How To Get More Learning With Less Experimentation:

Streamline Your Work Processes With Multivariate Analysis And Design Of Experiments

by Dr. Frank Westad

CAMO
Bring data to life

Experimentation informs every part of the biologic lifecycle. But it is costly and time consuming — especially when you are using outdated methods. As you strive for more efficiency from your scientists and engineers, can you streamline your work processes to get more learning with less experimentation?

Both multivariate analysis (MVA) and design of experiments (DoE) methods have numerous applications for simplifying the learning from large data sets and experimentation. However, many scientists and engineers still perceive these methods to be complex. Today's newer, intuitive software applications make these techniques user-friendly for even non-statisticians. Early adopters are seeing decreased time to market, reduced development and production costs, and improved quality and reliability.

Too Much Data … Not Enough Information

Rather than focus on more data, you can use the data you already have to make better decisions. This starts with defining questions that need answers, and then analyzing the data to answer the questions. Analyzing data sets has not been easy in the past — and now we have more data than ever. Fortunately, better decisions can be delivered using solutions based on methods that provide both sound interpretation and predictions at the individual sample level. Despite proven value, the industry as a whole has been slow to implement these techniques, often delayed by misinformation. Let's consider the top four myths one by one.

Common Misconceptions About MVA And DoE

Myth #1: Statistical expertise is required to use MVA and DoE.

This myth is the only one that is actually rooted in truth. Prior to software such as The Unscrambler® (for MVA) and Design-Expert® (for DoE), the methods required a high level of statistical expertise. These newer software applications are now designed with an intuitiveness that leads scientists and engineers through the process. In fact, some software applications are specifically developed for use by scientists. Therefore, instead of having to request help from a data expert, your scientists and engineers can use the applications with very minimal training and without waiting for someone else to get results. An important aspect is that the owner of the data needs to be involved in planning experiments and perform the analysis with the application specific knowledge in mind.

Myth #2: I already use DoE. I don't need MVA.

DoE is very useful for product and process development when important variables can be controlled, such as the amount of ingredients in a formulation or target values for process variables (e.g., temperature or pressure). Nevertheless, the DoE is often a starting point of a series of process steps such as mixing, chemical reaction, fermentation, drying and purification to produce the end product with a number of properties describing the quality.

MVA is needed to understand the chemical, biological, or physical changes leading up to the final product, and also in which process steps the deviation from the normal operating condition occurs.

Predictive modeling with MVA (using in-line instruments such as spectroscopy) has been used for decades in many industries (e.g., food, agriculture, biopharma, chemistry). However, these "measurements" may not always be recognized as MVA applications.

Another application is finding the reason for reduced product quality, where numerous cases have shown that looking at the individual variables is not enough. There may also be different reasons why production is out-of-spec (e.g., raw material quality, process settings, and external influence such as outside temperature and humidity).

Myth #3: I already use MVA. I don't need DoE.

Although MVA is the most effective way to analyze data to find underlying patterns and detect out-of-spec situations, the data collected are in most cases taken from production itself. In most processes, some variables (also called parameters) are controlled to certain set points. If the process is stable, it may seem from analysis of the historical data that these variables are not important. Obviously, if there are changes, the process will not be running under stable conditions. Thus, you cannot draw all the right conclusions and know the causality of the system from historical data alone. You would have to rely on empirical correlations that may be due to indirect relations.

DoE offers a method for testing hypotheses derived from the MVA and confirmation of the hypotheses. Or, MVA followed by DoE can also lead to an unexpected finding, leading to innovation.

DoE is also needed for setting up the most cost-effective experimental plan to screen many variables in the fewest number of experiments, followed by optimization. This is not only for the actual practical

experiments, but also for investigating robustness of simulation models in chemistry, physics, and more. By changing the internal model parameter for complex simulation systems, using effective DoE one can model the system once and for all, thereby "metamodeling."

Myth #4: MVA and DoE are not compatible.

The combination of DoE and MVA fits the improvement cycle process, both for product quality and process optimization. The paragraphs above refer to some aspects of this, but there are many more, depending on the industrial application. MVA and DoE can lead to improved knowledge about the system under observation, confirming the background knowledge and theory. The combination also offers the possibility to find something that was not known previously, which may lead to a higher yield in the process and innovation. To demonstrate the combined power of MVA and DoE, we offer the case study that follows.

Case Study

How To Get More Learning With Fewer Experiments Using MVA And DoE

This case study demonstrates the use of MVA and DoE together to span the variation in a data table with a subset of the samples and then plan designed experiments to get the most learning with the least number of experiments.

The goal of the study was to maximize the yield of a chemical reaction. We started with 45 organic solvents as shown in Figure 1 on the next page. To prepare the data for analysis, we scaled the data to unit variance so that we could look at the variables on a common basis.

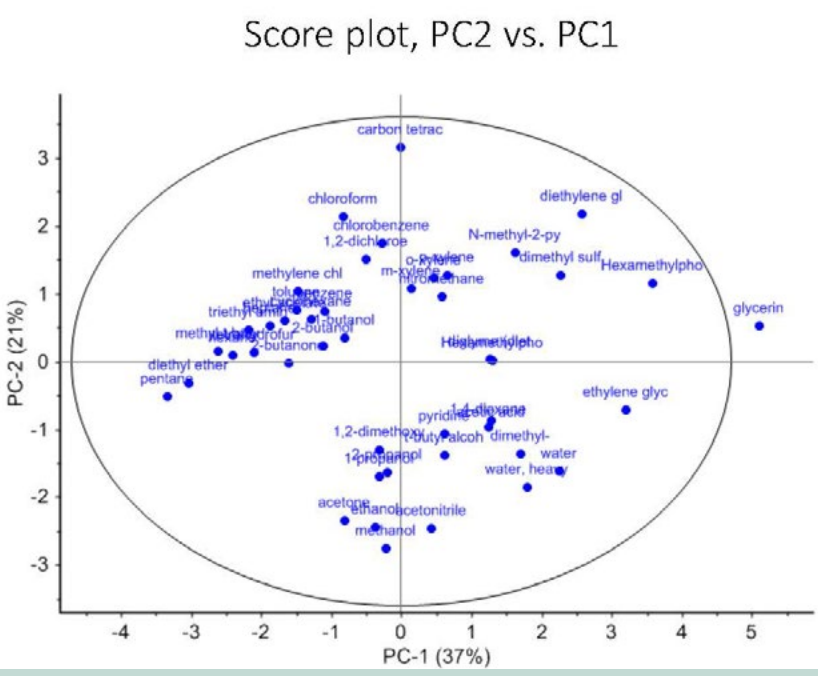
Part 1: Multivariate Analysis

We first applied a multivariate analysis method called Principal Component Analysis (PCA) with the objective of finding a new coordinate system that describes the underlying chemistry. Modeling the variance helps us to visually see "information" in the data that is not obvious in data tables. PCA can confirm our background knowledge, but it may also provide new insights that open the opportunity for innovation.

Figure 1

Solvent	formula	MW	boiling point (°C)	melting point (°C)	density (g/mL)	solubility in water (g/100g)	Dielectric Constant[3,4]	flash point (°C)
acetic acid	$C_2H_4O_2$	60.052	118	16.6	1.0446	Miscible	6.20	39
acetone	C_3H_6O	58.079	56.05	-94.7	0.7845	Miscible	21.01	-20
acetonitrile	C_2H_3N	41.052	81.65	-43.8	0.7857	Miscible	36.64	6
benzene	C_6H_6	78.11	80.1	5.5	0.8765	0.18	2.28	-11
1-butanol	$C_4H_{10}O$	74.12	117.7	-88.6	0.8095	6.3	17.8	37
2-butanol	$C_4H_{10}O$	74.12	99.5	-88.5	0.8063	15	17.26	24
2-butanone	C_4H_8O	72.11	79.6	-86.6	0.7999	25.6	18.6	-9
t-butyl alcohol	$C_4H_{10}O$	74.12	82.4	25.7	0.7887	Miscible	12.5	11
carbon tetrachloride	CCl_4	153.82	76.8	-22.6	1.594	0.08	2.24	--
chlorobenzene	C_6H_5Cl	112.56	131.7	-45.3	1.1058	0.05	5.69	28
chloroform	$CHCl_3$	119.38	61.2	-63.4	1.4788	0.795	4.81	--
cyclohexane	C_6H_{12}	84.16	80.7	6.6	0.7739	<0.1	2.02	-20
1,2-dichloroethane	$C_2H_4Cl_2$	98.96	83.5	-35.7	1.245	0.861	10.42	13
diethylene glycol	$C_4H_{10}O_3$	106.12	246	-10	1.1197	10	31.8	124
diethyl ether	$C_4H_{10}O$	74.12	34.5	-116.2	0.713	7.5	4.267	-45
diglyme (diethylene glycol dimethyl ether)	$C_6H_{14}O_3$	134.17	162	-68	0.943	Miscible	7.23	67
1,2-dimethoxy-ethane (glyme, DME)	$C_4H_{10}O_2$	90.12	84.5	-69.2	0.8637	Miscible	7.3	-2
dimethyl-formamide (DMF)	C_3H_7NO	73.09	153	-60.48	0.9445	Miscible	38.25	58
dimethyl sulfoxide (DMSO)	C_2H_6OS	78.13	189	18.4	1.092	25.3	47	95
1,4-dioxane	$C_4H_8O_2$	88.11	101.1	11.8	1.033	Miscible	2.21(25)	12
ethanol	C_2H_6O	46.07	78.5	-114.1	0.789	Miscible	24.6	13
ethyl acetate	$C_4H_8O_2$	88.11	77	-83.6	0.895	8.7	6(25)	-4
ethylene glycol	$C_2H_6O_2$	62.07	195	-13	1.115	Miscible	37.7	111
glycerin	$C_3H_8O_3$	92.09	290	17.8	1.261	Miscible	42.5	160

Figure 2



PCA also informs the decisions for the next phase of analysis or design. The PCA plots efficiently help us decide which of many possible individual relationships between variables should be explored further in experiments.

Thus, PCA helps to evaluate and visualize relationships between variables. A variety of plots can be used in PCA to map variables and interpret explained differences for variables and their correlations. The score plot (Figure 2) shows the clusters of plotted data. From these, it is important to pick samples that are dissimilar, not similar, for the DoE to achieve broader learning.

Together, the score plot and loading plot give an overview of all samples and variables. The variables closer to the outer circle are the most influential in describing the difference between the samples positioned on the score plot, as shown in Figure 3 on the next page. Once again, this is visually more obvious in the PCA plots than in tables.

The key takeaway from the first part of this study is that PCA provides a simple, visual representation that shows a map of the samples as well as the variables. The similarities between the samples are

visualized and the variables with the highest impact are identified. With that learning, you can select samples that will help you learn the most from your designed experiments.

Part 2: Design of Experiments and Optimization

To achieve meaningful experimental results, you need a good design. The PCA method helps you define the best subset for the experimental design. With DoE you want to use the right combination of variables to get the maximum value out of the experiment. With experiments being time consuming and expensive, optimizing experiments helps you get more learning (value) with less experimentation (cost) than traditional experimental methods.

Good experimental design has specific features:

- Even coverage inside and around surface of space
- Excess unique design points for testing lack of fit
- Replicates.

Replicates are multiple experimental runs with the same factor settings (levels). Replicates are subject to the same sources of variability, independently of each other. You can replicate combinations of factor levels, groups of factor level combinations, or entire designs.

Looking at the 44 organic solvents plotted in Figure 4, we used I-optimality to select optimal subset of six points to generate a model that spans the whole "universe" of solvent points. The six model points

Figure 3

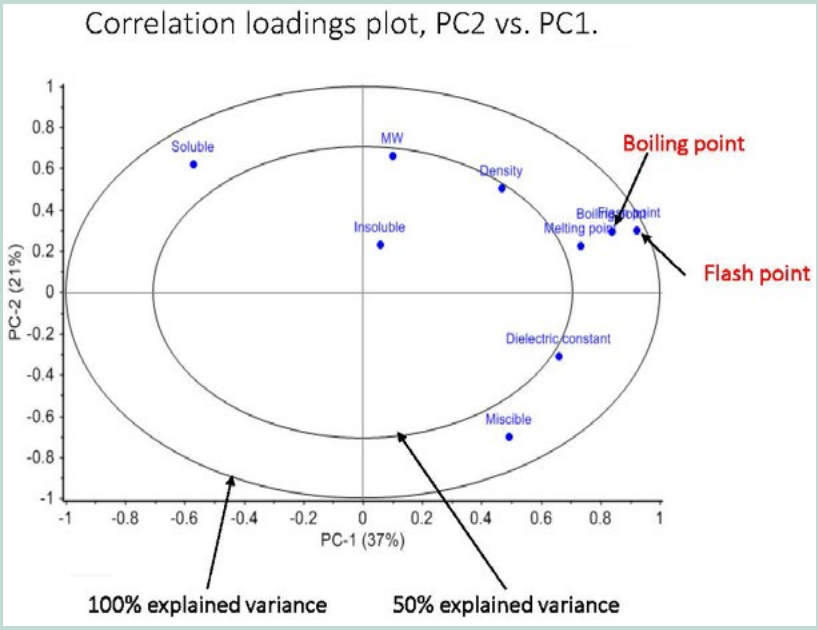


Figure 4

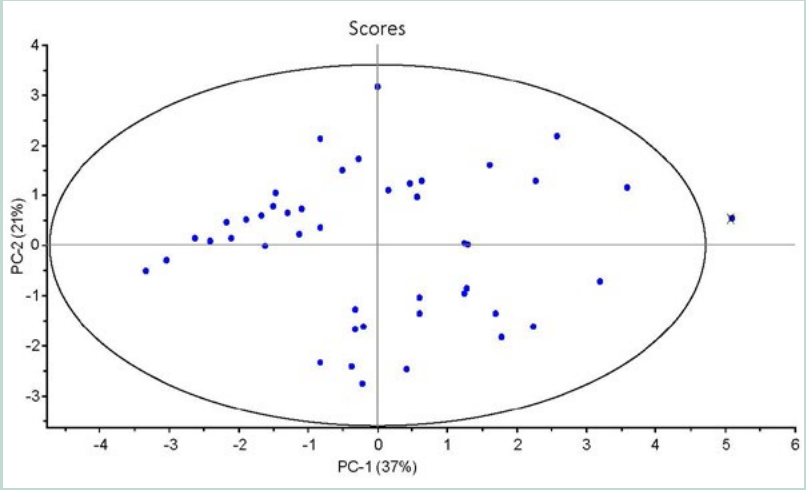


Figure 5

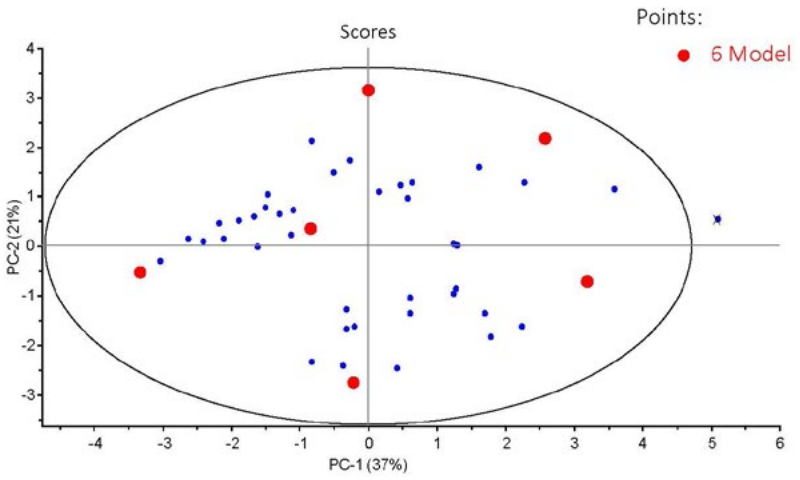


Figure 6

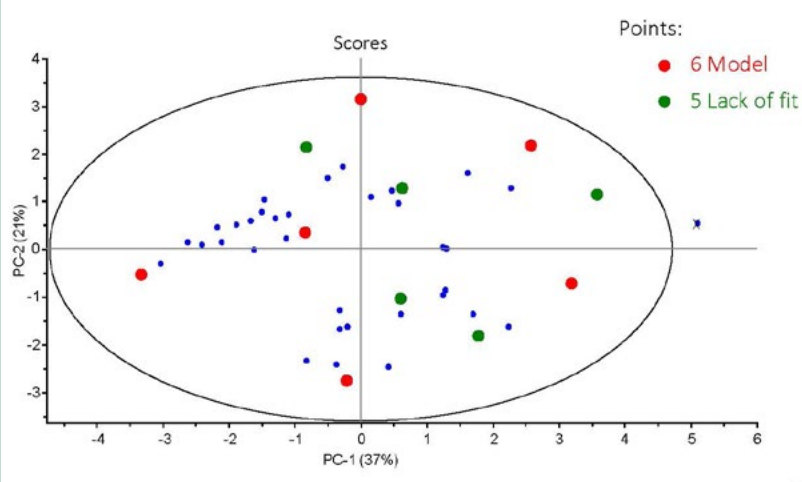
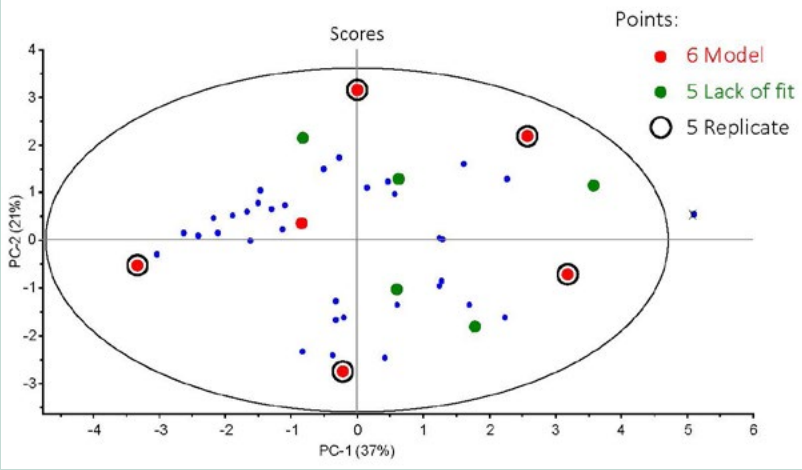


Figure 7



are shown in Figure 5. Referring to our "good experimental design" checklist mentioned previously, this achieves the goal of even coverage inside and around surface of space.

From there, we needed to add some extra points to support a quality design:

- Five lack of fit points are needed to fill in the gaps in design (see Figure 6)
- Five replicates, shown in Figure 7, were selected using model optimality criteria (I-optimality). Replicates are important for estimation of pure error used in the analysis of variance (ANOVA) table to be able to make the best decision about which design factors are significant and the goodness of the model.

Finally, the summary of design (see Figure 8 on the next page) shows the 11 solvents and five replicates to be used in the DoE.

Figure 9 also shown on the next page, shows the list of the optimal number of experiments to perform for making a model for the yield of the process. The next step is to find the optimum.

Next we started the experiment, see Figure 10. In this example, the second step (Analyze) showed a significant **lack of fit**. This told us we were missing something in the design. Perhaps, testing just two components is not enough to describe the response.

In the third step, shown on the following page, we identified the optimum combination of PC1 and PC2 based on yield.

However, the first DOE didn't show a good fit with only two components tested (PC1 and PC2). We realized that there was also information in the third component. At this point, we ran a second DoE including PC3. The new design led to a better solution (higher yield).

From here, we developed the design for DoE2. Our results, after including

Figure 8

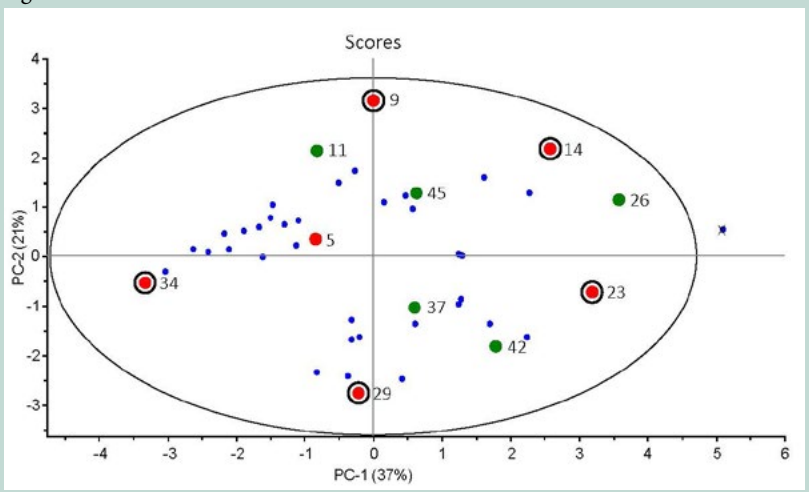


Figure 9

Run	Build Type	Comments	Factor 1 A:PC-1	Factor 2 B:PC-2
1	Replicate	pentane	-3.3312	-0.513686
2	Replicate	ethylene glycol	3.20478	-0.724302
3	Model	1-butanol	-0.813311	0.34074
4	Lack of Fit	Hexamethylphosphoramide	3.58336	1.14463
5	Model	carbon tetrachloride	-0.00216314	3.15579
6	Replicate	diethylene glycol	2.57854	2.17936
7	Lack of Fit	chloroform	-0.821893	2.12634
8	Model	pentane	-3.3312	-0.513686
9	Model	ethylene glycol	3.20478	-0.724302
10	Replicate	carbon tetrachloride	-0.00216314	3.15579
11	Lack of Fit	pyridine	0.612041	-1.05889
12	Lack of Fit	water, heavy	1.7929	-1.84868
13	Model	methanol	-0.212514	-2.7572
14	Lack of Fit	p-xylene	0.651025	1.27055
15	Model	diethylene glycol	2.57854	2.17936

Figure 10

Run	Comments	Factor 1 A:PC-1	Factor 2 B:PC-2	Response 1 Yield gms
1	pentane	-3.3312	-0.513686	12.75
2	ethylene glycol	3.20478	-0.724302	18.12
3	1-butanol	-0.813311	0.34074	36.89
4	Hexamethylphosphoramide	3.58336	1.14463	33.49
5	carbon tetrachloride	-0.00216314	3.15579	41.18
6	diethylene glycol	2.57854	2.17936	24.49
7	chloroform	-0.821893	2.12634	36.49
8	pentane	-3.3312	-0.513686	9.185
9	ethylene glycol	3.20478	-0.724302	18.02
10	carbon tetrachloride	-0.00216314	3.15579	34.89
11	pyridine	0.612041	-1.05889	45.34
12	water, heavy	1.7929	-1.84868	17.95
13	methanol	-0.212514	-2.7572	18.85
14	p-xylene	0.651025	1.27055	60.92
15	diethylene glycol	2.57854	2.17936	27.64
16	methanol	-0.212514	-2.7572	8.828

Figure 11

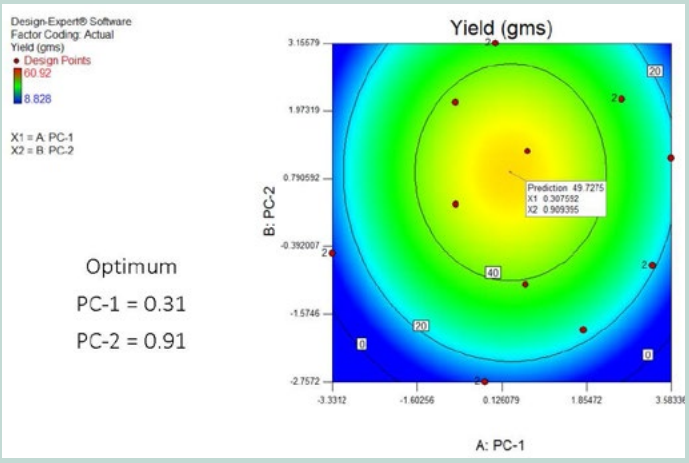


Figure 12

PC3 showed a much better fit and predicted maximum than DoE1 (see Figure 12. The predicted yield increased from 50% to 63%.

Value Of Using MVA With DoE

What was the value of adding MVA? If we had tried this study using only DoE, starting from the original table of solvents would have required many more experiments. Plus, the relationship between the variables would not have been visualized as in the results from the PCA. Further, the grouping of the solvents would not have been elucidated when looking at the original table of solvents or by scatter plot of the variables.

MVA methods help efficiently select the variables to get the most value out of DoE with the least number of experiments. Together, these methods can improve the efficiency of your people and harness the information of your data assets.

Dr. Frank Westad, Chief Scientific Officer, CAMO Software

Frank has served as a Research Scientist at SINTEF, Consultant with IDT GmbH (Germany), Application Specialist with CAMO, Senior Research Scientist at the Norwegian Food Research Institute (MATFORSK) and GE Healthcare. His experience includes numerous scientific papers, presentations at international conferences and teaching statistics, chemometrics and multivariate methods for the industry. He received his MSc in Chemistry and Data analysis in 1988, and completed his PhD in 2000 with the Norwegian University of Science and Technology.

ABOUT

CAMO Software

Founded in 1984 by a group of Norwegian scientists, CAMO Software are pioneers and an established leader in multivariate data analysis. CAMO Software develops software and solutions for analyzing large, complex data sets quickly, easily and accurately. The Unscrambler® is recognized for its powerful and user-friendly multivariate methods, easy data importing options, insightful visualization and the extensive and intuitive design of experiments tools, Design-Expert®. The models created with The Unscrambler® can be integrated directly into scientific instruments or in process monitoring solutions to perform multivariate analysis in real time. The Unscrambler® Process Pulse provides real-time process monitoring and can predict, identify, and help correct deviations in a process to drive continuous improvement. For more information, please visit **www.camo.com**.

CAMO

Bring data to life

LEARN MORE NOW