



ACM SIGMOD Programming Contest 2021

Team UKN (Runner up)

Yinan Mei, Haoyu Wang, Dongming Zhao, Shaoxu Song (Advisor)
{myn18@mails, wanghy20@mails, zdm20@mails, sxsong@}.tsinghua.edu.cn
DIQGroup @ Tsinghua University, China

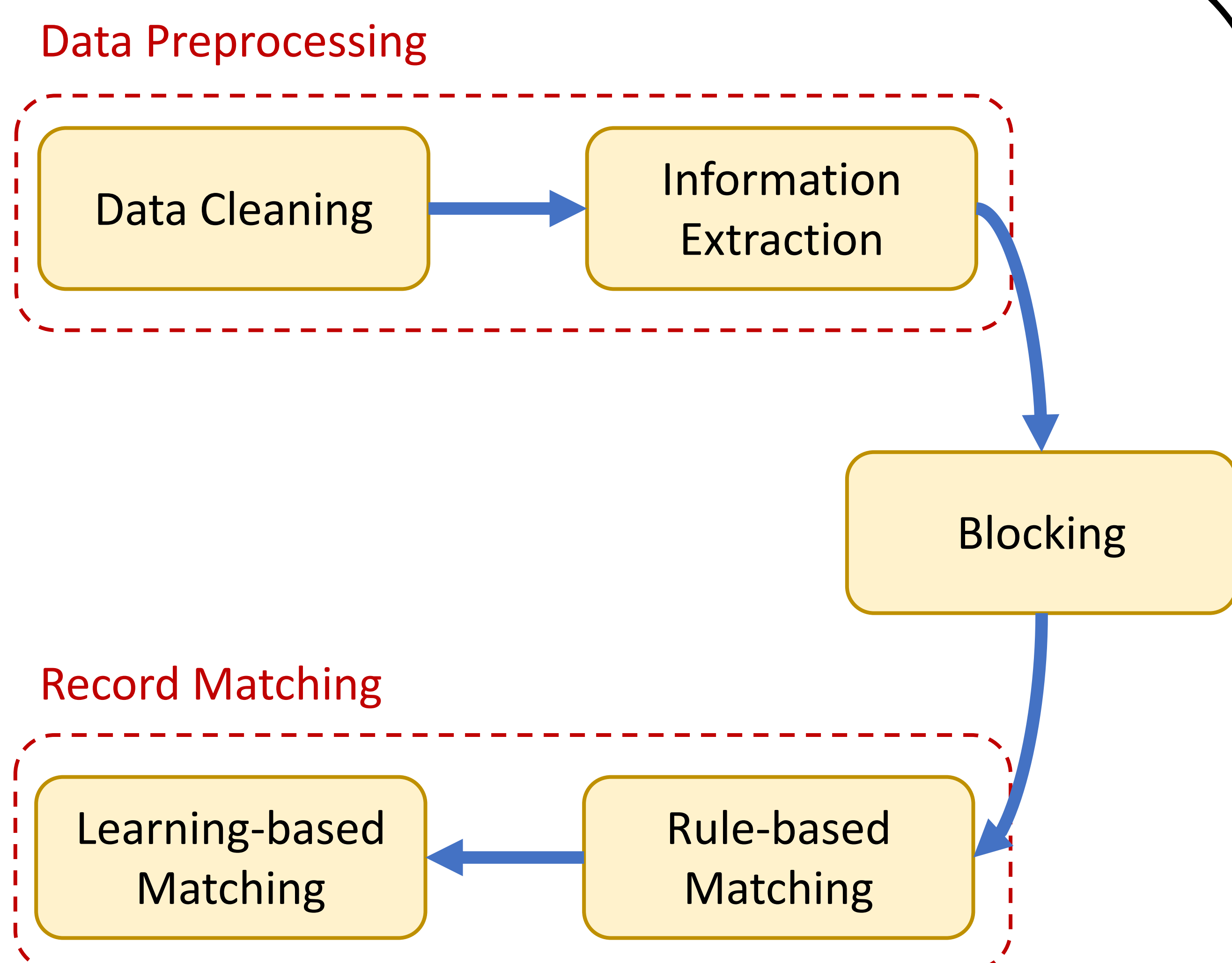
Contest Overview

- **Task:** solve Entity Resolution on several datasets, each one containing a different type of object (e.g., people, products, movies, etc.) and different distributions of data and noise.
- **Implementation:** Python 3.8

Blocking

- A loose blocking method guarantees a high recall.
- Notebook & NotebookLarge:
 - Group all the entities by extracted brand
- Altosight
 - Group all the entities by extracted brand & size

Framework



Rule-based Matching

- Notebook & NotebookLarge:
 - Match Criteria: Same brand, modelno, cpu_model, cpu_type, ram
- Altosight:
 - Match Criteria: Same brand, type, size, modelno (optional)

Learning-based Matching

- Post-processing uncertain pairs left by rule-based method
- Model: Xgboost
- X: similarity vectors on “title”; one-hot encoding for other attributes
- Y: matched=1; unmatched=0

Data Preprocessing

- Lowercase
- Remove Stop Words
 - Remove “amazon.com :”, “vology”, ...
- Remove Punctuations
 - in. -> in
- Synonyms Standardization
 - mémoire -> memory; memoria -> memory

Information Extraction

- Only part of the information counts.
- Notebook & NotebookLarge:
 - cpu_brand, cpu_model, cpu_type, cpu_freq, modelno, ram
- Altosight:
 - brand, size, type, modelno
- Approach: Regular Expressions

Results

Dataset	Precision	Recall	F-measure
Notebook	99.5	97.2	98.4
NotebookLarge	98.7	97.3	98.0
Altosight	89.0	92.5	90.7

Conclusion

- In the real world, building a fully automated Entity Resolution system is still a long way to go.
- We have to spend most of time on data cleaning.
- The rule-based approaches are still very effective.