# Title: Overview

Dedupe (Entity Resolution/Record Linkage)

![](data:text/html;base64,)

Similar products

1. <https://senzing.zendesk.com/hc/en-us/articles/360009902713-Senzing-App-Quickstart>
   1. Unsupervised
   2. Outputs fields which are similar
2. [Dedupe.io](https://dedupe.io/)
   1. Asks for sample similar records
3. <https://dedupe.ly>
   1. <https://www.youtube.com/watch?v=yr2XXBvZaB8>
4. [Acuma](https://www.acuma.co.uk/partners/saksoft/dedupe/)
5. [Hubspot](https://knowledge.hubspot.com/articles/kcs_article/contacts/how-does-hubspot-deduplicate-contacts)
6. [Openrise](https://www.openprisetech.com/dedupe/)
7. <https://www.workwisellc.com/crm-software/tools/data-deduplication/>

Technical consideration

1. Pairwise matching is exhaustive
   1. If we do pairwise comparison for 2000 records we will have ~2M comparison
   2. Current freshworks account has ~3M records
2. Building ML based (supervised) approach needs training data
   1. Create blocks of similar records and do exhaustive matching within them
   2. Current merged records are based on string matches
3. Real time matching
   1. NLP based methods using Elasticsearch
   2. Using boosting to get good ranking on search results

Available libraries

1. <https://pypi.org/project/dedupe/>
2. <https://github.com/seatgeek/fuzzywuzzy>
3. More linked [here](https://docs.google.com/spreadsheets/d/1N0g13p6ps2DC8_HU9z4RJOuRgGs_IvcMtzz9Z_5HR7Y/edit?usp=sharing)

Related articles

1. Overview - <http://users.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf>
   1. ![](data:text/html;base64,)
   2. ![](data:text/html;base64,)
   3. ![](data:text/html;base64,)
2. <https://medium.com/district-data-labs/basics-of-entity-resolution-with-python-and-dedupe-bc87440b64d4>
3. <https://www.slideshare.net/BenjaminBengfort/a-primer-on-entity-resolution>![](data:text/html;base64,)
4. <https://pdfs.semanticscholar.org/174d/4ad613a21cb27f73efc4c556522c8c761e71.pdf?_ga=2.174806948.727998850.1545201788-1131716626.1545201788>
5. <https://people.cs.umass.edu/~mccallum/papers/kdd289-hall.pdf>
6. <http://infolab.stanford.edu/serf/>
7. <https://github.com/daqcri/DeepER>

Appendix

Questions:

1. Rules that define deduping a lead (Eg: fields that should be exact match, nearest match, etc.)
2. Decision on product features based on different competitors.
3. Data access
4. Who decides mandatory fields? How to handle leads with a lot of empty fields?
5. Decision on initial set of fields and the corresponding weights to start with. Info provided by product team
6. When does dedupe system run? Eg: Real time during lead create, update

Workflow for Supervised approach:

One time dedupe setup:

1. Request user to select the data(rows) to be deduped
2. Request the user to select the fields to be used for dedupe and if any specific comparison method has to be used
3. Request user to label a minimum dataset (suggested by the system). The data presented is a pair of records and the user action is same group or not. ‘Label’ used for training is a uniquely generated groupid.
4. Train the model based on the training set and predict for the rest of the data and present the results to the user.
5. Request user to verify the results(user can blindly trust results, blindly reject and review every group and give feedback)
6. Merge records based on user feedback. Need to find an algo but not a ML prob. Should we take user inputs for merging as well?
7. Validate the results presented by the system based on user action and add these records to training set and retrain the model.
8. For any record which is not a part of any group, run prediction and see if it belongs to any group. Suggest to user.
9. Repeat step 7 if any new data was presented at step 8
10. Show the final data of groups identified by the system along with the user who reviewed the records and the reason for merging. Give an option to remove a particular record from the group or merge groups manually.

Real time dedupe:

1. Once dedupe is setup, any new entry(from any source) is checked whether it belongs to one of the existing groups. The results are presented to the user(who will be the reviewer for automated entries?). Based on the action taken by the reviewer, system adds the record to one of the groups if necessary. This entry is added to training set.
2. Model training happens once in x days.