Moving towards the standardized process of automatic statistical data editing using machine learning techniques

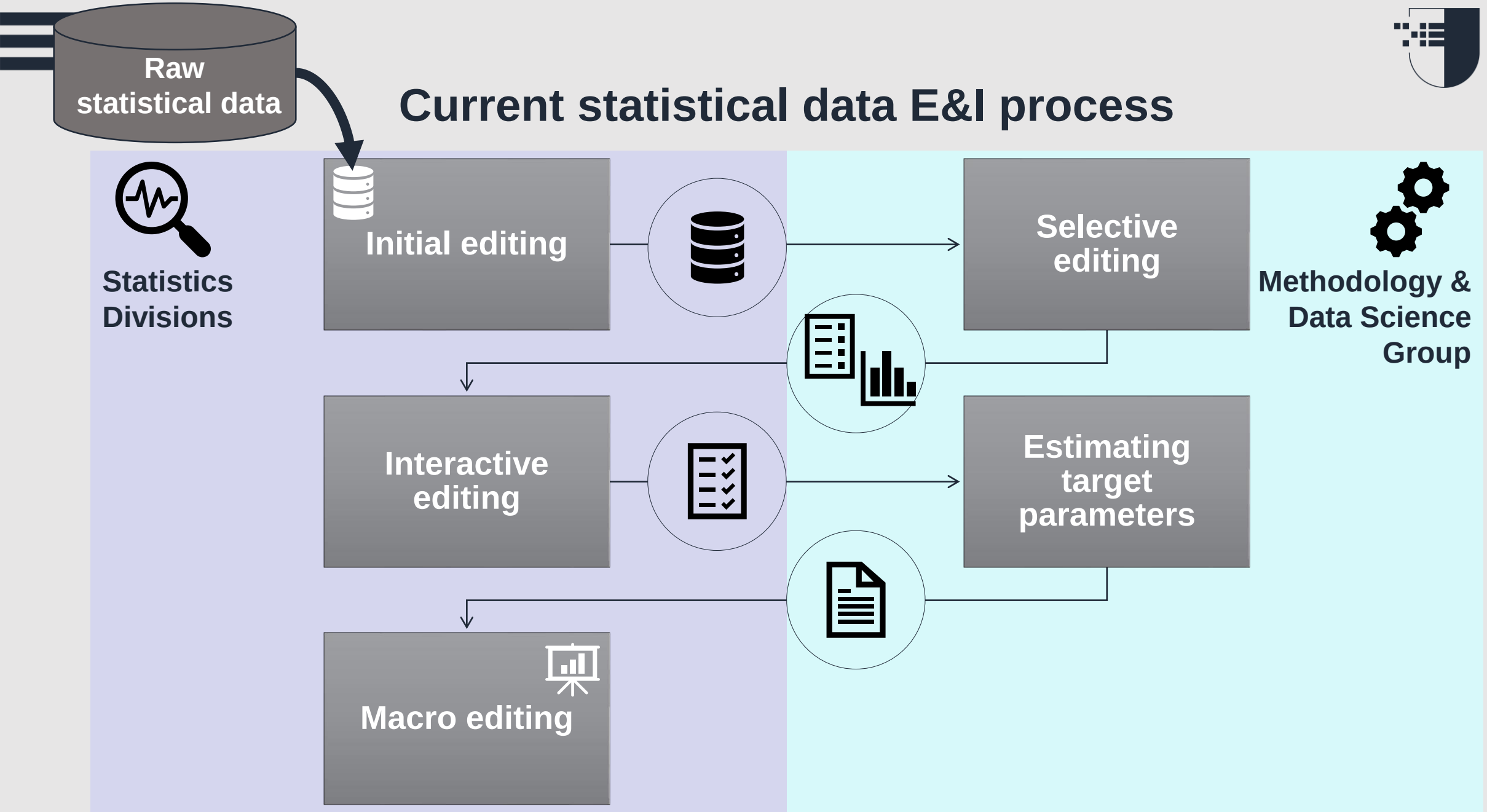
STATISTICS LITHUANIA
STATE DATA AGENCY

Ieva Burakauskaitė,
Adviser at Methodology and Data Science Group
Ieva.Burakauskaite@stat.gov.lt

Motivation

- Statistical data editing and imputation (E&I) is a significant but time-consuming process during the production of official statistics at National Statistical Institutes.
- In order to increase the efficiency of E&I, the Generic Statistical Data Editing Model (GSDEM) offers some valuable insights on various steps during the latter process such as the detection of the most influential errors using selective editing and the error treatment with either interactive or preferably automatic editing.
- Although the selective editing process step had already been developed and implemented at State Data Agency (Statistics Lithuania), the error treatment automatization part still needed to be refined.
- The migration of statistical surveys into the uniform platform has motivated Statistics Lithuania to re-evaluate the current E&I process, as such a platform offers numerous opportunities for the standardization and integration of the process.

Current statistical data E&I process



Our goal: Standardized automatic statistical data E&I process

- Integration of the E&I process according to the Generic Statistical Business Process Model (GSBPM), and standardization adopting the Generic Statistical Data Editing Model (GSDEM) as the reference framework.
- One platform for all E&I process steps – no data file exchange through e-mail / internal network.
- Convenient user interface for Statistics Divisions, and working E&I methods under the surface – no programming knowledge required of users.
- Only influential outliers flagged during selective editing – minimized data editing time and re-contact with respondents.

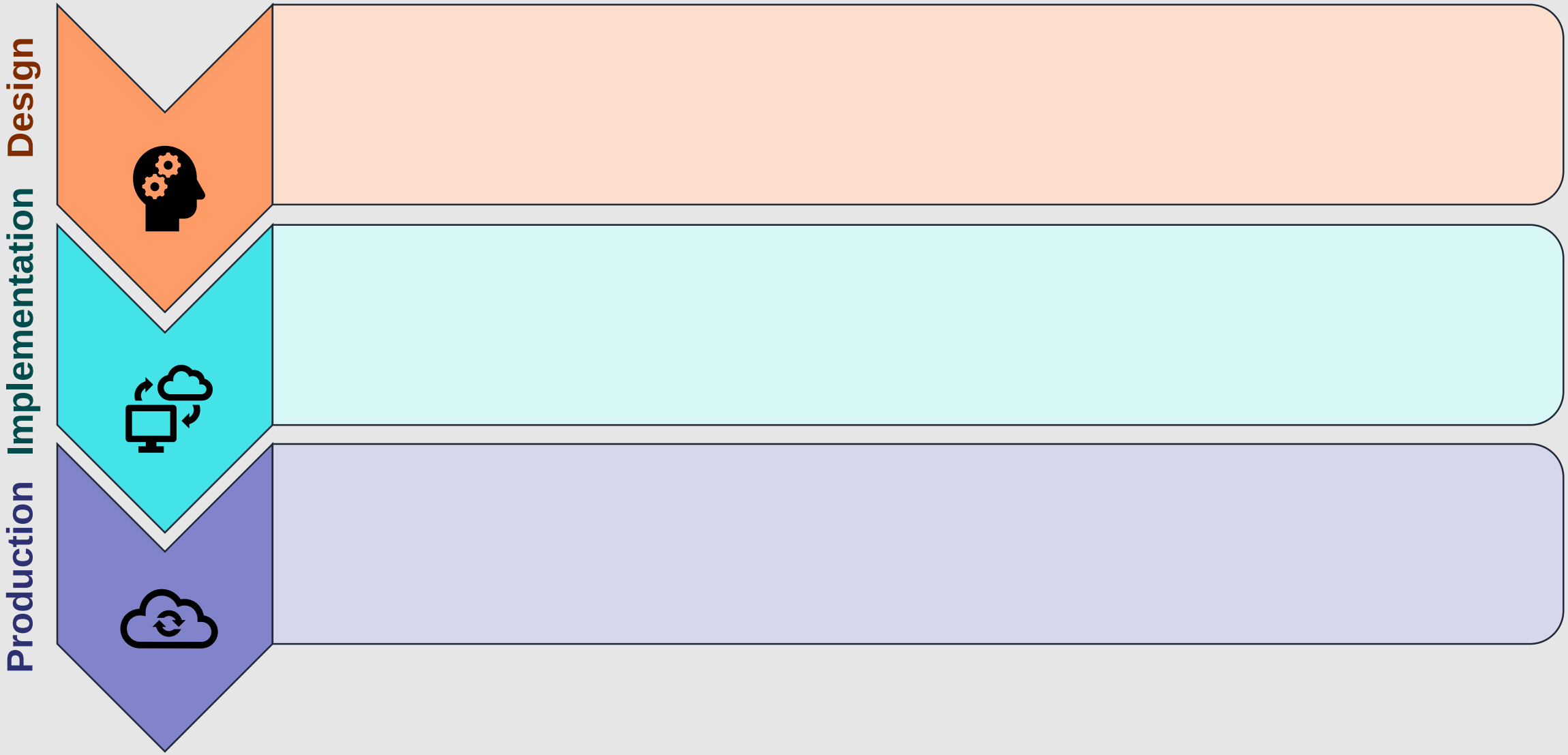
Desirable outcomes

- Efficient resource allocation: Employees are able to focus on data analysis, not on manual review / follow-up of the collected observations.
- Faster production of official statistics: Less time is spent on data editing, hence, on the preparation of statistical information in general.
- Lower response burden: Re-contact counts are minimized.
- Increase in quality of statistics: Less manual editing = less human error.

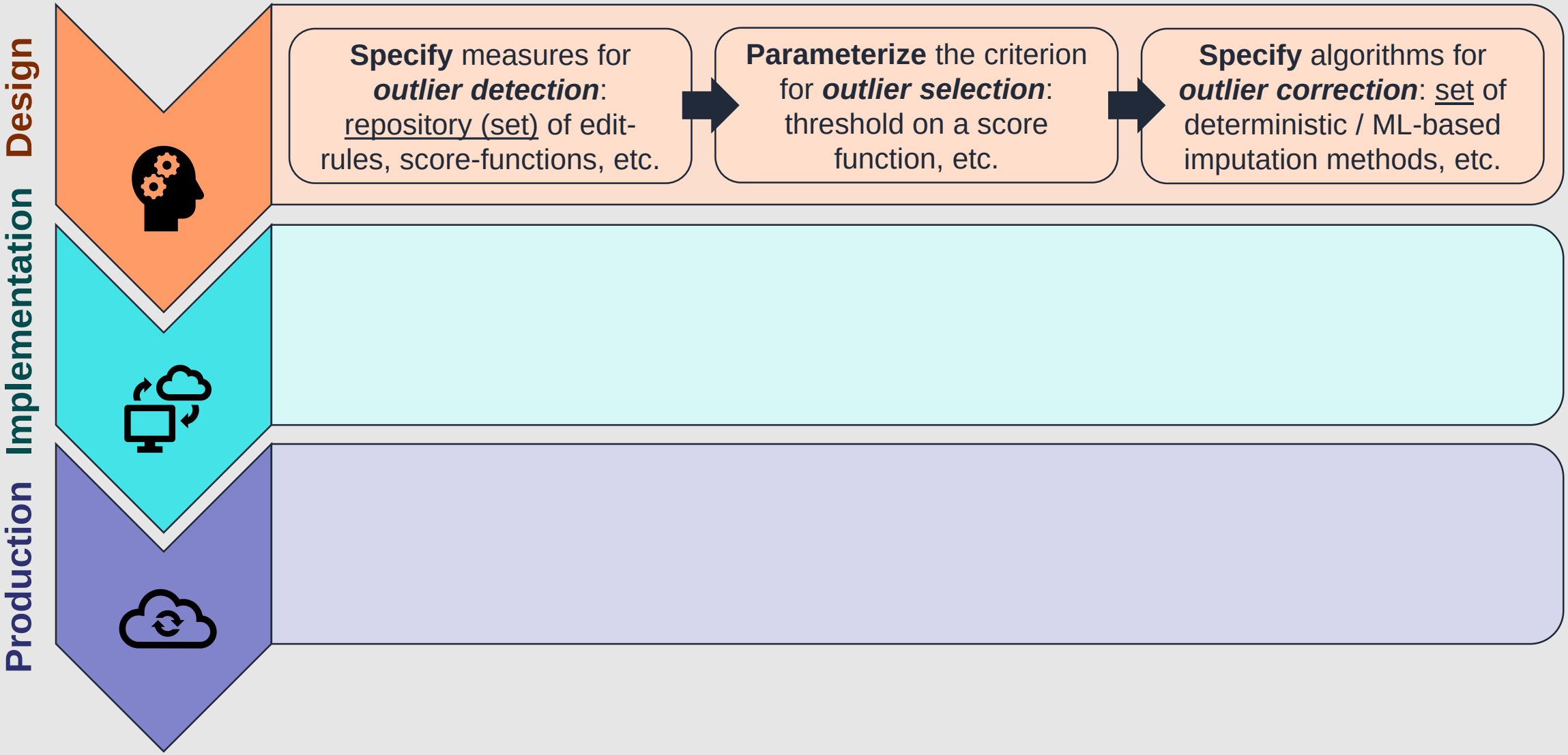
E&I function classification according to GSDEM

- **Review:** Functions that examine the data to identify potential problems.
- **Selection:** Functions that select units or fields within units that may need further treatment, i.e., to be adjusted or imputed.
- **Treatment:** Functions that change selected data values to improve the data quality.

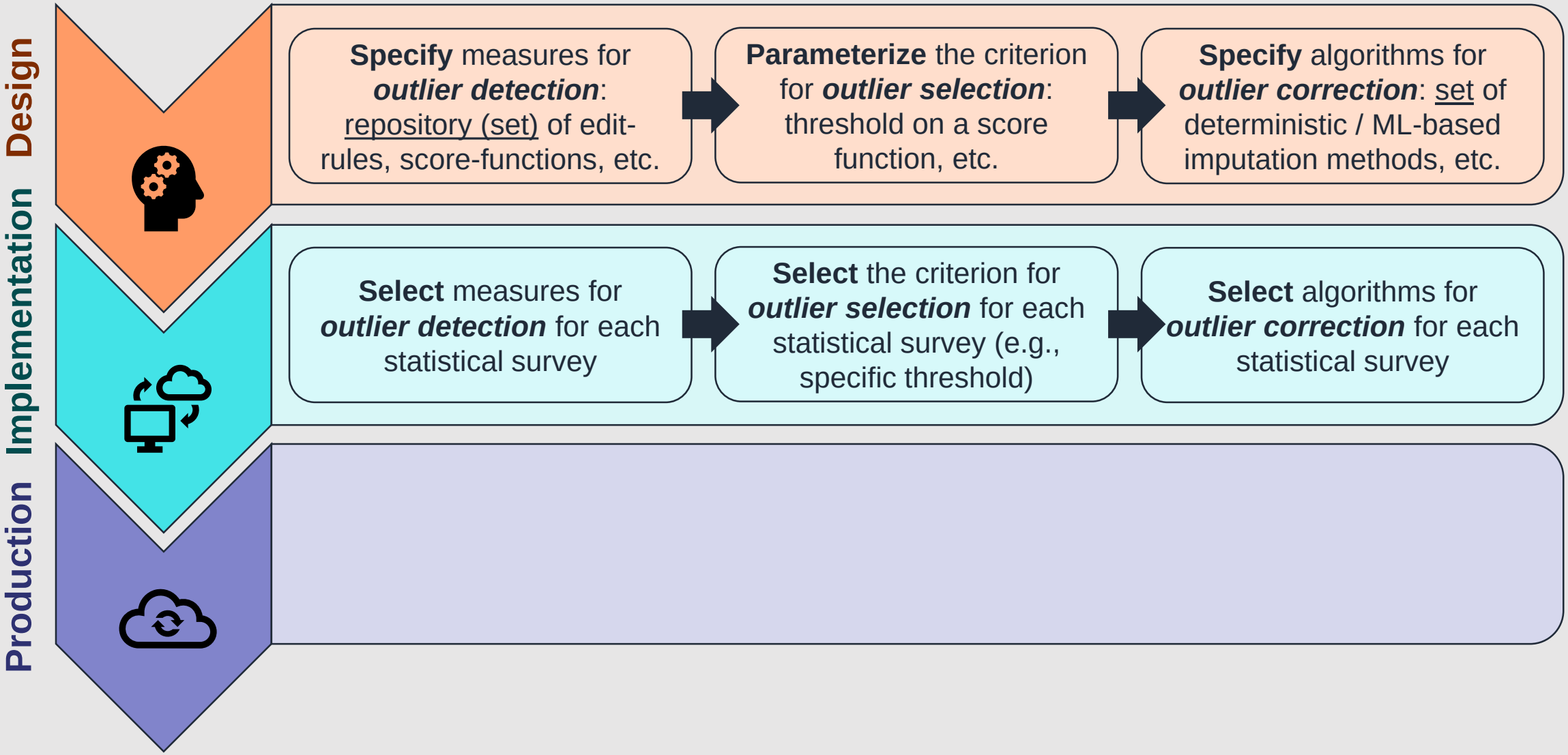
Standardization and automatization of E&I process



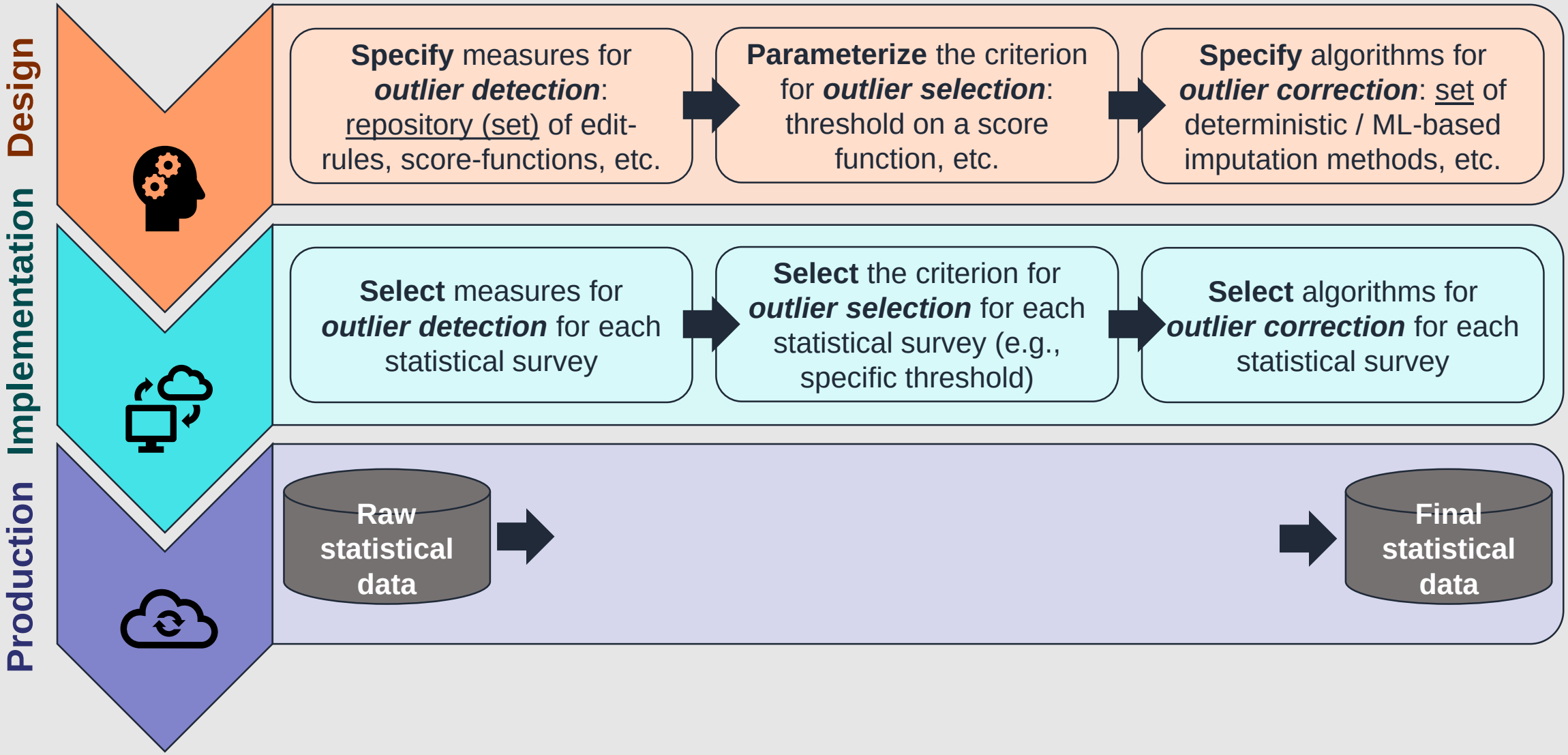
Standardization and automatization of E&I process



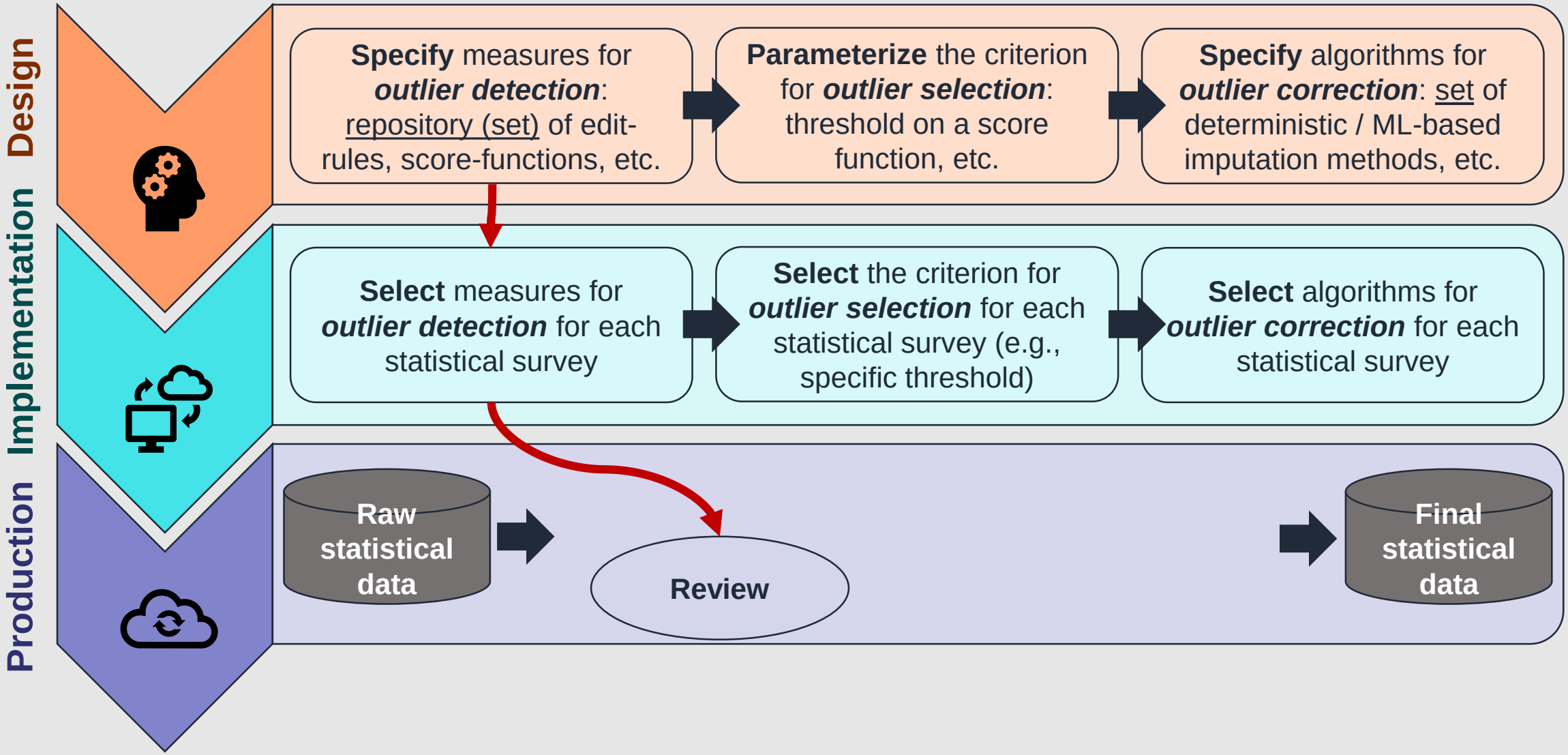
Standardization and automatization of E&I process



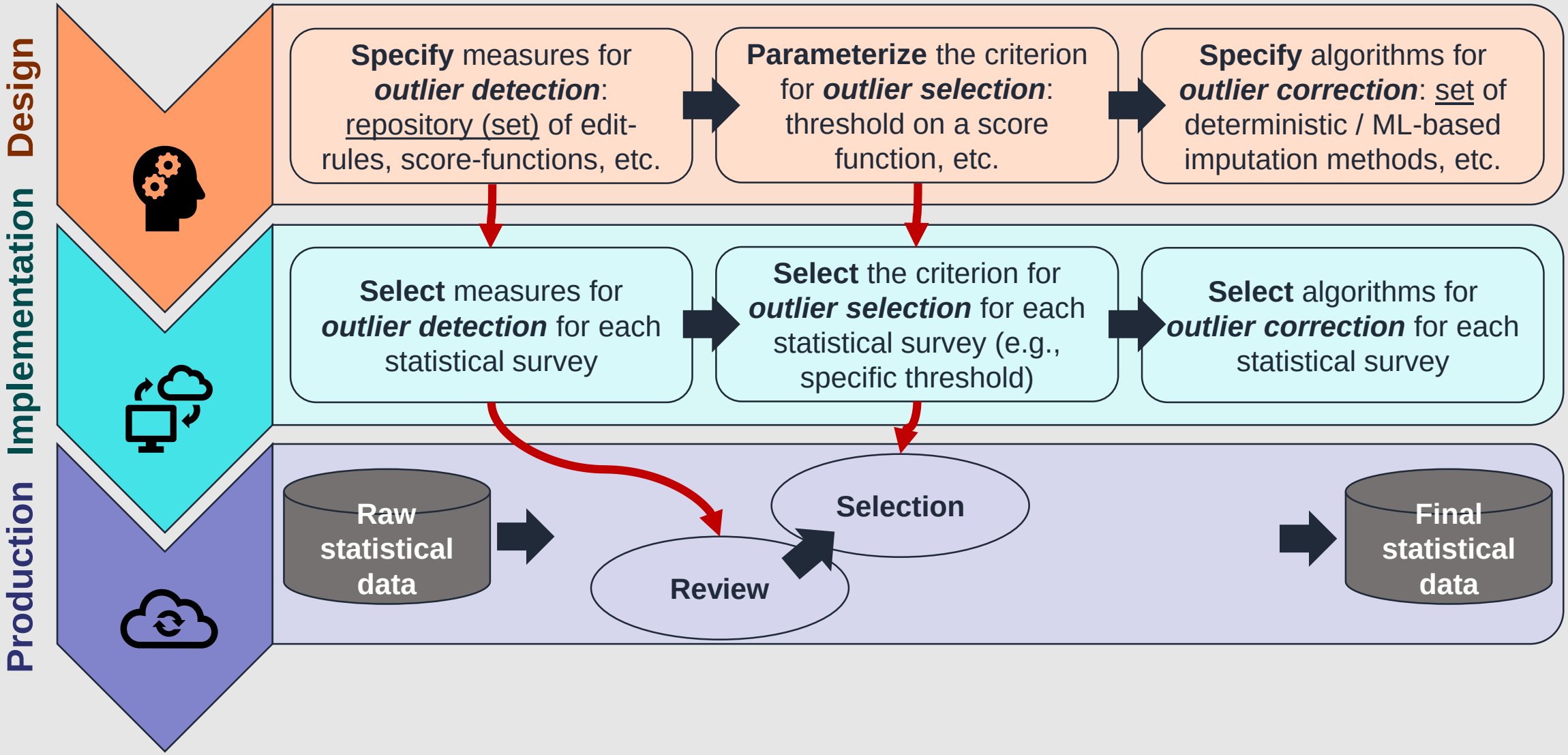
Standardization and automatization of E&I process



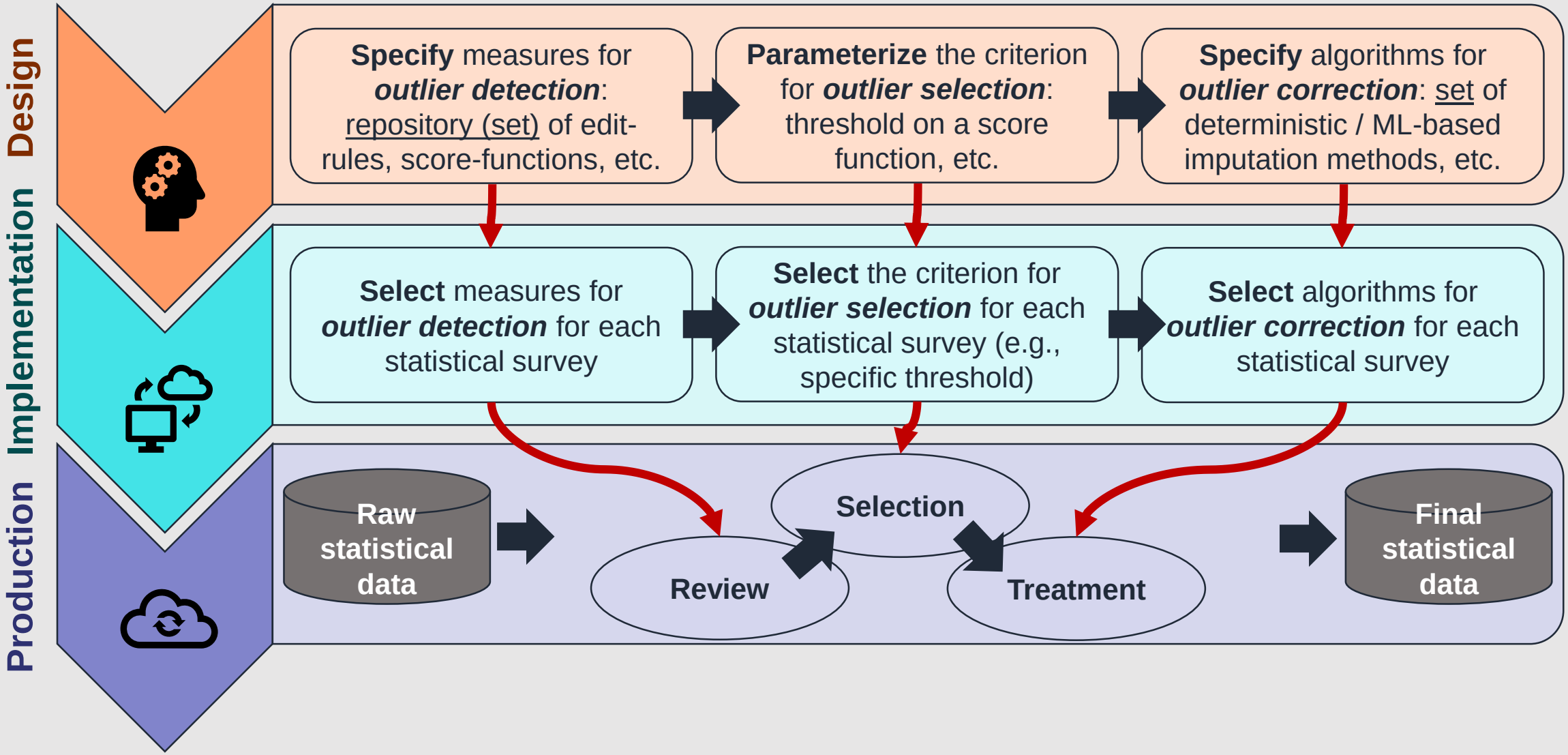
Standardization and automatization of E&I process



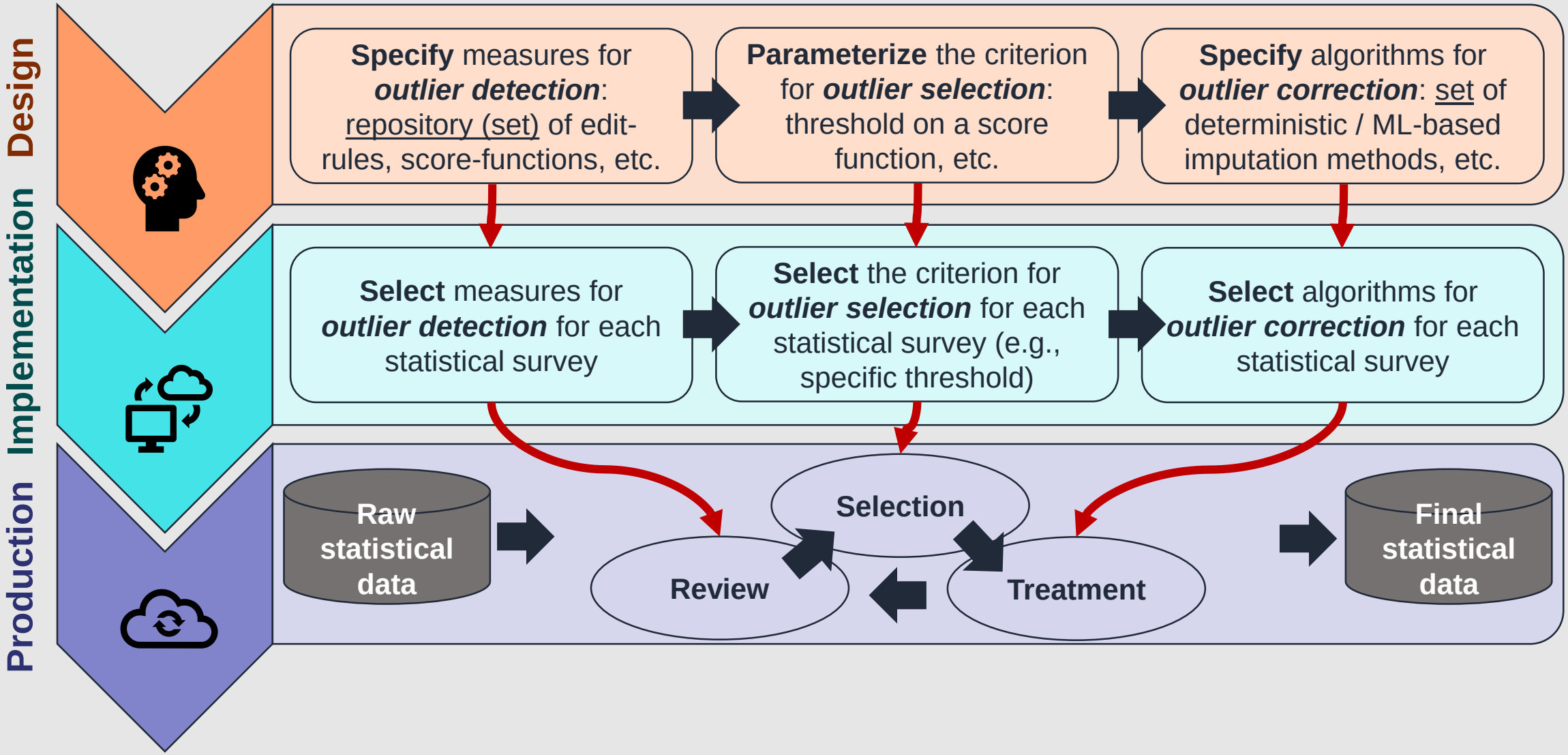
Standardization and automatization of E&I process



Standardization and automatization of E&I process



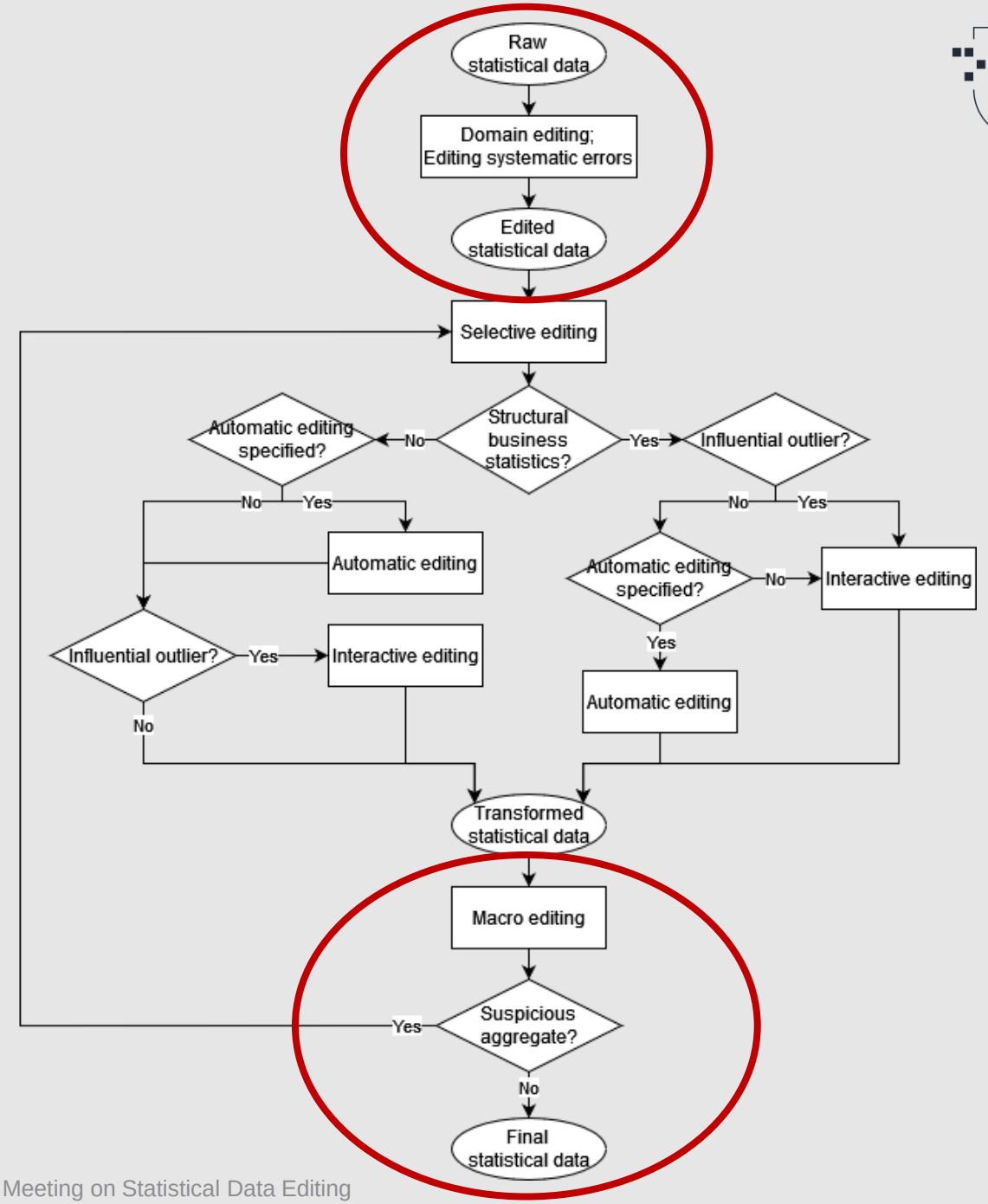
Standardization and automatization of E&I process



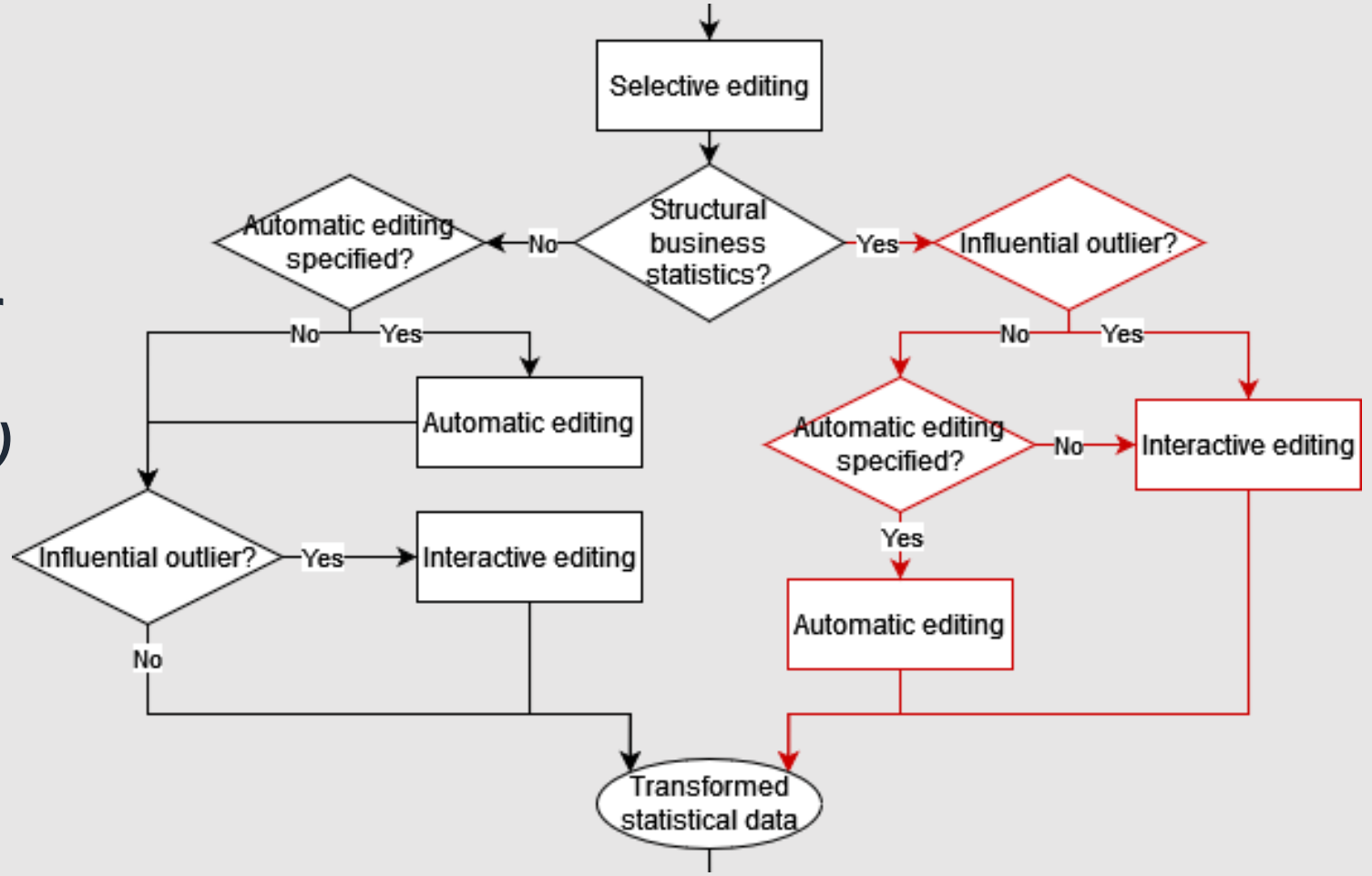
Example: E&I flow model for business statistics surveys
(based on suggestions in GSDEM)

Example: E&I flow model for business statistics surveys

(based on suggestions in GSDEM)

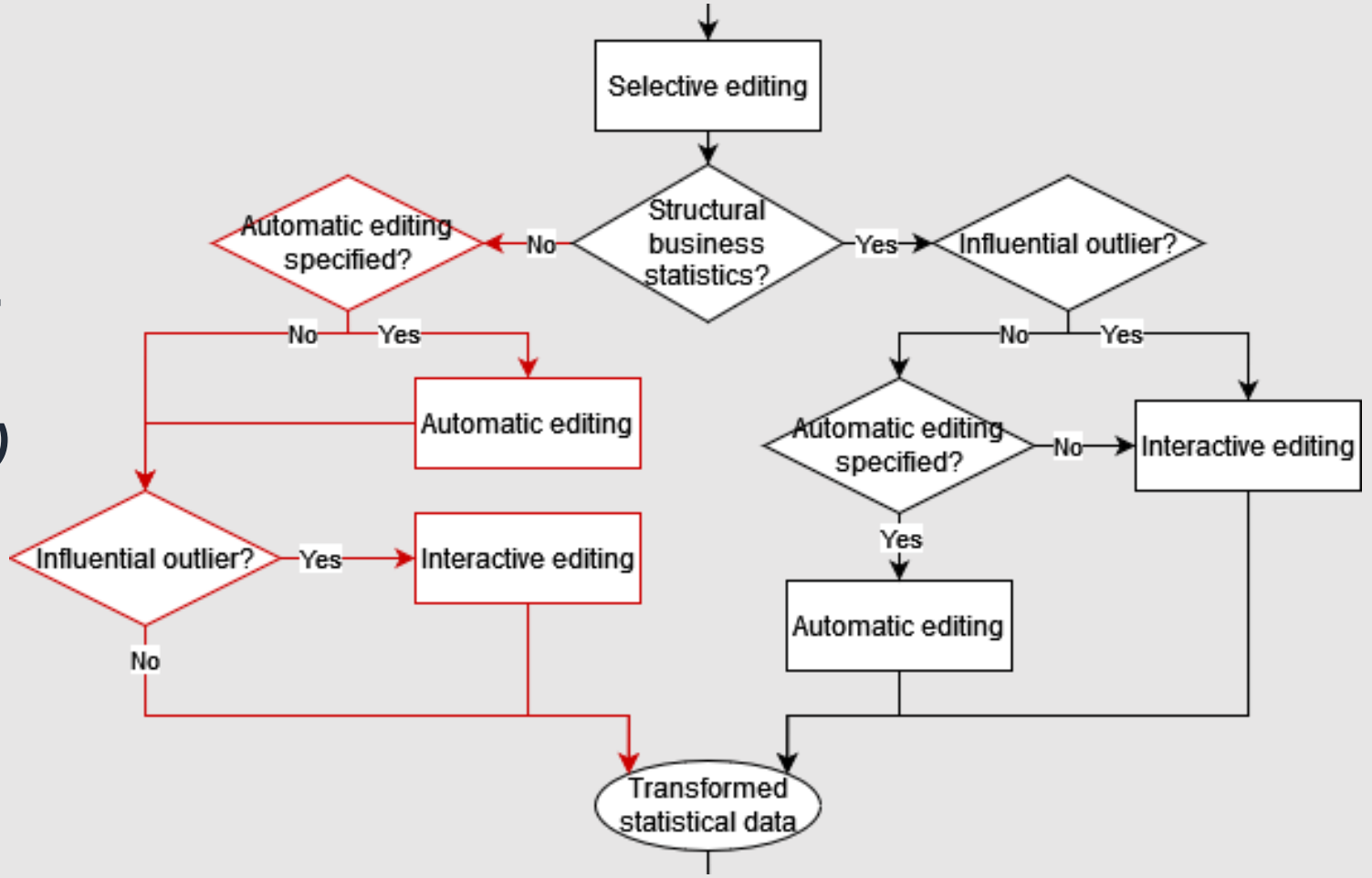


Example: E&I flow model for business statistics surveys
(based on suggestions in GSDEM)



Example: E&I flow model for business statistics surveys

(based on suggestions in GSDEM)



From *Design* to *Implementation* phase

Collaboration between IT Division, State Data Governance Information System Division, Methodology and Data Science Group, and various Statistics Divisions at Statistics Lithuania helps reaching the goal:

- IT infrastructure and software solutions needed to develop the working data editing tool and enable the automatization of the E&I process.
- Development of deterministic, mathematical methods and machine learning techniques needed to perform outlier detection and correction. To this end, collaboration with the academic community takes place.
- Expert knowledge needed to evaluate the quality of E&I machine learning techniques compared to the “traditional” methods.

Quality of statistical data and E&I process

- Standardized reports on statistical data quality as well as E&I process (using output metadata, i.e., paradata) may be generated for quality assessment, containing a set of quality indicators.
- Paradata contains indicators and measurements concerning the quality of the input, output or intermediate versions of the data set (e.g., imputation rates, number of edit failures and systematic errors).

Outlier detection study

Data validation for the *Quarterly Statistical Survey on Service Enterprises* of Statistics Lithuania (I)

Target variable: **Enterprise turnover** of the accounting period t (y_t).

Auxiliary variables: **Enterprise turnover** of the previous period (y_{t-1}), and of the same period of the previous year (y_{t-4}).

- Deterministic approach:

Edit-rules based on a comparison with overall trend of other observations belonging to the same subset of the population. The acceptance interval is constructed according to the interquartile range of observations.

- Outlier detection method:

Hidiroglou-Berthelot method based on the idea of acceptance boundary that varies according to the size of a unit. Here ratios from (i) y_{t-1}, (ii) y_{t-4} to y_t are compared to the corresponding overall trend of other observations belonging to the same subset of the population (Belcher, 2003).

Data validation for the *Quarterly Statistical Survey on Service Enterprises* of Statistics Lithuania (II)

Target variable: **Enterprise turnover** of the accounting period t (y_t).

Auxiliary variable: **Enterprise turnover** from VAT declarations (y_t^*).

- Selective editing method:

Based on the idea of only looking for influential outliers in order to focus the most accurate treatment on the corresponding subset of units to reduce the cost of the data editing, while maintaining the desired level of quality of the target estimates.

Case study by Burakauskaitė and Nekrašaitė-Liegė (2022)

- Predictions for the target variable (**quarterly turnover**) were obtained using the contamination model (Di Zio and Guarnera, 2013).
- The impact of the potential error on the target estimate was evaluated using the score function with a standard structure – the difference between the observed value of the target variable and its prediction multiplied by the sampling weight and a *suspicion component*.
- An impact of the suspicion component on the effectiveness of selective editing was evaluated by selecting (i) a *discrete* (Di Zio and Guarnera, 2013), and (ii) a *continuous* (Norberg et al., 2010) suspicion component.

Used software: R

Package ***SeleMix*** by Guarnera and Buglielli (2013) for performing selective editing.

Functions:

- ***ml.est*** – fitting the contamination model => estimating model parameters, predicting the “true” values of the target variable.
- ***sel.edit*** – identification of influential outliers.

Score function and discrete suspicion component

- General score function can be expressed as

$$S_i = \left| \frac{s_i w_i (y_i - \hat{y}_i)}{\hat{T}} \right|, \qquad i = 1, \dots, n,$$

where y_i denotes the *i*th observation of the target variable, $\hat{y}_i$ – prediction for the corresponding observation, w_i – sampling weight, s_i – suspicion component, and $\hat{T} = \sum_{i=1}^{n} w_i \hat{y}_i$ – estimate of the parameter of interest (for instance, sum of predictions).

- *Discrete suspicion component* is an indicator variable, denoting whether the corresponding observation is considered suspicious (possibly erroneous) by some edit-rule, or not, that is,

$$s_i = \begin{cases} 1 & \text{if } y_i \text{ is erroneos,} \\ 0 & \text{otherwise.} \end{cases}$$

Continuous suspicion component

- Suppose, we set such an edit-rule that flags an observation y_i as suspicious if $y_i \notin \left(\hat{y}^{(L)}, \hat{y}^{(U)}\right)$, $i = 1, \ldots, n$, for instance, $\hat{y}^{(L)}$ – 1st quartile, $\hat{y}^{(U)}$ – 3rd quartile of the vector of predictions $\hat{y}_i$.
- *Continuous suspicion component* depends on the deviation from the latter acceptance interval, that is,

$$\tilde{s}_i = \begin{cases} \dfrac{\hat{y}_i - \kappa\left(\hat{y}_i - \hat{y}^{(L)}\right) - y_i}{\max\{\hat{y}^{(U)} - \hat{y}^{(L)}, \alpha\hat{y}_i\}} & \text{if } y_i < \hat{y}_i - \kappa\left(\hat{y}_i - \hat{y}^{(L)}\right), \\ \dfrac{y_i - \hat{y}_i - \kappa\left(\hat{y}^{(U)} - \hat{y}_i\right)}{\max\{\hat{y}^{(U)} - \hat{y}^{(L)}, \alpha\hat{y}_i\}} & \text{if } y_i > \hat{y}_i + \kappa\left(\hat{y}^{(U)} - \hat{y}_i\right), \\ 0 & \text{if } \hat{y}_i - \kappa\left(\hat{y}_i - \hat{y}^{(L)}\right) \le y_i \le \hat{y}_i + \kappa\left(\hat{y}^{(U)} - \hat{y}_i\right), \end{cases}$$

and $s_i = \tilde{s}_i/(\tau + \tilde{s}_i)$, with parameters $\kappa \ge 0$, $\alpha > 0$ and $\tau > 0$, $i = 1, \ldots, n$.

Selective editing efficiency comparison (I)

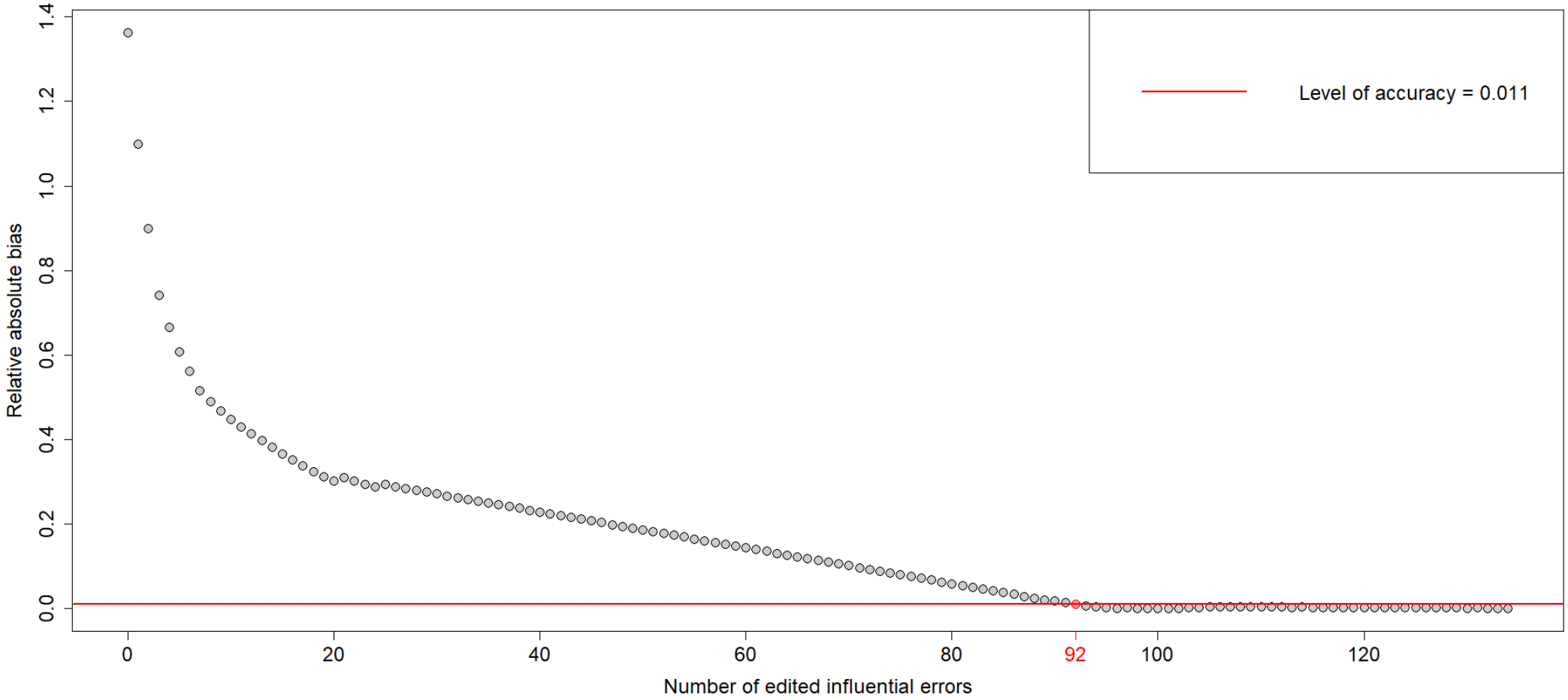


Figure 1: Relative absolute bias (RAB) dependency on the number of edited influential outliers using a *discrete suspicion component* (Burakauskaitė and Nekrašaitė-Liegė, 2022; Figure 1).

Selective editing efficiency comparison (II)

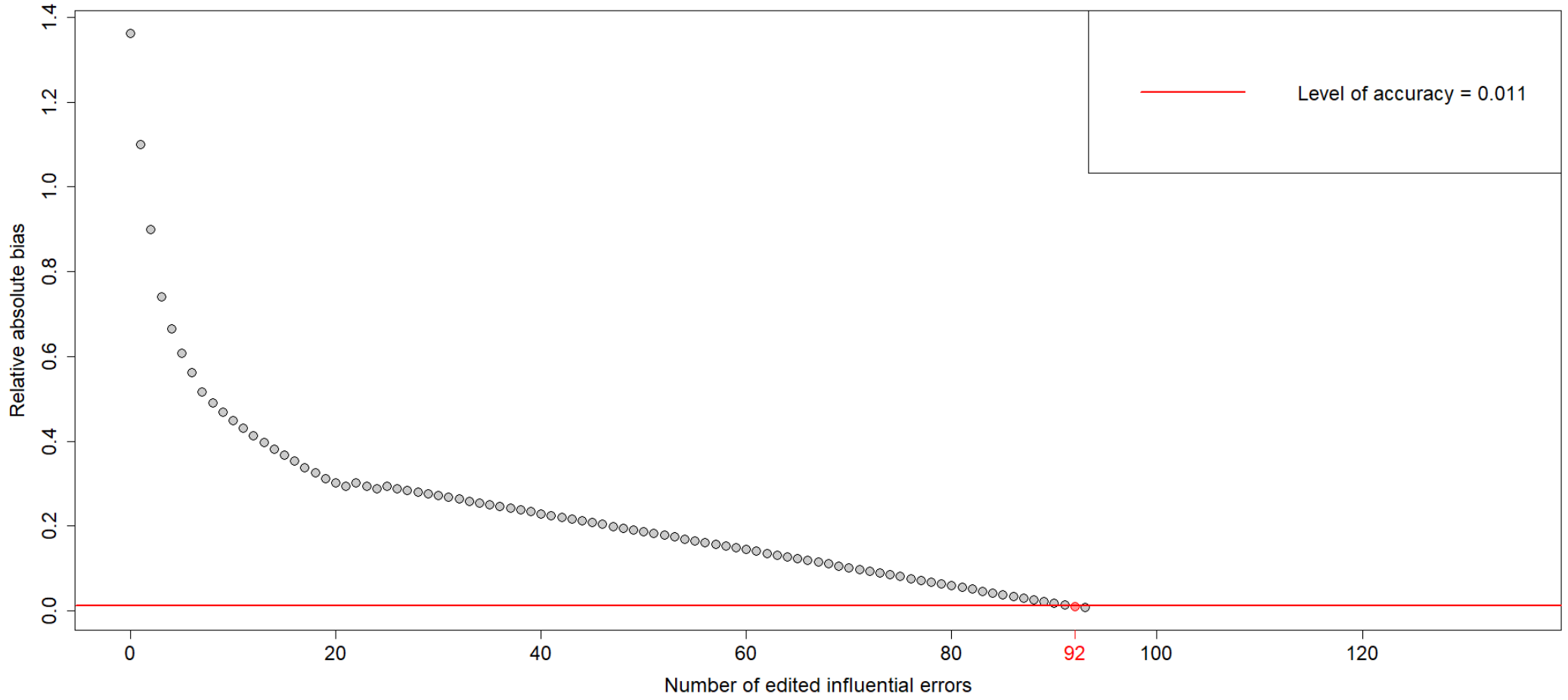


Figure 2: Relative absolute bias (RAB) dependency on the number of edited influential outliers using a *continuous suspicion component* (Burakauskaitė and Nekrašaitė-Liegė, 2022; Figure 2).

Outlier correction study

Case study by Uogelė (2023)

- The objective was to compare different missing data imputation methods on **monthly turnover** from the Monthly Statistical Survey on Trade and Catering Enterprises in Lithuania.
- Different degree of missingness was generated with missing completely at random (MCAR) and missing at random (MAR) mechanisms.
- Various machine learning based imputation methods were applied to impute the generated missing values, and the obtained results were compared according to such accuracy measures as normalized root mean squared error (NRMSE) and mean absolute error (MAE).

Used software: R

Packages:

- ***VIM*** (Kowarik and Templ, 2016) – *k*-nearest neighbor imputation according to the Gower's distance.
- ***mice*** (van Buuren and Groothuis-Oudshoorn, 2011) – imputation of multivariate data by chained equations.
- ***missForest*** (Stekhoven and Bühlmann, 2012) – imputation method based on the random forest algorithm. Alternative ***missRanger*** package by Mayer (2019) offers an option of using predictive mean matching, while *missForest* uses mean.
- ***Amelia*** (Honaker et al., 2011) – imputation based on a bootstrap expectation-maximization algorithm, producing multiple output data sets.

Missing value imputation for the *Monthly Statistical Survey on Trade and Catering Enterprises* of Statistics Lithuania (I)

Methods applied:

- K-nearest neighbor imputation (R function ***kNN***) as the reference method;
- Mean imputation in subsets of the population based on the enterprise size;
- Bayesian linear regression imputation (R function ***mice***, method “norm”);
- Stochastic regression imputation (R function ***mice***, method “norm.nob”);
- Predictive mean matching imputation (R function ***mice***, method “pmm”);
- Non-parametric missing value imputation using random forest (R function ***missForest***);
- Bootstrapping and expectation-maximization algorithm imputation (R function ***amelia***).

Missing value imputation for the *Monthly Statistical Survey on Trade and Catering Enterprises* of Statistics Lithuania (II)

Auxiliary information used for predicting enterprise turnover of the accounting period t (y_t):

- Enterprise turnover of the previous period (y_{t-1});
- Enterprise turnover from VAT declarations of the previous period (y^*_{t-1});
- Enterprise turnover from VAT declarations of the accounting period (y^*_t);
- Categorical variable of four-digit numerical code (classes) of economic activity group;
- Categorical variable for 7 enterprise size groups based on the number of employees.

Missing data imputation under MCAR assumption

	NRMSE				MAE ($\times 10^3$)			
Missing	**5%**	**10%**	**20%**	**30%**	**5%**	**10%**	**20%**	**30%**
Mean	0.83	0.77	0.75	0.70	212.08	160.98	145.31	159.84
MissForest	0.09	0.13	0.12	0.16	19.95	17.42	16.50	24.76
MissRanger	0.15	0.18	0.13	0.16	28.35	21.42	17.09	23.13

Table 1: NRMSE and MAE of monthly enterprise turnover imputation under MCAR response mechanism (Uogelė, 2023; Table 6).

Missing data imputation under MAR assumption

Missing	NRMSE				MAE ($\times 10^3$)			
	5%	10%	20%	30%	5%	10%	20%	30%
MICE-pmm	0.54	0.41	0.51	0.47	88.70	90.10	84.79	97.32
MICE-norm	0.11	0.07	0.24	0.10	25.11	22.98	30.50	24.43
MICE-norm.nob	0.12	0.07	0.14	0.09	24.67	21.28	25.40	24.90
kNN	0.56	0.43	0.43	0.47	120.55	118.75	92.18	112.65
MissForest	0.18	0.14	0.17	0.26	34.89	33.66	34.39	42.85
MissRanger	0.27	0.23	0.21	0.24	42.08	45.75	39.66	43.69

Table 2: NRMSE and MAE of monthly enterprise turnover imputation under MAR response mechanism (Uogelė, 2023; Table 21).

Final remarks

- The planned statistical data E&I process is subject to minor change due to project execution limitations, such as IT infrastructure, software solution and knowledge capacity.
- Possible future plans include continuous research on statistical data editing methods, as well as collaboration with academic community, and capacity building at Statistics Lithuania.

Any suggestions would be highly appreciated!

References (I)

- Belcher, R. (June 2003). *Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys*. Paper presented at the SSC Annual Meeting, Halifax, Nova Scotia, Canada.
- Burakauskaitė, I. and Nekrašaitė-Liegė, V. (2022). Selective Editing Using Contamination Model. *Romanian Statistical Review*, 1:55–65.
- Di Zio, M. and Guarnera, U. (2013). A Contamination Model for Selective Editing. *Journal of Official Statistics*, 29(4):539–555.
- Guarnera, U. and Buglielli, M. T. (2013). SeleMix: an R Package for Selective Editing.
- Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45:1–47.
- Kowarik, A. and Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74:1–16.

References (II)

- Mayer, M. (June 2019). missRanger: An R-package for Fast Imputation of Missing Values, Version 2.6.0, updated August 2024.
- Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P. and Nordberg, L. (2010). *A General Methodology for Selective Data Editing*. Stockholm: Statistics Sweden.
- Stekhoven, D. J. and Bühlmann, P. (2012). MissForest – Non-parametric Missing Value Imputation for Mixed-type Data. *Bioinformatics*, 28(1):112–118.
- Uogelė, J. (2023). *Missing Data Imputation Methods for Monthly Statistical Survey on Trade and Catering Enterprises* [Unpublished master's thesis]. Vilnius University.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45:1–67.