COMPUTER SCIENCE

(Almost) all of entity resolution

Olivier Binette[1] and Rebecca C. Steorts[2,3]*

Whether the goal is to estimate the number of people that live in a congressional district, to estimate the number of individuals that have died in an armed conflict, or to disambiguate individual authors using bibliographic data, all these applications have a common theme—integrating information from multiple sources. Before such questions can be answered, databases must be cleaned and integrated in a systematic and accurate way, commonly known as structured entity resolution (record linkage or deduplication). Here, we review motivational applications and seminal papers that have led to the growth of this area. We review modern probabilistic and Bayesian methods in statistics, computer science, machine learning, database management, economics, political science, and other disciplines that are used throughout industry and academia in applications such as human rights, official statistics, medicine, and citation networks, among others. Last, we discuss current research topics of practical importance.

INTRODUCTION

As commonly known in computer science and statistics, entity resolution is the process of taking large noisy databases and removing duplicate entities (often in the absence of a unique identifier) (*1*–*8*). This task has become increasingly more important in science given that the entity resolution task is critical to have more reliable analyses. Since then, entity resolution has been widely studied in many research fields such as statistics, computer science, machine learning (*5*, *6*, *8*–*12*), political and social science (*13*, *14*), medicine and epidemiology (*15*–*20*), official statistics (*21*–*26*), human rights statistics (*27*–*32*), author name disambiguation (*33*–*37*), and forensic science (*38*, *39*), among many other disciplines.

This review is motivated by the long history of entity resolution, its active development in the scientific literature over the years, and its growing relevance throughout many scientific domains. We aim to enable the broader community to understand entity resolution methodology, from its foundation to recent modern developments. Specifically, we focus on social science and official statistics applications of entity resolution such as the U.S. decennial census, casualty estimation in armed conflicts, voter registration data, and the analysis of coauthorship networks. These applications often require uncertainty quantification—rigorous statements of confidence regarding results—as well as principled and interpretable methods that can be subjected to scientific scrutiny. In contrast with other recent reviews of entity resolution, we therefore emphasize probabilistic record linkage as well as recent progress with Bayesian approaches, graphical modeling, and microclustering. Furthermore, given the breadth of the field, we attempt to explain differences between these communities in a meaningful way to bridge this gap. There has been an abundance of contributions due to the database, machine learning, and computer science communities [see surveys (*1*–*8*)]. In contrast to statistical approaches, their focus has typically been on rule-based approaches (*40*, *41*), supervised learning approaches (*42*), hybrid human-machine approaches (*43*, *44*), and scalability (*45*).

Our review is structured as follows. The "Overview of entity resolution" section provides an overview of entity resolution, including impactful applications, terminology and definitions, challenges, and differences among entity resolution disciplines. The "Deterministic record linkage" section discusses rule- and similarity-based approaches, which are popular due to their interpretability and scalability. The "Probabilistic record linkage" section introduces probabilistic record linkage methods that have led to many advancements and extensions. The "Modern probabilistic record linkage" section reviews advancements of modern probabilistic record linkage, which include extensions to the Fellegi and Sunter framework, Bayesian extensions, and semisupervised and fully supervised record linkage methods. Moreover, this is where the most contributions from the machine learning and computer science communities have recently evolved, which contrasts that of the statistical literature. The "Entity resolution as a clustering problem" section reviews clustering-based approaches to entity resolution. We cover clustering tasks that are specifically postprocessing steps, graphical entity resolution methods, and microclustering models. Discussion discusses open research problems, and Supplementary Text provides references to open source software and datasets.

Last, while we attempt to cover almost all of the entity resolution literature, its breadth makes it impossible to cover all of it. Our focus is solely on structured entity resolution, and our focus is not the other subtasks of the data cleaning pipeline (*5*, *6*, *8*). We focus on structured databases as these are more common in statistical applications, where a major goal is ascertaining uncertainty propagation of the structured task, which can be more difficult for unstructured databases. In addition, we seek to strike a balance between many diverse communities working on the same problem that approach the entity resolution task differently. We hope that this article will help build further understanding between the communities and bring them closer together.

OVERVIEW OF ENTITY RESOLUTION

One of the earliest references to record linkage (here considered synonymous to entity resolution) is from Dunn (*46*), who defined record linkage as the process of assembling pieces of information that refer to the same individual. In 1946, Dunn wrote the following:

"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

[1]Department of Statistical Science, Duke University, Durham, NC, USA. [2]Department of Statistical Science, Computer Science, Biostatistics and Bioinformatics, the Rhodes Information Initiative at Duke (iiD) and the Social Science Research Institute (SSRI), Duke University, Durham, NC, USA. [3]Principal Mathematical Statistician, United States Census Bureau, Washington, DC, USA.
*Corresponding author. Email: beka@stat.duke.edu

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."

In short, record linkage (or entity resolution) seeks to bring together all relevant information about a person, business, or entity. This problem has gathered interest from the scientific community, including in statistics, computer science, machine learning, database management, finance, fraud detection, political science, official statistics, and medicine, among many others. In this section, we provide an overview of some of the most important applications of entity resolution that have been used throughout science in recent years for structured data ("Impactful entity resolution applications" section). In addition, we review terminology in entity resolution and the data cleaning pipeline ("Terminology and definitions" section), as well as challenges in the entity resolution task ("Challenges of entity resolution" section). Last, we review critical differences between disciplines regarding how researchers approach the entity resolution task.

Impactful entity resolution applications

We now review applications across science that have motivated major developments in entity resolution.

Decennial census

One important and timely topic is one that faces the U.S. Census Bureau each decade when they attempt to count all the individuals in the population. This enumeration is used to allocate resources for roads, schools, projects, and apportion representation of legislators. Unfortunately, it has been shown difficult to accurately enumerate such a population using an optional census, and response rates are often quite low. Furthermore, some individual may be counted multiple times. For example, an individual that owns three houses might accidentally fill out three census forms. As another example, individuals in group quarters (such as universities and prisons) are often double counted by their "group" and a family member/parent/guardian (*47*). Deduplication is thus needed to obtain an accurate enumeration, with new methodology from the machine learning and statistical literature being recently proposed to this end (*48*). This methodology is scalable while providing exact error propagation throughout the blocking and the entity resolution task (*48*).

Human rights statistics

In this section, we review two case studies in human rights statistics.

Documented identifiable deaths in El Salvador. Between 1980 and 1991, the Republic of El Salvador witnessed a civil war. There are three databases available for this conflict, where duplications occur within and across each of the databases. The first two databases were collected during the conflict, whereas the third database was collected after the conflict. The first two databases contain reports on documented identifiable victims. The first source, El Rescate (ER-TL), a nongovernmental organization (based out of Los Angeles, CA), collected electronic data from published reports during the civil war (*49*). The second source, Comission de Derechos Humanos de El Salvador (CDHES), collected testimonials on violations from 1979 to 1991 (*50*). The third source contains reports on documented identifiable victims after the civil war. After the peace agreement in 1992, the United Nations created a Commission on the Truth (UNTC), which invited citizens to report war-related human rights violations. Hence, victims can be duplicated in these datasets.

Syrian conflict. One case study that has been of interest is the ongoing Syrian conflict. To our knowledge, the Human Rights Data Analysis Group (HRDAG) provided the first published work in this domain (*28*, *51*, *52*). There are four sources that collected data during the same time period—the Syrian Center for Statistics and Research (CSR-SY), the Syrian Network for Human Rights (SNHR), the Syria Shuhada website (SS), and the Violation Documentation Centre (VDC). Each source provides documented identifiable deaths in the conflict. Attributes available are full Arabic name, gender, death location, and date of death. HRDAG has labeled the dataset, as outlined in their paper (*51*).

Both applications have proven to be extremely important to the development of the entity resolution literature. The El Salvadoran application is important as it is an application to small-scale human rights data, where some attributes contain a great deal of noise due to the way the information was collected. Given uncertainty in the data, it is natural in this application to use fully unsupervised approaches. Furthermore, uncertainty propagation of the entity resolution process has been demonstrated to be important in recently published case studies (*30*–*32*, *50*, *53*). In addition, the Syrian dataset is also important given that researchers have proposed a near-linear time algorithm for unique entity estimation that provides uncertainty quantification of the number of documented identifiable deaths (*54*).

Estimation of voters in North Carolina

Another application that has been recently used in science has been voter registration databases, which are publicly available (and often online) in the United States. For example, the North Carolina State Board of Elections (NCSBE) publicly posts its voter registration database (www.ncsbe.gov/). This dataset contains rich information, such as first and last name, year of birth, phone number, and address. However, the voter registration number is often duplicated due to people moving, getting married, and various other reasons. See Table 1 for examples of public records from this dataset. The process through which voter registration records are matched with other official records can have a profound influence on one's ability to vote. Georgia's controversial "exact match" law (*55*), which was slightly changed in 2019, required an exact match between voter registration records and records from the Department of Driver Services or the Social Security Administration to validate voter registrations. For instance, typographical errors, different spellings of the same name, or outdated records could place a voter registration on hold. In 2017, about 670,000 registrations were canceled as a result (*56*). Enamorado (*57*) showed how this law could predominantly affect non-white voters [see also (*58*)]. This application illustrates the need for uncertainty quantification and fairness analyses.

Remark: While our motivating examples have highlighted individuals, we would like to note that entity resolution can apply to businesses and objects.

Inventor and author disambiguation

Author disambiguation is the problem of identifying documents that have been written by the same author. Disambiguated data are informative for the study of innovation and productivity (*59*). Many datasets have been used broadly throughout statistics, computer science, and machine learning to develop algorithms for this task. These are fully labeled (or partially labeled) benchmark datasets that tend to illustrate success on new methods. On the other hand, there are research questions of interest that can be posed from fully unsupervised author disambiguation using such datasets that do not contain any training labels, where the labels would also need to be provided in a principled manner that would not be biased toward any proposed algorithm.

Table 1. Example of public NCSBE records retrieved from for the county of Durham in North Carolina and corresponding to unique voter registration numbers. www.ncsbe.gov Some fields have been omitted for brevity, including ZIP code, phone number, and voter registration number. Street addresses have been permuted with other individuals to preserve some anonymity.

Name	Street address	Age	Sex	Race	Birth	Party
Domineck Q. AAshad Jr.	914 Monmouth Ave #3	26	M	B	–	LIB
Domineck Q. AAshad Sr.	1408 Auburndale Dr	55	M	B	NY	DEM
Xiomara A. Martinez	1715 Cole Mill Rd	31	F	O	HL	REP
Xiomara A. Martinez	2923 Forrestal Dr	31	F	O	HL	–
Virginia L. Mullinix	749 Ninth St #480	101	F	W	PA	REP
Jacqueline D. Fuller	141 Bagby LN	54	–	–	–	DEM
Jacqueline Fuller	905 Cook Rd	56	F	B	NC	DEM

Section S3 reviews recent benchmark datasets, where one can find a more comprehensive review of author disambiguation datasets and methodology in (*60*). Section S3 reviews research datasets, which typically do not have unique identifiers or went through an intensive and well-documented manual labeling process. In general, there does not appear to be a strong consensus within the author disambiguation community regarding a standard on benchmark datasets, which holds true for the rest of the entity resolution literature. In addition, most of the entity resolution datasets (including the hand labeled pairs) are not easily reproducible by authors, which presents a strong need for communities to set forth more clear standards regarding how such datasets should be created such that datasets do not favor one proposed method over another.

These applications have proven to be impactful as researchers and industry leaders have been able to consider many types of methods, such as deep learning models as well as semisupervised and fully supervised methods, among others, illustrating the strength of these approaches for this particular application.

Terminology and definitions

In this section, we introduce terminology that will be used throughout the article that is commonly known and used in the existing literature (*5*, *6*, *61*). In addition, we formally define the entity resolution problem.

A database (file) is a collection of records. A record in a database contains attributes (fields or features), such as given name, family name, date of death, and municipality. Here, we assume that each record refers to an entity (person, object, or event) with respect to which we want to aggregate relevant information. Entity resolution is the problem of identifying records that refer to the same entity, such as identifying individual victims among recorded deaths that contain duplication. Two records that refer to the same entity are said to be coreferent (a link) or to be non-coreferent otherwise (non-link). Entity resolution can be framed as clustering records according to the entity to which they refer or, equivalently, of identifying coreferent record pairs. This is also referred to as record linkage, deduplication, data matching, instance matching, and data linkage.

Remark: Performing entity resolution on a single database that contains coreferent records (duplicates) is often referred to as deduplication (or duplicate detection). This is the simplest case, where any structure corresponding to the source of each record is ignored. When the data are part of two databases, with duplication across but not within databases, the problem is referred to as bipartite record linkage.

The data cleaning pipeline

Entity resolution is usually thought of one stage in the data cleaning pipeline (*2*, *5*, *61*) represented below

$$\text{attribute alignment} \rightarrow \text{blocking} \rightarrow \text{entity resolution} \rightarrow \text{canonicalization} \quad (1)$$

In the first stage, attribute or schema alignment, records are parsed to identify a set of common attributes among the datasets. In the second stage, blocking, similar records are grouped into blocks. Only records appearing in the same block will then be compared; records that do not appear in the same block are automatically determined to be nonmatches. In the entity resolution stage, coreferent records are identified. Last, in the fourth stage, merging, data fusion, or canonicalization, entities resolved as matches in the third stage are merged to produce a single representative record. The focus of this review article is on structured entity resolution and not on the other stages of the pipeline. For surveys of the entire pipeline, we refer to (*5*, *6*, *8*).

Challenges of entity resolution

Entity resolution tasks face a trade-off between (i) scaling to large databases, (ii) providing uncertainty propagation of the entity resolution task through all stages of the data cleaning pipeline, and (iii) proposing methods that account for the distortions and errors found in the databases. For databases with a combined total of N records, there are $N(N-1)/2$ pairs of records that must be considered as being possibly coreferent. Evaluating each pair is therefore not scalable as the number of records grows. Most entity resolution methods avoid comparing all pairs by blocking. While this increases the computational speed, uncertainty cannot be propagated exactly from the blocking stage to the entity resolution stage, as shown in Eq. 1. Specifically, the entity resolution task will inherit any errors from the blocking task, some of which cannot be resolved. On the other hand, one can achieve exact uncertainty propagation by building a joint blocking and entity resolution model; however, this typically results in slower computational run times. Last, when

dealing with data that contain distortions, typographical errors, and noise, generative methods have seen success. Such models are typically more complex and thus typically may not scale to large databases. The trade-off between (i), (ii), and (iii) must be evaluated by the user based on each motivating application so that the most appropriate method can be chosen.

Differences in entity resolution disciplines

As previously stated, entity resolution is a broad interdisciplinary field of research. Given our motivating applications, our review paper is specifically concerned with structured entity resolution problems, where records are composed of clear attributes such as names and phone numbers. This can be contrasted with unstructured entity resolution, where entity instances may contain textual descriptions or images. In addition, we focus on entity resolution approaches that provide uncertainty quantification, such as probabilistic record linkage, Bayesian approaches, graphical modeling approaches, microclustering models, and semi- and fully supervised approaches, among others. These methods are needed in scientific applications where all sources of uncertainty that may affect the validity of results must be accounted for. The main challenge, in this case, is to properly quantify this uncertainty and to account for it in downstream analyses. This is a very different focus than what is found in most of the computer science, machine learning, and database management literature, where the main goal is to address big data challenges such as large amounts of data, continuously evolving databases, and streaming data. Without losing sight of our motivating applications, we have attempted to highlight approaches from both disciplines to bring these communities (and others) closer together.

DETERMINISTIC RECORD LINKAGE

In practice, the most commonly used entity resolution methods are based on a series of deterministic rules involving the comparison of record attributes. These are called deterministic, rule-based, and similarity-based approaches. A simple example is exact matching, where two record pairs are linked if they agree on all common attributes. This strict matching condition can be relaxed by allowing mismatch on a fixed number of attributes, by using disjunctions of exact matching rules, and by using similarity functions. These rules can also be learned from data, bringing us closer to the topic of probabilistic record linkage discussed in the "A theory of record linkage" and "Modern probabilistic record linkage" sections. However, the differentiating characteristic of deterministic approaches is that they do not account for uncertainty in the matching process. No probability model is used and no level of confidence is provided for the matching status of record pairs.

Record attributes are often distorted by noise (due to data entry errors, variant spellings, outdated records, etc.). Naturally, linkage rules should account for slight differences between attributes. One simple way to quantify such differences for names, addresses, and other textual attributes is via string distance functions. For westernized words, edit distances such as the Levenshtein distance (*62*, *63*) are used to account for deletions, insertions, and substitutions. The Jaro-Winkler distance (*21*, *64*) works well for the comparison of short strings such as name. In addition to edit distances and their variants, token-based similarity measures such as the Jaccard similarity and the cosine similarity are often used when dealing with unstructured text and longer strings. We refer the reader to (*65*, *66*) for more information regarding simple string distance functions.

In practice, rule-based systems are carefully crafted for the application at hand. As noted in (*67*, *68*), a large body of work in the computer science and database communities is devoted to rule-based methods. This includes the specification and learning of linkage rules (including learning string similarity functions) (*40*, *69*–*74*), the use and efficient computation of similarity functions (*75*–*77*), the use of indexing structures and algorithms for efficient execution of entity resolution in large databases (including blocking and filtering) (*78*–*81*), the use of clustering techniques to resolve linkage transitivity (see the "Clustering as a post-processing step" section) (*82*–*84*), and the integration of matching records (data fusion) (*61*, *85*). Notably, Benjelloun *et al.* (*86*) provide algorithms to minimize the number of pairwise comparisons when matching and merging records. This literature is thoroughly reviewed as part of (*1*, *6*–*8*). These methods allow the use of entity resolution at much larger scales than what is possible using probabilistic approaches of the kind discussed in the following sections. In addition, they can be used as part of a blocking stage to scale up other entity resolution methods that account for uncertainty.

While deterministic approaches are appealing for their simplicity, interpretability, and computational scalability, empirical studies comparing deterministic and probabilistic record linkage techniques used for epidemiological research have shown consistent improvements of probabilistic methods over deterministic approaches (*87*–*91*). Other literature has surveyed probabilistic versus deterministic methods more broadly in terms of comparisons, also finding improved performance with probabilistic methods (*92*, *93*). While being application specific, these evaluations showcase the potential of probabilistic approaches when linking noisy data, both accounting for uncertainty and providing good performance.

PROBABILISTIC RECORD LINKAGE

We now take a step back and turn to the earliest published works on probabilistic record linkage. These works have laid down foundations for the field, which we use as the basis of our discussion of modern methodology. First, we discuss the work of H. Dunn who defined record linkage, leading to the first algorithm solution by Newcombe *et al.* (*94*) ('Dunn's "Book of Life" and early references'). Next, we discuss in depth the Fellegi-Sunter record linkage framework (*95*) ("A theory of record linkage" section). This framework provides a principled statistical model for record linkage that is still widely used today. It has the notable property of requiring no training data for record linkage—it is entirely unsupervised. Since its introduction in 1969, the Fellegi-Sunter framework has been widely studied, extended, and even recently reinvented (*96*). These extensions, such as more flexible modeling, Bayesian propagation of uncertainty, and semisupervised learning, are discussed afterward in the "Modern probabilistic record linkage" section.

Dunn's "Book of Life" and early references

As previously mentioned, the concept of record linkage (here used as a synonym to entity resolution) first appeared in a paper by Dunn (*46*) in 1946. In the context of administering governmental programs and services, he defined record linkage as the process of assembling pieces of information that refer to the same individual.

Interestingly, Dunn framed record linkage as an entirely logistical problem: If birth certificate numbers were widely used, then any centralized index would effectively bind individual records into this "book of life." However, the reality is that birth certificate numbers, or other unique identifiers for that matter, are not widely used. Records collected from different organizations, at different times, and for different purposes usually cannot be trivially matched together. Record linkage thus becomes an algorithmic problem—what can best be used to identify records that refer to the same individual, given noisy, uncertain information?

This is a problem that Newcombe *et al.* (*94*) faced in 1959 when trying to match birth and marriage records for demographic studies (*97*–*100*). The authors proposed, to our knowledge, the first automated record linkage method, which did not require a unique identifier. Specifically, they used domain knowledge of last names, first initials, birthplaces, ages (of some records), and location of child birth/marriage events. While no single one of these pieces of information was entirely reliable, together they could be used for accurate record linkage. The idea of their method was quite simple and had two steps. In the first step, the authors used blocking. Specifically, to account for variations in spelling, records were blocked (indexed) on the basis of the Soundex coding of the names. Note that the Soundex coding scheme was introduced by M. K. Odell and R. C. Russell [see U.S. patents 1261167 (1918) and 1435663 (1922)]. It codifies names by the first letter and by a string of three numbers, with the property that phonetically similar names often share the same code. Table 2 provides an example of attribute information from compared marriage and birth records from the original (*94*). Second, when Soundex coding agreed between two records, they computed a likelihood ratio comparing the hypothesis that the record pair were a match to the hypothesis that they were not. If this likelihood ratio exceeded a threshold, then the two records were linked (declared coreferent); otherwise, they were not linked (declared non-coreferent). Studies of the accuracy of the linkage showed that about 98.3% of the true matches were detected and about 0.7% of the linked records were not actual matches. In terms of computational speed, 10 records could be linked every minute on the Datatron 205 computer.

Table 2. Example of attribute information from marriage and birth records. This table is adapted from table 1 of (*100*) and translated from French to English. AB and PE represent the Canadian provinces of Alberta and Prince Edward Island. Only the initials of the first and middle names are provided in these data.

Attribute information	Marriage record	Birth record
Husband's Soundex name code	A300	A300
Wife's Soundex name code	B600	B600
Husband's family name	**Ayad**	**Ayot**
Wife's family name	Barr	Barr
Husband's initials	J Z	J Z
Wife's initials	**M** T	**B** T
Husband's birth province	AB	AB
Wife's birth province	PE	PE

In short, the work of Newcombe *et al.* (*94*) introduced key ideas for record linkage in an application to demographic data, where blocking (indexing) was used to make the problem computationally tractable. They proposed an informal statistical approach based on a likelihood ratio test, where the pipeline was fully automated and required no training data.

A theory of record linkage

We now turn to the Fellegi-Sunter framework (*95*) introduced in 1969 and that formalizes the approach of Newcombe *et al.* (*94*) in a decision-theoretic framework. We will define the likelihood ratio of the type used by Newcombe *et al.* and we will see how records can be linked while controlling fixed error rates. Furthermore, we will review the Fellegi-Sunter probability model, its interpretation, and its underlying assumptions.

The decision model of the Fellegi-Sunter framework considers records in independent pairs. For a given pair of records, three possible actions are considered: to link, to possibly link, or to not link. The goal is to minimize the number of possible links while controlling for type I and type II error rates (false match rate and false nonmatch rate). In the Fellegi-Sunter framework, an optimal linkage procedure attains the specified error rates while minimizing the number of possible link assignments. A "fundamental theorem for record linkage" demonstrated by the authors shows that the optimal linkage procedure corresponds to thresholding a likelihood ratio.

The likelihood ratio is defined as follows. Let γ be the comparison vector used to represent the level of agreement/disagreement between two specified records. In practice, γ is usually decomposed as $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, where each γ_i corresponds to a comparison between a particular attribute (name, age, etc.) of the record pair. One can consider binary comparisons, where $\gamma_i \in \{0,1\}$ represents agreement or disagreement between record attributes, as well as more detailed comparisons involving the specific value for which there is an agreement (such as $\gamma_i =$ " initials agree and are J & M"). Now, let $m(\gamma)$ be the probability of observing the comparison vector γ for two records that are an actual match, and let $u(\gamma)$ be the probability of observing the comparison vector γ for two records that are not a match. The likelihood ratio is then defined as $m(\gamma)/u(\gamma)$, and its logarithm $W(\gamma) = \log(m(\gamma)) - \log(u(\gamma))$ is called the matching weight.

Two unsupervised methods are proposed by Fellegi and Sunter to estimate the m and u probabilities. In both methods presented below, the authors assumed conditional independence between the attribute comparisons $\{\gamma_i\}_{i=1}^{k}$ given the true underlying match/nonmatch status of the record pairs.

First, the authors considered detailed comparison vectors, which provide both an indication of agreement or disagreement for each attribute and a precise shared value in the case of an agreement. This allows one to exploit specific information about the record's attributes. For instance, two records agreeing on the less common name "Xander" are more likely to be a match than two records that only agree on the first name "John." In applications, it is often helpful to exploit such frequency information. Under this assumption, the authors used the frequency distribution of the record's attributes, together with prior information about error rates, to obtain estimates of the m and u probability distributions.

Second, the authors considered binary comparisons, where each γ_i is a binary variable indicating agreement or disagreement with the records' ith attribute. The distributions m and u can then be

estimated from the observed frequencies of agreement or disagreement between these fields. In particular, they derived analytical formulas to estimate m and u when only three fields are under comparison.

Winkler (*101*) extended the above estimation methods by proposing the use of the expectation-maximization (EM) algorithm to estimate the m and u distributions both in the context of detailed comparisons between fields (where particular agreement values are also taken into consideration) and binary comparisons. Independently, Jaro (*21*) proposed the EM algorithm for binary comparisons and considered its application for matching the 1985 test census (dress rehearsal) of Tampa, Florida, to an independent post-enumeration survey to evaluate the census coverage.

Interpretation of the probability model

While the Fellegi-Sunter approach was introduced in a decision-theoretic framework, it can be interpreted more easily through its underlying probability model, where the comparison vectors γ are distributed as the following mixture model

$$p(\gamma) = \lambda m(\gamma) + (1 - \lambda)\, u(\gamma)$$

where $\lambda > 0$ is the probability that a randomly chosen comparison vector corresponds to a matching pair of records. The methods proposed by Fellegi-Sunter, as well as the EM algorithm proposed by Jaro (*21*) and Winkler (*101*), provide estimates of the parameters λ, m, and u.

Denote a true match by M. Using the Bayes rule, one can express the probability that two records match given their comparison vector γ, as

$$p(M \mid \gamma) = \lambda m(\gamma)/p(\gamma) = 1 - \left(1 + \frac{m(\gamma)}{u(\gamma)} \frac{\lambda}{1 - \lambda}\right)^{-1} \quad (2)$$

The left-hand side of Eq. 2 is the posterior probability of a match, and the right-hand side shows how it can be obtained as a monotonous transformation of the Fellegi-Sunter likelihood ratio $m(\gamma)/u(\gamma)$. Therefore, as noted in (*102*), thresholding the posterior probability to assign links is equivalent to using a likelihood ratio test and the Fellegi-Sunter optimality result also applies in this context.

Assumptions of Fellegi-Sunter

The Fellegi-Sunter framework relies on crucial simplifying assumptions. The first assumption is that comparison vectors between the record pairs should be independent from one another. This is usually not satisfied in practice. For example, when Newcombe *et al.* (*94*) linked birth and marriage records, it was known that two different marriages could not result in the same birth. This constraint induces dependencies between comparison vectors, and applying the Fellegi-Sunter procedure can lead to impossible linkage configurations when this is not taken into consideration. Generally, any linkage that does not satisfy transitive closure is impossible—knowing that a links to b and that b links to c should entail that a also links to c.

The second assumption is that the m and u distributions are known or can be adequately estimated. To be practically feasible, their estimation relies on simplifying assumptions that usually do not hold. For one thing, the estimation methods discussed so far require conditional independence between the comparison of different record attributes, given the true match/nonmatch status of the record pairs. Smith and Newcombe (*103*) first remarked that this conditional independence assumption may not hold in practice. Thibaudeau (*104*) [see also (*105*–*107*)] proposed log-linear models with interaction terms to account for dependencies between field comparisons and showed improved performance in some applications. Given that the assumptions of Fellegi-Sunter are often not satisfied, this has led to many extensions in the literature, which we review in the "Modern probabilistic record linkage" section.

Modern probabilistic record linkage

In this section, we review modern probabilistic record linkage, which includes extensions to the Fellegi-Sunter framework, Bayesian variants of Fellegi-Sunter, as well as semisupervised and fully supervised classification approaches.

Extensions of Fellegi-Sunter

In many applications, the procedures neither of Fellegi-Sunter nor of Tepping are used to set classification thresholds. According to Belin and Rubin (*108*), for the matching of the 1990 Census with the post-enumeration survey, thresholds were set "by 'eyeballing' lists of pairs of records brought together as candidate matches." Part of the reason is that the error rates fixed in the Fellegi-Sunter framework, as well as the false match rates estimated using Eq. 2, are not attained in practice (*105*, *106*, *108*, *109*). This is due to the various simplifying assumptions and estimation errors involved in the application of such models. Therefore, methods using training data (classified record pairs) have been proposed to automate and improve the choice of tuning parameters in probabilistic record linkage.

For instance, Belin and Rubin (*108*) proposed to calibrate thresholds and error rates by using training data to fit a mixture model to the matching weight distribution. This allowed the authors to quantify uncertainty about the linkage's error rates and to calibrate the Fellegi-Sunter thresholds. Nigam *et al.* (*110*) showed how training data could be combined with unlabeled data to improve the estimation of the m and u distributions using the EM algorithm for text classification. Building on the same semisupervised framework, Winkler (*111*, *112*) and Larsen and Rubin (*102*) considered fitting more complex models, allowing dependencies between field comparisons. Last, Enamorado *et al.* (*14*) have scaled the seminal Fellegi-Sunter model to large databases, where they incorporated auxiliary information into the merge and post-merge analyses (*113*).

Bayesian Fellegi-Sunter

In the context of entity resolution, Bayesian methods provide a way to quantify and propagate uncertainty for the joint linkage structure of a set of records. Furthermore, Bayesian methods allow the incorporation of prior knowledge, such as linkage transitivity, into analyses. These properties have made them popular in inferential and scientific applications, where uncertainty must be taken into account to reach sound conclusions and where prior knowledge is often available. This section reviews Bayesian record linkage methodology, which extends the Fellegi-Sunter framework.

Fortini *et al.* (*23*) extended the seminal work of Fellegi and Sunter (*95*) in the special case of bipartite record linkage. The authors assumed a prior on the "matching pairs," a prior on the "matching configuration matrix," and a Dirichlet prior on the m and u distributions. Here, the matching configuration matrix, or coreference matrix, indicates the linkage structure between two databases. That is, if we

denote by i a record in the first database and j a record in the second database, then this matrix has entries $c_{i,j} \in \{0,1\}$, with $c_{i,j} = 1$ if records i and j are linked and $c_{i,j} = 0$ otherwise.

More recently, Sadinle (*30*) proposed the first Bayesian Fellegi-Sunter model, where he assumed two databases and a de-deduplication scenario. He assumed a likelihood similar to that of Fortini *et al.* (*23*). Sadinle (*30*) considered a partitioning approach that allows transitive closures to be satisfied. This allows quantification of uncertainty about the partition of records through a posterior distribution. In later work, Sadinle (*31*) extended the above framework for bipartite record linkage. In addition, the authors derive Bayes estimates under a general class of loss functions, which provides an alternative to the Fellegi-Sunter decision rule. Both the work of Sadinle (*30*, *31*) apply their proposed methodology with deterministic blocking rules to the case study on human rights in El Salvador. In addition to proposing new methodology, Sadinle (*30*) performed hand matching on a small set of the dataset such that pairwise evaluation metrics could be used. The work of Sadinle (*30*, *31*) has been extended by Wortman (*114*), where the author accounts for dependencies between fields and for heterogeneity in the comparison vector distribution.

One difficulty facing Bayesian Fellegi-Sunter is their computational burden. In other recent work, McVeigh *et al.* (*13*) considered this issue by proposing a blocking approach based on simpler probabilistic record linkage techniques. That is, the output of more simple non-Bayesian probabilistic record linkage is used to perform "post hoc blocking," after which a Bayesian Fellegi-Sunter method is used for coherent modeling and uncertainty quantification. This allows the authors to scale their proposed method to voter registration and census datasets with millions of entries.

Semi- and fully supervised classification approaches

The approaches of (*108*, *110*–*112*) and (*102*) discussed in the "Extensions of Fellegi-Sunter" section were semisupervised (*115*). Semisupervised methods use a relatively small amount of manually classified record pairs, known as labeled pairs, to improve upon unsupervised probabilistic record linkage. In this section, we review semisupervised methods and fully supervised methods, which focus on classifying record pairs as a first step to entity resolution.

Note that the use of training data in entity resolution can be a complex problem. In many statistical applications, such as with the El Salvadoran dataset, ground truth is not available. The closest available data for use in training might come from one or more human experts through costly review processes. However, despite best efforts, experts can be subject to errors, biases, and uncertainty. Furthermore, the sampling process from which labeled data are obtained is highly influential and must be accounted for. These issues are the topic of broad research on sampling/querying, crowdsourcing, active learning, and performance evaluation for entity resolution. Here, we only briefly touch on these topics, instead focusing on methods that assume a single set of reliable labels.

Semisupervised approaches

Following the terminology of Chapelle *et al.* (*115*), we consider three types of semisupervised approaches. First, generative semisupervised approaches target the joint likelihood of the labeled and unlabeled data as in the study of Nigam *et al.* (*110*) and Larsen and Rubin (*102*). Building on this framework, Enamorado (*116*) proposed an active learning algorithm that iteratively requests labels for specific record pairs. Other active learning approaches are proposed in (*117*–*121*). Second, change of representation semisupervised approaches use unsupervised learning as a first step to summarize the data (such as performing dimensionality reduction), before using a supervised algorithm for further analysis. For instance, Belin and Rubin (*108*) used the unsupervised Fellegi-Sunter framework to obtain univariate matching weights for all record pairs, before using labeled examples to fit a mixture model to the matching weights. This allows the authors to calibrate the model and potentially select better thresholds. Third, self-learning algorithms generalize the semisupervised EM algorithm considered in (*110*) and (*102*) to model-free classifiers. In this framework, Kejriwal and Miranker (*122*) combined self-learning and boosting of random forests and multilayer perceptrons to obtain good performances on entity resolution tasks using only small amounts of labeled pairs.

Fully supervised approaches

Fully supervised methods do not exploit information provided by unlabeled examples; instead, they rely on larger numbers of labeled pairs. Given the significant class imbalance when considering record pairs (very few pairs match), vast amounts of reliable training data or carefully selected training data are required for the use of these methods. These training data may come from crowdsourcing (*43*, *44*, *117*, *123*, *124*) or from extensive manual record linkage efforts (*125*–*127*), or it may be automatically generated using unsupervised methods to obtain an approximate training set (*128*–*130*). In practice, the amount of reliable training data necessary to train sophisticated learning algorithms such as deep neural networks (*42*, *131*–*136*) is not always available for entity resolution tasks. For instance, Kooli *et al.* (*134*) use more than 10 million examples of labeled record pairs (corresponding to more than 3000 resolved individual records) in an application to train deep neural networks. More recently, Kasai *et al.* (*135*) considered the issue of training deep neural networks for entity resolution with fewer labels, using active and transfer learning. The use of deep learning techniques in entity resolution is especially promising in application to unstructured or textual problems [see (*42*, *137*, *138*)], where, for instance, pretrained language models can be used (*139*). For structured entity resolution, simple classifiers [such as logistic regression, decision trees, random forests, Bayesian additive regression trees, and others (*140*)] are often preferred.

To give an example of how such methods are used in practice, consider the work of Ventura *et al.* (*141*), a case study for inventor disambiguation in the bibliographic database of U.S. Patent and Trademark Office (USPTO) patents. The authors proposed a supervised method based on random forests for deduplication. Their training data were constructed from the curriculum vitae of inventors in the field of optometrics as well as from a previous study on "superstar" academics in the life sciences (*126*, *142*). This allowed them to evaluate the performance of previous methods used in this application (*59*, *143*) and to train their random forest classifier on labeled comparison vectors of record pairs. Afterward, applying their entity resolution approach to other records in the USPTO database consisted of a four-stage pipeline. First, they use blocking where, in each block, they calculated comparison vectors for each record pair. Second, the authors calculated the predicted probability of a match using their random forest classifier applied to these comparison vectors. Third, the predicted probability was converted into an estimate of the dissimilarity between each pair of records. Fourth, the authors used single linkage hierarchical clustering corresponding to the dissimilarity scores in the previous step to

enforce transitive closures among record pairs. Clusters were determined by cutting the dendogram (tree) at a threshold. Last, all the clustering results were combined across blocks to obtain a final set of deduplicated records. Such clustering approaches to entity resolution, used either directly or as a follow-up to pairwise classification, are discussed next in the "Entity resolution as a clustering problem" section.

Active learning

Active learning models are based on semisupervised or fully supervised models. Instead of using a predetermined set of training data, however, they interactively request informative labeled data from an expert labeler. Their goal is to provide a balance between automation and human interaction, although this balance is difficult to quantify. These models perform well on many entity resolution applications—given the large class imbalance in entity resolution tasks, informative training data may be difficult to choose a priori (*144*). The active learning workflow for entity resolution is nicely reviewed in (*8*), where the authors consider early and modern approaches in the literature.

ENTITY RESOLUTION AS A CLUSTERING PROBLEM

The methods discussed so far focused on estimating the probability of a match between pairs of records given their comparison vector. This pairwise match probability provides a measure of uncertainty about specific links, where the corresponding false match and false nonmatch rates (or precision and recall) are pairwise evaluation metrics of performance. With the exception of Bayesian Fellegi-Sunter, these methods treat record pairs as being independent of one another, without accounting for the consequences of transitivity or other constraints on the linkage structure. This limits their practicality when linking more than two databases (and dealing with applications that have duplication across/within databases). Much of the literature has therefore advocated for a clustering-based approach to entity resolution and deduplication, which can integrate multiple databases (*30*, *31*, *48*, *82*, *145*–*152*). In this context, the goals shift. Instead of linking record to record, the goal is to cluster records to their true (unknown, latent) entity.

A large portion of this literature uses clustering as a second step to probabilistic record linkage to enforce transitivity of the output (*6*, *83*). Other clustering approaches are model-based, and in particular, we focus on graphical entity resolution in the "Graphical entity resolution" section. By probabilistically modeling the relationship of records to the latent entities to which they refer, these methods naturally provide uncertainty quantification regarding the clustering structure (*149*, *153*). Last, entity resolution can be viewed as what we refer to as a microclustering problem, meaning that the size of the latent clusters grows sublinearly as the number of records grows. This means that entity resolution does not experience power law (linear) growth as many traditional clustering tasks. We discuss microclustering in the "The microclustering property" section.

Clustering as a post-processing step

Many clustering approaches to entity resolution are based on pairwise similarities, pairwise match probabilities, or determined links and nonlinks. Therefore, these can be seen as post-processing the result of other pairwise record linkage procedures. They are used to resolve intransitivites in the linkage method and ensure a coherent output. There is a vast literature on the subject; we only review a selection of the proposed methodology for entity resolution. We refer the reader to (*3*, *5*, *83*, *154*) and (*6*) for more exhaustive reviews.

One of the first references in this area is Monge and Elkan (*82*), who framed entity resolution as a clustering problem. Specifically, they proposed that one should detect the connected components in the undirected graph of pairwise links [see also (*155*, *156*)] using a dynamic connectivity structure. Pairwise links were determined iteratively. At any given step, only records that were not in the same connected component were compared to determine the match/nonmatch status. It allowed the authors to resolve in transitivities in pairwise matching while avoiding superfluous comparisons. The idea of clustering through connected components is computationally efficient and has recently been exploited as part of a blocking stage in the study of McVeigh *et al.* (*13*). A more sophisticated technique, correlation clustering (*157*), maximizes the number of links within clusters plus the number of nonlinks across clusters. This approach was originally introduced in the context of document classification. The number of clusters is determined from an objective function, which is one advantage of this method. However, correlation clustering is NP-hard (*157*, *158*), and in practice, variants and approximate solutions are used (*83*, *159*–*161*). Another approach is hierarchical agglomerative clustering (*140*, *162*), which Ventura *et al.* (*147*) advocated in conjunction with ensemble classifiers for large-scale entity resolution. Ventura *et al.* (*141*) also applied this method for inventor disambiguation in the USPTO dataset as discussed in the "Semi- and fully supervised classification approaches" section.

Graphical entity resolution

We now turn to model-based clustering approaches, which allow quantification of uncertainty about the clustering structure. Bhattacharya and Getoor (*163*) built on the latent Dirichlet allocation (LDA) model to this end, where the goal in their application was to resolve individual authors in bibliographic databases. Their approach leveraged coauthorship groups (analogously to topics in LDA) to support the entity resolution process. They probabilistically modeled the unknown set of individual authors, the authors' group membership, and possible distortions in authors' names. Posterior inference was carried out using Gibbs sampling.

In a similar spirit, Tancredi and Liseo (*146*) proposed a new model for record linkage, which, instead of linking record to record, linked records to latent individuals. The authors used the hit and miss model of Copas and Hilton (*164*) as a measurement error model to explain possible distortions in the observed data. This deviates from the Fellegi-Sunter approach, as it does not use comparison data, instead working with the actual attribute information.

We refer to such approaches, where one recovers a bipartite graph linking records to reconstructed latent entities, as graphical entity resolution. More specifically, in (*148*) and (*153*), the authors developed a fully hierarchical Bayesian approach to entity resolution, using Dirichlet prior distributions over categorical latent attributes and assuming a data distortion model. They derived an efficient hybrid (Metropolis-within-Gibbs) Markov chain Monte Carlo algorithm for fitting these models. As with other Bayesian approaches, this allows full quantification of uncertainty regarding the number of latent individuals and the clustering structure of records into coreferent groups. In addition, Steorts *et al.* (*153*) showed that, for the proposed work and the work of Sadinle (*30*) and Tancredi

and Liseo (*146*), the use of a uniform prior on the set of links or nonlinks, in practice, leads to one having a biased estimation of the sample. This, in turn, led to the development of subjective priors on the linkage structure, which have appeared in (*151*). In addition to Bayesian models for categorical data, Steorts (*149*) extended the above work to both categorical and noisy string data by proposing a string pseudo-likelihood and an empirically motivated prior.

Motivated by the computational limitations of Steorts (*149*) and a case study of the 2010 Census, Marchant *et al.* (*48*) have proposed a scalable extension to this model. Their approach uses probabilistic blocking at the level of the latent entities, which enables distributed inference through a partially collapsed Gibbs sampler while accounting for blocking uncertainty.

The microclustering property

The work of Steorts (*149*) and Steorts *et al.* (*153*) led to interesting developments both in clustering and in entity resolution. The first is the formalization of the microclustering property, which describes the sublinear growth of clusters in entity resolution (and in other clustering tasks such as community detection). That is, one expects the size of the clusters to grow sublinearly as the total number of records also grows (*151*). On the other hand, traditional probabilistic, generative models for clustering, such as finite mixtures (*165*), Dirichlet processes (*166*–*168*), Pitman-Yor processes (*169*, *170*), and many others, assume a power law growth in the total number of records (data points) (*171*, *172*). Therefore, traditional probabilistic, generative models are misspecified for microclustering tasks, such as entity resolution, given that they have a sublinear growth rate. Therefore, applying a Bayesian nonparametric (BNP) model that favors large clusters makes little sense in the context where each cluster should correspond to a single true entity. The second development is the proposal of general BNP models that can satisfy the microclustering property. The authors also propose a more scalable algorithm, the chaperones algorithm, which allows computational speed-ups for entity resolution that are similar in spirit to the Split and Merge approach as also used by Steorts *et al.* (*153*).

The aforementioned work has led to considerations regarding the feasibility of entity resolution in the context of microclustering. There are only two papers addressing such implications in the literature to our knowledge. In recent work, Steorts *et al.* (*173*) provided the first quantitative bounds that, to our knowledge, can be used for entity resolution. Simulation studies offer guidance for when the bounds are tight and loose in practice. Johndrow *et al.* (*174*) showed that, unless the number of attributes (or features) grows with the number of records, entity resolution is not possible in certain situations.

DISCUSSION

Here, we have introduced the entity resolution problem as it relates to important social science issues such as the decennial census, human rights violations, voter registration, and inventor and author disambiguation. Applications are more widespread, dealing with medical, housing, and financial databases, among others. We have introduced the main terminology used in the literature, and we have provided the major challenges that researchers face within an entity resolution framework.

We have reviewed deterministic methods ("Deterministic record linkage" section) and seminal probabilistic record linkage methods, such as those proposed by Dunn, Newcombe, and Fellegi and Sunter ("Probabilistic record linkage" section), which led to many modern-day extensions ("Modern probabilistic record linkage" section). These extensions can be viewed as extensions of Fellegi and Sunter (frequentist and Bayesian) or as semi- or fully supervised classification approaches. The "Entity resolution as a clustering problem" section reviewed entity resolution methods that can be viewed as clustering tasks. These include methods where clustering is a post-processing step, graphical entity resolution, and microclustering models.

In the remainder of our discussion, we highlight a few remaining topics that are the subject of active research and that have important practical implications. First, we discuss the need to rigorously evaluate the performance of entity resolution methods in applications. Second, we discuss potential directions regarding scaling Bayesian entity resolution methods. Last, we discuss privacy issues surrounding the use of entity resolution.

Data fusion, merging, and canonicalization

Here, our focus has been on the structured entity resolution task. Of course, what happens after this task is equally as important, which is in the fourth stage—data fusion, merging, or canonicalization. Specifically, the entity resolution task provides one with potential records that may match to one another; however, it does not tell us which record is the true underlying entity. Canonicalization, merging, or data fusion is the task of merging groups of records that have been classified as matches into one record that represents the true entity (*5*, *175*). The earliest proposals of canonicalization were deterministic, rule-based methods, which were application specific and fast to implement (*177*). The existing literature assumes that training is available to select the canonical record, and authors have proposed optimization and semisupervised methods to finding the most representative record (*178*–*180*). For a full review of data fusion techniques, we refer to the study of Bleiholder and Naumann (*175*). This is an important area of future research as, to our knowledge, it is often unclear how researchers choose the canonical record or canonical dataset and how researchers proceed with downstream tasks, such as logistic regression or predictive analysis.

Evaluating entity resolution performance

Despite methodological advances, evaluating the performance of entity resolution remains a challenge for a number of reasons. Murray (*181*) expressed concerns regarding overreliance on simple (toy) datasets that may not be representative of real applications, as this could potentially lead methodological research astray. As a starting point, we review in section S2 some public datasets that can be used for comparisons/evaluations. However, given the wide range of fields of application of entity resolution, these datasets are comparatively few in number. We stress that, when using "benchmark datasets," it is crucial that researchers note the number of records under consideration, the level of noise in the data, its overall quality, and the reliability of the unique identifiers used for performance evaluation. In addition, one should not solely rely on toy datasets, but one should perform carefully thought out simulation studies to understand robustness to model misspecification to provide practitioners with a guide for using their method. In addition, extreme care should be taken regarding sensitivity of tuning parameters in proposed methods and sensitivity of the evaluated performance on choices of such parameters. Last, case studies should be considered, if possible, as this gives one an idea of how proposed methods work for "data in the wild."

In addition, many researchers advocate the use of expert-labeled data to help train entity resolution model and to evaluate their performance in applications. However, care should be taken as labeling errors and sampling procedures may introduce bias into estimates. Effectively eliciting expert-labeled data while accounting for such sources of bias is an active area of research and one that we consider to be its own field given the complexities involved.

Scaling entity resolution

Bayesian entity resolution algorithms have been successful in scaling to large datasets, as illustrated by McVeigh *et al.* (*13*) and Marchant *et al.* (*48*). McVeigh *et al.* (*13*) have scaled a Bayesian Fellegi-Sunter approach to roughly 57 millions of records using so-called post hoc blocks. The approach of Marchant *et al.* (*48*) is quite different as blocking and entity resolution are jointly modeled in a Bayesian framework, allowing uncertainty quantification about both parts of the pipeline. The authors scaled to roughly 1 million records using distributed computing. Further research is needed in this area to scale to larger datasets, such as census-size data or industrial-size datasets, while accounting for uncertainties encountered at all stages of the entity resolution pipeline.

Privacy issues

Entity resolution is fundamentally antithetic to data privacy—it is about gaining information about social entities through the integration of diverse databases. This raises ethical and legal questions for users of entity resolution as well as important privacy considerations (*182*). In particular, as more data are being collected, stored, analyzed, and shared across multiple domains, disclosure risks associated with (even anonymized) data releases become serious. For example, Sweeney (*183*) showed how a simple record linkage algorithm could be used to de-anonymize Netflix movie rankings data through the use of public IMDb profiles. Fienberg and Slavković (*184*) used public voter registration data to de-anonymize a health insurance dataset to showcase the need for stronger privacy measures. These are examples of linkage attacks, where an adversary uses background knowledge (such as voter registration files) to de-anonymize data or to gain information about individuals.

Data releases should therefore be managed through statistical disclosure control (SDC) systems, which aim to balance the utility of released data with privacy protections. To address these competing goals, many SDC techniques have been proposed and implemented such as top-coding, data swapping, data perturbation, and synthetic data generation, each potentially having its own measures of utility and risk properties; more details are those methods that can be found in (*185*, *186*). Furthermore, differential privacy (*187*) has emerged as a key rigorous definition of privacy. It provides a framework that can inform the design of privacy mechanisms with specified disclosure risks, in the presence of arbitrary external information.

As we have discussed, analyses often require or can benefit from the linkage of multiple databases. However, when databases are held by different organizations and contain private information that cannot be shared across them, record linkage should be done to ensure that (i) private information such as quasi-identifiers (name, date of birth, etc.) are not disclosed across organizations during the linkage process and (ii) only relevant summaries of the resulting linkage (usually a set of predetermined attributes of the linked records) are reported to manage disclosure risks. The achievement of these two goals is the subject of privacy-preserving record linkage (PPRL) (*188*–*190*). This is closely related to the problem of private multiparty data publishing under a vertical partitioning scheme (*191*–*194*). While progress has been made on point (i), the privacy implications of post-linkage data releases are difficult to analyze even under mild adversary models (*189*, *192*). Great care should be taken when using PPRL to ensure that all disclosure risks are properly assessed.

Unstructured and multimodal entity resolution

To our knowledge, one major area that has been explored in the computer science literature, but has been largely unexplored by statisticians, is the task of removing duplications from databases containing unstructured information, such as free-form text and images (*138*, *195*). One application of this is patent databases, which contain unstructured paragraphs (*59*). Another growing application is forensic science, which can involve the matching of fingerprints, ballistics, and other types of evidence from a crime scene (*38*, *39*). In these applications, structured and unstructured data must be combined for accurate linkage and inference. Last, it seems that there is a great need for algorithms that maintain individual-level privacy. However, it is not clear what methods would be most appropriate and if they would work well in practice. For example, are differential privacy measures too stringent? Case studies are needed to understand how such methods could be used in practice.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abi8021

REFERENCES AND NOTES

1. A. Doan, A. Halevy, Z. Ives, *Principles of Data Integration* (Morgan Kaufmann Publishers, 2012).
2. A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios, Duplicate record detection: A survey. *IEEE Trans. Knowledge Data Eng.* **19**, 1–16 (2007).
3. F. Naumann, M. Herschel, *An Introduction to Duplicate Detection* (Morgan & Claypool Publishers, 2010).
4. L. Getoor, A. Machanavajjhala, Entity resolution: Theory, practice & open challenges. *Proc. VLDB Endowment* **5**, 2018–2019 (2012).
5. P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection* (Data-Centric Systems and Applications, Springer-Verlag, 2012).
6. V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, K. Stefanidis, An overview of end-to-end entity resolution for big data. *ACM Computing Surveys* **53**, 1–42 (2021).
7. I. F. Ilyas, X. Chu, *Data Cleaning* (Association for Computing Machinery, 2019).
8. G. Papadakis, E. Ioannou, E. Thanos, T. Palpanas, *The Four Generations of Entity Resolution* (Morgan & Claypool Publishers, 2021).
9. T. Herzog, F. Scheuren, W. Winkler, *Data Quality and Record Linkage Techniques* (Springer, 2007).
10. W. E. Winkler, Matching and record linkage. *Wiley Interdiscip. Rev. Comput. Stat.* **6**, 313–325 (2014).
11. A. Jurek-Loughrey, P. Deepak, in *Semi-Supervised and Unsupervised Approaches to Record Pairs Classification in Multi-Source Data Linkage* (Springer, 2019), pp. 55–78.
12. J. Asher, D. Resnick, J. Brite, R. Brackbill, J. Cone, An introduction to probabilistic record linkage with a focus on linkage processing for wtc registries. *Int. J. Environ. Res. Public Health* **17**, 6937 (2020).
13. B. S. McVeigh, B. T. Spahn, J. S. Murray, Scaling Bayesian probabilistic record linkage with post-hoc blocking: An application to the california great registers. arXiv:1905.05337 [stat.ME] (14 May 2019).
14. T. Enamorado, B. Fifield, K. Imai, Using a probabilistic model to assist merging of large-scale administrative records. *Am. Polit. Sci. Rev.* **113**, 353–371 (2019).
15. E. Rogot, P. Sorlie, N. J. Johnson, Probabilistic methods in matching census samples to the National Death Index. *J. Chronic Dis.* **39**, 719–734 (1986).
16. N. Méray, J. B. Reitsma, A. C. Ravelli, G. J. Bonsel, Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J. Clin. Epidemiol.* **60**, 883–891 (2007).

17. M. A. Jaro, Probabilistic linkage of large public health data files. *Stat. Med.* **14**, 491–498 (1995).
18. R. Gutman, C. C. Afendulis, A. M. Zaslavsky, A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Am. Stat. Assoc.* **108**, 34–47 (2013).
19. M. Shan, K. Thomas, R. Gutman, A Bayesian multi-layered record linkage procedure to analyze functional status of medicare patients with traumatic brain injury. arXiv:2005.08549 [stat.ME] (18 May 2020).
20. E. Farley, R. Gutman, A Bayesian approach to linking data without unique identifiers. arXiv:2012.00601 [stat.CO] (1 December 2020).
21. M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Am. Stat. Assoc.* **84**, 414–420 (1989).
22. W. E. Winkler, Y. Thibaudeau, *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census* (U.S. Census Bureau, 1990), pp. 1–22.
23. M. Fortini, B. Liseo, A. Nuccitelli, M. Scanu, On Bayesian record linkage. *Res. Official Stat.* **4**, 185–198 (2001).
24. A. Chevrette, "G-link: A probabilistic record linkage system" (Technical Report, Statistics Canada, 2011).
25. A. Dasylva, R.-C. Titus, C. Thibault, Overcoverage in the 2011 Canadian census, in *Proceedings of Statistics Canada Symposium* (Statistics Canada, 2014).
26. A. Dasylva, Pairwise estimating equations for the primary analysis of linked data, in *Proceedings of Statistics Canada Symposium* (Statistics Canada, 2018).
27. K. Lum, M. E. Price, D. Banks, Applications of multiple systems estimation in human rights research. *Am. Statist.* **67**, 191–200 (2013).
28. M. Price, A. Gohdes, P. Ball, Documents of war: Understanding the Syrian conflict. *Significance* **12**, 14–19 (2015).
29. P. Sadosky, A. Shrivastava, M. Price, R. C. Steorts, Blocking methods applied to casualty records from the Syrian conflict. arXiv:1510.07714 [stat.AP] (26 October 2015).
30. M. Sadinle, Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Annal. Appl. Stat.* **8**, 2404–2434 (2014).
31. M. Sadinle, Bayesian estimation of bipartite matchings for record linkage. *J. Am. Stat. Assoc.* **112**, 600–612 (2017).
32. M. Sadinle, Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Annal. Appl. Stat.* **12**, 1013–1038 (2018).
33. R. Lai, A. D'amour, A. Yu, Y. Sun, L. Fleming, *Disambiguation and Co-Authorship Networks of the US Patent Inventor Database (1975–2010)* (Harvard Institute for Quantitative Social Science, 2011), vol. 2138.
34. G. Louppe, H. T. Al-Natsheh, M. Susik, E. J. Maguire, Ethnicity sensitive author disambiguation using semi-supervised learning, in *Proceedings of the International Conference on Knowledge Engineering and the Semantic Web* (Springer, 2016), pp. 272–287.
35. Y. Zhang, F. Zhang, P. Yao, J. Tang, Name disambiguation in aminer: Clustering, maintenance, and human in the loop, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, 2018), pp. 1002–1011.
36. S. Subramanian, D. King, D. Downey, S. Feldman, S2AND: A benchmark and evaluation system for author name disambiguation. arXiv:2103.07534 [cs.DL] (12 March 2021).
37. X. Liu, D. Yin, X. Zhang, K. Su, K. Wu, H. Yang, J. Tan, OAG-BERT: Pre-train heterogeneous entity-augmented academic language models. arXiv:2103.02410 [cs.CL] (3 March 2021).
38. X. H. Tai, Record linkage and matching problems in forensics, in *Proceedings of the IEEE International Conference on Data Mining Workshops* (IEEE, 2018), pp. 510–517.
39. X. H. Tai, W. F. Eddy, Automatically matching topographical measurements of cartridge cases using a record linkage framework. arXiv:2003.00060 [stat.AP] (28 February 2020).
40. W. Fan, X. Jia, J. Li, S. Ma, Reasoning about record matching rules. *Proc. VLDB Endowment* **2**, 407–418 (2009).
41. R. Singh, V. Meduri, A. Elmagarmid, S. Madden, P. Papotti, J.-A. Quiane-Riuz, A. Solar-Lezama, N. Tang, Generating concise entity matching rules, in *Proceedings of the 2017 ACM International Conference on Management of Data* (Association for Computing Machinery, 2017), pp. 1635–1638.
42. S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, Deep learning for entity matching: A design space exploration, in *Proceedings of the 2018 International Conference on Management of Data* (Association for Computing Machinery, 2018), pp. 19–34.
43. J. Wang, T. Kraska, M. J. Franklin, J. Feng, Crowder: Crowdsourcing entity resolution. *Proc. VLDB Endowment* **5**, 1483–1494 (2012).
44. C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, X. Zhu, *Corleone: Hands-off Crowdsourcing for Entity Matching* (Association for Computing Machinery, 2014), pp. 601–612.
45. G. Papadakis, J. Svirsky, A. Gal, T. Palpanas, Comparative analysis of approximate blocking techniques for entity resolution. *Proc. VLDB Endowment* **9**, 684–695 (2016).
46. H. L. Dunn, Record linkage. *Am. J. Public Health Nations Health* **36**, 1412–1416 (1946).
47. H. Hogan, P. J. Cantwell, J. Devine, V. T. Mule, V. Velkoff, Quality and the 2010 census. *Population Res. Policy Rev.* **32**, 637–662 (2013).
48. N. G. Marchant, R. C. Steorts, A. Kaplan, B. I. P. Rubinstein, D. N. Elazar, d-blink: Distributed end-to-end Bayesian entity resolution. arXiv:1909.06039 [stat.CO] (13 September 2019).
49. T. Howland, How El Rescate, a small nongovernmental organization, contributed to the transformation of the human rights situation in El Salvador. *Hum. Rights Q.* **30**, 703–757 (2008).
50. P. Ball, The Salvadoran human rights commission: Data processing, data representation, and generating analytical reports, in *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*, P. Ball, H. F. Spirer, L. Spirer, Eds. (American Association for the Advancement of Science, 2000), pp. 15–24.
51. M. Price, J. Klingner, A. Qtiesh, P. Ball, Full updated statistical analysis of documentation of killing in the Syrian Arab Repulic, in *Report by the Human Rights Data Analysis Group to the United Nations Office of the High Commissioner for Human Rights (OHCHR)* (Office of the UN High Commissioner for Human Rights, 2013).
52. M. Price, P. Ball, Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Rev. Int. Aff.* **34**, 9–20 (2014).
53. A. H. Green, P. Ball, Civilian killings and disappearances during civil war in El Salvador (1980–1992). *Demogr. Res.* **41**, 781–814 (2019).
54. B. Chen, A. Shrivastava, R. C. Steorts, Unique entity estimation with application to the Syrian conflict. *Annal. Appl. Stat.* **12**, 1039–1067 (2018).
55. J. Ax, Georgia lawsuit is latest blow in U.S. fight over voting rights (2018) [posted 12 October 2018; retrieved 17 July 2020].
56. B. Nadler, Voting rights become a flashpoint in georgia governor's race (2018) [posted 9 October 2018; retrieved 17 July 2020].
57. T. Enamorado, Georgia's 'exact match' law could potentially harm many eligible voters (2018) [posted 20 October 2018; retrieved 17 July 2020].
58. *Georgia Coalition For the Peoples' Agenda Inc. et al.* v. *Kemp*, Complaint for injunctive and declaratory relief (2018).
59. G. C. Li, R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, L. Fleming, Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Res. Policy* **43**, 941–955 (2014).
60. M.-C. Müller, F. Reitz, N. Roy, Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics* **111**, 1467–1500 (2017).
61. X. L. Dong, D. Srivastava, *Big Data Integration* (Morgan and Claypool Publishers, 2015).
62. V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Doklady* **10**, 707–710 (1966).
63. L. Yujian, L. Bo, A normalized levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1091–1095 (2007).
64. W. E. Winkler, String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, in *Proceedings of the Section on Survey Research, American Statistical Association* (American Statistical Association, 1990), pp. 354–359.
65. G. Navarro, A guided tour to approximate string matching. *ACM Comput. Surveys* **33**, 31–88 (2001).
66. W. W. Cohen, P. Ravikumar, S. E. Fienberg, A comparison of string distance metrics for name-matching tasks, in *Proceedings of the 2003 International Conference on Information Integration on the Web* (AAAI Press, 2003), pp. 73–78.
67. J. Wang, G. Li, J. X. Yu, J. Feng, Entity matching: How similar is similar. *Proc. VLDB Endowment* **4**, 622–633 (2011).
68. C. R. Rivero, D. Ruiz, Selecting suitable configurations for automated link discovery, in *Proceedings of the ACM Symposium on Applied Computing* (Association for Computing Machinery, 2020), pp. 907–914.
69. E. Ristad, P. Yianilos, Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 522–532 (1998).
70. H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita, "Declarative data cleaning: Language, model, and algorithms," thesis, INRIA (2001).
71. M. Bilenko, R. J. Mooney, Adaptive duplicate detection using learnable string similarity measures, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2003), pp. 39–48.
72. A. McCallum, K. Bellare, F. Pereira, A conditional random field for discriminatively-trained finite-state string edit distance. arXiv:1207.1406 [cs.LG] (4 July 2012).
73. M. Nentwig, M. Hartung, A. C. Ngonga Ngomo, E. Rahm, A survey of current link discovery frameworks. *Semantic Web* **8**, 419–436 (2016).
74. N. Andrews, J. Eisner, M. Dredze, Name phylogeny: A generative model of string variation, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Association for Computational Linguistics, 2012), pp. 344–355.
75. W. W. Cohen, Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inform. Syst.* **18**, 288–321 (2000).
76. T. Soru, A. C. N. Ngomo, Rapid execution of weighted edit distances. *Proc. Ontol. Matching Workshop* **1111**, 1–12 (2013).
77. H. Zhang, Q. Zhang, Embedjoin: Efficient edit similarity joins via embeddings, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2017), pp. 585–594.

78. J. Wang, J. Feng, G. Li, Trie-join: Efficient triebased string similarity joins with edit distance constraints. *Proc. VLDB Endowment* **3**, 1219–1230 (2010).
79. M. Yu, J. Wang, G. Li, Y. Zhang, D. Deng, J. Feng, A unified framework for string similarity search with edit-distance constraint. *VLDB J.* **26**, 249–274 (2017).
80. H. Wei, J. X. Yu, C. Lu, String similarity search: A hash-based approach. *IEEE Trans. Knowl. Data Eng.* **30**, 170–184 (2018).
81. G. Papadakis, D. Skoutas, E. Thanos, T. Palpanas, Blocking and filtering techniques for entity resolution: A survey. *ACM Comput. Surveys* **53**, 1–42 (2020).
82. A. E. Monge, C. P. Elkan, An efficient domain-independent algorithm for detecting approximately duplicate database records, in *Proceedings of the SIGMOD 1997 Workshop on Research Issues on Sata Mining and Knowledge Discovery* (DMKD, 1997), pp. 23–29.
83. O. Hassanzadeh, F. Chiang, H. C. Lee, R. J. Miller, Framework for evaluating clustering algorithms in duplicate detection. *Proc. VLDB Endowment* **2**, 1282–1293 (2009).
84. A. Saeedi, E. Peukert, E. Rahm, in *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution* (Springer International Publishing, 2017), pp. 278–293.
85. A. Heidari, G. Michalopoulos, S. Kushagra, I. F. Ilyas, T. Rekatsinas, Record fusion: A learning approach. arXiv:2006.10208 [cs.LG] (18 June 2020).
86. O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, J. Widom, Swoosh: A generic approach to entity resolution. *VLDB J.* **18**, 255–276 (2009).
87. S. B. Dusetzina, S. Tyree, A.M. Meyer, A. Meyer, L. Green, W.R. Carpenter, *Linking Data for Health Services Research: A Framework and Instructional Guide* (Agency for Healthcare Research and Quality, 2014).
88. S. Gomatam, R. Carter, M. Ariet, G. Mitchell, An empirical comparison of record linkage procedures. *Stat. Med.* **21**, 1485–1496 (2002).
89. K. M. Campbell, D. Deck, A. Krupski, Record linkage software in the public domain: A comparison of Link plus, the Link King, and a 'basic' deterministic algorithm. *Health Informatics J.* **14**, 5–15 (2008).
90. M. Tromp, A. C. Ravelli, G. J. Bonsel, A. Hasman, J. B. Reitsma, Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *J. Clin. Epidemiol.* **64**, 565–572 (2011).
91. T. Avoundjian, J. C. Dombrowski, M. R. Golden, J. P. Hughes, B. L. Guthrie, J. Baseman, M. Sadinle, Comparing methods for record linkage for public health action: Matching algorithm validation study. *JMIR Public Health Surveill.* **6**, e15917 (2020).
92. R. C. Steorts, S. L. Ventura, M. Sadinle, S. E. Fienberg, A comparison of blocking methods for record linkage, in *Privacy in Statistical Databases*, J. Domingo-Ferrer, Ed. (Springer, 2014), pp. 253–268.
93. J. S. Murray, Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *J. Privacy Confidential.* **7**, 3–24 (2016).
94. H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James, Automatic linkage of vital records. *Science* **130**, 954–959 (1959).
95. I. P. Fellegi, A. B. Sunter, A theory for record linkage. *J. Am. Stat. Assoc.* **64**, 1183–1210 (1969).
96. R. Wu, S. Chaba, S. Sawlani, X. Chu, S. Thirumuruganathan, ZeroER: Entity resolution using zero labeled examples, in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery, 2020), pp. 1149–1164.
97. H. B. Newcombe, P. O. W. Rhynas, Child spacing following stillbirth and infant death. *Eugen. Q.* **9**, 25–35 (1962).
98. H. B. Newcombe, The study of mutation and selection in human populations. *Eugen. Rev.* **57**, 109–125 (1965).
99. H. B. Newcombe, O. G. Tavendale, Effects of father's age on the risk of child handicap or death. *Obstet. Gynecol. Survey* **20**, 655–656 (1965).
100. H. B. Newcombe, Couplage de donnees pour les etudes demographiques. *Population* **24**, 653 (1969).
101. W. E. Winkler, Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, in *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 1988), pp. 667–671.
102. M. D. Larsen, D. B. Rubin, Iterative automated record linkage using mixture models. *J. Am. Stat. Assoc.* **96**, 32–41 (2001).
103. M. E. Smith, H. B. Newcombe, Methods for computer linkage of hospital admission separation records into cumulative health histories. *Methods Inf. Med.* **14**, 118–125 (1975).
104. Y. Thibaudeau, The discrimination power of dependency structures in record linkage. *Survey Methodol.* **19**, (1993).
105. J. Armstrong, J. Mayda, Estimation of record linkage models using dependent data, in *Proceedings of the Section on Survey Research Methodology* (American Statistical Association, 1992), pp. 853–858.
106. W. E. Winkler, Comparative analysis of record linkage decision rules, in *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 1992), pp. 829–834.
107. W. E. Winkler, Improved decision rules in the Fellegi-Sunter model of record linkage, in *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 1993), pp. 274–279.
108. T. R. Belin, D. B. Rubin, A method for calibrating false-match rates in record linkage. *J. Am. Stat. Assoc.* **90**, 694–707 (1995).
109. T. R. Belin, A proposed improvement in computer matching techniques, in *Statistics of Income and Related Administrative Record Research* (International Revenue Service, 1990), pp. 167–172.
110. K. Nigam, A. K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **39**, 103–134 (2000).
111. W. E. Winkler, Machine learning, information retrieval, and record linkage, in *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 2000), pp. 20–29.
112. W. E. Winkler, "Methods for record linkage and Bayesian networks" (Technical Report, Statistical Research Division, U.S. Census Bureau, 2002).
113. P. Lahiri, M. Larsen, Regression analysis with linked data. *J. Am. Stat. Assoc.* **100**, 222–230 (2005).
114. J. P. H. Wortman, "Record linkage methods with applications to causal inference and election voting data," thesis, Duke University (2019).
115. O. Chapelle, S. Bernhard, A. Zien, *Semi-Supervised Learning* (The MIT Press, 2006).
116. T. Enamorado, *Active Learning for Probabilistic Record Linkage* (Princeton University, 2019).
117. S. Sarawagi, A. Bhamidipaty, Interactive deduplication using active learning, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2002), pp. 269–278.
118. K. Bellare, S. Iyengar, A. G. Parameswaran, V. Rastogi, Active sampling for entity matching, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012), pp. 1131–1139.
119. Q. Wang, D. Vatsalan, P. Christen, Efficient interactive training selection for large-scale entity resolution, in *Advances in Knowledge Discovery and Data Mining* (Springer, 2015), pp. 562–573.
120. P. Christen, D. Vatsalan, Q. Wang, Efficient entity resolution with adaptive and interactive training data selection, in *Proceedings of the IEEE International Conference on Data Mining* (IEEE, 2015), pp. 727–732.
121. D. Firmani, B. Saha, D. Srivastava, Online entity resolution using an oracle. *Proc. VLDB Endowment* **9**, 384–395 (2016).
122. M. Kejriwal, D. P. Miranker, Semi-supervised instance matching using boosted classifiers, in *Proceedings of the European Semantic Web Conference* (Springer, 2015), pp. 388–402.
123. N. Vesdapunt, K. Bellare, N. Dalvi, Crowdsourcing algorithms for entity resolution. *Proc. VLDB Endowment* **7**, 1071–1082 (2014).
124. K. Frisoli, B. LeRoy, R. Nugent, A novel record linkage interface that incorporates group structure to rapidly collect richer labels, in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics* (IEEE, 2019), pp. 580–589.
125. M. Trajtenberg, G. Shiff, "Identification and mobility of Israeli patenting inventors" (Technical Report, Pinhas Sapir Center for Development, 2008).
126. P. Azoulay, J. S. G. Zivin, B. N. Sampat, The diffusion of scientific knowledge across time and space: Evidence from professional transitions for the superstars of medicine, *The Rate and Direction of Inventive Activity Revisited*, J. Lerner, S. Stern, Eds. (University of Chicago Press, 2012).
127. M. J. Bailey, C. Cole, M. Henderson, C. Massey, How well do automated linking methods perform? Lessons from US historical data. *J. Econ. Lit.* **58**, 997–1044 (2020).
128. V. I. Torvik, M. Weeber, D. R. Swanson, N. R. Smalheiser, A probabilistic similarity metric for medline records: A model for author name disambiguation. *J. Am. Soc. Inform. Sci. Technol.* **56**, 140–158 (2005).
129. P. Christen, A two-step classification approach to unsupervised record linkage, in *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics* (2007), pp. 111–119.
130. P. Christen, Automatic record linkage using seeded nearest neighbour and support vector machine classification, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2008), pp. 151–159.
131. R. D. Gottapu, C. Dagli, B. Ali, Entity resolution using convolutional neural network. *Procedia Comput. Sci.* **95**, 153–158 (2016).
132. M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, N. Tang, DeepER–Deep entity resolution, arXiv:1710.00597 [cs.DB] (2 October 2017).
133. M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, N. Tang, Distributed representations of tuples for entity resolution. *Proc. VLDB Endowment* **11**, 1454–1467 (2018).
134. N. Kooli, R. Allesiardo, E. Pigneul, Deep learning based approach for entity resolution in databases, in *Intelligent Information and Database Systems* (Springer International Publishing, 2018), pp. 3–12.
135. J. Kasai, K. Qian, S. Gurajada, Y. Li, L. Popa, Low-resource deep entity resolution with transfer and active learning, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistic, 2020), pp. 5851–5861.
136. N. Barlaug, J. A. Gulla, Neural networks for entity matching: A survey. *ACM Trans. Knowledge Discov. Data* **15**, 1–37 (2021).

137. B. Li, W. Wang, Y. Sun, L. Zhang, M. A. Ali, Y. Wang, GraphER: Token-centric entity resolution with graph convolutional neural networks, in *Proceedings of the 34th AAAI Conference on Artificial Intelligence* (AAAI, 2020), pp. 8172–8179.
138. Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, W. C. Tan, Deep entity matching: Challenges and opportunities. *J. Data Inform. Quality* **13**, 1–17 (2021).
139. Y. Li, J. Li, Y. Suhara, A. Doan, W. C. Tan, Deep entity matching with pre-trained language models. *Proc. VLDB Endowment* **14**, 50–60 (2020).
140. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2001).
141. S. L. Ventura, R. Nugent, E. R. Fuchs, Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Res. Policy* **44**, 1672–1701 (2015).
142. P. Azoulay, R. Michigan, B. N. Sampat, The anatomy of medical school patenting. *N. Eng. J. Med.* **357**, 2049–2056 (2007).
143. L. Fleming, C. King III, A. I. Juda, Small worlds and regional innovation. *Organization Sci.* **18**, 938–954 (2007).
144. A. Arasu, M. Götz, R. Kaushik, On active learning of record matching packages, in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (Association for Computing Machinery, 2010), pp. 783–794.
145. W. W. Cohen, J. Richman, Learning to match and cluster large high-dimensional data sets for data integration, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2002), pp. 475–480.
146. A. Tancredi, B. Liseo, A hierarchical Bayesian approach to record linkage and population size problems. *Annal. Appl. Stat.* **5**, 1553–1585 (2011).
147. S. L. Ventura, R. Nugent, E. R. Fuchs, Hierarchical linkage clustering with distributions of distances for large scale record linkage, in *Privacy in Statistical Databases*, J. Domingo-Ferrer, Ed. (Springer, 2014), pp. 283–298.
148. R. C. Steorts, R. Hall, S. E. Fienberg, SMERED: A Bayesian approach to graphical record linkage and de-duplication. *J. Mach. Learn. Res.* **33**, 922–930 (2014).
149. R. C. Steorts, Entity resolution with empirically motivated priors. *Bayesian Anal.* **10**, 849–875 (2015).
150. E. Rahm, The case for holistic data integration, in *Advances in Databases and Information Systems* (Springer International Publishing, 2016), pp. 11–27.
151. G. Zanella, B. Betancourt, H. Wallach, J. Miller, A. Zaidi, R. C. Steorts, Flexible models for microclustering with application to entity resolution, in *Proceedings of the 30th International Conference on Neural Information Processing Systems* (2016), pp. 1425–1433.
152. N. Monath, A. Kobren, A. Krishnamurthy, M. R. Glass, A. McCallum, Scalable hierarchical clustering with tree grafing, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2019), pp. 1438–1448.
153. R. C. Steorts, R. Hall, S. E. Fienberg, A Bayesian approach to graphical record linkage and deduplication. *J. Am. Stat. Assoc.* **111**, 1660–1672 (2016).
154. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques* (Morgan Kaufmann Publishers, 2011).
155. M. A. Hernández, S. J. Stolfo, The merge/purge problem for large databases, in *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery, 1995), pp. 127–138.
156. M. A. Hernández, S. J. Stolfo, Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining Knowledge Discov.* **2**, 9–37 (1998).
157. N. Bansal, A. Blum, S. Chawla, Correlation clustering. *Mach. Learn.* **56**, 89–113 (2004).
158. V. Filkov, S. Skiena, Integrating microarray data by consensus clustering, in *Proceedings of the International Conference on Tools with Artificial Intelligence* (ICTAI, 2003), pp. 418–426.
159. M. Charikar, V. Guruswami, A. Wirth, Clustering with qualitative information. *J.Comput. Syst. Sci.* **71**, 360–383 (2005).
160. N. Ailon, M. Charikar, A. Newman, Aggregating inconsistent information: Ranking and clustering. *J. Assoc. Comput. Mach.* **55**, 1–27 (2008).
161. A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation. *ACM Trans. Knowledge Discov. Data* **1**, 4-es (2007).
162. S. C. Johnson, Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
163. I. Bhattacharya, L. Getoor, A latent dirichlet model for unsupervised entity resolution, in *Proceedings of the Sixth SIAM International Conference on Data Mining* (Society for Industrial and Applied Mathematics, 2006), pp. 47–58.
164. J. B. Copas, F. J. Hilton, Record linkage: Statistical models for matching computer records. *J. R. Stat. Soc. A* **153**, 287–320 (1990).
165. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
166. C. E. Antoniak, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annal. Stat.* **2**, 1152–1174 (1974).
167. S. N. MacEachern, Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Stat. Simul. Comput.* **23**, 727–741 (1994).
168. S. N. MacEachern, Computational methods for mixture of Dirichlet process models, in *Practical Nonparametric and Semiparametric Bayesian Statistics* (Springer, 1998), pp. 23–43.
169. M. Perman, J. Pitman, M. Yor, Size-biased sampling of Poisson point processes and excursions. *Probability Theory Related Fields* **92**, 21–39 (1992).
170. J. Pitman, M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annal. Probability* **25**, 855–900 (1997).
171. J. F. C. Kingman, The representation of partition structures. *J. Lond. Math. Soc.* **s2-18**, 374–380 (1978).
172. T. Broderick, J. Pitman, M. I. Jordan, Feature allocations, probability functions, and paintboxes. *Bayesian Anal.* **8**, 801–836 (2013).
173. R. C. Steorts, M. Barnes, W. Neiswanger, Performance bounds for graphical record linkage, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (2017), vol. 54, pp. 298–306.
174. J. E. Johndrow, K. Lum, D. B. Dunson, Theoretical limits of microclustering for record linkage. *Biometrika* **105**, 431–446 (2018).
175. J. Bleiholder, F. Naumann, Data fusion. *ACM Comput. Surv.* **41**, 1–41 (2009).
176. S. Cohen, Y. Sagiv, An incremental algorithm for computing ranked full disjunctions, in *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (Association for Computing Machinery, 2005), pp. 98–107.
177. L. L. Yan, M. T. Ozsu, Conflict tolerant queries in aurora, in *Proceedings Fourth IFCIS International Conference on Cooperative Information Systems* (IEEE, 1999), pp. 279–290.
178. P. Bohannon, W. Fan, M. Flaster, R. Rastogi, 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* (Association for Computing Machinery, 2005), pp. 143–154.
179. A. Culotta, M. Wick, R. Hall, M. Marzilli, A. McCallum, Canonicalization of database records using adaptive similarity measures, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2007), pp. 201–209.
180. J. Murray, A unified framework for de-duplication and population size estimation (invited discussion). *Bayesian Anal.* **15**, 664–669 (2020).
181. J. Lane, V. Stodden, S. Bender, H. Nissenbaum, *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge Univ. Press, 2014).
182. A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in *Proceedings of the IEEE Symposium on Security and Privacy* (IEEE, 2008), pp. 111–125.
183. L. Sweeney, K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl-Based Syst.* **10**, 557–570 (2002).
184. S. Fienberg, A. Slavković, in *Data Privacy and Confidentiality* (International Encyclopedia of Statistical Science, Springer-Verlag, 2011), pp. 342–345.
185. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P. P. de Wolf, *Statistical Disclosure Control* (John Wiley & Sons, 2012).
186. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in *Theory of Cryptography Conference*, S. Halevi, T. Rabin, Eds. (Springer, 2006), pp. 265–284.
187. R. Hall, S. E. Fienberg, Privacy-preserving record linkage, in *Proceedings of the 2010 International Conference on Privacy in Statistical Databases* (Springer, 2010), pp. 269–283.
188. D. Vatsalan, P. Christen, V. S. Verykios, A taxonomy of privacy-preserving record linkage techniques. *Inform. Syst.* **38**, 946–969 (2013).
189. D. Vatsalan, Z. Sehili, P. Christen, E. Rahm, Privacy-preserving record linkage for big data: Current approaches and research challenges, in *Handbook of Big Data Technologies*, A. Y. Zomaya, S. Sakr, Eds. (Springer International Publishing, 2017), pp. 851–895.
190. W. Jiang, C. Clifton, A secure distributed framework for achieving k-anonymity. *VLDB J.* **15**, 316–333 (2006).
191. N. Mohammed, B. C. Fung, M. Debbabi, Anonymity meets game theory: Secure data integration with malicious participants. *VLDB J.* **20**, 567–588 (2011).
192. N. Mohammed, D. Alhadidi, B. C. Fung, M. Debbabi, Secure two-party differentially private data release for vertically partitioned data. *IEEE Trans. Dependable Secure Comput.* **11**, 59–71 (2014).
193. X. Cheng, P. Tang, S. Su, R. Chen, Z. Wu, B. Zhu, Multi-party high-dimensional data publishing under differential privacy. *IEEE Trans. Knowledge Data Eng.* **32**, 1557–1571 (2020).
194. M. Wilke, E. Rahm, Towards multi-modal entity resolution for product matching, in *Proceedings of the 32nd GI-Workshop on Foundations of Databases* (*Grundlagen von Datenbanken*) (GVDB, 2021).
195. H. Köpcke, E. Rahm, Frameworks for entity matching: A comparison. *Data Knowledge Eng.* **69**, 197–210 (2010).
196. F. Gregg, D. Eder, Dedupe (2015); https://github.com/dedupeio/dedupe [retrieved 29 July 2020].
197. J. de Bruin, recordlinkage 0.14 (2019); https://pypi.org/project/recordlinkage/ [released 1 December 2019; retrieved 29 July 2020].
198. P. Christen, Febrl—An open source data cleaning, deduplication and record linkage system with a graphical user interface, in *Proceedings of the 14th ACM International*

Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery, 2008), pp. 1065–1068.
199. Y. Govind, P. Konda, P. Suganthan, P. Martinkus, P. Nagarajan, H. Li, A. Soundararajan, S. Mudgal, J. R. Ballard, Entity matching meets data science: A progress report from the magellan project, in *Proceedings of the 2019 International Conference on Management of Data* (Association for Computing Machinery, 2019), pp. 389–403.
200. P. Konda, S. Das, P. Suganthan G. C., A. H. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, S. Prasad, G. Krishnan, R. Deep, V. Raghavendra, Magellan: Toward building entity matching management systems. *Proc. VLDB Endowment* **9**, 1197–1208 (2016).
201. M. Sariyar, A. Borg, The RecordLinkage package: Detecting errors in data. *R J.* **2**, 61–67 (2010).
202. M. Friedrichs, C. Webster, B. Marsh, J. Dice, S. Lee, fedmatch: Fast, flexible, and user-friendly record linkage methods (2021). R package version 2.0.3.
203. R. Linacre, S. Lindsay, splink: Probabilistic record linkage and deduplication at scale; https://github.com/moj-analytical-services/splink (2021).
204. L. Gagliardelli, G. Simonini, D. Beneventano, S. Bergamaschi, Sparker: Scaling entity resolution in spark, in *EDBT 2019: 22nd International Conference on Extending Database Technology* (PRT, 2019).
205. G. Papadakis, L. Tsekouras, E. Thanos, G. Giannakopoulos, T. Palpanas, M. Koubarakis, The return of JedAI: End-to-end entity resolution for structured and semi-structured data. *Proc. VLDB Endowment* **11**, 1950–1953 (2018).
206. K.-N. Tran, D. Vatsalan, P. Christen, Geco: An online personal data generator and corruptor, in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (Association for Computing Machinery, 2013), pp. 2473–2476.
207. M. Bilenko, R. Mooney, Riddle: Repository of information on duplicate detection, record linkage, and identity uncertainty (2006); www.cs.utexas.edu/users/ml/riddle/ [retrieved 29 July 2020].
208. B. Spahn, "Before the American voter," thesis, Stanford University (2019).
209. J. Tang, A. C. Fong, B. Wang, J. Zhang, A unified probabilistic framework for name disambiguation in digital library. *IEEE Trans. Knowledge Data Eng.* **24**, 975–987 (2012).
210. V. I. Torvik, N. R. Smalheiser, Author-ity 2009—Pubmed author name disambiguated dataset (2009).
211. J. Martin Montull, Inspire: Managing metadata in a global digital library for high-energy physics, in *Research Conference on Metadata and Semantic Research* (Springer, 2011), pp. 269–274.

Acknowledgments: We thank T. Enamorado, the deputy editor, and the reviewers for providing comments that have greatly improved the work. **Funding:** We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche du Québec—Nature et Technologies, NSF Career Award 1652431, and the Alfred P. Sloan Foundation. **Author contributions:** All authors wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests.

Submitted 31 March 2021
Accepted 4 February 2022
Published 25 March 2022
10.1126/sciadv.abi8021