Big Data-Driven Contextual Processing Methods for Electrical Capacitance Tomography

Andrzej Romanowski

***Abstract*—This paper presents a new approach to analyzing measurement records from industrial processes. The proposed methodology is based on the model of contextual processing and uses big data from experimental process tomography datasets. Electrical capacitance tomography is used for monitoring noninvasive flow and for data acquisition. The measurement data are collected, stored, and processed to identify process regimes and process threats. A specific physical modification was introduced into the pneumatic conveying flow rig in order to study flow behavior under extreme conditions, extending the available knowledge base. A support vector machine was applied for data classification. This study illustrates how contextual processing can facilitate data interpretation and opens the way for the development of methods for detecting pre-emergency flow patterns.**

***Index Terms*—Big data analysis, contextual processing, electrical capacitance tomography (ECT), industrial process tomography.**

I. Introduction

ALMOST 60% of all solid materials used in industry are processed in the form of bulk particulates [1]. Bulk solids are important and used widely in many fields, such as chemical engineering, food processing, and pharmaceutical production [2], [3]. Fully automating the operation and control of industrial processes involving bulk solids still presents a challenge in many cases [4], [2], [5], [6]. Process tomography techniques, especially electrical capacitance tomography (ECT), could be implemented in the control feedback loop [2], [4], [7], [8]. However, the integration of ECT into real-time monitoring and control systems has so far been limited, mainly due to inherent issues concerning data analysis and interpretation [4], [5], such as insufficient exploitation of measurement data related to process behavior [9], [10]. Therefore, this paper presents a methodology that extends the existing practice of modeling process monitoring, in order to increase the consistency and reliability of ECT data interpretation. This is achieved with the aid of the contextual data processing model (CDPM) depicted in Fig. 1. The

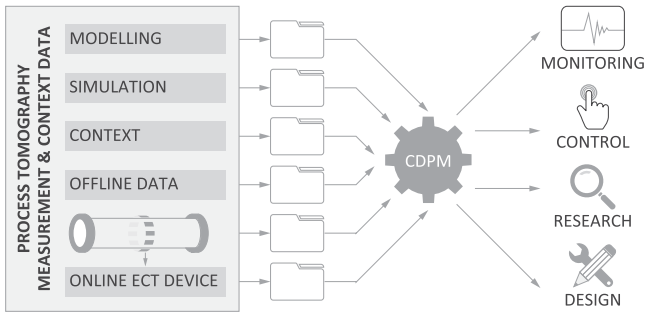

Fig. 1. CDPM workflow diagram. Left: variety of data and information input types; middle: CDPM inference engine; and right: possible output applications (monitoring, control, research, and design).

CDPM is able to support ECT with a data-driven system, which enhances its ability to understand process behavior [11]–[13]. Better understanding of an industrial process can lead to more efficient monitoring and control, better design, and eventually to its optimization [10].

The aim of this research was to improve the interpretation of ECT output, in order to recognize specific bulk flow regimes (e.g., risky phenomena), such as pipeline blockage threats, more effectively. The main difficulty with performing a comprehensive study of these phenomena is their infrequent occurrence, and hence the lack of sufficient experimental data [5]. Two schemes for increasing the volume of available quality data are therefore proposed:

1) Aggregation and joint processing of data captured in different settings, using typical techniques for big data analysis.
2) Provoking the occurrence of specific events and generating more empirical data, using the artificially generated input technique.

This study makes two major research contributions. The principal, practical contribution is its experimental demonstration of a new methodology for handling data produced using industrial process tomography, which contextually incorporates additional data into the process workflow. The main advantage of the proposed data-driven workflow is that it improves detection of critical flow regimes, which pose a risk to the industrial process. This enables preblockage flow patterns to be distinguished (c.a. 90% accuracy), even when there is such limited data available that this would not normally be possible. Current approaches to ECT process modeling do not support this methodology, but they could be extended to utilize the additional information of-

Manuscript received May 30, 2018; accepted June 29, 2018. Date of publication July 12, 2018; date of current version March 1, 2019. Paper no. TII-18-1375. This work was financed by the Lodz University of Technology, Faculty of Electrical, Electronic, Computer and Control Engineering as a part of statutory activity.

The author is with the Institute of Applied Computer Science, Lodz University of Technology, Lodz 90924, Poland (e-mail: androm@kis.p.lodz.pl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2018.2855200

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

fered by analysis of broader data sources, thereby improving performance. The methodology is validated using a binary support vector machine (SVM) classifier.

The second contribution is the CDPM itself: A theoretical model for data processing using a contextual approach. The application of this model expands the knowledge base that can be derived from monitoring process data, and hence augments the possibilities for industrial process modeling. The CDPM model is examined and validated in the case of ECT-monitored industrial pneumatic conveying, but it is easily scalable for use with other process tomography techniques and different applications.

This paper is structured as follows. Section II presents the key novelty of the paper, the concept, and rationale behind the CDPM. Four categories of context-related data are presented, which can be used for modeling and eventually monitoring flow processes. These include the big data approach and artificially generated inputs (AGIs), which are the main focuses of this study. Section III describes methods borrowed from the field of big data analysis and applied to ECT experimental data. It also presents the proposed data handling system, as well as providing an overview of the experimental workflow presented in later sections. Section IV covers verification of the proposed approach, using three computational use cases: 1) testing the adequacy of the methods for big data analysis; 2) evaluating the consistency of the concept of AGIs, and finally 3) cross-mixed comparative data analysis. The results are discussed alongside the subsequent stages of the verification process. Section IV presents a brief experiment predicting process failure in advance. Finally, possibilities for further research and conclusions are presented in Sections V and VI, respectively.

II. Contextual ECT Data Processing

Contextual data, contextual information, and contextual signal processing are different terms used to refer to methods and algorithms with context-aware, context-enabled, or context-driven features. These features can enhance computer systems and applications by broadening the input in comparison to classical standalone solutions [19]. Research related to context and context-awareness has been ongoing in computer science, cognitive science, artificial intelligence, and general engineering disciplines for over three decades, and is expected to remain central to continuing innovation over the coming years [15]. Many papers describe contextual computing as a fundamental requirement for interactive systems and mobile computing [16]. Some discuss the data-oriented perspective in broader ICT applications [17]. Yet, there has so far been less work utilizing contextual data processing techniques to solve engineering problems, such as issues relating to industrial control, measurement, or instrumentation [18], [16]. Numerous definitions of context and context-awareness have been offered, but the majority of studies have focused on the circumstances and settings that have an impact on the real meaning of the situation or activity. These conditions may be interpreted as external or complementary signals, parameters, and information sources, which assist the computerized systems to aggregate a more comprehensive input set for the job at hand. In other words, they help provide a wider picture of the ecosystem interface.

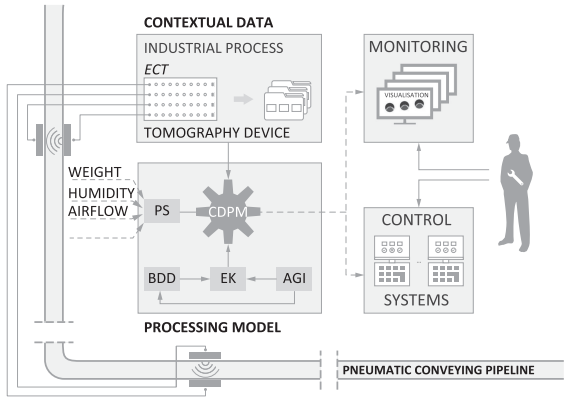

Fig. 2. Simplified schema of the CDPM for a pneumatic conveyor pipeline installation.

For the purposes of industrial process monitoring, the following categorization of context-related data (i.e., prior, external or contextual inputs, external with respect to the main measurement tool) is proposed:

1) Expert knowledge (EK) denotes the entire body of knowledge (derived from well-known rules, phenomena physics, expert observations, and theoretical models) regarding the investigated field. Inputs that fall into this category are independent of the experimental settings.
2) Peripheral sensor inputs (PS) are additional, extra contributions to the main computer system and complementary to the theoretically modeled behavior of the investigated object. These inputs can vary significantly in different experimental settings.

However, there may not be enough domain-related or expert knowledge available, in which case neither the first nor the second category will provide sufficient support. Moreover, modeling of related phenomena can too expensive or too difficult to conduct. Despite incorporating all available information from the EK or PS categories, the uncertainty surrounding the process will still be too high.

There are two other potential methods that can extend the desired capabilities of the system:

3) Big data-derived analysis (BDD) is a methodology taken from the broader field of big data. Tools and methods borrowed from the field of big data analytics can help to draw general conclusions regarding phenomena that are usually so peripheral that they are treated as statistically insignificant and negligible.
4) AGIs are introduced to provoke specific performance patterns. They are especially important when the patterns are of significant importance and/or if they are infrequent and hence difficult to capture and study outside the laboratory; in-the-wild. AGI are particularly helpful when studying the responses of a system, to learn the origins of particular phenomena that occur during normal operation.

Fig. 2 shows a schema of the monitoring and control system (on the right) based on ECT measurements (upper left), coupled with a CDPM in a pneumatic conveying process. The CDPM depicted in the lower left corner of Fig. 2 takes as its inputs ECT measurements and inferences based on EK, PS, BDD, and AGI

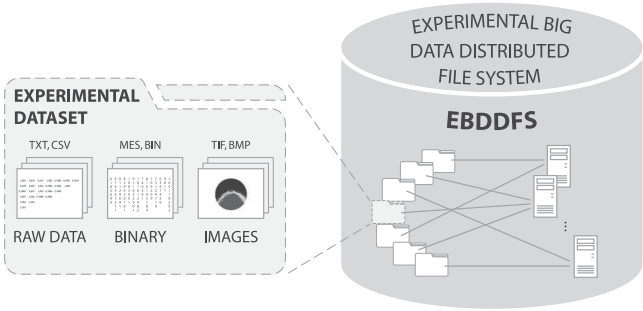

Fig. 3. EBDDFS (right) composed of experimental data from ECT measurements of various types (left).

elements. The complete CDPM configuration brings various benefits, such as regularization of inverse problem-solving based on PS and/or EK [5], [19], [20].

However, this paper is focused only on exploring the BDD and AGI elements. The BDD module is responsible for using the broader experimental data, using big data paradigms. The AGI element introduces the AGI, so that the behavior of the system (and how that behavior is reflected in the main measurements) can be investigated. It is planned in further research to deploy this approach for the construction (or alteration) of a theoretical model of the system, which could be applied in online monitoring strategies.

III. Exploring Big Data Paradigms for Processing Data From Experimental Process Tomography

Big data is sometimes described in terms of five Vs, which stand for velocity, volume, value, variety, and veracity. In this paper, we begin with volume and focus mainly on variety, in order to generate extra value. Velocity and veracity will be explored in future research. Big data usually refers to huge amounts of data, produced by humans both explicitly (documents, photos, videos, social network entries, etc.) and implicitly (as records of various activities), as well as by other systems connected to the Internet (such as sensors, transceivers, and webcams) [21], [22]. The term is also used to refer to some scientific data, such as astrophysical records or so-called big experiment data. Rarely is it applied in the context of conventional experiments and standard, moderately-sized research projects. The reasons are usually the rare and transient nature of the phenomena, the limited amount of data produced and the local character of the study.

The present work runs counter to this established approach. First, the size of the data produced during experiments with ECT can be substantial. Second, applying big data-related techniques such as data mining and machine learning to localized and disconnected data creates new opportunities. A specialized data aggregation and storage system for remote and "disconnected" experiments is also required [23], [24], [35].

Fig. 3 shows in a simplistic way how the measurement data from ECT-based experiments are aggregated. On the left, the basic experimental dataset is seen as a package of files arranged in sequences representing time series, documented in different representations. A single measurement sequence taken using the ECT device may contain three types of file: 1) raw measurement data; 2) binary data; and 3) reconstructed images accompanied by a single calibration file (omitted in Fig. 3 for simplicity). The raw measurements are usually stored as text files (.txt or .csv file formats). The binary files differ depending on the ECT equipment used (.mes, .bin or similar file formats). Images are saved as bitmaps (both uncompressed and compressed) or text values (.tif, .bmp, .jpg for rendered images and .txt, .csv, etc. for text-stored image-ready formats). In this study, we used only raw data and images. Such experimental ECT datasets can be 4 GB in size (for 300 s @ 200 fps for twin 16-electrode sensors) up to 1 TB+ [for two-dimensional (2-D) images and raw data] or even 100 TB+ (for 3-D).

The ultimate goal of this research is to create new opportunities for data mining, by linking incoherent experimental data. This can be accomplished using the proposed flexible computational environment, designed to explore the volume and variety of ECT experimental data (especially for joint exploration of experiments originating in different, remote settings) [23], [32]. The basic workflow for tomographic data analysis using the big data approach passes through the following consecutive steps:

1) Multiexperiment data linking into a MapReduce-based distributed framework; metadata-based indexing.
2) Classification stage: model definition (including data labeling for supervised learning classifiers) and training.
3) Verification of the proposed approach.

The first step is to set up a distributed computing environment, such as the Hadoop-based (or Apache-Spark) system, or a similar open source or proprietary networking/cloud computing service [24]. A simplified schema of how the experimental ECT data is organized into the Experimental Big Data Distributed File System (EBDDFS) is presented in Fig. 3. A variety of tomographic data types and files are reorganized across the network of computing resources. Efficient organization of data within the EBDDFS also depends on preprocessing and on extensive metadata description of the recorded files. This work used standard map-reduce arrangements, yet dedicated and more efficient BDD implementations such as [23] could be employed and tested in future versions [21], [35].

This specific distributed organization of data based on the map-reduce algorithm ensures proper handling and delivery of information, whenever it is necessary to call up particular features or patterns from the data. Initially, the system works on the available experimental data, but it is possible to add freshly captured data after system initiation virtually on-the-fly. In future, whenever the BDD "Velocity" factor begins to play the crucial role in the system, then so-called lambda architecture could be implemented [25]. Tests and development were conducted from Hadoop 1.0 up to Hadoop 2.7 over several years. Initial work was conducted on three workstations (1 server PC, 1 workstation PC, and 1 laptop) and the maximum size of the network was 26 PC workstations located in three LAN segments. The system is scalable. The tests presented in this paper were conducted within a distributed, yet locally managed, secure closed network at the Institute of Applied Computer Science, Lodz University of Technology.

The crucial path of the proposed big data computational approach workflow is the classification stage. This is where the EBDDFS system serves as a data container for further data manipulations. The machine-learning based module (SVM-based classification in this example) handles additional operations on the retrieved data, such as mining for particular features or patterns resident in the data that are not otherwise observable or detectable. The procedure for model training and deployment is in accordance with SVM best practices [26]. Data from the EBDDFS are firstly divided into training and testing sets. Classification algorithms are then used, according to standards commonly applied in the fields of pattern recognition and stock market trade predictions [27], [28]. The performance of the system is usually analyzed with the aid of a comparison between the output classification decisions and expert-labeled data, which is treated as a benchmark to assess the accuracy of the system [29], [30].

Domain experts, whether professionals or experimental researchers, tend to expect certain specific flow patterns associated with to emergency events, such as pipeline blockages, under particular flow rig settings, and current flow conditions. Nonetheless, little has been published beyond standard, normal operation flow regime maps [10], [31]. This topic has been addressed previously only in [32], which investigated the question of possible similarities in the flow regime patterns of bulk solid pneumatic conveyors in pre-emergency states. Unfortunately, no quantitative evidence was provided to support the conclusions of the study. It was conducted with the aid only of a visual data analysis system for joint comparison of ECT preblockage data [33]. The current approach extends this methodology with a computational system that could provide a basis for industrial applications.

In summary, so far most historic data related to process behavior has not been utilized [5], [9]. To our knowledge, there has been no attempt to jointly analyze ECT data originating in different experimental settings [10]. Using methods taken from the field of big data, we propose a new data-driven workflow, with the main aim of improving the interpretation of ECT data [9], [35]. This contextual approach extends the knowledge base, which is of particular importance to domain experts [4], [5] and branches of industry employing bulk solids processing in general [1], [18]. Furthermore, the AGI element of the contextual approach first introduced in [32] is quantitatively tested here for the first time. Thanks to the proposed EBDDFS system, data manipulation and mining is less intensive in terms of computational resources than classical approaches to processing ECT data [11], [12], [21].

IV. Results and Discussion

The three-stage verification procedure implemented in this study is presented schematically in Fig. 4. First, the big data paradigm (BDD) is applied to experimental data from two distinct, remote locations, captured on two different flow rigs, in order to computationally verify the feasibility of its application to ECT data and pneumatic conveyor flow (see Section IV-A). Second, AGIs are used to both improve understanding of the process and prepare extra data input for populating the big dataset used in computational experiments (see Section IV-B). Third, all comparative verification is conducted using the datasets from stages one and two (see Section IV-C).

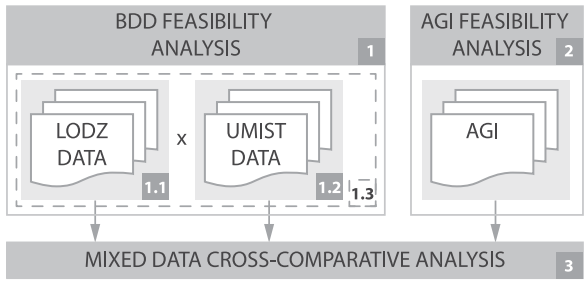

Fig. 4. Three-stage verification procedure: (a) BDD feasibility analysis for data from remote locations; (b) verification of AGI data; and (c) comparative analysis of data.

The first stage, BDD analysis, aims at proving the feasibility of the proposed concept, while the two later stages are designed to show how the proposed system could perform, as AGI contributes to both the "big data environment" and "expert knowledge" modules of the CDPM workflow model.

A. Data Pretreatment and Experimental Settings

In order to identify emergency situations, i.e., flow behavior that may lead to pipeline blockages, preliminary statistical data processing was conducted for each of the experimental dataset, preparing a set of features for classification. The following parameters were analyzed: mean concentration, autocorrelation SD, slug frequency and distribution, slug length, and slug length distribution. The data from EBDDFS were first pretreated from image sequences and raw data into discrete series of values, such as matrices and vectors. The selection of parameters was made based on previous work [11], [32], [31]. The minimal time window for analysis was set to 30 s. Tracking changes in these values within the assumed time window allows for classification of the labeled datasets using any standard binary classifier.

The very first tests were conducted using a Hadoop Mahout library and the Naïve Bayes classifier. Some initial results obtained using the Bayes classifier for pneumatic flow regime detection were presented previously. However, it was decided to change from Naïve Bayes to the SVM technique, since good results had been reported in a number of studies dealing with similar problems [19], [20], [34].

The following features were finally selected for SVM classification: mean concentration, average slug length, slug length distribution, slug occurrence frequency and distribution, and mean interslug gap SD. All of these parameters were calculated for a fixed 30 s time window. In order to avoid certain attributes from dominating others due only to their greater numerical ranges, linear scaling in the range of $[-1,1]$ was applied [26]. Tests and development were continued over five years, from Hadoop 1.0 to Hadoop 2.7 and in Mahout 0.11. While initial SVM development was conducted using MATLAB and LIBSVM, the final implementation was made using a custom Python-based plug-in for TomoKIS studio, a process tomography software suite [11].

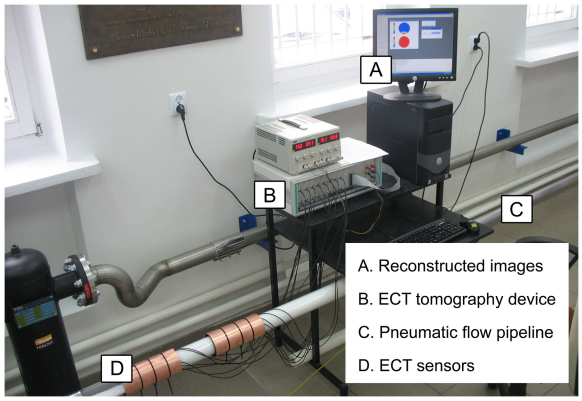

Fig. 5. Pneumatic conveyor flow installation in the Tom Dyakowski Process Tomography Lab, Lodz University of Technology.

TABLE I
BDD EXPERIMENTAL DATASETS: UMIST AND LODZ

	Dataset origin /label	ECT data acquisition speed [fps]	Approximate dataset length in frames/sec	No. of experimental datasets for normal operational conditions **L**	No. of experimental datasets ending in pipeline blockage **R1**	Total no. of experimental datasets **N**
1.	UMIST	50	14000/280	124 (L_U)	12 ($R1_U$)	136 (N_U)
2.	LODZ	66,7	20000/300	39 (L_L)	36 ($R1_L$)	75 (N_L)

The SVM models were developed for the radial basis function kernel with a coarse grid search for adjusting penalty (C) and kernel (γ) parameters [34].

The parameters/models were validated initially for 10-fold, and later for 5-fold cross-validation only, since the accuracy obtained exceeded 90% (maximum 96.74%) in all cases. The specific acceptable parameter values found were $C = 2$ and $\gamma = \{0.125; 0.008\}$. Performance was not analyzed in detail, in usual terms of sensitivity, specificity, positive predictive, and negative predictive values, since the focus was not the SVM algorithm itself. For this domain of application, only correct detection of possible pipeline blockages is important [10], [19].

The data were recorded during experiments using two different flow rigs in two different universities (the University of Manchester and Lodz University of Technology). The experimental data from these two locations will henceforth be referred to as UMIST and LODZ data, respectively. The experimental conditions were comparable in both cases: polyamide pellets of similar spherical shape were used as the conveyed medium; the pipeline diameter and airflow conditions were almost the same (in the same range of values). A segment of the pneumatic flow rig at Lodz University of Technology is pictured in Fig. 5.

No significant differences in data acquisition speed were observed for 50 fps and 67 fps, taking into account the dynamics of the pneumatic flow process of particulates [4], [7] (see Table I for details).

B. Initial Big Data-Driven Approach: Mining for Pipeline Preblockage Patterns in ECT Data From Two Different Flow Rigs

In the first stage of the investigation, a comparative analysis was conducted of the SVM classifications applied to the UMIST and LODZ experimental measurements. The main goal was to train the models for SVM-based classification of experimental datasets, to recognize those that tended to end with a pipeline "blockage," thereby detecting blockage threats in the analyzed data [11], [34]. Fig. 6 shows a simplified schema of the three combinations of data employed in the proposed verification approach. Results for model training and performance are given in three combinations: for distinct UMIST data (denoted as 1.1 in Fig. 6), for LODZ data (1.2 in Fig. 6) and for a MIXED dataset (1.3 on Fig. 6), using an aggregate of UMIST, and LODZ data together.

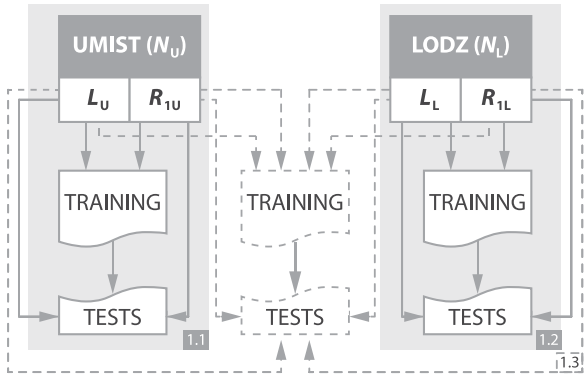

Fig. 6. Verification of experimental datasets: (1.1) UMIST data only; (1.2) LODZ data only, and (1.3) UMIST and LODZ data combined.

1) Model Training and Tests: Model training and testing was conducted for several different input dataset arrangements. Data gathered from EBDDFS for training constituted only a fraction (usually about 2/3) of all data (the other 1/3 was taken for model testing). As the total number of available experiments in this case is $N = 211$ (calculated as $N = N_U + N_L$), the fraction taken for model training is $K_1 = 140$. First, training was conducted in the following four configurations:

1) Full experimental datasets (from full length measurement sequences, i.e., 300 or 280 s for LODZ data and UMIST data, respectively);
2) 120 s measurement sequences corresponding to the very final 2 min of the flow period;
3) 60 s measurement sequences corresponding to the very final minute of the flow period; and
4) 30 s measurement sequences, corresponding to the final 30 s of the recorded flow period.

For each training dataset length, three different compositions of datasets were considered. Table II presents the composition of these datasets. They include experimental datasets ending with pipeline blockages. For example, row 1A shows classification results from a model trained with $K_1 = 140$ experimental datasets (out of a total $N = 211$ datasets from both the LODZ and UMIST sets). The fraction of "blockage" datasets is 16 and all came from LODZ experiments ($R1_L = 32$ "blockage" experiments available in LODZ data). This particular combination is indicated in the third column of Table II by the "16/32 LODZ" entry. Row 1B gives the results for UMIST data (4/40 UMIST) and so on.

2) Discussion of Outcomes: Each of the rows in Table II was populated with the average of five different results records. This means there were always five distinct training/testing composition sets, drawn randomly yet preserving a fixed ratio of R1/L

TABLE II
MODEL VERIFICATION FOR UMIST AND LODZ DATA

	Training set 140 (out of 211)		Classification results (correct [%] & SD)					
			UMIST data		Lodz data		Mixed (U+L)	
	Length of training dataset	Training set composition	%	SD	%	SD	%	SD
1A	Full meas.	16/32 LODZ	63.58	8.32	89.87	3.13	79.72	4.59
1B	Full meas.	4/40 UMIST	67.37	17.13	55.24	10.04	60.16	9.38
1C	Full meas.	16/46 MIXED	79.02	5.45	86.54	2.34	89.35	4.04
2A	Last 120 s	16/32 LODZ	82.10	8.82	91.43	3.49	86.63	5.59
2B	Last 120 s	4/40 UMIST	68.23	18.41	48.93	8.02	65.54	13.29
2B	Last 120 s	16/46 MIXED	84.32	6.51	87.76	5.69	89.75	5.41
3A	Last 60 s	16/32 LODZ	62.88	9.08	90.12	3.64	82.83	6.68
3B	Last 60 s	4/40 UMIST	49.12	17.29	44.02	20.91	69.16	14.21
3C	Last 60 s	16/46 MIXED	75.66	6.47	84.34	5.93	90.21	4.49
4A	Last 30 s	16/32 LODZ	57.53	9.92	93.23	2.91	79.42	9.37
4B	Last 30 s	4/40 UMIST	65.03	12.31	51.21	17.01	68.44	16.41
4C	Last 30 s	16/46 MIXED	79.31	7.76	76.73	7.83	81.01	7.04

from the entire set of experimental data. Fully randomized training sets gave poor results (c.a. 50% accuracy and below), due to the small number of R1, especially in UMIST data. The standard deviation (SD) was then calculated for these sets of five. During the stage of supervised learning, the model was trained with data manually labeled by the expert as falling into one of two categories: "blockage" (measurement datasets recorded just before a blockage in the pipeline) versus "other" (measurement datasets recorded under normal operating conditions).

The results presented in Table II show that the proposed approach performs at least sufficiently well with the experimental data to reach more than 90% accuracy. The best results for each combination are marked with a grey background. Classification results for the LODZ training data tested against LODZ (LODZ-LODZ) test data were better than for UMIST test data (LODZ-UMIST). Classification results for MIXED training data tested against UMIST (MIXED-UMIST) test data performed better than LODZ-UMIST and remained between LODZ-UMIST and LODZ-LODZ, as may be anticipated. The MIXED-MIXED configuration performed well in all cases and achieved in the best results around 90% accuracy. Models trained with 120 s as well as 60 s datasets showed better results than models trained with full and 30 s datasets. Distribution manifested by SD had smaller values for models trained with full measurement datasets and demonstrated the greatest variation for models trained with 30 s experimental datasets.

The most significant observation is that the MIXED-MIXED configuration performed consistently better than UMIST-UMIST and was very close or outperformed LODZ-LODZ arrangements (which were significantly better, by about 8%, only for 30 s datasets). This proves that the big data approach enables joint analysis of normally disconnected data (originating in remote locations in different experimental settings), bringing novel benefits and raising the quality of ECT industrial process data interpretation.

C. AGI Tests

The second stage of the verification phase employed an artificially introduced obstacle for contextual processing of data (see Fig. 5). The contextual approach based on AGI was first introduced in [32]. However, a proof-of-concept study, it did not conduct any extensive tests using AGI, either to reveal its utility for CDPM or provide any quantitative results.

TABLE III
MODEL VERIFICATION FOR AGI

	Training set 71 (out of 107)		Testing set 36 (out of 107)	Classification results (correct [%] & SD)	
				Lodz data	
	Length of training dataset	Training set composition	Testing set composition	%	SD
1A	Full measurement set	12/27/71	15/27/36	89.32	4.03
1B	Full measurement set	19/27/71	8/27/36	92.25	3.34
2A	Last 120 seconds	12/27/71	15/27/36	91.60	4.97
2B	Last 120 seconds	19/27/71	8/27/36	95.87	5.02
3A	Last 60 seconds	12/27/71	15/27/36	94.23	4.02
3B	Last 60 seconds	19/27/71	8/27/36	97.17	3.95
4A	Last 30 seconds	12/27/71	15/27/36	94.02	5.38
4B	Last 30 seconds	19/27/71	8/27/36	95.61	5.23

This study focused on an investigation of slug flow in the horizontal section of an industrial pneumatic conveyor. The settled layer occurs between two consecutive material aggregates (in contrast to plug flow, when looking at the vertical sections of pipeline installations). Therefore, the proposed modification forces a specific flow regime that is relevant to stable slug flow yet quite close to unstable slug flow (when looking at the experimental flow regime map) [10]. In other words, this modification alters the properties of the flow installation in such a way that it is easier to obtain a larger settled layer between the moving slugs. More details of the flow regime map and the rationale behind this routine can be found in [31].

The main experimental setup parameters were kept the same as in the previous experiments labeled LODZ (this time labeled AGI for ease of distinction). The data acquisition speed was 66.7 fps for the approximate duration of each experimental set of 300 s. The total number of $O = 107$ experiments including $R2 = 27$, which ended with a pipeline blockage was determined. As in the previous case, model training and testing were conducted for several different input dataset configurations. The fraction of experimental data for model training was this time $K_2 = 71$ out of $O = 107$. The same four configurations (full experimental datasets, 120, 60, and 30 s) were examined, and two different compositions of datasets are taken into account for each configuration. These are shown in Table III as pairs of rows, as in Table II. Consistent with the experimental results presented in Table II, each of the rows in Table III is populated with the average of five different results records. Five distinct training/testing composition sets were drawn randomly from the set of experimental datasets (yet preserving the fixed ratio of "blockage" datasets drawn from $R2$ to the rest of the datasets, as given in the "Training set composition" column). The SD was calculated for those sets of five. Other conditions, such as classification goals and preprocessing criteria, were the same as for Table II.

It can clearly be observed that the larger the fraction of $R2$ ("blockage") datasets in the model training data, the more accurate the results of the classification. This regularity is manifested by the differences between the classification results for

TABLE IV
SUMMARY OF EXTENDED MIXED COMPARATIVE DATASETS

	Label for Table 5	Dataset origins	Total no. of experimental datasets for safe operational conditions **ST**	No. of experimental datasets ending in pipeline blockage **RT**	Total no. of experimental datasets for model training **TT**	Total no. of experimental datasets **TE**
1.	BDD	UMIST & LODZ	163	48	140	211
2.	AGI	Lodz	80	27	71	107
3.	MIXED (U+L+A)	UMIST & LODZ + Lodz	243	75	211	318

corresponding pairs of rows A and B. However, it can be estimated that a content of $R2$ datasets below 10% (i.e., several fewer than 10 in absolute values) in the training set may lead to underfitting of the model. In addition, the SD factor is greater for larger $R2$ fractions in the model training data. To sum up, the overall results are satisfactory, since classification is maintained, in most cases, at 90% accuracy. Therefore, it can be concluded that AGI data constitute a promising option as an input for CDPM models. This method can be applied to enlarge available data samples, which is important for modeling and analysis of infrequent events, sparsely represented within datasets.

D. Extended Mixed Data Comparative Analysis

Finally, extensive comparative tests were performed of input datasets in different mixed configurations, in order to verify the feasibility of the proposed approach.

1) Model Training and Tests: Table IV shows details of the input data taken from EBDDFS. Model training and testing were conducted for several corresponding input dataset arrangements (full length, 120, 60, 30 s).

The performance of the proposed approach was verified using a combination of three differently trained models, tested with four permutations of the data. First, the model was trained for LODZ and UMIST experimental data (indicated as BDD in the "Training set" column of Table V, rows 1–4). Second, AGI-based data was used for training (indicated as AGI in the "Training set" column of Table V, rows 5–8). Next, the model was trained with a mix of both LODZ and UMIST plus AGI data (indicated as MIXED (U+L+A) in the "Training set" column of Table V, rows 9–12). Finally, model testing was conducted for test datasets drawn from the following four sources: 1) LODZ; 2) UMIST; 3) AGI, and 4) MIX of all. Note that 1) and 2) are elements of the BDD collection.

This time, only a single arrangement of the input datasets was applied for each training dataset period (i.e., there was no fixed ratio for the RT/ST fraction within the training set). However, each row shows results for ten distinct training/testing sets, drawn randomly from the entirety of the available data. This means that columns presenting the results show the average percentage. The SD was then calculated for each set of ten. The other conditions, such as classification goals and preprocessing criteria, as well as the number of datasets used for model training, were kept the same as for Tables II and III.

2) Mixed Data Tests Results Discussion: The data used for model training in Table V were randomly sampled from EBDDFS. As a consequence, there is no fixed fraction of "blockage" datasets within each training dataset used for model training, and the RT/TT ratio is therefore accidental. The reason behind this totally randomized selection of training datasets was the desire to more closely approximate future scale-up scenarios, in which it is assumed there will be a massive amount of available data to draw on. The results of this randomized model training (see Table V) are generally not as good as the classification results obtained using carefully prepared training data (see Tables II and III). For example, when looking at the BDD results for LODZ in Table II, accuracy reached 85%, whereas in Table V it drops for MIXED by about 16–17%. However, the SD remains at a comparable level in both cases, which in turn indicates that the general level of variation was much the same.

The AGI-based model performed well, except in the worst case with UMIST data. With UMIST data, accuracy dropped to almost 50%, yet some reasonable decrease in accuracy may be logically expected in this particular case. On the other hand, the AGI model obtained approximately 80% accuracy for MIXED data. The MIXED data results were the best or equally as good as the others in most cases. For both LODZ + UMIST and for MIXED (U + L + A) data, up to 90% accuracy was obtained. At the same time, the SD was also the greatest for MIXED data. It reached a maximum of 12% and dropped by 4–5% for separate classifications of UMIST and LODZ data, for which the accuracy was lower in general. An interesting feature is that the SD was smallest for AGI with MIXED data, while it was greater for LODZ + UMIST BDD and the largest for fully MIXED (U + L + A).

The most important results can be seen in the lower right corner of Table V, which show fully MIXED data tested with a model trained on fully MIXED data. This is the most accurate quadrant of the table; better than the other configurations by at least 2–3% and more on average. There are notable values, with accuracy exceeding 92% for 120 s sequences tested with 120 s and full models with an SD of just 7%. The 60 s and 30 s tests reached almost 90% and 87% accuracy, respectively, with SD within the range of 7–10%.

The best results in terms of both accuracy and stability were usually obtained for 120 s segments. A visual comparison is shown in Fig. 7 for the 120 s datasets. These results are taken from Table II (BDD analysis for carefully prepared training sets) and supplemented with others gathered using the same methodology, but for totally randomized training sets. The two types of results are assembled with results for MIXED (U + L + A) data.

As can be seen from Fig. 7, the results for MM (last bar on the right) obtained with totally randomized learning compositions yet composed of fully mixed data were at least equally as accurate (c.a. 92%) as those for LODZ-LODZ (c.a. 91%) and MIXED-MIXED (c.a. 89%). The darker bars show L-LODZ and M-MIXED sections. It is worth emphasizing that the MM results were even better than those shown using the lighter bars for the same compositions: LODZ-LODZ (c.a. 83%) and MIXED-MIXED (c.a. 81%) (L-LODZ and M-MIXED sections, respectively). Yet, the greatest advantage of the proposed methodology can be seen by comparing the MM and lighter bars in general, especially the MM and UMIST-UMIST data. These are randomized compositions of training sets for very limited blockage

TABLE V
EXTENDED COMPARATIVE VERIFICATION FOR MIXED DATA

	Training set		Classification results (correct [%] and standard deviation SD)																													
			UMIST data								LODZ data								Mixed (U+L)								MIXED (U+L+A)							
			Full	SD	120s	SD	60s	SD	30s	SD	Full	SD	120s	SD	60s	SD	30s	SD	Full	SD	120s	SD	60s	SD	30s	SD	Full	SD	120s	SD	60s	SD	30s	SD
1	BDD 140 (out of 211)	Full	57.39	3.57	60.34	3.92	62.02	3.78	65.31	4.12	70.18	2.93	73.06	3.11	72.20	4.92	71.15	5.06	68.47	1.92	74.57	2.08	72.20	2.11	70.06	2.10	72.19	2.46	88.91	2.26	69.44	2.50	68.05	2.49
2		120 s	58.16	3.53	60.34	3.52	64.55	5.40	66.30	4.63	74.58	3.16	78.77	3.28	85.53	4.65	78.69	4.28	75.08	1.99	82.12	2.21	73.39	2.31	74.68	2.39	76.04	2.11	86.07	2.34	85.32	2.61	84.66	2.53
3		60 s	60.70	5.05	61.66	4.14	63.34	5.12	66.21	4.93	72.43	3.29	75.21	3.03	86.55	2.82	77.53	3.61	70.06	2.13	73.46	2.08	84.66	1.95	73.76	2.39	69.40	2.31	68.86	2.80	79.95	2.79	74.57	2.64
4		30 s	61.13	3.91	63.00	5.62	66.59	4.90	67.64	5.21	75.94	5.60	77.24	4.54	80.38	3.47	84.30	3.79	76.85	2.16	74.13	2.09	81.32	2.09	79.34	2.24	71.22	2.02	74.38	2.92	75.08	2.68	79.82	2.81
5	AGI 71 (out of 107)	Full	53.08	5.39	53.94	6.03	59.33	7.85	53.76	6.40	72.04	3.61	71.32	3.78	74.05	3.66	70.73	3.72	71.22	3.76	80.20	3.90	69.55	3.84	69.53	2.14	71.22	1.63	72.72	1.90	71.11	1.72	69.55	2.61
6		120 s	54.00	4.85	55.06	4.91	54.75	6.91	54.23	7.28	68.33	3.61	80.22	3.28	84.16	4.04	76.64	3.67	76.30	3.91	69.14	4.12	74.12	4.17	72.12	3.92	69.45	1.73	70.10	1.59	75.49	1.81	80.99	2.19
7		60 s	54.93	6.31	54.41	5.72	53.51	6.06	56.72	7.12	71.21	3.72	76.33	4.12	84.23	3.76	76.02	3.83	74.57	4.12	79.82	4.62	82.53	4.23	79.34	3.88	74.52	1.87	77.24	1.77	80.45	4.41	75.83	2.31
8		30 s	52.44	6.87	53.27	7.47	53.82	6.41	55.06	6.58	76.85	4.05	75.36	4.01	81.66	3.84	83.08	3.90	72.00	5.03	72.12	4.28	70.11	4.30	78.69	4.21	72.96	1.90	76.37	2.11	80.77	5.90	82.53	2.40
9	MIXED (U+L+A) 212 (out of 318)	Full	61.12	4.97	58.32	6.14	62.33	5.83	57.89	5.89	79.86	3.48	81.22	3.61	75.08	2.92	69.55	4.20	74.03	6.37	77.90	6.59	84.33	6.92	76.29	7.04	86.21	8.28	92.85	7.97	76.22	3.91	79.09	8.69
10		120 s	62.05	4.46	65.72	5.76	64.56	6.01	60.40	6.83	80.53	3.71	94.04	3.59	93.46	3.38	74.78	3.98	76.31	6.72	84.05	5.98	89.14	7.38	74.52	6.92	74.12	9.32	92.01	6.54	88.74	7.23	81.42	7.94
11		60 s	60.19	5.60	62.65	5.14	63.01	5.79	58.73	7.09	69.48	4.27	72.11	3.80	96.01	3.55	89.23	4.03	70.67	6.82	79.58	6.82	77.60	7.44	77.03	7.63	63.29	9.32	73.32	8.99	89.04	10.06	86.47	9.42
12		30 s	59.72	6.93	58.88	6.17	59.12	5.94	60.79	5.94	68.22	3.85	71.67	4.27	73.83	4.11	79.35	3.94	74.10	6.59	72.04	6.79	78.61	7.50	78.43	7.48	81.03	8.77	79.52	9.39	81.06	12.11	86.94	8.62

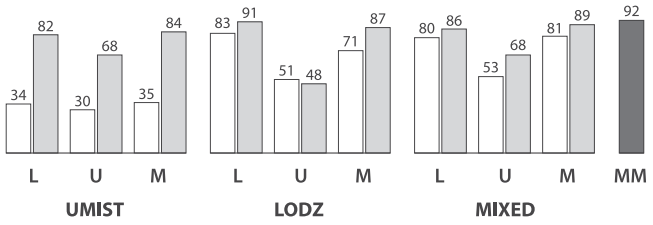

Fig. 7. Visual comparison of results for 120 s segments. L, U, and M stand for Lodz, UMIST, and MIXED testing datasets, while the words below describe training datasets. Each left-hand bar indicates results for randomized training data and each right-hand bar (darker) indicates results from Table II and carefully prepared training data. The last, darkest, bar on the right with MM subscript shows the results for cross-comparative mixed data from Table V. Numerical values on top of the bars indicate percentage accuracy.

TABLE VI
TOWARD DETECTING PROCESS THREATS

	Verification with data shifted from the end by lenght V					
Models trained with datasets of length L	30s	SD	60s	SD	120s	SD
30s	83.14	9.56	68.13	7.13	58.10	12.25
60s	86.32	8.44	82.60	6.72	76.59	9.28
120s	69.82	6.87	79.45	8.68	70.61	11.33

data, which result in the inability to correctly recognize the process regime at all (see c.a. 30% results for UMIST randomized training datasets). Similar outcomes were observed for 30 s and 60 s segments as well as for Full datasets. The best results varied by 1–2% for MM and LODZ-LODZ data, but the overall pattern remained the same.

Based on these results, the proposed CDPM concept has been verified. In particular, the results obtained for cross-comparative experiments with fully MIXED data validate the model's efficacy for ECT-monitored pneumatic conveying of bulk solids.

E. Toward Predicting Process Threats

The final experiment to be presented in this paper was designed to investigate whether the validated CDPM approach could be further developed to predict process threats. Models were trained with 30, 60, and 120 s flow fragments (as shown in Table V). They were tested on data composed of flow fragments taken from earlier moments of flow, before a "blockage." This means that these fragments were associated with flow periods shifted back in time with respect to the end of the dataset. Note that the data were randomly drawn from EBDDFS (as shown in Table V). Table VI shows the results of classification for test data taken from flow fragments ending 30, 60, and 120 s before the flow end (or before the "blockage" occurred), respectively.

The results show that the best performance for detecting process threats was achieved by 60 s-trained models (82–86% accuracy), especially when the data was shifted 30 s or 60 s before the end of the flow. Similar scores were obtained for the 30 s-trained model, but only for data shifted 30 s ahead. The 120 s-trained model performed acceptably (∼80%) with a 60 s shift only. The results for 60 s/60 s showed the lowest SD, but still reached SD = 6.72. The greatest SD = 12.25 was found for 30 s/120 s. Generally, the distribution of results in Table VI is rather heterogeneous, and much more thorough research is required. However, there is clearly potential for the future development of these methods for predicting pre-emergency states.

V. DIRECTIONS FOR FUTURE WORK

The initial experiment to predict process failure presented in Section VI needs further testing, including using larger data samples, before it can be considered an efficient fault forecasting method. When the data "velocity" factor increases and whenever the requirement for real-time dynamic data processing becomes important, the implementation of recent advancements in big data will be necessary, such as lambda architecture [21] and parallel analytics [36]. The methodology presented in this paper is scalable, thanks to the proposed EBDDFS. However, some specific improvements in terms of the computational environment are possible, by switching to state-of-the-art big data processing methods, such as those proposed in [25].

The proposed CDPM model, which aggregates multiple data processing methods, needs more research for other tasks, such as regularization of inverse problems and direct determination of process parameters (quantitative analysis). An interesting area that could be explored in future is definition of the specific kernel parameters used in such prediction methods for SVM classification. This should be correlated with analysis of the SD of results obtained for different time windows. It seems that longer periods (such as 120 s) have less variance, so γ parameters with lower values fit better, while higher γ values fit better

prediction models with shorter time windows (such as 30 s), which have higher variance in comparison to full length data.

The most promising avenue for future development is combination of the computational approach presented here with human-intelligence-based data processing methods, such as crowd sourcing analysis of scientific data [29], [33], in order to develop flexible and more effective data processing systems. Tests could be conducted to adjust the measurement datasets obtained from processes with various dynamics and using different sampling frequencies. This might involve additional preprocessing with respect to the time domain and the development of methods for cross-modality and data fusion in future distributed systems [12], [18].

VI. Conclusion

This work presents the first attempt to use a big data-driven, contextual data computational environment for ECT-monitored industrial process modeling. This can be achieved with the aid of a dedicated data processing system, such as the EBDDFS presented here, which helped improve the efficiency of measurement dataset analysis for a pneumatic conveying process. The primary impact of this paper is derived from the CDPM model for the verification of contextual data, which is aided by computational experimental tests using SVM-based supervised learning models. The big data-driven methodology presented in this work helps, in particular, to jointly analyze datasets that are normally treated as incoherent. Such data may come from diverse origins, but as long as it retains a common denominator it is possible to be collectively interpreted, employing the methods described in this work.

The other key contribution of this research is verification of the use of AGI as a legitimate element of the CDPM concept. It was shown that AGI can be a reliable resource and enhance the available knowledge base regarding infrequent phenomena. Moreover, the proposed CDPM model, which uses the big data approach and SVM classification, verified a framework for future data mining of measurement records, as well as for predicting process behavior. The presented approach has been shown to be scalable in terms of data volumes. Demonstrated for ECT and the pneumatic conveying process, it is generalizable to other process tomography modalities, as well as to other multiphase processes [19], [22].

Acknowledgment

The author would like to express his gratitude for continuous support acquired from Krzysztof Grudzień; for insightful machine learning discussions with Tomasz Jaworski; and for valuable comments concerning the art of writing received from Morten Fjeld.

References

[1] C. R. Woodcock and J. S. Mason, *Bulk solids handling: An introduction to the practice and technology*. New York, NY, USA: Springer, 2012.

[2] D. Neuffer *et al.*, "Control of pneumatic conveying using ECT," in *Proc. 1st World Congr. Ind. Process Tomography*, 1999, vol. 14–17, pp. 71–76.

[3] D. Sankowski and J. Sikora, Eds., *Electrical capacitance tomography: Theoretical basis and applications*. Electrotechnical Inst., Press, Warsaw, 2010, p. 308.

[4] W. Zhang, C. Wang, W. Yang, and C. H. Wang, "Application of electrical capacitance tomography in particulate process measurement – A review," *Adv. Powder Technol.*, vol. 25, no. 1, pp 174–188, 2014.

[5] A. Romanowski, K. Grudzien, R. G. Aykroyd, and R. A. Williams, "Advanced statistical analysis as a novel tool to pneumatic conveying monitoring and control strategy development," *Particle Particle Syst. Characterization* , vol. 23, no. 3/4, pp. 289–296, 2006.

[6] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, reconfiguration methods," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 3, pp. 636–653, May 2010.

[7] V. Mosorov, "Phase spectrum method for time delay estimation using twin-plane electrical capacitance tomography," *Electron. Lett.*, vol. 42, no. 11, pp. 630–632, 2006.

[8] J. Kryszyn, P. Wróblewski, M. Stosio, D. Wanta, T. Olszewski, and W. T. Smolik, "Architecture of EVT4 data acquisition system for electrical capacitance tomography," *Measurement*, vol. 101, pp. 28–39, 2017.

[9] S. Yin, S. X. Ding, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.

[10] R. A. Williams, "Landmarks in the application of electrical tomography in particle science and technology," *Particuology*, vol. 8, pp. 493–497, 2010.

[11] R. Banasiak, R. Wajman, T. Jaworski, P. Fiderek, P. Kapusta, and D. Sankowski, "Two-phase flow regime three-dimensonal visualization using electrical capacitance tomography – algorithms and software," *IAPGOS*, vol. 7, no. 1, pp. 11–16, 2017.

[12] T. Rymarczyk, P. Tchórzewski, P. Adamkiewicz, K. Duda, and J. Szumowski, "Practical implementation of electrical tomography in a distributed system to examine the condition of objects," *IEEE Sensors J.*, vol. 17, no. 24, pp. 8166–8186, Dec. 2017.

[13] K. Grudzień, Z. Chaniecki, and L. Babout, "Study of granular flow in silo based on electrical capacitance tomography and optical imaging," *Flow Meas. Instrum.*, to be published.

[14] G. Vert, A. Chennamaneni, and S. Iyengar, "Potential application of contextual information processing to data mining." *Inf. Knowl. Eng.*, 2010, pp. 317–325.

[15] A. Schmidt, M. Beigl, and H-W. Gellersen, "There is more to context than location," *Comput. Graph.*, vol. 23, no. 6, pp. 893–901, 1999.

[16] C. Bettini *et al.*, "A survey of context modelling and reasoning techniques," *Pervasive Mobile Comput.*, vol. 6, no. 2, 161–180, 2010.

[17] C. Bolchini, C. A. Curino, E. Quintarelli, F. A. Schreiber, and L. Tanca, "A data-oriented survey of context models," *ACM SIGMOD Rec.*, vol. 36, no. 4, pp. 19–26, 2007.

[18] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE Commun. Surveys. Tuts.*, vol. 16, no. 1, pp. 414–454, Jan.–Mar. 2014.

[19] P. Fiderek, J. Kucharski, and R. Wajman, "Fuzzy inference for two-phase gas-liquid flow type evaluation based on raw 3D ECT measurement data," *Flow. Meas. Instrum.*, vol. 54, pp. 88–96, 2017.

[20] B. Franke *et al.*, "Statistical inference, learning and models in big data," *Int. Statist. Rev.*, vol. 84, pp. 371–389, 2016.

[21] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, Challenges, and Future Research Topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017.

[22] V. Moreno *et al.*, "Applicability of big data techniques to smart cities deployments," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 800–809, Apr. 2017.

[23] V. Chang, "Towards a big data system disaster recovery in a private cloud," *Ad Hoc Netw.*, vol. 35, pp. 65–82, 2015.

[24] M. Zaharia *et al.*, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[25] P. Basanta-Val, N. C. Audsley, A. J. Wellings, I. Gray, and N. Fernández-García, "Architecting time-critical big-data systems," *IEEE Trans. Big Data*, vol. 2, no. 4, pp. 310–324, Dec. 2016.

[26] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification." Tech. Rep., Department of Computer Science, National Taiwan University, pp. 1–16, 2003.

[27] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern Recog.*, vol. 35, no. 4, pp. 945–965, 2002.

[28] M. Skuza and A. Romanowski, "Sentiment analysis of twitter data within big data distributed environment for stock prediction," in *Proc. FedCSIS*, 2015, pp. 1349–1354.

[29] C. Chen *et al.*, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans Intell. Syst. Technol.*, vol. 7, no. 4, 2016, Art. no. 52.

[30] A. G. M. Soares, C. G. Resque dos Santos, and S. Mendonca, "A review of ways and strategies on how to collaborate in information visualization applications," in *Proc. 20th IV Conf.*, Lisbon, Portugal, 2016, pp. 81–87.
[31] K. Ostrowski, S. P. Luke, M. A. Bennett, and R. A. Williams, "Real time visualisation and analysis of dense phase powder conveying," *Powder Technol.*, vol. 102, pp. 1–13, 1999.
[32] A. Romanowski, K. Grudzień, and P. Woźniak, "Contextual processing of ECT measurement information towards detection of process emergency states," in *Proc. 13th Int. Conf. Hybdrid Intell. Syst.*, 2013, p. 292–298.
[33] I. Jelliti, A. Romanowski, and K. Grudzien, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *Proc.Federated Conf. Comput. Sci. Inf. Sys.*, 2016, vol. 8, pp. 1613–1619.
[34] C. Cortes and V. Vapnik, *Support-Vector Networks Machine Learning*. Norwell, MA, USA: Kluwer, 1995, vol. 20, pp. 273–297.
[35] V. Chang, "Towards data analysis for weather cloud computing," *Knowl.-Based Syst.*, vol. 127, pp. 29–45, 2017.
[36] P. Basanta-Val, "An efficient industrial big-data engine," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1361–1369, Apr. 2018.

Andrzej Romanowski received the Ph.D. degree in computer science, addressing the issue of algorithms for inverse problem solving in electrical industrial process tomography, from Lodz University of Technology (TUL), Lodz, Poland, in 2008.

He is currently an Assistant Professor at the Institute of Applied Computer Science and is the Vice Dean of the Faculty of Electric, Electronic, Computer, and Control Engineering, TUL. He is a pioneer in problem-based learning and design thinking at the TUL. His current research focuses on data science for industrial and medical applications. He explores methods and tools that help utilize big data to reveal hidden patterns, new connections, and extra value. Additionally, as a practice-oriented academic with industry experience, he is primarily interested in stimulating user development through interactive systems. His goal is to establish coherent practices that use human–computer interaction (HCI) in industrial contexts to create knowledge and innovation.