Improving Data Validation using Machine Learning

Source: CC0 Public Domain

Team ‘Plausi++’:

Christian Ruiz
Christine Ammann Tschopp
Elisabeth Kuhn
Laurent Inversin
Mehmet Aksözen
Stefan Rüber

Source: Packard Bell Computer, 1964

Overview

Part I: Introduction

Part II: Basic idea of Plausi++ based on prediction

Part III: Feedback mechanism based on explanation

Introduction

-Background FSO Data Innovation Strategy

-Very helpful contribution by Prof. Diego Kuonen

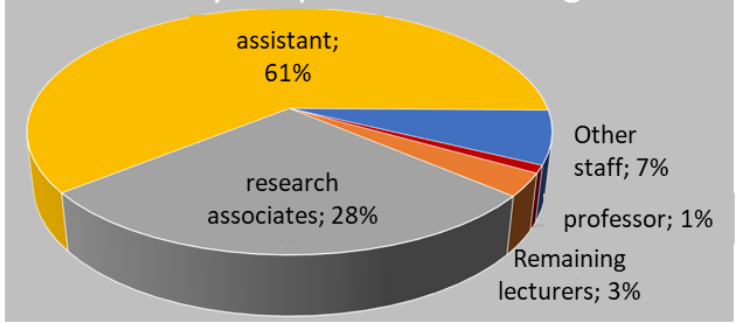
Employee: A. Meyer
Age: 21
Category: professor

Delivery to HEI

Prediction

1

Employee: A. Meyer
Age: 21
Probability for personnel categories:



Feedback from HEI

2

Employee: A. Meyer
Most probable personnel category at this age is assistant

3

4

Data validation / «Plausi»

- Manual
  (Different solutions)
- Based on rules
  (Different solutions)
- Idea: Automatic recognition
  (Enhancing other types of data validation)

Source: CC0 Public Domain

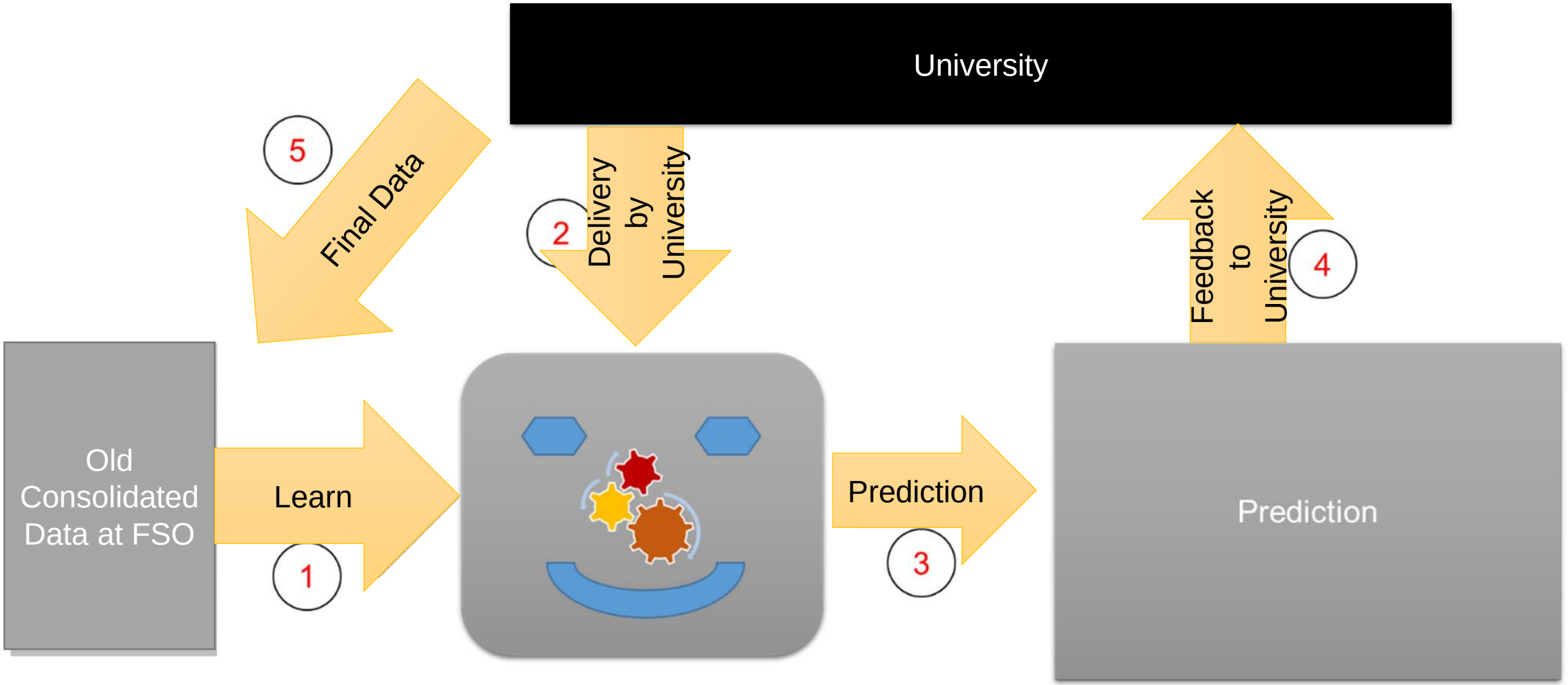
Aims

-Higher resource efficiency

-Higher velocity

-Less administrative burden for our data suppliers

-Higher data quality


University
5
Final Data
2
Delivery by University
Feedback to University
4
Old Consolidated Data at FSO
Learn
1
Prediction
3
Prediction

Part II: Basic idea of Plausi++

1) Selection of variables (from a FSO data set)
2) Prediction by ML algorithms of a ‘dependent variable’
3) Comparison between predicted and received data

If deviations: mistake or outlier («something seems odd», e.g. 21 years old prof.)

Example: Staff working at higher education institutions (HEI)

Staff category	explained by	sex, FTE, field, age, nationality, university

Dependent variable has 4 classes

P: Professors

U: Lecturers

A: Research assistants

D: Administrative employees

Sic: Partly we use 5 classes in the slides (+W)

Examples

Sex	FTE	Field	Age	Swiss	Uni	p(A|·)	p(D|·)	p(P|·)	p(U|·)	Observed
M	0.75	4.Exact	27	Yes	Yes	0.89	0.11	0.00	0.00	A ✓
F	0.80	5.Med.	26	No	Yes	0.66	0.34	0.00	0.00	A ✓
F	0.56	6.Techn.	57	No	No	0.06	0.07	0.35	0.52	P ⊘

Only hypothetic data is shown

	Delivered personal category	Probability of delivered personal category	Predicted personal category	Probability of predicted personal category
Person 1	A	3.4%	W ⊘	88.9%
Person 2	P	0.3%	U ⊘	99.4%
Person 3	P	4%	W ⊘	94.%
Person 4	U	76.6%	U ✓	76.6%
Person 5	W	6%	U ⊘	89.5%

Distribution of mistakes found between old and new data validation

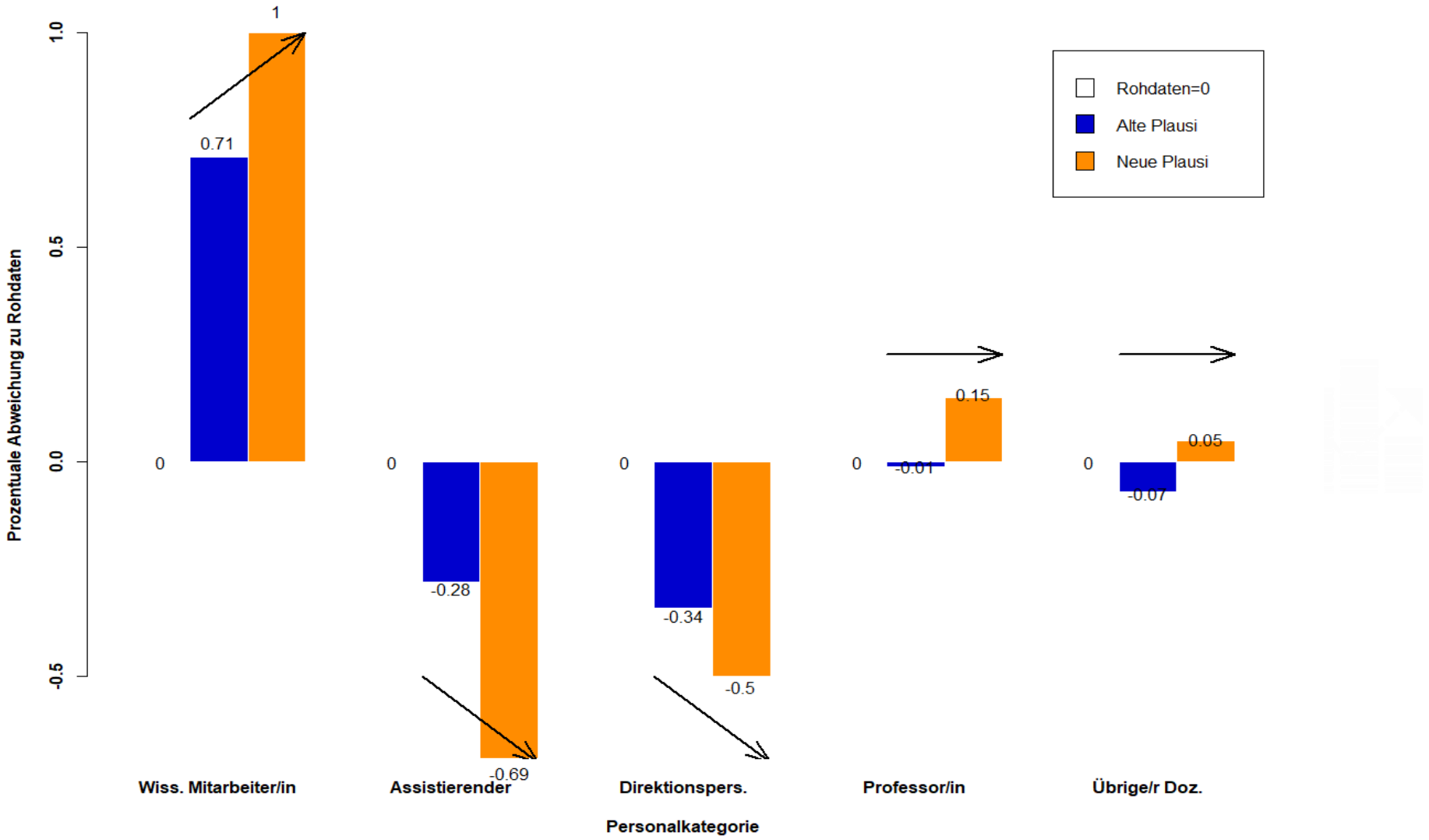
	NEW FALSE	NEW TRUE
OLD FALSE	75.07%	5.58%
OLD TRUE	17.94%	1.41%

- Both plausis complement each other!
- New plausi found mistakes in 7036 cases!
- 93% Accuracy
- 1000 variables (joining other datasets and past values)
- 2 tree algorithms: Gradient Boosting Machine (GBM) and Random Forest (RF)

Aggregated effects per University

	A	D	P	U	W
Hochschule A	-0.15	-0.42	0.15	0.07	0.37
Hochschule B	-0.08	-0.63	0.07	-0.11	0.76
Hochschule C	-3.13	0	0.09	0.71	2.34
Hochschule D	-3.84	-0.29	0.29	3.54	0.3
Hochschule E	0.43	-1.62	0.38	0.04	0.77
Hochschule F	-3.69	0.98	-0.21	-0.14	3.05
Hochschule G	0.42	-1.33	0.27	0.51	0.12
Hochschule H	0.84	-0.33	0.08	-0.02	-0.56
Hochschule I	0.19	-0.91	0.42	-0.63	0.93
Hochschule J	-1.66	0.55	0.62	-0.41	0.9
Hochschule K	0.62	-0.37	0.25	0.18	-0.68
Hochschule L	-3.32	1.83	0.54	1.8	-0.85
Hochschule M	-0.23	-1.57	0.12	0.07	1.6

Tabelle: Differenz in Prozentpunkten zwischen Rohdaten und Plausi++ Vorhersagen


Rohdaten=0
Alte Plausi
Neue Plausi
Prozentuale Abweichung zu Rohdaten
1.0
0.5
0.0
-0.5
1
0.71
0
0
-0.28
-0.69
0
-0.34
-0.5
0
-0.01
0.15
0
-0.07
0.05
Wiss. Mitarbeiter/in
Assistierender
Direktionspers.
Professor/in
Übrige/r Doz.
Personalkategorie

Part III: Feedback mechanism

Necessity of explanation and interpretability

Data suppliers are central

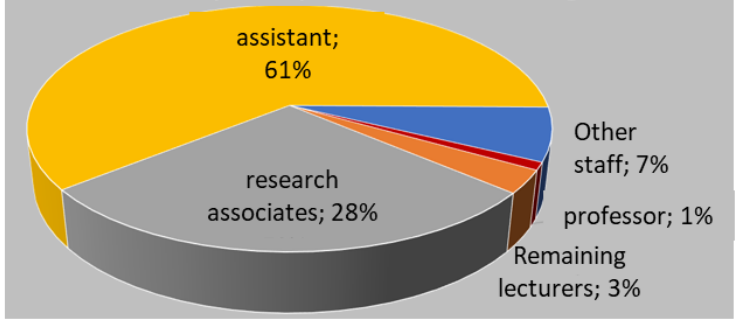
-> Higher data quality and less administrative burden

Employee: A. Meyer
Age: 21
Category: professor

Delivery to HEI

Prediction

Employee: A. Meyer
Age: 21
Probability for personnel categories:



Feedback from HEI

Employee: A. Meyer
Most probable personnel category at this age is assistant

The person is probably not prof. but assistant

But why?

Hall, Gill and Meng, June 26 2018, O'Reilly

"

So why isn't everyone just trying interpretable machine learning?

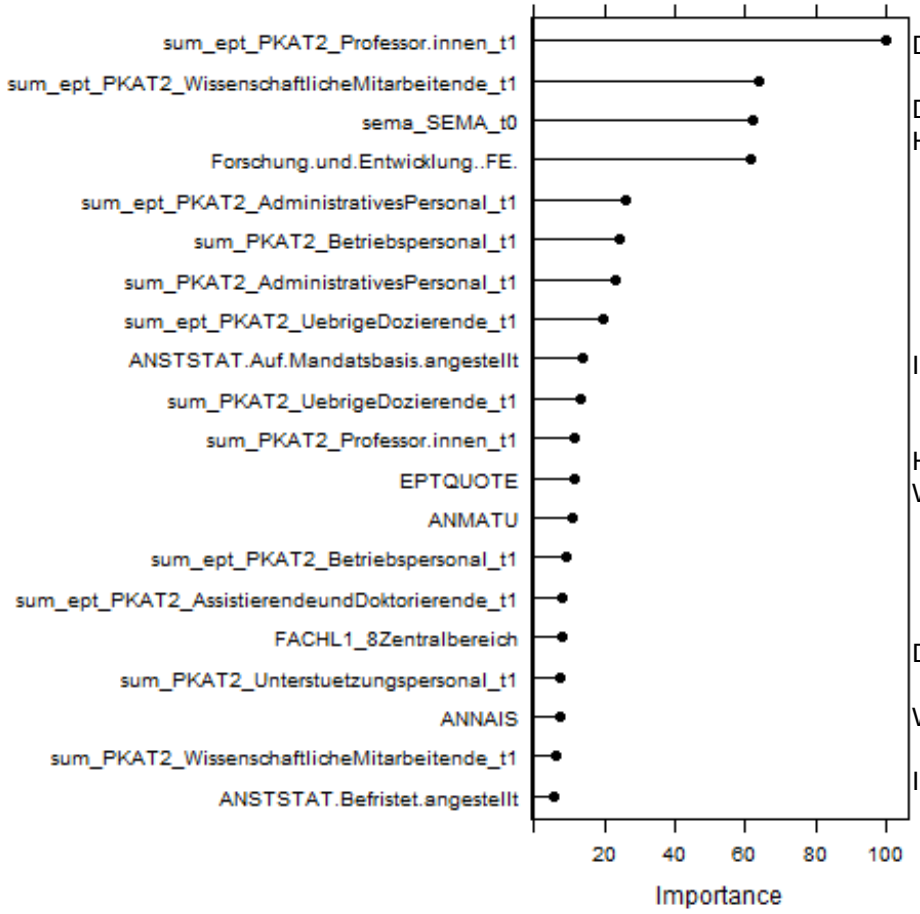
Simple answer:

it's fundamentally difficult,

and in some ways, a very new field of research.

"

Global explanation: Variable Importance (GBM-Model)



Did the person work as a prof. Last year?

Does/did the person study in that year?
How many % of the work are spent in R&D?

Is the person employed on a mandate/contract?

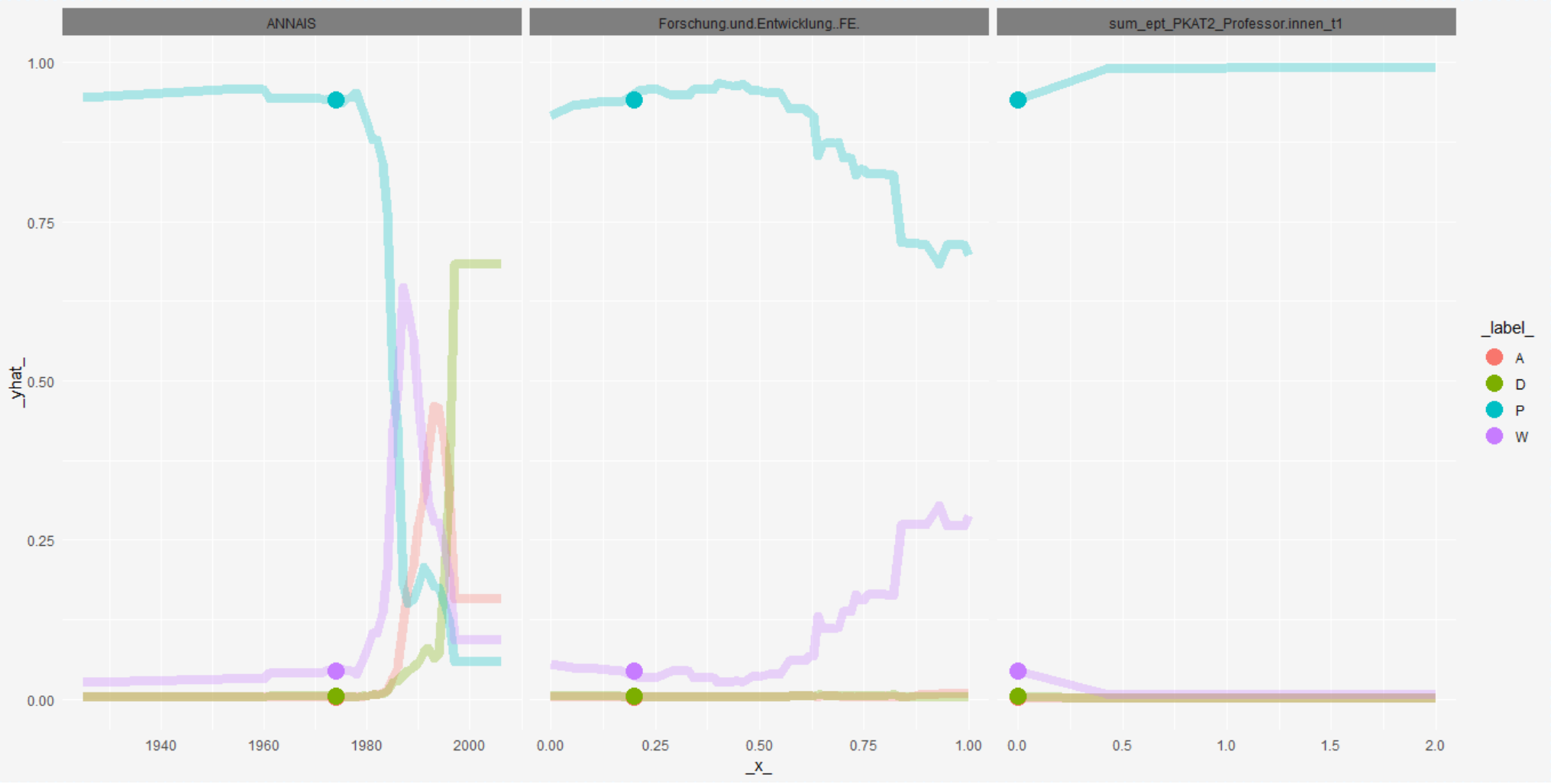
How many FTE did this person work overall?
When did the person finish A-levels?

Does the person work in the central administration?

When was the person born?

Is the person temporarily employed?

DALEX Partial Residual Plot: Outcome = P, Case = 101426

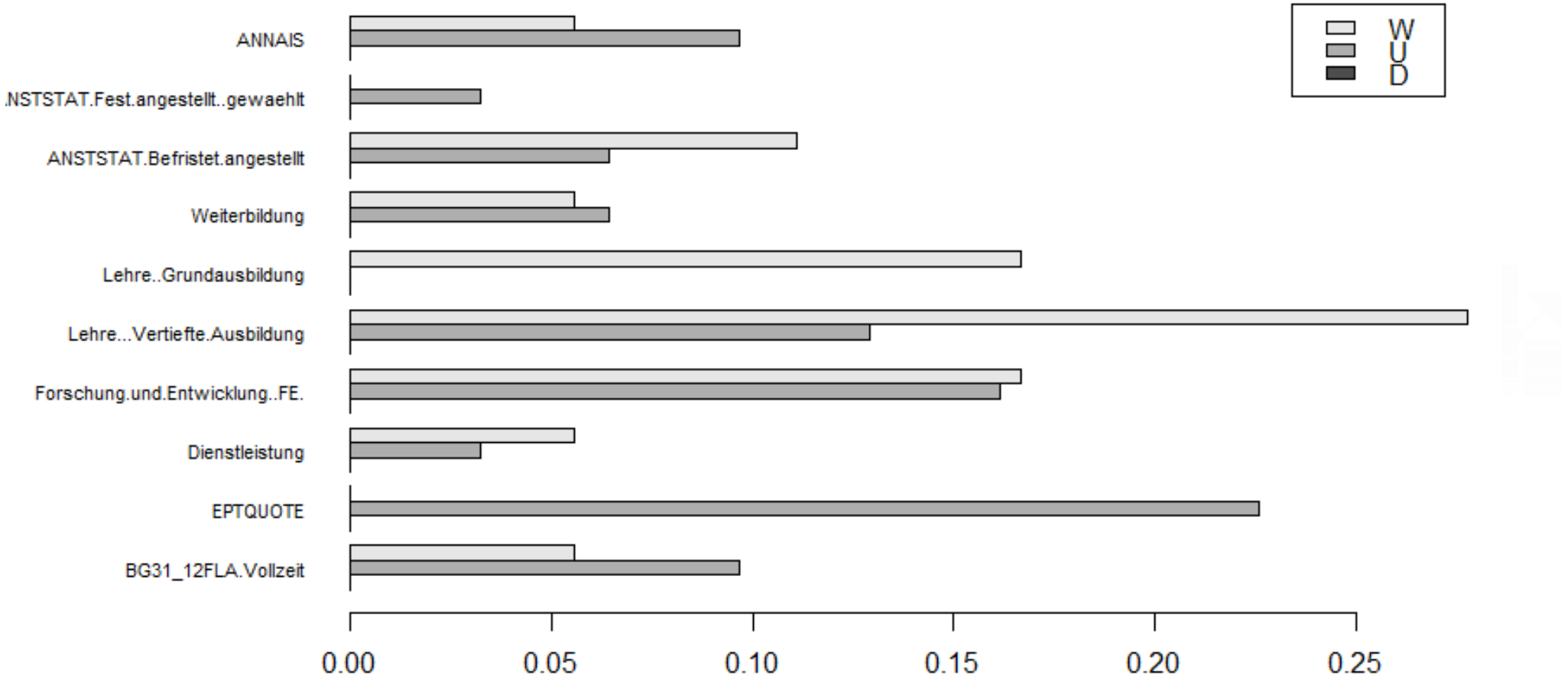


Sic: Without U for Better overview

DALEX Partial Residual Plot: Outcome = P, Case = 100269

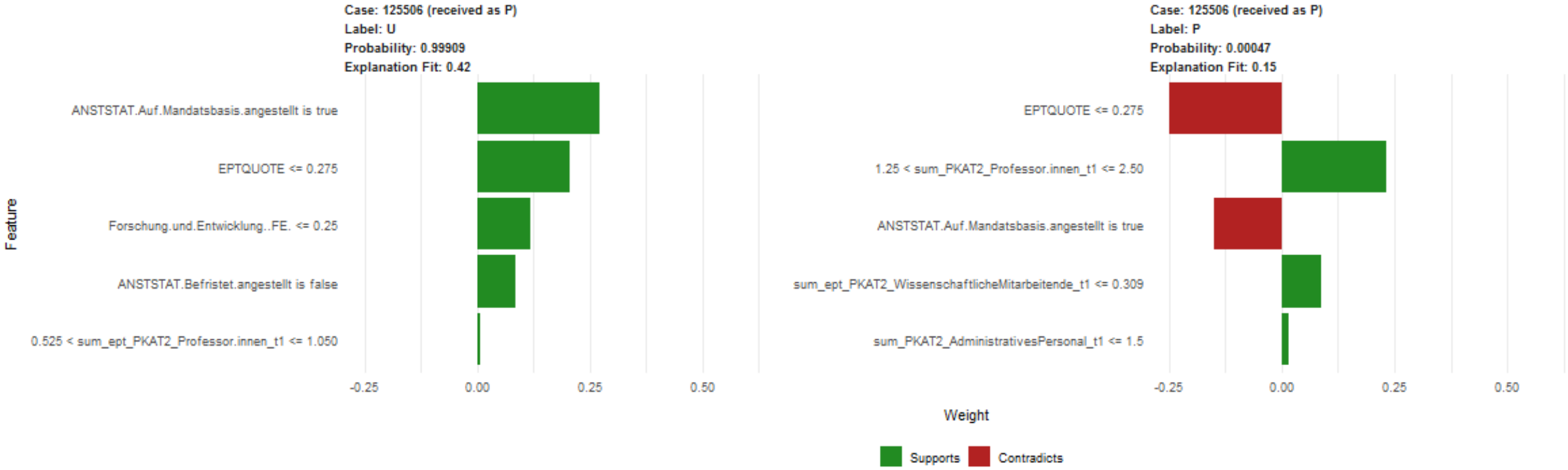
Sic: Without U for Better overview

Relative Likelihood of certain candidates – delivered = P



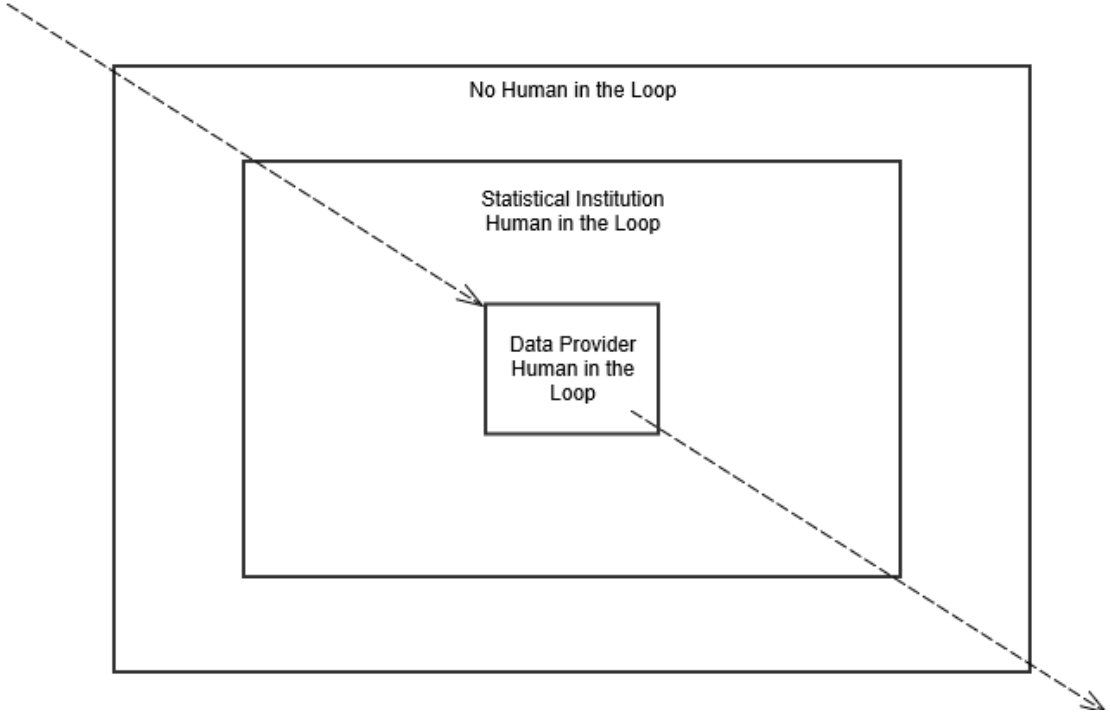
	Feedback an Datenlieferanten
Person 1	Das Alter (in Kombination mit den anderen Variablen) spricht dafür, dass es womöglich falsch ist oder sich um einen wissenschaftlichen Mitarbeitenden statt einem Assistierenden handeln könnte.
Person 2	Die niedrige VZÄ-Quote (in Kombination mit den anderen Variablen) deutet darauf hin, dass es sich hier um einen übrigen Dozierenden handelt oder die VZÄ-Quote zu tief ist.
Person 3	Die niedrige Anteil Lehre (in Kombination mit den anderen Variablen) macht es wahrscheinlicher, dass es sich hier um einen wissenschaftlichen Mitarbeitenden handelt oder der Anteil Lehre falsch ist.
Person 4	OK ✓
Person 5	Ein Anteil Forschung von 0% (in Kombination mit den anderen Variablen) deutet nicht auf einen wissenschaftlichen Mitarbeitenden hin. Entweder ist der Anteil der Forschung oder die Personalkategorie falsch.

LIME: Local Interpretable Model-Agnostic Explanations



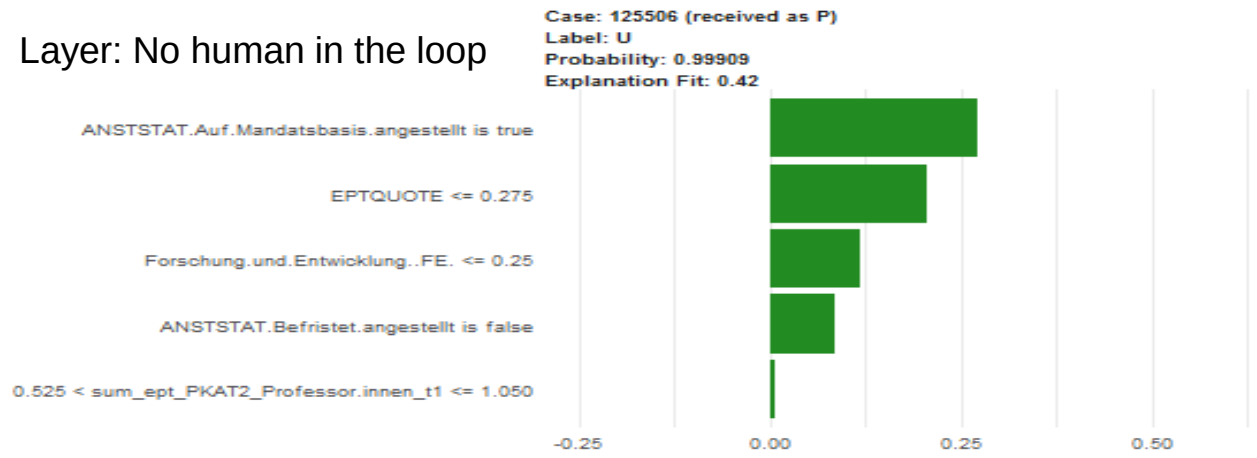
Onion model

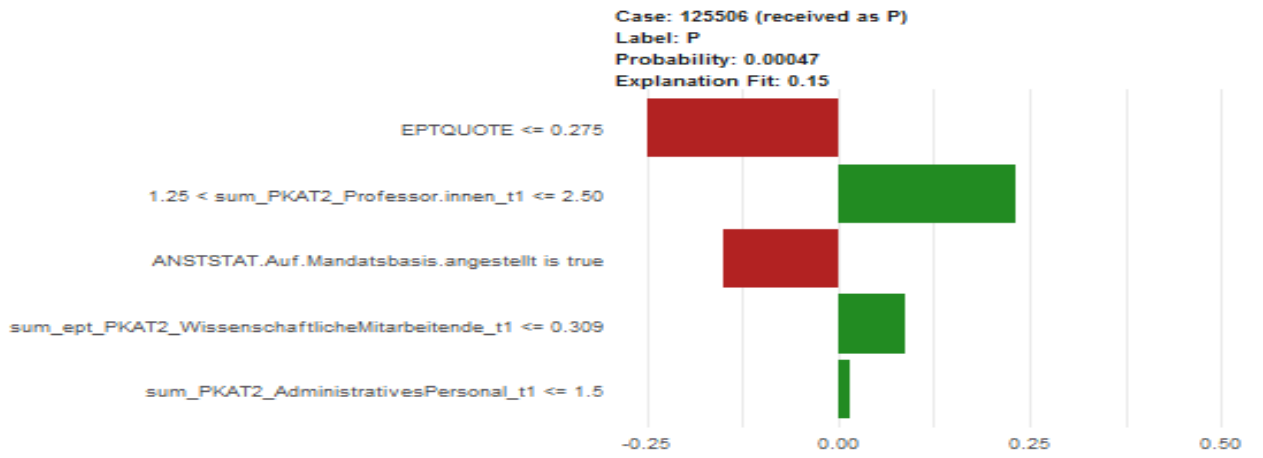
The deeper we get, the more interpretable the result is.
The variables in the innermost layer correspond to the delivered variables.



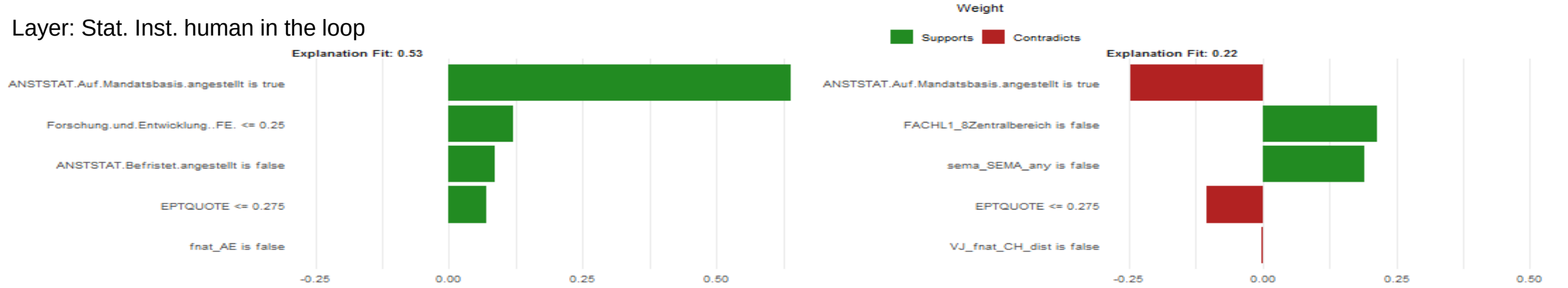
The larger the distance from the innermost layer, the more complex and less interpretable the result. However, the prediction becomes better.

Layer: No human in the loop

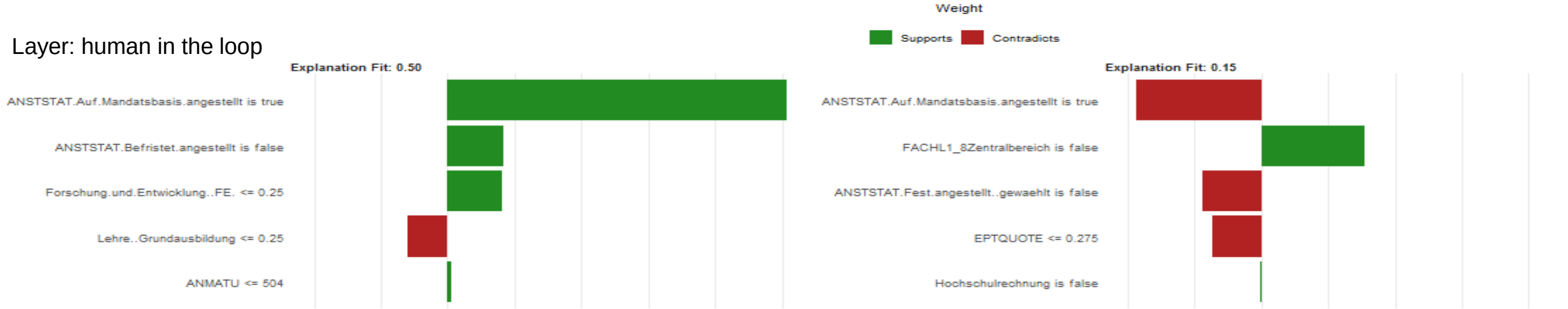




Layer: Stat. Inst. human in the loop



Layer: human in the loop



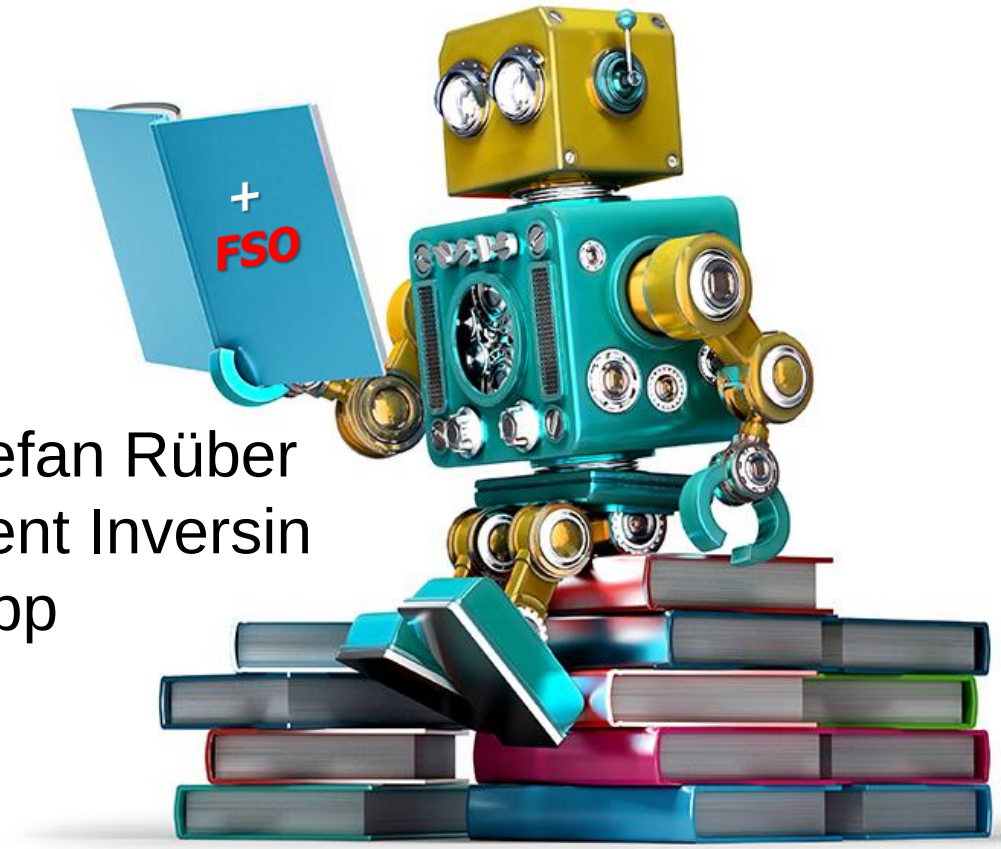
Conclusion / Next Steps

- Prediction works well. Accuracy currently over 93%
- Not anymore 6 but around 1000 variables
- Explanation part pioneering work and challenging!

- Pilot project until June 2019
- Putting to production / Mini pilot project
- Shapley values etc. not tried (due to time constraints)
- Team Change

Thank you very much for your attention!

Thanks to «Team-DALEX»: Mehmet Aksözen and Stefan Rüber
Thanks to «Team-LIME»: Elisabeth Kuhn and Laurent Inversin
Thanks to «Team-IT» : Christine Ammann Tschopp
Thanks to our advisor: Prof. Dr. Diego Kuonen



Source: CC0 Public Domain with modifications