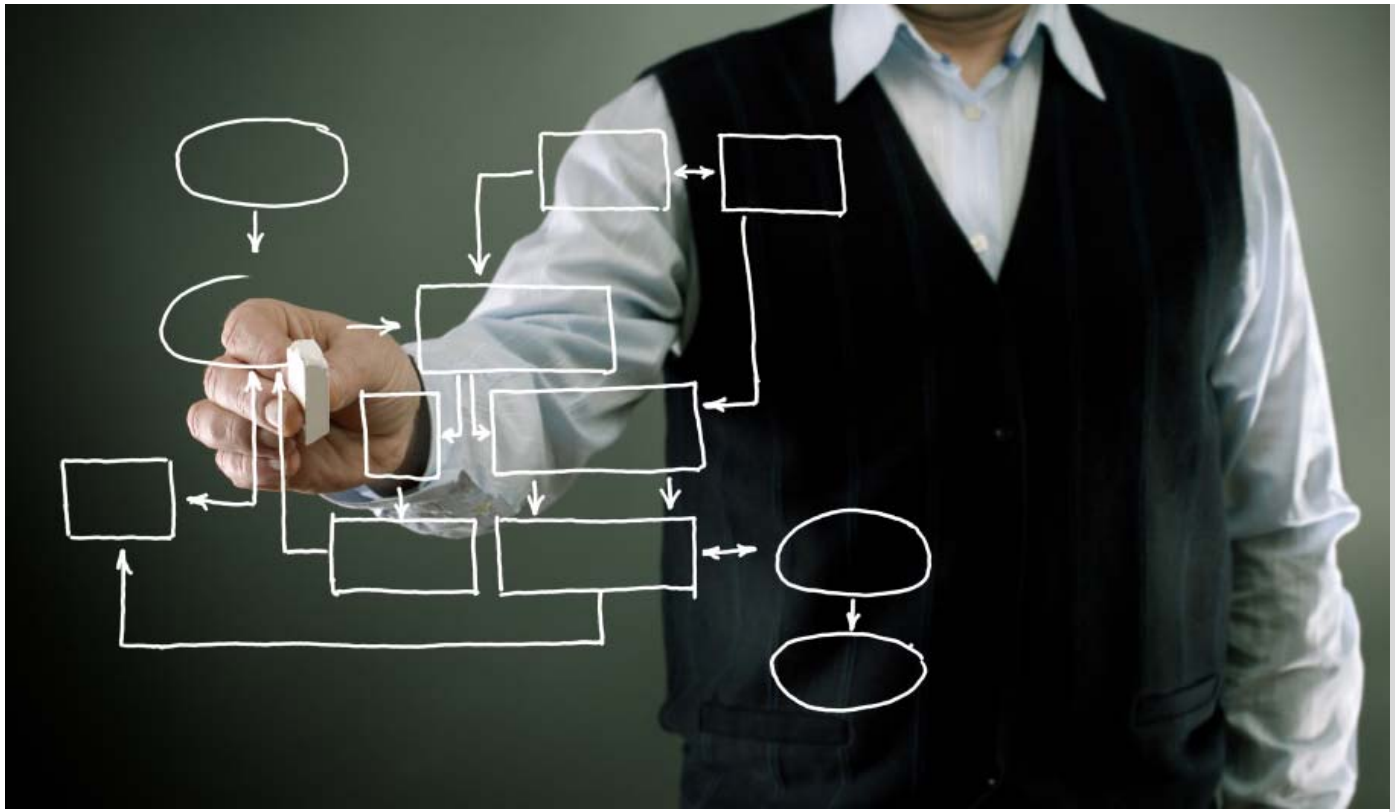


Gaining Deeper Insights in Symbolic Regression: Theoretical and Practical Issues

Michael Affenzeller

Head of research group HEAL

HEAL

HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

Contact:
Dr. Michael Affenzeller
FH OOE - School of Informatics, Communications and Media
Heuristic and Evolutionary Algorithms Lab (HEAL)
Softwarepark 11, A-4232 Hagenberg

e-mail:
michael.affenzeller@fh-hagenberg.at
Web:
http://heal.heuristiclab.com
http://heureka.heuristiclab.com

Content

- Softwarepark Hagenberg, Research group HEAL
- Research focus and infrastructure
- Analysis of algorithm dynamics
- Enhanced interpretability of symbolic regression models
- Applications

Softwarepark Hagenberg
Research group HEAL

Campus Hagenberg

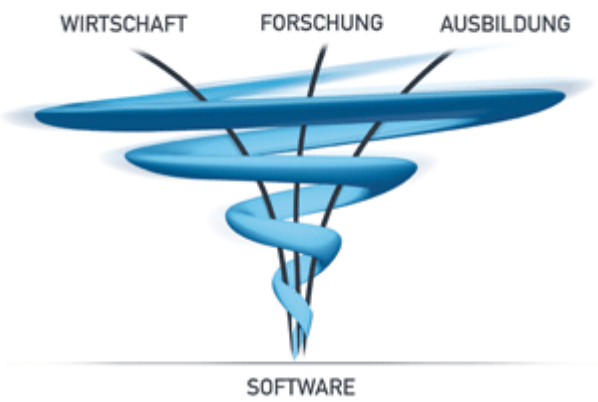
- Communication, Software, and Media
- 1400 students
- Research Center for Software Technologies and Applications

Softwarepark Hagenberg

- Education, Research, Economy
- Center of Software Competence at Hagenberg
- Several companies connected to IT-business

Softwarepark Hagenberg

- 1989: RISC moves from Linz to Hagenberg
- Since then: Development of the Softwarepark Hagenberg now: 800 employed, 1400 students
- 1993: Start of Upper Austria University of Applied Sciences, School of Informatics, Communication, and Media
- Center of Software Competence at Hagenberg
- Several companies connected to IT-business



Heuristic and Evolutionary Algorithms Research Group

Research Group

- 4 professors
- 7 PhD students
- Various interns, Master and Bachelor students

Research Focus

- Problem modeling
- Process optimization
- Data-based structure identification
- Supply chain and logistics optimization
- Algorithm development and analysis

Industry Partners (excerpt)

Scientific Partners

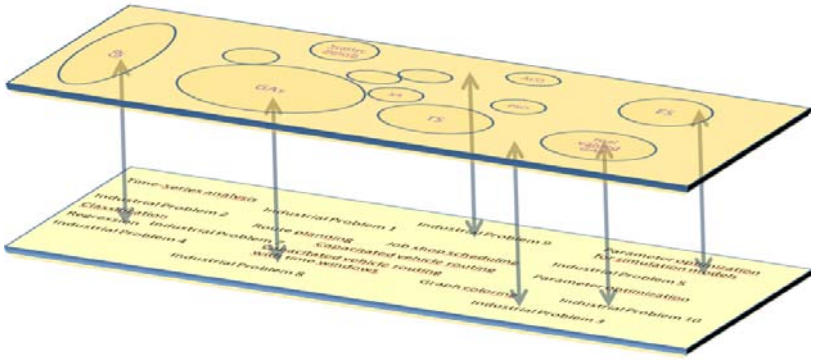
Metaheuristics

Metaheuristics

- Intelligent search strategies
- Can be applied to different problems
- Explore interesting regions of the search space (parameter)
- Tradeoff: computation vs. quality
  - Good solutions for very complex problems
- Must be tuned to applications

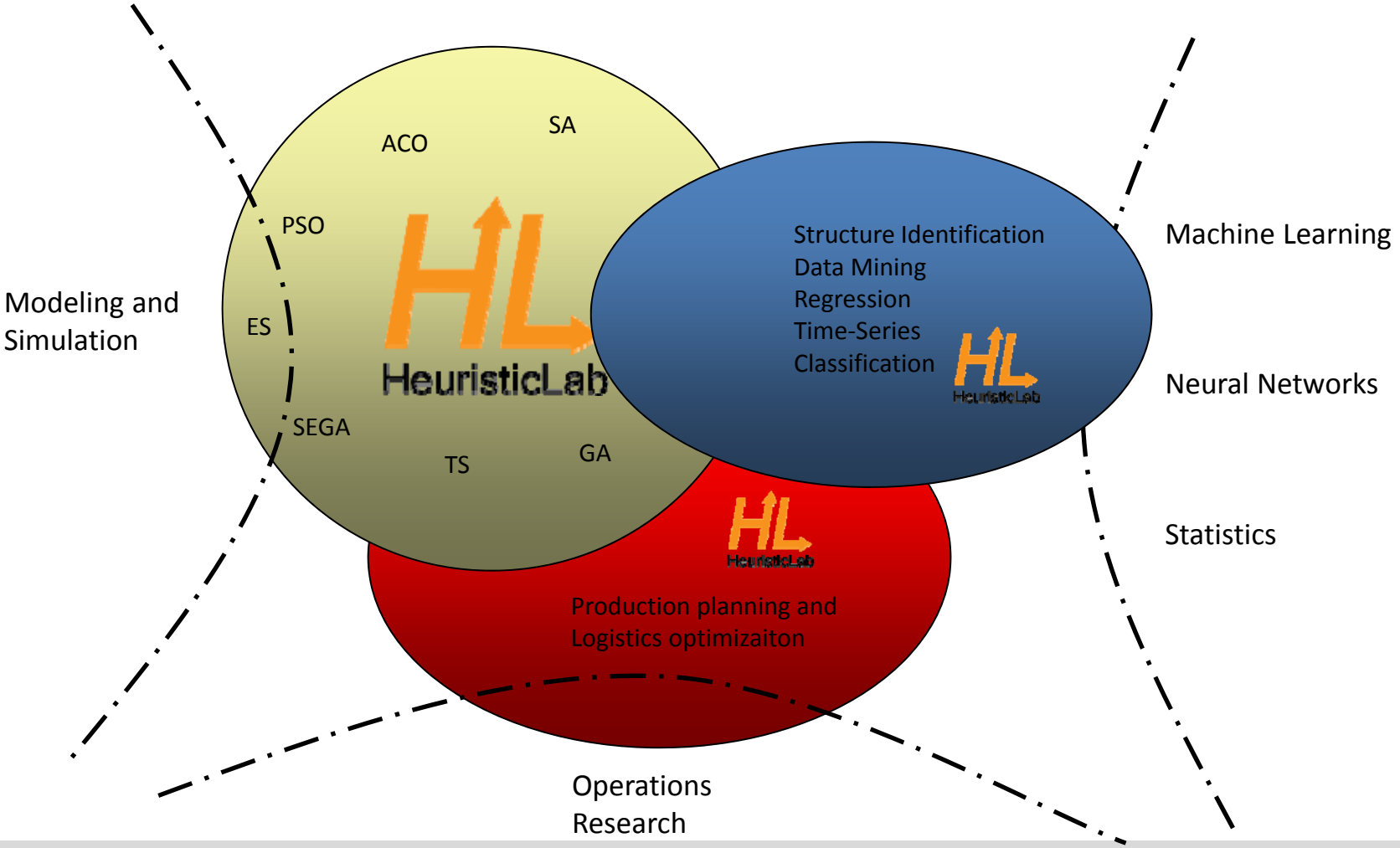
Challenges

- Choice of appropriate metaheuristics
- Hybridization



Finding Needles in Haystacks

Research Focus



HeuristicLab

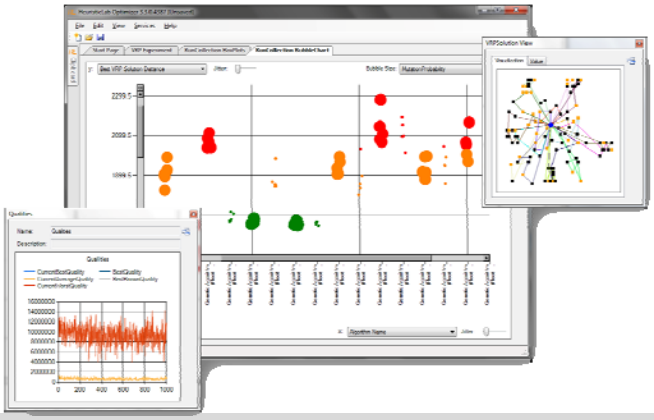
Open Source Optimization Framework HeuristicLab

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

http://dev.heuristiclab.com

Motivation und Goals

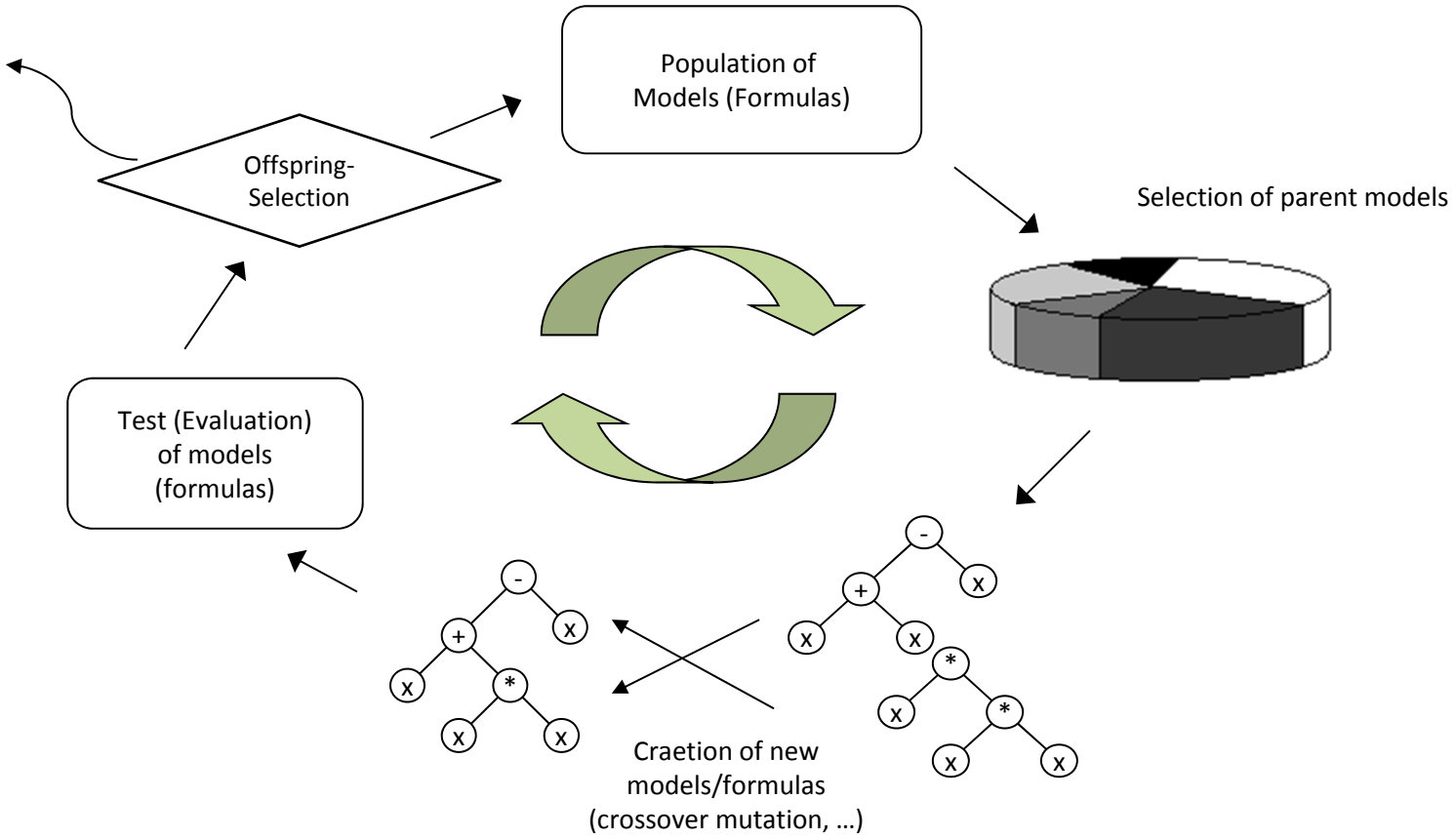
- Graphical user interface for interactive development, analysis and applicat
- Numerous optimization algorithms and optimization problems
- Support for extensive experiments and analysis
- Distribution through parallel execution of algorithms
- Extensibility and flexibility (plug-in architecture)



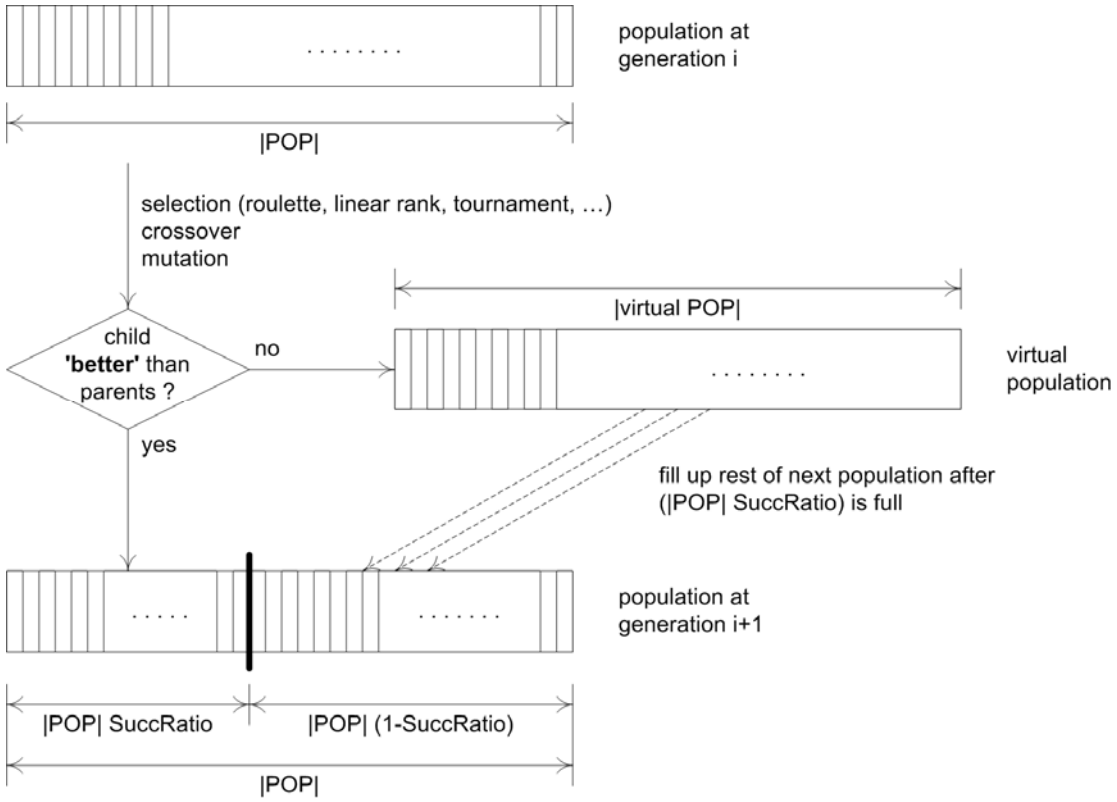
Cluster at campus Hagenberg

- Research cluster (since March 2006) with 14 cores
- Dell Blade system (since January 2009) with 112 cores
- 200-300 lab computers at campus Hagenberg (2011)

Integration of Offspring Selection in GP Workflow

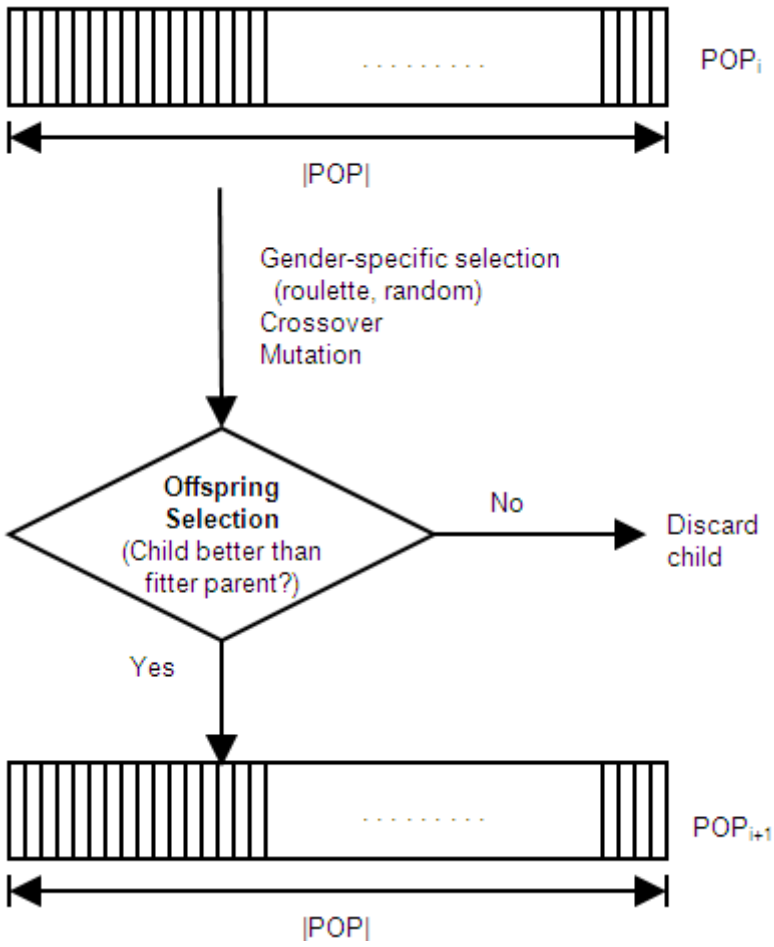


Integration of Offspring Selection in GP Workflow



- Originally developed for standard GAs for combinatorial optimization and parameter optimization
- Motivated by population genetics and evolution strategies
- Self-adaptive selection pressure
- Dynamic termination (migration in parallel variant)
- Generic extension of GA/GP

Strict Offspring Selection



- SuccRatio = 1.0
- CompFactor = 1.0

- Selection of genes
- Adaptive selection pressure
- Dynamic termination

- Possibility to use multiple operators in parallel
  - Dynamic adaptation of operator usage
  - Better results than the best opertor could provide
- More compact active gene pool

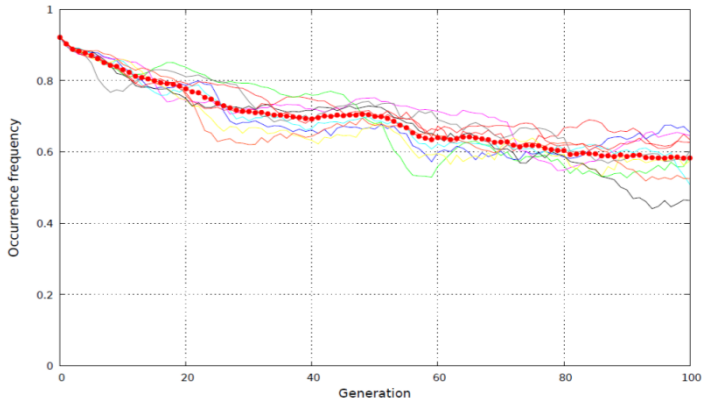
Observing Population Diversity: Showcase Standard GP

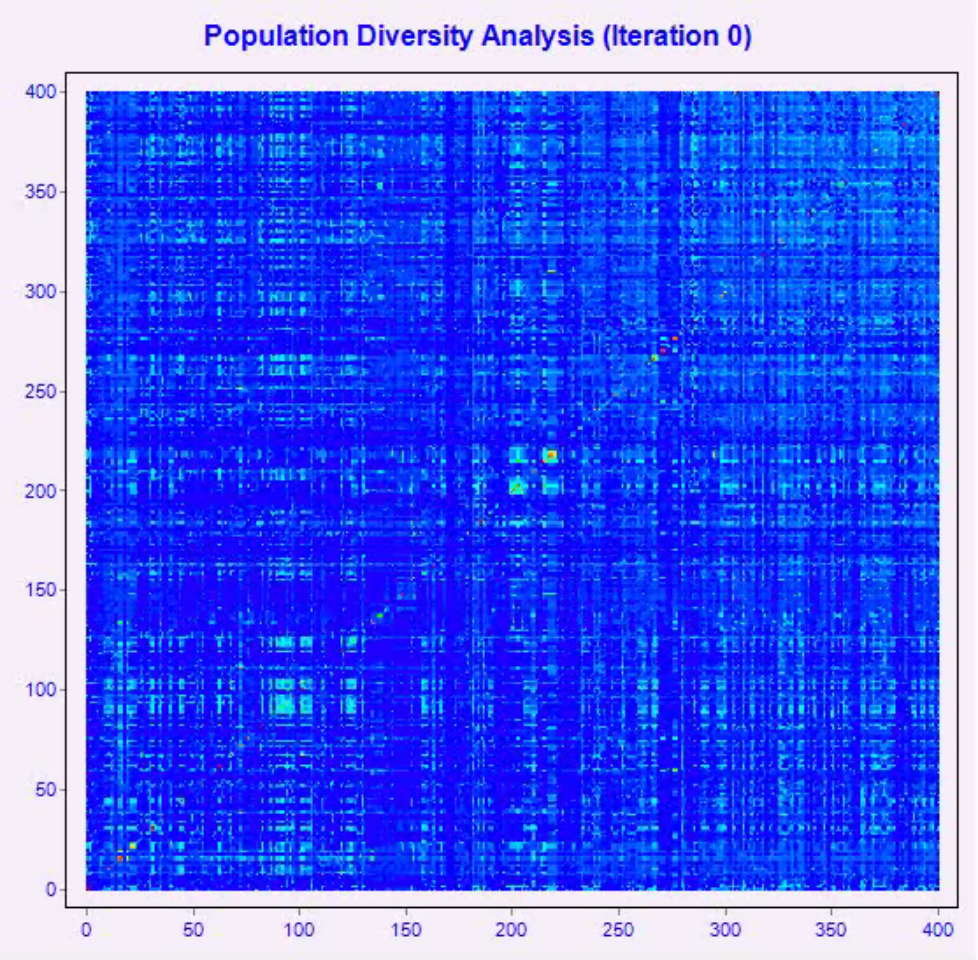
Pairwise similarities

$$Sim(t_1,t_2) = \frac{2 \cdot |MaximumCommonSubtree(t_1,t_2)|}{|t_1| + |t_2|}$$

Population diversity

$$Div(T) = 1 - \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} Sim(t_i,t_j)}{N(N-1)/2}$$





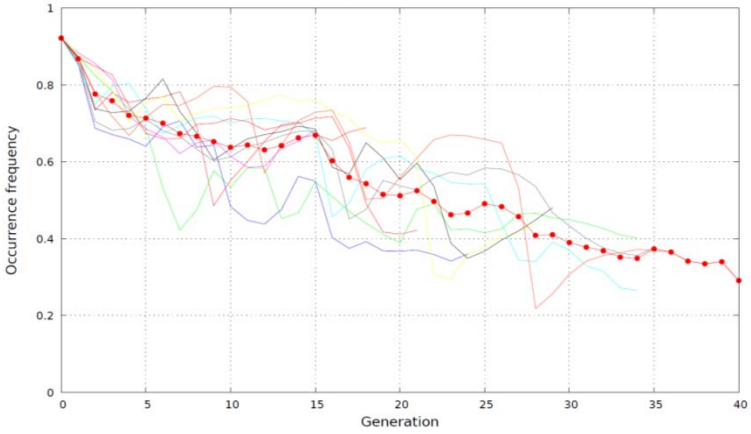
Observing Population Diversity: Showcase OSGP

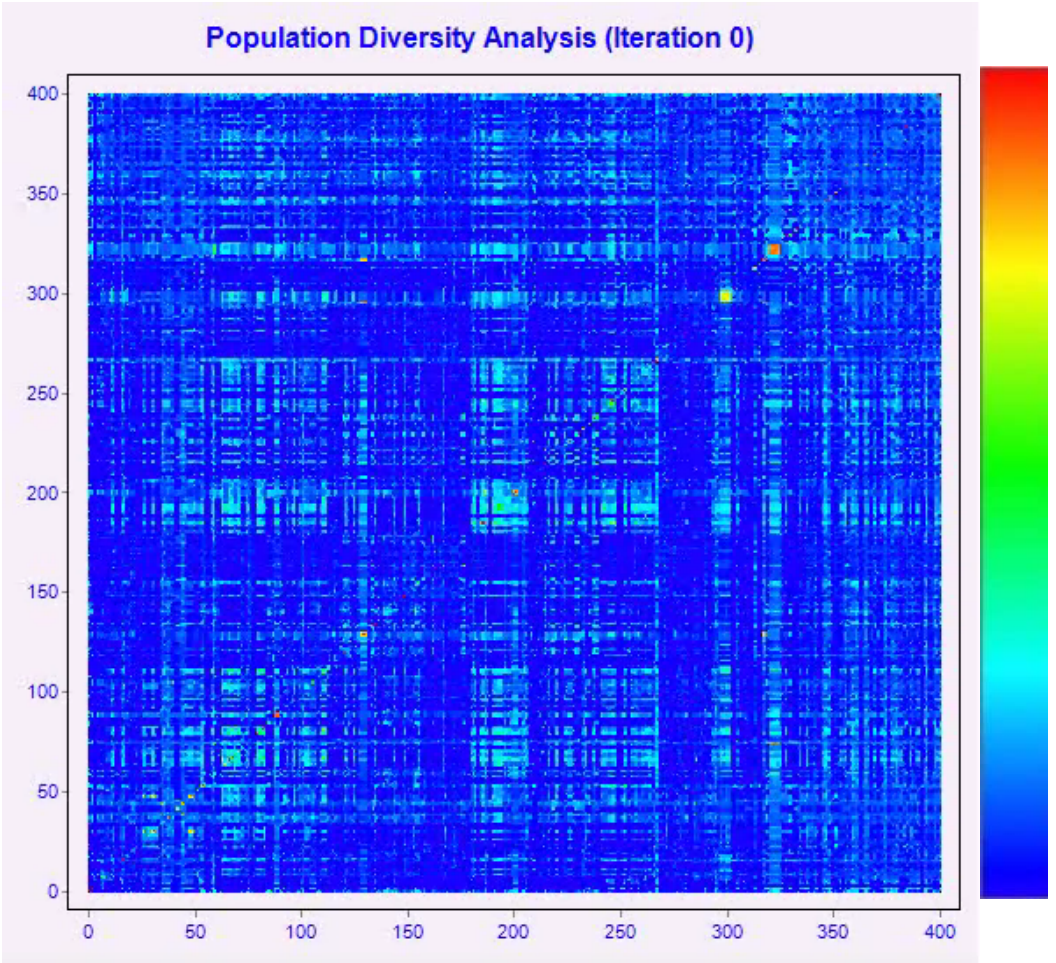
Pairwise similarities

$$Sim(t_1, t_2) = \frac{2 \cdot |MaximumCommonSubtree(t_1, t_2)|}{|t_1| + |t_2|}$$

Population diversity

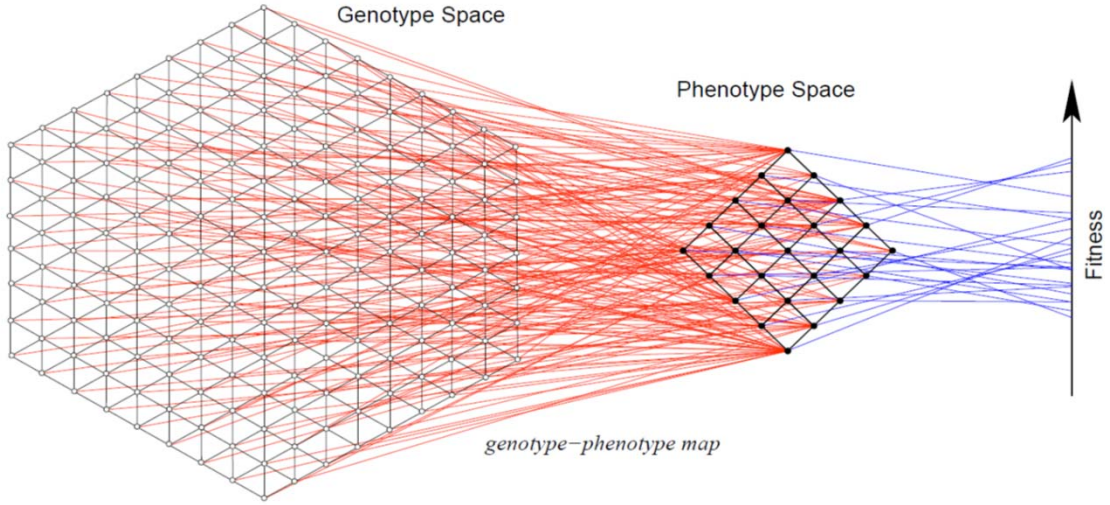
$$Div(T) = 1 - \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} Sim(t_i, t_j)}{N(N-1)/2}$$





Genealogy Analysis

Complex mapping between genotype and phenotype



How to gain deeper insights concerning

- Evolvement of relevant building blocks
- Influence of genetic operators
- Funtional basis

Interactive analysis concerning

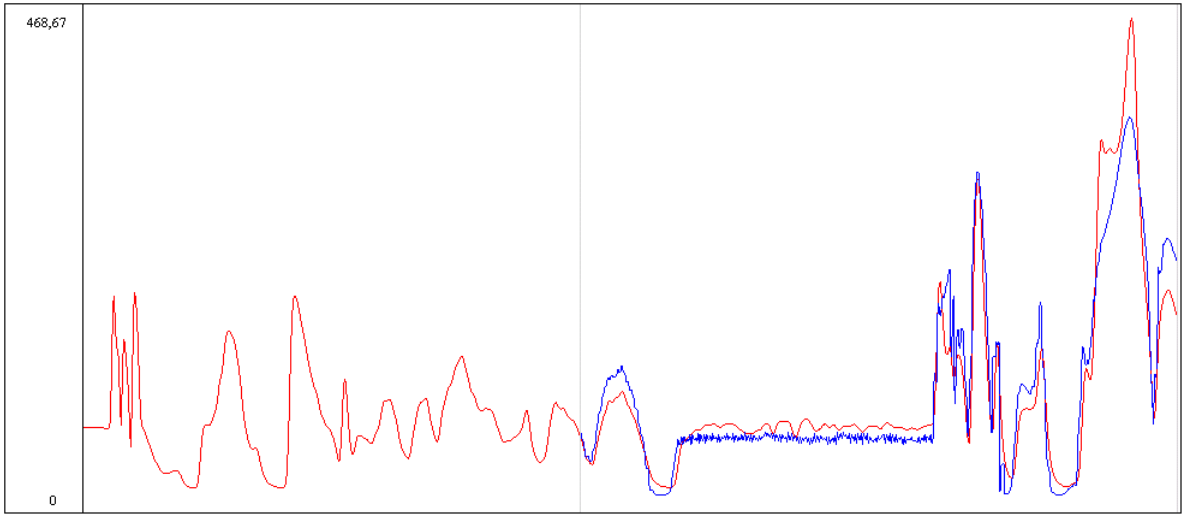
- Common substructures
- History of partial solutions

Genealogy Analysis: Demo video

Population graph

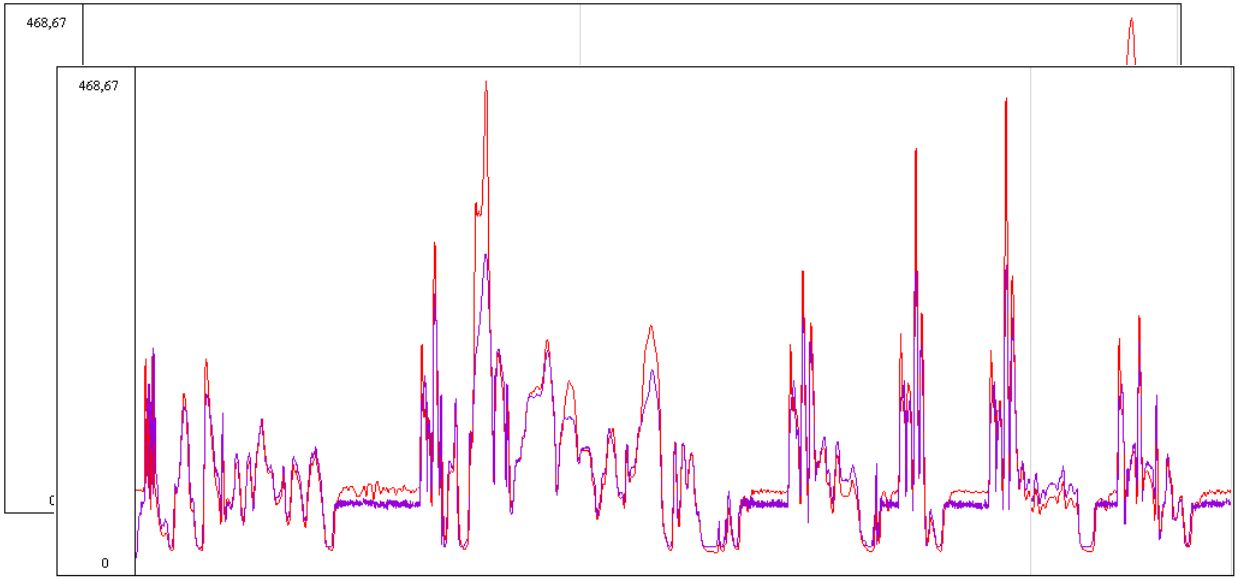
Distribution of fitness values

New Algorithmic Approaches: On-Line GP



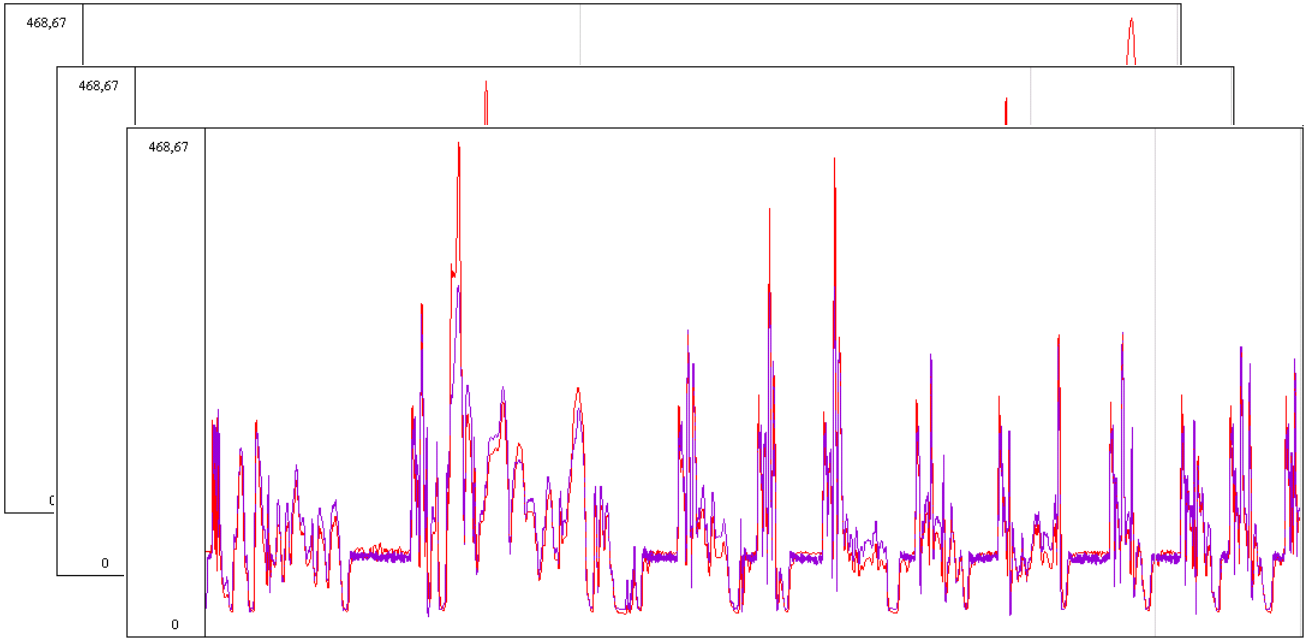
2 min

New Algorithmic Approaches : On-Line GP



5 min

New Algorithmic Approaches : On-Line GP



15 min

New Algorithmisch Approaches : On-Line GP

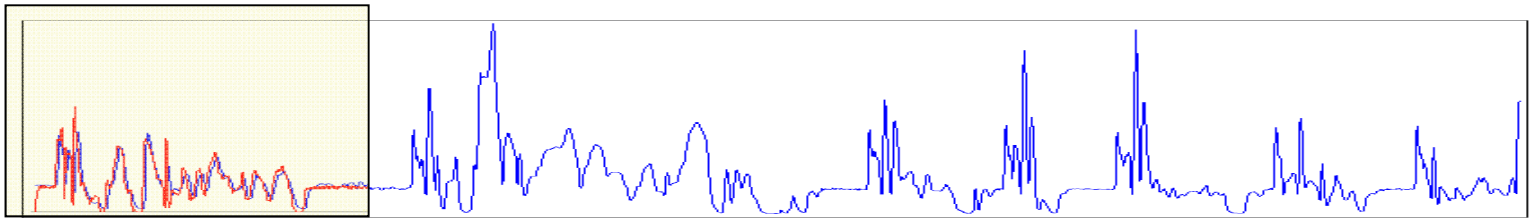
Properties:

- More compact models
- Less overfitting
- Speed-Up

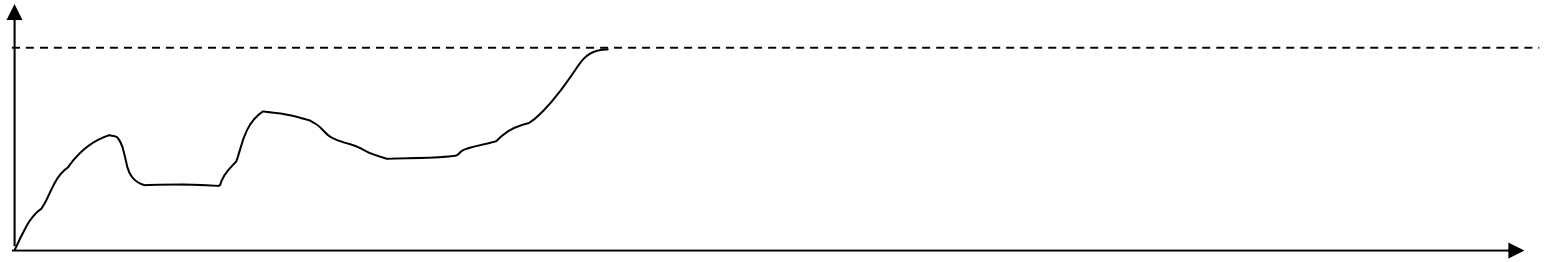
Interpretation of possible reasons:

- Changing environments
- Many models are able to explain parts of the data; due to changing focus those parts remain that are able to explain the entire system
- Speed-Up simply due to analysis of less data

New Algorithmic Approaches : Sliding-window OSGP

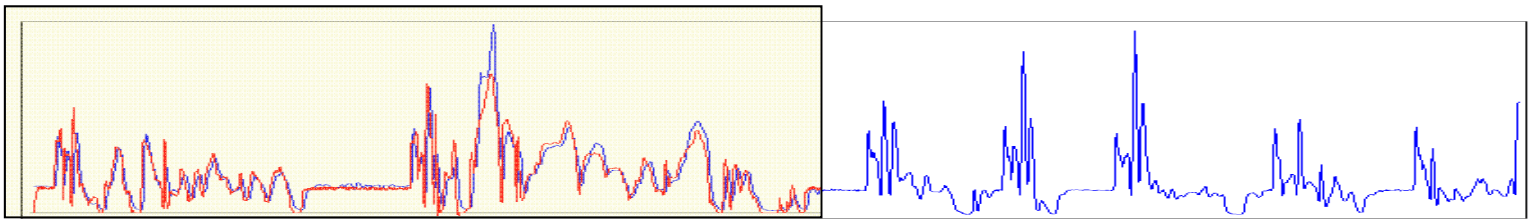


Selection pressure:

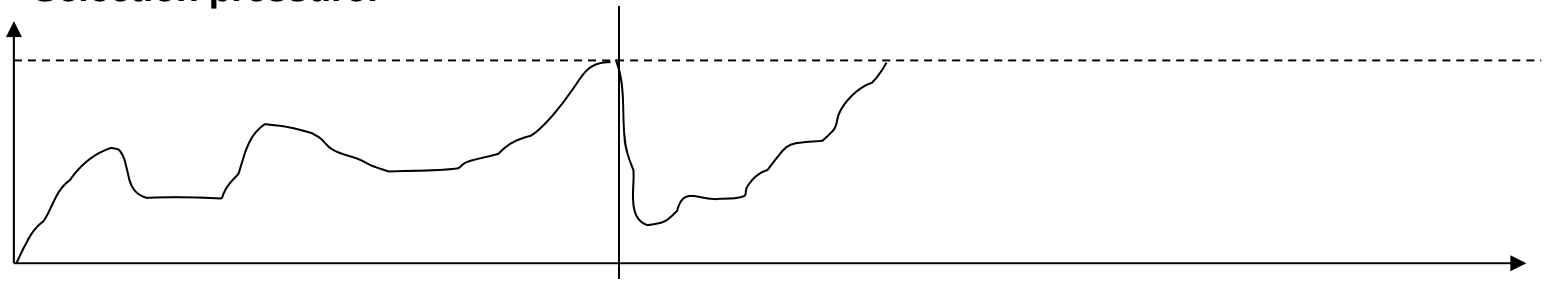


Identification Algorithm Execution Time

New Algorithmic Approaches : Sliding-window OSGP

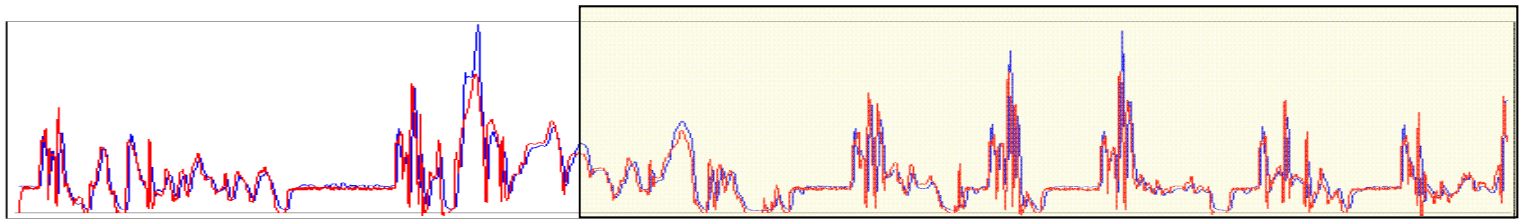


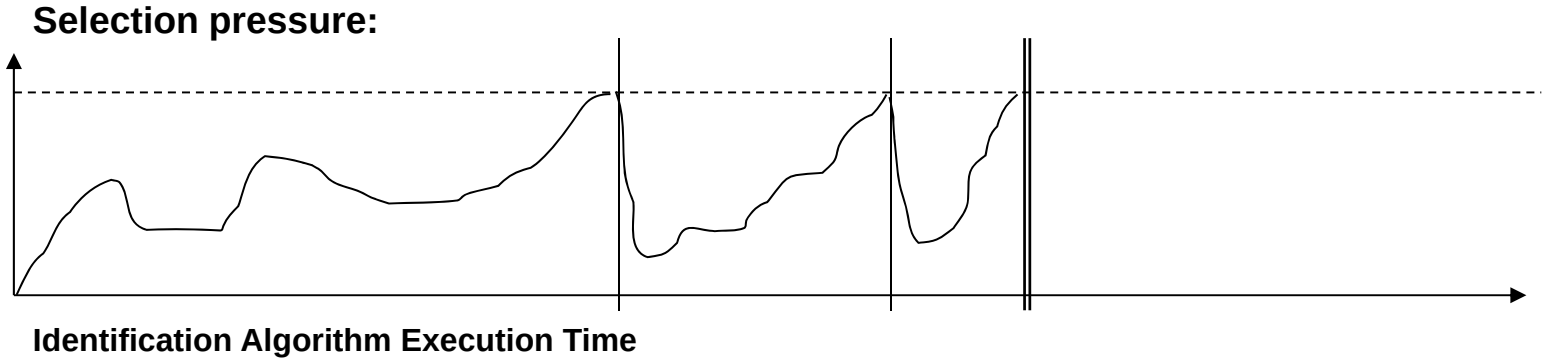
Selection pressure:



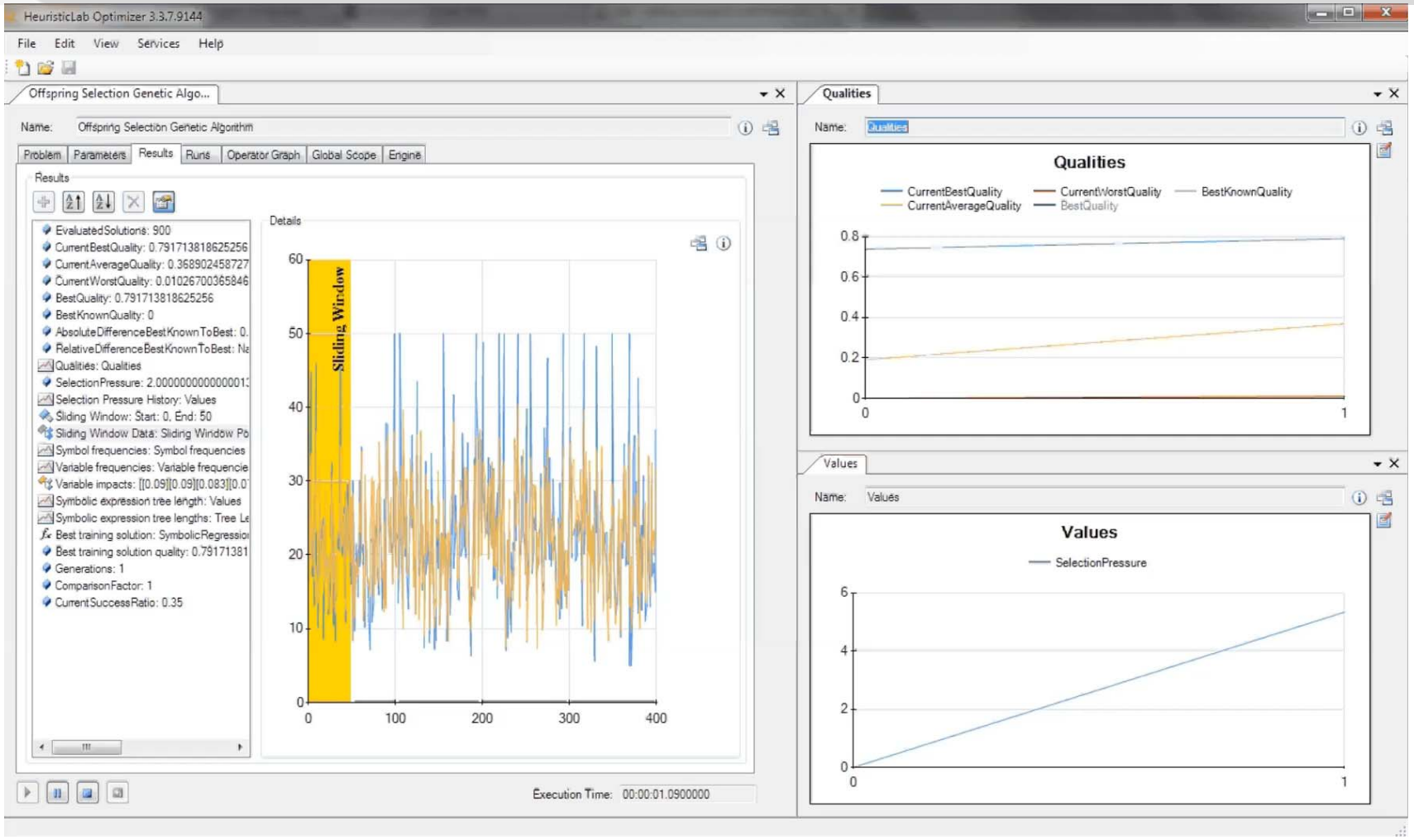
Identification Algorithm Execution Time

New Algorithmic Approaches : Sliding-window OSGP





New Algorithmic Approaches : Sliding-window OSGP Demo



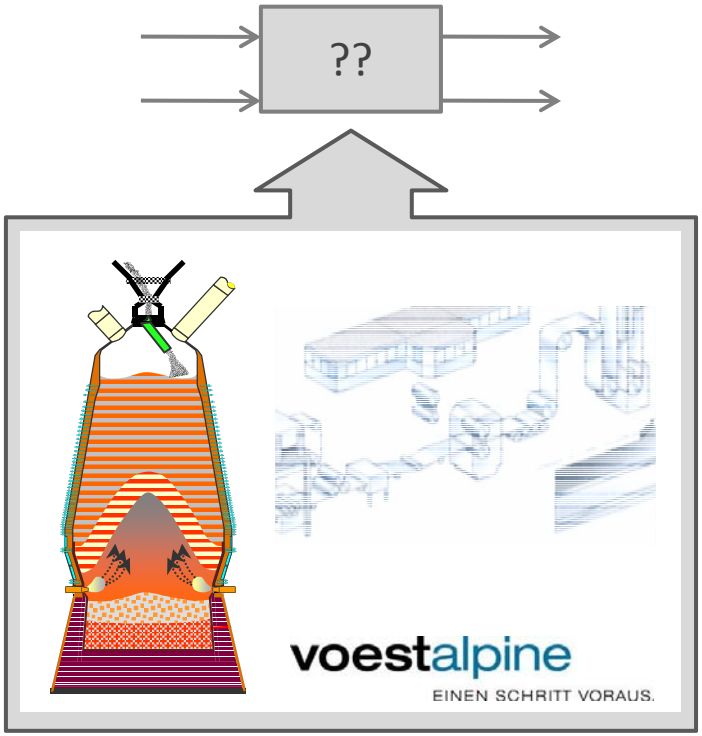
Enhanced Interpretability of Symbolic Regression Models

- Relevance of variables
- Model simplification
- Network analysis

Data Based Modeling: Starting point

Goal: Mathematical models that describe system behavior

System = Engine, human body, financial data etc.



Analysis of steel production processes

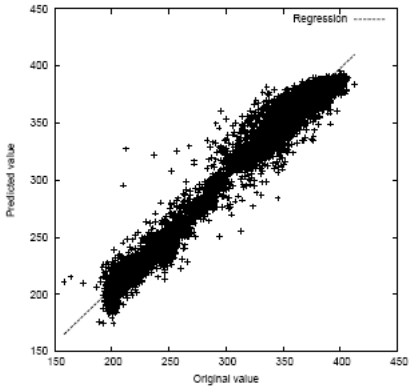
Medical data analysis

Data Based Modeling: Black-Box vs. White-Box Modeling

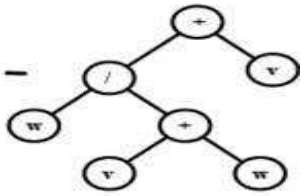
- Instead of **Black Box Models** (ANN, SVM) identification of model structure (White Box Modelle) (symbolic regression/classification with Genetic Programming)

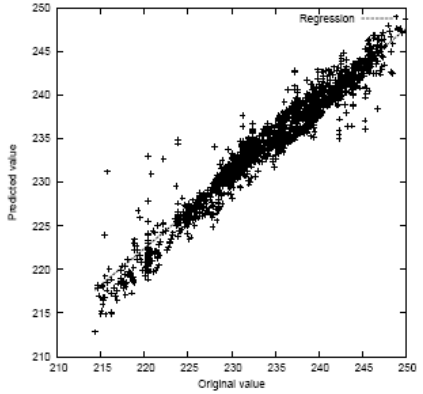
Black Box Model

?



White Box Model





Data Based Modeling: Genetic Programming (Model analysis)

- Algorithms quite robust
- So far: little **support for the analysis of results**
- Code bloat and introns lead to overcomplex models
- Problem: Equivalent GP Models often different regarding their structure
  - Implicit correlations of input variables
  - Same functional correlations explained differently due to stochastic process

- **Goals**
  - Simple analysis of relevance of imput variables
  - Visualization of model properties
  - Manual model manipulation
  - Interactive model-analysis
  - Export/Import of models (Mathematica, Excel, Tex …)
  - …

Relevance of Variables

Implicit feature selection of symbolic regression

$$\text{relevance}_{\text{freq}}(x_i) = \frac{1}{G}\sum_{g=1}^{G} \text{freq}_{\%}(x_i, \text{Pop}_g)$$

$$\text{freq}_{\%}(x_i, \text{Pop}) = \frac{\sum_{s \in \text{Pop}} \text{RefCount}(x_i, s))}{\sum_{k=1}^{n} \sum_{s \in \text{Pop}} \text{RefCount}(x_k, s)}$$

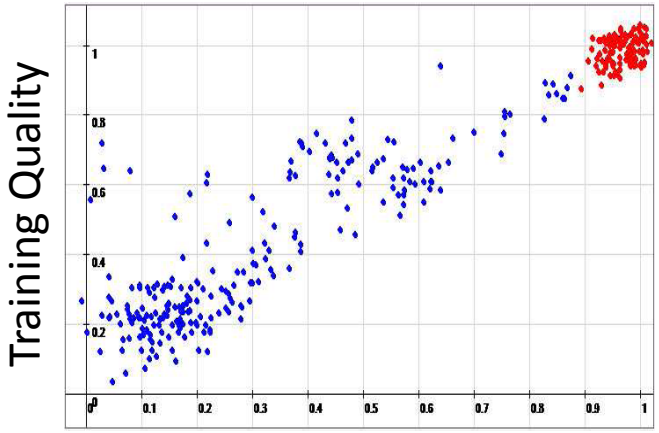
Limitations:

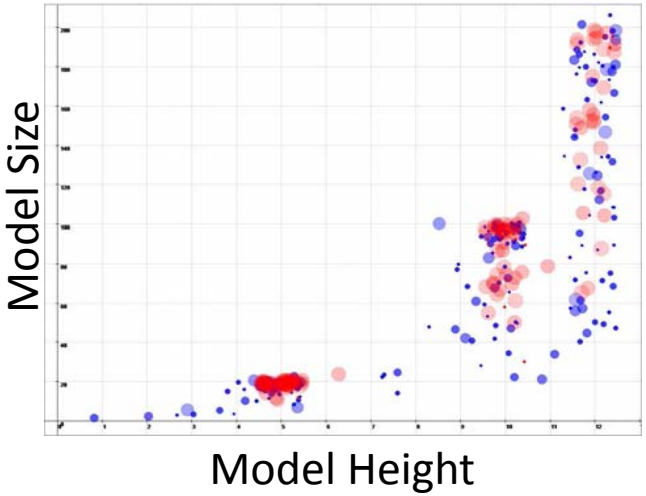
- Bloat
- Introns
- Ambiguity of models

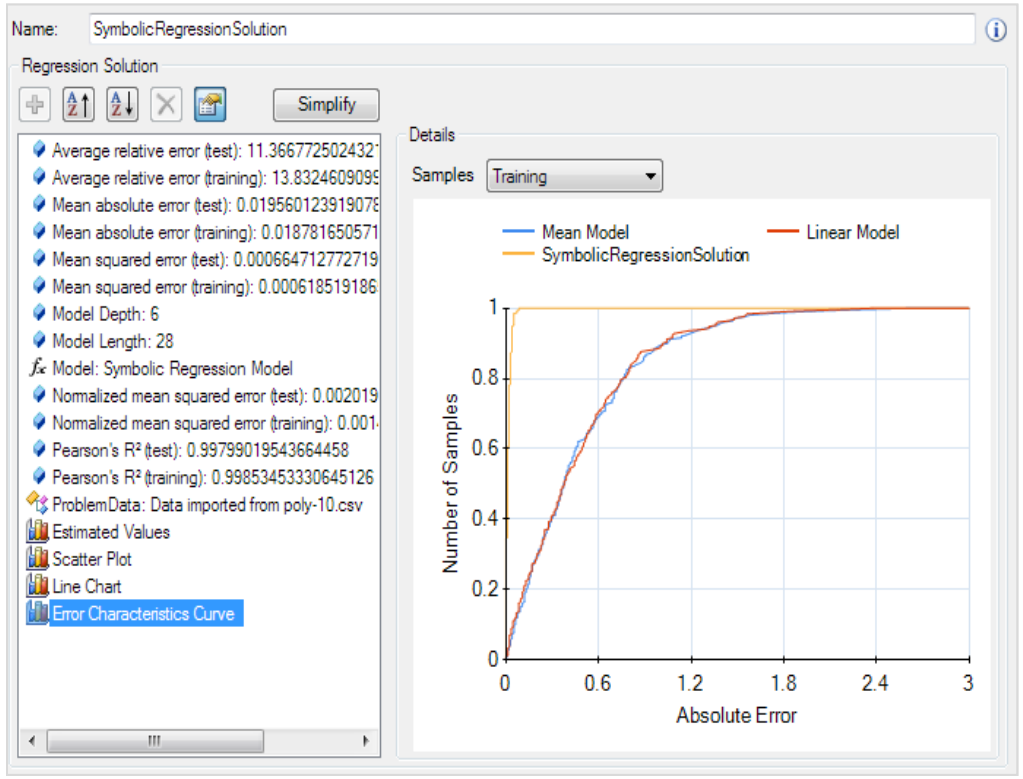
Related Variables		Unrelated Variables	
X_1	0.284	X_{10}	0.049
X_2	0.210	X_9	0.031
X_3	0.170	X_8	0.027
X_5	0.108	X_6	0.019
X_4	0.091	X_7	0.012

Impacts of variables for the Friedman 10 problem

Selection of Interpretable Models







Selected Model

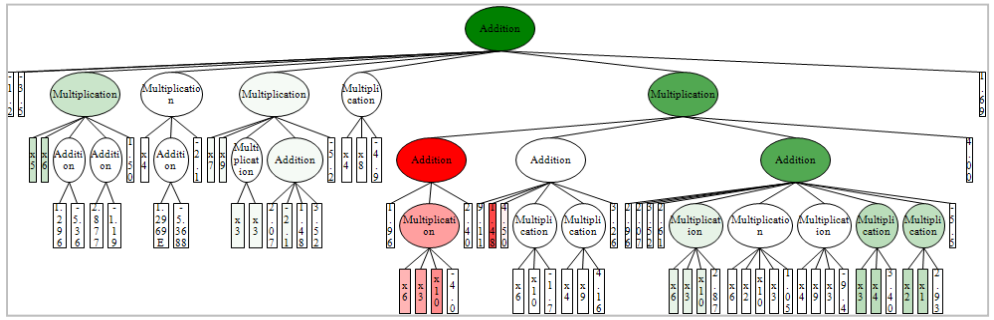
Model Simplification and Interpretation

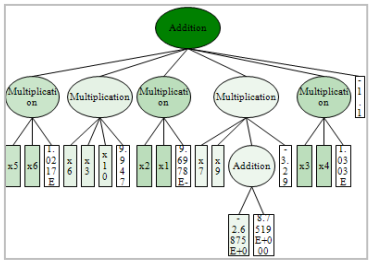
Simplification Methods

- Mathematical transformation
- Remove Nodes
- Constant optimization
- External optimization

Export

- Textual Export
- Latex, MatLab,
- Graphical Export





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

Variable Network Analysis

Starting point:

- Many models with similar quality and complexity
- Ambiguity of models caused also by dependencies among the input vatiables
- Which input variables can be explained by others

Analysis

- Model each input variable separately using smbolic regression
- Identify main contributing variables

Goals:

- More unique odels
- Better understanding of entire data set

Variable Network Analysis

Estimate target variable value from input variable values

$$y = f(x1, \ldots, x5) + e$$

	x1	x2	x3	x4	x5	x6	y
Training	52.0085	58497.7	26.9956	24.0391	12.9005	66.2633	4.027
	52.9874	59181.1	27.0061	23.7483	12.7239	66.8367	4.154
	54.0505	58879.2	26.8679	23.5869	12.5877	67.6773	4.047
	55.3182	58982.3	27.1852	23.7869	12.6229	67.7128	4.118
	46.5388	59270.3	26.4421	24.2601	12.5892	65.7425	4.743
	56.0167	58234.5	26.1072	23.4412	12.4224	67.9462	4.042
	57.0168	58021.7	26.2513	23.3085	12.3036	69.4392	3.721
	57.0139	59228.5	26.2285	23.4041	12.3685	68.3322	3.593
	62.5214	49855.7	26.3076	22.7812	12.0711	70.3331	3.785
	69.9885	53534.4	25.5916	21.3448	11.0299	73.7138	3.028
	70.0156	55058.2	25.5759	21.4617	11.0915	72.9846	3.137
Test	70.0446	56099.4	26.059	21.4652	11.0684	73.2351	?
	70.0335	55891.6	26.1834	21.3318	10.6288	74.8925	?
	70.0164	56841.1	26.1791	21.2119	10.0354	74.8419	?
	70.0147	58254.6	26.7381	21.3367	10.9244	73.2341	?
	70.0061	54444.9	26.5786	21.2395	10.6505	74.9187	?
	70.0325	48858.2	26.3799	21.4135	10.8558	74.0993	?

Variable Network Analysis

All variables are considered as target variable

$$x1 = f(x2, \ldots, x5, y) + e$$

	x1	x2	x3	x4	x5	x6	y
Training	52.0085	58497.7	26.9956	24.0391	12.9005	66.2633	4.027
	52.9874	59181.1	27.0061	23.7483	12.7239	66.8367	4.154
	54.0505	58879.2	26.8679	23.5869	12.5877	67.6773	4.047
	55.3182	58982.3	27.1852	23.7869	12.6229	67.7128	4.118
	46.5388	59270.3	26.4421	24.2601	12.5892	65.7425	4.743
	56.0167	58234.5	26.1072	23.4412	12.4224	67.9462	4.042
	57.0168	58021.7	26.2513	23.3085	12.3036	69.4392	3.721
	57.0139	59228.5	26.2285	23.4041	12.3685	68.3322	3.593
	62.5214	49855.7	26.3076	22.7812	12.0711	70.3331	3.785
	69.9885	53534.4	25.5916	21.3448	11.0299	73.7138	3.028
	70.0156	55058.2	25.5759	21.4617	11.0915	72.9846	3.137
Test	?	56099.4	26.059	21.4652	11.0684	73.2351	4.118
	?	55891.6	26.1834	21.3318	10.6288	74.8925	4.743
	?	56841.1	26.1791	21.2119	10.0354	74.8419	4.042
	?	58254.6	26.7381	21.3367	10.9244	73.2341	3.721
	?	54444.9	26.5786	21.2395	10.6505	74.9187	3.593
	?	48858.2	26.3799	21.4135	10.8558	74.0993	4.118

Variable Network Analysis

All variables are considered as target variable

$$x2 = f(x1, \ldots, x5, y) + e$$

	x1	x2	x3	x4	x5	x6	y
Training	52.0085	58497.7	26.9956	24.0391	12.9005	66.2633	4.027
	52.9874	59181.1	27.0061	23.7483	12.7239	66.8367	4.154
	54.0505	58879.2	26.8679	23.5869	12.5877	67.6773	4.047
	55.3182	58982.3	27.1852	23.7869	12.6229	67.7128	4.118
	46.5388	59270.3	26.4421	24.2601	12.5892	65.7425	4.743
	56.0167	58234.5	26.1072	23.4412	12.4224	67.9462	4.042
	57.0168	58021.7	26.2513	23.3085	12.3036	69.4392	3.721
	57.0139	59228.5	26.2285	23.4041	12.3685	68.3322	3.593
	62.5214	49855.7	26.3076	22.7812	12.0711	70.3331	3.785
	69.9885	53534.4	25.5916	21.3448	11.0299	73.7138	3.028
	70.0156	55058.2	25.5759	21.4617	11.0915	72.9846	3.137
Test	70.0446	?	26.059	21.4652	11.0684	73.2351	4.118
	70.0335	?	26.1834	21.3318	10.6288	74.8925	4.743
	70.0164	?	26.1791	21.2119	10.0354	74.8419	4.042
	70.0147	?	26.7381	21.3367	10.9244	73.2341	3.721
	70.0061	?	26.5786	21.2395	10.6505	74.9187	3.593
	70.0325	?	26.3799	21.4135	10.8558	74.0993	4.118

Variable Network Analysis

All variables are considered as target variable

$$x3 = f(x1, \ldots, x5, y) + e$$

	x1	x2	x3	x4	x5	x6	y
Training	52.0085	58497.7	26.9956	24.0391	12.9005	66.2633	4.027
	52.9874	59181.1	27.0061	23.7483	12.7239	66.8367	4.154
	54.0505	58879.2	26.8679	23.5869	12.5877	67.6773	4.047
	55.3182	58982.3	27.1852	23.7869	12.6229	67.7128	4.118
	46.5388	59270.3	26.4421	24.2601	12.5892	65.7425	4.743
	56.0167	58234.5	26.1072	23.4412	12.4224	67.9462	4.042
	57.0168	58021.7	26.2513	23.3085	12.3036	69.4392	3.721
	57.0139	59228.5	26.2285	23.4041	12.3685	68.3322	3.593
	62.5214	49855.7	26.3076	22.7812	12.0711	70.3331	3.785
	69.9885	53534.4	25.5916	21.3448	11.0299	73.7138	3.028
	70.0156	55058.2	25.5759	21.4617	11.0915	72.9846	3.137
Test	70.0446	56099.4	?	21.4652	11.0684	73.2351	4.118
	70.0335	55891.6	?	21.3318	10.6288	74.8925	4.743
	70.0164	56841.1	?	21.2119	10.0354	74.8419	4.042
	70.0147	58254.6	?	21.3367	10.9244	73.2341	3.721
	70.0061	54444.9	?	21.2395	10.6505	74.9187	3.593
	70.0325	48858.2	?	21.4135	10.8558	74.0993	4.118

Variable Network Analysis

All variables are considered as target variable

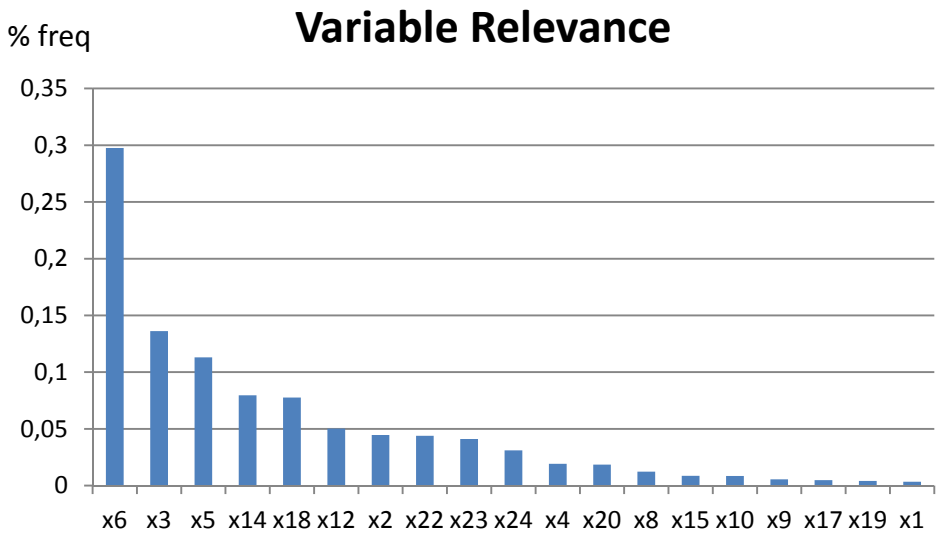
$$x3 = f(x1, \ldots, x5, y) + e$$

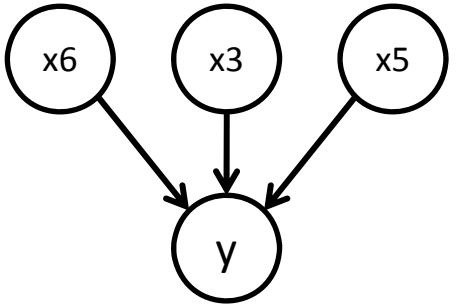
	x1	x2	x3	x4	x5	x6	y
Training	52.0085	58497.7	26.9956	24.0391	12.9005	66.2633	4.027
	52.9874	59181.1	27.0061	23.7483	12.7239	66.8367	4.154
	54.0505	58879.2	26.8679	23.5869	12.5877	67.6773	4.047
	55.3182	58982.3	27.1852	23.7869	12.6229	67.7128	4.118
	46.5388	59270.3	26.4421	24.2601	12.5892	65.7425	4.743
	56.0167	58234.5	26.1072	23.4412	12.4224	67.9462	4.042
	57.0168	58021.7	26.2513	23.3085	12.3036	69.4392	3.721
	57.0139	59228.5	26.2285	23.4041	12.3685	68.3322	3.593
	62.5214	49855.7	26.3076	22.7812	12.0711	70.3331	3.785
	69.9885	53534.4	25.5916	21.3448	11.0299	73.7138	3.028
	70.0156	55058.2	25.5759	21.4617	11.0915	72.9846	3.137
Test	70.0446	56099.4	26.059	?	11.0684	73.2351	4.118
	70.0335	55891.6	26.1834	?	10.6288	74.8925	4.743
	70.0164	56841.1	26.1791	?	10.0354	74.8419	4.042
	70.0147	58254.6	26.7381	?	10.9244	73.2341	3.721
	70.0061	54444.9	26.5786	?	10.6505	74.9187	3.593
	70.0325	48858.2	26.3799	?	10.8558	74.0993	4.118

Variable Network Analysis

Based on variable relevance create network of strong dependencies

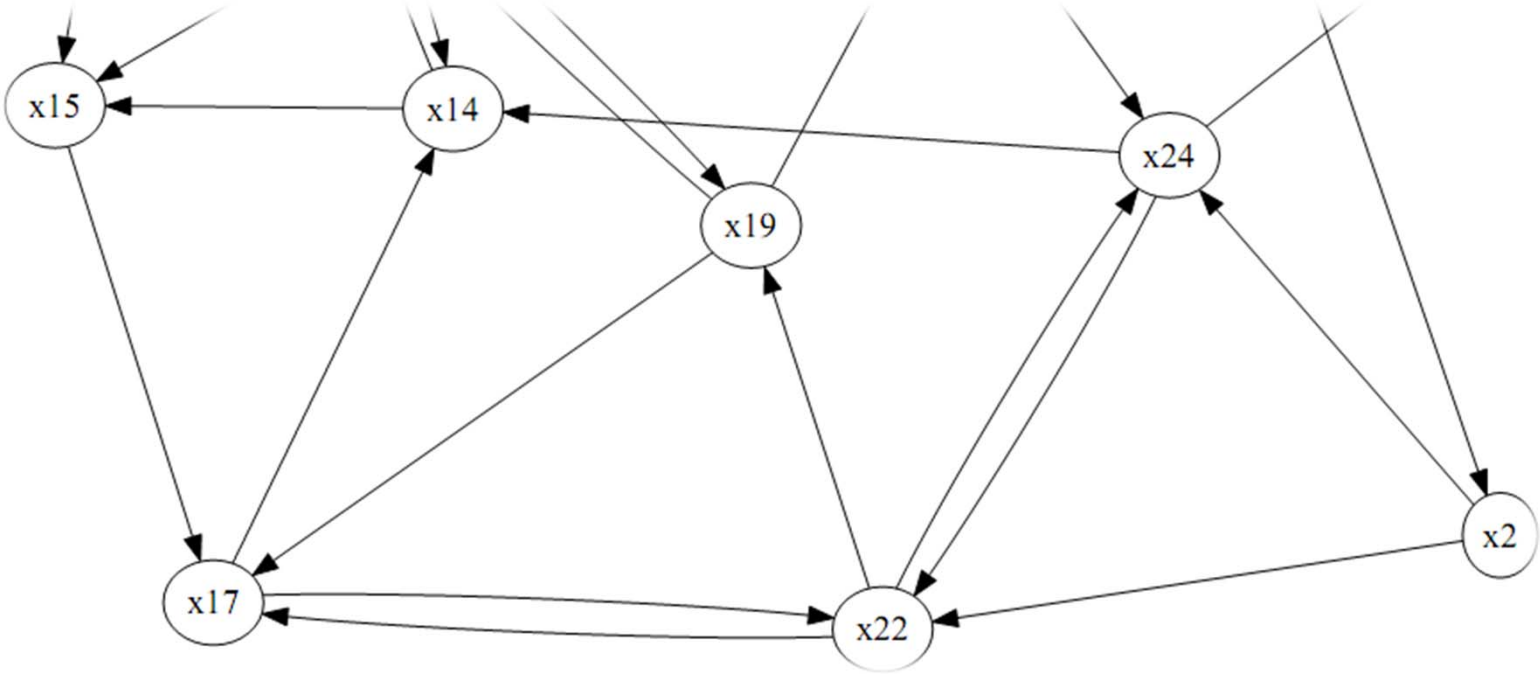
- Top three most important variables by relevance
- Example:
  - "*x6* → *y*" means *x6* is relevant to estimate *y*





Variable Network Analysis

For all target variables → network of variable interactions



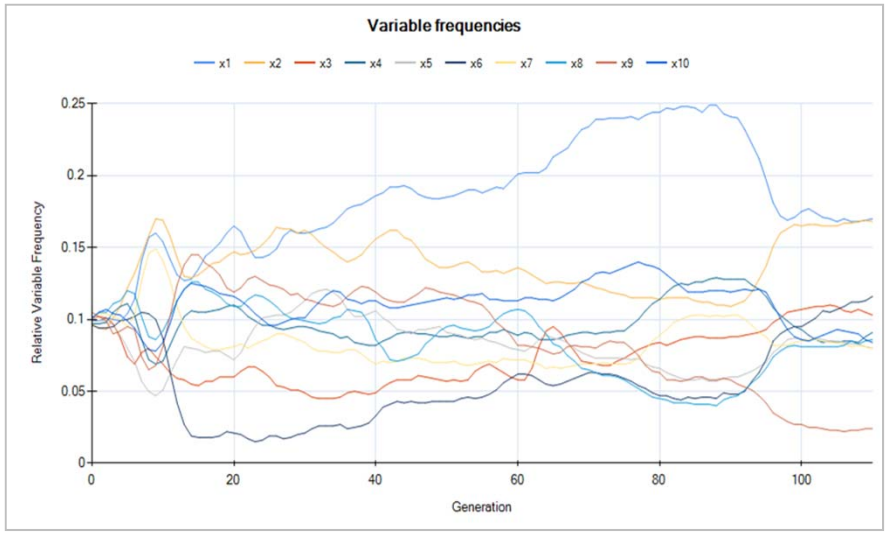
Variable Network Analysis

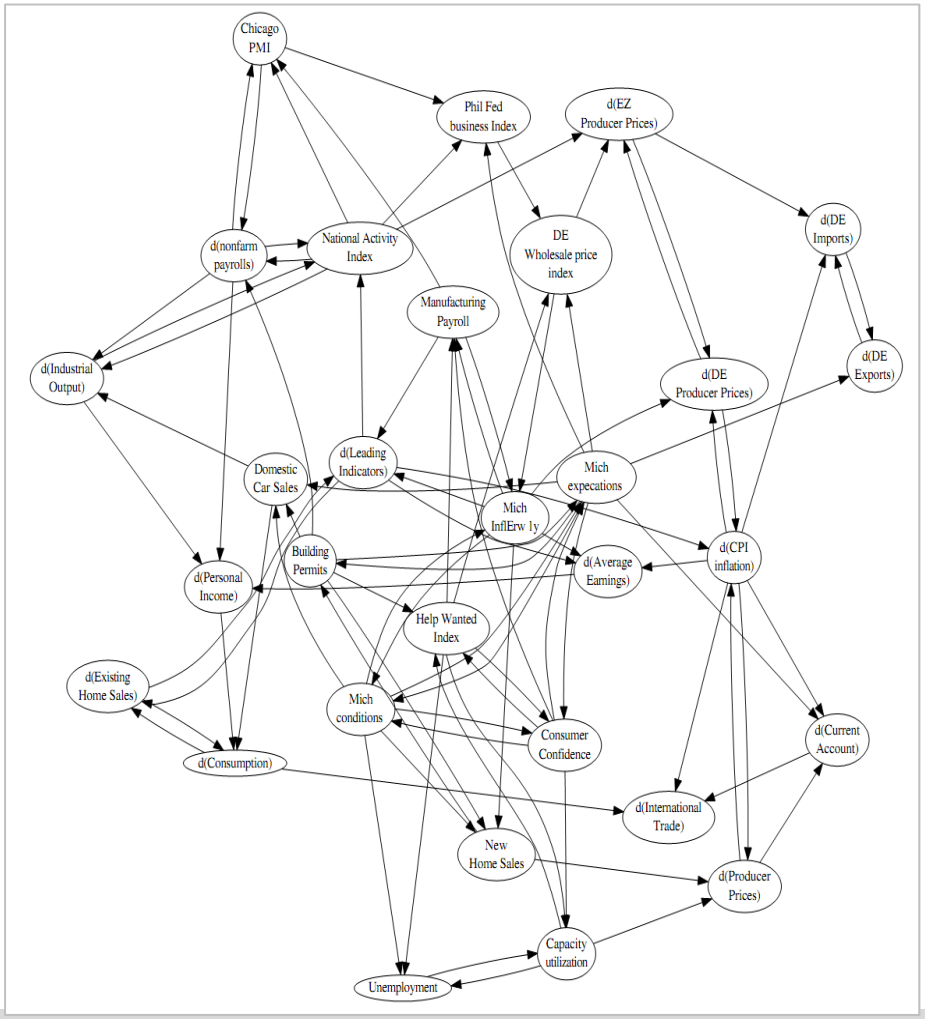
Variable Interaction Networks

- Reveals non-linear correlations

Variable Frequencies

- Analyzed during the algorithm run

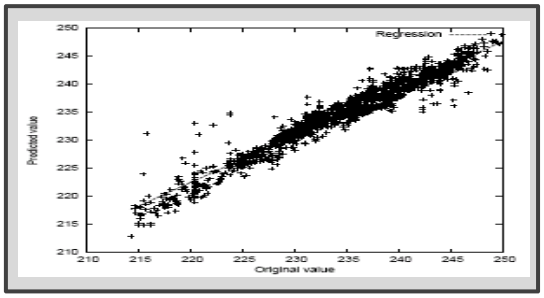


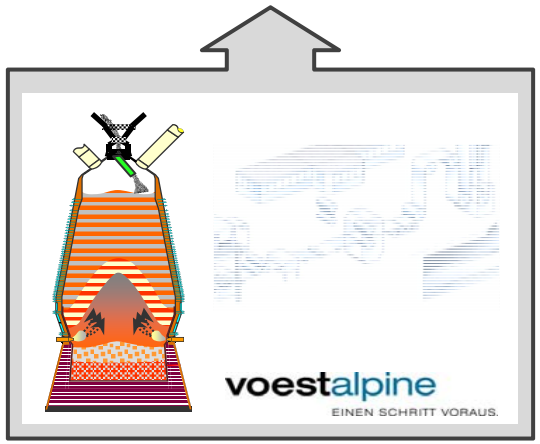


Data Based Modeling: Results

Results:

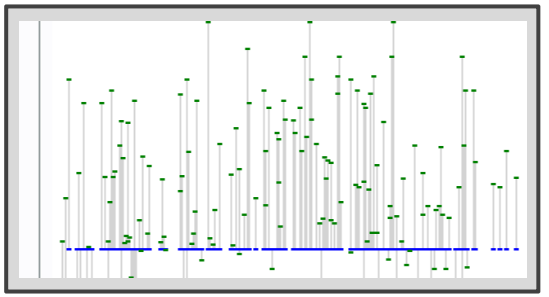
MeltingRate(t)= f(x3$_{(t-2)}$, x1$_{(t-3)}$, …)





Improved process understanding for steel production

C125(p780641) = (chol, GGT, x58, …)

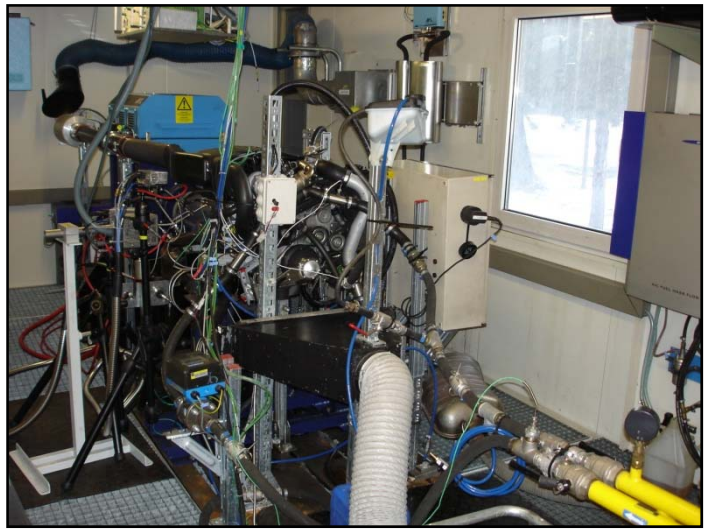


Virtual tumor markers for cancer diagnosis

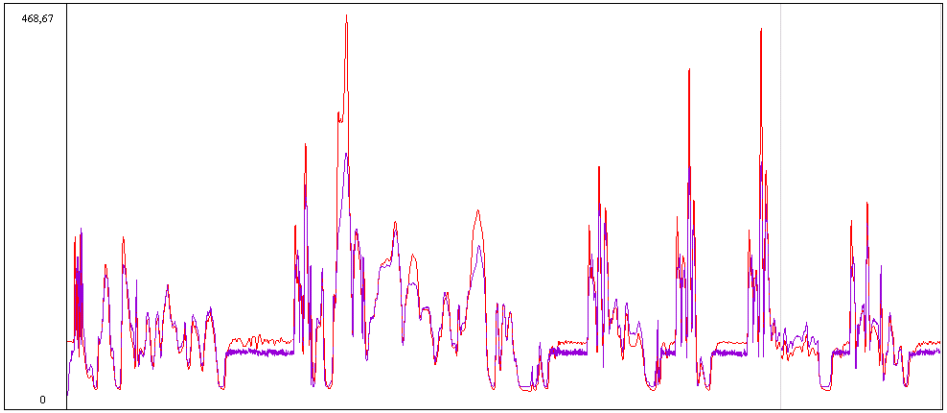
Example:
Virtual Sensors for Modeling Exhaust Gases

Motivation:

- High quality modeling of emissions (NO_x 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 O_2 in exhaustion etc.

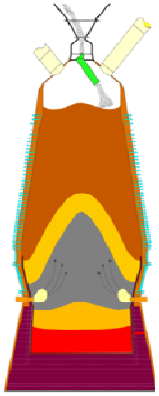


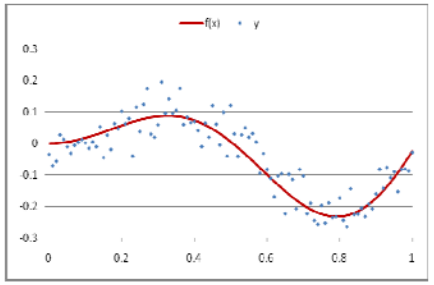
$$NO_x(t) = f(x1_{(t-7)}, x2_{(t-2)}, \ldots)$$



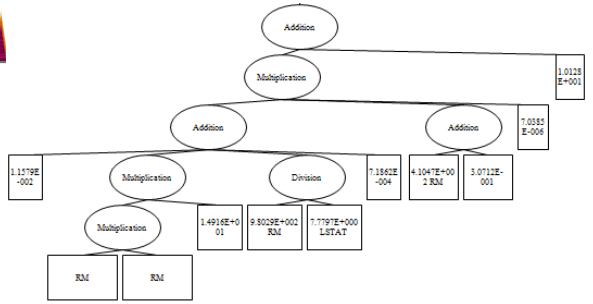
Example: Blast furnace modeling

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	?





Model → *f(x)* → Prognosis



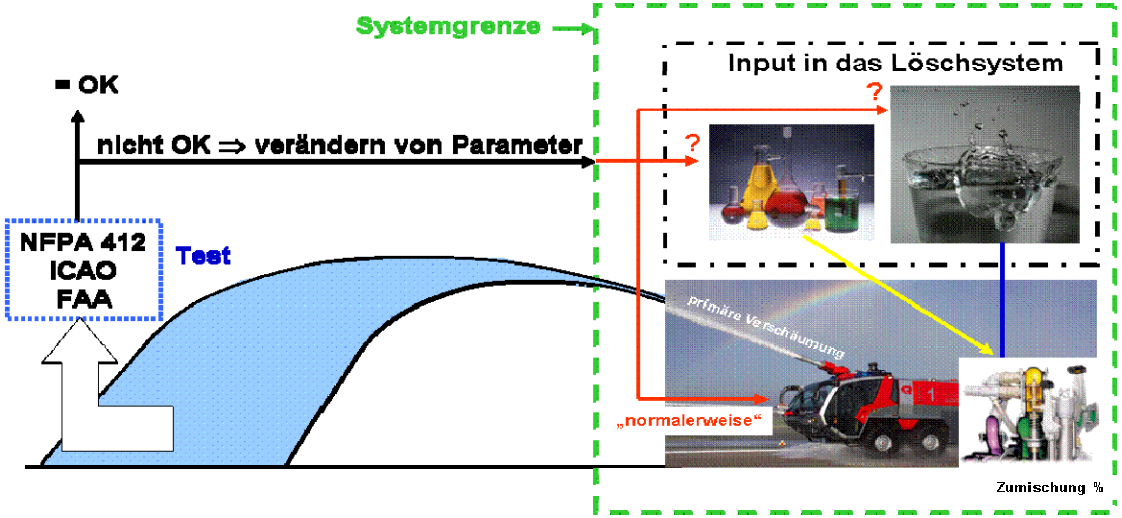
Innovations:

- Results as formulas → Domain experts can analyze, simplify and refine the models
- Integration of prior physical knowledge into modeling process
- Special interest of domain experts in model simplification and variable impact analysis

Example: Extinguishing systems

Goals

- Detect **relevant impact factors** and potential relationships between foam parameters with respect to throw range and foam quality
- **Model throw range and foam quality**
- Configure extinguishing systems for optimal throw range and foam quality



Example: Medical Diagnosis

Motivation:

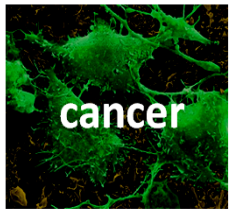
- **Research goal:** Identification of mathematical models for cancer diagnosis
- **Tumor markers:** substances found in humans (especially blood and / or body tissues) that can be used as indicators for certain types of cancer.

Data

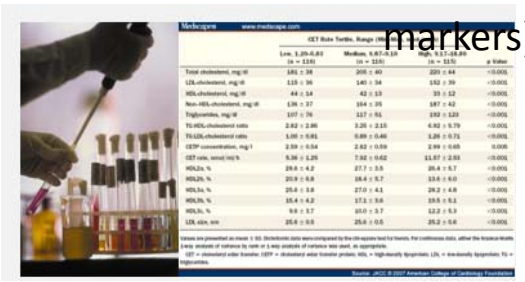
- Medical database compiled at the central laboratory of the General Hospital Linz, Austria, in the years 2005 – 2008
- Total: Blood values and cancer diagnoses for 20,819 patients

Modeling Scenarios

- Model virtual tumor markers using normal blood data
- Develop cancer diagnosis models using normal blood data
- Develop cancer diagnosis models using normal blood data and (virtual) tumor markers



Effects seen in data (blood examinations, tumor markers)



Cancer Diagnosis Estimation

Example: Medical Diagnosis: Data Preprocessing for Cancer Prediction

Patient	Date	[…Measured values…]	Diagnosis
1001	01.01.2004	[…]	-
1001	01.01.2005	[…]	-
1001	01.01.2006	[…]	-
1001	29.05.2007	[…]	-
1001	01.06.2007	[…]	-
1001	02.06.2007	[…]	-
1001	15.06.2007	[…]	C61
1001	02.07.2007	[…]	C61
1001	15.07.2007	[…]	C61
1001	02.09.2007	[…]	C61
1001	01.01.2008	[…]	C61
1001	05.02.2008	[…]	-
1001	01.03.2008	[…]	-
1001	17.03.2008	[…]	-
1001	01.01.2009	[…]	-

Healthy

Relevant measurements

Falsified (treatments, …)

Healthy (?)

Example: Medical Diagnosis: Symbolic Classification Ensembles

- Generate numerous forecasting models
- Reduce error caused by varianve
- Robustification of modeling results
- May be combined with various modeling techniques (e.g. random forests)

- May be combined with regression as well as with classification modeling
  - Averaging or prediction values for regression
  - Majority voting for classification

- Basis for enhanced confidence interpretation

Example: Medical Diagnosis: Symbolic Classification Ensembles

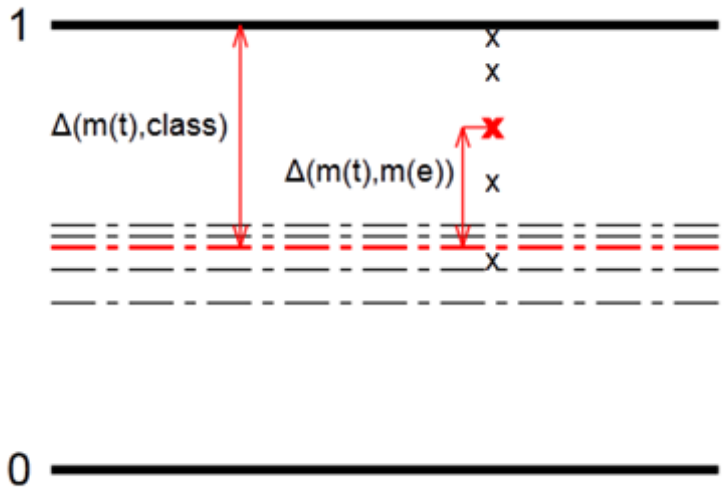
- **Based on clearness of majority voting**

$$cm_1 := 2\left(\frac{|votes(winning\ class)|}{|votes|} - 0{,}5\right)\ \epsilon\ [0, 1]$$

- **Assuming for example 300 models for a 2-class classification problem:**
  - Vote of 300:0 results in $cm_1 = 1.0$
  - Vote of 150:150 results in $cm_1 = 0.0$
  - Vote of 200:100 results in $cm_1 = 1/3$
  - Vote of 250:50 results in $cm_1 = 2/3$

Example: Medical Diagnosis: Symbolic Classification Ensembles

- Based on clearness of voting and on closeness of predictions



$$cm_2 = \min\left(\frac{\Delta(m(t),m(e))}{\Delta(m(t),class)}, 1\right) \in [0,1]$$

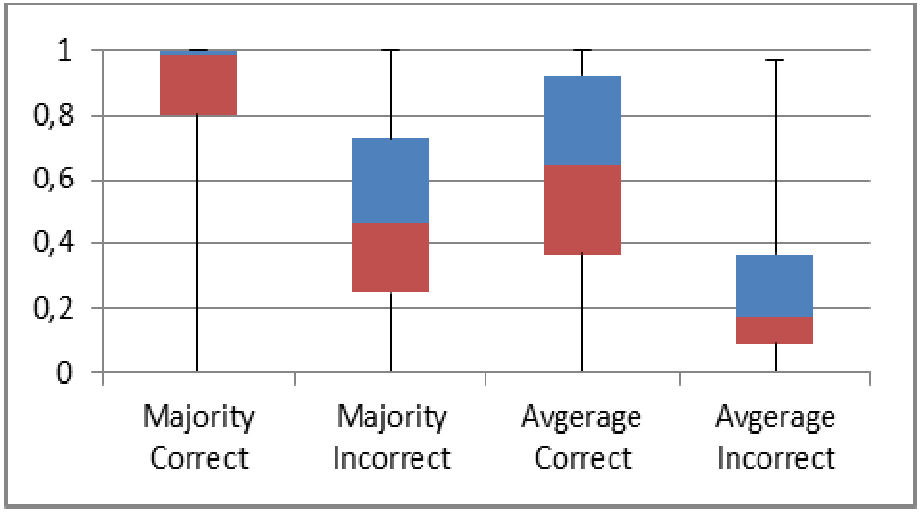
Example: Medical Diagnosis: Symbolic Classification Ensembles

Standard offspring selection GP modeling results

	Avg. training accuracy of 100 best models	Avg. test accuracy of 100 best models	Training accuracy of best model	Test accuracy of best training model
Breast with TM	83.17%	77.89%	84.74%	79.33%

Ensemble modeling results

	Majority Vote	Average Threshold
Accuracy training	84.99%	84.99%
Accuracy test	81.44%	81.44%
Average Confidence Correct Classified	$cm_1 = 0.8500$	$cm_2 = 0.6182$
Average Confidence Incorrect Classified	$cm_1 = 0.4806$	$cm_2 = 0.2449$
Confidence Delta	0.3694	0.3733

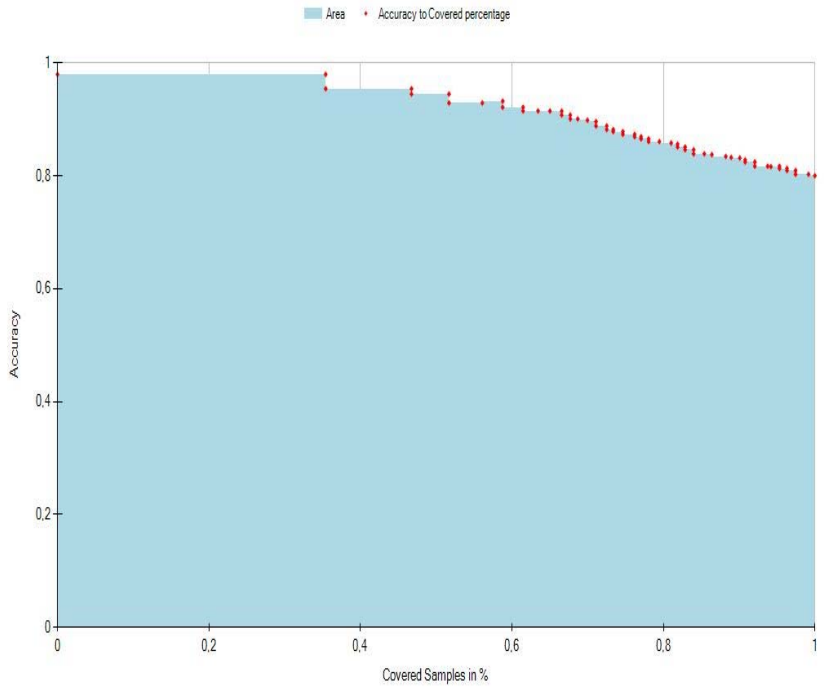


Example: Medical Diagnosis: Symbolic Classification Ensembles

- Test accuracy of best training model for breast cancer: 79.33%
- Test accuracy of ensemble model for breast cancer: 81.44%

Majority Vote

Test accuracy	Covered samples	Confidence
81.72%	95.33%	0.1
83.47%	88.24%	0.3
86.57%	78.05%	0.5
89.88%	69.97%	0.7
92.93%	56.09%	0.9
95.45%	46.74%	0.95
98.00%	35.41%	1



Contact

Prof.(FH) Priv.-Doz. DI Dr.

Michael Affenzeller

Heuristic and Evolutionary Algorithms Lab (HEAL)

FH OOE - School of Informatics, Communications and Media

Softwarepark 11, A-4232 Hagenberg

Phone: +43 (0)7236 3888 2031

Fax: +43 (0)7236 3888 2099

Email: michael.affenzeller@fh-hagenberg.at

Web

HEAL: http://heal.heuristiclab.com/

Josef Ressel Centre Heureka!: http://heureka.heuristiclab.com/

HeuristicLab: http://dev.heuristiclab.com/