DATA LINKAGE WORKFLOWS

John 'Mac' McDonald

Centre for Longitudinal Studies

Institute of Education

Data linkage

Data linkage (also known as record linkage)

- **for organising ONE dataset**
  - data cleaning
  - removing duplicates
- **for merging TWO OR MORE datasets**
  - merging individual-level datasets
  - adding census data to survey data
- **for master data management**
  - linking new transactions/events to master records

Identification of Duplicates Given Name, Address, Age

Matching Information

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17
Javier Martinez	49 E Applecross Road	33
Haveir Marteenez	49 Aplecross Raod	36
Gillian Jones	645 Reading Aev	22
Jilliam Brown	123 Norcross Blvd	43

Record linkage . . .

"[is] a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events (said to be matched)."

Fellegi IP & Sunter AB (1969) A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210

Problem of record linkage

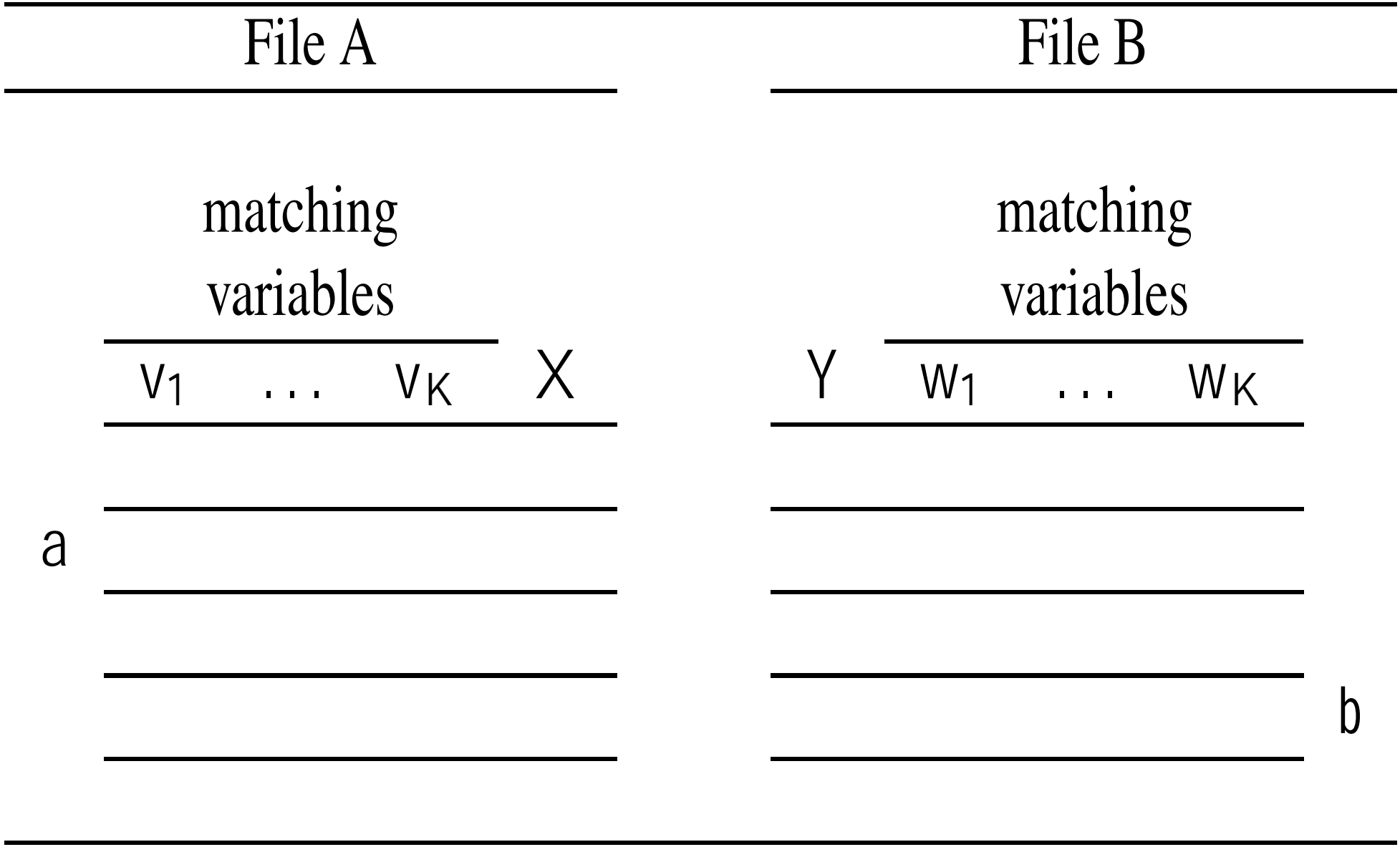
problem - **quickly and accurately** determining if pairs of records describe the same entity, but unique IDs to bring together the matching records are lacking

records must contain some common identifying information (**keys or matching variables**)

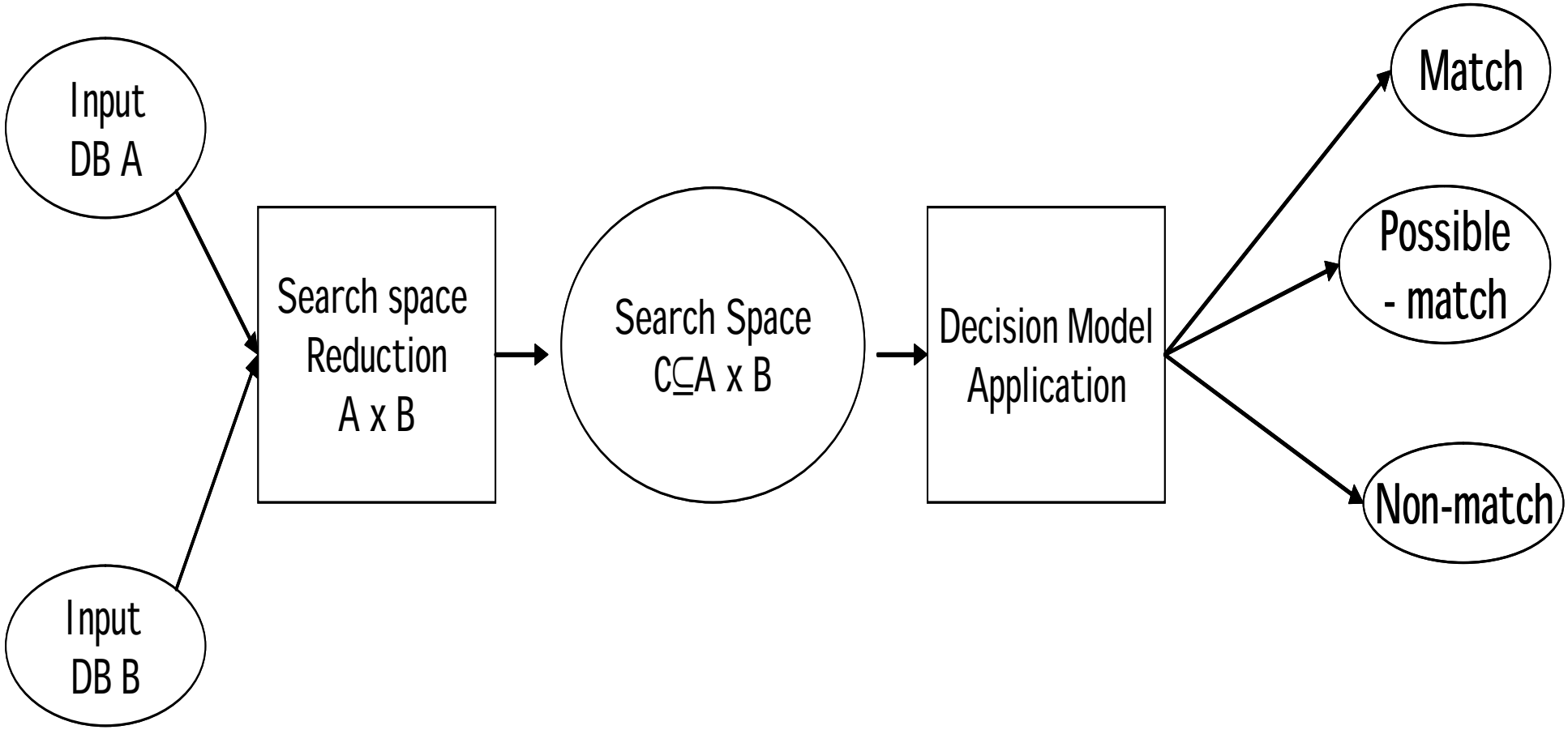
- unique identifier (ideal in theory)
- name and/or address
- age (DOB) and sex

N.B. for very large databases, processing time and accuracy are concerns and blocking can be used to reduce the total number of record pairs compared

Files A & B, record a in A & record b in B



Phases of record linkage



Data linkage

- **data linkage is context specific**, e.g. US addresses, scientific bibliography entries
- one-to-one linkage vs one-to-many linkage vs many-to-many linkage
- **no universal best method for data linkage**
- **linkage protocol used should be documented**

Linkage projects typically have three phases

- **pre-linkage**
  - data cleaning
  - processing data fields to recognize similarity
- **linkage phase: deciding whether two records are a**
  - **duplicate**
  - **match (link)**
- **post-linkage**
  - manual/clerical review of unlinked records
  - research using the linked data

Data linkage . . .

Data linkage is a challenging problem because of

- errors, variations and missing data on the information used to link records
- differences in data captured and maintained by different databases, e.g. age versus DOB
- data dynamics and database (DB) dynamics as data regularly and routinely change over time
  - name changes due to marriage & divorce
  - address changes

Methodology of record linkage

- **two distinct methodologies for data linkage**
- **deterministic linkage** methods involve exact one-to-one character matching of linkage variable(s)
- **probabilistic linkage** methods involve the calculation of linkage likelihood or linkage weights estimated given all the observed agreements and disagreements of the data values of the linkage variable(s)
- **probabilistic linkage methods can lead to much better linkage than simple deterministic linkage methods**

Methodology of record linkage . . .

- methods from computer science, statistics and operations research
- methods primarily implemented by computer scientists
- **general purpose versus domain specific**, e.g. US addresses, scientific bibliography entries
- **software for standardizing and parsing names and addresses that are used in the matching identifies**
  - **standardizing: replacing words with consistent abbreviations**, e.g. street = ST
  - **parsing: decomposing a string into a set of string components** which are individually compared

Data problems

- typos/mispelling
- letters or words out of order
- fused or split words
- missing or extra letters
- incomplete words
- extraneous information
- incorrect or missing punctuation
- abbreviations
- multiple errors

Methodology of record linkage . . .

- deterministic algorithms - unique key(s)
- probabilistic algorithms - model based
- data mining techniques, e.g. neural networks
- Bayesian methods
- fuzzy methods, e.g. search engine/wild cards
- Boolean or other rule based methods
- linguistic rules (names from different cultures)
- combination of algorithms

Deterministic linkage

- **simplest method of matching** - sort/merge
- exact matching ONLY works well if the linking data are perfect and present in all the databases you want to link
- works best when there is a single unique identifier (key)
- otherwise, matching based on sets of identifiers predetermined by the researcher
- identifiers have equal weight
- identifiers chosen by researcher or by availability
- **works best with high quality data, but yields less success than probabilistic linkage**

Deterministic linkage . . .

- **deterministic matching links records**
  - using a fixed set of matching variables
  - **exact one-to-one character matching of linking variables**
- sometimes only the first few characters of a field are used with a wildcard substituted for later characters
  - primitive, but widely implemented, approach to tolerating errors
  - Martin versus Martinez

Deterministic linkage . . .

- brings together record pairs very efficiently by sorting both files using common identifier(s), which is the idea of a key
- keys associated with concepts of sorting/indexing
- example keys: surname, first name and DOB
- **problem**
  - **offer no unique, known and accurate ID**
  - **missing values and partial agreements are common**

Surname	Name	Day of B	Year of B	freq
0	0	0	0	414138
0	0	0	1	5321
0	0	1	0	14004
0	0	1	1	168
0	1	0	0	3090
0	1	0	1	43
0	1	1	0	102
0	1	1	1	9
1	0	0	0	969
1	0	0	1	17
1	0	1	0	22
1	0	1	1	19
1	1	0	0	14
1	1	0	1	9
1	1	1	0	6
1	1	1	1	513

Probabilistic linkage

- **each matching variable is compared and assigned a score (weight) based on how well it matches**
- frequency analysis of data values is important
- uncommon value agreement stronger evidence for linkage, e.g. Rumplestilskin versus Smith
- **calculates a score for each field that indicates, for any pair of records, how likely it is that they both refer to the same entity**
- **sum the scores over fields**
- sort record pairs in order of their scores (weights)

Probabilistic linkage . . .

- cut off values for scores (weights) are used to distinguish between matches and non-matches
- **above a certain threshold, everything is a match** (link)
- **below a certain threshold, nothing is a match** (nonmatch or nonlink)
- **in between (grey area), possible match needs manual/clerical review**

Probabilistic linkage . . .

- **total score for a link between any two records is the sum of the scores generated from matching individual fields**
- score assigned to a matching of individual fields
  - is based on the probability that a matching variable agrees given that a comparison pair is a match
  - M-probability - similar to "sensitivity", i.e. the proportion of actual positives which are correctly identified

Probabilistic linkage . . .

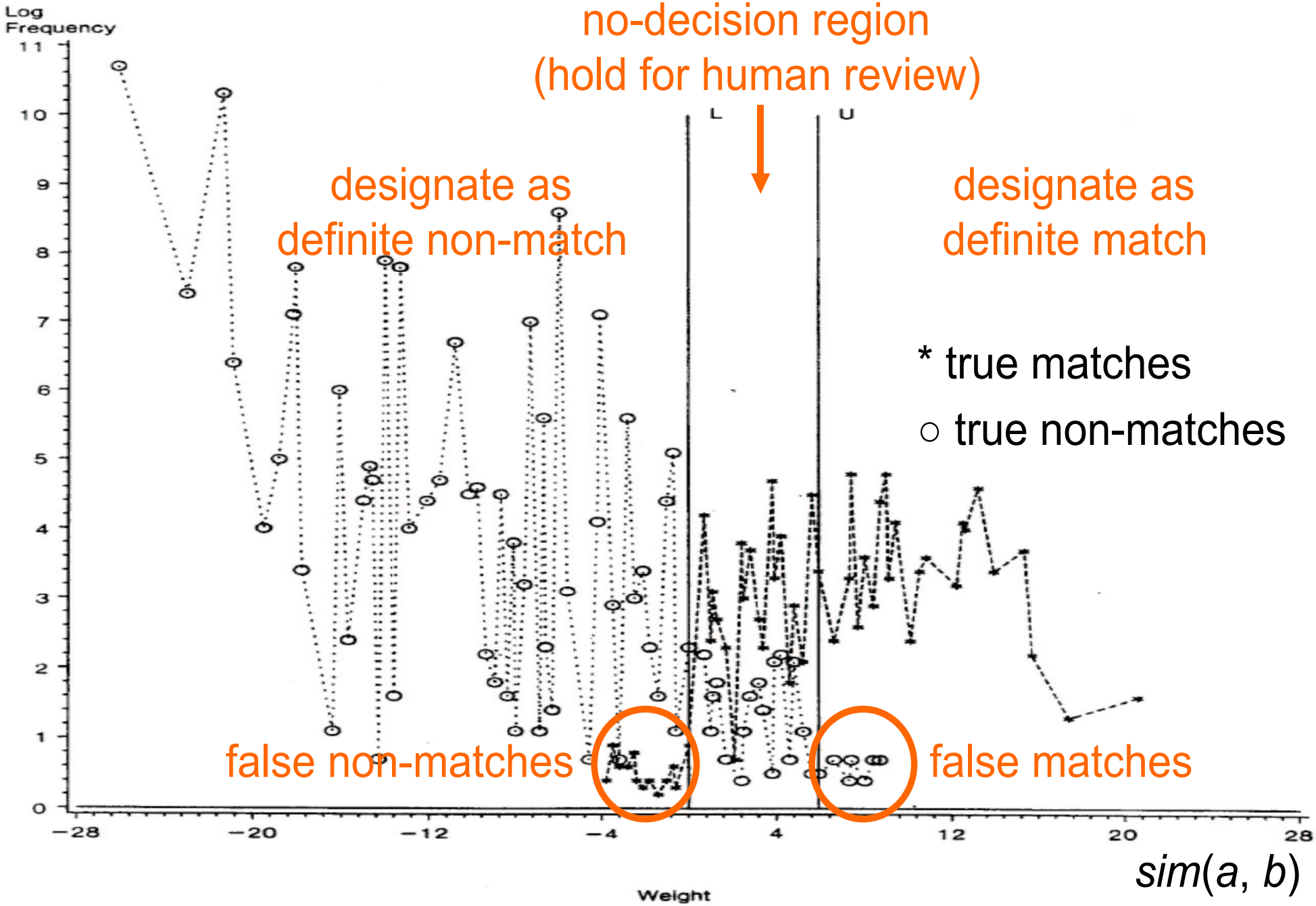
- score assigned to a matching of individual fields
  - reduced by the probability that a matching variable agrees given that a comparison pair is not a match (U = unmatched)
  - U-probability - similar to "specificity", i.e. the proportion of negatives which are correctly identified
- **agreement argues for linkage**
- **disagreement argues against linkage**
- full agreement stronger evidence for linkage than partial agreement

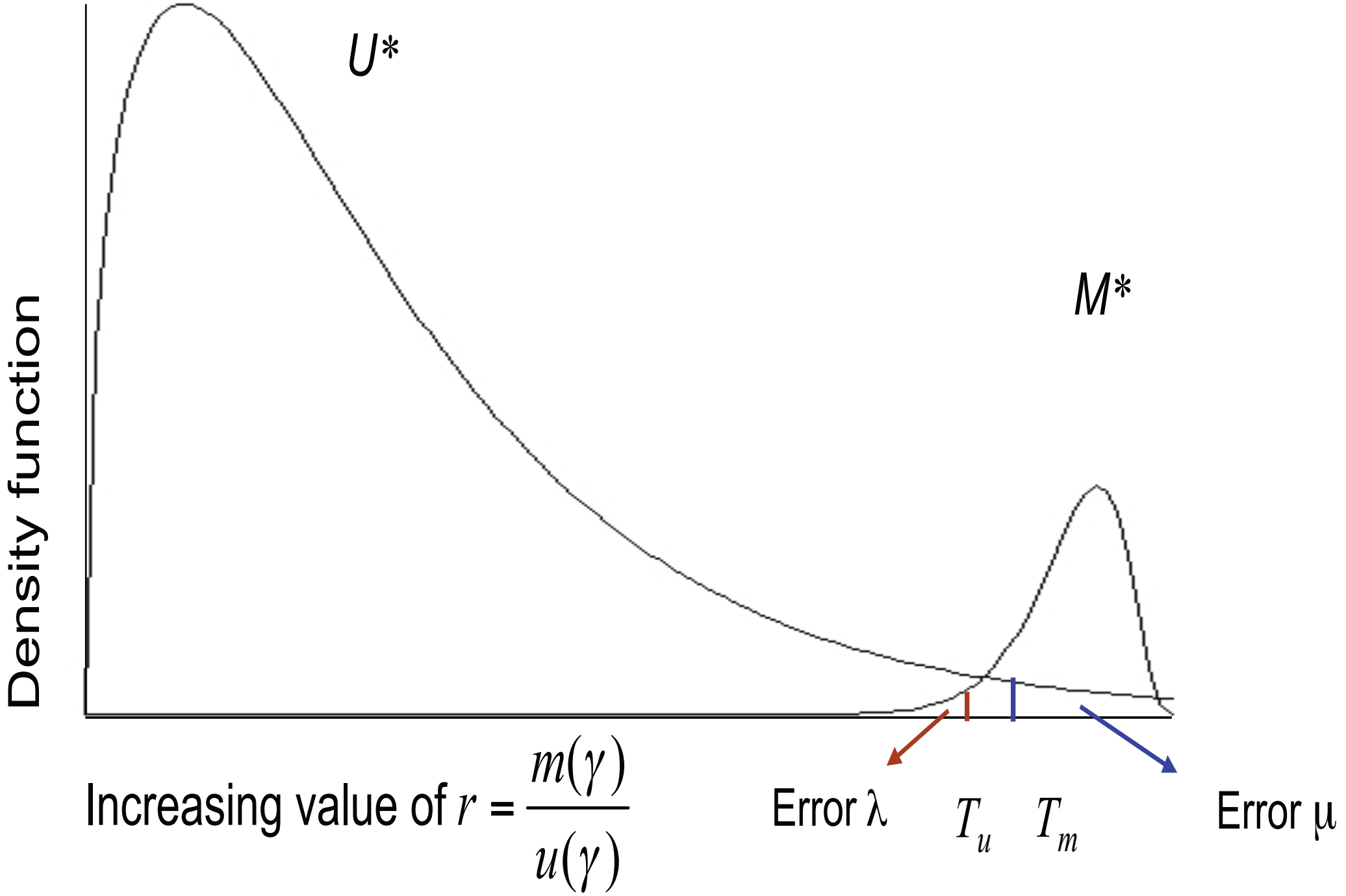
Probabilistic linkage . . .

- based on the probabilities of agreement or disagreement between the identifiers
- all identifiers do not have equal weight
- **accurate linkage is mainly dependent on the amount of discriminating power inherent in the variables common to the records that need to be matched and 'good' data**

Fellegi IP & Sunter AB (1969) A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210

Fellegi-Sunter model



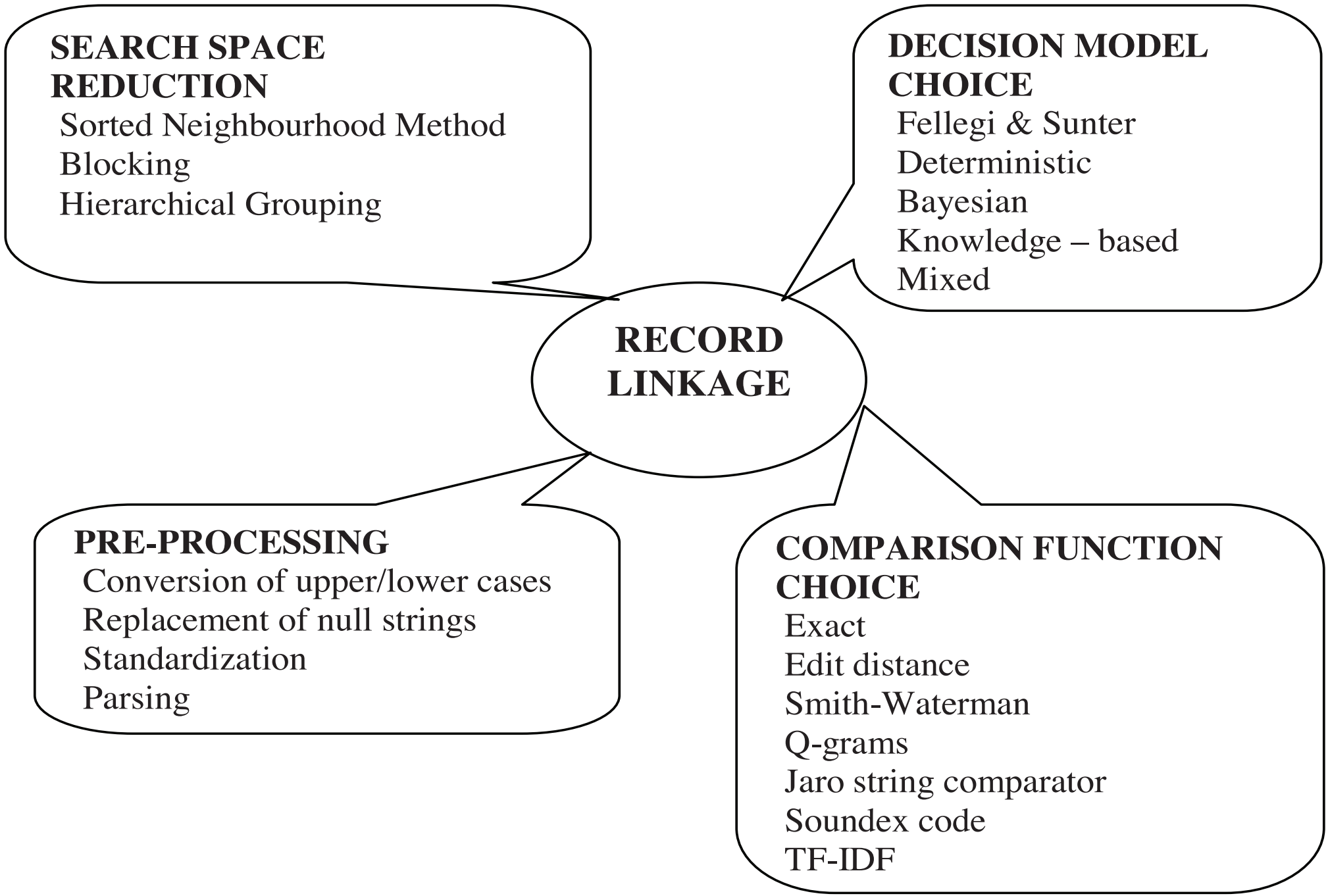

U*
M*
Density function
Increasing value of $r = \frac{m(\gamma)}{u(\gamma)}$
Error λ
T_u
T_m
Error μ

RELAIS

- RELAIS (Record Linkage At IStat) toolkit
- **an open source toolkit for building record linkage workflows**
- JAVA based
- statistical methods implemented in R

http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/

Figure 1: *The record linkage complexity*



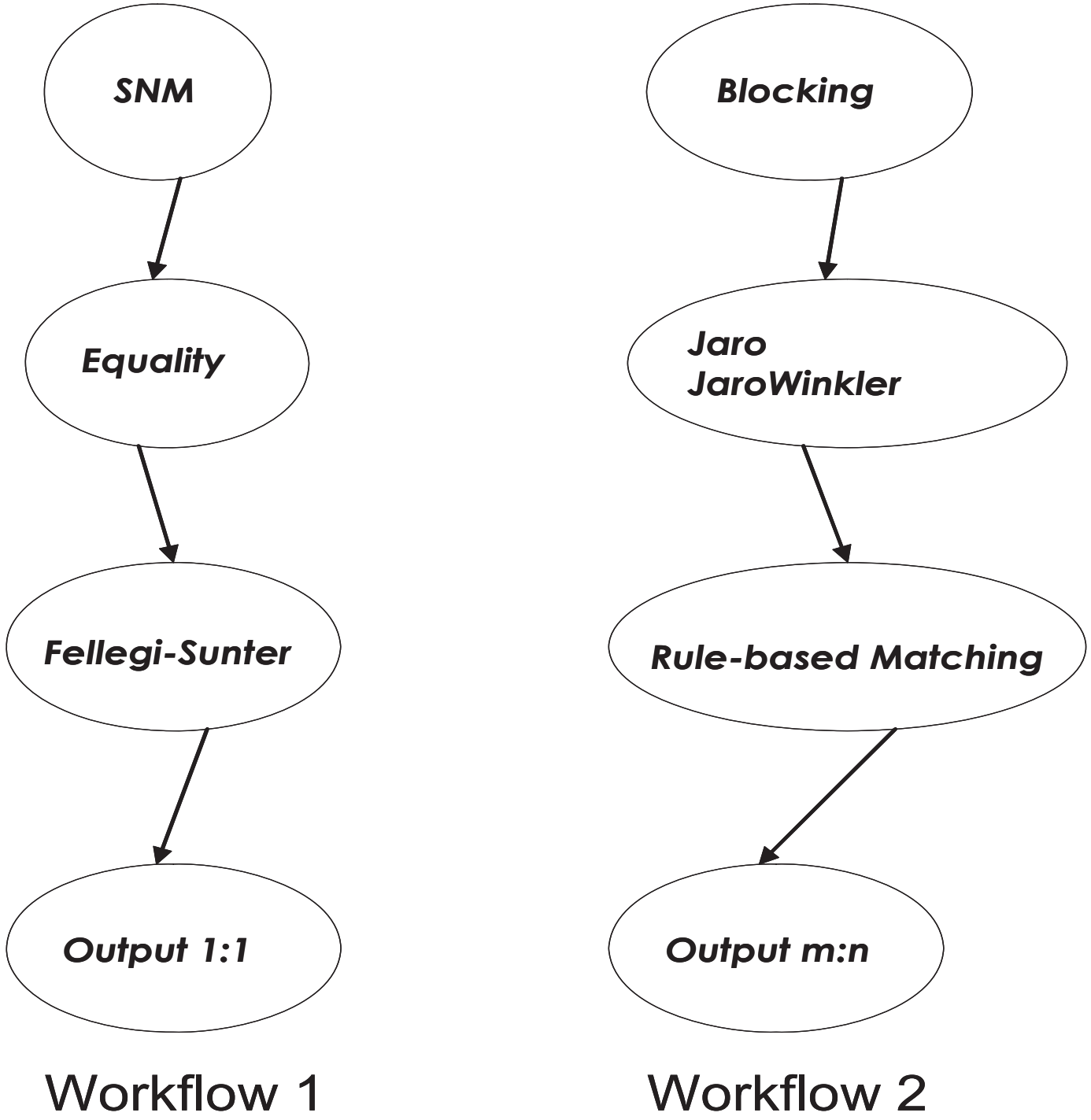
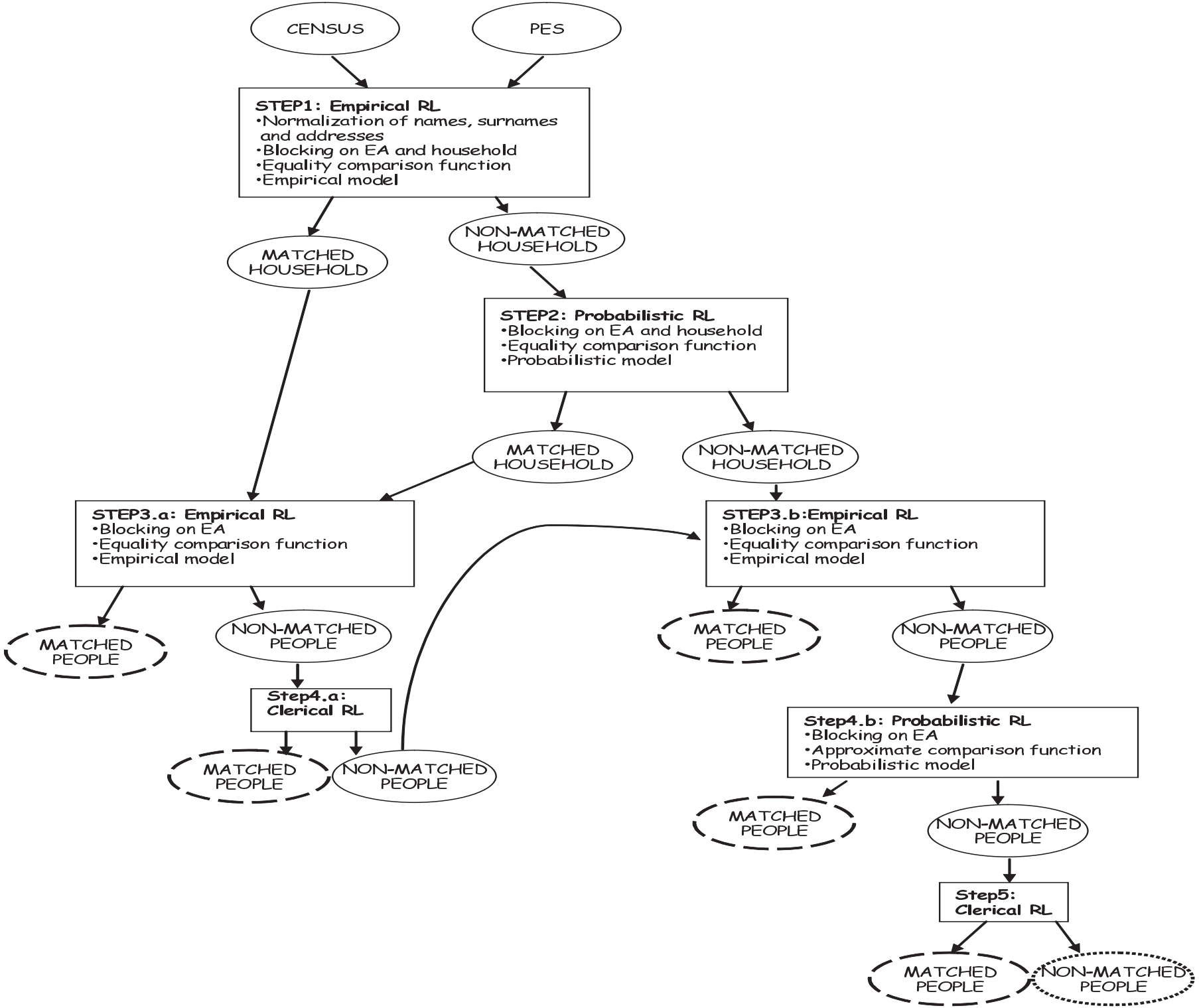


Figure 2: Examples of record linkage workflows

Requirement		Choice
Data requirement	Hierarchical structure	Workflow iteration: • Higher level (household) • Lower level (person)
	High quality	Equality comparison function on most of the phases
	Large data set	Blocking and phase iteration
Application requirement	Not significant errors in matching process	Probabilistic model and clerical review phase

Figure 7: An example of a pattern for building record linkage workflows


CENSUS
PES
STEP1: Empirical RL
•Normalization of names, surnames and addresses
•Blocking on EA and household
•Equality comparison function
•Empirical model
MATCHED HOUSEHOLD
NON-MATCHED HOUSEHOLD
STEP2: Probabilistic RL
•Blocking on EA and household
•Equality comparison function
•Probabilistic model
MATCHED HOUSEHOLD
NON-MATCHED HOUSEHOLD
STEP3.a: Empirical RL
•Blocking on EA
•Equality comparison function
•Empirical model
STEP3.b:Empirical RL
•Blocking on EA
•Equality comparison function
•Empirical model
MATCHED PEOPLE
NON-MATCHED PEOPLE
Step4.a: Clerical RL
MATCHED PEOPLE
NON-MATCHED PEOPLE
MATCHED PEOPLE
NON-MATCHED PEOPLE
Step4.b: Probabilistic RL
•Blocking on EA
•Approximate comparison function
•Probabilistic model
MATCHED PEOPLE
NON-MATCHED PEOPLE
Step5: Clerical RL
MATCHED PEOPLE
NON-MATCHED PEOPLE

Results using deterministic approach

1°Merge : (1,1,1,1) + on the 1°Merge-residuals 2°Merge : (1,0,1,1)

+ on the residuals 3°Merge : (1,1,0,1)

X	P(*X*=1|*M*)	P(*X*=1|*U*)
Surname	0.9853	0.0023
Name	0.9650	0.0074
Day of birth	0.9825	0.0327
Year of birth	0.9889	0.0127

Observed FMR=0.005

Observed FNMR=0.06

True Linkage Status

Results of the Linkage Procedure	***Matched***	***Not Matched***	***Total***
Matched	**538**	**3**	**541**
Not Matched	**35**		
Total	**573**		

Results under local independence assumption

		True Linkage Status		
		Matched	***Not Matched***	***Total***
Results of the Linkage Procedure	***Matched***	**567**	**10**	**577**
	Not Matched	**6**		
	Total	**573**		

The linkage results are “appreciable” but the linkage errors are not well estimated

Observed FMR=0.017 vs the expected 0.001

Observed FNMR=0.010 vs the expected 0.0001

Scheuren and Winkler (1993)

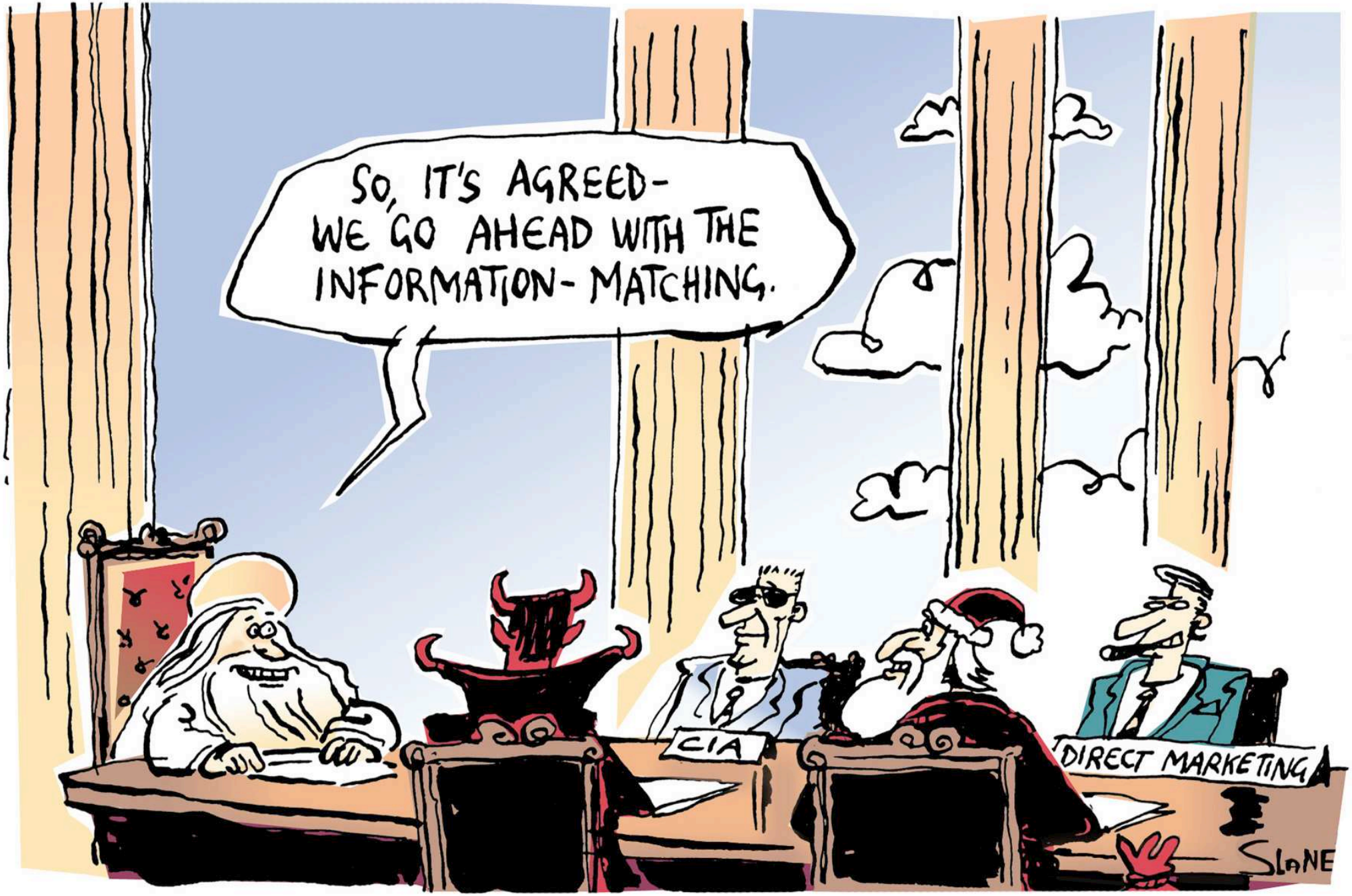
- **What should the linker do to help the analyst?**
- **What should the analyst know about the linkage and how should that information be used?**
- **In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analysis done.**

Scheuren F & Winkler W E (1993) Regression analysis of data files that are computer matched - part 1. Survey Methodology 19, 39-58

What should the matching variable(s) be?

Jenkins S et al (2006) **The feasibility of linking household survey and administrative record data**: New evidence for Britain. International Journal of Social Research Methodology 11, 29-43

- **IDs are subject to problems of survey item non-response and measurement error**
- 5 linkages: respondent-supplied NINO & 4 linkages using different combinations of sex, name, address and DOB
- as many linkages were made using non-NINO-based matches as were made using matches on NINO
- former were also relatively accurate when assessed in terms of false-positive and false-negative linkage rates.


SO, IT'S AGREED- WE GO AHEAD WITH THE INFORMATION- MATCHING.
CIA
DIRECT MARKETING
SLANE