Data Cleaning Pipeline

Rebecca C. Steorts

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

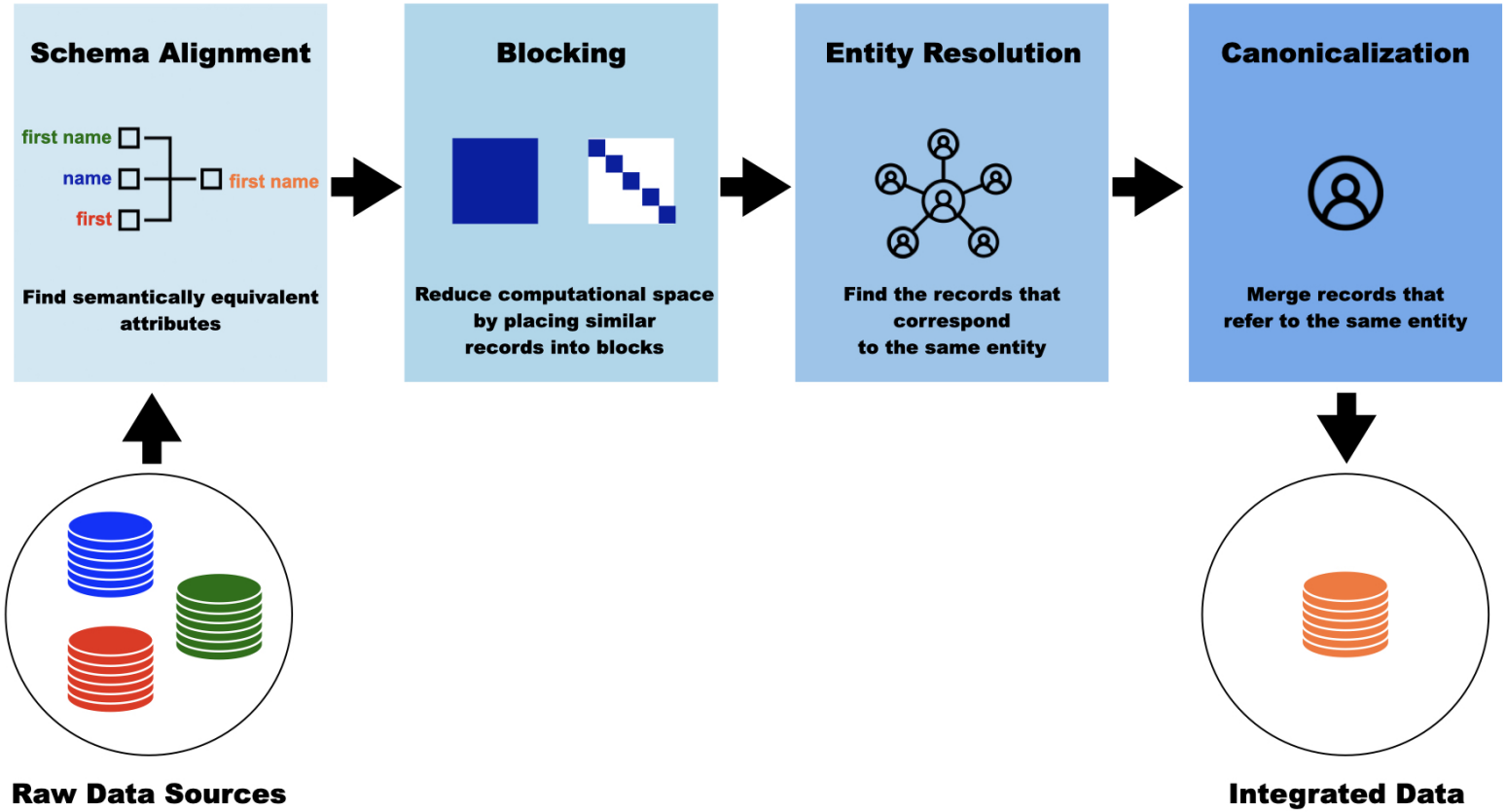
This work is partially supported by NSF CAREER Award 1652431. The views are of the author and not any agency, etc.

August 28, 2024

Goals

1. Enumerating a census.
2. Enumerating those that have died in a conflict (such as Syria).
3. Predicting those in poverty in small regions from survey data.
4. Predicting results of elections from voter registration data.
5. Predicting housing/rental prices from Zillow data.

Each task may contain duplicated information, which is problematic for the underlying task at hand.


Schema Alignment
first name
name
first
first name
Find semantically equivalent attributes
Blocking
Reduce computational space by placing similar records into blocks
Entity Resolution
Find the records that correspond to the same entity
Canonicalization
Merge records that refer to the same entity
Raw Data Sources
Integrated Data

1. The most important information in the pipeline is known as the profile or the record.

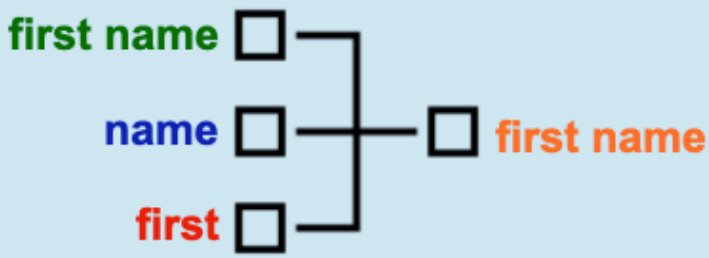
2. Each profile or record is a collection of attributes/fields about a person, organization, or object.

3. Commonly collected attributes about people are name, address, phone number, gender, among other types of information.

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

Entity 1	Entity 2	Entity 3	Entity 4	Entity 5
d1 d2	d3 d4	d5	d6	d7 d8 d9 d10

Schema Alignment



Find semantically equivalent attributes

1. It is important that we align attributes when our schemata are disparate.
2. The goal is to create alignments of attributes based upon the following:
    1. Similarity
    2. Structure
    3. Attributes Present

Formally, this is known as identifying "semantically equivalent attributes", such as first name, first, and name.

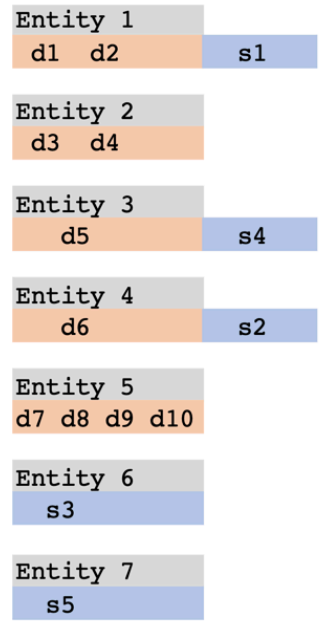
[Bernstein et al., 2011, Madhavan et al., 2001].

1. This stage leverages the attribute values from the records/profiles.
2. Schema knowledge is used (if available).
3. The goal is to learn attribute mappings between the data sources.
4. The goal is to also find "transformations, correspondences, or rules between the attributes." [Tejada et al., 2002, Yan et al., 2001].
5. Common transformations are used, such as: "Dr." to "Drive" or "3rd" to "third" [Active Atlas, Tejada et al., 2002].

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

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

(a)



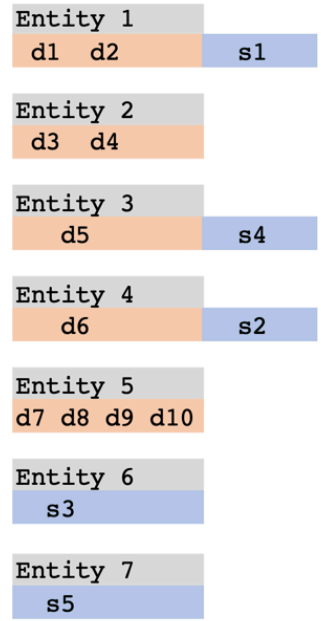
(b)

Figure: An example two databases: (a) the input databases and (b) the corresponding entities.

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

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

(a)



(b)

Figure: An example two databases: (a) the input databases and (b) the corresponding entities.

Alignment rules: first and last/name; sex and gender.

1. It is important that the schema are coded for all databases in the same way.
2. The naming structured should be well organized and documented in a relational database.
3. More information can be found in Papadakis et. al (2021) for more information and other illustrations.

Blocking

Reduce computational space by placing similar records into blocks

1. Blocking operates in a schema-aware fashion, assuming that the input data adheres to a known schema or to aligned schemata.
2. Based on this assumption and respective domain knowledge, the most suitable attributes are used for extracting one or more representative signatures from each profile.
3. These signatures are called blocking keys and are composed of (combinations of) parts of values from the most informative attributes.
4. Assuming that these keys reflect the overall similarity of profile pairs, profiles with identical or similar keys are placed into the same block to be compared in the entity resolution stage.

1. Standard Blocking (SB) [Fellegi and Sunter, 1969] requires an expert to manually define a part or a transformation of one or more attribute values as the single blocking key of each profile.
2. Every profile is then placed in the block corresponding to its blocking key.
3. To increase its robustness, a multi-pass functionality is applied in practice, i.e., SB is combined with several different definitions of blocking keys.

1. One common type of blocking is using q-grams [Christen, 2012b, Papadakis et al., 2015].
2. This converts SB keys into sub-sequences of q characters (q-grams) and defines a block for every distinct q-gram.

There are multiple extensions to these in the computer science and database management literature.

How might we define a blocking criteria for these data sources?

Define the blocking key the concatenation of the following three pieces of information:

1. (i) {"Name," Last2Characters},
2. (ii) {"Address," Last2Characters},
3. and (iii) {"Gender," FirstCharacter}.

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

(a)

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

(b)

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

(c)

thet, hetM
d1 d2
rsay, sayF
d3 d4
thrt, hrtM
d7 d8 d9 d10

(d)

Figure: (a) the input data source with bolded information used in blocking keys, (b) the blocking keys via SB, (c) the blocking keys of 4-grams blocking, and (d) the blocks of 4-grams blocking.

There are many other ways that blocking criteria can be defined and many options are reviewed in Papadakis et. al (2021).

Entity Resolution

Find the records that correspond to the same entity

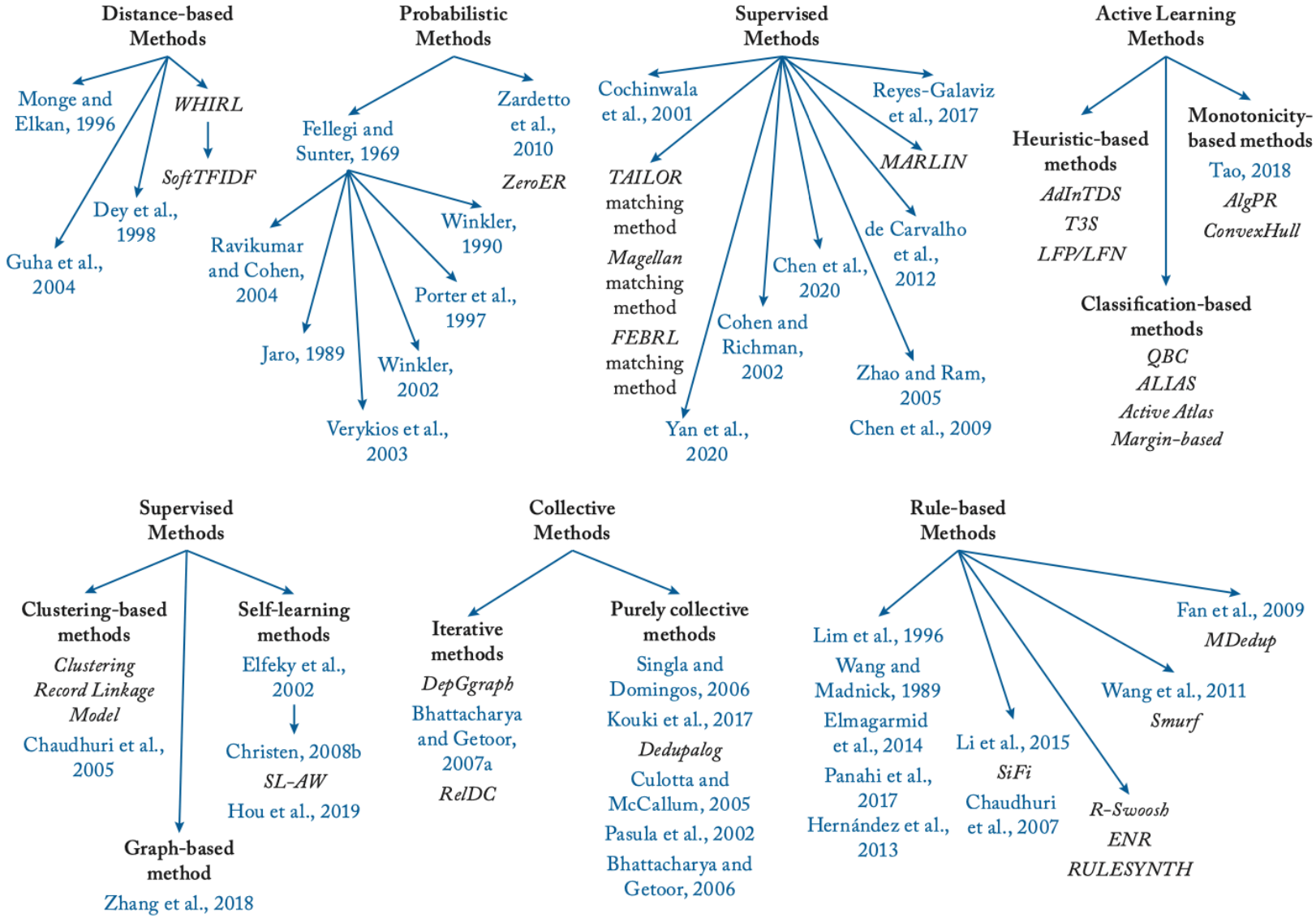


Figure: Citation: Papadakis et. al (2021).

Canonicalization

Merge records that refer to the same entity

In summary, after all the stages the output is an integrated data set with unique identifiers that can be used in statistical analyses.

1. How can we work together to enable that these systems work well?
2. How should these system be implemented? (scala, java, queries that work with scala/java)
3. Should we avoid scripting languages such as python or R?
4. How do we get students/collaborators involved in the building of complex pipelines described?
5. Are there limiting resources at play?

Thank you!

Questions?

Contact: beka@stat.duke.edu, rebecca.carter.steorts@census.gov

`https://github.com/resteorts/record-linkage-tutorial`

`https://www.science.org/doi/10.1126/sciadv.abi8021`

`https://github.com/cleanzr`

Thank you to Anup Mathur, Krista Park, Kristen Olsen, and Jenny Thompson for conversations or feedback that led to this presentation.