



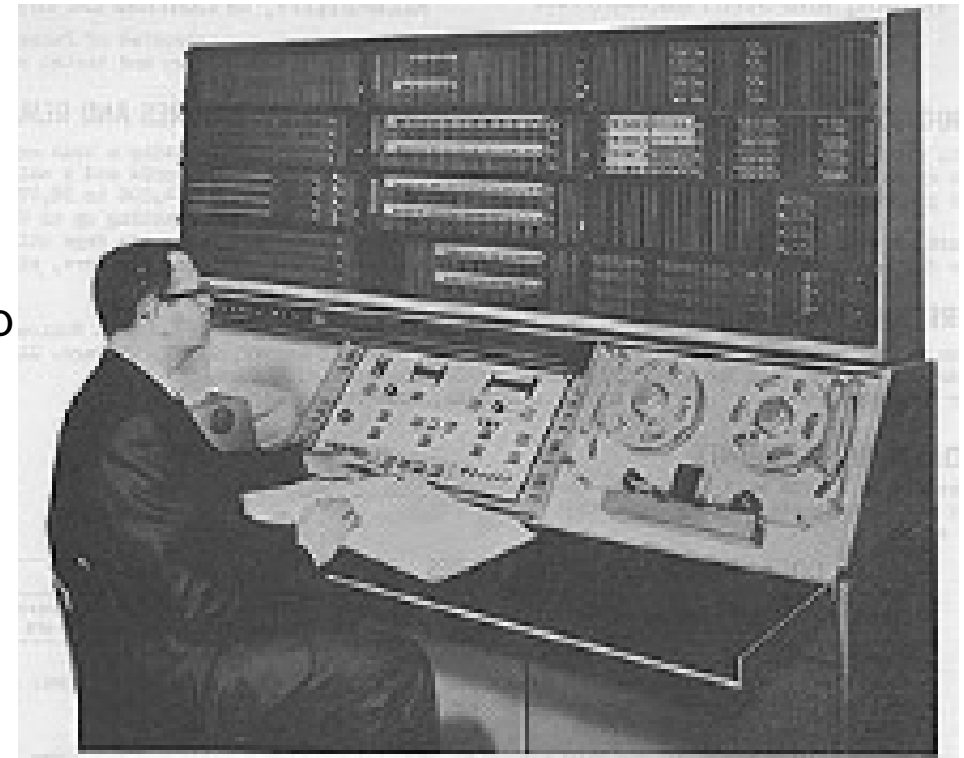
# Improving Data Validation using Machine Learning



Team 'Plausi++':

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

Source: CC0 Public Domain



Source: Packard Bell Computer, 1964



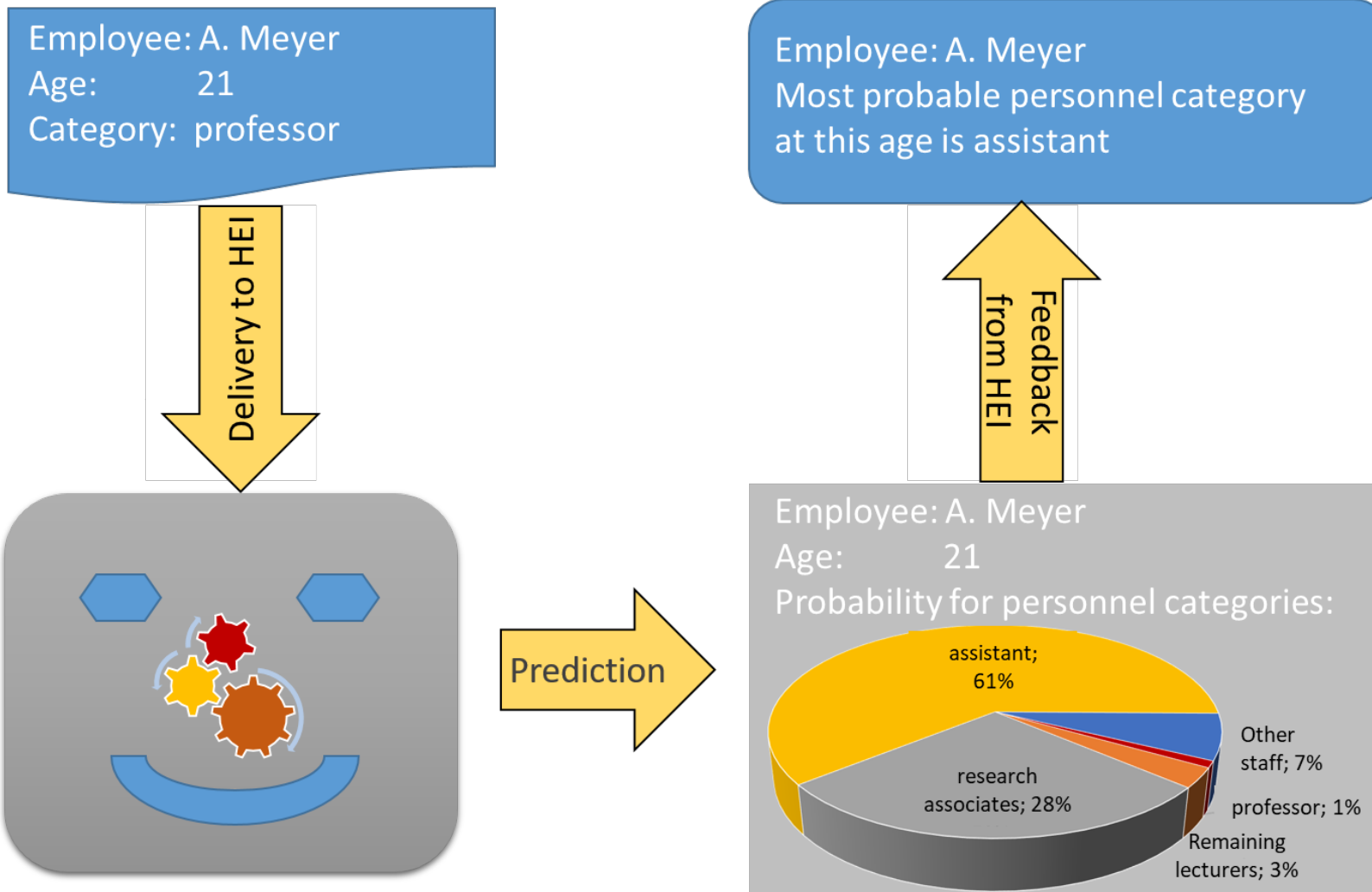
## Delimitation

The pilot project runs since April 2018 and is part of the «data innovation strategy» of the Swiss Federal Statistical Office (FSO). The pilot project will end in June 2019. The following presentation does not present the end results of the pilot project.



# Overview

- Part I: Introduction
- Part II: Basic idea of Plausi++
- Part III: Feedback mechanism





# 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

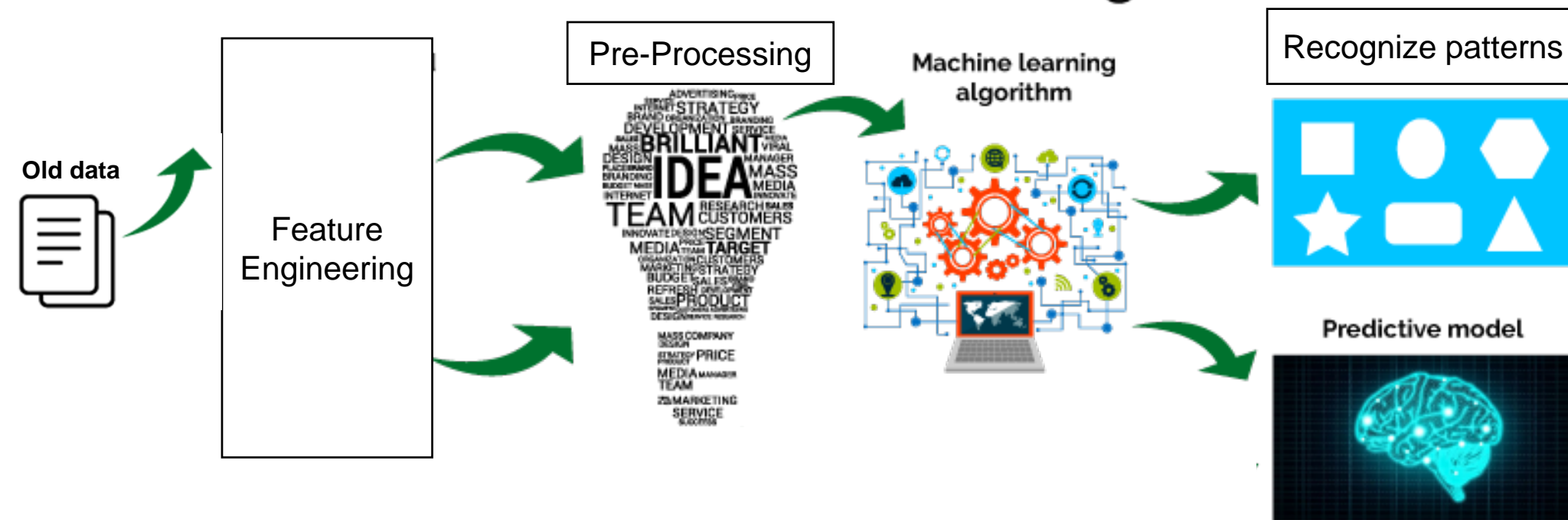
- Higher resource efficiency
- Higher velocity
- Less administrative burden for our data suppliers
- Higher data quality



## PHASE: TRAINING

# Machine Learning

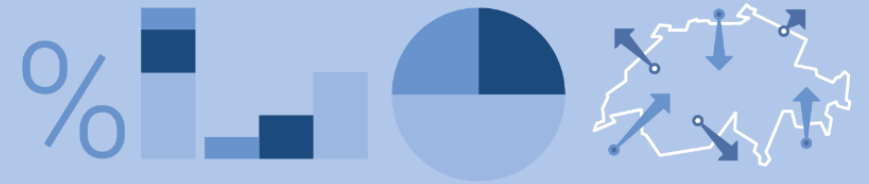
## Supervised learning



## Learn?

- Feed a large amount of data
- Recognize patterns
- Prediction of values
- Prediction is not equivalent to explanation or interpretation

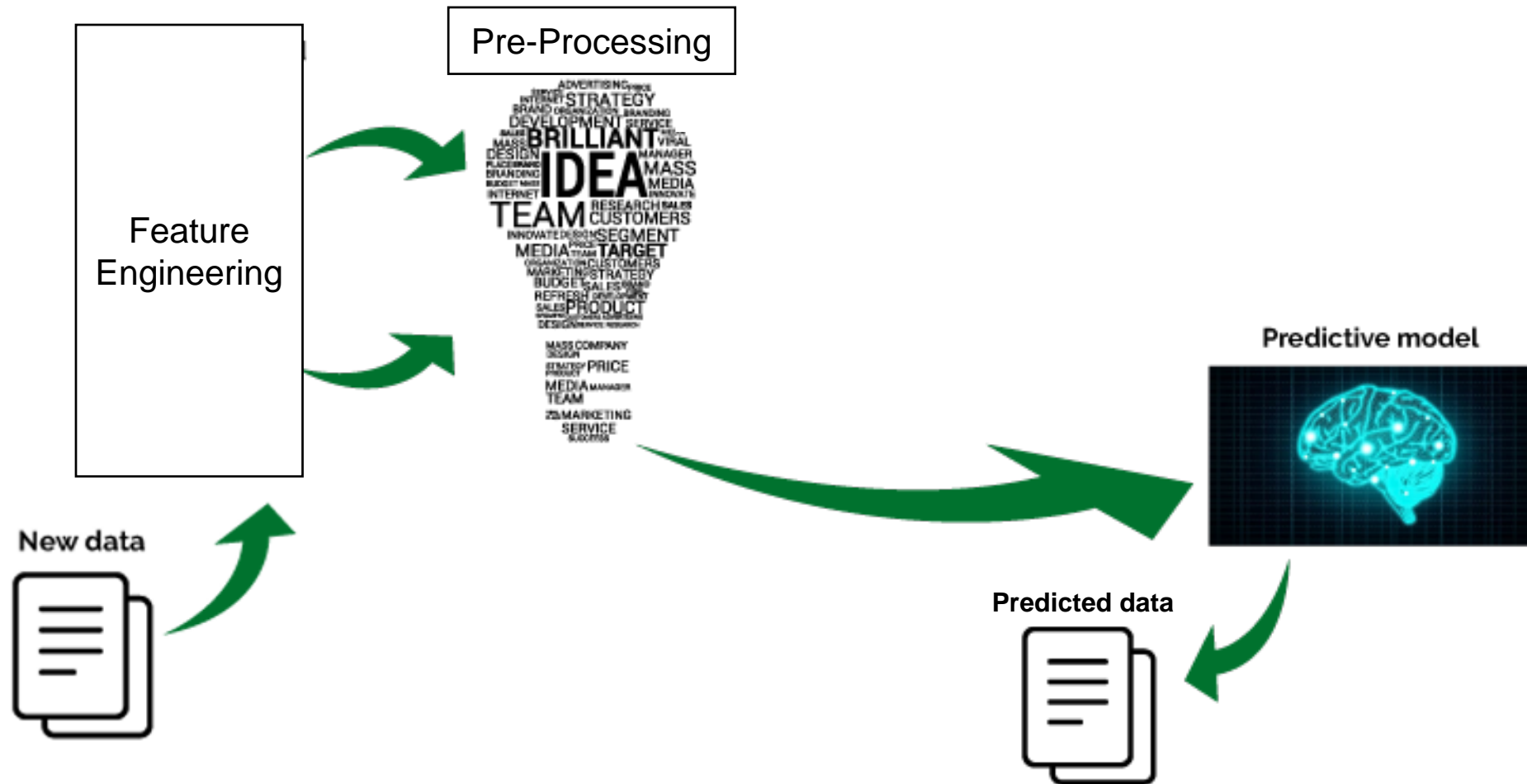
(Source: CC0 Public Domain with modifications)



## PHASE: TESTING

# Machine Learning

## Supervised learning



(Source: CC0 Public Domain with modifications)





## More human than machine

Human: Comprehension of the data

Human: Feature Engineering and pre-processing

Human: Choice of **appropriate** algorithms

Machine: Calculation

Machine: Preparation for final user

Human: Calibration and decisions

Human: Integration into the production environment



## 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



Sex	FTE	Field	Age	Swiss	Uni	Observed	Predicted	p(obs ·)	p(pred ·)
F	1.000	5. Medicine	18	TRUE	TRUE	A	D	0.002	0.996
M	1.000	5. Medicine	20	TRUE	TRUE	A	D	0.002	0.996
F	1.000	Other	18	TRUE	TRUE	A	D	0.007	0.989
F	1.000	Other	19	TRUE	TRUE	A	D	0.008	0.988
M	1.000	5. Medicine	21	TRUE	TRUE	A	D	0.009	0.988
F	0.200	8. Central	34	TRUE	FALSE	A	D	0.009	0.987
F	1.000	8. Central	22	FALSE	TRUE	A	D	0.01	0.987
F	0.900	8. Central	61	TRUE	TRUE	P	D	0.007	0.981
M	0.600	6. Technical	26	FALSE	FALSE	D	A	0.011	0.985
F	0.400	8. Central	31	FALSE	FALSE	A	D	0.011	0.984
F	1.000	5. Medicine	30	FALSE	TRUE	A	D	0.012	0.982
F	1.000	2. Economy	56	FALSE	TRUE	A	P	0.005	0.974
M	0.058	2. Economy	72	TRUE	TRUE	A	U	0.004	0.972
F	0.600	4. Exact	25	FALSE	FALSE	D	A	0.014	0.982
F	1.000	2. Economy	55	FALSE	TRUE	A	P	0.005	0.971
F	0.028	Other	55	TRUE	TRUE	D	U	0.008	0.973
M	0.700	4. Exact	25	FALSE	FALSE	U	A	0.002	0.967
F	0.021	Other	56	TRUE	TRUE	A	U	0.004	0.968
M	1.000	2. Economy	49	FALSE	TRUE	A	P	0.006	0.970
M	0.800	8. Central	36	FALSE	FALSE	A	D	0.016	0.980

Only hypothetic data is shown



## Results

- ✓ Currently over 94% of accuracy
- ✓ Instead of 6 variables we have about 1000 variables
- ✓ 5 classes instead of 4

## Required

- ✓ A sufficiently high amount of cases
- ✓ Relationships between the variables
- ✓ Meaningful relationships between the variables
- ✓ Sufficiently 'good' data quality in train set



## Part III: Feedback mechanism

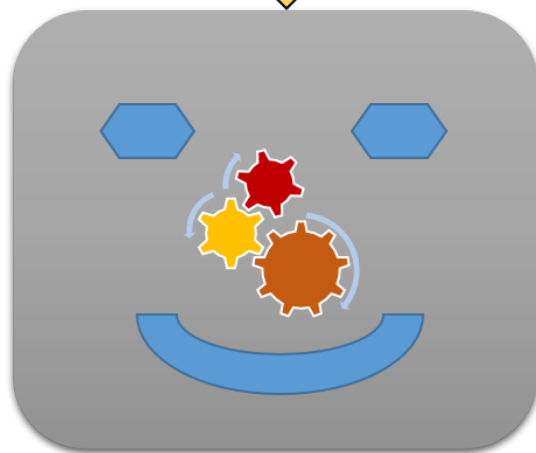
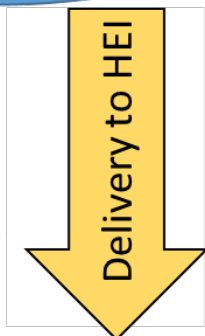
### Necessity of explanation and interpretability

Data suppliers are central

-> Higher data quality and less administrative burden



Employee: A. Meyer  
Age: 21  
Category: professor

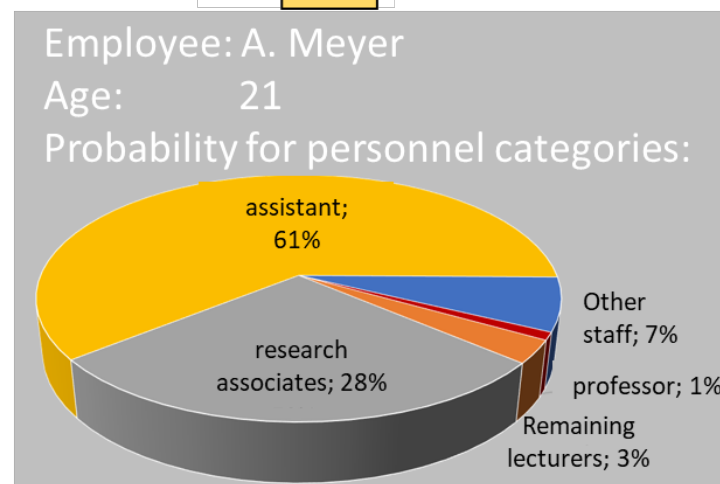


Employee: A. Meyer  
Most probable personnel category  
at this age is assistant



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:

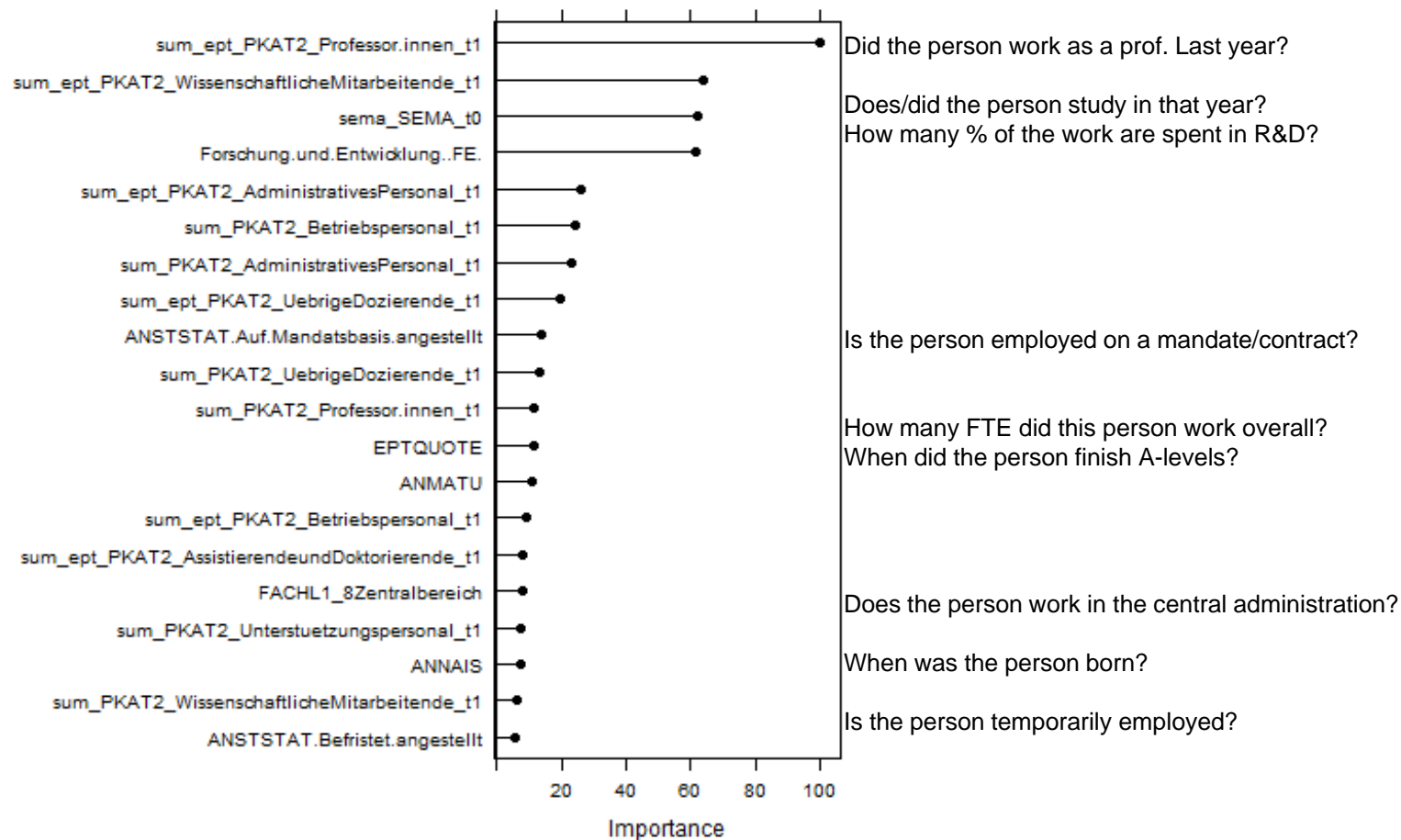
it's fundamentally difficult,

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

“

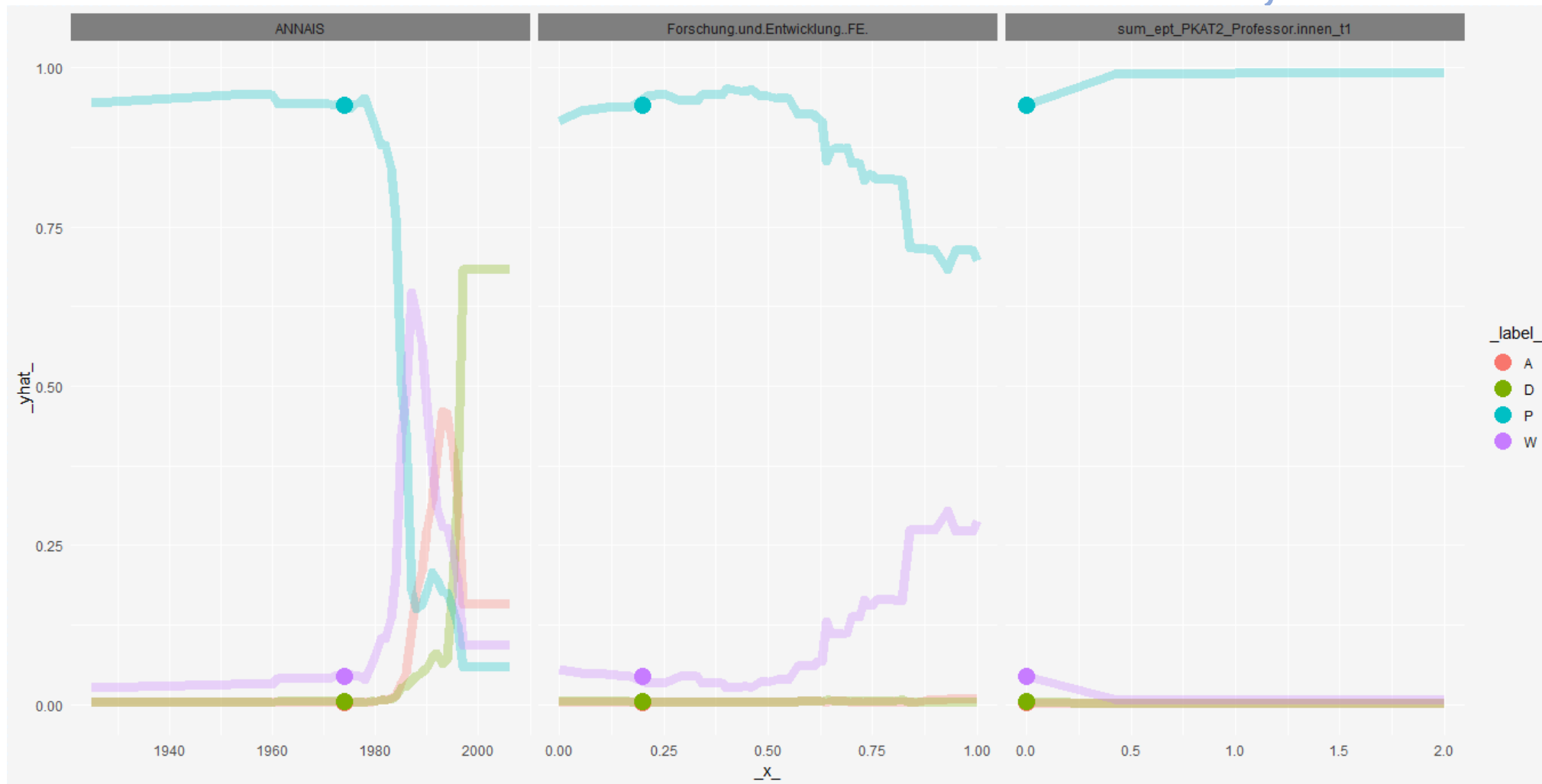


# Global explanation: Variable Importance (GBM-Model)





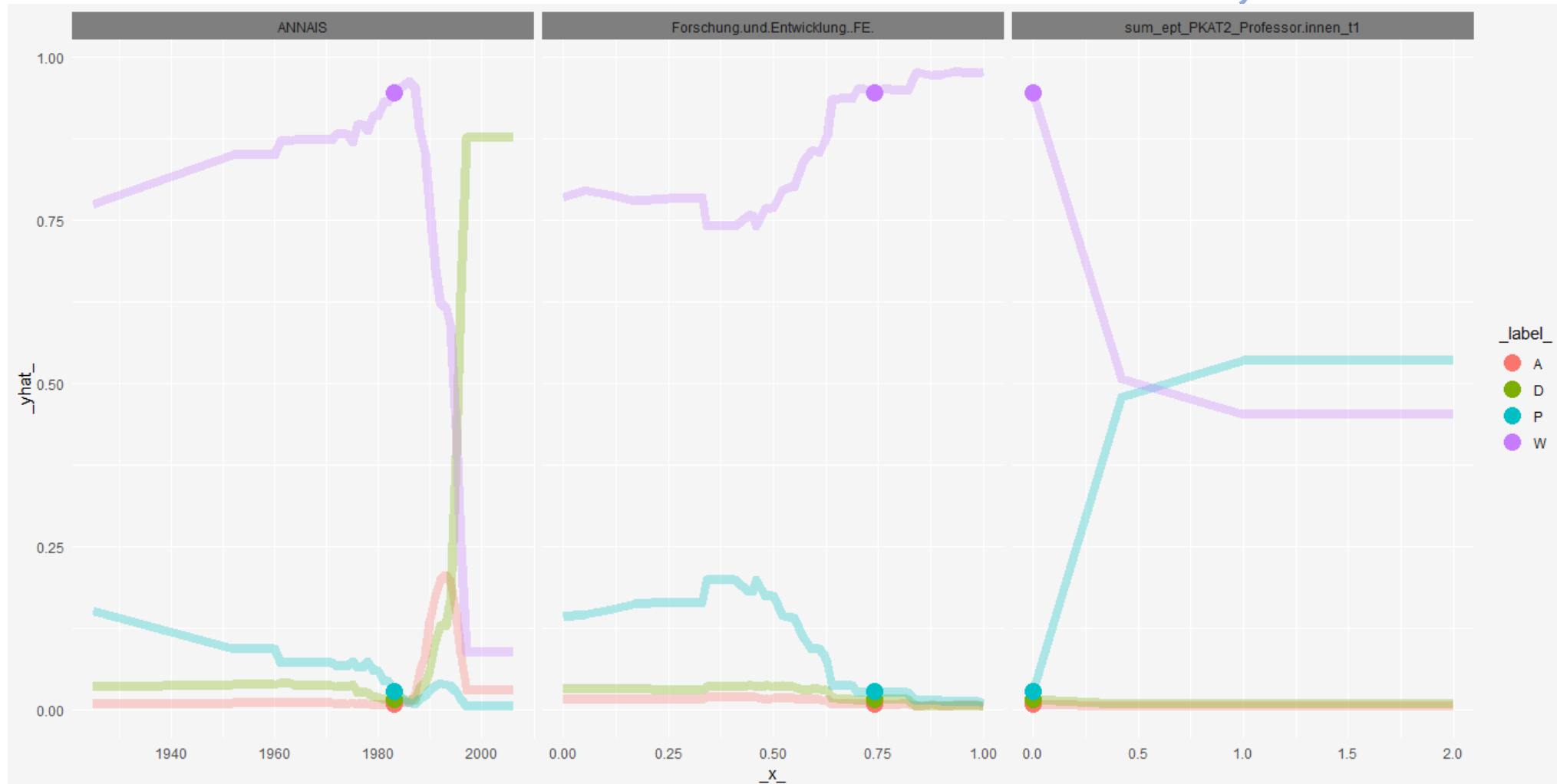
# DALEX Partial Residual Plot: Outcome = P, Case = 101426



Sic: Without U for  
Better overview



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

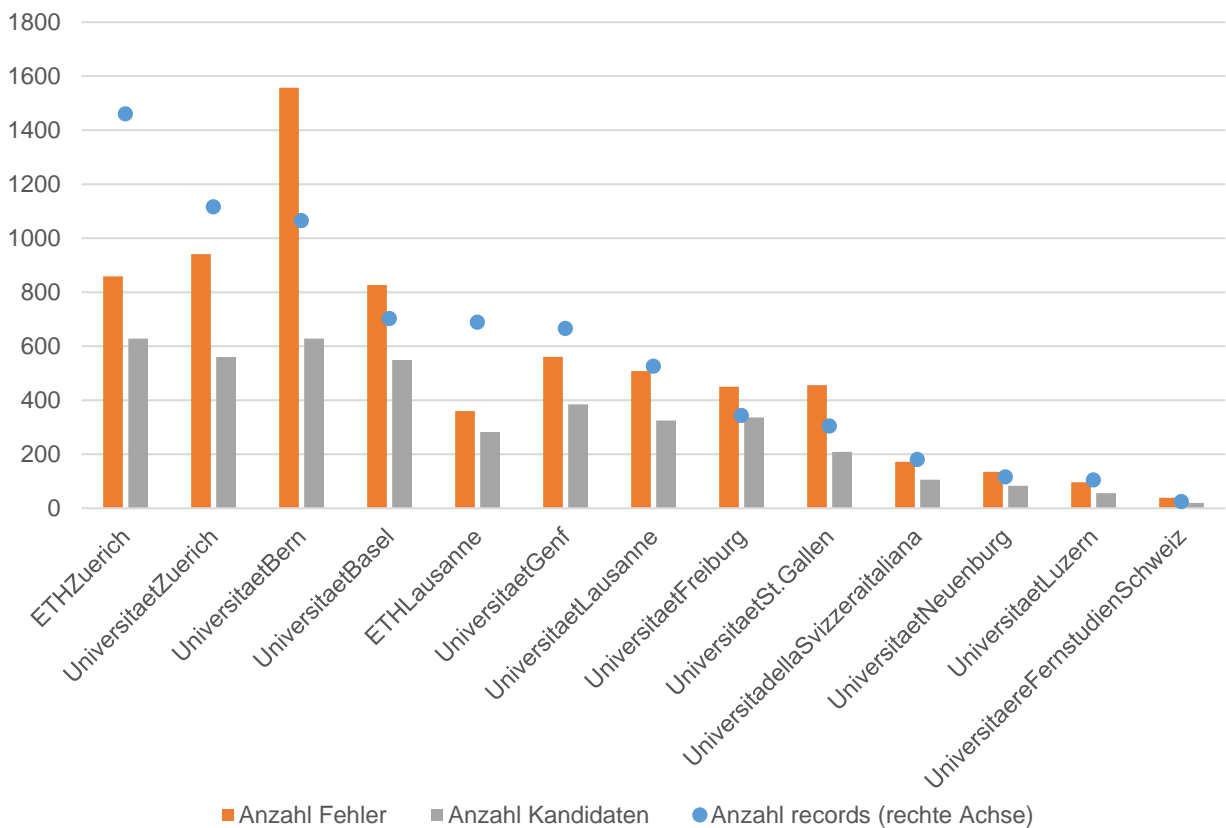


Sic: Without U for Better overview

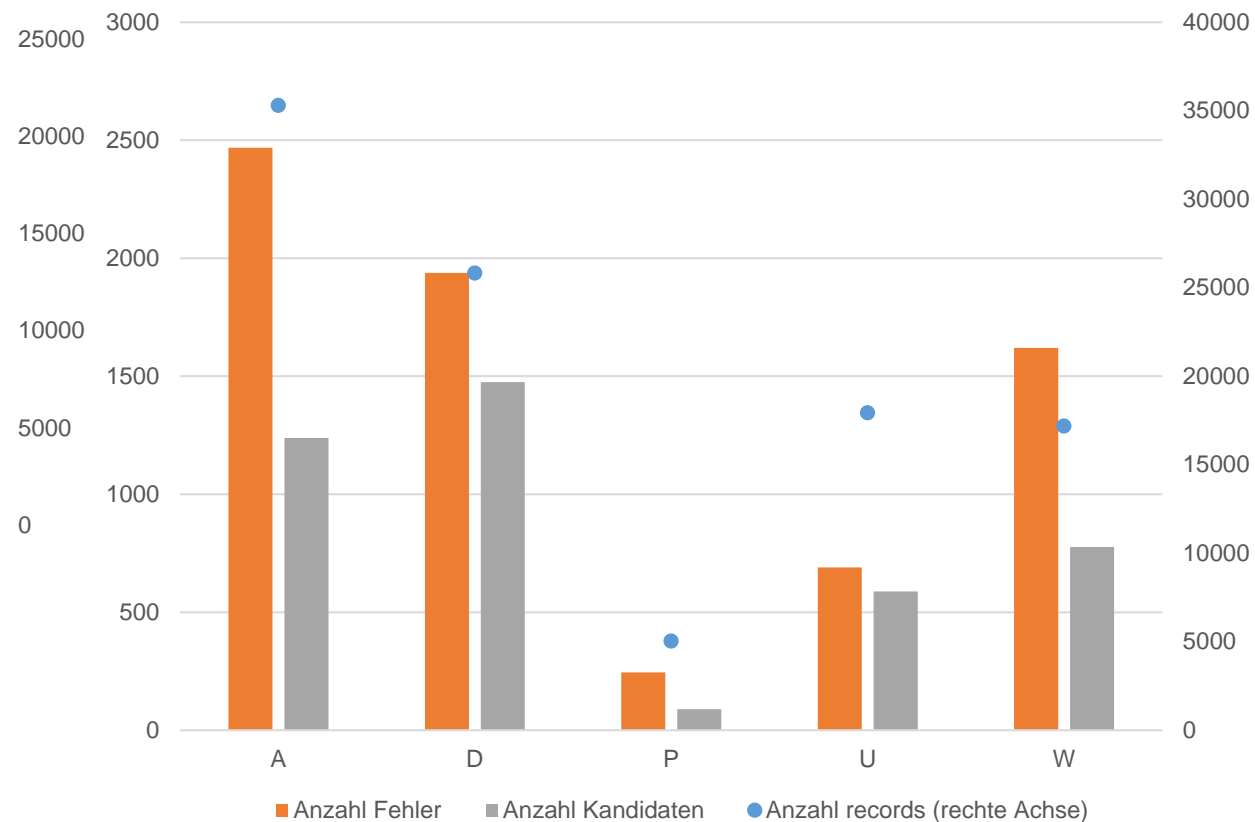


# Latest

Mögliche Erklärungen durch ceteris paribus-Regel



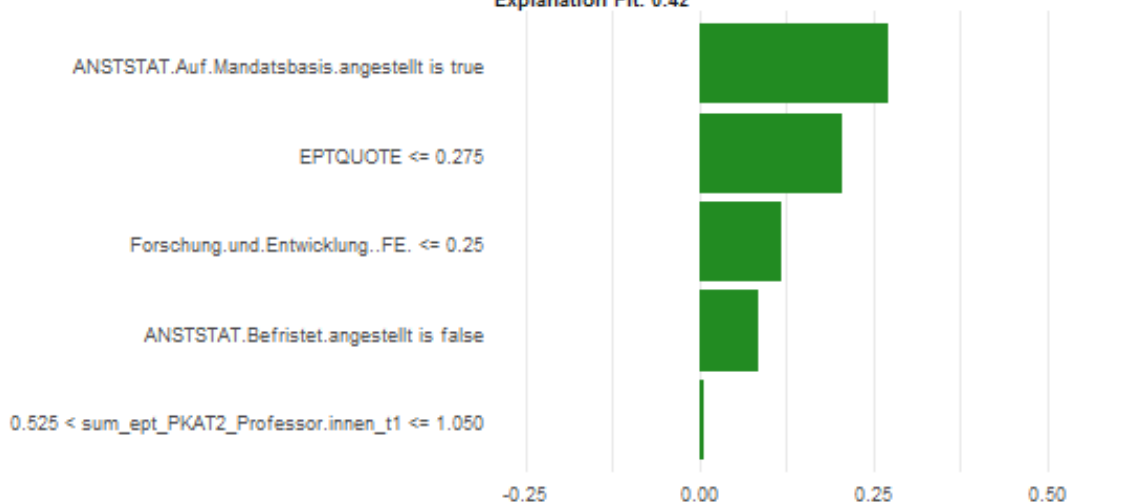
Mögliche Erklärungen durch ceteris paribus-Regel



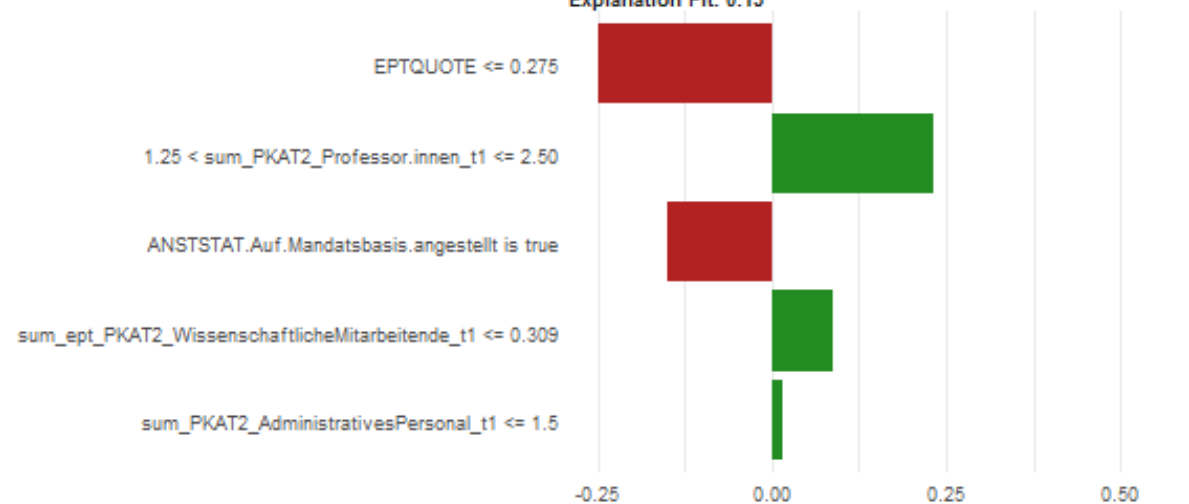


# LIME: Local Interpretable Model-Agnostic Explanations

Case: 125506 (received as P)  
Label: U  
Probability: 0.99909  
Explanation Fit: 0.42



Case: 125506 (received as P)  
Label: P  
Probability: 0.00047  
Explanation Fit: 0.15



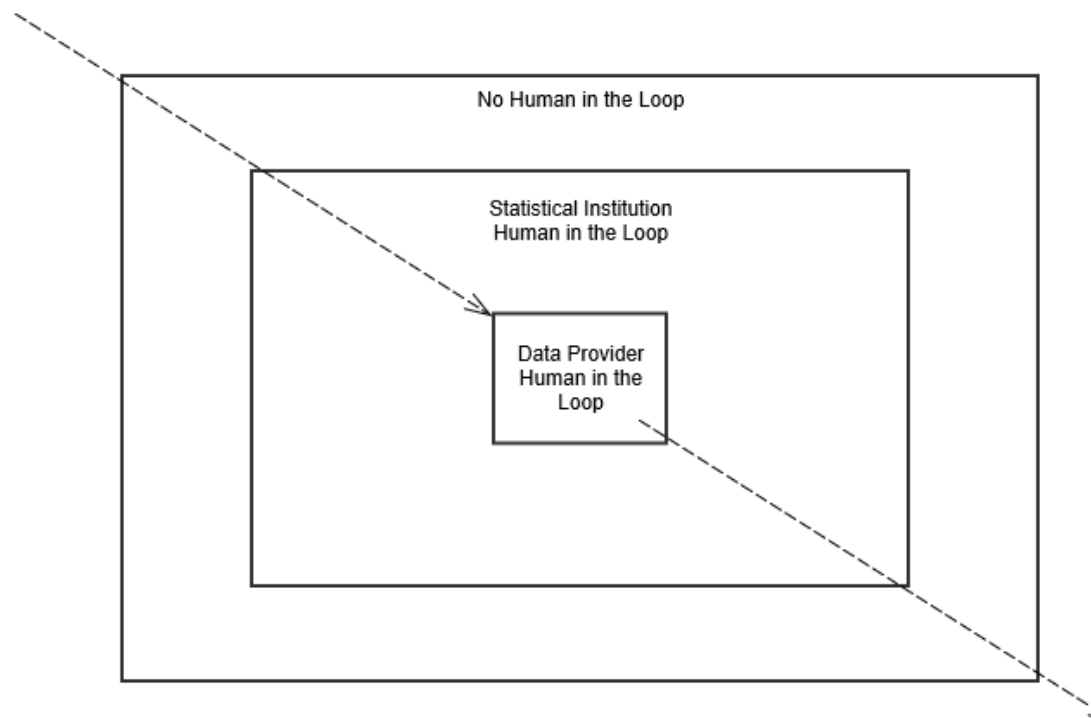
Weight

Supports Contradicts



# Onion model

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



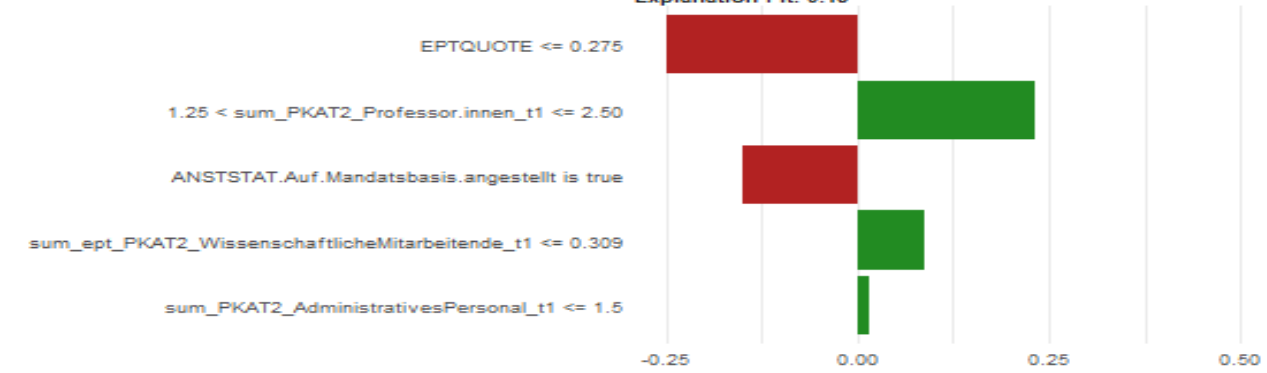
The larger the distance from the innermost layer, the more complex and less interpretable the result. However, the prediction becomes better.

### Layer: No human in the loop

Case: 125506 (received as P)  
Label: U  
Probability: 0.99909  
Explanation Fit: 0.42

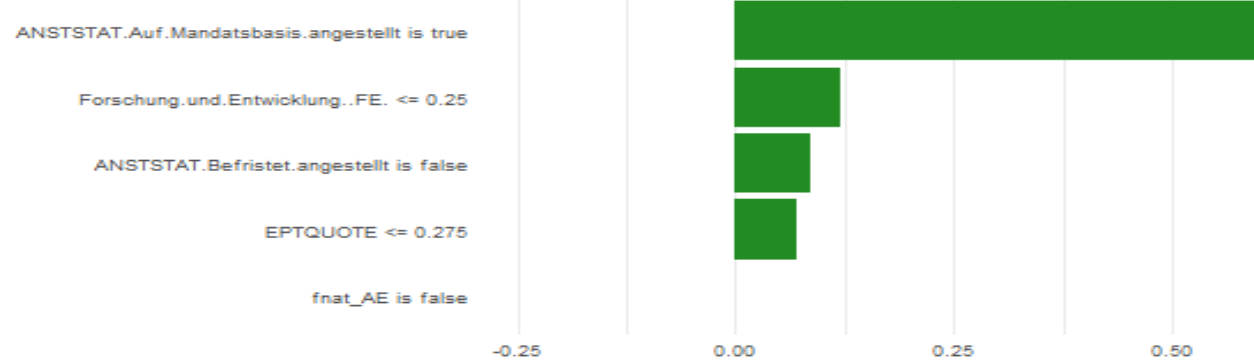


Case: 125506 (received as P)  
Label: P  
Probability: 0.00047  
Explanation Fit: 0.15

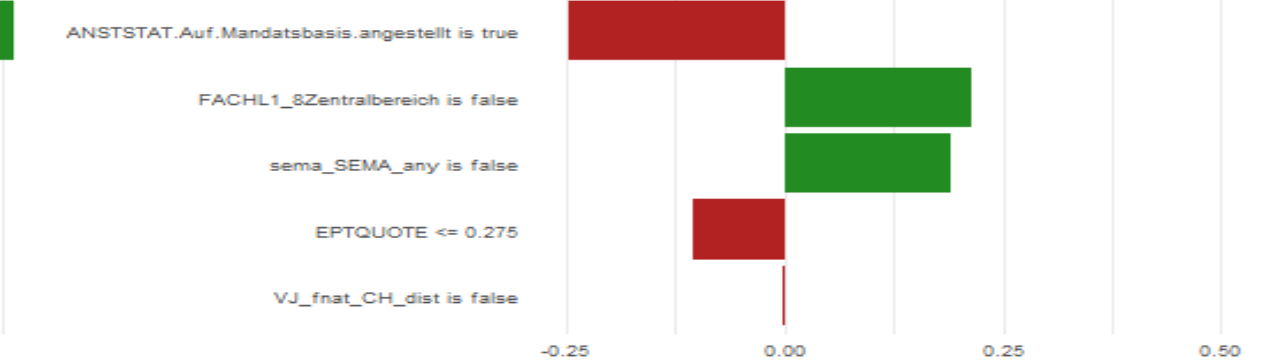


### Layer: Stat. Inst. human in the loop

Explanation Fit: 0.53



Explanation Fit: 0.22

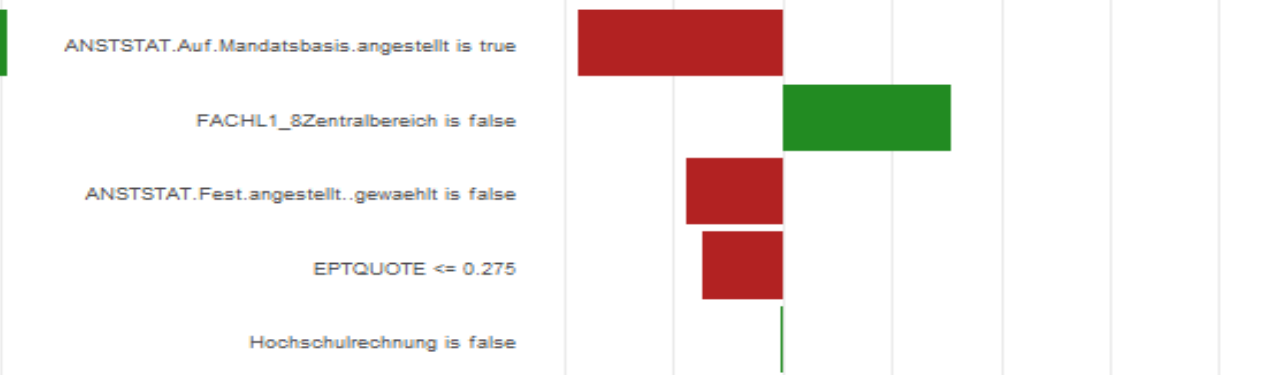


### Layer: human in the loop

Explanation Fit: 0.50



Explanation Fit: 0.15







## Conclusion

- Prediction works well. Accuracy currently over 94%
- Not anymore 6 but around 1000 variables
- Explanation part ongoing and pioneering work!
- Pilot project until June 2019
- Difficult challenges linked to feedback-mechanism
  
- Feedback highly appreciated



# 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



## Distribution of mistakes found

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

->New plausi found mistakes in 7036 cases!



# Aggregated effects of the plausis

