Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 13: Data cleaning for record linkage and blocking (1)

(Lecturer: Peter Christen)

Lecture outline

- Short review: The record linkage process
- Data pre-processing for record linkage
- Why blocking / indexing?
- Basic blocking techniques

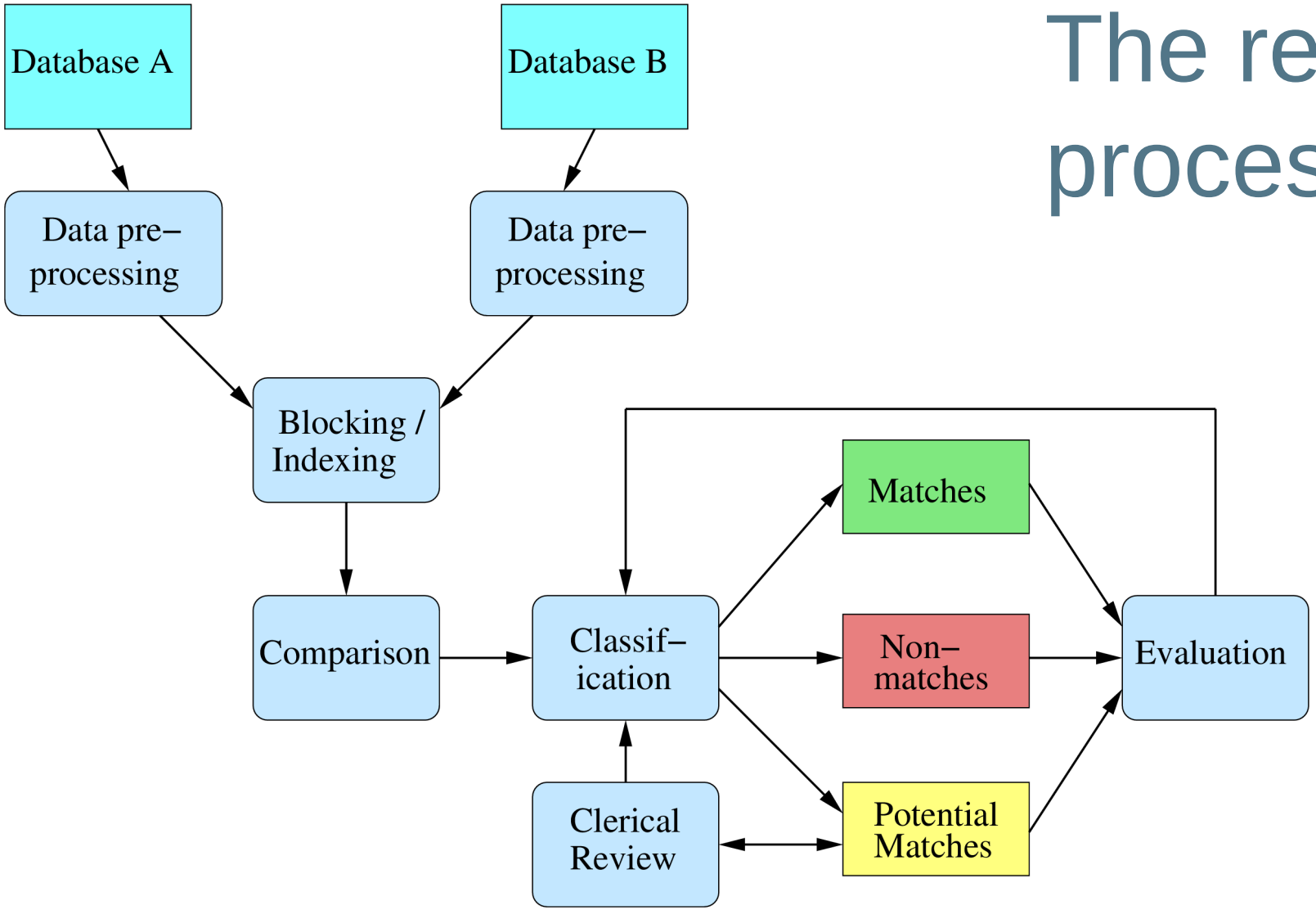
What is record linkage?

- The process of linking records that represent the same entity in one or more databases (patients, customers, businesses, products, publications, etc.)
- Also known as *data linkage*, *data matching*, *entity resolution*, *duplicate detection*, *object identification*, etc.
- Major challenge is that **unique entity identifiers** are not available in the databases to be linked (or if available, they are not consistent or not stable)

- For example, which of these records represent the same person?

Dr Smith, Peter 42 Miller Street 2602 O'Connor
Pete Smith 42 Miller St 2600 Canberra A.C.T.
P. Smithers 24 Mill Rd 2600 Canberra ACT

The record linkage process



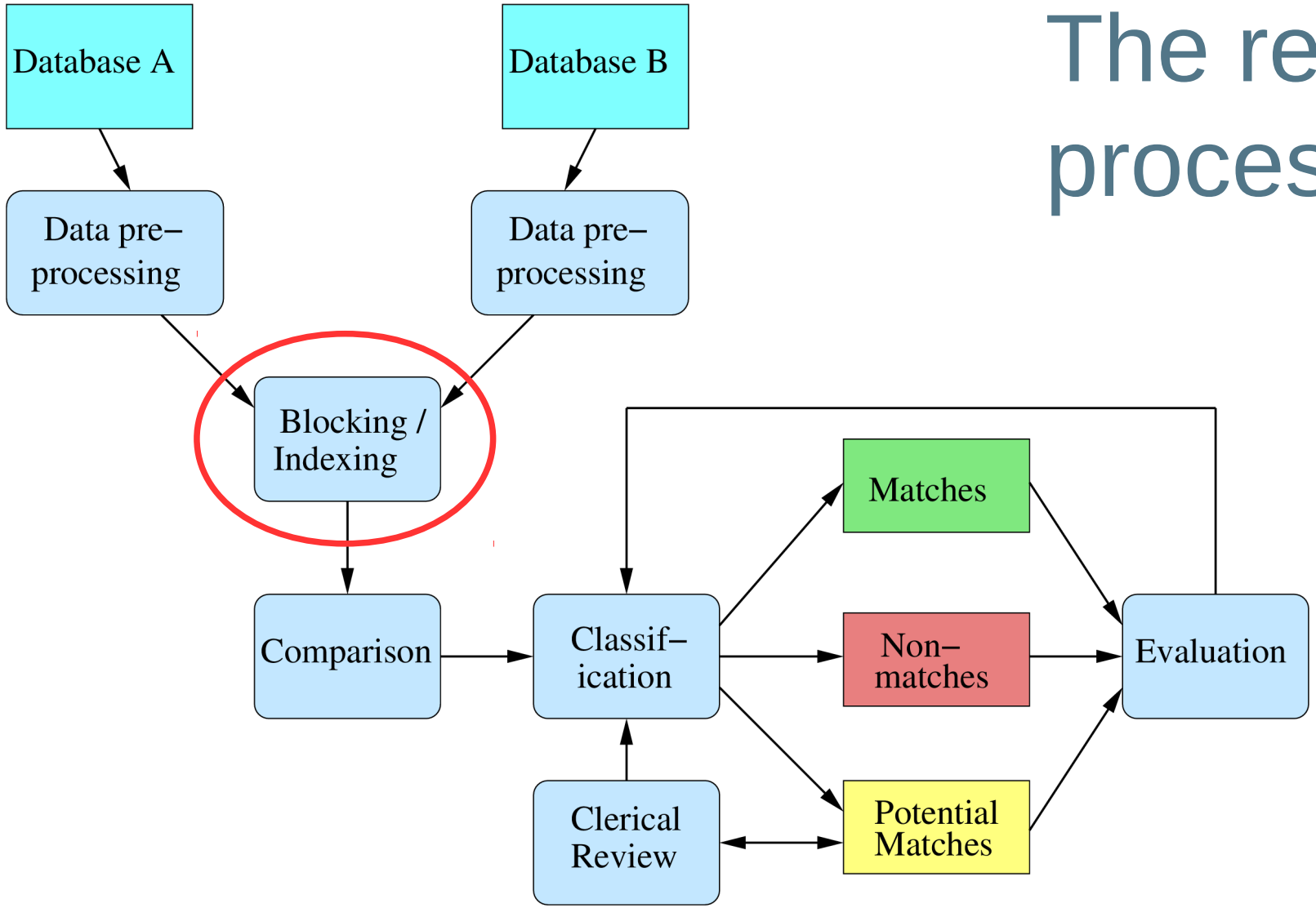
Data pre-processing for record linkage

- It is important to ensure the databases to be linked have similar structure and content
- Real world data are often dirty
  - Typographical and other errors
  - Different coding schemes
  - Missing values
  - Data changing over time
- Names and addresses are especially prone to data entry errors
  - Scanned, hand-written, recorded over telephone, and manually typed
  - Same person often provides her/his details differently
  - Different correct spelling variations for proper names (like ‘Gail’ and ‘Gayle’)

Data pre-processing steps

- Clean input:
  - Remove unwanted characters and words (“missing”, “n/a”, etc.)
  - Expand abbreviations and correct misspellings (“Wolllongonng” → “Wollongong”, “St” → “Street”, etc.)
- Segment names and addresses into well defined output fields (such as *title*, *first name*, *middle name*, *last name*, etc.)
- Verify correctness if possible
  - Do gender and first name match?
  - Does an address (or parts of it) exists in the real-world? (such as “Sydney NSW 7000”)

The record linkage process



Why blocking? (1)

- The number of record pair comparisons equals the product of the sizes of the two data sets

- For example: Matching two data sets containing 1 and 5 million records will result in 1,000,000 × 5,000,000 = 5 x 10^{12} record pair comparisons (one trillion = one million million)

- **Question**: *How many record pair comparisons are there for the deduplication of a single database with 1 million records?*

Why blocking? (2)

- Performance bottleneck in a record linkage system is usually the (expensive) detailed comparison of attribute (field) values between record pairs
  (such as approximate string comparison functions)
- Blocking / indexing / searching / filtering techniques are used to reduce the large amount of record pair comparisons
- **Aim of blocking**: Cheaply remove candidate record pairs which are obviously not matches (i.e. are *non-matches*)

Traditional blocking (1)

- Traditional blocking works by only comparing record pairs that have the same value for a *blocking key* variable
- One or more attributes in the databases to be linked are selected as blocking keys (and attribute values are concatenated)
- All records that have the same *blocking key value* (BKV) will be inserted into the same block
- Candidate record pairs will be formed from blocks where their BKV occurs in both databases

- **Questions**: *What are the criteria for good blocking keys?*

Traditional blocking (2)

- Problems with traditional blocking:
  - An erroneous value in a blocking key variable results in a record being inserted into the wrong block
  - Attributes that are know to change will mean true matching record pairs are potentially missed (for example, *surname* or *postcode*)
  - Missing values mean BKVs cannot be generated
  - Frequency distributions of BKVs influence number of record pairs generated
- **Question**: *Assume the number of surname values "Smith" in two databases is 10,000 and 25,000; while the number of "Dijkstra" is 4 and 5. How many record pairs are generated from each?*

Traditional blocking (3)

- Attributes selected as blocking keys should:
  - Not change over time (i.e. be constant)
  - Be accurate (no errors or variations in them)
  - Be complete (no missing values)
  - Have a frequency distribution close to uniform
  - Have a reasonable number of unique different values

- **Questions**: *Which one is better as a blocking key?*
  Gender or social security number?
  What are examples of attributes suitable as blocking keys?