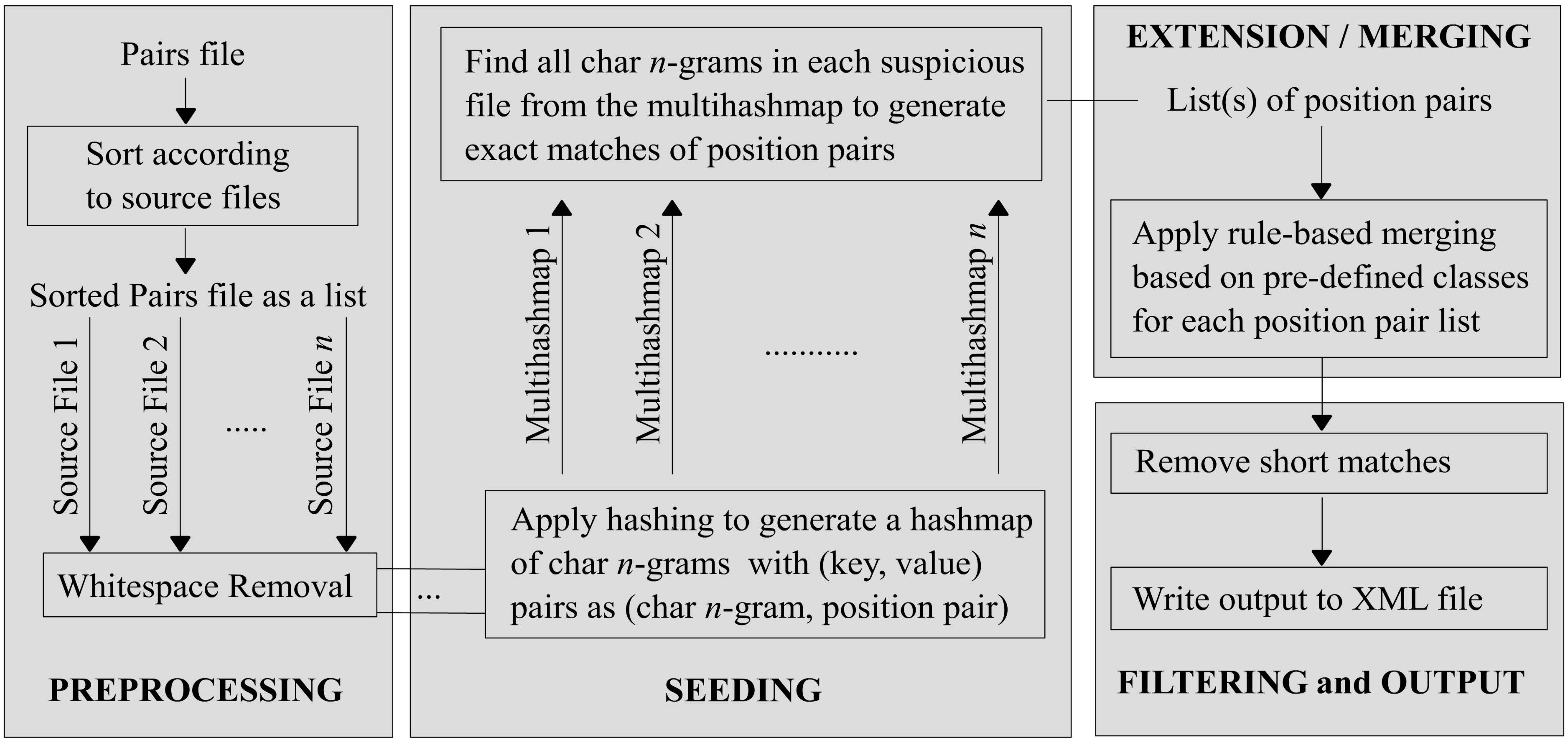
Hashing and Merging Heuristics for Text Reuse Detection

Faisal Alvi, Mark Stevenson, Paul Clough

King Fahd University of Petroleum and Minerals, Saudi Arabia, & University of Sheffield, United Kingdom.

Software Design: We use the three step approach of seeding, merging and filtering along with some preprocessing.



PREPROCESSING

Pairs file → Sort according to source files → Sorted Pairs file as a list → Source File 1, Source File 2, ..., Source File *n* → Whitespace Removal

SEEDING

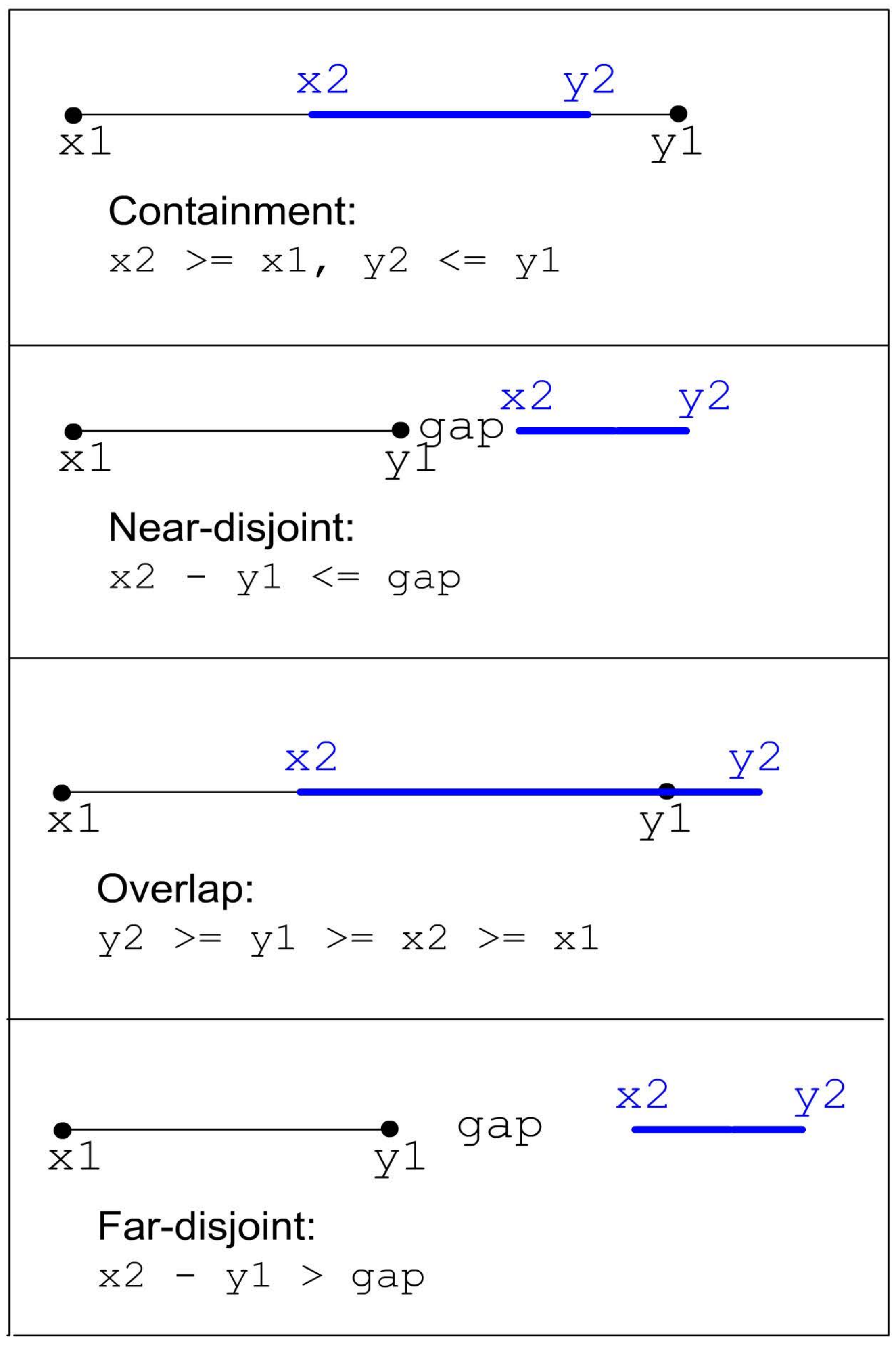
Apply hashing to generate a hashmap of char *n*-grams with (key, value) pairs as (char *n*-gram, position pair) → Multihashmap 1, Multihashmap 2, ..., Multihashmap *n* → Find all char *n*-grams in each suspicious file from the multihashmap to generate exact matches of position pairs

EXTENSION / MERGING

List(s) of position pairs → Apply rule-based merging based on pre-defined classes for each position pair list

FILTERING and OUTPUT

Remove short matches → Write output to XML file



Containment:
x2 >= x1, y2 <= y1

Near-disjoint:
x2 - y1 <= gap

Overlap:
y2 >= y1 >= x2 >= x1

Far-disjoint:
x2 - y1 > gap

Details: For seeding we use character 20-grams with Rabin-Karp Algorithm for multiple pattern search using a multihashmap (below). The seed-pairs are classified into four types as shown (left). A list of rules is then used to merge the individually found seed-pairs into contiguous passages (far below).

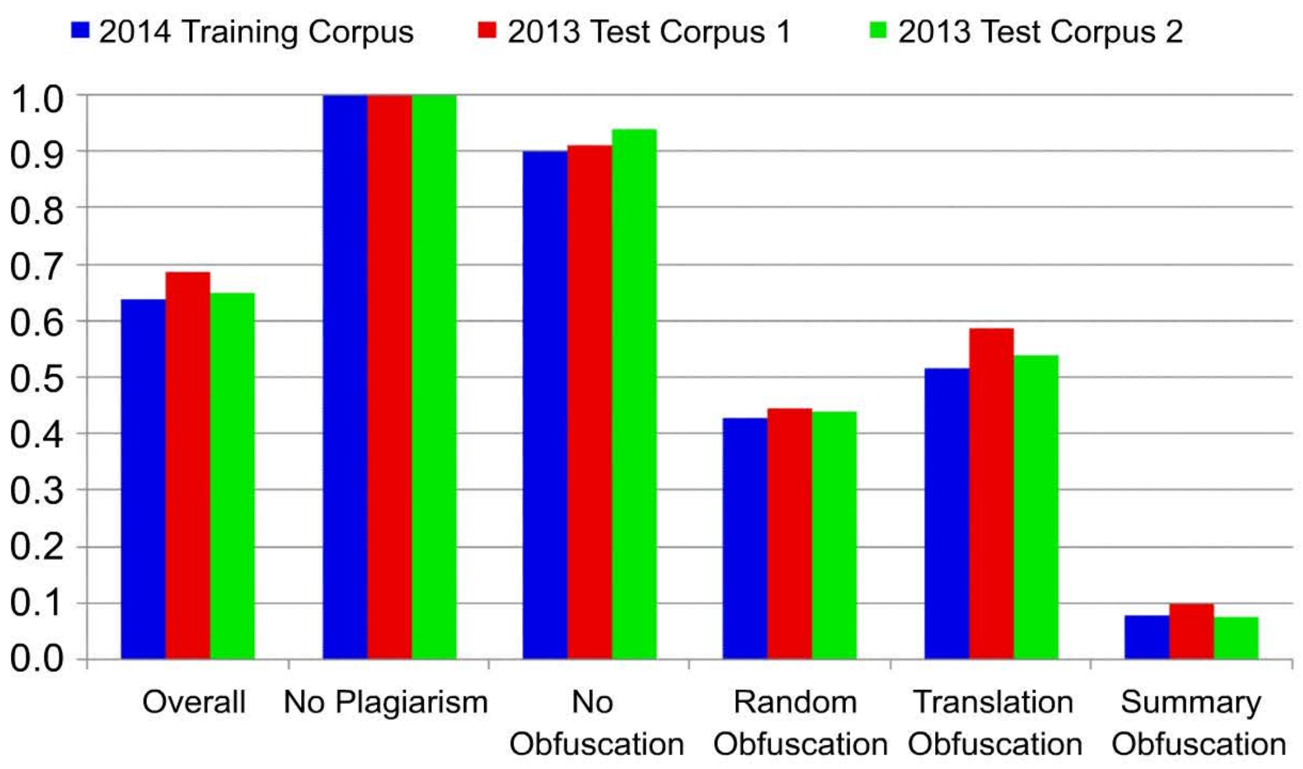
source-document01999

What are the **symptoms of arrhythmia**s? The effects on the body are often the same, however, whether the heartbeat is too fast, too slow, or too irregular. Some **symptoms of arrhythmia**s include, but are not limited to: weakness The **symptoms of arrhythmia**s may resemble other conditions.

Mutiple Valued Hash Map

KEY	VALUE
char n-gram	(start, end, size)
symptomsof arrhythmia	$(x_1, y_1, size_1)$, $(x_2, y_2, size_2)$, $(x_3, y_3, size_3)$.
theeffects onthebodya	....

Results: The software system scored an overall plagdet score of **0.65954** on the 2014 evaluation corpus and **0.73416** on the supplementary corpus.



Relationship in Source $(x_1, y_1, s_1), (x_2, y_2, s_2)$	Relationship in Suspicious $(a_1, b_1, s_1'), (a_2, b_2, s_2')$	Replacement Action Replacement 3-tuple
Containment: (x_1, y_1, s_1) contains (x_2, y_2, s_2)	Containment	$(x_1, y_1, s_1) \rightarrow (a_1, b_1, s_1')$
	Overlap	$(x_1, y_1, s_1) \rightarrow (a_1, b_2, b_2 - a_1)$
	Near-disjoint	No change (Term Repetition likely)
	Far-disjoint	Merging not possible
Overlap: (x_1, y_1, s_1) overlaps (x_2, y_2, s_2)	Containment	$(x_1, y_2, y_2 - x_1) \rightarrow (a_1, b_1, s_1')$
	Overlap	$(x_1, y_2, y_2 - x_1) \rightarrow (a_1, b_2, b_2 - a_1)$
	Near-disjoint	No change (Term Repetition likely)
	Far-disjoint	Merging not possible
Near-disjoint: $x_2 - y_1 \leq gap$	Containment	No change (Term Repetition likely)
	Overlap	No change (Term Repetition likely)
	Near-disjoint	$(x_1, y_2, y_2 - x_1) \rightarrow (a_1, b_2, b_2 - a_1)$
	Far-disjoint	Merging not possible
Far-disjoint $x_2 - y_1 > gap$	Containment	Merging not possible
	Overlap	
	Near-disjoint	
	Far-disjoint	