http://bit.ly/mlmu-fb-group

http://bit.ly/mlmu-slack

CEAi

IT VALLEY KOŠICE

T · · Systems ·

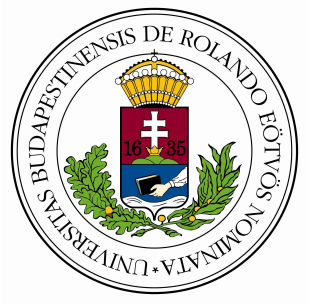
Thank you for the support!

Ke 14 | 14 Eastcubator

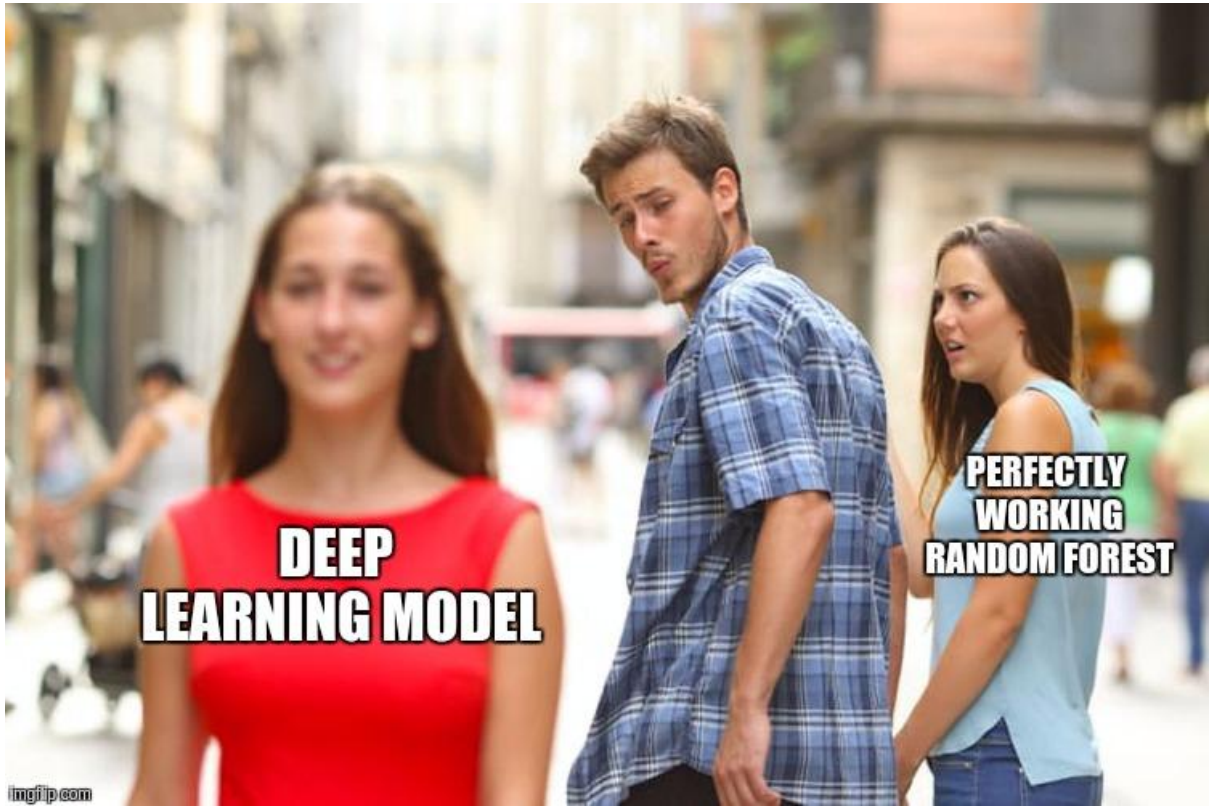
Automation of Data Science

Tomáš Horváth

Talk at the Machine Learning Meetup Košice, Slovakia
February 25, 2019



Overview



Questions

- Which model to use?
- How to tune its HPs?

Buzzwords

- Metalearning
- Hyper-parameter tuning

Contents

- Main approaches
- No technology demos
  - sorry...

Shortly about me

2002: MSc @ UPJŠ

2008: PhD @ UPJŠ

- Fuzzy Relational Learning, User Modeling

2009 - 2012: Post-doc @ University of Hildeseim, Germany

- Recommendation Techniques, GPR Image Processing

2015 - 2016: Post-doc @ University of São Paulo, Brasil

- Meta-learning, Hyper-parameter Tuning

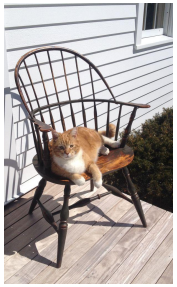
2016 - now: department chair @ T-Labs, Eötvös Loránd University of Budapest, Hungary

- Time-series analytics, Text Mining
- Applied Data Science

Thinking about a Data Science project...



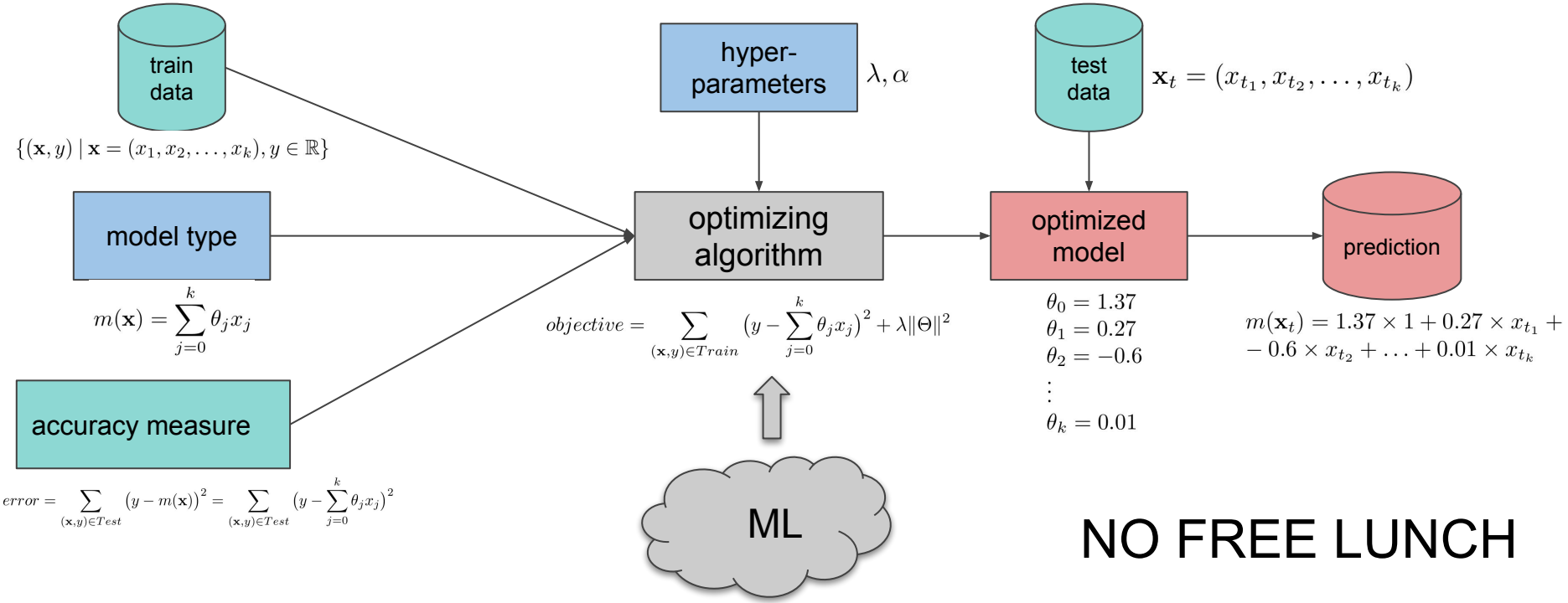


Business Under-standing → Data Understanding → Data Preparation → Modeling → Evaluation → Deploy-meny

these are, usually, quite time-consuming

Thinking about a Machine Learning Task...



NO FREE LUNCH

Question: Do we use the right model?

“Of course, it’s cool and everyone else is using it.”

- hackathons became bit boring for me, lately...

How does a data scientist choose the model?

- based on knowledge and past experience

Metalearning

Let's **focus on the data** instead of the model and ask some questions

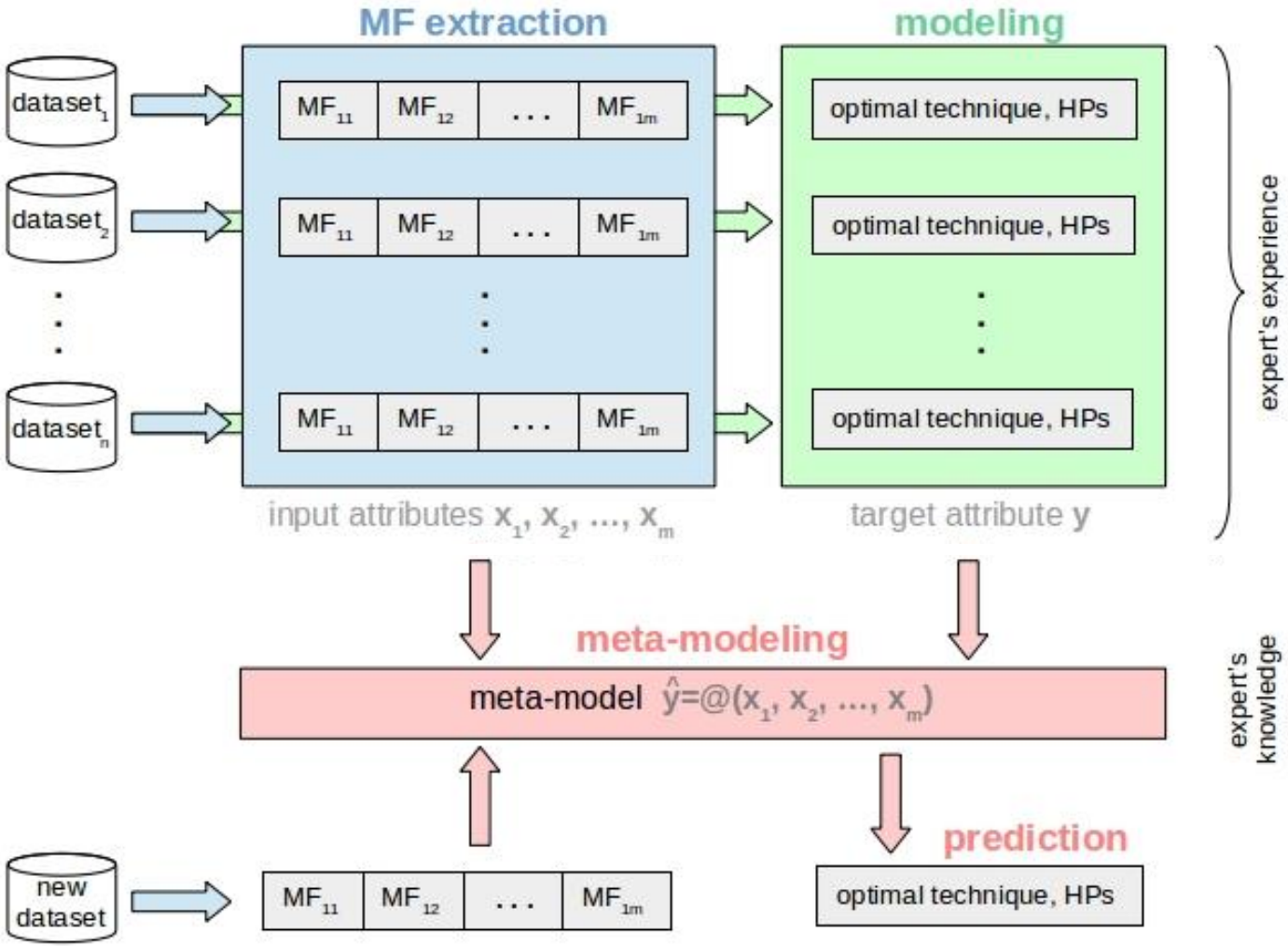
- How would a certain model perform on the data?
- What is the expected runtime of tuning of that model on the data?
- Do the hyper-parameters of that model need tuning on the data or using the deafult setting is ok?

Expert's knowledge

- performance limits of models
- time and space complexity of models
- interpretability of models

Expert's experience

- recorded choices and their outcomes during previous projects


MF extraction
modeling
dataset1
dataset2
datasetn
MF11 MF12 . . . MF1m
optimal technique, HPs
expert's experience
input attributes x1, x2, ..., xm
target attribute y
meta-modeling
meta-model ŷ=@(x1, x2, ..., xm)
expert's knowledge
new dataset
prediction

Metafeatures

Table 1: The main MF types, and their abbreviations (Abbr), identified in the literature and used in the experiments carried out in tis study.

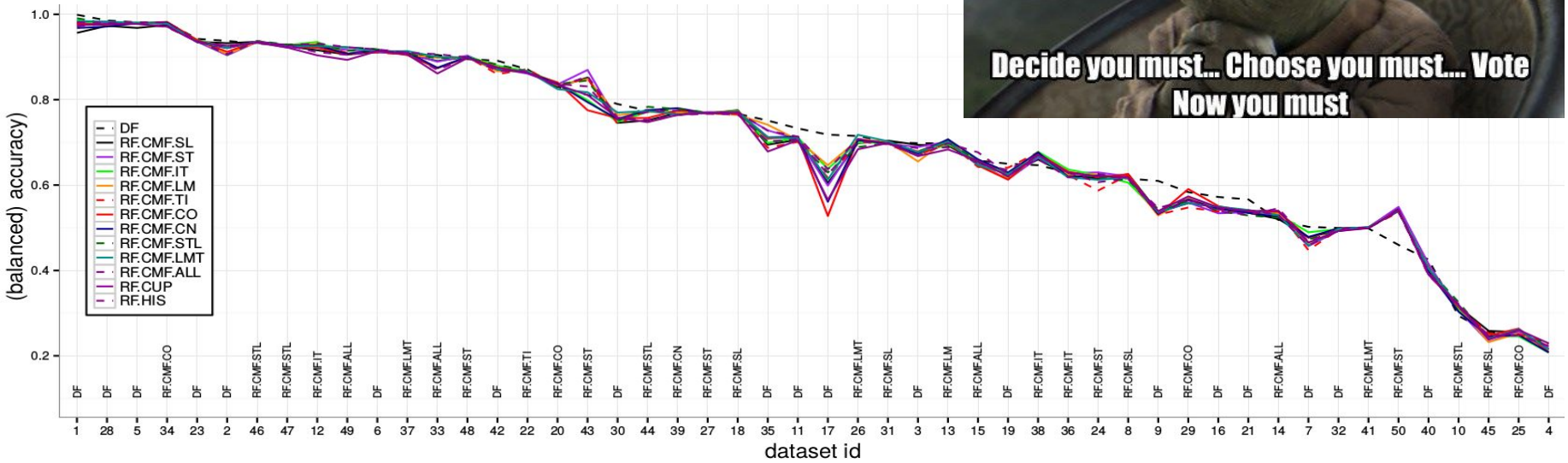
MF type	Abbr	#	Description
Simple	SL	17	Simple measures
Statistical	ST	7	Statistics measures
Inf. theoretic	IT	8	Information theory measures
Landmarking	LM	9	Performance of some ML algorithms
Model-based	MB	17	Features extracted from decision trees
Time	TI	5	Execution time of some ML algorithms
Complexity	CO	14	Measures analyzing complexity
Complex Network	CN	9	Complex network property measures

Table 3: MF used in the experiments.

Type	MF
SL	# classes, # attributes, # numeric attributes, # nominal attributes, # samples, dimensionality, % numeric attributes, % nominal attributes, min # levels, max # levels, mean # levels, std # levels, sum # levels, % samples in minority class, % samples in majority class, avg # samples by class, std samples by class
ST	skewness, skewness of pre-processed data, kurtosis, kurtosis of pre-processed data, absolute correlation, canonical correlation, fraction of canonical correlation
IT	class entropy, normalized class entropy, attribute entropy, normalized attribute entropy, joint entropy, mutual information, equivalent attributes, noise-to-signal ratio
LM	Näive Bayes performance, LDA performance, Decision Stump min performance, Decision Stump max performance, Decision Stump mean performance, std of Decision Stump performance, Decision Stump min gain ratio, Decision Stump random performance, 1-NN performance, std of 1NN performance
MB	# nodes, # leaves, nodes per attribute, nodes per instance, leaf corroboration, min # levels, max # levels, mean # levels, std # levels, min # branches, max # branches, mean # branches, std # branches, min # attributes, max # attributes, mean # attributes, std # attributes
TI	Decision Tree execution time, Näive Bayes execution time, LDA execution time, Decision Stump execution time, 1-NN execution time
CO	Fisher's discriminant ratio, Directional-vector Fisher's discriminant ratio, overlap of per-class bounding boxes, max individual feature efficiency, collective feature efficiency, L1 error of a linear classifier, L2 error of a linear classifier, non-linearity of a linear classifier, % points lying on decision boundary, avg intra/inter class NN distances, leave-one-out error, rate of 1-NN, non-linearity of 1-NN, % max covering spheres on data
CN	# edges, avg degree of the network (degree), density, max component, closeness, betweeness, clustering coefficient (clsCoef), # hubs, avg path

Case study: Which metafeatures to use?

- each MF group has some pros and cons
- choice might be data dependent



Question: Do we use the model with optimal setup?

“Think so, we always did it in this way.”

- or *“We have used the default settings, they are pretty good.”*

Hyper-parameter tuning

- usually underrated
- lack of HP tuning, joined with poor data and domain understanding and/or bad data pre-processing, can result in very “interesting” models

Popular choices

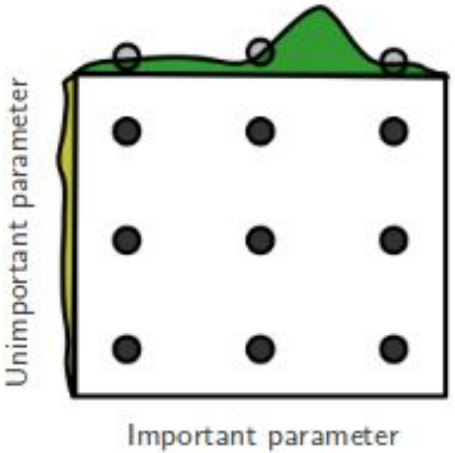
Grid search and **Random search**

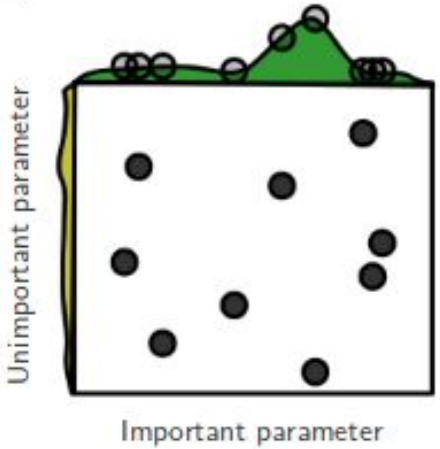
Easy to implement and understand

Work pretty well in case of

- small number of hyper-parameters
- many local optima of HP settings
- larger computational budgets and fast learning algorithms

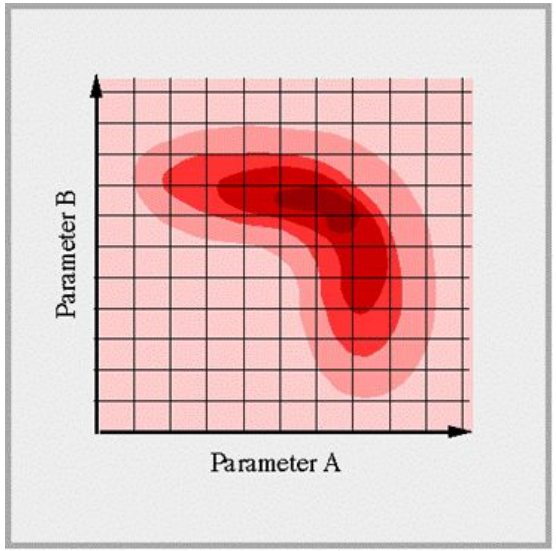
However, there are more sophisticated approaches



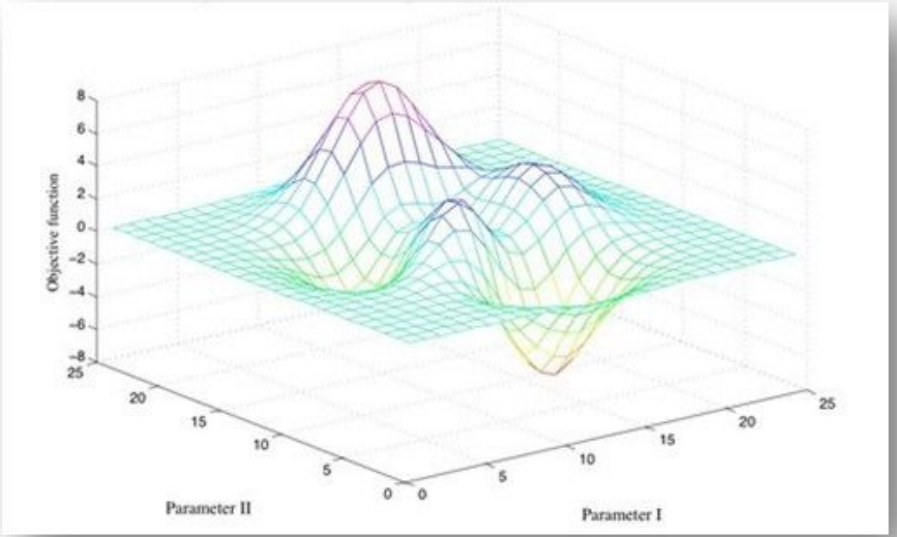


Hyper-parameter landscape

Some nice



Some not so pretty

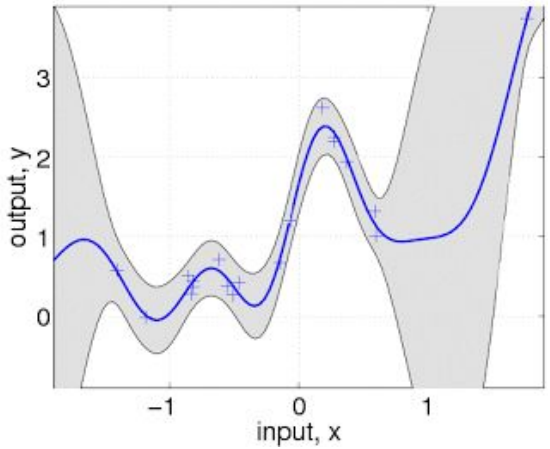


How often do we “dig into the characteristics” of the landscape during ML?

Black-box optimization

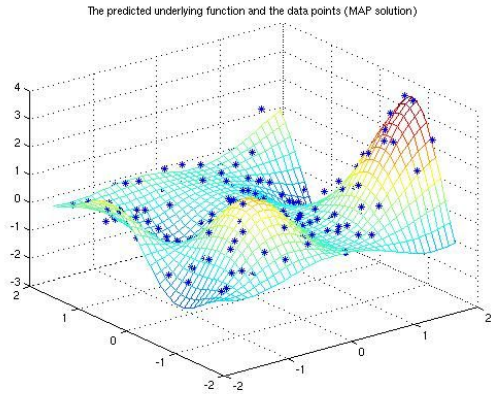
Sequential Model-based Optimization (SMBO)

- popular approach
- implemented in Auto-Weka, MLR, scikit-learn
- time consuming for larger number of HPs



Population-based approaches

- Evolutionary Algorithms (EA)
- Particle Swarm Optimization (PSO)
- Estimation of Distribution Algorithm (EDA)
- Iterated F-Race (Irace)
- ...



Case study: Tuning HPs of DT learners

J48 (9 HPs)

- 1 real, 2 integer, 6 boolean
- dependencies

CART (6 HPs)

- 1 real, 2 integer, 1 nominal, 1 boolean

CTree (6 HPs)

- 2 real, 3 integer, 1 boolean

94 public datasets

- binary and multi-class classification
- number of attributes from 3 to 1300
- number of instances from 100 to 45000

6 HP tuning techniques

- SMBO, Irace, PSO, EDA, EA, RS
- 30 repetitions, budget 900

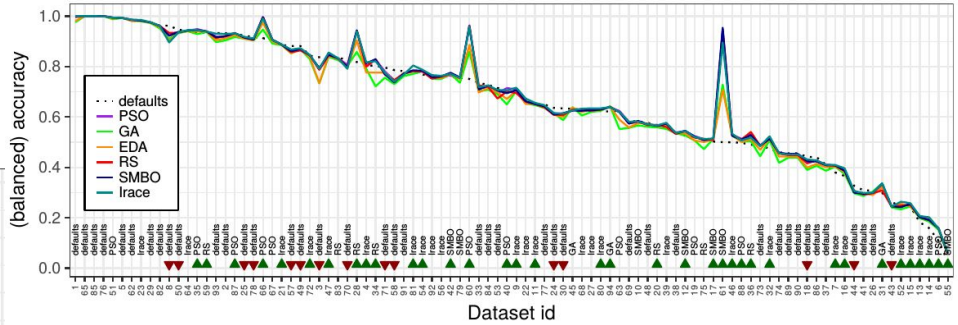
Nested cross-validation

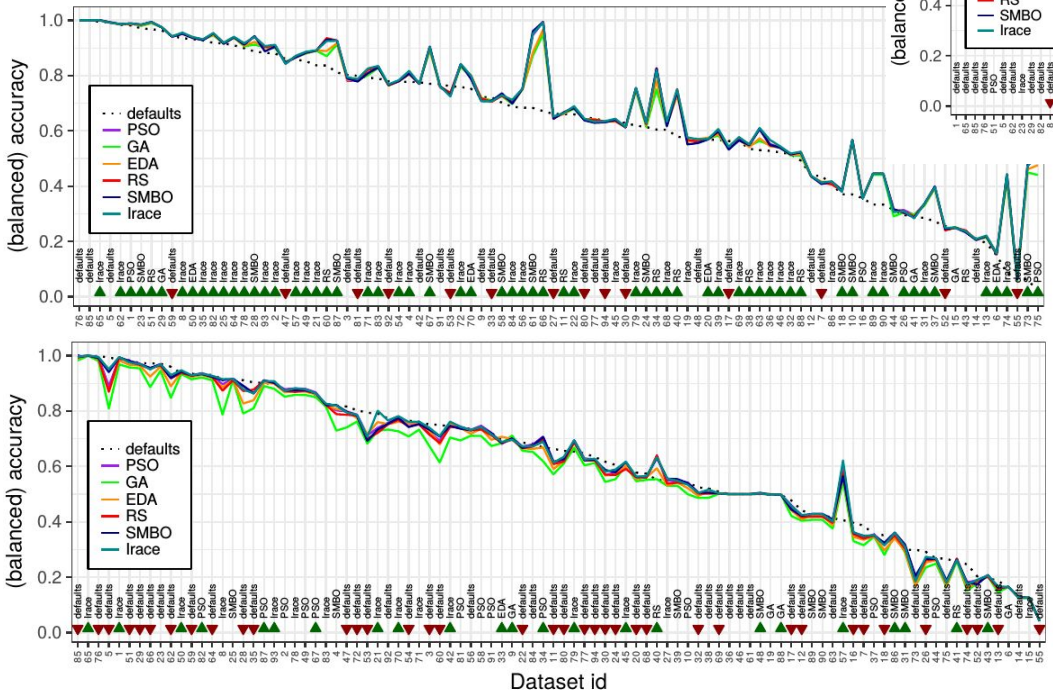
- 3 inner folds, 10 outer folds

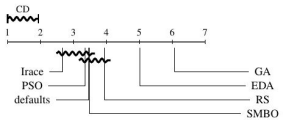
Around 1.3 billion learning runs

Which HP tuning approach to choose?

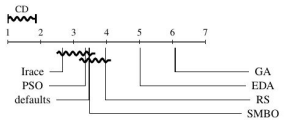
- each has some pros and cons
- choice might be data dependent



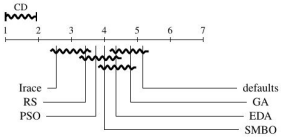




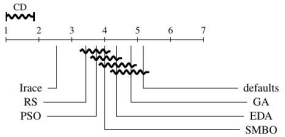
(a) J48 CD diagram with $\alpha = 0.05$.



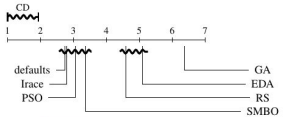
(b) J48 CD diagram with $\alpha = 0.1$.



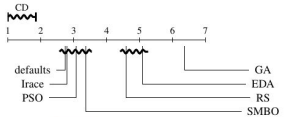
(c) CART CD diagram with $\alpha = 0.05$.



(d) CART CD diagram with $\alpha = 0.1$.

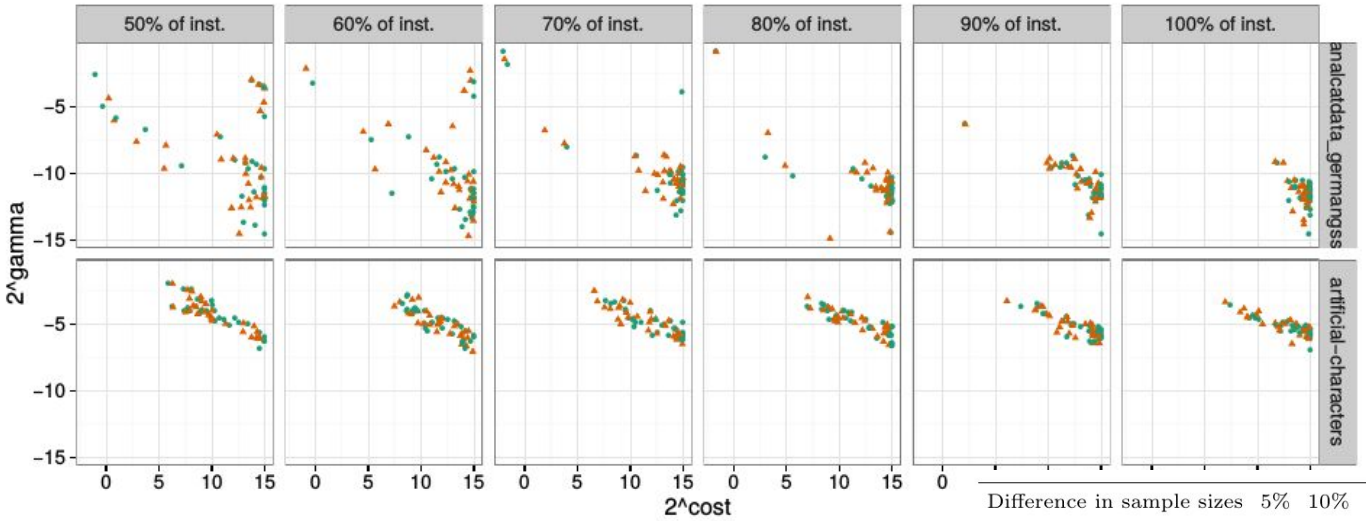


(e) CTree CD diagram with $\alpha = 0.05$.



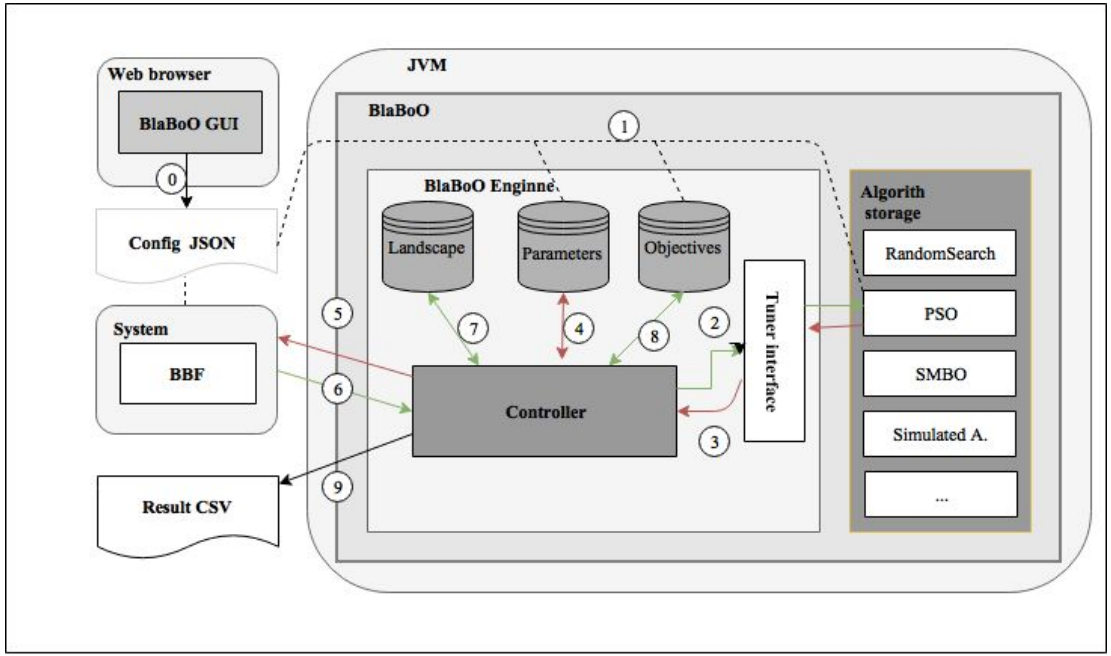
(f) CTree CD diagram with $\alpha = 0.1$.

What about sampling?



Difference in sample sizes	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
5%	–					●	●	●	●	●
10%		–				○	○	●	●	●
15%			–			○		○	●	●
20%				–					○	●
25%					–				○	●
30%	●	○				–				
35%	○	○					–			
40%	●	●	●	○				–		
45%	●	●	●	●	●	○			–	
50%	●	●	●	●	●	○				–

A student project: BlaBoO



- **easy** to install
- **Java-based**, i.e. platform independent
- various **types** of black-box optimizers
- **GUI** & **command line** mode
- easily **extendable**

https://github.com/kppeterkiss/BlackBoxOptimizer

Final remarks

We should choose the right model based on the data at hand.

Kind of a “no free lunch” in case of the HP tuning algorithms.

Default HP settings are often working quite well but we should always consider HP tuning (even if only on a limited budget).

Learners for meta-models as well as black-box optimizers have also their own hyperparameters ;)

It is not so hard to implement your own metalearning approach.

Be careful with “shiny” results in scientific publications or frameworks’ PR.

Final question: Can be the job of a data scientist fully automated?

The challenge is accepted, the quest is on the move, but we are not there yet.

Thanks for your attention!

Also, thanks for the creators of used images.

`http://t-labs.elte.hu/`