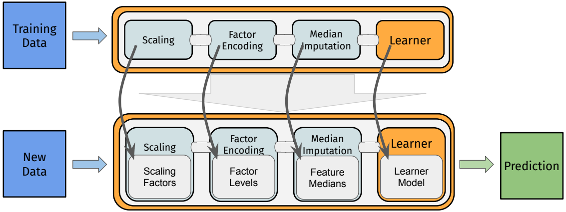
# AutoML: Hyperparameter Optimization

## Practical Problems

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer

# Linear Pipelining

- Applying preprocessing to the whole dataset leads to data leakage
- Preprocessing should have train and predict steps, too
- Can add it to learner, and embed it in CV
- Note: Preprocessing has hyperparameters
- Optimize pipeline jointly: $\mathbf{\Lambda} = \mathbf{\Lambda}_{\text{preproc}} \times \mathbf{\Lambda}_{\mathcal{I}}$
- Still HPO, not much different than for single learner
- $\mathbf{\Lambda}$ "simply" of higher dimension

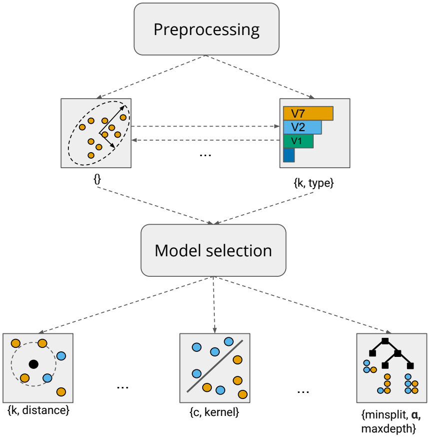


# Nonlinear Pipelines

Ideal to let HPO choose automatically:

- preprocessing
- feature extraction
- learner

$\rightarrow$ $\boldsymbol{\Lambda}$ becomes hierarchical search space!



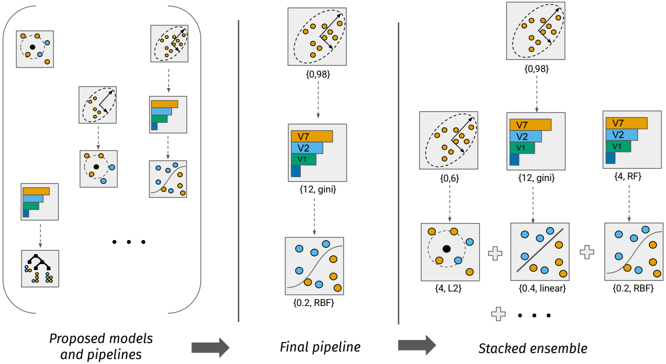

Suitable optimizers:

- TPE [Berstra et al. 2011]
- Random search, hyperband with sampler that follows the hierarchy
- BO with RF (imputation or hierarchical trees) [Hutter et al. 2011]
- Evolutionary approaches (similar to NAS)

# Obtaining Final Model

Options:

- Choose the optimal path as linear pipeline.
- Build ensemble of best configurations (e.g. [Feurer et al. 2015], [LeDell and Poirier. 2020]).



# Current Benchmark on Tabular Data [Gijsbers et al. 2019]

| Framework: | auto-sklearn | Auto-WEKA | H2O AutoML | RandomForest | TPOT |
|---|---|---|---|---|---|
| Binary tasks: | | | | | |
| adult | 1.045 | 1.000 | 1.049 | 1.000 | 1.048 |
| airlines | 1.403 | 1.016 | 1.435 | 0.997 | 1.343 |
| albert | 1.009 | | 1.115 | 1.001 | 0.981 |
| amazon_employee... | 0.972* | 0.886 | 1.048 | 1.003 | 1.012 |
| apsfailure | 1.000 | 0.985 | 1.001 | 1.000 | 1.001 |
| australian | 1.010 | 1.015 | 0.909 | 1.010 | 1.011 |
| bank-marketing | 1.012 | 0.950 | 1.015 | 1.000 | 1.008 |
| blood-transfusion | 1.495 | 1.379 | 1.532 | 0.985 | 1.149 |
| christine | 1.072 | 0.998 | 1.048 | 0.988 | 1.029 |
| credit-g | 0.970* | 0.829 | 0.991 | 1.004 | 0.924 |
| guiellermo | 1.004 | 0.934 | 1.024 | 0.999 | 0.878 |
| higgs | 1.018* | 0.845 | 1.041 | 0.999 | 1.005 |
| jasmine | 0.987 | 0.939 | 1.001 | 0.998 | 1.004 |
| kc1 | 0.999* | 0.934 | 0.992 | 0.987 | 1.013 |
| kddcup09_appetency | 1.181* | 1.043 | 1.176 | 1.016 | 1.134 |
| kr-vs-kp | 1.000* | 0.959 | 1.000 | 0.999 | 0.999 |
| miniboone | 1.008 | 0.957 | 1.010 | 0.999 | 1.001 |
| nomao | 1.002 | 0.973 | 1.002 | 1.000 | 1.001 |
| numerai28.6 | 1.679 | 1.544 | 1.730 | 1.042 | 1.428 |
| phoneme | 0.993* | 0.998 | 1.005 | 1.000 | 1.015 |
| riccardo | 1.000 | 0.996 | 1.000 | 0.999 | 0.992 |
| sylvine | 1.013 | 0.985 | 1.011 | 0.999 | 1.023 |
| Multi-class tasks: | | | | | |
| car | 1.030 | 0.906 | 1.060 | 0.878 | 1.060 |
| cnae-9 | 1.069 | 0.541 | 1.076 | 0.999 | 1.057 |
| connect-4 | 1.184 | -1.565 | 1.409 | 0.954 | 1.276 |
| covertype | 0.976 | -0.361 | 0.856 | 0.944 | 0.933 |
| dilbert | 1.182 | 0.459 | 1.205 | 0.979 | 1.111 |
| dionis | 0.580 | 0.590 | | 1.002 | |
| fabert | 1.026 | -5.235 | 1.049 | 1.004 | 1.005 |
| fashion-mnist | 0.995 | 0.717 | 1.052 | 0.993 | 0.841 |
| helena | 0.660 | -18.420 | 1.905 | 0.970 | 1.676 |
| jannis | 1.083 | -1.989 | 1.065 | 0.973 | 0.987 |
| jungle_chess... | 1.299 | -3.309 | 1.235 | 0.933 | 1.459 |
| mfeat-factors | 1.059* | 0.789 | 1.053 | 0.992 | 1.018 |
| robert | -0.001 | | 1.545 | 1.000 | 0.640 |
| segment | 1.004 | 0.808 | 1.012 | 0.992 | 1.008 |
| shuttle | 1.000 | 0.979 | 1.000 | 1.000 | 1.000 |
| vehicle | 1.102 | -4.630 | 1.166 | 0.986 | 1.099 |
| volkert | 1.002 | -5.585 | 1.111 | 0.954 | 0.945 |

Table 2: Performance of AutoML frameworks, scaled between a constant class prior predictor (=0) and a tuned random forest (= 1). Missing values mean that no results were returned in time. *: the task was also included in meta-learning models.

- On some datasets AutoML yields big performance improvements
- On many datasets RF is equally good
- Need more and diverse benchmarks