cbs

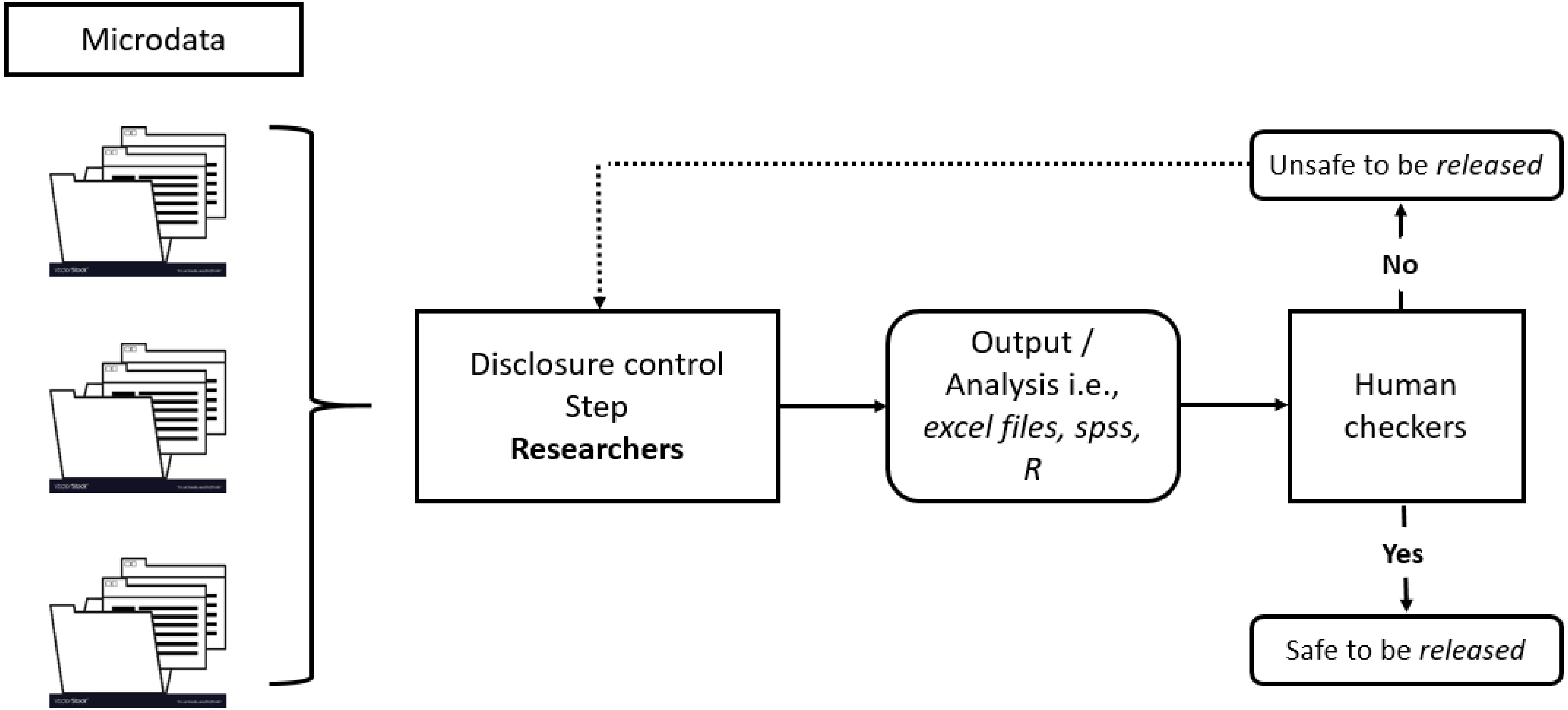
COACH: Computer-Assisted output Checking with Human-in-the-Loop

Manel Slokom, Jel Vankan, Peter-Paul De Wolf, Martha Larson

UNECE Expert meeting on Statistical Data Confidentiality 2023, Wiesbaden, Germany

September 2023

Introduction - Context



Introduction - Problems

- Traditional output checking are based on **manual** inspection
  - Time consuming
  - Resource intensive
- Green et al., (2021): ACRO (Automatic Checking of Research Output)

- To build machine learning models capable of predicting whether an output is safe for release or not:
  - Domingo et al., (2021)

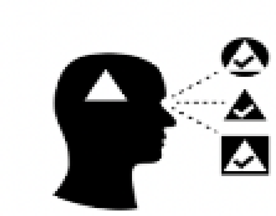
Green, E., F. Ritchie, and J. Smith (2021). Automatic checking of research outputs (acro): A tool for dynamic disclosure checks. ESS Statistical Working Papers 2021 Edition.
Domingo-Ferrer, J. and A. Blanco-Justicia (2021). Towards machine learning-assisted output checking for statistical disclosure control. In Proceedings of 18th International Conference on Modeling Decisions for Artificial Intelligence: MDAI 2021

Research Questions

- *How can we **semi-automate** output checking using machine learning ?*
    - ➢ How can we extend Domingo et al., (2021) work, i.e., on real data?
    - ➢ How can we involve **human checkers** in the process of training machine learning models?

Solution: COACH

Facilitate output checking process



Reduce human bias

Include human-in-the-loop

Experimentation Setup

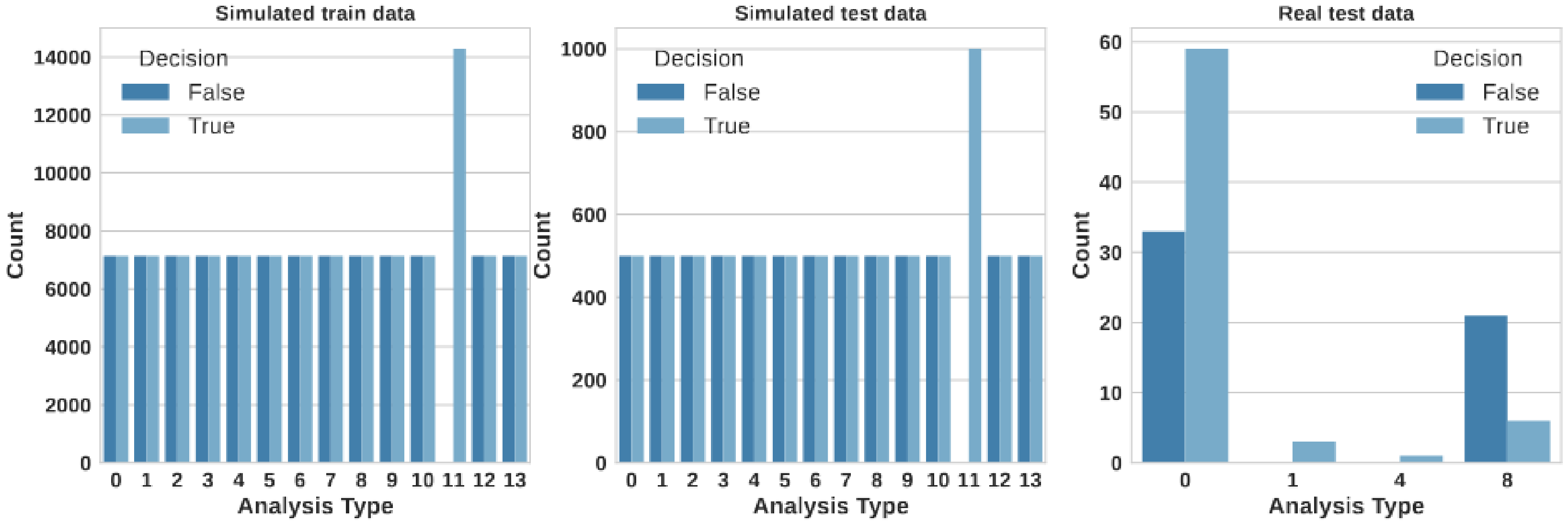
Experimental setup

Data Sets

- Simulated data (following Domingo et al., (2021)):
  - 14 rules of thumbs
  - 200K records in the training data and 14K records in the test data
  - Every rule has approx. 14700 records in the training data and 1000 in the test data

- Real data
  - 125 records dominated by frequency table, magnitude table, and regression model
  - Pre-processing step

Experimental Setup



Experimental setup

- Neural network
- LightGBM (LGBM)
- Random classifier

Experimentation Results

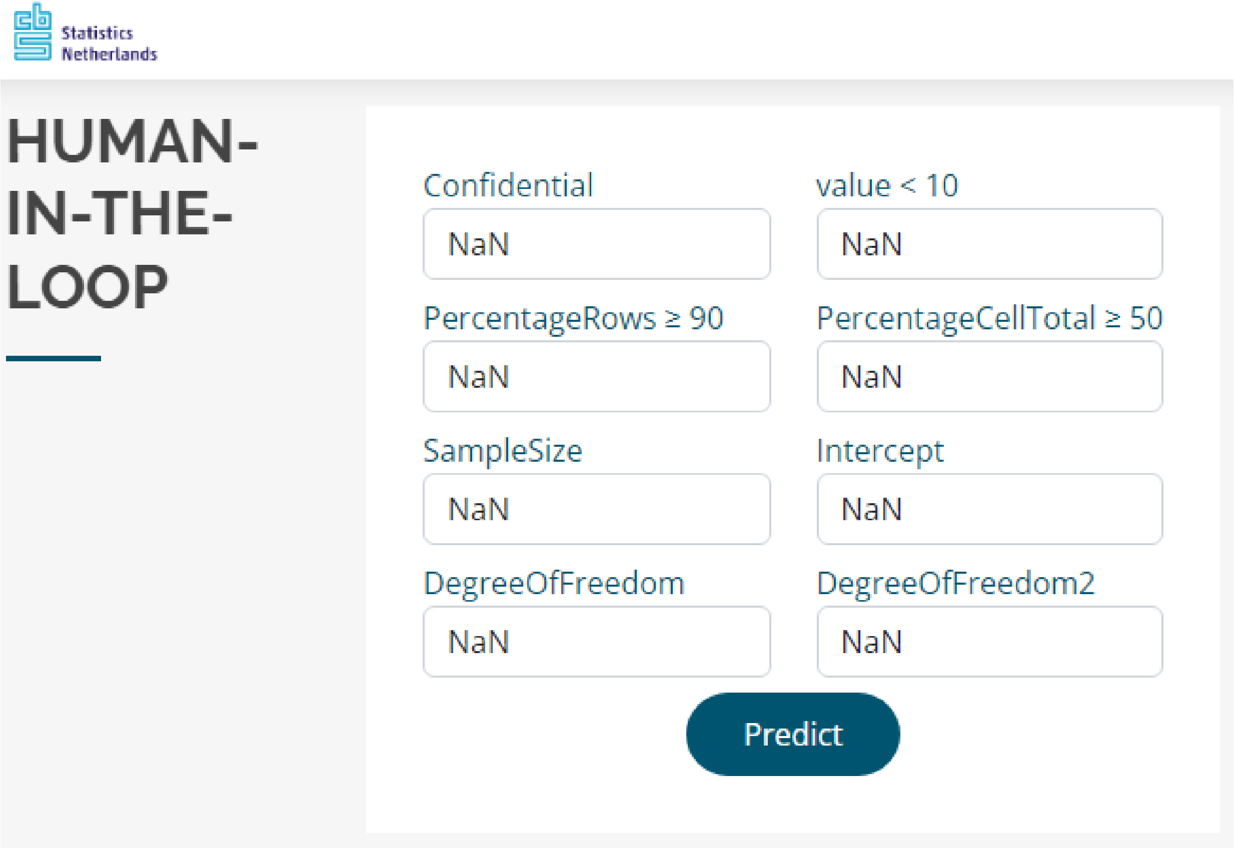
Prediction performance

Data Sets	*HITL*	Classifier	**F1 (Macro)**	**MCC**	**G-Mean**	**TP**	**FP**	**TN**	**FN**
Simulated Data	*None*	*Random*	0.3488	0.0000	0.5000	0	6500	0	7500
		LGBM	0.8489	0.7376	0.8599	6500	0	2101	5399
		Neural Network	0.9421	0.8838	0.9421	6123	377	433	7067

Evaluation of model trained on simulated data applied to real test data

Data Sets	*HITL*	Classifier	**F1 (Macro)**	**MCC**	**G-Mean**	**TP**	**FP**	**TN**	**FN**
Simulated Data	*None*	*Random*	0.3488	0.0000	0.5000	0	6500	0	7500
		LGBM	0.8489	0.7376	0.8599	6500	0	2101	5399
		Neural Network	0.9421	0.8838	0.9421	6123	377	433	7067
Real test data	*None*	*LGBM*	0.6139	0.4052	0.6409	16	**38**	1	68

Incorporating Human-in-the-loop with COACH



Incorporating Human-in-the-loop with COACH

Your output is

" Safe "

Ready to be released!

Summary

Confidential	0.0	value < 10	0.0
PercentageRows ≥ 90	nan	PercentageCellTotal ≥ 50	nan
SampleSize ≥ 90	nan	Intercept	nan
Degree of freedom 1	nan	Degree of freedom 2	nan

Feedback

Agree

Remark

Remarks...

Save

Your output is

" Unsafe "

A second protection is strongly required!

Summary

Confidential	1.0	value < 10	1.0
PercentageRows ≥ 90	1.0	PercentageCellTotal ≥ 50	nan
SampleSize ≥ 90	nan	Intercept	nan
Degree of freedom 1	nan	Degree of freedom 2	nan

Feedback

Agree

Remark

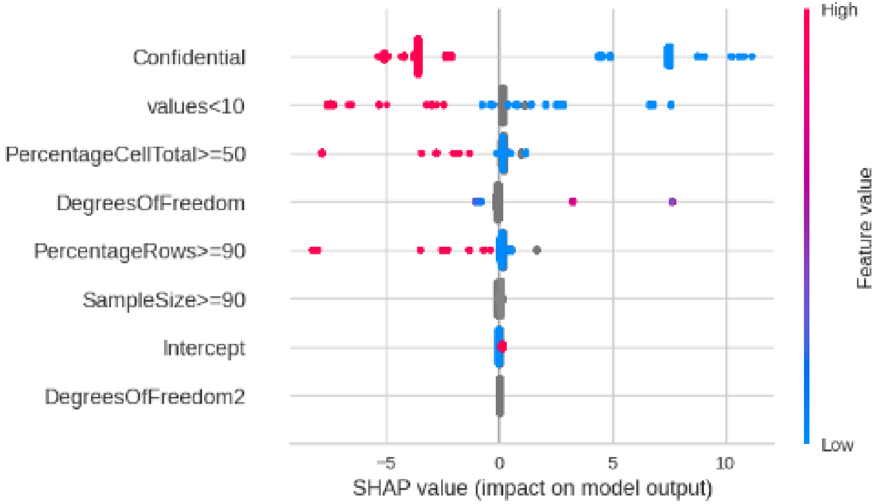
Remarks...

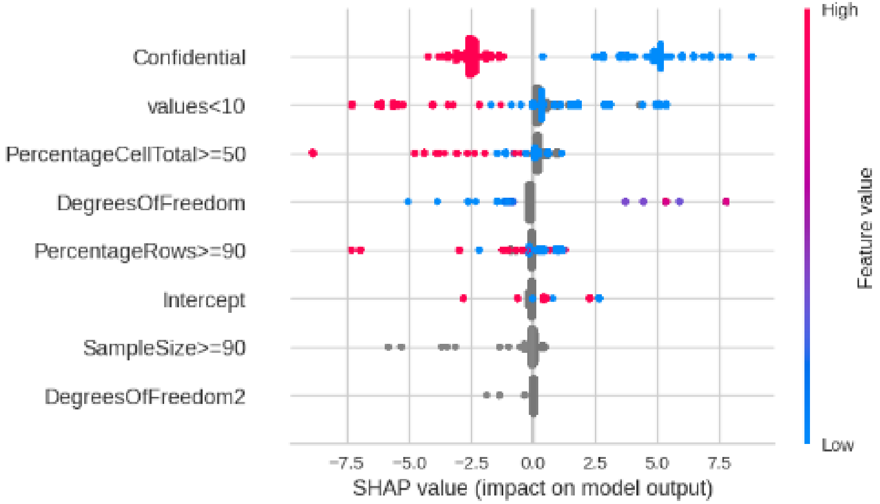
Save

Incorporating Human-in-the-loop with COACH

Data Sets	*HITL*	Classifier	**F1 (Macro)**	**MCC**	**G-Mean**	**TP**	**FP**	**TN**	**FN**
Simulated Data	*None*	*Random*	0.3488	0.0000	0.5000	0	6500	0	7500
		LGBM	0.8489	0.7376	0.8599	6500	0	2101	5399
		Neural Network	0.9421	0.8838	0.9421	6123	377	433	7067
Real test data	*None*	*LGBM*	0.6139	0.4052	0.6409	16	38	1	68
Simulated Data	*With*	*Random*	0.3488	0.0000	0.5000	0	6500	0	7500
		LGBM	0.8489	0.7376	0.8599	6500	0	2101	5399
Real test data	*With*	*LGBM*	0.9099	0.8229	0.9143	51	3	8	61

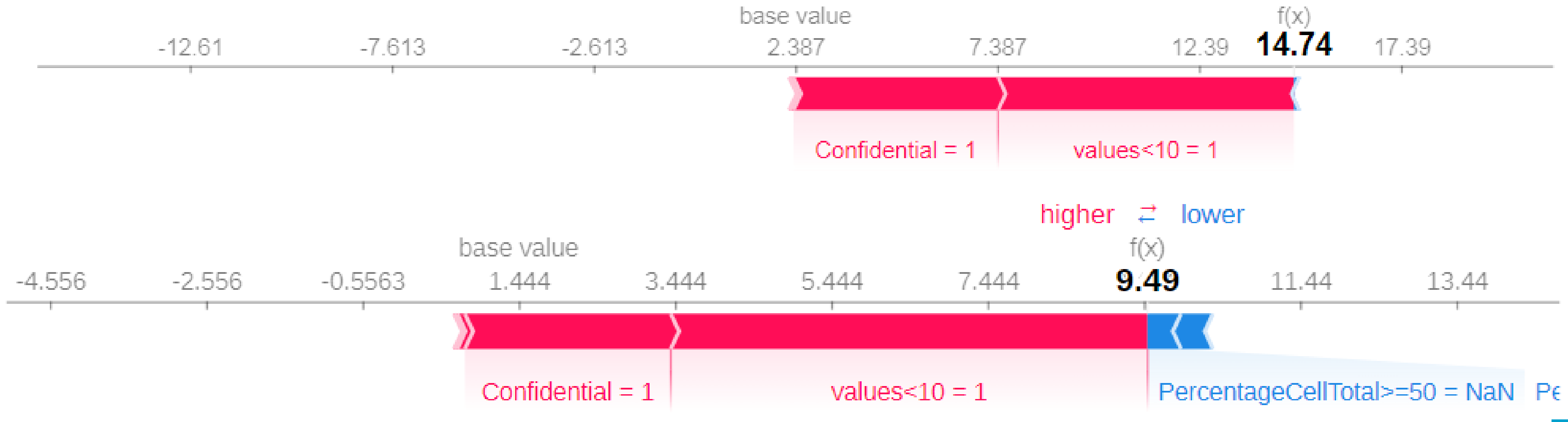
Utilizing Global SHAP Values





Global interpretability: the collective SHAP values show how much each predictor contributes, either positively or negatively, to the target variable.

Utilizing Local SHAP Values



Local interpretability using reasoning plots for an individual case in test data.

Conclusion & Future Work

Conclusion

- Extend Domingo et al, (2021).

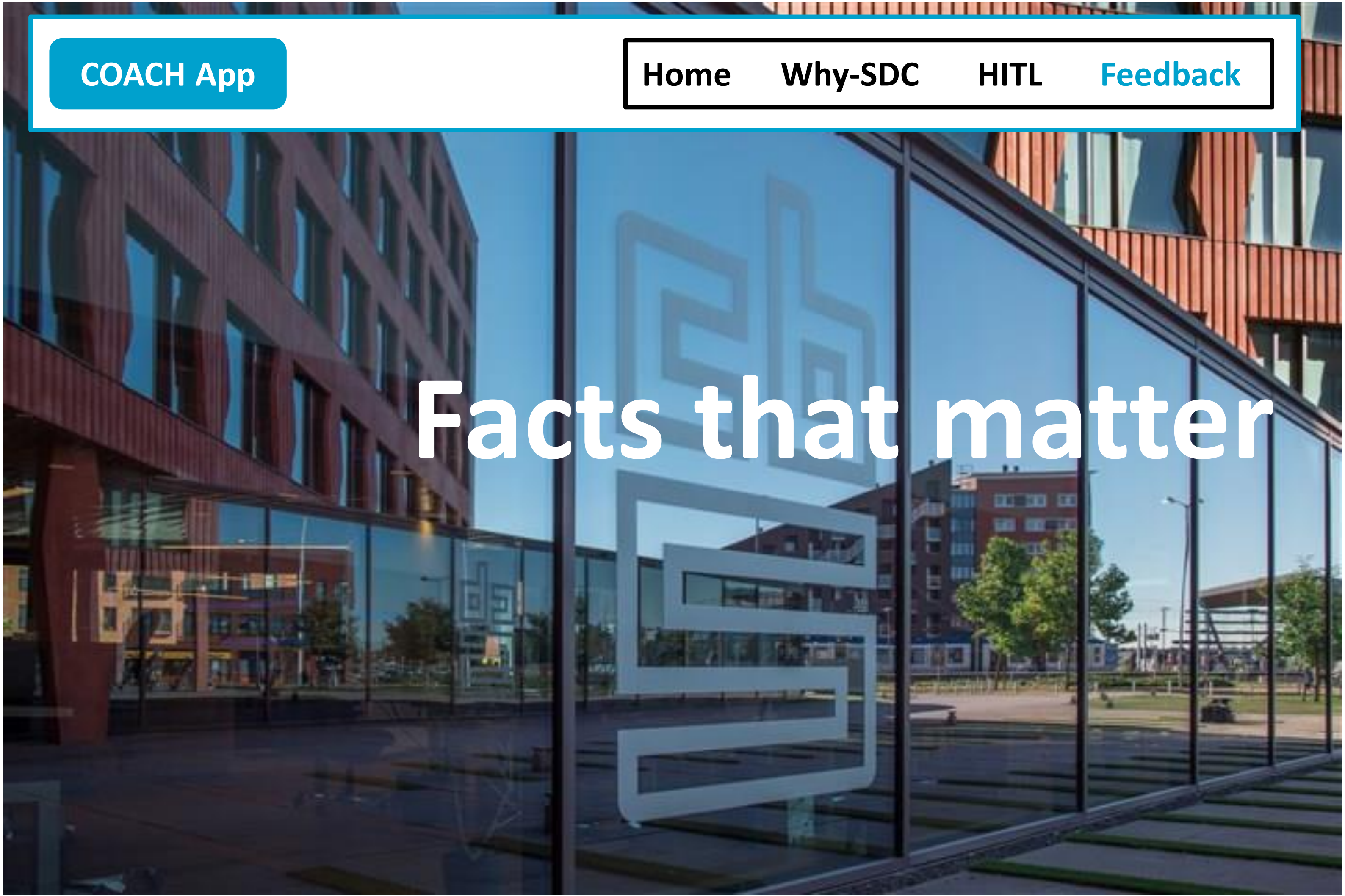
- Create **COACH**: a novel approach to semi-automate output checking
  - Human checkers are in-the-loop

- Utilize global and local SHAP values for *explainability*

Future work

- Improving COACH with AOCH (Assisted Output CHecking)

- Extending COACH
  - Explore other types of input data, e.g., features and pre-processing
  - Cross-platform: other statistical offices


COACH App
Home
Why-SDC
HITL
Feedback
Facts that matter