Sensor-data Analytics in Cyber Physical Systems
From Husserl to Data Mining

Paul O'Leary, Matthew Harker and Christoph Gugg
Chair of Automation, University of Leoben, Leoben, Austria

Keywords: Data Mining, Inverse Problems, Data Analytics, Phenomenology, Embedded Simulation.

Abstract: This position paper proposes a discussion of the need for a solid philosophical basis for mining sensor-data based on phenomenology. Additionally it is proposed that, when considering cyber physical systems, the solution of inverse problems is a prerequisite if the results are to have physical meaning. A prototype lexical analysis tool for sensor-data is presented and its application to knowledge discovery in large mechatronic systems is demonstrated.

1 INTRODUCTION

There are many different definitions for what constitutes a cyber physical system (CPS) (Baheti and Gill, 2011; Geisberger and Broy, 2012; Spath et al., 2013a; NIST, 2013; NIST, 2012; Park et al., 2012; IOSB, 2013; Spath et al., 2013b; Lee, 2008; Tabuada, 2006). The most succinct and pertinent to this paper is the definition given by the IEEE (Baheti and Gill, 2011) and ACM[1]: *A CPS is a system with a coupling of the cyber aspects of computing and communications with the physical aspects of dynamics and engineering that must abide by the laws of physics. This includes sensor networks, real-time and hybrid systems.*

Mining sensor-data from such system and performing knowledge discovery is significantly different from mining static databases. The fundamental question in any philosophy is: *What is a valid source of knowledge[2] and how do we acquire this knowledge?*

Consequently, any person developing a data analytics system which includes knowledge discovery must answer this philosophical question and the philosophy must be implemented within the software. There is no avoiding the truth of this statement. Most authors, however, do not explicitly address this issue in sufficient depth. Fundamentally, there is no information is the data stream itself. It is the association of some form a model with the data and the parameters of this model when applied to the data which reveal the information. Or more precisely enable an inference with respect to the cause of the observation[3].

Virtually all literature and standard works on data mining, e.g., (Han, 2005, Chapters 2 and 3), deal with data models, predominantly statistical data models, as a means of revealing correlations. In the case of mining sensor-data emanating from technical systems we are observing data related to specific phenomena which is influencing a system. In such cases if we are to make any statement about the phenomena it is necessary to consider the system model and not just data models. Unfortunately, inverse problems associated with such systems and sensor-data are not in general addressed, see for example (Aggarwal, 2013) where only data models are addressed.

If we are to perform sensor-data analytics, in a manner such that the results have physical meaning, i.e. obeying the laws of physics, then it will be necessary to solve inverse problems in real time. Furthermore, the use of multiple sensors in a large system will require compact descriptions of complex events. Humans use language to express complex sensory experience and associated perceptions, since a description of the sensory excitation would be too longwinded and not understandable. But more importantly, most of the information is to be found in the common understanding of the word used; and is

[1] ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS) (iccps.acm.org)

[2] Sometime the word truth is used instead of knowledge.

[3] If we have no model there can be no information. The numbers in a computer have no meaning. The attachment of meaning to a number is fundamentally adding a model. Many working in sensor networks do this implicitly, we prefer to do it explicitly.

O'Leary P., Harker M. and Gugg C..
Sensor-data Analytics in Cyber Physical Systems - From Husserl to Data Mining.
DOI: 10.5220/0005328601760181
In *Proceedings of the 4th International Conference on Sensor Networks* (SENSORNETS-2015), pages 176-181
ISBN: 978-989-758-086-4

not to be found in the sensor-data, but is associated with the sensor-data through the use of the word.

In this *position paper* we wish to initiate a serious and in depth discussion of the following issues:

1. Can the philosophy of phenomenology (Husserl and Hardy, 1999; Merleau-Ponty, 2002) be a support during the design and partitioning of a sensor-data mining system?
2. Do we need to consider the relevance of the development of natural languages (Hmc, 2000, Appendices I and II) as a possible model for generating symbolic representations of sensor-data when considering more complex events?
3. There is a very clear need to implement real-time solutions of inverse problems if we are to extract physically relevant information and knowledge. Statistical data models are not sufficient.
4. There is a need for embedded simulation to improve the overall efficiency of real-time sensor-data mining.

To support our hypothesis we demonstrate results from a prototype implementation of sensor-data mining systems. The system is motivated by the Indo-Asian model for phenomenology and implements the real-time solution of inverse problems and embedded simulation. The system has been applied to the analysis of two different large pieces of machinery.

2 THE POSITION

This section presents a discussion of the elements contained within this position paper. These are the main topics on which we wish to initiate an in depth discussion.

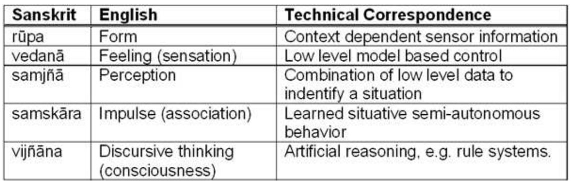
2.1 Phenomenology

Phenomenology as a branch of physics is defined as: *a body of knowledge that relates empirical observations of phenomena to each other, in a way that is consistent with fundamental theory, but is not directly derived from theory*. This is exactly what we wish to do with data mining and knowledge discovery. However, statistical data models alone will not essentially deliver results which are consistent with the fundamental theory of the system, since they make no attempt to model the system being observed. Without solving the inverse problems associated with the system we can not guarantee consistency with the fundamental theory.

If we look closer at knowledge discovery we must also consider phenomenology as a branch of philosophy. Edmund Husserl[4] is considered to be the founder of the philosophy of phenomenology which is normally defined as: *the study of the development of human consciousness and self-awareness as a preface to or a part of philosophy*. The origins, however, are much older in the 18th century, the Swiss German mathematician and philosopher Johann Heinrich Lambert applied the concept to that part of his theory of knowledge that distinguishes truth from illusion and error. Phenomonology was later extended by Martin Heidegger who introduced the concepts of present-at-hand and ready-to-hand. Which proposed that we gain all knowledge only by studying our“average-everyday” understanding of the world gained from direct interaction with objects. Later Maurice Merleau-Ponty with his best know work “Phenomenology of Perception” (Merleau-Ponty, 2002) analyzed how we perceive as a result of experiencing phenomena, his concept of *embodiment*, with a little over simplification, can be summarized as: *we perceive phenomena first, then reflect on them* - quite the opposite to “I think therefore I am”.

The East-Asian view of phenomenology, as summarized by the five aggregates (five Skandhas, see table 1) (Lusthaus, 2002) is quite different. It describes the path from sensory excitation to discursive thought in five distinct steps, i.e. the five aggregates. The aggregates are only recently finding reception in the western cognitive sciences (Davis and Thompson, 2013)[5]. The five steps are interesting since the assentation is that we are never in direct contact with objects in the world, but always in contact with a model for the world; with which we are back with the inverse problems and models for the system being observed.

Table 1: Abbreviated presentation of the five aggregates (skandhas) and their possible technical interpretation.



Sanskrit	English	Technical Correspondence
rūpa	Form	Context dependent sensor information
vedanā	Feeling (sensation)	Low level model based control
samjñā	Perception	Combination of low level data to indentify a situation
samskāra	Impulse (association)	Learned situative semi-autonomous behavior
vijñāna	Discursive thinking (consciousness)	Artificial reasoning, e.g. rule systems.

It is beyond the scope of this text to enter the de-

[4] Edmund Husserl (1859-1938) graduated from the University of Vienna with a doctoral dissertation in theoretical mathematics on the calculus of variations.

[5] Some care must be taken when considering this text since it is a western interpretation of an East-Asian philosophy which is based on contemplative practice. The authors are scholars and not contemplatives.

tails of the five Skandhas and how there are being received within the cognitive sciences. It suffices to say, that independent of their final truth or not, they do offer a good basis for discussion on the partitioning of a system and the suitability of which techniques at which level in a cognitive system.

To summarize the issue of phenomenology:

1. system models are required and the corresponding inverse problems must be solved if we are to extract conclusions from sensor-data which are based on physical systems.
2. there is much discussion on cognitive systems. We believe it would be more appropriate to use the term *perceptive systems*. Both the western and eastern view is that perception is the product of sensors activity and context, but not the product of thinking. Thinking is the process which leads to an understanding of the perceived.

The data-mining system presented later in this position paper implements the real-time solution of an inverse problem for each and every sensor channel and the possibility of an embedded simulation for every actuator channel.

2.2 Natural Language

The Yogachara (Lusthaus, 2002)[6] school asserts that language is generated when repetition is cognised in sensory data. It is the repetition which is considered to be characteristic and not the thinking about the repetition, this process leads to a naming. At this point we may not understand but we do perceive. In this manner the use of a word is considered to be a metaphor which points at a sensory experience rather than describing the experience directly. Monosyllabic words are considered to represent simpler sensory experiences, while polysyllabic words tend to describe more abstract experience which are commonly the result of complex multi-sensory experience.

When designing symbolic representations for sensor-data which should enable complex symbolic queries, see for example (Lin et al., 2003), we should consider the metaphoric nature of the symbols and associate human readable text with the symbols. This will facilitate greatly the process of knowledge discovery and support the automatic generation of derived *polysyllabic* symbols by combination of events, each of which was associated with a single channel sensor-data.

As will be seen later we propose a method of symbolic representation which at the scanner level consists of tokens and predicates extracted from the sensor data. The tokens and predicates can then be combined to define syntax. This opens the door to using concepts such as those embodied in *Lex* and *Yacc* (Johnson, 1975) to enable BNF definitions for the sensor-data analysis[7] and queries. With this the step has been made from natural language to computer implementable language.

2.3 Inverse Problems and Embedded Simulation in Real-time

The details of a method and framework for the implementing real-time solutions to inverse problems and embedded simulation can be found in (O'Leary and Harker, 2012) and the required automatic code generation in (Gugg et al., 2014). For this reason we shall not go into more detail on the method at this point. The solution approach requires the application of a linear differential operator to the data stream.

3 THE PROTOTYPE IMPLEMENTATION

In this section we present the initial design and implementation of a prototype sensor-data mining system motivated by the position presented in this paper.

3.1 Single Channel Lexical Analyzer

The schematic diagram for the single channel lexical analyzer (SCLA) is shown in Figure 1. It consists of a linear differential operator (LDO) which is applied to the sensor data stream. Each sensor channel has its own linear operator associated with it. The LDO can be used either to implement the solution to an inverse problem or to an embedded simulation, i.e. it can solve ODEs in both a forward and an inverse manner. The lexical analysis element corresponds to symbol aggregate approximation (SAX) (Lin et al., 2003). However, it is not applied directly to the sensor data but to the processed sensor data. As an extension of the SAX approach k-means clustering is applied to the processed signal to identify repetitive clusterings

[6] Yogachara is a philosophical school founded by Vasubandhu and Asanga in the 4$^{\text{th}}$ century in India. It was carried to China where it was adopted and known under the name Chan. A part of the philosophy is associated with the *Tridhatu*, i.e., the relationship between the realms of body, language and mind.

[7] It is important to differentiate between markup languages for sensor data, e.g. sensorML, and markup for the sensor-data analysis and queries.

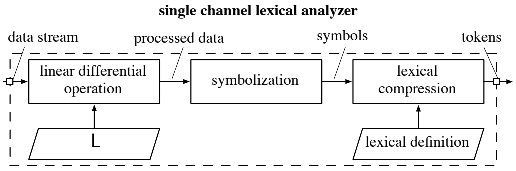


Figure 1: Schematic of the single channel lexical analyzer.

of values[8]. Each cluster is assigned a symbol and a human readable text describing the meaning of this cluster. Finally the lexical compression generate the tokens and predicates, i.e. each persistent symbol is replaced by a single token with a length predicate and pointer which link the token to its start and end positions in the real data. This enables backtracking of the symbolic queries to real sensor data values.

The SCLA can be defined in a simple structure and an array of such structures is used to manage the definitions of all signals and lexical analyzers. In this manner the implementing is not in the form of dedicated code, but as a generic processing engine. Its input is the set of sensor-data and the lexical definitions and it returns the token-tables for each SCLA and their combinations. In this manner applying the system to a new machine only required the definition of a new set of SCLA parameters and combinations.

3.2 Multi Channel Lexical Analyzer

The schematic for the muiti-channel lexical analyzer is shown in Figure 2. The outputs of the SCLAs are grouped according to their relevance to specific operations. The token tables are merged automatically, and the statistics for each token combination are computed during this task. This enables the identification of repetitive events, events which are seldom and of events which do not occur. The output is itself once again a token table. This is the first level of automatic knowledge discovers, since it characterizes the operations of the machines.

Symbolic queries are implemented as regular expressions which are applied to the token tables and their predicates. In this manner specific sequences of events can be located. The lexical compression implicitly implements dynamic time warping (DTW), and there is a positive semi definite distance metric associated with the tokens and their predicates. This ensures that it is possible to do sequence similarity measurements.

The final stage of the MCLA is the generation of a dictionary. Since all symbols are associated with human readable text it is possible to generate dictionary entries automatically which explain the meaning of tokens derived from combining single token tables. The combination of multiple tokens is considered akin to the formation of polysyllabic words.

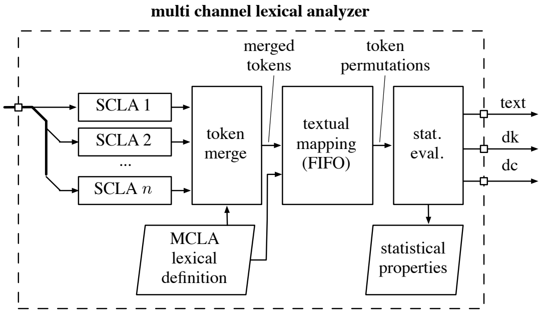


Figure 2: Schematic of the multi channel lexical analyzer.

4 DEMONSTRATION

The method has been applied to a number of different machines to evaluate the feasibility. Here we present the results of the application of the system to a large scale reclaimer, see Figure 3.

4.1 SCLA Demonstration

The slewing data from the machine is used to demonstrate a simple single channel lexical analyser. The data channel consists of $n = 840000$ samples. We are interested in determining the rate of change of slewing, i.e., the speed and direction of slewing. For this reason the linear differential operator implements a regularizing differentiation (O'Leary and Harker, 2010). The slew data and the result of the computation are shown in Figure 4.

Figure 3: The large scale reclaimer used to demonstrate the concepts presented in this paper.

[8] K-means clustering was chosen because the histogram exhibits clear peaks and the data is well modelled by a sum of gaussian distributions.

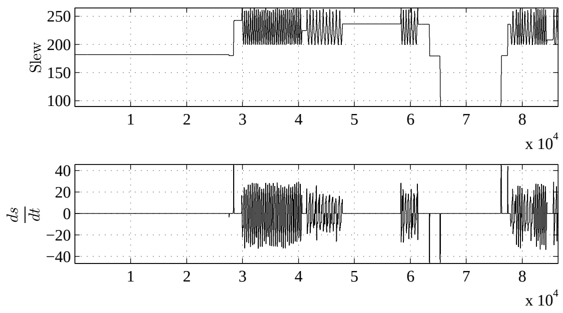


Figure 4: Top: the slew data and bottom: the rate of change of slew. The data consists of $n = 86225$ samples, a day's production sampled at $t_s = 1s$.

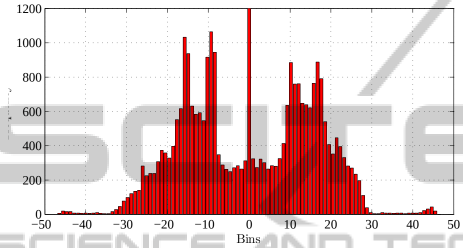


Figure 5: Histogram of the rate of change of slew. This is used to identify the levels for the symbolic aggregate approximation.

During the exploratory phase of operation, a k-means clustering is applied of the histogram for the rate of slewing, see Figure 5. For simplicity of demonstration three clusters are used and assigned the symbols $[1,2,3]$ with the associated human readable texts [slewing-left, no-slewing, slewing-right].

The generated token table for the slew data now enables symbolic queries. The result of a symbolic query for an interrupted right slew movement is shown in Figure 6.

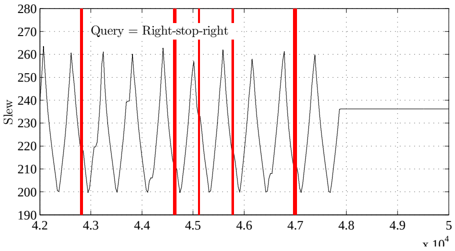


Figure 6: A symbolic query for an interrupted right slewing of the machine, abbreviated as right-stop-right. The occurrences of the query in the data are marked in red.

4.2 MCLA and Knowledge Discovery

Here we demonstrate the implementation of SCLAs for the *slew*, *long-position* and *tonnage* of this machine, each with a different LDO, see Figure 7 for a visual representation. Then the token tables for the three channels are merged automatically to identify more complex machine operations.

The data set has $n = 86225$ samples for each sensor, i.e. a full day of operation samples at $t_s = 1s$. The combined token table has $m = 578$ entries, only these entries need to be searched during a symbolic query - this corresponds to a very high degree of compression. Additionally the system generates a statistical analysis for the identified tokens, see Table 2. It can be seen that certain combinations of events never occur, this is knowledge discovery.

To get a measure for performance data was taken from a second machine with a sampling time $t_s = 20ms$. The data acquisition was started at $t_1 = 22 - Jan - 2014 10:39:09$ and stopped $t_2 = 24 - Jan - 2014 22:18:17$, that is for a time period of approximately 60 hours. The generation of the token table for a single channel required $t_c = 0.84sec$ this includes the solution of the inverse problem. A symbolic query on this data set required between $t_q = 20 \dots 30ms$ depending on the query. This computation was performed on a $2.4GHz$ dual core processor.

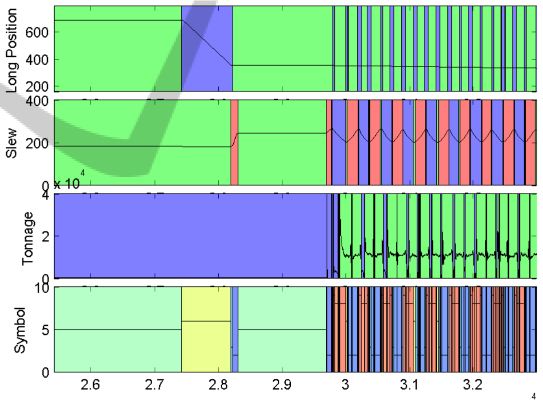
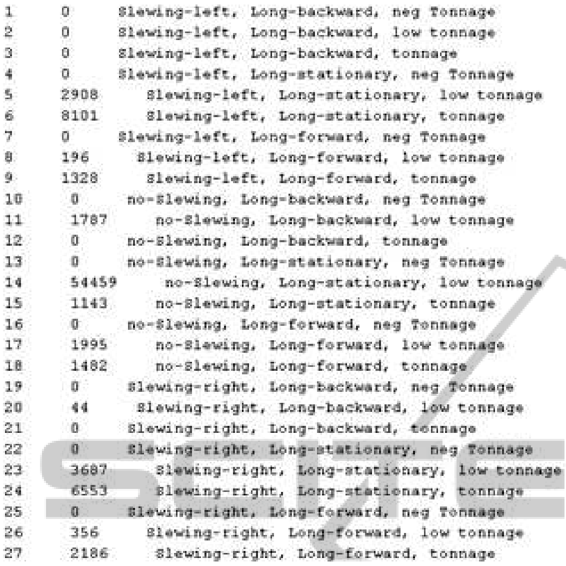


Figure 7: Three SCLA channels and the automatically generated *vocabulary* for the combination of these channels. This is only for a small zoomed portion of the data. The data set has $n = 86225$ samples for each sensor, i.e. a full day of operation samples at $t_s = 1s$. The combined token table has $m = 578$ entries, only these entries need to be searched during a symbolic query.

5 CONCLUSIONS

It is concluded that the concepts of the philosophy of phenomenology are of great support when designing sensor-data analytics, data mining and knowledge discovery systems. If complex events are to be automatically discovered and identified it will be necessary to consider the use of language to implement metaphoric abbreviations, i.e. terms which abbreviate the description of complex sensorial events. Furthermore, it has

Table 2: Statistical summary of the result of merging the token tables for *slew*, *long-position* and *tonnage*. Note some operations are automatically identified as not occurring.



```
1      0      Slewing-left, Long-backward, neg Tonnage
2      0      Slewing-left, Long-backward, low tonnage
3      0      Slewing-left, Long-backward, tonnage
4      0      Slewing-left, Long-stationary, neg Tonnage
5      2908      Slewing-left, Long-stationary, low tonnage
6      8101      Slewing-left, Long-stationary, tonnage
7      0      Slewing-left, Long-forward, neg Tonnage
8      196     Slewing-left, Long-forward, low tonnage
9      1328      Slewing-left, Long-forward, tonnage
10      0      no-Slewing, Long-backward, neg Tonnage
11      1787      no-Slewing, Long-backward, low tonnage
12      0      no-Slewing, Long-backward, tonnage
13      0      no-Slewing, Long-stationary, neg Tonnage
14      54459      no-Slewing, Long-stationary, low tonnage
15      1143      no-Slewing, Long-stationary, tonnage
16      0      no-Slewing, Long-forward, neg Tonnage
17      1995      no-Slewing, Long-forward, low tonnage
18      1482      no-Slewing, Long-forward, tonnage
19      0      Slewing-right, Long-backward, neg Tonnage
20      44     Slewing-right, Long-backward, low tonnage
21      0      Slewing-right, Long-backward, tonnage
22      0      Slewing-right, Long-stationary, neg Tonnage
23      3687      Slewing-right, Long-stationary, low tonnage
24      6553      Slewing-right, Long-stationary, tonnage
25      0      Slewing-right, Long-forward, neg Tonnage
26      356     Slewing-right, Long-forward, low tonnage
27      2186      Slewing-right, Long-forward, tonnage
```

been argued that the solution of inverse problems is indispensable if the results are to have physical meaning. The demonstration of the application of the system to a large mechatronic system has demonstrated the functionality of the concept.

The issue of scale space has still not been addressed. That is, there are features which are relevant at different time scales: some are relevant in millisecond ranges and some in hours. We are presently investigating a structured decimation of the data.

REFERENCES

Aggarwal, C. C. (2013). *Managing and Mining Sensor Data*. Springer Publishing Company, Incorporated.

Baheti, R. and Gill, H. (2011). Cyber-physical systems. *The Impact of Control Technology*, pages 161–166.

Davis, J. H. and Thompson, E. (2013). *From the Five Aggregates to Phenomenal Consciousness, in A Companion to Buddhist Philosophy (ed S. M. Emmanuel)*. John Wiley and Sons, Ltd, Chichester, UK.

Geisberger, E. and Broy, M. (2012). *agendaCPS: Integrierte Forschungsagenda Cyber-Physical Systems*, volume 1. Springer.

Gugg, C., Harker, M., O'Leary, P., and Rath, G. (2014). An algebraic framework for the real-time solution of inverse problems on embedded systems. *CoRR*, abs/1406.0380.

Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hmc, O. (2000). *The American Heritage Dictionary of the English Language*. Houghton Mifflin.

Husserl, E. and Hardy, L. (1999). *The Idea of Phenomenology*. Husserliana: Edmund Husserl – Collected Works. Springer Netherlands.

IOSB, F. (2013). Industry 4.0 information technology is the key element in the factory of the future. Press Information.

Johnson, S. C. (1975). Yacc: Yet another compiler-compiler. Technical report, Computing Science Technical Report No. 32, Bell Laboratories, Murray hill, New Jersey.

Lee, E. A. (2008). Cyber physical systems: Design challenges. In *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on*, pages 363–369. IEEE.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, pages 2–11, New York, NY, USA. ACM.

Lusthaus, D. (2002). *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Cheng Wei-shih Lun*. Curzon critical studies in Buddhism. Routledge Curzon.

Merleau-Ponty, M. (2002). *Phenomenology of Perception*. Routledge classics. Routledge.

NIST (2012). Cyber-physical systems: Situation analysis of current trends, technologies, and challenges. Technical report, National Institute of Standards and Technology (NIST).

NIST (2013). Strategic r&d opportunities for 21st century cyber-physical systems. Technical report, National Institute of Standards and Technology (NIST).

O'Leary, P. and Harker, M. (2010). Discrete polynomial moments and savitzky-golay smoothing. In *Waset Special Journal*, volume 72, pages 439–443.

O'Leary, P. and Harker, M. (2012). A framework for the evaluation of inclinometer data in the measurement of structures. *IEEE T. Instrumentation and Measurement*, 61(5):1237–1251.

Park, K.-J., Zheng, R., and Liu, X. (2012). Cyber-physical systems: Milestones and research challenges. *Computer Communications*, 36(1):1–7.

Spath, D., Ganschar, O., Gerlach, S., Hämmerle, M., Krause, T., and Schlund, S. (2013a). *Produktionsarbeit der Zukunft-Industrie 4.0*. Fraunhofer IAO Stuttgart.

Spath, D., Gerlach, S., Hämmerle, M., Schlund, S., and Strölin, T. (2013b). Cyber-physical system for self-organised and flexible labour utilisation. *Personnel*, 50:22.

Tabuada, P. (2006). Cyber-physical systems: Position paper. In *NSF Workshop on Cyber-Physical Systems*.