Data-driven predictive maintenance and time-series applications

Kefalas, M.

Chapter 3

Prognostics and Health Management

This chapter[1] aims at introducing to the reader the fields of prognostics and health management (PHM) and predictive maintenance (PdM). We start by discussing the importance of the field through examples and we present definitions and terminology that is used. Following that, data-driven applications of PdM in the aerospace industry are discussed. The reason for the selection of this specific industry is based on the high level of criticality that maintenance has in it. As a result, *timely* maintenance is an integral and indispensable aspect of aerospace which is reflected in the strict safety regulations, high availability expectations by airlines and clients, as well as maintenance costs. Finally, this chapter is summarized and directions for future work are presented.

3.1 Introduction

At 11 : 03 Eastern Daylight Time (EDT) on April 17 2018, Southwest Airlines Flight 1380 from New York to Dallas, was at flight level (FL) 320 (an altitude of approximately 32,000 feet or 9.8 km) and climbing when it experienced a left engine failure. As a result, most of the engine inlet and parts of the cowling broke off. Fragments from the inlet and cowling struck the leading edge of the wing and fuselage, causing a rapid depressurization and the death of a passenger. After investigations, the reason was found to be a failure of a single fan blade, due to a fatigue crack [2]. A similar event took place on 30th September 2017 when Air France Flight 66 from Paris to Los Angeles suffered an uncontained engine failure and made an emergency landing at Goose Bay Airport, Canada. Post-flight investigations indicated that the engine's fan hub had detached and dragged the air inlet with it during the flight [93]. These examples are by no means exhaustive. They show, however, how essential maintenance is, especially in

[1]Contents of this chapter are based on [230]; Duc van Nguyen, Marios Kefalas, Kaifeng Yang, Asteris Apostolidis, Markus Olhofer, Steffen Limmer, and Thomas Bäck. A Review: Prognostics and Health Management in Automotive and Aerospace. International Journal of Prognostics and Health Management, 10(2):35, 2019. PHM Society.

the aerospace industry.
Maintenance comprises a large percentage of operating costs of industries. According to the annual reports published by the Royal Dutch Airlines (KLM)[2], the maintenance costs from 2017 to 2020 are 941, 947, 882 and 738 million euro, respectively. These correspond to about 14% to 21% of the operational costs (total external expenses). DHL has estimated that an AOG (Aircraft On Ground) due to technical reasons, for an A380 Airbus, costs as much as 925.000 euro per day[3]. In the worst cases, the consequent costs could not be fully evaluated if the equipment failure led to a bad accident.
An approach that has surfaced in recent years, in order to mitigate the large maintenance costs is that of Prognostics and Health Management (PHM). PHM goes beyond CBM, as correct predictions of the future may allow avoiding failure and other large disturbances [133]. PHM includes a set of methodologies and actions that aim at minimizing maintenance costs by the assessment, diagnosis, prognosis, and health management of engineered systems. This allows a shift from traditional maintenance strategies, such as reactive maintenance (RM) and preventive maintenance (PM) to what is known as *predictive maintenance (PdM)*. PdM estimates when maintenance should take place and thus, increases safety, maximizes usability by avoiding immature maintenance. As a consequence reduces operation and maintenance costs and mitigates logistic bottlenecks. With an increasing prevalence of smart sensing, the progress of artificial intelligence, and with more powerful computing and increased storage, PHM has been gaining popularity across a growing spectrum of industries such as aerospace, smart manufacturing, transportation, and power generation [57]. Regardless of the field, one common expectation of PHM is its capability to translate raw data into actionable information to facilitate maintenance decision making [109]. PHM is also referred to as system health management (SHM), integrated systems health management (ISHM), vehicle health management system (VHMS) or engine health management (EHM).
PHM systems are designed in such a way that they can detect incipient component or system faults and/or failures, perform failure diagnostics, failure prognostics, and general health management. Among the tasks that a PHM system is expected to perform failure prognostics is the most significant one and lies in the core of PHM. Failure prognostics refers specifically to the phase involved with predicting future behavior and the system's useful lifetime left in terms of current operating state and the scheduling of required maintenance actions to maintain system health [227]. The useful lifetime left is often called the 'Remaining Useful Life (RUL)'. RUL is typically a random variable and unknown, and as such it must be estimated from available sources of information such as the information obtained in condition and health monitoring [208][4].

[2]https://www.klm.nl/en/information/corporate/publications
[3]Source: Airbus China
[4]In this dissertation we will be using the terms RUL *prediction* and RUL *estimation* interchangeably, unless otherwise stated.

More formally, based on [21], the RUL can be defined as follows:

Definition 3.1 (Remaining Useful Life (RUL))**.** Let $t \in \mathbb{R}_{\geq 0}$ be an instance of time at which we predict the RUL of an asset. Then,

$$RUL(t, \mathbf{D}_t) = \inf\{s \in \mathbb{R}_{\geq 0} : s \geq t \wedge \mathbb{1}_{\mathbb{S}^\complement}(CI(s, \mathbf{D}_t))\} - t \ , \tag{3.1}$$

where *inf* represents the infimum of a set and $\mathbb{1}$ is the indicator function. $\mathbb{S}$ is a user-defined system operating envelope. The operating envelope, $\mathbb{S}$, is a collection of boundary limits, that when exceeded put the integrity of an asset at risk. *CI* represents a user-specified condition index, which monitors if the asset has exceeded it's operating constraints. In this case the *CI* lies in the complement of $\mathbb{S}$ ($\mathbb{S}^\complement$), which indicates that the system must be repaired or maintained. $\mathbf{D}_t$ represents the data generated by an asset used for the RUL prediction of that asset. Most commonly $\mathbf{D}_t$ is sensor measurements recorded in time (time-series e.g., pressure, temperature) accompanied by event labels (e.g., times-to-failure), up until time t. In principle though, $\mathbf{D}_t$ can be any type of data, structured or not, that can facilitate the estimation.

The quantity $\inf\{s \in \mathbb{R}_{\geq 0} : s \geq t \wedge \mathbb{1}_{\mathbb{S}^\complement}(CI(s, \mathbf{D}_t))\}$ in Equation 3.1 can also be referred to as the *end-of-life* (EoL), to mark that the system's "life", based on user-defined criteria, has come to an end. Ultimately the estimation of RUL amounts to the approximation of the EoL. We should note that the EoL does not necessarily mean that the system has gone through a catastrophic failure but might operate sub-optimally according to user-defined criteria.

Finally, from a *data-driven* perspective, the estimation of the RUL of an asset involves creating a model which is trained on data from the same type of assets. In the work presented in this dissertation, the data used are (multivariate) time-series (see also Definition 2.1). In more detail, for the RUL estimatiom, let U be the set of training data. Each instance $u \in U$ is presented as a multivariate time-series of sensor readings $\boldsymbol{X}_u = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{T(u)}]^T \in \mathbb{R}^{m \times T(u)}$, with $T(u)$ time-steps where the last time-step corresponds to the end-of-life (EoL) of the unit u. Each point $\boldsymbol{x}_t \in \mathbb{R}^m$ is an m-dimensional vector corresponding to readings from m sensors at time t.

The main implementation steps for PHM consist of; i) defining critical component(s), ii) appropriate sensor selection for condition monitoring, iii) prognostics feature evaluation under data analysis and iv) prognostics methodology and tool evaluation metrics [16].

In the next section, we will briefly present diagnostics and prognostics and discuss the relation between these two notions.

3.2 Diagnostics and Prognostics

Diagnostics and prognostics are related processes of assessment of a system's health. Diagnostics aims at detecting, isolating, and identifying a fault or failure, whereas prognostics is the process of prediction of future states or RUL estimation based on current and/or historic information of the monitored asset [227, 138]. A main distinguishing factor between them is that diagnostics aims to diagnose or detect a fault or failure after, momentarily or *shortly* before it takes place, whereas prognostics focuses on anticipating these, within an actionable horizon. Prognostics is based on the understanding that equipment fails after a period of degradation, which when estimated (e.g., RUL) can lead to actionable information (e.g., maintenance planning and logistics management).

In more detail, diagnostics is enabled through the collection of data and/or other information. The different stages of diagnostics can be explained as follows [227]. Fault detection determines whether an abnormal operating condition exists. If it is detected, then it is reported. Fault isolation locates the fault to a specific component, sub-component or system that is failing or has failed. Fault identification deals with what is called the root cause or the basic event of the fault or failure. Fault symptoms are signatures that allow identifying the possible faults or failures. Prognostics is by nature an even more challenging task compared to diagnostics. Inheriting its name from the Ancient Greek *progignoskein*, which means knowing in advance, its task, as previously stated, is the assessment of the future health of a system. More specifically, prognostics is a CBM estimation of the RUL in order to make better-informed maintenance decisions [138]. Even though RUL estimation lies at the core of prognostics, it should not be considered the same task. Besides the RUL prediction, a comprehensive prognostics framework should be able to quickly and efficiently isolate the root cause of failures. In this sense, if fault/failure predictions can be made, the allocation of replacement parts or refurbishment actions can be optimally scheduled to reduce the overall operational and maintenance logistic footprints.

Finally, before we proceed, it is important to disambiguate two terms, which we will be seeing a lot and are (often) a source of confusion. These are fault and failure. The former implies that a system under observation is still operational, but cannot continue operating without any maintenance action, otherwise, it will cease operating, resulting in a failure.

3.3 PHM approaches

Prognostic approaches are most commonly classified into four types [59], namely i) reliability-based approaches, ii) model-based approaches, iii) data-driven approaches, and iv) hybrid approaches. While all approaches have their advantages and limitations, model-based, data-driven, and hybrid approaches are the most prominent and modern. These approaches can reason about individual assets, whereas reliability-based approaches rely on collective knowledge from a fleet of similar items.

In the following subsections, we will briefly dive into the different prognostics approaches and present some of their representative methods. Figure 3.1 summarizes the range of possible

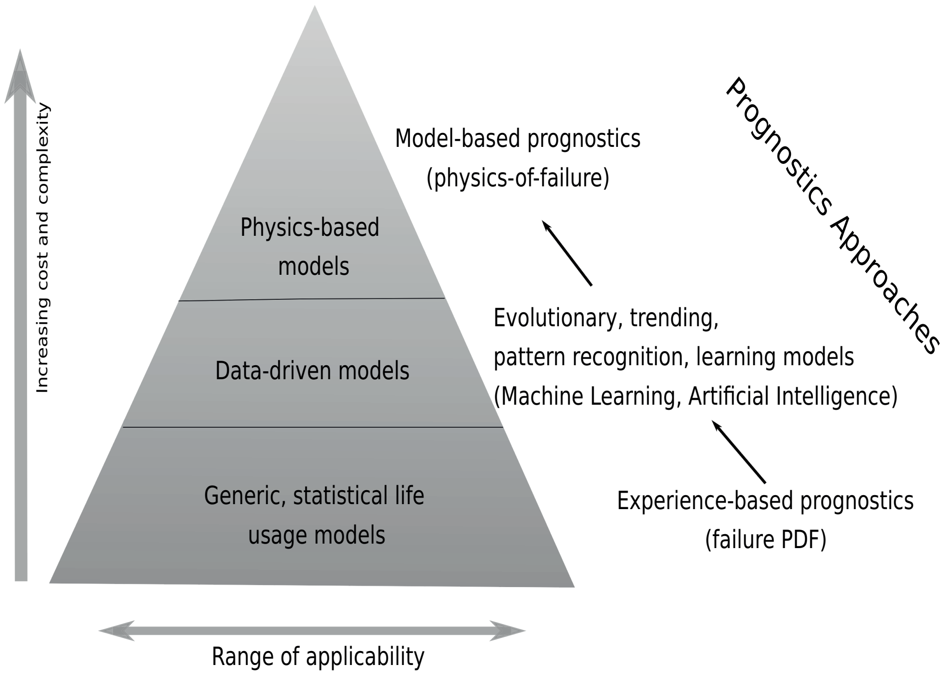

Figure 3.1: Prognostics approaches. Adapted from [227].

prognostics approaches and presents them as a function of applicability and implementation costs.

3.3.1 Reliability-based Approaches

Experienced-based prognostics, life usage model, statistical reliability-based approaches and probability-based prognostic techniques are all different terms describing a similar set of approaches in prognostics [59, 227]. Reliability-based approaches are some of the oldest and simplest forms of fault prognostics. They take into account the data and the knowledge that has accumulated by the experience during usage of industrial systems [155]. They require (massive) historical data and specifically, times-to-event records from a population of *identical* items. By "event" we mean failure or fault events or other significant maintenance events. Reliability-based approaches rely heavily on the assumption that the temporal information of these events follows specific distributions, which when fit can be used to infer times-to-event of a new, *identical*, asset. These distributions usually include but are not limited to, the exponential, the Weibull, the log-normal, and normal distributions and the Poisson distribution.
The advantages of such approaches in prognostics are the simplicity of their usage and the minimal need for domain knowledge. What is more, these statistical-based approaches can provide confidence intervals, which can be important in decision making, as they give a feeling

of accuracy and precision for the predictions [227, 216]. However, these methods require a large amount of historical repair and failure data in order to determine the parameters that faithfully model the life cycle of the system in question [155]. Consequently, such methods can be inaccurate when applied to newly developed assets, as reliability-based methods require (massive) historical repair and failure data, which in this case might be scarce [59, 155]. Furthermore, these approaches ignore component-specific conditions and do not take into account any indications (e.g., from sensor measurements) generated by the assets themselves in order to assess their health condition. Ignoring such (progressive) degradation phenomena can lead to premature or late maintenance [59, 83]. As a consequence, these prognostic methods can be used for scheduled or preventive maintenance (PM) and are used mainly for non-critical, components, that are not monitored and that are usually mass-produced.

3.3.2 Model-based Approaches

Model-based methods (or physics-based methods) utilize explicit physical models of the monitored components/systems for prognostics and RUL estimation. These physical models describe the degradation processes of the systems in question through mathematical models based on the failure mechanisms and first principles of damage [59, 138]. To establish this model, however, a thorough understanding of the system's physics is required [59]. For example, physics-based fatigue models have been extensively used to represent the start and propagation of structural anomalies [55]. These approaches are deterministic and allow for the estimation and the prediction of the dynamical states of the system in question. Moreover, these methods often use residuals as their features, where the residuals are defined as the results of consistency checks between the sensed measurements of a real system and the outputs of a mathematical model [55]. The premise here is that the residuals are large in the presence of failure/faults, and small or non-significant in the presence of normal disturbances (e.g., due to transient conditions), noise, and other modeling errors [55]. In this case, when the difference between the model and reality exceeds a user predefined threshold, an alert is generated. In this view, they can detect shifts from the nominal conditions of the underlying process when a simulation that is based on the model is executed in parallel to the real-time process. Of course, the latter demands a model that is developed based on the design-point (normal) conditions and which represents the ideal behavior of the system. Since these methods incorporate a physical understanding of the system, in many situations, these shifts are closely related to model parameters [8]. Furthermore, these models, do not require a large amount of data and are ideal when failure data do not exist or are scarce. What is more, physics-based approaches are very descriptive and interpretable, as the modeling relies on mathematical equations and established laws. This allows for such approaches to be efficiently validated and certified [59]. Additionally, these approaches can be used to simulate component failures for a better understanding of the system in question [8].

If the comprehension of the system's behavior together with the fidelity of the models is sufficient, these models allow for high accuracy and precision [59]. Moreover, if this understanding improves, the models can be adapted to increase their accuracy and address subtle performance problems [8, 55].

Model-based approaches, however, rely heavily on a thorough understanding of the system's physics and underlying processes, something which can be prohibitively costly in terms of time and money. As a result such a model's reliability often decreases as the system complexity increases, since there can exist underlying processes which have not been taken into account during model development. A direct consequence of this is that physics-based approaches can be successfully applied to those systems that a thorough understanding of their physics is possible, as well as of their failure modes and degradation behavior. Also, the developed physics models are usually component/system-specific and as a result, their reusability is very limited to other similar cases [59].

Examples of model-based approaches that are developed based on physical principles/laws are Kalman filters (KF) and their extensions. Namely, extended Kalman filters (EKF), unscented Kalman filters (UKF), and particle filters (PF).

Kalman filters (KF) were introduced as fault isolation and assessment technique for relative aircraft engine performance diagnostics in the late 1970s and early 1980s [210]. More widely used by engineers and other physical scientists, filtering problems are mathematical models for state estimation in signal processing and related domains. The main idea is to determine an estimate of some true value of a system from noisy and incomplete observations. Kalman filters or linear quadratic estimation as they are also known as take into account measurements recorded over time to make inferences about an unknown variable of interest (the state variable). Kalman filters work in a two-step process. In the first step, the prediction step, the Kalman filter produces an estimate of the current state, along with its probability distribution. Once the outcome of the next measurement is observed, the previously produced estimates are updated. It is a recursive procedure, which means that it only needs the present observations and the previously calculated state and its uncertainty matrix, to estimate the current state variable. The latter hands them the advantage of running in real-time.

Kalman filters (KF), however, are linear model-based estimators, which means that they assume linearity of the underlying dynamical system [157]. That is, they assume linearities in either the process model or the observation model, or both. Real-life systems can, however, be highly complex and as a result nonlinear. In order to overcome this assumption of KF and address the non-linearities, variants of KF have been created. More notable are the EKF and the UKF. The former assumes that the nonlinear functions are differentiable and linearizes about an estimate of the current mean and covariance [219]. The latter, instead, uses deterministic sampling to form a new mean and covariance estimate [219] with a sampling technique known as the unscented transform (UT) to determine a minimal set of sample points (sigma points) around the mean. UKF perform better than EKF when the prediction and update steps are highly nonlinear. This is a result of the linearization of the covariance.

The most popular model-based method is particle filters (PF) [41]. Their popularity arises from the fact that in contrast to KF, EKF, and UKF both linear and nonlinear state process and measurement models can be used. What is more, PF allow representing state estimates with arbitrarily shaped probability distributions. In more detail, PF methods or Sequential Monte Carlo (SMC) techniques are a class of algorithms that are used to approximate the optimal Bayesian filtering by representing the posterior probability density function (PDF) of the states discretely, with a population of particles with associated importance weights [50]. The particles are simply random samples from the unknown state space, representing possible realizations of the state sequences and the weights are the corresponding discrete probability masses [193]. As the filter iterates, the particles are propagated according to the system state transition model, while their weights are updated based upon the likelihoods of the measurement given the particle values [193]. For more details regarding PF, we refer the interested reader to [14].

3.3.3 Data-driven Approaches

In some instances, one might have only historical fault/failure or maintenance data leading up to a major maintenance event. Other times, the system in question cannot be adequately modeled due to its complexity. In such cases, one might use data-drive methods. These methods take their name from the fact that they rely (almost) entirely on data generated from the monitored asset(s). These data are performance parameters, such as but not limited to, pressure, temperature, speed, vibration, current, and acceleration [55]. These data or features derived from them are subsequently used to create an algorithmic model that correlates these measured parameters and/or features to the system health, degradation and fault progression, and RUL estimation [59].

As opposed to model-based approaches, data-driven methods can be developed and deployed much faster as they do not call for (a lot of) expert knowledge. The low cost of algorithm development and little knowledge required about the physics of the studied system makes this approach preferable by PHM practicioners [258] and applicable to a wider audience. However, data-driven methods depend largely on the size and quality of the acquired data. They require large amounts of historical data, something that is not always possible, especially for newly developed systems or other fielded applications, due to safety and data privacy concerns. In addition to that, there is also commonly a lack of a procedure to obtain the training data and there is further, a lack of run-to-failure data, for the methods to learn from. Due to this, applications in the literature usually make use of experimental data for model training, and thus these approaches may have wider confidence intervals than others [55]. Furthermore, this dependence on data dictates that the developed algorithms must also be robust to artifacts in the data (e.g., noise) [59, 55]. In principle, the more failure events are included in the data, the higher the accuracy of the estimation obtained. What is more, data-driven methods are more often than not “black box” approaches, in the sense that their internal decision making is not transparent [216]. As a consequence, the results of such a model are not always intuitive, due to the lack of physical knowledge about the system. This together with the fact that data-driven methods are based on approximations makes uncertainty quantification and management of the results an extra challenge.

Data-driven approaches mainly rely on techniques in the field of artificial intelligence (AI) and machine learning (ML). This includes, but is not limited to, decision trees (DT), random forests (RF), support vector machines (SVM), relevance vector machines (RVM), and deep neural networks (DNN). Often enough, statistical (parametric and non-parametric) approaches are also used to detect the presence of anomalies in the data [216] and can be considered as data-driven methods for prognostics. A non-exhaustive list of such techniques is multivariate statistical methods, partial least squares (PLS), signal analysis (e.g., Fast Fourier Transformation) hypothesis testing, analysis of variance (ANOVA), maximum - likelihood (ML) estimation, expectation-maximization (EM), Wilcoxon - Mann - Whitney test, Gaussian mixture models, and histogram - based approaches [216, 55].

For a more thorough overview on data-driven approaches, readers can find more details in [209, 224, 216, 55, 59, 227].

3.3.4 Hybrid Approaches

Fusion or hybrid-based prognostic methodologies combine the strengths of the model-based and data-driven approaches, to estimate the RUL under both operating and non-operating life cycle conditions [176]. By taking advantage of the two respective methods’ strengths, hybrid models can achieve robust health prediction results that can lead to a more reliable RUL approximation as compared to only model-based methods. Also, due to the use of a mathematical model, the amount of data required for training purposes are relatively lower than that needed in pure data-driven methods [49]. A hybrid model thus combines both data-driven methodologies with the knowledge of the system under study. It is a promising method, due to the fact that it can compensate for the lack of knowledge about the system’s physics and the lack of data [10, 87, 23].

3.4 PHM in the Aerospace Industry

Due to the high availability expectations from aircraft operators and clients and the high costs incurred for maintenance, when an aircraft is out of service [234] or Aircraft On Ground (AOG), as well as the supportability, testability, and reliability of modern aircraft [248], PHM systems play a significant role in the aerospace industry, from which it originated in the first place.

Nowadays, it is very challenging for the industry to keep its costs as low as possible and to generate maximum revenue, since the last decade has been turbulent for the aviation industry owing to the unprecedented rise in its commodities due to inflation [175], as well as due to the fluctuation in the price of fuel. Regarding the latter, IATA published that in 2017 the airline industry's estimated fuel bill reached 149 billion USD, more than 3 times the figure of 2003 (estimated at 44 billion USD) [99]. The industry has to ensure that its asset utilization is optimum and therefore, the maintenance management system of the existing aircraft needs to be precise to ensure that the aircraft spends maximum time in the air to make the best use of its machinery [175]. This is because maintenance is extremely expensive, mainly due to the price of spare parts. As a result, one wants to maximize the use and exploit the remaining life of the installed parts, keeping them in operation by maintaining and repairing them until they exhaust their life limit and need to be replaced. Apart from safety and costs-saving, this also enhances sustainability. This is the role of PHM; to make sure that this happens and that no part is exchanged prematurely. The notice of pending equipment failure allows for sufficient lead-time so that necessary personnel, equipment, and spare parts can be organized and deployed, thus minimizing both equipment downtime and repair costs [204], and optimizing maintenance. Integration is one of the trends of PHM systems, which means that PHM systems of the engine and other aircraft parts are integrated with aircraft PHM system [180]. To the best of our knowledge, however, there is no generic PHM framework and architecture enabling communication and integration with the various contributing systems [140], as well as no uniform design framework of aviation PHM systems between countries [248] and even between carriers/operators. In addition, a systematic method has yet to be established for developing and deploying a PHM system, as the current ones are application or equipment specific [134].

Among all the frameworks the most mature system is that of the F35 aircraft, which constitutes the double-deck architecture. Using this multilayered framework, the system integrates the airplane airborne information and sends the necessary information to the ground controls. This integrated health management system determines the safety of the aircraft and allows for the state management and maintenance guarantee [141]. Another predictive maintenance system, for a wide range of helicopters flown by the military (rotocrafts), is called HUMS (Health and Usage Monitoring Systems), developed by UTC Aerospace Systems. This system can detect several different types of issues using vibration analysis, ranging from shaft unbalance to gear and bearing deterioration. In civil aviation, the typical representatives are the Airplane Health Management (AHM) system of Boeing [248], the AIRcraft Maintenance Analysis (AIRMAN) system of Airbus, and a more recent addition, namely, aircraft real-time health monitoring system (AiRTHM) [248]. For more detailed information on these specific systems we refer the interested reader to [243] (Boeing) and [92], [56], [102] (Airbus).

There is also a lack of standards for PHM system development, data collection and analysis

methods, and data management, although the PHM4SMS (Prognostics and Health Management for Smart Manufacturing Systems) of NIST (National Institute of Standards and Technology) serves in designing such standards [235]. Particularly, in the aircraft industry, the published standard for the guidance for PHM systems development is MSG-3, developed by the Maintenance Steering Group (MSG) of the Air Transport Association (ATA) and is titled "Operator / Manufacturer Scheduled Maintenance Development". It is used for developing maintenance plans for aircraft, engines, and systems (Air Transport Association of America, 2013) before the aircraft is in service and it also helps in improving safety while at the same time reducing unnecessary maintenance tasks [235].
This chapter is intended to familiarize the reader with the PHM systems in the aerospace industry, by introducing concepts, presenting examples, and discussing research opportunities.

3.4.1 Classification of Sensors of the Gas Turbofan Engine

Here, we briefly classify the most common and informative measurements of a turbofan engine. An exhaustive list of sensor measurements of the entire airframe and of the stations of a turbofan engine is out of the scope of this chapter. The authors decided to emphasize the turbofan engine alone, due to the fact that it is the core of the aircraft and one of the most, if not the most, expensive assets of the airframe. Furthermore, this is a starting point for researchers in the quest for informative measurements. In the rest, we classify them by type and by function.
In Table 3.1, we provide a classification of the most common turbofan sensors based on their type, and in Table 3.2 we present a classification based on their application. We should note here that in Tables 3.1 and 3.2 N3, which is the speed in 3-spool turbofan engines (e.g., Rolls-Royce), is not applicable to all engines.

3.4.2 PHM Methods in the Aerospace Industry

In this section, we will give an overview of various PHM methods used in the aerospace industry. To be more specific, as stated in the introduction, CBM systems are founded upon the ability to infer equipment conditions using data collected from sensors on monitored systems. In aerospace, these systems could be engines, thrust reversers, avionics, flight controls, fly-by-wire, landing gear, braking, environmental control systems (ECS), electrical systems, and auxiliary power units, to name a few. For each system, there are also numerous sensors, which reflect their components' state and the overall system health. For example, the current Airbus A350 model has a total of around 6,000 sensors across the entire plane and this number will increase as big data analytics software and broadband links become more affordable [206].
In the following sections, we discuss prognostic and diagnostic methods used in aviation as they are crucial for safety, customer satisfaction, and airline revenue. We will emphasize more on prognostic applications in the industry, as this type of predictive analytics is common across

Table 3.1: Turbofan sensors classified by type.

Types	Sensors
Temperature	Oil temperature, Total air/gas temperature Static air/gas temperature, Nacelle temperature, Exhaust gas temperature (EGT)
Vibration	Core vibration, Fan vibration, Core phase angle, Fan phase angle
Pressure	Total air/gas pressure, Static air/gas pressure, Oil pressure
Spoll Speed	Core speed (N2), Fan speed (N1), N3
Miscellaneous	Fuel flow, Oil quantity, Altitude, Mach number, Variable bleed valve (VBV) position, Nacelle Anti-ice, Wing Anti-ice, Variable stator blades (VSV) position

Table 3.2: Turbofan sensors classified by application.

Functions	Sensors
Gas Path	Total air/gas pressure, Static air/gas pressure, Total air/gas temperature, Static air/gas temperature
Engine Oil	Oil temperature, Oil pressure, Oil quantity
Engine Balance	Core vibration, Fan vibration, Core phase angle, Fan phase angle
Stalling/Surging	VBV position, Wing anti-ice, Nacelle anti-ice, VSV position
Thrust Setting	Engine pressure ratio (EPR), Fan speed (N1), Core speed (N2), N3, Fuel flow
Exhaust	Exhaust gas temperature (EGT)
Flight Envelope	Altitude, Mach number

all fields of industry, but is particularly valuable in commercial aviation. In addition, as mentioned previously, diagnostics are included in prognostics and thus, we can consider prognostics as a natural extension of diagnostics. After all, one needs the latter to find the former [209]. Thus, we can consider the term prognostics to have a broader definition and enclose activities such as supervising, monitoring, detect and determining initial degradation, as well as making fault/failure predictions. Finally, in the following subsection, we will discuss *only* the data-driven methods used in PHM in aerospace, as these are the main topic of this dissertation. The reviewed literature is by no means exhaustive, but it serves as proof for the big appeal and interest for data-driven solutions in PHM in general, predictive maintenance, and RUL estimation. For a more thorough overview of other approaches (e.g., model-based), including applications in the automotive industry, we refer the interested reader to our original publication [230].

Applications of Prognostics Data-driven Methods in the Aerospace Industry

Neural Networks Neural networks allow the investigation of complex systems without the need for any knowledge or assumption about system structure. They are sophisticated modeling techniques capable of modeling problems that are analytically and inherently difficult and for which conventional approaches are not practical, including complex physical processes with nonlinear, high-order, and time-varying dynamics [8]. Recently, in [257], Zhang et al., designed a back-propagation, feedforward neural network to assess the starter degradation of the APU using its gas-path measurements. Feedforward NNs are the simplest form of artificial neural networks where information moves in only one direction from input nodes to output nodes. In a recent paper by Ma et al. [152] the authors proposed an effective deep learning method, termed stacked denoising autoencoder (SDA), for health state classification of aircraft engines considering the environmental noise. SDA proved to be effective in terms of cognitive computing and pattern classification theory. Furthermore, the proposed method beats its rivals, in terms of feature extraction due to the benefits of its deep architecture with a data destruction process that is effective for robust feature representation, where high-order features and shared representations can be learnt from the input samples by unsupervised self-learning. The feasibility of the proposed method was demonstrated using the 2008 PHM challenge datasets (see [183]). In [264], Ke-Xu et al. designed a particle-swarm optimized NN for spacecraft prognostics.

Other types of NNs that have gained popularity in the field of PHM are recurrent neural networks (RNNs). RNNs, developed in the 80s, are a class of NNs that capture time dynamics. RNNs and their variants, namely the long short term memory (LSTM) and the gated recurrent unit (GRU) networks differ from the traditional feedforward NNs, in that they can process information across time, making them ideal for sequential data, such as time-series. Specifically, due to their internal state (memory), they can process a sequence of inputs, granting them the ability to model temporal dependencies and are thus suited for tasks in which input and/or output consist of sequences of points that are not independent. For a more thorough understanding of RNNs and their variants, we urge the interested reader to [145, 98]. In this direction, Zhong et al. [261] designed a gated recurrent neural network (GRU network) to predict the exhaust gas temperature (EGT) of a turbofan aero-engine. The temperature of the exhaust gases of an engine has evolved to become the standard industrial indicator of the health of an aircraft engine [236]. This is because it can capture the cumulative effect of deterioration in the isentropic efficiency of gas path components. Their method could address the time-series and nonlinear characteristics simultaneously by the GRU blocks. The proposed algorithm was compared to five other single prognostic methods, namely, an artificial NN (ANN), support vector regression (SVR), extreme learning machine (ELM), and ensemble prognostic methods random forests-based ELM (RF-ELM) and average aggregation ELM (Avg-ELM). The proposed method achieved the best prediction accuracy and acceptable prediction stability. In [233], Vatani et al. predicted the degradation trends of a gas turbine engine by studying their effects on sensored data (i.e. temperature) by using an RNN as a first approach, as well as a nonlinear autoregressive model with exogenous input (NARX) neural network architecture. In [260] and [94], the authors developed an LSTM network for the estimation of RUL. In a similar manner, the method proposed in [245] uses an LSTM and proposes a dynamic differential technology to extract inter-frame information to cope with complex operating conditions.

Another type of NN, namely the convolutional neural network (CNN) has also gained recognition in the field of prognostics. CNNs generally differ from RNNs in that they are designed to effectively process spatial data. They are also very often used in the analysis of visual imagery, that exploits the local dependencies of visual information [145]. For a more thorough understanding of CNNs and their mathematical formulation, we direct the reader to [244, 98]. In [143], Li et al. use a deep convolution NN (DCNN) for estimating the RUL and they demonstrate the effectiveness of their method using the C-MAPSS dataset [198] for aero-engine unit prognostics. In [195], the authors present the first attempt for estimating the RUL using CNN-based regression. The deep architecture allows the network to learn features that provide a higher-level abstract representation of low-level sensor signals, by employing the convolution and pooling layers to capture the salient patterns of the sensor signals at different time scales. However, considering that the collected machinery features are usually from different sensors, the relationship between the spatially neighboring features is not significant. In [143], Li et al. address this issue by proposing to use 1-dimensional convolution filters in their CNN. In [242], Wen et al. propose a CNN with an added Residual Building Block (RBB), in order to tackle the vanishing/exploding gradient problem in artificial neural networks with gradient-based learning methods and backpropagation. Zhang et al. [256] investigated the use of CNN with an extended time window to tackle the RUL estimation problem under varying operating conditions. Furthermore, to improve the prognostic robustness and avoid the sensitivity to the abnormal data, CNN and extreme gradient boosting (XGB) are fused with model averaging (CNN-XGB). NNs might be powerful, however, they do not take into account uncertainty bounds arising from different sources like process noise, measurement noise, and an inaccurate process model. Uncertainty quantification is useful in prognostics as it gives an estimate of confidence on the prediction of the RUL or general health estimation of an asset. This in turn can help avoid overly confident decisions and further allows the end-user or decision-maker to make a better-informed choice.

In contrast to NNs, relevance vector machines (RVM) and Gaussian process (GP) regression take into account the width of the uncertainty bounds in addition to providing damage trajectories [82]. RVM [222] is a Bayesian formalism representing a generalized linear model of the identical functional form of the support vector machine (SVM). Although SVM [232] is a state-of-the-art technique for classification and regression, RVM is able to generate probabilistic outputs in a Bayesian framework that make more sense in RUL estimation applications and furthermore uses a lot of kernel functions for comparable generalization performance [82]. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. The distribution of a GP is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g., time, or space. In [82], the authors evaluate the NN-based approach, RVM and GPR for their prognostic capabilities on a test stand involving rotating equipment in an aerospace setting. In the paper, however, there is no clear winner, since each of the algorithms came up with its current state estimates which were not close to each other. The conclusion states that even though these algorithms can learn the dynamics of the process from sparse and noisy data fairly well, the RUL estimates depend significantly on the current state estimation.

Time-Series Analysis Other approaches used are methods from time-series analysis. The autoregressive moving average (ARMA) model forms a class of general linear models used in modeling and forecasting of time-series. It is comprised of two parts, namely one for the autoregression (AR) and the second for the moving average (MA). It is a powerful forecasting methodology that is able to capture trends found in a time-series and projects its future values. In a recent paper by Baptista et al. [22], the authors integrate the ARMA methodology with data-driven techniques, to predict fault events on a real industrial case of unscheduled removals of the engine bleed valve (EBV), based only on life-usage data (maintenance event data). EBV is used in most designs as a regulator for the flow that goes to the ECS and the anti-icing systems of the aircraft. The authors proposed a method in which they feed the entire past fault event history into the ARMA model and the output is then used as a feature that integrates with the data-driven model. The data-driven modeling gives further insight into the forecasting outcome from ARMA and improves its accuracy and efficiency. From the data-driven methods they used, in addition to ARMA (NN, k-nearest neighbors (KNN), random forest (RF), support vector regression (SVR), generalized linear regression (GLM)) the SVM produced the best overall results. In a similar manner, Su et al. used in [212] least squares support vector regression with sliding ARMA forecasting to model the nonlinear time-series. They demonstrate their method on a practical case study for the US-made F-16 fighter.

However, ARMA models are applied in cases where data show evidence of a stationary stochastic process. This means that the time-series' statistical properties are all constant over time. A stationary series has no trend. That is, its variations around its mean have a constant amplitude, and it "wiggles" in a consistent fashion, i.e., its autocorrelations remain constant over time. Equivalently, short-term random time patterns always look the same in a statistical sense. If the contrary stands, its generalization, the Autoregressive Integrated Moving Average (ARIMA) model can be adopted. The "Integrated" indicates that the data values have been replaced with the difference between their values and the previous values, to transform the time-series to a stationary one. In a recent paper by Ordóñez et al. [170], the authors combine time-series analysis methods (ARIMA) to forecast the values of the predictor variables with machine learning techniques to predict the RUL of aircraft engines for more than one period ahead of those variables.

Graphical Models Another important category of data-driven models used are graphical models, which denote the conditional independence structure between random variables [41]. In a recent paper, [20], Banghart et al. utilize Bayesian networks (BN) to estimate the risk of the landing gear system, cockpit warning/caution annunciator panel, and the environmental control system turbine assembly of the Northrop Grumman EA-6B Prowler military aircraft. BN is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Nodes represent variables, while arcs represent probabilistic relationships. For example, engine blade damage impacts non-mission-capable time, thus an edge/arc is drawn from the respective nodes. It is a combination of graph theory and probability theory. It is a representation of a joint probability distribution defined on a finite set of random variables that can be discrete or continuous. From a knowledge modeling standpoint, Bayesian networks can be seen as a special knowledge representation system. The advantage of BN lies in the fact that it does not rely the on explicit understanding of causal connections within the system(s) under observation, nor the identification of sequences of events leading to failure. Furthermore, given their probabilistic nature, BNs prove to be a suitable technique to address the inherent uncertainty of RUL estimation. In the same view, Ferreiro et al. in [67] use BN as a predicting technique and demonstrate their effectiveness by representing a physical model for aircraft brake wear, originally developed by British Aerospace Systems. They fit it to the available data (aircraft weight, landing velocity, brake operation during landing, flap position, and initial brake temperature) from flight conditions extracted from the operational plan of the aircraft. Although in this example the causal connections are based on understanding a physical system, the general idea is that BN can be successfully used in prognosis also, instead of diagnosis. A subclass of BN is the so-called dynamic Bayesian networks (DBN), which relate variables to each other, over adjacent time steps. They can be considered simply as BN for the modeling of time-series data [79]. A specification of DBN is the hidden Markov models (HMM), which have been applied to prognostic problems in aviation.

Hidden Markov Models HMM is a stochastic process model, characterized by a doubly embedded stochastic process with an underlying hidden stochastic process that can be observed through some probabilistic behavior. The latter justifies the word "hidden". It is also a powerful tool for RUL estimation. HMM is furthermore a parametric model with some distinct characteristics: it can *not* only reflect the randomness of machine behavior (i.e., sensor measurements) but also reveal hidden states and changing processes [8]. For a more thorough understanding, we direct the reader to [73]. In this view, in [24], the authors investigate the use of a Hidden semi-Markov model (HSMM) to predict the RUL of the shaft of utility helicopters until failure. The difference between an HSMM and an HMM is that the latter assumes that the sojourn time in the (hidden) state process follows a geometric distribution (most likely with parameter 1) in the discrete case and an exponential distribution. In contrast to that in semi-Markov processes, an upcoming transition's distribution is described by a product of an arbitrary PDF for the waiting time and a categorical distribution for the next state. The arbitrary condition for the PDF removes the memorylessness property of the process and as such, the process is Markovian only at the specified jump instants. Dong et al. in [54], proposed a HSMM for fault classification application for UH-60A Blackhawk main transmission planetary carriers and prognosis of a hydraulic pump health monitoring application. They compare HSMM with HMM and conclude that the former is capable of identifying the faults under both test cell and on-aircraft conditions while the performance of the HMM is not comparable with that of the HSMM. At the same time, the HSMM-based methodology can be used to estimate the RUL of equipment. However, HMM have some inherent limitations. One is the assumption that successive system behavior observations are independent and the other is that the Markov assumption that the probability in a given state at time t only depends on the state at time $t-1$ is clearly untenable in practical applications [8].
In the same context, time-series analysis methods, which we referred to before, have been combined with HMM. Specifically, AR models have been combined with HMM, in what is called an autoregressive hidden Markov model (ARHMM) [179], initially proposed for speech recognition. Here the observations are drawn from an autoregression process (linear prediction) [181]. In this view, Juesas et al. [107] developed a variant of ARHMM, named autoregressive partially-hidden Markov model (ARPHMM) for fault detection and prognostics of equipment based on sensors' data. The authors considered a modification of the learning procedure of the ARHMM, by integrating prior knowledge on latent variables. Their method was demonstrated on an instance of the C-MAPSS dataset [198]. They compared their approach, on the aforementioned

dataset, against RULCLIPPER (Remaining Useful Life estimation based on impreCise heaLth Indicator modeled by Planar Polygons and similarity-basEd Reasoning) [183], and SWELM (Summation Wavelet Extreme Learning Machine) [104]. The results are promising, in the sense that they are comparable to RULCLIPPER, and, on average, better than SWELM. It is interesting, however, as the authors point out, that using an ensemble between ARPHMM and SWELM outperforms RULCLIPPER, showing a direction towards developing ensemble approaches made of complementary and advanced prognostics algorithms.

Neuro-Fuzzy Systems Finally, a method with which we would like to conclude this section is the use of Neuro-Fuzzy systems (NF) for prognostics. NF systems are neural-network-based fuzzy systems, with the latter being a nonlinear mapping of an input data vector with a scalar output. Fuzzy logic is based on Zadeh's fuzzy set theory [111]. For a better understanding of fuzzy logic and neuro-fuzzy systems, we refer the reader to [189] and [5], respectively. In [40], the authors propose an integrated adaptive neuro-fuzzy inference systems (ANFIS) and high-order particle filtering, which forecasts the time evolution of the fault indication and estimates the probability density function of RUL. The ANFIS is used to model the fault propagation trend and the high-order particle filtering integrates the ANFIS, as an m-th-order hidden Markov model, to carry out long-term predictions and estimate the RUL PDF via a set of particles with associated weights. They apply their method on vibration data from the main gearbox of a UH-60 helicopter subjected to a seeded carrier plate crack fault and show that its prediction accuracy is higher than that of both the conventional ANFIS predictor and the particle-filter-based predictor where the fault growth model is a first-order model that is trained via the ANFIS.

3.5 Uncertainties in Prognostics

Before concluding, we must comment on uncertainty quantification (UQ) and management, as this is an indispensable part of PHM. Accounting for uncertainties is of paramount significance in prognostics. Uncertainties arise from various sources such as modeling uncertainties, measurement uncertainties, operating environment uncertainties, future load uncertainties, input data uncertainties. Such information is crucial for any prognostic estimate, otherwise, the prognostic results might be of limited use and cannot be incorporated in mission-critical applications. By accounting for the uncertainties the researcher or end-user can determine if, for example, the training data is not representative of the task, are too noisy, or if the selected model is poorly trained (i.e., underparameterized NN). The reason for this is that the single point estimates that we described, assume a deterministic algorithm or additional reasoning. Due to all the sources of uncertainty though, it is crucial that there must be confidence around the prediction. There are numerous ways for this, such as probability distributions of the RUL instead of a single-point RUL estimate. In [197] and [196], the authors discuss in a very concise and detailed manner the uncertainty issues and propose solutions by modifying PHM metrics and recommend suitable ways of graphically representing these metrics.

3.6 Discussions and Conclusions

Prognostics and health management (PHM) is a fairly new discipline that goes beyond condition-based maintenance (CBM) by predicting the future (health) states of an asset and estimating its remaining useful life (RUL). This process provides actionable information, enabling intelligent decision-making for improved performance, safety, reliability, and maintainability of engineered systems. This is achieved by using real-time and historical state information of subsystems and/or their components. Aside from the aforementioned, the significance of PHM lies in that these early warnings grant the user the horizon to design a timely maintenance schedule and follow all the required procedures for the logistics. In this chapter, we introduced PHM, presented the notion of RUL, and gave detailed explanations of the four prevailing PHM approaches, namely reliability-based, model-based, data-driven, and hybrid approaches. In addition, we briefly considered the significance of uncertainty quantification in prognostics. We, further, emphasized data-driven approaches in the aerospace industry where we gave an overview of recent tools and methods that have been used in the field.

The overview shows that methods other than traditional time-series analysis are gaining popularity in the data-driven prognostics in aerospace, such as graphical methods and NNs. The latter, specifically, are emerging as they propel data-driven solutions by not requiring (a lot of) engineering knowledge and more importantly by alleviating the need for explicit feature (predictor/parameter) construction. In addition to that, NNs lend themselves naturally to a multitude of sophisticated and automated methods for hyperparameter optimization, reducing the need for manual tuning. Moreover, NNs have the ability to model complex, highly non-linear systems without the need for any knowledge or assumption about system structure. Despite, however, the recent overall data-driven success in prognostics, in general, but also in aerospace there are still challenges and practical issues that need to be addressed. Below, we present some of these needs.

Need for securely obtaining more data. In detail, even though data-driven methods have been developed to counter the increasing complexity of systems and components, there is still no standard procedure to obtain data, in terms of a protocol or system. Data are either not integrated centrally, but scattered around different systems, or cannot be disclosed due to security and privacy issues and competition. This means that future work should not only emphasize algorithmic performance but also data quality and the drawing up of certain conventions per industrial field that govern data quality.

Regarding the issue of data sovereignty, it is important that future research takes into account and builds on approaches of collaborative learning, such as federated learning (FL) [6]. FL serves two purposes. On one side it allows for data augmentation by providing data from different data owners. In aviation this is significant and beneficial as a predictive maintenance model will not be trained on data from only one OEM or airline but from multiple, thus enabling multimodal learning. After all, a model trained on data from a turbofan engine in Northern Europe will exhibit different degradation patterns to a model trained on data from a turbofan engine in the Middle East, where high temperatures add more to the stress of the engine. On the other side, FL allows for data protection by training on the data from each party involved and simply aggregating and updating the weights/parameters of a *global* model. Since only partial model weights are shared with the global model from each party involved, privacy can be preserved and, this way, the data is less exposed to model inversion [142].

On top of that, there is also the possibility that data do not even exist due to the underlying cost of acquiring them, as for example run-to-failure data of a turbofan engine or other expensive asset. Future research should, thus, emphasize generating data when it is not available (e.g. through high fidelity simulations), taking into account that in field applications theoretical predictions and methods developed must be verified and validated first before practical applications become possible.

Need for real-time prognostics For field applications, another crucial challenge is the real-time (online) RUL estimation. The issue lies in the fact that the developed methods need intensive computational resources, which is in direct contradiction with hardware conditions of onboard computers, such as on cars and aircraft. Future directions should, therefore, investigate more in this direction, such as in edge computing [6]. For example, *safety-critical* computations, such as reliability and health prognostics, could take place on the edge, as edge computing allows for low latency since the data are processed closer to their source allowing thus, for accelerated insights. Furthermore, edge computing can increase model accuracy, especially in fields where the network bandwidth is too low or expensive, such as in aviation. Such issues are typically mitigated by reducing the size of data used in a (predictive) model. This results in information loss that could have otherwise been useful. When deployed at the edge, for example, data feedback loops can be used to improve AI model accuracy and multiple models can be run in parallel [249]. On the other hand, computations of *operations-critical* applications that deem no immediate result (e.g., fuel/energy consumption) can take place offline.

Need for explainable and interpretable data-driven PdM Data-driven methods for PdM, such as NNs are by construction "black box". This term refers to processes which lack interpretability of their internal workings and can be viewed only in terms of their inputs and outputs. This means that these models do not explain their predictions/outputs in a way that is understandable by humans, and as a result, this lack of transparency and accountability can have severe consequences [191], especially in operations-critical, or safety-critical systems, like aerospace. Interpretability of PdM methods can assist decision-makers in asserting the feasibility of the model logic, as well as in the troubleshooting of the developed methods, thus putting confidence in the process. This is why future research should invest in explainable PdM.

Need for uncertainty quantification (UQ) Finally, another topic that is of great importance in PdM is UQ. Uncertainties arise from various sources such as modeling uncertainties, and input data uncertainties. Quantifying uncertainty is crucial for any prognostic estimate, otherwise, it is of limited use and cannot be incorporated in safety-critical or operations-critical applications. By accounting for the uncertainties the researcher or end-user can determine if, for example, the training data is not representative of the task or too noisy (i.e., measurement uncertainties, operating environment uncertainties, future load uncertainties, input data uncertainties) or if the selected model is poorly selected (i.e., underparameterized NN). Especially when it comes to NNs, UQ is a rising topic of interest, given the fact that NNs are being industrially employed. Although in recent years there has been work done on UQ in NNs and some work on UQ in PdM, the field is still young with no consensus on how to measure this uncertainty. Furthermore, while the majority of model-based prognostic methods quantify the associated uncertainty, only a few studies in the data-driven domain address this matter, despite its importance [26]. Therefore, we recommend this as an important and exciting direction for further research in data-driven PdM.