EFFIZIENTERE DATENGETRIEBENE PROJEKTE DURCH AUTOMATISIERTES MASCHINELLES LERNEN

Dr. Janek Thomas

janek.thomas@scs.fraunhofer.de

Machine Learning

One goal of ML is to replace human decision processes, with their preferences, bias and unreliability, with an optimal automatic data driven process.*

*Optimal w.r.t. to a human defined performance measure

The Case for Automating Machine Learning

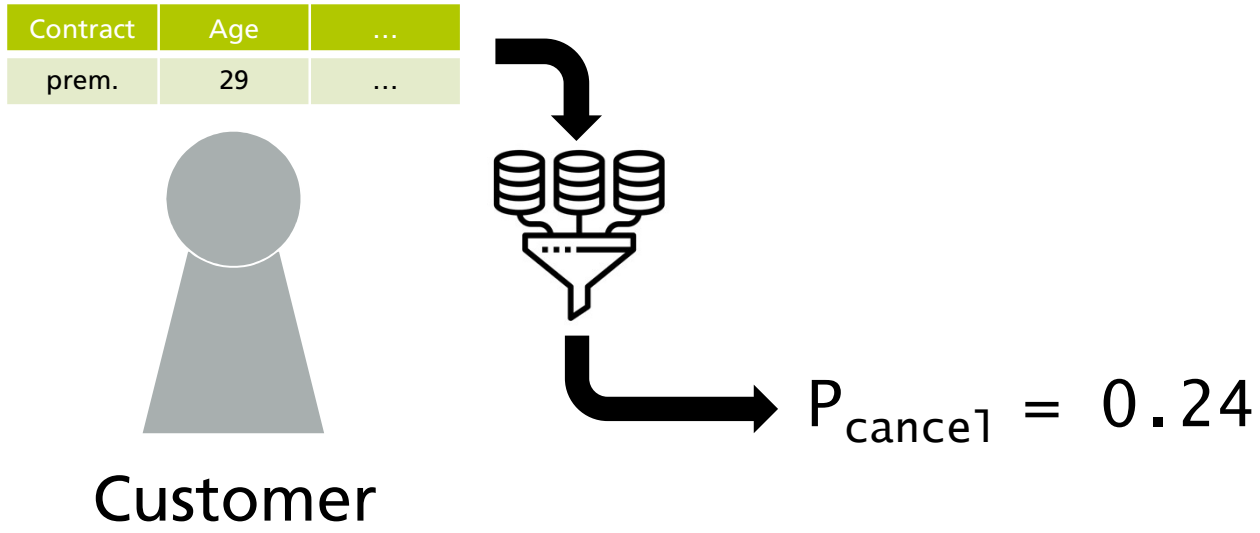
- In many businesses predicting **customer churn** is an essential metric.

> ***Customer attrition***, *also known as* ***customer churn***, ***customer turnover***, *or* ***customer defection***, *is the loss of clients or customers. …*
>
> *Detect soon which customers are about to abandon … and put into practice personalized retention plans.*

https://en.wikipedia.org/wiki/Customer_attrition

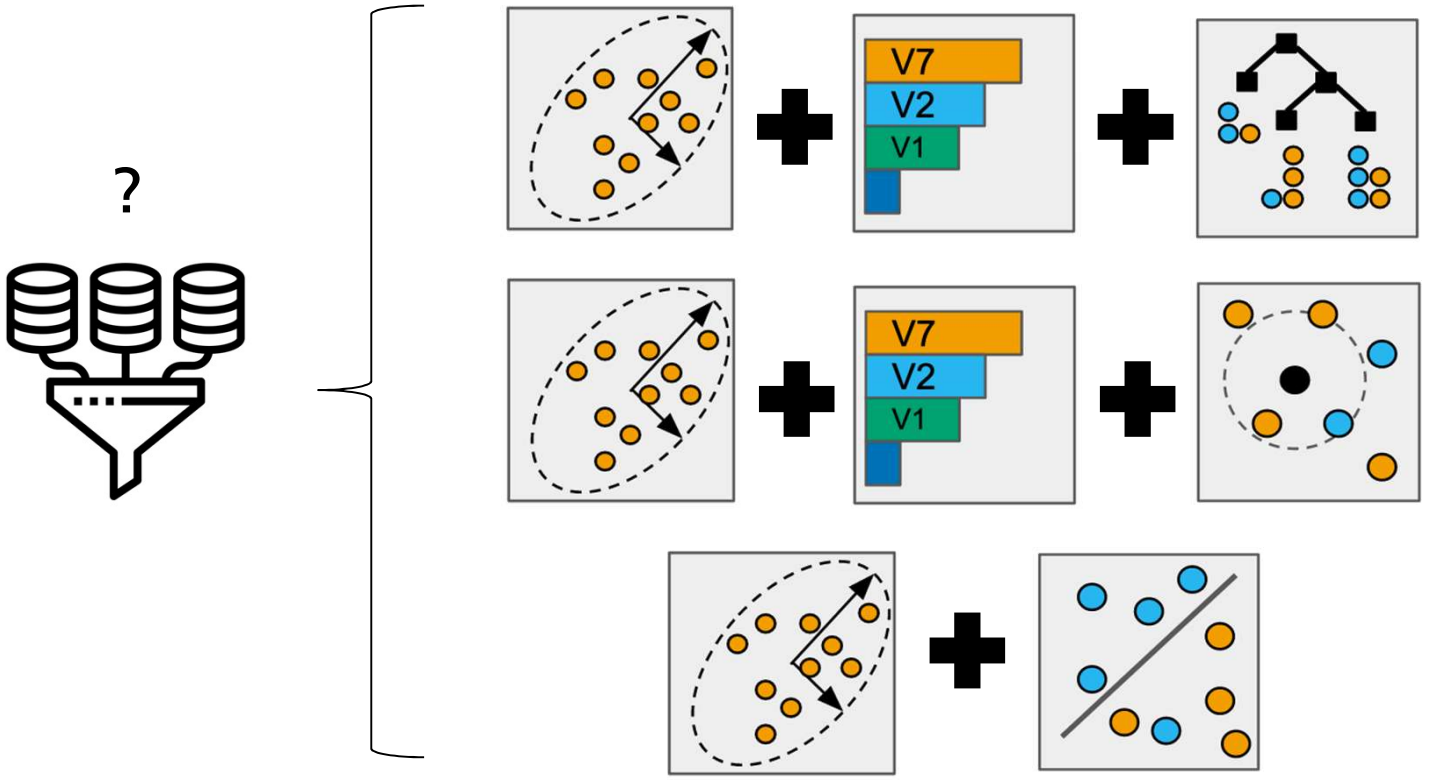
The Case for Automating Machine Learning

- In many businesses predicting **customer churn** is an essential metric.



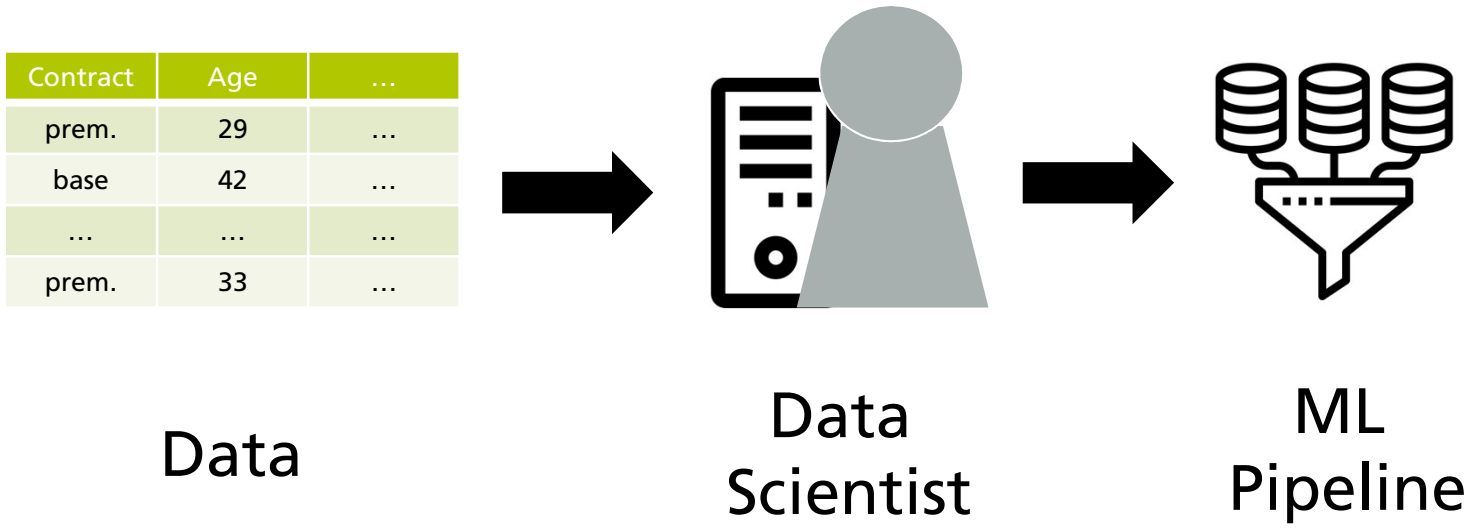
The Case for Automating Machine Learning

- Machine learning pipelines are composed of many interchangable steps



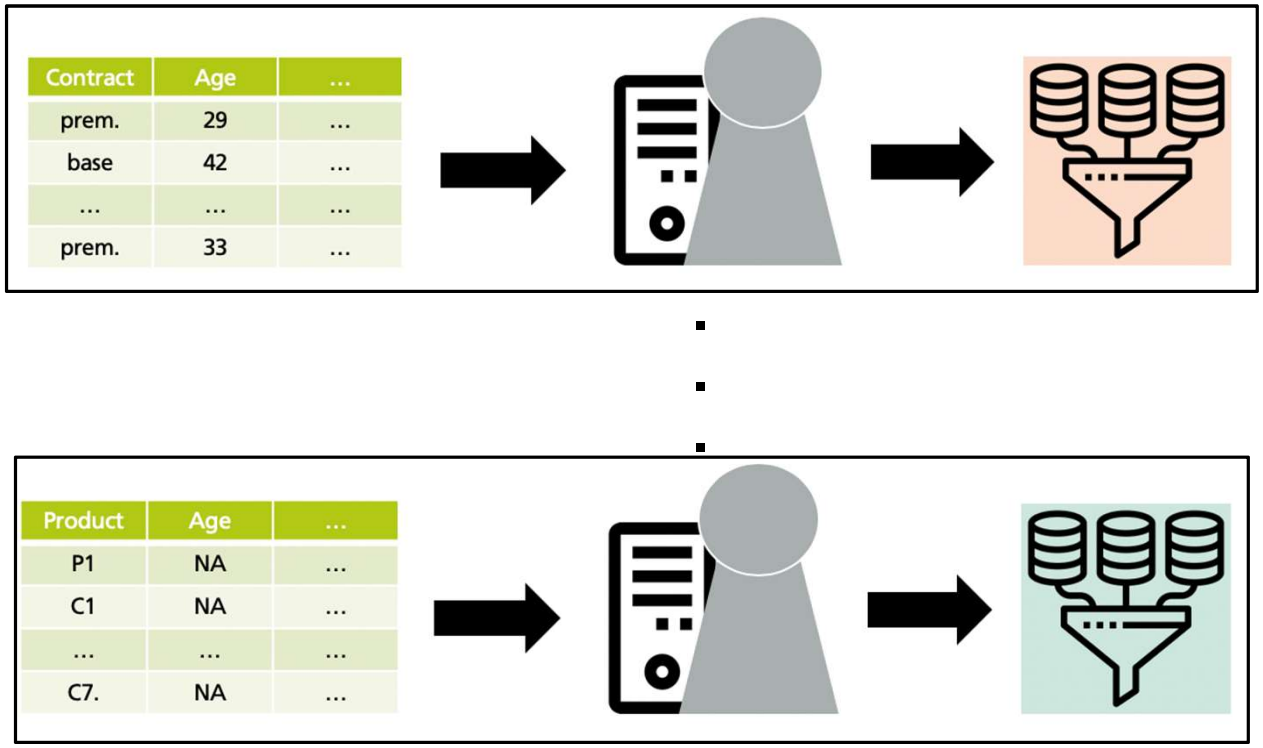
The Case for Automating Machine Learning

- Data Scientists select machine learning pipelines by trial and error, hyperparmeter tuning, preferences and experience based on the structure of the problem.



The Case for Automating Machine Learning

- This process is done for many different products, customer segments, time periods, …



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
ADA LOVELACE CENTER
Fraunhofer IIS

The Case for Automating Machine Learning

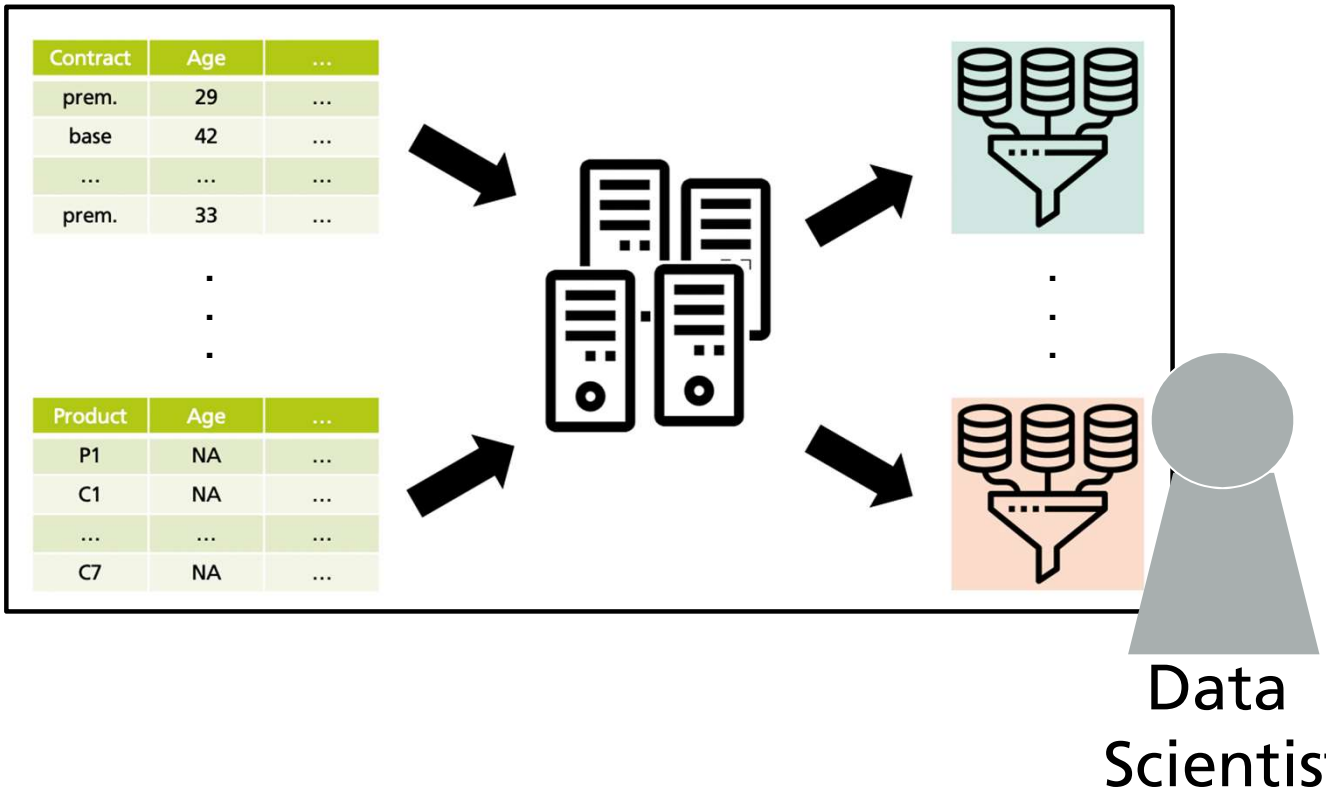
- These problems can differ in the
  - amount data
  - available features
  - data quality
  - *„difficulty“*
  - *…*

➢ The optimal ML pipeline will be different in many cases

➢ ML is still guided by human preferences

Automatic Machine Learning (AutoML)

- Can algorithms be trained to automatically build end-to- end machine learning systems?



Automatic Machine Learning (AutoML)

Use machine learning to do better machine learning!

Turn

```
Solution = data + manual exploration + computation
```

into

Solution = data + computation (x100)

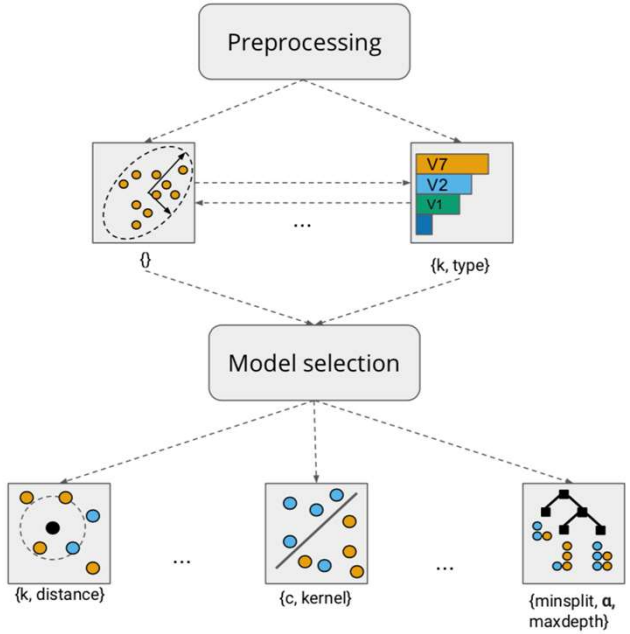
Automatic Machine Learning (AutoML)

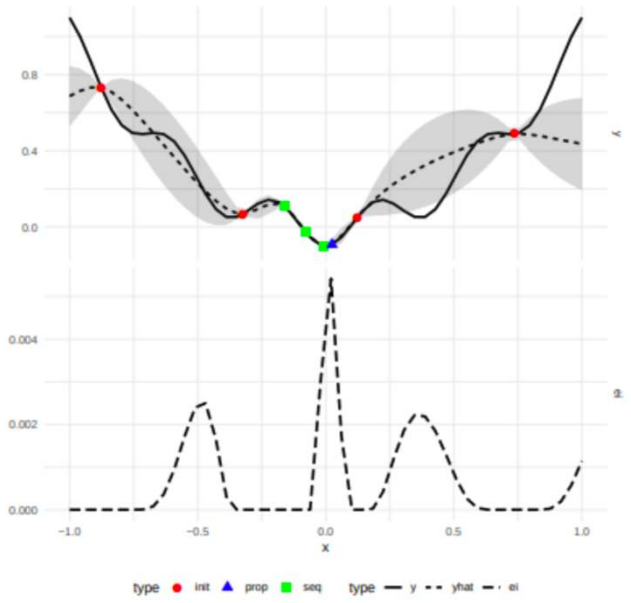
Not about automating data scientists

- Efficient exploration
  - Automate the tedious aspects
  - Make every data scientist a *super data scientist*
- Democratisation
  - Allow individuals and small companies to use machine learning effectively at lower cost
- Data Science
  - Understand algorithms, develop better ones
  - Self-learning algorithms

Automatic Machine Learning (AutoML)

Automatic Machine Learning is the **optimization** over a space of **machine learning pipelines** to minimize a cross validated performance measure.





Representation

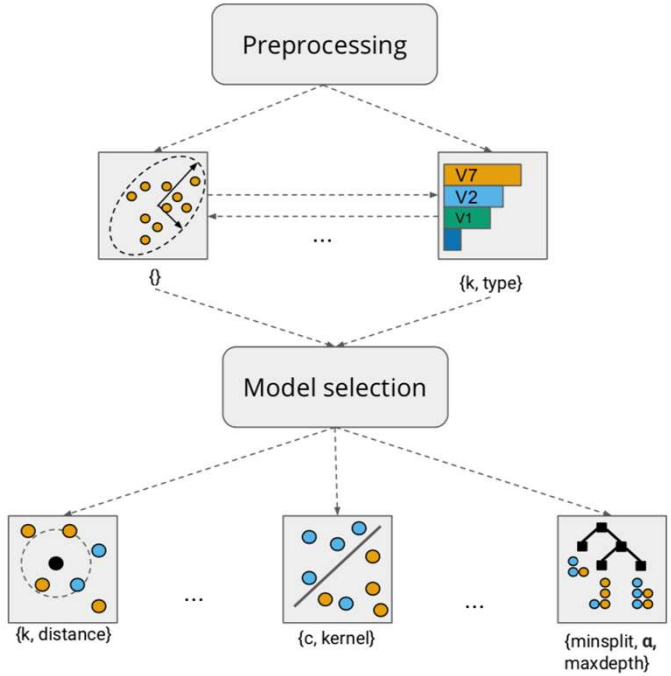
Optimization

Representation

Machine learning pipelines describe the full process of modeling:

- Preprocessing
- Modeling
- Postprocessing

Each step of these steps can be parameterized.



➢ Pipelines induce a hierarchical, mixed (continous and discrete) search space.

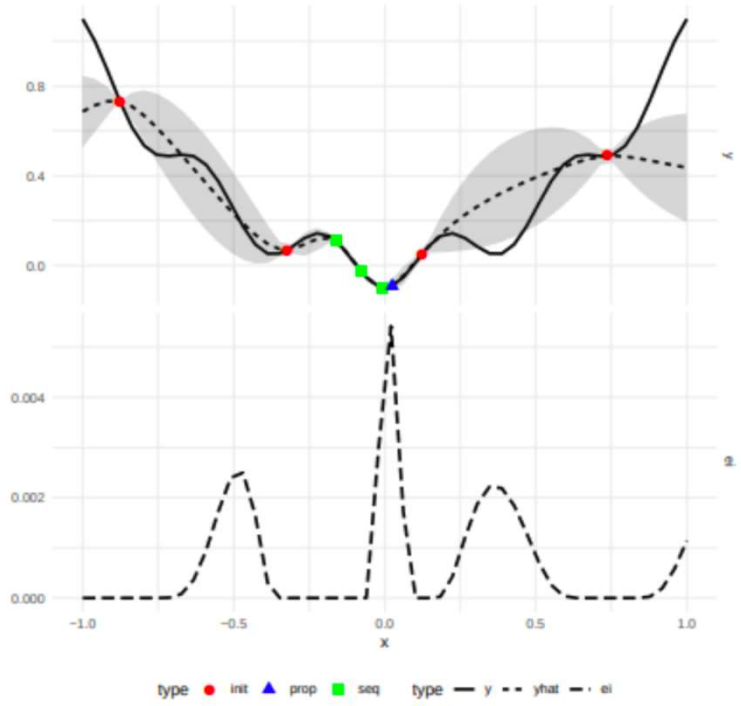
Optimization

Optimization strategies:

- Random search
- Bayesian optimization
- Genetic algorithms
- Reinforcement learning
- …

Sped up by:

- Meta-learning
- Early stopping
- Data subsampling



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
ADA LOVELACE CENTER
Fraunhofer IIS

Current Challenges of AutoML

Very large and complex search spaces

- How can the search space be simplified?
- How much performance is lost by restricting the search space?

Lack of flexibility

- How to consider sparseness, model size, fairness?
- Multiple objectives or additional constraints?

Finding a good Search Space

- Collect data about machine learning
  - 24 hours on 300.000 CPUs
  - Random exploration
  - 5tb of ML meta-data

Leibnitz Rechenzentrum
SuperMUC NG

- What can we do with this data?
  - Train surrogate models
  - Learn about important hyperparameter
  - Learn parameter ranges
  - …

Kühn, D., Probst, P., Thomas, J., & Bischl, B. (2018). Automatic Exploration of Machine Learning Experiments on OpenML. *arXiv preprint arXiv:1806.10961*.

Multi-Objective AutoML

- Current AutoML approaches are very good at optimizing predictive performance!

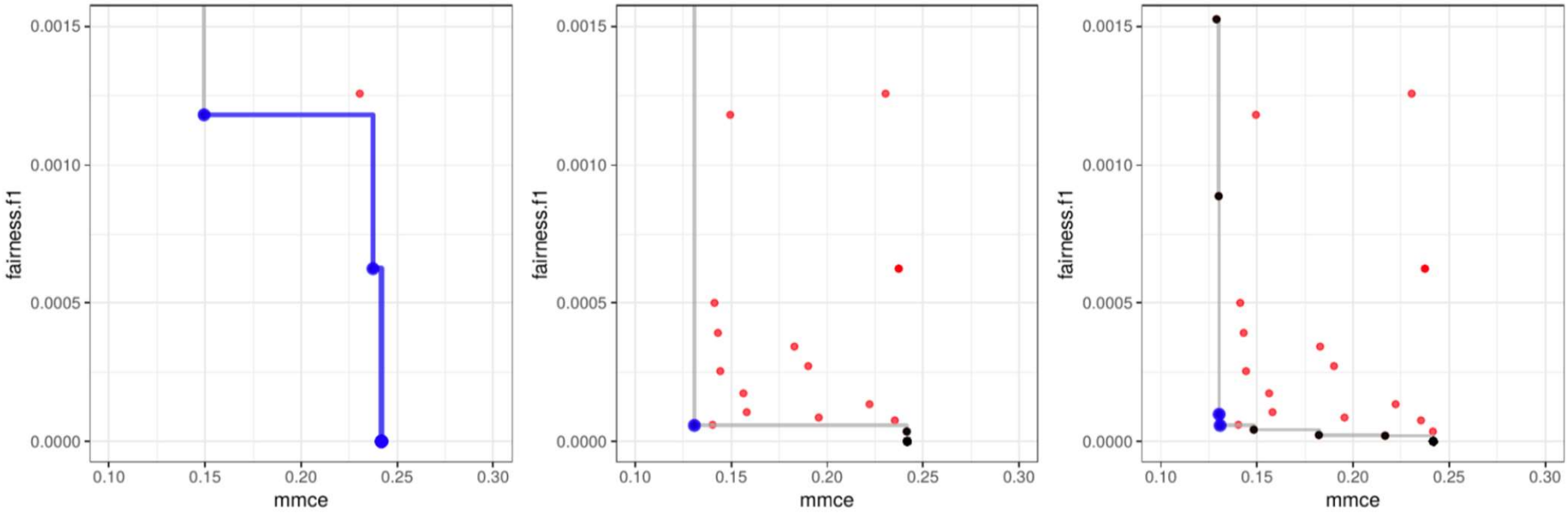
- Many real-world applications require models that *trade of* or *are good with respect to* multiple objectives.

Problem:

- Too narrow focus on a single measure for predictive performance!

- Users either use AutoML without considering other objectives, or do analysis manually!

Multi-Objective AutoML



Pareto front of fairness and predictive performance (mmce) after 20, 70 and 120 iterations of `AutoxgboostMC`.

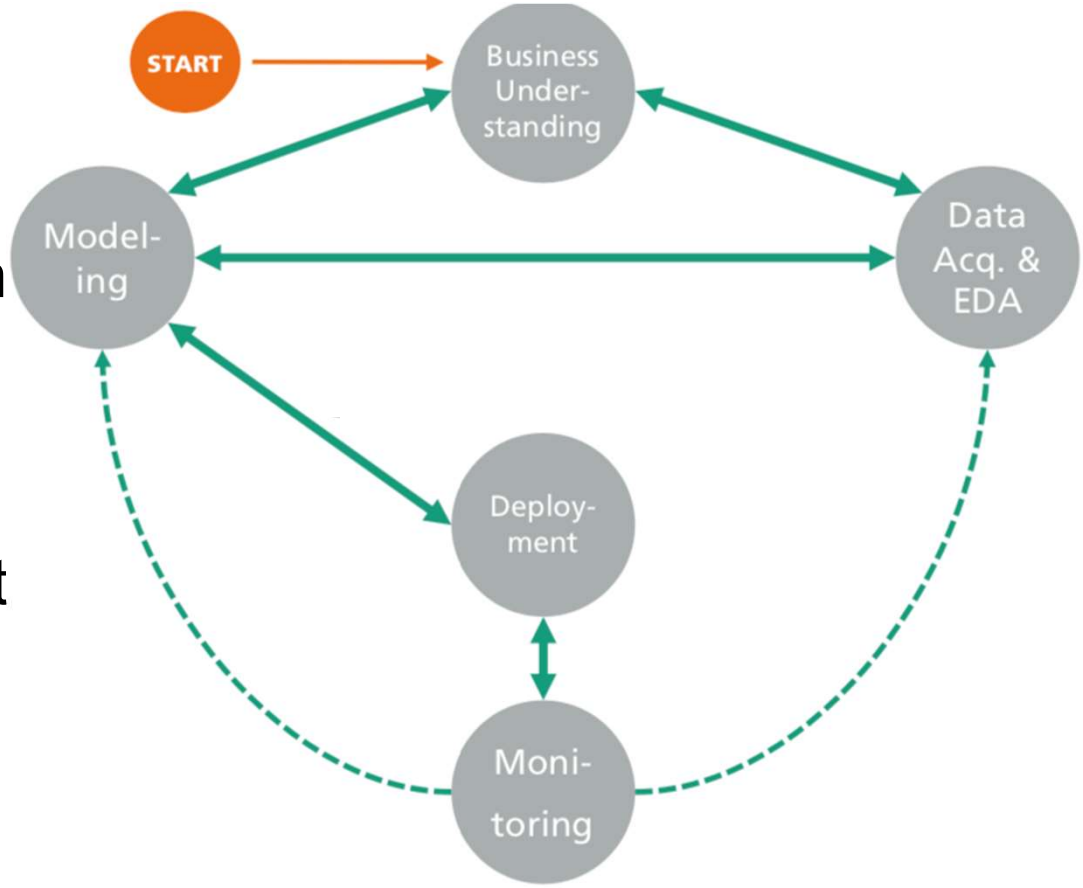
Pfisterer, F., Coors, S., Thomas, J., & Bischl, B. (2019). Multi-Objective Automatic Machine Learning with AutoxgboostMC. *ECMLPKDD Workshop on Automating Data Science (ADS)*

Automating Data Science

How do we move from **AutoML** to **AutoDS**?

Automation (potential)
of the full machine learnin
lifecycle:

- Automated EDA
- Automated deployment
- Automated monitoring



➢ *ECMLPKDD Workshop on Automating Data Science 2019*

Thank You!

janek.thomas@scs.fraunhofer.de

Kühn, D., Probst, P., Thomas, J., & Bischl, B. (2018). Automatic Exploration of Machine Learning Experiments on OpenML. *arXiv preprint arXiv:1806.10961*.

Pfisterer, F., Coors, S., Thomas, J., & Bischl, B. (2019). Multi-Objective Automatic Machine Learning with AutoxgboostMC. *ECMLPKDD Workshop on Automating Data Science (ADS)*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ADA LOVELACE CENTER

Fraunhofer IIS