Knowledge Discovery and Analysis in Manufacturing

Mark Polczynski, PhD - Marquette University, College of Engineering, Milwaukee, Wisconsin, USA
Andrzej Kochanski, PhD – Warsaw University of Technology, Institute of Manufacturing Technology, Warsaw, PL

Abstract: *The quality and reliability requirements for next generation manufacturing are reviewed, and current approaches are cited. The potential for augmenting current quality/reliability technology is described, and characteristics of potential future directions are postulated. Methods based on knowledge discovery and analysis in manufacturing (KDAM) are reviewed, and related successful applications in business and social fields are discussed. Typical KDAM applications are noted, along with general functions and specific KDAM-related technologies. A systematic knowledge discovery process model is reviewed, and examples of current work are given, including description of successful applications of KDAM to creation of rules for optimizing gas porosity in sand casting molds. Finally, directions in KDAM technology and associated research requirements are described, and comments related to application and acceptance of KDAM are provided.*

Introduction

Industries across the globe are pursuing "next generation manufacturing" (NGM) as a tactic for meeting rapidly-expanding global needs for high performance, low cost, high quality products and processes, and as a strategy for revitalizing companies and industries which have become non-competitive over time [1,2]. Within the pursuit of NGM lies a specific question: How will manufacturers achieve "next generation quality and reliability" for new products and processes? Demanding global customers simply will not settle for current best-in-class performance.

About fifty years ago, the quality and reliability of products began to exhibit rapid improvements. Commonly traced to the seminal work of Shewhart [3], fundamentally different approaches to quality and reliability were initiated, primarily in Japan after WW II. Referred to by terms such as total quality control [4] and total quality management [5], wide-scale application of these approaches in the U.S. occurred the 1980's, initially in semiconductor fabrication. These approaches now permeate manufacturing across the globe, with the effects most obvious to the public in areas such as increased automobile quality and reliability. Only an owner of an automobile manufactured in the 1970's can fully appreciate the profound impact of this quality and reliability revolution.

The approaches that have enabled these improvements can be characterized by two fundamental concepts. The first is that quality is best achieved by controlling inputs (processes) vs. inspecting outputs (products). In addition to improving the quality of products, this approach minimizes shipment of the 10-15% defective products typically not caught by inspecting product [6]. The second fundamental concept is that the focus of output quality/reliability improvement efforts should be on minimizing variations in inputs through statistical analysis of samples of input data. Thus was born statistical process control (SPC) with its now-ubiquitous X-bar and R charts [7], and related statistics-based approaches now commonly applied to manufacturing, such as analysis of variance (ANOVA) [8], Taguchi methods [9], design for six sigma (DFSS) [10], and design of experiments (DOE) [11].

Current Situation

If we think of the set of statistics-based quality and reliability tools as a technology, we can assume that this technology follows a classic technology s-curve [12], where incremental improvements accumulate slowly over time as the technology is introduced, then increase rapidly with technology improvements and widespread application, and finally level off as the full potential of the technology is realized. We can further assume that the general pattern of technology s-curve progression [13,14] also applies here, i.e., when an incumbent technology's contribution to improvement (value) levels off, the technology is ripe for augmentation or replacement by a new technology, as illustrated in Figure 1.

If this situation holds for quality and reliability technology, we can reasonably ask: Has the statistics-based quality/reliability paradigm reached maturity, and if so, what technology will follow the s-curve progression[1]?

[1] When considering technology s-curves, it is essential to note that it is not necessarily the level of contribution of an incumbent technology that declines over time, it is the return-on-investment in technology improvements that declines. The question here is: Which technologies should I invest in to reach the next level of product and service quality and reliability?

Assuming that the technology s-curve model holds here, we can begin to answer the question of technology succession in quality and reliability by postulating three characteristics of next generation quality/reliability technology.

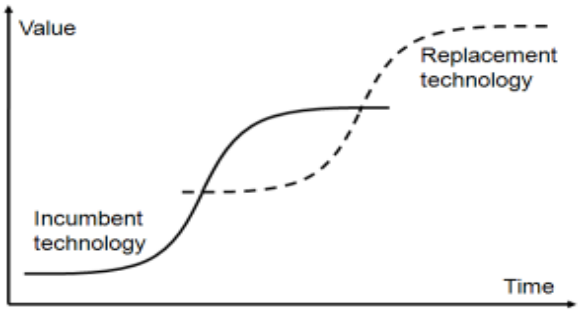

Figure 1: Technology s-curve progression

Generally, increasingly powerful and available computer and communication capacity is generating an ever-expanding sea of data. The manufacturing environment mimics this general trend, with the content of manufacturing-related databases potentially extending far beyond the information scope of current statistics-based quality and reliability approaches to include areas such as warrantee data, sales and marketing information, financial data, etc. These databases typically consist of many unrelated sets of data aggregated by many different entities within and outside of an organization, with each database geared toward supporting different organizational functions. We can predict that next generation quality/reliability technology will be based on integration of these massive extended databases.

A common refrain heard from those directly involved in making next generation manufacturing a reality is that they are drowning in this rising sea of data. Traditional analysis approaches to identifying underlying patterns and structures in data are producing diminishing returns relative to the growth in available data (the tail of the technology s-curve). We can therefore assume that next generation quality/reliability technology will include finding useful patterns and structures in data currently unperceivable using common statistical approaches.

The task of identifying useful data patterns and structures in massive extended databases implies a third characteristic of next generation quality and reliability technology, which is the ability to effectively utilize highly coherent, noisy, and corrupted data with missing field and record entries. This is a natural consequence of using data obtained from a variety of internal and external sources collected for purposes other than improving product designs and manufacturing processes.

Knowledge Discovery and Analysis

Given these characteristics, the next logical question is: Does such a technology currently exist? The answer is (of course): Yes. The set of tools and techniques grouped under terms such as data mining, machine learning, and knowledge discovery in data (KDD) [15-17] are designed specifically to provide the functions basic to next generation quality and reliability requirements, i.e. integration of massive extended databases to find useful but currently unperceivable patterns and structures in noisy data. Here, we will refer to the application of these and related technologies in manufacturing as *KDAM - knowledge discovery and analysis in manufacturing*.

Currently, research and application of these technologies occurs most commonly in social and business fields, and is most apparent to the public in sales and marketing applications. A commonly-encountered example is the amazon.com book recommendation system. When a customer logs in at amazon.com, a personalized home page appears recommending a number of books of potential interest to the customer categorized in several different ways, e.g., books similar to books that the customer has purchased, other books by authors of books the customer has purchased, books that other customers have purchased in addition to the books that the customer has purchased, etc. These recommendations are based on sophisticated data mining technologies applied to customer transaction data. Netflix, the world's largest on-line movie rental service, provides similar recommendations, but includes an on-line customer survey of preferences designed to increase the hit-rate of movies selected from the recommended titles[2].

Nielsen Claritas (www.claritas.com) uses the technologies referenced here to provide a consumer segmentation system that combines demographic, consumer behavior, and geographic data to help marketers identify, understand and target their customers and

[2] Recently, Netflix has offered the Netflix Prize (www.netflixprize.com), a $1 million award to any person or organization that produces a movie recommendation algorithm ten percent better than the existing Netflix algorithm [18]. An internet search on "netflix dataset" will provide the reader with interesting insights into this application of data mining and machine learning, a snapshot of how the Netflix Prize competitors are doing, and links to the actual Netflix dataset.

prospects with customized products and communications. For example, the Claritas PRIZM NE product classifies households in terms of 66 demographically and behaviorally distinct types, which are further segmented into social groups and "LifeStage" groups. These types and groups can be linked to specific geographical areas (zip codes), which can, for example, assist in identifying likely new store locations.

Amazon and Netflix provide examples of applications targeted at the personal level. Claritas provides an example of an application at the group social segmentation level. The SPSS Clementine software suite (www.spss.com/clementine) provides an example of a high-level application that performs enterprise-wide "predictive analytics", which SPSS defines as including: 1) analysis of past, present, and projected future outcomes using a range of technologies including data mining and related technologies, and 2) decision optimization algorithms for determining which actions will drive the optimal outcomes.

An example of Claritas PRIZM-NE classification illustrates the manner in which data mining can be used in business and social applications. The Claritas "Urban Uptown" class (one of the 66 major PRIZM-NE classes reference above) is defined as being "…home to the nation's wealthiest urban consumers. Members of this social group tend to be affluent to middle class, college educated and ethnically diverse, with above-average concentrations of Asian and Hispanic Americans. Although this group is diverse in terms of housing styles and family sizes, residents share an upscale urban perspective that's reflected in their marketplace choices. Urban Uptown consumers tend to frequent the arts, shop at exclusive retailers, drive luxury imports, travel abroad and spend heavily on computer and wireless technology".

One of the five groups within this Claritas class is the "Young Digerati", described as "the nation's tech–savvy singles and couples living in fashionable neighborhoods on the urban fringe. Affluent, highly educated and ethnically mixed, Young Digerati communities are typically filled with trendy apartments and condos, fitness clubs and clothing boutiques, casual restaurants and all types of bars – from juice to coffee to microbrew."

Clearly, this type of characterization of geographic areas is quite useful for applications such as new store site selection. Examples such as these illustrate that the technologies referenced here are well-suited to and well-established in social and business fields[3].

It is interesting to note that while the statistical approaches developed initially for the shop floor are now receiving significant attention in office environments [19,20], technology flow in the opposite direction (business application to shop floor) is occurring for the approaches focused on here.

KDAM Applications

Successful applications of KDAM technology do exist [21-28]. Common applications includes:

- Detection of root causes of deteriorating product quality,
- Identification of critical and optimal manufacturing process parameters,
- Prediction of effects of manufacturing process changes,
- Identification of root causes and prediction of equipment breakdown.

Of course, areas such as these have always been of intense interest, and statistical approaches such as SPC, DFSS, and DOE have proven and will continue to be extremely effective in supporting continuous improvement in these and related areas. So, why the need for an additional approach? Because further improvements in these areas are starting to rely more and more on identifying increasingly-obscure patterns and discovering increasingly-complex structures in data obtained on the shop floor. Today, it is becoming increasingly difficult to "see the forest for the trees". This is exactly the environment that KDAM is designed to work in.

KDAM and its fundamental technology elements (data mining, machine learning, etc.) encompass a wide range of functions, tools, techniques, etc. Some of the functions particularly relevant to KDAM include:

1. Regression: Defining functional relationships between outputs of interest and multiple and possibly dependent inputs. An example is predicting the dimension of a plastic molded part given typical ranges of molding process variables.

[3] The reader is encouraged to look up the Claritas description of their own zip code at: http://www.clusterstaging.claritas.com/MyBestSegments/Default.jsp.

2. Classification: Grouping of objects (products) into classes given previously known input/output (process/product) classifications. An example is association of acceptable and defective parts (two classifications) with the particular production conditions under which the parts were manufactured.
3. Clustering: Grouping of objects (products) by characteristics where there exist no previously known associations. An example is discovering that a particular employee operating a particular machine tends to produce parts with dimensions on the high side of the target value.

A wide variety of specific technologies have been applied to perform these functions. For purposes of illustration, a very brief list of commonly applied approaches is provided here:

- Artificial neural networks [29,30],
- Self-organizing maps [30],
- Genetic algorithms [31],
- Decision trees [32],
- Bayesian classifiers [33]
- Multivariate statistical projection [34-37].

Knowledge Discovery Process Models

In addition to advances in tools and algorithms and the means to test and compare approaches, a key ingredient to moving data mining and machine learning technologies from laboratory curiosities to real-world applications was the development of systematic knowledge discovery processes. Due to the variety of potentially-useful technologies, the magnitudes of the databases analyzed, the complexity of the problems being addressed, and the rate at which research in this area is proceeding, standardization of the knowledge discovery process was critical to widespread application of these technologies.

The knowledge discovery process has been incorporated into a number of formal process models which can generally be grouped as academic research models and industrial application models [38]. Roiger and Geatz [39] provide an informative review of these models, and compare them with the scientific method of problem solving. The CRISP-DM (Cross-Industry Standard Process for Data Mining – www.crisp-dm.org) provides an industrial model particularly well-suited to KDAM applications.

The CRISP-DM model is shown in Figure 2. The model has six primary steps:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling (data mining)
5. Evaluation
6. Deployment

The CRISP-DM model is characterized by an easy-to-understand vocabulary and good documentation (available on-line). The model divides the six primary steps into detailed sub-steps, and encourages iterative development through feedback among steps[4].

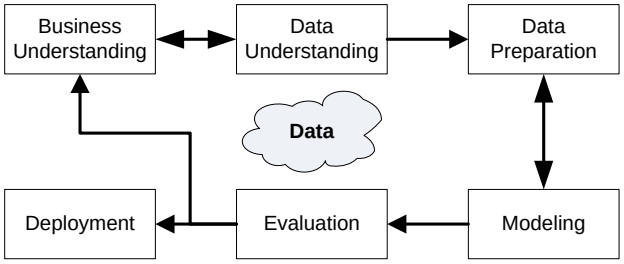

Figure 2. CRISP-DM process model with feedback loops.

KDAM Example – Metal Casting

Researchers at Warsaw University of Technology's Institute of Manufacturing Technology[5] are applying KDAM in foundry production and metal cast part

[4] CRISP-DM is comparable to the DMAIC improvement cycle [38] commonly associated with six-sigma approaches. DMAIC is composed of five phases: define, measure, analyze, improve, and control, which are generally comparable to the six CRISP-DM steps. Because KDAM approaches typically rely on real-time on-line production data gathered during regular process operation (vs. data obtained during designed and controlled experiments), the data can be quite noisy. Thus, the CRISP-DM model places significant emphasis on data understanding and preparation. Beyond this, examination of the details of the process steps reveals that differences among these and related methodologies often lies more in origins rather than intent, with approaches such as CRISP-DM being perhaps more closely associated with information technology, and DMAIC and related methodologies having origins in engineering.

[5] Contact Dr. Andrzej Kochanski, Institute of Manufacturing Technologies, Warsaw University of Technology, Warsaw, Poland, akochans@wip.pw.edu.pl.

manufacturing, including: 1) detection of causes of gas porosity in steel castings; 2) optimization of cast iron heat treatment parameters; 3) green molding sand formulation; and 4) prediction and improvement of melt quality and casting properties such as strength, elongation, and hardness [40-49].

Cast part manufacturing is challenging due to several aspects of the casting process. First, the casting alloy can consist of over a dozen chemical components. Second, not only does the casting material experience physical changes during the melting and cooling phases of casting, the chemical composition of the material also changes. Additionally, casting requires the parallel manufacture of other objects, such as the mold, which are typically made of completely different materials. In the case of sand casting, the properties of the mold also change physically and chemically during casting. Further, the sand casting mold materials (which have changed composition during molding) are re-cycled to make additional molds. Finally, the physical environment of a foundry may not be well-controlled (to say the least). Figure 2 shows the various elements of a sand casting system.

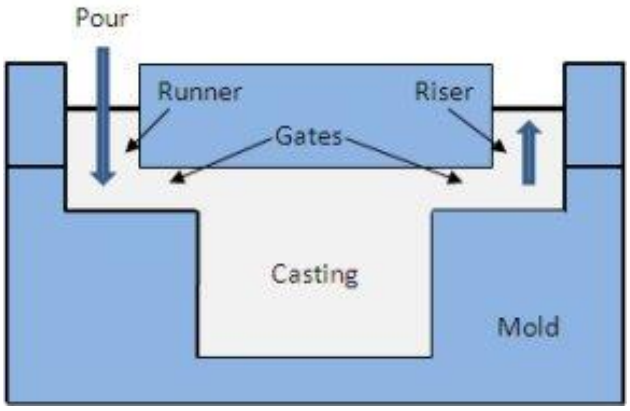

Figure 2: Elements of the sand casting process.

Critical cast part characteristics include part dimensions, surface finish, and feature shapes, and also material properties such as tensile strength, elongation, hardness, etc. These characteristics can be extremely important, especially for applications where human life depends on component quality and reliability. Figure 3 shows half of a sand casting mold. Figure 4 shows filled molds waiting for the cast parts to cool.

This brief overview of sand casting illustrates that a number of critical output characteristics rely in a complex manner on a number of process inputs that interact in complex ways and experience significant variation over time. This makes prediction of output characteristics based on mathematical modeling of chemical and physical process that occur during melting, pouring, and cooling very difficult, and therefore makes this process a prime candidate for analysis using KDAM technologies.


Figure 3: Sand casting mold.

Of the tools typically associated with KDAM, artificial neural networks (ANNs) have probably received the most use for foundry applications (see [40] and [42] for extensive bibliographies of ANN usage in foundries). Research results by the Warsaw group related to mold gas porosity [40] illustrates the type of problem that ANNs are well-suited to solving.

In one application by the Warsaw researchers, production data on gas porosity was collected for over 2000 parts produced in a steel foundry during five months of normal production. About 4% of the parts exhibited gas porosity. The final number of part samples used for ANN training was 170.

Table I shows 39 factors that were assumed to have an influence on mold gas porosity. Through the training and analysis of an artificial neural network, eleven were found to have relatively low impact.

Training of the neural network on the final 28 variables resulted in automatic generation of a set of rules of the form:

If water content in molding sand at molding time is high
And the time from molding to pouring is high
And the environment temperature is low
And the air humidity is high
Then the probability of excessive gas porosity is high.

Obviously, these rules follow common sense, even for a person that knows little about the molding process! So what is the benefit of using ANNs to characterize this

situation? After automatically generating rule sets, the neural network was used to create a series of response curves relating gas porosity to major process variables. The result was a quantitative description of safe ranges of variables within which the desired molding properties can be obtained. Interestingly, a related study by this research group [42] revealed that the primary cause of casting defects was the molder, a human variable.

Figure 4: Filled sand casting molds.

KDAM Example – Surface Finishing

Researchers at Marquette University's College of Engineering[6] are working with industry partners to initiate KDAM research related to surface finishing of machined metal parts.

Surface finish constitutes a critical product characteristic in a wide variety of manufacturing operations. Consistently achieving a specified surface finish while minimizing cost-related factors such as cycle time, material waste, and tool wear is essential to maintaining economically viable finishing processes.

Surface finish depends on a number of process variables such as tool speed, material feed rate, and tool condition. Further, there is a wide variety of finishing processes to choose from (e.g. grinding, honing, etc.). This can make it quite a challenge to optimize process variables to achieve a specified finish at minimal cost.

One application of surface finishing technology being investigated relates to production of internal combustion engines. Reducing engine exhaust emissions is a goal of increasing importance. One factor influencing emissions is combustion of engine oil due to leakage of oil between the piston ring and the cylinder wall. This problem

Table I – Foundry production parameters used as neural network inputs for detection of casting defects.

Parameter	Final
Metal content variables	
Scrap amount	Y
Scrap quality index	Y
Fe/Mn/Si amount	Y
Fe-Si amount	N
Fe/Ca/Si amount	Y
Lime amount	Y
% C change	Y
% S change	Y
% Mn change	Y
Final % Al	N
Final % Si	Y
Final % P	N
Mold variables	
Molding sand moisture content	Y
Molding sand permeability	Y
Molding sand tensile strength	Y
Mold quality index	N
Core sand code	Y
Core coating code	N
Molding sand code	N
Mold coating code	N
Core glue code	Y
Bar test casting porosity	N
Process variables	
Melting time	Y
Time from molding to pouring	Y
Tapping temperature	Y
Melting furnace number	Y
Ladle number	Y
Pouring order	N
Days from ladle repair	Y
Days from current furnace repair	Y
Days from previous furnace repair	Y
Pouring quality index	Y
Nozzle supplier code	N
Environmental variables	
Air humidity	Y
Environment temperature	Y
Environment temperature before pouring day	N
Human variables	
Melting team number	Y
Molding team number	Y
Assembling team number	Y

[6] Contact Dr. Mark Polczynski, Engineering Management Program, Marquette University, Milwaukee, WI, USA, mark.polczynski@marquette.edu.

is exacerbated under two conditions: 1) if the cylinder and piston/ring fit poorly, oil leaks through the gaps into the cylinder, and 2) if the cylinder and piston/ring fit too tightly, rings hydroplane over the oil film on the cylinder walls. Thus, reduction in this source of engine emissions relies heavily on maintaining an optimal fit between the cylinder and piston (Figure 5). This research focuses primarily on optimizing this fit.

Figure 5: Cut-away view of piston and cylinder

Being in its initial stages, this research is pursuing a different path than the metal casting research. In KDAM terminology, the metal casting problem is one of *classification*: given certain combinations of conditions that produce products classified as "acceptable" or "rejects", the goal is to discover rules that predict process conditions that consistently produce acceptable parts. For the surface finishing work, virtually all the parts analyzed are good parts, and the initial analysis focuses on *clustering*. The goal is to find clusters of parts with dimensional measurements near target values with low variance, and then identify various combinations of process variables associated with these clusters. Thus, an example cluster description might read something like:

- Part number 12345
- Machine number 10
- Operator XYZ
- Second shift
- Summer season
- Surface finish as specified, with low variance

After characterizing clusters with desirable product characteristics, the emphasis of the analysis will shift to discovering why these particular conditions produce parts with especially good surface finish. If the reasons can be identified, this poses the opportunity to propagate best practices across production.

Table II shows seven cylinder boring process variables and seventeen cylinder dimensions being used for initial clustering analysis. Note that in addition to surface measurements, a number of macroscopic bore dimensions are included. Since the data is already available, and since clustering (and classification) algorithms automatically generate outputs, and given that most good data mining and machine learning tools include means of identifying which types of data (attributes) are irrelevant to the analysis, there is generally no reason to not include as much data as is available in an initial analysis such as this.

Table II. Variables and dimensions used in cylinder boring clustering analysis.

Process variables
Part number
Boring process
Machine operator
Transfer line
Production date
Production time
Elapsed time
Bore dimensions
Piston bore diameter (6 measures)
Piston bore roundness (3 measures)
Piston bore chamfer (1 measure)
Surface finish
Ra
Rmax
Rpk
Rvk
Mr1
Mr2
Vo

As an example of the type of data being collected, Figure 6 shows a time plot of R_a surface roughness data for 4973 cylinder bore production samples measured over 140 days. The most commonly used measure of surface roughness, R_a is the arithmetic mean of the magnitude of the deviation of the surface profile from the mean line along the surface [50]. Each point in the figure has been

standardized by subtracting the average value and dividing by the standard deviation[7].

Marquette researchers are performing their analyses using Weka, a comprehensive tool bench for machine learning and data mining. Weka is free open source software developed as part of the Weka Machine Learning Project at the University of Waikato in New Zealand, and can be downloaded along with extensive documentation from the project's web site (www.cs.waikato.ac.nz/ml/weka). Part of the Marquette research involves evaluation of this software for producing robust tools that can be used on the shop floor.

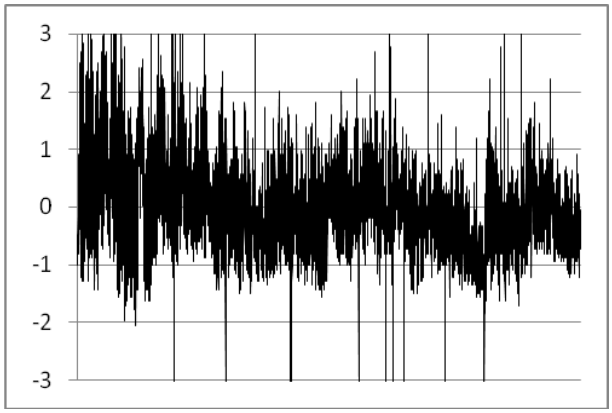

Figure 5: Normalized cylinder R_a measurements.

KDAM Directions

Aside from the examples cited above, is there any evidence of trends toward broad acceptance of KDAM technology? The answer is (again): Yes. Consider the situation at Toyota Motor Corporation.

Toyota is viewed as a world leader in the development of advanced manufacturing strategies [51], and the Toyota Production System [52] is applied as a model by many manufacturers around the world. A strong innovator in the field of Total Quality Control and the application of statistical methods, Toyota has relied heavily on approaches such as design of experiments, multivariate analysis, the Taguchi method, and robust design to improve product quality and reliability and reduce manufacturing costs. But the manner in which these tools are being used at Toyota are evolving. According to Hino [53]:

What Toyota has done is to accumulate past examples and data from the entire Toyota group in a systematic, stratified, and electronic format and then use computer ("office automation") analysis of these to derive answers to most problems and issues. With two weeks provided for the resolution of problems, the time-consuming experimental design and Taguchi methods are hardly ever used.

Hino also notes that "the company has been spending five times what other companies do to collect data, with the result that nearly all problems can now be solved by using past data." This approach constitutes, in essence, an enterprise-wide application of data mining to improve a wide range of functions within Toyota, including product quality and reliability.

Another significant example of this trend is provided by Davenport and Harris [54], who describe activities at Honda Motor Company as follows:

Honda instituted an analytical "early warning" program to identify major potential quality issues from warrantee service records. These records are sent to Honda by dealers, and they include both categorized quality problems and free text. Other text comes from transcripts of calls by mechanics to experts in various domains at headquarters and from customer calls to call centers. Honda's primary concern was that any serious problems identified by dealers or customers would be noticed at headquarters and addressed quickly. So Honda analysts set up a system to mine the textual data coming from these different sources.

As world leaders in manufacturing, Toyota and Honda have taken aggressive steps to improving manufacturing processes[8]. But typically, manufacturers will be somewhat more cautious when first applying KDAM technology. One low-risk path to adopting KDAM is to begin applying this technology in support of widely-accepted statistics-based quality and reliability approaches.

An example of this is the work of Perzyk et. al. [43], who propose utilization of classification tools such as

[7] In addition to potentially facilitation analysis, normalization "anonymizes" data, thus easing concerns about security of sensitive process and quality/reliability information.

[8] If these cases at Toyota and Honda are indicative of an east-first spread of the application of data mining and machine learning technology to improving product quality and reliability, then the reader might take a moment to reflect on history's penchant for repeating itself.

decision trees to determine the significance of trends in SPC control charts. To use SPC control charts, measurements are first taken on production samples at appropriate intervals of time, and the sample mean is calculated. Then, alarm and warning limits are calculated. Samples are measured during production and values are plotted vs. time on a chart showing upper and lower warning and alarm limits [7]. Figure 8 shows an example of an SPC control chart[9]. Equivalent charts using ranges of values within sample sets vs. sample mean are also commonly utilized.

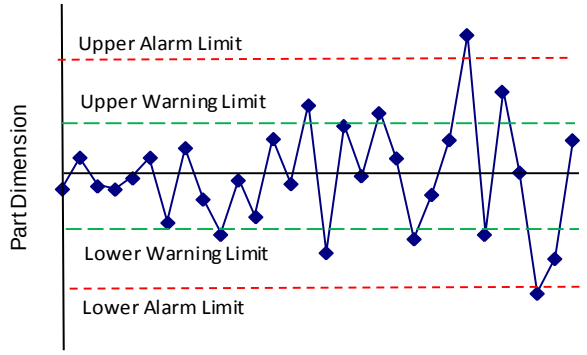

Figure 8: SPC control chart with limits.

Having prepared the control chart and recorded the results over some period of production, the question now becomes: What course of corrective action is appropriate, given the pattern of sample points vis-à-vis the warning and alarm limits?

Various heuristics have been developed over time to specify appropriate courses of action based on trends in the sample means with respect to the warning and alarm limits. An example would be the following rule set, which resembles the form of the rule set example given earlier which was automatically derived by an ANN:

Stop the process when one sample falls outside of an alarm limit,

Or when two samples fall outside the same warning limit,

Or after a run of six continuously rising samples within the warning limits.

[9] This univariate control chart is used for purposes of illustration. Latent-based multivariate control charts [33-38] constitute a more sophisticate approach.

For the approach recommended by Perzyk et. al., heuristics are replaced by using sample mean data as input to a decision tree which has been developed through analysis of the causes and effects of process variations. The decision tree thus continually and automatically "learns" appropriate responses to various sample data patterns based on a continuous flow of process data and production results.

Research Requirements

This example of the integration of a common data mining tool with a widely-applied statistical approach illustrates one route to greater use and acceptance of KDAM technology. However, significant research still needs to be conducted to establish wide-spread acceptance of KDAM in the manufacturing environment.

An example of the direction of KDAM research is provided by Perzyk, et. al. [49], who compared the effectiveness of artificial neural networks with the naïve Bayesian classifier for foundry casting processes. While ANNs have been successfully applied for such applications, this approach has shortcomings such as the need for a complex and time-consuming training process, and the ambiguity of results[10]. Although the naïve Bayesian classifier approach has not been widely used in this application, these researchers have demonstrated that this approach provides results that are comparable to those obtained from artificial neural networks, but is simpler to use, provides unique solutions, and can be developed with less data.

This example illustrates the primary research challenge for widespread application of KDAM technology in manufacturing. While significant time and energy has been applied in business and social fields to match and improve specific approaches to particular applications, much work needs to be done in matching and optimizing specific KDAM approaches with particular manufacturing applications before a general appreciation of the value of this technology can emerge.

To stimulate current and future KDAM-related activities, the authors have initiated creation of a KDAM consortium (www.technologyforge.net/KDAM) focused on providing a forum for the free and frequent exchange

[10] When applying neural networks, a particular problem can typically be effectively solved by networks with different architectures, topologies, and network weights, raising the question: Which is the "right" network?

of data mining and machine learning technology, datasets, and application results to rapidly and effectively improve the capabilities and efficiencies of manufacturing processes and the quality and value of manufactured products. Based on this ambitious goal, the mission of KDAM is to create and maintain an industry and academic community of interest to:

1. Identify and develop technologies suitable for achieving the KDAM goal, including technology reduction-to-practice in manufacturing environments,
2. Identify and classify particular manufacturing processes that have proven to benefit from or could benefit from the application of specific KDAM-related technologies,
3. Share technology developments and communicate application results and findings among Consortium partners, including application best/worst practice guidelines,
4. Where practical, share databases among Consortium members, thereby enabling development, verification, and comparison of KDAM-related technologies.

Next Generation Quality and Reliability

A basic premise of this discussion is that current approaches to quality and reliability represent a technology which can be expected to follow a typical technology s-curve progression. So-called KDAM technology is proposed as an approach capable of augmenting the current technology to provide next generation quality and reliability. If this premise is accepted, then two questions arise at this point in the discussion: What will this amalgamation of technologies ultimately look like, and how can the transition to new approaches be facilitated?

Regarding the first question, consider the rise of what Davenport and Harris define as "analytics" [54]:

... the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions. The analytics may be input for human decisions or may drive fully automated decisions. Analytics are a subset of what has come to be called business intelligence: a set of technologies and processes that use data to understand and analyze business performance.

Davenport and Harris go on to provide detailed descriptions of what analytics is and is becoming. It is suggested here that next generation manufacturing quality and reliability based on an amalgamation of statistics-based technologies with data mining and machine learning technologies conforms to their description of analytics.

In reference to the second question posed above, it is important to note some fundamental differences between statistics-based and KDAM-related approaches. In general, statistics-based approaches can be termed *confirmatory*, meaning that a pattern is *hypothesized* to exist in the data, and analysis confirms or denies its presence. On the other hand, KDAM-related approaches can be thought of as being *exploratory*, focusing on *discovering* "interesting" or "unusual" patterns in the data [55,57].

Another way to view these fundamental differences is to recall the words of George Box, a primary source for design of experiments technology: "To find out what happens with a process when you interfere with it, you have to interfere with it, not passively observe it". A data miner would not disagree with this basic premise, but might point out that nature regularly interferes with processes all by itself in an often disturbingly vigorous manner. If you collect enough data and analyze it appropriately, the patterns generated by natural interference may reveal themselves without human encouragement.

Probably the most productive way to view KDAM and current statistics-based methods is as complementary approaches. As an exploratory approach focusing on discovering interesting or unusual patterns in data, KDAM-type methods can form an effective front-end for improving the efficiency of statistics-based confirmatory approaches by reducing the cost and turn-around time to conduct design-of-experiment investigations to firmly establish causal relationships among variables.

Of course, should KDAM-related research indicate that this approach provides the potential for significant improvements in quality and reliability, the research must be followed by development of robust, inexpensive tools easily used by production workers on the factory floor. One enabler of the widespread successful use of tools such as SPC has been the proliferation of robust software tools. This condition must be replicated if KDAM is to contribute to next generation quality and reliability.

Given fundamental differences such as these, it can be expected that resistance to acceptance of alternate viewpoints will occur (e.g., which is the preferred problem-solving model, DMAIC or CRISP-DM?). An

old adage may apply here: Never underestimate the power of incumbency. Significant time and energy has been invested in statistics-based approaches, and spectacular improvements in quality and reliability have been and continue to be achieved. This calls to mind another old adage: If it ain't broke, don't fix it. The implication is that first-adopters of KDAM need to show spectacular results to establish the merits of this technology.

That being said, the technology s-curve progression model has proven to be applicable over a broad range of situations. Practitioners of both incumbent and alternative approaches are encouraged to peruse the literature which deals directly with comparison and contrast of statistics-based approaches and KDAM-related technologies [55,56]. These and future discussions will be vital to the development of next generation quality and reliability.

Summary

Based on the overview provided here, the current status and future of the deployment of data mining, machine learning, and related technologies in the field of manufacturing can be summarized as follows:

- Acknowledging the success of current statistical approaches to improving manufacturing quality and reliability, the need exists to develop next generation approaches to meet increasingly stringent global market quality, reliability, and cost requirements.
- A strong base of data mining and related tools, techniques, and processes has been developed to identify increasingly-obscure patterns and discover increasingly-complex structures in the types of data being generated on factory floors.
- These approaches are widely-applied in business and social applications, and research programs, tools, and examples of successful implementation are available to support these applications.
- Successful applications of knowledge discovery and analysis in manufacturing technology (KDAM) do exist, indicating the potential for transfer of technology into this field,
- Significant work needs to be conducted to measure, compare, and improve the effectiveness of particular KDAM technologies for specific manufacturing applications.
- Significant work needs to be done to establish the usefulness of KDAM vis-à-vis incumbent approaches, and to integrate relevant aspects of each into a useful toolset capable of supporting next generation quality and reliability.

Regarding the last two points, the amount of effort required to augment the incumbent quality/reliability technology should not be under-estimated. However, the demands of next generation manufacturing are rigorous, and the potential rewards to organizations that successfully transition to next generation quality and reliability levels are significant. This situation warrants efforts to vigorously assess the potential of KDAM to meet the challenges of next generation manufacturing.

References

1. J. Jordan, F. Michel, *Next Generation Manufacturing Methods and Techniques*, Wiley, 2000.
2. N. Comrades, M. Lystlund, "The vision of next generation manufacturing", *Int. Mfg. Sys*, vol. 14, no. 4, pp. 324-333, 2003.
3. W. Shewhart, *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Company, 1931.
4. A. Fiegenbaum, *Total Quality Control*, Mc-Graw-Hill, 1991.
5. S. George, A. Weimerskirch, *Total Quality Management: Strategies and Techniques Proven at Today's Most Successful Companies*, Wiley, 1998.
6. J. ReVelle, Manufacturing Handbook of Best Practices: *An Innovation, Productivity, and Quality Focus*, CRC Press. 2002.
7. J. Oakland, *Statistical Process Control*, William Hinemann, 1986.
8. H. Scheffe, *The Analysis of Variance*, Wiley-Interscience, 1999.
9. G. Taguchi, S. Chowdhury, Y. Wu, *Taguchi's Quality Engineering Handbook*, Wiley-Interscience, 2004.
10. K. Yang, B. El-Hak, *Design for Six Sigma*, McGraw-Hill Professional, 2008.
11. G. Box, J. Hunter, W. Hunter, "Statistics for Experimenters: Design, Innovation, and Discovery", 2nd Edition, Wiley, 2005.
12. E. Rogers, *Diffusion of Innovations*, Free Press of Glencoe, Macmillan Company, 1962.
13. D. Sahal, *Patterns of Technological Innovation*, Addison Wesley, 1981.
14. R. Foster, *Innovation: The Attacker's Advantage*, Summit Books, 1986.
15. D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, Massachusetts Institute of Technology, 2001.
16. I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 1999.
17. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley and Sons, 2005.

18. J. Ellenberg, "The Netflix Challenge", *Wired*, vol. 16, no. 3, pp 114-122, March 2008.
19. R. Munro, *Six Sigma for the Office*, ASQ Quality Press, 2002.
20. J. Lowenthal, *Defining and Analyzing a Business Process: A Six-Sigma Pocket Guide*, ASQ Quality Press, 2002.
21. A. Kusiak, "Data mining: Manufacturing and service applications", *Int J Prod Res*, vol. 44, pp 4175-4191, 2006.
22. D. Braha, *Data Mining for Design and Manufacturing - Methods and Applications*, Kluwer Academic Publications, 2001.
23. J. Harding, M. Shahbaz, R. Srinivas, A. Kusiak, "Data mining in manufacturing: A review", *J. Manuf. Sci. Eng. Trans. ASME*, vol. 128, no. 4, pp 969-976, 2006.
24. K. Wang, "Applying data mining to manufacturing", *J. Intell. Manuf.*, vol. 18, pp 487-495, 2007.
25. M. Perzyk, "Statistical and visualization data mining tools for foundry production", *Archives of Foundry Engineering*, 7(3), 111-116, 2007.
26. M. Perzyk, R. Biernacki, J. Kozlowski, "Data mining in manufacturing: Significance analysis of process parameters", *Proc. Inst. Mech. Eng., Part B, J. Eng. Manuf.*, in print, available from author: m.perzyk@acn.waw.pl.
27. M. Perzyk, R. Biernacki, J. Kozlowski, "Data mining in manufacturing: methods, potentials, limitations", in Advances in Production Engineering Conference, Warsaw University of Technology, Poland, 13-16 June 2007, pp. 147-156 (Publishing and Printing House of the Institute for Sustainable Technologies – NRI, Radom, Poland).
28. P. Bhagat, *Pattern Recognition in Industry*, Elsevier, 2005.
29. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.
30. D. Fogel *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence*, IEEE Press, 2000.
31. D. Coley, *An Introduction to Genetic Algorithms for Scientists and Engineers*, World Scientific, 1999.
32. L. Rokach, O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing Co., 2008.
33. J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.
34. T. Kourti, J. MacGregor, "Multivariate SPC Methods for Process and Product Monitoring", *J. Qual. Tech*., 28, 409-428, 1996.
35. J. MacGregor, "Using On-Line Process Data to Improve Quality: Challenges for Statisticians", *International Statistical Rev*., 65, 309-323, 1997.
36. J. MacGregor, T. Kourti, "Statistical Process Control of Multivariate Processes", *Control Eng. Practice*, 3, 403-414, 1995.
37. P. Lewicki, T. Hill, "Multivariate quality control", *Quality Magazine*, April, 2007, p 38-45.
38. K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer Science+Business Media, 2007.
39. R. Roiger, M. Geatz, Data Mining: A Tutorial-Based Primer, Addison Wesley, 2003.
40. M. Perzyk, A. Kochanski, "Detection of causes of casting defects assisted by artificial neural networks", *Proc. of the Inst. Mech. Eng. Part B - J. Eng. Manuf.*, vol. 217, no. 9, pp 1279-1284, 2003.
41. M. Perzyk, A. Kochanski, "Prediction of ductile cast iron quality by artificial neural networks", *J. Mat. Proc. Tech.* vol. 109, no. 3, pp 305-307, 2001.
42. M. Perzyk, A. Kochanski, "Modelling of foundry processes by artificial neural networks", *Advances in Manufacturing Science and Technology*, 25(4), 2001.
43. M. Perzyk, A. Soroczynski, R. Biernacki, "Possibilities of decision trees applications for improvement of quality and economics of foundry production", 2008, available from author: m.perzyk@acn.waw.pl.
44. M. Perzyk, "Data mining in foundry production: Research in Polish metallurgy at the beginning of XXI century", *Committee of Metallurgy of the Polish Academy of Sciences*, ed. K. Swiatkowski, 2006, available from author: m.perzyk@acn.waw.pl.
45. A. Kochański, *Predicting of Ductile Cast Iron Properties by Artificial Neural Networks*, PhD thesis, Warsaw University of Technology, Faculty of Production Engineering, 1999 (in Polish).
46. A. Kochański, M. Perzyk, "Ductile cast iron classification by combined modeling", *Acta Metalurgica Slovaca*, vol. 7, n. 3, pp 50-55, 2001 (in Polish).
47. A. Kochanski, M. Perzyk, "New applications of artificial neural networks in foundry", *Acta Metallurgica Slovaca*, 7(3), 380-384, 2001 (in Polish).
48. A. Kochański, "Detection model of cast defects formation reasons (Wspomaganie wykrywania przyczyn powstawania wad w odlewach)", *Polska Metalurgia w Latach 2002-2006*, ed., Świątkowski K., Committee of Metallurgy of the Polish Academy of Science (Komitet Metalurgii PAN), Krynica, Poland (in Polish).
49. M. Perzyk, R. Biernacki, A. Kochanski, "Modeling of manufacturing processes by learning systems: The naive Bayesian classifier versus artificial neural networks", *J. Mat. Proc. Tech.*, Vol. 164-165, pp1430-1436, 2005.
50. D. Whitehouse, *Surfaces and Their Measurement*, Taylor and Francis, 2002.
51. J. Liker, *The Toyota Way*, McGraw-Hill, 2004.
52. S. Shingo, *A Study of the Toyota Production System*, Productivity Press, 1989.
53. S. Hino, *Inside the Mind of Toyota*, Productivity Press, 2006.
54. T. Davenport, J. Harris, *Competing on Analytics*, Harvard Business School Press, 2007.
55. S.J. Cunningham, "Machine learning and statistics: A matter of perspective", *New Zealand J Computing*, 6(1a):69-73, August 1995.
56. K. Parsaye, M. Chignell, *Intelligent Database Tools and Applications*, John Wiley & Sons, 1993.