W3 Record Linkage

Why we link records

Motivation
• Data are collected everywhere: public sector, private sector, researchers, individuals
• electronic health records, financial records
• loyalty cards, credit cards, scanner data
• tax data, social security records, census records
• social media traces, text messages, blogs
• Hope ➔ New insight! Maybe even at a lower cost

Longitudinal Employer-Household Dynamics (LEHD)
• LEHD is part of the center for economics studies at the U.S. Census Bureau
• U.S. states share unemployment insurance earnings data and the quarterly census of employment and wages
• Administrative and survey data added by Census Bureau
• Mission: dynamic information on workers, employers and jobs and no additional data collection burden

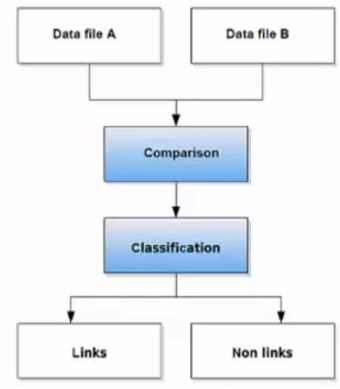
Discussions at Agencies around the World
• U.S. committee on National Statistics report
  - How can administrative and survey data be combined?
  - How can we make more use of administrative data?
  - How can data from other sources be combined with administrative and survey data?
• With the hope of
  - improving the quality of whatever is the already existing data source
  - increase the amount of information available for a unit of interest (more variables)
  - answer new research questions

Challenges
• data from different sources come in very different formats
• records are of different quality
• Often no unique identifier across data from different sources (who belongs together?)
• Different privacy regulations in different sources
• Data sets often very large, finding matches is costly

Ref)
Data Matching – concepts and techniques for record linkage, entity resolution, and duplicate detection (Peter Christen)
Data Quality and Record Linkage Techniques (Thomas N. Herzog, Fritz J. Scheuren, William E. Winkler)

Gentle Introduction: Application, Challenges

Gentle Intro
• What is record linkage?
  - term used to describe the combination of information for the same entity (person, firm etc.) in one or more data source
  - data matching, entity resolution, object identification, duplicate detection etc.



Applications
• Why would one want to link records?
- removing duplicates in one data set

Matching Information

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17
Javier Martinez	49 E Applecross Road	33
Haveir Marteenez	49 Aplecross Raod	36
Gillian Jones	645 Reading Aev	22
Jilliam Brown	123 Norcross Blvd	43

- merging two or more data files

Multiple rationales

- Follow - up of cohorts (for example death registries)
- Merging panel waves
- Validating answers in surveys: Comparing individual provided information with registry or other data
- Bias-detection in surveys: Analysing data for nonrespondents
- Use of external data for imputation or weighting of survey data
- Adding contact information to survey-samples

- identifying the intersection of the two data sets

Multiple rationales

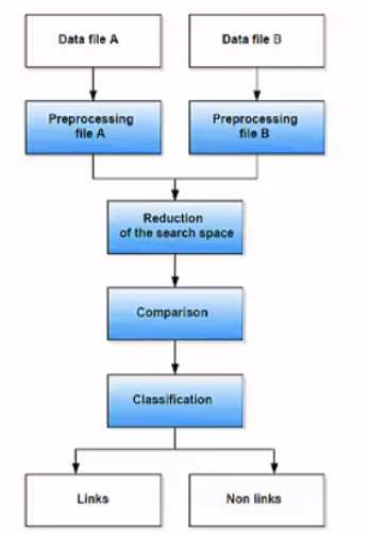
- Discovery of undercoverage within a census or sample
- Estimation of population size through capture-recapture
- Examination of reidentification risks of micro data files
- Discovery of underreporting in registries (e.g. linkage with mortality registry)
- Dropping duplicates as part of data cleansing

- updating data files (with the data row of the other data files) and imputing missing data

- Update of sampling frames
- Update of registries (e.g. new registrations in the cancer registry)

Traditional Challenges
•Lack of Unique identifiers
• dirty data
- typographical errors
- variations
- missing and out-of-date values
- different coding schemes (i.e. dates)
• Privacy
- sensitive data (names, address)
- informed consent
- use

New Challenges
• Scalability
- naïve comparison of all record pairs is quadratic
- remove likely no-matches as efficiently as possible
- need efficient etchniques
• More linked / multi-relational
- need to infer relationship
• No training data in many linkage applications
- no record pairs with known true match status

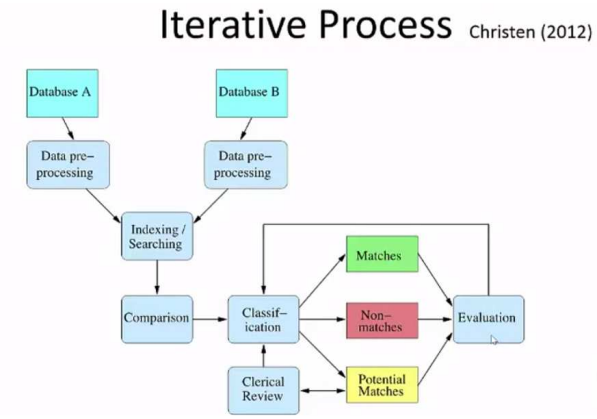


Shares of effort within linkage process
• 5% matching and linking efforts
• 20% checking that the computer matching is correct
• 75% cleaning and parsing the two input files
(Gill 2001, p. 31)

Importance of Preprocessing
• "In situations of reasonably high-quality data, preprocessing can yield a greater improvement in matching efficiency than string comparators and 'optimized parameters'. In some situations, 90% of the improvement in matching efficiency may be due to preprocessing." (Winkler 2009, p. 370)
• "Inability or lack of time and resources for cleaning up files in preparation of matching are often the main reasons that matching projects fail." (Winkler 2009, p. 366)

Note of Identifiers
• Typical identifiers:
  - People: first and last name, address, birth date, sex
  - establishments / firms: name, legal form, address
• The higher the number of different manifestations of an identifier, the better its suitability for a comparison
• Complex identifiers should be parsed into its separate components
• Means of getting identifiers in the first place

Benefits of variation in identifiers
• Variations within a given unit possible in almost every variable
• Variation can arise almost everywhere
  - many reasons (like marriage with change of name, nickname)
  - multiple variations can help provide link to the other data set
• Always keep all available variations and apply them!



Key Linkage Techniques

3 key techniques
• Deterministic linkage
• probabilistic linkage
• "Computer Science" methods

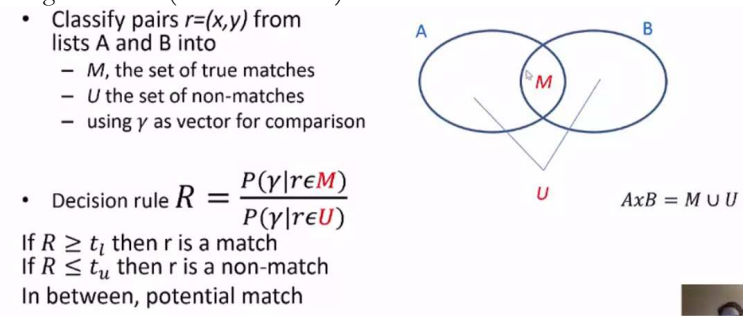
Deterministic (rule based) linkage
• simplest method of matching
• sort/merge exact match
• works best with single unique identifier (key)
• Identifiers
  - have equal weight
  - chosen by researcher or availability
• works well if keys are perfect and present in all datasets you want to ink. In example one key not enough (missing data), next key (name) has errors in it

Data Set	#	SSN	Name	DOB	Sex	ZIP
Set A	1	000956723	Smith, William	1973/01/02	Male	94701
	2	000956723	Smith, William	1973/01/02	Male	94703
	3	000005555	Jones, Robert	1942/08/14	Male	94701
	4	123001234	Sue, Mary	1972/11/19	Female	94109
Set B	1	000005555	Jones, Bob	1942/08/14		
	2		Smith, Bill	1973/01/02	Male	94701

Example: https://en.wikipedia.org/wiki/Record_linkage

Probabilistic Record Linkage
• Basic Idea by Newcombe & Kennedy (1962), Theory Fellegi & Sunter (1969)
• Allows wider range of potential identifier
• Computing weights based on estimated ability to match or not match (frequency ratios)
• Unlike rule based method, algorithms can be trained
• Problem: estimating errors and thresholds, independence assumption, clerical review
• Often used:
  - M probability: probability that a field agrees given that the pair of records a true match
  - U probability: probability that a field agrees given that the pair of records is NOT a true match

Felligi & Sunter (in Science 1969)



Assumption Behind
• Conditional Independence Assumption (CIA): given a pair of records representing the same entity (true match), we assume that agreement in each field is independent of agreement in other fields
• CIA is a mathematical convenience only
• In reality, record fields won't be conditionally independent. Improve match discrimination by eliminating covarying fields (more fields not always better matching)
  - area codes covary with geography
  - first name may covary with sex

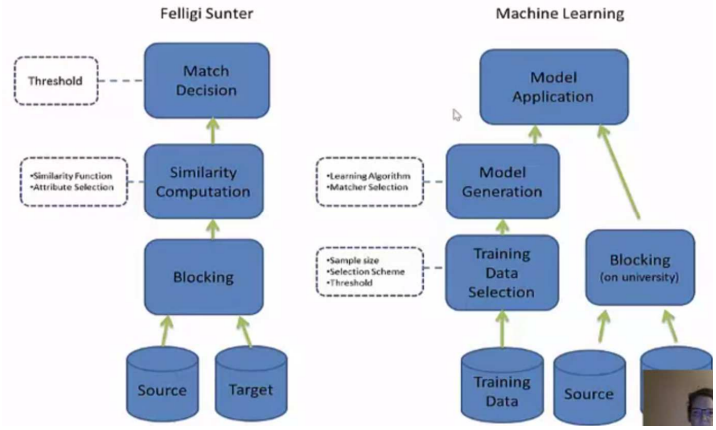
Probabilistic Matching
• Each matching variable is compared and assigned a score (weight) based on how well it matches
• Frequency analysis of data values is important
• Uncommon value agreement stronger evidence for linkage (e.g. Julia Lane v.s. Frauke Kreuter)
• Calculates a score for each field that indicates, for any pair of records, how likely it is that they both refer to the same

entity
• sum the scores over fields
• sort record pairs in order of their scores (weights)
BLOCKING really critical

Issues
• which models should be used to estimate m- and u-probabilities? (modelling issue)
• how do we set the thresholds? (Evaluation)
• which variables/features are used for comparison? (Feature engineering)
• How do we avoid looking at all possible pairs? (Efficiency) … [all pairs means if A=100 and B=1,000 then 100 * 1,000 = 100,000 possible comparisons]

On efficiency…
• Remember "Strata" from earlier modules?
• same idea of can be applied!
• Choose smaller set of pairs likely to contain all matches
  - simple blocking: compare all pairs that "hash" to the same value (e.g. same city, same birth year)
  - extensions:
    - block on multiple attributes and take union of all pairs found
    - pick ordered attributes and sort (e.g. sort on last name). The pick all pairs that appear "near" each other in the sorted order

Make Connections!
• Problem similar to other clustering and prediction problems!
• Propensity score matching (and the methods used there) are design to find similar cases (!)
• Clustering algorithms are design to find similar cases …..



Summary: Record Linkage Technique (Christen 2015)
• Deterministic matching
  - rule-based matching (complex to build and maintain)
• Probabilistic record linkage (Fellegi and Sunter, 1969)
  - use available attributes for linking (often personal information, like names, addresses, dates of birth etc.)
  - calculate match weights for attributes
• "Computer science" approaches
  - based on machine learning, data mining, database on information retrieval techniques

Software

- **Link Plus**, developed by the Centers for Disease Control, www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm
  - Graphical user interface - easy to use
  - Beginner-level knowledge of the linkage process.
  - Best for linkage < 1 million records
- **The Link King**, developed by Washington State's Division of Alcohol and Substance Abuse www.the-link-king.com.
  - Requires a license for base SAS
  - Easy to use graphical user interface
  - Requires first and last name, as well as Social Security Number or date of birth
- **ChoiceMaker 2** (developed by ChoiceMaker Technologies www.sourceforge.net/projects/oscmt/)
- **FEBRL** (developed by the ANU Data Mining Group and available at www.sourceforge.net/projects/febrl/)
- **Merge ToolBox (MTB)** is a Java application developed by the German RLC http://record-linkage.de/-Downloads--software.htm
  - Easy to use interface
  - Allows for privacy preserving record linkage