

BASICS OF SEMI-AUTOMATED RECORD LINKAGE

REVIEW THESE SLIDES AT YOUR OWN PACE

Record Linkage: Same or Different People?

Given multiple databases, determine if records refer to the same real world people or not

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	

Maybe
Father/Son

Probably
data error

Maybe
Twins

Common Issues with Data about People

Make Record Linkage Difficult to do Fully Automatically

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)

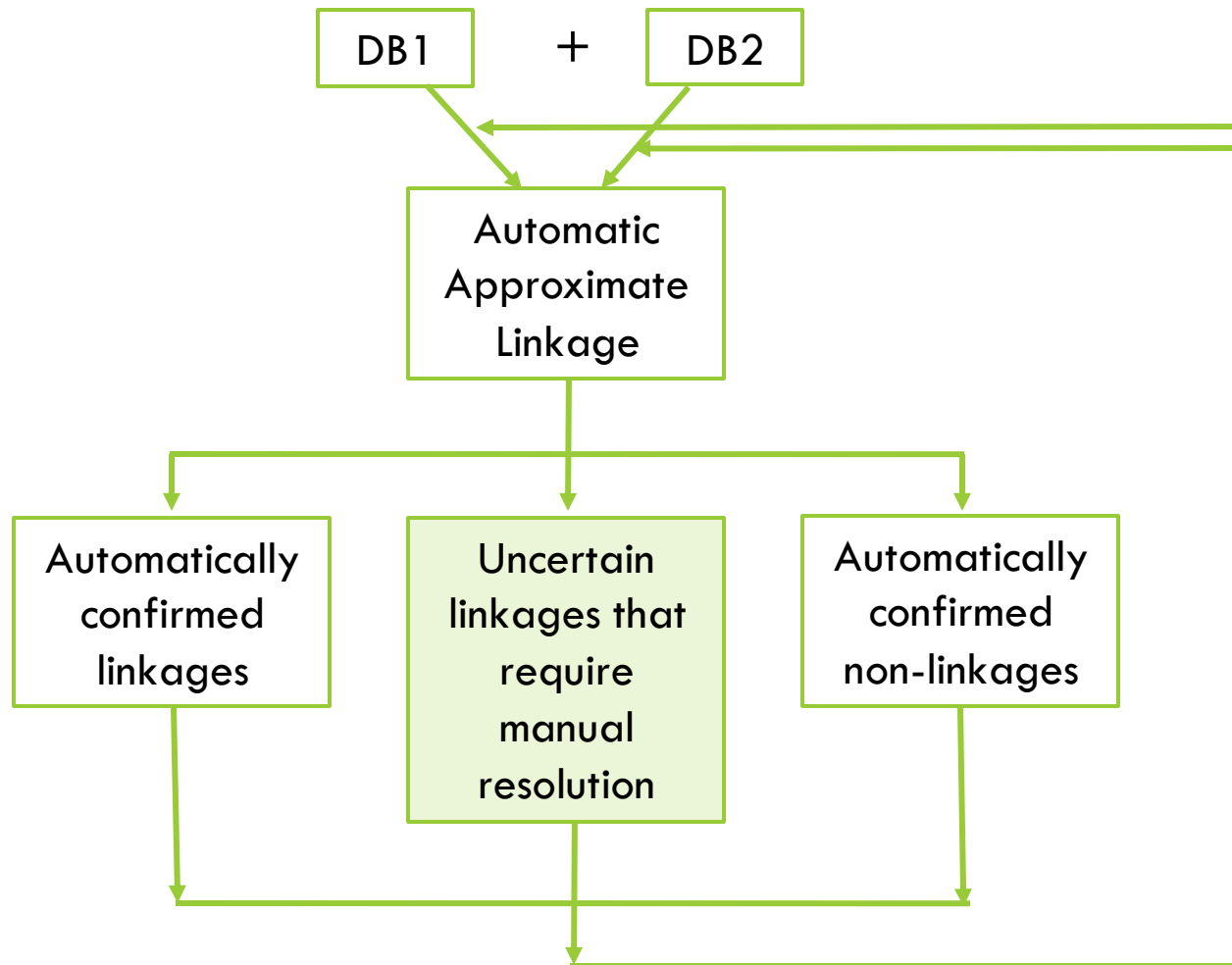
Data are sometimes missing

- 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

Approximate Record Linkage Human-Computer System (semi-automated system)



- Human Interaction With Data for
 - Standardize
 - Clean Data
 - Tuning parameters for automatic algorithms
 - Build Training Data
- 75%-80% automatics
- 15%-25% manual resolution

Status Quo: Show everything

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	

Are there ways to enhance privacy during the human interaction with data for semi-automated record linkage ?

1. USE VISUAL MARK UP

- **TO SHARE META DATA ABOUT THE DIFFERENCES BETWEEN RECORDS**
- **TO FACILITATE LINKAGE DECISIONS**

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

Different Characters

Mistyping can lead to certain characters replacing others

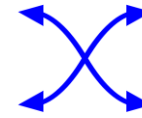
Pair	ID	First name	Last name	DoB (M/D/Y)	Sex	Race
3	9000018540 	SAL	BYRD 	04/07/1960 	F	W
	9000018870	SAL	BIRD	04/09/1960	F	W

Switched Characters

Two characters can be interchanged by mistake

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



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

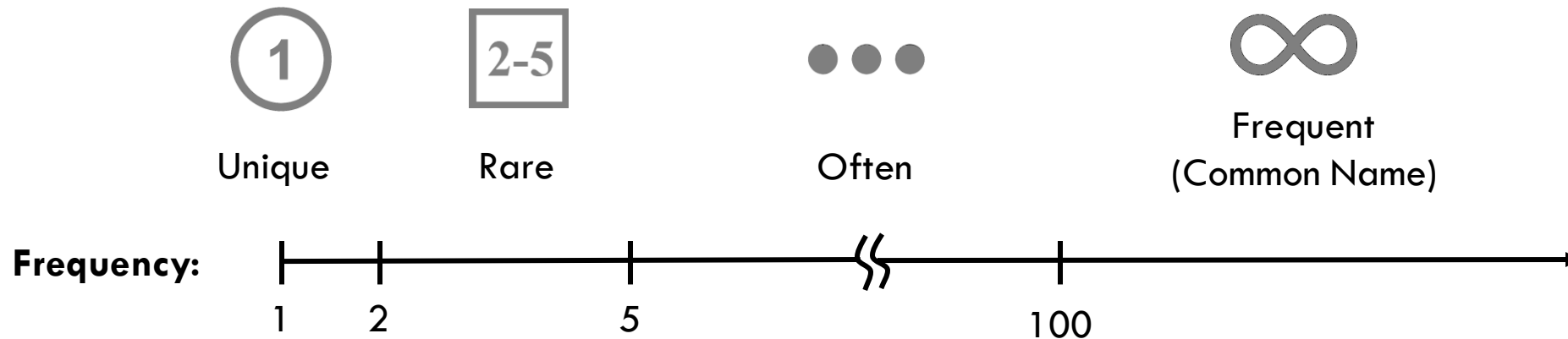
Michelle Williams ?

Name Frequency

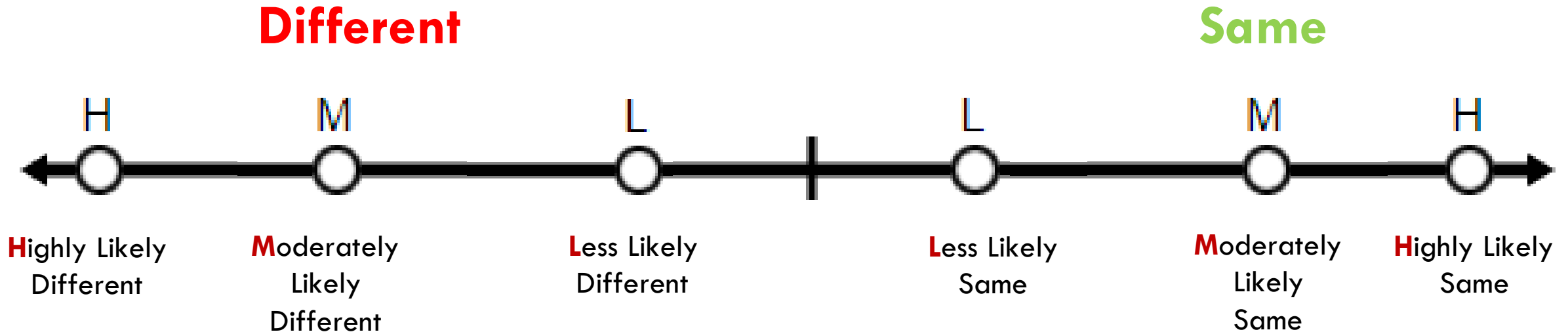


It would not be surprising for two people to have the same **common name**, but it might be unlikely for two people to have the same **rare names**.

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



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.

2. USE VISUAL MASKING FOR PRIVACY

- USE VISUAL MASKING TO HIDE INFORMATION**

Checkmarks: Identical Values

Identical values are shown as checkmarks.

Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
1	1234567891	...	BRIAN	TIPTON	∞	09/09/1960	F	W
	1234561291	<div>2-5</div>	BRIANNA	TIPTON	∞	09/09/1960	F	W



Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
1	*****@*	...	*****	✓	∞	✓	✓	✓
	*****&*	<div>2-5</div>	*****&	✓	∞	✓	✓	✓

Stars Used in Similar IDs

When two items are similar, **stars** (they look like this ***) are used for characters that are the **same**.

@@@ and &&& show the characters that are **different**.

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	1234567891	...	BRIAN	TIPTON	∞	09/09/1960	F	W
	1234561291	<div>2-5</div>	BRIANNA	TIPTON	∞	09/09/1960	F	W



Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	*****@@*	...	*****	✓	∞	✓	✓	✓
	*****&&*	<div>2-5</div>	*****&&	✓	∞	✓	✓	✓

*** for Missing Values

When one of the values in a pair is missing, the other one is represented by ***

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



Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
7	*****	...	✓	✓	...	**/**/****	✓	✓
	?	...	✓	✓	...	?	✓	✓

Different Items



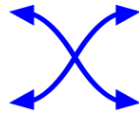
When two items are very different, they are shown as @@@ and &&&

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
13	6556368585 DIFF	①	WILL DIFF	GREENE	...	07/03/1950	M	B DIFF
	1092091430	①	DAVE	GREENE	...	07/03/1950	M	W



Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
13	@@@@@@@@@@ DIFF	①	@@@@ DIFF	✓	...	✓	✓	@ DIFF
	&&&&&&&&&	①	&&&&	✓	...	✓	✓	&

Swaps



When columns have swapped values, the swapped parts are shown by &&& and @@@@

Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
5	0000006947	①	BRYANT	MADLINE	①	02/05/1962	F	W
	0000006947	2-5	MADLINE	BRYANT	...	05/02/1962	F	W



Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
5	✓	①	&&&&&	@@@@@@@@	①	@@/&&/****	✓	✓
	✓	2-5	@@@@@@@@	&&&&&	...	&&/@@/****	✓	✓

Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
1	8000002767 ✗	①	JUDE	WILLIAM +	①	09/09/1906 ↔	M	W
	8000003567	①	JUDE	WILLIAM JR	①	09/09/1960	M	DIFF B
2	0000006947	①	BRYANT	MADELINE	①	05/02/1962	F	W
	0000006947	25	MADELINE	BRYANT	...	05/02/1962	F	W
3	9000018540 DIFF	...	SALLY DIFF	BYRD	...	07/04/1960 X	F DIFF	W
	6000008928	∞	JOHN	BYRD	...	04/07/1960	M	?

Pair	ID	FFreq	First name	Last name	LFreq	DoB (M/D/Y)	Sex	Race
1	*****@** ✗	①	✓	***** +	①	**/**/**@ ↔	✓	@ DIFF
	*****&*	①	✓	***** &&	①	**/**/**&&	✓	&
2	✓	①	&&&&&&	@@@@@@@@	①	✓	✓	✓
	✓	2-5	@@@@@@@@	&&&&&	...	✓	✓	✓
3	@@@@@@@@@@@@ DIFF	...	@@@@@ DIFF	✓	...	@@/@@/***** X	@ DIFF	*
	&&&&&&&&&	∞	&&&	✓	...	&&/@@/*****	&	?



3. INTERACTIVE ON-DEMAND INTERFACE

- TO FIND THE OPTIMAL BALANCE BETWEEN PRIVACY AND HIGH QUALITY RECORD LINKAGE RESULTS



Interactive On-Demand Interface

That was hard, wasn't it?

Sometimes, data masking can hide data that might be essential for record linkage.
What if **you could open up the masked data as you need to see more?**

Over the next few pages, we will walk you through an interactive on-demand interface for record linkage.