Blocking vs Matching

| **Aspect** | **Blocking** | **Matching** |
| --- | --- | --- |
| **Stage in ER** | **Pre-processing** | **Decision-making** |
| **Primary Goal** | Quickly reduce the O(n2)O(n^2) pairwise search space by grouping (or “blocking”) similar records | For each candidate pair, decide if the two records truly refer to the same real-world entity |
| **Input** | Raw dataset of nn records | Candidate pairs output by blocking |
| **Output** | A set of record blocks or “canopies,” each containing a small subset of records | A labelled set of pairs (match / non-match), often with confidence scores |
| **Typical Techniques** | • **Standard keys:** e.g. first character of last name, zip code• **Sorted Neighborhood**• **Canopy Clustering**• **Locality-Sensitive Hashing (LSH)** | • **Rule-based comparisons:** e.g. exact token, Jaro-Winkler thresholds• **Machine-learning classifiers:** logistic regression, random forest, SVM, neural nets• **Similarity scoring functions:** cosine on TF-IDF, edit-distance, embedding distance |
| **Computation** | Very cheap per record (hashing, sorting, simple distance) | Relatively expensive (feature extraction + classifier/scoring) |
| **Recall vs. Precision** | Targets **high recall** (few true matches dropped), may admit many false positives | Targets **high precision** (few false matches), trades off recall if thresholded too strictly |
| **Error Risk** | • **Missed matches** if two true-matching records never share a block | • **False positives** if classifier or thresholding isn’t tight enough |
| **Evaluation Metrics** | • **Pair completeness (PC):** proportion of true matches retained in blocks• **Reduction Ratio (RR):** fraction of candidate pairs eliminated | • **Precision / Recall / F1 / ROC-AUC** over the labelled pairs |
| **Role in Workflow** | 1. **Block** raw data → get candidate sets2. **Match** within each block → decide actual duplicates | — |