NIST Advanced Manufacturing Series 100-7

Summary of the Symposium on Data Analytics for Advanced Manufacturing

Anantha Narayanan
Ronay Ak
Yung-Tsun Tina Lee
Rumi Ghosh
Sudarsan Rachuri

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AMS.100-7

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Advanced Manufacturing Series 100-7

Summary of the Symposium on Data Analytics for Advanced Manufacturing

Anantha Narayanan
Ronay Ak
Yung-Tsun Tina Lee
Systems Integration Division
Engineering Laboratory

Rumi Ghosh
Robert Bosch, LLC

Sudarsan Rachuri
Department of Energy

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AMS.100-7

April 2017

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Kent Rochford, Acting NIST Director and Under Secretary of Commerce for Standards and Technology

DISCLAIMER

Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

Summary of the Symposium on Data Analytics for Advanced Manufacturing

Anantha Narayanan, Ronay Ak, Yung-Tsun Tina Lee
Systems Integration Division, Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

Rumi Ghosh
Robert Bosch, LLC
Palo Alto, CA 94304, USA

Sudarsan Rachuri
Office of Energy Efficiency and Renewable Energy, Advanced Manufacturing Office
Department of Energy
Washington, DC 20585, USA

Abstract

The Symposium on Data Analytics for Advanced Manufacturing was held in conjunction with the 2016 IEEE International Conference on Big Data in Washington D.C. on December 6-7, 2016. The symposium was jointly organized by Advanced Manufacturing Office, Office of Energy Efficiency and Renewable Energy, Department of Energy and Systems Integration Division, Engineering Laboratory, National Institute of Standards and Technology, Department of Commerce. The Symposium's goal was to address the manufacturing industries' issues, needs, and challenges related to data analytics-based solutions for advanced manufacturing. The Symposium brought industry, government, and academic organizations together to exchange ideas and provide a common platform to build partnerships. This report summarizes the keynote presentations, technical paper presentations, and panel discussions of this symposium. The Symposium also hosted a manufacturing data challenge, which is summarized in this report. This report elicits the lessons learned from this symposium, and presents action items for future research on data analytics for advanced manufacturing.

1. Introduction

Advanced manufacturing includes not only the assembly of products on the shop floor, but also a wide variety of complex processes involving quality control, predictive maintenance, resource management, manufacturing-system control, and customer engagement. The manufacturing industry is a vital part of all economies and thus manufacturing companies seem to push themselves to increase manufacturing quality and simultaneously control variable production costs. A major challenge to the manufacturing industry is to improve its efficiency across the product life cycle (design, manufacturing, and use and post-use). Thus, an advanced manufacturing system requires capabilities and technologies for designing and improving the overall system performance [1-3].

Smart manufacturing systems have the ability to transform data gathered from a variety of manufacturing processes into actionable knowledge for decision making. Today, there exist in a large number of data analysis methods, algorithms, and tools – collectively known as data analytics (DA) – which are used extensively in many domains, but are yet to pervade the manufacturing domain [2-5]. The use of DA in manufacturing can lead to new insights and knowledge discovery that can help enhance quality, productivity, flexibility, reduce costs, and gain competitive advantages. However, DA tools are often complex and difficult to apply effectively. Manufacturers must identify the DA tools that best fit their objectives. The availability of DA tools, such as scikit-learn[1] and OpenScoring[2], on the internet is increasing access and greatly reducing cost.

The big data era has just emerged; manufacturers in a range of industries are now collecting real-time shop-floor data including data from Computer Numerical Control (CNC) machines. Many organizations are also capable of generating shop-floor-like data using simulation models. Continuous improvements in sensor technologies, data acquisition systems, and data mining techniques will allow the manufacturing industry to effectively and efficiently collect and analyze large and diverse volumes of data in real time or timely manner. These data, through DA, could reveal valuable insights and support performance improvements across multiple levels of manufacturing system functionality. For example, manufacturing managers can use DA to take a deep dive into historical process data, identify patterns and relationships among discrete process steps and inputs, and then optimize the factors that prove to have the greatest effect on yield [1].

Having great potential for converting raw data into information assets, DA could be a key competitive differentiator for smarter decision making during design, manufacturing, use and post-use. However, DA tools are difficult to integrate into existing manufacturing systems to control processes considering that many manufacturers do not have the interfaces as well as the required data analytics expertise. Integrating data with DA tools is often an issue because many DA tools have proprietary interfaces and have specialized data formats for input and output. Thus, there is a need for open standards and communication protocols.

In recent times, many researchers and industry practitioners from different engineering disciplines have been contributing to the manufacturing data analytics efforts. However, a special interest research group for this domain is still lacking. There is a need for bringing together leading practitioners and researchers from the industry, academia, and government to address common issues and specific challenges, and collaborate and nurture manufacturing innovation using DA for design, manufacturing, post-manufacturing, and reuse phases for advanced manufacturing systems. This need has led to the initiation of the "Symposium on Data Analytics for Advanced Manufacturing," as a part of the 2016 International Conference on Big Data held at Washington, D.C. In the following pages we detail the symposium.

[1] http://scikit-learn.org/
[2] https://github.com/openscoring/openscoring

This report is organized as follows: Section 2 presents an overview and objective of the Symposium. The keynote presentations, technical paper presentations, panel presentations and discussions are summarized in Sections 3-5, respectively. Section 6 introduces the manufacturing data challenge. Finally, Section 7 provides concluding remarks and discussion for future work.

2. Overview of the Symposium

The Symposium on Data Analytics for Advanced Manufacturing was held in conjunction with the 2016 IEEE International Conference on Big Data (IEEE Big Data) in Washington D.C. on December 6-7, 2016 [6]. The theme of the Symposium was “From Sensing to Decision-Making.” The primary objective of the Symposium was to bring together researchers and industry practitioners from the manufacturing, information science, and data science disciplines to identify and discuss issues, needs, and challenges related to manufacturing DA. Through the Symposium, the participants discussed common issues and specific challenges related to DA, and shared their research results and practical design and development experiences in manufacturing DA applications. The Symposium fostered collaboration among academia, industry, and governmental organizations to help improve efficiency, product quality, and sustainability in the manufacturing industries. About 45 to 50 people attended the Symposium; nearly half of them were from industry while the other half were from universities, research and government institutes.

The Symposium was expanded from the Special Session of “From Data to Insight: Big Data and Analytics for Smart Manufacturing” of the 2014 IEEE Big Data and 2015 IEEE Big Data. Both of the 2014 and 2015 events were well attended and hence the conference highly supported having such symposium held at their 2016 conference.

The Symposium was organized and held by Dr. Sudarsan Rachuri from the U.S. Department of Energy (DOE), Dr. Ronay Ak and Tina Lee from the National Institute of Standards and Technology (NIST), Dr. Anantha Narayanan from the University of Maryland (UMD), and Dr. Rumi Ghosh from the Robert Bosch LLC (Bosch). Their detailed biographical information can be found in APPENDIX A. The Symposium was a two-day program and consisted of four main parts: 1) keynote presentations, 2) a panel session, 3) technical paper presentations, and 4) the Manufacturing Data Challenge session (see APPENDIX C for the complete agenda).

The Symposium began with Dr. Rachuri’s opening remarks; he stressed the importance of smart manufacturing, the role of DA, and the needed standards and best practices to support DA. Ms. Lee gave opening remarks on the second day where she summarized the first day’s events. Dr. Ak and Dr. Narayanan chaired technical paper presentation sessions in Day 1 and Day 2, respectively. Dr. Ghosh served as the chair for the “Manufacturing Data Challenge” session.

The detailed information about the keynote talks, paper presentations, panel discussion, and the Manufacturing Data Challenge are provided in the following sections.

3. Keynote Presentations

The Symposium featured three keynote presentations from leaders and experts in government and industry. In addition to the three keynote presentations, the Symposium organizers invited Dr. Mark Johnson from the Advanced Manufacturing Office at the U.S. Department of Energy, to deliver a keynote at the general conference. This effort is intended to bridge gap between the manufacturing DA community and the larger general big data analytics community. Dr. Johnson gave a review of the Advanced Manufacturing Offices work to leverage high performance computing, smart manufacturing approaches for the U.S. clean energy manufacturing sector—through targeted R&D in modeling and simulation and partnerships with industry, academia, technology incubators and other stakeholders. The details on Dr. Johnson's talk "Leveraging High Performance Computing to Drive Advanced Manufacturing R&D at the US Department of Energy" are available on the conference website [6].

To get a better understanding of public-private partnerships and research efforts, the symposium had three invited talks. Dr. Frank Gayle, Deputy Director of the interagency Advanced Manufacturing National Program Office (AMNPO), NIST, gave an overview of Manufacturing USA program to set the stage for public private partnerships. This was followed by two industry presentations. Dr. Matteo Bellucci, Engineering Leader, GE Additive, GE Global Research, gave an overview of GE Brilliant Factory efforts and the digital transformation at GE. In the next industry talk, Dr. Rumi Ghosh, Senior Data Mining Engineer, Robert Bosch, LLC gave an overview of data analytics activities at Bosch. Dr. Ghosh also planned and ran the Manufacturing Data Challenge, hosted on Kaggle. See Section 6 for further details. We summarize below the three keynote presentations of the Symposium. The detailed bios of the keynote speakers can be found in APPENDIX B.

3.1. **The National Network for Manufacturing Innovation**, *Dr. Frank Gayle, Deputy Director of the interagency Advanced Manufacturing National Program Office (AMNPO), NIST*

Dr. Gayle started his presentation by introducing the interagency Advanced Manufacturing National Program Office. He described the importance of manufacturing to national well-being, and provided information about the U.S. manufacturing industry's current situation in the global market. He pointed out that the U.S. is losing its competitive advantage in the advanced-technology based products market; he claimed that one of the main reasons is that the products invented in the U.S. are being manufactured elsewhere, and that innovation tends to go where products are manufactured. He talked about how the Manufacturing USA program, including 14 manufacturing innovation institutes, has been put in place to rejuvenate manufacturing in the U.S. The vision of the program is to establish global leadership in advanced manufacturing with a mission of connecting people, ideas, and technologies to solve industry-relevant advanced manufacturing challenges, thereby enhancing industrial competitiveness and economic growth, and strengthening the U.S. national security. Dr. Gayle stated the four goals of the program are to (i) increase competitiveness, (ii) facilitate technology transition, (iii) accelerate the manufacturing workforce, and (iv) ensure stable and sustainable manufacturing. He continued his presentation with the institute's design and up-to-date progress of the program. He concluded his speech by focusing on America Makes and its success in developing processes and accelerating adoption of additive manufacturing technologies in the U.S. manufacturing sector, yielding innovative products and increased domestic manufacturing competitiveness.

3.2. The General Electric (GE) Brilliant Factory, *Dr. Matteo Bellucci, Engineering Leader, GE Additive, GE Global Research*

Dr. Bellucci began with an introduction to the "Brilliant Factory" concept being developed at GE. He and his team are leading most of the R&D Brilliant Factories efforts across the company, spanning virtual validation of new factories as well as processes such as casting and additive manufacturing. He pointed out that the GE Brilliant Factory envisions an approach to enable digital manufacturing that brings total digital integration within and between every part of the value chain, starting from the design phase all the way to supply chain and service. The digital thread connects product development and design, manufacturing system and process design, material flow systems, manual and automated fabrication and assembly processes, quality verification, distribution, service, and life cycle management. The Brilliant Factory concept allows for value to be realized in each of these parts of the value chain. Collaboration between designers, manufacturing engineers and operations, is enabled by a "Digital Thread". He claimed that the use of Brilliant Factory tools is essential for simulating individual manufacturing processes and the total manufacturing system. By driving compatibility between the product design and the manufacturing plants, these virtual tools and methods enable the early optimization of cost, quality, and time to help achieve integrated products, process and resource design, and affordability. Dr. Bellucci also mentioned about GE Predix [7], GE's operating system for the Industrial Internet. Predix, the world's first-ever Industrial Internet platform, provides real-time, big data analytics capabilities. He also mentioned that GE is investigating in additive manufacturing (AM) and GE Additive will be their new business line. He concluded by pointing out the role of and the need for standards for transferring data and models online between different DA environments. He commented that validating predictions both from design of analytics and also in measurements are critical for accurate decision-making.

3.3. From Sensors to Sensing- Industrial Data Mining at Bosch, *Dr. Rumi Ghosh, Senior Data Mining Engineer, Robert Bosch, LLC*

Dr. Ghosh began with an introduction to Bosch and their data mining activities. She talked about the business sectors in which data mining is widely used. She gave information about Bosch sensor technology, and said that 50% of smartphones worldwide already use Bosch sensors. She identified where big data analytics adds value in the context of product life cycle analytics, and presented some use cases showcasing some of their key research challenges. These use cases included (i) root cause analyses for scrap reduction where top parameters influencing the scrap are identified, (ii) the test time reduction tool which aims at reducing test efforts without external support, and (iii) classification problem in the presence of highly imbalanced dataset. She presented a novel perspective to the classification problem with imbalanced data, and demonstrated her technique on a manufacturing application. She identified barriers to the progress of industrial data mining with a special emphasis on the need for standardization. She claimed that significant investment in infrastructure is needed, and interoperability and standardization of data collection, storage, retrieval, and presentation are required. She highlighted Bosch's effort in this direction and demonstrated the use of the leading data mining standard, the Predictive Model Markup Language (PMML) [8]. She concluded that (i) PMML provides freedom of development tools for data scientists, that is PMML allows data scientists to easily share predictive analytic models between different applications; and (ii) PMML standard's model coverage capability should be increased and error handling can potentially be improved.

4. Technical Presentations

The Symposium received seventeen (17) paper submissions. Through a rigorous peer review process, ten (10) papers were selected for full paper presentations, representing an acceptance rate of 59%. Four of these ten papers were related to the Manufacturing Data Challenge, hosted on Kaggle [9]. The four papers will be described in detail in Section 6. The presentations of the other six papers are summarized below.

4.1. Evaluation of a PMML-Based GPR Scoring Engine on a Cloud Platform and Microcomputer Board for Smart Manufacturing, *Max Ferguson, Kincho Law, Raunak Bhinge, Yung-Tsun Tina Lee, and Jinkyoo Park*

This paper presented a scoring engine for Gaussian Process Regression (GPR) based on PMML. Note that GPR is included in the PMML's latest version, v4.3, and NIST contributed to the PMML/GPR standardization effort. This paper evaluated deployment strategies for GPR models in a manufacturing context by comparing cloud platforms and local micro-computers. The Raspberry Pi platform was chosen as the local micro-computer platform. The paper concluded that under experimental settings, the cloud platform achieved a lower response time and higher prediction rate than the local microcomputer. The question on network latency affecting response time in real manufacturing situations in practice was discussed. Evaluating the performance of different techniques to deploy standards-based analytics solutions is important for widespread acceptance and use of data analytics in the manufacturing industry. This paper was selected as the best paper of the Symposium.

4.2. Cloud-Based Machine Learning for Predictive Analytics: Tool Wear Prediction in Milling, *Dazhong Wu, Connor Jennings, Janis Terpenny, and Soundar Kumara*

This paper presented a tool wear prediction algorithm based on a novel parallelized random forest algorithm. The algorithm was implemented on a public cloud based on the MapReduce framework. Data analytics promises new possibilities for machine prognostics and health management, where traditional physics based models are difficult to formulate and apply. However, ability to process large volumes of data and computational efficiency are important challenges. The techniques (e.g., parallel random forest) presented in this paper addressed those challenges.

4.3. A System and Architecture for Reusable Abstractions of Manufacturing Processes, *Alexander Brodsky, Mohan Krishnamoorthy, William Bernstein, and M. Omar Nachawati*

This paper proposed an architecture to manage a repository for conducting analysis and optimization on manufacturing performance models. The architecture supports a repository of unit manufacturing process (UMP) models, as well as a knowledge base ecosystem for manufacturing users. The goal is to promote the rapid performance-related assessment of a manufacturing system through the reuse of existing manufacturing performance models. As new algorithms and tools are developed to support data analytics, it is important to have a structured approach to facilitate the selection and application of reusable models for manufacturing analysis. This paper focuses on fulfilling this requirement.

4.4. Convergence and Divergence in Academic and Industrial Interests on IOT Based Manufacturing, *Srinivasan Radhakrishnan and Sagar Kamarthi*

This paper presented a knowledge map of terms related to the Internet of Things (IoT) appearing in academic research, and compared it with work in the industry. The terms were retrieved from a large database of published technical articles. The paper used statistical techniques to evaluate the direction of research in academia and industry. IoT is a quickly evolving area. It promises great technological capabilities, but it is important to understand the relationships between the technology challenges and business requirements. This paper presented valuable insights into these relationships.

4.5. Complexity-Entropy Feature Plane for Gear Fault Detection, *Srinivasan Radhakrishnan and Sagar Kamarthi*

This paper uses the Complexity-Entropy Causality Plane technique to classify machine condition as faulty or normal based on vibration signals. The approach requires minimal preprocessing of signals, and is insensitive to external noise, non-stationarity, and trends. The method generates two feature vectors that can be graphed to visually distinguish normal and faulty conditions. Data analytics techniques usually require some expertise to interpret the results, which makes it difficult for average manufacturers. The technique presented in this paper is easier to apply and makes it easy to interpret the results.

4.6. Advancing Additive Manufacturing Through Visual Data Science (an invited talk), *Dr. Chad A. Steed, Team Leader and Senior Researcher in the Computational Data Analytics Group at the Oak Ridge National Laboratory (ORNL)*

This presentation introduced the Advanced Manufacturing (AM) activities at the Oak Ridge National Laboratory (ORNL) Manufacturing Demonstration Facility (MDF). As advances in AM lead to unprecedented results in 3D printing, big data analytics are crucial to understanding the synthesis process and improving the quality of the build results. With this aim in mind, an interdisciplinary team of researchers based out of ORNL MDF are pursuing multiple threads of research to improve the efficacy of 3D printing through the application of interactive big data analytics. One research thread involves the application of information visualization and visual analytics techniques, which allow researchers to explore large and complex 3D printer log files. Another thrust leverages advanced computer vision algorithms that automatically detect microstructure properties (e.g., porosity, hot spots) in near infrared imagery captured during 3D printer builds. The AM paradigm and the role of the advanced big data analytics in analyzing the large volume of data produced from AM processes were presented. The Falcon [10], a temporal visual analytics system for exploring long, complex, and multivariate time series data was introduced. The value of big data analytics in AM using Falcon was also demonstrated. The presentation concluded with a case study that involves real 3D printer data and ties the two thrusts together.

5. Panel Session

The Symposium included a two-hour panel session, in which experts from industry, academia, and government led a discussion with the attendees on topics related to the application of data analytics in smart manufacturing. The panelists and their affiliations are listed below:

- Dr. Valerie Coffman, Chief Technology Officer, Xometry.
- Sivaramakumar Gopalasundaram, Manager of Department of Data Analytics, Cognizant.
- Dr. Mateo Bellucci, Engineering Leader of GE Additive, GE Global Research.
- Dr. Soundar Kumara, Professor of Industrial and Manufacturing Engineering, The Pennsylvania State University.
- Dr. Sankaran Mahadevan, Professor of Civil and Environmental Engineering, Vanderbilt University.
- Dr. Ram D. Sriram, Chief of the Software and Systems Division, Information Technology Laboratory, NIST.

The session was moderated by Dr. Rachuri. More information about the panelists can be found in APPENDIX B.

The following topics are given in advance for panel discussion:

- Current state of the art and technology trends in manufacturing data analytics
    - Smart sensors and data acquisition
    - High performance and distributed computing for data analytics and workflow
    - Network and cloud infrastructure, enterprise data management, and IT/Operations integration
- Advanced manufacturing applications
    - Time scale and "real-time" data analytics
    - Issues in translating data into actions and results
    - Investment strategy and innovation across life cycle (design, manufacturing, use and post-use)
- Role of standards
    - Validation and uncertainty quantification (UQ) of prediction models
    - Standards, protocols, and infrastructure for manufacturing analytics

The panelists began the session with a short introduction and a summary of what they felt were important technical challenges to be addressed. This was followed by a group discussion session with attendees asking questions. Here, we summarize the questions and discussions of the panel session.

Question: From the manufacturing perspective, can you comment on good practices for communication between manufacturing engineers and data scientists?

Answers:
Dr. Bellucci recommended mission-based themes – take someone from each skill, have them spend weeks/months/days in direct meetings formulating the problem. He observed that co-location helps much more than teleconferencing.

Dr. Sriram noted that there are two schools of thought – teaching fundamentals of the factory floor to the data scientists, or teaching data science to the manufacturing engineers, where the latter being the more holistic approach.

Dr. Kumara stated the importance of giving the ownership to the people in the plant (such as a foreman), and working with the plant employees to build an atmosphere of trust and confidence.

Question: What are the challenges in formally representing the capabilities of a process or machine?

Answers:
Dr. Bellucci stressed that finding a right representation from the end user and design engineer's perspective is important. Many smart manufacturing tools are available; the challenge is finding and matching the right tools. He also stressed that communication and convincing is required. In the beginning, codify the knowledge of the experts, and once a critical mass is achieved, rules can be automatically inferred. As more users start making decisions new rules can be created. Dr. Bellucci also observed that typically the flow of information is fragmented, and no single person who is clear about how their decision affects the whole flow. It is important to give the owner of the design immediate visibility.

Dr. Coffman mentioned that at Xometry, they define their own database of processes, and each material is matched to a process using machine learning. Classifying parts is more difficult. The challenge is in knowing the geometric features that become machine learning features, so that part to process matching may be automated.

Dr. Kumara brought attention to the Defense Advanced Research Projects Agency (DARPA) component, context, and manufacturing model libraries (C2M2L) [11] which captures the data structures and data elements needed to build these libraries. This can become a standard model as it is open source.

Question: Xometry uses machine learning to classify parts and match clients. You may find out you need custom algorithms for specific problems. What are the opportunities for new machine learning techniques here, beyond basic feature construction?

Answers:
Dr. Coffman mentioned that one open topic is developing a quantitative recommendation engine to answer questions, such as "how likely is a machine shop to take a job?"

Dr. Kumara expressed interest in explainable models – machine learning models that can quantitatively explain why a result is achieved. He pointed to DARPA's work on explainable artificial intelligence [12].

Dr. Sriram raised the question of how algorithms can be implemented in a novel manner, to best fit a particular application domain. He suggested finding new learning mechanisms that are more symbolic learning focused, rather than purely statistical learning (including deep learning) techniques.

Question: How should we describe/annotate processes so that when we find our UMP we can easily find a model that fits our needs?

Answers:
Dr. Sriram suggested that meta-data can be used to record information, and scenarios may be re-created from meta-data.

Dr. Mahadevan observed that the same ideas for matching part features to manufacturers may be applied for annotating models. It also depends on the context, and may vary with the type of manufacturing.

Dr. Kumara brought attention to Roger Schank's script theory [13-15] which can help in formalizing the meta data.

Question: What are your comments about human in the loop?

Answers:
Dr. Mahadevan observed that a lot of work on human performance, human error, etc. come from transportation and nuclear power industries. In the latter, there are sustained efforts towards human risk assessment to be included in probabilistic risk assessment of nuclear power plants. What is missing is how human error contributes to manufacturing quality. Errors of omission and commission are serious. How do you incorporate that into overall quality assessment? Current research should work on quantifying the effects of human errors and incorporate the resulting uncertainty into decision making process. Multiple factors affect human performance, and Bayesian probabilistic analysis and evidence theory can be used to quantify the effects of these factors on manufacturing system performance and product quality.

Dr. Bellucci raised the question of accountability – humans are not accurate, but they are accountable and trusted. Before algorithms can be held accountable they need to be "trusted" therefore significant efforts in development and validation effort is required before business decision will be delegated.

Question: Are there any classes of machine learning or data mining algorithms that are particularly useful for manufacturing?

Answers:
Dr. Mahadevan mentioned that Bayesian algorithms are very useful in manufacturing, because of their ability to fuse heterogeneous data sources. Neural networks for deep learning are another avenue for exploration. He also observed that while new models are being developed every day, it is important to be able to verify and validate these algorithms, and set the context for their proper use. He mentioned that it is important to look at standardized procedures for verification and validation of data driven models, and establishing uncertainty and confidence when extrapolating them to real world problems.

Dr. Sriram stated that it is a challenge to validate unsupervised learning models.

Question: How do you see the evolution of artificial intelligence (AI) in manufacturing, compared to other domains?

Answers:
Dr. Sriram stated that considerable progress has been made in applying AI techniques on the shop floor activities, but this progress has been slow in conceptual design. A few domains such as the financial domain have seen considerable application of AI techniques, while domains such as healthcare are making slow progress.

Dr. Bellucci observed that there is a need to increase productivity in manufacturing, which leads to investment in smart technologies. How it is implemented in day-to-day operations is a challenge.

6. Manufacturing Data Challenge

The Manufacturing Data Challenge was sponsored by Bosch, and coordinated by Dr. Rumi Ghosh. Bosch is one of the world's leading manufacturing companies and has an imperative to ensure that the recipes for the production of its advanced mechanical components are of the highest quality with high safety standards. Bosch achieves this goal by closely monitoring their parts as they progress through the manufacturing processes. Because Bosch records data at every step along its assembly lines, they have the ability to apply advanced analytics to improve these manufacturing processes. However, the intricacies of the data and complexities of the production line pose problems for current methods.

Bosch released its huge dataset consisting of anonymized records of measurements and tests made for each component along the assembly line in a Kaggle competition. Kaggle [9] was founded in 2010 as a platform for predictive modelling and analytics competitions, on which companies and researchers post their data, and statisticians and data miners from all over the world compete to produce the best models. In this competition, Bosch challenged Kagglers to predict internal failures using thousands of measurements and tests made for each component along the assembly line. This would enable Bosch to bring quality products at lower costs to the end user.

The following attributes made this competition uniquely challenging:

- A very large data set: Around 3 million records were provided.
- A high dimensional feature space: The data included 2,140 categorical, 968 numerical and 1,157 date features.
- Lack of domain knowledge: Because the features were anonymized, Kagglers could not apply domain knowledge to tune their solutions. The importance of the features had to be discovered by the algorithms.
- A rare event prediction: In addition to being one of the largest datasets (in terms of number of features) ever hosted on Kaggle, the ground truth for this competition is that the data set is highly imbalanced. The failure rate was ~ 0.6 % which means less than 1 % of the population falls under one of the classes. Therefore, the traditional classification algorithms fail to provide accurate prediction results for the test set.

The objective of the competition was to process these features into something meaningful so they can be used to build a predictive model. The competition was launched on August 11, 2016 and the final submission deadline was November 11, 2016. Within these 3 months, around 1,300 teams submitted their solutions on Kaggle.

6.1. Data sets

The data for this competition represent measurements of parts as they move through Bosch's production lines. Each part has a unique Id. The goal is to predict which parts will fail quality control (represented by a 'Response' = 1).

The dataset contains an extremely large number of anonymized features. Features are named according to a convention that tells you the production line, the station on the line, and a feature number. For example, L3_S36_F3939 is a feature measured on Line 3, Station 36, and is Feature number 3939.

On account of the large size of the dataset, the Bosch team has separated the files by the type of feature they contain: numerical, categorical (including 500 are multi-value, 1,490 single value, and 150 empty), and finally, a file with date features. The date features provide a timestamp for each measurement taken.

For each Id in the test set, the Kagglers are expected to predict a binary prediction for the response variable.

6.2. Evaluation

Submissions are evaluated on the Matthews Correlation Coefficient (MCC) [16] between the predicted and the observed response. MCC is a measure of binary classifications. It takes into account all elements of the confusion matrix (true and false positives and negatives). The MCC is given by:

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \quad ,$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. MCC takes a value between -1 and +1. A value close to +1 represents perfect prediction, 0 is no better than random, and -1 represents total disagreement between prediction and observation.

6.3. Travel Grant by Bosch

Bosch encouraged Kagglers to submit their solutions to the Symposium as a technical paper and paper presentation by offering a travel grant. We received 11 full paper submissions related to this Challenge. All the papers were peer reviewed by experts in the field. Bosch provided a travel grant of $2000 for the top three papers that received the highest review scores among all submissions. Based on the review scores, the following papers were chosen as the finalists:

- Bayesian Optimization for Predicting Rare Internal Failures in Manufacturing Processes (by Abhinav Maurya)
- Using Big Data to enhance the Bosch Production Line Performance: A Kaggle Challenge (by Ankita Mangal and Nishant Kumar)
- Machine Learning, Linear and Bayesian Models for Logistic Regression in the Failure Detection Problems (by Bohdan Pavlyshenko).

6.4. Technical Paper Submissions Related to Data Challenge

Four papers, which are related to the Kaggle manufacturing data challenge, were presented at the Symposium. These presentations are summarized in Sections 6.4.1-6.4.4.

6.4.1. **Machine Learning, Linear and Bayesian Models for Logistic Regression in the Failure Detection Problems**, *Bohdan Pavlyshenko*

This paper presented a solution based on logistic regression, using three different methods. The paper presented a machine learning model based on (i) the XGBoost classifier, a generalized linear model, and (ii) a Bayesian model. The XGBoost model gave high scoring failure detection, while the Bayesian model provided statistical distributions of model parameters, which are useful for risk assessment.

6.4.2. **Predict Failed Product Using Large-scale Data: A Two-stage Approach with Clustering and Supervised Learning**, *Darui Zhang, Bin Xu, and Jasmine Wood*

This paper presented a two-stage approach to solve the challenge problem. The first step clustered the data into groups based on the manufacturing process, and the second step used supervised learning to predict product failure in each cluster. The random forest technique was used to build the prediction model.

6.4.3. **Bayesian Optimization for Predicting Rare Internal Failures in Manufacturing Processes**, *Abhinav Maurya*

This paper presented a Bayesian optimization technique to optimize the MCC to solve the challenge problem. The paper used Gradient Boosting Machine as the base classifier. It applied a meta-optimization algorithm that directly maximized MCC by learning optimal weights for the positive and negative classes. This approach significantly improved classification performance for highly imbalanced, high dimensional data sets.

6.4.4. **Using Big Data to Enhance the Bosch Production Line Performance: A Kaggle Challenge**, *Ankita Mangal and Nishant Kumar*

This paper first used different data mining techniques to understand the anonymized features in the high dimensional data set, producing interesting insights into the data set. The paper deduced a set of important features, and identified interesting observations such as weekly periodicity in the observations recorded per day. The paper used an XGBoost model to make the final prediction.

7. Concluding Remarks

The Symposium was the third successive year that we have conducted manufacturing-oriented events in conjunction with the IEEE Big Data Conference. This 2016 event was expanded to a symposium, to increase participation and enrich the content. The Symposium had an attendance of 45 to 50 people comprising a mixture of industry, academia, and government. The Symposium discussed some of the most critical challenges and opportunities in applying data analytics and related standards to improve smart manufacturing systems. Participants at the Symposium presented some of their cutting edge research, and discussed some of the most challenging problems in this area. This report summarized the contributions presented at the Symposium.

The organizers received encouraging feedback from the attendees during and at the end of the Symposium. An informal survey conducted at the end of the Symposium commented that the quality of the technical presentations was good, and the amount of time allotted to them was appropriate. The survey also indicated that the attendees generally found the Symposium a good learning experience. Some of the feedback suggested that it would be beneficial to have more industry involvement in the Symposium. Planned industry visits and a separate industry track were recommended for the next edition of the Symposium. The manufacturing data challenge garnered a lot of attention, and was a great success. In the future, more attention and resources will be devoted towards such challenge problems, as we believe solving these problems will greatly benefit to both the research community and the practitioners in the field. Two keynote speakers emphasized the necessity of high quality standards bridging data analytics and manufacturing applications. It is worth mentioning that some of the technical presentations demonstrated the usefulness of such standards (such as PMML) in practical manufacturing applications.

The lessons learned from this symposium will be used to improve the organization of the next edition of the symposium. We plan to hold an improved symposium collocated with the 2017 IEEE

Big Data Conference, and build on this continuing tradition. Our aim is to make the *Symposium on Data Analytics for Advanced Manufacturing* the venue of choice for cutting edge research and industry participation in the area of data analytics for advanced manufacturing.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Steve Eglash from Stanford University, for his valuable inputs towards organizing this Symposium. We owe special thanks to Prof. Xiaohua Tony Hu from Drexel University, for organizing the IEEE International Conference on Big Data, and providing us a venue to host the Symposium. We would like to thank Dr. Vijay Srinivasan, Chief of the Systems Integration Division at NIST, for his support and encouragement for conducting this Symposium. We also thank the keynote speakers and the authors who submitted their papers to the Symposium, for their valuable technical contributions. Finally, we are grateful to all the participants who attended the Symposium.

References

1. Auschitzky, E., Hammer, M., and Rajagopaul, A., July 2014, How Big Data Can Improve Manufacturing, McKinsey & Company.
2. Manufacturing and the Data Conundrum, 2014, A report by The Economist Intelligence Unit Online. Available: https://www.eiuperspectives.economist.com/sites/default/files/Manufacturing_Data_Conundrum_Jul14.pdf.
3. MacDougall, W., July 2014, Industrie 4.0-Smart Manufacturing for the Future, Germany Trade & Invest, Berlin, Germany, Online. Available: https://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Brochures/Industries/industrie4.0-smart-manufacturing-for-the-future-en.pdf.
4. Dornfeld, D. and DeVries, M. F., 1990, "Neural Network Sensor Fusion for Tool Condition Monitoring," *CIRP Annals-Manufacturing Technology*, vol. 39, no. 1, pp. 101-105.
5. Bukkapatnam, T. S., Lakhtakia, A., and Kumara S., 1997, "Chaotic Neurons for On-Line Quality Control in Manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 13, no. 2, pp. 95-100.
6. IEEE International Conference on Big Data (IEEE BigData 2016), 2016, Online. Available: http://cci.drexel.edu/bigdata/bigdata2016/.
7. PREDIX, Online. Available: https://www.ge.com/digital/predix.
8. The Predictive Model Markup Language (PMML), Online. Available: http://dmg.org/.
9. Kaggle, Online. Available: https://www.kaggle.com/.
10. Falcon, Online. Available: http://csteed.com/projects/.
11. DARPA C2M2L Program, Online. Available: http://cps-vo.org/group/avm/c2m2l.
12. Explainable Artificial Intelligence (XAI), DARPA, Online. Available: http://www.darpa.mil/program/explainable-artificial-intelligence.
13. Script Theory, Online. Available: http://www.instructionaldesign.org/theories/script-theory.html.
14. Schank, R.C., 1975, Conceptual Information Processing. New York: Elsevier.

15. Schank, R.C., 1982, Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press.
16. Matthews, B. W., 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2, pp. 442-451.

APPENDIX A – The Organizers of the Symposium

Ronay Ak is a Research Associate in the Systems Integration Division of the Engineering Laboratory (EL) at NIST. She received her PhD in the Energy & Power Systems (Engineering) Discipline in the Chair on Systems Science and the Energetic Challenge-EDF Chair which is shared between two top engineering schools- Ecole Centrale Paris (ECP) and Ecole Superieure d'Electricite (SUPELEC) in Paris, France. Her research interests include manufacturing process monitoring and control, advanced data analytics, multi-objective optimization, and reliability analysis.

Rumi Ghosh is a senior data mining (DM) engineer at Robert Bosch, LLC. Before joining Bosch, her research focused on graph analysis, connections between people and social media analysis. When she joined Bosch, she forayed into internet of things. Her responsibilities at Bosch include development of algorithms for real data mining and machine learning problems for a wide spectrum of domains ranging from connected industries to connected mobility to connected homes. She has also been involved in conversion of point solutions for DM projects to self-serving automated DM tools in the area of industrial internet of things.

Yung-Tsun Tina Lee is a computer scientist in the Systems Integration Division of the Engineering Laboratory at NIST. She leads the Data Analytics for Smart Manufacturing Systems Project of the Smart Manufacturing Systems Design and Analysis Program. She was the co-editor of the Simulation Interoperability Standards Organization (SISO) Standards of Core Manufacturing Simulation Data (CMSD) and is currently the chair of CMSD Product Support Group.

Anantha Narayanan is a Research Associate at the University of Maryland, and works as a Guest Researcher at the Systems Integration Division at the National Institute of Standards and Technology (NIST). Anantha received his PhD in Computer Science from Vanderbilt University in Nashville, Tennessee. His research interests include data analytics, domain specific modeling, model transformations, and smart manufacturing. He is currently working on the Data Analytics for Smart Manufacturing project at NIST.

Sudarsan Rachuri is currently the Federal Program Manager for clean energy manufacturing, Innovation Institute of Smart Manufacturing which was funded up to $70 million through the Advanced Manufacturing Office (AMO) in the Office of Energy Efficiency and Renewable Energy (EERE). Prior to joining DOE, Dr. Rachuri spent 20 years at NIST and initiated several programs including the Sustainable Manufacturing Program and Smart Manufacturing Systems Design and Analysis Program.

APPENDIX B – The Keynote Speakers and Panelists of the Symposium

Mark Johnson serves as the Director of the Advanced Manufacturing Office (AMO) in the Office of Energy Efficiency and Renewable Energy (EERE). AMO is focused on creating a fertile innovation environment for advanced manufacturing, enabling vigorous domestic development of new energy-efficient manufacturing processes and materials technologies to reduce the energy intensity and life-cycle energy consumption of manufactured products. Previously, Dr. Johnson served as a Program Director in the Advanced Research Projects Agency–Energy (ARPA-E) where he had the longest tenure in that post—from ARPA-E's formation in 2010 to mid-2013. At ARPA-E, Dr. Johnson led initiatives to advance energy storage and critical materials, as well as projects in small business, advanced semiconductor, novel wind architectures, superconductors and electric machines. He also served as the Industry and Innovation Program Director for the Future Renewable Electric Energy Delivery and Management (FREEDM) Systems Center. This is a National Science Foundation Gen-111 Engineering Research Center targeting the convergence of power electronics, energy storage, renewable resource integration and information technology for electric power systems. He joins EERE on assignment from North Carolina State University, where he is an Associate Professor of Materials Science and Engineering. His research has focused on crystal growth and device fabrication of compound semiconductor materials with electronic and photonic applications. He also taught in the Technology, Entrepreneurship and Commercialization program jointly between the NC State Colleges of Management and Engineering. In addition to his academic career, he is an entrepreneur and early stage leader in Quantum Epitaxial Designs (now International Quantum Epitaxy), EPI Systems (now Veeco) and Nitronex (now GaAs Labs). Dr. Johnson has a bachelor's degree from MIT and a Ph.D., from NC State, both in Materials Science and Engineering.

Matteo Bellucci is one of the Engineering Leaders at GE Additive since January 2017. He led the Process System Lab at GE Global Research Center from November 2011 to January 2017. He and his team were leading most of the R&D Brilliant Factories efforts across the companies, spanning virtual validation of new factories as well as processes such as Casting and Additive. His team also understand how to optimize maintenance, and increase automation and throughput of various manufacturing processes. Dr. Bellucci obtained his PhD in Aerospace Engineering from University of Naples.

Valerie Coffman is a graduate of Johns Hopkins and received a PhD in Physics from Cornell where she wrote software for studying the fracture properties of materials. After graduation, she spent 5 years at the National Institute of Standards and Technology (NIST) writing open source software for materials science research. Valerie joined Xometry as Chief Technology Officer in 2014.

Frank Gayle is Deputy Director of the interagency Advanced Manufacturing National Program Office (AMNPO) which is headquartered at the National Institute of Standards and Technology (NIST). Before joining the AMNPO in December 2012, Dr. Gayle worked in the NIST Metallurgy Division in positions from research metallurgist to Division Chief. As Division Chief, he was responsible for broad support of industry needs for measurements, standards, and data in the application of metals. As Deputy Director, Dr. Gayle is responsible for the operations of the AMNPO, and leads efforts to carry out the Congressionally mandated development of the Manufacturing USA program.

Sivaramakumar Gopalasundaram working as manager in Cognizant in the department of Data Analytics. Dr. Siva has a doctorate degree in Adaptive systems from Indian Institute of Science, Bangalore, India. He has worked in the area of Supply chain optimization, business forecasting and data analytics in various industrial sectors such as chemical processes, discrete manufacturing, health care, semiconductor, retail, automotive and communication.

Soundar Kumara is the Allen, E., and Allen, M., Pearce Professor of Industrial Engineering at Penn State. He also holds a joint appointment with the Department of Computer Science. Has an affiliate appointment with the school of Information Sciences and Technology. His research interests are in Manufacturing Process Monitoring, IOT in Manufacturing and Service Sectors, Health Analytics, Graph Analytics and Large Scale Complex Networks. He is a Fellow of Institute of Industrial Engineers (IIE), Fellow of the International Academy of Production Engineering (CIRP), and Fellow of American Association for Advancement of Science (AAAS), and American Association of Mechanical Engineers (ASME). He has more than 250 publications to his credit and several of his papers have won best paper awards. 52 Ph.D., and 54 MS students graduated under his tutelage. The Label Propagation Algorithm developed by his team is currently one of the fastest community detection algorithms (near linear time) and is part of many software libraries (R, Python, and iGraph).

Sankaran Mahadevan is Professor of Civil and Environmental Engineering at Vanderbilt University, Nashville, Tennessee, where he has served since 1988. He also has a joint appointment as Professor of Mechanical Engineering. His research interests are in reliability and uncertainty analysis methods, material degradation, structural health monitoring, design optimization, and model uncertainty. The methods have been applied to civil, mechanical and aerospace systems. This research has been funded by NSF, NASA (Glen, Marshall, Langley, Ames), FAA, U. S. DOE, U. S. DOT, Nuclear Regulatory Commission, U. S. Army Research Office, U.S. Air Force, U. S. Army Corps of Engineers, General Motors, Chrysler, Union Pacific, Transportation Technology Center, and the Sandia, Los Alamos, Idaho and ORNL. Prof. Mahadevan has directed 30 Ph.D. dissertations and 20 M.S. theses, taught several industry short courses on reliability methods, and authored more than 300 technical publications, including two textbooks and 120 peer-reviewed journal articles.

Ram D. Sriram is currently the chief of the Software and Systems Division, Information Technology Laboratory, at the National Institute of Standards and Technology. Before joining the Software and Systems Division, Dr. Sriram was the leader of the Design and Process group in the Manufacturing Systems Integration Division, Manufacturing Engineering Laboratory, where he conducted research on standards for interoperability of computer-aided design systems. He was also the manager of the Sustainable Manufacturing Program. Prior to joining NIST, he was on the engineering faculty (1986-1994) at the Massachusetts Institute of Technology (MIT) and was instrumental in setting up the Intelligent Engineering Systems Laboratory. Dr. Sriram has a B.Tech. from IIT, Madras, India, and an M.S. and a Ph.D. from Carnegie Mellon University, Pittsburgh, USA.

APPENDIX C – The Agenda of the Symposium

December 6, 2016

Time	Event
9:45 – 10:45	Conference Keynote Speech: *Leveraging High Performance Computing to Drive Advanced Manufacturing R&D at the US Department of Energy* Mark Johnson, Advanced Manufacturing Office, U.S. Department of Energy
10:45 – 11:05	*Coffee Break*
11:05 – 11:15	Opening Remarks: Sudarsan Rachuri, DOE
11:15 – 12:00	Symposium Keynote Speech: *An overview of Manufacturing USA Innovation Institutes and Collaboration Network* Dr. Frank W. Gayle, Advanced Manufacturing National Program Office (AMNPO), NIST
12:00 – 12:45	Symposium Keynote Speech: *The GE Brilliant Factory* Dr. Matteo Bellucci, GE Global Research Center, Niskayuna, NY
12:45– 14:00	*Lunch*
14:00 – 16:05	Panel: Big Data Analytics for Advanced Manufacturing: Challenges and Opportunities *Panelists:* Dr. Matteo Bellucci (GE), Dr. Valerie R. Coffman (Xometry), Dr. Sivaramakumar Gopalasundaram (Cognizant), Prof. Soundar Kumara (Penn State), Prof. Sankaran Mahadevan (Vanderbilt University), Dr. Ram Sriram (NIST) *Panel Moderator*: Dr. Sudarsan Rachuri (DOE)
16:05 – 16:25	*Coffee Break*
16:25 – 18:05	Session 1 *(Session Chair: Dr. Ronay Ak)*
16:25 – 16:50	Max Ferguson, Kincho Law, Raunak Bhinge, Yung-Tsun Tina Lee, and Jinkyoo Park, *Evaluation of a PMML-Based GPR Scoring Engine on a Cloud Platform and Microcomputer Board for Smart Manufacturing*
16:50 – 17:15	Dazhong Wu, Connor Jennings, Janis Terpenny, and Soundar Kumara, *Cloud-Based Machine Learning for Predictive Analytics: Tool Wear Prediction in Milling*
17:15 – 17:40	Alexander Brodsky, Mohan Krishnamoorthy, William Bernstein, and M. Omar Nachawati, *A System and Architecture for Reusable Abstractions of Manufacturing Processes*

December 7, 2016

Time	Event
8:45 – 9:45	Conference Keynote Speech: *Harnessing the Data Revolution: A Perspective from the National Science Foundation* Dr. Chaitanya Baru, National Science Foundation
9:45 – 10:45	Conference Keynote Speech: *On the Power of Big Data: Mining Structures from Massive, Unstructured Text Data* Prof. Jiawei Han, University of Illinois at Urbana-Champaign, USA
10:45 – 11:05	*Coffee Break*
11:05 – 11:10	Opening Remarks: Tina Lee, NIST
11:10 – 11:55	Symposium Keynote Speech: *From Sensors to Sensing- Industrial Data Mining at Bosch* Dr. Rumi Ghosh, Robert Bosch LLC
11:55 – 12:45	Session 2 *(Session Chair: Dr. Anantha Narayanan)*
11:55 – 12:20	Srinivasan Radhakrishnan, and Sagar Kamarthi, *Convergence and Divergence in Academic and Industrial Interests on IOT Based Manufacturing*
12:20 – 12:45	Srinivasan Radhakrishnan, and Sagar Kamarthi, *Complexity-Entropy Feature Plane for Gear Fault Detection*
12:45 – 14:00	*Lunch*
14:00 – 15:45	Session 3: Bosch Big Data Challenge *(Session Chair: Dr. Rumi Ghosh)*
14:00 – 14:05	Introduction to Bosch Data Challenge – Dr. Rumi Ghosh
14:05 – 14:30	Bohdan Pavlyshenko, *Machine Learning, Linear and Bayesian Models for Logistic Regression in the Failure Detection Problems*
14:30 – 14:55	Darui Zhang, Bin Xu, and Jasmine Wood, *Predict Failed Product Using Large-scale Data: A Two-stage Approach with Clustering and Supervised Learning*
14:55 – 15:20	Abhinav Maurya, *Bayesian Optimization for Predicting Rare Internal Failures in Manufacturing Processes*
15:20 – 15:45	Ankita Mangal and Nishant Kumar, *Using Big Data to Enhance the Bosch Production Line Performance: A Kaggle Challenge*
15:45 – 16:10	Invited Talk: *Advancing Additive Manufacturing Through Visual Data Science* Dr. Chad Steed, ORNL
16:10 – 16:25	*Coffee Break*
16:25 – 16:40	Closing Remarks
16:40	Closure and Open Discussion