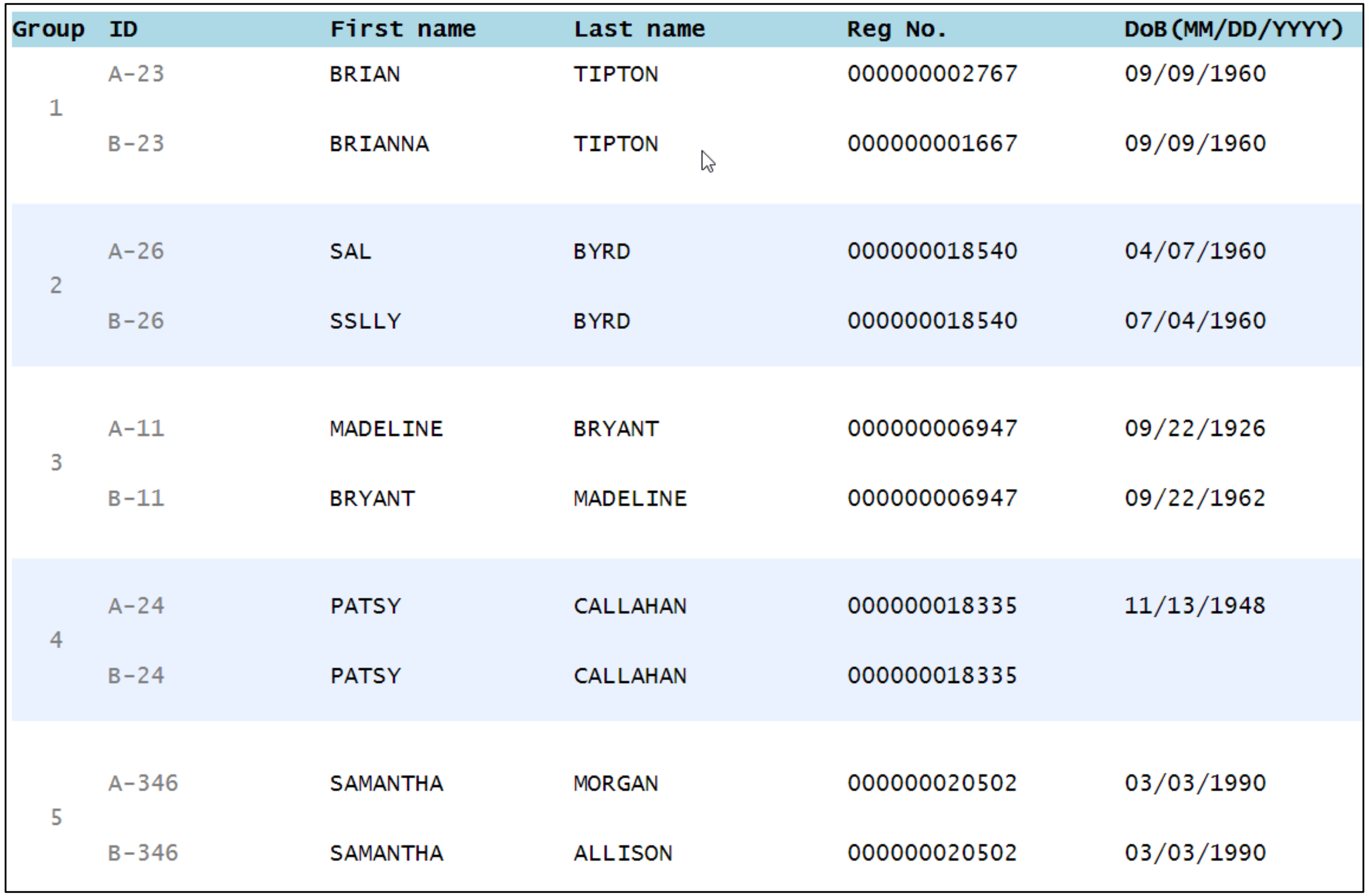
Record Linkage

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	ID	First name	Last name	Reg No.	DoB (MM/DD/YYYY)
1	A-23	BRIAN	TIPTON	000000002767	09/09/1960
	B-23	BRIANNA	TIPTON	000000001667	09/09/1960
2	A-26	SAL	BYRD	000000018540	04/07/1960
	B-26	SSLLY	BYRD	000000018540	07/04/1960
3	A-11	MADELINE	BRYANT	000000006947	09/22/1926
	B-11	BRYANT	MADELINE	000000006947	09/22/1962
4	A-24	PATSY	CALLAHAN	000000018335	11/13/1948
	B-24	PATSY	CALLAHAN	000000018335	
5	A-346	SAMANTHA	MORGAN	000000020502	03/03/1990
	B-346	SAMANTHA	ALLISON	000000020502	03/03/1990

Inherent Nature of Real Data

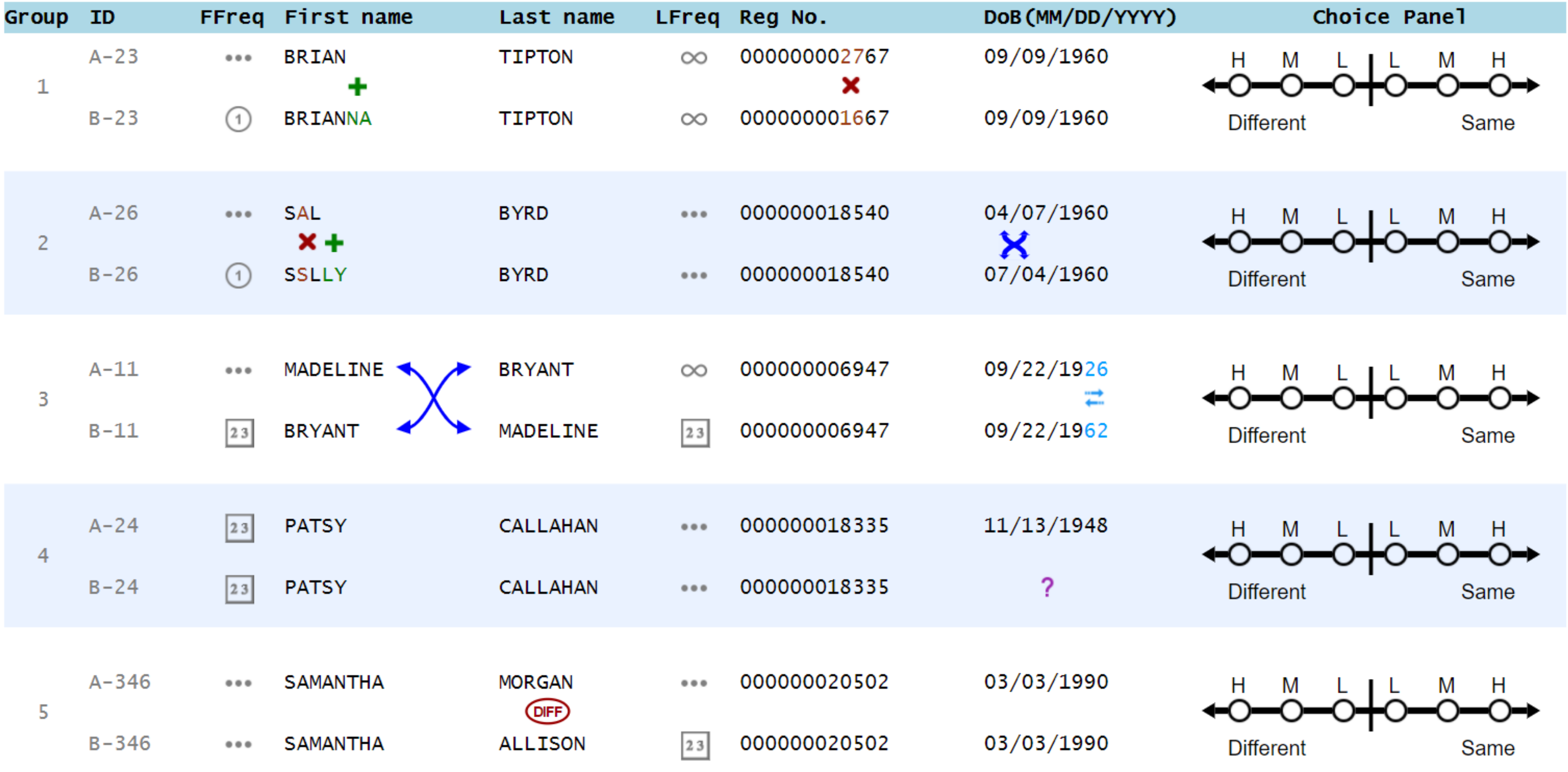
There an inherent problems in real data that make RL difficult
(maybe consider building one page per bullet. TBD after deciding on Monday)

- 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.

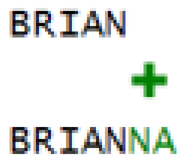
Icons to help you spot the differences



Group	ID	FFreq	First name	Last name	LFreq	Reg No.	DoB (MM/DD/YYYY)	Choice Panel
	A-23	•••	BRIAN	TIPTON	∞	000000002767	09/09/1960	H M L L M H
1			+			✗		←○—○—○—|—○—○—○→
	B-23	①	BRIANNA	TIPTON	∞	000000001667	09/09/1960	Different Same
	A-26	•••	SAL	BYRD	•••	000000018540	04/07/1960	H M L L M H
2			✗ +				✗	←○—○—○—|—○—○—○→
	B-26	①	SSLLY	BYRD	•••	000000018540	07/04/1960	Different Same
	A-11	•••	MADELINE	BRYANT	∞	000000006947	09/22/1926	H M L L M H
3			⇄	⇄			⇄	←○—○—○—|—○—○—○→
	B-11	23	BRYANT	MADELINE	23	000000006947	09/22/1962	Different Same
	A-24	23	PATSY	CALLAHAN	•••	000000018335	11/13/1948	H M L L M H
4								←○—○—○—|—○—○—○→
	B-24	23	PATSY	CALLAHAN	•••	000000018335	?	Different Same
	A-346	•••	SAMANTHA	MORGAN	•••	000000020502	03/03/1990	H M L L M H
5				DIFF				←○—○—○—|—○—○—○→
	B-346	•••	SAMANTHA	ALLISON	23	000000020502	03/03/1990	Different Same

Indel

- Describes an insertion (or deletion) of characters.



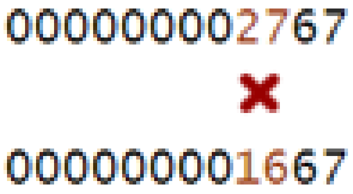
BRIAN

+

BRIANNA

Replace

- When characters are used in the place of another.

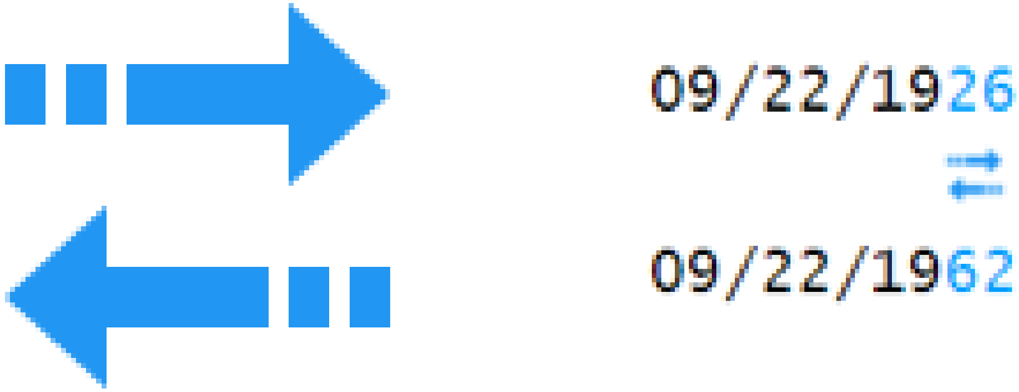


000000002767

000000001667

Transpose

- When the 2 characters are interchanged



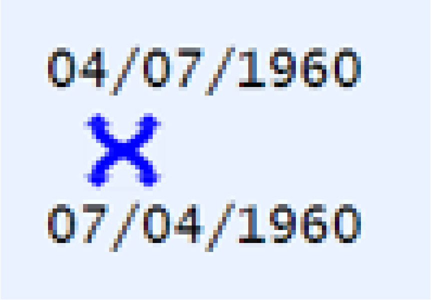
09/22/1926

⇄

09/22/1962

Swaps

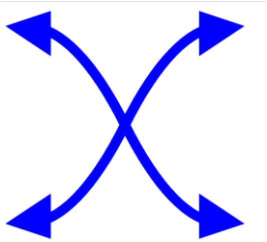
- Sometimes whole values are swapped.
- Date Swaps:

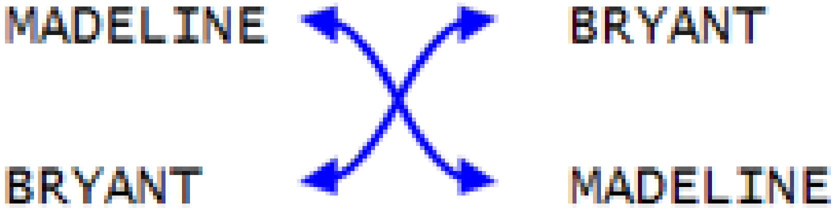


04/07/1960

07/04/1960

- Name swaps:



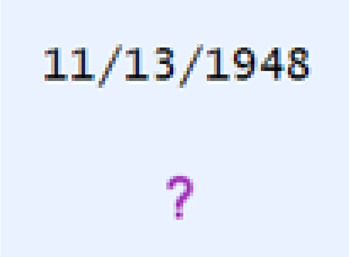


MADELINE BRYANT

BRYANT MADELINE

Missing

- When data is missing



11/13/1948

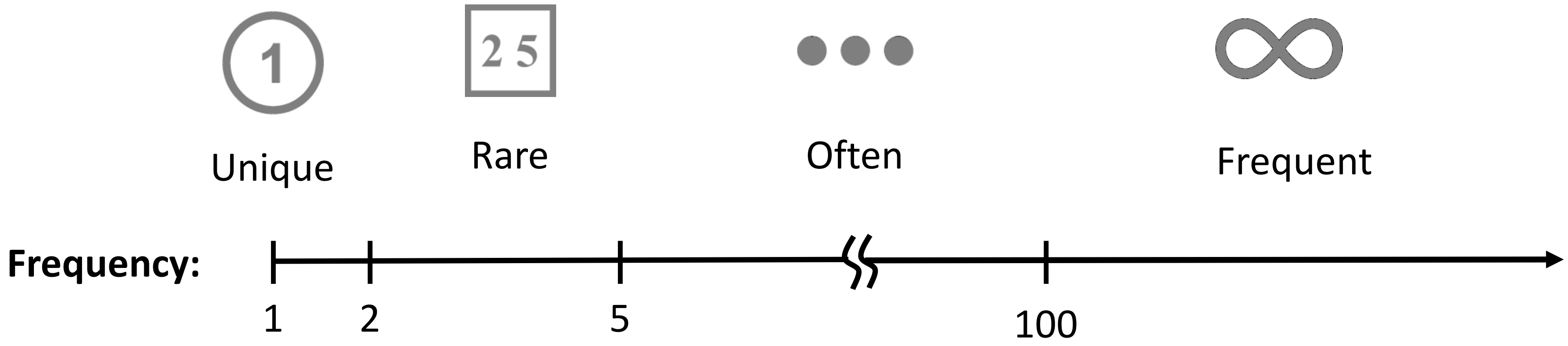
?

Different

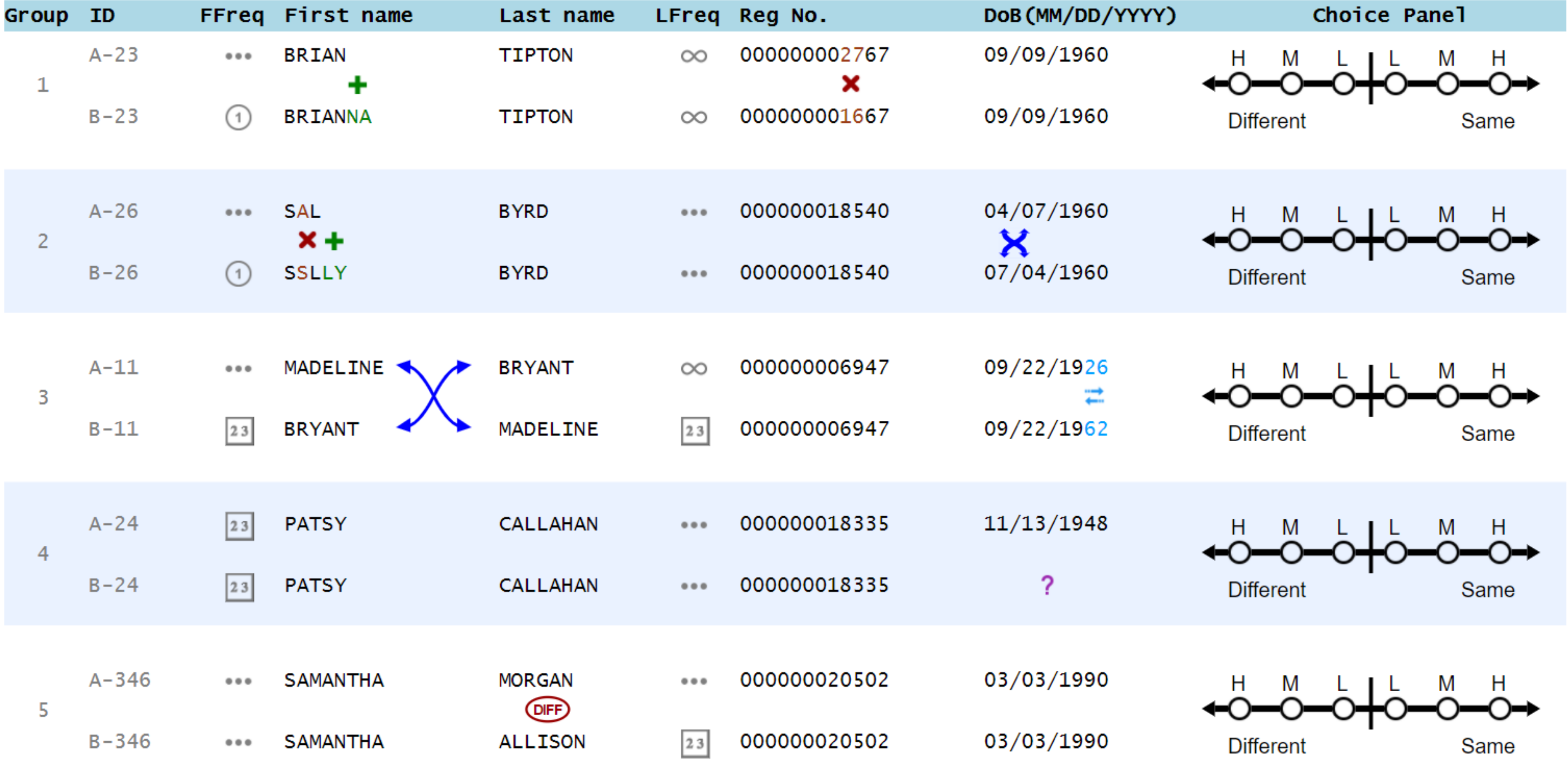
- Sometimes, things are completely different.

Frequency Icons

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



Now the icons will make more sense to you…



Group	ID	FFreq	First name	Last name	LFreq	Reg No.	DoB(MM/DD/YYYY)	Choice Panel
	A-23	•••	BRIAN	TIPTON	∞	000000002767	09/09/1960	H M L | L M H
1			✚			✖		○–○–○–|–○–○–○
	B-23	①	BRIANNA	TIPTON	∞	000000001667	09/09/1960	Different Same
	A-26	•••	SAL	BYRD	•••	000000018540	04/07/1960	H M L | L M H
2			✖ ✚				✖	○–○–○–|–○–○–○
	B-26	①	SSLLY	BYRD	•••	000000018540	07/04/1960	Different Same
	A-11	•••	MADELINE	BRYANT	∞	000000006947	09/22/1926	H M L | L M H
3			⤨	⤨			⇄	○–○–○–|–○–○–○
	B-11	23	BRYANT	MADELINE	23	000000006947	09/22/1962	Different Same
	A-24	23	PATSY	CALLAHAN	•••	000000018335	11/13/1948	H M L | L M H
4								○–○–○–|–○–○–○
	B-24	23	PATSY	CALLAHAN	•••	000000018335	?	Different Same
	A-346	•••	SAMANTHA	MORGAN	•••	000000020502	03/03/1990	H M L | L M H
5				DIFF				○–○–○–|–○–○–○
	B-346	•••	SAMANTHA	ALLISON	23	000000020502	03/03/1990	Different Same

Here is where things get difficult…

- One important thing to note is that privacy comes at a cost and hence, some data may be hidden from you.
- The upcoming pages will give you an understanding of how the data is disclosed on a need to know basis.