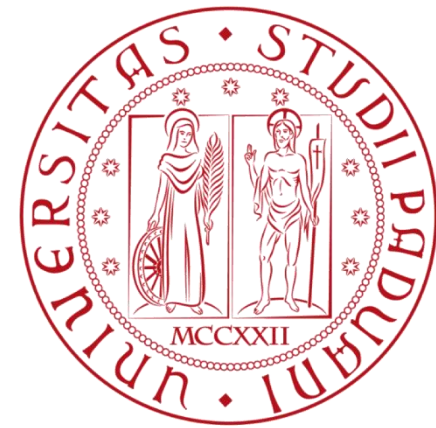




Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Record linkage

Deterministico vs. probabilistico

Record Linkage

- Procedura di integrazione di dati provenienti da fonti diverse. E' una tecnica algoritmica con lo scopo di **identificare le coppie di record appartenenti a due base dati che corrispondono alla stessa unità statistica** (persone, eventi..ecc..)

Identificazione delle coppie:	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

Un esempio di identificazione delle coppie

	firstname	lastname	dob	ssn
1	Jane	Johnson	05/05/1985	1000000005
2	Amy	Miller	11/25/1985	1000000006
3	Mary	Smith	02/08/1985	1000000001
4	Amy	Miller	08/05/2000	1000000007
5	Elizabeth	Jones	05/05/1985	1000000003
6	Catherine	Johnson	05/05/1985	1000000002
7	Maria	Sanchez	01/01/1983	
8	Jane	Doe	01/05/1985	1000000000

File A

File B

	Firstname	lastname	dob	ssn
1	Jane	Doe	01/06/1985	1000000000
2	Mary	Smoth	02/07/1985	
3	Katie	Jonson	05/05/1985	1000000002
4		Jones	05/05/1985	1000000003
5	Maria	Sanchez-Martinez	01/01/1983	1000000004
6	Jane	Johnson	05/05/1985	1000000005
7	Anne	Miller	05/01/1980	2000000007

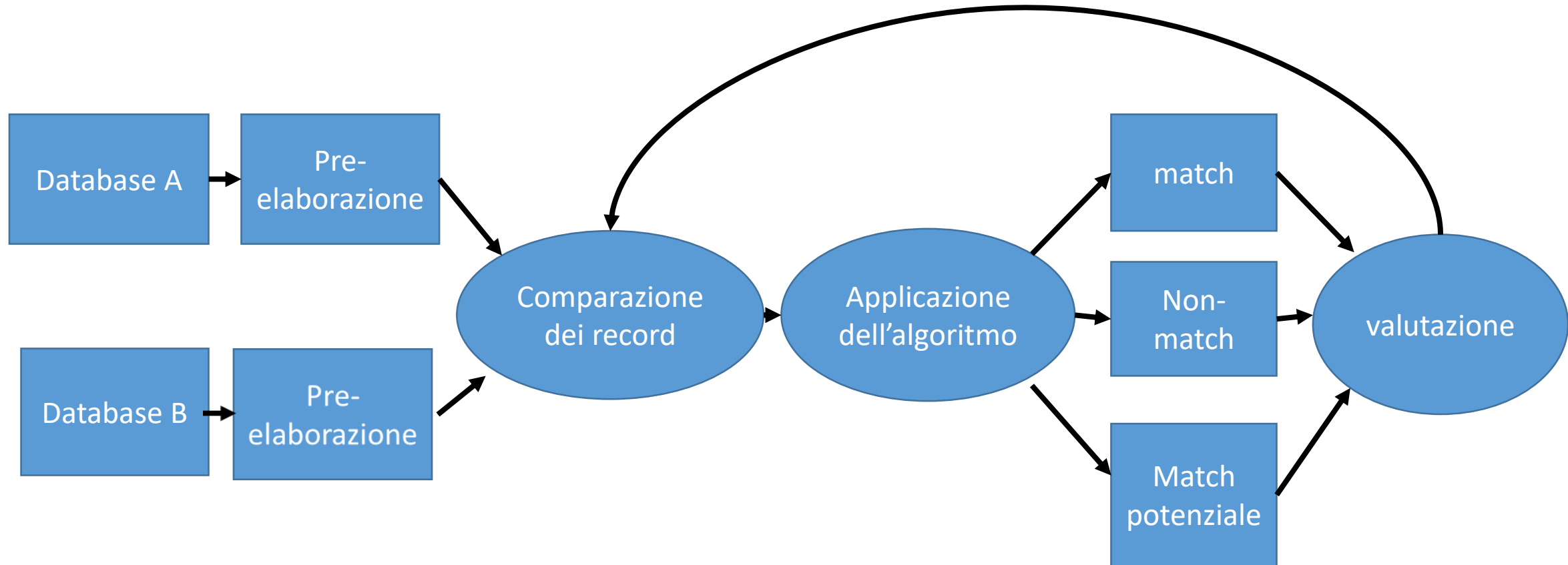
Un esempio: dopo il linkage (long format)

	_matchID	fileid	firstname	lastname	dob	Ssn
Matched	1	A	Jane	Doe	01/05/1985	1000000000
	1	B	Jane	Doe	01/06/1985	1000000000
	2	A	Mary	Smith	02/07/1985	1000000001
	2	B	Mary	Smoth	02/08/1985	
	3	A	Catherine	Johnson	05/05/1985	1000000002
	3	B	Katie	Jonson	05/05/1985	1000000002
	4	A	Elizabeth	Jones	05/05/1985	1000000003
	4	B		Jones	05/05/1985	1000000003
	5	A	Maria	Sanchez	01/01/1983	
	5	B	Maria	Sanchez-Martinez	01/01/1983	1000000004
Not matched	6	A	Jane	Johnson	05/05/1985	1000000005
	6	B	Jane	Johnson	05/05/1985	1000000005
		A	Amy	Miller	08/05/2000	1000000007
		A	Amy	Miller	11/25/1985	1000000006
		B	Anne	Miller	05/01/1980	2000000007

A cosa serve?

- Costruire liste complete su cui basare un campionamento.
 - Imprese presenti in più base dati incomplete: registro delle imposte (Finanza), registro delle imprese locali (Camera di Commercio), registro INPS.
- Avere più informazioni in un unico file.
 - Collegare informazioni su forze lavoro con quelle sulle cause di morte (ad es.)
 - Collegare le schede ospedaliere (per ogni evento) a residenza e occupazione.
 - Cercare un sospettato di un crimine all'interno dei registri della polizia
 - Creare profili di consumatori (informazioni dello shopping online)

Le fasi del record linkage



Applicazione dell'algoritmo

In fase di applicazione abbiamo 2 distinti metodi:

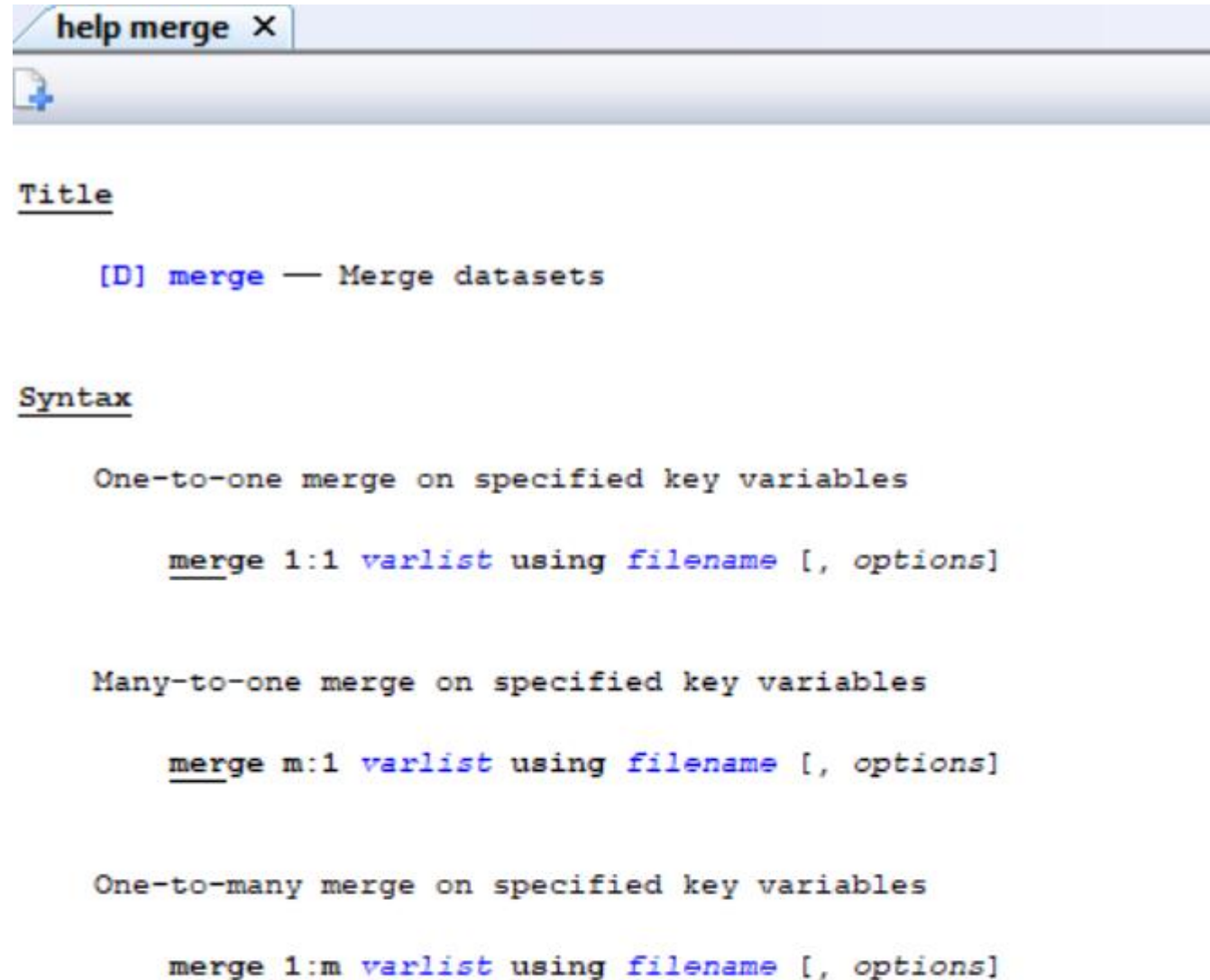
- *Linkage Deterministico*: è un matching esatto che coinvolge una coppia di osservazioni che può essere identificata grazie all'utilizzo di una o più variabili.
 - *Es*: nome e data di nascita identificano esattamente gli individui in entrambi i file.
- *Linkage Probabilistico*: calcolo delle probabilità stimate da tutte le corrispondenze e non-corrispondenze dei valori assunti dalle variabili chiave inserite nel modello.
 - *Es*: la probabilità di abbinare 2 osservazioni appartenenti ai file A e B date le variabili nome e data di nascita è uguale a p

Linkage deterministico

- Si può applicare quando abbiamo una variabile chiave identificativa dei soggetti (ID dell'individuo o codice fiscale) oppure quando la combinazione di più variabili identifica esattamente le unità nei diversi databases.
- Le variabili identificative assumono un peso identico nel determinare le coppie di unità da abbinare.
- Si applica quando abbiamo dati molto accurati, di alta qualità, puliti.
 - E' problematico quando abbiamo errori di trascrizione e registrazione, valori mancanti, oppure non abbiamo un identificativo delle unità accurato

Record Linkage deterministico

- In STATA: merge
- Help merge ----->



The screenshot shows a STATA help window titled 'help merge'. The window has a title bar with a close button (X) and a standard icon. Below the title bar, the word 'Title' is underlined. The main text reads '[D] merge — Merge datasets'. Below this, the word 'Syntax' is underlined. There are three sections of text describing different merge types: 'One-to-one merge on specified key variables' with the command 'merge 1:1 varlist using filename [, options]', 'Many-to-one merge on specified key variables' with the command 'merge m:1 varlist using filename [, options]', and 'One-to-many merge on specified key variables' with the command 'merge 1:m varlist using filename [, options]'. The command 'merge' is underlined in each example.

```
help merge X

Title

[D] merge — Merge datasets

Syntax

One-to-one merge on specified key variables

merge 1:1 varlist using filename [, options]

Many-to-one merge on specified key variables

merge m:1 varlist using filename [, options]

One-to-many merge on specified key variables

merge 1:m varlist using filename [, options]
```

HELP – spiegazione della sintassi

- Merge joins corresponding observations from the dataset currently in memory (called **the master dataset**) with those from filename.dta (called the **using dataset**), matching on one or more key variables. merge can perform match merges (***one-to-one, one-to-many, many-to-one, and many-to-many***), which are often called 'joins' by database people.
- Merge is for **adding new variables** from a second dataset to existing observations. You use merge, for instance, when combining hospital patient and discharge datasets. If you wish to add **new observations to existing variables, then see [D] append**. You use append, for instance, when adding current discharges to past discharges.
- By default, merge creates a **new variable, _merge**, containing numeric codes concerning the source and the contents of each observation in the merged dataset. These codes are explained below in the match results table.

Tipi di Linkage deterministico

- Codice identificativo univoco per individui, famiglie, e paesi
 - Merge one-to-one (es. individuo – individuo)
 - Merge one-to-many (es. famiglia – individuo)
 - Merge many-to-one (es. individuo – famiglia)
- Le variabili che hanno lo stesso nome in entrambi i dataset vengono mantenute uguali a quelle originali
(come se non stessimo abbinando i casi)

Merge one-to-one (1)

- L'esempio tipico è quando vogliamo abbinare casi (identificati dalla variabile ID) in quanto i due dataset contengono informazioni diverse. Tuttavia il matching è parziale perché i casi 5 e 4 sono contenuti in un solo file

```
. merge 1:1 id using filename
```

in memory
master

+

in filename.dta
using

=

merged result

id	age
1	22
2	56
5	17

id	wgt
1	130
2	180
4	110

id	age	wgt
1	22	130
2	56	180
5	17	.
4	.	110

(matched)

(matched)

(master only)

(using only)

Merge one-to-one (2)

- In alcuni dataset l'identificativo dell'individuo è ripetuto in quanto viene intervistato più volte nel tempo (time). Dovremo identificare ogni caso attraverso le variabili ID e time.

```
. merge 1:1 pid time using filename
```

master + *using* = *merged result*

pid	time	x1
14	1	0
14	2	0
14	4	0
16	1	1
16	2	1
17	1	0

pid	time	x2
14	1	7
14	2	9
16	1	2
16	2	3
17	1	5
17	2	2

pid	time	x1	x2	_merge
14	1	0	7	3
14	2	0	9	3
14	4	0	.	1
16	1	1	2	3
16	2	1	3	3
17	1	0	5	3
17	2	.	2	2

Merge one-to-many

- Nel secondo database abbiamo informazioni ad un livello diverso (regionale) che si riferiscono a tutti coloro che vivono in quelle regioni. Es. tasso di disoccupazione regionale può predire la difficoltà con cui un individuo trova il suo primo impiego.

```
. merge 1:m region using filename
```

master

+

using

=

merged result

region	x
1	15
2	13
3	12
4	11

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

region	x	id	a	_merge
1	15	2	29	3
1	15	5	24	3
2	13	1	26	3
2	13	3	22	3
3	12	4	21	3
4	11	.	.	1
5	.	6	20	2

Merge many-to-one

- Viceversa...

```
. merge m:1 region using filename  
      master      +      using      =
```

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

region	x
1	15
2	13
3	12
4	11

merged result

id	region	a	x	_merge
1	2	26	13	3
2	1	29	15	3
3	2	22	13	3
4	3	21	12	3
5	1	24	15	3
6	5	20	.	1
.	4	.	11	2

Risultati – massimizzare i match veri

numeric code	equivalent word (results)	description
1	<u>master</u>	observation appeared in master only
2	<u>using</u>	observation appeared in using only
3	<u>match</u>	observation appeared in both
4	<u>match_update</u>	observation appeared in both, missing values updated
5	<u>match_conflict</u>	observation appeared in both, conflicting nonmissing values

Codes 4 and 5 can arise only if the update option is specified.
If codes of both 4 and 5 could pertain to an observation, then 5 is used.

Append (aggiungere casi)

- Un caso di record linkage in cui il match è uguale a 0

```
. use even  
(6th through 8th even numbers)  
. list
```

	number	even
1.	6	12
2.	7	14
3.	8	16

```
. use odd  
(First five odd numbers)  
. list
```

	number	odd
1.	1	1
2.	2	3
3.	3	5
4.	4	7
5.	5	9

```
. append using even  
. list
```

	number	odd	even
1.	1	1	.
2.	2	3	.
3.	3	5	.
4.	4	7	.
5.	5	9	.
6.	6	.	12
7.	7	.	14
8.	8	.	16