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)	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	B
2	0000006947	BRYANT	MADELINE	05/02/1962	W
	0000006947	MADELINE	BRYANT	05/02/1962	W
3	9000018540	SALLY	BYRD	07/04/1960	W
	6000008928	JOHN	BYRD	04/07/1960	

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.

	0000018335	∞	PATSY	CALLAHAN	•••	11/13/1948	B
7	?	∞	PATSY	CALLAHAN	•••	?	B

Insertions & Deletions

Insertion (or deletion) of characters are common typing errors

	8000001276	1	JAYDEN	TIPTON	∞	09/09/1960	W
1	+		+				
	8000002768	1	JADEN	TIPTON	∞	09/09/1960	W

Replace ✖

Mistyping can lead to certain characters replacing others

	9000018540	1	SAL	BYRD	∞	04/07/1960	W
3	✖			✖		✖	
	9000018870	1	SAL	BIRD	1	04/09/1960	W

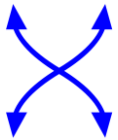
Transpose

Two characters can be interchanged by mistake

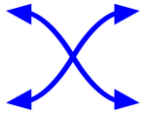
11	1719582520	1	ROGRES	HYLEMON	∞	07/15/1924	W
	1719852520	1	ROGERS	HYLEMON	∞	07/15/1942	W

Swaps

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

Date Swaps 

9	0000020502	∞	SAMANTHA	MORGAN	∞	02/11/1958	1
	0000020502	∞	SAMANTHA	MORGAN	∞	11/02/1958	1

Name Swaps 

5	0000006947	1	BRYANT	MADELINE	1	09/22/1926	W
	0000006947	25	MADELINE	BRYANT	∞	09/22/1926	W

Different DIFF

Sometimes the values are different

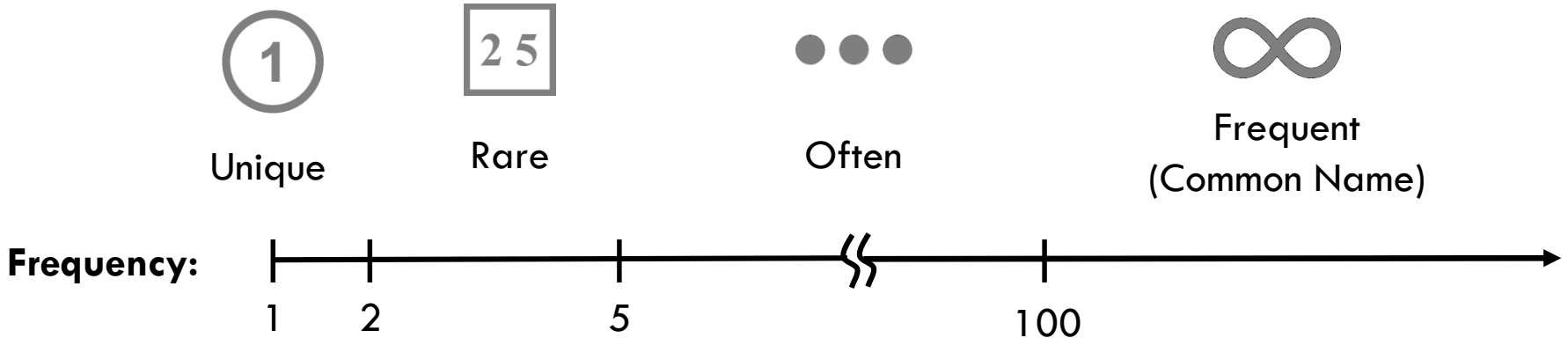
13	6556368585 DIFF	1	WILL DIFF	GREENE	∞	07/03/1950	B DIFF
	1092091430	1	DAVE	GREENE	∞	07/03/1950	W

Name Frequencies

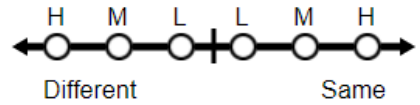
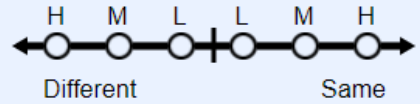
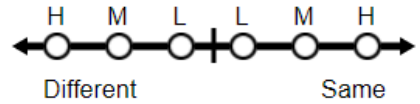
Another important information is how common or unique a given name is.

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

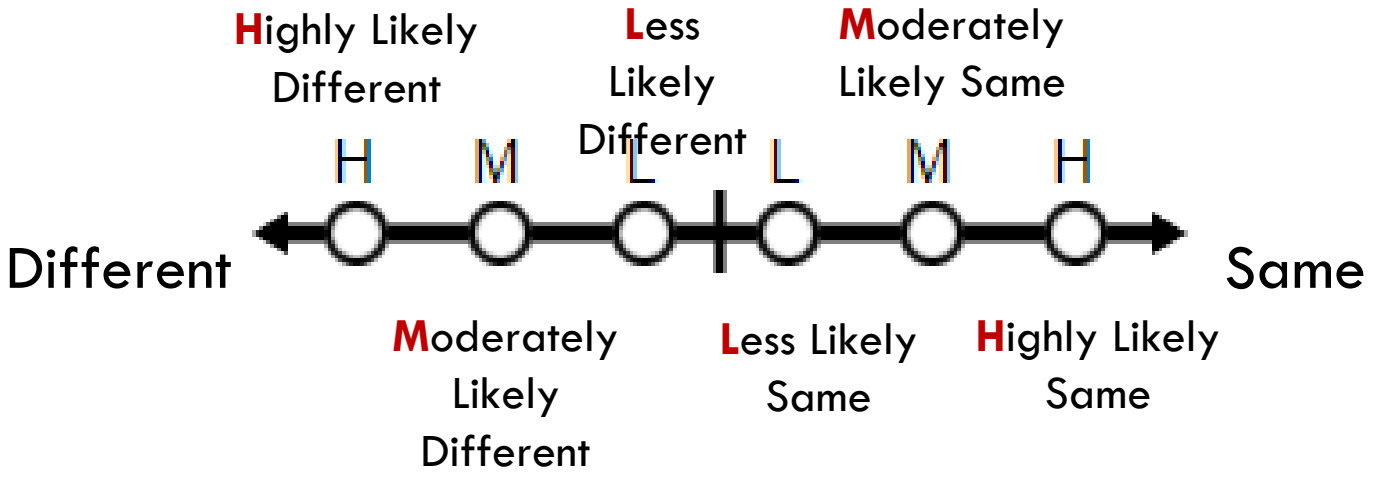
Frequency Icons indicate how many times a given name occurred in the source data



Record Linkage Task

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Race	Choice Panel
1	8000002767 ✖	1	JUDE	WILLIAM ✚	1	09/09/1906 ⇄	W DIFF	
	8000003567	1	JUDE	WILLIAM JR	1	09/09/1960	B	
2	0000006947	1	BRYANT	MADELINE	1	05/02/1962	W	
	0000006947	25	MADELINE	BRYANT	∞	05/02/1962	W	
3	9000018540 DIFF	•••	SALLY DIFF	BYRD	∞	07/04/1960 ✖	W	
	6000008928	∞	JOHN	BYRD	∞	04/07/1960	?	

The Answer Submission 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