



DATA LINKAGE FLOW CHART AND GUIDE

sdp.cepr.harvard.edu
sdp@gse.harvard.edu

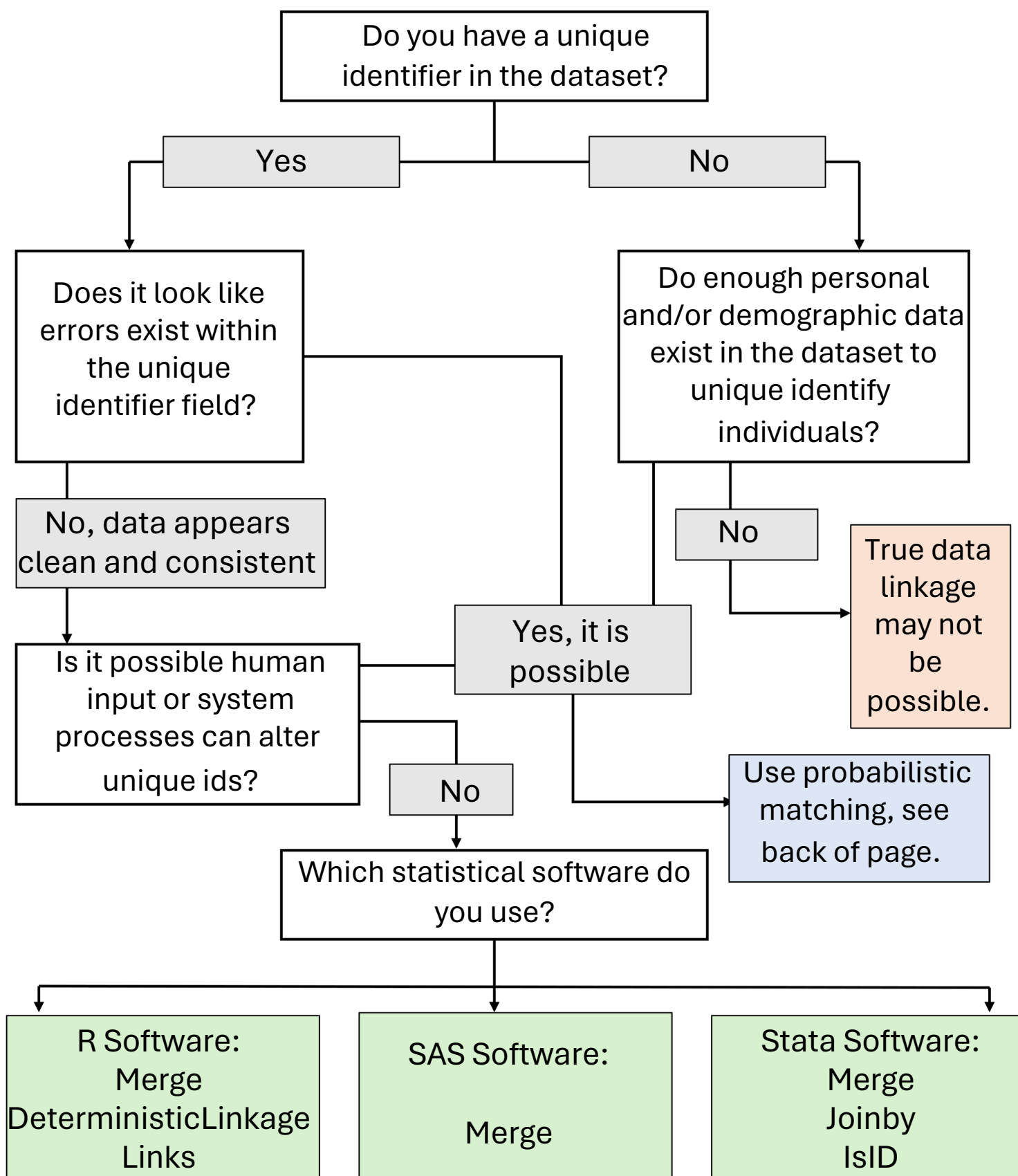
This tool is part of [Preschool Supply & Demand](#), a collaboration between the Strategic Data Project and representatives from the California Department of Education, Denver Public Schools, Denver Preschool Program, South Carolina First Steps, Chicago Early Childhood Integrated Data System, National Head Start Association, Massachusetts Department of Early Education and Care, and Washington Education Research and Data Center. It is supported by the Gates Foundation.



Center for Education Policy Research
HARVARD UNIVERSITY

Probabilistic or Deterministic Data Linkage?

Answer the following questions to determine the best type of data linkage approach based on available data, and to find code commands for data linkage in R, SAS, and Stata.



Probabilistic Matching Methods

Record Linkage

- Record linkage commands calculates matching probabilities based on data similarities across multiple fields.
- This approach can be customized with weighting, thresholds, etc.
- Both quantitative and/or qualitative data points can be used.

Distance Metrics

- This approach compares string variables, such as names or institutions, and calculates similarities across string sequences using various distance metrics. A common example of a distance metric is Levenshtein Distance.

Supervised Machine Learning Algorithms

- This probabilistic matching approach can be used to mine datasets for matches, especially when data with known matches is available for training the algorithm.

Fine-tuning probabilistic matching

- **Weighting:** Adjust the data linkage equation so that different variables are assigned different weights to reflect their importance and reliability.
- **Thresholds:** A threshold is a cut-off point for the matching score or probability and should be chosen carefully to balance the trade-off between retaining false positives and losing true matches.
- **Matching Variables:** Less can be more when it comes to the number of variables used for record linkage. Determine the most necessary data points for true matches and exclude any variables that would only add noise to the linkage equation.

Probabilistic matching command codes by software:

R Software:

RecordLinkage
MatchIt
FastLink
StringDist

SAS Software:

The Link King
Matching Process
Macros

Stata Software:

Reclink2
Clrevmatch
DtaLink
Ultimatch
Strdist