Harvard Data Science Review • Issue 2.2, Spring 2020

Data Science in Heavy Industry and the Internet of Things

Michael Horrell¹ Larry Reynolds¹ Adam McElhinney²

¹**Uptake, Chicago, Illinois, United States of America,**

²**Paylocity, Schaumburg, Illinois, United States of America**

Published on: Apr 30, 2020

DOI: <https://doi.org/10.1162/99608f92.834c6595>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Increasingly cheap and available sensors enable new applications of data science for heavy industries. From locomotives to wind turbines to solar farms, data scientists close to these industries will be the first to attempt to turn the data collected on these machines into consistent sources of economic value. This article discusses an approach to framing industrial analytics problems and goes into detail on one problem in equipment reliability, predictive maintenance. We discuss a host of challenges associated with building and implementing analytics using equipment data, and we give recommendations on how to surmount these challenges through careful data analysis, data collection, and communication. We also discuss training and getting started on industrial analytics problems.

Keywords: Internet of Things (IoT), industrial internet, failure prediction, prognostics, telematics, machine learning, predictive maintenance

1. Introduction

The stereotypical environment for a data scientist is decidedly not heavy industrial. Sleek workplaces furnished with beanbags may seem like a far cry from a factory, a mine, or other environments relying on heavy equipment to do work. However, the Internet of Things (IoT)—a term referring to the estimated billions of devices that can collect data with sensors and transmit that data—will change this image for some practitioners. Heavy equipment can contain hundreds or thousands of sensors, and, with the rise of IoT, the data collected by these sensors can be accumulated and analyzed to create economic value.

Increased connectivity of heavy equipment, and, more generally, connectivity of any device with a sensor, is driven by a number of factors. Decreasing costs of bandwidth, accessibility of Wi-Fi and cellular networks, and robust cloud infrastructures are making sensor data collection, transmission, storage, and analysis easier; see the study by Goldman Sachs (2014). This study estimated there were around two billion connected devices in 2000 and a projected 28 billion connected devices by 2020. Consumer products such as exercise bracelets and smart thermostats may be the most visible examples of this phenomenon; however, this same study estimated the opportunity for IoT in the industrial space alone to be \$2 trillion in 2020. These estimates, of course, are based on assumptions and data collected in 2014; therefore, some caution is warranted when interpreting these numbers. However, more recent estimates in Gartner (2018), IoT Analytics (2018), and Ericsson (2016), further indicate the market for IoT is large and growing.

Given these new opportunities, traditional industrial companies, tech companies, and a host of startups are competing for space in the industrial IoT market. To do this, many are relying on data scientists to analyze, visualize, and create predictions from these new data streams. Uptake, the company we work for, is one startup that focuses on equipment reliability and productivity. This article focuses on our experience building data

science solutions for industrial IoT applications. We first present our approach to framing problems in industrial IoT. Next, we discuss predictive maintenance, a method of using IoT data to improve maintenance practices. In particular, we use predictive maintenance to highlight the challenges present in working with sensor data and describe our approaches to overcoming these challenges. Finally, we discuss training for aspiring industrial data scientists.

2. A Top Down Approach to Creating Value From Industrial IoT Data

Sensor data from heavy equipment are, materially, no different from many other data sources. For example, GPS measurements on construction vehicles could be used in a consumer application to provide motorists with more accurate traffic predictions. However, the new opportunity presented by these data is to improve the efficiency and operation of businesses within traditional industries. Industry analysts have written that the Fourth Industrial Revolution will be enabled in part by the data availability that comes with the industrial IoT (Hanley et al., 2018). We focus our discussion on a small piece of this transformation. Specifically, we discuss the question: If a company relies on heavy equipment to be productive, how can a data scientist use sensor data to enhance that productivity? We describe an overall approach to answering this question that can be applied to individual companies or industries.

To begin solving problems in industrial IoT, we encourage data scientists to start from the basics of a company's business. Our view is that it is critical for data scientists to understand the details of how a company creates value and, more generally, the key performance indicators (KPIs) that companies often measure themselves on. Data scientists then measure their performance by showing improvement on appropriate KPIs.

For example, in the rail industry, failures per locomotive year (FLY) is a core KPI that gets tracked. Mechanical failures not only result in expensive repairs, but the associated unplanned downtime can be even more costly. Revenue lost due to unplanned downtime has been estimated at \$160,000 per locomotive per year, and it has been estimated that Class 1 railroads (those generating a minimum of around \$400 million in revenue per year) can realize an annual savings of \$80 million if only 10% of unplanned maintenance is converted to planned maintenance (Predikto, 2017). Reduced FLY lowers both maintenance and unplanned downtime costs by catching failures before they get serious and before they affect the overall operation of a rail network. Data scientists in this area can then be confident they are creating value by focusing on reducing failures.

Data scientists won't necessarily be asked to tie their work to specific KPIs—a data scientist working in a purely consultative capacity may simply need to solve a set of problems already defined by a stakeholder—however, we believe there are a number of advantages to proactively defining problems and solutions in this way. First, a data scientist's work is clearly tied to a company's mission and bottom line. Second, focusing on business drivers can provide self-evident success criteria for the project and can improve communication across all stakeholders. And third, issues of scale and solvability tend to be surfaced earlier, potentially saving data scientists and others time and effort.

An example in electrical power transmission illustrates the third point. System Average Interruption Duration Index (SAIDI) is a measure of power outages and severity (Institute of Electrical and Electronics Engineers Standards Association [IEEE SA], 2012). However, outages and equipment failures in this industry are frequently caused by squirrels (American Public Power Association [APPA], 2017). The American Public Power Association has even written a tongue-in-cheek “Open Letter to Squirrels” as a tribute to their ubiquity (APPA, 2019). It may be possible for data scientists to estimate spatiotemporal averages of ‘squirrel-risk’ as an attempt to protect against squirrel-related outage events, but of course a data scientist cannot, on any given day, predict whether such an event will happen.

With the hype around both data science and the Internet of Things, data scientists will be under extra pressure to create compelling solutions. We believe a concerted focus on the mechanics of how solving a data science problem leads to business value will help ensure that the problems attempted are realistically solvable and valuable.

3. Predictive Maintenance

Machines and components inevitably wear, degrade, and break. Per the examples in Section 2, this can be costly and negatively affect key productivity measures. Companies using heavy equipment create reliability strategies to deal with wear-out and breakages while simultaneously maintaining productivity. Periodic oil changes, for example, are part of a reliability strategy. A decision to run a piece of equipment to failure is a valid reliability strategy as well. Given that reliability is important for companies across many different industries, we provide a deeper dive on predictive maintenance, one of the more technical ways data scientists can help improve equipment reliability.

3.1. Background

The math behind predictive maintenance is referred to in some literature as Prognostics (Lee et al., 2014; Lei et al., 2018; Roemer et al., 2006; Sikorska et al., 2011; Wheeler et al., 2009). Leading up to a machine failure, signatures of the impending failure—for example, an increasing temperature or a dropping pressure—can sometimes be captured by sensor data. The prognostic/failure prediction model focuses on detecting these signatures as soon as possible. If a problem is caught early enough, repairs may be as minor as tightening a bolt. The longer a potential failure goes undetected, often the more expensive it is to repair.

Figure 1 is adapted from Blann (2013) and shows this rough phenomenon: a failing component’s performance and condition degrades as it reaches a total failure point. Along the way it hits a couple important points in relation to predictive maintenance. Point S is the start of a failing component. Point P is the point where it is observed in the data. And point F indicates a completely failed component. Notably P is different from S, indicating the actual onset of a failing component may occur significantly earlier than it is actually detectable in what data is available. The period between when a problem is detected (P) and when a component completely fails (F) is what is labeled the PF interval. It is in this period that attempted corrective actions can

be taken to reduce overall costs. Different components will have different curves and thus different PF intervals. For example, a single failing bolt on some machines may not be detectable before failure. In this case, point P is on top of point F. Predictive maintenance is concerned with moving point P to as early in time as possible.

Figure 1. Adapted from Blann (2013), the theoretical PF curve describes important points in a predictive maintenance problem. Point S indicates the start of a failing component. Point P indicates the point at which a failing component is observed given existing data streams, and point F indicates the point of functional failure. The condition of a component and timing around these points may vary from component to component.

3.2. Data and Implementation Challenges

As with many data science problems, the core of solving a predictive maintenance problem involves gathering data, conducting analysis, building and deploying a model, and tracking outcomes and feedback to ensure the model is performing appropriately. A host of technical and statistical issues make this challenging. We enumerate a set of challenges here and refer to them in the following subsections containing our recommendations.

(A) Data quality is difficult to guarantee

A1 Lack of complete and quality failure information is perhaps the most difficult problem to solve.

Unlike sensors that collect data automatically, documentation of failures and their fixes usually depends

on mechanics in a shop doing this documentation. Unsurprisingly, data quality for data science is not a priority for most mechanics. Some shops work on paper records as well. This adds another layer of complexity to getting the right data into the hands of data scientists.

A2 Many mobile machines rely on cellular or satellite connections to transmit data. For older nonmobile machines, sending data to the cloud often means retrofitting to older hardware. In both cases, drops in data can occur and connection can be expensive. This creates data that can be spotty and out of order, contain duplicates, and can force tradeoffs to be made on what data to collect even before a data scientist has seen data samples. In addition, critical or erratic machine operation can also cause issues for sensors, creating a scenario where data gaps exist precisely during the critical periods where data is needed.

A3 Outside of connectivity, sensor configurations also cause headaches because not all sensors are installed in precisely the same way on even the same types of machines. This causes modeling issues because some type of central calibration may be needed before a model can be applied confidently at the desired scale. In addition, not all components have sensors that can be used for predictive maintenance. Many parts won't have sensors and other parts may have sensors that do not serve predictive maintenance purposes.

(B) Clean data isn't always revealing or easy to work with

B1 Replacements are not equivalent to failures. For example, planned maintenance, such as changing oil every 3,000 miles, results in replacements without failure. Part failures also cause working parts to fail. For example, a flat tire could cause a collision, further causing other part replacements. The consequence of this is that even a perfect record of part replacements may not provide a consistent target to train against when building a machine learning model.

B2 High value failures are rare. Machine prognostics mirror medical prognostics and survival analysis where events are rare or censored (Ambler et al., 2012 Buyske et al., 2000). While this is good news for companies operating these machines, data scientists may find difficulty when gathering even years of data yield only a handful of failure examples to work with. In addition, for the most complicated machines, there are a wide range of failure types. This can mean that the value of preventing a single failure type may be negligible, but value grows significantly as more failures are prevented.

B3 In contrast to rare failure data, sensor signal data can be enormous. Vibration sensors collect data many thousands of times per second. Nonvibration sensors will collect data once per second or more frequently. This puts a strain on computation when doing exploratory analysis, and in some cases, practitioners will need to work only on summaries of the underlying data as opposed to the raw data itself.

B4 Many sources of data are highly dependent. In a statistical sense, all data coming from a single piece of equipment are dependent. All data coming from groups of equipment in the same geographic area are

dependent. Even data generated by different pieces of equipment but with the same operator will be dependent. Data dependencies affect and may limit modeling and validation approaches.

B5 Machine context matters. Machines age, operate in hot and cold climates, go into tunnels and through mud and work in otherwise very extreme conditions. Each of these modes of operation can change the signatures of data coming off a machine.

(C) A perfect prediction alone doesn't directly translate into value

C1 If failure signatures are detected by a model, acting on a model prediction requires manual work and logistics. For example, to replace a failing part on a machine, the right new part must be available at the right maintenance shop. Inventory management presents tremendous challenges on its own; see, for example, Williams and Tokar (2008) for an overview. Creating the right prediction and delivering it in such a way that enables the right follow-up workflow can be a challenge.

C2 Predictive maintenance problems can be “high-stakes” problems (Rudin, 2019). High dollar amounts—and in some cases, human safety—are connected to actions both taken and not taken. Consumers of predictions must be able to trust a prediction in order to confidently take the right actions.

Successful approaches to predictive maintenance and prognostics will confront some of these issues head-on and side-step others.

3.3. Recommendations for Model Building

Prognostics models may be as simple as creating a rule—for example, a simple low-fuel indicator is a rule that helps operators prevent fuel outages—or may involve complex physical simulations to determine acceptable bounds for mechanical parameters (Lei et al., 2018; Sikorska et al., 2011). Machine Learning approaches fall somewhere in the middle of these extremes in terms of complexity and focus directly on developing functions of the data to optimize empirical performance metrics. We give recommendations around machine learning model building based on our experience. Given many of the data challenges discussed in Section 3.2, our recommendations involve collecting the right data to enable model building and handling that data carefully so that the right conclusions from the data can be drawn.

1. *Focus on good cross-validation.* Cross-validation is generally good practice for modeling (see for example, Taddy, 2019, and Draper, 2013). However, points A1, A2, A3, B1, and B5 make it especially difficult to trust a predictive maintenance model on training performance alone. Dependency in the data (point B4) creates additional overfitting concerns. Various forms of blocking (grouping certain data together so dependent data doesn't end up in both training and testing procedures) are practical ways to deal with dependent data and will ensure offline performance metrics more accurately reflect performance of a deployed model (Burman et al., 1994). Machine learning methods that are more robust to overfitting—for

example, random forests (Breiman & Cutler, 2005)—are not a replacement for good cross-validation in our experience. Glickman et al. (2019) show similar findings.

2. *Gather information from subject matter experts (SMEs).* Given that failure records may not be dependable for model building (points A1 and B1), input from SMEs can help cover gaps in records. Their input such as understanding of physical properties or operational context can also help modelers make better sense of high dimensional data and rare failures (B2 and B3). SMEs' input won't address all informational and data gaps; for example, we have gone into the field to directly gather the right data in some cases. However, leveraging the holistic experience offered by many SMEs can help data scientists build and contextualize their models faster (Berinato, 2019; Kozyrkov, 2019).
3. *Systematically gather contextual data.* Per A3 and B5, individual machines can experience a wide range of conditions. Collecting data on these conditions—and making that data available at model runtime—allows models or modelers to account for different modes in the data. One interesting example we encountered involved gathering data on locomotive tunnels. Running a locomotive inside a tunnel causes average temperatures to rise and creates spikes in other signals. Under nontunnel conditions, these signatures could indicate impending part failures. After determining the existing model would not be able to properly differentiate between problems and tunnels, our team (1) built a map of all tunnels in the associated rail networks and (2) made this information available as a feature to control our model. The additional contextual data in this example helped reduce false positives from our model. Contextual data can be used to enhance visualizations as well.
4. *Seek out or create data sources that measure direct component degradation or performance.* As a simple example, consider again the fuel gauge. If running out of fuel is a failure condition, the fuel gauge provides a direct measurement of remaining operational life. In this example, solving the predictive maintenance problem is almost as simple as checking the fuel gauge. Machine components will rarely have measures this direct (point A3), but if they do, they should be found and used. If a degradation measure does not exist, it can be created in some cases. Vibration sensors, for example, are added to equipment to co-indicate degradation of bearings and other components (Lei et al., 2018). Using vibration data, root mean square (usually called RMS), a measure of vibration energy, can be measured and trended to find systems or components that are not operating properly (Lei et al., 2018).

The final points associated with turning model predictions into real-world value (C1, C2) may be best addressed through clear model interpretations and communication.

3.4. Recommendations for Communication

Good predictions alone do not immediately translate into value (points C1 and C2). Building trust in a single prediction requires clear interpretations and clear evidence. Building trust in a set of predictions may additionally require experimentation and A/B testing.

When communicating a single predictive maintenance prediction, we prefer to express predictions in binary terms. We also attempt to automatically present clear evidence in support of both positive and negative predictions. We pair our predictions with written interpretations beginning with a phrase like, ‘There is evidence of a problem,’ with the evidence presented in well-thought-out figures or a series of plots. Or, to communicate a negative prediction, we might write ‘there is no evidence of a problem,’ a statement that should be readily verified with accessible data. More generally, our aim is to present model predictions as simply another data source. Like any data source, it relies on assumptions and can be wrong. Like any data source, consumers should be familiar with the assumptions and premises resulting in a prediction. We believe thinking through and expressing predictions in this way—even in cases where a significant amount of uncertainty about a prediction exists—empowers consumers to evaluate predictions themselves so that ultimately the right actions can be taken. As potential ‘high stakes’ predictions, we believe this approach also aligns with approaches discussed in Rudin (2019), which call for deeper interpretability of models used in scenarios like these.

The value of clear, interpretable predictive maintenance predictions will be self-evident in many contexts. When it is not, an ideal way to quantify value is to conduct an experiment (Taddy, 2019). Understandably, industry stakeholders do not jump at the chance to have data scientists run experiments involving their multimillion-dollar assets. Likewise, data scientists should not feel completely free to conduct any experiment they like, since they will not bear the full cost of running those experiments. We find that when experimentation is possible, impactful experiments depend on trust built with stakeholders.

As one example, our team was able to run an A/B test to identify a large set of mis-calibrated machines for one customer. To start this experiment, we worked with the manager of these machines to identify a subset to give a special calibration as a treatment. This subset was chosen to maximize measurement capability and minimize potential impact on operations. The machines outside this subset were left untreated. By tracking productivity of these machines posttreatment, we proved that the special calibration improved output. Consequently, all machines were given the calibration, creating a measurable bump in output for that population of machines. This was a great outcome for both parties. We achieved this by building trust through quality communication and finding an acceptably small but measurable way to get to our goal.

4. Training and Getting Started

For prospective data scientists looking to add IoT to their expertise, the usual data science skillset remains extremely relevant (see, for example, Berthold, 2019). Industrial data scientists should be strong in math and statistics, adept at executing quality cross-validation, experts in developing software in the core languages R and Python, and able to communicate analyses effectively to many audiences, whether it is a mechanic or an executive (McElhinney, 2019).

Following our recommendations in Section 3, prospective data scientists will need additional focus in interacting with experts. Industrial analytics isn't the only place where interfacing with experts is important. For example, data scientists working on medical applications may communicate directly with medical doctors. However, given the data challenges described in Section 3, and especially for predictive maintenance problems, it can be absolutely critical for data scientists to interface with an industry expert. Importantly, they must be able to do this while maintaining overall control of how a problem is being solved. To borrow a phrase from Meng (2018), data scientists should strive to be “Proactive co-investigators/partners, not passive consultants” (p. 51). For the practicing data scientist, this means bringing data and plots to conversations with specific research questions in mind. Conversely, data scientists should avoid statements like: ‘the expert said X, so I did X’; or questions like: ‘does the expert want Y in the model?’ Follow-up questions like ‘was X justified by the data and our understanding of the problem?’ or, ‘does Y lead to any substantive improvements in the model?’ will help data scientists create better solutions.

Given the importance of interaction with subject matter experts, data scientists with additional experience with machines and mechanics can significantly speed up model building as well. We have seen many cases where just knowing relative locations of components on a machine has potentially saved weeks of model-building time. To jumpstart this process, we have sent data scientists to formal training events intended for mechanics and other heavy equipment analysts.

For those looking to get their hands on sample data, NASA (2014) collects a number of data sets that track devices as they fail in either simulations or lab experiments. Turbofans (Saxena & Goebel, 2008), bearings (Lee et al., 2007), and batteries (Saha & Goebel, 2007) are some examples of data sets that are open to the public. These data sets are great for practicing cross-validation and playing with methods to find early patterns of failure in these devices. However, many of the data issues we mentioned in Section 3 may not be present in lab experiments. Practitioners getting started with these data sets should keep this in mind to make sure their methods do not become ineffective in real data scenarios.

Industry analysts have high hopes that IoT will bring transformations to many traditional industries. Using IoT data to change how heavy equipment is operated and maintained is a part of this expected transformation. While heavy equipment may not be many data scientists' traditional area of application, a passive approach to solving problems in this area may ultimately fall short of creating transformation. Data scientists will be successful in helping realize this future if they play proactive roles in defining the right problems, gathering the right data, and taking the lead in communication.

Disclosure Statement

Michael Horrell, Larry Reynolds, and Adam McElhinney are currently or were previously employed by Uptake, a company specializing in analytics for Heavy Industry.

References

- Ambler, G., Seaman, S., & Omar, R. Z. (2012). An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in medicine*, 31(11–12), 1150–1161.
<https://doi.org/10.1002/sim.4371>
- American Public Power Association. (2017). *Defending against outages: Squirrel tracker*.
<https://www.publicpower.org/blog/defending-against-outages-squirrel-tracker>
- American Public Power Association. (2019). *An open letter to squirrels*.
<https://www.publicpower.org/blog/open-letter-squirrels>
- Berinato, S. (2019). Data science and the art of persuasion. *Harvard Business Review*, (January–February 2019), 126–137. <https://hbr.org/2019/01/data-science-and-the-art-of-persuasion>
- Berthold, M. R. (2019). What does it take to be a successful data scientist? *Harvard Data Science Review*, 1(2).
<https://doi.org/10.1162/99608f92.e0eaabfc>
- Blann, D. (2013). *Maximizing the P-F interval through condition-based maintenance*. Maintworld.
<https://www.maintworld.com/Applications/Maximizing-the-P-F-Interval-Through-Condition-Based-Maintenance>
- Breiman, L., & Cutler, A. (2005). *Random forests—Classification description*. Random Forests.
https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Burman, P., Chow, E., & Nolan, D. (1994). A Cross-validatory method for dependent data. *Biometrika*, 81(2), 351–358. <https://doi.org/10.2307/2336965>
- Buyske, S., Fagerstrom, R., & Ying, Z. (2000). A class of weighted log-rank tests for survival data when the event is rare. *Journal of the American Statistical Association*, 95(449), 249–258.
<https://doi.org/10.1080/01621459.2000.10473918>
- Draper, D. (2013). Bayesian model specification: Heuristics and examples. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 409–431). Oxford University Press.
- Ericsson. (2016). *Internet of Things forecast*. <https://www.ericsson.com/en/mobility-report/internet-of-things-forecast>
- Gartner. (2018). *Gartner identifies top 10 strategic IoT technologies and trends* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2018-11-07-gartner-identifies-top-10-strategic-iot-technologies-and-trends>

Glickman, M., Brown, J., & Song, R. (2019). (A) Data in the life: Authorship attribution in Lennon-McCartney songs. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.130f856e>

Goldman Sachs. (2014). *The Internet of Things: Making sense of the next mega-trend*. <https://www.goldmansachs.com/insights/pages/internet-of-things/iot-report.pdf>

Hanley, T., Daecher, A., Cotteleer, M., & Sniderman, B. (2018). *The Industry 4.0 paradox*. Deloitte. <https://www2.deloitte.com/us/en/insights/focus/industry-4-0/challenges-on-path-to-digital-transformation/summary.html>

Institute of Electrical and Electronics Engineers Standards Association. (2012). *1366-2012-IEEE guide for electric power distribution reliability indices*. <https://standards.ieee.org/standard/1366-2012.html>

IoT Analytics. (2018). *State of the IoT 2018: Number of IoT devices now at 7B—Market accelerating*. <https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/>

Kozyrkov, C. (2019). What great data analysts do—and why every organization needs them. In *Strategic analytics: The insights you need from Harvard Business Review* (Advance Edition). Harvard Business Review Press.

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 42(1–2), 314–334. <https://doi.org/10.1016/j.ymssp.2013.06.004>

Lee, J., Qiu, H., Yu, G., Lin, J., & Rexnord Technical Services. (2007). *Bearing data set*. NASA Ames Prognostics Data Repository, NASA Ames Research. <http://ti.arc.nasa.gov/project/prognostic-data-repository>

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104, 799–834. <https://doi.org/10.1016/j.ymssp.2017.11.016>

McElhinney, A. (2019, March 25). Developing a data science career framework. *Uptake Tech Blog*. <https://medium.com/uptake-tech/developing-a-data-science-career-framework-2d2c61e7f6a1>

Meng, X.-L. (2018). Conducting highly principled data science: A statistician's job and joy. *Statistics & Probability Letters*, 136, 51–57. <https://doi.org/10.1016/j.spl.2018.02.053>

NASA. (2014). *Prognostics center—Data repository*. <https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/>

Predikto. (2017). *Railroad and transit*. <https://www.predikto.com/industries/rail>

- Roemer, M., Byington, C., & Kacprzynski, G. (2006). An overview of selected prognostic technologies with application to engine health management. In *Proceedings of GT2006, ASME Turbo Expo 2006: Power for Land, Sea, and Air* (pp. 707–715). <https://doi.org/10.1115/GT2006-90677>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Saha, B., & Goebel, K. (2007). *Battery data set*. NASA Ames Prognostics Data Repository, NASA Ames Research Center. <http://ti.arc.nasa.gov/project/prognostic-data-repository>
- Saxena, A., & Goebel, K. (2008). *Turbofan engine degradation simulation data set*. NASA Ames Prognostics Data Repository, NASA Ames Research Center. <http://ti.arc.nasa.gov/project/prognostic-data-repository>
- Sikorska, J., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5), 1803–1836. <https://doi.org/10.1016/j.ymssp.2010.11.018>
- Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw-Hill Education.
- Wheeler, K. R., Kurtoglu, T., & Poll, S. D. (2009, August 30–September 2). A survey of health management user objectives related to diagnostic and prognostic metrics. In *ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. 1287–1298). American Society of Mechanical Engineers Digital Collection. <https://doi.org/10.1115/DETC2009-87073>
- Williams, B., & Tokar, T. (2008). A review of inventory management research in major logistics journals: Themes and future directions. *The International Journal of Logistics Management*, 19(2), 212–232. <https://doi.org/10.1108/09574090810895960>

©2020 Michael Horrell, Larry Reynolds, and Adam McElhinney. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.