Assessing the Linked Data Quality

Issues to Consider Regarding Maintenance of Data Quality

Creep
Inconsistencies and Precedence
Historical Memory and Metadata

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine		Sampson	0.95
113	EHDI	Kathy		Sampson	0.95

- Birth Defects Registry contributes an individual,Catherine A. Sampson

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	**0.90**

- Link the Electronic Birth Certificate
  - Name is Catherine A. Simpson
  - Are these the same person?
  - Perform probabilistic match
    - Require .90 probability of a match to conclude two similar records are the same
    - Probability is .90: We conclude they're the same person

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	0.90
113	EHDI	Kathy	A	Simpson	**0.90**

- Link Newborn Hearing Data
  - Is Kathy A. Simpson the same person?
  - Perform probabilistic match (require .90)
    - p=.90 that it's the same as Catherine A Simpson
    - Probability is .90, we conclude they're the same person

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	0.90
113	EHDI	Kathy	A	Simpson	**0.81**

- If we compare to Catherine A. Sampson
  - P_{Match}=.81
  - Conclude they are NOT the same individual
  - Would not assign same ID
- Which is correct?

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	0.90
113	EHDI	Kathy	A	Simpson	**0.81**

- $p_{Match}=\alpha$ is the minimal prob required to conclude that two records belong to the same individual
- Even if $p_{Match}<\alpha$, two records can be linked through a sequential pairing of statistically intermediate records
  - A matched with B, B matched with C, C matched with D…

Linkage Creep

ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	0.90
113	EHDI	Kathy	A	Simpson	**0.81**

- A sequential series of paired records (A and B, B and C, C and D, etc.) each have $P_{Match}=\alpha$
- The probability of two records at opposite ends of this sequential pairing belonging to the same person will be $<\alpha$, and possibly $<<\alpha$

Linkage Creep

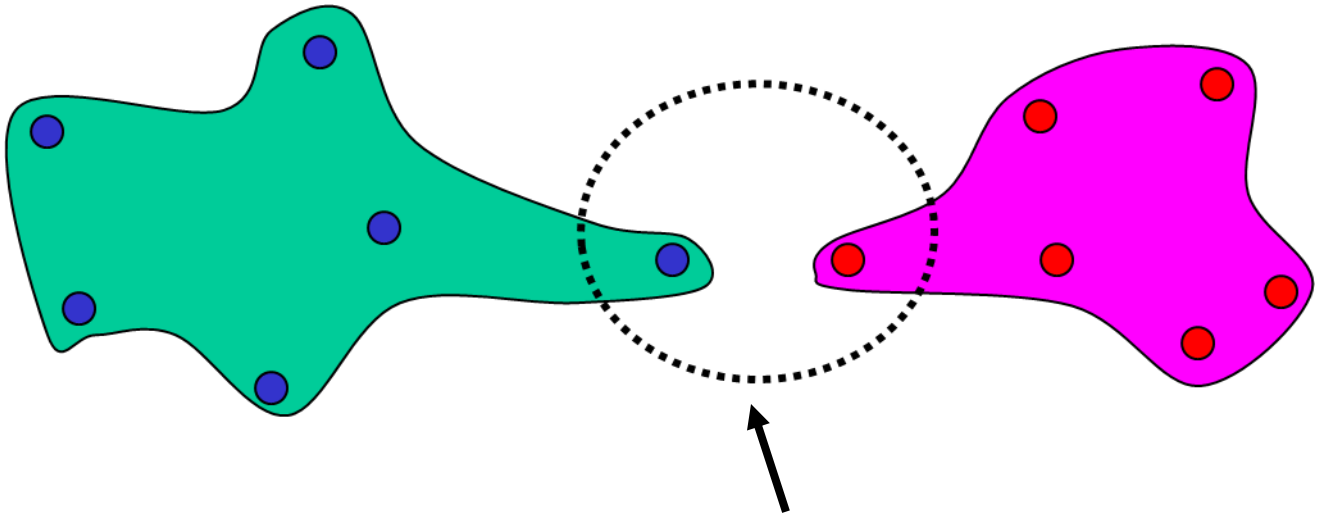
ID	Source	FirstName	MiddleInitial	LastName	P_{Match}
113	BDR	Catherine	A	Sampson	
113	EBC	Catherine	A	Simpson	0.90
113	EHDI	Kathy	A	Simpson	**0.81**

- Maximum probabilistic distance between records
  - $p_{Match}=\alpha_i$ is the probability that two successive records in such a sequence are the same individual
  - Minimal probability that two extreme records in the series will belong to the same individual

$$\mathrm{Min}\left(P_{Match}\right)=\prod a_i$$

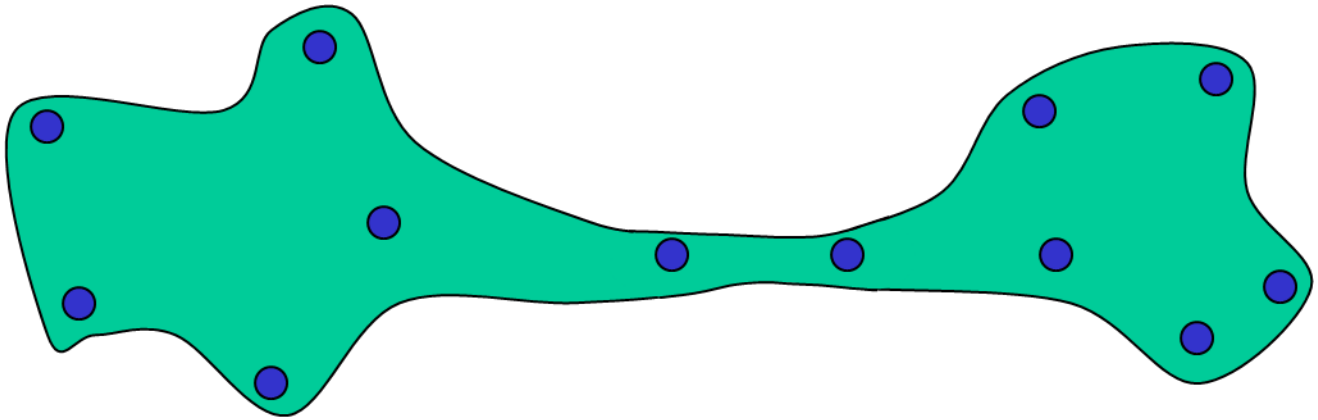
Linkage Creep

- When is this a problem?
  - Over time, two distinct individuals may project "tendrils" composed of combinations of identifiers that statistically overlap in probabilistic space



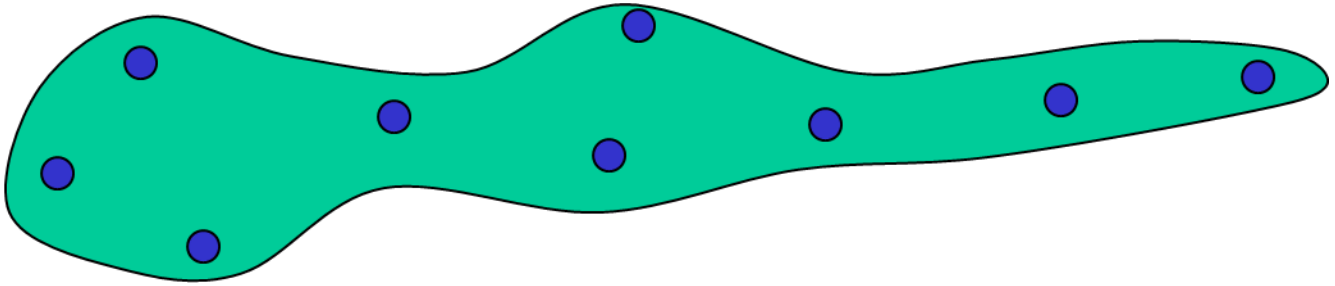
Linkage Creep

- When is this a problem?
  - Linkage creep will result in the two distinct individuals being erroneously combined under a single ID



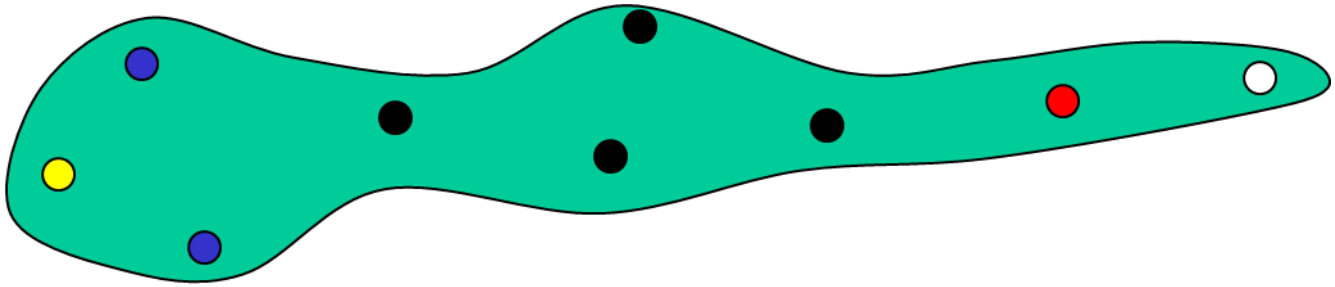
Linkage Creep

- When is this not problem?
  - Over time, certain key identifiers for an individual are expected to change
  - This phenomenon will increase as a historical database grows, and as additional sources are input into a centralized system



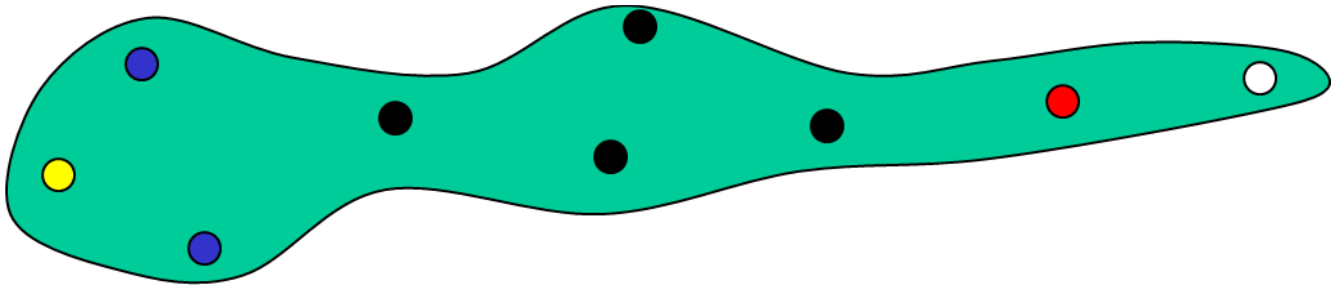
Linkage Creep

- Complexity of "creep" in longitudinal datasets
  - Black records are related to all records at $P_{Match} \geq \alpha$
  - Yellow and Blue records are NOT related to White record at $P_{Match} \geq \alpha$
  - Yellow record is also not related to Red record at $P_{Match} \geq \alpha$

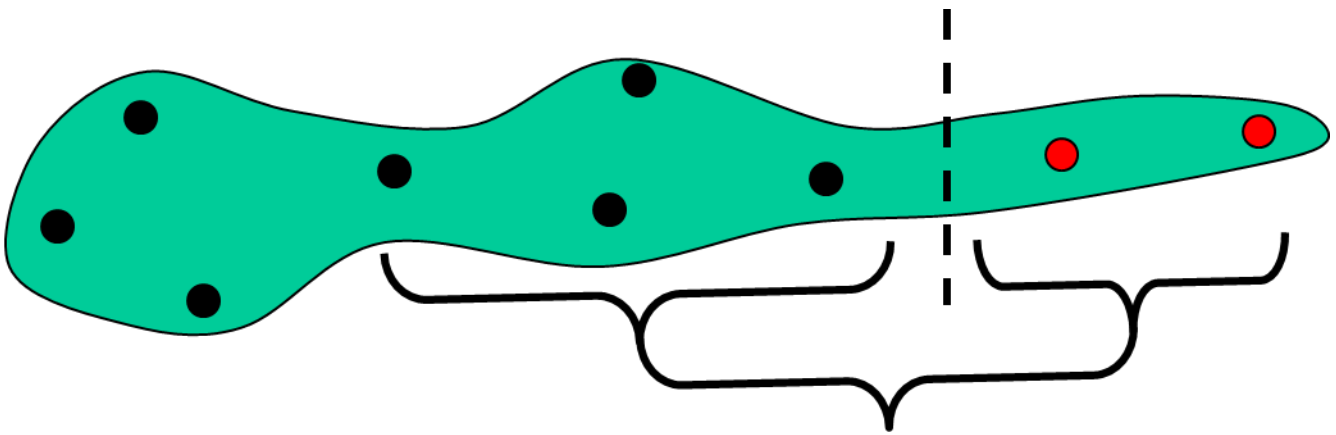


Linkage Creep

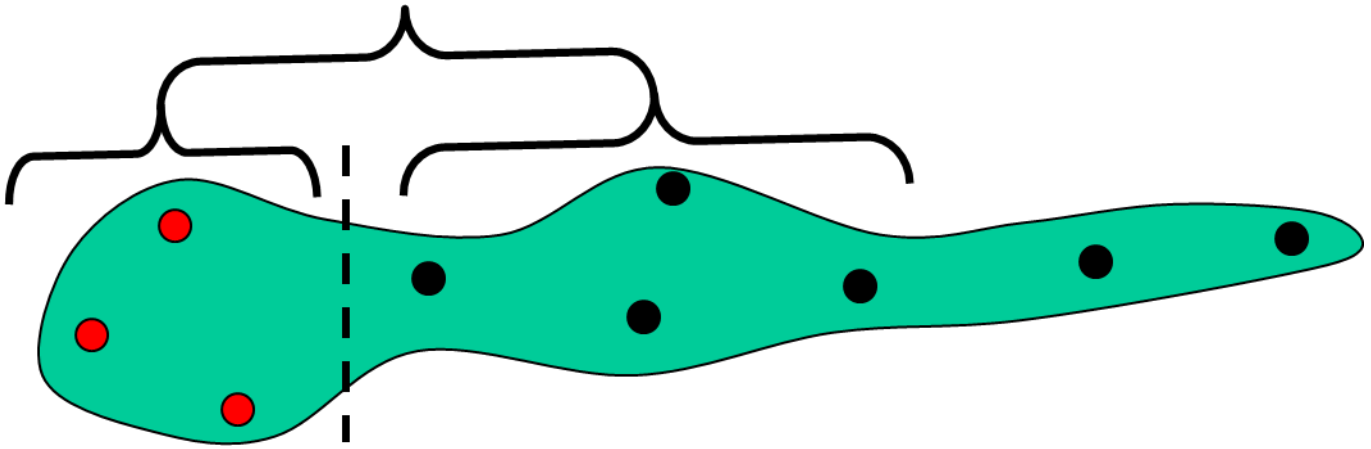
- Forbidding “creep” will result in a single individual being divided into two IDs over time
- Further challenge—where to divide records into additional IDs?



Linkage Creep



$P_{Match} \geq \alpha$: Should be same ID



Pride and Precedence

- Linkage can create inconsistencies regarding IDs or individuals across systems
- Winifred Szamick and Winafred Szamick
  - EHDI classifies as two variations of the same person
  - The Birth Defects program classifies as different people
  - Each proudly insists they are correct
- Which (if either) source takes precedence?
  - May dramatically impact your results
  - Particularly vulnerable when systems contain multiple records for each person over time

Linkage-Based Deduplication

NEW.ID	A.ID	A.FIRST	A.MI	A.LAST	B.ID	B.FIRST	B.MI	B.LAST
1001	1	Craig	A	Mason	432	Craig	A	Mason
1001	1	Craig		Mason	444	Craig		Mason
1003	2	Tao	A	Mason	212	Tao	A	Mason
1003	2	Tao		Mason	212	Tao		Mason
1005	5	Chris		Mason	551	Chris		Mason
1006	6	Christopher		Mason	551	Christopher		Mason
1007	7	Jim		Mason	318	Jim		Mason
1008	8	James		Mason	318	James		Mason
1008	8	James	C	Mason	122	James	C	Mason

- File A Takes Precedence
  - All cases with same A.ID share the same NEW.ID
  - If File:A says two records are the same individual, assume they are the same individual
  - Regardless of whether File:B says they are different people

Linkage-Based Deduplication

NEW.ID	A.ID	A.FIRST	A.MI	A.LAST	B.ID	B.FIRST	B.MI	B.LAST
1000	1	Craig	A	Mason	432	Craig	A	Mason
1001	1	Craig		Mason	444	Craig		Mason
1003	2	Tao	A	Mason	212	Tao	A	Mason
1003	2	Tao		Mason	212	Tao		Mason
1005	5	Chris		Mason	551	Chris		Mason
1005	6	Christopher		Mason	551	Christopher		Mason
1007	7	Jim		Mason	318	Jim		Mason
1007	8	James		Mason	318	Jim		Mason
1008	8	James	C	Mason	122	James	C	Mason

- File B Takes Precedence
  - All cases with same B.ID share the same NEW.ID
  - If File:B says two records are the same individual, assume they are the same individual
  - Regardless of whether File:A says they are different people

Linkage-Based Deduplication

NEW.ID	A.ID	A.FIRST	A.MI	A.LAST	B.ID	B.FIRST	B.MI	B.LAST
1001	1	Craig	A	Mason	432	Craig	A	Mason
1001	1	Craig		Mason	444	Craig		Mason
1003	2	Tao	A	Mason	212	Tao	A	Mason
1003	2	Tao		Mason	212	Tao		Mason
1005	5	Chris		Mason	551	Chris		Mason
1005	6	Christopher		Mason	551	Christopher		
1007	7	Jim		Mason	318	Jim		Mason
1007	8	James		Mason	318	James		Mason
1007	8	James	C	Mason	122	James	C	Mason

- Collapse
  - If either File:A or File:B indicate two records are the same individual, all records with either the corresponding A.ID or B.ID are given the same NEW.ID
  - In essence, this assumes that one file knows something that the other file does not…

Linkage-Based Deduplication

NEW.ID	A.ID	A.FIRST	A.MI	A.LAST	B.ID	B.FIRST	B.MI	B.LAST
1001	1	Craig	A	Mason	432	Craig	A	Mason
1002	1	Craig		Mason	444	Craig		Mason
1003	2	Tao	A	Mason	212	Tao	A	Mason
1003	2	Tao		Mason	212	Tao		Mason
1005	5	Chris		Mason	551	Chris		Mason
1006	6	Christopher		Mason	551	Christopher		Mason
1007	7	Jim		Mason	318	Jim		Mason
1008	8	James		Mason	318	James		Mason
1009	8	James	C	Mason	122	James	C	Mason

- Expand
  - If either File:A or File:B indicate two records are the NOT the same individual, **both** records are given different NEW.IDs
  - Expand could be used to create a new file that would then be checked to see if there are statistically duplicate individuals

Historical Memory and Metadata

- Over time changes may be made to linked data
  - Records may be initially matched, and then determined to not be a true match
  - Inconsistencies may appear and be eliminated
- Future linkages with the same or other data
  - Problems may be fixed and then recreated
  - Millions of records, billions of comparisons to track
  - Must be automated in data

Yikes!

- A parent is erroneously told their child has a birth defect due to a probabilistic linkage that is statistically valid, but nevertheless erroneously links two records
- This error is corrected in the linked database, but a subsequent de-duplication or the linkage of a new dataset results in this erroneous link once again being made through a probabilistic match
- The parent will almost certainly be less forgiving when contacted a second time and mistakenly told their child has a birth defect

Metadata

- **Metadata**: *Data about the data*
- Metadata regarding **linkages**
  - God-field: These records should never (or should always) be classified as belonging to the same person
    - For example, result of name change
  - Iterations in which records were matched
  - Probability for match (w_t probably meaningless)

Metadata

- Metadata regarding **individual fields**
  - What is the source of information
    - Same info from multiple sources that do not agree
    - Precedence: But do you automatically trust some sources regardless of any other information
  - What is the history of values for a field
    - Analyses of metadata to identify problems
    - Jimmy was screened, no he wasn’t, yes he was…

Assessing the Quality of a Linkage Project

Matching Protocols

- How do we know the quality of a linked data set
  - How many errors do we have?
  - Missed matches we should have made
  - Records we matched that are wrong
- Underdeveloped area
  - Strategies poorly defined
  - No clear “best practices”

Percentage Matched

- Percentage of records matched
  - Sometimes the approximate theoretical percentage that should match is known
  - If unknown, determining an “adequate” match may be subjective
- Percentage of records matched can indicate that you are in trouble
  - It doesn’t necessarily indicate that you are safe
  - Just because two records match, doesn’t make it right

Hand Matched Comparison

- Manually verify a subset of matches
  - Directly evaluate linked records
  - Manually “rematch” a subset of data
    - Possibly a subset of questionable matches
- Hand-matched comparison may not be correct
  - Different people use different criteria
  - Problematic at the large-scale

Other Measures

- Uncertain matches
  - How many possible matches are ***NOT*** classified as either a predicted match or predicted non-match?
  - Require further review
- Extent of agreement on fields ***not*** used in matching
  - Agreement on middle initial, etc.
- Rule-based or iterative solutions
  - How many different rule sets or iterations were required to obtain a given result
  - Many iterations may introduce room for inconsistencies

Estimated Probabilities

- Probabilities in probabilistic matching provide a potential tool for evaluating linkages
  - Not ask “are two records the same person?” Yes/No
  - Estimate how likely two records are the same person
- Estimate the number of erroneous linkages
- Possible to conduct a detailed examination of quality by ignoring very strong and very weak pairings, and only focusing on pairings that are ambiguous
  - Estimate the proportion of errors within ranges of w_t

Simulated Data

- Create simulated population that is tracked across multiple generations
- Large number of parameter inputs
  - Ethnic composition of population
  - In- and out-migration rates
  - Birth rates in and outside of marriage
  - Marriage-stability factor, marriage/divorce rates
  - Life-span for healthy adults
  - Accidental death and illness rates

Simulated Data

- Once population is created, datasets for various "official records" can be created
  - Birth certificates, marriage records, etc.
- Various types of errors and missing data combinations can be applied to datasets
  - Percentage of births to unmarried mothers with no father listed
  - Spelling errors across datasets
  - Name changes, particularly for mothers

Simulated Data

- Linkage algorithms then applied to the simulated datasets
  - User will **know** if a linkage is correct
  - Assess ability to recreate family patterns
  - Assess impact of different types of issues, such as no father listed on birth certificate or no access to one type of records, such as marriages
- Useful for understanding algorithms and data needs or consequences

Summary

- Important to evaluate linkage results
- Quality of linkage will increasingly be a concern as more systems start to “talk” with each other
- Area for future growth and research
  - Guidelines and best practices
  - New methodological approaches…