"Record Linking 101"

Combining Files without a Common Identifier

SAMHSA Integrated Database Project

Washington, DC

December 5, 2003

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Treatment
Center for Mental Health Services
www.samhsa.gov

SAMHSA

SAMHSA Integrated Data Project

- Center for Mental Health Services
- Center for Substance Abuse Services
- Contractors
  - The MEDSTAT Group, Inc.
  - National Association of State Mental Health Program Directors Research Institute (NASMHPD)
  - National Association of State Alcohol and Drug Abuse Directors, Inc. (NASADAD)

Introductions

Agenda

- Match-merge linking
- Probabilistic and deterministic linking
  - Identifying variables
  - Comparisons
- IDB probabilistic record linking
  - Calculating weights
  - Determining Links

Match-Merge Methods

- Familiar concept in data processing
- Uses keys (identical variables) on each file
  - Records are combined (merged) when the respective keys on each file are the same (match)
  - Records are not combined when the keys are different
- Keys can be simple or complex

Match-Merging Related Files

- Match-merging files from a single authority is usually very accurate
- Agency specific identifiers are often used as file keys
- Identifiers from a single authority are generally reliable – errors are rare

Problems with Match-Merge Methods

- Problems when keys are incorrect
- May occur because of omissions and errors
- Two outcomes:
  - Records do not match when they should
  - Records match when they should not

Match-Merging Unrelated Files

- There are often problems merging files from separate agencies – even with a common identifier (i.e., SSN)
- Omissions and errors are more prevalent with identifiers that are not specific to an agency
- Manual review of SSN match-merges reveal many errors
  - – records that should be linked but are not
  - – incorrect links

A Linking Tangent – Background

File *A*	
ID	**Var1**
1	E
2	F
3	G

File *B*	
ID	**Var2**
1	X
3	Y
4	Z

- Two simple files:
  - File A
  - File B
  - Both with a variable “ID” as their key

A Match-Merge

- Match-merge on ID

Match-merge Results		
ID	Var1	Var2
1	E	X
3	G	Y

Probabilistic and Deterministic Linking

- Related techniques – overcome limitations of match merging
- Makes linking possible, even with
  - Missing information
  - Errors in data
- Uses multiple criteria
- More work than match-merging

Terminology

- ***Record-pair*** – a combination of records from two files such that one half of each pair is derived from the first file and the remainder is from the second file

A Conceptual Linking Example

File *A*	
ID	**Var1**
1	E
2	F
3	G

File *B*	
ID	**Var2**
1	X
3	Y
4	Z

- Two simple files:
  - – File A
  - – File B
  - – Both with a variable “ID” as their key

More Terminology

- ***Links*** – record-pairs that represent the same person or entity (a.k.a. linked). In match merging, the "matched" records are links
- ***Non-links*** – record-pairs that do not represent the same person or entity

More Terminology

- ***Joined Records*** – a collection of record-pairs: all the joined records (the sum of all links and non-links)
- ***Decision Space*** – the complete set of record-pairs that are evaluated to determine links

More Terminology

- ***Cartesian Product*** – a set of joined records constructed from two files such that each record from the first file is paired with every record from the second file, as depicted below

		File 2		
	Record	**X**	**Y**	**Z**
File 1	**L**	*L-X*	*L-Y*	*L-Z*
	M	*M-X*	*M-Y*	*M-Z*
	N	*N-X*	*N-Y*	*N-Z*

A Record-Pairing

- An alternate method to link files
  - **Cartesian product**
  - Evaluate keys: A.ID = B.ID
  - Keep pairs where the IDs are the same

Combined Files			
A.ID	**B.ID**	**Var1**	**Var2**
1	1	E	X
1	3	E	Y
1	4	E	Z
2	1	F	X
2	3	F	Y
2	4	F	Z
3	1	G	X
3	3	G	Y
3	4	G	Z

A Record-Pairing

- An alternate method to link files
  - Cartesian product
  - **Evaluate keys: A.ID = B.ID**
  - Keep pairs where the IDs are the same

Combined Files			
A.ID	**B.ID**	**Var1**	**Var2**
1	1	E	X
1	3	E	Y
1	4	E	Z
2	1	F	X
2	3	F	Y
2	4	F	Z
3	1	G	X
3	3	G	Y
3	4	G	Z

A Record-Pairing

- An alternate method to link files
  - Cartesian product
  - Evaluate keys: A.ID = B.ID
  - **Keep pairs where the IDs are the same**

Combined Files			
A.ID	**B.ID**	**Var1**	**Var2**
1	1	E	X
3	3	G	Y

Extending the Conceptual Example

- Decision rule: a function of A.ID and B.ID
  - Keep (or true or 1) if A.ID=B.ID
  - Remove (or false or 0) if A.ID≠B.ID
- With “match-merging,” ID is a single variable
- Extend concept for deterministic and probabilistic linking
  - ID is a collection of variables

Identifying Variables

- Information that can identify a person
  - Directly
  - Indirectly
- Linking requires identifying variables
- Used in “decision rules” to determine links

More Terminology

- ***Identifying Variables*** – information that can be used to identify a person. This includes direct identifiers such as name and indirect identifiers such as date of birth and race

Examples of Identifiers

- ID numbers
- Name
- Gender
- Address

Weak and Strong Identifiers

- Some identifiers are “weak”
  - By themselves, they do not directly identify a person
  - They must be used with other information to work as an identifier
- Other identifiers are “strong”
  - They can be used to directly identify a person

Useful identifiers

- As a general rule, strong identifiers are better than weak ones
- But strong is not the same as good or useful
- A useful identifier is available on all files
- An identifier found on a single file cannot be used for comparisons

Comparing Identifiers

- Record-pairs are evaluated with comparisons
- Compare each and every set of identifying variables
- Look for
  - Agreement
  - Disagreement

More Terminology

- ***Comparison Variables*** – identifying variables used in comparing the two halves of a record-pair
- ***Comparisons*** – the result of equating comparison variables from a record-pair. Record-pairs typically contain a mixture of comparisons in both agreement and disagreement. Comparisons are part of the process of evaluating record-pairs to determine links

Types of Comparisons

- Comparisons can be dichotomous or continuous
- Dichotomous – true or false
  - No gray areas, unforgiving of errors and mistakes
  - Example: gender – either the same or different
- Continuous – a continuum
  - Indicates the degree of agreement
  - Forgiving of mistakes/errors

More Terminology

- ***Dichotomous Comparisons*** – comparisons which evaluate as either true or false – agreement or disagreement
- ***Continuous Comparisons*** – comparisons resulting in a numeric score that reflects partial agreement ranging from complete disagreement to complete agreement.

Comparing Names

- Misspellings often occur with names
  - Anderson and Andersen
  - Whalen and Whelan
- Phonetic equivalents sometimes used to account for differences in spelling
  - Russell Soundex
  - New York State Identification Information System (NYSIIS)

Phonetic Names

Name	Soundex	NYSIIS
Whalen	W45	WALAN
Whelan	W45	WALAN
Graber	G616	GRABAR
Gerber	G616	GARBAR
Aijian	A25	AJAN
Askam	A25	ASCAN
Haskens	H252	HASCAN
Haskant	H253	HASCAN

Approximate String Matching

- A continuous comparison
- Calculates degree of agreement
  - Additions/Deletions/Changes
  - Percentage based on name lengths
- WHALEN & WHELAN
  2 changes → 66.7% agreement
- AIJIAN & ASKAM
  3 changes + 1 addition (or deletion)
  → 26.7% agreement

Weights

- A comparison variable’s overall importance in determining links is quantified as a *comparison weight*
- Weights signify the relative importance of variables
  - Higher points for more important information
  - Lower points for less important information
- Links made for records pairs with a point total over a predefined threshold

More Terminology

- ***Weights*** – numeric values that indicate the overall importance of a comparison relative to other comparisons. The discriminating power of each comparison variable – its importance in determining links – is expressed as a weight.

Types of weights

- Deterministic weights
  - Arbitrarily determined before the linking process
- Probabilistic weights
  - Calculated from the relative probabilities of agreement (or disagreement)

  Weight = $\log_2$ [Pr(agree | link) / Pr(agree | non-link)]
- Agreement and disagreement weights
  - For each comparison variable
  - Not symmetrical

Points for Deterministic Agreements

- Points for “agreements” should reflect the relative importance of that agreement
  - Higher points for more important information (i.e., SSN)
  - Lower points for less important information (i.e., gender)
- Negative points for disagreements are also possible, but not often employed

A Deterministic Linking Example

- Comparison Points
  - 20 points for a complete SSN agreement, or 10 points for agreement on the last four digits of the SSNs
  - 15 points for an agreement on last name
  - 8 points for an agreement on first name
  - 5 points for a date of birth agreement
  - 1 point for a gender agreement, or

    10 points if gender does not agree
- Linking Threshold: 25 or more points

More Terminology

- ***Score*** & ***Scoring*** – the sum of the products of all the comparisons with the associated weights. The score is used to evaluate record-pairs and determine links and non-links. When weights are applied and summed into scores, the scores for record-pairs that should be linked are generally higher than scores for the record-pairs that should not be linked.

Deterministic Linking Problems

Setting Points and Thresholds

- Appropriate points for agreement
- Effective point thresholds
- How should it be done?
  - Trial and Error?
  - Intuition?
- *Recall: point values should reflect the relative importance of an agreement (or disagreement)*

Deterministic Linking (continued)

- A clear improvement over match merging
- Record linkage is possible even with errors, or missing information
  - When SSN is not available
  - Errors in SSN do not necessarily cause incorrect links

Probabilistic Linking

- Similar to deterministic linking
  - Multiple criteria/comparisons
  - Scores to determine links
- Differences from deterministic linking
  - Points and scoring not known beforehand
  - Commonly uses disagreements as well as agreements
  - More complex

Determining Probabilistic Weights

- Which comes first, weights or links?
  - Weights needed to divide record-pairs into links and non-links
  - Link/non-link division necessary to calculate weights
- Solutions
  - Sample files
  - Iterative process

More Terminology

- ***Scaling*** – adjusting the weight for a comparison variable to reflect the relative frequency of a specific value.

Scaling Weights

- Some weights are scaled – adjusted up or down for specific values
- Scaling factors inversely related to the relative frequency of the identifier’s value
- Not an issue for evenly distributed identifiers (i.e., SSN and gender)
- Used for identifiers not evenly distributed (i.e., last-name and race)

Other Linking Issues

- File size
- Blocking
- Scores

File Size Issues

- Recall that the initial decision space for linking two files is the Cartesian product of those two files
- As file size increases
  - The decision space increases exponentially
  - The proportion of potential links decreases

			Potential Links	
File *A*	File *B*	Decision Space	Number	Proportion
100	100	10,000	100	1.000%
1,000	1,000			
10,000	10,000			
100,000	100,000			

File Size Issues

- Recall that the initial decision space for linking two files is the Cartesian product of those two files
- As file size increases
  - The decision space increases exponentially
  - The proportion of potential links decreases

			Potential Links	
File *A*	**File *B***	**Decision Space**	**Number**	**Proportion**
100	100	10,000	100	1.000%
1,000	1,000	1,000,000	1,000	0.100%
10,000	10,000	100,000,000	10,000	0.010%
100,000	100,000	10,000,000,000	100,000	0.001%

Blocking

- Blocking is the process of creating record-pairs only when there is some evidence for linking the two records
- Blocking decreases the decision space, reducing the number of comparisons necessary
  - Eliminates record-pairs with no linking evidence
  - Results in more efficient search for links
- The importance of blocking increases as the size of files increase

More Terminology

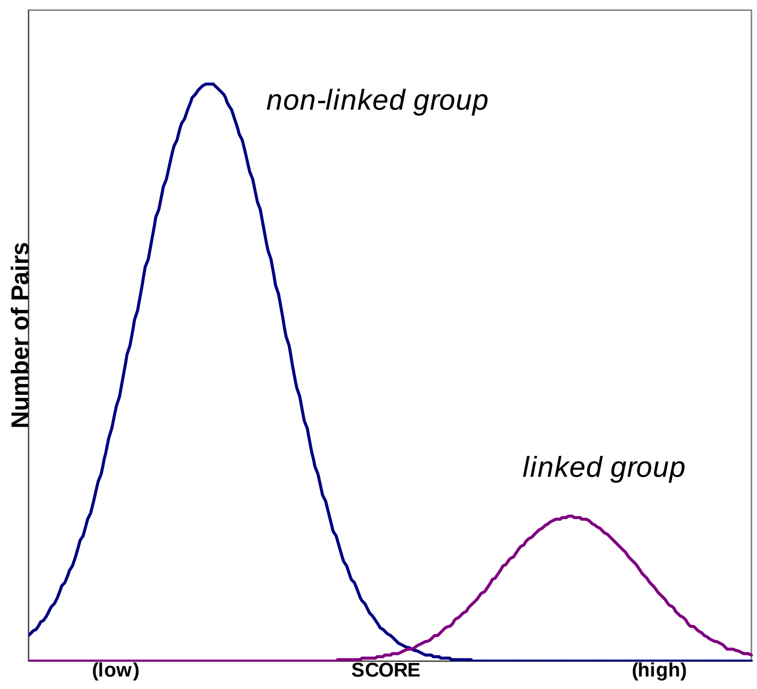
- ***Blocking*** – a technique to limit the decision space to a manageable size without eliminating potential links

Linking Scores

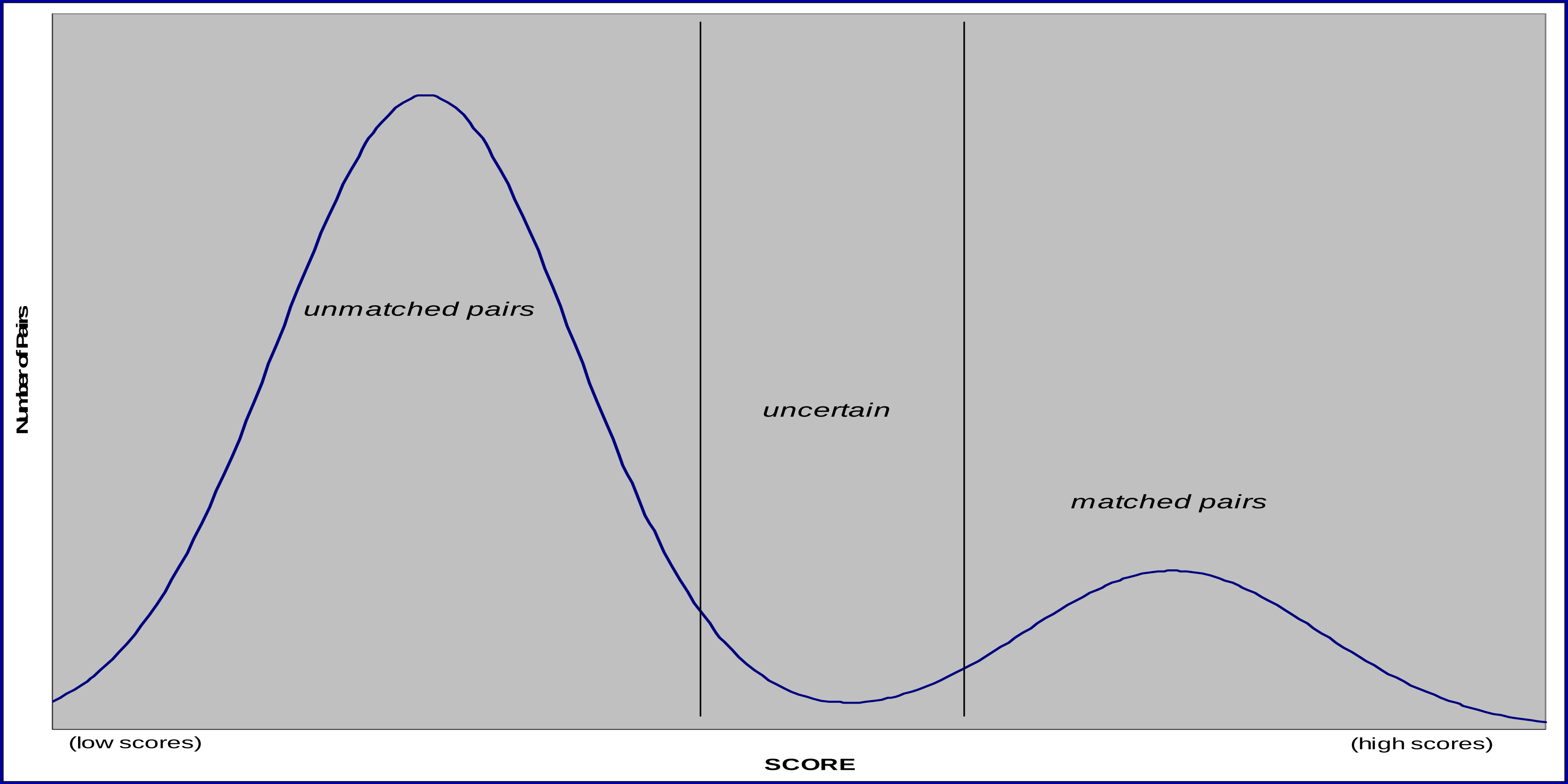
- Scores are the sum of all comparisons
  - Agreements
  - Disagreements
- Combines comparisons, weights, and scaling factors
- For each comparison variable
  - Agreement | disagreement weight
  - Plus any scaling factor for the variable’s value
  - Multiplied by the comparison result

Record-Pair Scores

- Scores for record-pairs will vary
- Scores for links are generally higher than scores for non-links



More Record-Pairs and Scores



More Terminology (continued)

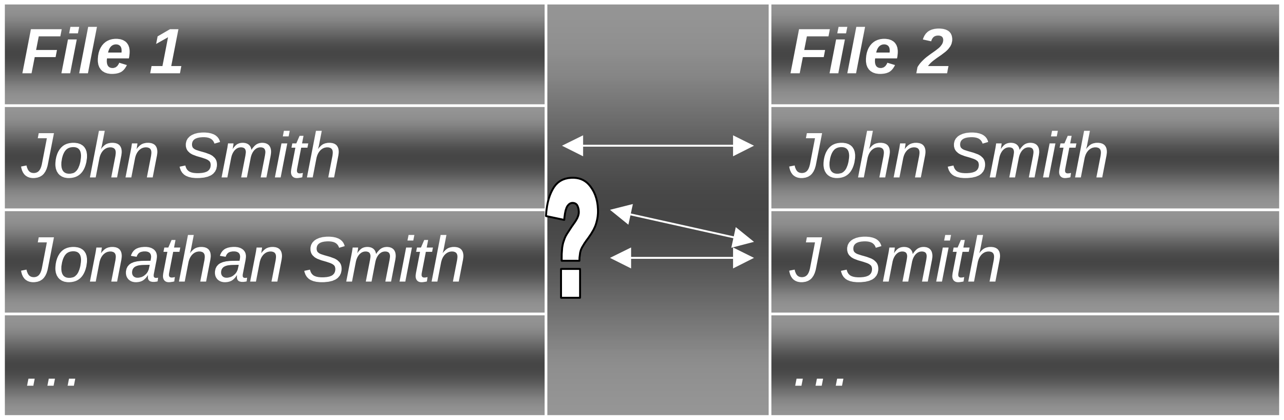
- ***Decision Groups*** – the division of the decision space into groups based on scores for the purpose of deciding which records should be linked. Record-pairs can be classified as links, non-links, and uncertain pairs.
- ***Uncertain pairs***– record-pairs for which a link or non-link determination cannot be made.

IDB Probabilistic Record Linking

- Linking
  - – Concatenating Data
  - – Scaling Factors
  - – Blocking, Joining, and Comparing
  - – Initial Links – Deterministic
  - – Probabilistic Iterations
- Manual review
- Mapping of IDs

Duplicated Client Records

- Accurate linking assumes at least one source of unduplicated data
- Duplication creates ambiguous results



File 1		File 2
John Smith	↔	John Smith
Jonathan Smith	?	J Smith
…		…

Linking Data with Duplicate Records

- “Typical” approach
  - Unduplicate first file
  - Link second file to the unduplicated records from the first file
  - Unduplicate any records from the second file not linked to the first file
- Unduplicating is similar to Linking
  - Same procedures and evaluation criteria
  - Unduplicating a file = Linking a file to itself

Unduplicating – Cartesian Product

Record		File 1			File 2		
		L	M	N	X	Y	Z
File 1	L	L-L	L-M	L-N	L-X	L-Y	L-Z
	M	M-L	M-M	M-N	M-X	M-Y	M-Z
	N	N-L	N-M	N-N	N-X	N-Y	N-Z
File 2	X	X-L	X-M	X-N	X-X	X-Y	X-Z
	Y	Y-L	Y-M	Y-N	Y-X	Y-Y	Y-Z
	Z	Z-L	Z-M	Z-N	Z-X	Z-Y	Z-Z

Unduplicating – Decision Space

Record		File 1			File 2		
		L	M	N	X	Y	Z
File 1	L		L-M	L-N	L-X	L-Y	L-Z
	M			M-N	M-X	M-Y	M-Z
	N				N-X	N-Y	N-Z
File 2	X					X-Y	X-Z
	Y						Y-Z
	Z						

Concatenating Data

- Concatenate all data and unduplicate/link
- Combines steps of unduplicating data and linking files
  - – Reduces the number of processing steps
  - – Less set-up time
  - – Saves review time
- Works with any number of data sources

Scaling Factors

- Recognizes that agreements on uncommon values are more important than agreements on common values
- Associated with specific values of a variable
  - One scaling factor for the Last Name “Whalen”
  - Separate scaling factor for “Smith”
- Inversely related to a values relative frequency

Scaled Variables

- First Name (NYSIIS phonetic)
- Middle Initial
- Last Name (NYSIIS phonetic)
- Birth Year
- Race
- ZIP Code

Blocking and Joining

- Creates the decision space for linking
- Subset of Cartesian product of the concatenated data
  - “Upper” triangle
  - Some evidence for linking the joined pair

Blocking

- SSN agreement
- DOB agreement plus agreement on NYSIIS phonetic of last name
- DOB and gender agreement plus agreement on NYSIIS phonetic of first name
- Gender agreement plus agreement on NYSIIS phonetic of first and last names

Comparisons

- On identifying variables
- Made once – at start of the process
  - Time consuming / resource intensive
  - Results saved for later iterations
- Mixture of dichotomous and continuous comparisons

Comparisons of Names and IDs

- Continuous comparisons
  - SSN – approximate string matching
  - Medicaid ID – approximate string matching
  - First Name – approximate string matching
  - Last Name – approximate string matching
- Dichotomous comparisons
  - Middle Initial

Comparisons of Demographic Data

- Date of Birth – continuous comparison
  - If two of the three components agreed (Year, Month, Day)
  - Based on the days difference
- Race – dichotomous comparison
- Gender – dichotomous comparison
- ZIP Code – continuous comparison
  - Based on the distance between centroids of the ZIP Codes

Initial Links – Deterministic

- First link / non-link determination
- Used to develop the first set of probabilistic weights and thresholds

Deterministic Criteria

- SSN agreement, Medicaid ID agreement, DOB agreement, and Gender agreement
- SSN agreement, DOB agreement, Gender agreement and one of the following:
  - At least 80% agreement for First Name
  - At least 90% agreement for Last Name
  - Agreement on Middle Initial
- Medicaid ID agreement, DOB agreement, Gender agreement and one of the following:
  - At least 80% agreement for First Name
  - At least 90% agreement for Last Name
  - Agreement on Middle Initial

More Deterministic Criteria

- At least 80% agreement for first name, At least 90% agreement for last name, DOB agreement, gender agreement and one of the following:
  - ZIP Code agreement
  - Race agreement
- At least 80% agreement for first name, at least 90% agreement for last name, DOB agreement, and at least 90% agreement for SSN or Medicaid ID
- At least 80% agreement for first name, at least 90% agreement for last name, DOB agreement, and agreement on middle initial

Probabilistic Iterations

- Classify record-pairs as links or non-links
  - – First iteration – deterministic criteria
  - – Following iterations – probabilistic scores and thresholds
- Calculate new weights
  - – Agreement and disagreement weights
- Compute scores
- Determine thresholds for classifying links, non-links, and uncertain record-pairs

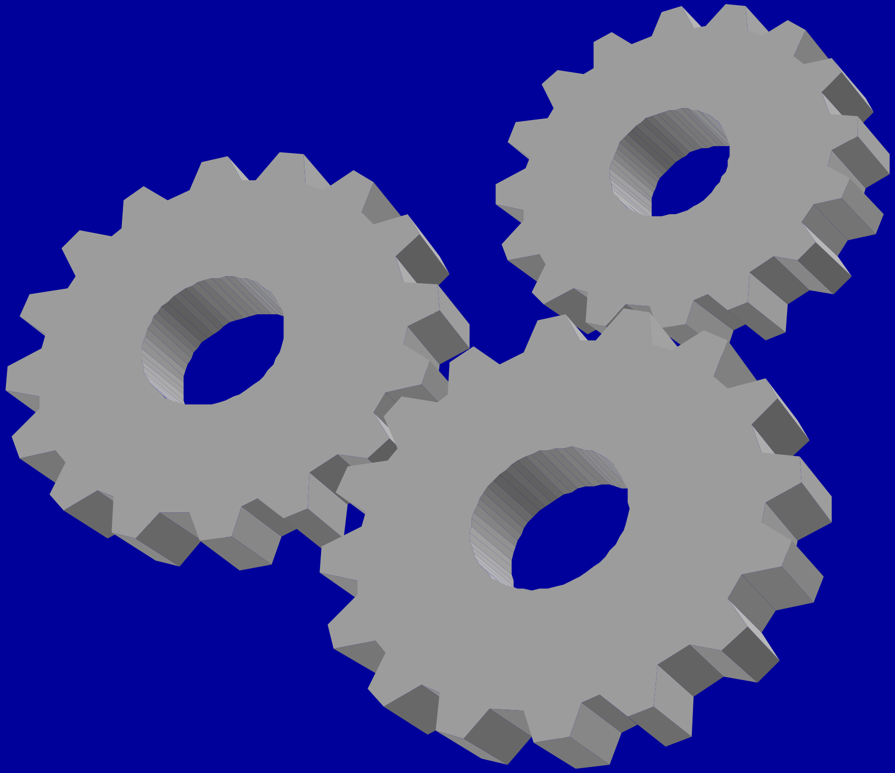
Manual review

- Review of uncertain record-pairs
- Print from the final iteration
  - Uncertain record-pairs
  - Link record-pairs that might be twins

Mapping of IDs

- Gathers all record-pairs link
  - Automatic links from the iterations
  - Links from the manual review
- Assigns synthetic ID for IDB
- Each new ID is associated with one or more IDs from the source data

Mapping Links for Data Integration



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Treatment
Center for Mental Health Services
www.samhsa.gov

SAMHSA

General Comparison of Linking

- Based on analysis of two states
- Relative to overlap from Probabilistic Population Estimate (Pandiani & Banks)
- Links found
  - Probabilistic linking: 80-86%
  - Match merge: 51-72%
  - Deterministic links: 59-76%

Conclusion/Discussion