This work is supported by NSF CAREER Award 1652431 and  
the Alfred Sloan Foundation (DRB #: CBDRB-FY20-309).

June 27, 2021

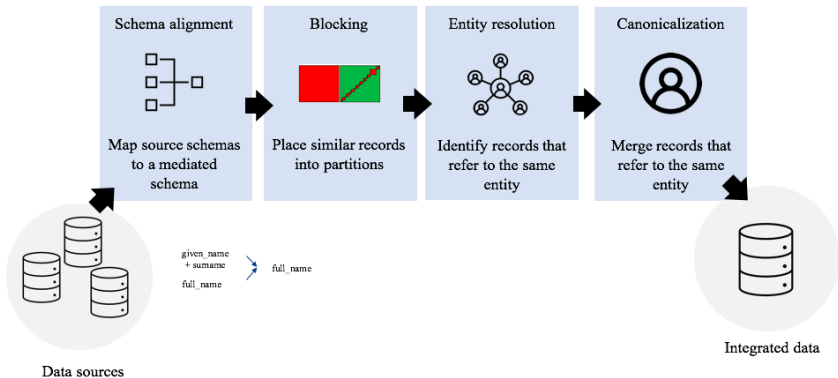




What do these datasets have in common?

- 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.

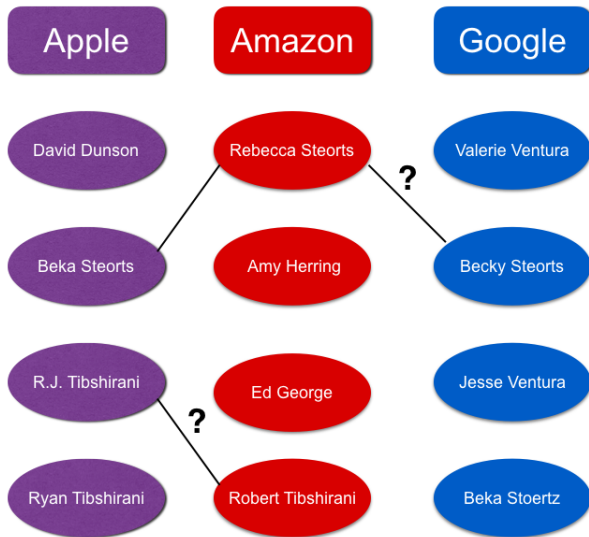
# Foundations and Terminology

# 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

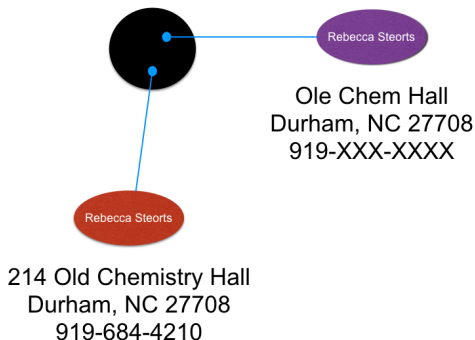


Becky Steorts

213 Main Street  
Charleston, WV  
304-XXX-XXXX

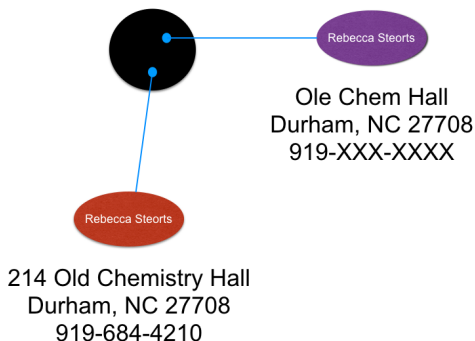
# Goal of Entity Resolution

This is a cluster of size 2



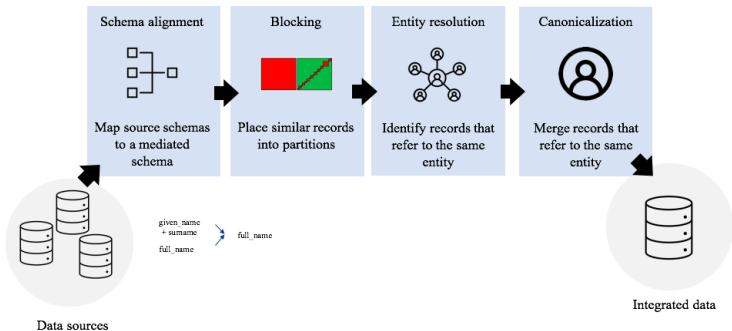
# Goal of Entity Resolution

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 (2021)]

# 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.



# The History of Probabilistic Record Linkage



# Record Linkage\*

HALBERT L. DUNN, M.D., F.A.P.H.A.

*Chief, National Office of Vital Statistics, U. S. Public Health Service,  
Federal Security Agency, Washington, D. C.*

## **Halbert L. Dunn (1896-1975):**

- Chief of the National Office of Vital Statistics from 1935-1960.
- A “leading figure in establishing a national vital statistics system in the United States”.

*Record linkage* is the task of assembling together all important pieces of information which refer to the same individual.

# Record Linkage\*

HALBERT L. DUNN, M.D., F.A.P.H.A.

*Chief, National Office of Vital Statistics, U. S. Public Health Service,  
Federal Security Agency, Washington, D. C.*

EACH person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

The Book has many pages for some

the various important records of a person's life.

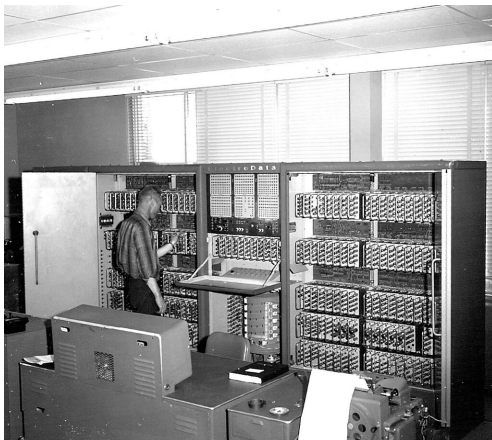
The two most important pages in the Book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, "death clearance."

# Automatic Linkage of Vital Records\*

**Computers can be used to extract “follow-up”  
statistics of families from files of routine records.**

**H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James**

Proposed a probabilistic record linkage method and implemented it on the Datatron 205 computer.



## A THEORY FOR RECORD LINKAGE\*

IVAN P. FELLEGI AND ALAN B. SUNTER

*Dominion Bureau of Statistics*

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

# Fellegi and Sunter (1969), JASA

The authors formalized Newcombe et al. (1959) in a decision-theoretic framework.

One determines if two records are a match using a likelihood ratio test exceeding a threshold.

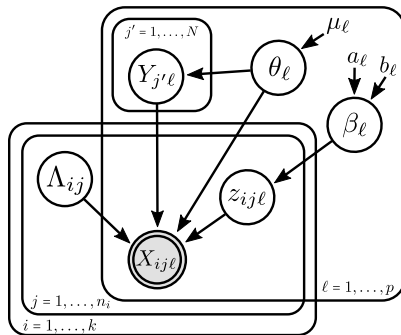
Many advancements have been made to the original paper, however, I will focus on more modern approaches known as graphical entity resolution that my group has worked on.

# Graphical Entity Resolution



# Graphical Bayesian ER

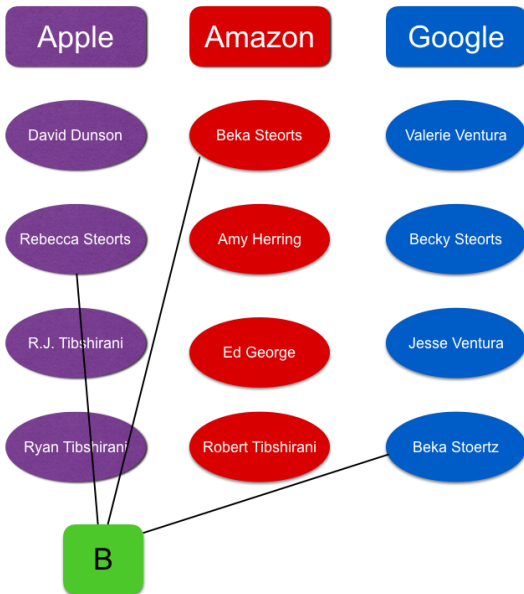
Builds off Copas and Hilton (2001), Tancredi and Liseo (2011).

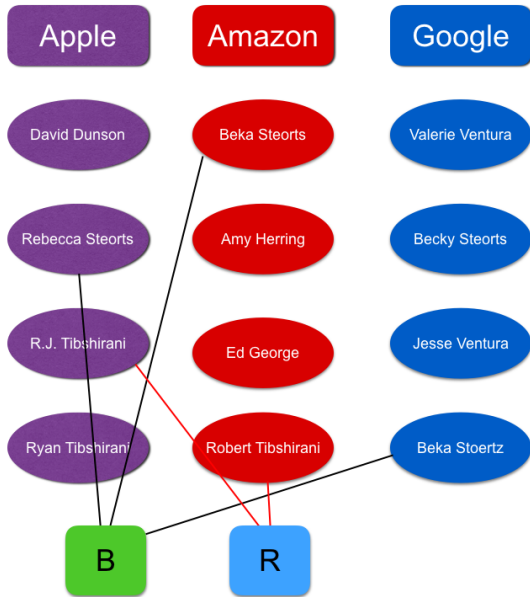


[Steorts+ 2014 AISTATS, Steorts+ 2016 JASA, Steorts 2015 BA]

# Why Graphical Bayesian ER

- 1 Handles any number of databases simultaneously
- 2 Handles both categorical and textual data
- 3 Handles missing data
- 4 Uncertainty quantification is natural
- 5 Transitive closures are nearly free
- 6 Has sound theoretical properties
- 7 Can scale to databases that contain millions of records
- 8 Generalizes to a wide variety of applications
- 9 Has equivalent or better performance than alternatives
- 10 All software is open source and freely available to non-profits





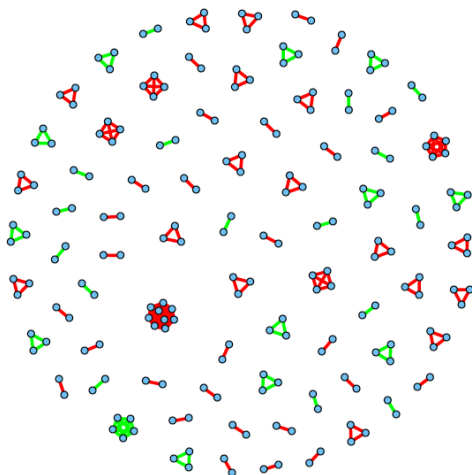


Figure 1: Removing duplications of a longitudinal medical data set (60K).

[Steorts+ JASA 2016]

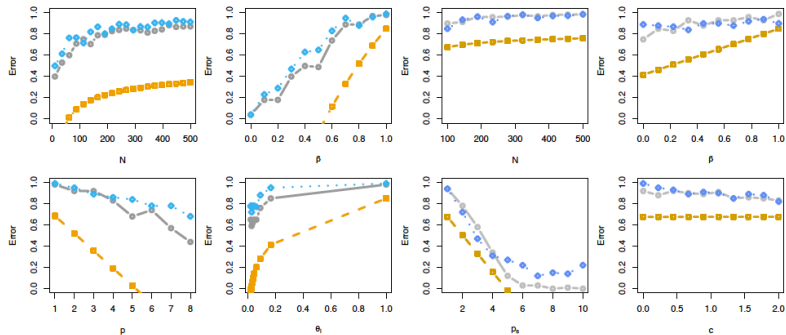


Figure 2: Bayesian ER models have tight performance bounds.

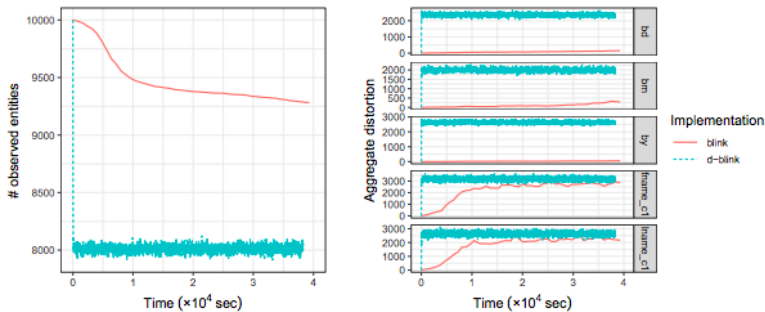
[Steorts+ AISTATS 2017]

## Our Goal

To scale Bayesian ER methods to millions of records without sacrificing accuracy and provide uncertainty of the ER task

We propose a scalable joint (Bayesian) model for blocking and performing entity resolution, where the error from this joint task is measured exactly.





**Figure 3:** Comparison of convergence rates for d-blink and blink. The summary statistics for d-blink (number of observed entities on the left and attribute distortions on the right) rapidly converge to equilibrium, while those for blink fail to converge within 11 hours.

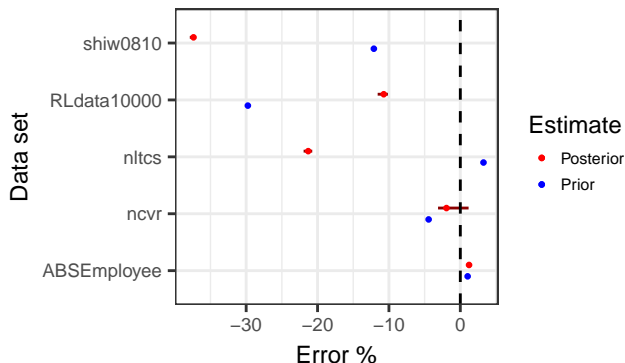
[Marchant+ (2021) JCGS]

**Table 1:** Assessment of the pairwise linkage performance for dblink and FS method as our baseline. We note that FS is supervised and does not propagate the entity resolution error exactly compared to dblink.

Data set	Method	Pairwise measure		
		Precision	Recall	F1-score
ABSEmployee	dblink	<b>0.9943</b>	<b>0.8867</b>	<b>0.9374</b>
	Fellegi-Sunter (100)	0.9964	0.9510	0.9736
	Fellegi-Sunter (10)	0.4321	0.6034	0.9736
NCVR	dblink	<b>0.9179</b>	<b>0.9654</b>	<b>0.9411</b>
	Fellegi-Sunter (100)	0.8989	0.9974	0.9456
	Fellegi-Sunter (10)	0.8989	0.9974	0.9456
NLTCs	dblink	<b>0.8363</b>	<b>0.9102</b>	<b>0.8717</b>
	Fellegi-Sunter (100)	0.7969	0.9959	0.8853
	Fellegi-Sunter (10)	0.1902	0.9999	0.3196

[Marchant+ (2021) JCGS]

# Posterior Bias Plot



**Figure 4:** Error in the posterior and prior estimates for the number of observed entities for d-blink. The results show that the posterior estimate is very sharp and typically underestimates the true number, which is consistent with Steorts, Hall, Fienberg (2016).

[Marchant+ (2021) JCGS]

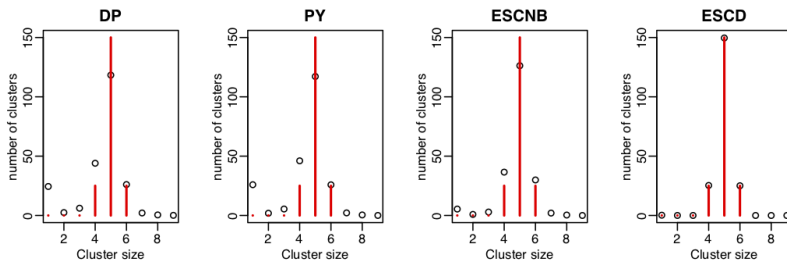
# Case Study Applied to the 2010 Decennial Census

**Table 2:** Results for ER of 2010 Census and Numident data in Wyoming. Pairwise evaluation measures are computed using ground truth identifiers available for a subset of the records, where the unadjusted count was reported to be 563,626.

Pairwise measures			Posterior population size	
Precision	Recall	F1-score	Mean	Std. error
0.97	0.84	0.90	616,000	5,000

[Marchant+ (2021) JCGS]

# The Microclustering Property



**Figure 5:** Illustrating that infinite dimensional BNPs are often misspecified for ER tasks.

[Zanella+ NIPS 2016, Betancourt+ 2021 JASA]

# Case Studies Applied to Human Rights Applications

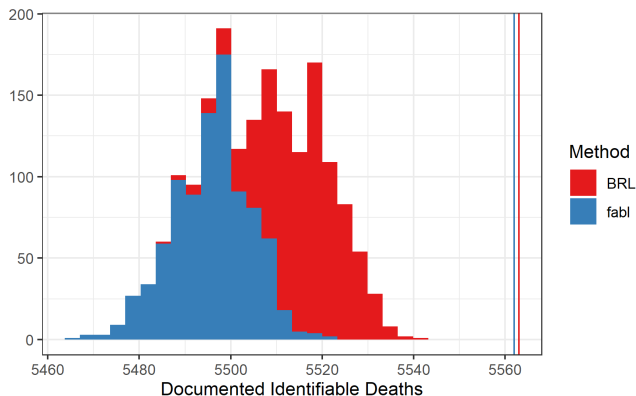


**Figure 6:** Case study of the number of casualties in Syria (March 2011 – April 2014) with standard error in real time, matching results of the Human Rights Data Analysis Group (HRDAG).

[Chen+ 2018 AoAS]



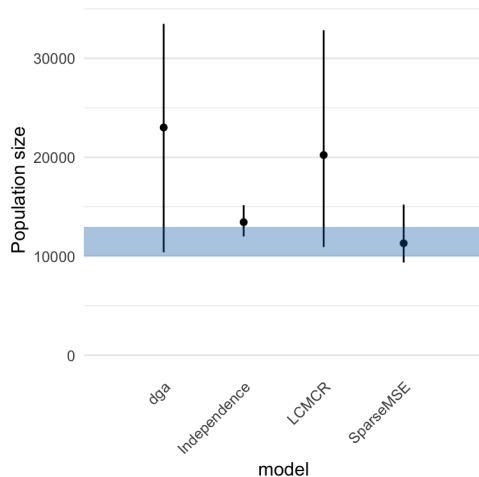
# Case Study in El Salvador



**Figure 7:** Comparison of fabl (our method) compared to Sadinle (2017), known as BRL.

[Kundinger, Reiter, Steorts (2021), In Preparation]

# Case study of human trafficking data in the UK



[Binette and Steorts (2021), Submitted]

# This Short Course

- ① An Overview of Entity Resolution (Just Completed)
- ② A Deep Dive Into Entity Resolution
- ③ Distributed Bayesian Entity Resolution
- ④ Distributed Bayesian Entity Resolution with Demos

# This Short Course

- ① An Overview of Entity Resolution (Just Completed)
- ② A Deep Dive Into Entity Resolution
- ③ Distributed Bayesian Entity Resolution
- ④ Distributed Bayesian Entity Resolution with Demos

There are other following materials that may be of interest after completion of this short course:

- ① [Binette and Steorts Review Article](#)
- ② [Entity Resolution Software](#)
- ③ [Longer Entity Resolution Tutorial](#)

Thank you!  
Questions?

Contact: [beka@stat.duke.edu](mailto:beka@stat.duke.edu)  
Webpage: [resteorts.github.io](https://resteorts.github.io)