57. Computational Intelligence in Industrial Applications

Ekaterina Vladislavleva, Guido Smits, Mark Kotanchek

In this chapter, we review the progress and the impact of computational intelligence for industrial applications sampled from the last 10 years of our personal careers and areas of research (all authors of this chapter do computational modeling for a living). This chapter is structured as follows. Section 57.2 introduces a classification of data-driven predictive analytics problems into three groups based on the goals and the information content of the data. Section 57.3 briefly covers most frequently used methods for predictive modeling and compares them in the context of available a priori knowledge and required execution time. Section 57.4 focuses on the importance of good workflows for successful predictive analytics projects. Section 57.5 provides several examples of such workflows. Section 57.6 concludes the chapter.

57.1 Intelligence and Computation

Developments in computational intelligence (CI) are driven by real-world applications. Over the years a lot of CI has become ubiquitous to the average user and is deeply interwoven into the way modern design, research and development is done.

In our view, CI is human intelligence assisted and (dramatically) enhanced by computational modeling. Intelligence is the capability to predict, and, in theory, there are two directions to get to prediction through computing – data-driven modeling and first principle modeling. In reality though, since even fundamental models and theories have to be validated by data, everything is data driven. For this reason, from now on we will focus on data-driven computational modeling, and say that it exists to enhance predictive capabilities of the human or business. While prediction is the ultimate goal and computational modeling is the means to achieve this goal, we will use concepts of predictive analytics and (data-driven) computational modeling as if they were the same.

Computational modeling methods allow us to generate various hypotheses about a specific problem based on observations in an objective way. The mental models that the scientists develop during this process help them to filter through these hypotheses and come up with new experiments that either support or falsify some of the previous hypotheses or lead to new ones. This process supports the scientific method and significantly accelerates technological development and innovation.

There are many examples of new computational methods empowering problem solving in the areas of material science, energy management, plant optimization, sensory evaluation science, broadband technology, social science (economic modeling), infectious disease prevention, etc. And while success in many cases is undeniable, two main challenges still remain.

First, there is an education gap to bridge before modern CI techniques can reach their full potential, are widely accepted, and become as natural as performing experiments in the lab. While many engineering educational programs are embracing these techniques and help raise awareness of the useful methods in data-driven modeling and computational statistics, the majority of programs in pure sciences tend to ignore them for the most part. There is still a considerable (psychological, cognitive, educational) barrier for experimental scientists – biologists, chemists, physicists, computer scientists – to fully exploit the potential of CI. People will happily save an hour of computing time by spending an additional week in the lab, while in some cases it makes much more sense to spend a week of computing time to spare one experiment in the lab (consider, e.g., an expensive car *crash-test*). We appeal to educational programs to nurture the interest in computation among graduates and facilitate the joint projects of academia with industry targeted at the use and further development of computational intelligence methods for real-world problems.

Second, there is a development gap in the production of scalable off-the-shelf CI algorithms. The parallelization bottleneck seems to affect most CI methods when they are executed on massively parallel architectures. The fact that computational advances in hardware (exa-scale computing) happen at a much faster pace than advances in the design of scalable CI algorithms raises the question: *Up to which moment can we get more intelligence, i. e., more predictive capability, with more computational power?*

57.2 Computational Modeling for Predictive Analytics

While many barriers remain in improving the incorporation of CI in classical education, in solving the new (previously unthinkable) challenges, and in further innovating CI technology, the current time is a perfect moment to make this happen.

First of all, the realization for the indispensability of CI across all industries and all sciences grows as does the number of required CI practitioners (computational statisticians, data scientists, modelers). The report of *Manyika* et al. on Big Data [57.1] predicts a potential gap of 50–60% (300 000 people) in demand relative to the supply of well-educated analytical talent in the USA by 2018. The *data science* and *big data* movement have grown in the last decade to become a buzz-word omnipresent in scientific magazines, technology reviews, and business offerings.

While we are happy that the attention of the average user is being focused on the importance and impact of computational modeling, we are also concerned with the fact that too many details are omitted and almost everything (business strategies, CI methods, targets for predictive analytics, etc.) gets thrown onto one pile.

While Big Data is occupying the minds of future engineers, data scientists, and business majors as a *next big thing to watch* and a synonym of predictive analytics, we want to balance the story some more and provide a full picture of what we think constitutes predictive analytics by computational modeling. While business and industry is striving to become data driven these days, it seeks CI strategies to compete, innovate, and capture value. Success and impact of CI will be generated only if the right strategies are used in the right place.

Success of CI in industry will be awarded to methods that create impact measured in attaining the new level of understanding and knowledge, in units of dollars. In Fig. 57.1 we sketch a relation between the degree of intelligence and the level of competitive advantage from [57.2]. Further on, we will use the terms predictive analytics and computational modeling (for predictive analytics to sustain human intelligence) as if they were the same.

We distinguish three pillars of computational modeling for predictive analytics: business analytics applied to big data (millions to billions of records, dozens to hundreds of variables), process analytics applied to medium-sized data (tens of thousands of records, hundreds of variables), and research analytics applied to precious data (tens to thousands of records, dozens to hundreds to thousands of variables) (Fig. 57.2).

57.2.1 Business Analytics

Business analytics is the part of predictive analytics associated with big data. In recent years, other sci-

ences also created big data problems, so the field could be called big data analytics. The distinguishing feature of business analytics is the fact that it is used to inspect big data streams to provide a quick and simple analysis with immediate value reliably and consistently. Because of the size, big data already offers tremendous challenges in stages preceding analytics – in storage, retrieval, and visualization. These imply that the predictive goals can only be modest, except when big computing facilities and specialized data bases are available (like it happens in environmental and biological research, Internet search, smart grids, etc.). Main goals here are:

- Visualization (e.g., dashboards).
- Recommendation (e.g., studying customer habits and preferences to recommend a new suitable product item). Recommendation uses network analysis to select relevant or similar items.
- Identification of (simple) trends to enhance customer experience and increase surplus. Trends are typically found using time series analysis.
- Binary classification to distinguish out-of-the-ordinary data points from the prototypes following the trends (credit risk analysis, fraud detection, spam identification).

Because of the memory limitations, the challenge in business analytics is to quickly give an answer to simple questions with the main focus on algorithms for in- and out-of-memory computation and visualization. Industries benefitting most from business analytics are retail, banking, insurance, health-care, telecommunications, and social networks.

For example, at large multinational manufacturing companies, business analytics predominantly revolves around the multivariate forecasting of supply and demand. The expected prices and volumes of feedstocks and raw materials as well as the expected demand for various products are important to minimize risk and optimize production as well as the supply chain. Classical statistical forecasting techniques are the main workhorse for this area and the main challenges consist of being able to gather the required data, dealing with possibly large numbers of candidate inputs and outputs for the models and properly dealing with the hierarchies that exist, e.g., products-markets-industry resulting in an explosion in the number of models that have to be built and maintained.

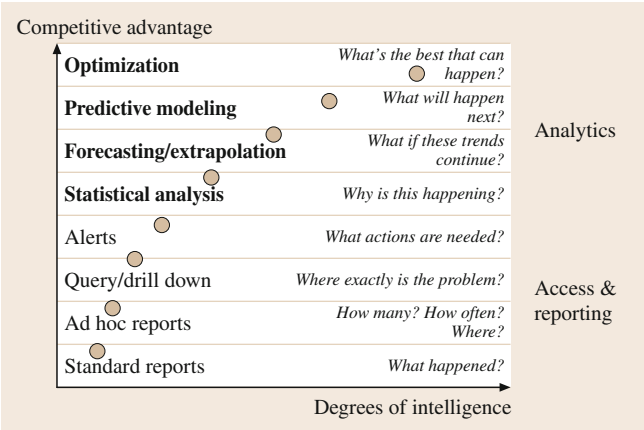

Fig. 57.1 *Davenport* and *Harris* [57.2] have wonderfully adapted the graphics from SAS software. The *graph above* eloquently explains why to use predictive modeling to excel, compete, and capture value

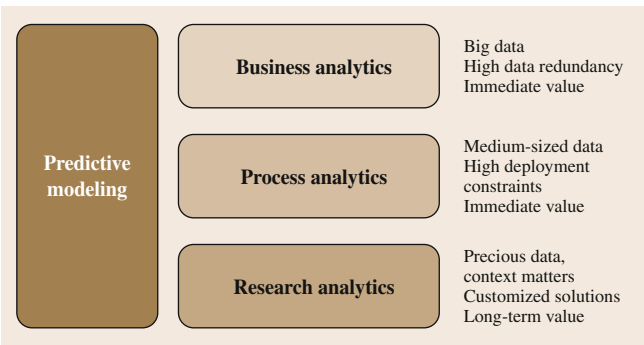

Fig. 57.2 Predictive modeling has three components: Business analytics, predictive analytics, and research analytics

57.2.2 Process Analytics

Process analytics exploits medium-size data to generate time-sensitive prediction of an industrial process (e.g., manufacturing, process monitoring, remote sensing, etc.) with immediate value.

Process analytics models must be very robust, simple (mostly linear), and concise to be deployed in real industrial processes.

This well understood and probably most conservative area of predictive analytics has experienced a big change in the last years. A couple of decades ago, process optimization and control groups had more people and less pressure. Nowadays, pressure for integrating production workflows has increased together with the

need to meet tighter quality specifications, much tighter emission thresholds, requirements to reduce production, operation, and energy costs, and to maximize throughput. The sensor's side has changed – sensors have become much more sophisticated and much more abundant. The human interference has also decreased due to (sometimes exaggerated) drive to automation, and cost reduction.

All these factors have dramatically increased the demand for reliable optimization and control. In general, process analytics models must be very robust, simple (mostly linear) and concise to be deployed in real industrial processes. The main challenge for this industry is to integrate more sophisticated models and adopt new computational methods for process analytics to adapt to the changing world of new requirements while maintaining robustness over a wide process range. At the time when this chapter was written data-driven modeling was still considered exotic for the field of process analytics, and model deployment is still heavily constrained.

The main goals in process analytics are process forecasting and process optimization and control.

The challenge in process forecasting is to build models that hit the tradeoffs between model interpretability and their long-term (real-time) predictive power. The technological challenge of successful CI methods is the capability to identify driving features in a large set of correlated features. For example, think of a problem of predicting the quality of a manufactured plastic using the smallest subset of available factors controlling the production process – pressures, temperatures, flows. Robust feature selection is as important as good prediction accuracy – models that are too bulky will never be accepted by process engineers.

The main challenge in process control is the multiobjective nature of control specifications and subsequent optimization problems. Consider an example of manufacturing and wholesaling thin sheets of metal. The thickness of the sheet is an important quality characteristic that should not fall below a predefined minimum, or the product will be considered off-spec. If due to the processing condition the thickness variability is high (sheets are several meters wide and tens of meters long, rolled at high speeds, high temperatures), penalty for off-spec material is high, and costs for raw steel are also high – the manufacturer faces a delicate problem of making the sheet thicker than the allowed minimum to keep the clients happy but not too thick to keep the production costs down. These competing objectives usually require a multiobjective approach to process optimization.

Process analytics relies on a rich data set coming from the many sensors in a typical plant. Mature platforms exist that store this, often high-frequency, sensor data in databases and plant information systems. The primary intent for this data is to run the various plant control and quality control systems but archived data are often available for predictive modeling as well. The use of models that predict the aging and lifetime of catalysts and the associated changes in optimal settings for the plant are good examples.

Another example is the building of the so-called soft sensors that link difficult measurements, such as, e.g., grab samples that need to be brought to the lab for analysis with results only becoming available after some time to some of the easier high-frequency measurements, such as, e.g., temperatures, flows, and pressures. These models then serve as substitutes for the difficult measurements at a high frequency and can be calibrated if needed when the slow measurements become available. There are also many opportunities to use the demand and supply forecasting models from the business analytics side to optimize production and product mix that is most optimal for a given scenario. As an extreme example, it may be cheaper to shutdown a plant for a while vs continued production when demand is forecasted to be very weak. The level and amount of coupling that is possible between demand–supply forecasts and actual production can vary significantly and depends on many factors, but it is clear that much more is possible in this area.

Examples of industries employing process analytics are manufacturing, chemical engineering, energy, environmental science.

57.2.3 Research Analytics

Research analytics is used to accelerate the development of new products and systems. This task is fundamentally different from all the ones mentioned previously as it is usually applied to small, complex, and precious data, is heavily dependent on problem context and provides long-term value without immediate rewards. (By *small* we mean any data set where the number of record is comparable or even smaller than the number of features. In this way, gene expression data with thousands of variables taken over dozens of individuals is small.)

Research analytics provides very customized solutions and requires a close collaboration between analysts/modelers and subject matter experts.

Research analytics is by nature much less generic and becomes very dependent on the specific product that is being developed. In research, once you have predictive analytics, then there is only a small step to make from optimization of existing products to the design of new ones. One example of a research analytics success story is the development of an application to predict the exact color of a plastic part based on the composition of the colorants and the specific grade a plastic being used, see [57.3]. Robust color prediction models led to the capability of actually designing colorant compositions in silico directly from customer specifications. The models also provided the specifications that were necessary to even let the customer produce that part himself.

How far one is able to take this depends on the fidelity of the models as well as the quality of the data that is available. Another example of research analytics at work is the design of new coatings and catalysts based on high throughput experimentation where all the available data is being used to build models on the fly. These models are than used to design the next experiments such that the information gain is maximum. The requirements for the modeling process are quite high because everything is embedded in a high-throughput workflow but the benefits are also huge. Significant speedups in the total design time as well as the performance of new products can be achieved this way.

We stress that because research analytics is an enhancement to human intelligence for the development of new products and systems, the benefits of its application scale proportionally with the size of the problem and the impact of that particular product or system. For big enough problems the benefits quickly get into the hundreds of thousands to millions of dollars.

Research analytics can help drive innovation in all industry segments, particularly in materials science, formulation design, pharmaceuticals, engineering, simulation-based optimization in research, bio-engineering, healthcare, telecommunications, etc. In the coming 10 years, we will continue to see the trend of innovation enabled to a large extent by predictive modeling.

57.3 Methods

Over many years of exercising process and research analytics, we built up a practice of using predictive modeling as the integration technology for real-world problem solving. In the last 8–10 years, predictive modeling for computational intelligence has evolved from the solution of last resort to the main stream approach to industrial problem solving (prediction, control, and optimization). It is technology that glues together fundamental modeling and domain expertise, high-performance computing and computer science, empirical modeling and mathematics – a heaven for an inquiring mind and interdisciplinary enthusiast.

Predictive modeling is a bridge that connects theory and facts (data) to enable insight and system understanding. The theory for poorly understood problems is often based on simplifying assumptions, on which the fundamental models are built. The facts, or empirical evidence, are often affected by high uncertainty and a limited observability of the system's behavior.

Predictive modeling applied iteratively to a growing set of facts tests the theory against the data and *extrapolates* models build on the data to confirm or adjust the theory until the theory and facts start to agree. The validation always lies in the hands of a subject matter expert who in the case of success accepts both the theory and the designed predictive models as plausible and interesting. While the real validation comes when models are deployed and keep generating value, without the first step of intriguing the subject matter expert the project does not have a chance to succeed.

To clear any obstacles toward the acceptance of models by the domain expert the models should be:

1. Interpretable
2. Parsimonious
3. Accurate
4. Extrapolative
5. Trustable, and
6. Robust.

In an industrial setting, the capability of having a trustable prediction of the output within and outside the training range is as important as interpretability and the possibility of integrating information from first principles, low maintenance and development costs with no (or negligible) operator interference, robustness with re-

spect to the variability in data, and the ability to detect novelties in data to attune itself toward changes in system's behavior.

There is no single technique producing models that are guaranteed to fulfill all of the requirements above, but rather there is a continuum of methods (and hybrids) offering different tradeoffs in these competing objectives.

Commonly used predictive modeling techniques include linear regression, and nonlinear regression [57.4], boosting, regression random forests [57.5], radial-basis functions [57.6], neural networks [57.7], support vector machines (SVM) [57.8, 9], and symbolic regression [57.10, 11] (see more in [57.12]).

In Fig. 57.3, we place some of the most common methods for predictive modeling for process and research analytics in the objective space of development time versus the level of a priori knowledge about the problem. When identifying which methods to use other objectives (like interpretability and extrapolative capability) must also be taken into account. The time axis is depicted on a log scale, and the exact development time depends on implementation and a particular algorithm flavor.

Support vector machines and ensemble-based neural networks lose to linear, nonlinear, and regularized regression in interpretability, but have advantages for problems where little a priori information is known about the system, and no assumptions on model structures can be made (see Fig. 57.3).

Regression random forests, and symbolic regression [57.13–15] have further advantages for problems where not only model structures but also the variable drivers (significant factors) are unknown.

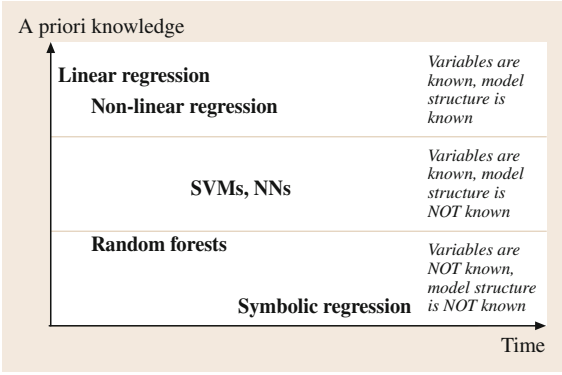

Fig. 57.3 Predictive modeling methods as competing tradeoffs in development time versus the level of a priori knowledge about the problem

Random forests proved to be robust and very efficient for predicting the response within the training range and for identifying the most significant variables, but because they do not possess extrapolative properties they can only be used in problems where no extrapolation is necessary. Recent studies [57.16] indicate that variable selection information obtained by random forests can loose meaning if correlated variables are present in the data and affect the response differently.

In business analytics, when the speed of model development is the main goal, linear regression and regularized learning are the only remaining options. (Recent developments for predictive modeling for big data are also focusing on the feature generation problem, where the set of original data variables gets expanded to a much larger set of new features – transformations of the original variables, for which regularized linear regression is applied. Much like in support vector regression).

In process analytics when the driving input factors are known – ensemble-based neural networks, support vector regression, and ensemble-based symbolic regression are the modeling alternatives.

If very little is known about the process or system, experiments are demonstrating correlations among input variables, and concise interpretable models are required – symbolic regression is the only resort, which comes at a price of a higher development time (Fig. 57.3).

We stress the importance of using ensembles of predictive models irrespective of which modeling method is used. Ensemble disagreement used as a trustability measure defines the confidence of prediction and is crucial for reliable extrapolation. (It cannot be stressed enough that all prediction in a space of dimensionality above 3 is mostly extrapolation, even when evaluated inside the training range.)

We deal mostly with process and research analytics. In our experience, the aspect of trustablility via ensembles of global transparent models, coupled with the massive algorithmic efficiency gains and the ability to easily handle real-world data with spurious and correlated inputs has led to symbolic regression largely replacing neural networks and support vector machines in our industrial modeling. Our experience also is that symbolic regression models tend to extrapolate well as well as provide warning of that extrapolation.

The reason for symbolic regression being successful for process and research analytics is the fact that all real-world modeling problems we have seen up to now contained only a dozen of relevant inputs (never more

than 25 variables, in most cases less than 10) which were truly significantly related to the response. Because symbolic regression searches for plausible models in a space of all possible structures from the given set of potential inputs, and allowed functional transforms, the computational complexity increases nonlinearly with the dimensionality of the true design space. For this reason, symbolic regression effortlessly identifies dozens of driving variables among tens to hundreds of candidates (albeit using hours of multicore computing time). But it should not be used for problems where hundreds of inputs are significantly related to the response and should be filtered out of thousands of candidates. We claim though that no methods are available to solve the latter kind of problems because the necessary amount of data capturing true input–output relationships will never be collected.

Although tremendous progress has been made over the past decade in terms of the efficiency and quality of symbolic regression model development, we also have made corresponding advances from a holistic perspective encompassing the overall modeling workflows from data collection through model deployment.

57.4 Workflows

Although there is no universal solution for predictive modeling and no size fits all, especially for research analytics, nothing is as important for a successful solution as a good modeling workflow.

We would like to make a case for the utmost importance of workflows and the need to nurture and actively proliferate them through all CI projects. In the next section, we give an example on how a successful workflow developed in a project from flavor science could be seamlessly applied to a project in video quality prediction. And because predictive modeling for CI will soon be used in nearly all industry segments and research domains, we believe that it is the responsibility of CI practitioners to facilitate innovation through proliferation and popularization of (interpretable) workflows allowing straightforward application in new domains.

The most general approach to practical predictive modeling is depicted in Fig. 57.3.

We view this generic framework as an iterative feedback loop between three stages of problem solving (just as it usually happens in real-life applications):

1. Data generation, analysis and adaptation
2. Model development, and
3. Problem analysis and reduction.

An important observation is made in the *Toward 2020 Science* report edited by *Emmott* and *Rison* [57.17]:

> *What is surprising is that science largely looks at data and models separately, and as a result, we miss the principal challenge – the articulation of modelling and experimentation. Put simply, models both consume experimental data, in the form of the context or parameters with which they are supplied, and yield data in the form of the interpretations that are the product of analysis or execution. Models themselves embed assumptions about phenomena that are subject of experimentation. The effectiveness of modeling as a future scientific tool and the value of data as a scientific resource are tied into precisely how modelling and experimentation will be brought together.*

This is exactly the challenge of predictive modeling workflows – a holistic approach to bring together data, models, and problem analysis into one generic framework. Ultimately, we want to automate this iterative feedback loop over data analysis and generation, model development, and problem reduction as much as possible, not in order to eliminate the expert, but in order to free as much thinking time for the expert as possible.

This philosophical shift away from human replacement in the modeling workflow toward human augmentation has been very important in the last decade. A successful workflow must offer suites which mine the developed models to identify driving factors, variable combinations, and key variable transforms that lead to insight as well as robust prediction.

57.4.1 Data Collection and Adaptation

Very often, especially in big companies, and especially for process analytics, CI practitioners do not have access to data creation and experiment planning. This gap is a typical example of a situation, where multivariate data is given and there is no possibility to gather better sampled data.

In other situations, there is a possibility to plan the experiments, and gather new observations of the response for desired combinations of input variables, but the assumption always is that these experiments are very expensive, i. e., require long computation, simulation, or experimentation time. Such a situation is most common in research analytics and meta modeling for the design and analysis of simulation experiments.

The questions to ask at the data collection and adaptation stage are: How to design experiments within the available timing and cost budget to optimally cover the design space (possibly containing spurious variables)? How can available data and developed models guide design-space exploration in the next iterations? Is available data well sampled? Is it balanced? What is the information content of performed experiments? Is there redundancy in the data and how to minimize it?

57.4.2 Model Development

In model development, the focus is on automatic creation of collections of diverse data-driven models that infer hidden dependencies on given data and provide insight into the problem, process, or system in question.

Irrespective, of which modeling engines are used at this stage, the questions on how to best generate, evaluate, select, and validate models given particular data features (size and dimensionality) are of great importance. Model quality, in general, i. e., generalization, interpretability, efficiency, trustworthiness, and robustness is the main focus for model analysis leading to the next stage.

57.4.3 Problem Analysis and Reduction

The stage of problem analysis and reduction supposes that developed models are carefully scrutinized, filtered, and validated, to infer preliminary conclusions on problem difficulty. The focus is on driving inputs, assessment of variable contribution, linkages among variables, dimensionality analysis, and construction of trustable model ensembles. The latter if defined well will contribute to intelligent data collection in the style of active learning.

With a goal to augment human intelligence by computation, we emphasize the critical need for a human, an inquiring mind who will test the theory, the facts (data) and their interpretations (models) against each other to iteratively develop a convincing story where all elements fit and agree.

57.5 Examples

57.5.1 Hybrid Intelligent Systems for Process Analytics

A good example of a unified workflow for process analytics is the hybrid intelligent systems framework popularized at the Core R&D department of the Dow Chemical Company in the late 1990s.

The methodology was developed to improve soft sensor performance (performance of predictive models), to shorten its development time, and minimize maintenance. It employed different intelligent system components – genetic programming, support vector machines, and analytic neural networks [57.18].

The process analytics in this framework consists of three steps following data collection:

1. Data preprocessing and compression. Support vector regression using the ϵ-insensitive margin is used to identify and remove data outliers and compress data to a representative set of prototypes (support vectors). The result is a clean and compressed data set.
2. Preliminary variable selection using ensemble-based stacked analytic neural networks [57.19]. The result of this step is a ranking of input variables and quantification of variable contribution based on iterative input elimination and re-training.
3. Convolution parameter estimation to identify relevant time-lags of significant inputs using appropriate convolution functions.
4. Development of transparent predictive models using symbolic regression via genetic programming and final variable selection using symbolic regression models.
5. Model selection and analytical function validation.
6. Online implementation.
7. Soft sensor maintenance to guarantee robustness of process prediction.

Examples of the use of this workflow for reactor modeling can be found in [57.18].

We all practiced the hybrid intelligent systems workflow in the past, but the massive algorithmic effi-

ciency gains in ensemble-based symbolic regression via genetic programming of the last decade [57.14, 20] have led us to simplify the workflow and largely eliminate steps one and two to replace them by direct application of symbolic regression.

57.5.2 Symbolic-Regression Workflow for Process Analytics

The major modeling engine breakthrough was the incorporation of a multiobjective viewpoint; this introduced orders of magnitude improvements in model development speed while simultaneously allowing the analyst to choose the proper balance between complexity and accuracy post facto. In essence, the data could now define the appropriate model structure and driving inputs, which became the main reason for symbolic regression's success for predictive modeling.

Other conceptual advances ordinal genetic programming, interval arithmetic, Lamarckian evolution and secondary optimization objectives, such as age, model dimensionality, nonlinearity, etc., have brought us to the current situation where we can largely inject data into a (properly designed) symbolic regression engine and interesting and useful models will emerge.

The symbolic regression workflow has become as depicted in Fig. 57.4, but with model development done using Pareto-aware symbolic regression [57.14].

Distillation Tower Example

The dataset comes from an industrial problem on modeling gas chromatography measurements of the composition of a distillation tower and is available online at http://www.symbolicregression.com.

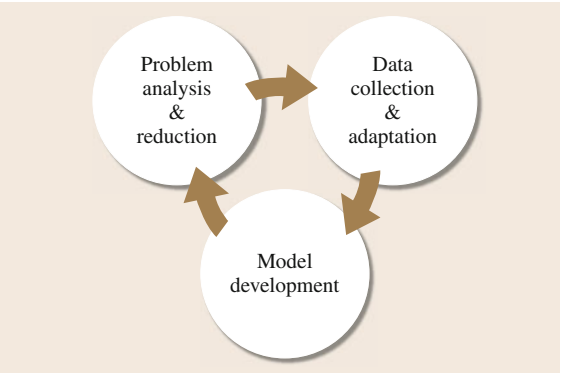

Fig. 57.4 Generic iterative model-based problem solving workflow (after [57.15])

A chemical reaction typically generates a variety of chemicals along with the one (or several) of interest. One method of isolating the mixture coming from the reactor into various purified components is to use a distillation column. The (hot gaseous) input stream is fed into the bottom and on the way to the top goes through a series of trays having successively cooler temperatures. The temperature at the top is the coolest. Along the way, different components will condense at different temperatures and be isolated (with some statistical distribution on the actual components). With vapors rising and liquids falling through the column, purified fractions (different chemical compounds) can be retrieved from the various trays. The distillation column is very important for the chemical industry because it allows continuous operation as opposed to a batch process and is relatively efficient.

This distillation column problem contains nearly 7000 records and 23 potential input variables – mixture of flows, pressures, and temperatures – in addition to the quality metric and material balance. The response variable is the concentration of a purified component at the top of the distillation tower. This quality variable needs to modeled as a function of relevant inputs only. The range of the measured quality metric is very broad and covers most of the expected operating conditions in the distillation column.

To design the test data, we sorted the samples by their response values and selected every third and seventh samples for the validation set and every fourth and eight samples for the test set. The remaining points formed the training set.

Many input variables in the data are heavily correlated. Because symbolic regression can deal with correlated variables, we used all 23 inputs in the first round of modeling to perform initial variable importance analysis.

The workflow that follows exploratory data analysis is described below:

1. *Initial modeling*: We allocated 2 hours of computing time on a quad-core machine to perform 24 20-minute independent runs of symbolic regression by genetic programming using Evolved-Analytics' DataModeler [57.14]. All symbolic regression runs used basic arithmetic operators augmented by a negation and a square as primitives. All models were stored on disk, and all other settings set to default settings of the symbolic regression function of [57.14]. In total, more than 3000 symbolic

regression models were generated during 24 independent runs.

2. *Variable importance analysis*: For all models presence-based importances were computed. Figure 57.5 demonstrates that only a handful of variables is identified as drivers ([57.14] suggests to use importance threshold of 20%).
3. *Variable combination analysis*: All developed models were analyzed for dimensionality and most frequent variable combinations. In Fig. 57.6, one can see model subsets niched according to constituting variable combinations. The bottom graph suggests that variables colTemp1, colTemp3, and colTemp5 might be sufficient for describing the response, since they cover the *knee* of the Pareto front in complexity vs. accuracy space.
4. *Variable contribution analysis*: Models were simplified by identifying and eliminating the least contributing variable. Variable combination analysis was repeated for simplified models and resulted in identifying colTemp1 and colTemp3 as new candidates for a sufficient subspace.
5. *New runs performed on a subset of input variables identified as drivers*: The new batch of independent symbolic regression runs was applied to the same data but only using colTemp1 and colTemp3 as the candidate input variables. As expected, models generated in this experiment demonstrated that the same complexity–accuracy tradeoffs can be achieved in only two-dimensional rather than 23-dimensional input space.
6. *Ensemble generation using developed models and a validation set*: Final model ensemble was generated automatically using developed symbolic regression models and validation data set. It was augmented by quadratic and cubic models on two variable drivers.
7. *Ensemble prediction validation using test data*: Ensemble prediction and ensemble disagreement were finally evaluated on the test data. Initial requirements for prediction accuracy to not exceed 5–7% of standard deviation were met by all ensemble models. Ensemble prediction is graphed in Fig. 57.7.

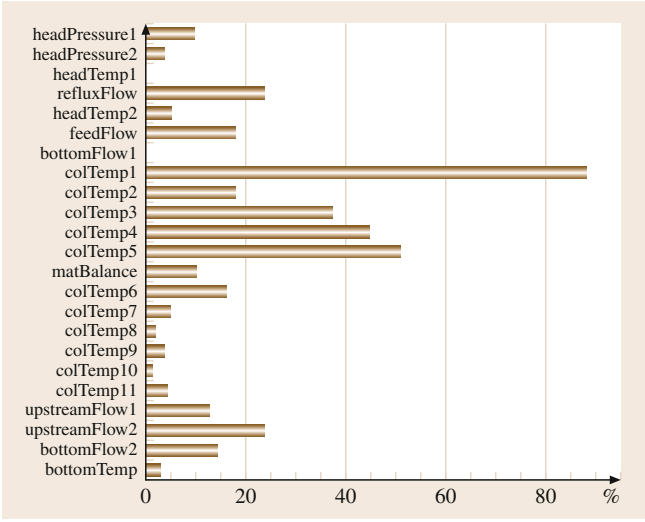

Fig. 57.5 Variable presence in developed symbolic regression models

This example demonstrates the use of a good model development workflow. An ensemble similar to the one described here has been deployed for controlling a gas chromatography measurement in a real distillation column.

57.5.3 Sensory Evaluation Workflow for Research Analytics

A flavor design case study is an example of a more specialized workflow [57.21]. In sensory evaluation, scientifically designed experiments are used to define a small set of mixtures that can be presented aromatically to evaluators to identify the ingredients that drive hedonic response (positively or negatively) of a target panel of consumers. Each panelist is asked how much they like the flavor, ranging from like extremely to dislike extremely with 9 distinctions. Details of the study can be found in [57.21]. Our focus here is the workflow that allowed to evaluate the consistency of liking preferences in the target population and gain insight into how to design or identify flavors that most consumers would consistently like.

The data for this project was provided by the Givaudan Flavors Corp. It falls into a category of precious data. It consists of sensory evaluation scores of 36 mixed flavors containing seven ingredients evaluated by 69 human panelists. In other words, data has seven input variables (flavor ingredients), 36 records (flavors), and 69 response measurements per record (Fig. 57.8).

Because of the high variability of response values per flavor, panelist responses were modeled individually. Because transparent and diverse input response models were required to approximate this challeng-

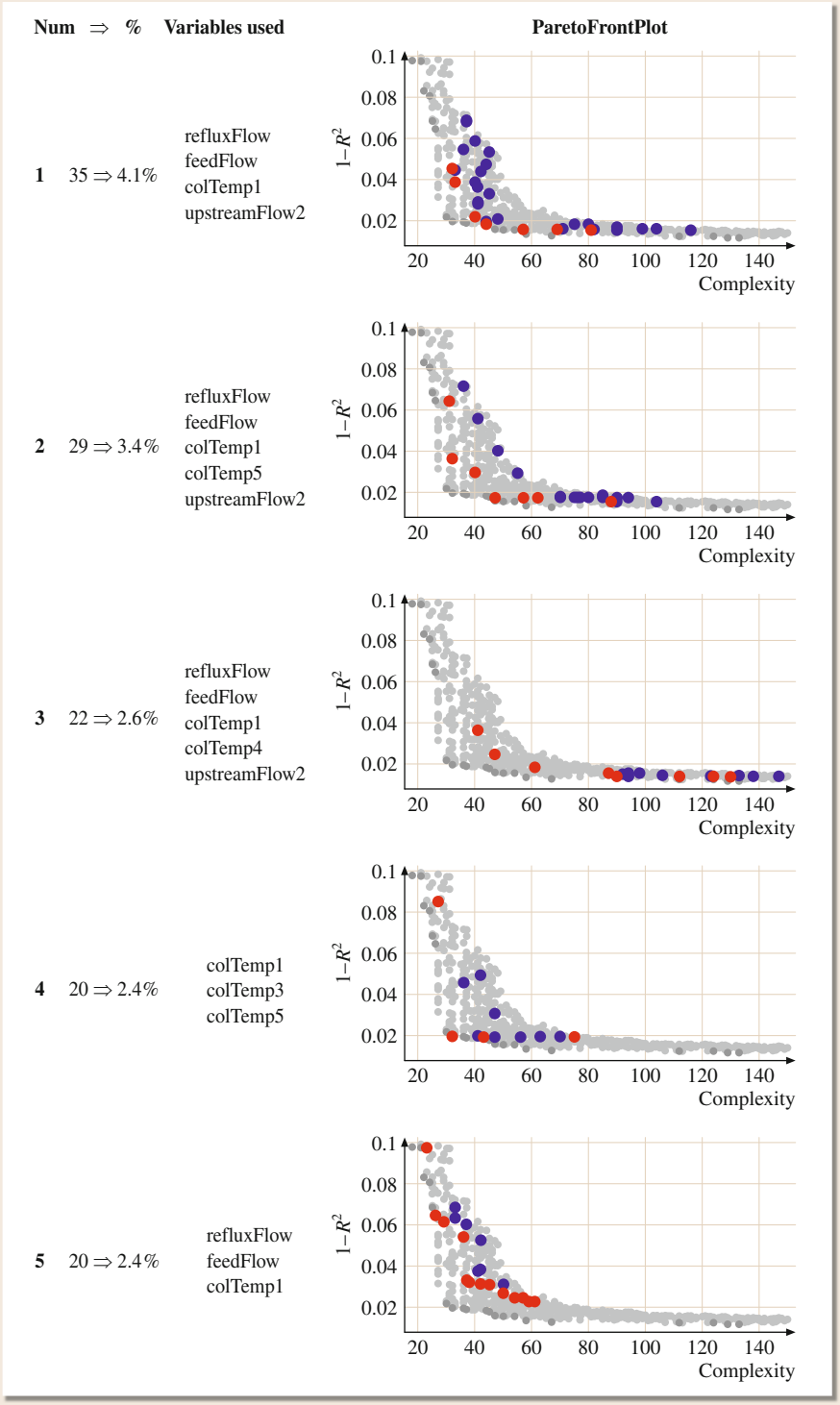

Fig. 57.6 Complexity–accuracy tradeoffs for most frequent variable combinations in the distillation column example

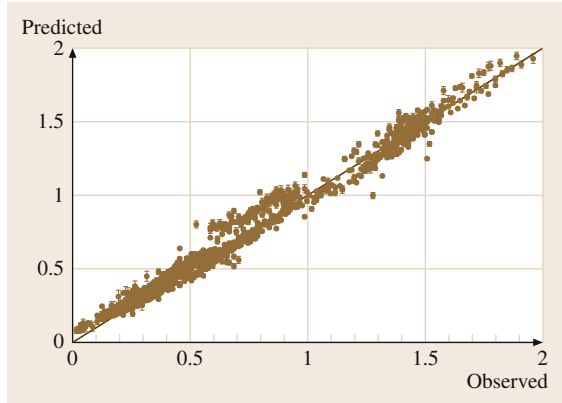

Fig. 57.7 Prediction of the final ensemble of symbolic regression models on test data. All models seem to agree on unseen test data set. This should not be surprising, because the training, validation, and the test set were designed to cover the full range of operating conditions ►

ing data set, modeling was done using ensemble-based symbolic regression.

For each panelist, a standard workflow was applied to identify driving ingredients which changes in panelist's liking [57.22].

When developed, model ensembles predicting individual responses could be bootstrapped to a richer set of virtual mixtures (tens of thousands of flavors instead of the available 36). The bootstrapped responses

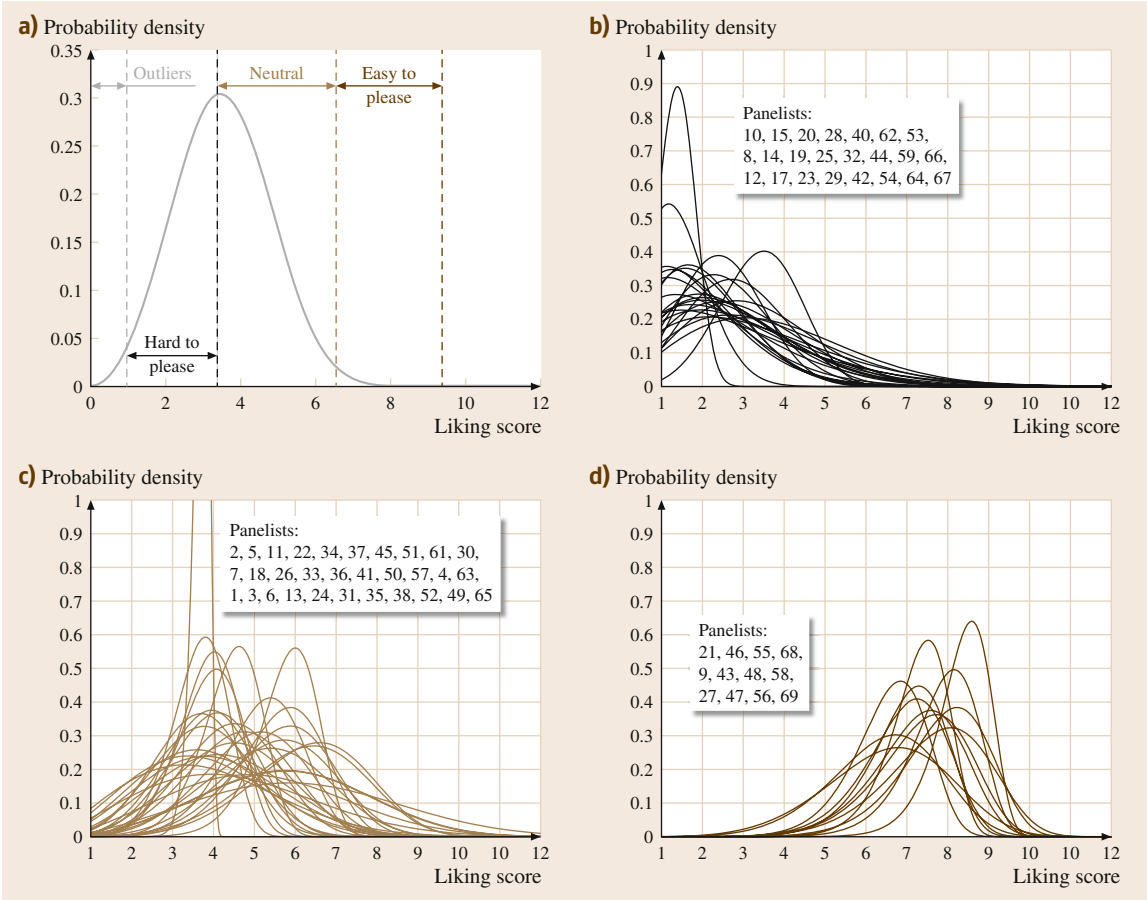

Fig. 57.8a–d Example of panel segmentation by propensity to like from [57.22]: (**a**) Decision regions for evaluating cumulative distribution for liking score density model (**b**) hard to please panelist (**c**) neutral panelists, (**d**) easy to please panelists

were used to cluster the target population into three segments: easy to please – (cyber)individuals who consistently give high ratings to most flavors, hard to please – individuals that consistently use a low range of scores for all flavors, and neutral panelists whose preferential range is centered around the medium score – *neither like, nor dislike*. Such segmentation of the target population by people's propensity to like products turned out to be very useful in several other applications beyond flavor design. It focuses product development by giving insight into the fundamental variability in the preferences of the target audience.

The standard workflow for variable importance estimation applied to model ensembles forecasting the scores of individual panelists also allowed to segment the target population by ingredients that drive liking in the same direction. Such segmentation of the consumer market combined with the cost analysis for new product design offers visualization and analysis of beneficial tradeoffs for product specialization.

The third outcome of this study was the development of a model-guided optimization workflow for designing optimal virtual mixtures. Multi-objective optimization using swarm intelligence was used to find tradeoffs in the flavor design space that simultaneously maximize the average liking score and minimize variance in the liking across virtual panelists.

Such model-guided optimization workflow combined with the standard ensemble-based modeling workflow presents a strong motivation for the development of a targeted data collection system for designing new products.

We should point out that despite a very custom design and specialized domain of sensory evaluation in food science, the workflow could successfully be applied in the very different domain of video quality prediction. Ensemble-based symbolic regression was used to model the perceived quality of perturbed video frames and results were used to predict customer satisfaction and segment the representative population of video viewers by propensity to notice perturbations and sensitivity to particular perturbations [57.23].

57.6 Conclusions

In this chapter, we discussed how computational intelligence leads to predictive analytics to produce business impact. We identified three main areas of predictive analytics: business analytics that deals mainly with visualization and forecasting, process analytics which aims to improve optimization and control of manufacturing processes, and research analytics which aims at speeding up and improving product and process design. All three areas have the potential to save and earn many millions of dollars but deal with very different data sources, context, information content, amount of available domain knowledge, and time and prediction requirements for value generation. Driven by different motivations, the areas are subsequently employing different predictive modeling methods.

We presented several predictive modeling methods in the context of different prediction requirements, solution development, and deployment constraints. We emphasized that there is no single method that fits all problems, but rather there is a continuum of methods, and each problem dictates selection of a method by specific time requirements and the amount of available a priori subject-matter knowledge (Fig. 57.3).

We stressed the importance of good and stable predictive modeling workflows for success in CI projects and provided several examples of such workflows for process and research analytics, illustrating that research analytics projects require highly customized approaches.

We point out that successful CI projects are amplifiers, that necessarily keep the human in the loop and vastly enhance her/his capabilities. Because of this, integrating CI in the various process and business workflows is essential!

It is clear that our ability to generate data as well as our ability to analyze it and produce actionable knowledge are quickly expanding. The challenge remains on how to develop scalable CI algorithms that keep up with the ever rising tide of data, given that computational advances in hardware (massive parallelization, exa-scale computing) are developing at a much faster pace than the CI algorithms.

A question that still puzzles us is: *Can we get more intelligence with more computational power, and where (and whether) it stops?* Undoubtedly, the right answer lies in the development of new algorithms that can tackle the new challenges – advanced material design,

problems in bio-informatics, complex-system modeling in social sciences, and social networks. We expect the largest impact of predictive modeling to happen in the areas of research and process analytics – in design of new products and new processes. Examples of design problems that can be assisted by data-driven CI methods for research analytics are the development of advanced materials – photovoltaic cells, alternative fuels, bio-degradable replacements for paints and plastics, composite materials, sustainable food sources. From the process analytics side, we would like to see CI methods used for optimization of water purification, emission control in combustion processes, simulation-based optimization of social events on a world scale (terror attacks, revolutions, pandemics spread), efficiency optimization of manufacturing cycles, garbage minimization, and recycling.

It cannot be stressed enough that the dynamics around CI is changing – instead of CI being an optional addition to the arsenal of problem solving tools and methods, CI is becoming indispensable to deal and make progress with this new breed of real-world problems. The only way for CI practitioners to bring CI to prime time is to develop scalable algorithms, proliferate good workflows, and implement them in great applications.

References

57.1 M. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers: Big data: The next frontier for innovation, competition, and productivity, available online at http://www.mckinsey.com/mgi (2011)

57.2 T.H. Davenport, J.G. Harris: *Competing on Analytics: The New Science of Winning*, 1st edn. (Harvard Business School, Boston 2007)

57.3 J.C. Torfs, G.J. Brands, E.G. Goethals, E.M. Dedeyne: Method for characterizing the appearance of a particular object, for predicting the appearance of an object, and for manufacturing an object having a predetermined appearance, which has optionally been determined on a basis of a reference object, WO Patent Ser 20 0204 2750 A1 (2004)

57.4 R.A. Johnson, D.W. Wichern: *Applied Multivariate Statistical Analysis* (Prentice Hall, Englewood Cliffs 1988)

57.5 L. Breiman: Random forests, Mach. Learn. **45**, 5–32 (2001)

57.6 M.D.J. Powell: Radial basis functions for multivariable interpolation: A review. In: *Algorithms for Approximation*, ed. by J. Mason, M.G. Cox (Clarendon, Oxford 1987) pp. 143–167

57.7 S. Haykin: *Neural Networks and Learning Machines*, 3rd edn. (Pearson Educ., Harlow 2008)

57.8 V. Vapnik: *Estimation of Dependences Based on Empirical Data* (Springer, Berlin, Heidelberg 1982)

57.9 V. Vapnik: The support vector method, Proc. 7th Int. Conf. Artif. Neural Netw. (1997) pp. 263–271

57.10 J.R. Koza: *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT, Cambridge 1992)

57.11 R. Poli, W.B. Langdon, N.F. McPhee: *A Field Guide to Genetic Programming* (Lulu, Raleigh 2008)

57.12 A.K. Kordon: *Applying Computational Intelligence: How to Create Value* (Springer, Berlin, Heidelberg 2010)

57.13 W. Banzhaf, P. Nordin, R.E. Keller, F.D. Francone: *Genetic Programming: An Introduction on the Automatic Evolution of Computer Programs and its Applications* (Morgan Kaufmann, San Francisco 1998)

57.14 M. Kotanchek: Evolved Analytics LLC: *DataModeler Release 8.0* (Evolved Analytics LLC, Midland 2010)

57.15 E. Vladislavleva: *Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming* (Tilburg Univ., Tilburg 2008)

57.16 S. Stijven, W. Minnebo, K. Vladislavleva: Separating the wheat from the chaff: On feature selection and feature importance in regression random forests and symbolic regression, Proc. 13th Annu. Conf. Companion Genet. Evol. Comput. (2011) pp. 623–630

57.17 S. Emmott, S.Rison: *Towards 2020 Science*, Microsoft, Cambridge (2006)

57.18 A.K. Kordon, G.F. Smits: Soft sensor development using genetic programming, Proc. Genet. Evolut. Comput. Conf. (2001) pp. 1346–1351

57.19 A.K. Kordon, G.F. Smits, A.N. Kalos, E.M. Jordaan: Robust soft sensor development using genetic programming. In: *Nature-Inspired Methods in Chemometrics: Genetic Algorithm and Artificial Neural Networks*, ed. by R. Leardi (Elsevier, Amsterdam 2003) pp. 69–108

57.20 M. Kotanchek: Real-world data modeling, Proc. 12th Annu. Conf. Companion Genet. Evol. Comput. (2010) pp. 2863–2896

57.21 K. Veeramachaneni, E. Vladislavleva, U.-M. O'Reilly: Knowledge mining sensory evaluation data: Genetic programming, statistical techniques, and swarm optimization, Genet. Progr. Evol. Mach. **13**(1), 103–133 (2012)

57.22 K. Vladislavleva, K. Veeramachaneni, U.-M. O'Reilly: Learning a lot from only a little: Genetic programming for panel segmentation on

sparse sensory evaluation data, Proc. 13th Eur. Conf. Genet. Progr. (2010) pp. 244–255

57.23 N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, P. Demeester: Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression, IEEE Trans. Circuits Syst. Video Technol. **23**(8), 1322–1333 (2013)