Experimental Evaluation

- **Data Sets**
  - NC: a real voter registration data set from North Carolina that contains names and addresses of over 8 million voters, as well as their voter registration numbers.
  - CCA: a confidential commercial database which contains names and addresses of tens of millions of individuals. Subsets with various sizes were extracted.
  - OZ-1: Corrupted data set generated from an Australian telephone directory. One attribute is corrupted and modified.

- **Results**
  - **Quality**: using multiple trees has a noticeable improvement on matching quality with a trivial increase in query time when using sorting keys based on concatenated attribute values.
  - **Scalability**: the approach is scalable to large databases.

Research Problem

- Entity resolution (ER) is an important step in data cleaning.
- ER is the process of identifying groups of records in multiple data sources that refer to the same real-world entity.
- Most existing ER techniques link multiple data sources offline and work only with static databases. Many organizations need to link data records in real-time.

There is a need for **real-time** ER techniques that work with **dynamic** databases.

Dynamic Sorted Neighborhood Index (DSNI)

- Most available indexing techniques are designed for batch.entity resolution algorithms that resolve all records in a database rather than resolving those relating to a single query record. Such techniques are not suitable for real-time entity resolution on dynamic databases which are updated constantly.
- We propose a tree-based sorted neighborhood index that facilitates resolving a stream of query records against a large and dynamic database in real-time.

![Image of DSNI](image)

The table on the left shows a small example data set. Record \( r_{10} \) is a query record. The right figure represents one tree in the F-DySNI built for records from this data set after inserting the query record.

A Forest-based DySNI (F-DySNI)

- A dynamic sorted neighborhood index that uses multiple index trees with different sorting keys to facilitate real-time ER for dynamic databases.
- This technique aims to reduce the effect of errors and variations in attribute values on matching quality by building several distinct index trees.

![Image of F-DySNI](image)

Indexing techniques are generally used to efficiently extract a set of candidate records from that are similar to a query record, and that are to be compared with the query record in more detail.

Indexing can reduce the number of comparisons significantly.