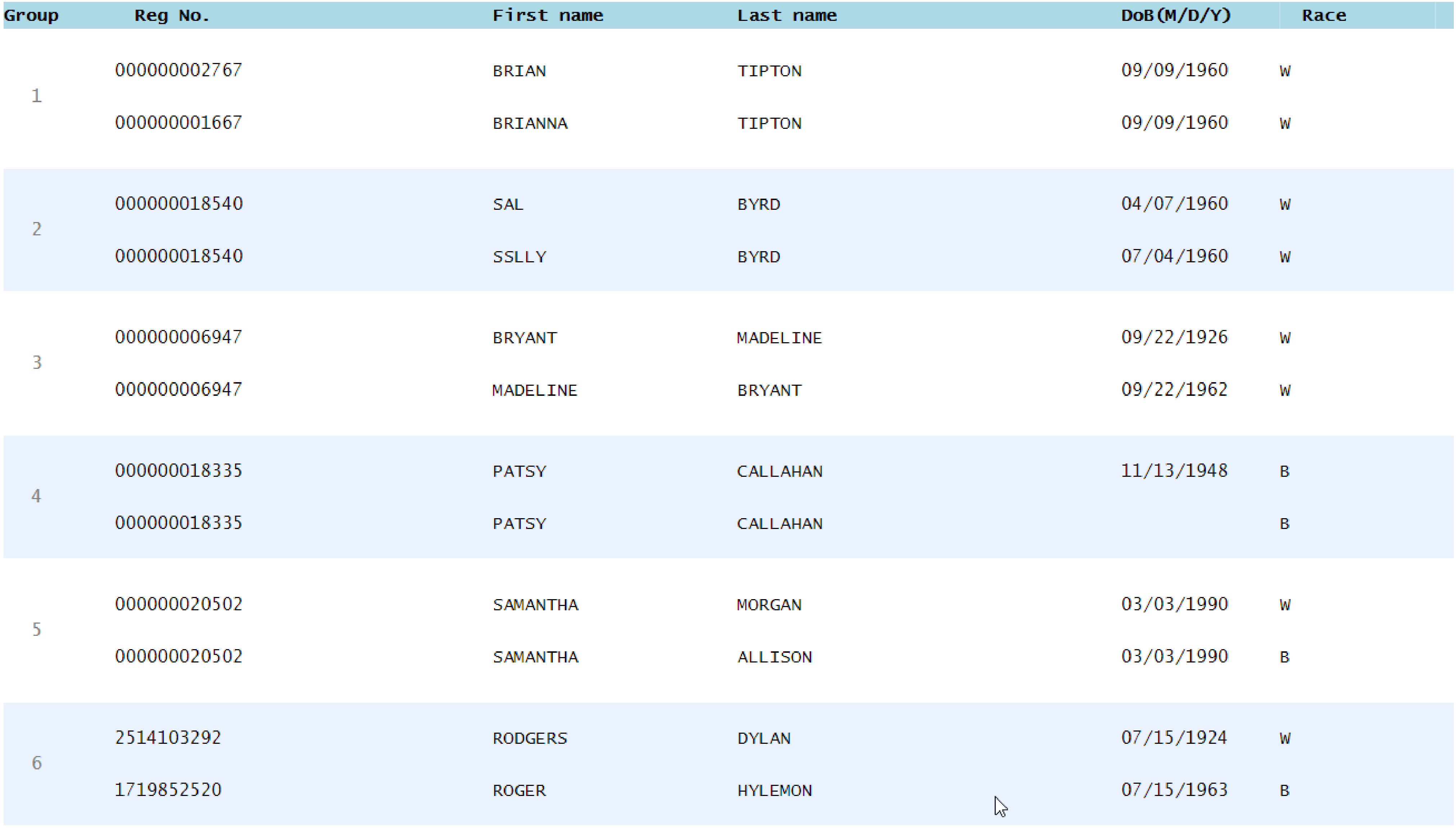
Record Linkage

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.



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
5	000000020502	SAMANTHA	MORGAN	03/03/1990	W
	000000020502	SAMANTHA	ALLISON	03/03/1990	B
6	2514103292	RODGERS	DYLAN	07/15/1924	W
	1719852520	ROGER	HYLEMON	07/15/1963	B

Inherent Nature of Real Data

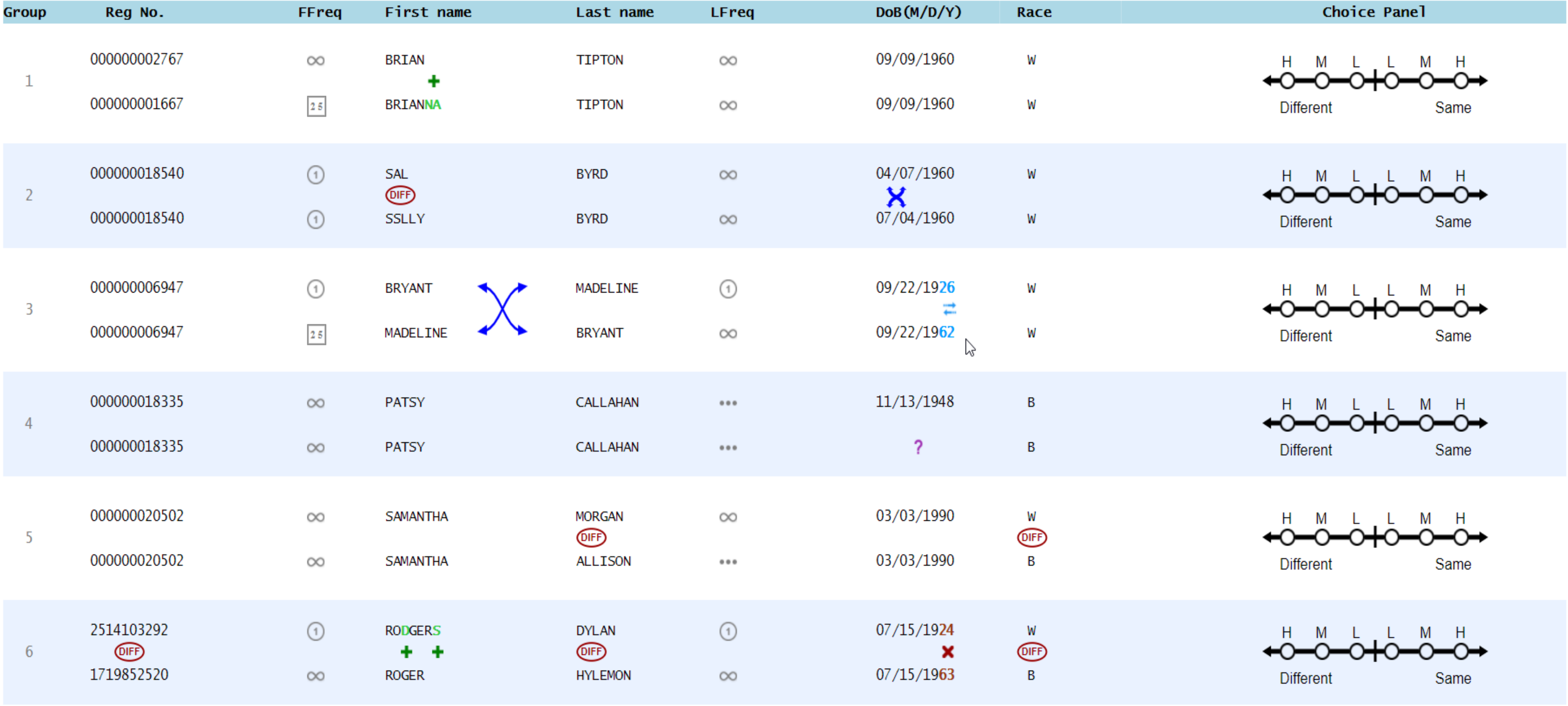
There an inherent problems in real data that make RL difficult
(maybe consider building one page per bullet. TBD after deciding on Monday)

- 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
- Missing Data
    - ssn are often missing
- Errors in Data
    - 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

Intervention Icons

- When data is in the raw form as shown previously, it is very hard to spot differences and to what extent the records are different.
- So there are icons of many kinds to help direct your attention towards what is actually different between the two records.

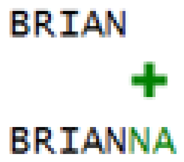
Icons to help you spot the differences



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

Indel

- Describes an insertion (or deletion) of characters.



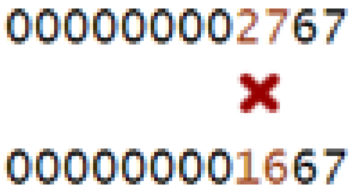
BRIAN

+

BRIANNA

Replace

- When characters are used in the place of another.



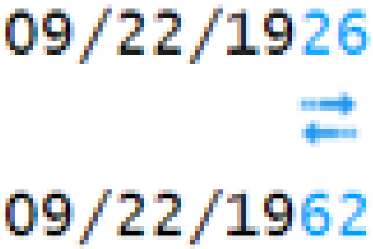
000000002767

✗

000000001667

Transpose

- When the 2 characters are interchanged



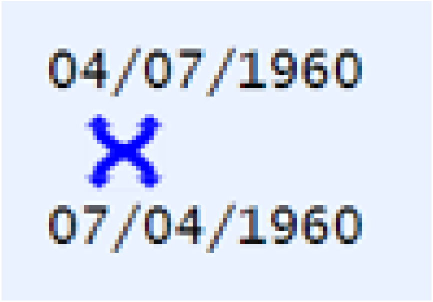
09/22/1926

⇄

09/22/1962

Swaps

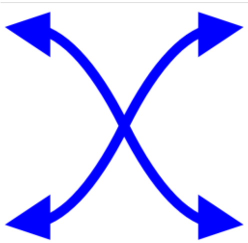
- Sometimes whole values are swapped.
- Date Swaps:

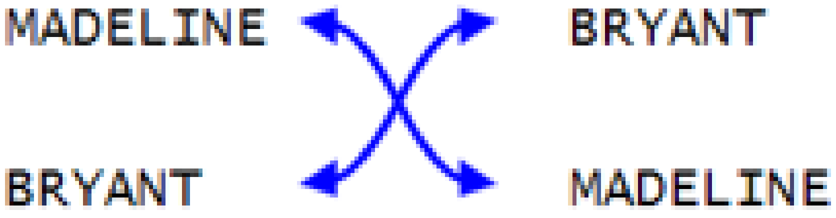


04/07/1960

07/04/1960

- Name swaps:



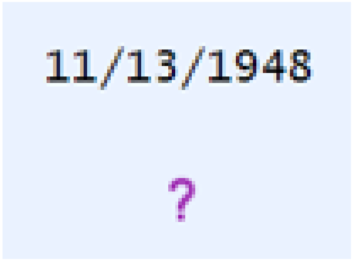


MADELINE BRYANT

BRYANT MADELINE

Missing

- When data is missing



11/13/1948

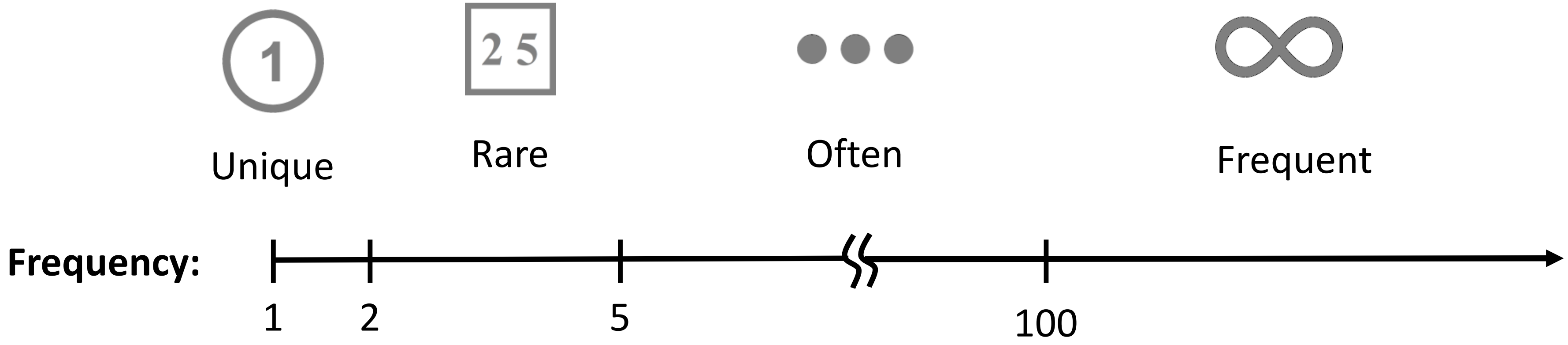
?

Different

- Sometimes, things are completely different.

Frequency Icons

- Frequency Icons indicate how many times a given name occurred in the source data. This information can also be used link records.



Unique

Rare

Often

Frequent

Frequency:

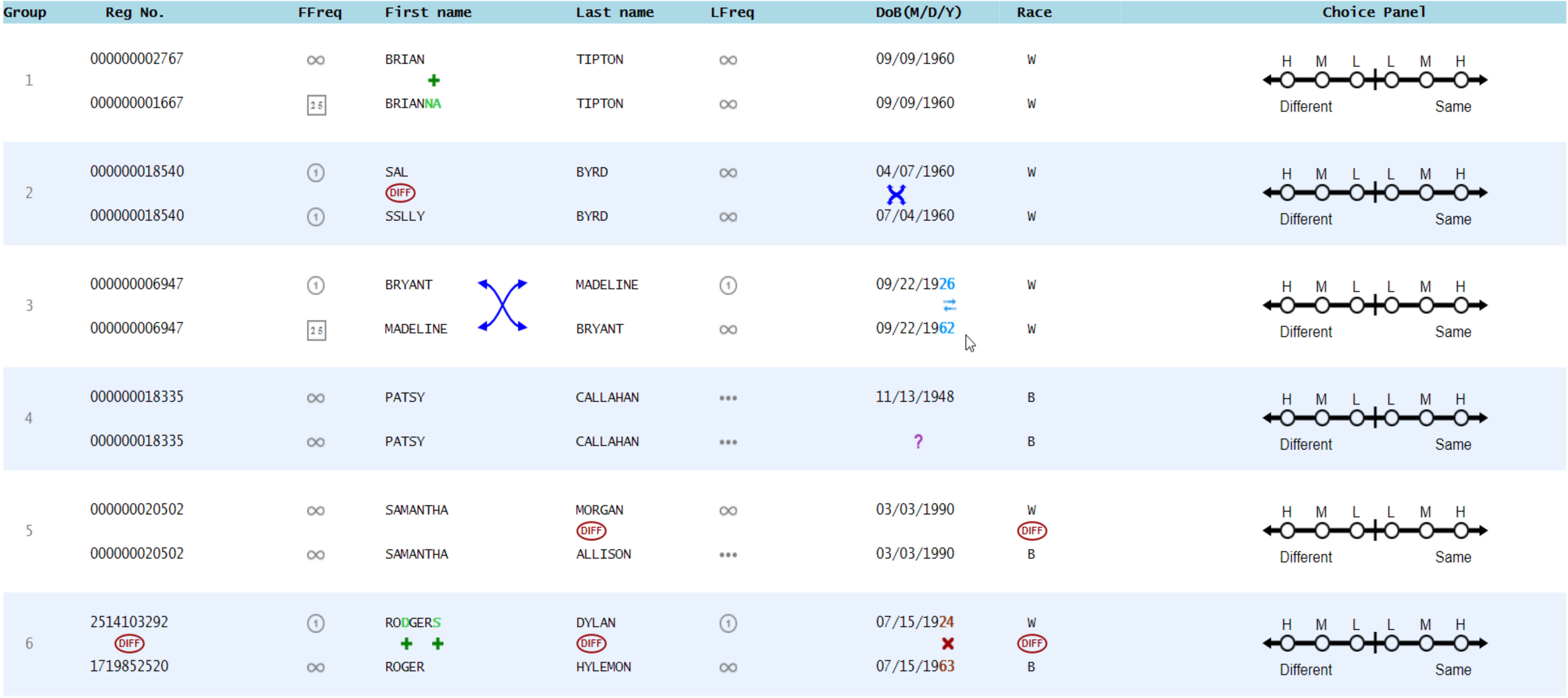
1

2

5

100

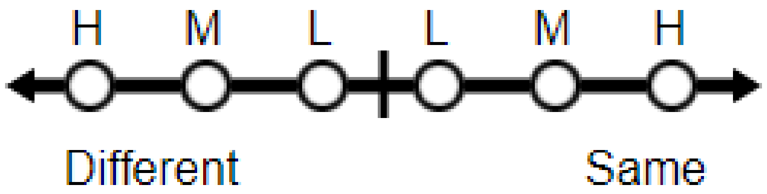
Now the icons will make more sense to you…



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

- You would have noticed a panel on the right:



- This is where your answers go in.
- The right most button meaning you think the 2 records are most definitely the same and left most one meaning they are most definitely different.
- The scales in between represent the various degrees of certainty between them