Qualifying Evaluations from Human Operators: Integrating Sensor Data with Natural Language Logs

Michael P. Brundage[1], Michael Sharp[1], and Radu Pavel[2]

[1] *National Institute of Standards and Technology, Gaithersburg, MD, 20899, United States*
michael.brundage@nist.gov
michael.sharp@nist.gov

[2] *TechSolve Inc., Cincinnati, OH, 45237 United States*
pavel@TechSolve.org

ABSTRACT

Even in the increasingly connected world of smart manufacturing and the Industrial Internet of Things (IIoT), there will always be a need for human operators and evaluations. When creating equipment condition monitoring models and heuristics, the observations from human operators are often difficult to quantify or track. This situation can lead to the observations being underutilized, misunderstood, or ignored completely if autonomous sensors are employed. This work seeks to highlight the untapped potential for augmenting numeric data from sensors and control systems with human input and vice versa, by integrating documented natural language reports with data collection technology in a novel and intuitive way. It is a first-step experiment and seeks to establish a link between human-generated data and sensor-driven information to motivate, justify, and guide future endeavors. This is an exploratory work that utilizes an experimental setup with a limited and controlled accelerated aging setup where human observations were recorded at regular intervals alongside streaming sensor data. The goal is to validate the relationship between observers' natural language, quantified sensed values, and some ground truth knowledge about the state of the tool. We provide recommendations for follow-on work and extensions of the performed analysis as part of a next steps outline.

1. INTRODUCTION

Currently, many companies, industries, and organizations take advantage of smart manufacturing concepts to design, define, and promote the next generation of digital manufacturing and enterprise capabilities. Large companies are already at the forefront of the development and deployment of digital technologies that enable connectivity and automation (Jin, Weiss, Siegel, & Lee, 2016). However, small- and medium-sized manufacturers (SMMs) lack the resources that big Original Equipment Manufacturers (OEMs) have to research and implement these new concepts (Software, 2018). Without a proper understanding of how emerging technologies and the integration of these technologies can make them more competitive, small manufacturers tend to delay the change of their traditional ways to new, digitally-integrated strategies.

The changes from the digital revolution in manufacturing are profound and pose a real challenge, but also an opportunity to manufacturers of all sizes. To avoid being left behind, companies need to be proactive and develop strategies to exploit the opportunities of digitalization, improve existing processes, and develop new business models.

One of the most rapidly growing trends associated with digital manufacturing is the use of monitoring and data collection systems. These provide visibility and actionable information with respect to machine utilization, capacity, and overall equipment effectiveness. This, in turn, can inform condition-based or predictive maintenance. An increasing number of OEMs leverage technologies such as these to assess the current and future states of equipment, machine tools, manufacturing cells, supporting subsystems, and even manufacturing processes (Software, 2018).

2. BACKGROUND AND MOTIVATION

Researchers have spent considerable time crafting methods to collect and analyze data coming from industrial equipment (Kunche, Chen, & Pecht, 2012; Li, Verhagen, & Curran, 2018). The majority of the research focus is on the analysis of sensor data to improve maintenance operations. Multiple publications provide techniques and results based on numeric data analysis. Kothamasu, Huang, and VerDuin (2006) reviewed the strategies and techniques of monitoring and predicting machine health that focuses on improving reliability and reducing unscheduled downtime. Djurdjanovic, Lee, and Ni (2003) introduced a toolbox of data-driven algorithms and presented applications in mechanical systems prognostics. Katipamula and Brambley (2005) completed a representative review of the research and practices of fault detection, diagnostics, and prognostics for building systems. Lu, Li, Wu, and Yang (2009) summarized wind turbine condition monitoring and fault diagnosis activities. Venkatasubramanian, Rengaswamy, Yin, and Kavuri (2003) presented a review of quantitative model-based methods for chemical process fault detection and diagnosis. More recent methods leveraging machine learning (ML) and artificial intelligence (AI) techniques continue to heavily rely on numeric data collected from sensors, PLCs, and machine controls.

Analyzing human-generated text from the shop floor offers another promising avenue to improve operations. This area of research is called Technical Language Processing (TLP) and centers around using Natural Language Processing (NLP) methods on technical text data (Brundage, Sexton, Hodkiewicz, Dima, & Lukens, 2021). In particular, using TLP on Maintenance Work Orders (MWOs) has been an area of budding research (Brundage et al., 2021; Ho, 2015; Lukens, Naik, Saetia, & Hu, 2019). Information within MWOs provides a health history of an asset that is rich with quantitative and tacit knowledge within the text. Previously, researchers have analyzed this information to capture key information about maintenance operations (Ho, 2015; Lukens et al., 2019; Sexton, Brundage, Hoffman, & Morris, 2017). Multiple efforts successfully calculated the Mean Time Between Failure (MTBF) by only using MWOs (Ho, 2015; Sexton et al., 2017). Other works have created maintenance Key Performance Indicators (KPIs) from MWOs to help analysts understand the performance of their maintenance operations (Brundage, Morris, Sexton, Moccozet, & Hoffman, 2018).

An underexplored area of research is the merging of these two major data streams: (1) data coming directly from the equipment and (2) the human-generated text data. The human-generated test data (i.e. natural language) can add significant value to the maintenance analysis since observations by humans on the floor (e.g., “the tool is smoking” or “there is a banging noise in the machine”) provide context to the sensor data. Hybrid data can improve predictive maintenance capabilities by providing ground truth to the “state” of the component being monitored. As an example, if an estimate shows that a bearing will fail in 5 days but the technician pulls it from the floor with zero damage, an analyst can update their model based on these observations. Similarly, if the sensor data indicates zero problems with the bearing, but the technician observes heavy smoke and noise, this can also improve the predictive model.

Hybrid data has significant potential for artificial intelligence (AI) techniques that can achieve improved accuracy and classification based on information captured through natural language. In addition, text data can complement numeric data by providing information about subsystems that may be outside the reach or sensitivity of the sensor-based systems (e.g., a tube that disconnects every few hours of equipment utilization, or a filter that needs to be changed so the surface of a component does not get contaminated).

This paper aims to illustrate the importance of this hybrid dataset to improve maintenance operations. First, we describe an experimental setup to generate real-world data that combines both text-based data via human observations and sensor data to investigate tool wear. This experiment can be run within any laboratory testbed or manufacturing environment. Second, we provide methods on how to merge this data and prepare it for analysis. This leads to initial insights about analyzing the data and how it can be used to improve maintenance operations. Lastly, we discuss improvements to the experiment and future extensions to this work.

3. Testbed Setup

This experimental work uses TechSolve’s M. Eugene Merchant Technology Development Center machining lab to develop a dataset as close to live manufacturing environments as possible. The facility has modern Computer Numerical Control (CNC) machines, a full array of measuring and analysis equipment (including force dynamometers, profilometers, Coordinate-measuring machine (CMM) equipment, microscopes, and inspection devices), and a variety of standard shop machines and equipment. The selected testbed consists of an instrumented machine tool. The experiments use the tool to run cutting tests under the close observation of experienced machine operators. The primary goal of the setup was to rapidly degrade a series of cutting tools under increased workloads while recording both instrument readings and periodic human observations via free-form text. This setup allows establishing relations between the human-generated and sensor-driven data. Periodic direct measurements of the tool-wear were also taken and are used as a “ground truth” basis for the health of the tool-piece in the results section of this paper.

The experimental setup included the following elements:

- Machine: Milltronics HMC35, instrumented with sensors, data acquisition system and tool condition monitoring system
- Cutting tools: Carbide end mill with 4 flutes, 0.5” diameter and 1” length of cutting zone
- Metalworking fluid: Water-based (Trimsol 206)
- Workpiece: 4140 steel block of 6” x 4” x 4”
- Fixturing: The workpiece, clamped into a vise with 6” long jaws

The data collection system enabled simultaneous acquisition of machine control data and data from the added sensors. The following data was collected from the Fanuc 0i Controller: date and time; the axes positions in absolute, machine, and relative coordinates; distance to go on each axis; actual spindle speed; actual feed rate; spindle load; spindle motor speed; exes loads; servo delays; and the acceleration or deceleration delays. The added sensors included:

- A three phase hall effect sensor, monitoring the power drawn by the motor of the spindle with an analog output of 0 to 10VDC corresponding to 0HP to maximum HP
- Uniaxial accelerometers ((Integrated Circuit-Piezoelectric (ICP) type), placed on the housing of one of the ball-screw bearings of each axis
- A tri-axial accelerometer on the spindle housing, an Integrated Electronics Piezo-Electric (IEPE) sensor with a higher sensitivity comparing to the accelerometers on the feed axes
- J-type thermocouples on each axis and each axis motor, on the spindle, and in the metalworking fluid tank

The machining center has other sensors and instrumentation, which were not used for this experiment.

4. Experimental Design

Using the setup described above, the experiment design focused on linking sensor values with simultaneous human observations. The experiment focused on the degradation of the cutting tools rather than the degradation of the machine itself, allowing repeatable and timely observations. The approach included the following steps:

1. Create a long cut machining program of at least 1.5 hours.
2. Modify the program such that some of the cuts exhibit chatter or higher than normal vibration – to create controlled process failures.
3. While the CNC program is running, collect data from sensors and controllers and have a technician, other than the one that created the machining program, come to observe the test periodically (e.g., 15 minutes) and take notes relative to the status of the cut, tool and machine (similar to creating maintenance logs).

To create the cutting program, TechSolve engineers identified the test conditions for this experiment through a series of exploratory cutting tests. These tests validated the physical combination of material, tooling, and cutting parameters, and helped establish the intervals for collecting human-generated information during the experiment. The results defined a CNC program that would enable a cut of approximately 3 hours (continuous cutting) including regions with chatter or increased vibration for the tool to simulate real-world observable events. The exploratory cutting tests also established a criterion for the end of tool life. The engineers observe the exploratory tests and use these observations to measure tool wear after machining one complete surface (one layer). A cutting test was considered complete after removing 6 layers of material.

The engineers use climb milling to machine the 6” x 4” surface of the workpiece (the steel block) in a transversal direction. A spindle tap test determined the chatter lobe diagram and the stability zone for the tool-spindle assembly. Based on this information, the team identified cutting conditions that would generate chatter. The CNC program’s design induced chatter to introduce abnormal cutting conditions in the experiment. Chatter was achieved by increasing the radial depth of cut and slightly increasing the feed rate. The team used a limited number of test scenarios to validate chatter generation and to observe the effects on the cutting tool and machined surface. Eventually, the team selected the cutting conditions presented in Table 1 for both the normal and abnormal (chatter) cuts.

The tests were planned and conducted as follows. One technician (Technician A) participated in the development of the test and of the CNC program, which included randomly inserted abnormal (chatter) cutting parameters. Another technician (Technician B) conducted observations on the cutting process at periodic intervals.

Each cutting test followed the procedure below:

1. The machine setup was prepared and the workpiece was clamped in the vise by Technician A.
2. Technician A took pictures of the fresh cutting tool prior to starting the test; both the end and lateral surfaces engaging the workpiece have been photographed.
3. Technician A started the CNC program, while Technician B observed and took notes of the process condition.
4. The machine ran linear climb milling cuts according to the CNC program created by Technician A.
5. While Technician B was free to move away from the machine, they were instructed to make text entries about the operation every 15 minutes. The technician was unaware prior to any given cutting pass if it was being performed with the chatter parameters. At 30 minute intervals, once a layer was removed from the steel block, Technician B had the option to pause the program and observe the tool and workpiece condition. A table was created in Excel to record various characteristics of the process. The idea was to emulate a data collection log similar to what is applied for MWOs. Table 2 lists a selection of the recorded characteristics.
6. A test was considered complete after 6 layers were removed from the steel block. A layer consisted in a volume of 6” x 4” x 0.5”, removed through the milling process with the 0.5” axial depth of cut.

The experiment automatically collected data from the sensors and machine tool control for each pass. Technician B was the main technician for observing the tests and collecting notes.

Table 1. Cutting Parameters

Parameter	Normal Cut (Stable)	Abnormal Cut (Chatter)
Axial Depth of Cut	0.5 in	0.5 in
Radial Depth of Cut	0.04 in	0.12 in
Tool Diameter	0.5 in	0.5 in
Cutting Speed	435 sfm	435 sfm
# of Teeth	4	4
RPM	3323 rpm	3323 rpm
Feed per Tooth	0.0018 in/tooth	0.002 in/tooth
Feed Rate	24.00 in/min	26.59 in/min
Feed Stroke	4.75 in	4.75 in

Table 2. Human Recorded Process Characteristics

Recorded Characteristic	Description
Test No. and Workpiece No.	Indication Value [1 - 6]
Various Test and Observation Times	Time Records
Test Status	Short Descriptor
Layer No.	Indication Value [1st / 2nd / etc]
Tool Gauge length (in)	Value
Type of observation	Short Descriptor
Process condition	Short Descriptor [Normal / Abnormal / other]
Tool condition	Short Descriptor
Tool flank wear - END (in)	Value [0 - 0.005 in]
Tool flank wear - LATERAL (in)	Value [0 - 0.005 in]
Tool wear - RAKE FACE (in)	Value [0 - 0.005 in]
Surface finish - face (μ in)	Value [$\sim$15 - 50 Ra]
Surface finish -lateral (μ in)	Value [$\sim$15 - 25 Ra]
Did you hear chatter?	Yes / No
Anything out of ordinary?	Yes / No
Out of ordinary description	Short Description
General Notes/Comments	Free Form Text

However, if Technician B was unavailable due to other tasks, Technician A would take notes. In general, the technicians were instructed to write down observations and, if needed, stop the cycle to observe what happened if they heard sounds or observed unusual behavior of the cut. Irrespective of observing irregularities or not, the test was stopped at the end of each machined layer (approximately 30 minutes) to allow Technician B to measure tool wear. Either technician would then start the process where it was left off, and the process of collecting numeric data and periodical observations continued.

5. RESULTS

We aim to establish a quantitative link between the recorded human observations and the sensed values of the test pieces. Establishing this link requires translating the human observations into some ordinal scaled value. Although we expect this to eventually be automated, we accomplished translation by having independent experts read the human-generated text. After the tests were completed, the experts read the entries to infer a level of damage between None to Imminent Failure. We presented text entries in the *Recorded Characteristics* table (Table 2) to 7 independent experts to interpret. They were instructed to disregard any measured wear or finish quality values and focus exclusively on the text entries. Figure 1 shows an example of the interpreted damage from the 7 experts. We use the average value of the expert interpretations to calculate the quantitative link between the sensor values, “ground truth”, and the human observations.

The manually-measured values of the tool wear are used as the “ground truth” level of degradation on each machine tool. Figure 2 shows that there is a strong correlation between human-interpreted damage level and the measured wear of the tool. The average correlation between the interpreted damage and the measured wear across all tests is 0.74, with a maximum of 0.97 and a minimum of 0.42. This includes all statistically significant correlation coefficients (p ¡ 0.05) for all of the three measured wear features: rake face, flank lateral, and flank end wear. Not all values were measured at all points. When there were not enough recordings to establish statistical significance, we omitted the correlation coefficient.

The next step focuses on quantifying the relationship between sensed values and human interpretations. Due to the asynchronous nature of the two sets of data, we employed a method for determining asynchronous correlation. This simple process uses dual signal interpolation to find the estimated overlap values for each data stream, then concatenates those values to determine a correlation coefficient. We present the breakdown

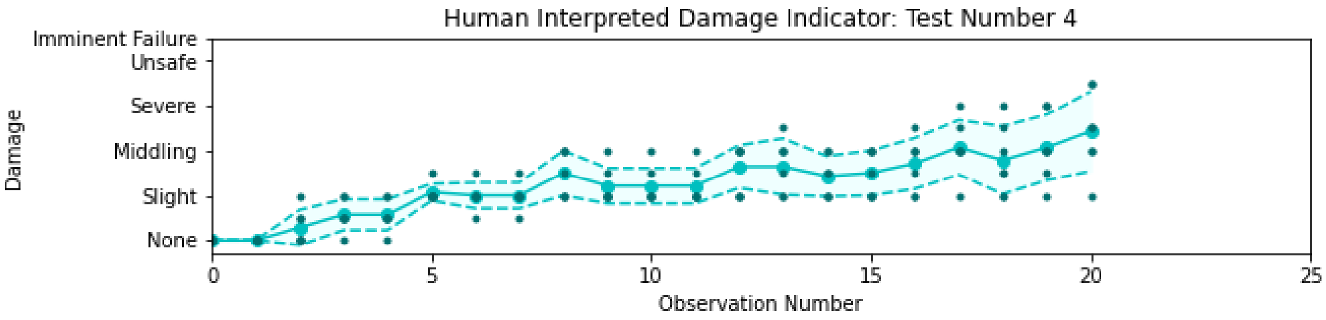


Figure 1. Results of Human Interpreted Damage for Test Piece 4

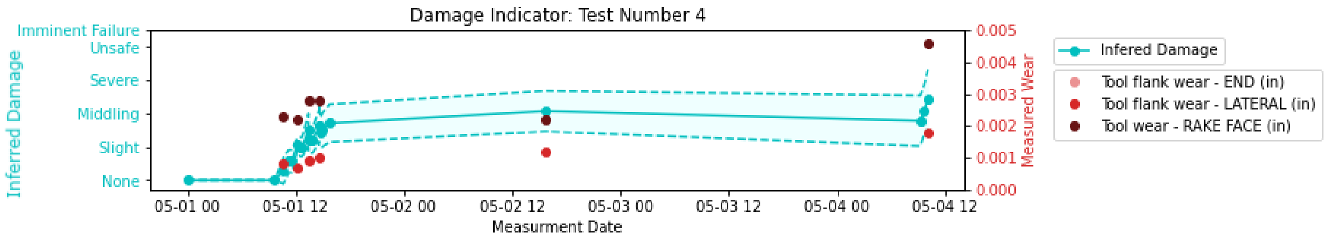


Figure 2. Relationship Between Human Interpreted Values and Ground Truth Wear
(END: Corr = nan, P = nan) (LATERAL: Corr = 0.80, P = 0.00) (RAKE FACE: Corr = 0.66, P = 0.00)

of that process as pseudo code below.

Algorithm 1: Pseudocode

```
With S1 = Signal 1 Values, S1t = Signal 1 Locations
With S2 = Signal 2 Values, S2t = Signal 2 Locations
# Find Points of Overlap
OL1 = t < max(S2t) & t >= min(S2t) for t in S1t;
OL2 = t < max(S1t) & t >= min(S2t) for t in S2t;
# Interpolate Sig X at Sig Y locations
if len(S1)>1 then
    F1 = interpolate(S1t,S1, kind = interpkind);
    S1est = F1(S2t(OL2));
else
    S1est = S1[0] for x in S2t(OL2);
end
if len(S2)>1 then
    F2 = interpolate(S2t,S2, kind = interpkind);
    S2est = F2(S1t(OL1));
else
    S1est = S2[0] for x in S1t(OL1);
end
# Concatenate and Calculate Correlation
S1cat = concatenate((S1(OL1),S1est),axis = 0) ;
S2cat = concatenate((S2(OL2),S2est),axis = 0) ;
return Correlation(S1cat,S2cat)
```

The goal of this work was not to establish an optimal degradation inference method from sensor signals, but to show that correlating the human observations with information derived from sensed values is not only possible, but may yield more robust methods for decision support than either could provide alone. Raw sensor data rarely is used in live settings for decision support, and so the decision was made to use a simplistic form of information extraction for the recorded sensors. This method compressed standard-sized windows of the raw signals into one root mean square value per window. This step both acts as a pseudo information extraction algorithm and makes correlating the signals more visually appealing with less computational demand due to the reduction of raw values. Future work may focus on more robust information extraction methods to connect the sensor values to the ground truth degradation of the tool wear.

Because of this simplistic data compression/ information extraction method, we expect a noticeable level of disconnect between the recorded values and the ground truth. However, the authors feel there is sufficient correlation in enough of the signals to establish that concurrent information exists.

This work focused on signals whose activity was expected to directly correspond to the wear of the tool. This included the power signal, the tri-axial accelerometer on the spindle, three accelerometers on the feed axes, and a suite of temperature sensors described in the previous section. The figures below show an example of the processed signals and their correlations between other sensor signals.

Inspecting Figures 1, 2, 3, we can visually identify a pattern of increasing values over time occurring in the human observations, the "ground truth", and the sensed values. Asynchronous correlation between the respective values confirms this visual cue.

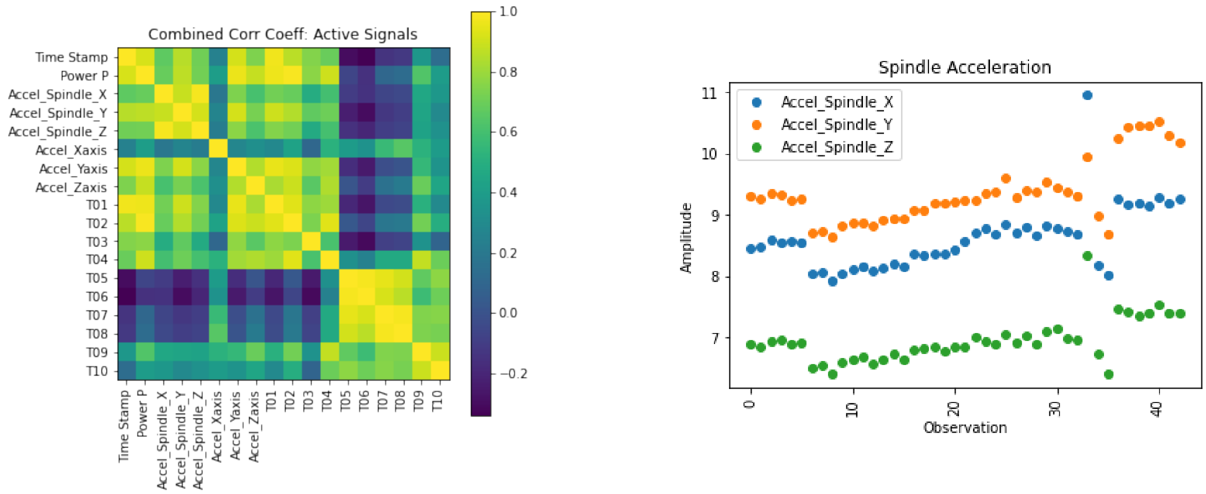


Figure 3. Relationships of Sensed Signals

Figure 4 shows the relationship for both the power consumption and the spindle acceleration to the human observations. These exhibit strong asynchronous correlations of 0.92 and 0.91, respectively.The average correlation between each group of signals and the human assessment can be found in Table 3.

Looking at this table, it is very important to note that most of the negative or low correlation values are due to the imprecision of the information extraction method applied to the sensors and the disregard for confounding factors. This is especially pertinent in the cases of the temperature signals where long delays between starts and stops can have a profound effect on the pattern of the signal. See Figure 5 as an example of this.

Notice that even in these examples that the signals visually trend with the human assessment inferred damage values. These trends, in turn, follow the ground truth pattern of the measured wear values and establish a clear link between the wear of a tool and the human assessment of that tool in a live working environment.

The visual confirmation, the calculated correlations, and the intuitive understanding of how humans process information provides ample evidence to support the idea that asynchronous human input should be used to supplement sensor readings in a live environment. This work justifies the exploration of deeper topics relating to the combination of heterogeneous data sources, both from human and mechanical sources. The value of using human agents as supplemental sensors in areas that are under-sensed or hard to evaluate makes intuitive sense. However, there is also value to adding that information to a well-sensed environment or asset. Humans are natural pattern identifiers and can add valuable insight or tacit information that a limited sensor set might not be able to. Future work will explore methods for incorporating this information and quantifying the value return of its use.

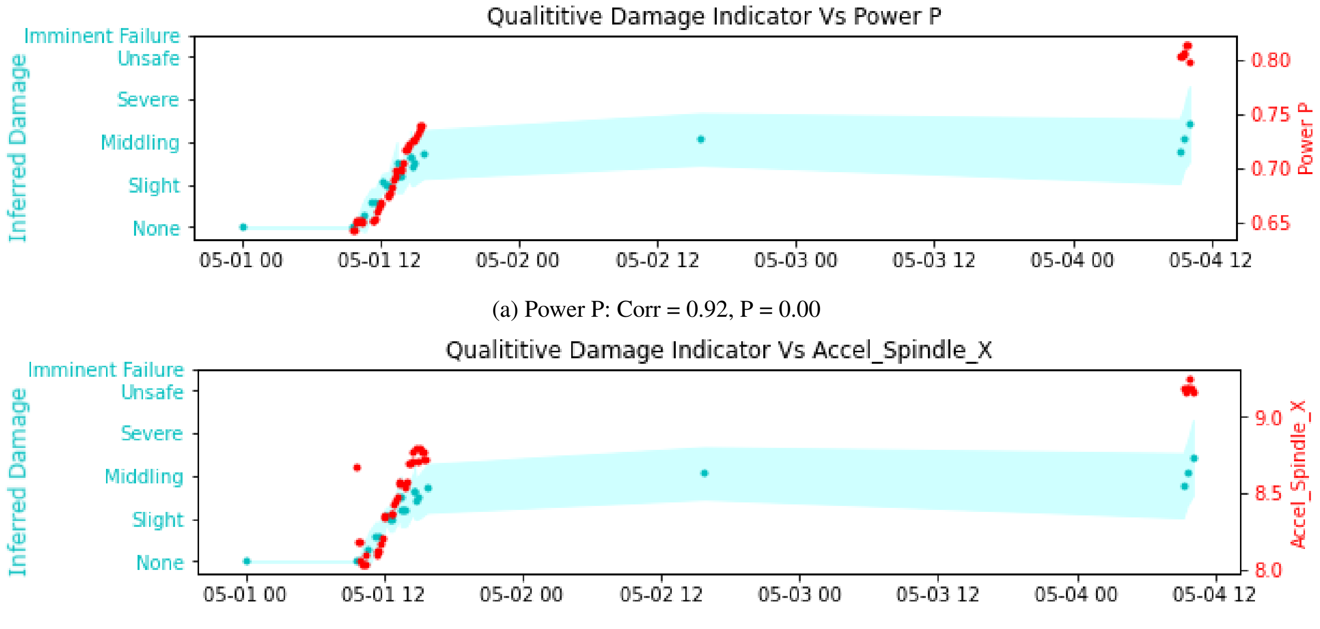
6. Discussion

This experimental setup and initial results provided some significant observations and potential avenues for further improvements.

Adding More Human Observers The major focal point for creating this dataset is the linking of the human observations to mechanically-sensed values. Of the two, the highest levels of variations will arise from the human observations, and thus, a greater number of human observations will allow for better relational development with the more rigorously sensed data. Although a high level of human observations or redundant observations may not be expected in a real-world scenario, adding more human observers would greatly enhance the development and validation of methods and technologies to best use this information in a real-world setting.

The current setup involved one human observer at any given observation of a machine. Future experiments can explore having multiple humans to simulate a real maintenance operation. The ideal dataset would have multiple people independently evaluating and recording regular observations of each machine through time. This would allow for developing better uncertainty bounds and expected variations between human agents.

Future investigations might also seek to find the saturation point for inserting human observations. Intuitively, humans will not effectively notice subtle changes over time if the change is gradual enough; the same is true for some computer-based monitoring systems. However, allowing a human agent to step away from the system, then check back on it after some interval may circumvent this problem. Identifying the interval of greatest return



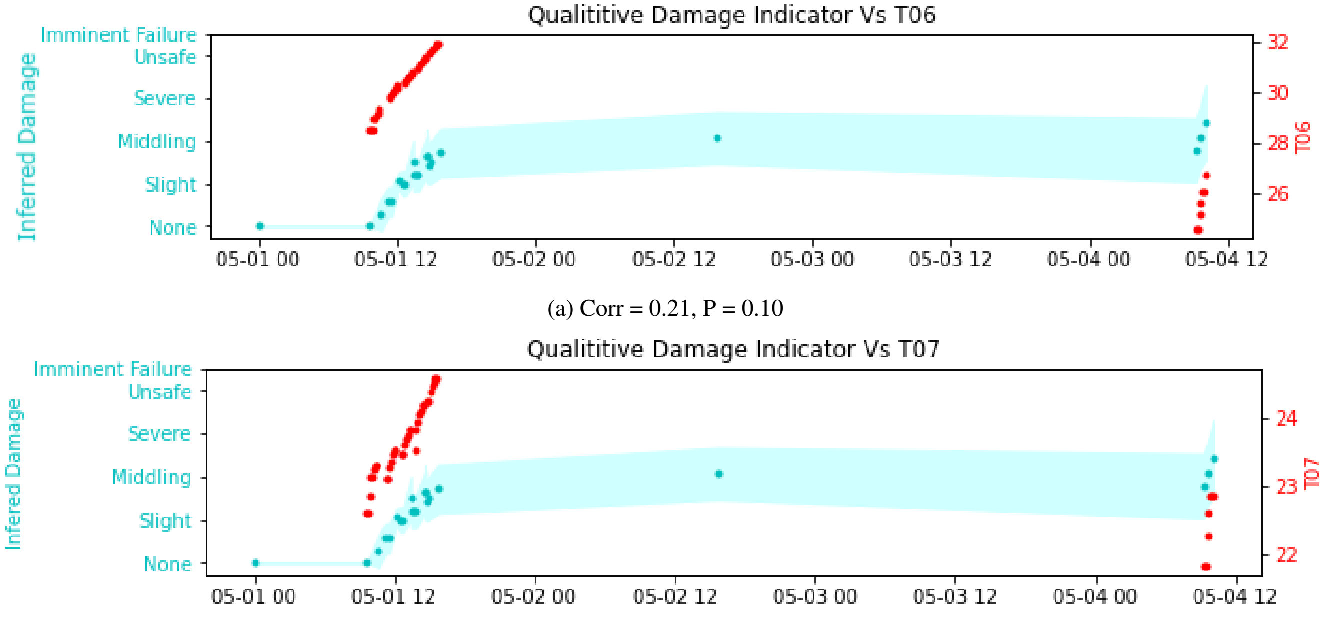
(a) Power P: Corr = 0.92, P = 0.00

(b) Accel Spindle X: Corr = 0.91, P = 0.00

Figure 4. Relationship Between Sensed Signals and Human Assessment

Table 3. Asynchronous Correlations Between Sensors and Human Assessment

	Correlation to Human Damage Assessment		
Sensor Group	**Average**	**Max**	**Min**
Power	0.44	0.92	-0.69
Spindle Accelerometers	0.66	0.91	0.22
Machine Accelerometers	0.64	0.85	0.39
Temperature	0.11	0.95	-0.79



(a) Corr = 0.21, P = 0.10

(b) Corr = 0.09, P = 0.50

Figure 5. Example of Temperature Shift Caused by Delayed Restart

from human observation could help to schedule regular 'walk-through' style checkups on various processes and machines. This process would be especially useful for assets that are not fully equipped with mechanical sensors.

Incorporating NLP/TLP Humans are best suited to express their observations through free-form text. Although Likert scales and similar can be useful in analyses, they are prone to inconsistency and lack much of the contextual information that free-form text can provide Hodge and Gillespie (2003). Additionally, much of the human-collected information already available in industrial settings (e.g., maintenance work orders) does not intrinsically contain this type of structured information. This motivates and necessitates a focus on natural language processing (NLP) or technical language processing (TLP) as a means to automate and capture a more full scope of any human observation.

By its very nature, free-form text is somewhat inconsistent, and as such, difficult to definitively confirm ground truth. To help circumvent this, this work created a post-collection Likert scale value for each entry from a human observer. These were made by aggregating the responses of multiple experts, but fundamentally may or may not have captured the original observer's intent. An ideal test setup would prompt the human observers to provide some Likert-style value as well as the free-form text. This setup, coupled with the post-collection interpretations, could be used to train and validate any developed NLP/TLP models as well as provide more information about the expected uncertainty within the data.

Developing state-of-the-art language processing models could enable deeper and more rich use of currently available data. These would eventually preclude the need for Likert-style assessment from the human operators. These advances can allow observers to express their assessment more efficiently. In turn, this procedure increases the chances that the observer will provide useful information.

Focus on Temporal Asynchronicity Facilities often sense and record values at unaligned intervals. Adding in irregular and inconsistent observations and recordings from human agents creates a strong need to address methods for merging asynchronous data streams. This work showed visual alignment, as well as one method for interpolating expected values to create alignment. However, these are not the only, nor the best possible methods.

Addressing asynchronicity can be done in a multitude of ways that will largely be dictated by the desired results and the preferred models. Although it would take very little effort to find a solution for a given application, future work may want to focus on finding the optimal methods to aggregate, incorporate, merge, or align heterogeneous data sources that are largely asynchronous.

Experimental Scope The scope of any future experiments should build upon those that have come previously. The current experiment focused on a single type of machine tool, specifically a CNC milling center. Conversely, a future experiment could focus on a series of machines to provide scenarios similar to a full manufacturing line. Such a setup would allow researchers to not only understand process parameters but manufacturing system dynamics as well.

Future experiments should first recreate the scope of the experiment described in this work with a larger number of test units. This work's experiment involved a single type of failure on a single type of asset. Next, experiments should progress to multiple types of failures on a single asset. Later experiments could address multiple failure types amongst multiple assets. We recommend this progression of experiments because it allows for steady validation of developed tools while retaining focus on the human-supplied portions of the experimental data.

In any experiment, a minimum number of assets should be subjected to trials to develop statistical significance and allow for variation across the human observations. Although needs may vary depending on specific setups, the authors suggest a minimum of 50-100 entries confirmed by some 'ground truth' as a starting point. NLP or TLP tools may require hundreds or even thousands of entries Brundage et al. (2021).

Real-World Data Concerns Although controlled testing and environments greatly ease the process of defining and developing tools and technologies for incorporating human-derived information, the perspective that these will ultimately be used in a live industrial setting should not be lost. Whenever possible, steps should be taken to ensure that the types and formats of data ultimately reflect those that could or would be acquired in an industrial setting.

Unfortunately, obtaining real-world data, particularly industrial data, can be difficult due to proprietary restrictions and fear of losing competitive advantage. Whenever possible, providing real-world datasets can help researchers advance their analysis methods by verifying their tools and test datasets against real industrial data. Making reference datasets - both laboratory and real-world - available to the general public can accelerate the development of applicable tools, best practices, and standards.

7. Conclusions

This paper discusses the need and a methodology to create a dataset with both sensor-based and human-generated data.

Our initial analysis illustrates the value of capturing this information within the scope of maintenance operations. Our results show a strong correlation between a human interpretation of the system, ground truth measurements, and hard sensor values captured during the experiment.

This work provides the initial motivations and justifications for further developing rigorous methods to utilize human-derived data in the traditionally-incompatible environment of sensor-driven technologies. We provide discussions on successes and challenges faced during this experiment, along with a loose guide on improvements for future work.

Skilled humans will always be one of the most accurate tools for assessing the broadest intake of direct and indirect information about a system. Sensing equipment can provide more consistent and objective precision than any single human. Each is well suited to provide incredibly useful information for their respective areas of excellence. The challenge we highlight and begin to address is the development of datasets that show this. Datasets facilitate the development of tools and technologies that capture and capitalize on both types of valuable information. The future of industry lies at the intersection of humans and technology.

Acknowledgement

The authors would like to thank Timothy Sprock and Moneer Helu from NIST for their guidance in crafting the experiment.

NIST Disclaimer

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

References

Brundage, M. P., Morris, K., Sexton, T., Moccozet, S., & Hoffman, M. (2018). Developing maintenance key performance indicators from maintenance work order data. In *International manufacturing science and engineering conference* (Vol. 51371, p. V003T02A027).

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, *27*, 42–46.

Djurdjanovic, D., Lee, J., & Ni, J. (2003). Watchdog agent—an infotronics-based prognostics approach for product performance degradation assessment and prediction. *Advanced Engineering Informatics*, *17*(3-4), 109–125.

Ho, M. (2015). *A shared reliability database for mobile mining equipment* (Unpublished doctoral dissertation). University of Western Australia.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to likert scales. *Social Work Research*, *27*(1), 45–55.

Jin, X., Weiss, B. A., Siegel, D., & Lee, J. (2016). Present status and future growth of advanced maintenance technology and strategy in us manufacturing. *International journal of prognostics and health management*, *7*(Spec Iss on Smart Manufacturing PHM).

Katipamula, S., & Brambley, M. R. (2005). Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i. *Hvac&R Research*, *11*(1), 3–25.

Kothamasu, R., Huang, S. H., & VerDuin, W. H. (2006). System health monitoring and prognostics—a review of current paradigms and practices. *The International Journal of Advanced Manufacturing Technology*, *28*(9-10), 1012–1024.

Kunche, S., Chen, C., & Pecht, M. (2012). A review of phm system's architectural frameworks. In *The 54th meeting of the society for machinery failure prevention technology, dayton, oh.*

Li, R., Verhagen, W. J., & Curran, R. (2018). A functional architecture of prognostics and health management using a systems engineering approach. In *Proc. eur. conf. phm soc* (pp. 1–10).

Lu, B., Li, Y., Wu, X., & Yang, Z. (2009). A review of recent advances in wind turbine condition monitoring and fault diagnosis. In *2009 ieee power electronics and machines in wind applications* (pp. 1–7).

Lukens, S., Naik, M., Saetia, K., & Hu, X. (2019). Best practices framework for improving maintenance data quality to enable asset performance analytics. In *Annual conference of the phm society* (Vol. 11).

Sexton, T., Brundage, M. P., Hoffman, M., & Morris, K. C. (2017). Hybrid datafication of maintenance logs from ai-assisted human tags. In *2017 ieee international conference on big data (big data)* (pp. 1769–1777).

Software, S. P. (2018). "the connection between production monitoring and oee" [Computer software manual]. (Available at `https://www.engineering.com/story/40841`. Accessed 04-05-21)

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering*, *27*(3), 293–311.