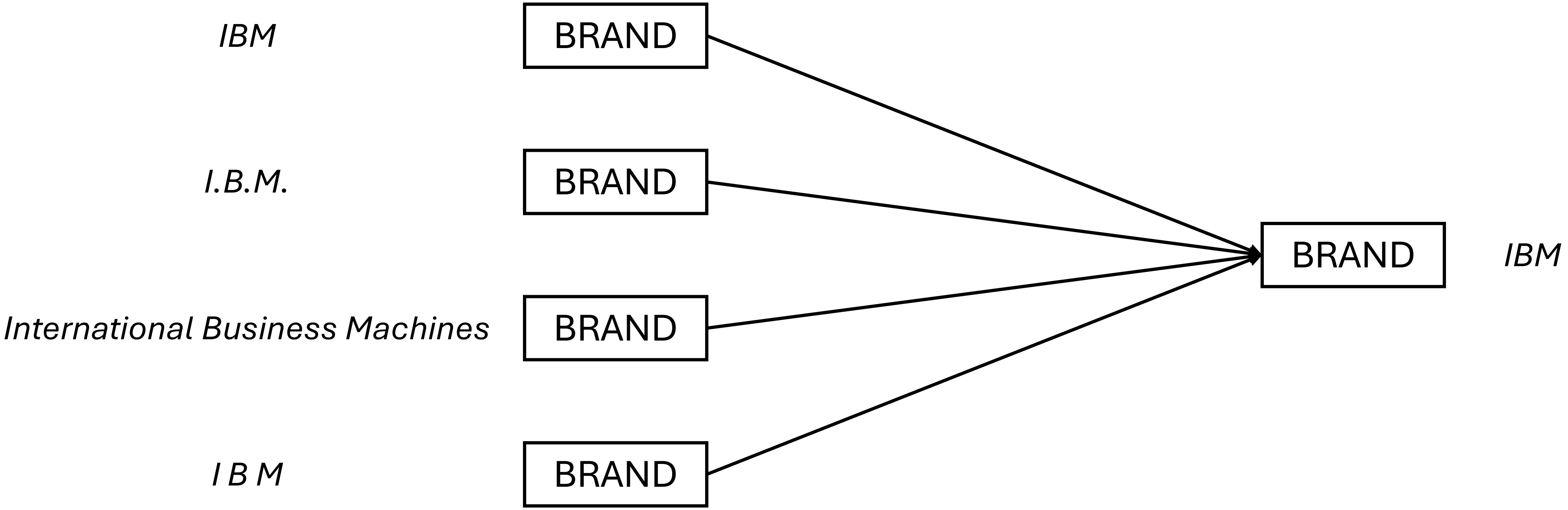
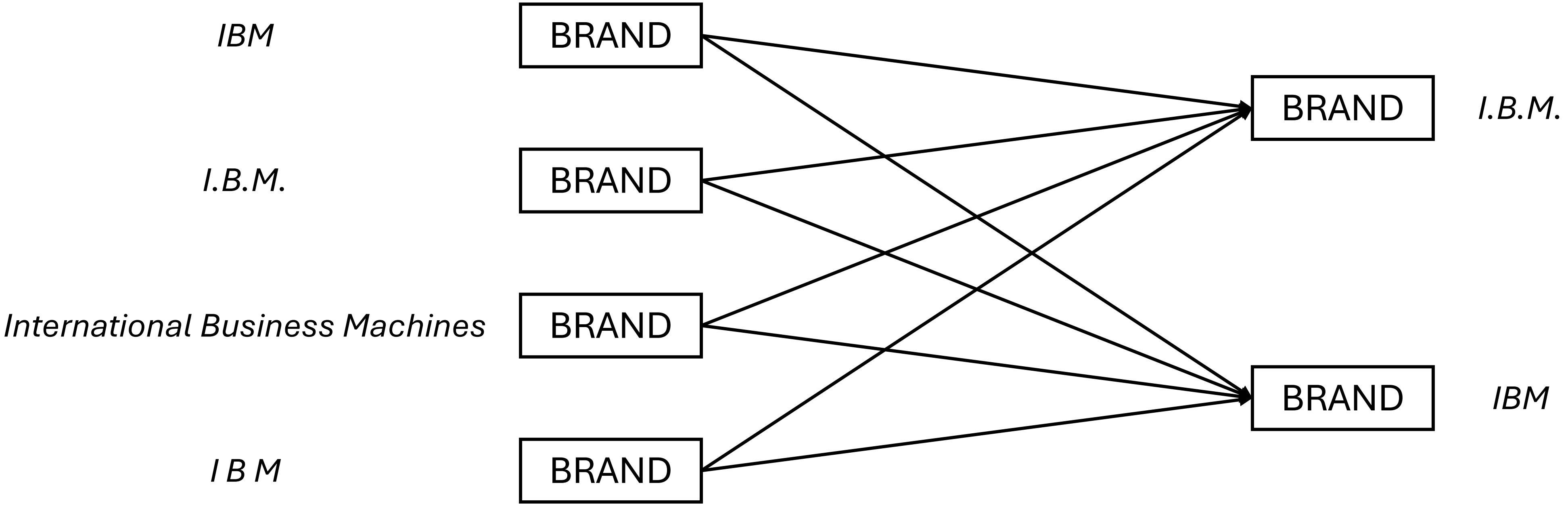
Entity name resolution

- Entity name resolution is the process of aggregating a number of similar strings representing the same entity and associating them with a single name (i.e., canonical and preferred)



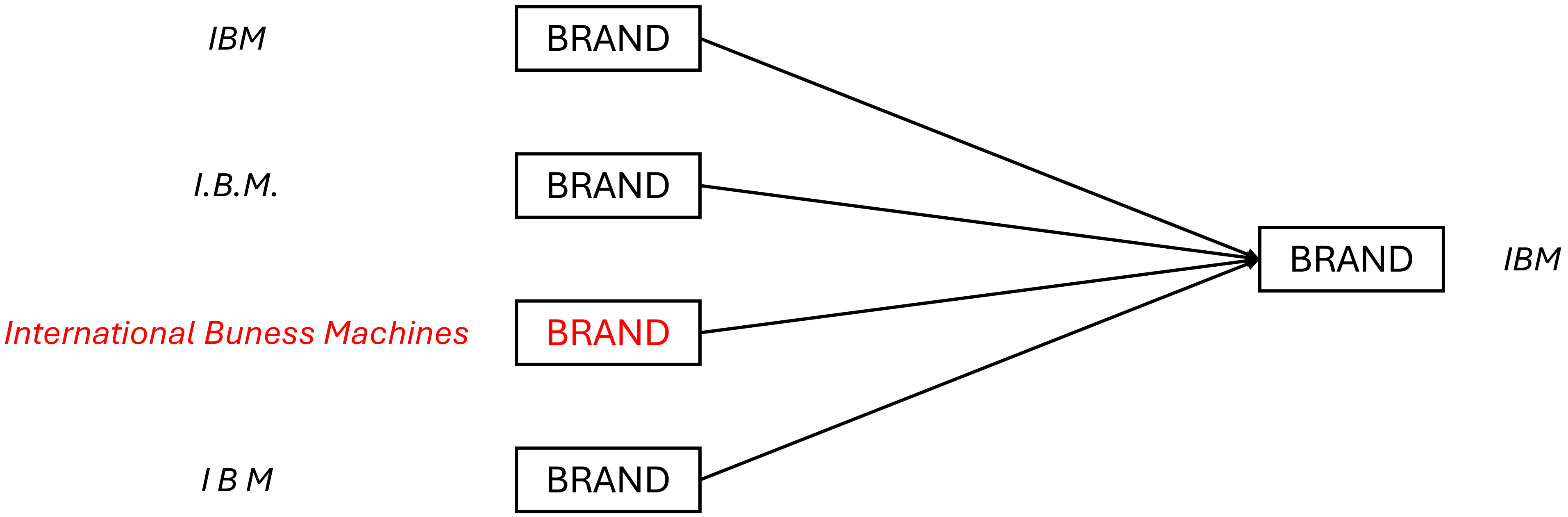
Entity name resolution problems (1/4)

- Canonical ambiguity



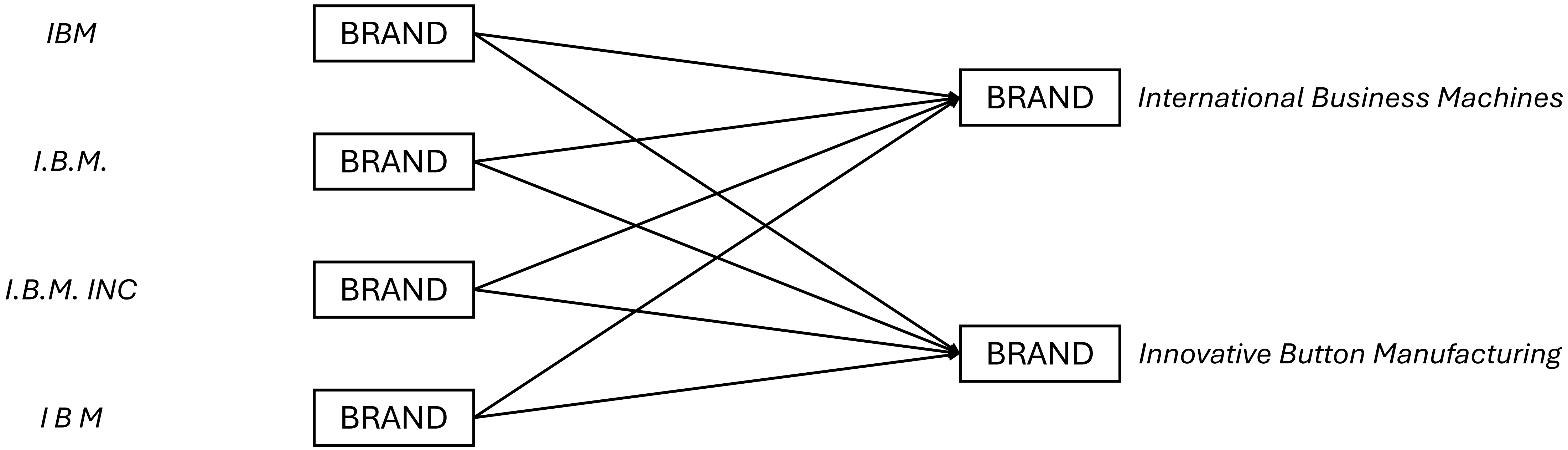
Entity name resolution problems (2/4)

- Misspelled entities



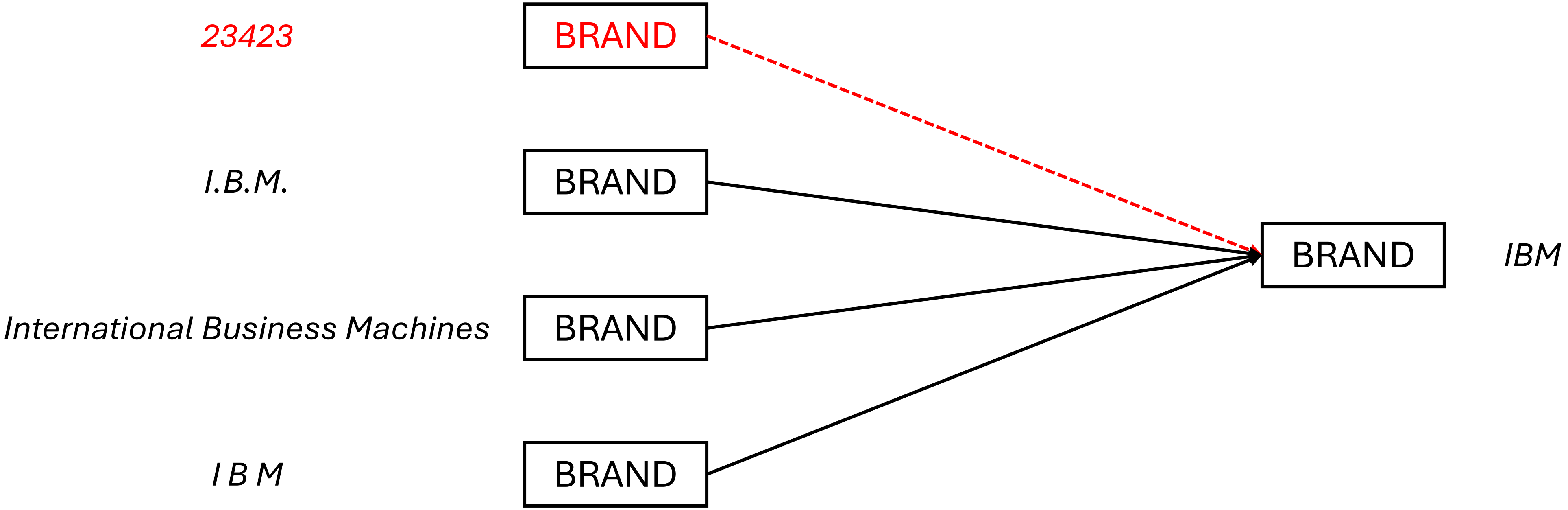
Entity name resolution problems (3/4)

- Resolution ambiguity


IBM
BRAND
I.B.M.
BRAND
I.B.M. INC
BRAND
I B M
BRAND
BRAND
International Business Machines
BRAND
Innovative Button Manufacturing

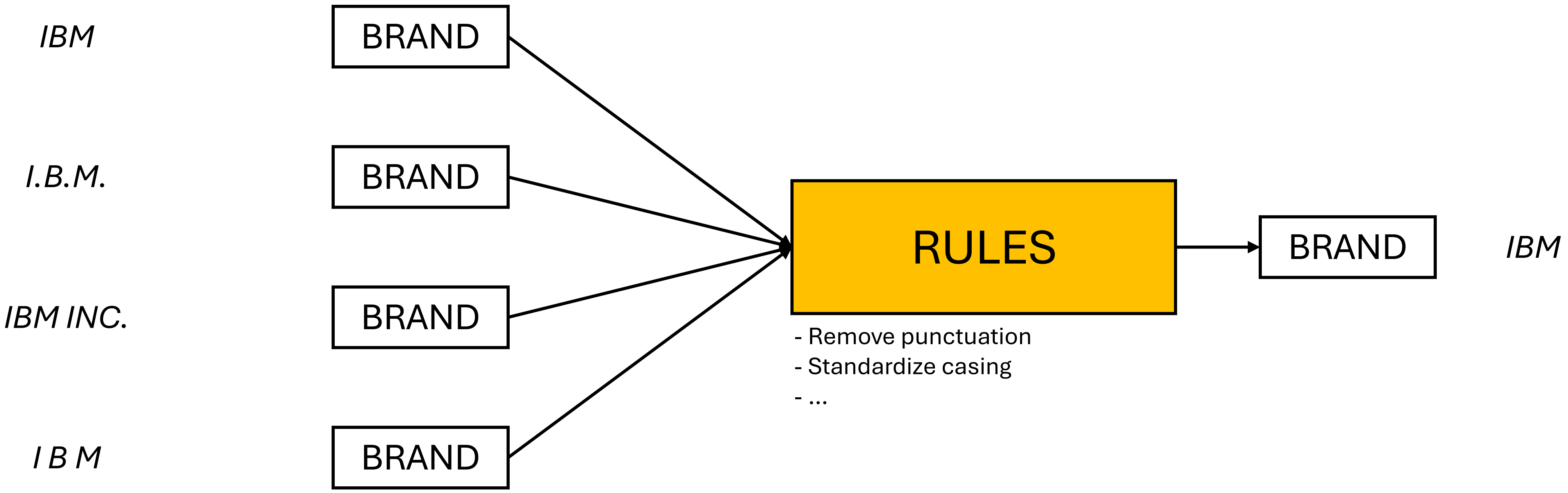
Entity name resolution problems (4/4)

- Invalid data



1. Canonicalization

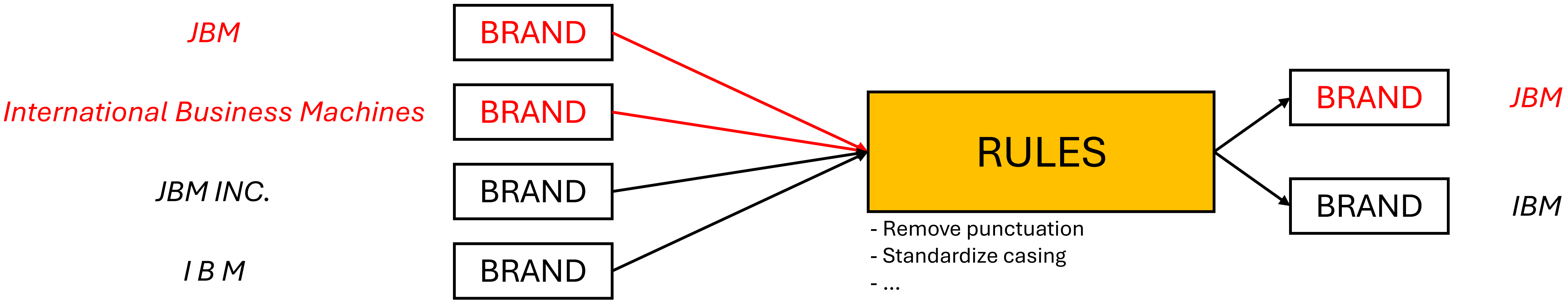
- Canonical clustering maps variations to a single entity name
- To this end, a series of rules can be applied to standardize names



- Remove punctuation
- Standardize casing
- ...

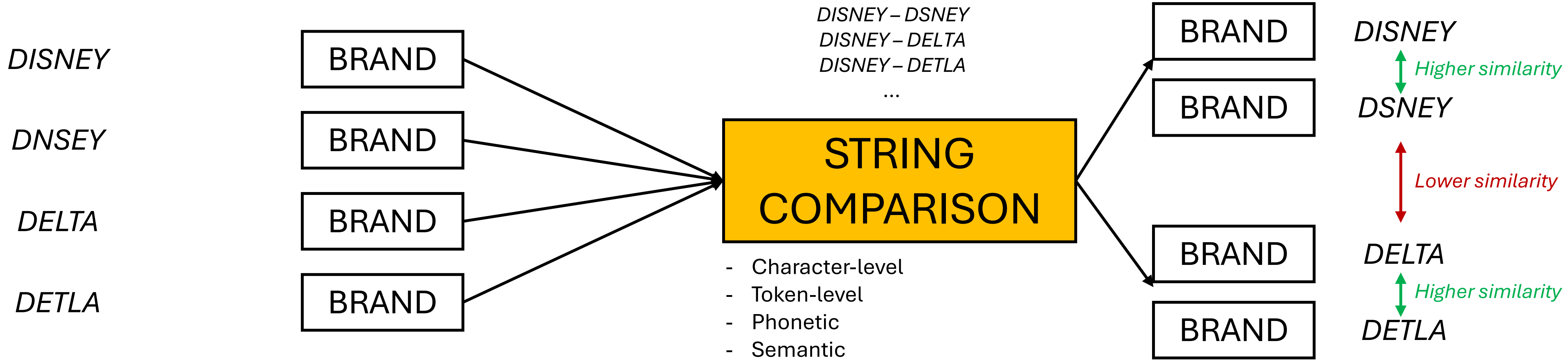
1. Canonicalization: problems

- Canonical clustering maps variations to a single entity name
- To this end, a series of rules can be applied to standardize names
- Rules fail to capture spelling errors
- Semantic matching requires additional knowledge
- Canonical clustering does not solve entity resolution ambiguity



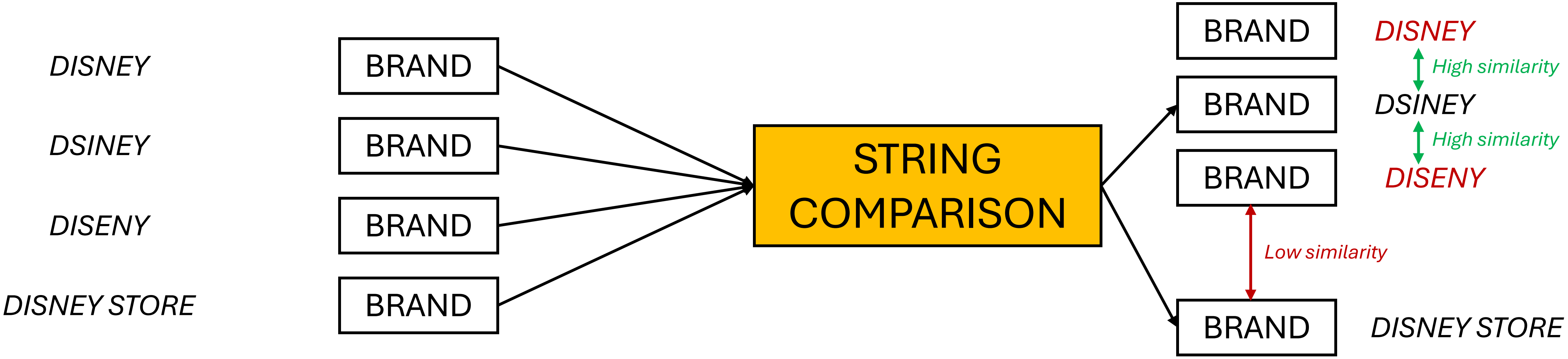
2. Similarity calculation

- Compares pairs of entities and outputs a similarity score
- Useful for resolving spelling errors and aggregating entities



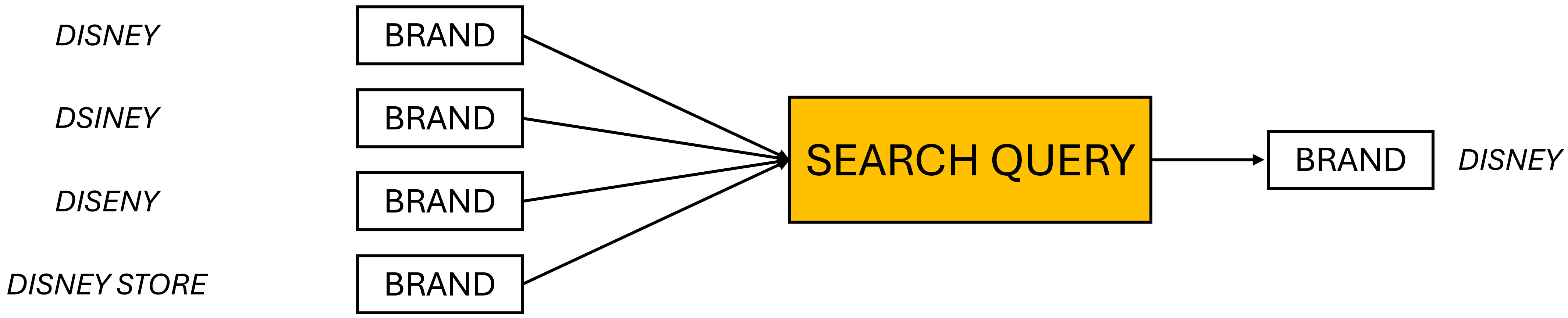
2. Similarity calculation: problems

- Compares pairs of entities and outputs a similarity score
- Useful for resolving spelling errors and aggregating entities
- Semantic matching requires additional knowledge
- Canonical clustering does not solve entity resolution ambiguity



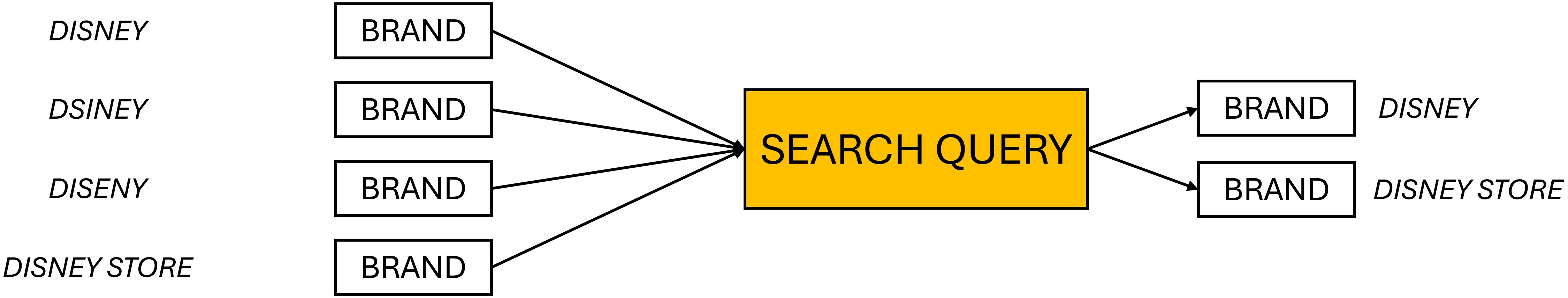
3. Validation

- Decide which entity name is the golden (i.e., actual) one
- Uses external knowledge (e.g., databases, search engines)



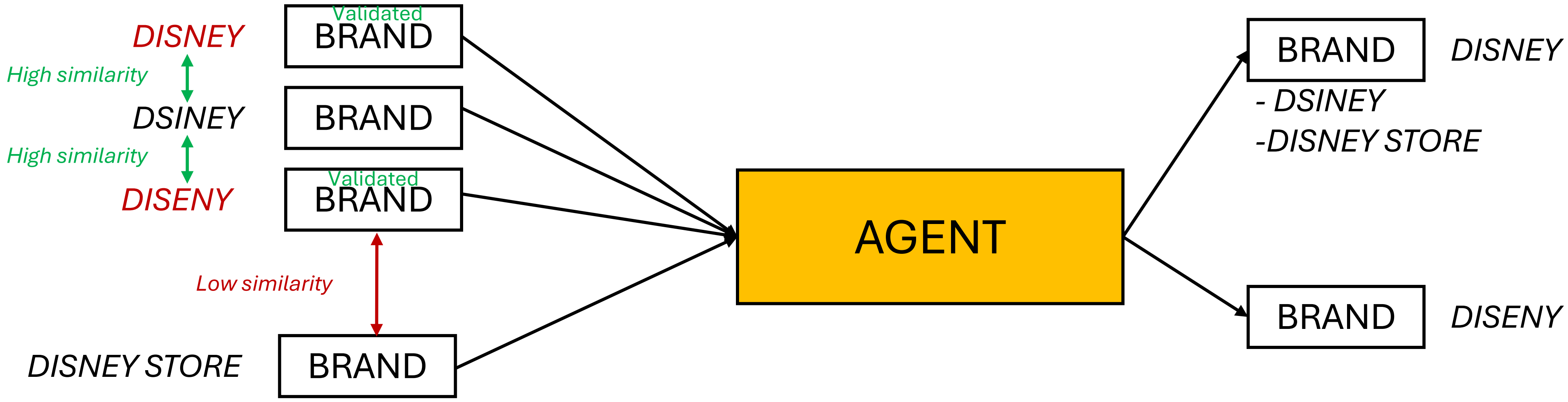
3. Validation: problems

- Decide which entity name is the golden (i.e., actual) one
- Uses external knowledge (e.g., databases, search engines)
- Databases can contain ambiguous information
- Results might be different from preferred display name

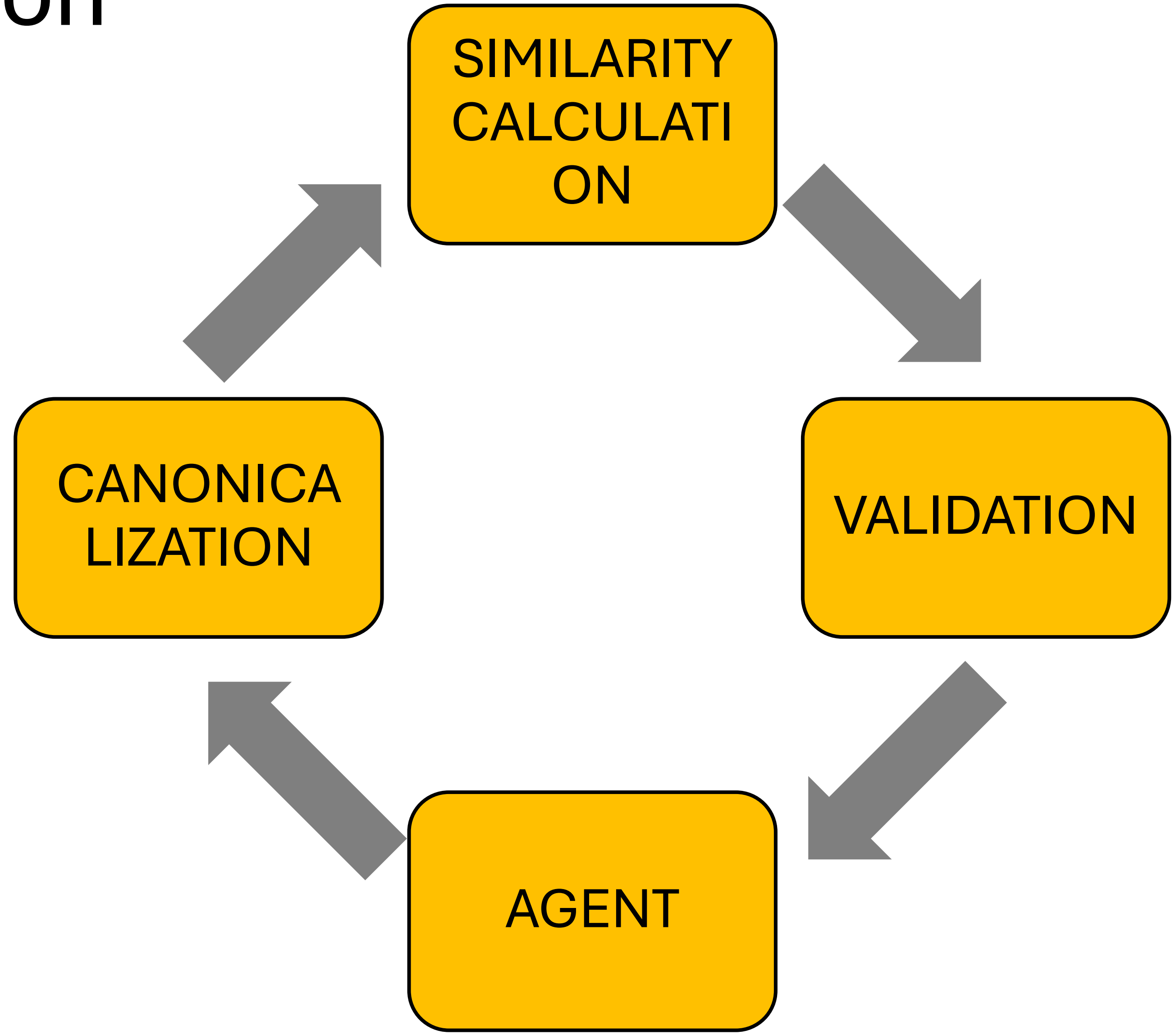


4. Manual/Automated review

- Use data from the previous steps to make informed decisions
- Time consuming
- Databases can contain ambiguous information
- Results might be different from preferred display name



5. Iteration


SIMILARITY CALCULATI ON
VALIDATION
AGENT
CANONICA LIZATION