Why it's helpful

- Most citation managers catch a lot of duplicates, but it is impossible to catch all duplicates using exact matching on key fields (e.g., author, title, date, pages).
- This process provides a list of certain duplicates that do not need to be confirmed manually and a list of probable duplicates identified using machine learning. Overall, fewer duplicates require manual review and less effort is needed to identify probable duplicates compared to traditional methods.

Tool in use
Endnote & Excel, with the DeDupe package in Python

Phase 1: Maximum Precision
- Create up to three .RIS files from your search results in priority order. Highest priority is typically given to database with best quality data (e.g., PubMed).
- New fields for First Author's Last Name and Starting Page Number are added to .RIS file (Subsidiary Author and Section fields respectively).
- Duplicates identified based on exact match of title, year, first author's last name, and starting page #. Records marked as duplicates do not require manual review.

Phase 2: Maximum Recall
- ‘Keeps’ (or items not marked as duplicates in Phase 1; 9,797 in screenshot) move to Phase 2. Machine learning algorithm predicts sets of probable duplicates among these records.
- User manually reviews subset of records containing probable duplicates (i.e., 874 in screenshot).
Phase 1: Maximum Precision

Input: RIS files – separated in priority order

Two new fields added:
Starting page # & First author’s last name

Dupes removed:
Match title, year, 1st author last name, starting page

Output:
Two RIS files: Keeps & Dupes (no checking needed)

Phase 2: Maximum Recall

Input: ‘Keeps’ from Phase 1

Dupes removed:
Uses machine learning to predict likely duplicates

Cluster ID added:
Matches groups of duplicates to one cluster ID #

Output:
Two RIS files: Unique Records & To Review