PACMEL

Process-aware Analytics Support based on Conceptual Models for Event Logs (**PACMEL**)

- **CHIST-ERA 2017 BDSI NCN (Unisono)**
- **Project Leader:** prof. Grzegorz J. Nalepa
- **Partners:** prof. Diego Calvanese (Free University of Bozen-Bolzano), prof. David Camacho (Universidad Politécnica de Madrid)
- **Start time:** 01.04.2019
- **End time:** 31.03.2022
- **Duration:** ~~24~~ 36 months
- **www:** http://pacmel.geist.re
- **chistera www:** external webpage

Motivation

Nowadays great attention is paid to the Industry 4.0 concept, whose central idea is the exploitation of large amounts of data generated by different kinds of sensors, to enact highly automatized, robust processes and to develop high quality process monitoring systems that support intelligent semi-autonomous decision making. At the same time, big data analytics as core competency and a process-oriented management approach are very often indicated as pillars of any modern company. Towards this, the main objective of PACMEL was to develop a process-aware analytics framework for analyzing data from sensors and devices, enabling its use for process modeling and analysis, with the aim of improving the business processes according to the BPM cycle. The framework can be applied to the data system of smart factories to support the business process management activities in the scope of process modeling and analysis. On the one hand, it will allow the creation of conceptual models of particular industrial processes being executed in the factory, taking into account the various abstraction levels of the collected data. This will be achieved by combining knowledge extraction techniques with semantic technologies such as ontology-based data access and integration. On the other hand, it will support model mapping methods and visualization techniques that allow one to relate the interpreted sensor data to the process models for process analysis. We will use a real dataset, related to a very complex and specific process, from an industrial domain (mining). The complexity of the considered process is a consequence both of its intrinsic characteristics, and of the conditions under which it is realized. Working with this challenging example will bring valuable insights and results that can be applied across various industrial domains including aeronautics and manufacturing.

Project structure

Consortium

PACMEL was planned as a 2 year project, extended to 3 years. The project consortium included 3 partners:

1. AGH, Poland (AGH University of Science and Technology), PI: Grzegorz J. Nalepa,
2. UNIBZ, Italy (Free University of Bozen-Bolzano), PI: Diego Calvanese,
3. UPM, Spain (Universidad Politecnica de Madrid), PI: David Camacho.

Work packages

The workplan included 6 work packages, further partitioned into tasks, briefly described below.

WP1: Identification of requirements for smart factories, 9 m. (M1-9), Led by: AGH The first objective of this WP was the analysis of case studies with respect to Industry 4.0/Smart Factory to identify the industrial requirements that would guide the development of a general process-aware analytics framework. To this end, an exploratory analysis of the related industrial datasets was conducted. Finally, an analysis of the needs of various domains to create opportunities for further applications of our framework was planned.

WP2: Knowledge extraction and data mining, 12 months (M4-15), Led by: UPM The main objectives of this WP included: the study and analysis of dimension reduction of the considered data sets and, the selection of the appropriate granulation and abstraction level for the data analysis with respect to sensor readings. The data analysis and filtering was provided to solve existing data quality problems, such as missing values and data redundancy, and to identify sources of unique states of factory machines/devices with respect to sensor readings. Finally, an analysis and selection of appropriate machine learning algorithms, and data mining techniques, for data clustering and classification were performed to allow for the extraction and discovery of new patterns.

WP3: Ontology driven interpretation, 14 months (M5-18), Led by: UNIBZ The objective of this WP was to provide a conceptual and technological framework for the interpretation of the data produced as a result of WP2, in terms of the semantically meaningful elements that constitute the knowledge about the domain of interest. Such knowledge, suitably encoded in an ontology, would provide the basis for addressing the number of fundamental problems. A work on extension of the ontology-based data access (OBDA) paradigm (with its techniques for efficient automatic query transformation and evaluation) to deal with the type of data at hand was planned.

WP4: Process-aware analytics framework, 15 months (M9-M23), Led by: UPM The main objectives included: (1) the provisioning of model mapping methods and visualization techniques that allow one to relate the interpreted sensor data (provided by methods in WP3) to the process models; (2) implementation of proof-of-concept software tools; and (3) the creation of feasibility studies for possible applications of the developed framework in mining as well as in other domains.

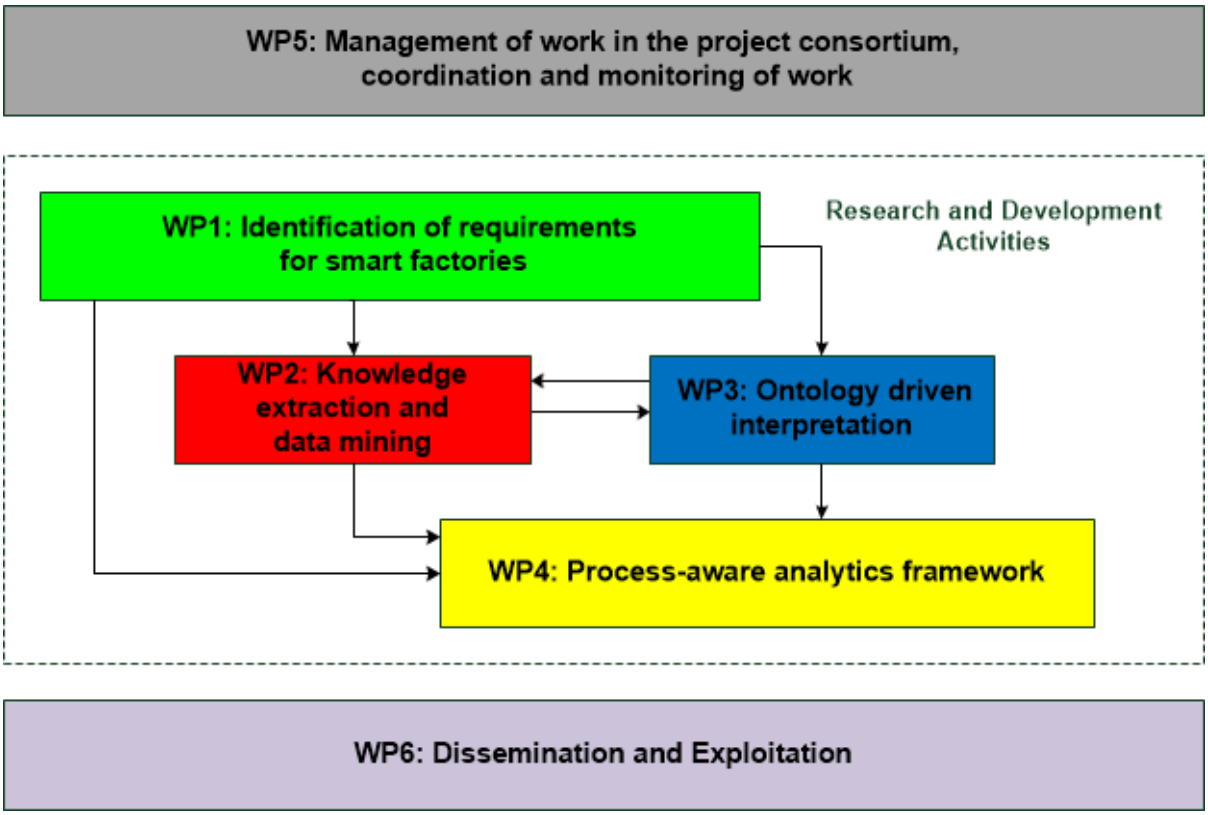
WP5: Management, 24 months (M1-24), Led by: AGH The activities included the organization and coordination of telecons and project meetings. A set up and maintenance of an on-line work repository, based on a GitLab installation used by AGH was planned. We will guarantee the participation in the yearly CHIST-ERA meetings, along with monitoring of the project progress and reporting. Finally, risk monitoring, management and mitigation techniques would be implemented.

WP6: Dissemination and exploitation, 24 months (M1-24), Led by: UNIBZ/AGH The overall goal of the systematic efforts in this WP was to raise awareness and to foster participation in the related scientific communities and among industry stakeholders as well as to disseminate knowledge to research teams beyond the project consortium. Activities in this WP would include: the project

website setup and maintenance; the scientific publications based on collaborative research work; the participation in scientific conferences, and events, network and exchange with contiguous projects and initiatives; the organization of industrial workshops and, the publication of tools and methods developed in the project as well as of a subset of the studied dataset to allow reproducibility of project results.

Dependencies

The dependencies between the Work Packages in the project are presented below.



Project team

AGH, Poland

- Principal Investigator: Grzegorz J. Nalepa
- Main investigator: Szymon Bobek
- Co-investigators: Edyta Brzychczy, Marcin Szpyrka
- Supportinng investigators: Aneta Napieraj
- PhD Students: Maciej Szelążek, Agnieszka Trzcinkowska, Paweł Jemioło, Michał Kuk, Maciej Mozolewski

UPM, Spain

- Principal Investigator: David Camacho
- Main investigator: Victor Rodriguez-Fernandez

- Co-investigators: Antonio Gonzalez-Pardo
- Students: David Montalvo

UNIBZ, Italy

- Principal Investigator: Diego Calvanese
- Main investigator: Marco Montali
- Co-investigators: Roberto Confalonieri

Main results

The project targeted four general main outcomes:

1. Improved process mining and knowledge extraction techniques;
2. The integration of various heterogeneous data sources by employing ontologies to support semantically enriched search of information about industrial processes extending knowledge modelling opportunities;
3. Process design support by recommendation methods of process modelling notation according to industrial needs;
4. Process monitoring support using time-series aware process mining methodologies that will use directly the raw timed data from the industrial domain, and visualization tools for the purpose of process-aware analytics expanding process exploratory analysis.

Implementation of new advanced scientific methodologies

Thanks to the cooperation with FAMUR and PGG (Polska Grupa Górnicza – Polish Mining Group), AGH developed a complete description of the mining use case, including the specific requirements. We built a repository of sensor data from 5 distinct longwalls from several polish mines owned by PGG and using equipment from FAMUR. Moreover, we performed exploratory data analysis of the sensor data, which allowed us for dimensionality reduction. Finally, we built a hierarchical formal model of the longwall shearer, as published in [3].

In WP2 UPM and AGH designed a new approach to conformance checking by adapting its basic elements to the paradigm of time-based data and time-aware processes. Instead of event logs, time series logs are defined and used. In the same way, the use of WF-Nets to represent the process model has been replaced by Workflow Net with time series (TSWF-net). This changes the perspective of the conformance checking methods completely, and thus, a new method is proposed based on these new ideas. The achievements made here may open up a new way for researchers to investigate how to adapt other families of process mining techniques, such as discovery and enhancement, to the use of time series data [1].

In WP4 UPM developed new model mapping methods based on raw data instead of ontologies (due to the lack of progress in WP3). In this way, deep learning architectures were used to map the raw data to a compressed representation which can be easily visualized. A software tool, called DeepVATS was created to explore and visualize cyclic patterns and outliers from raw multivariate time series data, displaying them in a 2D space which is suitable for a domain expert. This tool was released as an open-source platform in 2022Q2. As part of WP4 the DeepVATS tool was evaluated using different use cases, such as the longwall shearer operation in an underground coal mine. The patterns found in this

tool, which are completely data-driven and unsupervised, will were compared with expert knowledge and with the rules found by process mining techniques, as in [1,3].

In 2020 work of AGH included the implementation of the formal model for conformance checking of a longwall shearer process. This work was described in the paper in the Energies journal. The approach uses place-transition Petri nets with inhibitor arcs for modelling purposes. We used event log files collected from the mining use case. One of the main advantages of the approach is the possibility for both offline and online analysis of the log data. In the paper we presented a detailed description of the longwall process, an original formal model we developed, and the implementation in the TINA software (http://projects.laas.fr/tina). The tool can be used for model development, interactive model simulation and formal model analysis, see [3] for results of practical experiments.

In [13] we performed detailed survey on different explanation mechanism and knowledge integration approaches with data mining pipeline, and application to real-life industrial use-cases. This survey along with first attempts to combine domain knowledge with visual analytics and data mining models resulted in works published in [12], where we present the Immersive Parallel Coordinates Plots (IPCP) system for Virtual Reality (VR). The system provides data science analytics built around a well-known method for visualization of multidimensional datasets in VR. The data science analytics enhancements consist of importance analysis and a number of clustering algorithms to automate the work previously done by the experts manually.

The results and experience from these works were foundation for implementation of PACMEL framework and tools that form it: KnAC [17], ClAMP [16] and DeepVATS [21].

Development of innovative software

The source code of the implementation of the methodology presented in [1] is publicly available in Github (https://github.com/vrodriguezf/tscc). A second relevant contribution from UPM related to the development of innovative software has been the design and development of a software framework for high-dimensionality time series analysis later developed into DeepVATS [21]. In 2019 UniBZ carried out some foundational work that is relevant for PACMEL. Specifically, it developed techniques for verification of data-aware processes using an approach based on Satisfiability Modulo Theories (SMT). The considered setting is rich, and in the specification of the behaviour of the modelled systems data is not abstracted away. In general, this leads to an infinite-state system, so that specific techniques are necessary to avoid undecidability of verification. The developed techniques show decidability of verification of safety problems (reachability) and rely on the MCMT model checker for array-based systems for an actual implementation. The work is reported in an article [2].

KnAC [17] is a tool for expert knowledge extension with a usage of automatic clustering algorithms. It allows to refine expert-based labelling with splits and merges recommendations of expert labelling augmented with explanations. The explanations were formulated as rules and therefore can be easily interpreted incorporated with expert knowledge. It was implemented as an open-source software available at https://github.com/sbobek/knac, along with publicly available benchmark dataset concerning text clusterization. The additional evaluation and feasibility study on data from PACMEL project (from coal mine) was also performed and will be made available upon request due to confidentiality commitments to the company that delivered the data. The tool is written in Python programming language with compliance to the scikit-learn convention, which allows for seamless integration of the tool within ML/DM pipelines.

ClAMP toolkit [16] is a mechanism that is complement to KnAC and in some cases can be used within

it. It aims at helping experts in cluster analysis with human-readable rule-based explanations. The developed state-of-the-art explanation mechanism is based on cluster prototypes represented by multidimensional bounding boxes. This allows to represent an arbitrary shaped clusters and combine strengths of local explanations with the generality of the global ones. The main goal of our work was to provide a methods for cluster analysis, that will be agnostic with respect to the clustering method and classification algorithm and will provide explanations in a form of executable and human-readable form.

DeepVATS [21] is an open-source tool that supports domain experts in the analysis and understanding of time series, especially when these are of long duration, due to they entail a high information overload. It works by presenting the domain expert with a plot containing the projection of the latent space of a Masked Time Series Autoencoder trained to reconstruct partial views of the input dataset. The intersection of goals of DeepVATS and KnAC was a motivation to integrate both tools into one framework. This module is responsible for implementing the interaction layer between human users and the remaining system modules. The DeepVATS provides tools for time-series analytics, dimensionality reduction, clustering and visualization, while KnAC is responsible for helping in analysis of obtained results. This allows not only to discover new patterns and visualize them, but also understand the difference between discovered clusters with rule-based explanations delivered by KnAC. The above mentioned integration of tools allowed us to provide the final “PACMEL framework” toolset as described in the [22] paper. The code of the main tools was made public, as mentioned below, with some representative data samples as part of the benchmark.

New ideas, new knowledge, new interpretative models of complex phenomena

Working towards provisioning to domain experts we developed a concept of Explanation-driven model stacking, presented in [6]. Practical usage of explainability methods in AI (XAI) is limited nowadays in most of the cases to the feature engineering phase of the data mining process. In our work we argue that explainability as a property of a system should be used along with other quality metrics such as accuracy, precision, recall in order to deliver better AI models. We developed a method that allows for weighted ML model stacking [7]. The code is made public on GitHub (https://github.com/sbobek/inxai). In [18] we proposed an approach for evaluation of selected imperative and declarative models to decide which mode is more appropriate from a practical point of view for process monitoring. We formulated quantitative and qualitative criteria for models comparison. We then performed an analysis using the PACMEL mining case study. We used sensor data, and our approach consists of the following stages: (i) event log creation, (ii) process modelling, and (iii) process mining. We used the selected models in conformance checking tasks, with the use of a real event log. Evaluation of the created models indicated that in the case of the longwall mining the declarative model is more appropriate for practical use.

Tools

As described above three main tools developed during the project are available publically

- KnAC: https://github.com/sbobek/knac
- ClAMP: https://github.com/sbobek/clamp
- DeepVATS: https://github.com/vrodriguezf/deepvats

Public benchmark

Representative prototypes of computational procedures with data samples where prepared, tested, stored and finally published in the project GitLab repository:
https://gitlab.geist.re/pml/x_benchmark-with-selected-datasets

Papers

- [1] V. Rodriguez-Fernandez, A. Trzcionkowska, A. Gonzalez-Pardo, E. Brzychczy, G. J. Nalepa, and D. Camacho. Conformance Checking for Time Series-aware Processes. IEEE Transactions on Industrial Informatics. 17(2): 871-881 (2021)
- [2] D. Calvanese, S. Ghilardi, A. Gianola, M. Montali, and A. Rivkin. SMT-based Verification of Data-Aware Processes: A Model-Theoretic Approach. Mathematical Structures in Computer Science. 2020. 30(3): 271-313 (2020)
- [3] M. Szpyrka, E. Brzychczy, A. Napieraj, J. Korski, G. J. Nalepa, Conformance Checking of a Longwall Shearer Operation Based on Low-Level Events, Energies 2020, 13, 6630.
- [4] M. Szelążek, S. Bobek, A. Gonzalez-Pardo, G. J. Nalepa, “Towards the Modeling of the Hot Rolling Industrial Process. Preliminary Results”. In: International Conference on Intelligent Data Engineering and Automated Learning (pp. 385-396). Springer, 2020.
- [5] S. Bobek, A. Trzcinkowska, E. Brzychczy, G. J. Nalepa: Cluster Discovery from Sensor Data Incorporating Expert Knowledge, Workshop of Knowledge Representation & Representation Learning KR4L, ECAI 2020 in Santiago de Compostela, June 2020 https://smartdataanalytics.github.io/KR4L
- [6] S. Bobek and G. J. Nalepa. Augmenting automatic clustering with expert knowledge and explanations. In Computational Science – ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part IV, page 631–638, Cham, 2021. Springer International Publishing
- [7] S. Bobek, M. Mozolewski, and G. J. Nalepa. Explanation-driven model stacking. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, editors, Computational Science – ICCS 2021, pages 361–371, Cham, 2021. Springer International Publishing.
- [8] F. Piccialli, F. Giampaolo, E. Prezioso, D. Camacho, and G. Acampora, “Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion,” Information Fusion, vol. 74, pp. 1–16, Oct. 2021
- [9] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, and D. Camacho, “Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis,” Future Generation Computer Systems, vol. 122, pp. 220–233, Sep. 2021.
- [10] J. Huertas-Tato, A. Martín, J. Fierrez, and D. Camacho, “Fusing CNNs and statistical indicators to improve image classification,” Information Fusion, vol. 79, pp. 174–187, Mar. 2022.
- [11] A. I. Torre-Bastida, J. Díaz-de-Arcaya, E. Osaba, K. Muhammad, D. Camacho, and J. Del Ser, “Bio-inspired computation for big data fusion, storage, processing, learning and visualization: state of the art and future directions,” Neural Comput & Applic, Aug. 2021.
- [12] S. Bobek, S. K. Tadeja, Struski, P. Stachura, T. Kipouros, J. Tabor, G. J. Nalepa, and P. O. Kristensson. “Virtual reality-based parallel coordinates plots enhanced with explainable AI and data-science analytics for decision-making processes.” Applied Sciences, 12(1), 2022
- [13] G. J. Nalepa, S. Bobek, K. Kutt, and M. Atzmueller. “Semantic data mining in ubiquitous sensing: A survey.” Sensors, 21(13), 2021.
- [14] M. Kuk, S. Bobek and G. J. Nalepa, “Explainable clustering with multidimensional bounding

boxes," 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, pp. 1-10.
- [15] S. Bobek, M. Kuk, J. Brzegowski, E. Brzychczy, and G. J. Nalepa. "KnAC: an approach for enhancing cluster analysis with background knowledge and explanations." CoRR, abs/2112.08759, 2021
- [16] S. Bobek, M. Kuk, G. J. Nalepa, "Enhancing cluster analysis with explainable AI and multidimensional cluster prototypes," in IEEE Access, (submitted, under review)
- [17] S. Bobek, M. Kuk, J. Brzegowski, E. Brzychczy, and G. J. Nalepa. "KnAC: an approach for enhancing cluster analysis with background knowledge and explanations." Applied Intelligence (submitted, under second round review)
- [18] E. Brzychczy, M. Szpyrka, J. Korski, G. J. Nalepa, Imperative vs. declarative modelling of industrial process. The case study of the longwall shearer operation, 2021. (Submitted to IEEE Access).
- [19] K. Kutt, G. J. Nalepa, Loki – the Semantic Wiki for Collaborative Knowledge Engineering, 2021 (Submitted to Expert Systems with Applications).
- [20] M. Szelążek, S. Bobek, G. J. Nalepa, Semantic Data Mining Based Decision Support for Quality Assessment in Steel Industry, 2022 (Submitted to Expert Systems).
- [21] V. Rodriguez-Fernandez, D. Montalvo, F. Piccialli G. J. Nalepa, D. Camacho, DeepVATS: Deep Visual Analytics for Time Series (Submitted to Knowledge-Based Systems)
- [22] S. Bobek, V. Rodriguez Fernandez, M. Szpyrka, E. Brzychczy, M. Mozolewski, D. Camacho, G. J. Nalepa, Framework for Process-aware Analytics for Industrial Processes Based on Heterogeneous Data Sources and Domain Knowledge, 2022 (Submitted to Engineering Applications of AI)

Project-related events

- SEDAMI Semantic Data Mining Workshop at IJCAI2021
- PRAXAI - Practical applications of explainable artificial intelligence methods 2021 hosted at the 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2021
- Special issue on Machine Learning challenges and applications for Industry 4.0. Expert Systems: The Journal of Knowledge Engineering.
- Special issue on Effective and Efficient Deep Learning based Solutions. Neural Computing and Applications.
- Industry Meets Academia session at the School of Underground Mining Conference in 2021
- Special Session: Practical Applications of Deep Learning
- Industry meets Academia session with FAMUR at SEP 2020
- MIEL 2019 at BPM 2019

Go back to → projects