

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

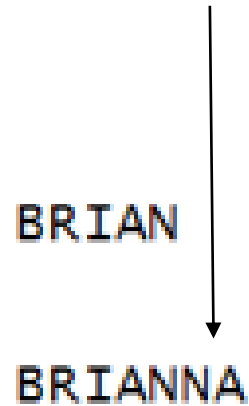
- Data are expressed differently
  - nick names (Elizabeth & Beth)
- Data change over time
  - Women get married and change their last name
- Data are not unique attributes
  - John Smith (there are different people that have the same name)
  - Twins & Family members have similar identifying information 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

# Errors

- In the next few slides, we'll see a quick overview of the different error types that are common in data.

# Indel

- Describes an insertion (or deletion) of characters.



A vertical arrow points from the word "BRIAN" to the word "BRIANNA". The word "BRIAN" is in blue capital letters. The word "BRIANNA" is in blue capital letters, with the "I" and "N" characters highlighted in red. This illustrates an insertion of the characters "I" and "N" into the original string.

BRIAN

BRIANNA

# Replace

- When characters are used in the place of another.

000000002767



000000001667

# Transpose

- When the 2 characters are interchanged

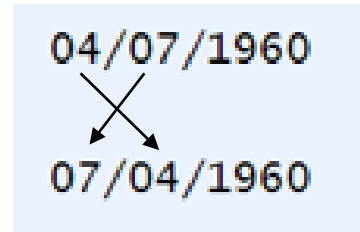
09/22/1926



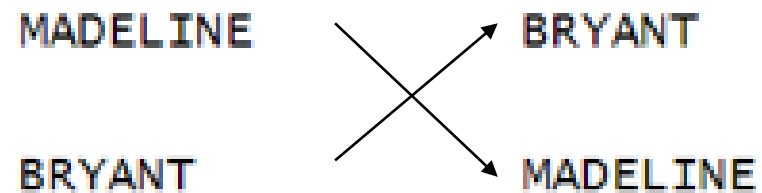
09/22/1962

# Swaps

- Sometimes whole values are swapped.
- Date Swaps:



- Name swaps:



# Missing

- When data is missing

11/13/1948



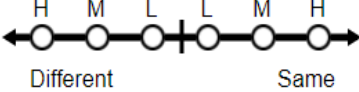
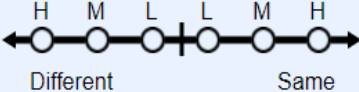
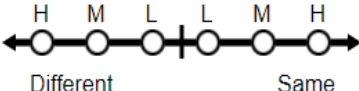
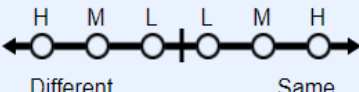
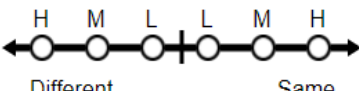
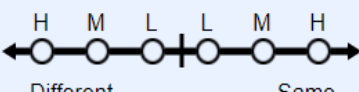
# Different

- Sometimes, things are completely different.

MORGAN

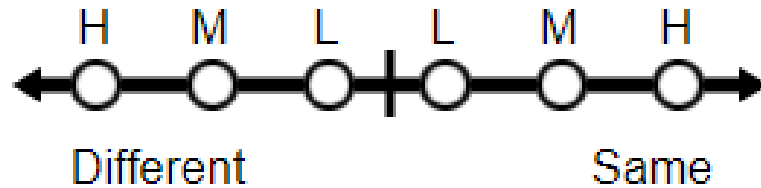
ALLISON

# So this will be your screen...

Group	Reg No.	First name	Last name	DoB (M/D/Y)	Race	Choice Panel
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	

# 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