Study: Complete Record Linkage 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 .

Your job during the study will be **to look at pairs of records about people** and **determine the likelihood that the pair refers to the same real world person**.

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B
2	0000006947	BRYANT	MADELINE	05/02/1962	F	W
	0000006947	MADELINE	BRYANT	05/02/1962	F	W
3	9000018540	SALLY	BYRD	07/04/1960	F	W
	6000008928	JOHN	BYRD	04/07/1960	M	

Common Issues in Real Data

Learn and watch out for common issues in data the during record linkage task

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

Missing

Data are sometimes missing.

7	0000018335	PATSY	CALLAHAN	11/13/1948	F	B
		PATSY	CALLAHAN		F	B

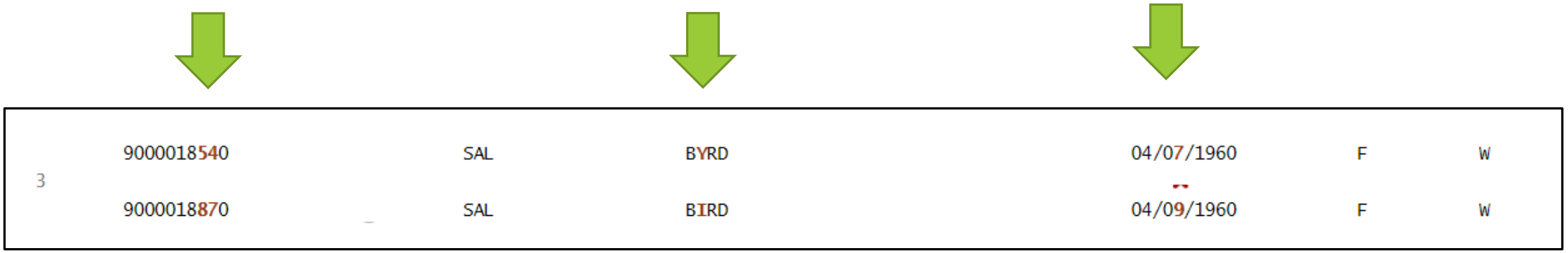
Insertions & Deletions

Insertion (or deletion) of characters are common typing errors

	8000001276	JAYDEN	TIPTON	09/09/1960	M	W
1						
	8000002768	JADEN	TIPTON	09/09/1960	M	W

Replace

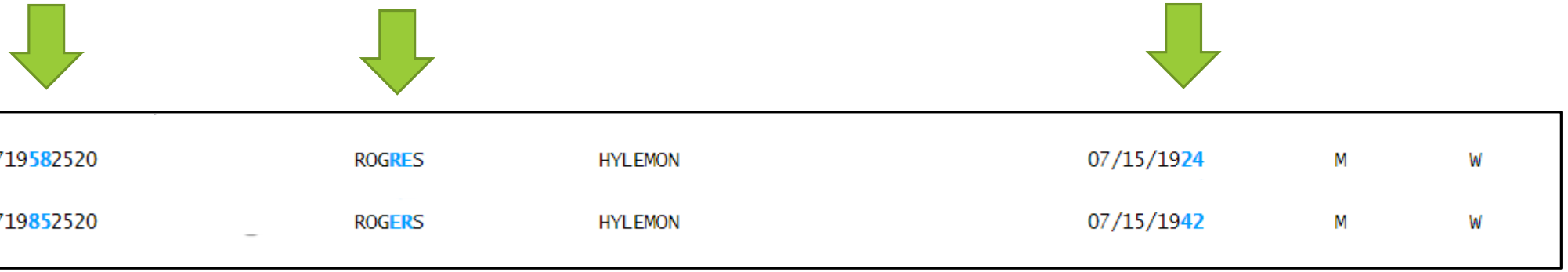
Mistyping can lead to certain characters replacing others



3	9000018540	SAL	BYRD	04/07/1960	F	W
	9000018870	SAL	BIRD	04/09/1960	F	W

Transpose

Two characters can be interchanged by mistake

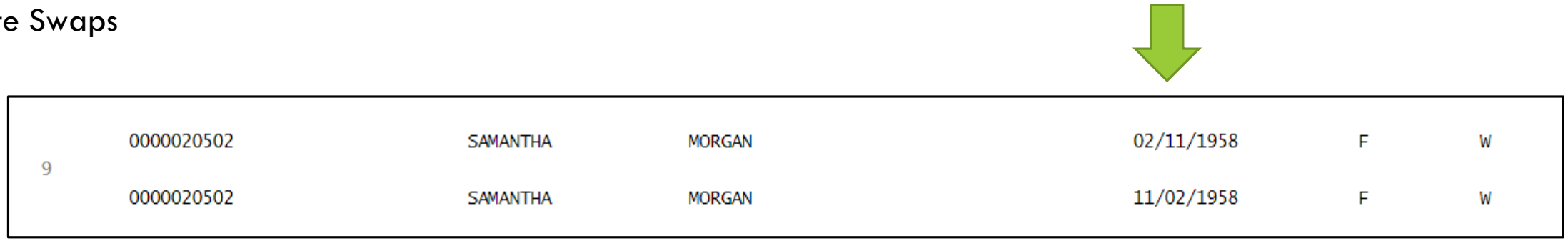


11	1719582520	ROGRES	HYLEMON	07/15/1924	M	W
	1719852520	ROGERS	HYLEMON	07/15/1942	M	W

Swaps

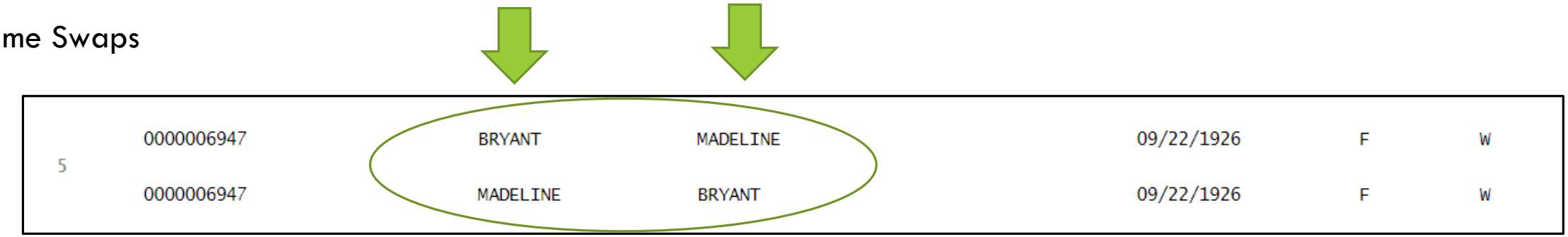
Due to mix up, sometimes whole values are swapped as well

Date Swaps



9	0000020502	SAMANTHA	MORGAN	02/11/1958	F	W
	0000020502	SAMANTHA	MORGAN	11/02/1958	F	W

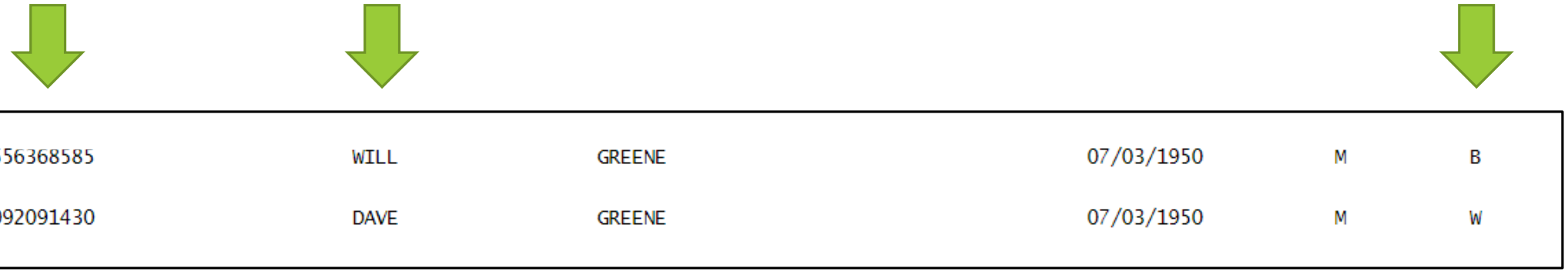
Name Swaps



5	0000006947	BRYANT	MADELINE	09/22/1926	F	W
	0000006947	MADELINE	BRYANT	09/22/1926	F	W

Different

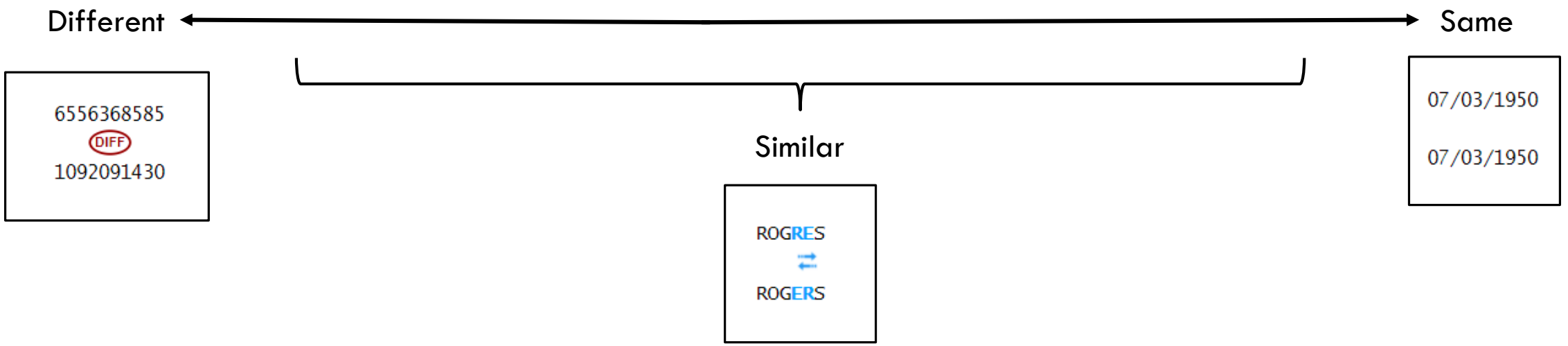
Sometimes the values are different



13	6556368585	WILL	GREENE	07/03/1950	M	B
	1092091430	DAVE	GREENE	07/03/1950	M	W

Comparing values

Roughly, **think of the similarity of 2 values being in a continuum**. On one end, we have values that are very different and on the other end, we have values that are exactly the same. Values with a few operations (like transpose, insert, replace etc.) fall in between.



The Decision Making Process

The answer to a record linkage problem is not a simple yes or no. It is a process that requires you to **think in terms of chances**. Let's take an example:

1	1489599505	JOHNATHON	BUCKNER	02/15/1989	M	W
	1489599505	JONATHON	BUCKNER	02/15/1989	M	W

Asking "Do these 2 records refer to the same person?" is approaching it wrong.

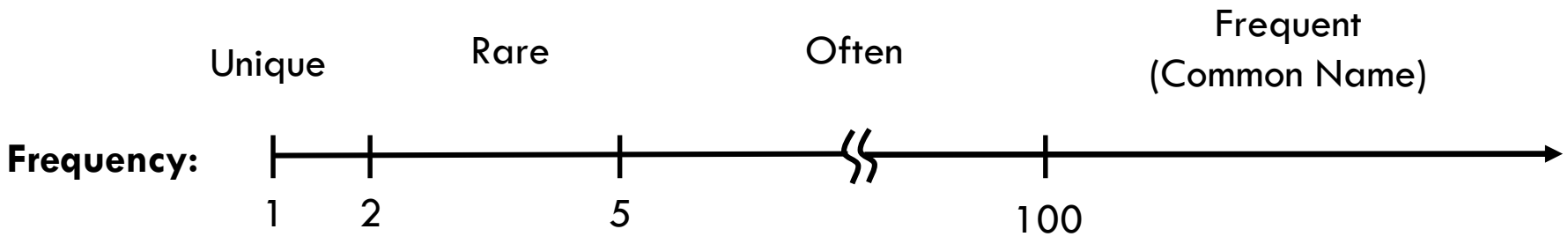
The question to ask is "**What are the chances** that 2 records with the same information in all fields except the first name which is pretty similar**, refer to the same person?**".

You still cannot be 100% sure that they refer to the same person but the chances are pretty high.

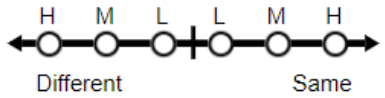
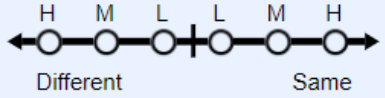
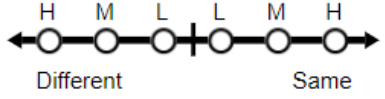
Name Frequencies

How common or unique a given name is important information you can use.

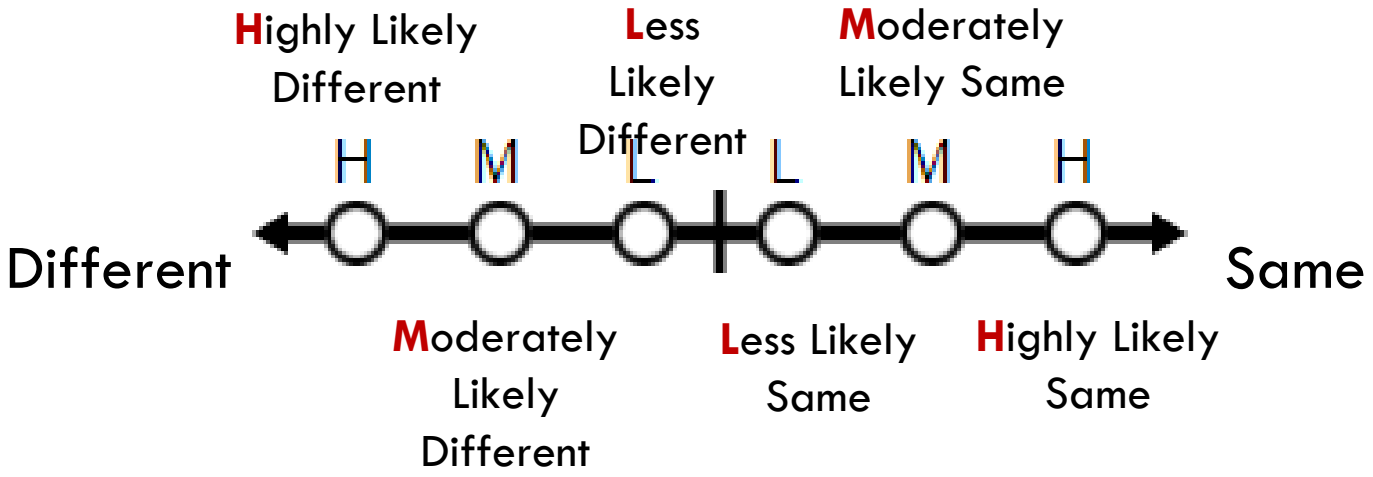
Intuitively, two **common names** might often refer to **different people** while two **rare names** often refer to the **same person**



How to do a Record Linkage Task

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race	Choice Panel
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W	
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B	
2	0000006947	BRYANT	MADELINE	05/02/1962	F	W	
	0000006947	MADELINE	BRYANT	05/02/1962	F	W	
3	9000018540	SALLY	BYRD	07/04/1960	F	W	
	6000008928	JOHN	BYRD	04/07/1960	M		

The Response Panel



You should answer if you think the given pair

- 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.

Ready?

Click on button below to confirm you have gone through all the slides, and the next button to move onto the practice problems.