Using Digital Twin Technology to Ensure Data Quality in Transport Systems

Björkqvist, Jerker; Manngård, Mikael; Lund, Wictor

Published in:
Proceedings of TRA2020, the 8th Transport Research Arena: Rethinking transport – towards clean and inclusive mobility

Published: 01/01/2020

Document Version
Accepted author manuscript

Document License
Publisher rights policy

Rethinking transport
27–30 April 2020

TRA 2020 TRANSPORT RESEARCH ARENA HELSINKI

Proceedings of 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland

Using Digital Twin Technology to Ensure Data Quality in Transport Systems

Mikael Manngård, Wictor Lund, Jerker Björkqvist

Åbo Akademi University, Domkyrkotorget 3, 20500 Turku, FINLAND

Abstract

Autonomous transport systems are dependent on real-time data from their environment. The data is read from a multitude of sensors ranging from sensors measuring temperature, pressure, and flow to advanced sensor systems such as GPS and Lidar systems. When the sensor data is used as input to an automation system, measurement errors and sensor faults may lead to invalid situation awareness or even catastrophic failure.

Digital Twins (DTs) are digital representations of a physical systems, where the physical rules of a system are expressed in a mathematical form. When feeding sensor data to the DT, sensor values can be checked against the rules of the physical system. This enables feasibility checking of sensor values, and together with sensor redundancy, faulty measurements can be corrected.
In this paper, we present the DT for a marine vessel energy subsystem and use the DT for validating and cleansing of representative data for that subsystem.

Keywords: Digital Twin, Data Quality, Energy Systems, Heat Exchanger

1. Introduction

In recent years we have seen an exponential growth in the number of sensors applied to the environment around us. The sensors are connected to the internet of things (IoT), the sensor data is stored in data centers (Big Data) and the data can be processed centrally (Cloud Computing) in order to gain new insights. The market value of internet giants such as Google and Facebook come from the fact that they have access to large data sets which are gathered from social media and from mobile devices. The spread of the IoT to new areas will generate new data sets which can give value to new markets.

Data from sensors attached to physical devices and systems – factories, production lines, work machines, cars, cranes, ships – has since long been used for supervisory control applications such as SCADA systems. However, with the emerging sensor technology, the connectivity and cost developments has made it possible to install and collect data in a way that was not possible previously. In 2018, the number of IoT sensors was estimated to 7 billion and for 2025 the predication is 22 billion sensors (Lueth, 2018). With these numbers of sensors and data availability, a new market is emerging for using the data to increase situation awareness and to optimize operations.

When sensor data from the IoT is used as input to algorithms, the outcome of the algorithm is dependent on the data quality. If the data input is wrong, in most cases, any conclusions from the output of the algorithm is most

likely also wrong. The unfortunate recent example is the Boing 737 Max passenger jet, where invalid sensor data was input to the "Maneuvering Characteristics Augmentation System" (MCAS) software system (Travis, 2019), leading to the crash of two brand-new passenger jets. This is an extreme example of what happens when invalid data is fed to an algorithm that performs actions on the data. Invalid sensor data can also have less immediately catastrophic effects such as confusion and mistrust in digital support systems.

When illegal sensor data is inspected by a human, in many cases, they can conclude that there is something wrong with the data. This is because we have a picture of the total system and an understanding of where the sensor is installed. Based on the full picture, we can decide what kind of data should be generated by the sensor. What if these capabilities of humans to understand and validate sensor data based on system understanding could be copied to the digital world?

A Digital Twin is a model of the physical system in a digital form, i.e. a model which can be directly used by computer to perform calculations. The Digital Twin is a virtual replica of a physical entity which enables a live digital replica driven by data from the physical asset. In the paper by (Parris, et al., 2016), the authors present how the company General Electrics (GE) utilize Digital Twins in their operations. The concept of Digital Twin can be used in many ways such as anomaly detection, optimal operation and detailed monitoring of physical systems that cannot be directly observable.

In this paper, the objective is to show how a Digital Twin on the core energy system of a ship can be used to validate data from sensors installed in the energy system. The core energy system model used in this case study consist of the diesel engine, which is the energy source, the PTI/PTO, the propulsion system, the generator/electrical motor, the core electrical grid and the load generated by the passengers. The Digital Twin is a model of this energy system, based on the physical rules that apply to the energy conversions that take place. As the sensor values are used for controlling the operation of the ship energy system, it is essential for correct and energy efficient operation that the sensor readings are reliable. The applied Digital Twin technology enables validation and cleansing of sensor data.

2. Related work

Data quality is essential for good decision making and analysis. There are vast databases available, for instance business data, economic data, environmental data and nowadays of course data gathered by Internet companies. When it comes to data quality, we want to ensure that the data is useful for the next processing steps, and that it represents the phenomena we want to observe with enough precision and reliability.

The term *data quality* is often used in the context knowledge discovery in databases (KDD). There are plenty of methodologies for ensuring and improving data quality in offline data sets, in general (Batini, et al., 2009), for time series (Esling & Agon, 2012), and in the context of the IoT (Chen, et al., 2015)

In (Sharma, et al., 2010) the authors analyze how to ensure sensor time series data by dividing faults in the categories Constant, Short and Noise. To detect the faults, rules based, estimation based and learning based methods are used.

Sensor fusion is used in many sensor systems where there are many sensors measuring weak or incomplete signals. For example, multiple sensors with different characteristics can be merged to enable good enough positional awareness to perform navigation tasks (Kam, et al., 1997). Sensor fusion resembles much the idea of Digital Twin but is usually not using the multitude of sensor values that can be used simultaneously today.

Machine learning, including neural networks have been used for data quality assessment. In (Rahman, et al., 2014) the authors use classifiers, trained using classified source data provided by domain experts to identify raw time series data in four classes from "good" to "bad". In (Napolitano, et al., 1998) the authors elaborate of having a neural network implemented directly on microprocessor to do sensor validation on a Grumman F-14 fighter jet.

3. Process description

To demonstrate our Digital Twin based methodology, we will first create a Digital Twin using a mathematical

model of the cooling system of a marine vessel energy system. The main purpose of a cooling system is to extract heat from the engines, the lubrication oil and the charge air. Although internal combustion engines used for marine applications have one of the highest thermal efficiency among combustions engines, only about 50% of the fuel energy is converted into mechanical power (Zou, et al., 2013). Thus, to improve the energy economy, as much as possible of the remaining waste heat should be extracted via heat exchangers to be utilized for other energy consuming tasks on-board.

A typical cooling system for a marine internal combustion is illustrated in Fig.1. The system consists of a high temperature (HT) circuit and a low temperature (LT) circuit. The main purpose of the HT circuit is to cool the engine jacket while the purpose of the LT circuit is 1) to transfers heat away from the HT circuit and 2) to cool the lubrication oil (LO). The engine air intake temperature is controlled by charge air coolers (CAC). Waste heat can be recovered through a waste heat recovery (WHR) heat exchanger (HE), and excess heat is dumped out into the sea.

In order to achieve the optimal energy efficiency, the recovered waste heat must be utilized efficiently. This can be done by making optimal control decisions to schedule tasks such as fresh water production, heating of the passenger cabins or the swimming pool or running the air conditioning, at time instances when as much as possible waste-heat is available. In order to make good decisions about the heating of the ship, the measurements concerning the current consumption and production rates of heat must be available. It is possible to further improve these decisions if it is possible to predict the future energy demands and availability. For prediction purposes, a Digital Twin of the relevant components can play a key role, but cannot be utilized if the available measurements are not adequate. Although measurements from operation critical locations in the ship's cooling system are available from most installations, but less critical points of the process are typically not hooked up to the data acquisition system. Furthermore, the available data is often corrupted by sensor faults such as noise, outliers and stuck or missing data. To address this data unavailability problem, we propose that a Digital Twin model should be utilized to automatically detect faulty data and to correct measurement errors, increasing the data's adequacy. To illustrate how this could be done in practice, a heat exchanger Digital Twin is derived in Section 3. In Section 4, we demonstrate through a simulation case study how data from a heat exchanger can be reconstructed using a simple heat-exchanger model and a sparse-optimization framework.

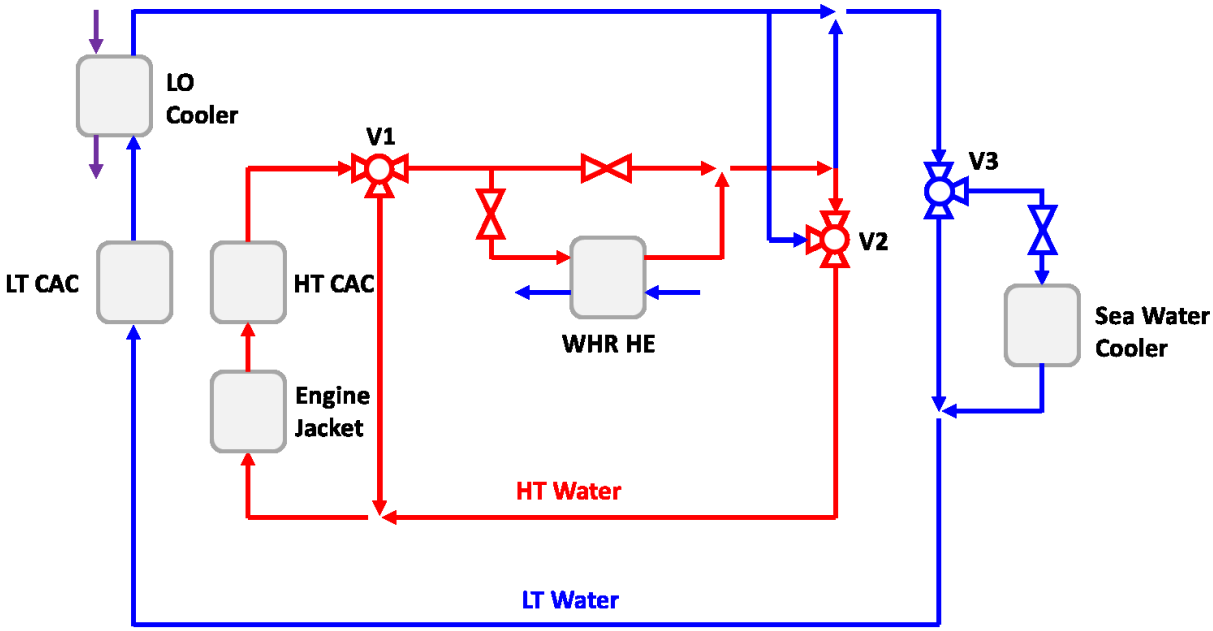

Fig. 1. Schematics of the LT and HT circuits of a marine engine cooling system.

4. Heat exchanger Digital Twin

A Digital Twin is a virtual model of a physical system that is used to simulate and verify the behavior of a true-world process (Tao, et al., 2018). Examples of such processes might be mechanical components, plants and manufacturing supply chains. In the shipping industry, Digital Twins are predicted to have a key role in improving the overall energy efficiency of ships. A few potential areas of impact are in design of energy optimized control systems, in optimal route planning and in determining the operation strategies of hybrid vessels. In this case study, a Digital Twin of a heat exchanger used in marine engine cooling and waste-heat-recovery system is developed.

The heat exchanger model has been derived from a finite elements approximation by discretizing the length of the heat-exchanger in n elements as illustrated in Fig. 2. Perfect mixing within each element is assumed. As the number of elements n grows large, the model is consistent with the heat equation governed by the laws of thermodynamics. The following set of differential equations describes the heat transfer of an element i of the heat-exchanger model

$$\frac{d}{dt}T_{\mathrm{C,out}}^{(i)}(t) = \frac{\dot{m}_{\mathrm{C}}(t)}{M_{\mathrm{C}}}\left(T_{\mathrm{C,in}}^{(i)}(t) - T_{\mathrm{C,out}}^{(i)}(t)\right) + \frac{\dot{Q}_i(t)}{c_{\mathrm{p,C}}M_{\mathrm{C}}}, \tag{1}$$

$$\frac{d}{dt}T_{\mathrm{H,out}}^{(i)}(t) = \frac{\dot{m}_{\mathrm{H}}(t)}{M_{\mathrm{H}}}\left(T_{\mathrm{H,in}}^{(i)}(t) - T_{\mathrm{H,out}}^{(i)}(t)\right) - \frac{\dot{Q}_i(t)}{c_{\mathrm{p,H}}M_{\mathrm{H}}}, \tag{2}$$

where $c_{\mathrm{p,C}}, c_{\mathrm{p,H}}$ are the specific heat capacities in the cold and hot side of the heat exchanger and $M_{\mathrm{C}}, M_{\mathrm{H}}$ are the masses of the media at the cold and hot side. The heat transfer rate $\dot{Q}$ is given by

$$\dot{Q}(t) = K\left(T_{\mathrm{H,out}}(t) - T_{\mathrm{C,out}}(t)\right),$$

with $K = Ah$, with A the area of heat transfer in an element and h is the heat transfer coefficient.

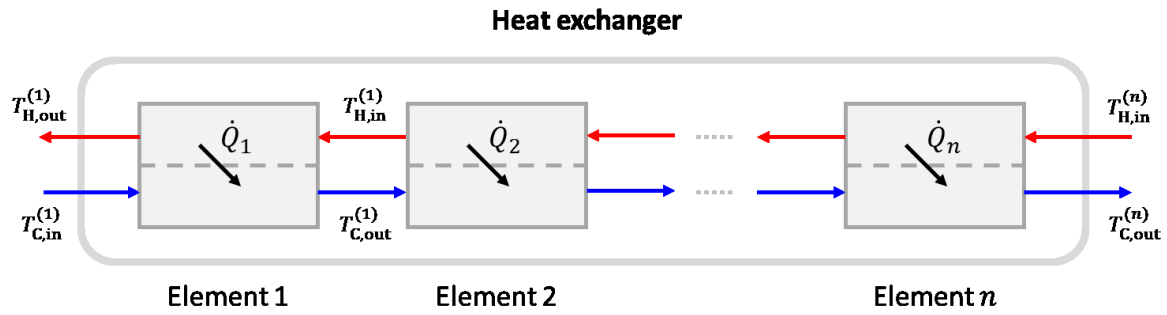

Fig. 2. Finite-element approximation of a heat exchanger.

5. Automatic cleansing of data

In this case study we use the heat-exchanger digital twin presented in Section 3 to ensure that the measurements from sensors in the system are sound from a physical standpoint. While doing data cleansing, it is important to ensure that the properties from the physical process such as mass and energy balances are satisfied. Since the measurements from the process have physically related properties, the data from different sensors cannot be manipulated independently. If we make a correction in one of the signals, every other related signal must be changed accordingly such that the physical laws are still satisfied. In Section 4.1. A few common types of sensor faults are defined and in Section 4.2 the data cleansing problem is formulated as a sparse-optimization problem.

5.1. Sensor data faults

Real-world sensor data is sometimes corrupted by faults. In this section, we describe the properties of the most

commonly occurring types of sensor faults occurring in process data and how they can be encoded mathematically. The sensor data faults considered in this paper are

- *Noise* – Sensor noise is always present in raw data. It is commonly assumed that the noise is zero-mean and uncorrelated with the true underlying signal. A noise corrupted signal $y(t)$ is then described as $y(t) = x(t) + e(t)$, where $x(t)$ is the true signal and $e(t)$ is the noise.

- *Missing data* – When a sensor fails, data will be missing. Best practice in case of a sensor failure would be to simply log a binary variable indicating whether data is missing. However, in practice it is common that when a sensor fails, only the last observed value is logged. This will manifest itself as a 'stuck value' in the logged data. Although it is easy for a human to detect these types of faults by inspection, it is often time consuming to sort out data manually.

- *Stuck values* – Stuck values typically occur when a sensor fails and is hence a product of sensor failure. Although stuck values are easily detected by a human by inspection, in practice, is can be hard to distinguish a faulty stuck sensor value from an accurately logged sensor value at times when the process is idle. Thus, in the following section, we propose to use a digital twin model to automatically detect and correct stuck sensor values.

- *Outliers* – Outlier data is when a single or few consecutive data points are anomalous compared to the other data points. Outliers most often stem from measurement errors but can also be rare and unexpected measurements from the process. Outliers are common in real-life data and are typically easy to spot by a human by inspection but may be time consuming to correct for large data sets. For a sequence of data $\{y(t)\}_{t=0,1,\dots,N}$ data containing outliers, the signal can be expressed mathematically as $y(t) = x(t) + d(t)$, where $d(t)$ is a disturbance signal describing the outliers. If an outlier occurs at time point t', $d(t) = 0$ for $t \neq t'$ and $d(t') = a$ where $a \in \mathbb{R}$ is a constant describing the magnitude of the outlier. If the number of outliers is small compared to the size of the data set, the signal $d(t)$ is sparse, i.e. mostly zero.

- *Trends* – Slowly drifting trends occur in data as an effect of either external factors or changing process properties. For example, dirt accumulating in a heat exchanger (fouling) would manifest itself as a trend in data since the heat transfer coefficient would change slowly by time. Slowly drifting trends ca be modelled as piece-wise linear or smooth functions of time. For a signal $y(t) = x(t) + d(t)$ with a disturbance $d(t)$ describing a slowly drifting trend, the trends can be separated from the true signal $x(t)$ by constraining the second difference of $d(t)$, see e.g. (Kim, et al., 2009) (Manngård, et al., 2017) (Shirdel, et al., 2016) (Hodrick & Prescott, 1997).

In the following section the data cleansing problem is automated by formulating it as a convex optimization problem.

5.2. Problem formulation

The goal in this section is to formulate an optimization problem that automatically detects sensor faults in the heat-exchanger data by utilizing the model described in Section 3. We begin by considering the general case where the signal $y(t)$ is corrupted by sensor faults. When this is established, we will continue with the more specific problem of correcting sensor faults in heat-exchanger data is tackled.

Assume that the observed set of measurements $\{y(t)\}_{t\in\mathcal{T}}$ with vectors $y(t) \in \mathbb{R}^m$ contain sensor faults and can be expressed as

$$y(t) = x(t) + d(t) + e(t).$$

where $d(t)$ describes the outliers and the stuck values, $e(t)$ is the measurement noise and $\mathcal{T}$ is the set of points in time. Note that missing data is not modeled in this formulation. Also note that trends have been left out for illustrational purposes but can easily be added back if required. Filtering of trends is a well-studied topic and has for example been covered in (Kim, et al., 2009) (Manngård, et al., 2017) (Shirdel, et al., 2016) (Hodrick & Prescott, 1997). According to what we established in Section 4.1, a signal of outliers $d(t)$ is sparse. If we assume that the

probability for the event that $d(t)$ represents a stuck value is low, we can handle the stuck values in the same way as we handle outliers. This enables us to use sparse optimization to separate $d(t)$ from the true signal. Sparse optimization is a field of combinatorial optimization which originates from compressive sensing (Baraniuk, 2007) (Candes, et al., 2006), and has been used in a wide range of applications, including image reconstruction (Candes, et al., 2006) (Elad & Aharon, 2006), trend filtering (Kim, et al., 2009) (Manngård, et al., 2017) (Shirdel, et al., 2016) and model reduction (Manngård, et al., 2018). A key result within the field of sparse optimization is that a signal constrained with respect to the ℓ_1 norm tends to be sparse (Candes, et al., 2006). Hence it would be desirable to formulate an optimization problem that minimized the convex cost function

$$J(x,d) = \sum_{t\in\mathcal{T}}\sum_{i=0}^{m} \left(y_i(t) - x_i(t) - d_i(t)\right)^2 + \lambda|d_i(t)| \tag{3}$$

where λ is a positive scalar constant acting as a weight between the goodness of fit and sparsity of d_i. Unfortunately, direct minimization of the cost $J(x,d)$ in (3) would result in an ill-posed problem and additional constraints on $x(t)$ should be imposed for a unique solution to exist. Conveniently, having a digital twin available which relates the signal $x_i(t)$ to other signals and measurements in the process can be used to constrain problem the problem.

6. Case study: cleansing of heat-exchanger data

Let's consider the problem of removing sensor faults from heat-exchanger data collected from a system as described in Sections 2. If we assume that all inlet and outlet temperatures are measured (may contain missing data), i.e. we have a data set containing $T_{\mathrm{C,in}}^{(1)}(t), T_{\mathrm{C,out}}^{(n)}(t), T_{\mathrm{H},in}^{(n)}(t)$ and $T_{\mathrm{H,out}}(t)$ for $t \in \mathcal{T}$ where the set $\mathcal{T}$ contains the time indices of the observed data. Defining the vector of measurements

$$y(t) = \left[T_{\mathrm{C,in}}^{(1)}(t)\ T_{\mathrm{C,out}}^{(n)}(t)\ T_{\mathrm{H},in}^{(n)}(t)\ T_{\mathrm{H,out}}(t)\right]^T$$

for $t \in \mathcal{T}$, outliers and stuck values can be corrected in the measured data, missing data can be reconstructed, and noise can be removed from the measurement by minimizing (3) subject to constraints (1)-(2) imposed by the digital twin. Note that, in practice, since signals are measured in discrete-time, the system model should be converted into a discrete-time model for example using zero-order hold.

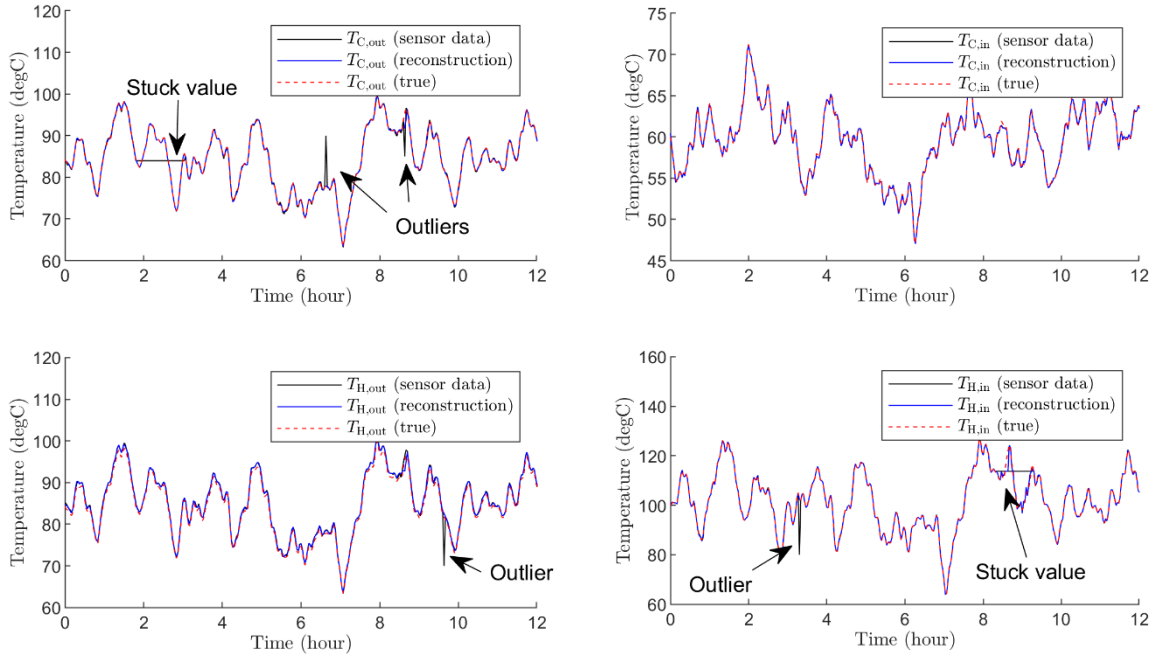

Fig. 3. Sensor data with faults introduced and reconstructed time-series data.

Heat exchanger temperature data was generated using the heat exchanger model presented in Section 3. Measurements where corrupted with zero-mean white noise with standard deviation 0.2 and sensor faults in the form of stuck values and outliers where introduced. In Fig. 3 a measured (simulated) signals containing stuck values and outliers together with reconstructed signals using the proposed method is presenter with weight parameters $\lambda = 0.1$. The outliers and stuck values were automatically detected and corrected. Since the digital twin model, which is based on mass and energy balances of the physical system, was included as constraints in the optimization problem, the reconstructed signals are consistent basic laws of nature.

7. Conclusions

We have presented a case study where we use a Digital Twin to do online error detection and error correction of data from a heat exchanger. The Digital Twin is created using a physics-based model of a heat exchanger. The Digital Twin uses its internal model to check whether the current input signals are reasonable in terms of a convex cost function. The cost function validates the input signal when it shows a good fit to the Digital Twin model. In addition to validation of the data, the cost function can, together with a convex solver correct the input signals.

This case study is with a heat exchanger on a cruise ship, but the methodology can be used in other engineering applications where physics-based models are available. The authors ongoing projects on making Digital Twins of the full heating system of the cruise ship, a mechanical driveline and a diesel engine.

8. References

Baraniuk, R. G., 2007. Compressive sensing. *IEEE signal processing magazine,* 24(4), pp. 118 - 121.
Batini, C., Cappiello, C., Francalanci, C. & Maurino, A., 2009. *ACM Computing Surveys,* 41(3).
Candes, E., Romberg, J. & Tao, T., 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory,* 52(2), pp. 489-509.
Chen, F. et al., 2015. Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks,* 11(8).

Elad, M. & Aharon, M., 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing,* 15(12), pp. 3736-3745.
Esling, P. & Agon, C., 2012. Time-series data mining. *ACM Computing Surveys,* 45(1).
Hodrick, R. J. & Prescott, E. C., 1997. Postwar US business cycles: an empirical investigation. *Journal of Money, credit, and Banking,* 29(1), pp. 1-16.
Kam, M., Zhu, X. & Kalata, P., 1997. Sensor fusion for mobile robot navigation. *Proceedings of the IEEE,* 85(1).
Kim, S. J., Koh, K., Boyd, S. & Gorinevsky, D., 2009. l_1 Trend Filtering. *SIAM review,* 51(2), pp. 339-360.
Lueth, K. L., 2018. *State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating.* [Online] Available at: https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/ [Accessed 5 6 2019].
Manngård, M., Böling, J. M. & Toivonen, H. T., 2017. Subspace identification for MIMO systems in the presence of trends and outliers. *Computer Aided Chemical Engineering,* Volume 40, pp. 307-312.
Manngård, M., K. J. & Böling, J. M., 2018. Structural learning in artificial neural networks using sparse optimization. *Neurocomputing,* Volume 272, pp. 660-667.
Napolitano, M. et al., 1998. Sensor validation using hardware-based on-line learning neural networks. *IEEE Transactions on Aerospace and Electronic Systems,* 34(2), pp. 456-468.
Parris, C. J., Laflen, B., Grabb, M. L. & Kalitan, D. M., 2016. The future for industrial services: the Digital Twin. *Infosys insights*, pp. 42-49.
Rahman, A., Smith, D. V. & Timms, G., 2014. A Novel Machine Learning Approach Toward Quality Assessment of Sensor Data. *IEEE Sensors Journal,* 14(4), pp. 1035-1047.
Sharma, A. B., Golubchik, L. & Govindan, R., 2010. Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks,* 6(3).
Shirdel, A. H., Böling, J. M. & Toivonen, H. T., 2016. System identification in the presence of trends and outliers using sparse optimization. *Journal of Process Control,* Volume 120-133, p. 44.
Tao, F. et al., 2018. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology,* 94(9-12), pp. 3563-3576.
Travis, G., 2019. How the Boeing 737 Max Disaster Looks to a Software Developer. *IEEE Spectrum*, 18 April.
Zou, G. et al., 2013. *Modeling ship energy flow with multidomain simulation.* Shanghai, China, In Proc. 27th CIMAC World Congress on Combustion Engines. International Council on Combustion Engines..