

Webinar at INSEE

# RELAIS: a statistical toolkit for record linkage

Tiziana Tuoto, Mauro Scanu

12 April 2021

# The RELAIS philosophy

- Decompose a RL project in its constituting phases
- The best solution for all cases does not exist : choose the most appropriate technique for each phase, depending on application and data requirements, not only on practitioner's skill
- Dynamically build ad-hoc workflow for each RL problem

# Short History of RELAIS

RELAIS = Record Linkage at Istat

The RELAIS background: record linkage in Istat before 2006

- Wide use of record linkage in different production processes: first experiences date back to '80s
- Common practice was to develop *ad hoc* linkage procedures for each project, basically via deterministic techniques
- Little awareness of linkage errors in further analyses of linked data

The RELAIS background: record linkage in Istat before 2006

- Only a few official experiences with probabilistic approach
  - ❑ longitudinal estimates for LFS (Labour Force Survey)
  - ❑ building of the Active Enterprises Statistical Archive
  - ❑ post-enumeration surveys for the population Censuses in 1991 and 2001
- Decennial studies on the Fellegi-Sunter methodology with the EM algorithm

# Starting the Relais idea

- ❑ Methodological unit for micro-data integration
  - Test methods for RL
  - Census and PES data
  - Problems: a lot of work for developing procedures with different software

- ❑ IT unit for internal software development
  - Developing, testing, migrating software for the Istat processes
  - Attention forward the open source

## The Relais Project

- ❑ Joint informal group in march 2006 (5 people)
- ❑ International Conference IQIS (June 2006)
- ❑ Alfa version (January 2007)

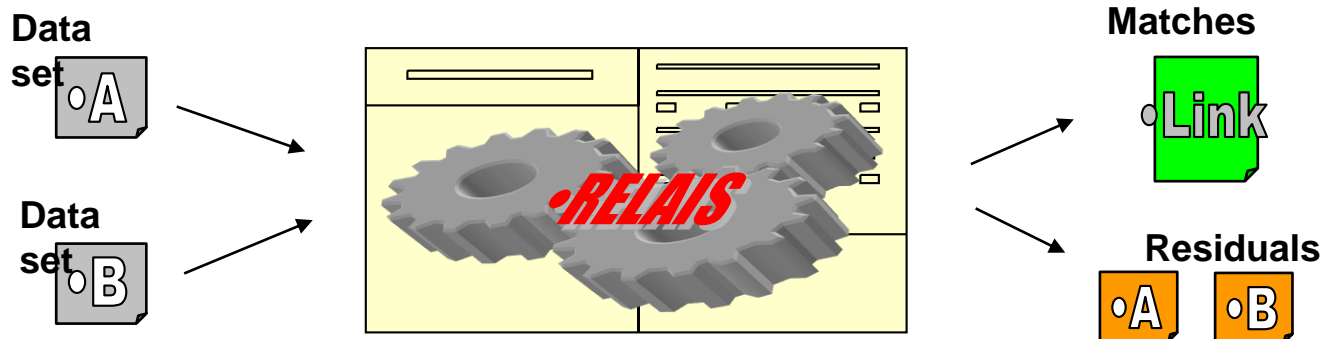
# RELAIS milestones

- ❑ First version RELAIS 1.0 released in February 2008
  - ❑ probabilistic method
  - ❑ architectural structure based on file
  
- ❑ Version RELAIS 2.0 released in June 2009
  - ❑ adding deterministic methods
  - ❑ architectural structure based on database
  
- ❑ Version RELAIS 2.1 released in May 2010
- ❑ Version RELAIS 2.2 released in May 2011
- ❑ Version RELAIS 2.3 releases in May 2012
- ❑ Version RELAIS 3.0 released in July 2015

# Decompose RL in phases

1. Pre-processing of the input files
2. Creation-Reduction of the search space of link candidate pairs
3. Choice of the matching variables
4. Choice of the comparison function
5. Choice of the decision model
6. Identification of unique links
7. RL evaluation

# The Main Idea



Phase  
Decomposition



Search Space  
Creation



Comparison  
Phase



Decision Model



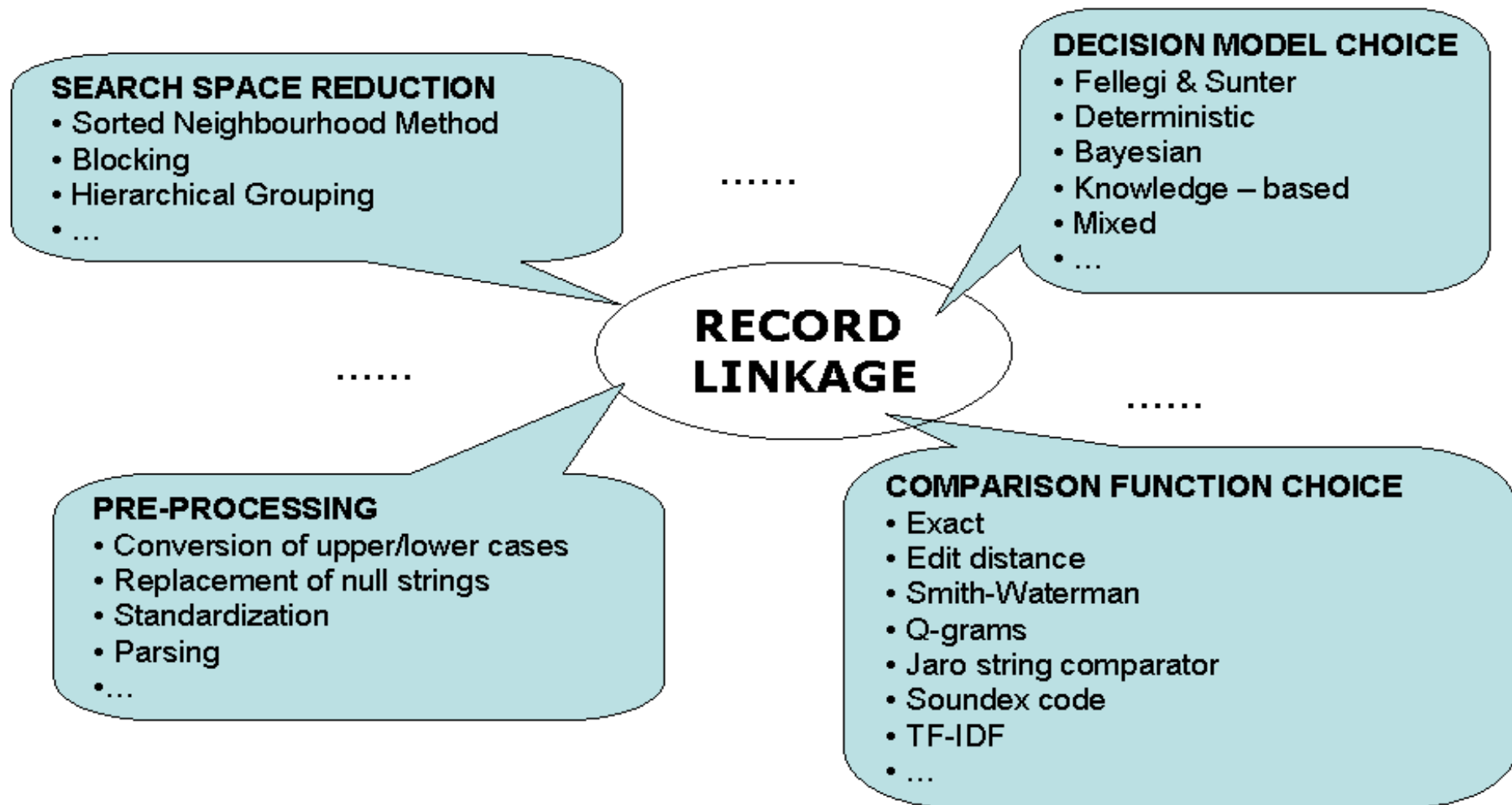
Thresholds  
Choice



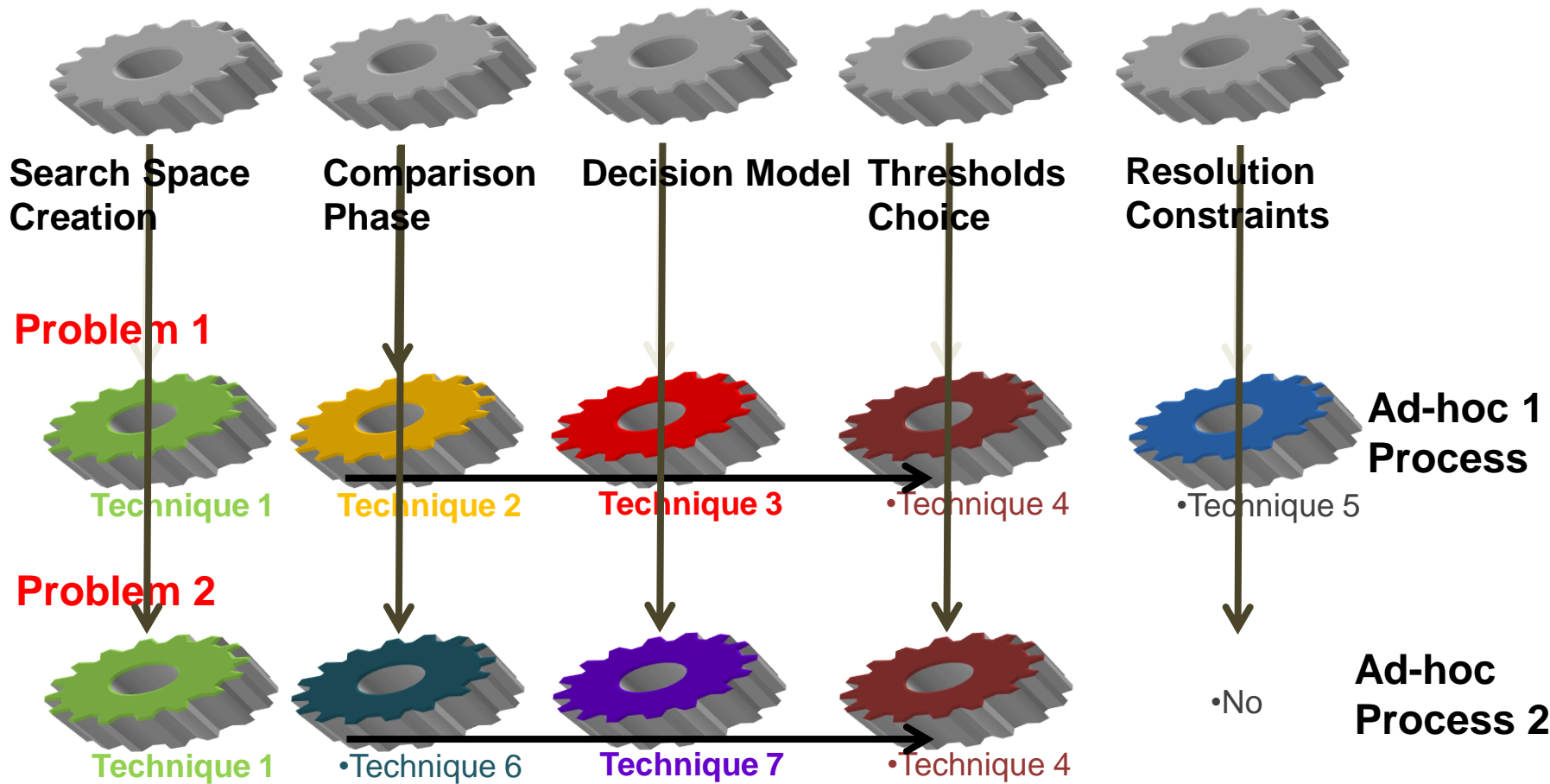
Resolution  
Constraints



# Choose the most appropriate techniques



# Collection of Techniques



# 1. Pre-processing of the input files

In this phase we deal with:

- 'null' values and strings
- abbreviations
- punctuation marks
- UPPER/lower cases
- variations due to pronunciations
- free fields
- parsing
- standardization
- reconciliation of descriptive, semantic and structural conflicts

## 2. Reduce the number of comparisons

No reduction: Cross product of input file

Methods for reduction:

- Blocking
- Blocking Union
- Sorted Neighbourhood
- Blocking + Sorted Neighbourhood (Nested Blocking)
- SimHash
- Blocked SimHash

# 3. Selection of matching variables

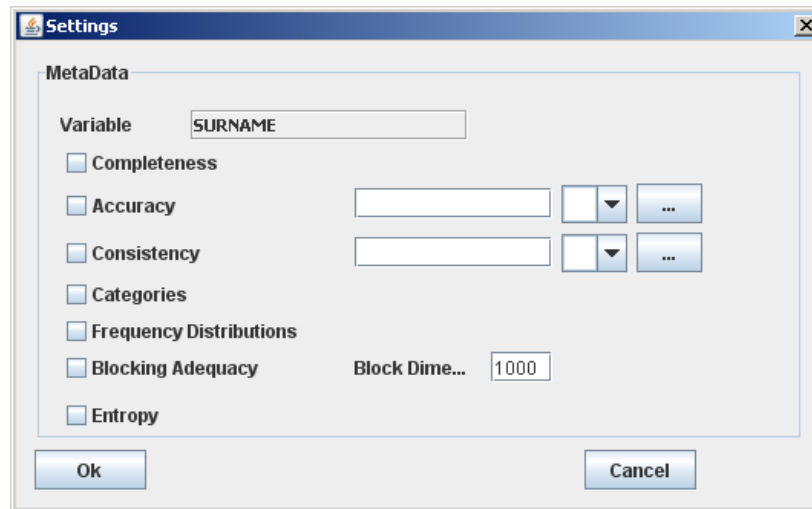
In RELAIS the following indicators are available in order to guide the selection of matching and blocking variables. They are calculated directly on the data:

- ❑ Completeness
- ❑ Accuracy
- ❑ Consistency
- ❑ Number of categories
- ❑ Frequency distribution
- ❑ Entropy

Available in Data profiling Menu

# Data profiling

- Data profiling (i.e. metadata) is evaluated on the data for the proposed variables
- This can help in :
  - Choosing matching variables for the decision model
  - Choosing blocking/sorting variables for the reduction of the search space



## 4. String comparison function

The similarity-closeness of the compared pairs is determined applying a string comparison function on the matching variable

The most widespread functions are:

- equality
- window equality
- “edit” based distance e.i. Levenstein, Jaro, Jaro-Winkler, ...
- “pronunciation” based distance e.i. Soundex, ...
- “token” based distance e.i. Dice, Monge-Elkan, 3Grams, ...
- hybrid distance

# 5. Choice of the decision model

- Exact Linkage
- Deterministic Linkage
- Probabilistic Linkage based on Fellegi-Sunter model



## 5. Choice of the decision model (2)

### Exact Linkage

There is a single rule and the only possible comparison among the matching variables is the equality

- not need the search space creation
- speed even for large data sets
- not 1:1 linkage

## 5. Choice of the decision model (3)

### Deterministic Linkage 'Rule based'

- ❑ The user specifies the rules to classify the results of comparison among variables
- ❑ A rule is composed by a set of simple equivalent sub-rules

**Rule =**

**Sub-rule1 OR Sub-rule2 OR Sub-rule3**

- ❑ An interface guides the choices of rules and matching variables

## 5. Choice of the decision model (4)

Probabilistic approach, according to Fellegi-Sunter

Pros and cons

- It allows to identify matches even if the matching variables are affected by errors
- It allows to evaluate the quality of the linkage results and its effects on the further elaborations on the linked data
- It is more complex to manage by non-expert users

## 5. Choice of the decision model (5)

Probabilistic approach, according to Fellegi-Sunter

The parameter estimation is based on all the considered data (other methods use info from previous studies or analyses on sample from data)

- Loglinear model with latent class
- EM algorithm for the parameter estimation
- Conditional independence hypothesis

At least 3 matching variables are needed

# Output of the decision model

## Contingency table

$\gamma$

SURNAME	NAME	LASTCODE	STREET	FREQUENCY
0	0	0	0	159053
0	0	0	1	6155
0	0	1	0	7788
0	0	1	1	315
0	1	0	0	982
0	1	0	1	64
0	1	1	0	45
0	1	1	1	12
1	0	0	0	642
1	0	0	1	515
1	0	1	0	48
1	0	1	1	97
1	1	0	0	33
1	1	0	1	59
1	1	1	0	46
1	1	1	1	154

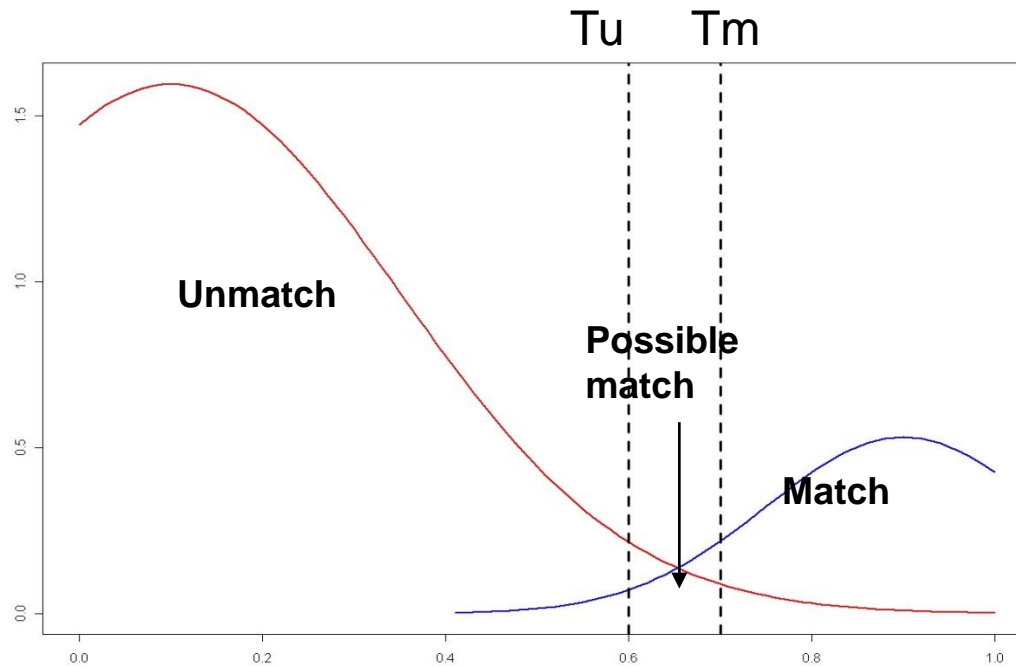
# Output of the decision model

Table with estimated m and u probabilities

$\gamma$

surname	name	lastcode	street	f_m	f_u	m	u	r	p_post
0	0	0	0	131.17182	158921.82...	0.06508	0.91338	0.07125	8.2E-4
0	0	1	0	36.17792	7751.82208	0.01795	0.04455	0.4029	0.00465
0	0	0	1	141.03532	6013.96468	0.06998	0.03456	2.02452	0.02291
0	1	0	0	29.61471	952.38529	0.01469	0.00547	2.68442	0.03016
0	0	1	1	36.87934	278.12066	0.0183	0.0016	11.44736	0.11708
0	1	1	0	6.72896	38.27104	0.00334	2.2E-4	15.17865	0.14953
0	1	0	1	30.02064	33.97936	0.0149	2.0E-4	76.27116	0.46907
0	1	1	1	9.99853	2.00147	0.00496	1.0E-5	431.26398	0.83321
1	0	0	0	641.84389	0.15611	0.31846	0.0	354944.77...	0.99976
1	0	1	0	47.99794	0.00206	0.02381	0.0	2006982.7...	0.99996
1	0	0	1	514.99559	0.00441	0.25552	0.0	1.0084881...	0.99999
1	1	0	0	32.99979	2.1E-4	0.01637	0.0	1.3372068...	0.99999
1	0	1	1	96.99985	1.5E-4	0.04813	0.0	5.7023471...	1.0
1	1	1	0	45.99995	5.0E-5	0.02282	0.0	7.5610382...	1.0
1	1	0	1	58.99999	1.0E-5	0.02927	0.0	3.7993438...	1.0
1	1	1	1	153.99999	1.0E-5	0.07641	0.0	2.1482827...	1.0

# Assignment of the thresholds



# Assignment of the thresholds

Missing Matches

Thresholds

surname	name	lastcode	street	f_m	f_u	m	u	r	p_post
0	0	0	0	131.17182	158921.82...	0.06500	0.91338	0.07125	8.2E-4
0	0	1	0	36.17792	7751.82208	0.01795	0.04455	0.4029	0.00465
0	0	0	1	141.03532	6013.96468	0.06998	0.03456	2.02452	0.02291
0	1	0	0	29.61471	952.38529	0.01469	0.00547	2.68442	0.03016
0	0	1	1	36.87934	278.12066	0.0183	0.0016	11.44736	0.11708
0	1	1	0	6.72896	38.27104	0.00334	2.2E-4	15.17865	0.14953
0	1	0	1	30.02064	33.97936	0.0149	2.0E-4	76.27116	0.46907
0	1	1	1	9.99853	2.00147	0.00496	1.0E-5	431.26398	0.83321
1	0	0	0	641.84389	0.15611	0.31846	0.0	354944.77...	0.99976
1	0	1	0	47.99794	0.00206	0.02381	0.0	2006982.7...	0.99996
1	0	0	1	514.99559	0.00441	0.25552	0.0	1.0084881...	0.99999
1	1	0	0	32.99979	2.1E-4	0.01637	0.0	1.3372068...	0.99999
1	0	1	1	96.99985	1.5E-4	0.04813	0.0	5.7023471...	1.0
1	1	1	0	45.99995	5.0E-5	0.02282	0.0	7.5610382...	1.0
1	1	0	1	58.99999	1.0E-5	0.02927	0.0	3.7993438...	1.0
1	1	1	1	153.99999	1.0E-5	0.07641	0.0	2.1482827...	1.0

False Matches



## 6. 1:1 or multiple linkages

- Types of matching:
- 1-1
- 1-n: file A of events (e.i. admission ), file B of individuals
- n-m: file A and file B of events
- deduplication: special type of n-m matching

To reduce to 1:1 matching

- Simplex algorithm
- Greedy Algorithm

# 7. Linkage evaluation

## Estimation of False Match Rate (FMR) and False Non-Match Rate (FNMR):

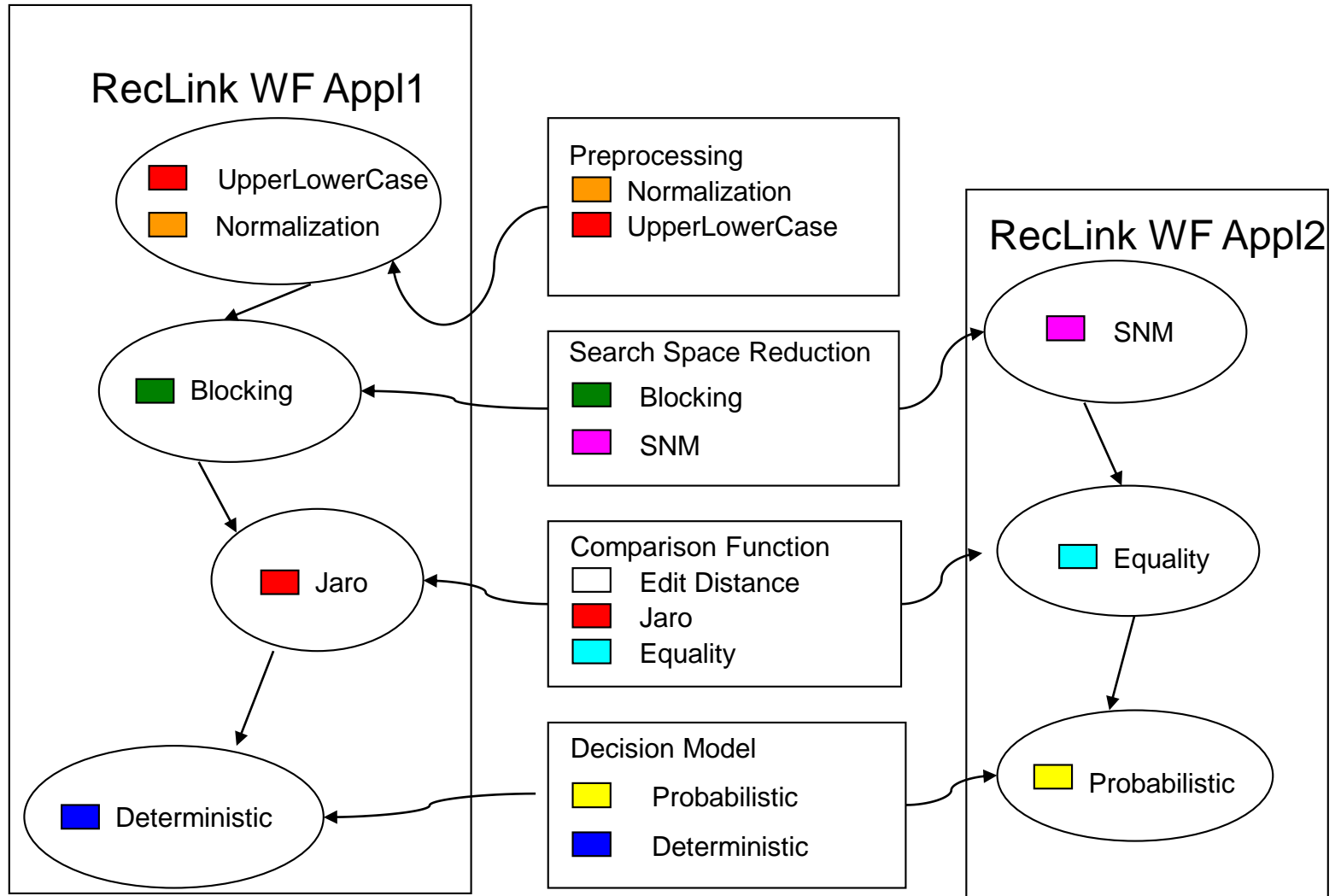
- Manual revision of a sample of pairs  
FMR=false matches/matches  
FNMR=missed matches/matches
- Statistical model based on “training sample”  
(Belin and Rubin, 1991)
- Probabilistic model result (if the model fits very well the data )  
(Fellegi and Sunter, 1969)  
(Torelli and Paggiaro 2001)

# Choice of the best RL strategy

RELAIS allows combining techniques for each of the record linkage phases, so that the resulting workflow is actually built on the basis of the requirements of the application at hand.

RELAIS is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the **best record linkage workflow**

# Build ad-hoc RL workflows



# Technological choices in RELAIS

- Modular structure: each phase is planned as a “module” of the toolkit, with an explicit interface with the other modules
- Top-down design: this allows to omit and/or iterate “modules” (phases) of the record linkage process
- Open Source Project:
  - Java (object-oriented language for the management of data), R (functional language for calculus), MySQL

## Advantages:

- parallel development of various techniques is allowed
- “dynamic” composition of record linkage processes
- design for Web service encapsulation in order to permit remote invocation
- techniques for each phase can be implemented and maintained very rapidly by relying on a community of developers

## Advantages of the relational database

- Simplify the interaction among the several phases with i easy-to-access permanent objects (tables)
- Immediate access to temporary results for expert user
- Temporary results are available also after the stop of the software
- Optimization of several techniques using data indexing and optimization of complex accesses

Easy to use also for non-expert users

- Graphical Users Interface
- Checks in building RL processes
  - The access menu for the different phases are made available only when the preparatory steps are successfully done
- Auxiliary tools
  - Data profiling for the input files
  - Open access to temporary results of previous phases



EUPL: European Union Public Licence

Winning choice of the open-source philosophy and of the overcoming of ad-hoc approaches

Sharing experiences and solutions with NSIs of Spain, Tunisia, Brazil, UK, Leetonia, Bosnia and Herzegovina

Thanks to the modular approach and the OS, adding new techniques to the pool already available is really easy

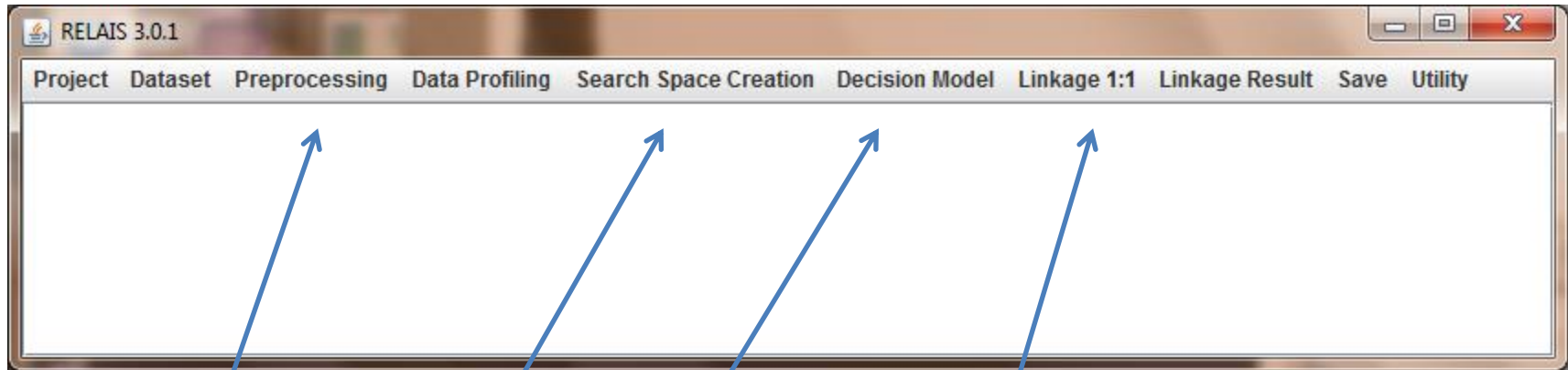
# Some future developments 1

- Modification of the probabilistic approach:
  - Not only binary comparison vector (in progress)
  - Allowing interactions between matching variables
  - Bayesian approach
- Alternative decisional models
  - Based on soft-computing techniques (e.g. genetic algorithm)
  - Other machine learning techniques

# Some future developments 2

- New algorithm for optimal 1:1 reduction (done)
- Improvement of GUI functionalities for output management and user interactions (manual review)
- Interfaces for clerical review
- Evaluation of the error rates with alternative methods

# Menu



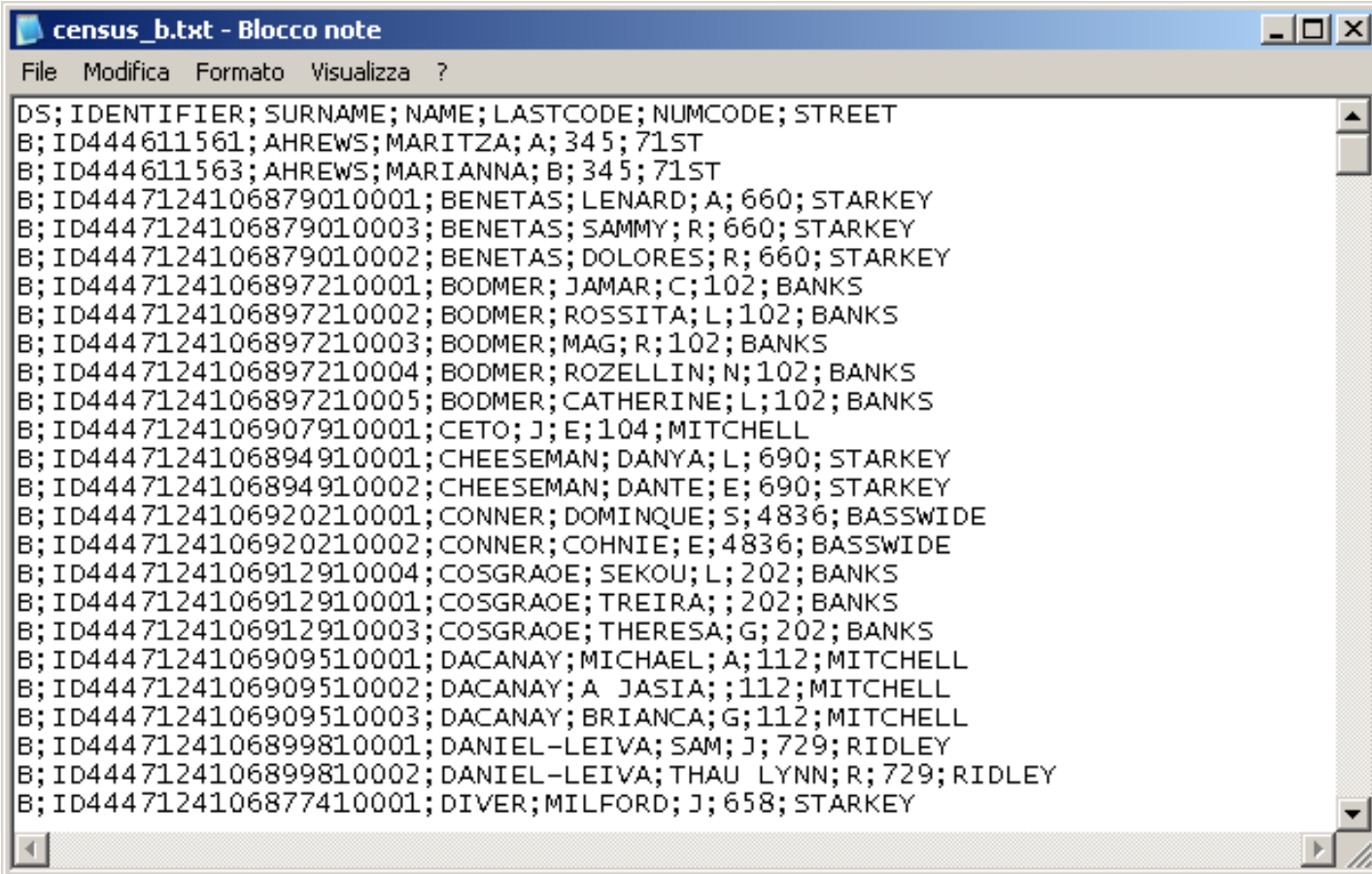
**1:1 constraints**

**Choice of: matching variables; comparison functions; decisional model**

**Search Space Creation**

**Preprocessing (optional)**

# Example of an Input File



```
census_b.txt - Blocco note
File Modifica Formato Visualizza ?
DS; IDENTIFIER; SURNAME; NAME; LASTCODE; NUMCODE; STREET
B; ID444611561; AHREWS; MARITZA; A; 345; 71ST
B; ID444611563; AHREWS; MARIANNA; B; 345; 71ST
B; ID4447124106879010001; BENETAS; LENARD; A; 660; STARKEY
B; ID4447124106879010003; BENETAS; SAMMY; R; 660; STARKEY
B; ID4447124106879010002; BENETAS; DOLORES; R; 660; STARKEY
B; ID4447124106897210001; BODMER; JAMAR; C; 102; BANKS
B; ID4447124106897210002; BODMER; ROSSITA; L; 102; BANKS
B; ID4447124106897210003; BODMER; MAG; R; 102; BANKS
B; ID4447124106897210004; BODMER; ROZELLIN; N; 102; BANKS
B; ID4447124106897210005; BODMER; CATHERINE; L; 102; BANKS
B; ID4447124106907910001; CETO; J; E; 104; MITCHELL
B; ID4447124106894910001; CHEESEMAN; DANYA; L; 690; STARKEY
B; ID4447124106894910002; CHEESEMAN; DANTE; E; 690; STARKEY
B; ID4447124106920210001; CONNER; DOMINQUE; S; 4836; BASSWIDE
B; ID4447124106920210002; CONNER; COHNIE; E; 4836; BASSWIDE
B; ID4447124106912910004; COSGRAOE; SEKOU; L; 202; BANKS
B; ID4447124106912910001; COSGRAOE; TREIRA; ; 202; BANKS
B; ID4447124106912910003; COSGRAOE; THERESA; G; 202; BANKS
B; ID4447124106909510001; DACANAY; MICHAEL; A; 112; MITCHELL
B; ID4447124106909510002; DACANAY; A JASIA; ; 112; MITCHELL
B; ID4447124106909510003; DACANAY; BRIANCA; G; 112; MITCHELL
B; ID4447124106899810001; DANIEL-LEIVA; SAM; J; 729; RIDLEY
B; ID4447124106899810002; DANIEL-LEIVA; THAU LYNN; R; 729; RIDLEY
B; ID4447124106877410001; DIVER; MILFORD; J; 658; STARKEY
```

# Example of Matches

```
Match.txt - Blocco note
File Modifica Formato Visualizza ?
DS; KEY_DS; IDENTIFIER; SURNAME; NAME; LASTCODE; NUMCODE; STREET; R; P_POST
A; 367; ID4445112314671410004; ROWEN; EARNEST; A; 7112; 2ND ; 5.83715653798434e+018;1
B; 307; ID4445112314671410004; ROWAN; ERNEST; A; 7112; 2ND ; 5.83715653798434e+018;1
A; 449; ID4449126128945310002; YETES; ALFREDRICK; S; 403; WOODHAVEN ; 5.83715653798434e+018;1
B; 392; ID4449126128945310002; YATES; ALFREDICA; S; 403; WOODHAVEN ; 5.83715653798434e+018;1
A; 72; ID4447124106890710004; MOSQUERA; ANDREA; G; 682; STARKEY ; 5.83715653798434e+018;1
B; 65; ID4447124106890710004; MOSGUERA; ANDRE; G; 682; STARKEY ; 5.83715653798434e+018;1
A; 292; ID4450127238361810001; RUDASILL; KIMBERLY; L; 122; WARE; 5.83715653798434e+018;1
B; 238; ID4450127238361810001; RUDAGILL; KIMBERY; L; 122; WARE; 5.83715653798434e+018;1
A; 386; ID4449126128936210003; BELTRAN; RUTH ANN; M; 114; BRANTWOOD ; 5.83715653798434e+018;1
B; 322; ID4449126128936210003; BELTAN; RUTHANN; M; 114; BRANTWOOD ; 5.83715653798434e+018;1
A; 34; ID4447124106883210002; DOMINGUEZ; SHAVON; M; 672; STARKEY ; 5.83715653798434e+018;1
B; 26; ID4447124106883210002; DOMINQUEZ; SHAVAN; M; 672; STARKEY ; 5.83715653798434e+018;1
A; 315; ID4450127238356810003; TONSTALL; HARRIETT; J; 201; MAIN; 5.83715653798434e+018;1
B; 257; ID4450127238356810003; TUNSTALL; HARRIET; J; 201; MAIN; 5.83715653798434e+018;1
A; 370; ID4445113314678910001; SAUNDERS; NICOLAS; G; 430; 72ND; 5.83715653798434e+018;1
B; 310; ID4445113314678910001; SANDERS; NICHOLAS; G; 430; 72ND; 5.83715653798434e+018;1
A; 348; ID4445112314672210003; HOPPER; CARLOS; C; 7100; 2ND ; 5.83715653798434e+018;1
B; 287; ID4445112314672210003; HOOPER; CARLO; C; 7100; 2ND ; 5.83715653798434e+018;1
```

## Statisticians:

Nicoletta Cibella

E-mail: [cibella@istat.it](mailto:cibella@istat.it)

Marco Fortini

E-mail: [fortini@istat.it](mailto:fortini@istat.it)

Tiziana Tuoto

E-mail: [tuoto@istat.it](mailto:tuoto@istat.it)

## Computer Scientists:

Monica Scannapieco

E-mail: [scannapi@istat.it](mailto:scannapi@istat.it)

Laura Tosco

E-mail: [tosco@istat.it](mailto:tosco@istat.it)

Luca Valentino

E-mail: [luvalent@istat.it](mailto:luvalent@istat.it)

<http://www.istat.it/it/strumenti/metodi-e-software/software/relais>