A Primer on the Data Cleaning Pipeline

Rebecca C. Steorts*

* Rebecca C. Steorts is an Associate Professor in the Department of Statistical Science at Duke University with affiliations in Computer Science, Biostatistics and Bioinformatics, the Social Science initiate at Duke, and the Rhodes Information Initiative at Duke. Steorts is a Mathematical Statistician at the United States Census Bureau.

Corresponding author:	Rebecca C. Steorts
	Old Chemistry Hall
	Durham, North Carolina 27708
	Email: beka@stat.duke.edu, rebecca.carter.steorts@census.gov
Word count:	4048 words excluding captions and references

A Primer on the Data Cleaning Pipeline

Abstract

The availability of both structured and unstructured databases, such as electronic health data, social media data, patent data, and surveys that are often updated in real time, among others, has grown rapidly over the past decade. With this expansion, the statistical and methodological questions around data integration, or rather merging multiple data sources, has also grown. Specifically, the science of the "data cleaning pipeline" contains four stages that allow an analyst to perform downstream tasks, predictive analyses, or statistical analyses on "cleaned data." This article provides a review of this emerging field, introducing technical terminology and commonly used methods.

Statement of Significance: The article reviews the data cleaning pipeline, introducing technical terminology and commonly used methods.

1 Introduction

The availability of databases, such as electronic health data, social media data, patent databases, administrative and commercial data, geodata, digital trace data, and surveys that are often updated in real time, among others, has grown rapidly over the past decade. This expansion of data has generated widespread interest in the statistical, methodological, and ethical issues surrounding the integration (or merging) of multiple data sources (Fienberg and Manrique-Vallier, 2009; Kaplan et al., 2022; Sadinle, 2018; Aleshin-Guendel and Sadinle, 2022; Wick et al., 2009; Culotta et al., 2007; Binette and Steorts, 2022b). Application areas of interest include statistical science, computer science, engineering, and machine learning (Herzog et al., 2007; Christen, 2012a; Winkler, 2014; Jurek-Loughrey and Deepak, 2019; Christophides et al., 2019; Asher et al., 2020; Papadakis et al., 2021; Binette and Steorts, 2022a), the social sciences (McVeigh et al., 2019; Enamorado et al., 2019a), health care (Shan et al., 2022; Rogot et al., 1986; Méray et al., 2007; Jaro, 1995; Gutman et al., 2013a; Shan et al., 2020; Farley and Gutman, 2020), official statistics and survey methodology (Jaro, 1989; William E Winkler and Thibaudeau, 1990; Fortini et al., 2001; Chevrette, 2011; Dasylva et al., 2014; Dasylva, 2018), human rights statistics (Lum et al., 2013; Price et al., 2015; Sadosky et al., 2015; Sadinle, 2014, 2018; Aleshin-Guendel and Sadinle, 2022), author name disambiguation (Lai et al., 2011; Louppe et al., 2016; Zhang et al., 2018; Subramanian et al., 2021; Liu et al., 2021), and forensic science (Tai, 2018; Tai and Eddy, 2020).

The main goal of merging databases is answering statistical analyses (downstream tasks) on "cleaned data." As such, one must perform the four stages of the "data cleaning pipeline" to have more reliable conclusions. Figure 1 provides a visual representation of the pipeline. For simplicity, we consider structured data, such as name, address, and zip code. We omit unstructured data, such as paragraphs or images, from our review. For more information regarding how to handle both structured and unstructured data, we refer to Papadakis et al. (2022).

The first stage of the data cleaning pipeline, known as *schema alignment*, takes multiple data sources and forces the attributes (fields or features) to be *structured and aligned*. For example,
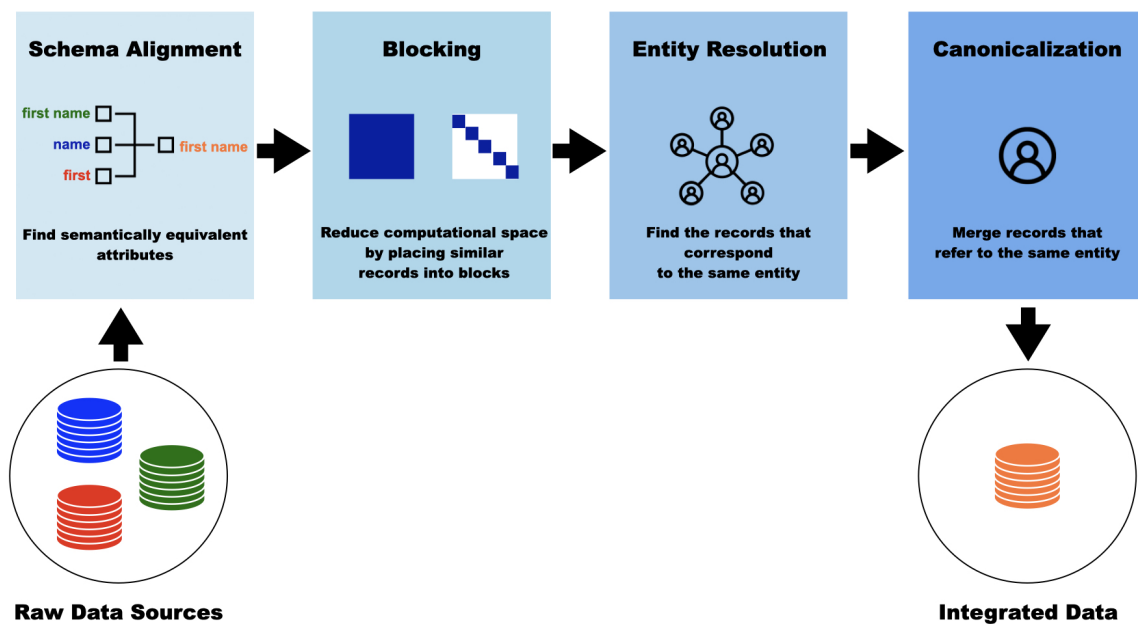
Figure 1: We illustrate the four stages of data cleaning pipeline: 1) schema alignment, 2) blocking, 3) entity resolution, and 4) canonicalization. Observe that the input to the schema alignment stage is multiple, raw data sources. The output after canonicalization is a single integrated database with unique identifiers for usage in downstream tasks and statistical analyses.

multiple data sources contain raw, messy data that needs to be converted such that the rows correspond to records and each column corresponds to an attribute.

In the second stage, the schema aligned data undergo *blocking*, which places similar records into "bins," called *blocks*. The goal of blocking is to avoid all-to-all record comparisons, which scales quadratically (with the number of records).

In the third stage, the blocked data undergo *entity resolution* (record linkage or de-duplication), which removes duplications within the blocks. The output of the entity resolution stage is clusters of records that correspond to the same entity.

In the fourth and final stage, each cluster of records that corresponds to a single entity undergoes *canonicalization*, *data fusion*, or *merging*, which identifies the most representative record in the cluster. Said differently, the canonicalization stage attempts to repair any errors from previous stages to output a single integrated data set, with a unique identifier for each record. Each stage builds upon the previous one; any errors made in a previous stage propagate into the next stage. Finally, the output from canonicalization is for a *downstream task*, which is a prediction, inference, or statistical analysis.

The main goal of the data cleaning pipeline is to be able to provide meaningful statistical analyses on the final integrated data source. Common examples of downstream tasks include multiple systems estimation (capture recapture analysis or population size estimation), small area estimation, composite estimation, or model-building exercises, such as regression. The remainder of the article focuses on reviewing the four stages of the data cleaning pipeline. Also, we review common approaches that are utilized after the data cleaning pipeline, where statistical analyses or downstream tasks are performed. We conclude with a summary.

2 A Review of the Data Cleaning Pipeline

In this section, we introduce and review the four stages of the data cleaning pipeline that allow one to perform downstream tasks and statistical analyses.

2.1 A Review of Schema Alignment

In this section, we review terminology and present simple approaches.

2.1.1 Terminology and Background

A *schema* (or relational database), which derives from the computer science literature, is an organizational structure for a database. In the context of this problem, a schema consists of all attributes for the input data sources. Disparate data sources may use different naming conventions for records that refer to the same attribute. For instance, the schema alignment stage attempts to identify that the attributes "first name," "name," and "first" all refer to a global (or common) attribute "first name." This stage replaces the aforementioned attributes with the global attribute "first name." Therefore, *schema alignment* is essential when the multiple input data sources lack attribute agreement.

More formally, this stage attempts to find "semantically equivalent attributes" (Papadakis et al., 2021) and/or leverage attribute information. For example, attribute mappings between the data sources provide information regarding "transformations, mappings, or rules" (Papadakis et al., 2021). For instance, it is beneficial to learn mappings or correspondences between acronyms of attributes, such as "North Carolina" versus "NC." Other examples include: "Street" versus "St." or "first" versus "1st."

Methods using a schema are *schema-based* or *schema-aware*. On the other hand, data sources that lack or ignore this information are *schema agnostic* or *schema-independent*. In many applications, an analyst may receive data that has already been schema aligned, and thus, would

lack information regarding the original raw data sources. In these cases, one would not need to perform this stage.

2.1.2 Approaches for Schema Alignment

Schema alignment requires the analyst to identify all attributes that lack agreement and identify a common naming convention. Next, the analyst updates the common naming convention, such as manually. While this approach is simple, it is not reproducible or easily documented. Importantly, given its manual nature, mistakes can easily occur. Figure 2 provides two disparate schema needing attribute alignment. We create a mapping between first/last and name. In addition, we map gender and sex.

profile	name	address	gender	state
d1	Alan Smith	123 Main Street	M	NC
d2	Alan Smith	123 Main Street	M	NC
d3	Ann Waters	155 Green Way	F	NC
d4	Anne Waters	155 Green Way	F	NC
d5	Sally Glines	18 Court Road	F	NC
d6	Matt Box	1871 Red Drive	M	NC
d7	Joe Smith	2971 Orchard Court	M	NC
d8	Joe Smith	2971 Orchard Court	M	NC
d9	Joe Smith	2971 Orchard Court	M	NC
d10	Joe Smith	2971 Orchard Court	M	NC

profile	first	last	sex	state	age
s1	Alan T.	Smith	M	NC	50
s2	Matt	Box	M	NC	
s3	Sammy	Smith	M	NC	23
s4	Sally	Glines	F	NC	
s5	Joe	Green	M	NC	34

(a)

Entity 1
d1 d2   s1

Entity 2
d3 d4

Entity 3
d5   s4

Entity 4
d6   s2

Entity 5
d7 d8 d9 d10

Entity 6
s3

Entity 7
s5

(b)

Figure 2: (a): Two disparate schema. (b): The corresponding profiles (or records) that correspond to each unique entity.

As already mentioned, some methods learn attribute mappings between the data sources, which can provide information regarding "transformations, mappings, or rules" of the attributes (Winkler et al., 2001; Tejada et al., 2002). Both Winkler et al. (2001) and Tejada et al. (2002) recommend rule-based systems, where the rules can differ depending on the linked data sources. For

example, a simple rule may require attribute alignment if the attributes are semantically equivalent. A concrete example of such a rule is to recognize that "first name," "name," and "first" all identify a person's first name. While rules or more automated methods can be utilized, they will always rely on manual review of an analyst. No automated system is perfect; it may be difficult to find errors for complex rules.

Many proposed methods for performing schema alignment exist within the computer science, machine learning, and database management literatures. As already mentioned, we only consider structured data. In the case of both structured and unstructured data, the first stage requires *schema refinement* (instead of alignment). We refer to Papadakis et al. (2021) for an introduction to this extension.

2.2 A Review of Blocking

In this section, we briefly review blocking. Blocking places similar records into "bins," providing a dimensionality reduction. The blocks can be overlapping or non-overlapping. Note that blocking does not refer to other definitions in other areas, such as in clinical trials or analysis of variance.

Typically, blocking operates (internally) in a schema-aware manner. As such, this step assumes that the data sources are schema'aligned. Thus, analysts must decide the most useful attributes that will lead to representative signatures, known as *blocking keys*. Users place similar or identical blocking keys into the same partition or bin (Christen, 2012b; Steorts et al., 2014b; Murray, 2016; Papadakis et al., 2021).

Formally, given data sources $d_1, \ldots, d_k$ a blocking method (with a key) B^{key} generates a set of blocks $\{b_1^{\text{key}}, \ldots, b_n^{\text{key}}\}$. Each blocking key b_i^{key} is a subset of records (or profiles) from the data sources that have the same (or similar) value for key. That is, $b_i^{\text{key}} \subset \cup_{j=1}^k d_j$. Papadakis et al. (2021) provides illustrations regarding blocking keys.

The simplest method for constructing blocks rely on attributes (e.g., gender or zip code). *Traditional blocking* methods place records in the same block if and only if they agree on the blocking attributes. In addition, traditional blocking, while computationally efficient, makes an unrealistic assumption that the blocking keys are error free (Steorts et al., 2014b; Papadakis et al., 2021). We can form more complex rules based upon the attributes that involve unions and intersections of the attributes, known as *conjunctions* (or disjunctions of conjunctions). These methods require an expert to define a transformation of one or more attributes as the blocking key of each record that occurs in a data source (or profile).

On the other hand, *probabilistic approaches* do not assume part of the attributes are fixed and error-free. Instead, these approaches utilize the entire records, resulting in a more data-driven approach. This involves replacing each record by a multi-set of length k contiguous substrings, called *k-grams*, *k-shingles*, or *k-tokens*. Figure 2.2 provides a simple illustration of one data source, where we consider two blocking criteria. See Steorts et al. (2014b); Papadakis et al. (2021) for more advanced probabilistic blocking.

One of the most widely used blocking methods is locality sensitive hashing (LSH), which is a probabilistic method for dimension reduction that has strong theoretical properties. Specifically, LSH is a quick way of finding approximate nearest neighbors, which is one way of creating blocks. Many variants of LSH for blocking and are discussed in the previous references. Finally, supervised, unsupervised, and hybrid approaches are appropriate for blocking; however, this is not optimal procedure for all applications (Christen, 2012b; Papadakis et al., 2021).

2.3 A Review of Entity Resolution

In this section, we review comparison-based methods (Fellegi-Sunter and supervised learning methods) and unsupervised learning methods (comparison- and direct-based methods).

profile	name	address	gender	state
d1	Alan Smith	123 Main Street	M	NC
d2	Alan Smith	123 Main Street	M	NC
d3	Ann Waters	155 Green Way	F	NC
d4	Anne Waters	155 Green Way	F	NC
d5	Sally Glines	18 Court Road	F	NC
d6	Matt Box	1871 Red Drive	M	NC
d7	Joe Smith	2971 Orchard Court	M	NC
d8	Joe Smith	2971 Orchard Court	M	NC
d9	Joe Smith	2971 Orchard Court	M	NC
d10	Joe Smith	2971 Orchard Court	M	NC

(a)

id	key
d1	thetM
d2	thetM
d3	rsayF
d4	rsayF
d5	esadF
d6	oxveM
d7	thrtM
d8	thrtM
d9	thrtM
d10	thrtM

(b)

id	key
d1	thet, hetM
d2	thet, hetM
d3	rsay, sayF
d4	rsay, sayF
d5	esad, sadF
d6	oxve, xveM
d7	thrt, hrtM
d8	thrt, hrtM
d9	thrt, hrtM
d10	thrt, hrtM

(c)

thet, hetM
d1 d2

rsay, sayF
d3 d4

thrt, hrtM
d7 d8 d9 d10

(d)

Figure 3: (a): Toy illustration of a data source with the standard blocking criterion bolded. (b): Standard blocking keys. (c): 4-grams based upon the standard blocking in (b). (d): The blocks from 4-grams, which are the same as for standard blocking with the exception that the blocking keys are different.

9

2.3.1 Fellegi-Sunter Based Methods

The earliest published work is that of Dunn (1946) who believed that it was important to document record keeping for official statistical agencies. Roughly ten years later, Newcombe et al. (1959) proposed one of the first algorithms for entity resolution. Fellegi and Sunter (1969) formalized the work of Newcombe et al. (1959) via a likelihood ratio test based upon agreement patterns of record pairs and classified the records as *links*, *non-links*, or *possible-links*. Their method — the Fellegi-Sunter (FS) method — was shown to be optimal under statistical decision theory. The authors provided two variants of their method, one exact and one approximate via the expectation–maximization algorithm, where the former is more scalable and practical. The FS method is one of the most popular methods utilized in statistical entity resolution (and other fields) due to its simplicity, statistical optimality, and scalability. Moreover, researchers have extended it from both methodological and computational viewpoints (Winkler, 2006; Sadinle and Fienberg, 2013; Enamorado et al., 2019b; McVeigh et al., 2019; Sadinle, 2014, 2018; Aleshin-Guendel and Sadinle, 2022; Tang et al., 2020; Armstrong and Mayda, 1992; Winkler, 1992, 1993; Thibaudeau, 1993; Smith and Newcombe, 1975; Belin, 1990; Armstrong and Mayda, 1992; Belin and Rubin, 1995; Nigam et al., 2000; Larsen and Rubin, 2001; Lahiri and Larsen, 2005; Wortman, 2019).

2.3.2 Supervised Entity Resolution

A supervised learning method requires a comparison vector consisting of record pairs — the same input as required as the FS method. Many packages have implemented supervised learning methods that require a binary comparison vector. More specifically, supervised learning methods require training data and testing data. The training data may come from manually labeled data of record pairs in the form of match or non-match. The user builds a classifier (using the training data) that predicts the match or non-match status (on the testing data). The resulting output is the match and non-match status for all record pairs.

Supervised entity resolution employs training data to develop classification methods (Trajten-

berg and Shiff, 2008; Azoulay et al., 2012; Bailey et al., 2017) or crowdsourcing (Sarawagi and Bhamidipaty, 2002; Wang et al., 2012; Vesdapunt et al., 2014; Gokhale et al., 2014; Frisoli et al., 2019). Fully supervised methods utilize large amounts of training data, which contrasts semi-supervised methods, which often assume a minimal amount of training data. Examples of supervised models include logistic regression, random forests, support vector machines, Bayesian additive regression trees, among other types of methods that originated in the machine learning literature (Hastie et al., 2001). Neural networks are popular for supervised entity resolution (Gottapu et al., 2016; Ebraheem et al., 2017, 2018; Mudgal et al., 2018; Kooli et al., 2018; Kasai et al., 2020; Barlaug and Gulla, 2021). For more information on supervised learning methods, we refer to Papadakis et al. (2021).

The aforementioned methods are useful when the training data is extremely accurate. Furthermore, these methods scale to extremely large databases. Supervised learning methods are sensitive to tuning parameters and the amount of training data needed, which can be seen using standard evaluation metrics used in the literature such as recall (true positive rate or sensitivity) and precision (positive predictive value). In addition, using too much training data can lead to biases in predictions or inferences downstream. Finally, most classification methods do not quantify uncertainty of the entity resolution task into later statistical analyses.

2.3.3 Unsupervised Entity Resolution

In this section, we review the two main approaches in the literature to unsupervised (or Bayesian) entity resolution. First, we review Bayesian FS methods, which extend the seminal work of Fellegi and Sunter (1969) to an unsupervised setting. This framework traditionally models the comparisons of the attributes between pairs of records through a mixture model (Larsen and Rubin, 2001; Jaro, 1989). This is known as the *comparison-based framework* (Sadinle, 2014, 2017, 2018; Aleshin-Guendel and Sadinle, 2022).

Second, we review unsupervised entity resolution methods, where the goal is to cluster records

11

to an unknown, latent entity. These methods directly model the attribute-level distortion, noise, and/or transcription errors of a record (Matsakis, 2010; Liseo and Tancredi, 2011; Tancredi and Liseo, 2011; Liseo and Tancredi, 2013; Gutman et al., 2013b; Steorts et al., 2014a; Tancredi and Liseo, 2015; Steorts, 2015; Steorts et al., 2016, 2018; Dalzell and Reiter, 2018; Tancredi et al., 2020; Steorts et al., 2017; Tang et al., 2020; Kaplan et al., 2018; Marchant et al., 2021). This is known as the *direct-modeling framework*.

Comparison-Based Methods   Comparison-based methods of record pairs became popular with the framework of Fellegi and Sunter (1969). This work led to many extensions previously mentioned. In this section, we review unsupervised (or Bayesian) extensions. Sadinle (2014) extended Fellegi and Sunter (1969), considering the case of *duplicate detection* or *de-duplication*, or the process of removing duplicate records within a single database.

One extension made was providing transitive closures regarding the match status. (Note that a linkage is transitive in the following sense: if records A and B are the same individual, and records B and C are the same individual, then records A and C must be the same individual as well.) The coreference matrix is a "partition of the data file that groups co-referent records together" (Sadinle, 2014). In addition, Sadinle (2014) included prior information, if available, on the match and non-match probabilities. This provided a posterior distribution on the co-reference matrix (or linkage structure), which provides a partition for representing linkages of records across and within databases. For example, assume two databases with record i in the first database and record j in the second database. More formally, define the entries of the co-reference matrix such that $c_{i,j} = 1$ if records i and j are the same record. Otherwise, $c_{i,j} = 0$. Furthermore, the proposed Bayesian methodology allows for uncertainty quantification regarding the match and non-match status.

Sadinle (2017) considered the setting of *bipartite record linkage*, where the goal is to merge two databases. Duplications could occur across databases but not within databases. This assumption

imposed is called a *one-to-one restriction* regarding the linkage across the two databases. In short, a record in the first database can match a maximum of one time to a record in the second database. Fellegi and Sunter (1969) (and most extensions) ignored this assumption. Sadinle (2017) proposed incorporating missing values handling partial agreements. To our knowledge, Sadinle (2017) provided the first extensions using generalized loss functions that allow for a rejection option, providing an alternative to the traditional FS decision rule.

Sadinle (2018) extended the bipartite record linkage framework to include the case of entity resolution, called joint duplicate detection and record linkage. This provided a general framework for duplications within and across all databases. *Linkage averaging* was introduced and propagates the uncertainty of the record linkage method to a downstream task, such as multiple systems estimation. The proposed methodology is illustrated on a case study for human rights violations in El Salvador.

Aleshin-Guendel and Sadinle (2022) considered Bayesian entity resolution models, where a structured prior is proposed on the coreference matrix. For more information, we refer to details in the paper regarding the construction of the prior distribution. Similar to earlier work, Aleshin-Guendel and Sadinle (2022) derived loss functions with an abstain option and illustrated their proposed methodology on real data and simulation studies.

Direct-Modeling Based Methods   Other approaches assume a model at the attribute-level of each record. These models cluster observed records (at the attribute-level) to a latent (unknown) entity, which avoids all-to-all record comparisons. A *linkage structure* partitions co-referent (or duplicated) records into the same group. (The linkage structure is also a co-reference matrix.) A user must define a likelihood for the observed records (at the attribute-level) in the data sources. One must place priors on any unknown parameters, such as the linkage structure. Such a specification provides for a posterior distribution on the linkage structure (Matsakis, 2010; Liseo and Tancredi, 2011; Tancredi and Liseo, 2011; Liseo and Tancredi, 2013; Gutman et al., 2013b;

Steorts et al., 2014a; Steorts, 2015; Steorts et al., 2016).

Many models assume the records contain distortions, noise, and transcription errors. As such, the likelihood function handles these "distortions." One way to model the distortion of the records is to use a latent variable model, where the distortion is a latent (or unknown, random) variable. For example, Tancredi and Liseo (2011) adapted the hit and miss model of Copas and Hilton (1990) to entity resolution, proposing a measurement error model for distortions.

In addition to the assumptions above, some researchers proposed that the attributes were only categorical-valued (Matsakis, 2010; Tancredi and Liseo, 2011; Steorts et al., 2014a, 2016). On the other hand, Tancredi and Liseo (2011) proposed continuous attributes under a normality assumption. Steorts (2015) extended this prior work to include both categorical and string-based attributes. Tancredi and Liseo (2011); Steorts et al. (2014a, 2016) provided for Bayesian decision rules under a given loss function. This approach allows for uncertainty quantification regarding the linkage structure (or match and non-match status) and functions of the linkage structure. One major limitation of both direct- and comparison-based methods has been the ability to scale to moderate to large databases.

Marchant et al. (2021) proposed one way to scale to larger databases containing over a million total records. Specifically, they proposed a joint blocking and entity resolution framework, extending the work of Steorts (2015). They proposed an *auxiliary variable representation* of Steorts (2015) which partitions the latent entities and records into blocks. The work of Marchant et al. (2021) differs from typical blocking in the literature as the assignments of latent entities and records to blocks are *random* and *jointly inferred* with other parameters of the model. Furthermore, they formally proved that the auxiliary variable representation preserves the marginal posterior distribution over the model parameters. This result implies that any inferences are independent of the blocking criteria. They utilized a distributed partially-collapsed Gibbs sampler to perform inference and other computational optimizations, where they investigated their model on many data sets as well as a case study from the United States Census Bureau.

This prior work has led to interesting discoveries both in clustering and in entity resolution. The *microclustering property* describes the sub-linear growth of entity resolution (and other clustering tasks, such as community detection). The size of the clusters grow sub-linearly as the total number of records grow (Zanella et al., 2016; Betancourt et al., 2022b,a). Therefore, applying a Bayesian nonparametric (BNP) model with a large number of data points (records) might tend to favor clusters comprising many entities, which makes little sense in the context where each cluster should correspond to a single true entity. Betancourt et al. (2022b) proposed a framework that overcomes many limitations of Zanella et al. (2016), namely a lack of interpretability, identifiability, and full characterization of model asymptotic properties. These papers all proposed subjective priors on the linkage structure that satisfy the microclustering property.

In addition, Betancourt et al. (2022a) proposed subjective priors on the linkage structure using *allelic partitions*. Betancourt et al. (2022a) stated that "allelic partitions are an equivalent representation of partitions which summarizes the number of clusters of each size." They considered the microclustering property in a potentially more natural manner. They provided both real and simulation studies illustrating their proposal are similar to previous findings. All the microclustering papers used the *chaperones algorithm*, which may perform more efficiently than traditional Gibbs sampling (Zanella et al., 2016).

Takeaways   We have reviewed entity resolution methods in the literature that may be useful to analysts and tried to point out varying flavors. Supervised methods are the most useful when users have access to training data that is accurate. Unsupervised methods are most useful in situations when users do not have access to unique identifiers, lack confidence in training data, and/or care about uncertainty quantification. The choice of when to use an entity resolution method is application specific. For example, in situations where unique identifiers are not very reliable, analysts would prefer unsupervised or semi-supervised methods over fully supervised methods.

2.4 Canonicalization

In this section, we review the fourth state of the pipeline, which is *canonicalization*, *data fusion*, or *merging*. This stage attempts to repair any errors from the entity resolution task to create a single integrated data set. The output of this stage is a single integrated data set such that statistical inferences, predictions, and analyses can be performed. It is important to note that each record in the integrated data set is assumed to be unique (with a corresponding unique identifier).

Early methods of canonicalization were rule-based methods as they are easy to implement and scale to large databases (Cohen and Sagiv, 2005). Much of the literature assumes training data is available to select the most representative record. In many proposals, authors considered decision theory techniques or supervised learning methods (Yan and Ozsu, 1999; Bohannon et al., 2005; Culotta et al., 2007). See Bleiholder and Naumann (2009) for a comprehensive review. In more recent work, Kaplan et al. (2022) proposed canonicalization methods to choose an integrated (or representative) data set from linked data. The methods are agnostic to the linked method and provide natural uncertainty quantification. They applied their methods to a pipeline approach of blocking, entity resolution, canonicalization, and statistical analysis on a case study of voter registration, illustrating the importance of most of the pipeline approach and its inner workings.

3 Summary

The ultimate goal after the data cleaning pipeline is to a perform downstream task or statistical analysis on the integrated data set. The goal of this article is to provide an introduction to the entire data cleaning pipeline in a non-technical manner.

The data cleaning pipeline consists of four stages – schema alignment, blocking, entity resolution, and canonicalization. First, schema alignment takes multiple data sources and ensures they are in the same cleaned, tabular format. Second, the aligned schema undergoes blocking, which reduces the dimensionality of the problem. Blocks contain similar records. Third, the analyst performs en-

tity resolution on the blocked data, removing duplications. This stage provides clusters of records that correspond to the same person or object. The fourth stage identifies the most representative record (after entity resolution) to form a single integrated data source (with unique identifiers). This step is essential for predictive and statistical analyses.

References

Aleshin-Guendel, S. and Sadinle, M. "Multifile partitioning for record linkage and duplicate detection." Journal of the American Statistical Association, 1–10 (2022).

Armstrong, J. and Mayda, J. "Estimation of Record Linkage Models Using Dependent Data." In Proceedings of the Section on Survey Research Methodology, 853 – 858. American Statistical Association (1992).

Asher, J., Resnick, D., Brite, J., Brackbill, R., and Cone, J. "An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries." International journal of environmental research and public health, 17(18):6937 (2020).

Azoulay, P., Zivin, J. S. G., and Sampat, B. N. "The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine." In Lerner, J. and Stern, S. (eds.), The Rate and Direction of Inventive Activity Revisited. University of Chicago Press (2012).

Bailey, M., Cole, C., Henderson, M., and Massey, C. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." National Bureau of Economic Research Working Paper Series (2017).

Barlaug, N. and Gulla, J. A. "Neural Networks for Entity Matching: A Survey." ACM Transactions on Knowledge Discovery from Data, 15(3):1–37 (2021).

Belin, T. R. "A Proposed Improvement in Computer Matching Techniques." In Statistics of Income and Related Administrative Record Research, 167–172. Washington, DC, USA: International Revenue Service (1990).

Belin, T. R. and Rubin, D. B. "A Method for Calibrating False-Match Rates in Record Linkage." Journal of the American Statistical Association, 90(430):694–707 (1995).

Betancourt, B., Sosa, J., and Rodríguez, A. "A prior for record linkage based on allelic partitions." Computational Statistics & Data Analysis, 172:107 – 474 (2022a).

Betancourt, B., Zanella, G., and Steorts, R. C. "Random partition models for microclustering tasks." Journal of the American Statistical Association, 117(539):1215–1227 (2022b).

Binette, O. and Steorts, R. C. "(Almost) all of entity resolution." Science Advances, 8(12):eabi8021 (2022a).

—. "On the reliability of multiple systems estimation for the quantification of modern slavery." Journal of the Royal Statistical Society Series A, 185(2):640–676 (2022b).

Bleiholder, J. and Naumann, F. "Data Fusion." ACM Computing Surveys, 41(1):1–41 (2009).

Bohannon, P., Fan, W., Flaster, M., and Rastogi, R. "A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification." In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, 143–154 (2005).

Chevrette, A. "G-LINK : A Probabilistic Record Linkage System." Technical report, Statistics Canada (2011).

Christen, P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag (2012a).

—. "A survey of indexing techniques for scalable record linkage and deduplication." IEEE Transactions on Knowledge and Data Engineering, 24(9):1537–1555 (2012b).

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. "End-to-End Entity Resolution for Big Data: A Survey." arXiv e-prints (2019). arXiv:1905.06397.

Cohen, S. and Sagiv, Y. "An Incremental Algorithm for Computing Ranked Full Disjunctions." In Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium

on Principles of Database Systems, 98–107. New York, NY, USA: Association for Computing Machinery (2005).

Copas, J. B. and Hilton, F. J. "Record Linkage: Statistical Models for Matching Computer Records." Journal of the Royal Statistical Society. Series A (Statistics in Society), 153(3):287–320 (1990).

Culotta, A., Wick, M., Hall, R., Marzilli, M., and McCallum, A. "Canonicalization of Database Records Using Adaptive Similarity Measures." In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 201–209. New York, NY, USA: Association for Computing Machinery (2007).

Dalzell, N. M. and Reiter, J. P. "Regression Modeling and File Matching Using Possibly Erroneous Matching Variables." Journal of Computational and Graphical Statistics, 0(0):1–11 (2018).

Dasylva, A. "Pairwise Estimating Equations for the Primary Analysis of Linked Data." Proceedings of Statistics Canada Symposium (2018).

Dasylva, A., Titus, R.-C., and Thibault, C. "Overcoverage in the 2011 Canadian Census." In Proceedings of Statistics Canada Symposium (2014).

Dunn, H. L. "Record Linkage." American Journal of Public Health and the Nation's Health, 36(12):1412–1416 (1946).

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., and Tang, N. "DeepER – Deep Entity Resolution." arXiv e-prints (2017). arXiv:1710.00597.

—. "Distributed representations of tuples for entity resolution." Proceedings of the VLDB Endowment, 11(11):1454–1467 (2018).

Enamorado, T., Fifield, B., and Imai, K. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." American Political Science Review, 113:353–371 (2019a).

—. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." American Political Science Review, 113(2):353–371 (2019b).

Farley, E. and Gutman, R. "A Bayesian Approach to Linking Data Without Unique Identifiers." arXiv e-prints (2020). arxiv:2012.00601.

Fellegi, I. P. and Sunter, A. B. "A Theory for Record Linkage." Journal of the American Statistical Association, 64(328):1183–1210 (1969).

Fienberg, S. E. and Manrique-Vallier, D. "Integrated methodology for multiple systems estimation and record linkage using a missing data formulation." AStA Advances in Statistical Analysis, 93(1):49–60 (2009).

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. "On Bayesian Record Linkage." Research in Official Statistics, 4(1):185–198 (2001).

Frisoli, K., LeRoy, B., and Nugent, R. "A Novel Record Linkage Interface That Incorporates Group Structure to Rapidly Collect Richer Labels." In 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 580–589 (2019).

Gokhale, C., Das, S., Doan, A., Naughton, J. F., Rampalli, N., Shavlik, J., and Zhu, X. Corleone: Hands-off crowdsourcing for entity matching, 601–612 (2014).

Gottapu, R. D., Dagli, C., and Ali, B. "Entity Resolution Using Convolutional Neural Network." Procedia Computer Science, 95:153–158 (2016).

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. "A Bayesian procedure for file linking to analyze end-of-life medical costs." Journal of the American Statistical Association, 108(501):34–47 (2013a).

—. "A Bayesian procedure for file linking to analyze end-of-life medical costs." Journal of the American Statistical Association, 108(501):34–47 (2013b).

Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer (2001).

Herzog, T., Scheuren, F., and Winkler, W. Data Quality and Record Linkage Techniques. New York: Springer (2007).

Jaro, M. A. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." Journal of the American Statistical Association, 84(406):414–420 (1989).

—. "Probabilistic linkage of large public health data files." Statistics in Medicine, 14(5-7):491–498 (1995).

Jurek-Loughrey, A. and Deepak, P. Semi-supervised and Unsupervised Approaches to Record Pairs Classification 55–78. Cham: Springer International Publishing (2019).

Kaplan, A., Betancourt, B., and Steorts, R. C. "Posterior Prototyping: Bridging the Gap between Bayesian Record Linkage and Regression." arXiv e-prints (2018). arXiv:1810.01538.

—. "A Practical Approach to Proper Inference with Linked Data." The American Statistician, 1–10 (2022).

Kasai, J., Qian, K., Gurajada, S., Li, Y., and Popa, L. "Low-Resource Deep Entity Resolution with Transfer and Active Learning." ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, 5851–5861 (2020).

Kooli, N., Allesiardo, R., and Pigneul, E. "Deep Learning Based Approach for Entity Resolution in Databases." In Nguyen, N. T., Hoang, D. H., Hong, T.-P., Pham, H., and Trawiński, B. (eds.), Intelligent Information and Database Systems, 3–12. Cham: Springer International Publishing (2018).

Lahiri, P. and Larsen, M. "Regression Analysis With Linked Data." Journal of the American Statistical Association, 100(469):222–230 (2005).

Lai, R., D'Amour, A., Yu, A., Sun, Y., Torvik, V., and Fleming, L. "Disambiguation and Co-Authorship Networks of the US Patent Inventor Database." Harvard Institute for Quantitative Social Science, Cambridge, MA, 2138 (2011).

Larsen, M. D. and Rubin, D. B. "Iterative Automated Record Linkage Using Mixture Models." Journal of the American Statistical Association, 96(453):32–41 (2001).

Liseo, B. and Tancredi, A. "Bayesian estimation of population size via linkage of multivariate Normal data sets." Journal of Official Statistics, 27(3):491–505 (2011).

—. "Some advances on Bayesian record linkage and inference for linked data." (2013).
URL http://www.ine.es/e/essnetdi_ws2011/ppts/Liseo_Tancredi.pdf

Liu, X., Yin, D., Zhang, X., Su, K., Wu, K., Yang, H., and Tang, J. "OAG-BERT: Pre-train Heterogeneous Entity-augmented Academic Language Models." arXiv e-prints (2021). arXiv:2103.02410.

Louppe, G., Al-Natsheh, H. T., Susik, M., and Maguire, E. J. "Ethnicity sensitive author disambiguation using semi-supervised learning." international conference on knowledge engineering and the semantic web, 272–287 (2016).

Lum, K., Price, M. E., and Banks, D. "Applications of multiple systems estimation in human rights research." The American Statistician, 67(4):191–200 (2013).

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I., and Steorts, R. C. "d-blink: Distributed end-to-end Bayesian entity resolution." Journal of Computational and Graphical Statistics, 30(2):406–421 (2021).

Matsakis, N. E. "Active Duplicate Detection with Bayesian Nonparametric Models." Ph.D. thesis, Massachusetts Institute of Technology (2010).

McVeigh, B. S., Spahn, B. T., and Murray, J. S. "Scaling Bayesian Probabilistic Record Linkage

with Post-Hoc Blocking: An Application to the California Great Registers." arXiv e-prints (2019). arXiv:1905.05337.

Méray, N., Reitsma, J. B., Ravelli, A. C., and Bonsel, G. J. "Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number." Journal of Clinical Epidemiology, 60(9):883 – 891 (2007).

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. "Deep Learning for Entity Matching: A Design Space Exploration." In Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, 19–34. New York, NY, USA: ACM (2018).

Murray, J. S. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." Journal of Privacy and Confidentiality, 7(1):3–24 (2016).

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. "Automatic Linkage of Vital Records." Science, 130(3381):954–959 (1959).

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. "Text Classification From Labeled and Unlabeled Documents Using EM." Machine Learning (2000).

Papadakis, G., Ioannou, E., Thanos, E., and Palpanas, T. The Four Generations of Entity Resolution. Morgan & Claypool Publishers (2021).

Price, M., Gohdes, A., and Ball, P. "Documents of war: Understanding the Syrian conflict." Significance, 12(2):14–19 (2015).

Rogot, E., Sorlie, P., and Johnson, N. J. "Probabilistic Methods in Matching Census." Journal of Chronic Diseases, 39(9):719–734 (1986).

Sadinle, M. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." The Annals of Applied Statistics, 8(4):2404–2434 (2014).

—. "Bayesian estimation of bipartite matchings for record linkage." Journal of the American Statistical Association, 112(518):600–612 (2017).

—. "Bayesian Propagation of Record Linkage Uncertainty Into Population Size Estimation of Human Rights Violations." The Annals of Applied Statistics, 12(2):1013–1038 (2018).

Sadinle, M. and Fienberg, S. E. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." Journal of the American Statistical Association, 108(502):385–397 (2013).

Sadosky, P., Shrivastava, A., Price, M., and Steorts, R. C. "Blocking Methods Applied to Casualty Records from the Syrian Conflict." arXiv e-prints, 1–25 (2015). arXiv:1510.07714.

Sarawagi, S. and Bhamidipaty, A. "Interactive Deduplication Using Active Learning." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 269–278 (2002).

Shan, M., Thomas, K., and Gutman, R. "A Bayesian Multi-Layered Record Linkage Procedure to Analyze Functional Status of Medicare Patients with Traumatic Brain Injury." arXiv e-prints (2020). arxiv:2005.08549.

Shan, M., Thomas, K. S., and Gutman, R. "A Bayesian multilayer record linkage procedure to analyze postacute care recovery of patients with traumatic brain injury." Biostatistics (2022).

Smith, M. E. and Newcombe, H. B. "Methods for Computer Linkage of Hospital Admission Separation Records Into Cumulative Health Histories." Methods of Information in Medicine, 14(3):118–125 (1975).

Steorts, R. C. "Entity Resolution with Empirically Motivated Priors." Bayesian Analysis, 10(4):849–875 (2015).

Steorts, R. C., Barnes, M., and Neiswanger, W. "Performance Bounds for Graphical Record Linkage." In Singh, A. and Zhu, J. (eds.), Proceedings of the 20th International Conference on

Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, 298–306. Fort Lauderdale, FL, USA: PMLR (2017).

Steorts, R. C., Hall, R., and Fienberg, S. E. "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of Machine Learning Research*, 33:922–930 (2014a).

—. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association*, 111(516):1660–1672 (2016).

Steorts, R. C., Tancredi, A., and Liseo, B. "Generalized Bayesian Record Linkage and Regression with Exact Error Propagation." In *International Conference on Privacy in Statistical Databases*, 297–313. Springer (2018).

Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. "A Comparison of Blocking Methods for Record Linkage." In *Privacy in Statistical Databases*, Lecture Notes in Computer Science, 253–268. Springer, Cham (2014b).

Subramanian, S., King, D., Downey, D., and Feldman, S. "S2AND: A Benchmark and Evaluation System for Author Name Disambiguation." *arXiv e-prints* (2021). arXiv:2103.07534.

Tai, X. H. "Record linkage and matching problems in forensics." *IEEE International Conference on Data Mining Workshops*, 510–517 (2018).

Tai, X. H. and Eddy, W. F. "Automatically matching topographical measurements of cartridge cases using a record linkage framework." *arXiv e-prints* (2020). arXiv:2003.00060.

Tancredi, A. and Liseo, B. "A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems." *Annals of Applied Statistics*, 5(2B):1553–1585 (2011).

—. "Regression analysis with linked data: problems and possible solutions." *Statistica*, 75(1):19–35 (2015).

Tancredi, A., Steorts, R., and Liseo, B. "A Unified Framework for De-Duplication and Population Size Estimation (with Discussion)." *Bayesian Analysis*, 15(2):33–682 (2020).

Tang, J., Reiter, J., and Steorts, R. C. "Bayesian Modeling for Simultaneous Regression and Record Linkage." In Privacy in Statistical Databases, Lecture Notes in Computer Science. Springer, Cham (2020).

Tejada, S., Knoblock, C. A., and Minton, S. "Learning domain-independent string transformation weights for high accuracy object identification." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 350–359 (2002).

Thibaudeau, Y. "The Discrimination Power of Dependency Structures in Record Linkage." Survey Methodology, 19(1) (1993).

Trajtenberg, M. and Shiff, G. "Identification and mobility of Israeli patenting inventors." Technical report (2008).

Vesdapunt, N., Bellare, K., and Dalvi, N. "Crowdsourcing Algorithms for Entity Resolution." Proceedings of the VLDB Endowment, 7(12):1071–1082 (2014).

Wang, J., Kraska, T., Franklin, M. J., and Feng, J. "CrowdER: Crowdsourcing Entity Resolution." Proceedings of the VLDB Endowment, 5(11):1483–1494 (2012).

Wick, M., Culotta, A., Rohanimanesh, K., and McCallum, A. "An entity based model for coreference resolution." In Proceedings of the 2009 SIAM International Conference on Data Mining, 365–376. SIAM (2009).

William E Winkler and Thibaudeau, Y. "An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census." Research Report, (April 2015):1–22 (1990).

Winkler, W. E. "Comparative Analysis of Record Linkage Decision Rules." In Proceedings of the Section on Survey Research Methods, 829–834. American Statistical Association (1992).

—. "Improved Decision Rules In The Fellegi-Sunter Model Of Record Linkage." In Proceedings of the Section on Survey Research Methods, 274–279. American Statistical Association (1993).

—. "Overview of record linkage and current research directions." Technical Report Statistics #2006-2, U.S. Bureau of the Census (2006).

—. "Matching and Record Linkage." Wiley Interdisciplinary Reviews: Computational Statistics, 6(5):313–325 (2014).

Winkler, W. E. et al. Record linkage software and methods for merging administrative lists. US Bureau of the Census (2001).

Wortman, J. P. H. "Record Linkage Methods with Applications to Causal Inference and Election Voting Data." Ph.D. thesis, Duke University (2019).

Yan, L. L. and Ozsu, M. T. "Conflict Tolerant Queries in AURORA." In Proceedings Fourth IFCIS International Conference on Cooperative Information Systems. CoopIS 99 (Cat. No. PR00384), 279–290. IEEE (1999).

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. "Flexible Models for Microclustering with Application to Entity Resolution." In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, 1425–1433. NY, USA: Curran Associates Inc. (2016).

Zhang, Y., Zhang, F., Yao, P., and Tang, J. "Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1002–1011 (2018).