*MIE2014, Istanbul*

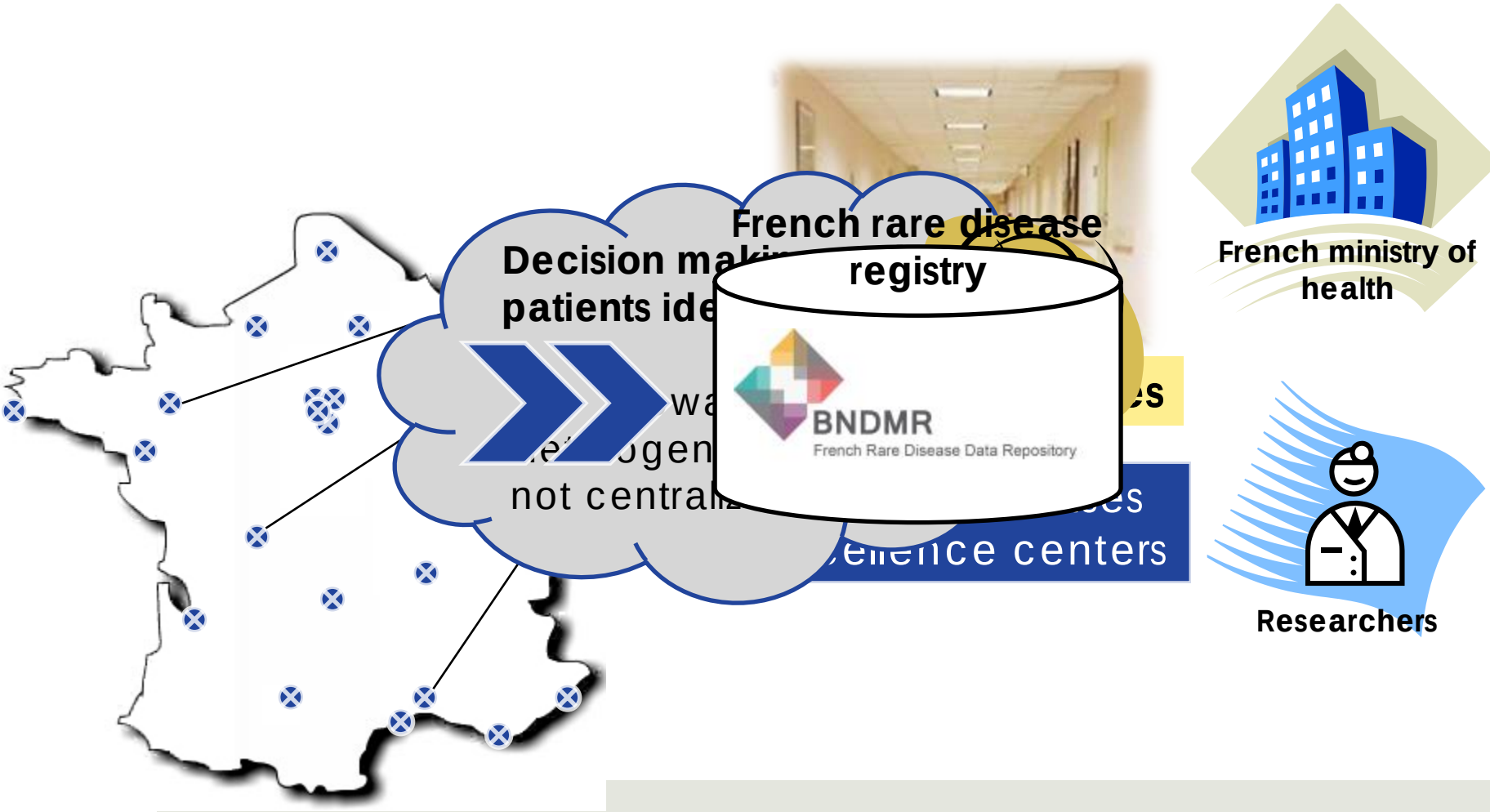


# Formalizing mappings to optimize automated schema alignment: application to rare diseases

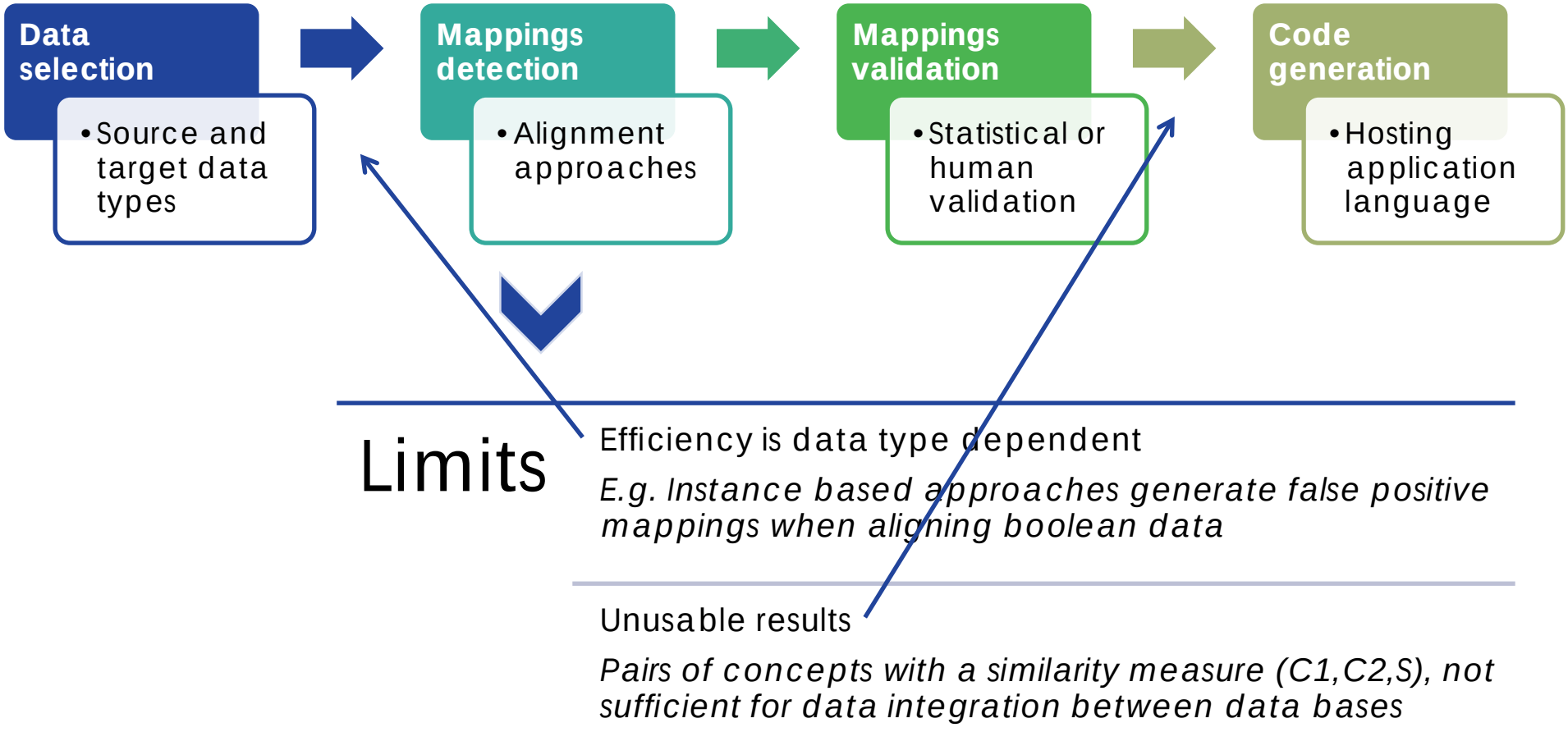
Meriem Maaroufi
Rémy Choquet
Paul Landais
Marie-Christine Jaulent

Paris, France

# French Rare Disease Organization



# Data integration process



# Hypothesis

(C1 2,S)

Characterizing mappings in a complete formalization will improve alignment results usability

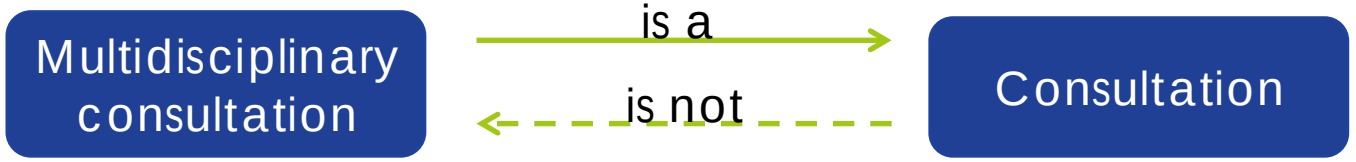
# 1st duality
# Data element – Value element

- Data element = the container
- Characteristics:
  - Label, definition
  - Data type
  - Value domain, restrictions…
- Notation: $\mathbf{E_i}$ (i=1..n;n=card(schema))

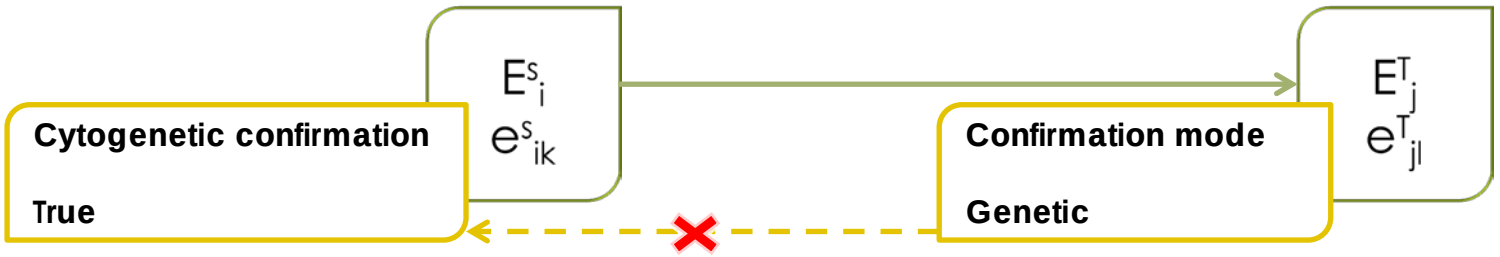
- Value element = the content
- An integer, a string, a Boolean value or an entry of a list
- Depends on
- Notation: $\mathbf{e_{ik}}$ (k=1..p;p=card($E_i$))

# 2nd duality
# Source element – Target element

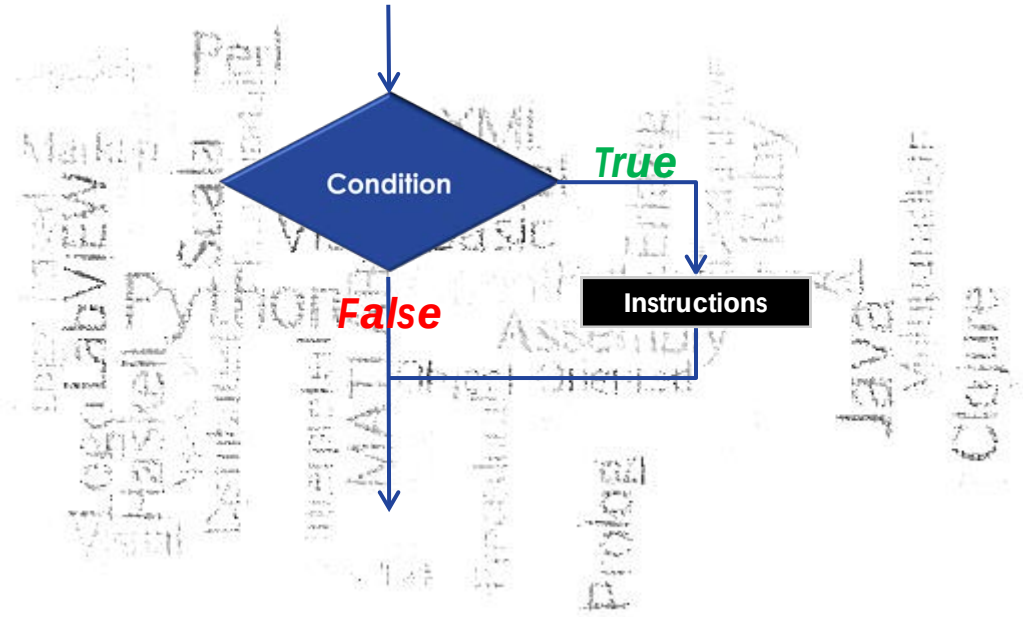
- A mapping is rarely a bijection.
  - It is often due to generalization/specification.




- A mapping has a **direction** : from source schema to target schema.



# Conditional structure: rules




**If… then…** formalism is supported by most programming languages

Well suitable for bi-level mappings

Can define exact mappings and data transformations

Result

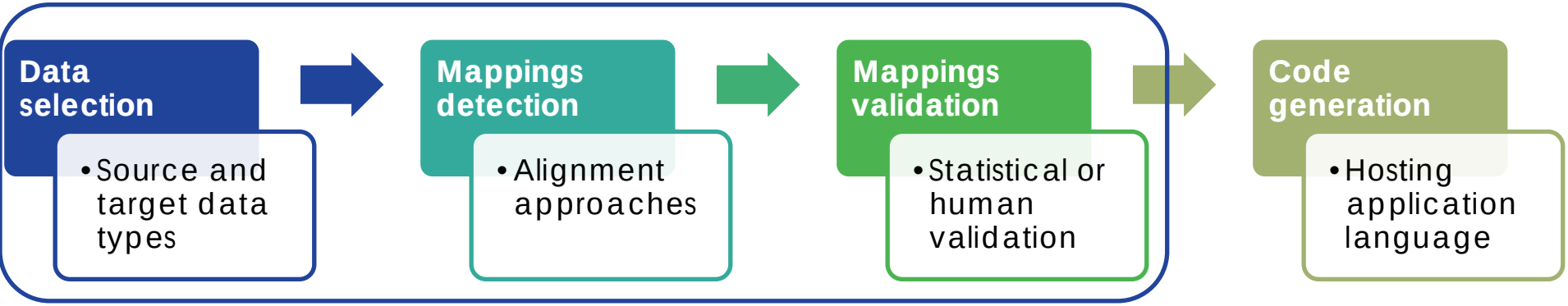
# Mapping formalization

$$Mapping = \{E^S_i\text{-}E^T_j ; e^S_{ik}\text{-}e^T_{jl} ; Rule(S\text{->}T)\}$$

A rule defines the relation between the involved source and target data elements and value elements

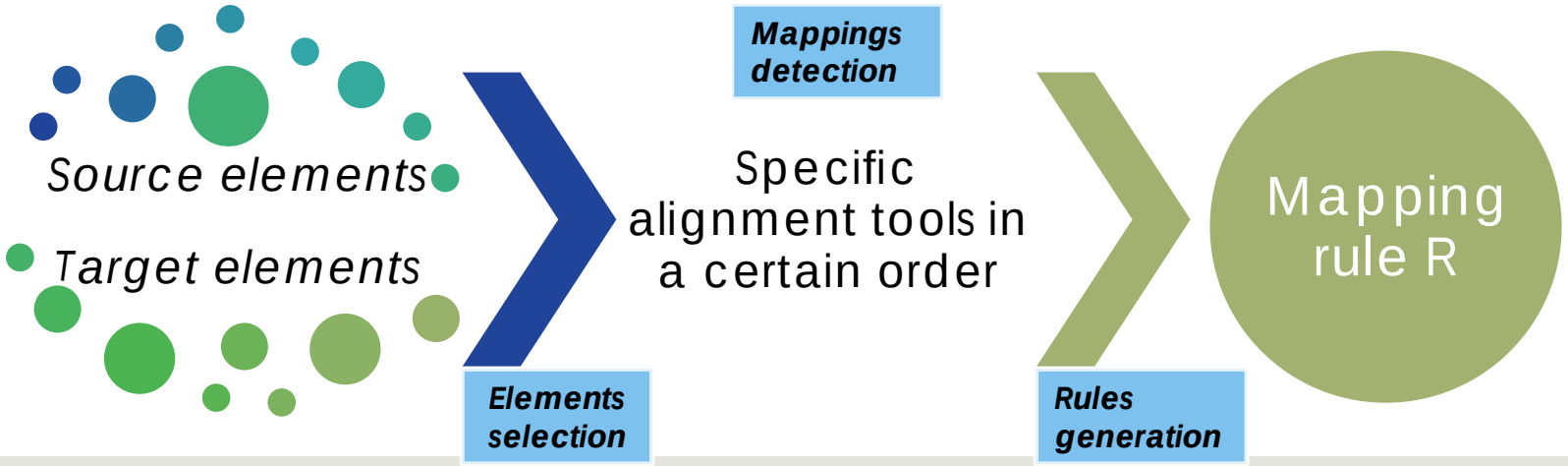
| $E^s_i$ $E^T_j$ | $e^s_{ik}$ $e^T_{jl}$ | Rule |
|---|---|---|
| Glycemia – Hypoglycemic state | integer – true | **If glycemia < $threshold then hypoglycemic state = true** |
| Act type – Participant profession | nurse intervention – nurse | **If Act type = nurse intervention then Participant profession = nurse** |

# Application to BNDMR context

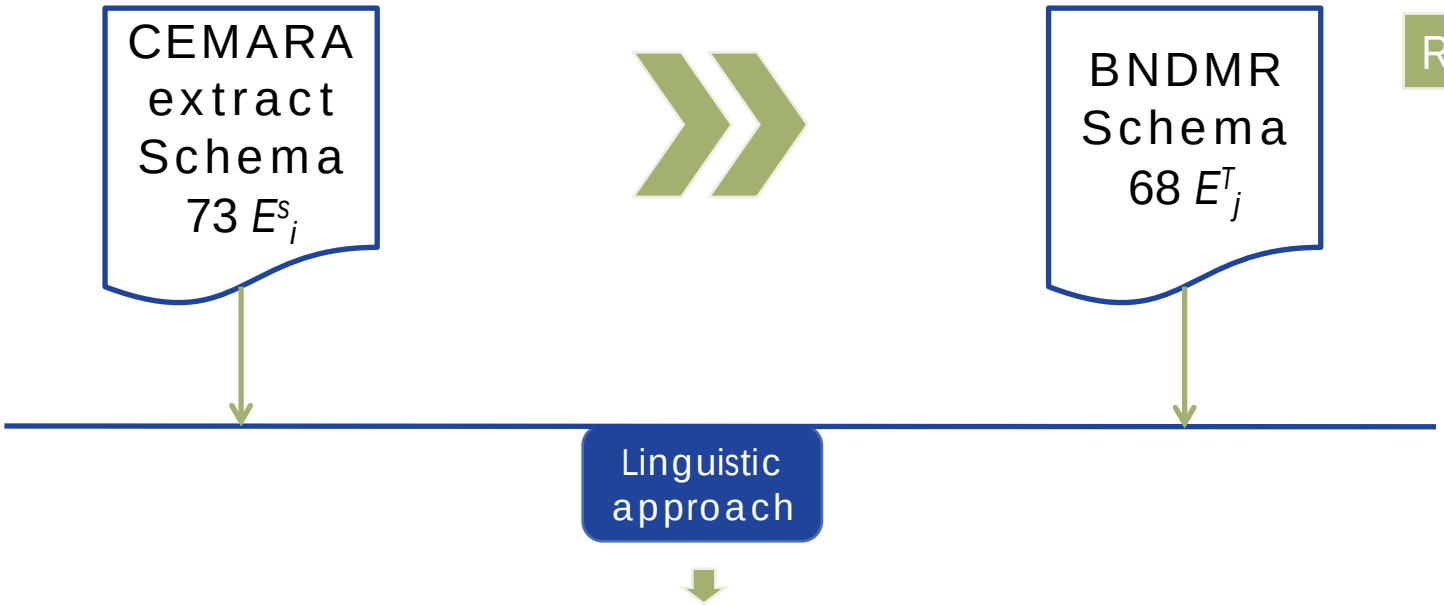


# Rules generation methodology

A **specific process** :
- is a workflow
- that involves some chosen alignment approaches
- operating in a given order
- on selected data elements
- To detect specific mappings: defined rules.

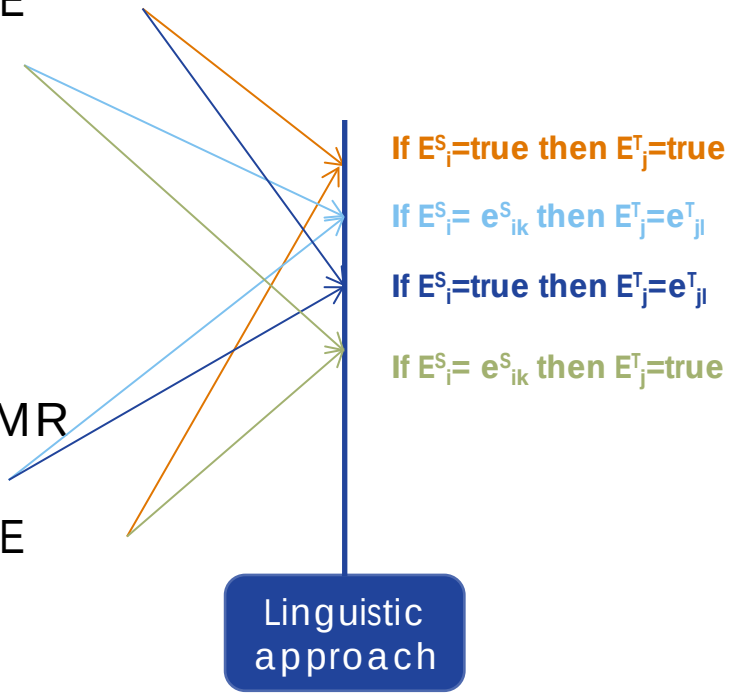


# Tools & experimentation



# Process example

Source CEMERA
43 boolean DE
56 lists VE




Target BNDMR
106 lists VE
16 boolean DE

| | Reference test | Experimentation |
|---|---|---|
| bool-bool | 3 | 3 |
| list-list | 6 (DE-DE) | 35 |
| bool-list | 1 (DE-DE) | 22 |
| list-bool | 0 | 1 |

If PropLink=propositus [source]
then Propositus=true [target]

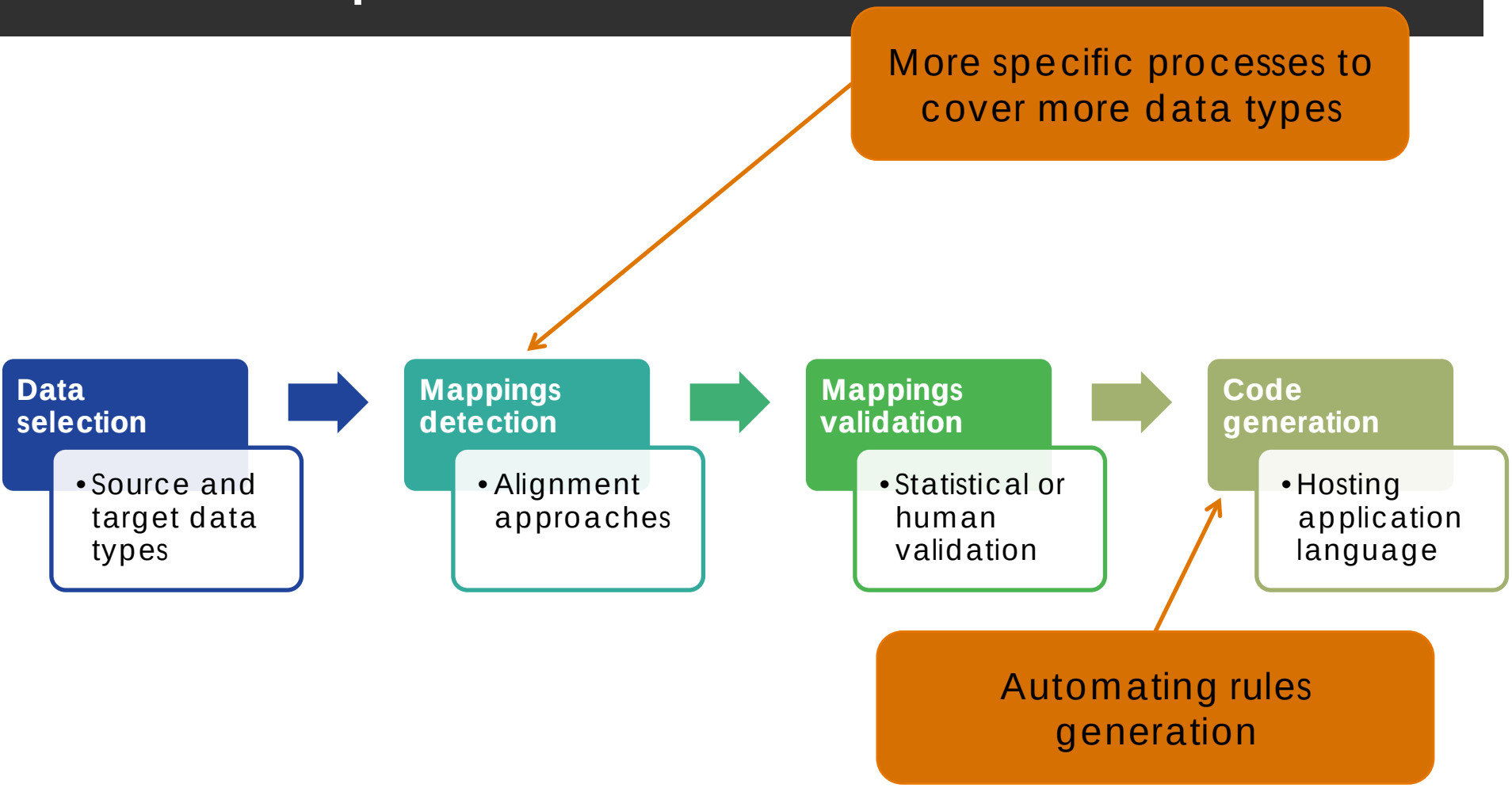
If ConfCyto=true [source]
then ConfirmationMode=cytogenetic [target]

***Elements selection*** ***Mappings detection*** ***Rules generation***

# Conclusion

- The proposed formalization "$mapping = \{E^S_i\text{-}E^T_j ; e^S_{ik}\text{-}e^T_{jl} ; Rule\}$" is well suitable to characterize simple and complex mappings.
- Mappings characterized by the proposed formalization can be directly used in data integration processes (e.g. ETL).
- Depending on input data types, processes for mappings detection will be different.

# Perspectives



meriem.maaroufi@bndmr.fr

# Thank you for your attention!

Special thanks to:

- BNDMR team
- INSERM UMR-1142 team LIMICS