Big Data Conceptual Modelling in Cyber-Physical Systems

Ada Bagozi[a], Devis Bianchini[*,a], Valeria De Antonellis[a], Alessandro Marini[a], Davide Ragazzi[a]

[a] Dept. of Information Engineering, University of Brescia, Italy

Abstract. *Management of large volumes of data, collected from modern Cyber-Physical Systems, is calling for models, tools and methods for data representation and exploration, in order to capture relevant properties of physical objects, and manage them in the cyber-space. In this context, the impact of big data disruptive characteristics (namely, volume, velocity and variety) on data modelling and information systems design needs further investigation. In particular, data exploration is assuming an ever growing relevance, being a way users/operators can learn from data by inspecting it according to different perspectives. In this paper, we use conceptual modelling for (big) data exploration in a dynamic context of interconnected systems. We rely on a multi-dimensional model, that is suited for properly providing data organization for exploration. Furthermore, we propose a model-driven approach that guides the design of multiple exploration strategies according to different objectives. The model-driven approach exploits a model of relevance, aimed at focusing the attention of the users/operators only on relevant data that are being explored. We describe the instantiation of the proposed concepts through some scenarios in the smart factory context, in order to show how conceptual modelling helps abstracting from implementation details and focusing on semantics of explored data.*

Keywords. Data Exploration • Conceptual Modelling • Big Data • Multi-Dimensional Data • Industry 4.0 • Cyber-Physical Systems

1 Introduction

A Cyber-Physical System (CPS) is characterized by the integration of physical devices (machines and sensors) with cyber components (computer, data and programs) to form a context-sensitive system apt to react to dynamic changes in real-world (Lee and Seshia 2017). CPS are widespread in several domains like smart grids, autonomous automobile systems, domotics, medical monitoring and, more recently, Industry 4.0 (Lee et al. 2015b). Relevant function in CPS is the collection of raw data from dynamic physical environments, integrated with many types of cyber-space resources, and its transformation into actionable knowledge in real-time. Data management is leading to new CPS capabilities and challenges: for example, in the Industry 4.0 scenario, data is emerging as a new industrial asset, creating opportunities for operations improvement and increased industrial value through the capitalization of immaterial assets, and promoting advanced functions like self-awareness, self-configuration and self-repairing of machines (Hou and Wang 2013). To this aim, models, tools and methods for the collection, organization and exploration of data are required in order to capture relevant properties of physical objects, and manage them in the cyber space through information/data management and analysis systems (Monostori 2014). In particular, data exploration is assuming an ever growing relevance, being a way users/operators can learn from data by inspecting it according to

* Corresponding author.
E-mail. devis.bianchini@unibs.it

different perspectives. Nevertheless, the disruptive characteristics of big data, namely, volume, velocity and variety, pose additional issues for those who are in charge of extracting knowledge from it.

In this context, conceptual modelling can play a fundamental role, given its capability of abstracting data representation from its implementation in physical systems by means of concepts, their properties and mutual relationships (Chen 1976; Fliedl et al. 2005; Karagiannis et al. 2016; Olivé 2007), in order to build information systems (Cabot et al. 2017). Embley and Liddle (Embley and Liddle 2013) expect conceptual modelling to address big data challenges by structuring information, making big data volume searchable: it may help to highlight the semantics of underlying data in a fast and automatic way, choosing the best representation to foster data exploration. On the other hand, velocity of data acquisition requires the integration of different models and techniques, apt to properly summarize data that are incrementally collected from monitored interconnected systems.

In this paper, we propose a conceptual model apt to provide a high level representation of a Cyber-Physical System through a set of "facets" or "dimensions", either flat or hierarchically organized. Aggregation of data according to different dimensions (e.g., time, monitored system), being related to the observed physical problems, can give proper semantics to the collected data. Moreover, multi-dimensional model enables data exploration by following the hierarchical structure of dimensions. The proposed multi-dimensional model is integrated with data summarisation techniques, in order to provide a synthetic representation over large volumes of data to be managed. Given the conceptual model, we define a model-driven data exploration approach, that relies on data relevance techniques, aimed to focus the attention of the user/operator on relevant data only and to guide multiple exploration strategies according to different objectives. We envisage the application of the proposed model-driven approach to some paradigmatic research scenarios in the smart factory context (Lee et al. 2015a), in order to show how conceptual modelling helps abstracting from implementation details and focusing on semantics of explored data. In particular, the first scenario concerns monitoring of a Cyber-Physical System for anomaly detection and adaptive recovery from damage conditions. The second scenario shows the potential utility of the approach for data-driven performance comparison across different physical systems. The two scenarios are conceived for smart factory maintenance operators; in the first scenario, operators may be interested in preventing downtimes of the monitored systems, while in the second one operators may want to understand which part of the monitored system isn't working properly in order to replace or repair it. With respect to exploratory data analysis (Tukey 1977) and Data Mining (Han and Kamber 2006), our approach aims at supporting exploration as a multi-step process, where the users/operators may iteratively improve focus on relevant data, by receiving suggestions based on the model of relevance. Compared to On Line Analytical Processing (Golfarelli and Rizzi 2009), we manage data that is incrementally collected, organized and analysed on-the-fly. Finally, with respect to traditional faceted search (Tunkelang 2009), we deal with high data volumes and velocity, that imply efficient techniques for storing and managing them.

In (Bagozi et al. 2017c) we introduced the summarisation and relevance techniques as ingredients to perform exploration of real time data in a dynamic context of interconnected systems. In (Bagozi et al. 2017b) we proposed IDEAaS (Interactive Data Exploration As-a-Service), a framework where innovative services are designed to enable data exploration. In (Bagozi et al. 2017a) we discussed the application of the multi-dimensional model, data summarisation and relevance evaluation techniques to support anomaly detection in collaborative systems in the context of Cyber-Physical Systems and Industry 4.0. This paper extends this research for what concerns the introduction of a model-driven approach based on the conceptual model. In particular, we discuss the application of the approach for performance comparison across different physical systems, introducing additional contributions with respect to anomaly detection issues and thus abstracting the characteristics of big data exploration over multiple scenarios.

The paper is organized as follows: Section 2 introduces the conceptual model with the help of a running example; Section 3 describes summarisation and relevance evaluation techniques engaged to support data exploration; in Section 4 we discuss the model-driven data exploration approach, taking into account different scenarios in the Industry 4.0 context; in Section 5 we highlight cutting-edge features of our approach compared to state of the art; finally, Section 6 closes the paper with some final remarks and future work.

2 Conceptual Model

2.1 Running example

To better explain concepts addressed in this paper, we will use the running example introduced in (Bagozi et al. 2017c). In the example, we considered an Original Equipment Manufacturer (OEM) producing multi-spindle machines for various industrial sectors: automotive, aviation, water industry, etc. A multi-spindle machine is a turning machine that allows multiple tools to cut pieces of material simultaneously and independently each other. A picture of the multi-spindle machine is shown in Figure 1. Each spindle is mounted on a unit moved by an electrical engine to perform X, Y, Z movements. The multiple spindles are carried in a precision rotating drum where raw material is positioned. The total number of operations needed to complete a manufacturing cycle are divided among the number of spindles, so that a cycle is completed with one full rotation of the drum. Each spindle is equipped by a cross-slide and end-slide tool. Tools are selected according to the instructions specified within the machine Part Program, that is, a set of instructions executed by the numerical control of the multi-spindle machine to manage its operations.

Spindle precision, working performances, minimization of tool breaks and machine downtimes are critical factors. Real time data collected from the multi-spindle machine concerns the spindle rotation as impressed by an electrical engine and its rotation speed as collected by the machine numerical control. For each spindle, we measure the velocity of the three axes (X, Y and Z) and the electrical current absorbed by each engine, the value of rpm (rotations per minute) for the spindle, the percentage of power absorbed by the spindle engine (charge coefficient). Hereafter, we will refer to the measured aspects as *features*. The aim of the OEM is to understand if it is possible to use real time data collected directly from the machine for monitoring the spindle rolling friction torque increase and the tool wear. With spindle rolling friction torque increase we refer to a specific behaviour of the spindle shaft that turns hard more and more due to different possible reasons: lack of lubrication and bearing wear that may lead to possible bearing failures. Tool wear monitoring is referred to possible tool usage optimizations in order to balance the trade-off between the number of tools used and the risk of breaking the tool during operations, that may lead to long downtimes. Spindle rolling friction torque increase and tool wear can be monitored by observing the spindle power absorption for similar rpm values. If a greater power absorption is detected, disregarding the tool that is being used, the spindle rolling friction torque increase could be identified as the possible anomaly that increases the energy request to perform the manufacturing operations. If the increase in absorbed power is related only to the usage of a particular tool, this can be recognized as a symptom of a possible exceeding tool wear. Therefore, machine components and tools, as well as time, represent multiple dimensions to perform data exploration. We will address such dimensions as first-class citizens of our conceptual model.

2.2 The model in a nutshell

The multi-dimensional model we propose for data exploration can be represented as an hypercube, as shown in Figure 2. In figure each node contains records of measures collected at a given time t_1. Measures are described through a timestamp and

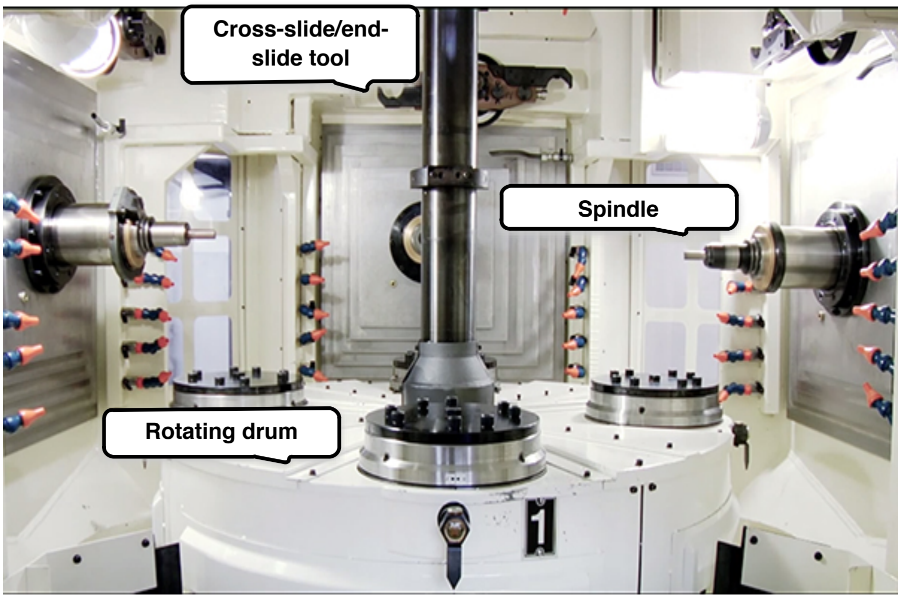


Figure 1: Multi-spindle machine considered for the running example.

the measured value. These measures are collected and organized according to several dimensions, such features/feature spaces and domain-specific-dimensions, representing hypercube axes.

Features and feature spaces. Measures are associated to `Features`, that is, the measured quantities, in turn collected into `Feature_Spaces`. A feature space conceptually represents a set of related features, that are jointly measured to observe a physical phenomenon. Multiple feature spaces might be observed, and the observation of a feature might be useful to monitor more than one feature space. In the considered running example, features are the speed over the three axes X, Y and Z, the electrical current, the value of spindle rpm and the percentage of absorbed power. The set composed of spindle power absorption and rpm features is an example of feature space used to monitor spindle rolling friction torque increase and tool wear. They can be observed by monitoring the spindle power absorption.

Domain-specific dimensions. These dimensions group together collected measures according to "facets", such as the observed machine or the tool used during manufacturing. Also domain-specific dimensions can be organized through hierarchies: tools can be aggregated into tool types (`Tool:Tool_type` axis in Figure 2), while monitored physical components (e.g., spindles) can be aggregated into the machines they belong to, in turn organized into plants and enterprises. Other dimensions might be the part program that is being executed by the numerical control of the monitored system and the working mode (G0, fast movement of the spindle, e.g., to catch the tool, or G1, slow movement of the spindle during the manufacturing). Among analysis dimensions, we always consider time.

In Figure 2 a subset of all possible dimensions is showed, achieved by slicing over the feature space $\mathtt{fs}_1$ and the part program $\mathtt{pp}_2$. For example, the node identified as "`A`" contains the records of measures collected at time t_1 for multi-spindle machine $\mathtt{m}_1$ (spindle $\mathtt{s}_3$), that is using tool $\mathtt{u}_3$, during working mode G_0, considering features in the feature space $\mathtt{fs}_1$, while running the part program $\mathtt{pp}_2$.

Collected measures could be meaningfully compared over domain-specific dimensions: for example, it makes sense to compare measures across different components/machines, for performance

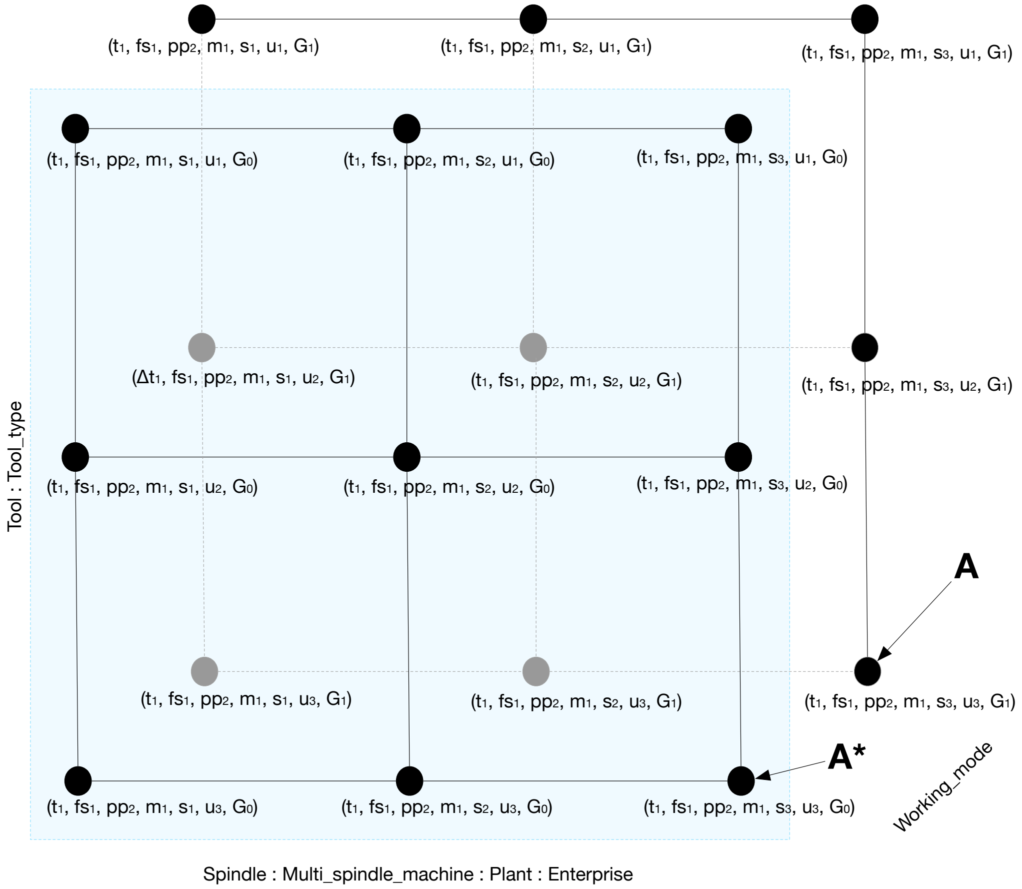

$(t_1, fs_1, pp_2, m_1, s_1, u_1, G_1)$
$(t_1, fs_1, pp_2, m_1, s_2, u_1, G_1)$
$(t_1, fs_1, pp_2, m_1, s_3, u_1, G_1)$
$(t_1, fs_1, pp_2, m_1, s_1, u_1, G_0)$
$(t_1, fs_1, pp_2, m_1, s_2, u_1, G_0)$
$(t_1, fs_1, pp_2, m_1, s_3, u_1, G_0)$
$(\Delta t_1, fs_1, pp_2, m_1, s_1, u_2, G_1)$
$(t_1, fs_1, pp_2, m_1, s_2, u_2, G_1)$
$(t_1, fs_1, pp_2, m_1, s_3, u_2, G_1)$
Tool : Tool_type
$(t_1, fs_1, pp_2, m_1, s_1, u_2, G_0)$
$(t_1, fs_1, pp_2, m_1, s_2, u_2, G_0)$
$(t_1, fs_1, pp_2, m_1, s_3, u_2, G_0)$
A
$(t_1, fs_1, pp_2, m_1, s_1, u_3, G_1)$
$(t_1, fs_1, pp_2, m_1, s_2, u_3, G_1)$
$(t_1, fs_1, pp_2, m_1, s_3, u_3, G_1)$
A*
Working_mode
$(t_1, fs_1, pp_2, m_1, s_1, u_3, G_0)$
$(t_1, fs_1, pp_2, m_1, s_2, u_3, G_0)$
$(t_1, fs_1, pp_2, m_1, s_3, u_3, G_0)$
Spindle : Multi_spindle_machine : Plant : Enterprise

Figure 2: The multi-dimensional data model for big data exploration.

comparison purposes, or across different tools, in order to detect anomalies.

Users/operators can move over hypercube nodes for data exploration. In this work, we also introduce exploration constrains to prevent meaningless comparisons.

Exploration constraints. There are some characteristics that should remain constant while performing any kind of comparison between measures. For instance, comparing measures collected during working mode G_0 with those collected during working mode G_1 makes no sense. The same considerations hold for the tool type, indeed comparing measures collected from a spindle while using tools of different types doesn't lead to any conclusion. These considerations lead to the introduction of the concept of *Exploration constraints*, i.e., a set of dimensions, that may be at different hierarchical levels, over which comparison between measures does not make sense. An example of exploration constraint is the dimension `Tool` at the hierarchical level `Tool_type` or the dimension `Working_mode`. In Figure 2 node "A" represents measures that should not be compared to measures represented by node "A*", because the two nodes are related to measures collected in a different working mode G_0/G_1.

2.3 Model definitions

Formal definitions of measures and exploration dimensions are given in the following.

Definition 1 (Feature) A feature represents a monitored variable that can be measured. A feature F_i is described as $\langle n_{F_i}, u_{F_i} \rangle$, where n_{F_i} is the feature name, u_{F_i} represents the unit of measure.

Let's denote with $F = \{F_1, F_2 \dots F_n\}$ the overall set of features.

Definition 2 (Measure) We define a measure for the feature F_i as a scalar value $X_i(t)$, expressed in terms of the unit of measure u_{F_i}, taken at the timestamp t.

Definition 3 (Feature space) We denote with $FS = \{FS_1, FS_2, \dots FS_m\}$ the set of feature spaces, where $FS_j \subseteq F$. Given a feature space $FS_j = \{F_1, \dots F_h\}$, we denote with the vector $\vec{X}_j(t)$ a record of measures $\langle X_1(t), \dots\ X_h(t)\rangle$ for the features in FS_j, synchronized with respect to the timestamp t. Feature spaces can be monitored independently each others.

Definition 4 (Domain-specific dimensions) We denote with $\mathcal{D}$ the subset of the multi-dimensional space created by p domain-specific dimensions $\mathcal{D}_1, \dots \mathcal{D}_p$, where $\mathcal{D} = \mathcal{D}_1 \times \dots \times \mathcal{D}_p$. Dimensions can be organized in hierarchies, at different levels. We denote with $\mathcal{D}_j^i$ the i-th level in the hierarchy of j-th dimension and with $d_i \in \mathcal{D}_i$ a single instance of the dimension $\mathcal{D}_i$.

Definition 5 (Exploration constraints) We define an exploration constraint ECX_i as a tuple $(\mathcal{D}_j, i)$, where i is the i-th level in the hierarchy over the j-th dimension $\mathcal{D}_j$. Comparison between different measures does not make sense over dimension $\mathcal{D}_j$ at the i-th hierarchical level. We denote with ECX the set of all possible exploration constraints $\{ECX_i\}$.

Definition 6 (Multi-dimensional model) We describe the multi-dimensional model as a set $\mathcal{V}$ of nodes and a set of exploration constraints ECX. Each node $v \in \mathcal{V}$ is described as

$$v = \langle \vec{X}_j(t), fs_j, d_1, d_2, \dots d_p \rangle \tag{1}$$

where $\vec{X}_j(t)$ represents a record of measures taken at time t, for the feature space fs_j and the values $d_1, d_2, \dots d_p$ of domain-specific dimensions $\mathcal{D}_1, \dots \mathcal{D}_p$, ECX represents the set of exploration constraints defined over the dimensions $\mathcal{D}_1, \dots \mathcal{D}_p$.

The conceptual model we defined to capture information collected from the Cyber-Physical System is shown in Figure 3 using ER notation. It reflects the typical structure of multi-dimensional models, where *facts* are represented by *measures*, as collected from the monitored physical system. Each feature presents physical limits (*bounds*), that should not be violated in order to avoid machine damages. We distinguish among warning and error bounds: (b) `warnings` identify anomalous conditions that may lead to breakdown or damage of machines; (c) `errors` identify unacceptable conditions in which a machine can not operate. Besides defining *features* bounds, we introduced the notion of *context* to specify contextual boundaries. A contextual bound represents the limit of a feature within a specific context where the feature is measured. The rationale is that, in a specific context (e.g., the working mode, the part program), when the Cyber-Physical System works normally, a feature should assume values within a specific range, that might be different from the overall physical limits for the same feature. These bounds are modelled through a relationship with attributes between the *feature* and the *context* entities in the model as shown in Figure 3.

Exploration constraints are not modelled in the conceptual model. They are defined as further meta-data that are used to guide/constrain the exploration.

3 Data summarisation and relevance evaluation

The characteristics of big data, namely, volume, velocity and variety, pose additional issues for data collection and organization. High volume calls for techniques and tools to provide a compact view over the large amount of collected data and to focus data exploration on relevant data only. Furthermore, when dealing with real time data, collected in Cyber-Physical Systems, data streams must be considered, where not all data are available since the beginning, but are collected in a fast and incremental way. To this aim, the conceptual

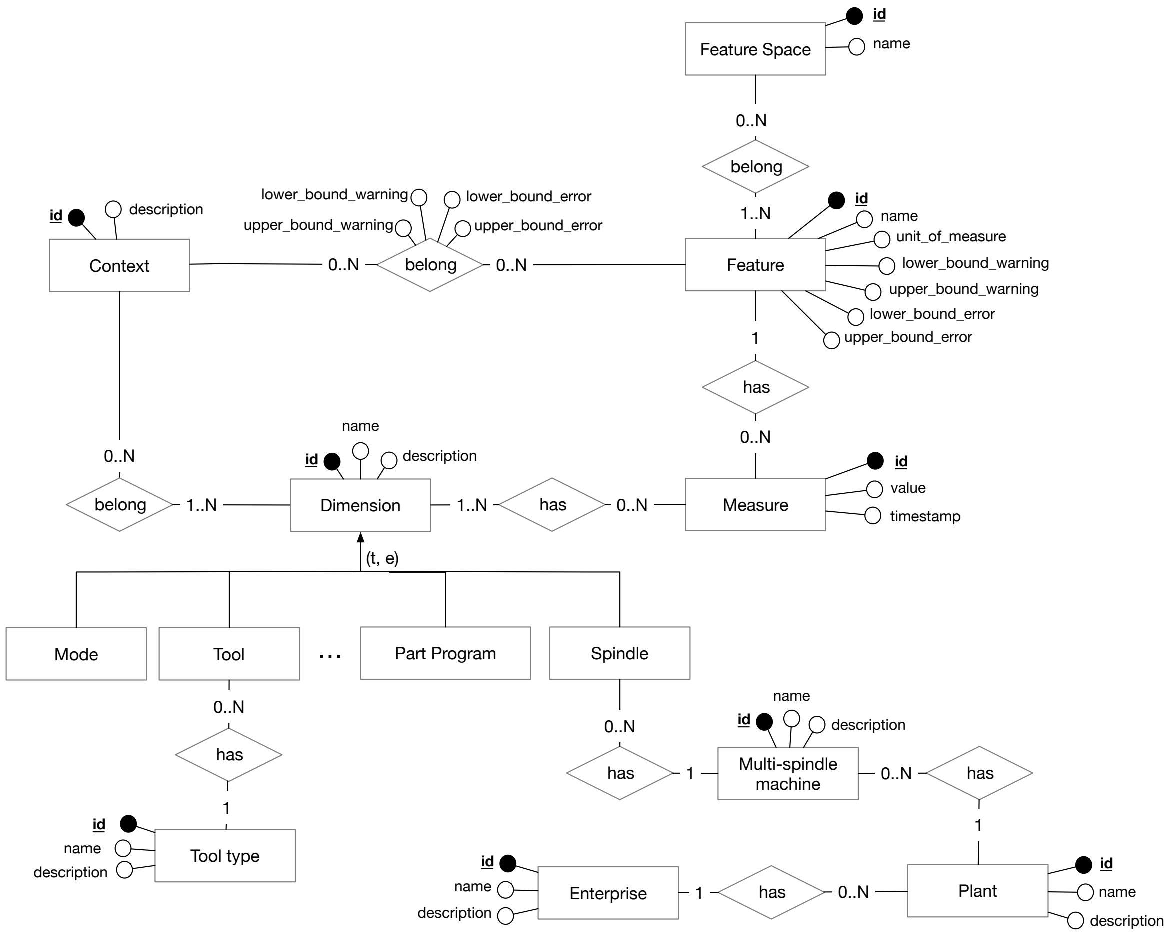


Figure 3: Conceptual model for exploration of data collected from Cyber-Physical Systems.

model proposed above is integrated with data summarisation and relevance evaluation techniques. These techniques have been detailed in (Bagozi et al. 2017c) and are shortly summarized in the following. The aim in this paper is to discuss the application of the conceptual model and relevance evaluation to different data exploration scenarios (see Section 4).

3.1 Clustering-based data summarisation

In our approach, data summarisation is based on clustering-based techniques. Clustering offers a two-fold advantage: (a) it gives an overall view over a set of measure records, using a reduced amount of information; (b) it allows to depict the behaviour of the system better than single records, that might be affected by noise and false outliers, in order to observe a given physical phenomenon.

The clustering algorithm is performed in two steps: (i) in the first one, a variant of Clustream algorithm (Aggarwal et al. 2003) is applied, that incrementally processes incoming data to obtain a *set of syntheses*; (ii) in the second step, X-means algorithm is applied (Pelleg and Moore 2000) in order to cluster syntheses obtained in the previous step. X-means algorithm does not require any a-priori knowledge on the number of output clusters. Syntheses are defined as follows.

Definition 7 (Synthesis) We define a synthesis S_j in a feature space fs_j as a tuple consisting of five elements, that is, $S_j = \langle N_j, \vec{LS}_j, SS_j, \vec{X}_j^0, R_j \rangle$, where: (i) N_j is the number of records included into the synthesis (from $\vec{X}_j(t_1)$ to $\vec{X}_j(t_N)$, where $t_N = t_1 + \Delta t$); (ii) $\vec{LS}_j$ is a vector representing the linear sum of measures in S_j; (iii) SS_j is a scalar representing the quadratic sum of points in S_j; (iv) $\vec{X}_j^0$ is a vector representing the centroid of the synthesis; (v) R_j is the radius of the synthesis. In particular:

$$\vec{LS}_j = \sum_{k=1}^{N_j} \vec{X}_j(t_k) \quad SS_j = \sum_{k=1}^{N_j} \vec{X}_j^2(t_k) \qquad (2)$$

$$\vec{X}_j^0 = \frac{\sum_{k=1}^{N_j} \vec{X}_j(t_k)}{N_j} \qquad (3)$$

$$R_j = \sqrt{\frac{\sum_{k=1}^{N_j} (\vec{X}_j(t_k) - \vec{X}_j^0)^2}{N_j}} \qquad (4)$$

The second step aims to cluster syntheses. Clustering is performed to minimize the distance between syntheses centroids within the same cluster and to maximize the distance between syntheses centroids across different clusters. Clusters give a balanced view of the observed physical phenomenon, grouping together syntheses corresponding to the same working status.

Definition 8 (Cluster) A cluster C is defined as follows: $C = \langle \vec{C}_0, \mathcal{S}_C \rangle$, where $\vec{C}_0$ is the cluster centroid, $\mathcal{S}_C$ is the set of syntheses belonging to the cluster. We denote with SC the set of identified clusters.

An incremental data-stream clustering algorithm has been developed, where the clustering algorithm is computed incrementally over time. The minimum granularity of the time dimension corresponds to the time interval over which clustering is performed. This means that, considering Δt as the time interval on which records of measures are grouped in syntheses, that in turn are clustered, every Δt seconds the clustering algorithm outputs a new cluster set SC built on top of the previous sets. Δt is chosen at configuration time such that 1/Δt is greater than the data acquisition frequency. Let's denote with $\Sigma(\vec{X}_j(\Delta t), fs_j, d_1, d_2, \dots d_p)$ the set of clusters obtained by applying clustering on measures collected during time interval Δt for dimensions $d_1 \in \mathcal{D}_1$, $d_2 \in \mathcal{D}_2 \dots$ $d_p \in \mathcal{D}_p$, for monitoring feature space fs_j; $\vec{X}_j(\Delta t)$ denotes the set of measures taken at a given time interval Δt, from $\vec{X}_j(t)$ to $\vec{X}_j(t + \Delta t)$. We include the output of summarisation procedure into the multi-dimensional model as follows, starting from Definition 6.

Definition 9 (Cluster-based multi-dim. model) We define the cluster-based multi-dimensional model as a set $\mathcal{V}'$ of nodes and a set of exploration constraints ECX. Each node $v' \in \mathcal{V}'$ is described as

$$v' = \langle \Sigma(\vec{X}_j(\Delta t), fs_j, d_1, d_2, \dots d_p) \rangle \qquad (5)$$

where $\Sigma(\cdot)$ represents the application of data summarisation procedure to $\vec{X}_j(\Delta t)$ for dimensions $d_1 \in \mathcal{D}_1$, $d_2 \in \mathcal{D}_2 \dots$ $d_p \in \mathcal{D}_p$, for monitoring feature space fs_j; ECX represents the set of exploration constraints defined over the dimensions $\mathcal{D}_1, \dots \mathcal{D}_p$.

3.2 Data relevance evaluation

Relevance-based techniques are used to detect components status over time. In literature, data relevance is defined as the distance from an expected status. The point is to define the expected status and how to compute such a distance. In (Bagozi et al. 2017c) we defined the expected status as the set of clusters computed during normal working conditions for the monitored system, denoted with $\hat{SC}$, and data relevance is based on the notion of *cluster distance* between the clusters set SC, that represents the current behaviour of the monitored system, and the set $\hat{SC}$. In the following, we will describe how the conceptual model enables data relevance evaluation in different considered scenarios.

4 Model-driven data exploration approach

Data exploration is performed on top of the multi-dimensional model in order to pursue different goals. In particular, in this section we discuss two possible exploration scenarios:

- *exploration for anomaly detection*, to promptly identify anomalies by monitoring and observing if collected data overtakes or gets closer to feature or contextual bounds, that represent physical limits of breakage of the monitored system; this kind of exploration may be implemented in a state detection service, and used by OEMs to prevent downtimes of monitored systems and by multi-spindle machine owner to plan supply chain activities;
- *exploration for performance comparison*, to compare performances of different monitored systems, while fixing the other analysis dimensions (e.g., using the same tools, performing the same manufacturing steps as codified within the part program); this kind of exploration can be used by OEMs to monitor a machine fleet over multiple clients and to offer remote configuration and optimization services.

Beyond these options, generic exploration of collected data is enabled by the multi-dimensional model. We assume that the user/operator formulates an explicit, albeit vague exploration request, by instantiating a subset of available dimensions (e.g., a specific spindle, tool or part program), and expects the system to suggest some promising data to explore. Data summarisation techniques contribute to reduce the complexity of the exploration by providing a compact view over underlying data. In the following, we will focus on the two exploration scenarios mentioned above, as generic exploration has been already described in (Bagozi et al. 2017c), where traversals inspired by operators in OLAP systems have been proposed for browsing data within the multi-dimensional space. We remark that the list of scenarios we will discuss here is not exhaustive. Our aim is to show how the conceptual model's features can be properly used to support model-driven data exploration mechanisms and can be configured for different scenarios.

4.1 Exploration for anomaly detection

The goal of a state detection service is to detect anomalies and send alerts concerning the system status. We consider three different values for the *status*: (a) `ok`, when the system works normally; (b) `warning`, when the system works in anomalous conditions that may lead to breakdown or damage; (c) `error`, when the system works in unacceptable conditions or does not operate. The warning status is used to perform an early detection of a potential deviation towards an error state. The migration of the system status from one value to the others raises an *alert* and occurs when one or more measures exceed a given bound, either a feature bound or a contextual bound, according to the conceptual model definitions given in Section 2. These bounds set the ranges for the three different values of the status: `ok`, `warning` and `error`. Feature bounds determine the *absolute status* of a feature, while contextual bounds determine the *contextual status* of a feature. The system status (either absolute or contextual) can be propagated to the whole feature space and along the hierarchy of monitored physical system, according to the following propagation rules.

a) *Propagation over the feature space.* Given a feature space $FS_j = \{F_1, F_2, \ldots F_h\}$, the value of the status associated to FS_j, given the status values for each feature $F_1, F_2, \ldots F_h$, is computed as follows:
   - `ok`, if the status of each feature $F_i (\forall i = 1 \ldots h)$ is `ok`;
   - `warning`, if the status of at least one feature $F_i (\forall i = 1 \ldots h)$ is `warning`;
   - `first level error`, if the status of at least one feature $F_i (\forall i = 1 \ldots h)$ is `error`;
   - `second level error`, if the status of each feature $F_i (\forall i = 1 \ldots h)$ is `error`.

b) *Propagation along the hierarchy of the monitored physical system.* The value of a feature status for a spindle is propagated to the highest level of the hierarchy (machine, plant, enterprise) as follows: the status of the machine is

- `ok`, if the status for all its spindles is `ok`;
- `warning`, if the status of at least one spindle is `warning`;
- `first level error`, if the status of at least one spindle is `error`;
- `second level error`, if the status of all its spindles is `error`.

The same applies for the status of the plant (resp., enterprise), computed starting from the status of its machines (resp., plants).

Relevance-based anomaly detection. For anomaly detection, the expected status is identified through the set $\hat{SC} = \{\hat{C}_1, \hat{C}_2, \ldots \hat{C}_n\}$ of clusters computed during normal working conditions. Relevant data are recognized when their clusters set differs from $\hat{SC}$. Let's denote with $SC = \{C_1, C_2, \ldots, C_m\}$ the current clusters set, where n and m do not necessarily coincide. We evaluate the distance between SC and $\hat{SC}$ by aggregating distances between each cluster belonging to SC and the closest cluster belonging to $\hat{SC}$ and vice-versa, for symmetry purposes. Formally, the distance is computed as:

$$\Delta(SC, \hat{SC}) = \frac{\sum_{i=1}^{m} d(C_i, \hat{SC}) + \sum_{j=1}^{n} d(SC, \hat{C}_j)}{m+n} \tag{6}$$

where $d(C_i, \hat{SC}) = min_{j=1,\ldots n} d_c(C_i, \hat{C}_j)$ and $d(SC, \hat{C}_j) = min_{i=1,\ldots m} d_c(C_i, \hat{C}_j)$ is the distance between clusters. To compute the distance between two clusters $d_c(C_i, \hat{C}_j)$, we combined different factors: (i) the distance between clusters centroids $d_{\vec{C}_0}(C_i, \hat{C}_j)$, to verify if C_i translates with respect to $\hat{C}_j$; (ii) the *intra-cluster distance* $d_c^{intra}(C_i, \hat{C}_j)$, to verify if there has been an expansion or a contraction of cluster C_i with respect to $\hat{C}_j$; (iii) the difference in number of syntheses contained in C_i and $\hat{C}_j$, denoted with $d_N(C_i, \hat{C}_j)$. The overall value of $d_c(C_i, \hat{C}_j)$ is given by:

$$\begin{aligned} d_c(C_i, \hat{C}_j) = \alpha \cdot d_{\vec{C}_0}(C_i, \hat{C}_j) \\ + \beta \cdot d_c^{intra}(C_i, \hat{C}_j) \\ + \gamma \cdot d_N(C_i, \hat{C}_j) \end{aligned} \tag{7}$$

where α, β and $\gamma \in [0, 1]$ are weights such that $\alpha + \beta + \gamma = 1$, used to balance the impact of terms in Equation (7). To set the optimal weights, a grid procedure can be performed over α and β (γ is set with $1 - \alpha - \beta$), with the value of each weight varying from 0 to 1. In our preliminary experiments, presented in (Bagozi et al. 2017c), we put $\alpha = \beta = \gamma = \frac{1}{3}$.

In particular, $d_{\vec{C}_0}(C_i, \hat{C}_j)$ is computed by applying the Euclidean distance ($D0$) between clusters centroids, according to the following formula:

$$D0 = \sqrt{(\vec{C}_0^i - \hat{\vec{C}}_0^j)^2} \tag{8}$$

where $\vec{C}_0^i$ and $\hat{\vec{C}}_0^j$ are centroids of C_i and $\hat{C}_j$, respectively. The computation of intra-cluster distance $d_c^{intra}(C_i, \hat{C}_j)$, performed on the sets of syntheses of C_i and $\hat{C}_j$, is similar to the computation of $\Delta(SC, \hat{SC})$ in Equation (6), that is:

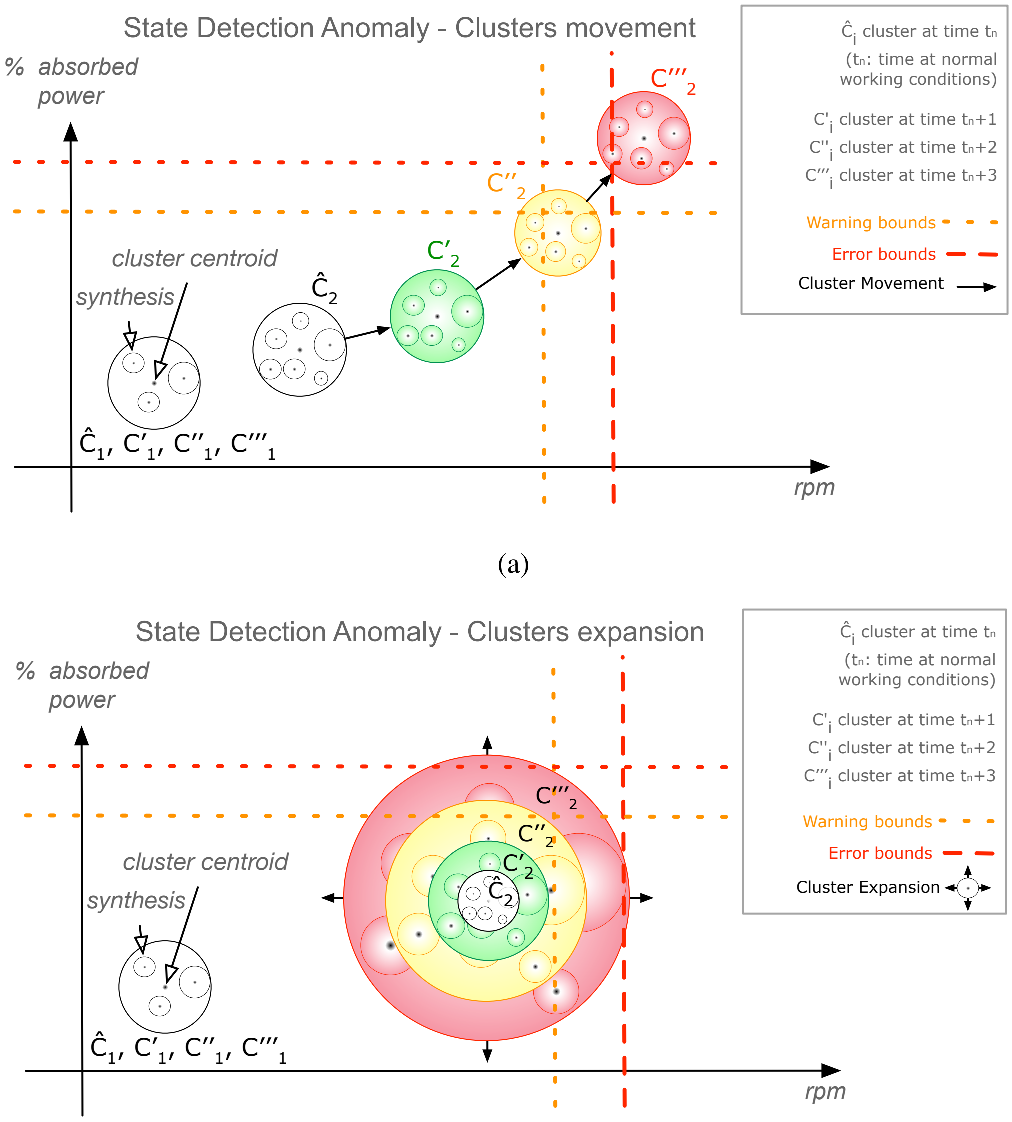
$$d_c^{intra}(C_i, \hat{C}_j) = \frac{\sum_{k=1}^{n_1} d_s(S_k, \hat{C}_j) + \sum_{h=1}^{n_2} d_s(C_i, S_h)}{n_1 + n_2} \tag{9}$$

where $S_k \in \mathcal{S}_{C_i}$, $S_h \in \mathcal{S}_{\hat{C}_j}$, $|\mathcal{S}_{C_i}| = n_1$, $|\mathcal{S}_{\hat{C}_j}| = n_2$, $d_s(S_k, \hat{C}_j) = min_{h=1,\ldots n_2} d_s(S_k, S_h)$ and $d_s(C_i, S_h) = min_{k=1,\ldots n_1} d_s(S_k, S_h)$. Term $d_s(S_k, S_h)$ represents the average inter-syntheses distance ($D1$):

$$D1 = \sqrt{\frac{\sum_{i=1}^{N1} \sum_{j=N1+1}^{N1+N2} (\vec{X}(t_i) - \vec{X}(t_j))^2}{N1N2}} \tag{10}$$

where $N1$ and $N2$ are the number of records in S_k and S_h, respectively.

The proposed relevance techniques enable to detect over time clusters movements, clusters contraction/expansion, changes in the number of clusters. Figures 4(a) and (b) show examples



(a)

(b)

Figure 4: Illustration of clusters sets changing over time for anomaly detection: (a) clusters movements; (b) clusters expansion. Clusters set $\hat{SC}$ is generated at time t_n at normal working conditions and consists of clusters $\hat{C}_1$ and $\hat{C}_2$.

of clusters changes for the anomaly detection purpose: changes in clusters set over time is detected due to spindle rolling friction torque increase, causing a decrease of rpm and an increase of the percentage of absorbed power. This exploration scenario is applied by fixing all the dimensions and observing evolution over time of measures within the feature space. In particular, the relevance techniques allow to identify what are the clusters that changed over time. Let's denote with $\{\overline{C_i}\}$ the set of such clusters. Data that is summarized in the clusters $\{\overline{C_i}\}$ is considered as relevant and, for each cluster in $\{\overline{C_i}\}$, the distance of cluster centroid from the warning and error bounds is computed. If this distance is equal or lower than the cluster radius, this means that a warning or error status has to be detected. Note that distance also helps to detect *potential* state changes. Consider for example Figure 4(a), that shows an example of clusters evolution over time for the smart factory case study. The figure shows how the cluster C_1 doesn't changed its position, as well as its size, from time t_n to t_{n+3}. On the other hand, cluster C_2 evolves from the wealth zone to the warning and error zones. At time t_{n+2} cluster C_2'' crosses the warning bound of rpm feature causing a warning alert, at time t_{n+3} cluster C_2''' moves into the error zone, crossing error bounds of both the considered features. At time t_{n+1} cluster C_2' still remains inside the wealth zone, however relevance techniques detected its change. Therefore cluster C_2' is recognised as relevant and monitored to detect warning or error state changes. This allows for better performance of the anomaly detection algorithm, that focuses only on potential state changes.

4.2 Exploration for performance comparison

The goal of this kind of exploration is to compare different machines in order to identify changes in working conditions. Therefore, this exploration scenario is applied by comparing clusters sets over the monitored system dimension and fixing all the other domain-specific dimensions. Let's consider the situation depicted in Figure 5. In the figure, two machines are compared considering how the clusters sets distance between two spindles evolves over time. At time t_1 the two spindles present a clusters sets distance equal to d_1. This distance is usually different from 0 since the two spindles work in different environments and the likelihood of having exactly the same measures for considered features over the compared physical systems is very low. We refer to distance d_1 as *baseline distance*, occurring when all the spindles are working in normal conditions. The baseline distance can be computed as follows:

$$\Delta_{baseline}(\mathcal{M}_1, \mathcal{M}_2) = \Delta(\hat{SC}_1, \hat{SC}_2) \qquad (11)$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the two considered spindles, $\hat{SC}_1$ (resp., $\hat{SC}_2$) is the clusters set obtained for spindle $\mathcal{M}_1$ (resp., $\mathcal{M}_2$) during normal working conditions.

At time t_{n+1} the two spindles present a distance $d_2 = d_1$. This means that the relative distance in terms of clusters sets between the two spindles is not changed. We remark here that, as shown in Figure 5, the condition $d_2 = d_1$ holds also if the two spindles changed their behaviours, and their respective clusters sets evolved accordingly. On the other hand, at time t_{n+2} the distance d_3 is changed compared to d_1 and d_2, meaning that the two spindles started behaving differently each other. The metric of relevance, in this case, aims at highlighting the difference between d_i and the baseline distance, denoted as $\Delta_{t_{n+2}}(\mathcal{M}_1, \mathcal{M}_2)$ and computed as follows:

$$\Delta(SC_1, SC_2) - \Delta_{baseline}(\mathcal{M}_1, \mathcal{M}_2) \qquad (12)$$

where SC_1 (resp., SC_2) is the clusters set obtained at time t_{n+2} for spindle $\mathcal{M}_1$ (resp., $\mathcal{M}_2$). These considerations raise some additional questions, namely: (a) what about if we consider more than two spindles? (b) what about if we are observing two spindles whose behaviour evolves accordingly (case d_2) towards anomalous conditions? (c) what are the conditions under which we can also identify what are the spindles whose performances decreased compared to the other one? For what concerns the latter question, consider the case d_3: what is the spindle with decreased performances among the two observed ones?

To solve the first question, we provide an extension of Equations (11) and (12): $\Delta_{baseline}(\mathcal{M}_i, \mathcal{M}_j)$ and $\Delta_{t_k}(\mathcal{M}_i, \mathcal{M}_j)$ are computed for each pair $\mathcal{M}_i$ and $\mathcal{M}_j$ of compared spindles and the average values are considered.

The second question can be answered by combining together different exploration scenarios: in case of distance $d_2 = d_1$, exploration for anomaly detection applied on one of the considered spindles might help to identify the case in which all spindles changed their behaviour accordingly. Finally, for what concerns the third question, the

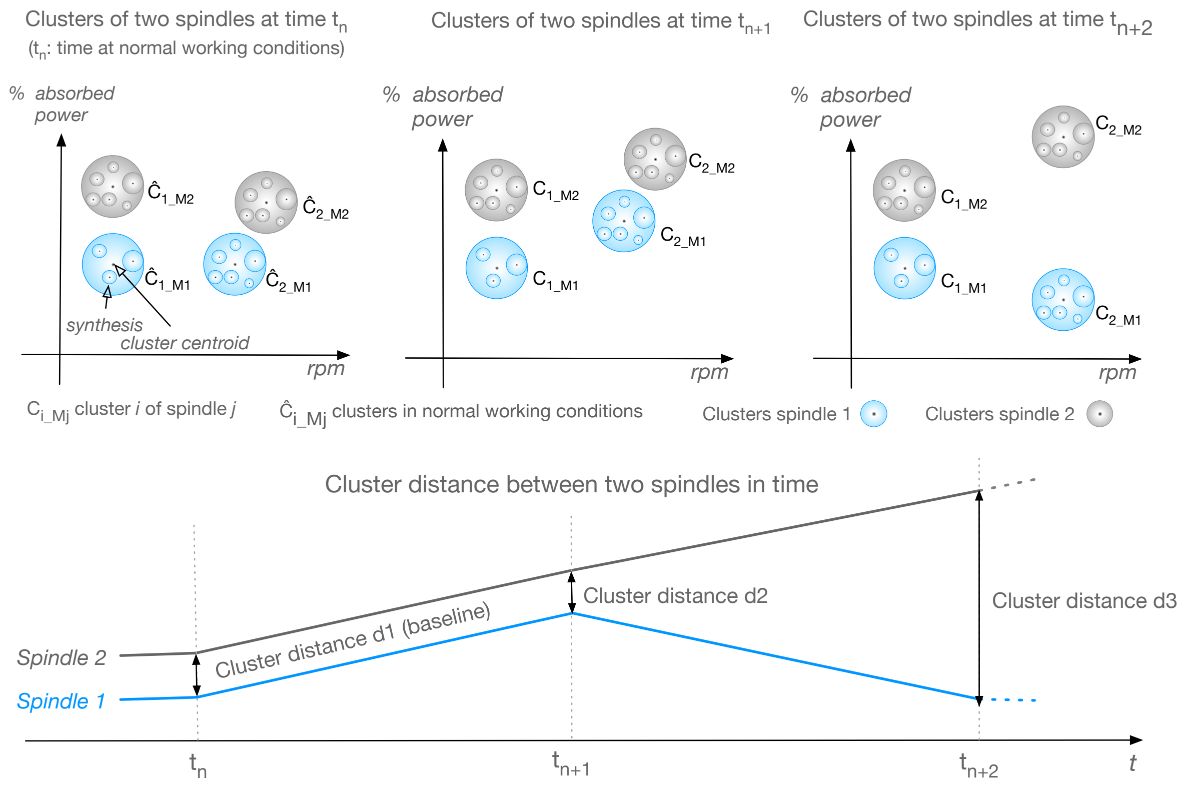

Clusters of two spindles at time t_n
(t_n: time at normal working conditions)
Clusters of two spindles at time t_n+1
Clusters of two spindles at time t_n+2
% absorbed power
rpm
Ĉ1_M2
Ĉ2_M2
Ĉ1_M1
Ĉ2_M1
C1_M2
C2_M2
C1_M1
C2_M1
synthesis
cluster centroid
C_i_Mj cluster i of spindle j
Ĉ_i_Mj clusters in normal working conditions
Clusters spindle 1
Clusters spindle 2
Cluster distance between two spindles in time
Spindle 2
Spindle 1
Cluster distance d1 (baseline)
Cluster distance d2
Cluster distance d3
t_n
t_n+1
t_n+2
t

Figure 5: Illustration of clusters sets changing over time for performance comparison. $\hat{C}_{i\_\mathcal{M}j}$ represents a cluster for spindle $\mathcal{M}_j$ while working in normal conditions ($j = 1, 2$).

identification of the spindle with decreased performances with respect to the other ones can be detected by applying anomaly detection on each spindle or, if we are considering more than two spindles, according to the "two out of the three" logics: the target spindle will present a distance from the other ones that is greater than the same computation made for all the spindles. In case of two spindles, the third question can be answered only by applying exploration for anomaly detection for all the considered spindles.

4.3 Discussion

Our work proposes a conceptual modelling approach to allow flexible big data exploration in Cyber-Physical Systems, specifically to enable detection of unexpected situations. We investigated the potential benefits of conceptual modelling to foster exploration scenarios that pursue different goals. In particular, we considered two scenarios in the Industry 4.0 context: exploration for anomaly detection, to promptly identify anomalies on the monitored Cyber-Physical System, and exploration for performance comparison across different Cyber-Physical Systems. Our examples show how conceptual modelling reveals to be useful in abstracting from specific characteristics of each scenario:

- it enables to specify the desired exploration dimensions (e.g., features/feature space, or domain-specific dimensions as spindles, tools, etc.) thanks to the multi-dimensional paradigm; for example, exploration for anomaly detection is applied by observing over time the evolution of measures collected on monitored features and fixing all the other dimensions; while exploration for performance comparison is applied by comparing clusters sets across different monitored systems and fixing all the other domain-specific dimensions;

- it is used to adapt the relevance techniques to the different cases, starting from an expected status and using the clusters sets distance analysis to identify changes: for example, in the case of exploration for anomaly detection, clusters sets distance analysis is based on the distance with respect to the clusters set $\hat{SC}$, generated at normal working conditions of the monitored system; in the exploration for performance comparison, clusters sets distance analysis is based on a baseline distance, obtained by considering normal working conditions of all the compared physical systems.

The abstraction enabled by conceptual modelling suggests the feasibility to provide a model-driven framework, where models and techniques can be adapted to different domains, beyond the Industry 4.0 one considered here.

5 Related Work

In this section we analyse some approaches that have been focused on data exploration and exploratory computing research fields, investigating the potential use of conceptual modelling to meet their goals. This comparison is summarized in Table 1.

The approach presented in (Kamat et al. 2014) deals with structured, multi-dimensional OLAP data, incrementally collected and organized in a cube structure, where axes correspond to facets to guide the exploration. The faceted cube exploration model is used to bound the space of possible queries. Data sampling techniques are applied to guess the next choices the user will likely to perform, thus reducing response times.

The approach presented in (Kalinin et al. 2014) treats multi-dimensional data, but it does not provide a conceptual modelling approach for exploration. The approach enables range queries (explicitly formulated by the user) on features, that must have a sortable numeric data type. Queries identify partially overlapping windows, shown to the user according to a cost-benefit criterion, that depends on the efforts required to collect data shown in the windows.

Also the approach in (Dimitriadou et al. 2016) handles multi-dimensional data already available for data exploration, but it does not provide a reference conceptual framework apt to be applied to different scenarios. The data shown to the user is fetched from the DBMS using sampling techniques. This approach builds clusters of objects using data mining techniques (k-means) and applies a classifier to infer data relevance.

The approach presented in (Costa et al. 2017) relies on multi-dimensional data incrementally collected. Moreover, a loss-less compression algorithm is used in order to minimise space consumption. In order to retrieve interesting events, authors exploit an occurrence frequency threshold: values whose frequency is below such threshold are properly highlighted.

Other approaches use multi-dimensional model to organize data, thus demonstrating the effectiveness of this kind of model to enable data exploration. Some of them also use summarisation/approximation techniques, to provide compact views over data and relevance evaluation techniques in order to guide exploration. The main contribution of our work compared to them mainly resides on the abstraction we performed in order to adapt the model and techniques to different scenarios. Existing approaches either do not mention any specific application scenario (making the proposal difficult to apply in a specific context such as the one of Cyber-Physical Systems) or are focused on very specific problems such as anomaly detection (Huber et al. 2016; Moghaddass and Zuo 2014; Stojanovic et al. 2016; Wang and Agrawal 2011). For what concerns anomaly detection approaches, the investigation of multi-dimensional modelling with summarisation and relevance evaluation techniques is limited. We refer to (Bagozi et al. 2017a) for a survey on this kind of approaches.

6 Concluding remarks

In this paper we discussed the application of conceptual modelling to provide a high level big data

	IDEAaS	[Kamat et al. 2014]	[Kalinin et al. 2014]	[Dimitriadou et al. 2016]
Multi-dimensional model	✓	✓	✓	✓
Multi-dimensional model construction	✓	✓		
Summarisation techniques	✓			✓
Approximation techniques		✓	✓	✓
Relevance techniques	✓			✓

	IDEAaS	[Costa et al. 2017]	[Stojanovic et al. 2016]	[Wang and Agrawal 2011]
Multi-dimensional model	✓	✓		✓
Multi-dimensional model construction	✓	(✓)		
Summarisation techniques	✓	✓	✓	
Approximation techniques				✓
Relevance techniques	✓	✓		

	IDEAaS	[Huber et al. 2016]	[Moghaddass and Zuo 2014]	
Multi-dimensional model	✓		✓	
Multi-dimensional model construction	✓			
Summarisation techniques	✓	✓		
Approximation techniques				
Relevance techniques	✓			

Table 1: Overview of approaches on (big) data exploration.

representation and enabling model-driven data exploration. In particular, we exploited the ability of conceptual modelling to abstract data representation from implementation details and to focus on data semantics, by considering multiple exploration scenarios for monitoring Cyber-Physical Systems. We proposed a data representation model structured over a set of dimensions, hierarchically modelled. The resulting multi-dimensional model has been further enriched with data summarisation techniques, to provide a compact view over large amount of data and therefore managing data complexity in terms of volume and acquisition speed (velocity). Given the conceptual model, we defined a model-driven data exploration approach, that relies on data relevance techniques, aimed to focus the attention of the operators on relevant data only and to guide multiple exploration strategies according to different objectives. We also introduced exploration constraints to prevent useless comparison between collected data.

Further research will be performed in order to expand the set of analysis dimensions, in order to properly correlate data collected from CPS with high level aspects, concerning product and process quality, energy consumption and manufacturing sustainability. The research can fruitfully exploit the flexibility of the proposed conceptual model. Additional scenarios will be also considered, and a model-driven tool to support configuration and set up of new scenarios will be investigated. Finally, future work will also be focused on analysing data visualization techniques, already addressed in approaches like (Kruiger et al. 2017) and (Saket et al. 2017), and developing a proper GUI specifically meant for big data exploration.

References

Aggarwal C., Han J., Wang J., Yu P. (2003) A framework for clustering evolving data streams. In: Proc. of 29th International Conference on Very Large Data Bases (VLDB), pp. 81–92

Bagozi A., Bianchini D., De Antonellis V., Marini A., Ragazzi D. (2017a) Big Data Summarisation and Relevance Evaluation for Anomaly Detection in Cyber Physical Systems. In: On the Move to Meaningful Internet Systems. OTM 2017 Conferences, pp. 429–447

Bagozi A., Bianchini D., De Antonellis V., Marini A., Ragazzi D. (2017b) Interactive Data Exploration as a Service for the Smart Factory. In: 2017 IEEE International Conference on Web Services (ICWS), pp. 293–300

Bagozi A., Bianchini D., De Antonellis V., Marini A., Ragazzi D. (2017c) Summarisation and Relevance Evaluation Techniques for Big Data Exploration: the Smart Factory case study. In: Proc. of 29th Int. Conference on Advanced Information Systems Engineering (CAISE2017), pp. 264–279

Cabot J., Gómez C., Pastor O., Sancho M., Teniente E. (2017) Conceptual Modeling Perspectives. Springer

Chen P. (1976) The entity-relationship model - toward a unified view of data. In: ACM Transactions on Database Systems 1(1), pp. 9–36

Costa C., Chatzimilioudis G., Zeinalipour-Yazti D., Mokbel M. F. (2017) Efficient Exploration of Telco Big Data with Compression and Decaying. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 1332–1343

Dimitriadou K., Papaemmanouil O., Diao Y. (2016) AIDE: An Active Learning-Based Approach for Interactive Data Exploration. In: IEEE Transactions on Knowledge and Data Engineering 28(11), pp. 2842–2856

Embley D., Liddle S. (2013) Big Data - Conceptual Modeling to the Rescue. In: Proc. of International Conference on Conceptual Modeling (ER), pp. 1–8

Fliedl G., Kop C., Mayr H. (2005) From textual scenarios to a conceptual schema. In: Data & Knowledge Engineering 55, pp. 20–37

Golfarelli M., Rizzi S. (2009) Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill

Han J., Kamber M. (2006) Data Mining: Concepts and Techniques Edition 2. (ed.). Morgan Kaufmann Publisher

Hou Z., Wang Z. (2013) From model-based control to data-driven control: Survey, classification and perspective. In: Information Science 235, pp. 3–25

Huber M., Voigt M., Ngonga Ngomo A.-C. (2016) Big Data Architecture for the Semantic Analysis of Complex Events in Manufacturing. In: INFORMATIK 2016 - the 46th Conference of the Gesellschaft für Informatik e.V. (GI), 2nd International Workshop on Big Data, Smart Data and Semantic Technologies (BDSDST), pp. 352–360

Kalinin A., Cetintemel U., Zdonik S. (2014) Interactive data exploration using semantic windows. In: Proc. of the ACM SIGMOD International Conference on Management of Data, pp. 505–516

Kamat N., Jayachandran P., Tunga K., Nandi A. (2014) Distributed and Interactive Cube Exploration. In: Proc. of 30th International Conference on Data Engineering (ICDE), pp. 472–483

Karagiannis D., Mayr H., Mylopoulos J. (2016) Domain-Specific Conceptual Modeling - Concepts, Methods and Tools. Springer International Publishing, pp. 1–594

Kruiger J. F., Hassoumi A., Schulz H.-J., Telea A., Hurter C. (2017) Multidimensional Data Exploration by Explicitly Controlled Animation. In: Informatics 4(3), pp. 1–21

Lee E. A., Seshia S. A. (2017) Introduction to Embedded Systems, A Cyber-Physical Systems Approach, Second Edition. MIT Press, ISBN 978-0-262-53381-2

Lee J., Ardakani H., Yang S., Bagheri B. (2015a) Industrial big data analytics and cyber-physical systems for future maintenance and service innovation. In: Proc. of Conference on Intelligent Computation in Manufacturing Engineering (CIRP) Vol. 38, pp. 3–7

Lee J., Bagheri B., Kao H. (2015b) A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. In: Manufacturing Letters 3, pp. 18–23

Moghaddass R., Zuo M. J. (2014) An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process. In: Reliability Engineering & System Safety 124(Supplement C), pp. 92–104

Monostori L. (2014) Cyber-physical production systems: Roots, expectations and R&D challenges. In: Proc. of the 47th CIRP Conference on Manufacturing Systems, pp. 9–13

Olivé A. (2007) Conceptual Modeling of Information Systems. Springer-Verlag Berlin Heidelberg

Pelleg D., Moore A. (2000) X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Proc. of 17th International Conference on Machine Learning (ICML), pp. 727–734

Saket B., Kim H., Brown E. T., Endert A. (2017) Visualization by Demonstration: An Interaction Paradigm for Visual Data Exploration. In: IEEE Transactions on Visualization and Computer Graphics 23(1), pp. 331–340

Stojanovic L., Dinic M., Stojanovic N., Stojadinovic A. (2016) Big-data-driven anomaly detection in industry (4.0): An approach and a case study. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1647–1652

Tukey J. (1977) Exploratory data analysis. Addison-Wesley Publishing Company Reading, pp. 1–688

Tunkelang D. (2009) Faceted Search (Synthesis Lectures on Information Concepts, Retrieval and Services). Morgan and Claypool Publishers, pp. 1–94

Wang F., Agrawal G. (2011) Effective and Efficient Sampling Methods for Deep Web Aggregation Queries. In: Proc. of Conference on Extending Database Technology (EDBT), pp. 425–436