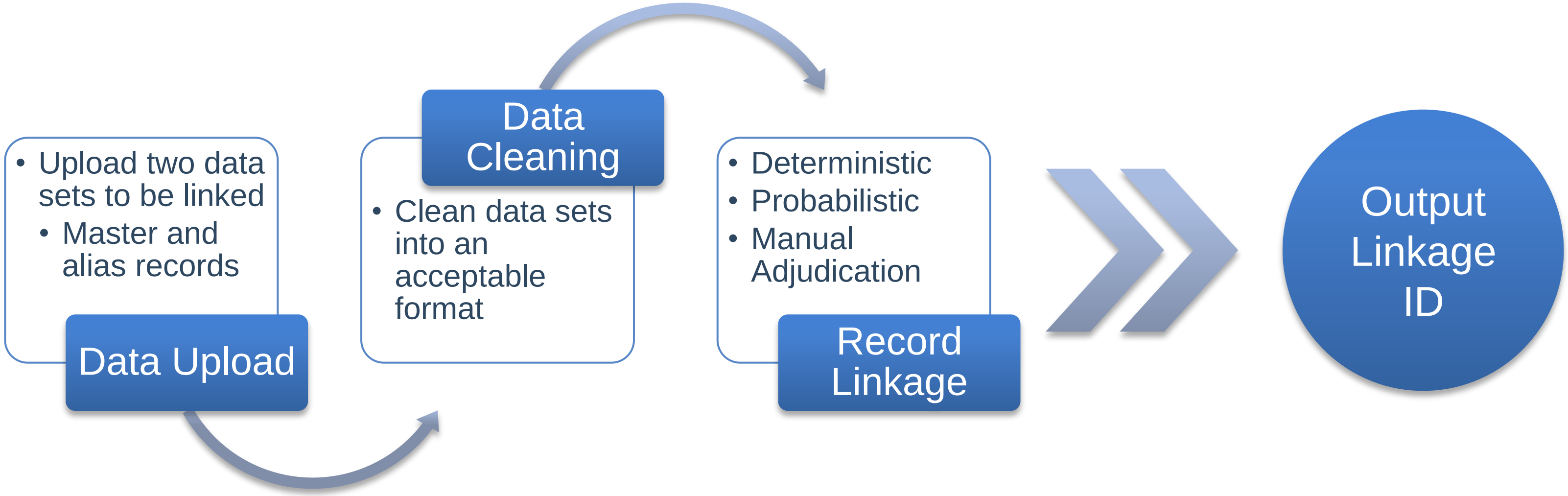
Next Level Analytics, Inc.

Web Application

Introduction to the Record Linkage Web Application

Introduction

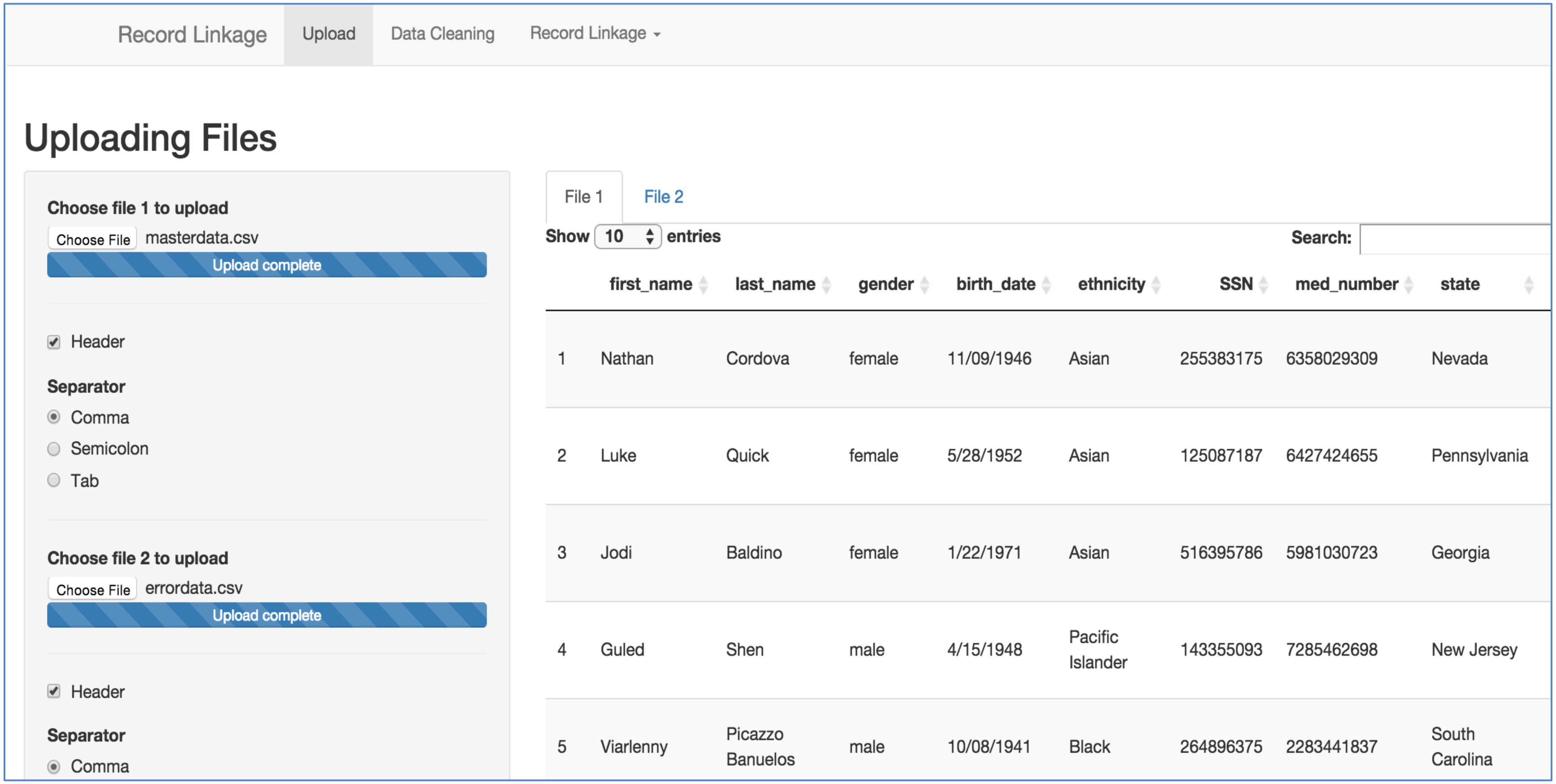
- The record linkage web application takes two data sets of person records and find matches between the two data sets.
- The first data set should be a master data set of unique original records, and the second should be a list of records to be linked to the master data set.
- The application provides a easy-to-use interface to complex linkage algorithms, the process flow is as follows:



Data Upload

Upload two .csv files containing the master data set and an alias data set to be linked.

Uploading Files



Uploading Files

Choose file 1 to upload

Choose File masterdata.csv

Upload complete

☑ Header

Separator

● Comma
○ Semicolon
○ Tab

Choose file 2 to upload

Choose File errordata.csv

Upload complete

☑ Header

Separator

● Comma

Show 10 **entries** **Search:**

	first_name	last_name	gender	birth_date	ethnicity	SSN	med_number	state
1	Nathan	Cordova	female	11/09/1946	Asian	255383175	6358029309	Nevada
2	Luke	Quick	female	5/28/1952	Asian	125087187	6427424655	Pennsylvania
3	Jodi	Baldino	female	1/22/1971	Asian	516395786	5981030723	Georgia
4	Guled	Shen	male	4/15/1948	Pacific Islander	143355093	7285462698	New Jersey
5	Viarlenny	Picazzo Banuelos	male	10/08/1941	Black	264896375	2283441837	South Carolina

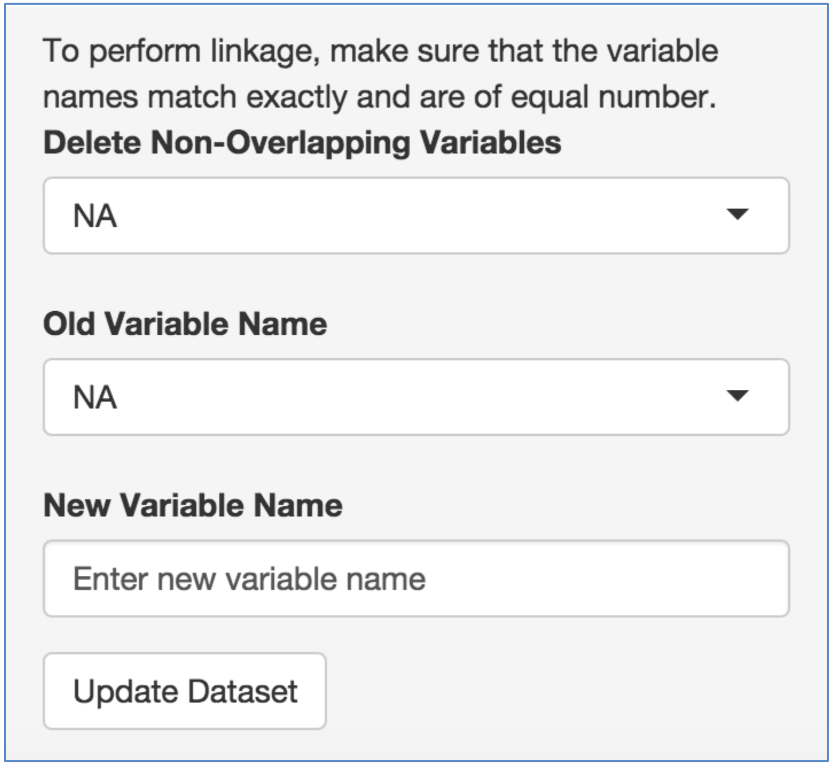
Cleaning Data

The web application offers some functionalities in terms of converting the data sets into desired format before performing record linkage.

Description

- The application requires that the two data sets have identical variable names. If this is not the case, the user can delete variables that are not found in both data sets by selecting “***Delete Non - Overlapping Variables***”.
- The user can also change variable names so they are represented the same way across the two data sets. For example, the user may change ***Old Variable Name*** “given_name” to ***New Variable Name*** “first_name”.

Application Screenshot



To perform linkage, make sure that the variable names match exactly and are of equal number.

Delete Non-Overlapping Variables

NA ▾

Old Variable Name

NA ▾

New Variable Name

Enter new variable name

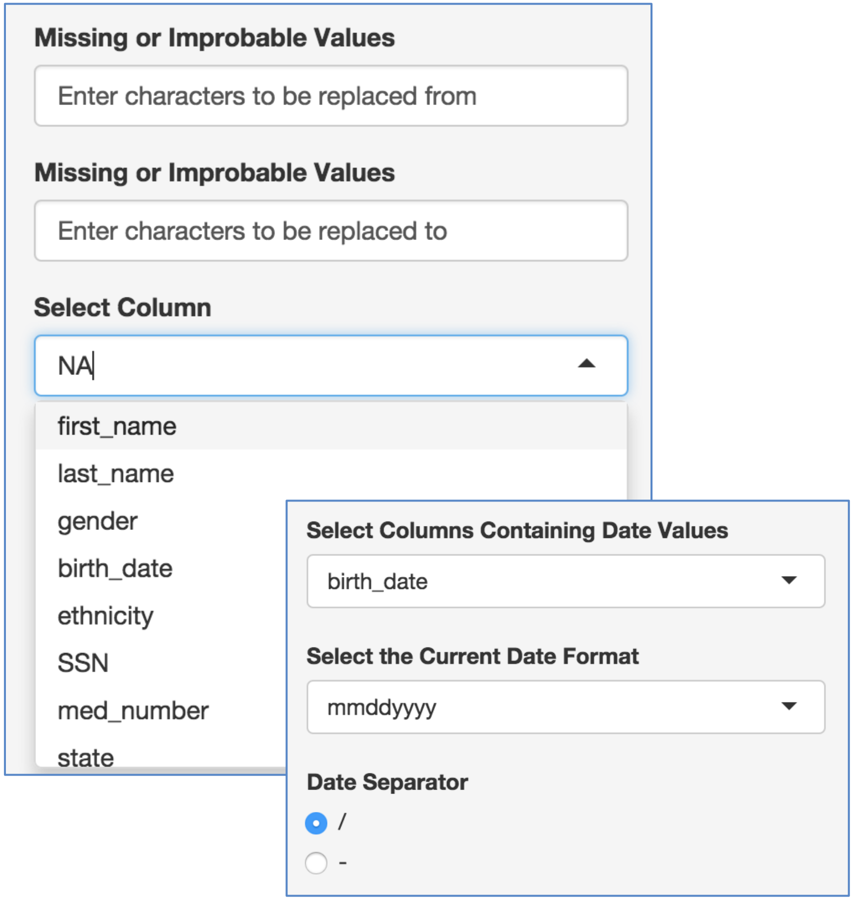
Cleaning Data

The web application includes tools to optimize how information is represented in the data sets for best linkage results.

Description

- Based on the data set, some values may be represented in undesirable ways. The user may convert all such values by specifying the values to be replaced from, what they should be replaced to, and the column they are found in.
- For example, missing values in SSN may be represented by improbable entry such as “999999”. This should be represented as a missing value, or NA, instead so that the algorithm does not take a missing value as an identical value.
- Dates can also be represented in different formats, thus we can convert them to the standard format of YYYY-MM-DD prior to linkage.
- The user can make multiple adjustments by clicking “***Update Dataset***” each time. Once the data sets are in the desired format, the user can move on to performing record linkage.

Application Screenshots



Missing or Improbable Values

Enter characters to be replaced from

Missing or Improbable Values

Enter characters to be replaced to

Select Column

NA

first_name
last_name
gender
birth_date
ethnicity
SSN
med_number
state

Select Columns Containing Date Values

birth_date

Select the Current Date Format

mmddyyyy

Date Separator

● /
○ -

Record Linkage

The application provides two methods of record linkage: the deterministic method and the probabilistic method. The user can select the desired method in the top menu.

Deterministic Linkage

- The deterministic approach matches records based on exact agreement of certain variables. Every record in data set 1 is compared to every record in data set 2 to find matches.
- The application offers two ways of choosing the parameters for linkage:
  - User can select the fields that must be matched for two records to be classified as a link;
  - User can also select a threshold for the maximum number of disagreeing variables that is allowed for two records to be classified as a link. This method will generate more links than the previous one as it does not matter which fields are a match as long as a certain number of fields do.

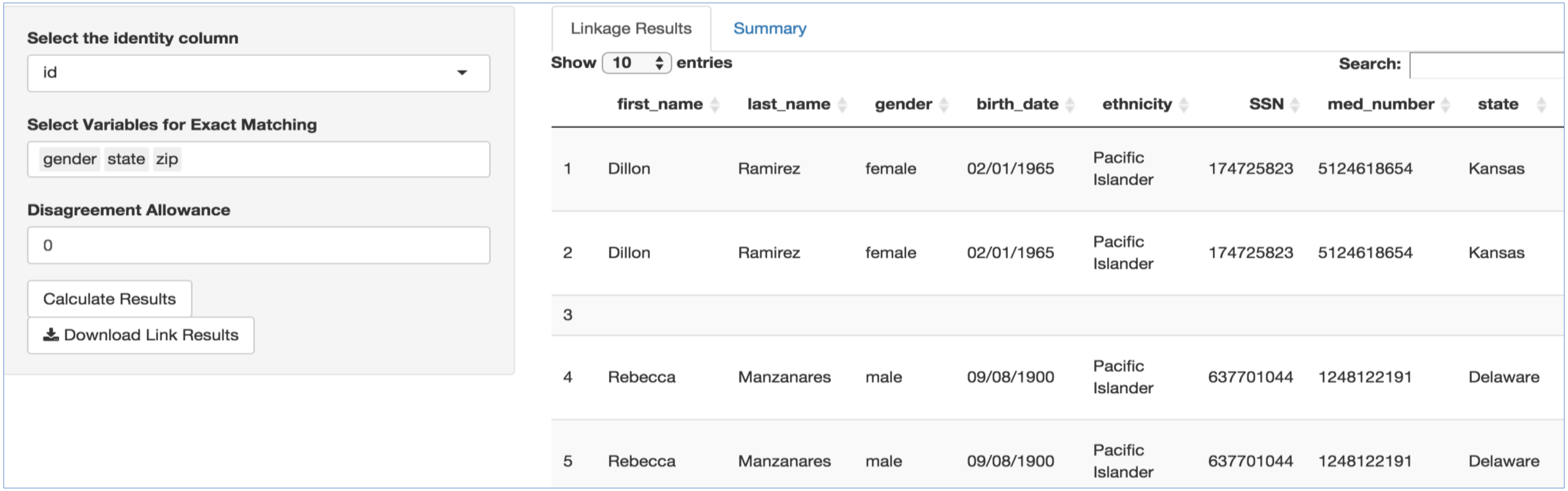
Probabilistic Linkage

- The probabilistic approach matches records based on a probabilistic weighting. The weights are calculated separately for each variable and are summed to arrive at the final probabilistic weighting for each comparison pair.
- The user is required to select appropriate thresholds, or weight cut-offs, to determine which records are classified as links.
- The user also have the option to manually adjudicate record pairs that have weights around the threshold.

Record Linkage - Deterministic

The deterministic method will link records based on exact agreements of variables.

Application Screenshot



Select the identity column: id

Select Variables for Exact Matching: gender state zip

Disagreement Allowance: 0

	first_name	last_name	gender	birth_date	ethnicity	SSN	med_number	state
1	Dillon	Ramirez	female	02/01/1965	Pacific Islander	174725823	5124618654	Kansas
2	Dillon	Ramirez	female	02/01/1965	Pacific Islander	174725823	5124618654	Kansas
3								
4	Rebecca	Manzanares	male	09/08/1900	Pacific Islander	637701044	1248122191	Delaware
5	Rebecca	Manzanares	male	09/08/1900	Pacific Islander	637701044	1248122191	Delaware

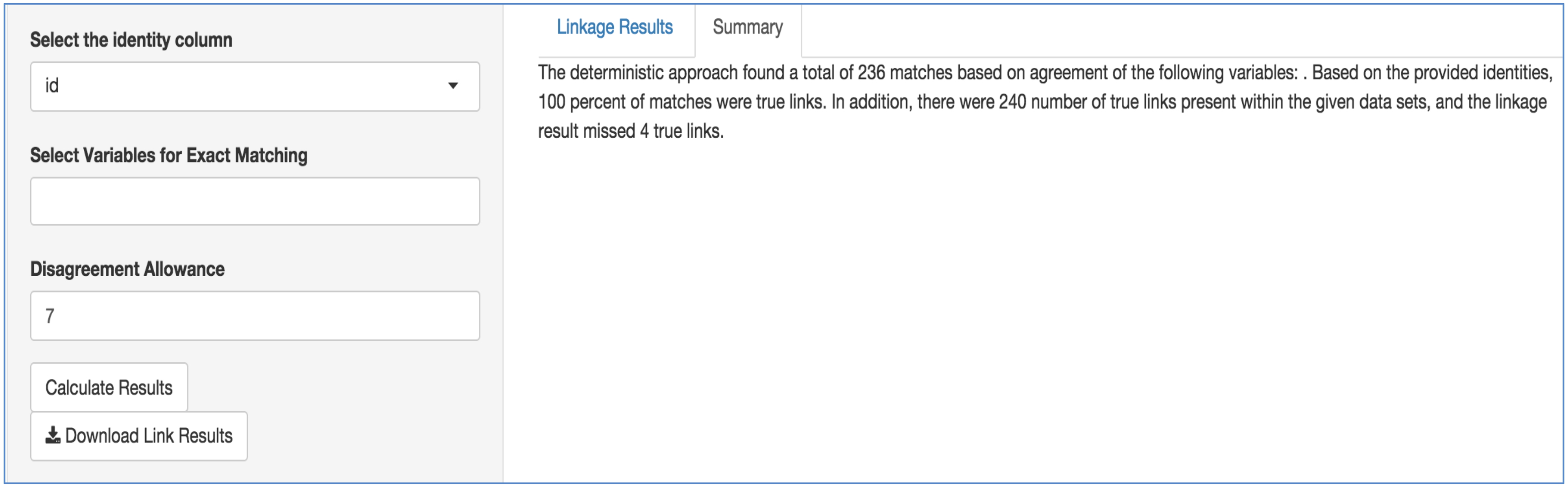
Description

- In "***Select Variables for Exact Matching***", the user may select the variables that must equal between two records for the algorithm to determine a match. If this field is left blank, all variables will be used in comparison.
- ***"Disagreement Allowance"*** is the maximum number of disagreeing variables allowed for the algorithm to return a match. For example, if this number is 3, records that have at most three non-identical variable values will be returned as a match.
- In the main panel, the user can visually review the records that have been matched. The results are displayed in match pairs.

Record Linkage - Deterministic

The deterministic method will link records based on exact agreements of variables.

Application Screenshot



The deterministic approach found a total of 236 matches based on agreement of the following variables: . Based on the provided identities, 100 percent of matches were true links. In addition, there were 240 number of true links present within the given data sets, and the linkage result missed 4 true links.

Description

- Sometimes, the data sets contain a unique identifier through which true match status can be obtained. In this case, the user can specify this identifier in "***Select the Identity Column***".
- The ***Summary*** tab provides a succinct report of the match results. If the identity column has been provided, the ***Summary*** tab also reports the accuracy of the match results.
- "***Download Link Results***" will download the linkage results as shown in the previous tab into a .csv file. The file contains the original row indices that can identify the linked records.

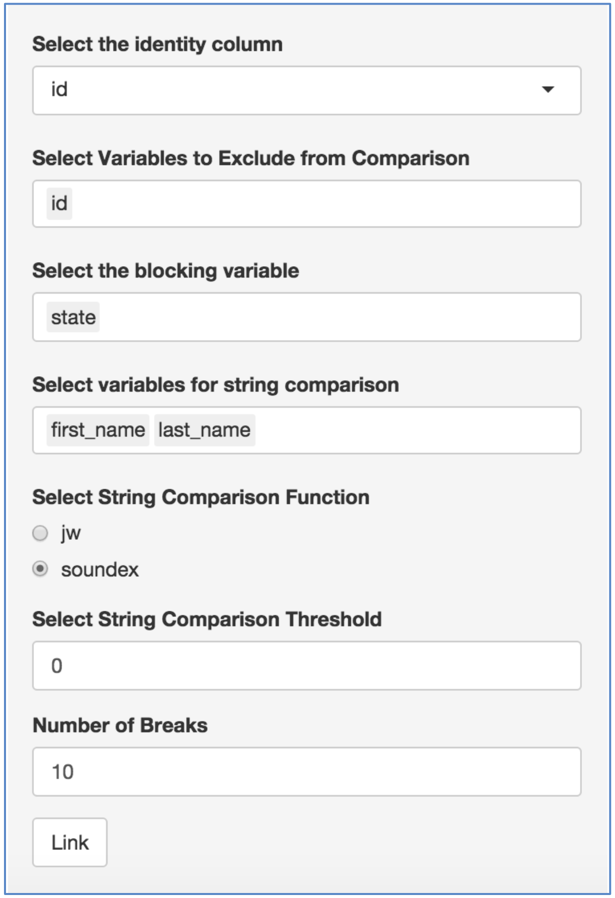
Record Linkage - Probabilistic

The probabilistic method will link records based on probability of match.

Description

- ***Select the Identity Column***: the user can specify the identity column if applicable in order to assess the performance of the probabilistic linkage algorithms.
- ***Select Variables to Exclude from Comparison****:* the user should specify variables that should be excluded from probabilistic linkage, the identity column is automatically excluded.
- ***Select the Blocking Variable***: if the blocking technique is desired, the user can specify the blocking variable here.
- ***Select Variables for String Comparison***: if the fuzzy matching techniques are desired, the user should select the appropriate variables here.
- If string comparison is used, the user can select the desired ***String Comparison Algorithm***.
- **Select String Comparison Threshold** will convert distance scores above the threshold into an exact match. The default value is 0.99.
- At this time, the user can move onto the next step by clicking ***Link***.

Application Screenshots



Select the identity column: id

Select Variables to Exclude from Comparison: id

Select the blocking variable: state

Select variables for string comparison: first_name last_name

Select String Comparison Function

○ jw

● soundex

Select String Comparison Threshold: 0

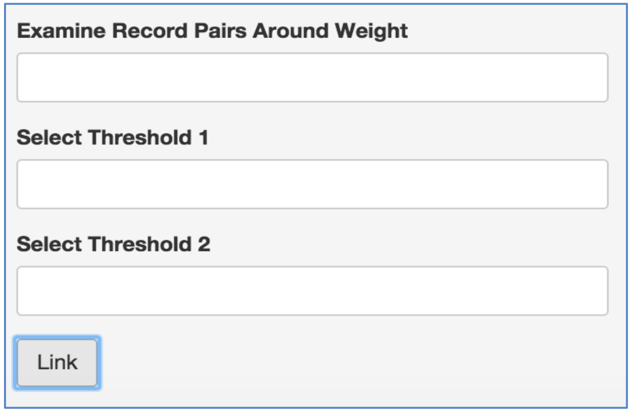
Number of Breaks: 10

Record Linkage - Probabilistic

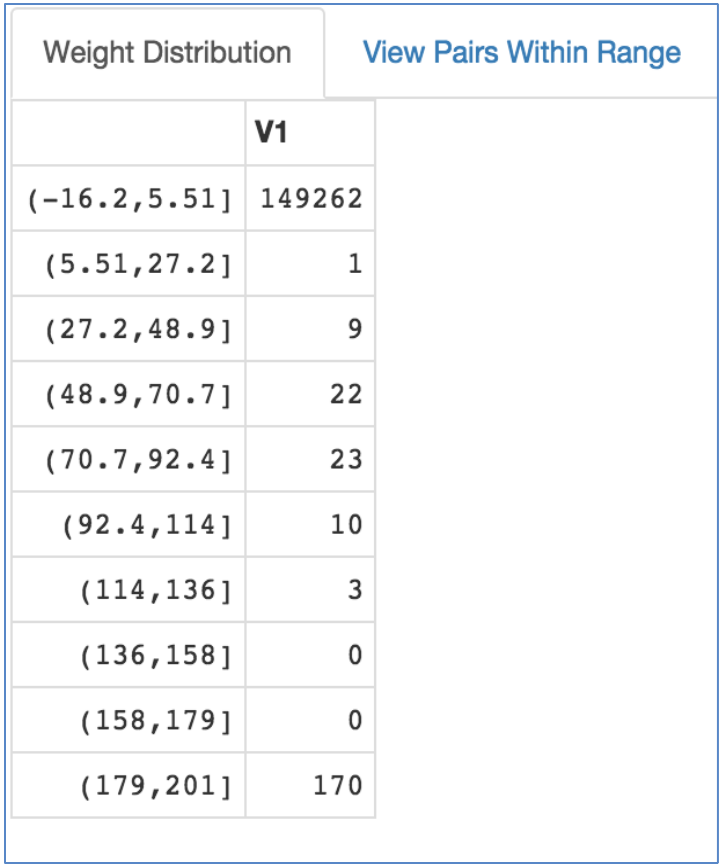
The weight distribution table gives the user a starting point for examination. Typically, the user should examine the weight brackets in the middle ranges where the number of records is low.

Description

- After clicking “***Link***”, the left panel will now show the ***Weight Distribution*** of the linkage results. The user should use this as a guide for choosing weight thresholds to examine and match. The “***Number of Breaks***” input allows the user to increase the granularity of the weight distribution table.
- After specifying a value for “***Examine Record Pairs Around Weight***”, the “***View Pairs Within Range***” tab will display record pairs around the weight threshold for manual inspection.



Application Screenshot

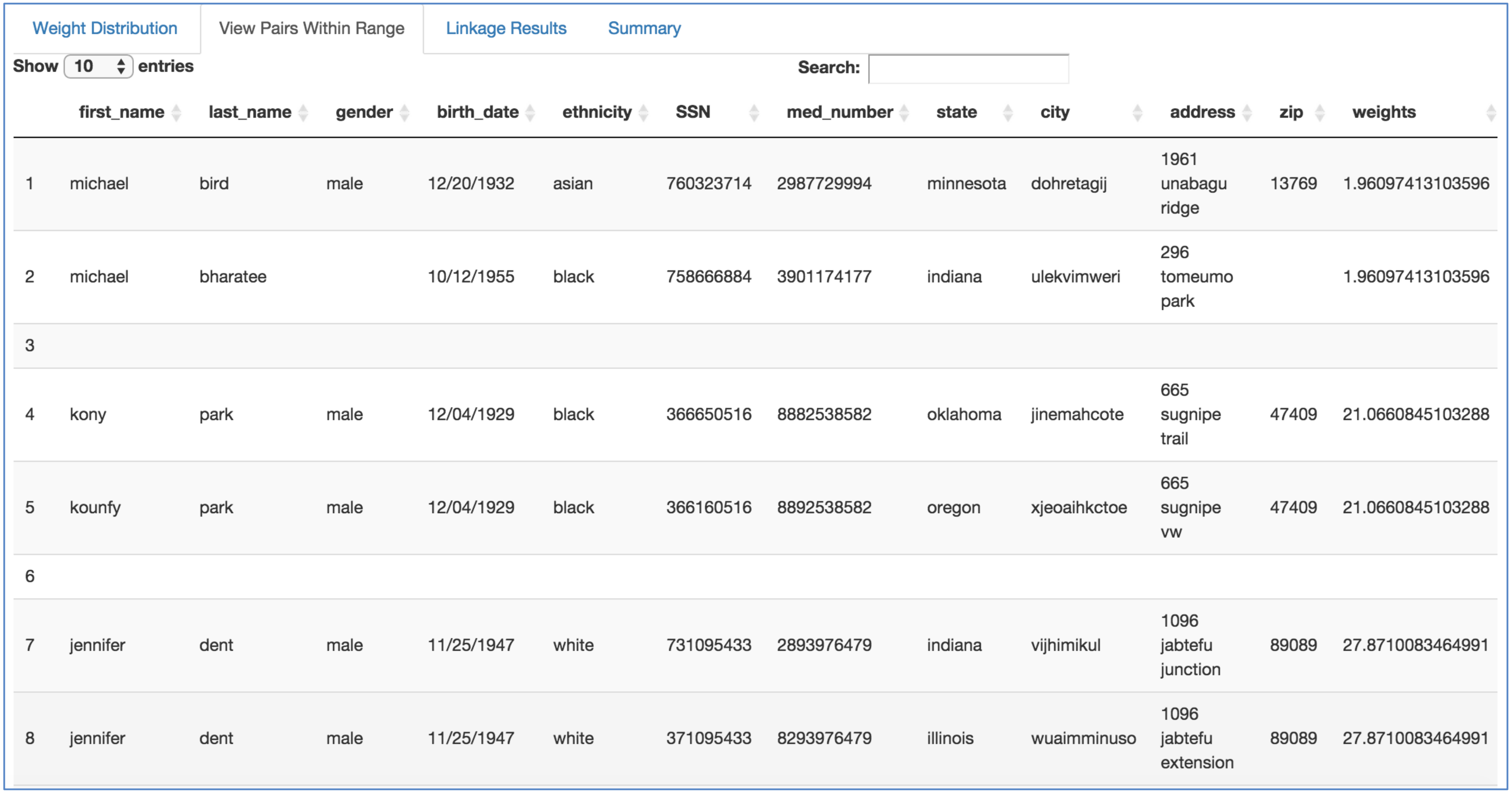


	V1
(-16.2,5.51]	149262
(5.51,27.2]	1
(27.2,48.9]	9
(48.9,70.7]	22
(70.7,92.4]	23
(92.4,114]	10
(114,136]	3
(136,158]	0
(158,179]	0
(179,201]	170

Record Linkage - Probabilistic

The user can examine multiple sets of records around a weight threshold. The user should be able to identify an approximate threshold above which record pairs are true matches, and below which are false matches.

Application Screenshot



	first_name	last_name	gender	birth_date	ethnicity	SSN	med_number	state	city	address	zip	weights
1	michael	bird	male	12/20/1932	asian	760323714	2987729994	minnesota	dohretagij	1961 unabagu ridge	13769	1.96097413103596
2	michael	bharatee		10/12/1955	black	758666884	3901174177	indiana	ulekvimweri	296 tomeumo park		1.96097413103596
3												
4	kony	park	male	12/04/1929	black	366650516	8882538582	oklahoma	jinemahcote	665 sugnipe trail	47409	21.0660845103288
5	kounfy	park	male	12/04/1929	black	366160516	8892538582	oregon	xjeoaihkctoe	665 sugnipe vw	47409	21.0660845103288
6												
7	jennifer	dent	male	11/25/1947	white	731095433	2893976479	indiana	vijhimikul	1096 jabtefu junction	89089	27.8710083464991
8	jennifer	dent	male	11/25/1947	white	371095433	8293976479	illinois	wuaimminuso	1096 jabtefu extension	89089	27.8710083464991

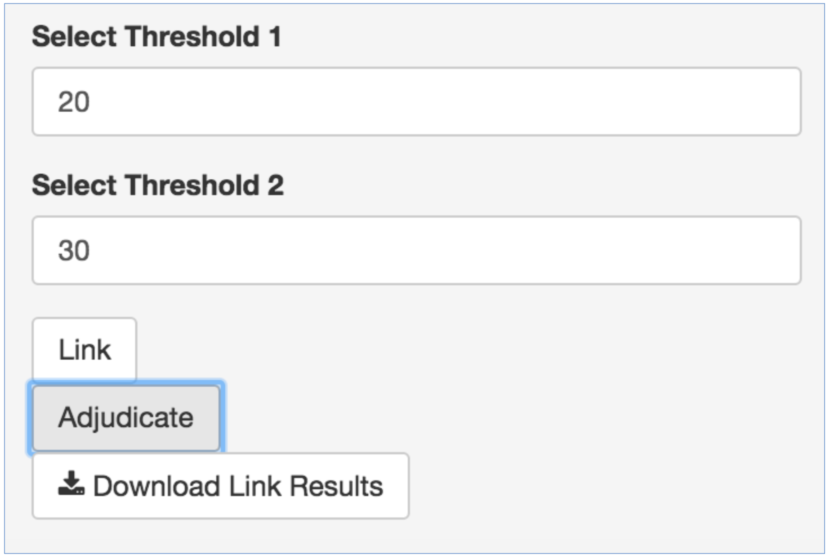
Record Linkage - Probabilistic

Threshold selection is a crucial step in the matching process; the user should select thresholds after carefully examining the record pairs while considering the number of possible pairs that can be reviewed manually

Description

- The user should first navigate the the "***Linkage Results***" tab before proceeding.
- Before classifying pairs as matches, the user must select the appropriate thresholds. ***Threshold 1*** is used to determine non-matches. If only threshold 1 is specified, all pairs above that threshold are classified as matches.
- ***Threshold 2*** is used to distinguish between matches and possible matches. If threshold 2 is provided, all pairs with weight between threshold 1 and threshold 2 are classified as possible matches.
- **Note:** threshold 2 must be higher than threshold 1.
- At this point, the button "***Adjudicate***" appears. This allows the user classify record pairs based on the threshold, and to manually adjudicate the match status of record pairs with weight between ***Threshold 1*** and ***Threshold 2***.
- A text editor will open up to facilitate the adjudication process.

Application Screenshot



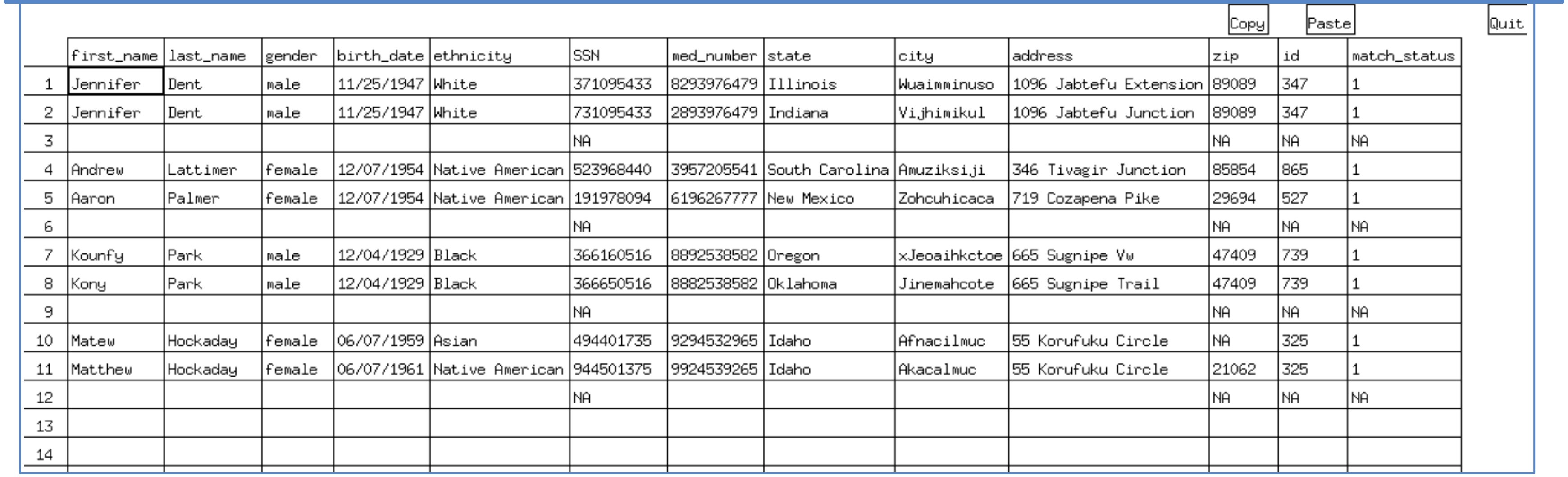
Select Threshold 1: 20

Select Threshold 2: 30

Record Linkage - Probabilistic

The application facilitates the manual review of record pairs that have a probability weighting within the two thresholds.

Application Screenshot



	first_name	last_name	gender	birth_date	ethnicity	SSN	med_number	state	city	address	zip	id	match_status
1	Jennifer	Dent	male	11/25/1947	White	371095433	8293976479	Illinois	Wuaimminuso	1096 Jabtefu Extension	89089	347	1
2	Jennifer	Dent	male	11/25/1947	White	731095433	2893976479	Indiana	Vijhimikul	1096 Jabtefu Junction	89089	347	1
3						NA					NA	NA	NA
4	Andrew	Lattimer	female	12/07/1954	Native American	523968440	3957205541	South Carolina	Amuziksiji	346 Tivagir Junction	85854	865	1
5	Aaron	Palmer	female	12/07/1954	Native American	191978094	6196267777	New Mexico	Zohcuhicaca	719 Cozapena Pike	29694	527	1
6						NA					NA	NA	NA
7	Kounfy	Park	male	12/04/1929	Black	366160516	8892538582	Oregon	xJeoaihkctoe	665 Sugnipe Vw	47409	739	1
8	Kony	Park	male	12/04/1929	Black	366650516	8882538582	Oklahoma	Jinemahcote	665 Sugnipe Trail	47409	739	1
9						NA					NA	NA	NA
10	Matew	Hockaday	female	06/07/1959	Asian	494401735	9294532965	Idaho	Afnacilmuc	55 Korufuku Circle	NA	325	1
11	Matthew	Hockaday	female	06/07/1961	Native American	944501375	9924539265	Idaho	Akacalmuc	55 Korufuku Circle	21062	325	1
12						NA					NA	NA	NA
13													
14													

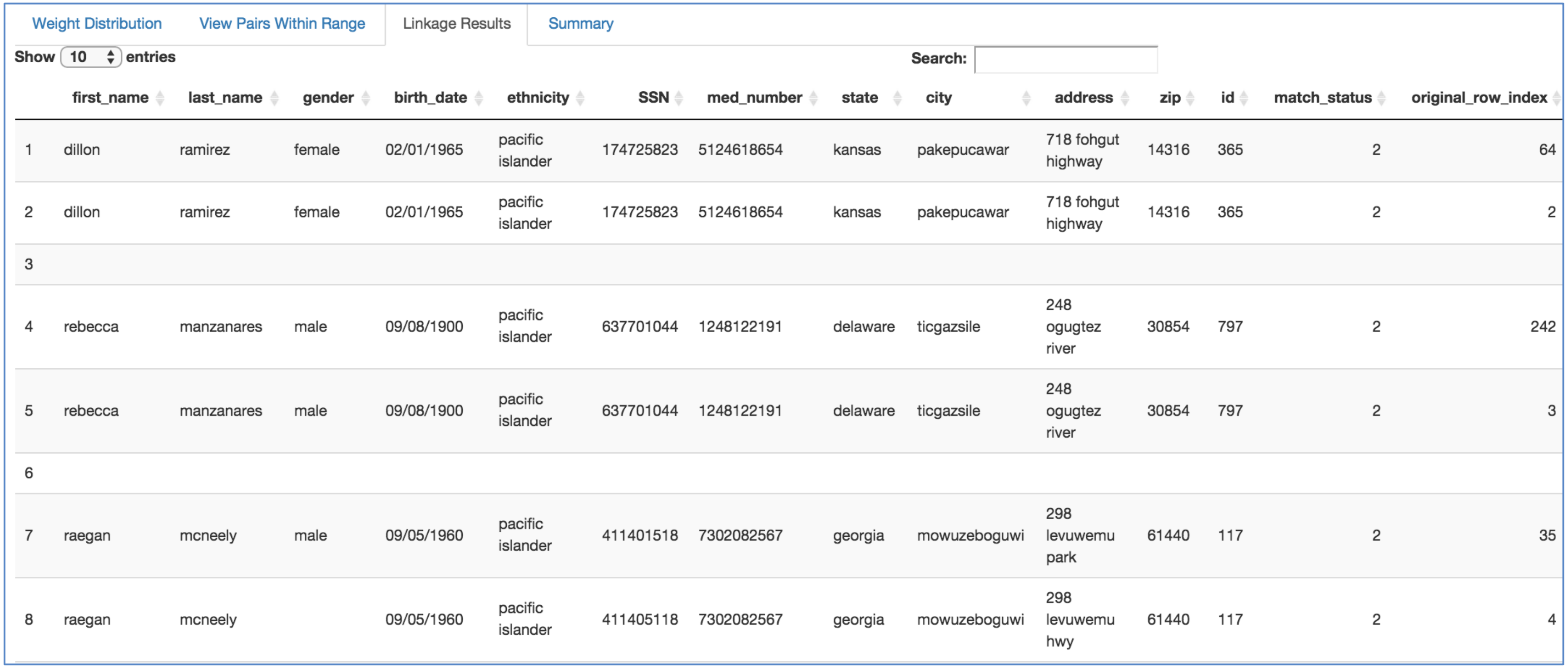
Description

- A **match_status** of **1** means a possible match, a **match_status** of **2** means a match, and a **match_status** of **0** means a non-match. The user can manually change the **match_status** in the text editor based on the visual inspection of pairs.
- The ability to adjudicate possible matches is beneficial because there will rarely be a weight threshold that clearly separates all true links from true non-links. The user can manually inspect pairs within the range where the most uncertainly lies, which will improve the accuracy of the linkage.

Record Linkage - Probabilistic

Once the record pairs have been adjudicated, the record pairs with a match status of 2 will be displayed on the "Linkage Results" tab.

Application Screenshot



Show 10 entries

Search:

	first_name	last_name	gender	birth_date	ethnicity	SSN	med_number	state	city	address	zip	id	match_status	original_row_index
1	dillon	ramirez	female	02/01/1965	pacific islander	174725823	5124618654	kansas	pakepucawar	718 fohgut highway	14316	365	2	64
2	dillon	ramirez	female	02/01/1965	pacific islander	174725823	5124618654	kansas	pakepucawar	718 fohgut highway	14316	365	2	2
3														
4	rebecca	manzanares	male	09/08/1900	pacific islander	637701044	1248122191	delaware	ticgazsile	248 ogugtez river	30854	797	2	242
5	rebecca	manzanares	male	09/08/1900	pacific islander	637701044	1248122191	delaware	ticgazsile	248 ogugtez river	30854	797	2	3
6														
7	raegan	mcneely	male	09/05/1960	pacific islander	411401518	7302082567	georgia	mowuzeboguwi	298 levuwemu park	61440	117	2	35
8	raegan	mcneely		09/05/1960	pacific islander	411405118	7302082567	georgia	mowuzeboguwi	298 levuwemu hwy	61440	117	2	4

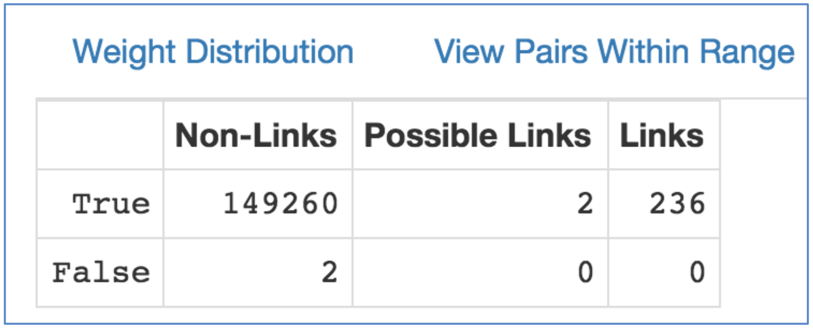
Record Linkage - Probabilistic

This concludes the record linkage tutorial.

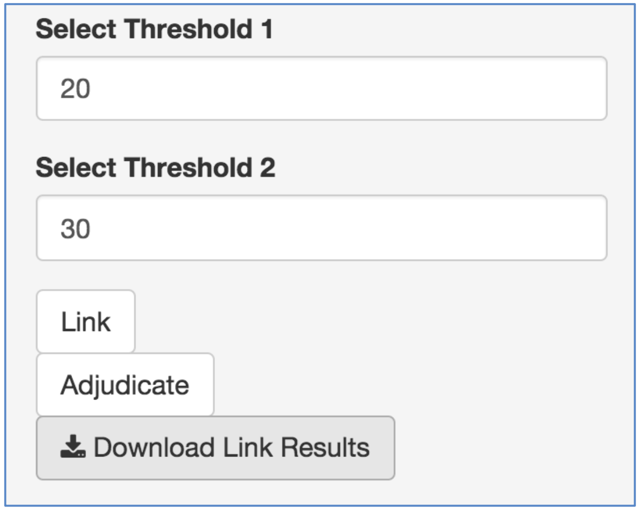
Description

- Finally, the "***Summary***" tab shows a confusion matrix based on the result of linkage. If an identity column has been provided, the linkage result of the algorithm is compared to the true match status of the data sets.
- User can download the linkage results by clicking "***Download Link Results***"; this will download the output as seen in the "***Linkage Results***" tab into a .csv file format.

Application Screenshot



	Non-Links	Possible Links	Links
True	149260	2	236
False	2	0	0



Select Threshold 1

20

Select Threshold 2

30