

Unique DB IDs

- ▶ Expect same person
 - ▶ Typos
 - ▶ Change in last name (females)
 - ▶ Nick names
- ▶ BUT, often twins can get IDs mixed up
 - ▶ Can you find the twins?

Sometimes Unique DB IDs have typos

- ▶ Expect same person
 - ▶ Typos
 - ▶ Change in last name (females)
 - ▶ Nick names
- ▶ BUT also, often twins have very similar IDs
 - ▶ Can you find the twins?

Thus, different DB IDs (name & dob)

- Expect different person
- Remember different people can at random share some identifying information. For example,
 - Have same first name and dob, especially for very common names
 - Have same last name and dob, especially for very common names
 - Have same first name and last name, especially for very common names
- Also, family members often share identifying information (last name)
 - Twins would have same last name and dob.
 - Jr/Sr would have same fname/lname but different DOB about 30ish years part
- This can be confusing when database often also have duplicate records for the same person
 - This means two different IDs
 - But same identifying (name, dob, race etc) information.
 - Or SIMILAR
 - Typos
 - Change in last name (females)
 - Nick names
- Can you find the SAME person?
 - Ignore different lastname for females
 - Same first name that is rare and same dob

Sometimes Unique DB IDs are missing. Need to use other info (name & dob)

- Expect different person
- Remember different people can at random share some identifying information. For example,
 - Have same first name and dob, especially for very common names
 - Have same last name and dob, especially for very common names
 - Have same first name and last name, especially for very common names
- Also, family members often share identifying information (last name)
 - Twins would have same last name and dob.
 - Jr/Sr would have same fname/lname but different DOB about 30ish years part
- This can be confusing when database often also have duplicate records for the same person
 - This means two different IDs
 - But same identifying (name, dob, race etc) information.
 - Or SIMILAR
 - Typos
 - Change in last name (females)
 - Nick names
- Can you find the SAME person?
 - Ignore different lastname for females
 - Same first name that is rare and same dob



Record Linkage

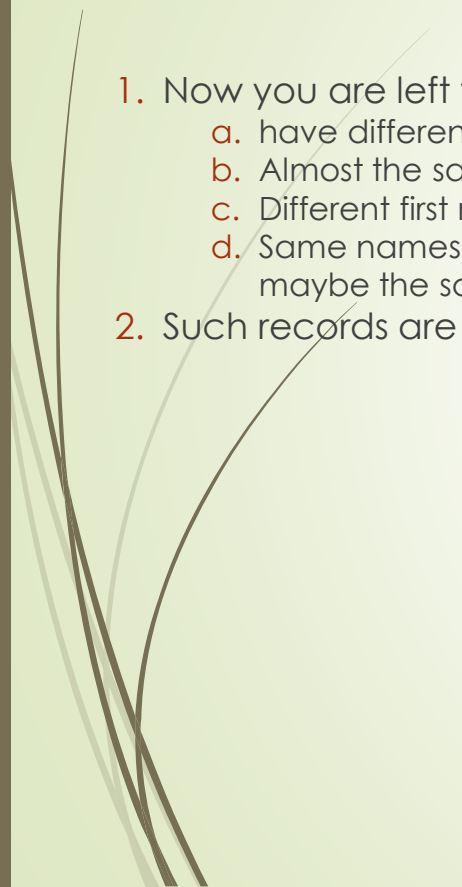
Selective Disclosure

What is Record Linkage?

1. Let's say you are given 2 digital telephone directories, one from 2010 and another from 2015 and you are asked to identify the same people in both of them.
2. As a first pass you probably will write a program that says "If the full name, the phone number and the addresses match, the records refer to the same person. Filter them out." and the computer does the heavy lifting.

Unique DB IDs

- ▶ Expect same person
 - ▶ Typos
 - ▶ Change in last name (females)
 - ▶ Nick names
- ▶ BUT, often twins can get IDs mixed up
 - ▶ Can you find the twins?

- 
1. Now you are left with records like this:
 - a. have different last names but all other fields match. Perhaps the person got married?
 - b. Almost the same name. "Roger" vs "Roge". Probably a typo?
 - c. Different first names but all other fields match. Twins maybe?
 - d. Same names but addresses and phone number differ. Two different people with same names or maybe the same person moved?
 2. Such records are tricky and need careful human inspection to be labelled same or different.

Our study

1. In the study we deal with a dataset that has 4 fields:
 - a. First Name
 - b. Last Name
 - c. Registration Number
 - d. Date of Birth
2. You are supposed to decide whether 2 records given to you are same or not based on the fields provided
3. The raw data behind the study will look like this:

First name	Last name	Reg No.	DoB (MM/DD/YYYY)
DAVID	CARROLL	000000015574	04/07/1960
DAVID	ICARROLL	000000015574	07/04/1960

Markers

1. The study uses markers to help you direct your attention and make the decision making process easier.
2. There are basically 4 kinds of markers.
 - a. Same and different marker
 - b. Swaps and Missing markers
 - c. Error markers
 - d. Frequency markers

Same and different markers

1. If the fields are exactly the same a check-mark is used to represent it in the place of the actual value.

First name
DAVID
DAVID

First name
✓

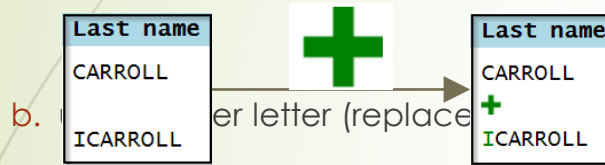
2. If the fields are different, we use a "DIFF" marker to point that they are different.

SAXTON
AUTREY

SAXTON
DIFF
AUTREY

Error Markers

1. Think of them as typos. Sometimes when typing, we add an
 - a. extra letter/forget a letter (insert/delete)



- c. type them in the wrong order (transpose)
- 2.