Considerations and Recommendations for Data Availability for Data Analytics for Manufacturing

Don Libes, Seungjun Shin, Jungyub Woo
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, Maryland, 20899 USA
{libes,seungjun.shin,jungyub.woo}@nist.gov

Abstract— **Data analytics is increasingly becoming recognized as a valuable set of tools and techniques for improving performance in the manufacturing enterprise. However, data analytics requires data and a lack of useful and usable data has become an impediment to research in data analytics. In this paper, we describe issues that would help aid data availability including data quality, reliability, efficiency, and formats specific to data analytics in manufacturing. To encourage data availability, we present recommendations and requirements to guide future data contributions. We also describe the need for data for challenge problems in data analytics. A better understanding of these needs, recommendations, and requirements may improve the ability of researchers and other practitioners to improve research and more rapidly deploy data analytics in manufacturing.**

Keywords- big data; challenge problems; data analytics; data quality; requirements; smart manufacturing

I. INTRODUCTION

Data analytics for the purpose of improving performance in the manufacturing enterprise is becoming increasingly attractive for manufacturers. This is motivated by several factors. First, "big data" characteristics (increasing volume, velocity, veracity, variety, and complexity) from models, sensors, etc., enables new approaches (e.g., algorithmic, visualization) that produce novel conclusions. These novel conclusions lead to a second factor -- an increasingly competitive environment. Manufacturers who do not use data analytics will not be competitive with manufacturers who do [1]. Competition can drive prices down, increasing costs for resources and additional regulation provides an additional factor for finding savings wherever possible. Thus, data analytics can be expected to play a critical role in smart manufacturing in the future.

While papers and books describe the promise and some successes of data analytics, few real-world descriptions of the difficulties, challenges, and pitfalls exist [2]. Data analytics is not a cookbook science. Granted, one can find plenty of papers describing success; one can find papers describing methodologies and models; and one can find papers comparing algorithms. However, in our literature search, we could not find papers specific enough to guide anyone through the complete process of deploying data analytics. Each enterprise traditionally makes its own decisions about what data are important and how these data should be stored and analyzed.

Complexities exist for many reasons such as uniqueness and diversity of manufacturing enterprises and goals, lack of standards or in some cases, multiple standards without compatibility. Work on these areas is hindered by a lack of widely-recognized challenge problems in data analytics for manufacturing applications. These problems must be meaningful and find wide acceptance by researchers across many fields and industries such that progress is possible in reasonable timeframes. Underpinning all of these is a corresponding need for real data. By its very essence, data analytics cannot function without data. For the same reason, research into data analytics also requires data.

The National Institute of Standards and Technology (NIST) is researching data analytics to support U.S. manufacturing. However, NIST is not a manufacturer and has no direct access to the amounts and types of big data that forms the basis for most of our research into data analytics [3][4]. NIST has therefore had to rely on industrial partners for data. Unfortunately, these data are often problematic in a number of ways which in turn impedes the research. In our own work, we have faced significant issues dealing with contributed data and had to expend significant time resolving these issues due to the lack of relevant guidelines for data acquisition, types of data, and the volume of data. This is the motivation and source for much of this paper.

This paper discusses the complexities of producing usable manufacturing data. The paper also provides recommendations on what can be done to 1) produce more data that will help researchers (such as NIST and academia) of data analytics in smart manufacturing, and 2) help manufacturers that want to supply their data to researchers. We describe the types and qualities of data needed. We describe other attributes of significance. We also describe the reliance that many researchers have on industry data, why this reliance presents difficulties, and suggestions on how these difficulties can be ameliorated. At the very least, we need to understand how critical data is to our current state of the art of data analytics. In that sense, we believe that our thoughts, observations, and recommendations in this paper may be helpful as guidance and toward stimulating further discussion.

While our own interest stems from manufacturing issues, that is not intended to limit applicability of this paper.

Readers may find that much of our discussion is widely applicable to other domains.

II. Data Availability Issues and Recommendations

Deciding what data to collect, how to collect, and how much to collect is a significant challenge. Such decisions include the source of data and their type, location, and frequency. Each of these may have additional complexities. For example, locations may change manually or automatically. Sensors may be mobile - for example, on-transport devices or in-process parts. It may be of value to measure all parts or just a percentage and frequency may be regularly periodic or event-driven. Each of these issues has further ramifications. For example, using data from multiple sources can be difficult when the sources are unsynchronized or use different output formats. This "data fusion" cuts across many of the issues we describe.

This section of the paper describes issues and potential recommendations that we are considering.

A. Data Use Restrictions

We recognize that there may be restrictions on the way certain data are used. While this may not be true in an experimental test environment, at a certain point, we want to reflect real world concerns. Restrictions arise in several ways that are further described here.

1) Proprietary Data

Significant amounts of data exist but are inaccessible to researchers as the data are proprietary. Many companies collect such data but only make it available internally or to a limited set of partners with confidentiality agreements.

The problem of proprietary data includes both generated (for example, from sensors) and intrinsic data. For example, the manufacturer of a milling machine may decline to provide data describing speed and torque relationships. Possibly, the manufacturer may not even have the kind of comprehensive data that researchers want. Manufacturers may have models and other descriptions of their products to the extent needed for their own manufacturing purposes but may be unwilling to release the data.

The unwillingness, for example, of machine tool manufacturers to share data is understandable. Models and performance data can be used by competitors to improve their own manufacturing processes and products. While the very largest buyers of such equipment may have the leverage to access (or can afford to pay for) proprietary information, it is inaccessible to the majority - especially to mid- and small-size manufacturers.

Without this proprietary information, researchers are left to make sample parts and carry out onerous experiments and measurements, all for limited purposes. With limited resources, many assumptions must be made since these experiments involve physical operations rather than virtual simulation. For this reason, we are left with researchers proposing equipment and other manufacturing decisions that are far from optimal despite the possibility of significant improvements with access to proprietary data.

Although we could recommend that equipment manufacturers make as much of their data as possible available to enable and encourage researchers, this statement is too vague. It may be more useful for them to, for example, make available complete data for a machine that is no longer sold but nonetheless widely used. Similarly, manufacturers can release data for manufactured products or process plans that are no longer used or critically proprietary.

2) Data Security

To thwart corporate espionage or even inadvertent data leaks, data at all levels must be protected. Traditionally, any user or application behind a firewall can access any and all data. For enhanced security, it is becoming common to compartmentalize access on a need-to-know basis [5]. This not only requires justification and requests, but also time limits. This can be problematic since need-to-know and on-demand philosophies conflict with the ability of algorithms that expect open access to all data. In part, this is because the algorithms themselves may not know in advance what data they need to explore.

This problem is exacerbated by the use and number of suppliers, shippers, and other external parties in the supply chain. These parties will likely not allow access to their internal data except in extremely limited ways. This significantly limits the types of, and the extent of both data discovery and analytics that can take place. In short, the traditional benefits of using third parties to gain efficiency and cost advantages are offset by these data-access issues and the less-than-optimal decisions that result.

Infrastructure deemed critical refers to elements crucial to such issues as national economic security, economic viability, and public safety. These may include both long-term and short-term issues that arise from deliberate attacks or disruptions from natural events [6]. One important strategy to provide security is to reduce risk by minimizing access. From a security perspective, the more that access is restricted, the more likely it will achieve the goals of reducing risk and defending infrastructure. Restrictive access thwarts data analytics where the more expansive and ready access to data, the more likely it will be possible to detect previously unknown patterns and conclusions from massive quantities of seemingly unrelated data.

Encryption may be used to shield data from inappropriate consumers. This introduces drawbacks such as extra processing which could be difficult at low levels with limited processing power. Encrypted data also generally take more space both during transmission and when resident. Finally, encryption provides yet another opportunity for data leakage as encryption keys provide another attack surface [7]. However, encryption may be useful when sharing data in a scenario that does not require real-time access. Many encryption variants exist. For example, homomorphic encryption allows computations directly on encrypted data. Decryption can be done at a later time in a secure environment and the result will match the results that would have occurred had the data analytics been done on the unencrypted data [8].

Data producers may find it useful to digitally sign data as proof of provenance and that the data have not been changed.

This latter can be useful as researchers want to be able to replicate results of others or want to assert improved results on the same body of data. Digital signatures can also be used with public-key encryption so that only partners with corresponding private keys can read the data.

Security-related recommendations for improving access to data do not generalize widely. We can only raise awareness of the tradeoffs.

B. Non-Existent or Missing Data

In some areas, data may not exist for a variety of reasons. Examples: Data may be too expensive to capture – installing sensors or buying models may cost more than the expected benefit. It may be impossible to capture data that are internal to a sealed system. A KPI (Key Performance Indicator) may be unavailable if it is too new or vaguely defined. Sensors, communications, or a database may have failed in a run that is too expensive to repeat.

Such non-existence of data, while frustrating, occurs in real scenarios. Sensors don't always work. Databases don't always record. For a variety of other reasons, it may be impossible to obtain data.

Workarounds include the use of simulated data, reference data, and reasoning techniques [9][10][11]. Simulated data are generated by models and algorithms rather than measured in situ. Reference data are any data acquired second-hand. Reference data can derive from a wide-variety of sources (design goals, for instance) as well as measured. However, while reference data are ideal in theory, it is rare to find different manufacturers who have identical independent variables in their manufacturing systems except in very simple processing. Even manufacturing operations of commodity items generally vary widely among manufacturers. Lastly, reasoning techniques such as Bayesian networks may provide suitable workarounds to address missing data.

C. Data Timing and Synchronization

Synchronization of data is a significant problem in several ways. An enterprise typically has many producers of information. To make use of such information from multiple sources, it is necessary to understand when that information was created. This is more difficult than simply timestamping data from a central time server.

Consider a sensor on a machine tool that is reporting force data. It is necessary to understand how this correlates with machining operations. Operations can be viewed as a hierarchy with a process plan at a high level and machine instructions describing speed and position at a low level.

There are two significant issues: synchrony of sensor timing, and synchrony of processing steps.

First, timing chips, while inexpensive, are not a part of many sensors [12]. The resulting data are unsynchronized or require retroactive synchronization to deal with relative timestamps. Absolute timing adds cost. While central-time service distribution is achievable, adding this functionality to all sensors increases costs. This is especially problematic because there will typically be hundreds of sensors in a work cell, and many thousands throughout a factory. Relative synchronization is cheaper but is non-trivial. A priori synchronization is not trivial nor is adding timestamps at a later stage (for example by an MTConnect-aware server that sends the results back to the database), which generally adds latency resulting in inconsistent time lags.

Second, processing steps are typically hidden inside the processing elements of a machine tool, and can be inaccessible to the external world. This makes it difficult to correlate these processing steps to sensor measurements. Even if these steps are known (for example from experimentation), there can be indefinite latency since operations and algorithms inside commercial controllers are proprietary and not intended to be examined nor guaranteed. This latency introduces mismatches between low-level data, making certain types of data analytics impossible or introducing a significant amount of uncertainty.

Most of what is available today are unsynchronized sensor data. We believe that many sensor manufacturers or integrators simply don't view this type of synchronization as a requirement. Hence it remains a challenge for researchers.

Recommending that all sensors and machine tool internals include absolute timing circuitry is simple but unrealistic. Instead, we recommend that such information be tagged at the lowest level possible with absolute timing information with millisecond precision or, if possible, sub-millisecond. Timestamps must be in a format that is directly machine processable such as ISO (International Organization for Standardization) 8601 [13]. Timings that are known or likely to be affected by lag should be tagged as such. This should include a measurement or estimate and probability of the lag.

D. Data Frequency

During manufacturing, there is a tradeoff over the frequency with which data are collected. While it is intuitive that additional data can lead to detection of additional patterns, increased accuracy, and the potential for otherwise undiscoverable insights, additional data also have costs.

Increased data means increased cost of processing time, data storage, access, network traffic, and database contention. Synchrony (see above) and the cost and accuracy of interpolation are also a concern as is algorithm performance. While an increase in data can lead to better results, algorithms will invariably take more time. Data mining algorithms already use heuristics to trade off exponential performance for suboptimal results so unnecessary data only leads to further deteriorating performance. This is especially problematic when it impedes the ability to draw conclusions and to make adjustments to processes in real-time.

One argument for collecting more data is that it is simpler. No knowledge is needed on how to pass judgment on data. So immediately after having taken and handed off a measurement, the sensor begins the cycle again with no waiting for any kind of clock cycle. Of course throttling may be induced indirectly. For example, the network may be incapable of absorbing data at the rate at which the sensor sends it. The database may not be fast enough. The architecture stack may induce throttling at many different

levels. Whether this throttling occurs directly (planned throttling) or indirectly is a significant concern. For example, if a database is thrashing due to the amount of data it receives or from demands by the analytics algorithms, gaps that appear between batches in the data may be large, causing poor performance of algorithms - as if less data were collected than actually was [14]. Understanding and managing these tradeoffs is beset by subtle choices and risks.

During research or pre-production, the problem of excess data is ameliorated to an extent. For example, real-time performance may be unnecessary. However, badly performing algorithms are still at risk with even minor increases in data size. And data mining by its very nature involves exploring data for unexpected results. Therefore, more data are better. However, not all data are of equal value. Being able to "help" algorithms by withholding data is a challenge. A related issue is that some algorithms perform worse with additional data. For example, for certain machine-learning algorithms, increasing the training set size can produce outputs that are inferior to those of smaller training set sizes. While this phenomenon is data dependent, it is a significant concern since this runs counter to expected outcomes that suggest more data are better [15]. This arises with other algorithms as well such as overfitting in neural network training. And, in practical terms, the world is moving to essentially unlimited data so this question of data size arises frequently.

The goal of replacing unlimited big data with limited ideal data is often called the "smart data" problem. However, it remains an open problem [16]. For now, we are stuck requiring, generating, storing, and analyzing "big data."

To assist with the problem of frequency, we recommend that data publishers produce two sets of data in situations where frequency is an issue. One set should be a complete set with all measurements. A second set should be limited to a more realistically representative amount that would normally be available in production. Additional sets may be provided as supplements but should be marked appropriately - for example, indicating extra sensors are used that would not normally be available.

E. Data Formats, Standards, and Specifications

Enterprises sit atop a vast collection of disparate data, often produced by a multitude of heterogeneous sensors, and often ultimately stored in files formatted according to a variety of standards, with varying degrees of compliance. Significant amounts of data may follow no standards whatsoever. Using standard specifications such as XML (eXtensible Markup Language) and JSON (JavaScript Object Notation) can help [17]. However, problems such as underspecification can remain that leave ambiguities. For example, using an XML attribute called "time" means little if there is no definition for how the string is to be interpreted -- absolute with respect to UTC (Coordinated Universal Time) or local time zone? Relative to what -- start of a process or something else?

While standards can be helpful, they are not panaceas. For example, equipment and software from different vendors may use different standards. Data, while still standard-compliant, can lose fidelity during interchange. Standards frequently have different levels of compliance that users may choose from. Even highly specific standards do not guarantee data that are usable. For example, the machine tool standard, MTConnect, covers only one direction of communication; so, correlation to commands may not be present or need reconstruction with timing uncertainties. Equally important, MTConnect does not cover all possible types of data that a machine tool can generate [18]. For example, MTConnect defines a fixed set of statistics for DataItems. Kurtosis, a measure of peakedness relative to a normal distribution, is in the set. Skewness, a measure of symmetry, is not in the set.

All of these choices have reasons for existence. For example, different vendors may have different reasons for their choices. These can include historical issues, expense in tracking developing standards, and interactions with other software. For example, the choice of OWL (Web Ontology Language) variants depends on how much need there is for expressiveness – the breadth of concepts that can be represented. But greater expressiveness brings with it a loss of computational guarantees [19].

Data standards may be descriptive (describing practices) or prescriptive (defining practices). Each carries with it downsides. For example, descriptive standards may prevent the use of innovative techniques that are too new to be incorporated in standards while prescriptive standards may be ignored when better technology solutions are discovered. These dilemmas are particularly apparent in rapidly changing and highly-competitive fields. To allow variations and technological advances, some standards intentionally leave areas of ambiguity with a resulting ambiguity in the data.

Some standards and specifications have been created to help with the issues mentioned in this section, such as PMML (Predictive Model Markup Language) and PFA (Portable Format for Analytics) [20][21]. For example, PMML and PFA can specify time offsets or transformations such as normalization.

We recommend that well-recognized standards be used at every level of data formatting and description. Metadata should be supplied to describe data formats and meaning. Proprietary and ad hoc standards should be avoided except when necessary. Standards and specifications such as PMML and PFA should be used instead of ad hoc programming language solutions to deal with likely conversion issues.

F. Data Uncertainty and Reliability

Enterprise data can have all the characteristics common to any collection of big data. These include uncertainty, reliability, and accuracy, among others. It is important that data collections address these types of characteristics in a realistic way.

Some of these characteristics should be dealt with by providing the appropriate metadata (data describing data). For example, uncertainty that is quantified should be expressed in metadata included with the data. In some cases, however, uncertainty may be unknown a priori and it is the task of the data analyst to deduce the uncertainty during the

analytics process. As in the former case, the latter possibility can and should be expressed in the metadata.

In the general sense, problematic data presents several choices. It is tempting to cull data before publishing. It may be similarly tempting to artificially induce problematic data whether programmatically or physically – such as by using an intentionally faulty or broken sensor. No matter which of these approaches are taken, metadata must indicate what is known about the data.

When there has been modification of data (removal, normalization, etc.), we recommend that multiple data sets be made available corresponding to the original and the modified data.

G. Data Access, Storage, and Processing

Data access and availability impact data analytics. In even simple cases, data is structurally changed when it is distributed to researchers. For example, distribution traditionally has meant serialization. However, this raises issues of how de-serialized data are intended to be stored and accessed. Alternatively, researchers might have direct access to the database of a producer, perhaps because the producer is generating information in real-time or, more commonly, because the data set size is extremely large and researchers only need access to a relatively small portion.

Our estimates of a small manufacturing shop scenario with a modest number of machine tools using Hadoop/Hive technology could generate daily data of 1 TB (25 MB/sec) based on 500 K/day sensor readings at 50 KB/reading with database overhead, data normalization, and other issues. With replication, compression, and caching for 250 workdays/year, this approaches 1 PB required to retain a year's worth of data. Our estimates may soon be seen as conservative given that the cost to store data continues to decrease and the ability to generate data continues to increase.

Data may be converted to an entirely different database type by, in essence, changing the solution. Data storage choices are significant since different database technologies have different strengths and weaknesses. For example, highly-regular relational data are optimally stored in a relational database. In contrast, data that are predominantly hierarchical such as geospatial (e.g., factory floor map) incur a significant extra expense when stored relationally. Another example is that a fixed data schema can significantly improve the performance of data mining by reducing the time to access data. In contrast, more agile databases have the flexibility to store new data types but only by increasing the expense of access.

A related problem is that some enterprises have limited database expertise with the result that inferior database solutions are employed. For example, data may not be normalized so data are redundant, contain inconsistencies, lack connections, etc. While this might not be a problem for ad hoc in-house use, distributing such databases causes significant and unnecessary extra effort and frustration.

More sophisticated enterprises will use multiple databases. Different machines, different caching, different bandwidths all affect availability. For example, joins within a database are generally easy; joins across databases on multiple machine often incur significant expenses. However, multiple databases are a more realistic portrayal of enterprise data since they can simplify the problems of rapidly changing applications and data analytics demands. This contrasts with traditional data distribution using a single file, however large.

We mentioned earlier that databases might process data before storage. As processing elements become increasingly powerful, it is common to find databases at the lower levels of the enterprise with sophisticated operations and storage capacities. Such low-level databases can be used to address certain problems such as avoiding network overload by only sending up significant results rather than raw data. We have hypothesized that stored programs may also be used to solve security issues noted earlier (see Proprietary Data) in that processing takes place inside of a controlled proprietary environment so that results are returned without exposing raw data [21]. However, we have not seen any evidence of this in practice and remain skeptical of it as a workable technique. Nonetheless, the use of low-level processing in general can have the inadvertent effect of making data availability less likely.

Multiple researchers who maintain the same data in different ways may be solving different problems. Therefore, we recommend that data sets identify how they are nominally intended to be accessed and, if necessary, stored. Researchers who opt to use different techniques should state those differences when releasing results so it is clearer when comparisons are being made using different access and storage technology as well as other aspects of the research.

H. Data Preparation Time and Other Resources

Assuming a manufacturer is supportive of releasing data, the time and effort required can be substantial. The potential for unaffordability can be a reason to prevent the release of data.

We have already mentioned many categories of issues. Each of these takes time. For example, a manufacturer must take the time to decide what is proprietary and what is not and how to separate them. Decisions may change due to multiple influences. For example, sensor measurements that may initially seem distributable may "give away" too much information if taken at a much higher frequency or with greater accuracy.

Having engineers devote limited time to selecting database fields, formatting and packaging data, finding a suitable repository, keeping it updated, and creating data descriptions, can be difficult to justify if it does not immediately feed back to the corporate bottom line. Similarly, legal requirements and concerns may force the use of additional resources, for example, for non-disclosure agreements that may vary across consumers and data sets.

We recommend that, before embarking on data distribution projects, manufacturers create a budget that incorporates all aspects of time and resources that will be needed. This not only protects a company from unwelcome surprises but is more likely to lead to a careful and useful release of data that can be sustained in the future.

III. Categories of Data

A. Application Domains

It is important to understand the categories of data that exist in an enterprise so that data may be properly acquired and used for the purpose of data analytics. We are in the process of enumerating a comprehensive list of application domains and needs of manufacturing. This section describes a few examples to show the general idea.

1) Supply-Chain Management Data

Supply-chain data include product types, scheduling, risks, cost accounting, efficiency, sustainability, and other factors. Many of these may include choices, for example between multiple suppliers or products and tradeoffs. Single-supplier, single-product data sets are useful but only for very limited purposes.

Access to these kinds of data would allow investigations into problems such as maximizing responsiveness to change, minimizing inventory and energy use, and balancing cost vs critical dependencies.

2) Production Scheduling Data

Production scheduling data include performance criteria, product data, work cell models, and resource descriptions. Data should include a variety of generated schedules under stated conditions (for example, fixed maximum total time or minimizing total time needed) and computed outputs (such as energy used, time expended, and other KPIs).

Access to these kinds of data would allow investigations into production efficiency, load leveling, dynamic rescheduling, and trade-offs such as inventory reduction vs total production throughput.

3) Process Planning Data

Process planning data include product data, resource descriptions, and fabrication techniques. Data should include a variety of process plans under stated conditions (for example, fixed maximum total time or minimizing total time needed) and output (such as energy used, time expended, and other KPIs).

Access to these kinds of data would allow investigations into work cell efficiency and trade-offs such as energy vs manufacturing time vs throughput.

4) Machine Tool Data

Machine tool data include tool models, energy, reliability, and tool wear. Data should describe a variety of raw materials, processes, tools, and outputs such as energy use, time expended and other KPIs).

Access to these kinds of data would allow investigations into comparisons of different processes, machines, and tools, as well as experiments trading off various factors such as material cost, cutting speed, tool wear, and mean time to failure.

B. Types of Data

Separately from the preceding domain categories, there are different types of data characterized, not by the application domain, but by their type. Each has significant concerns.

1) Key Performance Indicators

KPIs are a type of metric designed to guide decision-making, for sustainability or profitability for example. However, use of common KPIs can still not be taken for granted. For example, in a survey of manufacturers, 36.8 % of respondents did not use KPIs [23]. Reasons included "Unsure of what data to measure" (20.7 %) and "Not convinced that measuring KPIs adds value" (24.1 %). Even KPIs that intuitively sound useful may not be regarded so by many people. For example, less than half of respondents who use KPIs in the Schenck survey agreed that it was worth such straightforward metrics for equipment utilization (42.6 %) and downtime (40.4 %).

ISO 22400 defines standard key performance indicators for manufacturing operations management [24]. While we do not discourage collection and use of other KPIs and the merits of existing and new KPIs are a regular source of debate, we encourage the use of standard methods for KPIs collection where they exist.

2) Human Interaction

While minimizing manual activity is desirable due to tradeoffs of cost, consistency, and reliability, some human activity will remain. Unfortunately, each type of data collected about human activity comes with different and potentially expensive tradeoffs. For example, keyboard interaction is easily and inexpensively tracked, even to individual keystroke values and timing. In contrast, observing what a human is looking at, while trackable, is expensive. Eye movements can be recorded, screen videos can be saved, and the two can be synchronized. Other types of interactions are even more expensive ranging from short-term indicators (e.g., muscle effort, joint stress) to long-term indicators (e.g., endurance, concentration).

The justification of collecting such data may seem difficult for the very reason that such interactions are necessary in the first place. Namely, the interaction cannot be inexpensively and reliably automated in the first place. Whether it is human judgment, human vision, or some other type of interaction that is seemingly simple for humans but hard for machines, these types of issues make these data among the least easy to deal with.

At the same time, data from observing human interaction is intriguing. While high-level judgments may remain difficult, they have a significant potential to lead to more efficient human activity in the workplace. However, such observations are remarkably easy to overlook and likely to be misunderstood without data.

3) Data Models

Much of our focus on data collection has been on low-level data collected directly from sensors. However, understanding these data can be difficult or impossible without the underlying models on which the data are based.

Models are descriptions of designs, activities, and algorithms. They exist at every level of an enterprise – from supply chain to machine tool. Such models include a variety of information including requirements, architecture descriptions, interaction diagrams, and physical specifications. Use of models have several problems:

- Models may contain proprietary information which is stripped before distribution.

- Models may require integration with other models or it may be unclear how they relate to the data.
- Models may be more aspirational than accurate. Implementers may have deviated from the models.

4) High-level Data and Metadata

We have already mentioned metadata in several contexts earlier in the paper such as describing the format, timing, and source of data. However, metadata is a more general issue that deserves additional elaboration. As an example, many axioms or assumptions are metadata that should be documented.

Consider an investigation into energy use. While energy use may be readily available from power meters and fully recorded, hidden assumptions may not be. The price of energy may already have been considered when the enterprise was designed. A high-cost energy location will naturally have minimized the use of energy while a low-cost energy location will lead to data that is biased toward profligate energy use. Such assumptions may not make sense for researchers, depending on the focus of the research. As another example, consider a factory in a location where energy cost is time-sensitive. Time-sensitive energy pricing may come about from the power company or in-house sources such as solar power availability, which is only available during the day. A failure to record this metadata will lead to energy-use data that may look quite odd.

Other metadata omissions can be equally problematic. For example, consider a machine tool shared across production lines or one that is older and needs more frequent maintenance. These types of factors will influence the data. However, a lack of explanation will leave inexplicable impacts in the research conclusions. Similarly, data may reflect expenses related to machine tools due to requirements in the product line that are not evident in the data. Schedules (i.e., absences) related to trained personnel availability may also perturb data in similar ways.

5) Simulated Data

Some projects have addressed a lack of physical data by creating simulated sensor data complete with comparisons of simulated data to physical data. This is not necessarily easy since real-world artifacts will include errors for many reasons such as calibration drift or pseudo-random errors. These must be modeled (to some degree) for useful data. In some cases, simulated data may be a useful substitute. In others, simulated data may not be useful. Therefore, researchers must take care to appreciate the differences in requirements.

We believe there is a valid use for simulated data and encourage development in data analytics. However, such data must include appropriate metadata that explains its simulated nature and source. This may include assumptions, requirements, and source code for the simulated data generators.

IV. Data Repositories

There is strong motivation to create "big data" repositories to hold contributions from industry as well as academia. Two such repositories exist already.

- CO2PE! (Cooperative Effort on Process Emissions in Manufacturing) is an initiative with a number of objectives related to improving manufacturing processes. CO2PE! proposes to "develop a methodology that allows to provide data in a format useful for inclusion in LCI (Life Cycle Inventory) databases." [25] CO2PE! also expects that its partners will contribute to LCI data "as required for systematic LCA studies, covering the production stage..." While more specialized to specific fields and problems, the focus of CO2PE! is a strong indicator supporting the focus of our own work described in this paper.
- ecoInvent is an LCI database representing "human activity and its exchanges with the environment and with other human activities." [26] These exchanges include energy and other resources. These datasets represent the higher levels of data of an enterprise and go all the way to national tracking of resources. Some of the factors we described earlier are addressed by ecoInvent such as formats and uncertainty. ecoInvent also addresses other factors such as revision control.

Since these are specialized repositories, it is likely that there will be additional repositories created. We would like to see repositories that have the following attributes:

- Automated. We anticipate contributions will occur by connecting to a website and uploading data along with metadata that describes scenario information. In the case of direct access, information must also be provided along with certain best-effort promises of availability and any bandwidth restrictions (such as "no more than 1 GB/day without prior approval"). Researcher access can occur by manual or automatic download using Webservices.
- Secure. Digital signatures must be used to ensure authorship and to ensure changes are trackable.
- "Lightly" curated. Given the ever-increasing span and complexity of scenarios and rapidly evolving technology, it may be difficult to judge the value of most submissions. Submissions that attract little interest initially may turn out to be of significant value at a later time. Except for the space consumed, submissions that are of no interest have no impact on others.
- Generated data. While we seek contributions of real data, simulated data are also useful. Ideally, simulated data will include sources to the data generator with generation parameters so that distribution size can be minimized and further experimentation is possible while allowing research conclusions to cite the parameters for the particular generation run.

One particularly desirable type of repository is for challenge-problem data. Challenge problems are problems widely recognized as fundamental or critical to advancing research in their fields. In the field of data analytics in smart manufacturing, there are many issues associated with these

challenge problems including their description, availability, and relevance. Most importantly, challenge problems require data.

Challenge-problem data serve two purposes. First, such data are necessary for researchers who would otherwise have access to insufficient real data. Not all researchers have direct access to machine tools or work cells, for example. Second, having data common to all researchers addressing these problems allows their results to be evaluated and compared more easily. Without common data, trying to compare results is akin to comparing apples and oranges. It becomes impossible to do meaningful comparisons.

V. Conclusion

In this paper, we have described issues that presently hinder access to and usability of data. We have made some recommendations and observations with the intent of spurring more data contributions and providing guidance to make them more useful. In our own data analytics research using real manufacturing data, we have encountered many of these issues. Sometimes significant time is required for resolving problems. In the worst case, these data may not be usable at all. We hope that our experiences allow researchers to avoid these problems in the future.

We have also described the need for data for challenge problems. A better understanding of these needs, recommendations, and requirements may improve the ability of researchers and other practitioners to better study and more rapidly deploy data analytics in manufacturing. Lastly, we have described existing repositories and suggestions for future "big data" repositories for data analytics. We believe that all of these ideas will improve the opportunities for data analytics researchers and contributors.

VI. Acknowledgments

Thanks to Sudarsan Rachuri, KC Morris, Al Jones, Sharon Kemmerer, Sandy Ressler, Guodong Shao, and several anonymous reviewers for their insightful critiques, suggestions, and fruitful discussions. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] L. Columbus, "84 % of Enterprises See Big Data Analytics Changing Their Industries' Competitive Landscapes In The Next Year," Forbes, Oct 19, 2014.

[2] D. Libes, T. Lee, "Challenges of Data Analytics in Smart Manufacturing Systems," in draft.

[3] Tata Consultancy Services, "Manufacturing: Big Data Benefits and Challenges," http://sites.tcs.com/big-data-study/manufacturing-big-data-benefits-challenges, 2013.

[4] Tata Consultancy Services, "Emerging Big Returns on Big Data," http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf, 2013.

[5] K. Stouffer et al., "Guide to Industrial Control Systems (ICS) Security," NIST Special Publication 800-82, National Institute of Standards and Technology, Gaithersburg, Maryland, 2015.

[6] U.S. Dept of Homeland Security, National Infrastructure Protection Plan (NIPP) 2013: Partnering for Critical Infrastructure Security and Resilience, 2013.

[7] P. Manadhata, An Attack Surface Metric, CMU-CS-08-152, Carnegie Mellon University, Nov 2008.

[8] C. Gentry, "A Fully Homomorphic Encryption Scheme," https://crypto.stanford.edu/craig/craig-thesis.pdf, Stanford University, September 2009.

[9] R. Little, Statistical Analysis with Missing Data, ISBN 978-047118386, Wiley, 2002.

[10] M. Karthika, J. Pearl, "On the Testability of Models with Missing Data," Proceedings of AISTAT-2014, 2014.

[11] K. Mohan, G. Van den Brock, A. Choi, J. Pearl, "An Efficient Method for Bayesian Network Parameter Learning from Incomplete Data," Causal Modeling and Machine Learning Workshop, ICML-2014, 2014.

[12] C. Grimes, E. Dickey, and M. Pishko, Encyclopedia of Sensors, American Scientific Publishers, ISBN 1-58883-056-X, 2006.

[13] ISO, "Data Elements and Interchange Formats - Information Interchange - Representation of Dates and Times," ISO 8601:2004, http://www.iso.org/iso/catalogue_detail?csnumber=40874, 2004.

[14] P. Denning, "Thrashing: Its Causes and Prevention," Proceedings AFIPS, Fall Joint Conference 33: 915-922, 1968.

[15] W. Yousef, S. Kundu., "Learning Algorithms May Perform Worse with Increasing Training Set Size: Algorithm–Data Incompatibility," Computational Statistics & Data Analysis, Elsevier, vol. 74(C), pages 181-197, 2014.

[16] B. Rossi, "Is Big Data Dead? The Rise of Smart Data," Information Age, http://www.information-age.com/technology/information-management/123458486/big-data-dead-rise-smart-data, Sep 23, 2014.

[17] JSON.org, "JSON: The Fat-Free Alternative to XML," http://www.json.org/xml.html, undated.

[18] P. Warndorf, "MTConnect Institute Releases Version 1.3.0 of the MTConnect Standard," http://mtconnect.org/recent-news/mtconnect®-institute-releases-version-130-of-the-mtconnect-standard.aspx, 2014.

[19] World Wide Web Consortium, OWL Web Ontology Language Reference, http://www.w3.org/TR/owl-ref, 2004.

[20] A. Guazzelli, "Predictive Analytics, Big Data, Hadoop, PMML," http://www.predictive-analytics.info/2009/04/pmml-data-pre-processing-primer.html, 2009.

[21] J. Pivarski, "PFA: Portable Format for Analytics," http://scoringengine.org, 2015.

[22] Y. Bernier, "Latency compensating methods in client/server in-game protocol design and optimization," Valve, Kirkland, WA, http://web.cs.wpi.edu/~claypool/courses/4513-B03/papers/games/bernier.pdf, July 7, 2015.

[23] S. Schenk, "Using Key Performance Indicators," http://www.schencksc.com/2013mfgsurvey, 2014.

[24] ISO, "Automation Systems and Integration - Key Performance Indicators (KPIs) for Manufacturing Operations Management - Part 2: Definitions and Descriptions," ISO 22400-2:2014, http://www.iso.org/iso/catalogue_detail.htm?csnumber=54497, Jan 1994.

[25] Group T, CO2PE! (Cooperative Effort on Process Emissions in Manufacturing) Home, http://www.co2pe.org, International University College Leuven, 2015.

[26] B. Weidema et al, "Overview and methodology," ecoInvent Centre, http://www.ecoinvent.org/files/dataqualityguideline_ecoinvent_3_20130506.pdf, Swiss Centre for Life Cycle Inventories, 2013.