TALL SOM FORTELLER

Too good to be true?
Machine learning in the editing process

EVA CHARLOTTE BERNER AND SOLVEIG BJØRKHOLT

Statistisk sentralbyrå
Statistics Norway

Who are we?

- Eva Charlotte Berner
  - Statistics Norway
  - Mathematics, IT, Economics, Data Science
  - EvaCharlotte.Berner@ssb.no

- Solveig Bjørkholt
  - University of Oslo and Statistics Norway
  - Political Science, Data Science
  - Solveig.Bjorkholt@ssb.no

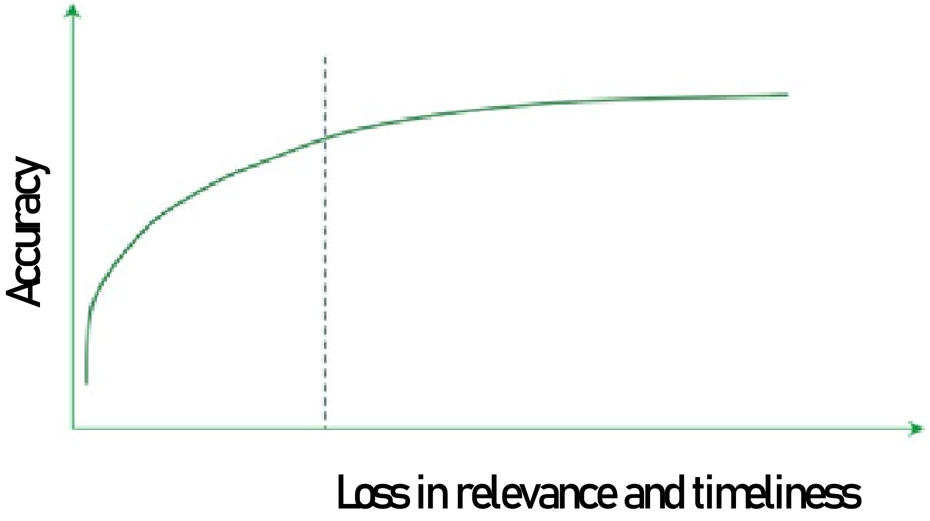
Machine Learning in Editing (MLiE)

- Research question: How can we use machine learning to improve the editing process?

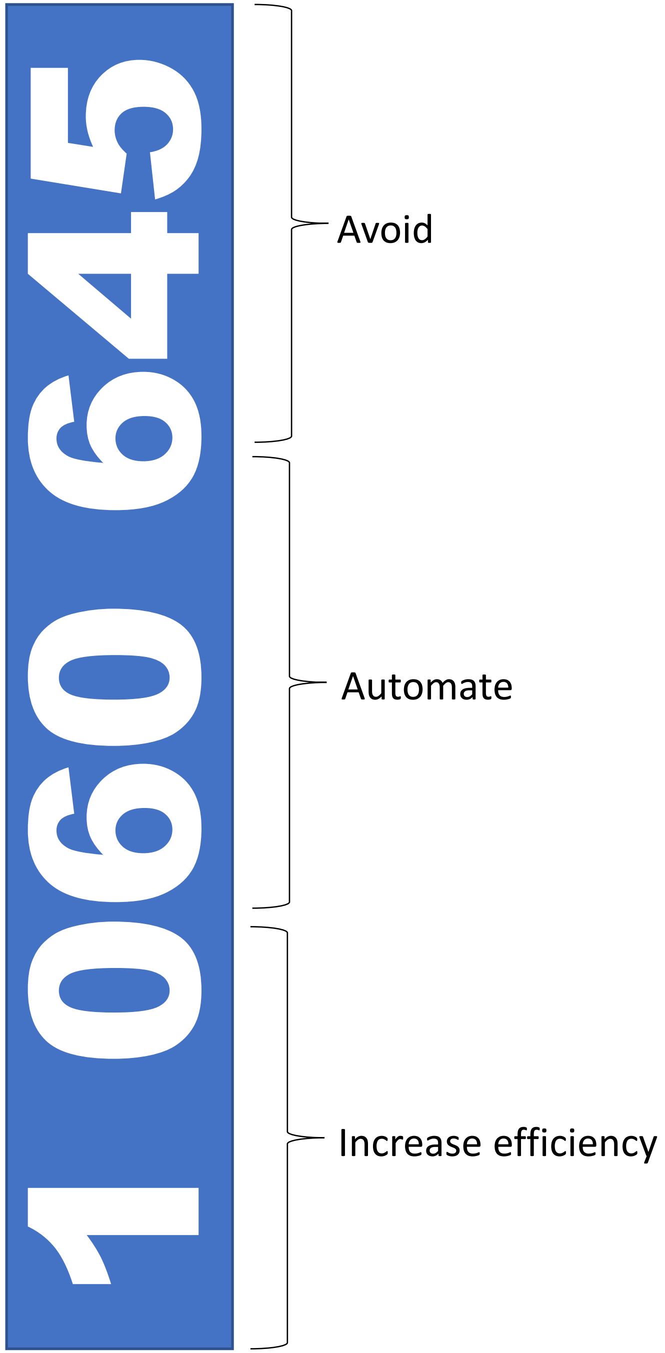
- Our case: Research and development (R&D) in the Norwegian Business Sector

The current editing process is…

- inefficient
  - Mostly manual
- hard to transfer
  - Built on a lot of undocumented know-how
- very hard to reproduce
  - All that exists is a huge log



Statistisk sentralbyrå
Statistics Norway



1 060 645

Avoid

Automate

Increase efficiency

Average time per editing	Hours	Man-hours
30 seconds	8 839	5
1 minute	17 677	10
2 minutes	35 355	20
5 minutes	88 387	50
10 minutes	176 774	101
20 minutes	353 548	202

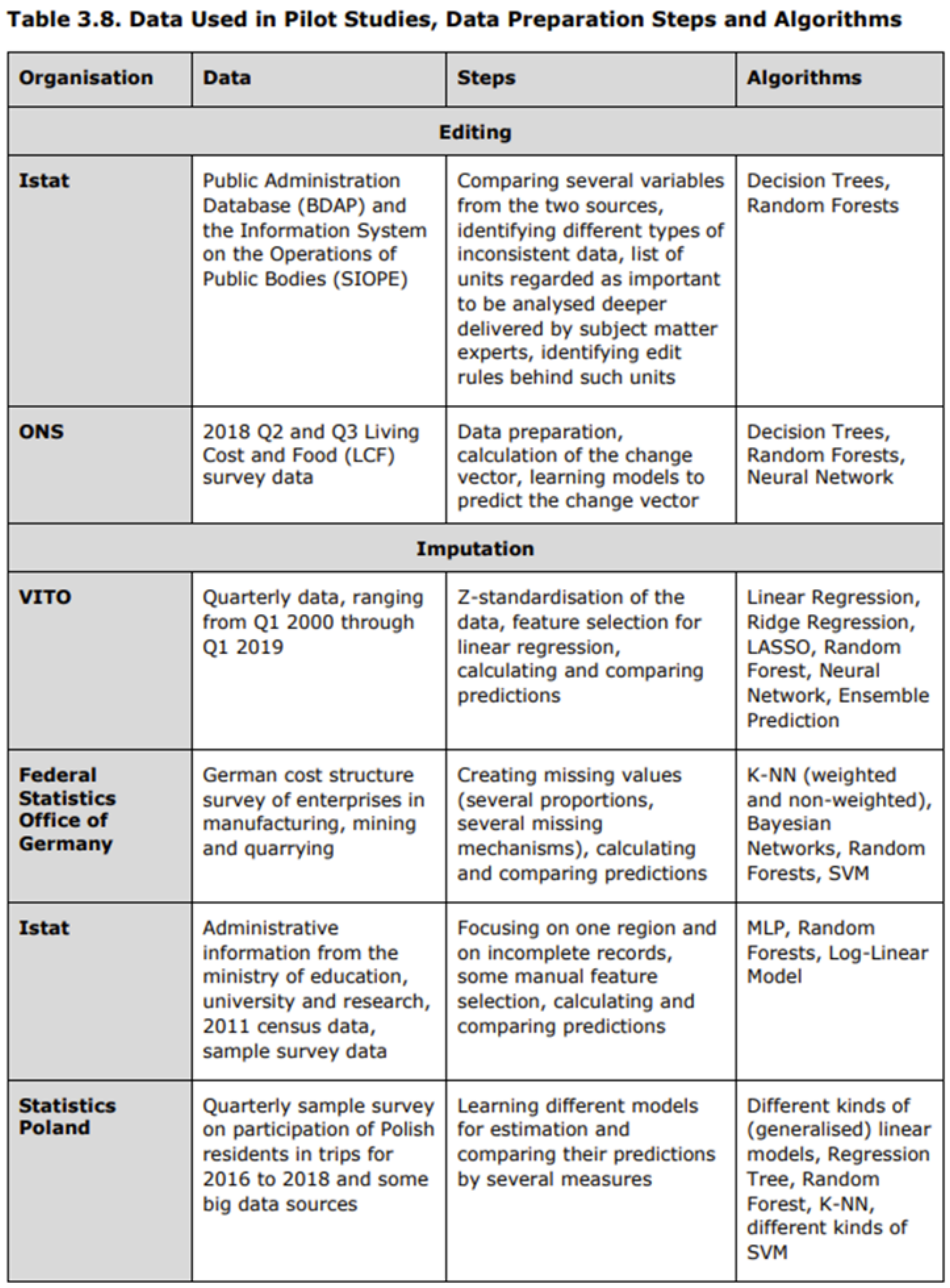
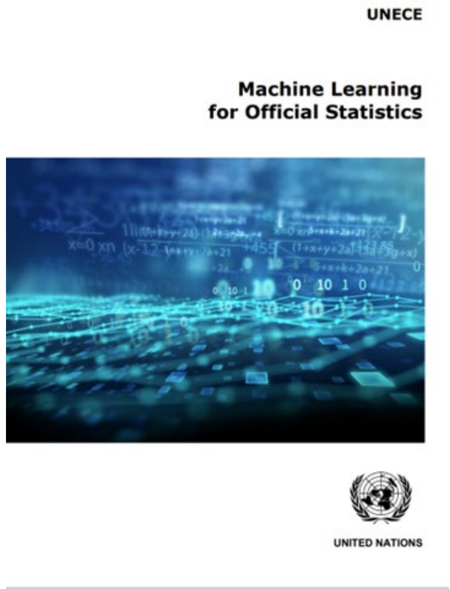


Table 3.8. Data Used in Pilot Studies, Data Preparation Steps and Algorithms

Organisation	Data	Steps	Algorithms
Editing			
Istat	Public Administration Database (BDAP) and the Information System on the Operations of Public Bodies (SIOPE)	Comparing several variables from the two sources, identifying different types of inconsistent data, list of units regarded as important to be analysed deeper delivered by subject matter experts, identifying edit rules behind such units	Decision Trees, Random Forests
ONS	2018 Q2 and Q3 Living Cost and Food (LCF) survey data	Data preparation, calculation of the change vector, learning models to predict the change vector	Decision Trees, Random Forests, Neural Network
Imputation			
VITO	Quarterly data, ranging from Q1 2000 through Q1 2019	Z-standardisation of the data, feature selection for linear regression, calculating and comparing predictions	Linear Regression, Ridge Regression, LASSO, Random Forest, Neural Network, Ensemble Prediction
Federal Statistics Office of Germany	German cost structure survey of enterprises in manufacturing, mining and quarrying	Creating missing values (several proportions, several missing mechanisms), calculating and comparing predictions	K-NN (weighted and non-weighted), Bayesian Networks, Random Forests, SVM
Istat	Administrative information from the ministry of education, university and research, 2011 census data, sample survey data	Focusing on one region and on incomplete records, some manual feature selection, calculating and comparing predictions	MLP, Random Forests, Log-Linear Model
Statistics Poland	Quarterly sample survey on participation of Polish residents in trips for 2016 to 2018 and some big data sources	Learning different models for estimation and comparing their predictions by several measures	Different kinds of (generalised) linear models, Regression Tree, Random Forest, K-NN, different kinds of SVM

«Editing is a very realistic task for machine learning algorithms»

- UNECE report 2022

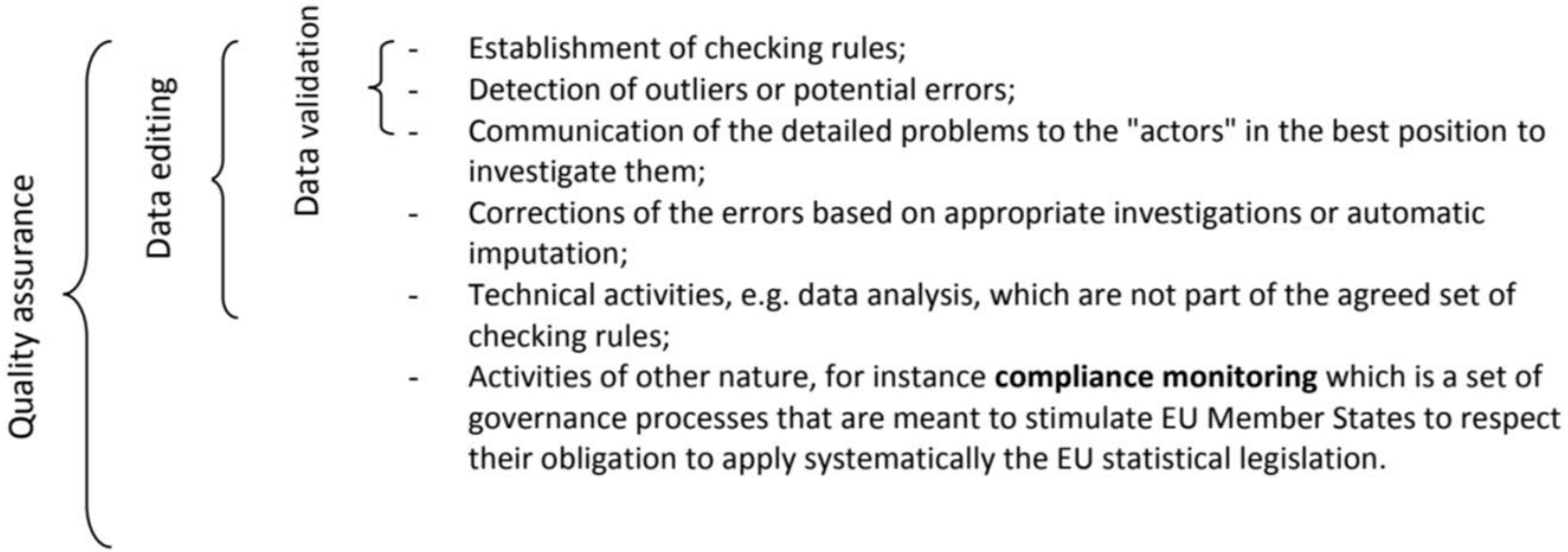


UNECE

Machine Learning for Official Statistics

UNITED NATIONS

Editing is Validation + Imputation



- Establishment of checking rules;
- Detection of outliers or potential errors;
- Communication of the detailed problems to the "actors" in the best position to investigate them;
- Corrections of the errors based on appropriate investigations or automatic imputation;
- Technical activities, e.g. data analysis, which are not part of the agreed set of checking rules;
- Activities of other nature, for instance **compliance monitoring** which is a set of governance processes that are meant to stimulate EU Member States to respect their obligation to apply systematically the EU statistical legislation.

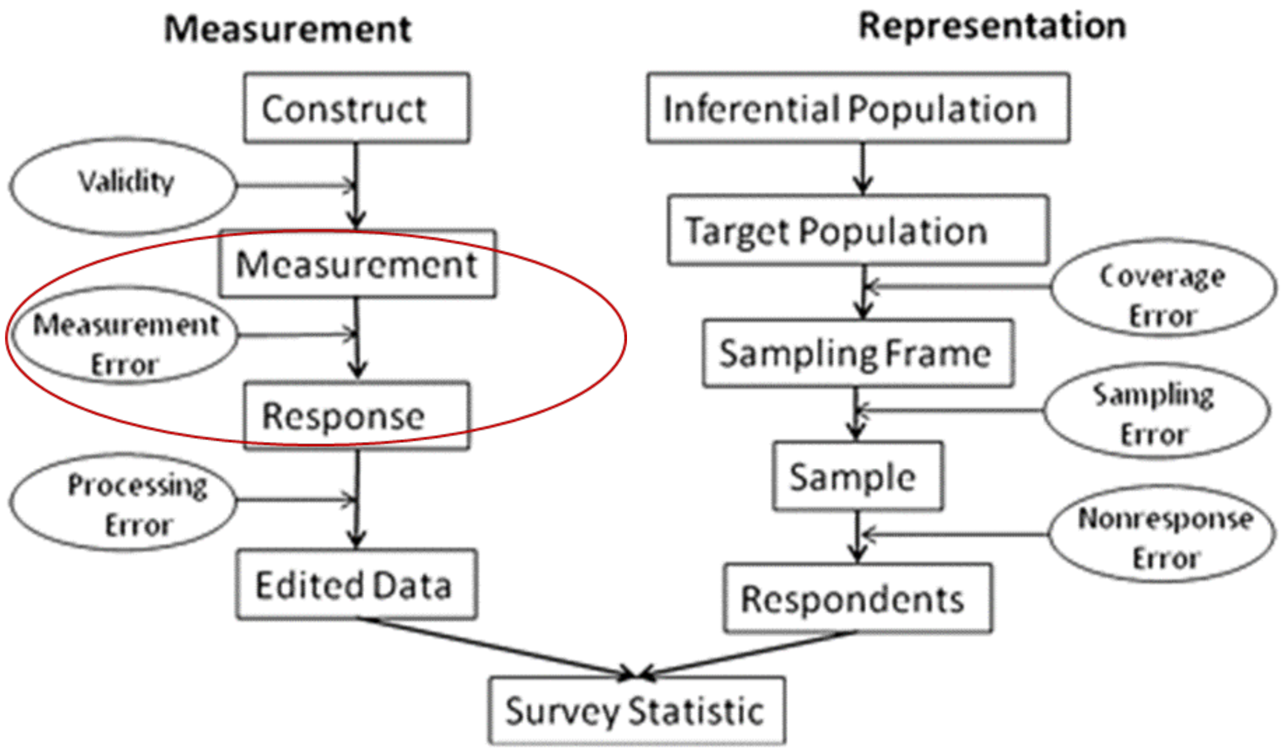
Methodology for data validation 1.0 (europa.eu)

Data editing removes measurement error

- Overall goal:
  - Eliminate all sources of error that makes the survey value deviate from the true value (Total Survey Error)

- Goal in our study:

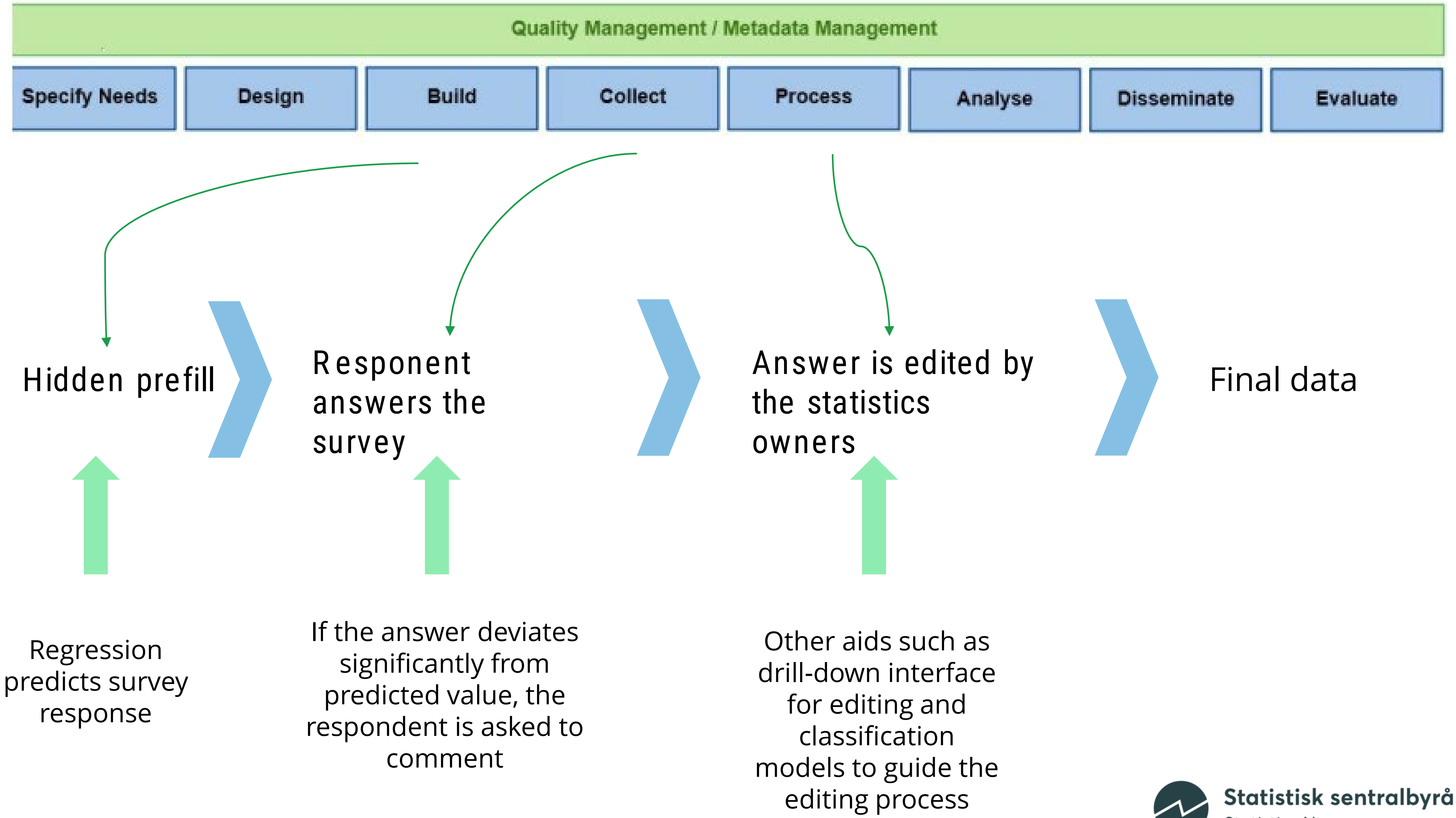
  Eliminate **measurement error**



Two machine learning problems constitute MLiE

- **The prediction problem**: Create a machine learning algorithm that predicts the individual responses to a survey.

- **The classification problem:** Create a machine learning algorithm that classifies that resulting survey dataset into *valid* and *not valid* responses.



Quality Management / Metadata Management

Specify Needs | Design | Build | Collect | Process | Analyse | Disseminate | Evaluate

Hidden prefill

Respondent answers the survey

Answer is edited by the statistics owners

Final data

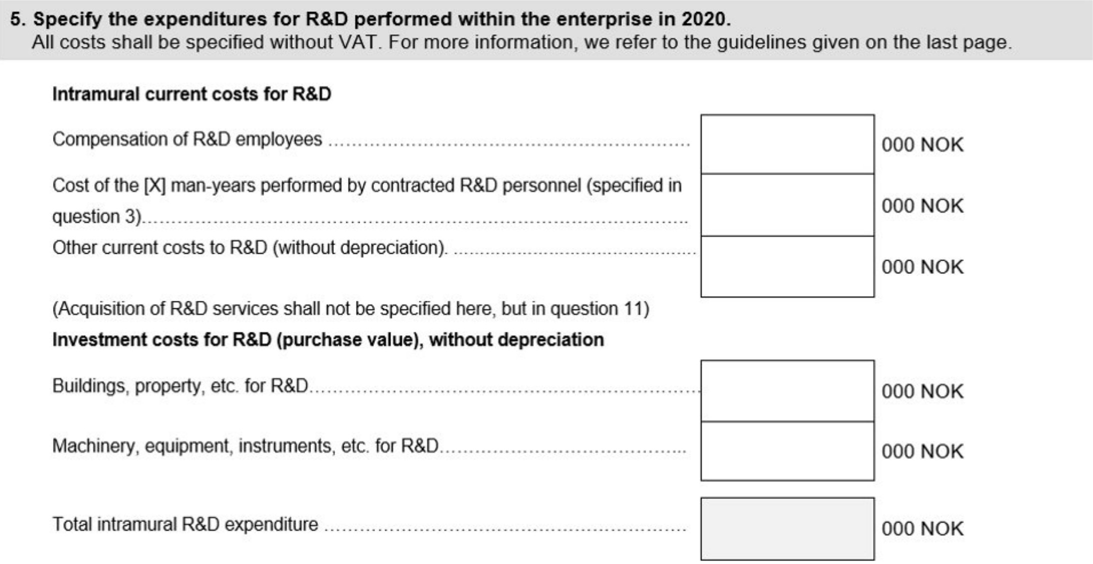
Regression predicts survey response

If the answer deviates significantly from predicted value, the respondent is asked to comment

Other aids such as drill-down interface for editing and classification models to guide the editing process

Statistisk sentralbyrå
Statistics Norway

The prediction problem



5. Specify the expenditures for R&D performed within the enterprise in 2020.
All costs shall be specified without VAT. For more information, we refer to the guidelines given on the last page.

Intramural current costs for R&D

Compensation of R&D employees		000 NOK
Cost of the [X] man-years performed by contracted R&D personnel (specified in question 3)..........		000 NOK
Other current costs to R&D (without depreciation).		000 NOK

(Acquisition of R&D services shall not be specified here, but in question 11)

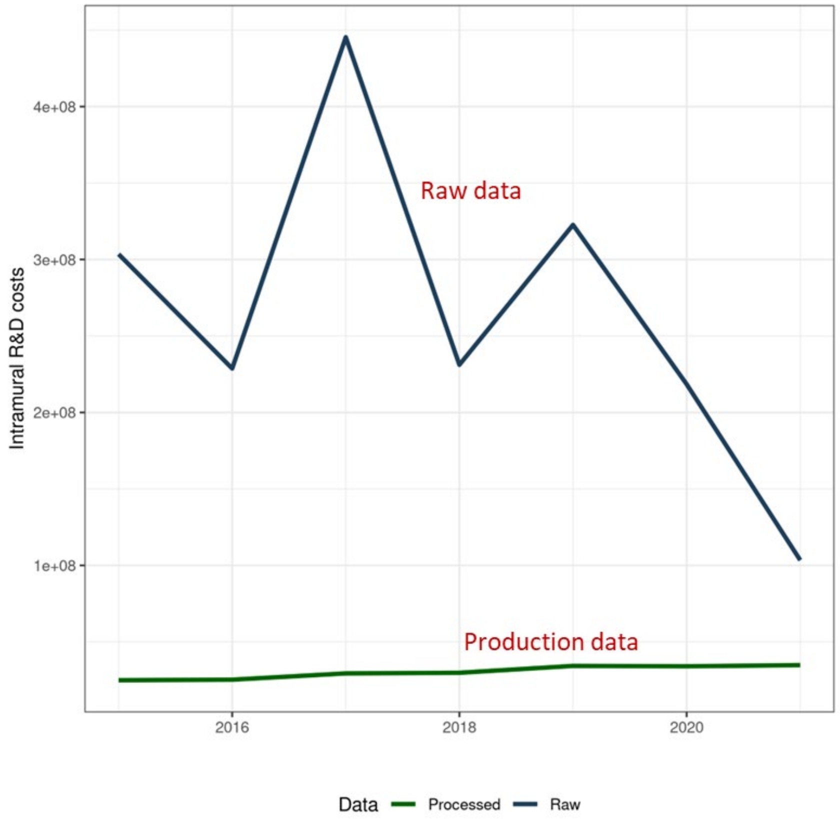
Investment costs for R&D (purchase value), without depreciation

Buildings, property, etc. for R&D..........		000 NOK
Machinery, equipment, instruments, etc. for R&D..........		000 NOK
Total intramural R&D expenditure		000 NOK

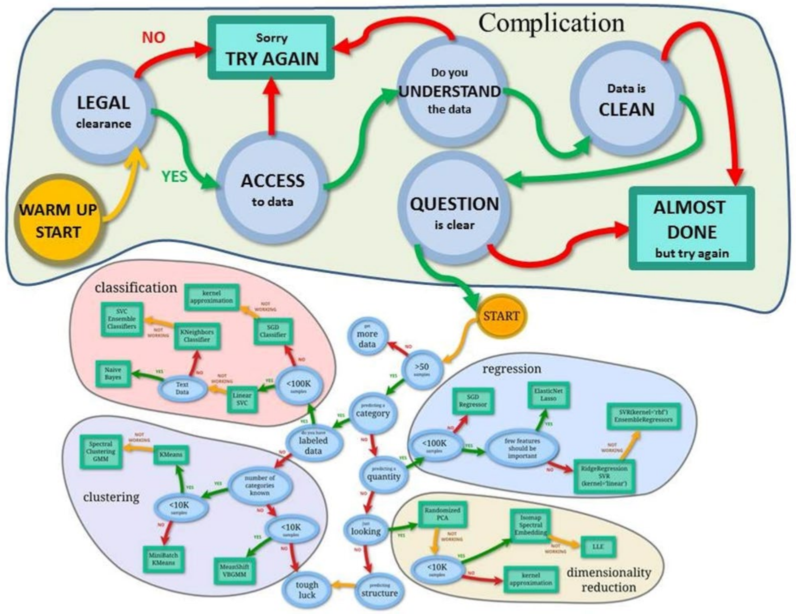
- Focus on the *intfou* variable
  - Intramural R&D expenditure
- One of the main variables in the R&D survey
- The total of:
  - Compensation to R&D employees
  - Cost to hired R&D personnel
  - Other current costs
  - Investments

Measurement error in the *intfou* variable can be categorized into:

- "Thousand-error"
- Report daughter companies' R&D
- Not having correct or enough information
- Misunderstand terminology (e.g. "own employees" vs. "hired personnel")



Data wrangling



- Collect and merge datasets
- Remove duplicates
- Identify and flag outliers
- Impute and/or remove missing values

Final: Survey timeseries 2011 - 2021

Research framework

- $Y_{ikt} = \text{function}(\text{predictors})$

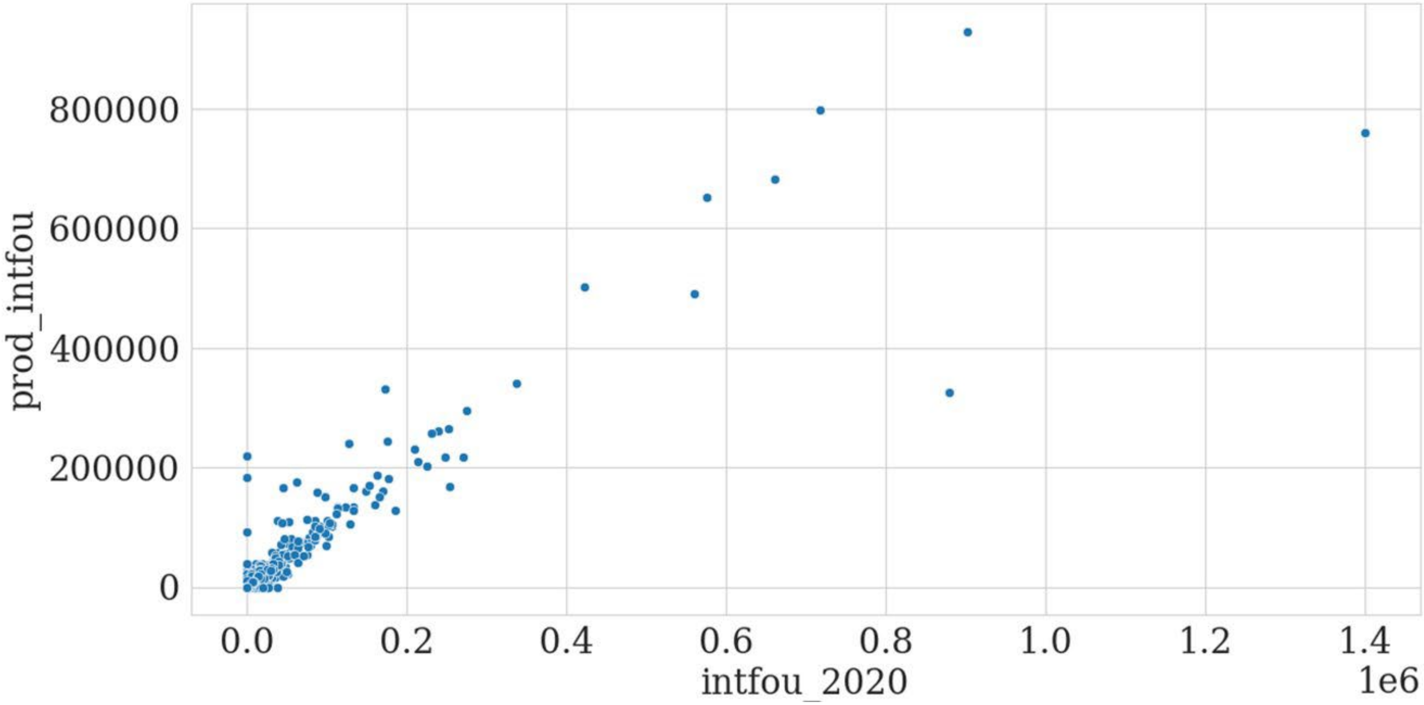
  - Y = expected value
  - for enterprise *i*
  - for variable *k*
  - for year *t*

Regression

Y = expected R&D expenditure

$$intfou_{it} = \alpha + \beta_1 intfou_{it-1} + \beta_2 intfou_{it-2} + \beta_3 intfou_{it-3} + \beta_4 intfou_{it-4} + \beta_5 intfou_{it-5} + \beta_6 intfou_{it-6} + \varepsilon$$

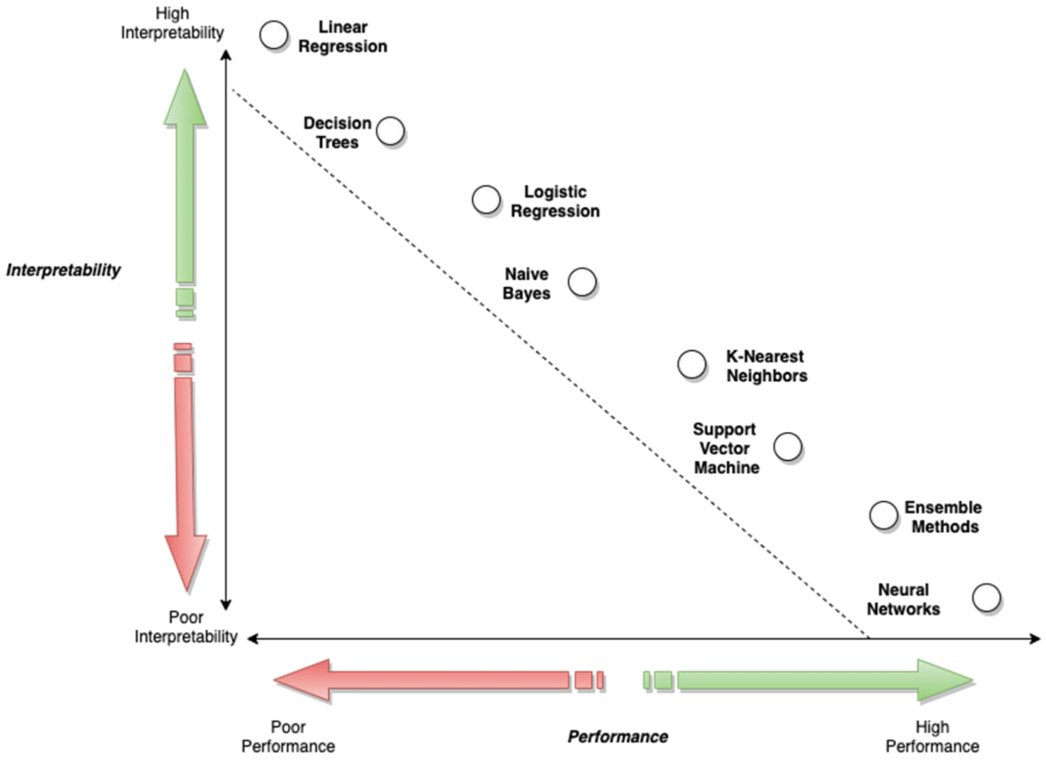
High correlation between last year and this year



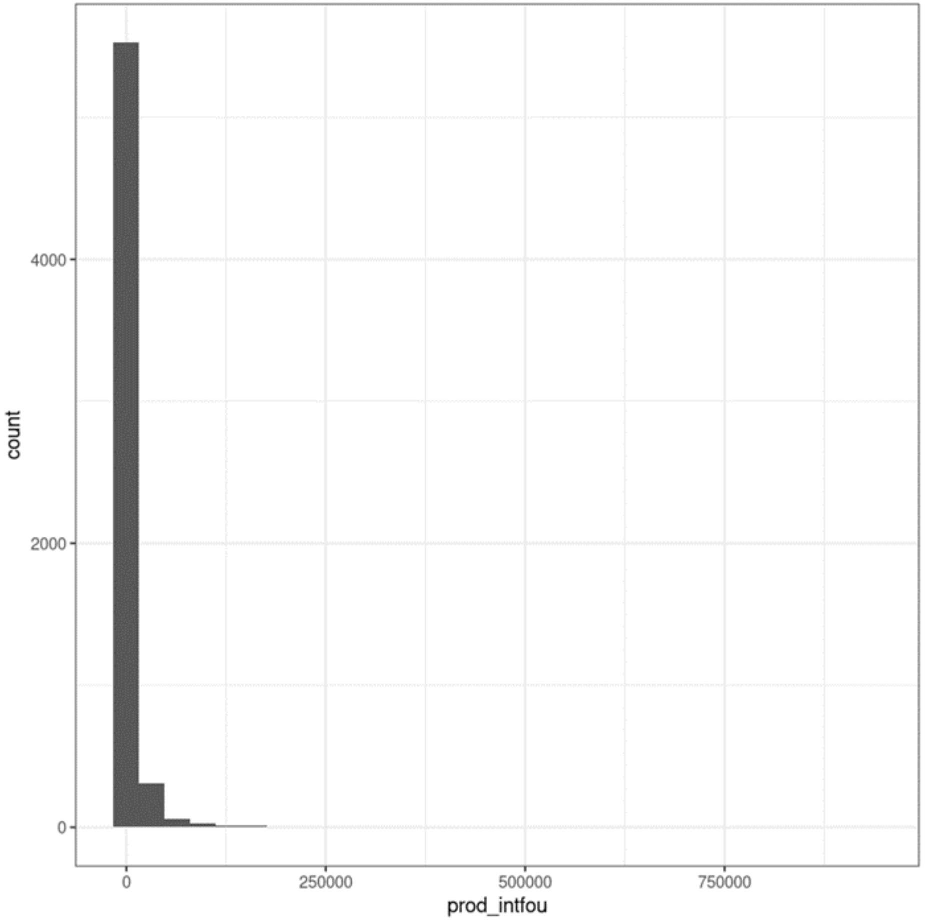
Deciding on the model

- Linear relationships allow for
  - OLS
  - Lasso
  - Ridge
  - ElasticNet

- Benefit: Easier to interpret than more complex models



Challenge: Very imbalanced data



Preliminary results

- If a company does not perform R&D, it is unlikely to perfom R&D next year.
- Historical values account for 60% explained variables for only R&D enterprises. Something else must be causing the other 40%.
  - From interviews we suspect structual data, M&As, Demergers may play role.

	Explained variance	RMSE
Linear	0.93	9664
Ridge	0.93	9723
Lasso	0.93	9651
Elastic Net	0.80	16296

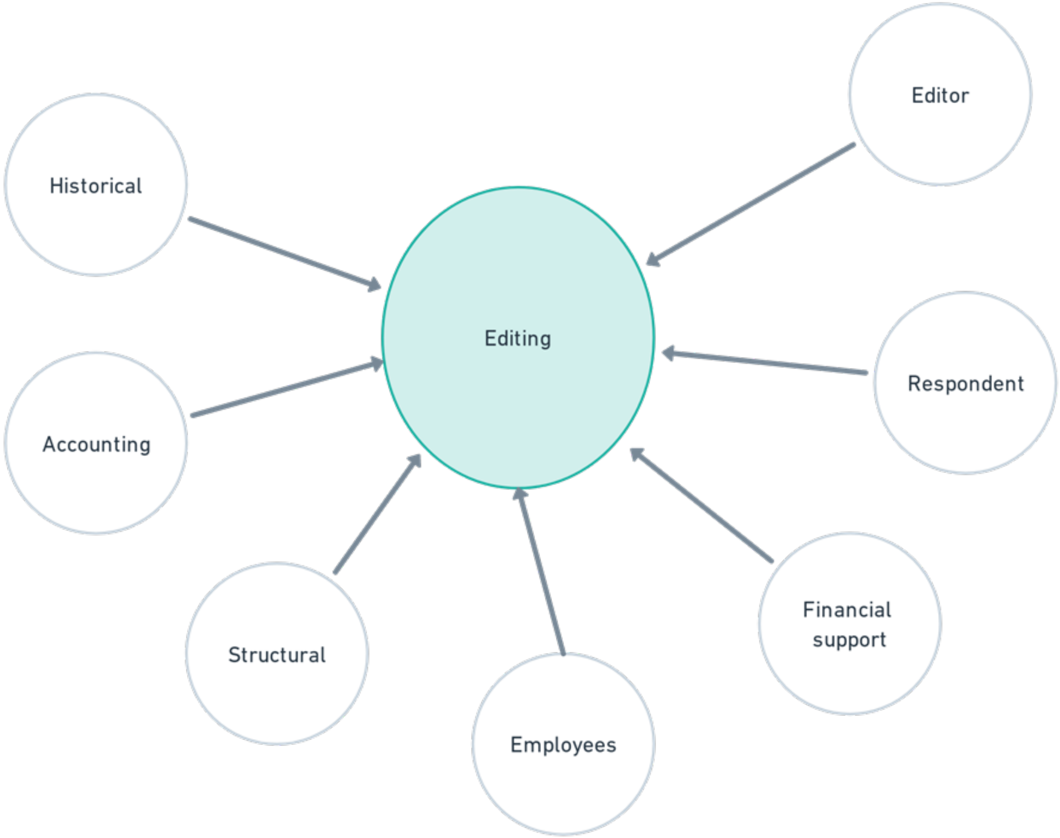
Tab 2: Model performance for all enterprises

	Explained variance	RMSE
Linear	0.60	25624
Ridge	0.60	25496
Lasso	0.60	25654
Elastic Net	0.67	23154

Tab 3: Model performance for R&D enterprises only

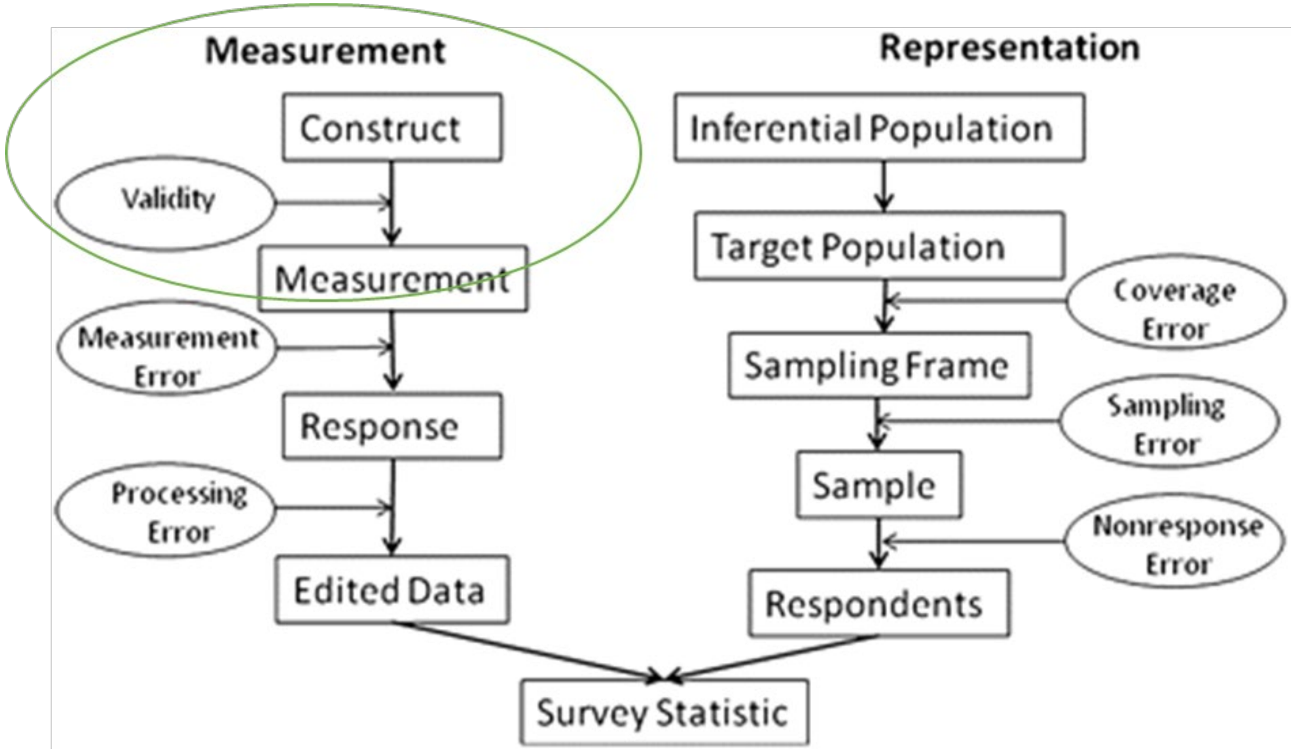
Future work

- Increase model performance for R&D enterprises only by include more variables. Specifically, M&A data.
- Develop classification algorithm.
- Generalise methodology.



Other survey errors:

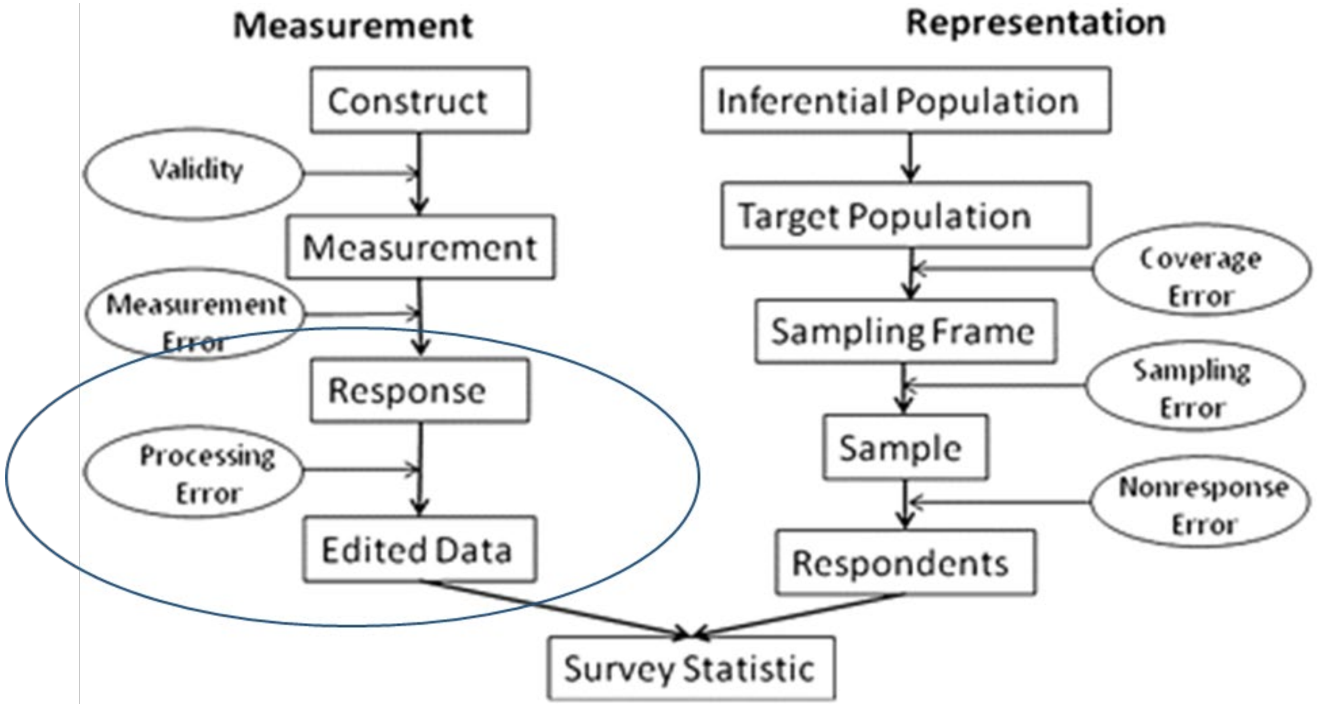
- **Measurement validity**
  - Clear definitions
  - Good examples
  - Visit and interview enterprises
  - Training in enterprises



Statistisk sentralbyrå
Statistics Norway

Other survey errors:

- **Processing error**
  - Machine learning algorithms reproduce processing errors from training data
  - Using other sources to double-check the training data, e.g. annual reports, call enterprises, read reports, do field research, etc.



Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Thank you!

Statistisk sentralbyrå
Statistics Norway