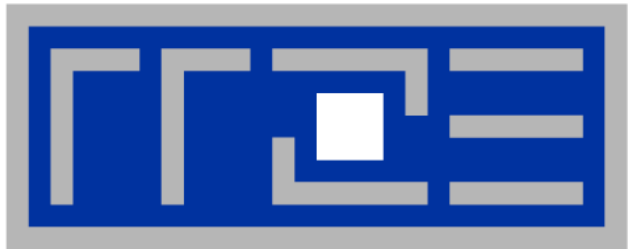


Data Linkage in IdM Systems - Revised

M.Sc. Wi.-Ing, M.Sc. CE

Krasimir Stoyanov Zhelev

Regionales RechenZentrum Erlangen

Chief Software Architect

Projects & Processes

06. October 2009

Content

- Data Linkage System
  - Overview
  - Problematic
  - Process
- Reporting
- Data Mapping
  - Standardization / Normalization
  - Rules
  - Data Sets
- Blocking
  - Overview
  - Types
- Statistics

Content II

- Matching
  - Attribute Comparison
  - Name Comparison
  - Similarity Functions
  - Process
  - Business Rule Engine
- Result Aggregation
- DaLi
  - Framework
  - Domain Model
  - DaLiG
- Conclusions

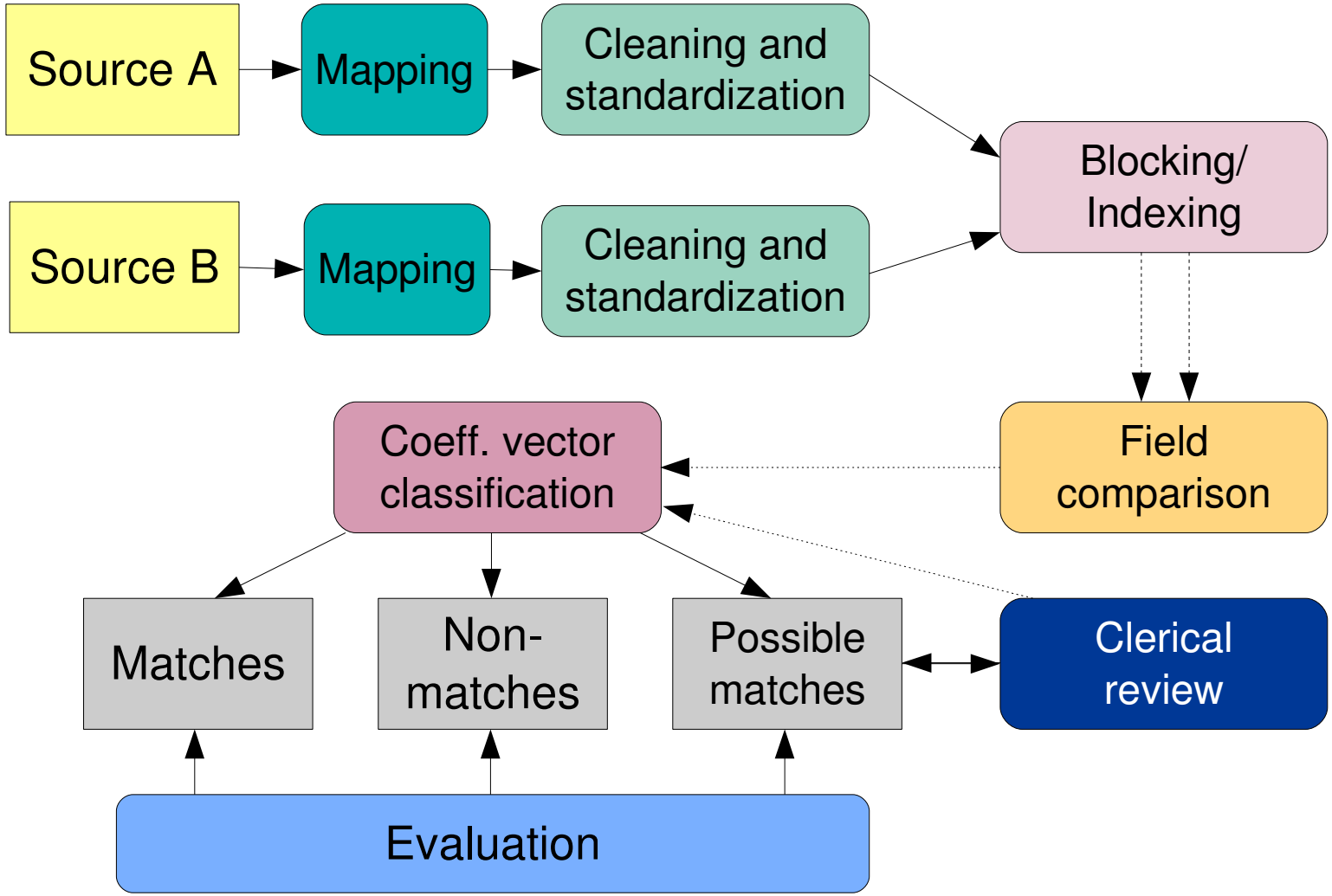
Data Linkage Systems - Overview

- **Goal:**
  - **Linking and/or aggregating data from the same or various sources that refers to the same entity in the case where no unique entities identifiers are available**
- **Reasons:**
  - **Internal de-duplication of data sources**
  - **Merging of different data sources**
  - **Improve data quality – clean up typos, …**
  - **Ensure data integrity – correct data in all systems**
  - **Extend existent data – fill in missing data from other systems**
  - **Provide basis for statistical evaluations - normalized**
  - **Support data mining**
  - **Geocode matching**

Problematic

- Unique identifiers are not available -> attributes matching
- Entity *mapping*:
  - Entities can have different cardinality
  - Attributes mapping is not always trivial – types, formats
- Large amounts of data should be processed
  - For two source A and B: $O(|A|x|B|)$
  - *Blocking* or *Filtering* has to be applied
- *Standardization, normalization* and *comparison* can be computationally expensive
- Classification of matching results - *matched* (confirmed match), *rejected* (confirmed reject), *unsure*, *pending*
- Automation is not feasible – exact matches do not exist
  - Black lists has to be maintained
- Privacy and confidentiality

Data Linkage Process

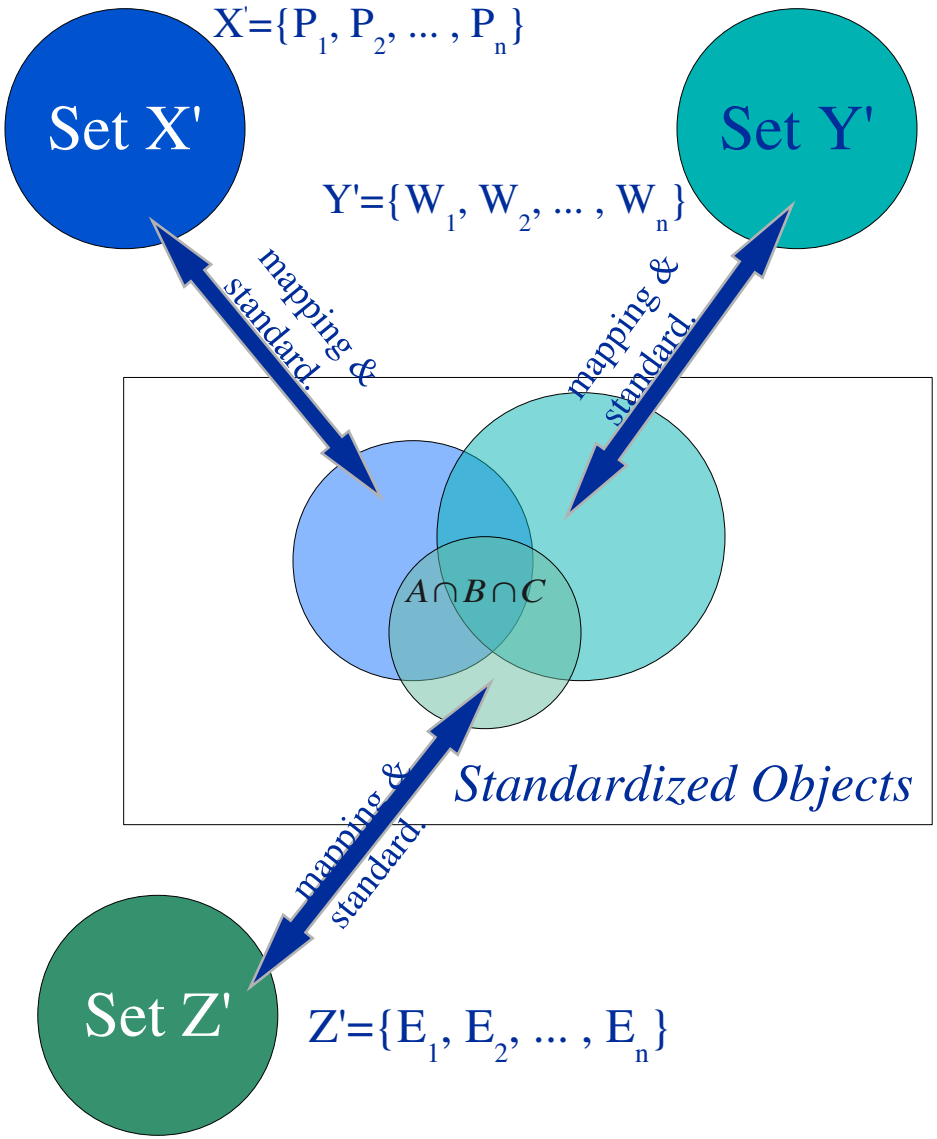


Reporting

- Statistics reports:
  - Frequency distribution reports – drill down
  - Frequency distribution reports pro source – drill down
- Internal duplicates:
  - Traditional blocking
  - Similarity blocking
- Attributes reports:
  - Empty values
  - Traditional mapping
  - Similarity blocking
- False positives and false negatives reports
  - Generated from clerk review lists
- Simulation results reports pro group
- Birt as a reporting engine

Mapping and Standardization(Ontology)

- **Different types of objects:**
  - **X: persons – P**
  - **Y: affiliations – W**
  - **Z: entitlements – E**
- **Mappings cardinality**
  - **one-to-one**
  - **one-to-many**
  - **many-to-one**
  - **many-to-many**
- **Different types of attributes**
  - **dates**
  - **names**
- **Data consistency**
  - **same semantics**
  - **same format**



Mapping & Standardization II

- Completeness Rule
  - as many attributes should be mapped as possible
  - allows cross system mappings
- Clarity Rule
  - Semantic definition of a Standardized Object
    - representation: $SO = \{A_1, A_2, \ldots, A_n\}$
    - usually by extending an existent type
    - proper attribute types should be selected
  - Attributes set definition
    - type – string, date, number
    - value – format and standardized form
  - Constraint definitions
    - imposed on the value of an attribute
    - related to the semantic meaning of the attribute
    - garbage data collection – date(01.01.1000), name

Data Sets

Ontology Overlapping

SO	ID	Source	A_1	A_2	...	A_n
Source X	121525	sos	Yes	Yes	...	No
Source Y	2118945	diapers	Yes	Yes	...	Yes

Weighted Ontology Overlapping

SO	ID	Source	A_1	A_2	...	A_n
Source X	121525	sos	0.9	0.75	...	0
Source Y	2118945	diapers	0.85	0.87	...	0.96

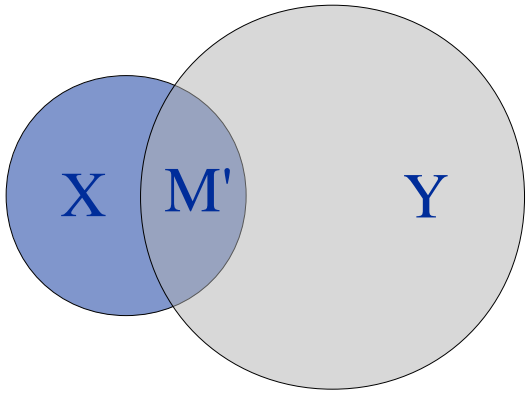
Case review

Typical case

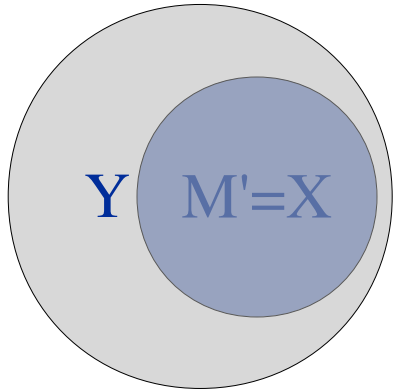
$$M = X \cap Y = \{SO_1, SO_2, \ldots, SO_m\}$$

Containment case

$$M = X ; M = X \cap Y = \{SO_1, SO_2, \ldots, SO_m\}$$



Typical

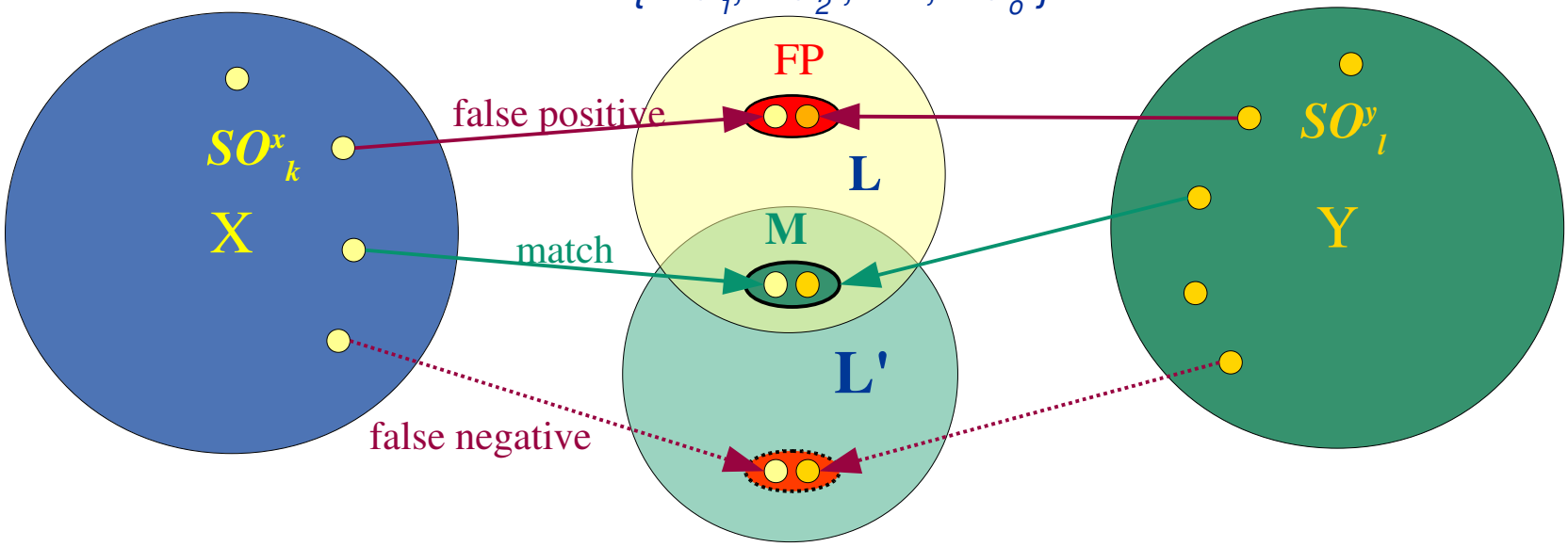


Containment

Data Sets Theoretically – Two Sources

- **Linkage Couple:** $LC = \{ SO^x_k , SO^y_l \}$
- **Subsets:** ***Matched*** **(M),** ***False Positives*** **(FP),** ***False Negatives*** **(FN)**
- **False Negatives can be found only by clerks**

$L = \{ LC_1, LC_2 , \dots , LC_o \}$

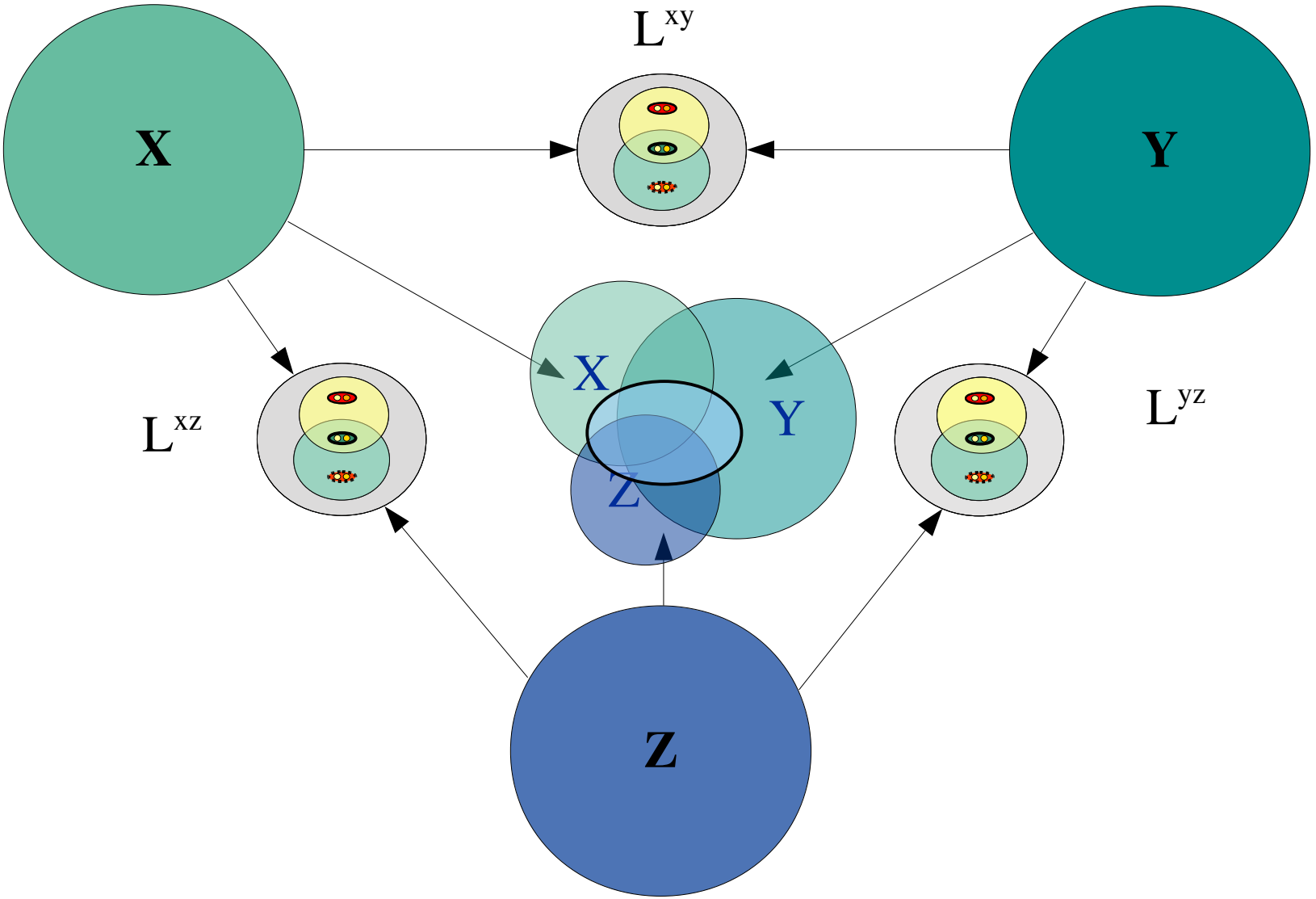


$X = \{ SO^x_1, SO^x_2 , \dots , SO^x_n \}$

$Y = \{ SO^y_1, SO^y_2 , \dots , SO^y_m \}$

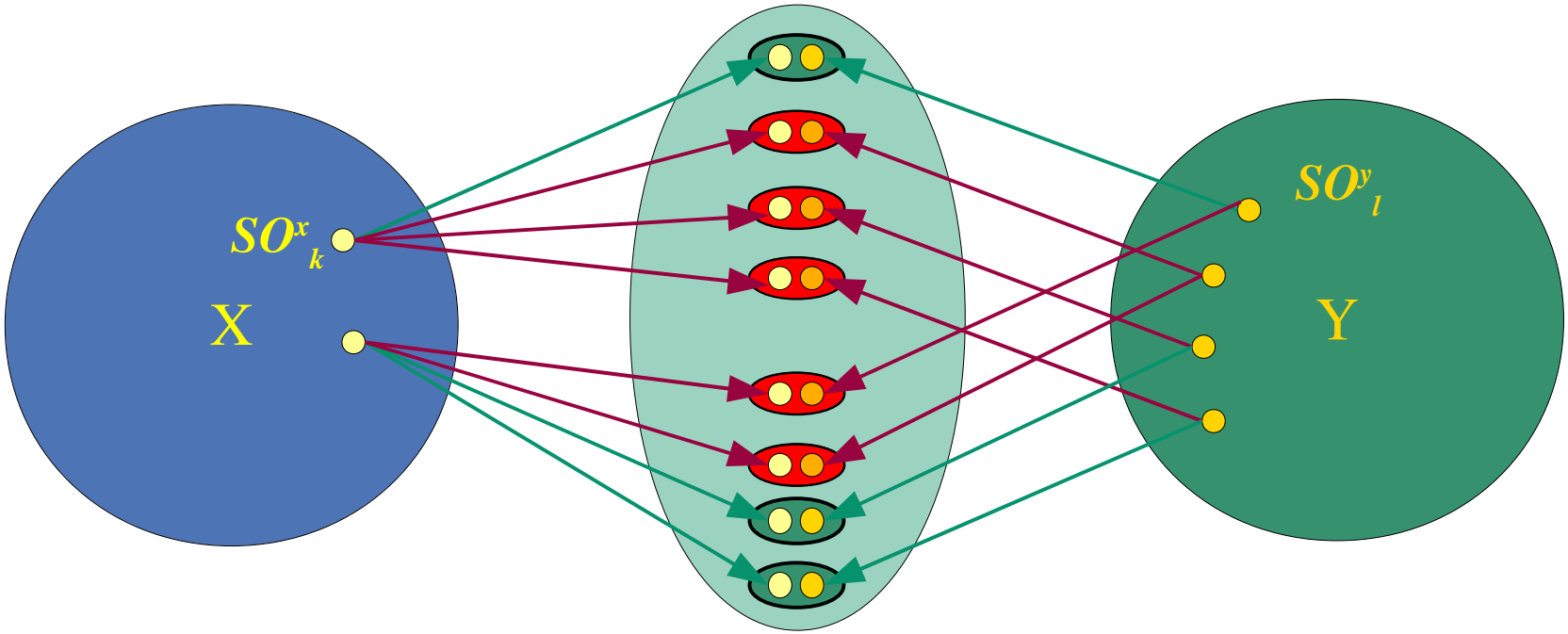
$L' = \{ LC_1, LC_2 , \dots , LC_p \}$

Data Sets Theoretically – Three Sources



Blocking - Overview

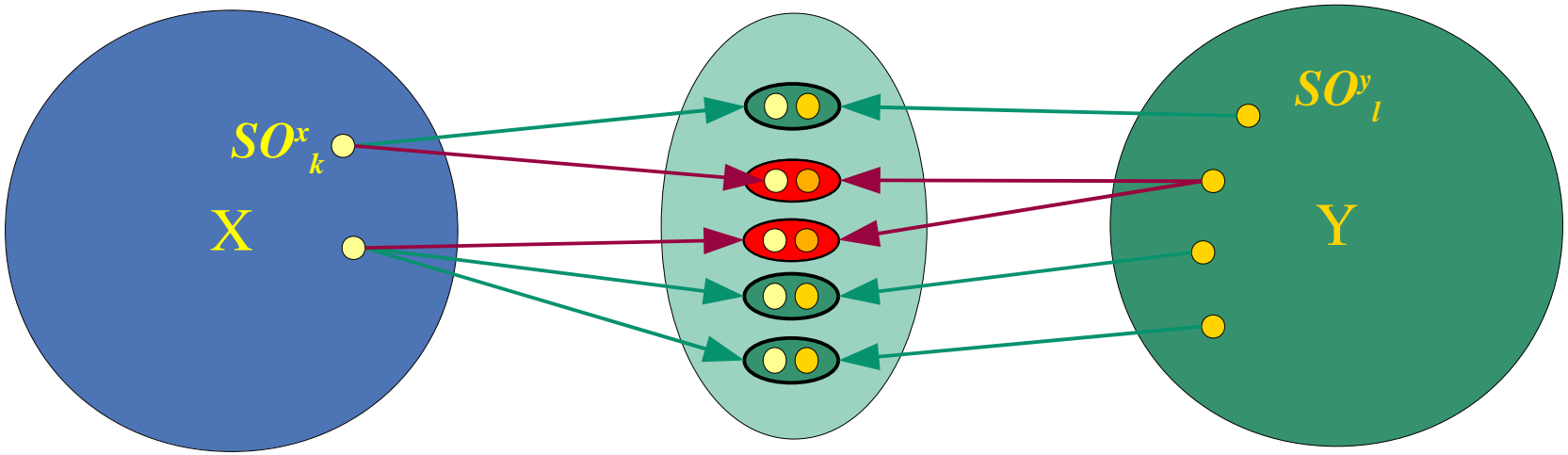
- **Blocking required because of large problem size - *O(|A|x|B|)***
- **Effectively reduce problem size by fast grouping/filtering**
- **Traditionally blocked variable(date of birth):**
  - **wrong value – groups entity in a wrong subset**
  - **uniformly distributed values**



Blocking - Types

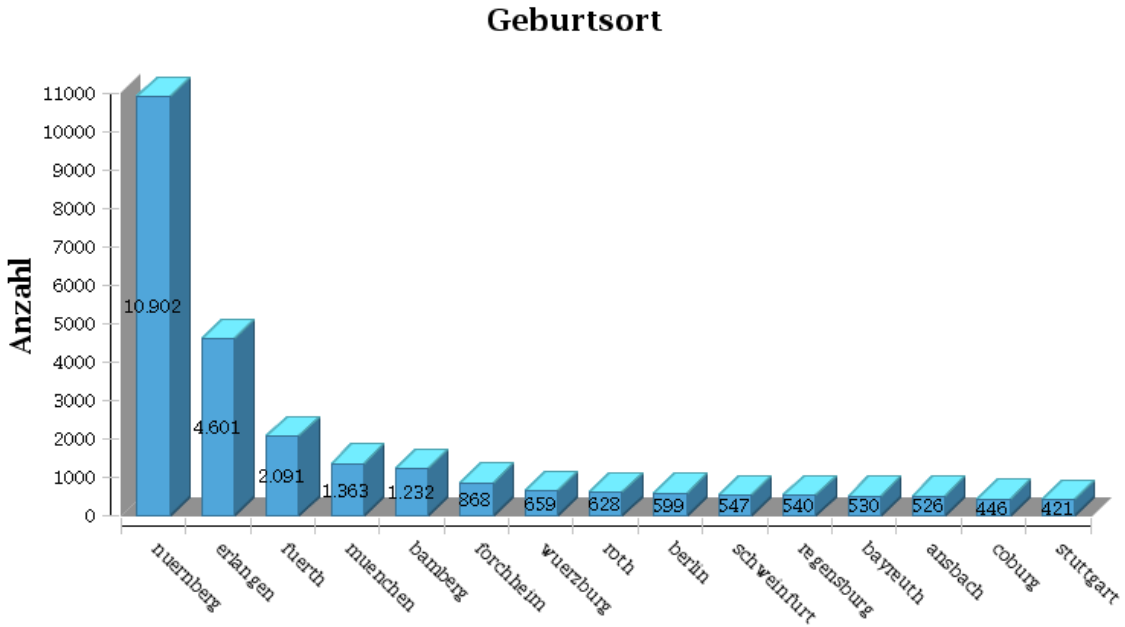
- Traditional blocking
- Sorted neighborhood blocking
- Q-gram blocking
- Similarity blocking

SIMILARITY_PLACEHOLDER(valueA, valueB) > THRESHOLD_PLACE_HOLDER



Statistics

- Not all attributes have same significance
- Generate frequency distributions:
  - from IDM system if such exists
  - from a leading system
  - pro source
- Normalized data should be used for statistics



Matching - Attribute Comparison

- Research shows: 80% of attribute errors are single errors
- Most common error types:
  - A letter was substituted for another letter
  - A letter is deleted
  - An extra letter is inserted
  - Two adjacent letters are transposed
- Errors according to data source
  - OCR – similar looking characters or sequences
  - keyboard – neighboring keys
  - telephone – assuming spelling
  - system limitations – max. length of input field
  - human factor – different reporting of data
- Different sources match worse

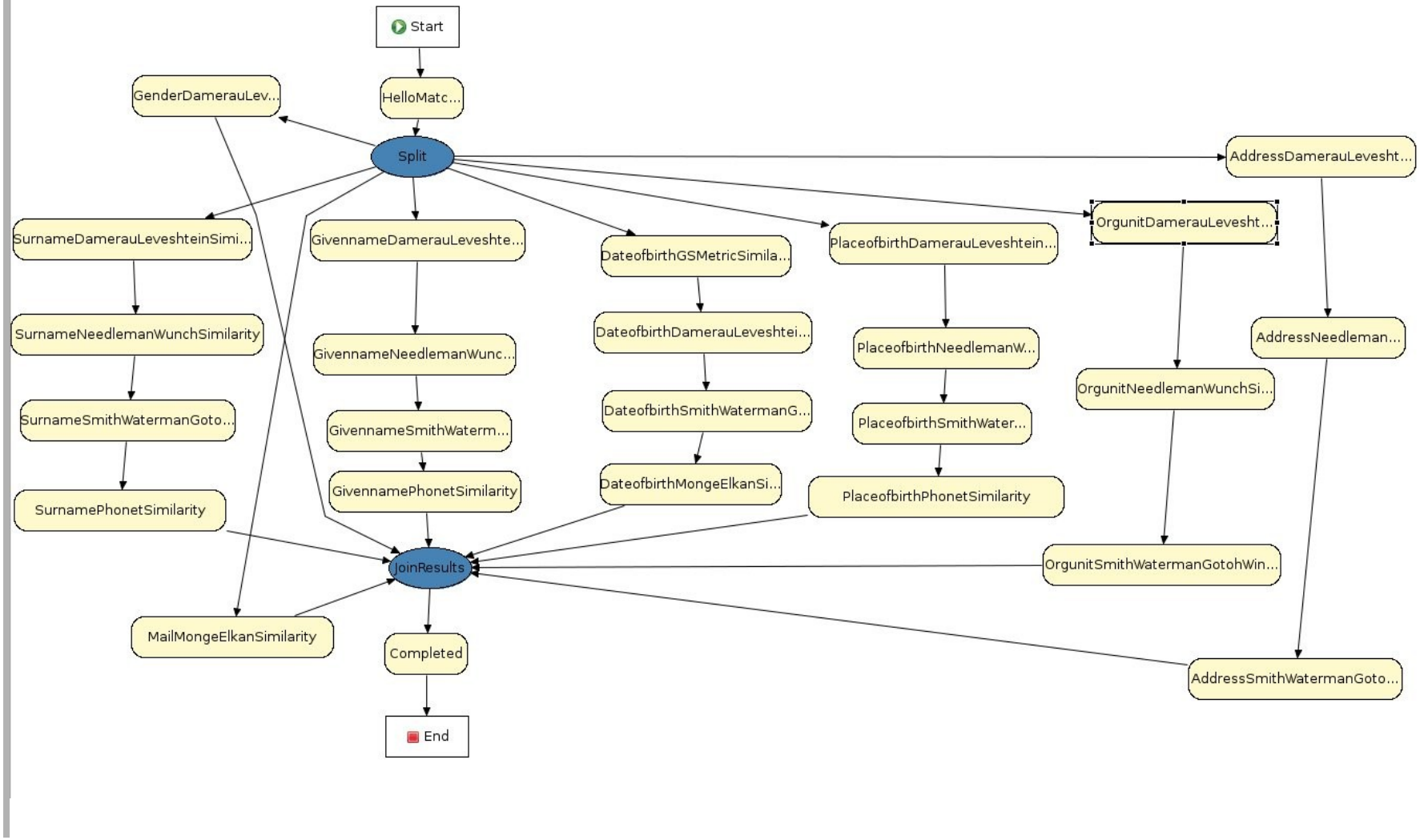
Matching - Name Comparison

- Generally there is no legislation on naming conventions
- Names have no correct spelling but rather a set of legitimate name variations
- Common problems:
  - Different spelling – Meier, Meyer, Maier
  - Different structure – middle name (Stoyanov, von …)
  - Nicknames, short names – (Wilhelm - Willi)
  - Names change – getting married, real name change
  - Compound names - (Hans-Peter)
  - Different transliterations – (Krassimir, Krasimir)
- Most important person related linkage attributes:
  - Name – first name, surname
  - Date of birth
  - Place of birth
  - Address

Matching - Similarity Functions

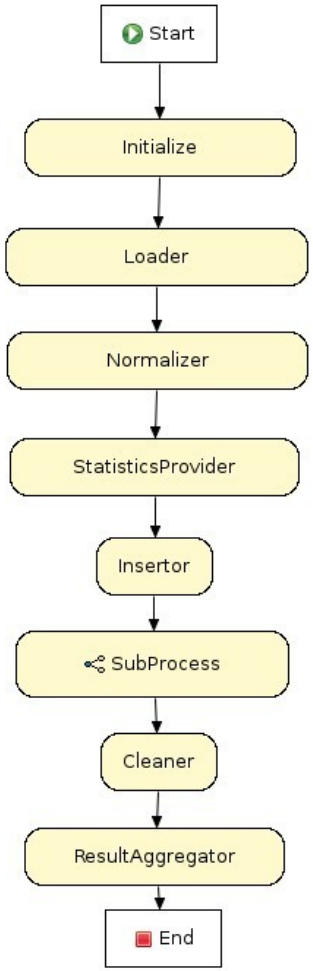
- Pattern Matching
  - Levenshtein – counts insertions, deletions and substitutions
  - Damerau-Levenshtein Distance – includes transpositions
  - Smith-Waterman – developed for DNA sequences
  - Jaro – also estimates transpositions
  - Jaro-Winkler – empirically improved Jaro for start of word
  - ...
- Phonetic Encoding
  - Soundex – keeps first letter encodes the others
  - Phonet – improved German version of Soundex
  - Phonix – different rules for start, middle, end of word
  - ...
- Combined
  - ...

Matching - Process



Matching - Business Rule Engine

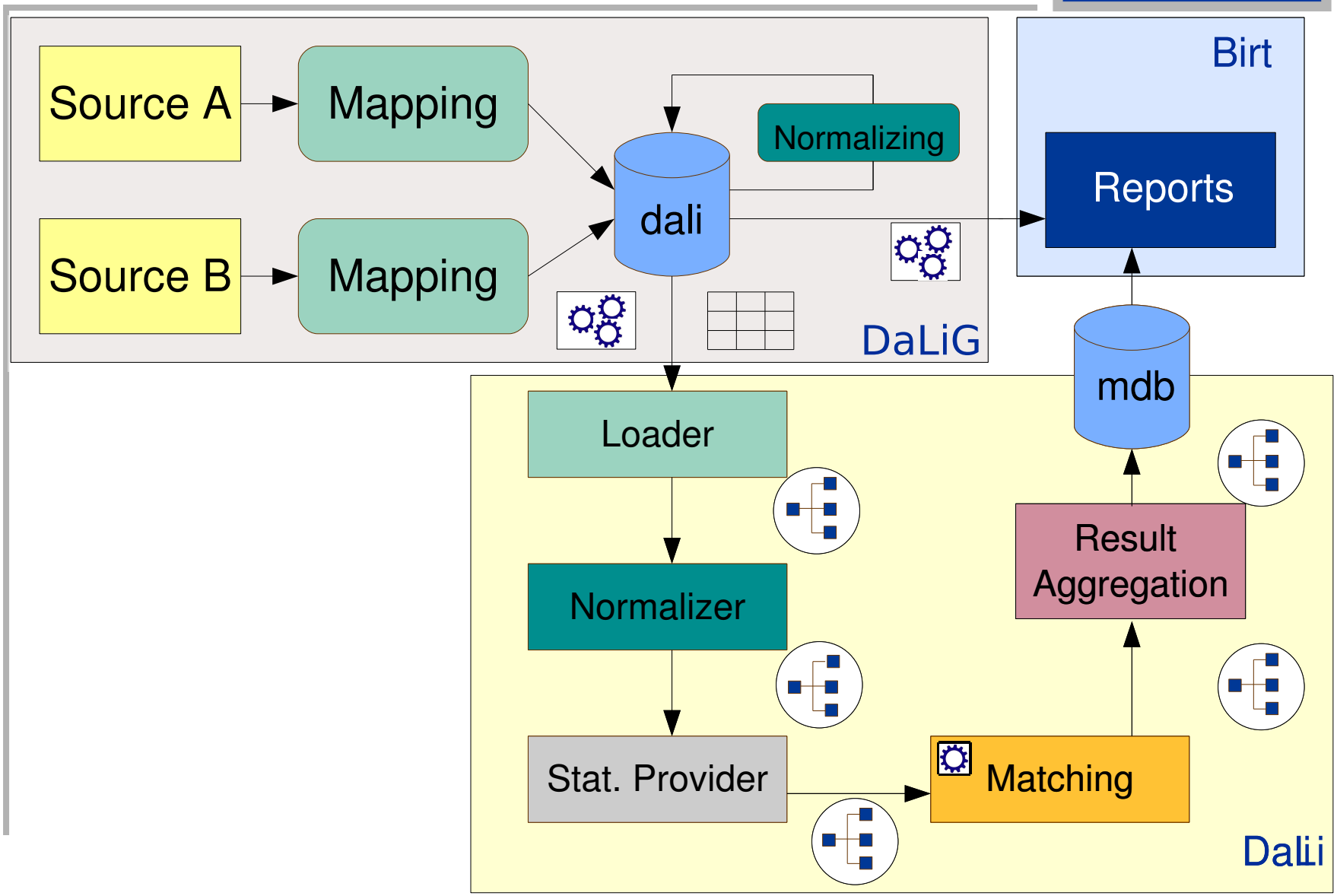
- **Business Rules Engine integration:**
  - **implementing a more complicated matching logic**
  - **investigating which *combinations* of similarity function is optimal on attribute basis**
  - **investigating which *order* of similarity function is optimal on attribute basis**
  - **rapid prototyping and evaluation of matching processes**
  - **evaluate blocking strategies**
  - **customization of the obtain results**
  - **appropriately handling system type – *initial load* or *realtime***



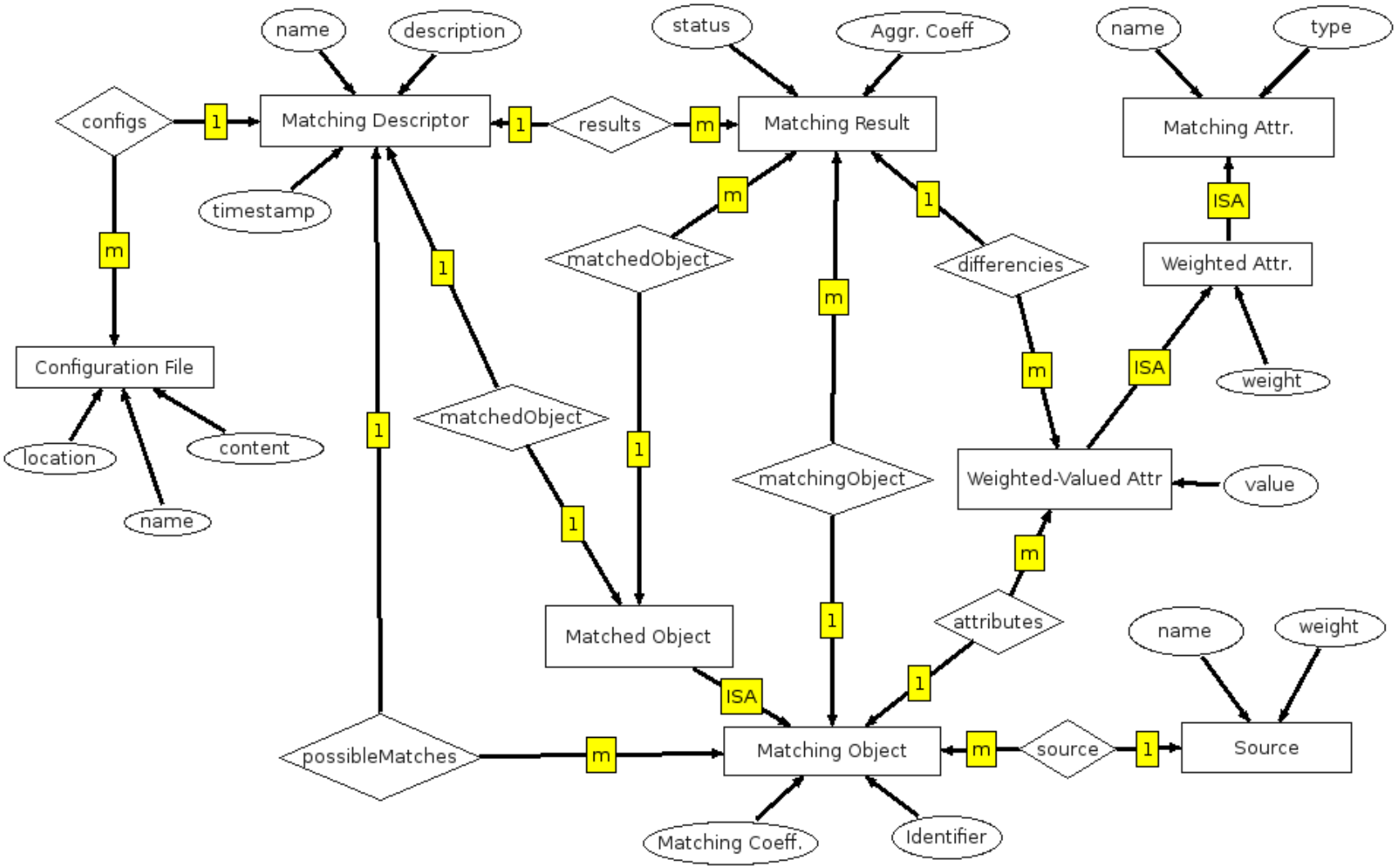
Result Aggregation

- Result aggregation can be complex:
  - data frequency distribution
  - weighting coefficients
  - number of errors
- Classification of matching results:
  - matched
  - rejected
  - unsure
  - confirmed match
  - confirmed reject
  - pending
- Clerk Lists:
  - Contain data for proven false positives
  - Contain data for proven false negatives

DaLi Framework



DaLi - Domain Model



Conclusion

- Data linkage is a complex and error prone process
- Gained experience so far:
  - It is important to know the specifics of the involved systems.
  - First fast approximation functions should be used to filter out possible negative positives.
  - Phonetic comparison should always be combined with an approximation function unless specifically searching for phonetic errors.
  - Data should be statistically enriched.
  - Significant effort should be allocated to tuning up thresholds and weighted coefficients
  - Business rule engine can be used to improve results.
- A framework is developed to allow the generation of various reports and testing of different scenarios

Thank You for the attention!

Contact

M.Sc. Wi.-Ing, M.Sc. CE

Krasimir Zhelev

Chief Software Architect

Projects & Processes

RRZE Martensstrasse 1

D-91058 Erlangen

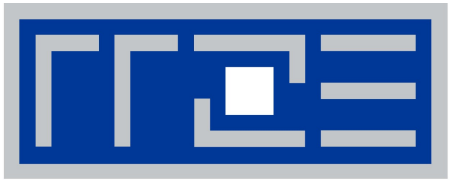
Tel.: +49 9131 85-28145

Fax: +49 9131 302941

krasimir.zhelev@rrze.uni-erlangen.de

http://www.rrze.uni-erlangen.de

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)



Regionales
RechenZentrum
Erlangen

Der IT-Dienstleister der FAU