Explainable AI:
Intelligent Data Preprocessing and
White Box Modeling by Symbolic Regression

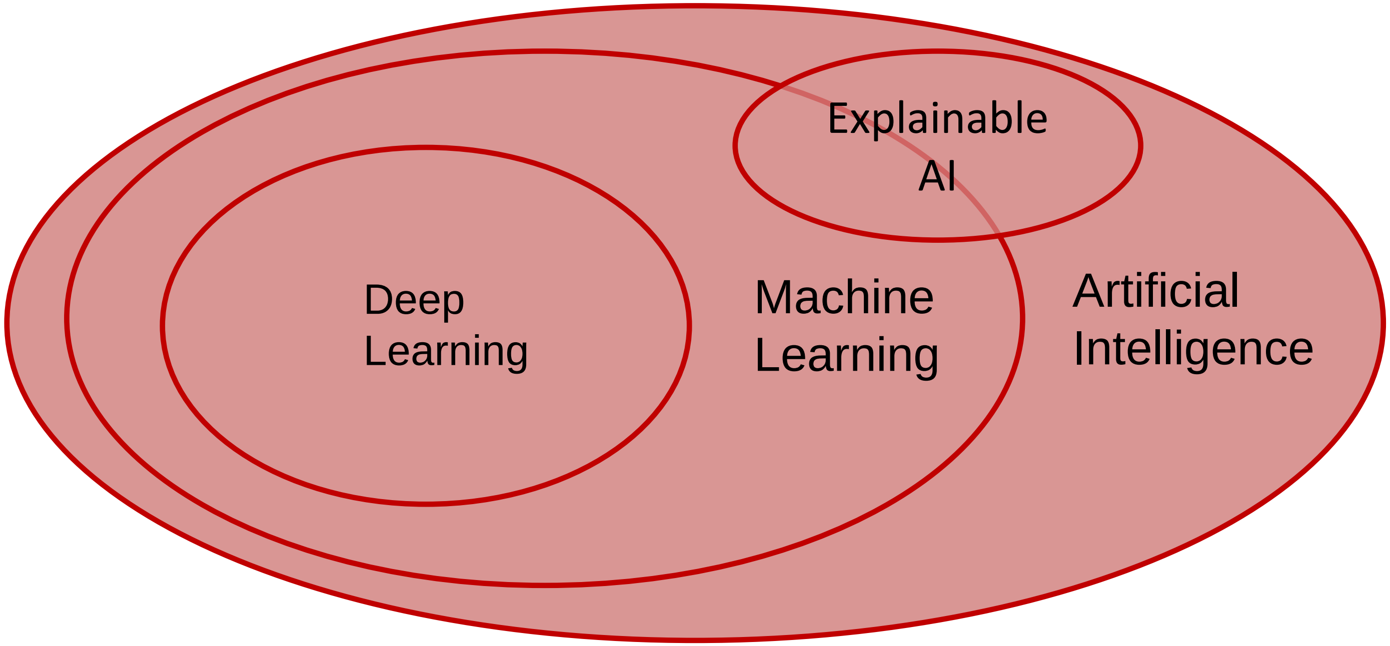
Susanne Schaller, Stephan M. Winkler
FH Oberösterreich, Campus Hagenberg
Softwarepark Hagenberg

HAGENBERG | LINZ | STEYR | WELS


X
AI

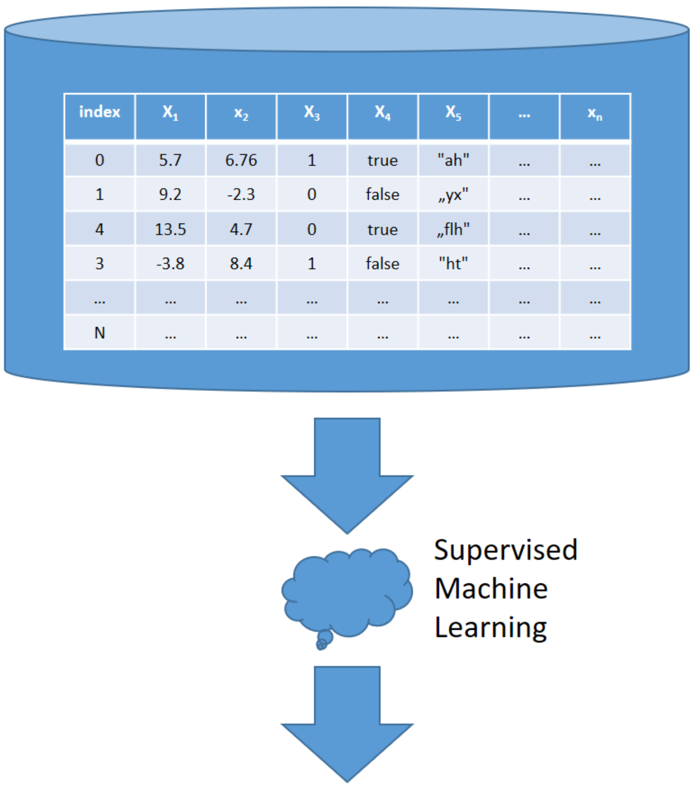
Explainable AI

Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the **results can be understood** by human experts. It contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision.



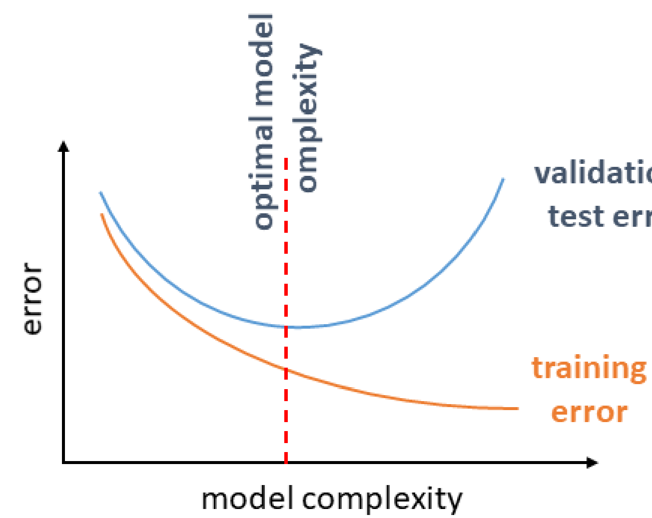
Systems Identification and Machine Learning

– Systems identification
– Modeling
– Process analysis
– Prediction
– Optimization



index	X_1	x_2	X_3	X_4	X_5	...	x_n
0	5.7	6.76	1	true	"ah"	...	...
1	9.2	-2.3	0	false	„yx"	...	...
4	13.5	4.7	0	true	„flh"	...	...
3	-3.8	8.4	1	false	"ht"	...	...
...	...	...	...	...	...	...	...
N	...	...	...	...	...	...	...

$y_1 = f_1(x_1, x_2, ..., x_n)$
$y_2 = f_2(x_1, x_2, ..., x_n)$ Models
$y_3 = f_3(x_1, x_2, ..., x_n)$



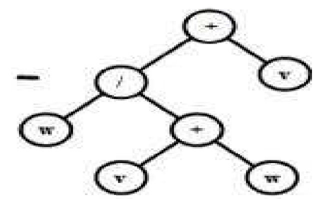
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

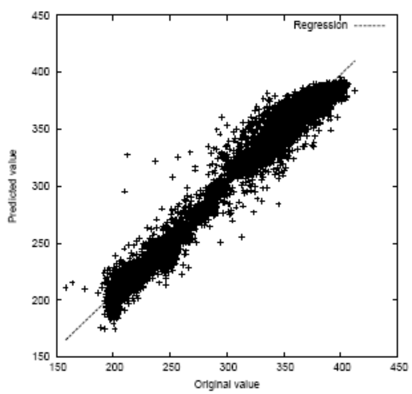
Black-Box vs. White-Box Modeling

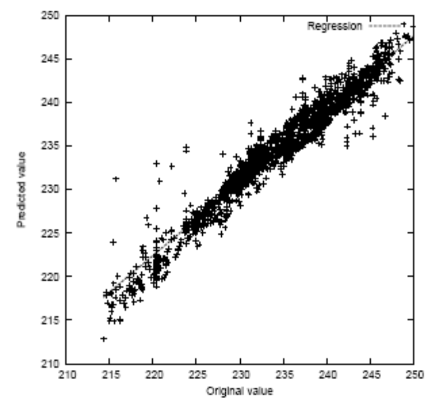
Black Box Model

?

White Box Model



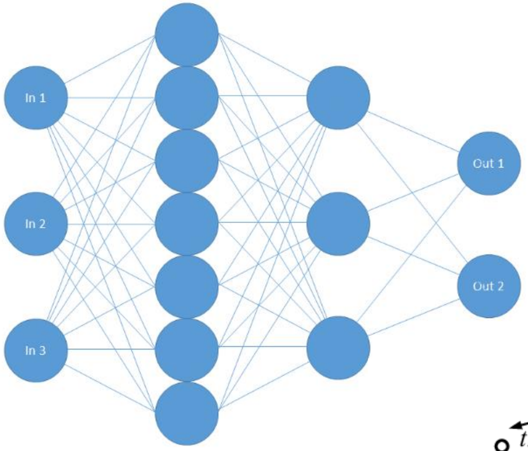


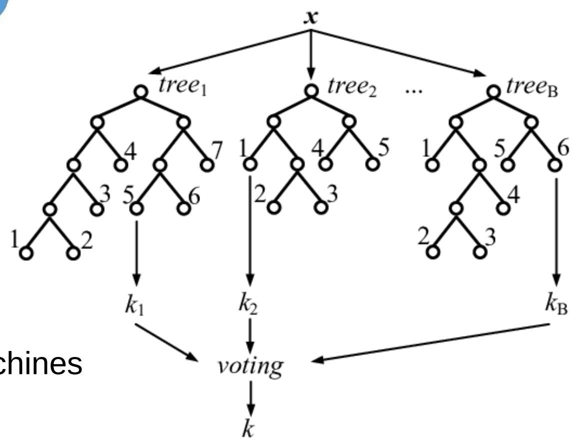


UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Black-Box vs. White-Box Modeling

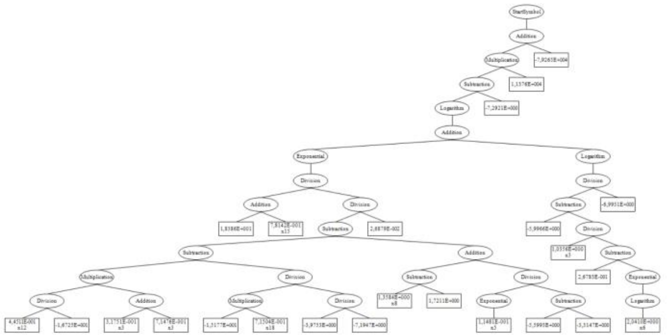
Black Box





- random forests
- support vector machines
- neural networks

White Box



$$y = \left(\left(\left(\left(\left(c_0 \cdot \mathbf{x31} + \left(\frac{(c_1 \cdot \mathbf{x22} + c_2 \cdot \mathbf{x17})}{c_3} - (c_4 \cdot c_5 + c_6 \cdot \mathbf{x49})\right)\right)\right) + (c_7 + c_8) \cdot \frac{c_9}{c_{10}}\right) + c_{11} \cdot \mathbf{x14}\right) \cdot c_{12} + c_{13}\right)$$

$$\begin{aligned} c_0 &= -0.15366 & c_5 &= -1.8194 & c_{10} &= -11.55 \\ c_1 &= 1.4153 & c_6 &= 1.3732 & c_{11} &= 1.353 \\ c_2 &= -0.090437 & c_7 &= -5.5435 & c_{12} &= -0.030191 \\ c_3 &= 18.821 & c_8 &= 4.6554 & c_{13} &= -5.7891 \\ c_4 &= -9.0013 & c_9 &= -11.338 & & \end{aligned}$$

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

White-Box Modeling

– **Genetic Programming**

› evolutionary process

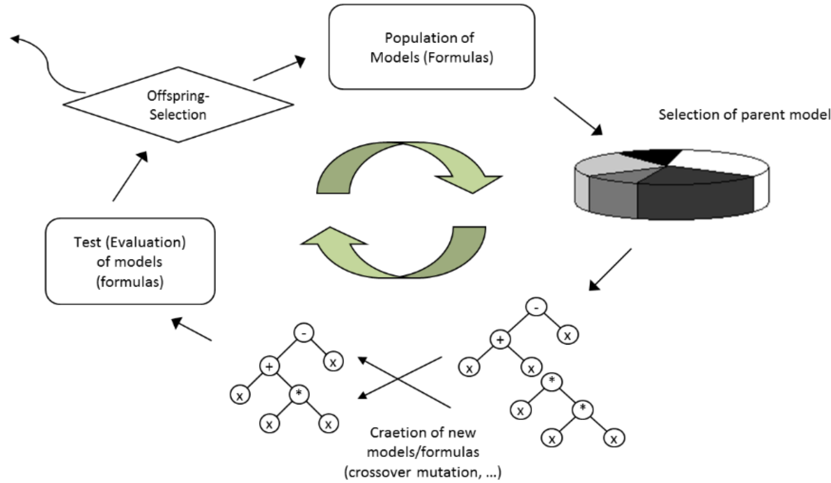
› implicit feature selection

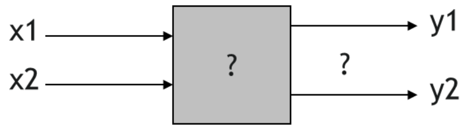
› optimizes model structure and parameters

› generates interpretable formulas

› results directly applicable

› assessment of variable relevance





y1(t+1) = y2(t) / (x1+x2)
y2(t) = x1(t)+3*exp(x2(t-2)) ?

$$y = c_0 \cdot x_1 \cdot (\log(c_1 \cdot x_2) + c_2) + c_3$$
$$c_0 = 0.500331886126962$$
$$c_1 = 2.3702293890766$$
$$c_2 = 1.28570833399083$$
$$c_3 = 4.91856540837157$$

White-Box Modeling

– **Genetic Programming**

› evolutionary process

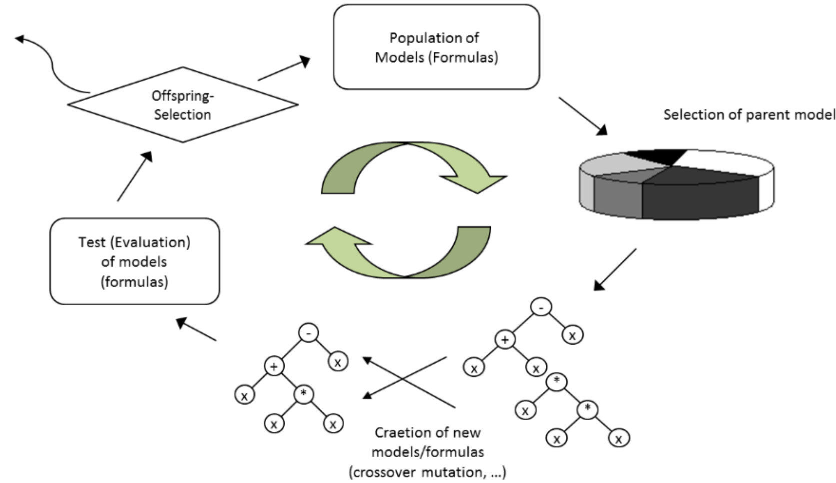
› implicit feature selection

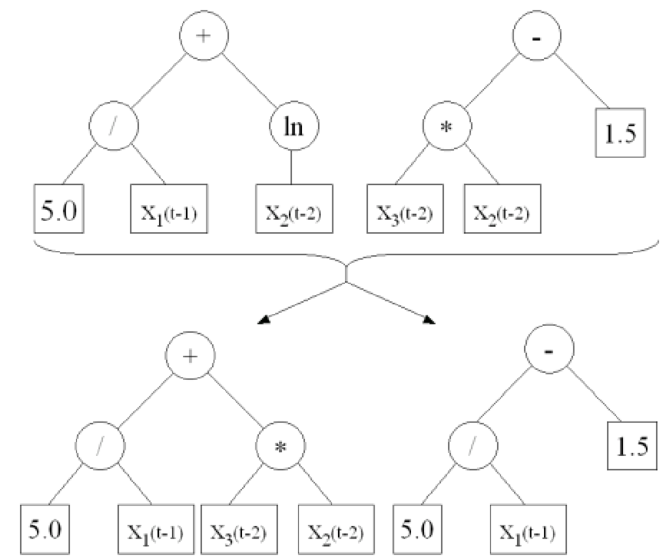
› optimizes model structure and parameters

› generates interpretable formulas

› results directly applicable

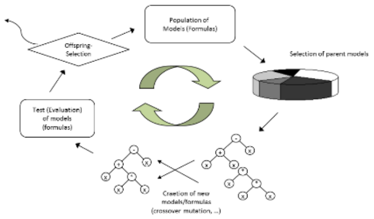
› assessment of variable relevance

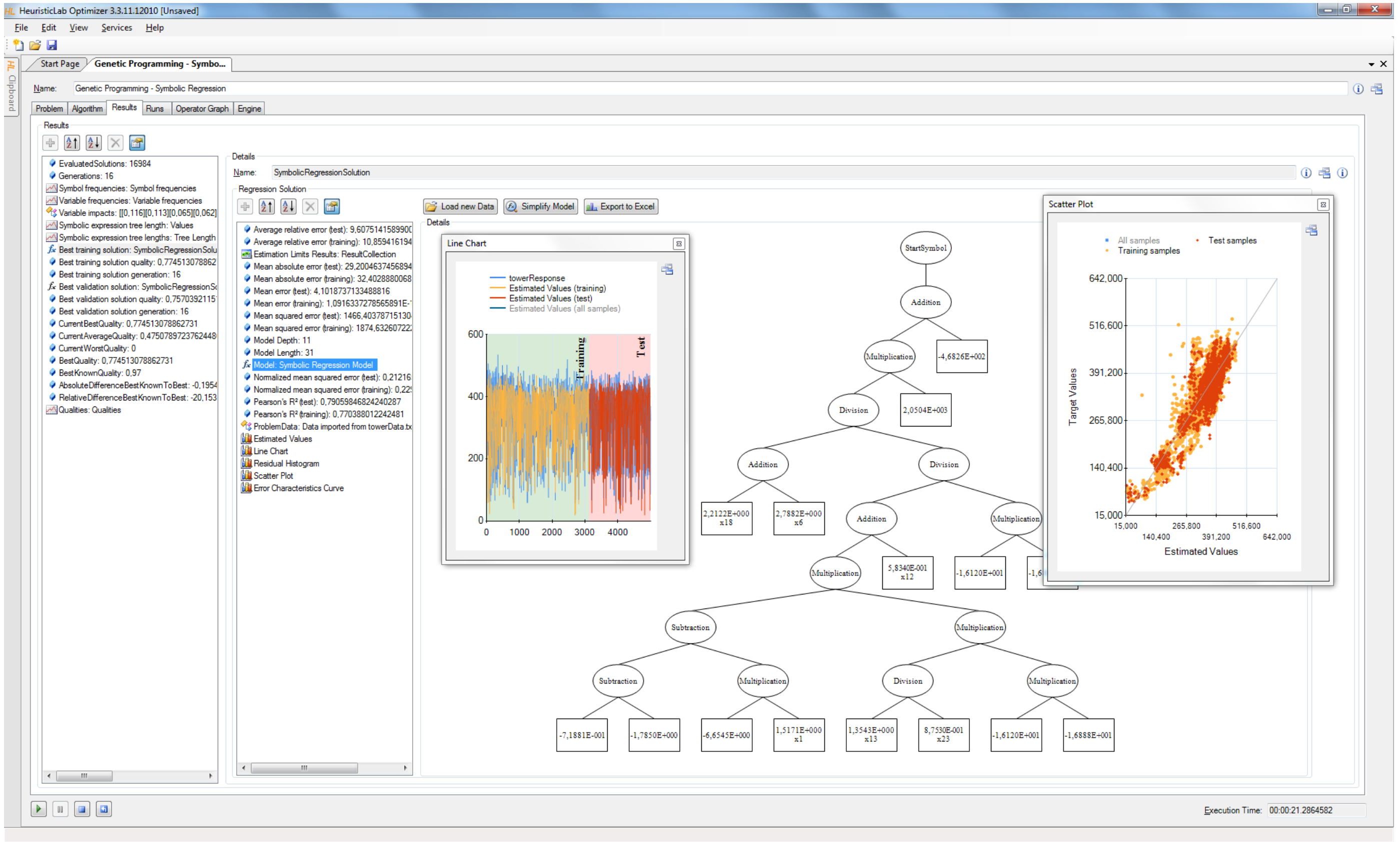




UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

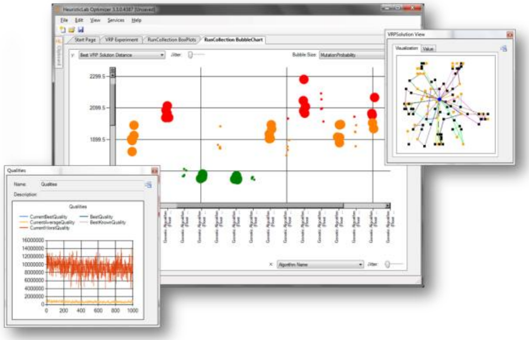
White-Box Modeling: Symbolic Regression in HeuristicLab





UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

HeuristicLab



Open Source Optimization Environment HeuristicLab

- developed since 2002
- basis of many research projects and publications
- 2nd place at *Microsoft Innovation Award 2009*
- HeuristicLab 3.3.x since May 2010 under GNU GPL

Motivation and Goals

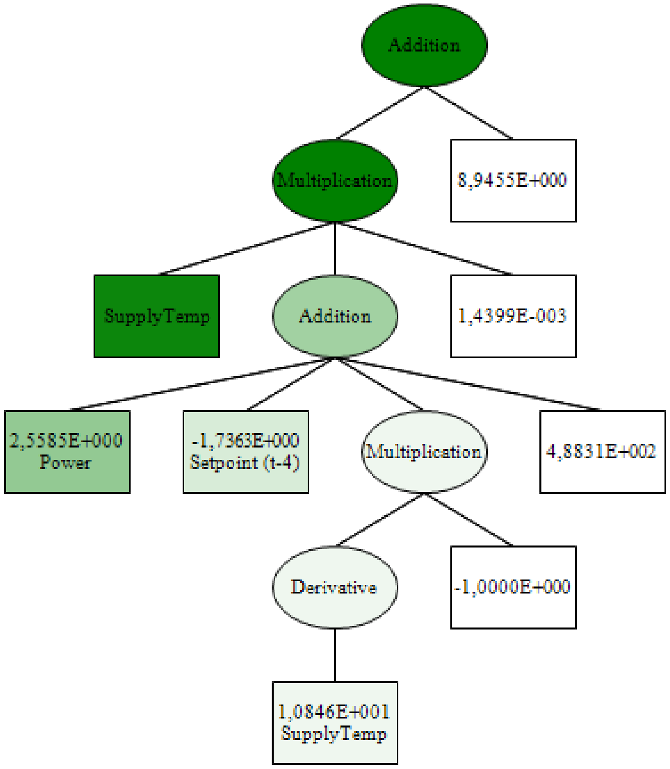
- graphical user interface for interactive development, analysis and application of optimizations methods
- numerous optimization algorithms and optimization problems
- support for extensive experiments and analysis
- distribution through parallel execution of algorithms
- extensibility and flexibility (plug-in architecture)

Distributed Computing with HeuristicLab Hive

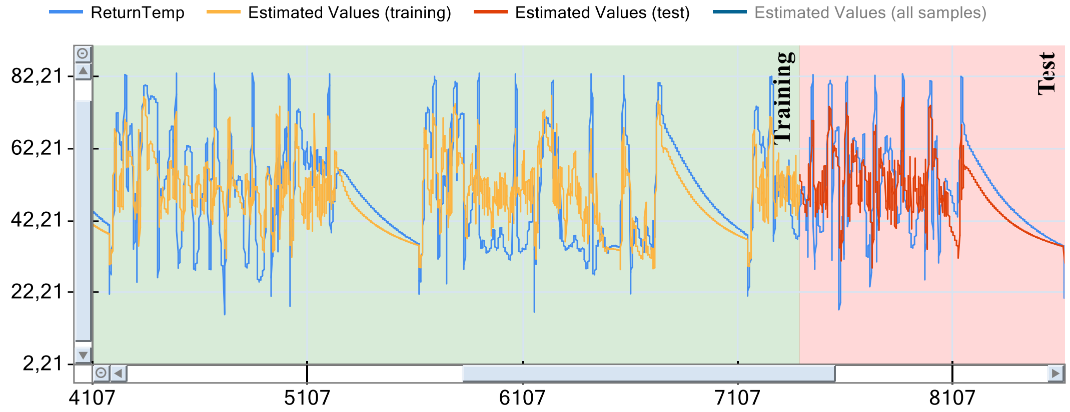
- framework for distribution and parallel execution of HeuristicLab algorithms
- compute resources at Campus Hagenberg
  - 2006 – 2011: research cluster 1 (14 cores)
  - since 2009: research cluster 2 (112 cores, 448GB RAM)
  - since 2011: lab computers (100 PCs, on demand in the night)
  - since 2017: research cluster 3 (448 cores, 4TB RAM)

White-Box Modeling

$$\text{SupplyTemp} \cdot \left(c_0 \cdot \text{Power} + c_1 \cdot \text{Setpoint}(t-4) + \frac{d(c_2 \cdot \text{SupplyTemp})}{dt} \cdot c_3 + c_4 \right) \cdot c_5 + c_6$$



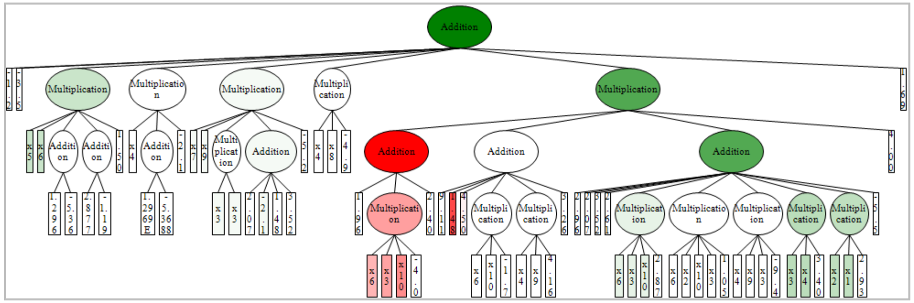
$$\begin{aligned}
c_0 &= 2.5585 \\
c_1 &= -1.7363 \\
c_2 &= 10.846 \\
c_3 &= -1.0 \\
c_4 &= 488.31 \\
c_5 &= 0.0014399 \\
c_6 &= 8.9455
\end{aligned}$$

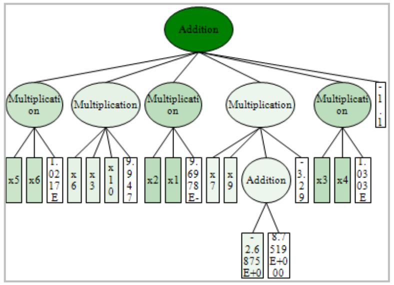


UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Model Simplification

- Simplification Methods
  - mathematical transformation
  - remove nodes
  - constant optimization
  - external optimization



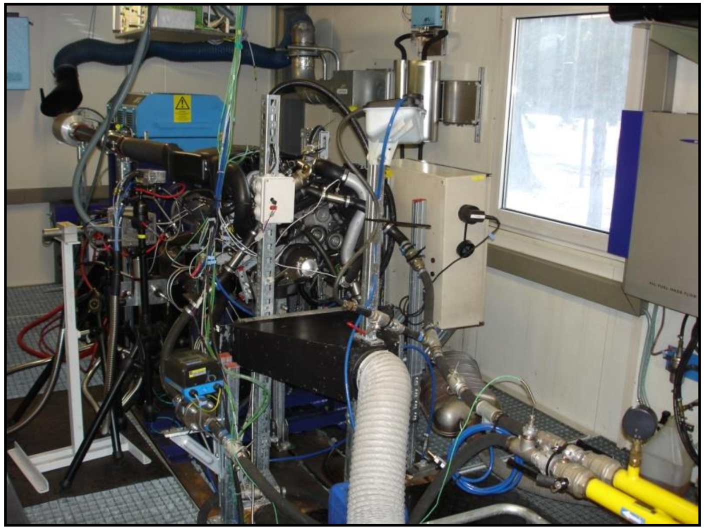


- Export
  - textual export
  - LaTeX, MATLAB
  - graphical export

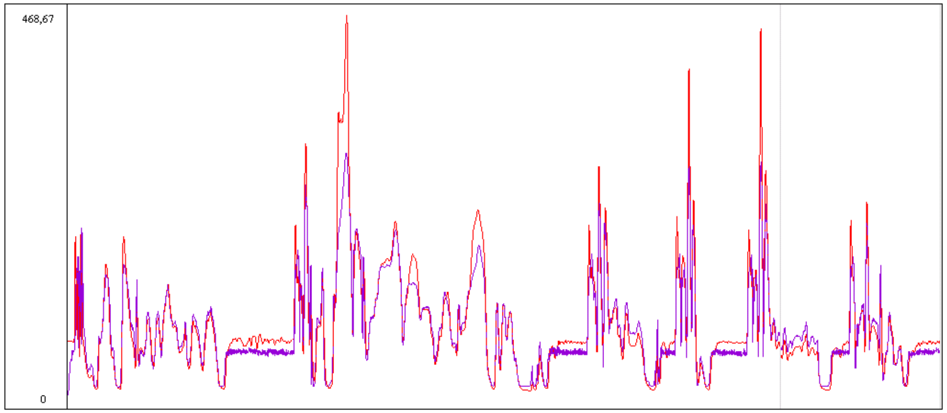
$$y = x_1 \cdot x_2 + x_3 \cdot x_4 + x_5 \cdot x_6 + x_1 \cdot x_7 \cdot x_9 + x_3 \cdot x_6 \cdot x_{10}$$

Example: Virtual Sensors for Modeling Exhaust Gases

- high quality modeling of emissions (NOx and soot) of a diesel engine
- virtual sensors: (mathematical) models that mimic the behavior of physical sensors
- advantages: low cost and non-intrusive
- identify variable impacts: injected fuel, engine frequency, manifold air pressure, concentration of O2 in exhaustion etc.

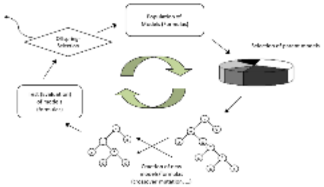


$NO_x(t) = f(x1_{(t-7)}, x2_{(t-2)}, \ldots)$



UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Example: Virtual Sensors for Modeling Exhaust Gases



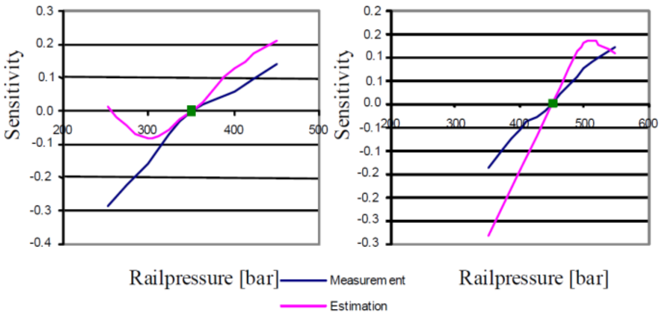


Figure 2: Poor performance of the same neural network model at other operating points

$$\left[NO_x^*(t)\right] = \frac{2.696 m_f^*(t-10) + 2.618 m_f^*(t-7)}{\log\left(0.029 N^*(t-10)\right)} + \frac{\left[1.798 m_f^*(t-5)\right]\left[7.536 W^*(t-5)\right]}{\left[0.027 N^*(t-9)\right]\log\left(0.031 N^*(t-3)\right)}$$

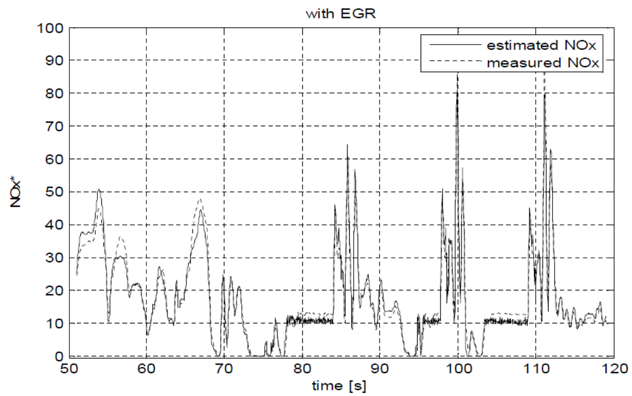
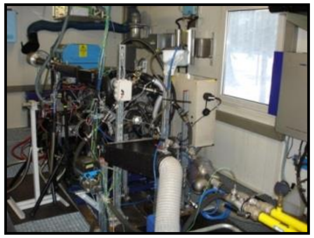
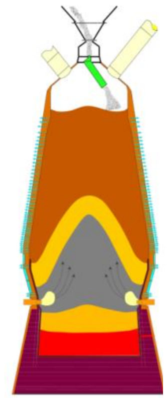


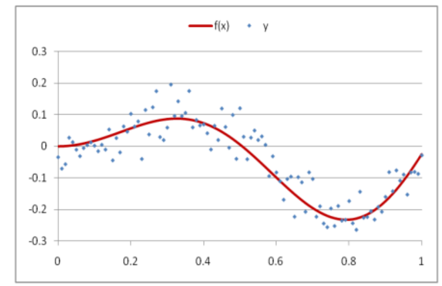
Figure 6: Estimated and Measured NOx (System with EGR)



Example: Blast Furnace Modeling

x1	x2	x3	x4	x5	y
28.07845	13.93902	87.63394	20.07777	63.00267	250.4028
27.95657	12.75236	87.05083	19.95878	63.00894	440.0825
25.43135	23.03532	88.32881	21.98374	74.99575	292.6644
28.5034	36.71041	87.59461	20.55528	75.01106	100.8683
23.03413	46.5804	79.38985	18.67402	80.31421	435.7738
20.97957	41.52231	73.32074	21.49193	79.98517	288.5032
28.07431	28.49076	106.4166	27.38095	79.97826	?
28.00494	36.33813	104.7173	27.99428	75.00266	?
28.0274	31.84306	102.277	28.81878	78.1752	?
26.503	27.67078	93.81539	21.29002	62.99904	?
23.869	27.25298	93.67531	24.54099	80.00291	?

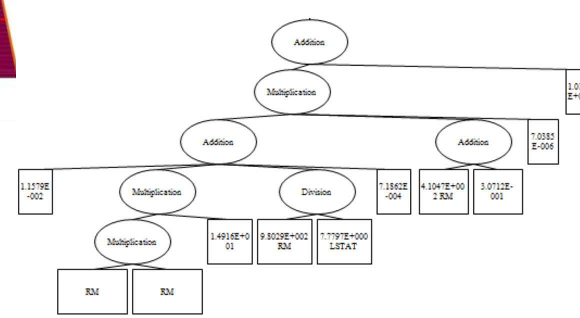




f(x)

Model

Prognosis



- results as formulas → domain experts can analyze, simplify and refine the models
- integration of prior physical knowledge into modeling process
- powerful data analysis tools: model simplification and variable impact analysis

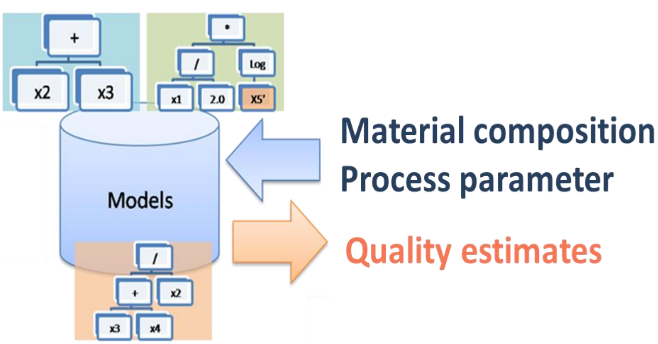
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Example: Plasma Nitriding Modeling

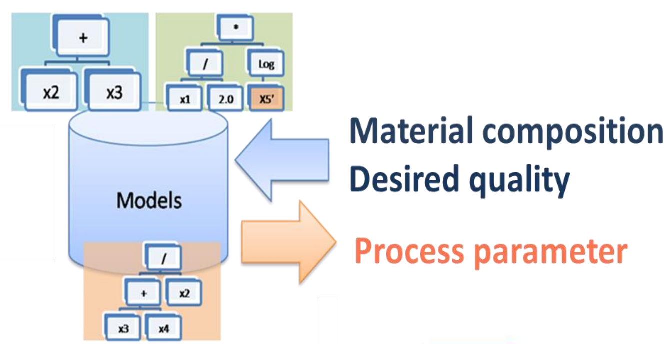
- Motivation
  - hardening of materials (e.g. transmission parts)
  - process parameter settings based on expert knowledge

- Modeling Scenarios
  a) prediction of quality values based on process parameters and material composition
  b) propose process parameter settings to reach the desired material characteristics

a)

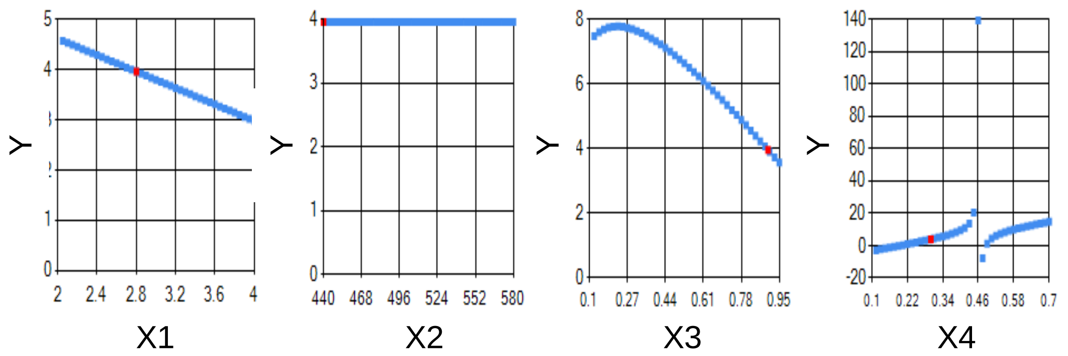


b)



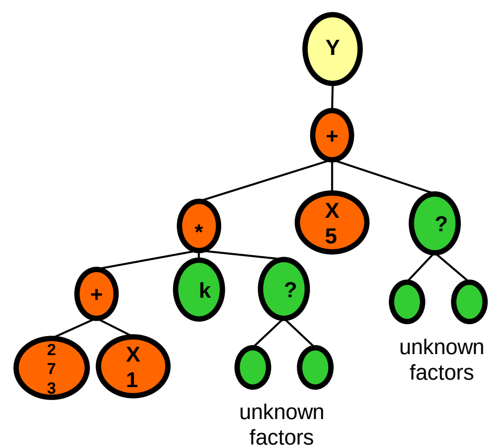
Integration of Expert Knowledge

Model Analysis



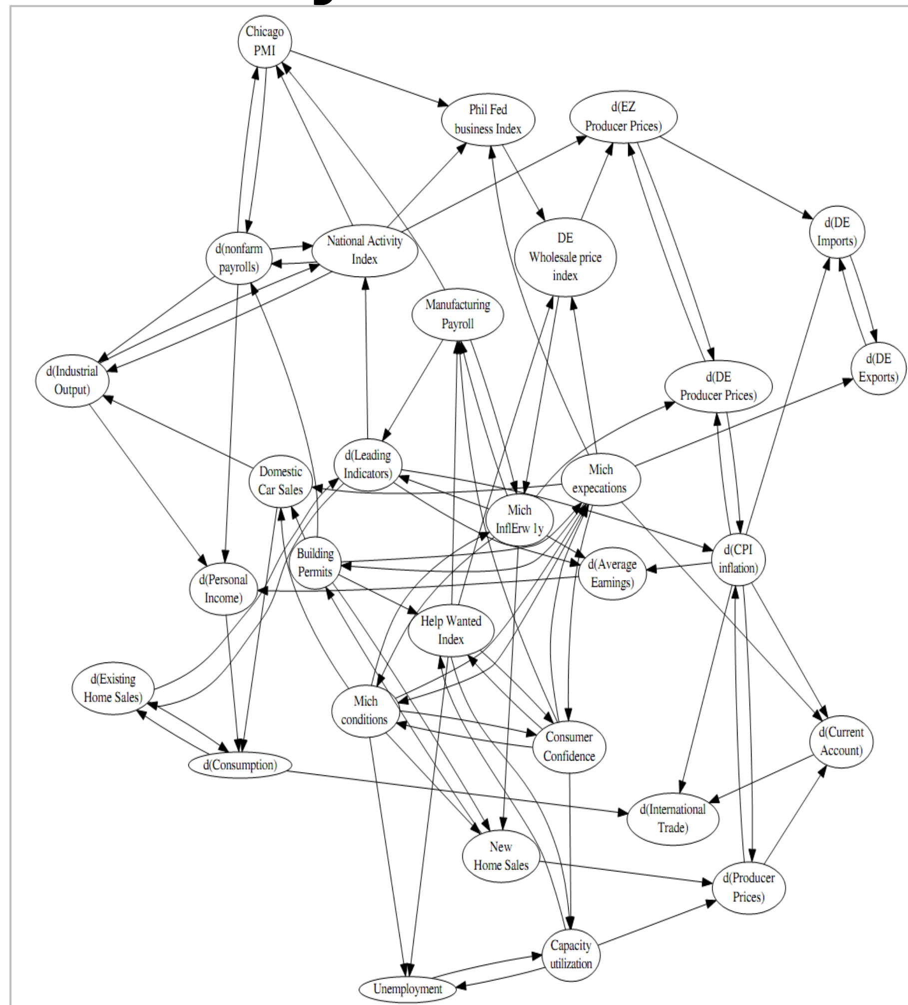
Knowledge Integration

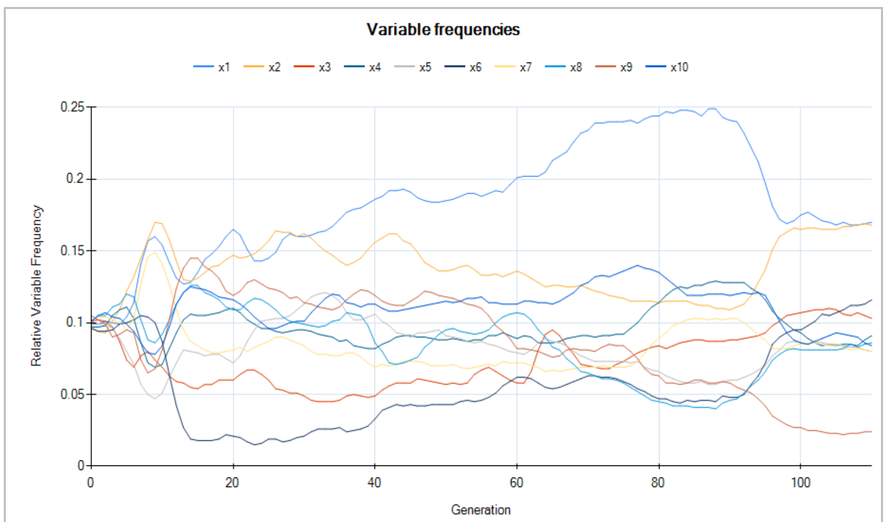
- specification of known correlations
- model extension through algorithm



Holistic Knowledge Discovery

- Variable interaction networks
  - reveals non-linear correlations
- Variable frequencies
  - analyzed during the algorithm run





Acknowledgements

Bioinformatics Research Group

Heuristic and Evolutionary Algorithms Laboratory

http://bioinformatics.fh-hagenberg.at/

https://heal.heuristiclab.com/