**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)

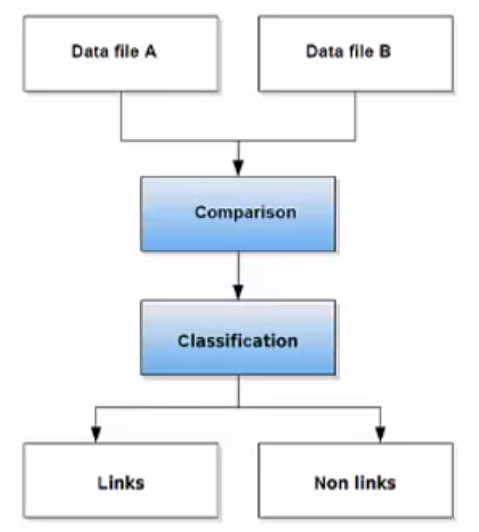
**Gentle Introduction: Application, Challenges**

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

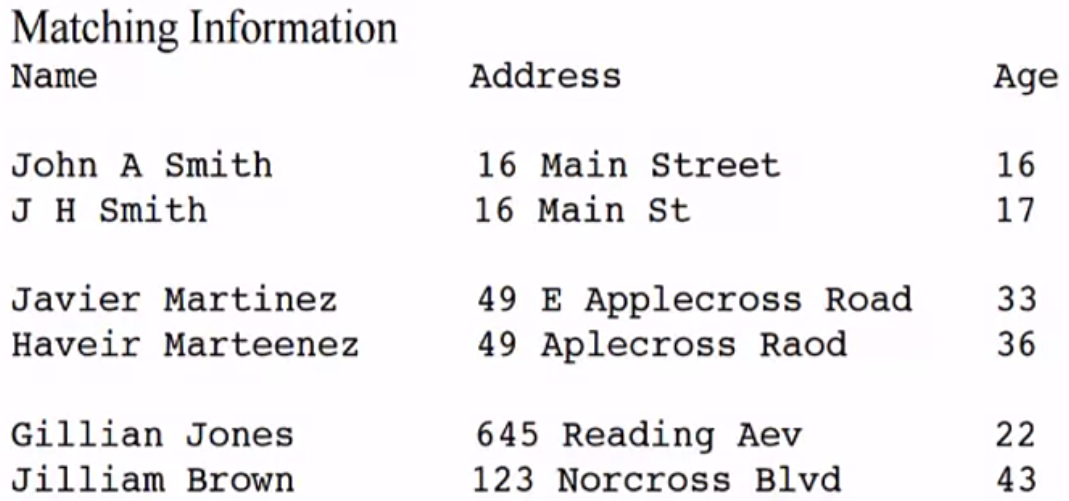
- data matching, entity resolution, object identification, duplicate detection etc.



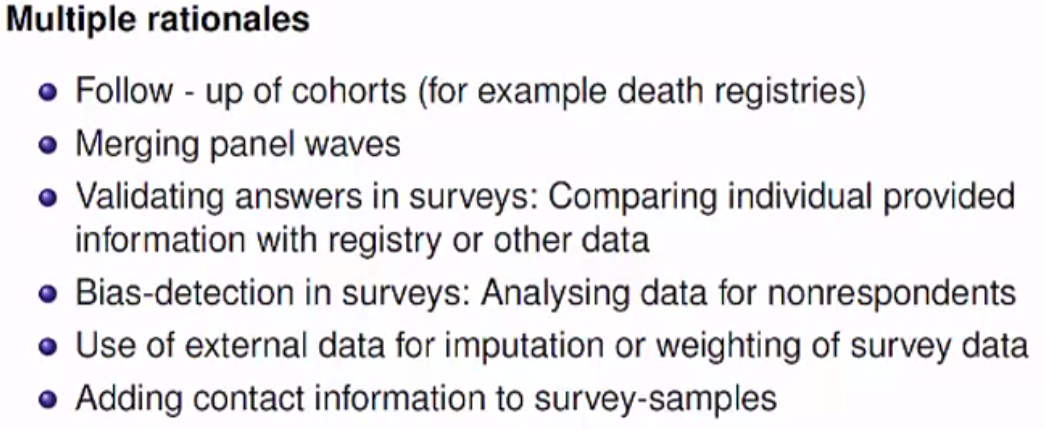
Applications

• Why would one want to link records?

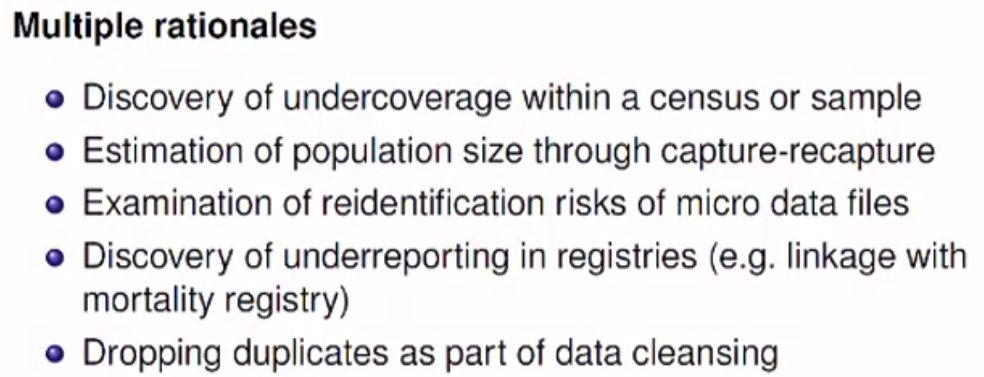
- removing duplicates in one data set



- merging two or more data files



- identifying the intersection of the two data sets



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

