Same or Different People?

Data entry collects information about people. Your job in this study is to:
1) Look at pairs of rows of data about people
2) Decide whether or not the pair refers to the same 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 with Data about People

Watch out for common issues

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
- Same names in Families with different suffix (Jr and Sr)

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 Values ?

Data are sometimes missing.

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
7	0000018335	PATSY	CALLAHAN	11/13/1948	F	B
	?	PATSY	CALLAHAN	?	F	B

Added or Deletions Characters

Insertion (or deletion) of characters are common typing errors

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
	8000001276	JAYDEN	TIPTON	09/09/1960	M	W
1						
	8000002768	JADEN	TIPTON	09/09/1960	M	W

Different Characters ✗

Mistyping can lead to certain characters replacing others

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
	9000018540	SAL	BYRD	04/07/1960	F	W
3	✗		✗	✗		
	9000018870	SAL	BIRD	04/09/1960	F	W

Switched Characters

Two characters can be interchanged by mistake

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
	1719582520	ROGRES	HYLEMON	07/15/1924	M	W
11	⇄	⇄		⇄		
	1719852520	ROGERS	HYLEMON	07/15/1942	M	W

Column Swaps

Sometimes whole values are swapped as well:

Date Swaps

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
9	0000020502	SAMANTHA	MORGAN	02/11/1958	F	W
	0000020502	SAMANTHA	MORGAN	11/02/1958	F	W

Name Swaps

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
5	0000006947	BRYANT	MADELINE	09/22/1926	F	W
	0000006947	MADELINE	BRYANT	09/22/1926	F	W

Different DIFF

This icon is shown if the values in a column are very different.

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
	6556368585	WILL	GREENE	07/03/1950	M	B
13	DIFF	DIFF				DIFF
	1092091430	DAVE	GREENE	07/03/1950	M	W

Common and Rare Names

It can be helpful to consider how common or unique a person's name is.

For example, consider how common these names might be in the United States:

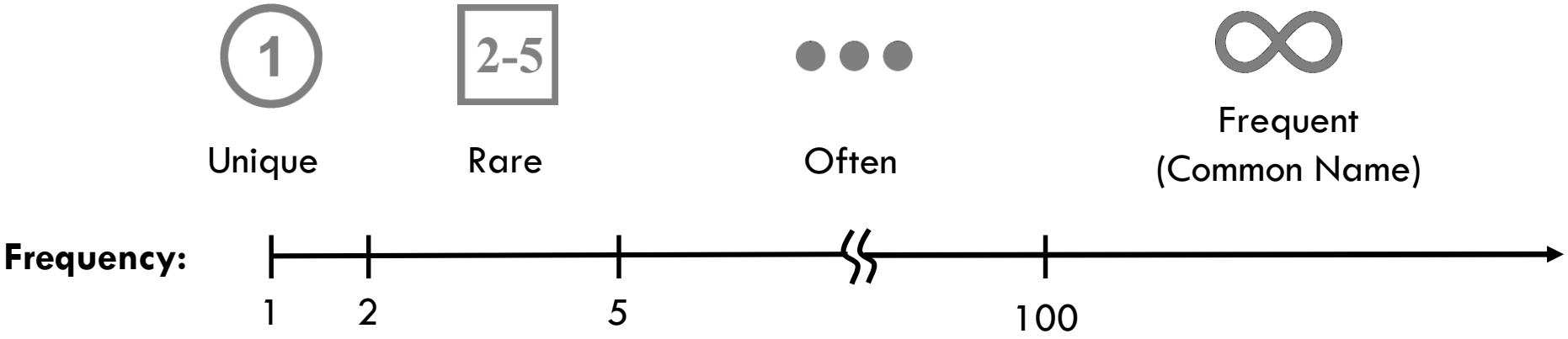
Very Common First Names	Uncommon First Names
Michael	Brooklynn
Matthew	Jamarion
Mary	Jaxson
Ashley	Araceli

Very Common Family Names	Uncommon Family Names
Smith	Febland
Jones	Pober
Jackson	Southwark
Williams	Raynott

Name Frequency

It would not be surprising for two people to have the same **common name**, but it might be unlikely for two people to have the same **rare names**.

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



How to use Name Frequencies

Ffreq is the frequency the name has occurred as a **First name.**

Lfreq is the frequency the name has occurred as a **Last Name**.

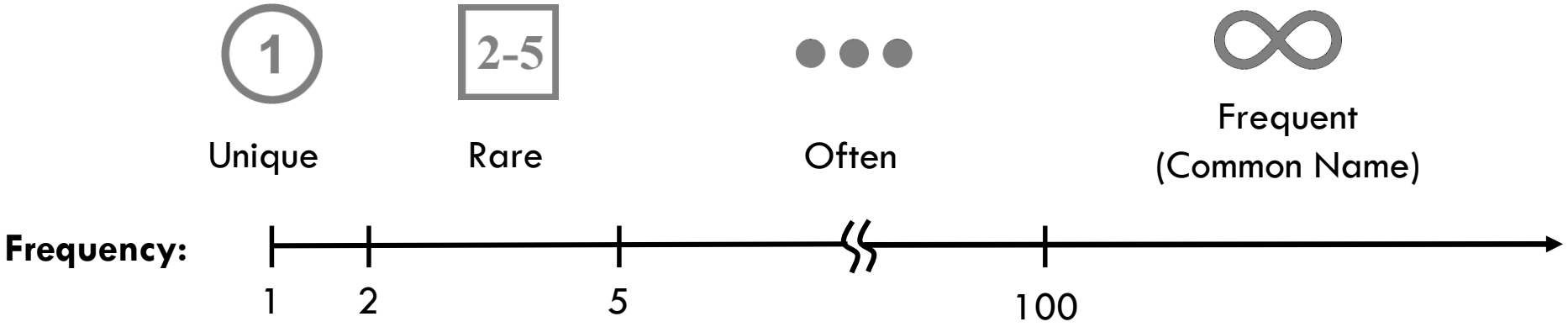
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767	∞	JOHN	SMITH	∞	09/09/1906	M	W
	✖							
	8000003567	∞	JOHN	SMITH	∞	09/09/1906	M	W

Since **John Smith** is a **common name**. Despite the ID being pretty similar, the chances that both the records refer to the same person are **low.**

Another Example of Frequencies

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
2	0000006847	Unique Name 1	DEQUAN	WAMBOLDT	1	05/02/1962	F	W
	0000006947	1	DEQUAN	WAMBOLDT	1	05/02/1962	F	W

Here, both first and last names are **unique**. The icons mean the first and last name only appear one time in both data sources. The chances that this is the same person are much **higher**!



Decision Making

Deciding if two rows are the same person is not a simple "yes" or "no" decision. You have to **think in terms of chance**. Let's take an example:

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000001276	1	JAYDEN	TIPTON	∞	09/09/1960	M	W
	+		+					
	8000002768	1	JADEN	TIPTON	∞	09/09/1960	M	W

We don't know for sure if these two rows refer to the same person. You should ask yourself:

> "**What are the chances** that two rows refer to the same person when the ID and the unique first names have small differences and all other info same?"

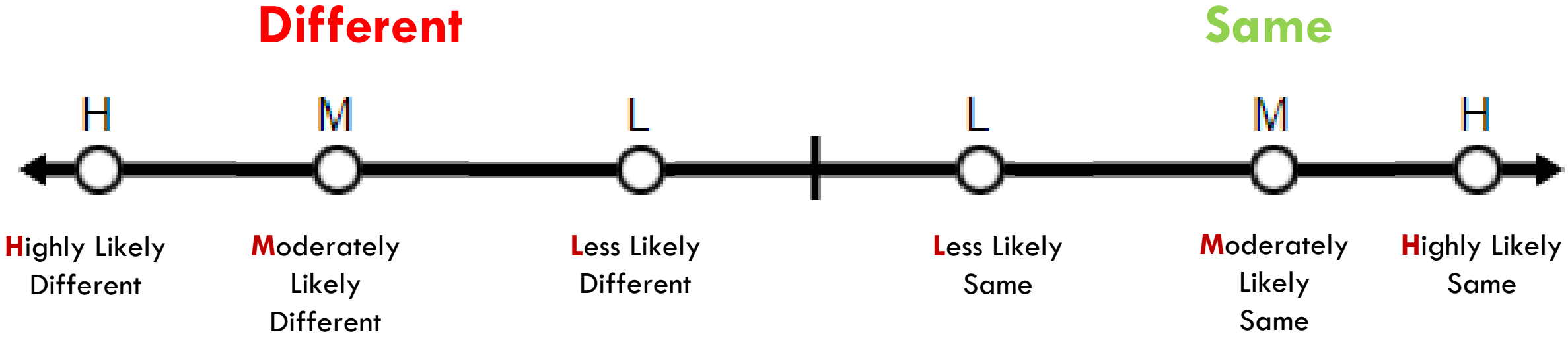
The chances are pretty high that this is the same person, but you still cannot be 100% sure.

How to Give Your Answer

Next, we explain how you will give you answer for each pair of people.

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

The Answer Buttons



If you think the rows are the **same person**, click one of the choices on the **right side**. Pick one of L, M, H depending on your confidence level.

If you think the rows are for **different people**, click one of the choices on the **left side**. Pick one of L, M, H depending on your confidence level.

Ready to Give it a Try?
Let's do Some Practice Problems

The next step will be to do some practice problems.

Ready to continue?

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