Record Linkage

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.

Group	ID	First name	Last name	Reg No.	DoB (MM/DD/YYYY)
1	A-23	BRIAN	TIPTON	000000002767	09/09/1960
	B-23	BRIANNA	TIPTON	000000001667	09/09/1960
2	A-26	SAL	BYRD	000000018540	04/07/1960
	B-26	SSLLY	BYRD	000000018540	07/04/1960
3	A-11	MADELINE	BRYANT	000000006947	09/22/1926
	B-11	BRYANT	MADELINE	000000006947	09/22/1962
4	A-24	PATSY	CALLAHAN	000000018335	11/13/1948
	B-24	PATSY	CALLAHAN	000000018335	
5	A-346	SAMANTHA	MORGAN	000000020502	03/03/1990
	B-346	SAMANTHA	ALLISON	000000020502	03/03/1990

Inherent Nature of Real Data

There an inherent problems in real data that make RL difficult
(maybe consider building one page per bullet. TBD after deciding on Monday)

- 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
- Missing Data
  - ssn are often missing
- Errors in Data
  - 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

Indel

- Describes an insertion (or deletion) of characters.

BRIAN

BRIANNA

Replace

- When characters are used in the place of another.

000000002767

000000001667

Transpose

- When the 2 characters are interchanged

```
09/22/1926

09/22/1962
```

Swaps

- Sometimes whole values are swapped.
- Date Swaps:

```
04/07/1960

07/04/1960
```

- Name swaps:

```
MADELINE        BRYANT

BRYANT          MADELINE
```

Missing

- When data is missing

```
11/13/1948
```

Different

- Sometimes, things are completely different.