Open-source data validation integrating *SDMX*, *R*, *Python*, *VTL* and machine learning

Violeta Calian and Ragnhildur Björg Konráðsdóttir

Statistics Iceland

Statistics Iceland

Goal

Statistics Iceland

Open-source implementation of validation processes

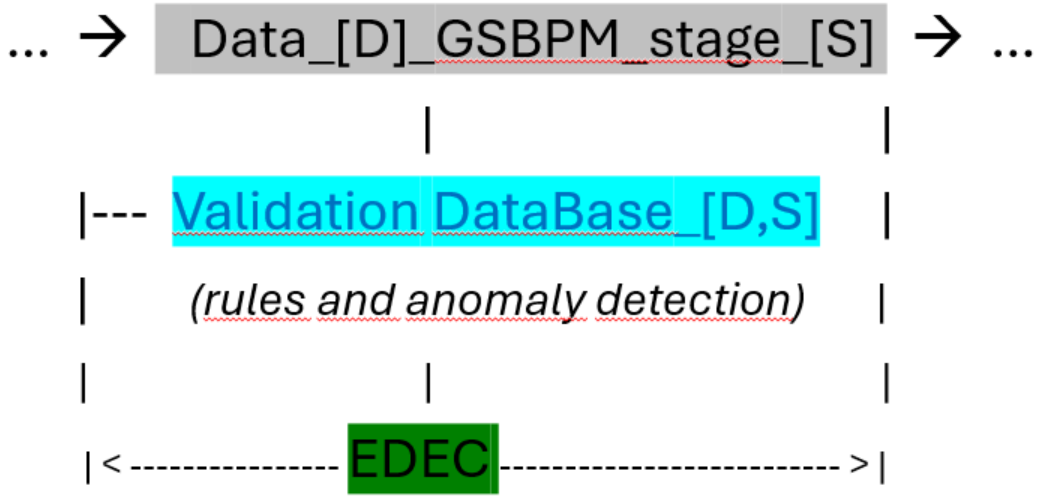
at any stage of GSBPM and any data set

by using

SDMX Global Registry

and Machine/Deep Learners,

with R, Python, VTL



github: https://github.com/violetacln/sdmx_ML_validation

Principle:

Statistics Iceland

*unique **method** of solution for data-dependent processes*

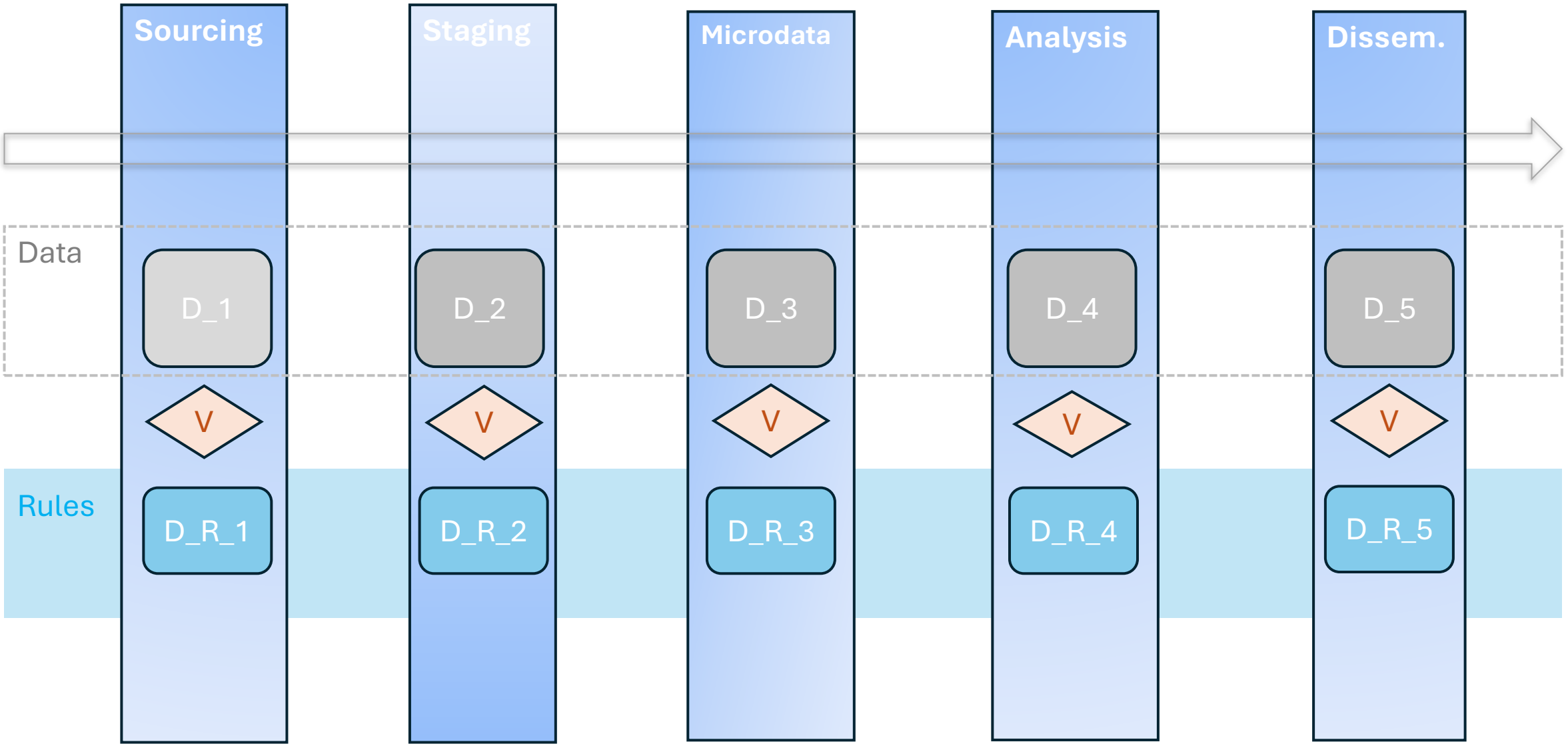
- Multiple sources of *validation rules* for each
  - data-set
  - data-instance

- Discovery/learning:
  - anomalous data-points
  - significant association rules in the data

by using open-source code

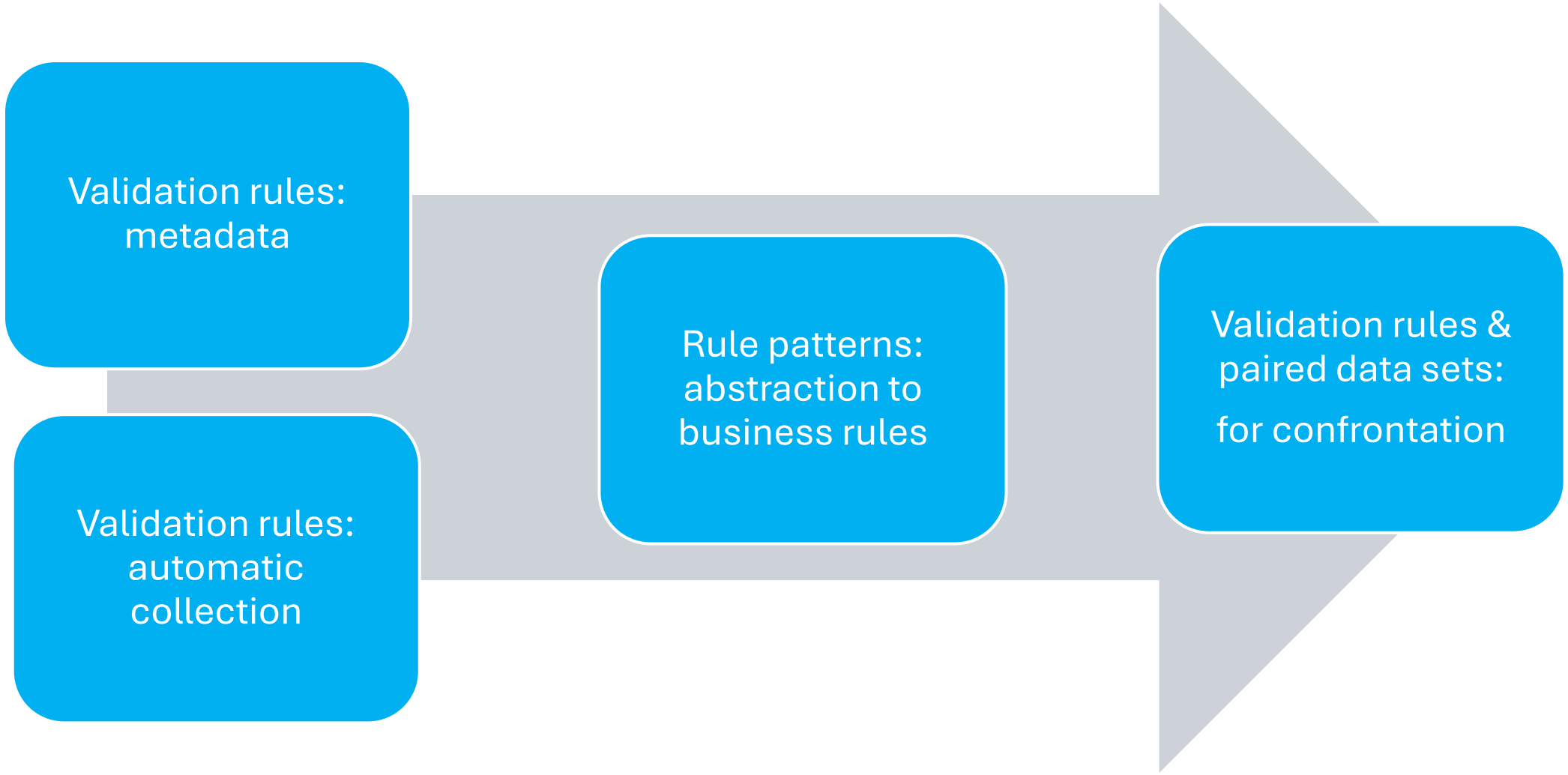
Data flow

Statistics Iceland



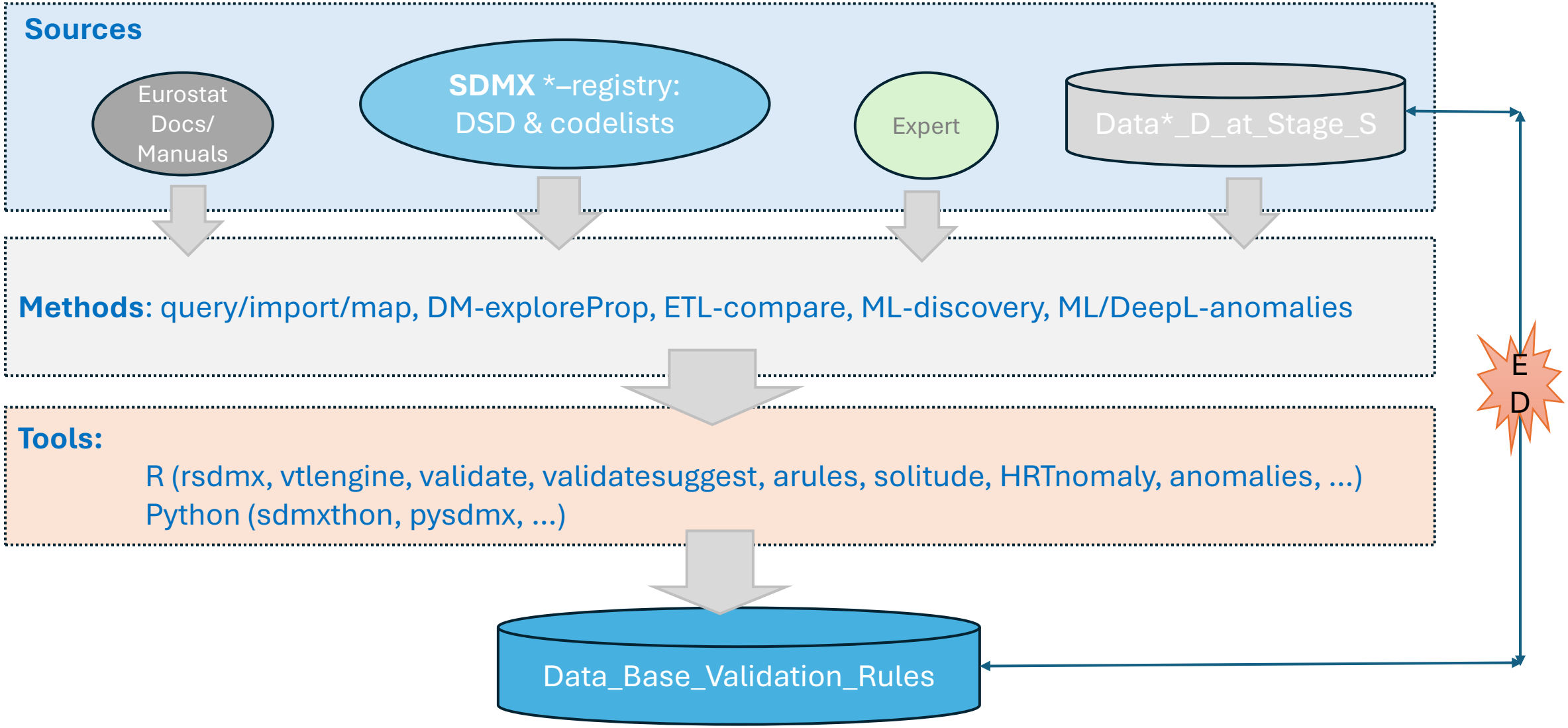
Rule flow

Statistics Iceland

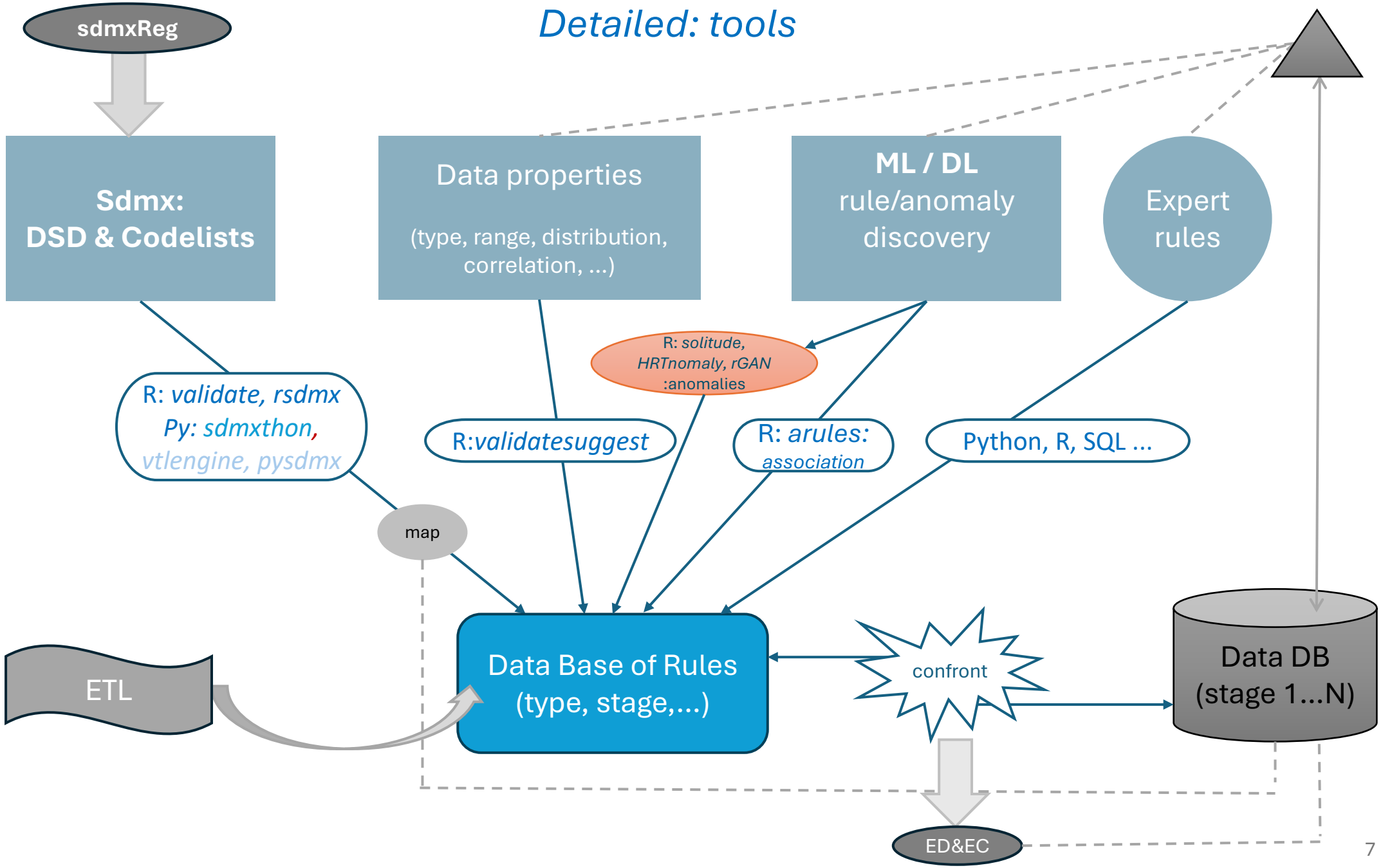


Automatic collection of validation rules

Statistics Iceland

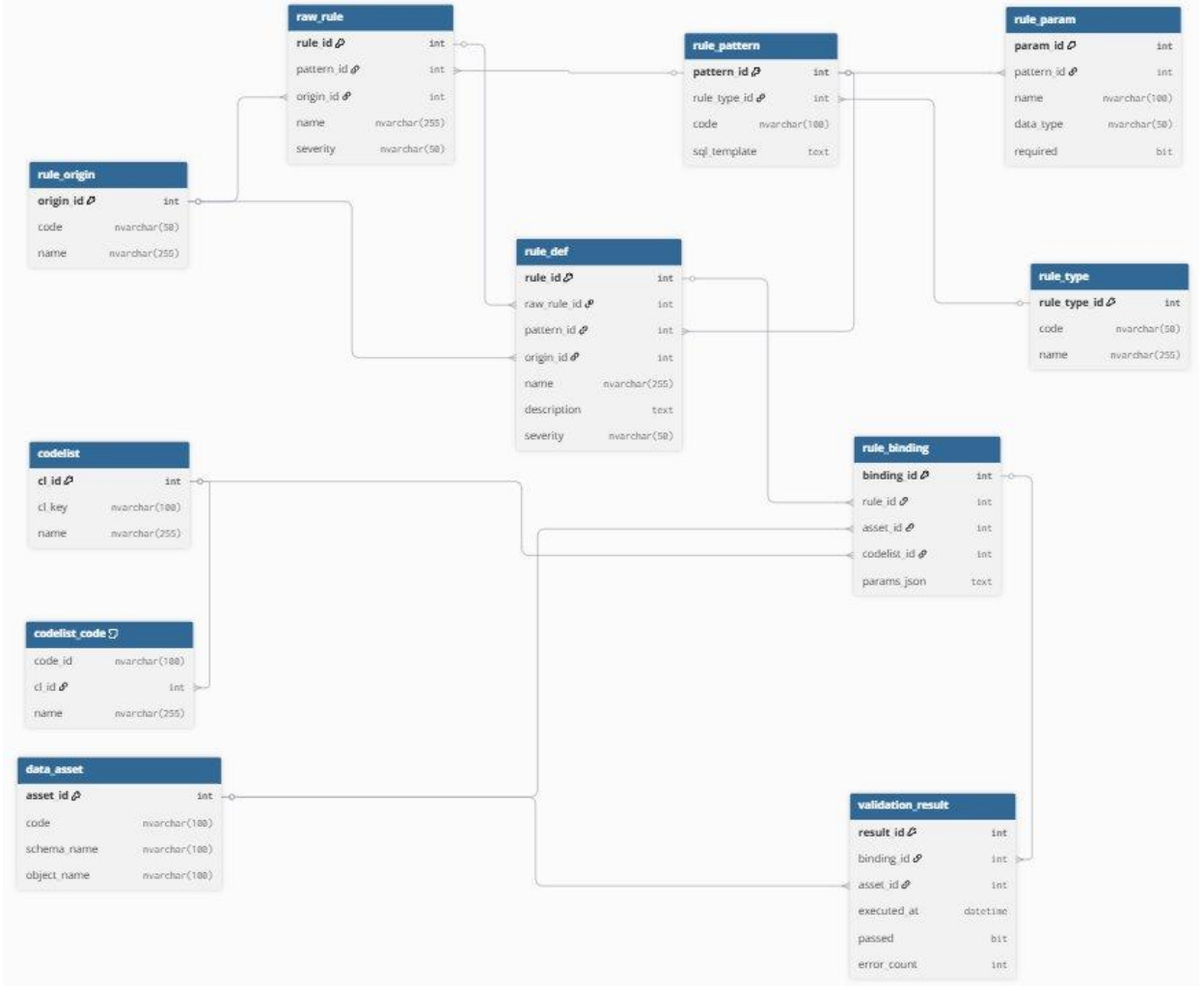


Detailed: tools



Database of rules

Statistics Iceland



Special cases

Statistics Iceland

- Dissemination stage (any data set): *SDMX* – special role

- Adding: disclosure control rules / anomaly detection derived from SDC-requirements

Simple code examples

Statistics Iceland

https://cran.r-project.org/web/packages/validate

```
rules <- validator_from_dsd(endpoint = sdmx_endpoint("ESTAT")
   , agency_id = "ESTAT", resource_id = "STSALL", version="latest")

length(rules)
[1] 13
rules[1]
Object of class 'validator' with 1 elements:
 CL_FREQ: FREQ %in% sdmx_codelist(endpoint = "https://ec.europa.eu/tools/cspa_services_global/sdmxr
egistry/rest", agency_id = "SDMX", resource_id = "CL_FREQ", version = "2.0")
Rules are evaluated using locally defined options
```

rules <- validatesuggest::suggest_all(data=d, vars = names(d), domain_check = TRUE, range_check = TRUE, pos_check = TRUE, type_check = TRUE, na_check = TRUE, unique_check = TRUE, ratio_check = TRUE, conditional_rule = TRUE)

rules_fromApriori <- apriori(data=tdata, parameter=list(support=0.3)) ; inspect(head(rules, n = 100, by = "confidence"))
data_test %>% solitude::iso$predict() %>% arrange(desc(anomaly_score))

summary(arsenal::comparedf(dfN, dfM), by=..., tol.vars=c(a=„A“, ...), tol.num.val=epsilon)

sdmxthon.parsers.data_validations.validate_data() sdmxthon.parsers.metadata_read.create_metadata()

Statistics Iceland

Thank you!