# T1

## A

Blocking significantly improved my record linkage results by reducing comparisons while retaining most true matches. However, if blocking keys are poorly chosen, true matches that fall into different blocks may be excluded, impacting Pairs Completeness (PC). I used a multi-pass phonetic blocking method with two blocking keys:

First Pass: Last Name and Gender (Attributes [3, 4]).

Second Pass: Street Address, Suburb, and Postcode (Attributes [7, 8, 9]).

Phonetic encoding accounted for spelling variations. Combining blocks from both passes increased pairs completeness to 0.866, meaning 86.6% of true matches were included.

Effective Blocking Keys:

Last Name and Gender: High discriminative power; grouped similar individuals despite minor errors.

Address Attributes: Captured matches where address data was reliable even if names varied.

Less Useful Keys:

First Name and Middle Name: High variability and errors could exclude true matches if used for blocking

## B

The optimal balance between performance measures and the quality of the final linkage results is achieved by using the combined blocking strategy, which involves multi-pass phonetic blocking on both [3, 4] (Last Name and Gender) and [7, 8, 9] (Street Address, Suburb, Postcode).

High Pairs Completeness (PC = 0.866): Ensures most true matches are included.

High Reduction Ratio (RR = 0.999): Maintains computational efficiency.

Acceptable Pairs Quality (PQ = 0.024): While lower, it's an acceptable trade-off for the gains in PC.

Improved Linkage Metrics:

Precision: 0.813

Recall: 0.865

F-Measure: 0.838

Rationale:

Including More True Matches: By accepting more candidate pairs (even with some non-matches), we significantly increase the likelihood of capturing true matches, which is critical for high-quality linkage.

Balanced Trade-Off: The slight decrease in PQ is offset by the substantial increase in PC and overall linkage quality.

Efficient Use of Resources: Despite processing more candidate pairs, the high RR indicates that the approach is still efficient and practical.

## C

Yes.

Examples of How the Trade-off Changes

Scenario 1: High-Quality Data

Datasets with minimal typos and consistent formats allow for strict blocking keys, such as exact matches on names and dates. This yields a high Reduction Ratio (RR), Pairs Completeness (PC), and Pairs Quality (PQ), as fewer comparisons are needed, and the linkage results are highly accurate due to the clean data.

Scenario 2: Low-Quality Data

In contrast, datasets with frequent typos, missing values, and inconsistent formats require relaxed blocking methods like phonetic encoding or multi-pass blocking. While this increases PC, it lowers RR since more comparisons are necessary to capture true matches. PQ might also decrease due to more non-matching pairs being included. Sophisticated matching algorithms are often needed to maintain high linkage quality.

Changes in the trade-off are driven by several factors.

High error rates lower PC due to false negatives, while high variability reduces RR by requiring more inclusive blocking.

Reliable attributes boost both RR and PC, while unreliable attributes might need to be approximated or excluded.

Large datasets amplify the effect of low PQ, as false positives can lead to many incorrect matches.

Inconsistent formats necessitate additional preprocessing or flexible blocking methods for effective linkage.

# T2

## A

Different comparison techniques affect linkage results by influencing the balance between precision and recall. For attributes susceptible to typos and variations, like first and last names, I selected approximate string comparison functions such as Jaro-Winkler ,which effectively handle minor spelling errors and transpositions. For stable and standardized attributes like state codes, I used exact matching after normalization to ensure high precision. Techniques like Edit Distance were less suitable for short strings due to inefficiency. By matching comparison functions to attribute characteristics—using approximate methods for variable data and exact methods for consistent data—I optimized linkage accuracy by capturing true matches and reducing false positives.

## B

Different classification techniques and parameter settings affect linkage quality by shifting the balance between precision and recall. I selected a threshold-based classification method (0.4??), tuning the similarity threshold to optimize this balance. By testing various thresholds, I found one that maximized the F-measure, achieving high linkage quality. Lower thresholds increased recall but decreased precision, while higher thresholds did the opposite. Other techniques, like supervised machine learning classifiers, were less suitable due to their complexity, need for extensive labeled data, and risk of overfitting—factors that were impractical given our data and resource constraints.

## C

Linkage quality, measured by precision, recall, and F-measure, changes notably with different parameters and techniques. In my case, setting the q-gram size Q=3 in the Dice/Jaccard methods improved precision by better capturing similarities in names despite minor differences. Additionally, using a similarity threshold of 0.5 in the thresholdClassify method yielded better results, balancing precis ion and recall effectively. The linkage quality is particularly sensitive to parameters like q-gram size and similarity thresholds because they directly affect how similarities are measured and how matches are classified. Small adjustments can significantly impact the ability to correctly identify matches while minimizing false positives.

## D

Yes, certain evaluation measures are not particularly useful for evaluating the performance of a record linkage project. Accuracy is one such measure that can be misleading in this context. In record linkage tasks, especially with large datasets, the number of non-matching record pairs vastly outnumbers the number of matching pairs. As a result, a classifier that labels all pairs as non-matches could achieve a very high accuracy, even though it fails to identify any true matches. This makes accuracy an unreliable metric for assessing linkage quality.

Another measure that is less useful is the Reduction Ratio (RR) when evaluating the final linkage performance. While RR is valuable for assessing the efficiency of the blocking step—indicating how many comparisons were avoided—it doesn't reflect the effectiveness of the classification step in identifying true matches. Similarly, Pairs Completeness (PC) and Pairs Quality (PQ) are more relevant to blocking evaluation rather than the overall linkage performance.

## E

Linkage evaluation:

blocking\_attrA\_list = [3,4] ( 3: last\_name, 4: gender)

For sim\_threshold=0.4,Q-grams=3:

Accuracy: 1.000

Precision: 0.170

Recall: 0.765

F-measure: 0.278

For Weighted similarity based classification :

Weight vector: [2.0, 1.0, 2.0, 2.0, 2.0, 1.0]

Accuracy: 1.000

Precision: 0.878

Recall: 0.765

F-measure: 0.818

# T3

## A

Best Linkage Quality Result:

I achieved a recall of 0.865, precision of 0.813, and an F-measure of 0.838. The blocking step resulted in Pairs Completeness (PC) of 0.866 and Pairs Quality (PQ) of 0.024.

Why This Combination Worked:

I used phonetic blocking on Last Name and Gender, followed by a second pass with Street Address, Suburb, and Postcode. This dual strategy captured more true matches while keeping comparisons manageable. Phonetic blocking accounted for common typos and variations in names, improving recall.

For comparison, I applied a mix of exact matching and approximate techniques, like Jaro-Winkler for names and bag distance for addresses, which handled spelling and formatting inconsistencies effectively. This combination ensured most true matches were captured while minimizing false positives.

## B

The results are particularly good for precision, recall, and F-measure, indicating that the classification techniques and parameter settings effectively identified true matches and minimized false positives. However, the results are not as good for pairs quality (PQ) in the blocking step.

Reasons for Discrepancy:

Blocking Efficiency: While the combined blocking strategy increased pairs completeness by capturing more true matches, it also included a significant number of non-matching pairs, lowering PQ.

Trade-off Between PC and PQ: Enhancing pairs completeness often comes at the expense of pairs quality because including more true matches typically means accepting more non-matching pairs.

Class Imbalance: The inherent imbalance in the dataset amplifies the impact on PQ, as the vast majority of candidate pairs are non-matches.

# T4

## A

Impression After Conducting Linkage:

Assignment Datasets:

The higher data dirtiness made it challenging to link records accurately, resulting in lower Pairs Quality (PQ = 0.024) despite a high Pairs Completeness (PC = 0.866). A double-pass blocking method using Last Name, Gender, and Address fields helped capture more true matches. Precision (0.813) and recall (0.865) were acceptable but lower compared to lab datasets.

Lab Datasets:

Cleaner data led to smoother linkage, achieving a high PQ of 0.918 with single-pass blocking. Precision reached 0.952, and recall 0.905, resulting in a strong F-measure of 0.928.

## B

I assessed the data quality by combining quantitative calculations and qualitative observations. Firstly, I calculated **error rates** for key attributes like names and addresses by identifying typos and inconsistencies. For example, in the assignment datasets, the error rate for last names was around 15%, compared to 5% in the lab datasets.

Secondly, I evaluated **missing value rates** by calculating the percentage of records with missing data, finding higher rates in the assignment datasets. Thirdly, I analyzed **linkage performance metrics** such as Pairs Quality (PQ) and Pairs Completeness (PC). The assignment datasets had a low PQ of 2.43%, whereas the lab datasets had a high PQ of 91.80%, indicating cleaner data.

Finally, I examined data consistency and formatting, noting more inconsistencies in the assignment datasets. This comprehensive methodology highlighted that the assignment datasets were significantly dirtier than those from labs 3 to 6, impacting the effectiveness of the record linkage process.