USER STUDY: Complete Record Linkage (RL) Tasks

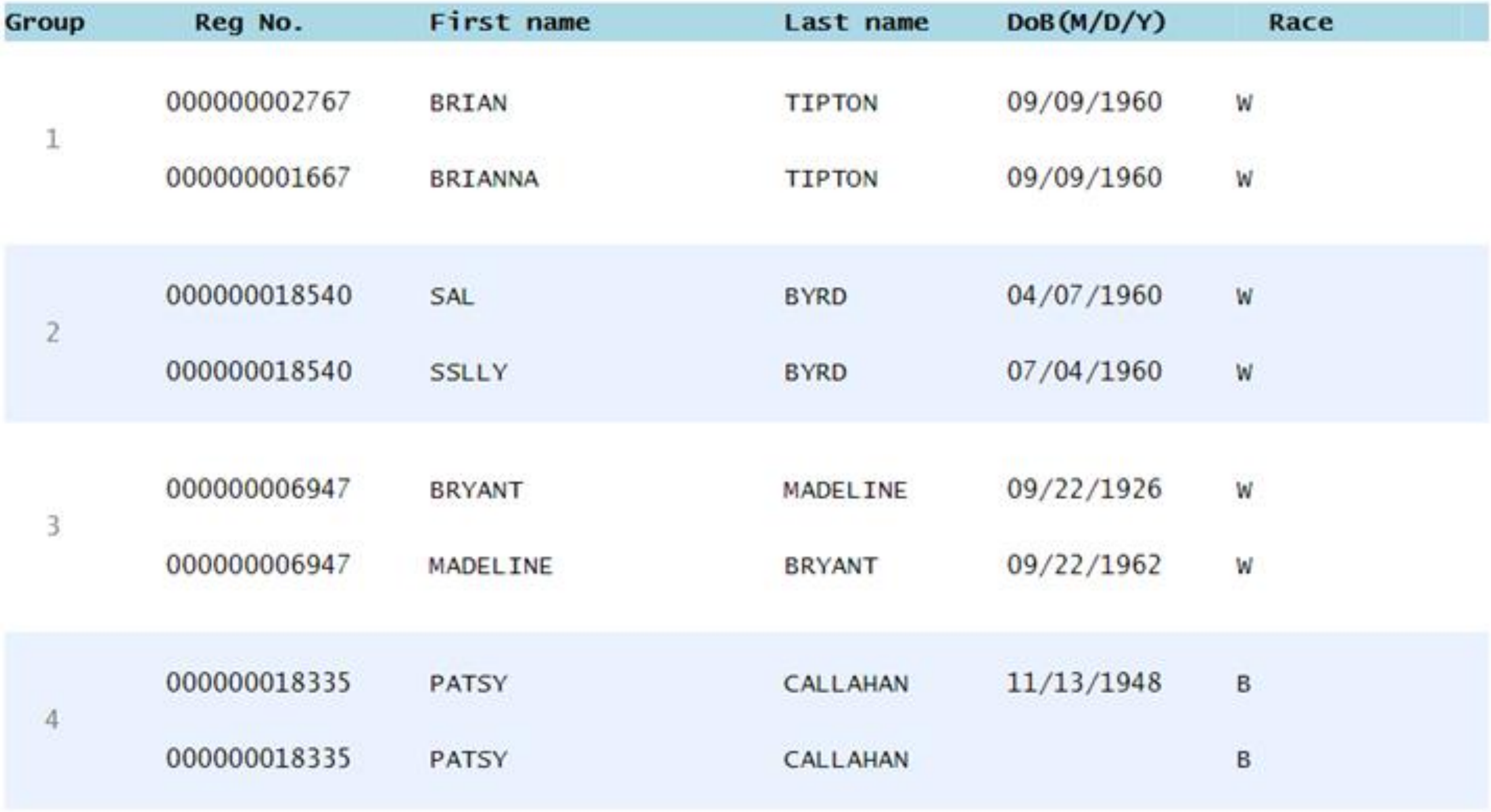
The task is to identify data records that refer to the same real world person

Example: Link two hospital databases to find patients that have visited both hospitals

On the right,

You see pairs of records that are potentially the same person

Your job during the study will be to look at pairs of such identifying information, and determine the likelihood that the pair refers to the same real world person



Group	Reg No.	First name	Last name	DoB(M/D/Y)	Race
1	000000002767	BRIAN	TIPTON	09/09/1960	W
	000000001667	BRIANNA	TIPTON	09/09/1960	W
2	000000018540	SAL	BYRD	04/07/1960	W
	000000018540	SSLLY	BYRD	07/04/1960	W
3	000000006947	BRYANT	MADELINE	09/22/1926	W
	000000006947	MADELINE	BRYANT	09/22/1962	W
4	000000018335	PATSY	CALLAHAN	11/13/1948	B
	000000018335	PATSY	CALLAHAN		B

Inherent Nature of Real Data

There are inherent problems in real data that make RL difficult

- Data are expressed differently
  - nick names (Elizabeth & Beth)
- Data change over time
  - Women get married and change their last name
- Data are not unique attributes
  - John Smith (there are different people that have the same name)
  - Twins & Family members have similar identifying information such as DOB & last name
- Data are sometimes missing
  - ssn are often missing
- Data have errors
  - Inserting/deleting extra characters
  - Typing in the wrong character
  - Transposing two characters
  - First name and last name are mixed up
  - Day and month is mixed up

Common ISSUES in Data

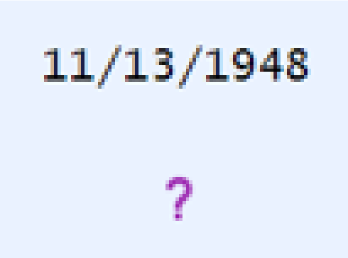
- When given a record linkage task,
  you need to learn and watch out for common issues in data

- The following slides discuss the most common issues in data.
  And how these issues are indicated in the user study

Missing

- Data are sometimes missing

?

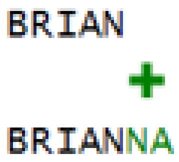


11/13/1948

?

Insertions & Deletions (Indel)

- Insertion (or deletion) of characters are common typing errors



BRIAN

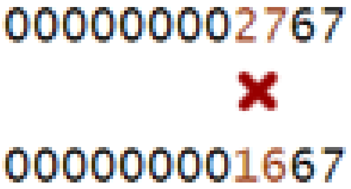
+

BRIANNA

Replace

- Mistyping can lead to certain characters replacing others

✗



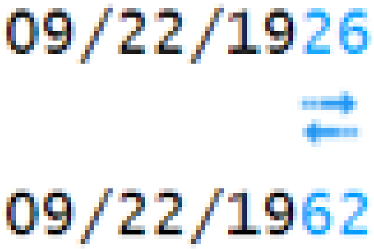
000000002767

✗

000000001667

Transpose

- Two characters can be interchanged by mistake



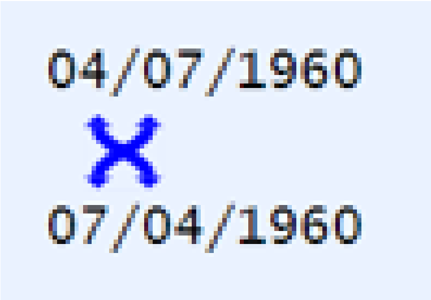
09/22/1926

⇄

09/22/1962

Swaps

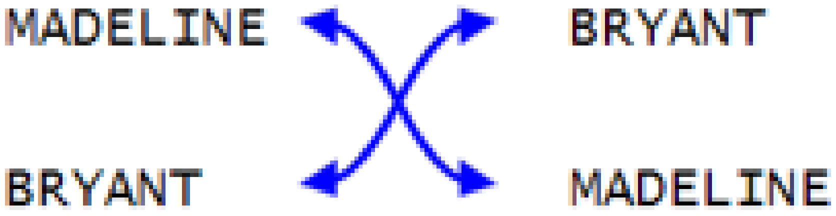
- Due to mix up, sometimes whole values are swapped as well
- Date Swaps:



04/07/1960

07/04/1960

- Name swaps:
-



MADELINE

BRYANT

BRYANT

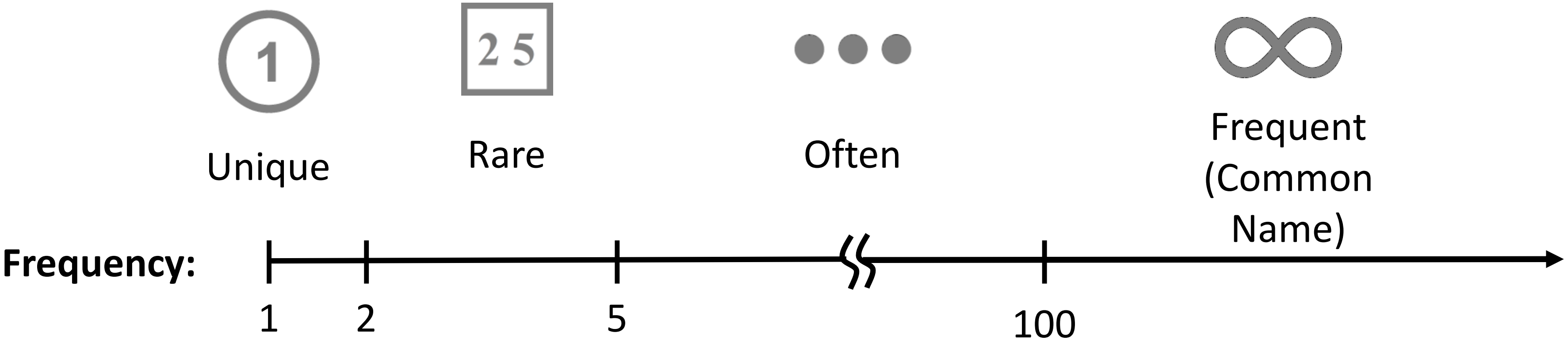
MADELINE

Different

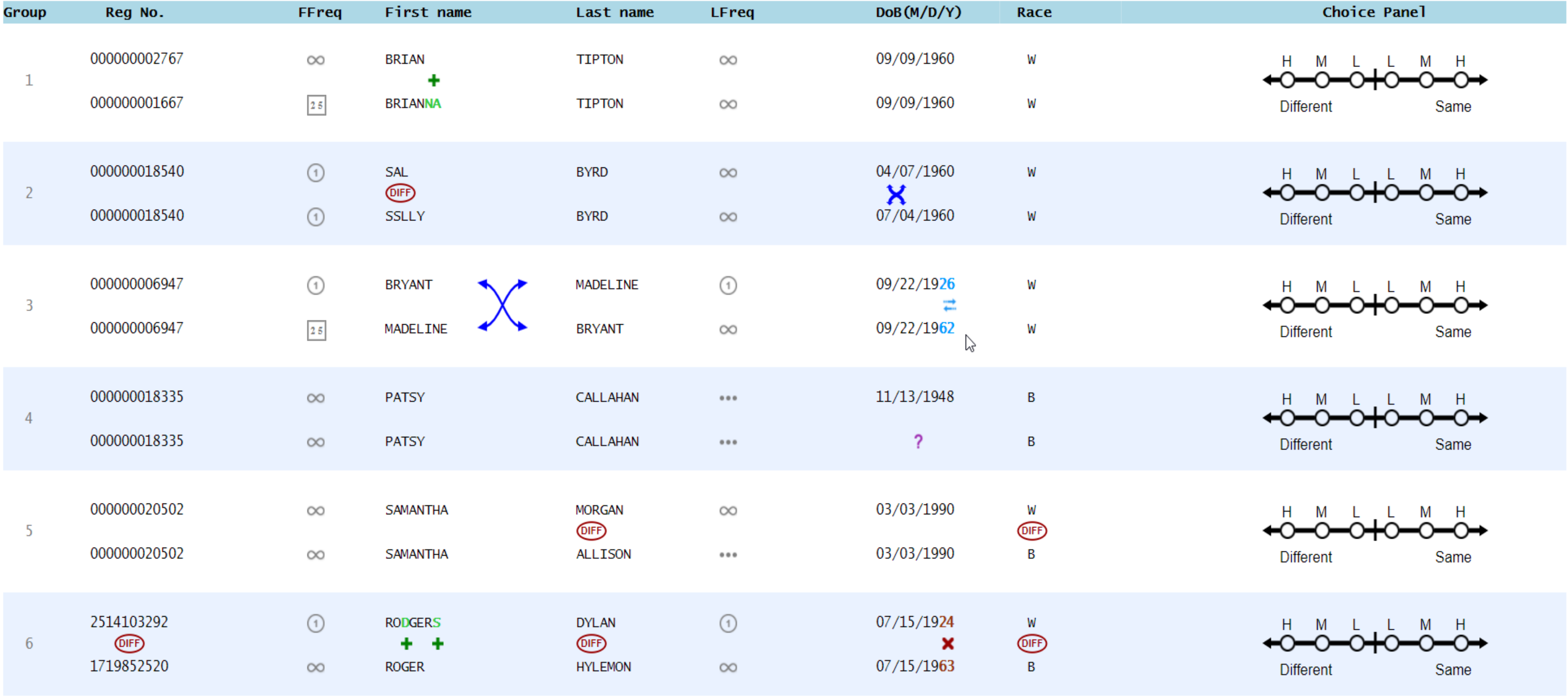
- Sometimes, data are completely different

Name Frequencies

- Another important information to consider in record linkage tasks is how common or unique the given name is.
- Intuitively, two common names often refer to different people while two rare names often refer to the same person
- Frequency Icons indicate how many times a given name occurred in the source data to help you make better record linkage decisions.



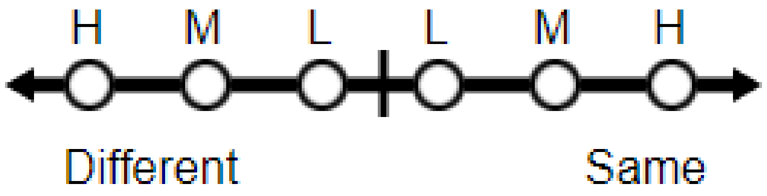
What you will See during the user study



Group	Reg No.	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Race	Choice Panel
1	000000002767 000000001667	∞ 25	BRIAN + BRIANNA	TIPTON TIPTON	∞ ∞	09/09/1960 09/09/1960	W W	H M L L M H ○ ○ ○ ○ ○ ○ Different Same
2	000000018540 000000018540	① ①	SAL DIFF SSLLY	BYRD BYRD	∞ ∞	04/07/1960 ✕ 07/04/1960	W W	H M L L M H ○ ○ ○ ○ ○ ○ Different Same
3	000000006947 000000006947	① 25	BRYANT ⤨ MADELINE	MADELINE BRYANT	① ∞	09/22/1926 ⇄ 09/22/1962	W W	H M L L M H ○ ○ ○ ○ ○ ○ Different Same
4	000000018335 000000018335	∞ ∞	PATSY PATSY	CALLAHAN CALLAHAN	••• •••	11/13/1948 ?	B B	H M L L M H ○ ○ ○ ○ ○ ○ Different Same
5	000000020502 000000020502	∞ ∞	SAMANTHA SAMANTHA	MORGAN DIFF ALLISON	∞ •••	03/03/1990 03/03/1990	W DIFF B	H M L L M H ○ ○ ○ ○ ○ ○ Different Same
6	2514103292 DIFF 1719852520	① ∞	RODGERS + + ROGER	DYLAN DIFF HYLEMON	① ∞	07/15/1924 ✕ 07/15/1963	W DIFF B	H M L L M H ○ ○ ○ ○ ○ ○ Different Same

The answer submission panel

- At the far right is the answer panel.



- You should answer if you think
  - Refers to the same person (pick one of L, M, H on the right depending on your confidence level)
  - OR refers to two different people (pick one of L, M, H on the left depending on your confidence level)

Ready to Give it a try?
Let’s do some Practice Problems

- Now let’s try to learn how to apply these concepts to make good record linkage decisions through some practice problems