Berlin Time Series Analysis Meetup

Berlin, Deutschland
634 Mitglieder · Öffentliche Gruppe

exXcellent solutions

Forecasting practitioners:

WHAT CAN WE LEARN FROM KAGGLE COMPETITIONS?

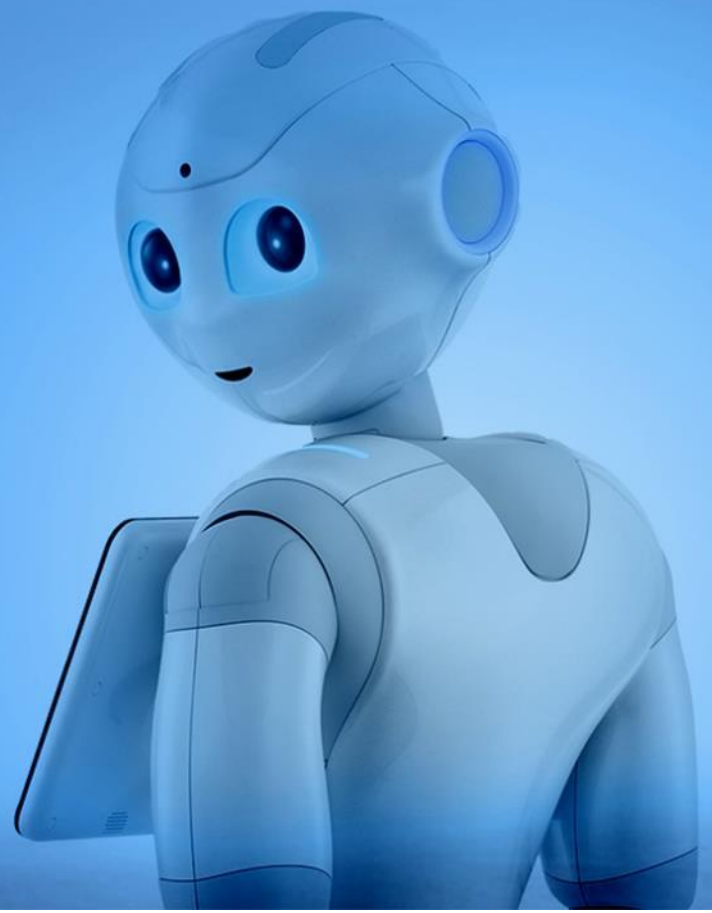
Thomas Bierhance
Practice Lead for Data Science & AI

@datenzauberai

the essence for your business.

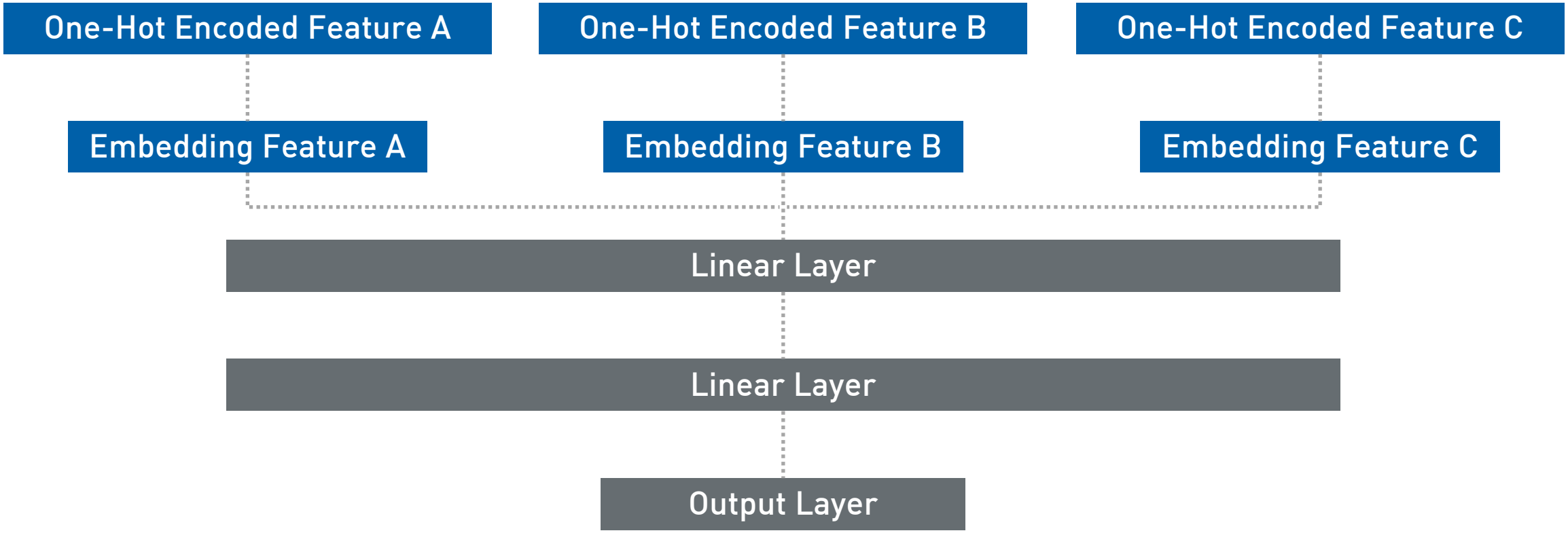

REAL LIFE PROJECTS ARE NOT LIKE A KAGGLE COMPETITION
HUNT FOR DATA LEAKS,
PROBE THE LEADERBOARD,
TUNE WITHOUT FEEDBACK FROM PEOPLE,
SQUEEZE THE LAST 0.5% OUT OF A HUGE ENSEMBLE

1. Embeddings for categorical features



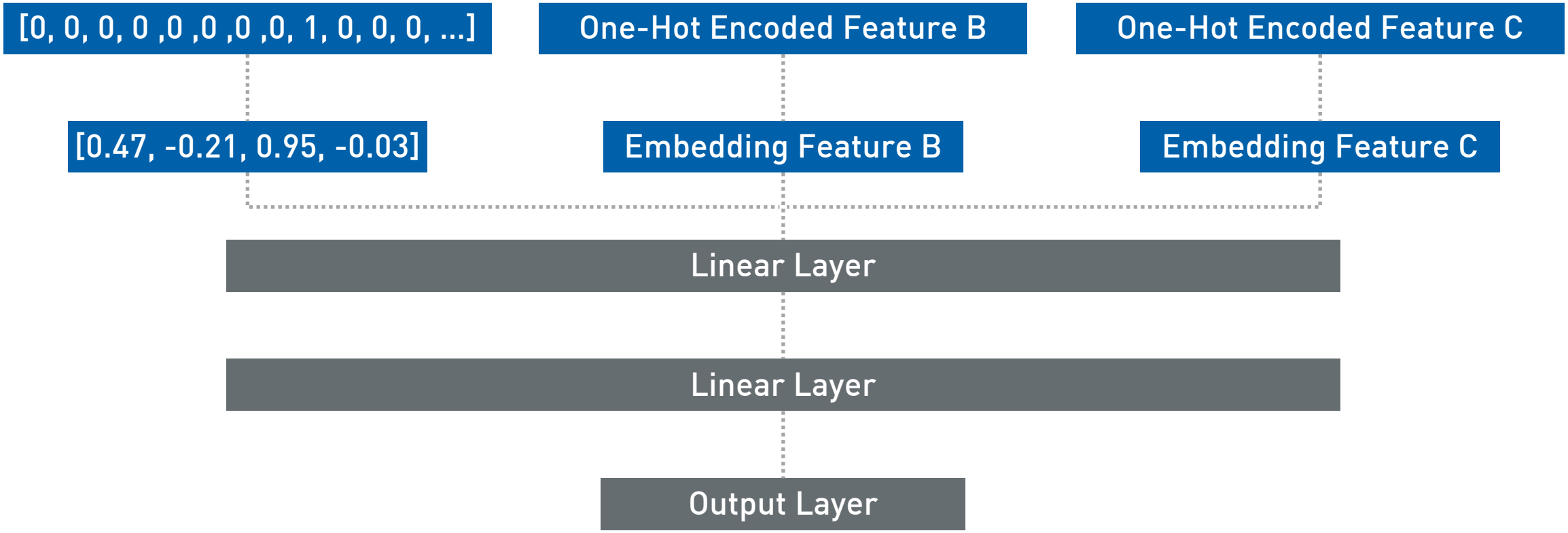
Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

Sample architecture for embeddings



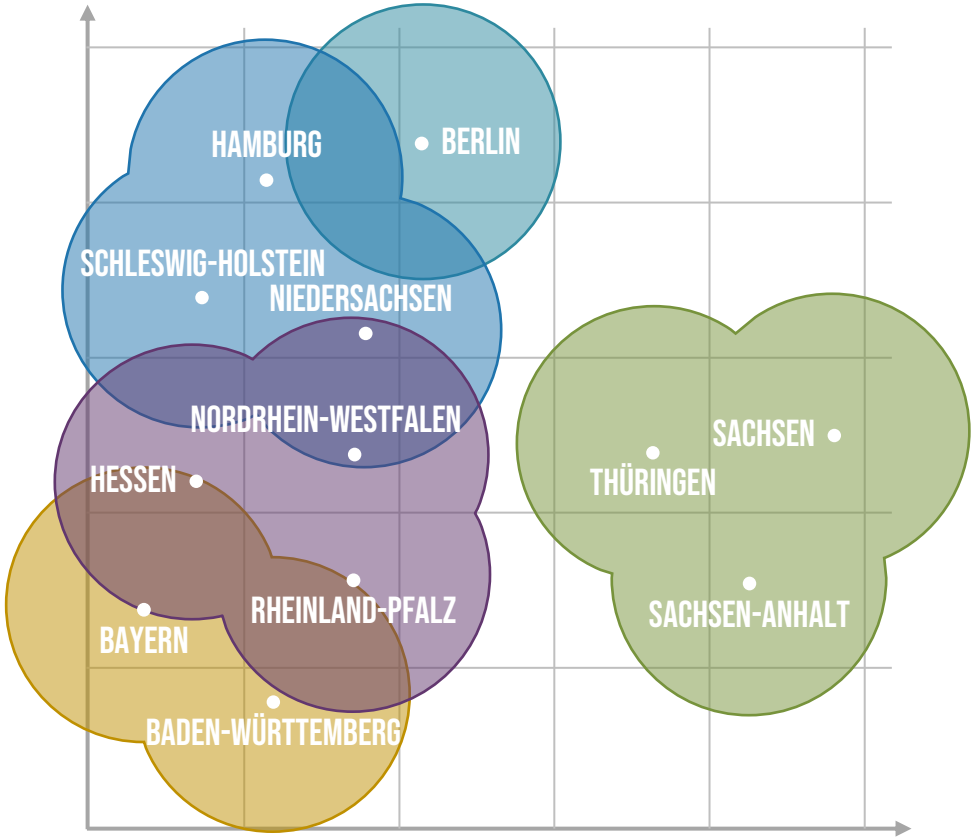
Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

Sample architecture for embeddings

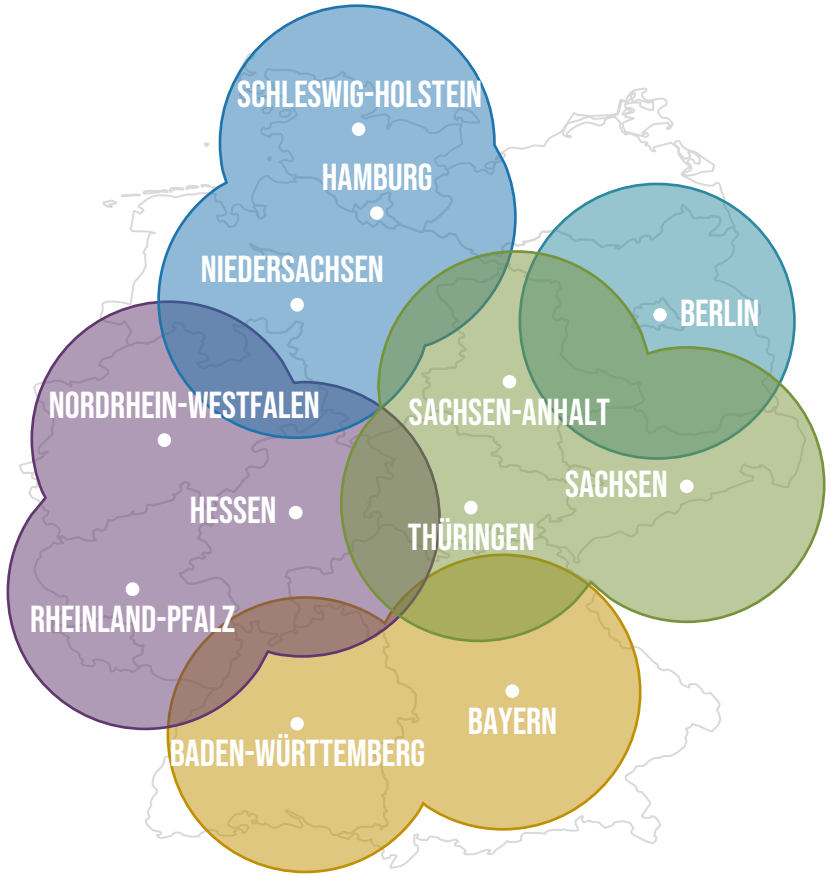


Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

Embeddings for federal states (Bundesländer)



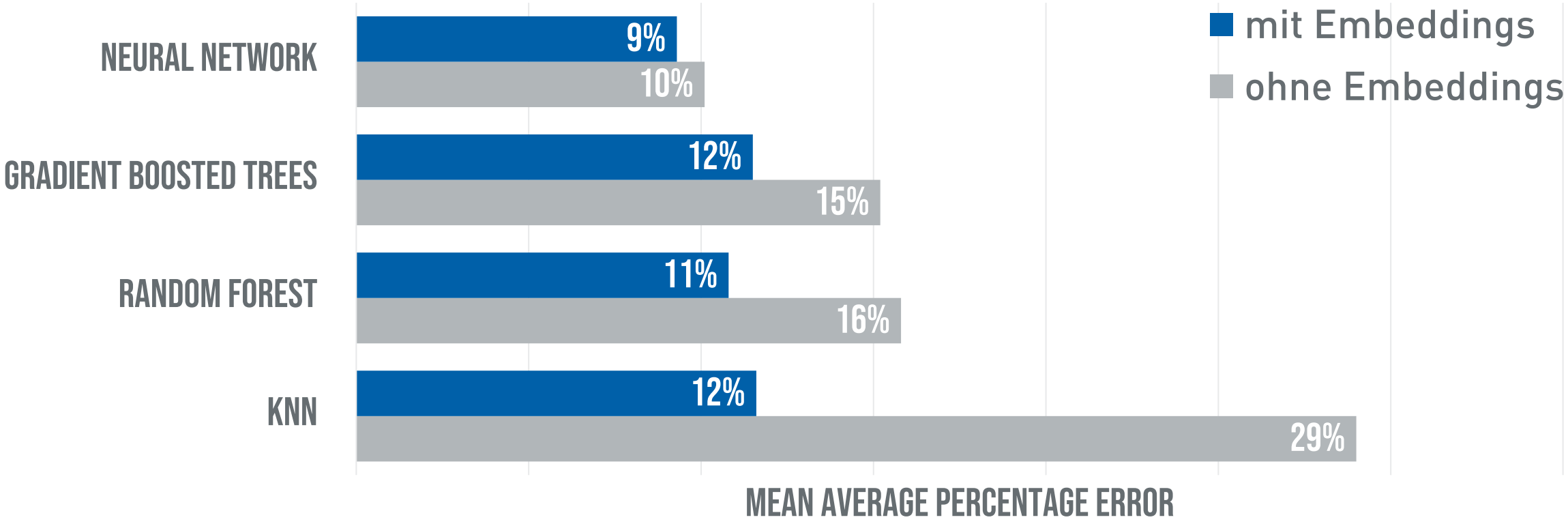
2D projection of the embeddings
(t-SNE)



federal states for comparison

Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

Embeddings with trees and KNN



Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

What about a new categories?

Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

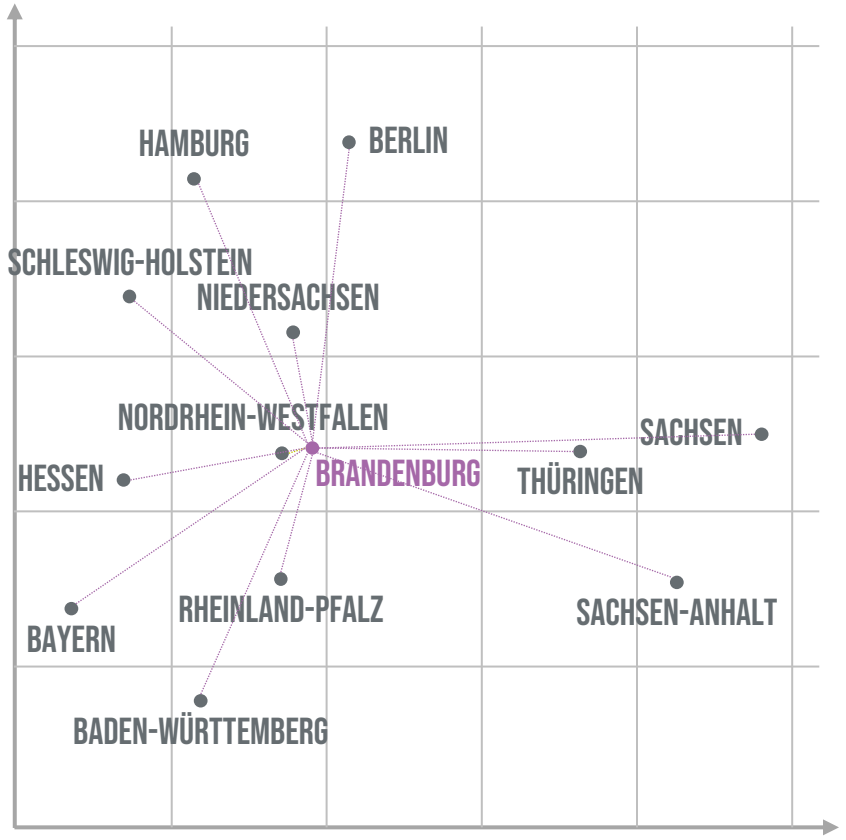
New categories: Best Guess

Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

New categories: Best Guess

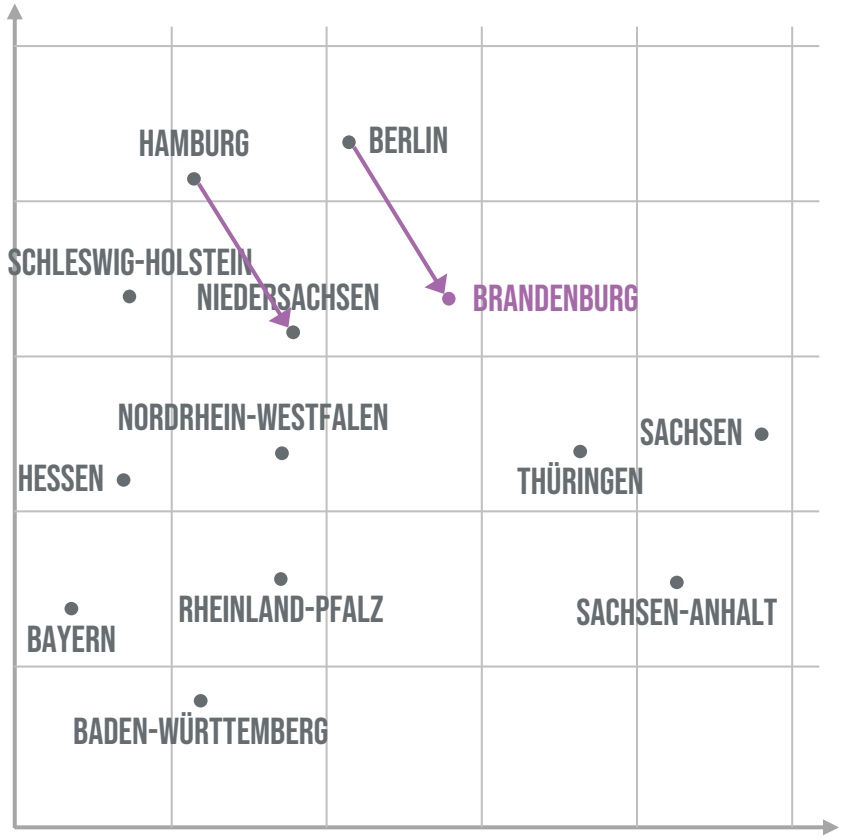
Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

New categories: Best Guess, Mean

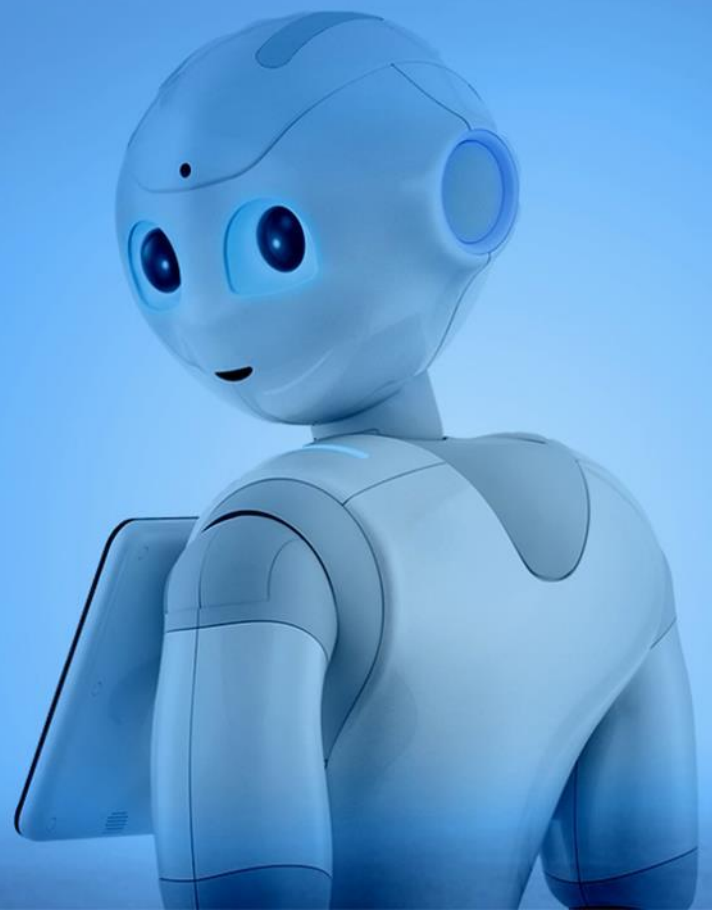


Rossmann Store Sales, 3rd place (Cheng Guo, Felix Berkhahn)

New categories: Best Guess, Mean, Difference



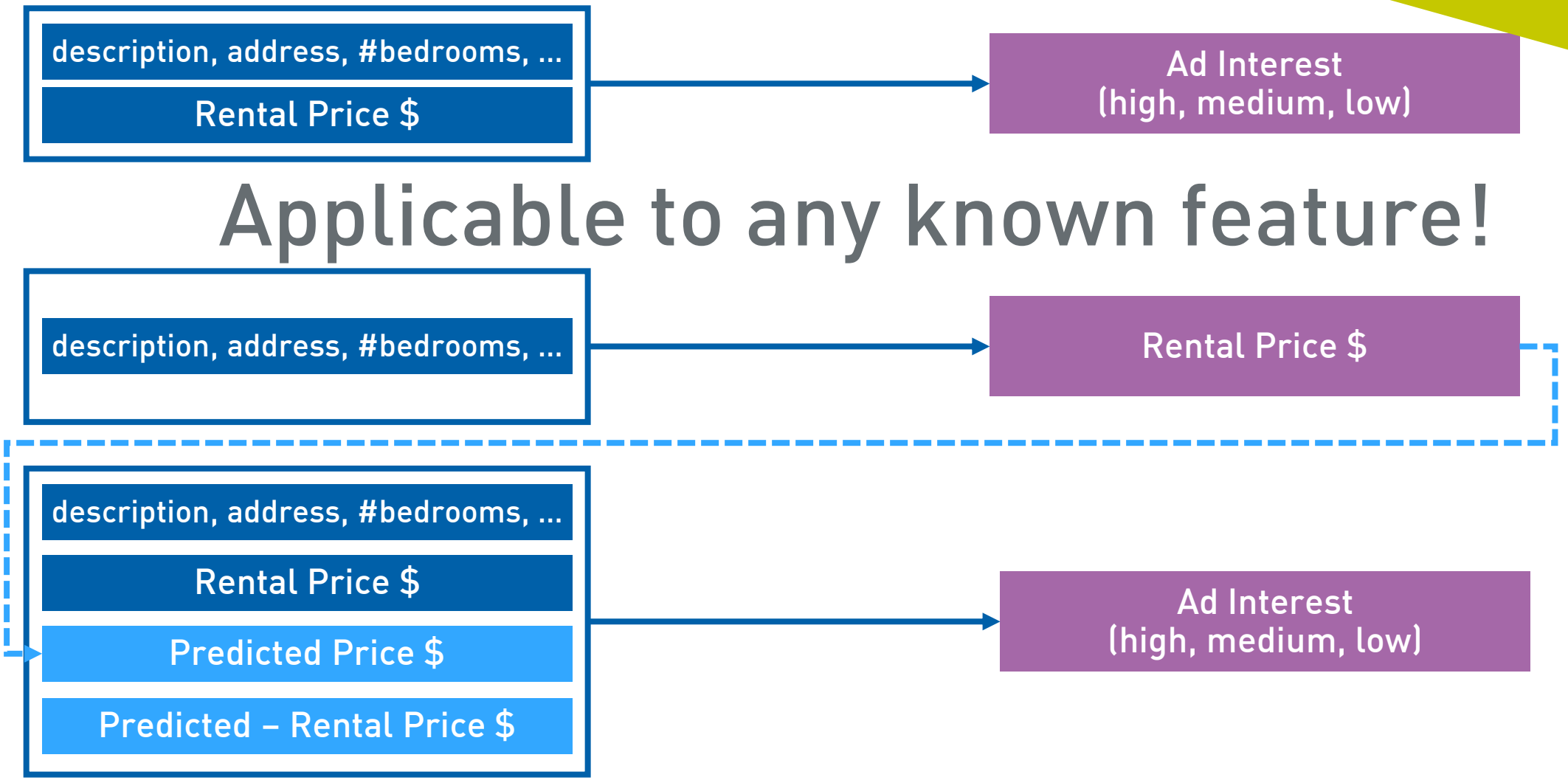
2. Predict the known



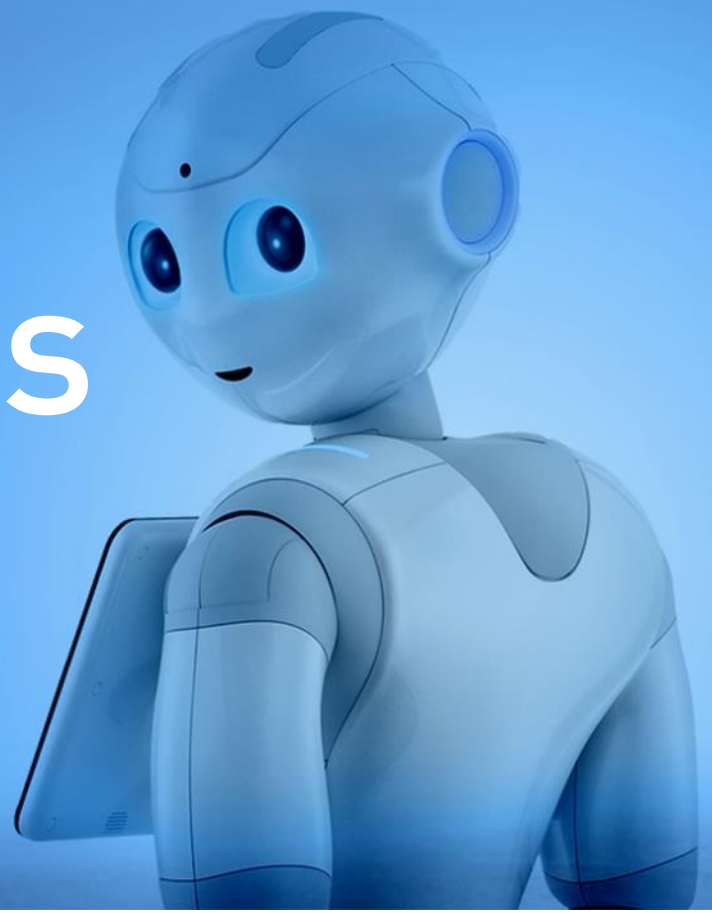
Two Sigma Connect, 26th place (Jean-François Puget)

Predict the known to predict the unknown

Highly recommended:
https://youtu.be/VC8Jc9_lNoY



3. Combine statistical algos w/ ML algos

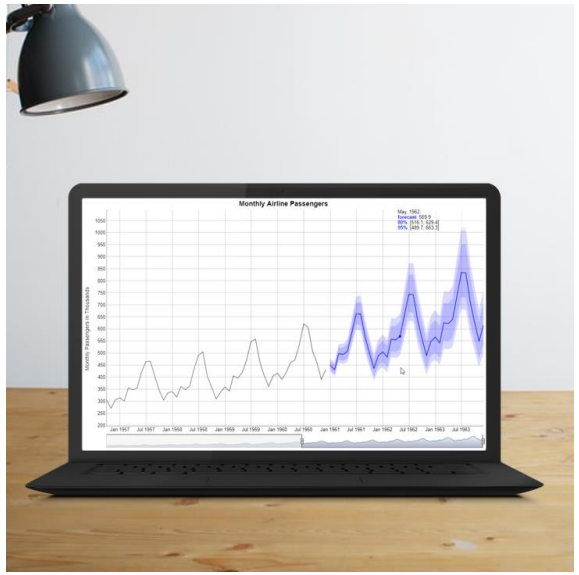


Advantages of statistical time series algos

Trend & seasonality are built-in

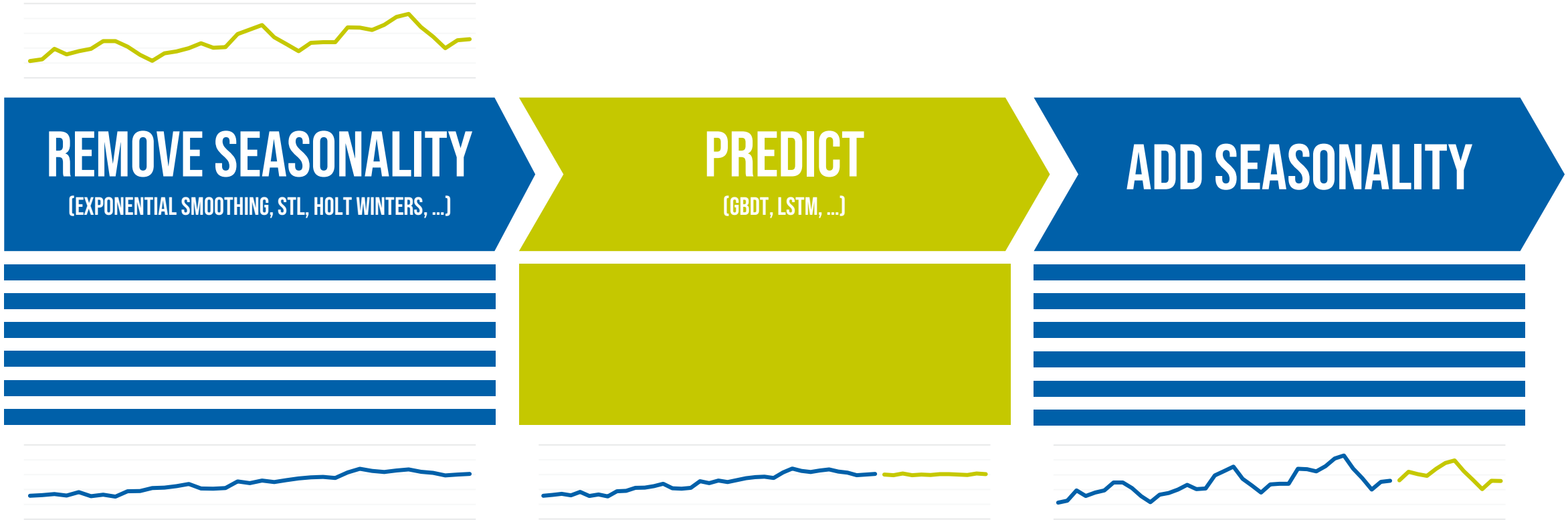
Squeeze everything out of few data

Are flexible to expand and combine



Able to extrapolate

Handling the seasonality



Extrapolation des Trends

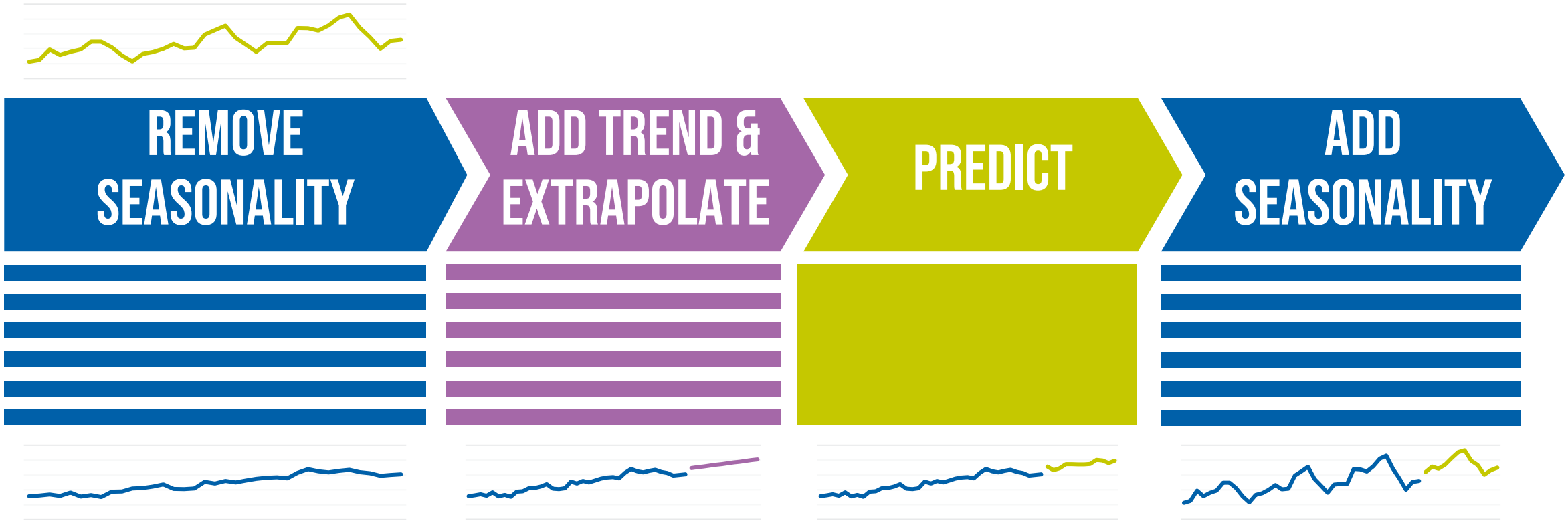
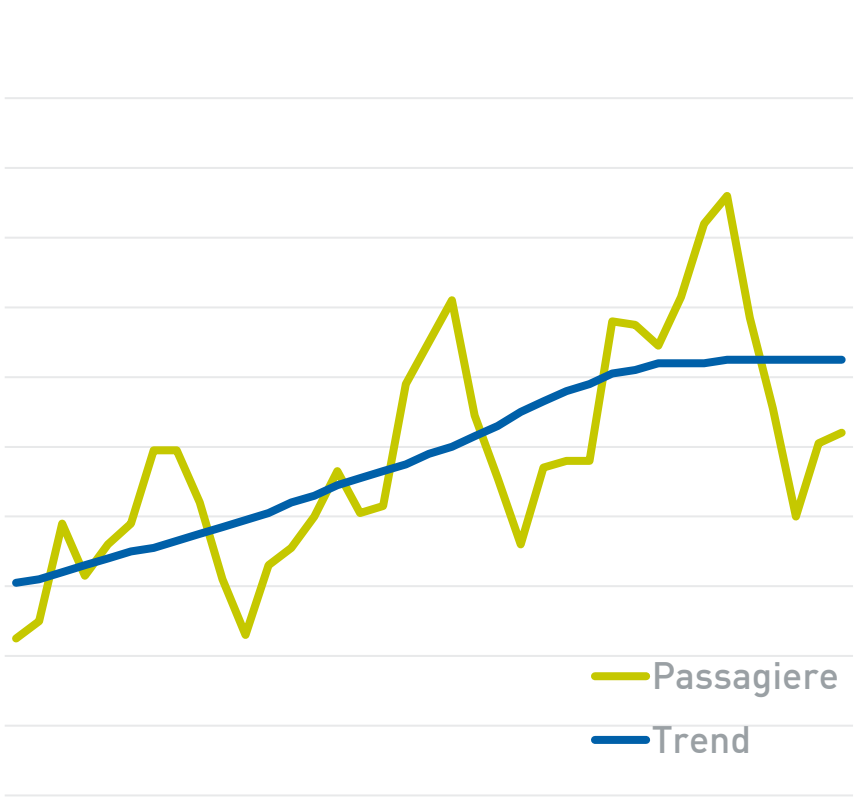
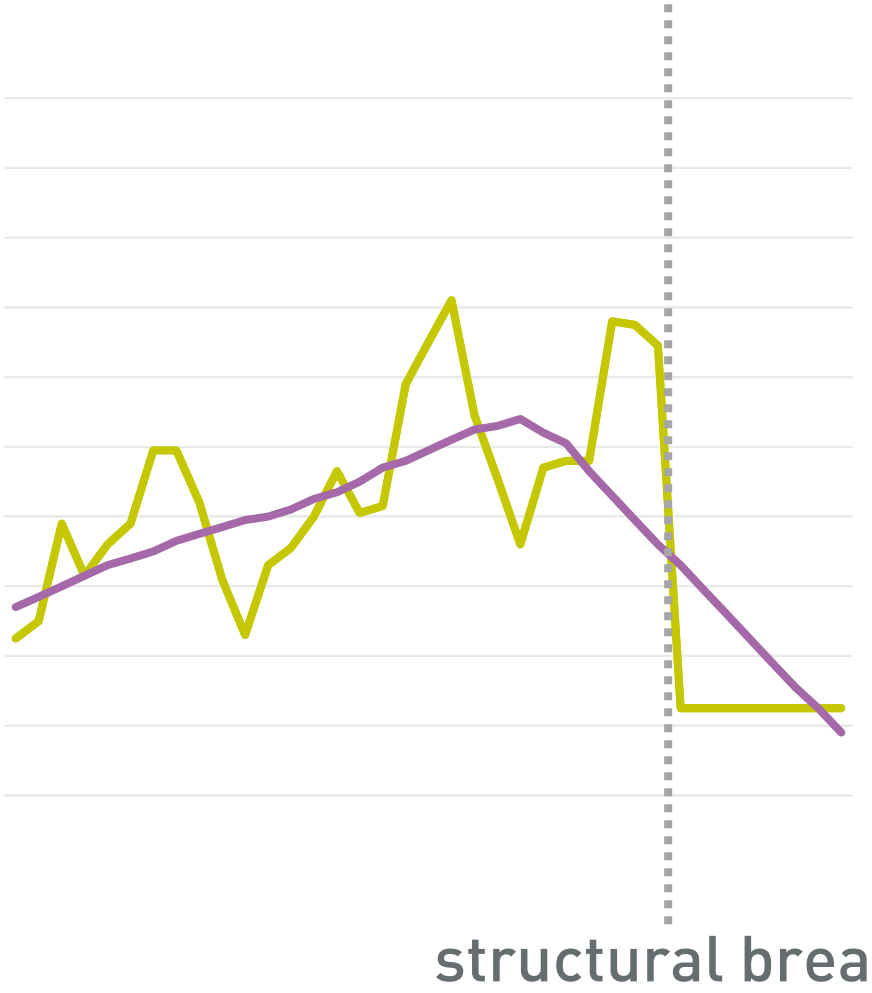


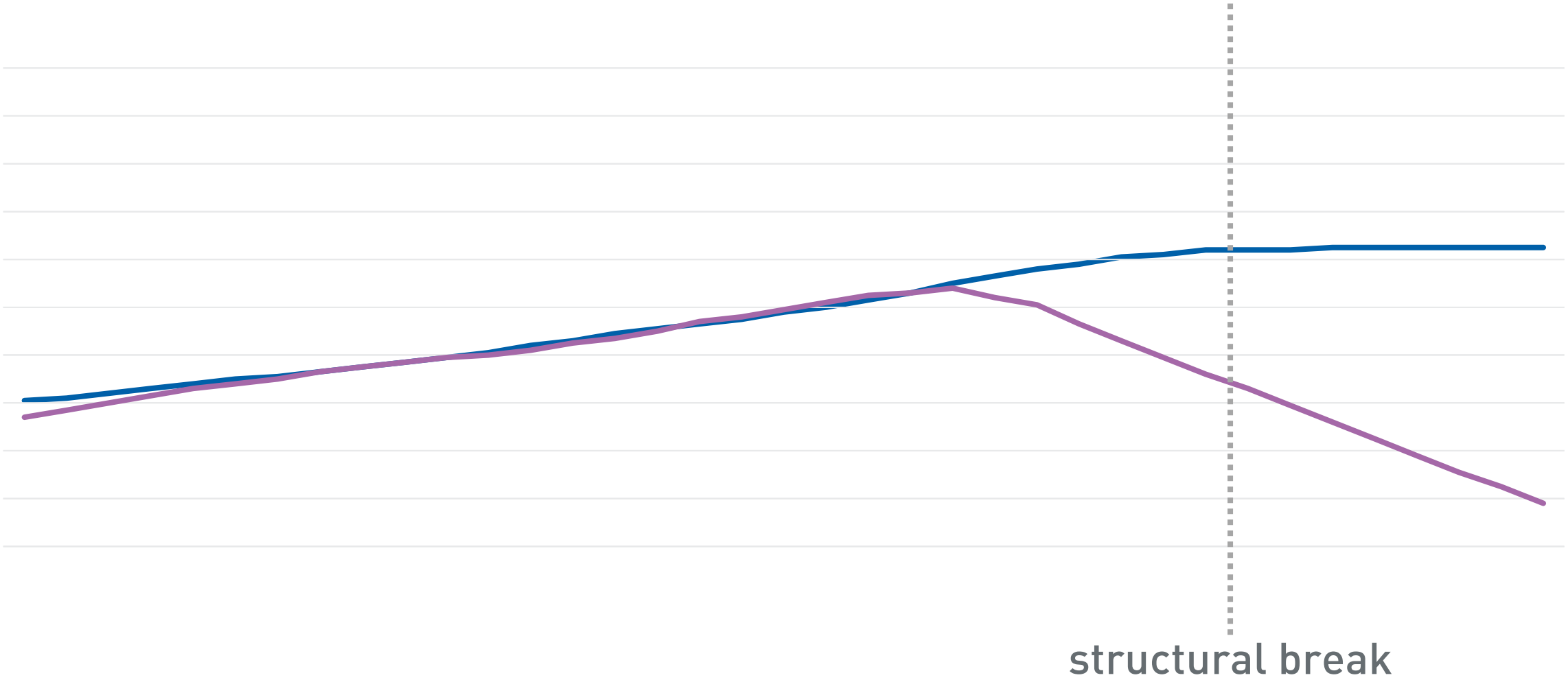
PHOTO BY JUSTUS MENKE ON UNSPLASH

Statistical algos: back to the future

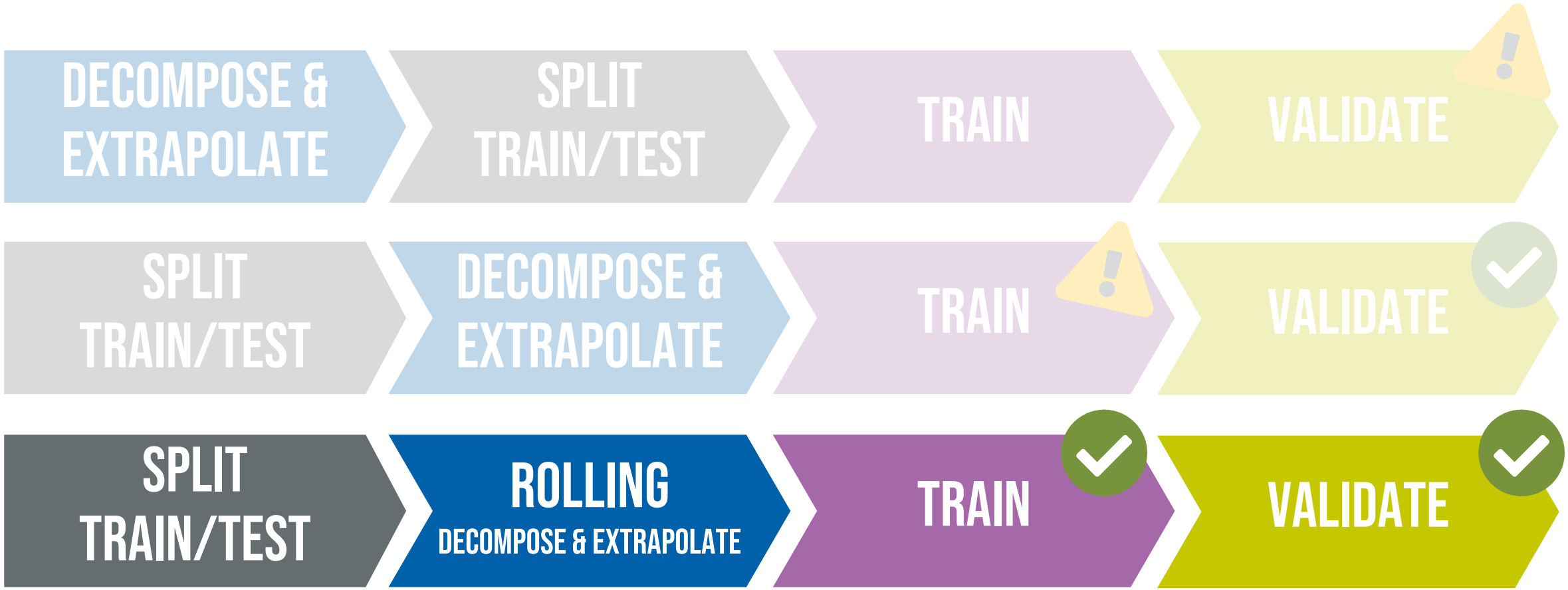




Statistical algos: back to the future



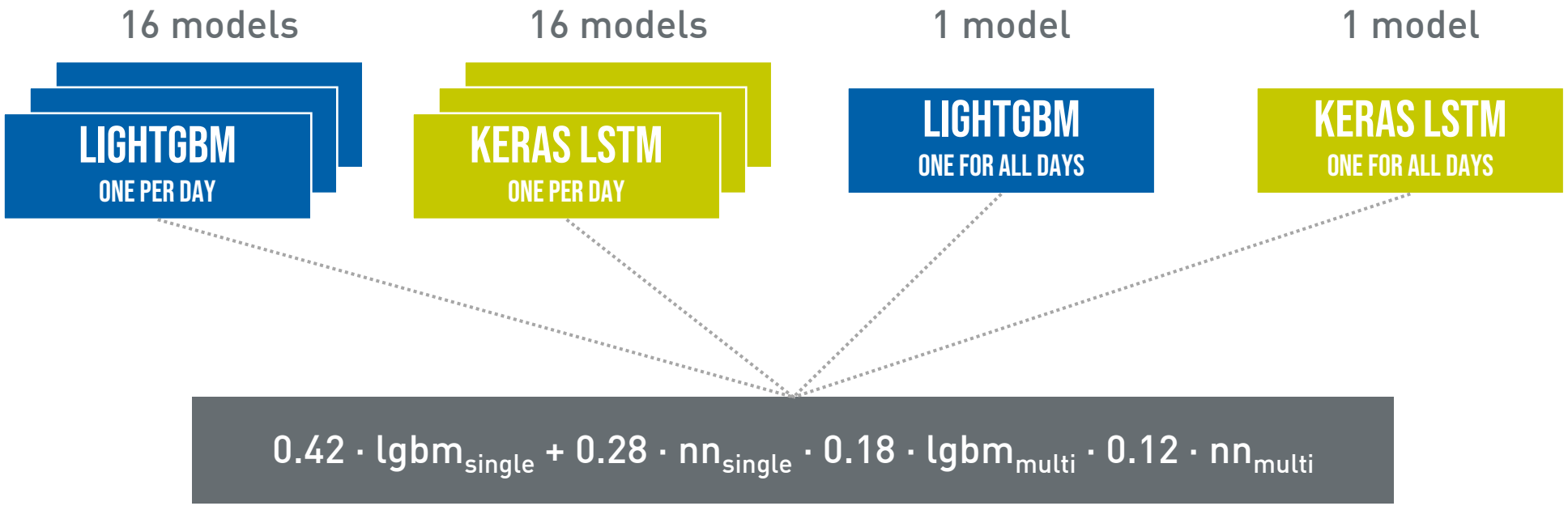
Strategies for Split, Feature Engineering and Training



4. Ensembles

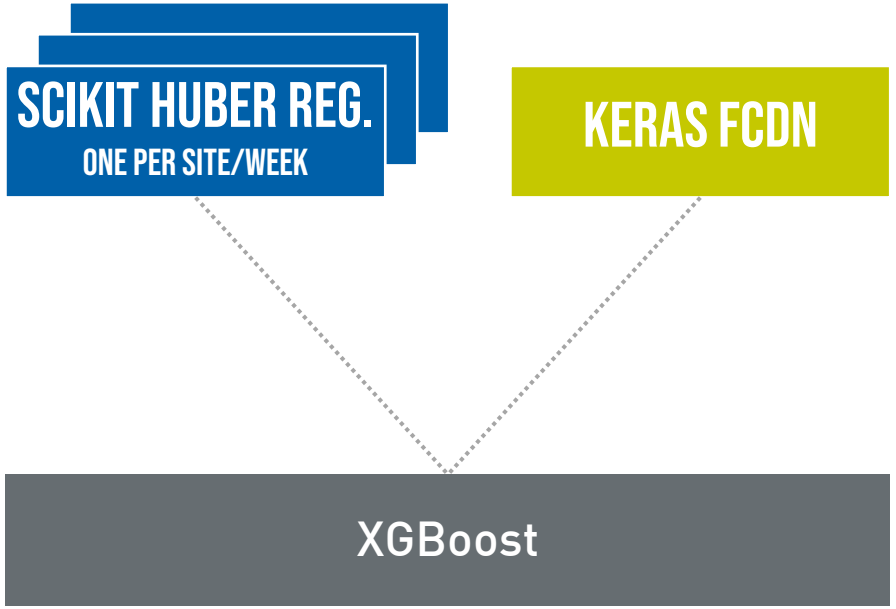
Corporación Favorita Grocery Sales Forecasting, 1st place (Eureka, Weiwei, infzero)

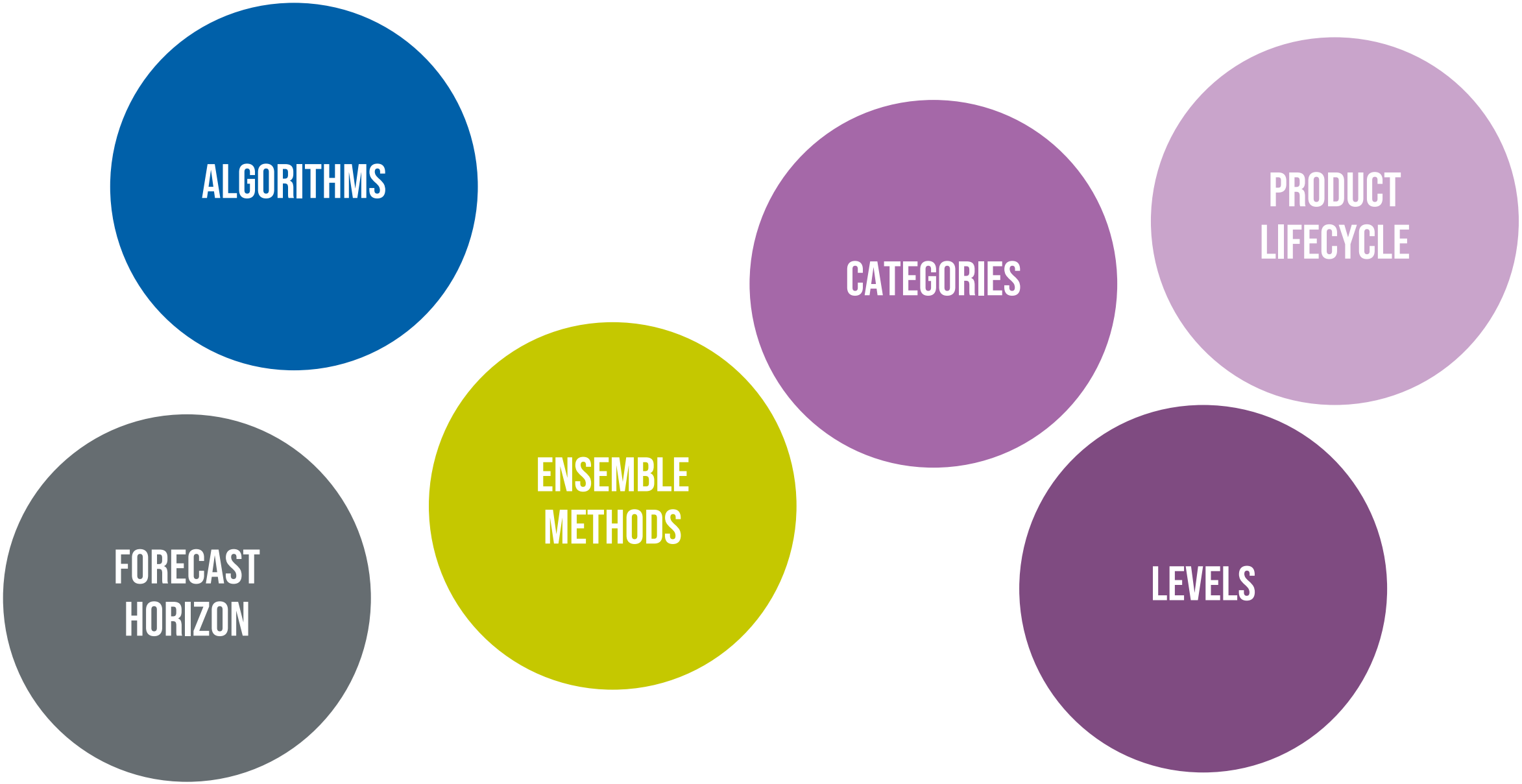
Simple combination of 4 models



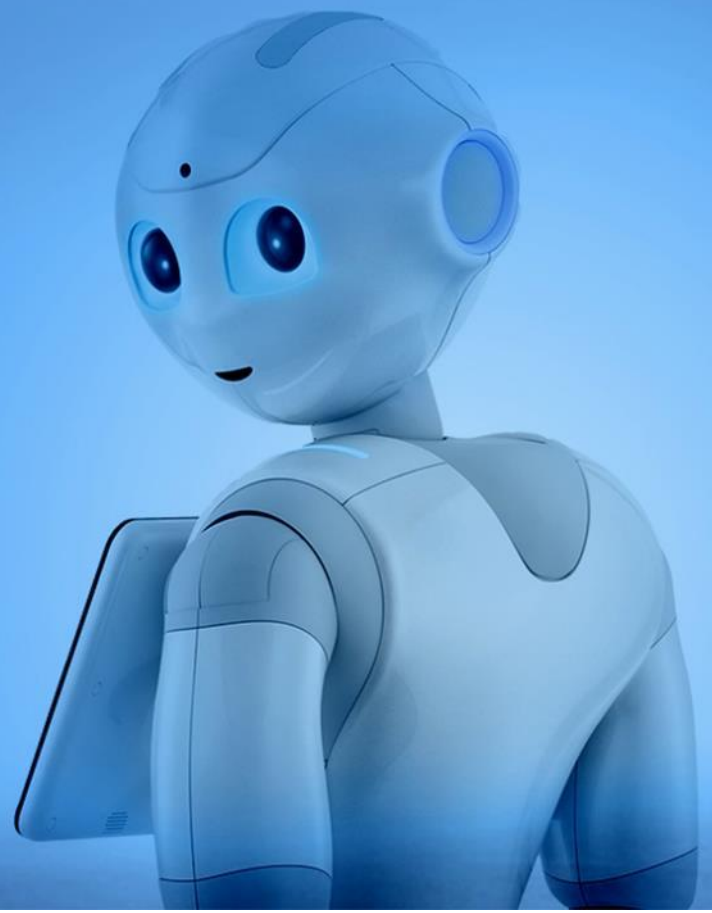
Web Traffic Time Series Forecasting, 2nd place (Jean-François Puget)

Stacking/Boosting mit XGBoost



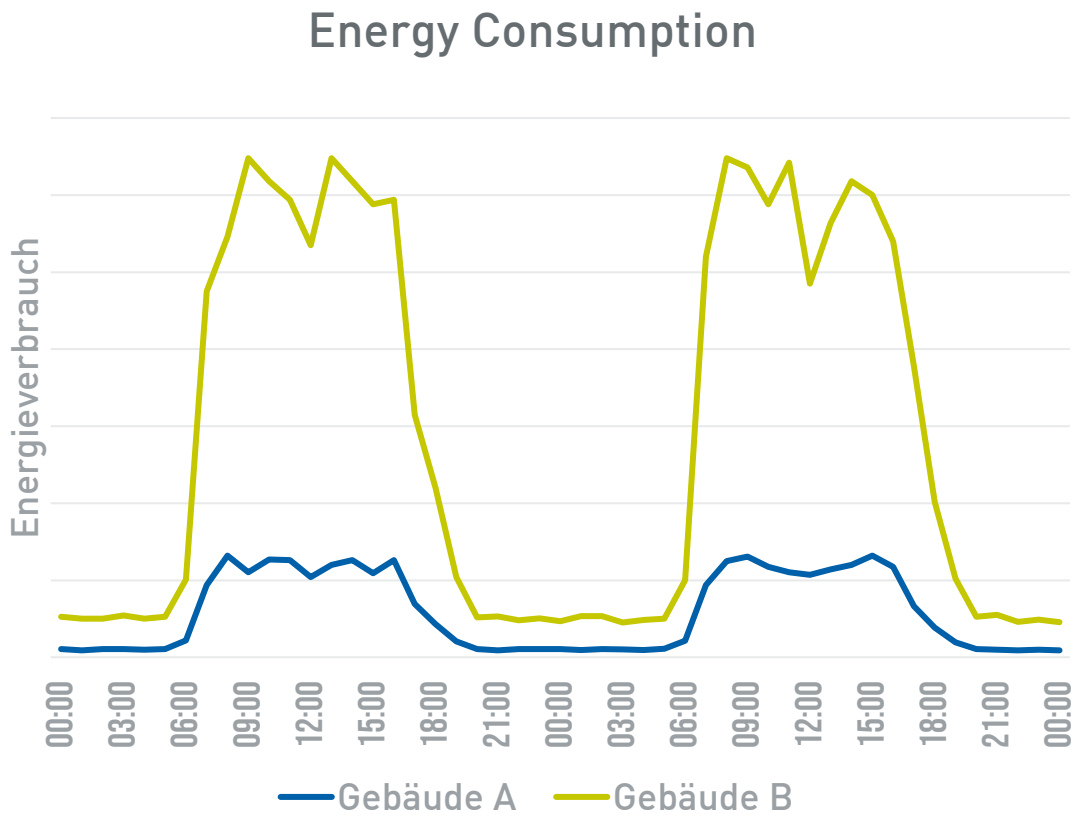

ALGORITHMS
CATEGORIES
PRODUCT LIFECYCLE
ENSEMBLE METHODS
FORECAST HORIZON
LEVELS

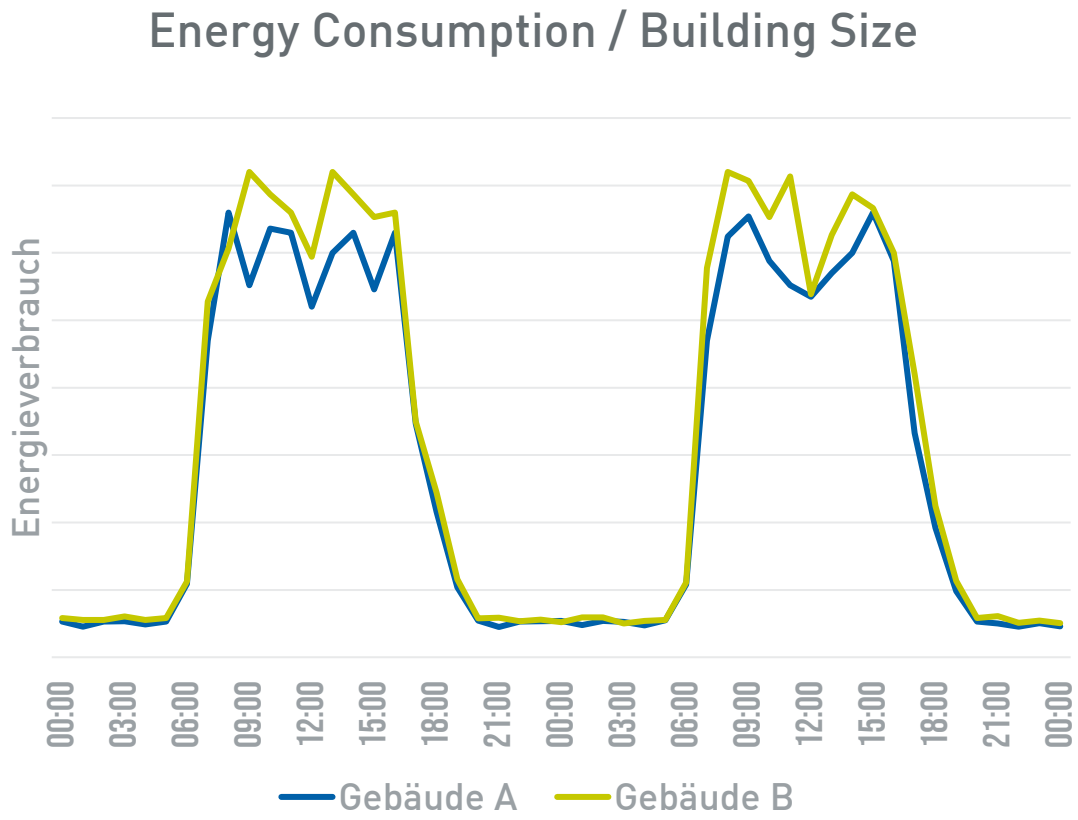
5. Transform the target variable



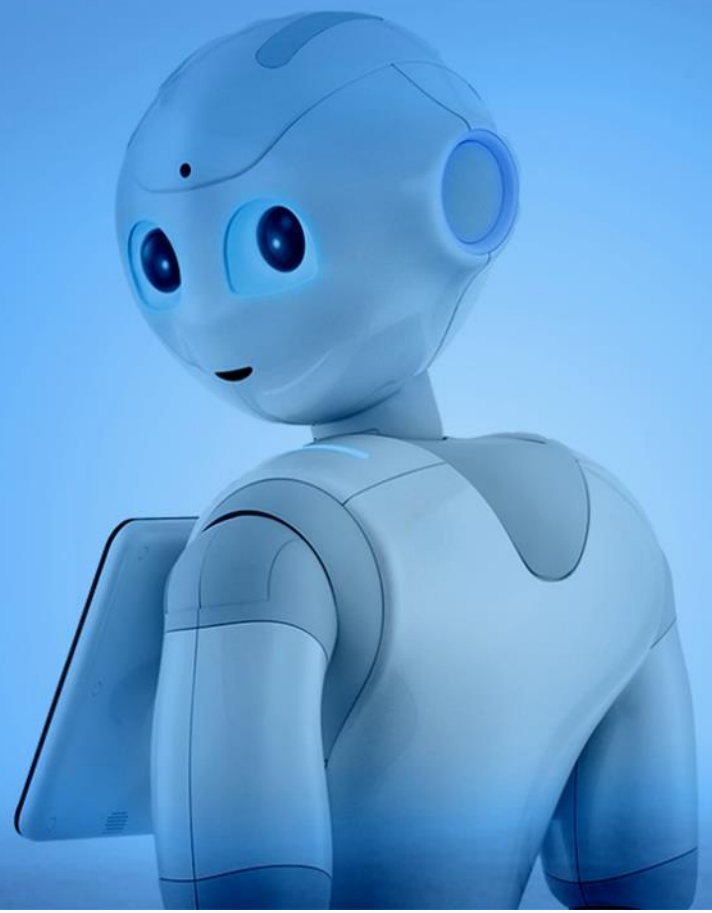
ASHRAE - Great Energy Predictor III, 1st place (Matt Motoki, Isamu Yamashita)

Transform the target variable





6. External Data



Learning from Kaggle's Forecasting Competitions (Casper Solheim Bojer, Peder Meldegaard)

Which external data is worthwhile?

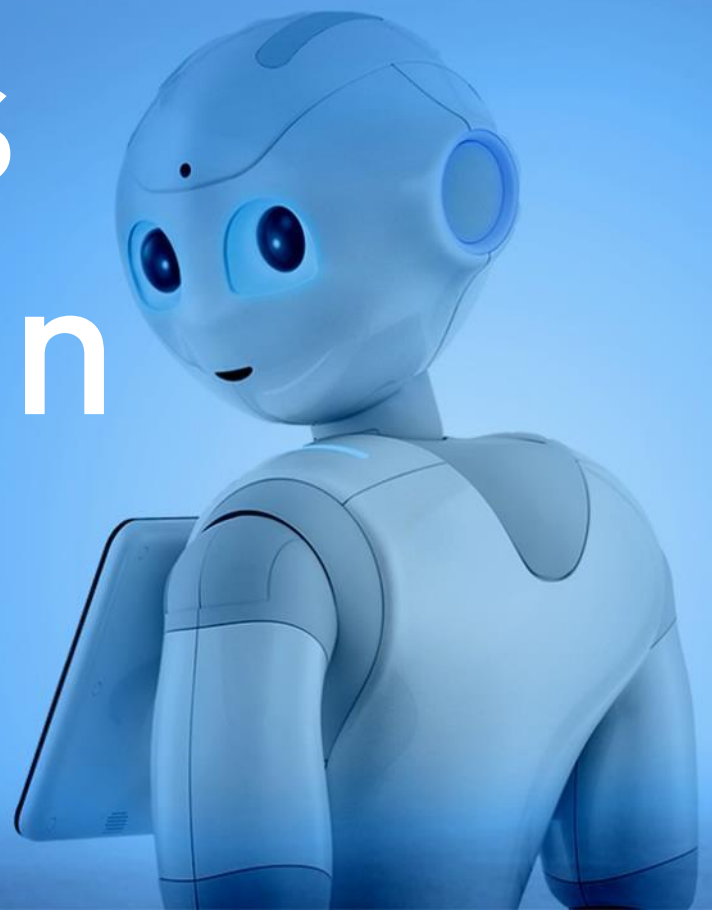
Data that do not require predictions are often very helpful.

Vacations and public holidays

Data that require predictions are difficult.

macroeconomic indicators, weather, ...

7. Simple models can actually win too



M5 Forecasting - Accuracy, 4. Platz (monsaraida)

It can be so simple...

Features

Sales features
- sales_lag_{s|s+1|…|s+14}
- rolling_mean_{7|14|30|60|180}
- rolling_std_{7|14|30|60|180}
- release

Calendar features
- tm_{d|dw|w|w_end|wm|m|y}
- moon
- event_name_{1|2}
- event_type_{1|2}
- snap_{CA|TX|WI}

Price features
- price_{max|mean|min}
- price_{std|norm|nunique}
- price_cent_{max|min}
- price_momentum_{d|m|y}

Id features
- item_id, cat_id, dept_id
- enc_item_id_{mean|std}
- enc_cat_id_{mean|std}
- enc_dept_id_{mean|std}

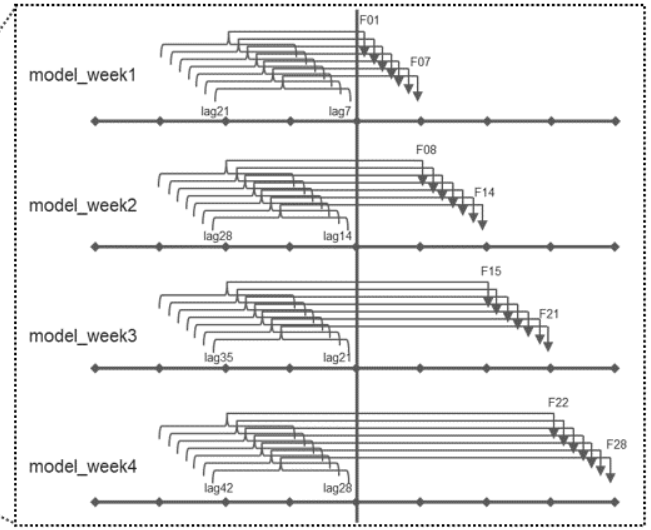
10 stores x 4 weeks = 40 LightGBM models
- objective=tweedie
- no early stopping

store=CA_1, store=CA_2, store=CA_3, store=CA_4, store=TX_1, store=TX_2, store=TX_3, store=WI_1, store=WI_2, store=WI_3

model_week1, model_week2, model_week3, model_week4

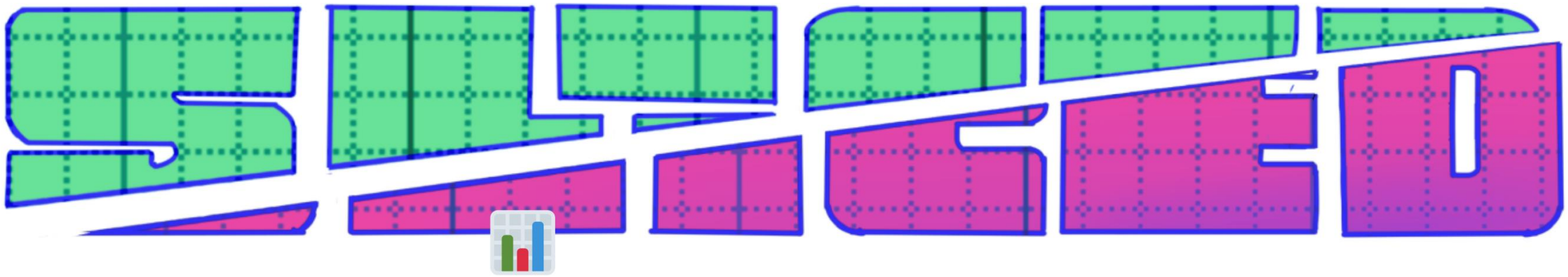
Practical/Simple solution
- no blending/stacking
- no recursive modeling
- no postprocessing/multiplier

Weekly models



More inspiration instead of another boring Netflix show if you can't sleep tonight...

Live Kaggling on Twitch, hosted by Nick Wan and Meg Risdal



SLICED Show

Streaming Tuesdays at 8:30PM ET starting June 2021!

nickwan_datasci - Twitch

Same nickwan as twitter.com/nickwan I stream variety and coding

https://www.twitch.tv/nickwan_datasci

SLICED is like the TV Show Chopped but for data science. Competitors get a never-before-seen dataset and two-hours to code a solution to a prediction challenge. Contestants get points for the best model plus bonus points for data visualization, votes from the audience, and more.