**Foster et al., Chapter 3**

Foster and colleagues provide a thorough introduction to the concept of record linkage and explain why it is a central task in data management and social science research. Record linkage refers to the process of identifying which records from different datasets refer to the same person, household, business, or organization. The authors describe common challenges such as inconsistent spelling of names, missing values, changes in addresses, and differences in formatting across administrative systems. These issues make it difficult to rely on simple identifiers and require more advanced methods to correctly match records. The chapter explains both deterministic matching, which uses exact agreement on key variables, and probabilistic matching, which evaluates the degree of similarity between records. Foster and colleagues emphasize that successful record linkage improves data completeness and reliability, which allows researchers to combine datasets, track individuals over time, and carry out more accurate analyses in areas such as health, education, and public policy.

**Enamorado, Fifield, and Imai (2019)**

Enamorado, Fifield, and Imai extend the discussion of record linkage by developing a probabilistic model specifically designed for merging large administrative databases. They argue that traditional approaches often fail when datasets are massive or contain measurement errors, because strict rules for matching may overlook true matches while also creating false links. Their model uses statistical methods to estimate the probability that two records correspond to the same entity based on multiple fields such as names, dates, and demographic characteristics. This approach helps researchers prioritize likely matches and reduce the uncertainty that arises from inconsistent or incomplete information. The authors also show how their method performs better than existing techniques by achieving higher accuracy in identifying correct pairs. Their contribution is significant for modern social science research because administrative records are becoming more widely available, and the ability to link these records efficiently allows researchers to study populations more comprehensively and to answer questions that require detailed, individual-level data.

**Class Discussion**

**Linkage Methods**

Linkage methods refer to the different ways researchers connect or merge records from separate datasets that belong to the same person, household, or entity. These methods can be based on how much agreement is required between fields, how uncertainty is handled, or how statistical relationships are used to identify matches. It can be:

i**. Deterministic Linkage**

Deterministic linkage relies on exact agreement between one or more key variables to decide if two records represent the same entity. Examples include matching Social Security numbers, full names with birthdates, or identical addresses. This method is simple and transparent, but it requires perfect consistency across datasets. When information is missing, misspelled, or formatted differently, deterministic linkage may fail to identify true matches, resulting in under linkage and incomplete merged data.

ii. **Probabilistic Linkage**

Probabilistic linkage allows for uncertainty by comparing the similarity of multiple fields rather than requiring perfect matches. Instead of treating records as identical or different, it assigns a probability that two records refer to the same individual based on evidence from several attributes. This method is more flexible when information is inconsistent, and it is especially useful for large or messy administrative datasets. Probabilistic linkage reduces both false matches and missed matches, and it often produces more accurate results in real-world applications.

iii. **Statistical Matching vs. Exact Matching**

Statistical matching uses estimated relationships between variables to infer links when no unique identifier is available, while exact matching depends on identical values for selected fields. Statistical matching tries to create the most plausible match based on shared characteristics, even if no exact agreement exists. Exact matching is more rigid and only works well when the data are clean and identifiers are reliable. Statistical matching can preserve more information in cases where exact agreement is rare, but it requires careful modeling and may introduce uncertainty into the final dataset.

## Linkage Algorithms

Linkage algorithms provide computational procedures that determine how records are compared and matched. Basic algorithms include pairwise comparison, where each record in one dataset is compared with all possible records in another dataset, but this becomes inefficient for large files. More advanced algorithms use blocking or indexing techniques to limit comparisons to only those records that share certain features, which reduces computational burden. Algorithms may also incorporate machine learning methods, similar scoring functions, and optimization routines to identify the most likely matches. The choice of algorithm affects both the accuracy and the speed of the linkage process, and it plays a crucial role when working with large-scale administrative data.