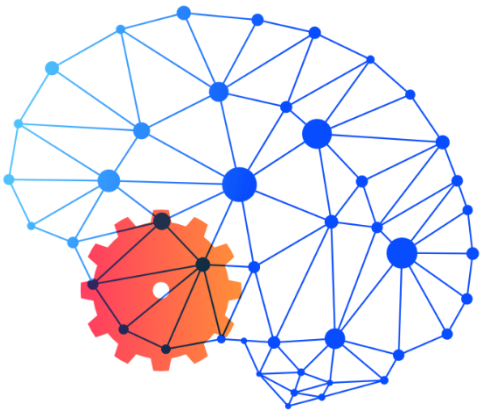

capri

Cognitive Automation Platform for European PRocess Industry digital transformation

Deliverable

D3.6 Reference Implementation of Cognitive Process Plants and Alignment with other cognitive initiatives

Deliverable Lead: Politecnico di Milano

Deliverable due date: 31/03/2022

Actual submission date: 31/03/2022

Version: 4.0

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Document Control Page	
Title	Reference Implementation of Cognitive Process Plants and Alignment with other cognitive initiatives
Lead Beneficiary	Politecnico di Milano
Description	Description of the cognitive solutions reference implementation with a focus on generalization for the process industry and the CAPRI addressed sectors. It will include an alignment of the Capri research work with the most recent advancements.
Contributors	Silvia Razzetti, Sergio Gusmeroli (POLIMI), Antonio Salis, Gabriele de Luca (ENG), Ana Pinto (AIMEN), Cristina Vega, Mireya de Diego, Laura Sanz, Anibal Reñones (CAR), Chritstoph Nölle, Norbert Holzknecht (BFI), Nenad Stojanovic (NISSA), Jakob Rehrl, Philip Clarke (RCPE), Asier Arteaga (SIDENOR)
Creation date	04/01/2022
Type	Report
Language	English
Audience	Public
Review status	☐ Draft ☒ WP leader accepted ☒ Coordinator accepted
Action requested	☐ to be revised by Partners ☐ for approval by the WP leader ☐ for approval by the Project Coordinator ☐ for acknowledgement by Partners

Document History			
Version	Date	Author(s)/ Reviewer(s)	Status
0.1	04/01/2022	Sergio Gusmeroli (POLIMI), Silvia Razzetti (POLIMI)	First ToC
1.0	07/02/2022	Silvia Razzetti (POLIMI)	Introduction
1.1	07/02/2022	Gabriele de Luca, Antonio Salis (ENG)	Chapter 2: TOC and assignments
2.0	22/02/2022	Silvia Razzetti (POLIMI)	Chapter 2 and 5
3.0	01/03/2022	Silvia Razzetti (POLIMI)	Chapter 3.1.1

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

3.1	08/03/2022	ALL	Chapter 2, 3, 4 Refinement of Chapter 5
3.2	15/03/2022	Silvia Razzetti (POLIMI)	Chapter 6 and deliverable finalisation
3.3	17/03/2022	Ana Pinto (AIMEN) Silvia Razzetti (POLIMI)	First Review
3.4	28/03/2022	Silvia Razzetti (POLIMI) Nenad Stojanovic (NISSA)	Finalisation of Chapter 4 and 5
3.5	28/03/2022	Ana Pinto (AIMEN)	Second Review
4.0	29/03/2022	Cristina Vega (CAR) Laura Sanz (CAR) Mireya de Diego (CAR)	Final Review

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Table of Contents

Table of Figures

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

List of Tables

List of Acronyms and Abbreviation	
Acronyms	**Description**
AI	Artificial Intelligence
API	Application Programming Interface
CAP	Cognitive Automation Platform
CS	Cognitive Solution
DoA	Description of Action
DT	Digital Twin
EDA	Exploratory Data Analysis
GUI	Graphical User Interface
IDE	Integrated Development Environment
IT	Information Technology
IoT	Internet of Things
IIoT	Industrial Internet of Things
MI	Mutual Information
ML	Machine Learning
MPC	Model Predictive Control
MQTT	Message Queue Telemetry Transport
OPC	Open Platform Communications
OS	Open Source
OSS	Open Source Solution

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

PCA	Principal Component Analysis
PI	Process Industry
PICO	Process Industry COgnitive
RA	Reference Architecture
RAP	Reclaimed Asphalt Pavement
SWOT	Strengths, Weakness, Opportunities and Threats
ToC	Table of Content
TSG	Trouble Shooting Guide
VIF	Variance Inflation Factor
WP	Work Package

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

DISCLAIMER

The sole responsibility for the content of this publication lies with the CAPRI project and in no way reflects the views of the European Union.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

EXECUTIVE SUMMARY / ABSTRACT SCOPE

Deliverable D3.6 – "Reference Implementation of Cognitive Process Plants and Alignment with other cognitive initiatives" is the document that accompanies the closure of WP3 – "Smart modules for cognitive process industry plants" at month M24. What will be presented in this deliverable corresponds to the achievement of milestone MS5 of technology validation of the Cognitive Automation Platform and related Modules. However, it's worth to mention that WP3's results will be further exploited in WP4 – "Cognitive technology solutions for process industry plants" and WP5 – "Prototype demonstrations of cognitive automation platform in CAPRI use cases", meaning that the activities don't really end at M24.

As presented in D3.1 – "CAPRI final reference architecture", the implementation of the CAPRI assets followed a bottom-up approach:

1. The 19 Cognitive Solutions (CSs) have been developed at laboratory level in WP3, for the three use cases (Asphalt, Steel, Pharma). This allowed to identify a number of requirements, in terms of infrastructure (functionalities, performance, analytics tools, cognitive capabilities) to implement the CSs at the production environment.
2. Taking into account the requirements, during WP3 the Reference Architecture of the CAPRI platform was defined to support the operations of the Cognitive Solutions.
3. WP4 aims to physically implement the CAPRI platform, starting from the Reference Architecture defined in WP3.
4. Finally, WP5 will validate the integration of the 19 CSs with the CAPRI platform and their deployment in the production environment.

Hence, the strong relationship occurring between WP3's activities and next WPs' tasks is quite evident, and the reason why in D3.6 we collected a number of recommendations and lessons learnt for WP4 and WP5 (but also for similar initiatives in the field of process industry).

The current document will not provide an overview of the Cognitive Solutions, as this is already presented in other deliverables (D3.2 – "CAPRI Industrial IoT Platform and Data Space", D3.3 – "CAPRI Industrial Analytics Platform and Data Space", D3.4 – "CAPRI Smart knowledge and semantic data models", D3.5 – "CAPRI Smart decision support"). Instead, D3.6 skips the point 1. of the above bullet list, and focuses on point 2., representing the natural conclusion of D3.1.

Actually, in the former deliverable (D3.1) we presented the final Reference Architecture defined at month M18 as a starting point for WP4's implementation activities inside CAPRI, while in D3.6 the objective is to assume a broader perspective in order to generalise the WP3 achievements, addressing also other process industry domains.

The initial step is the evaluation of the Open-Source components developed so far, since it guarantees more flexibility in terms of exploitation and adoption by external partners. Taking into account that the Cognitive Automation Platform (CAP) has been already designed as a series of interoperable open-source tools in D3.1, the focus of the current document is the Cognitive Solutions themselves. From this perspective, WP3 has analysed in detail the 19 CSs identifying three main scenarios: CSs implemented using an open-source language (Python mainly) and/or open-source components since the beginning; CSs implemented based on proprietary elements but whose features (all or some of them) can be easily translated as open-source; CSs that allow for the implementation of some open-source features on top.

Moving from the Architecture to the Data/Knowledge perspective, a number of different models to analyse data and extract insights from it have been developed, including data visualisation, correlation analysis, and anomaly detection. The approach followed consists in the application of analytical models to the CAPRI use case datasets, in parallel to the CSs development, running further investigation on them. Even when the analysis performed are not completely domain-agnostic, they can be easily applied to different datasets, also concerning other domains.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Finally, to complete the exploration of the different sectors, we focused on the SPIRE-06 cluster[1] projects aiming at identifying different approaches to be applied also in CAPRI use cases. A specific attention is paid to COGNITWIN project, in particular to the integration of the numerical models into StreamPipes[2] pipelines in the context of the creation of hybrid models. These models have been conceived to be easily customised for new model types, reason why it was chosen as a subject of study in CAPRI, aiming at understanding how it may be integrated in the CAP platform.

[1] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/dt-spire-06-2019

[2] https://streampipes.apache.org/

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

1 Introduction

1.1 Scope of Deliverable

Deliverable D3.6 – "Reference Implementation of Cognitive Process Plants and Alignment with other cognitive initiatives" aims at describing the CAPRI Cognitive Solutions Reference implementation with a focus on the generalization for the process industry and the CAPRI addressed sectors. As a sequel of D3.1, where the Reference Architecture has been defined leveraging on open-source components, D3.6's goal is to identify the open-source assets developed so far on top of the Cognitive Automation Platform. This corresponds to the achievement of milestone MS5 of technology validation of the Cognitive Automation Platform and Modules for the four levels of sensor, control, operation, and planning.

It includes the detailed analysis of the Cognitive Solutions, that are now mature enough to allow identifying relevant features to be implemented as open-source in order to guarantee a larger applicability outside CAPRI's borders.

Additionally, still in the perspective of the asset's openness, another goal of D3.6 is to provide a preliminary analysis towards the generation of Open Data, in order to identify the tools available for producing synthetic data and the CAPRI pilots' databases suitable for this purpose.

Last, but not least, D3.6 aims also to provide some recommendations and lessons learnt to take into account in WP4 and WP5, both in terms of features to be further evaluated and improved in next developments, and regarding innovative aspects implemented in WP3. This exercise is useful also for future projects and similar activities.

1.2 Audience

Deliverable D3.6 is a public report accessible to anyone interested in the topic. However, it is a technical document, and as so, addressed mainly to a reader with technical background related to Data Architecture and Dataset analysis that is familiar with the process industry domain (precisely, Asphalt, Steel and Pharma).

To properly understand this document, it is required to be accustomed with the concepts of architectural components and Open Source solutions, as well as a basic understanding of data analysis and data pipeline.

To avoid repetition of the previous deliverables, in D3.6 the 19 Cognitive Solutions are mentioned without providing detailed description. Hence, it is our recommendation to the reader to get acquainted with the CAPRI CSs before reading the document.

Within CAPRI project, D3.6 is addressed to:

- Technology providers involved in WP4 and WP5 that are expected to inherit WP3's achievements to further develop them. Besides the section about recommendations and lessons learnt, which is specifically addressed to next WPs, the entire document lists a number of activities to keep on or to begin doing, in the attempt of making assets open-source (both in term of cognitive solutions and data);
- Three project pilots, since they are directly involved in all activities regarding the implementation of the CAP, the deployment of the cognitive solutions and the exploitation of datasets;
- WP7, specifically to T7.1 – "Exploitation and business plan development" and T7.2 – "Open data of CAPRI cognitive solutions", since it deals with Open Data and the preliminary analysis run in D3.6 represents a fundamental starting point.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

1.3 Relationship with other deliverables

D3.6 is strongly related with D3.1 since it is its conceptual sequel. If the latter provides the detailed description of the Reference Architecture and the results of the preliminary analysis on the pilot's dataset, the former complements it by: i) presenting the openness approach on each cognitive solution, ii) further deepening the analysis of data, and iii) investigating the generation of Open Data.

As mentioned, D3.6 does not contain the detail description of the CSs, so to have a complete picture of the activities run in WP3 it is required to take into account other deliverables, such as D2.2 – "Use case requirements for cognitive technologies applications in use cases" (of type Report, delivered before the implementation of the CSs) or D3.2, D3.3, D3.4 and D3.5 (of type Other, that contain the final results).

On the other hand, due to the evident relationship of WP3 with WP4 and WP5, D3.6 is expected to impact the deliverables of those WPs, which will base their tasks on the guidelines provided in D3.6.

1.4 Document Structure

The document is organized in four main chapters (Section 2 – Section 5), beside the introductive chapter (the current **Section 1**, where the purpose of document, the target audience and the structure are described) and the conclusive one (**Section 6**, summarizing main achievements and addressing future activities).

Section 2 – CAPRI Data and Knowledge Reference implementation provides an overview of the analysis conducted on the pilot's datasets, specifying the techniques and approached used, besides the implementation of the CSs. Additionally, a preliminary investigation regarding Open Data is provided, leveraging on the generation of synthetic data on top of the pilots' datasets, identifying which are the existing tools and the suitable databases to perform this task.

Section 3 – CAPRI Open Source Reference implementation analyses the 19 Cognitive Solutions considering three different "Open Source scenarios", which will be followed during the implementation phase of the CSs, integrating them in the CAP and deploying them in the pilots' plants.

Section 4 – Data Pipelines and Process Industry – COGNITWIN collaboration aims at describing the approach followed in COGNITWIN project to manage data, based on Data Pipelines, and at identifying its points of interest for CAPRI use cases, in order to make it applicable also in different scenarios.

Section 5 – Recommendation and lessons learnt lists several suggestions for the platform and architecture (WP4) and for the demonstration in the pilots (WP5) (but also for other projects). These suggestions were collected by the project partners with the purpose of identifying the strengths of the current Reference implementation, as well as the weaknesses to be turned in opportunities in the next WPs.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

2 CAPRI Data and Knowledge reference implementation

In parallel to the development of the Cognitive Solutions and the definition of CAP Reference Architecture, the CAPRI use cases represent fertile scenarios to test new cognitive approaches in the field of Data and Knowledge.

In D3.1 – "CAPRI final reference architecture", the PICO (Process Industry Cognitive) architecture based on D2Lab was presented, providing an overview of its main features and mapping them with the CAP layers, in order to introduce an additional cognitive component at disposal of the three pilots (and of the process industry in general). The final objective was to include into the CAP an additional cognitive process applied to the pilots' data, able to reproduce three fundamental human-processes: **Cognitive Perception** (to get data and information from the system), **Fast Thinking** (to identify unexpected variations and promptly react) and **Slow Thinking** (to evaluate complex situations and to support decision making). For more details, see D3.1.

So far, an exploratory approach was adopted at the level of the "Sensor layer", mainly leveraging on raw data coming directly from the sensors, to identify the most suitable type of analysis that can be performed, especially in the context of Fast Thinking; the final objective is to connect the PICO platform with the CAP in order to get data as input and return visualisation alarms as output.

This analysis is still on-going, but some interesting results are available. The following paragraphs detail the investigation run so far, updating the preliminary results shown in D3.1.

Finally, due to the high importance of Data in the development of Cognitive processes, an analysis about Open Data is provided.

2.1 Pilots' dataset analysis

The Data and Knowledge reference implementation is performed at domain level: so far, for the Asphalt use case, it is strongly related to the implementation of the Cognitive Solution CAO1 (predictive maintenance of baghouse); while for the Steel and Pharma domains, further analysis is being conducted, independently of the CSs.

2.1.1 Asphalt use case

An extensive **Exploratory Data Analysis** (EDA) of real data of the **predictive maintenance of baghouse (CAO1)** was performed, being a continuation of the initial analysis performed in D3.1 - "CAPRI final reference architecture". The EDA was employed in order to understand the latent trends in the baghouse data using temporal dimensions. For this purpose, a variety of types of visualizations such as bar plot, box plot, distribution plots were used, as well as a comprehensive study regarding feature dependencies and collinearity. The sensor's data was collected from April 2021 to November 2021, providing information on:

- Baghouse filters drop pressure
- Drying drum drop pressure
- Baghouse temperature
- Electric Power blower
- AMPERAGE Exhauster
- ELECTRICITE Exhauster
- Information about production orders like: Formula (Recipe) Code (SPA0200), production flow (SPA0400), Formula (Recipe) Name (SPA0300), Final Product Temperature (ºC) (SPA0600), dosification setpoint (SPB0105), bitumen percentage (SPB0110) etc.)
- Maintenance history data

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Notice that the obtained data did not include any information regarding

- Exhaust power (%)
- Exhaust gases pipe drop pressure
- Dust emission in the clean gas chamber
- Combustion gases analyser

, as the appropriate sensors have not been installed yet.

Initially, the data was cleaned, transformed and synchronized in order to develop a single intermediate dataset containing all the sensors information. An example of the developed dataset is presented in Figure 1.

Date	Drying drum drop pressure	Baghouse filter drop pressure	Baghouse temperature	ELECTRICITE Exhausteur (KWH)	AMPERAGE Exhausteur (A)	SPA0400	SPB0105
2021-04-05 06:12:35	6.801694	146.284723	71.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:12:40	6.801694	149.007440	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:12:45	5.435237	150.363700	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:12:50	5.435237	150.363700	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:12:55	5.435237	149.007440	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:13:00	5.435237	146.284723	71.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:13:05	5.435237	146.284723	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:13:10	5.435237	147.651180	70.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:13:15	5.435237	149.007440	71.0	4.4392	459.7083	150.0	10.0
2021-04-05 06:13:20	5.435237	151.730157	71.0	4.4392	459.7083	150.0	10.0

Figure 1: Intermediate data set (CAO1)

From Figure 2 to Figure 6 the line plot of the numerical features are presented: on Figure 2 the “Baghouse filter drop pressure”, on Figure 3 the “drying drum drop pressure”, on Figure 4 the “Baghouse temperature”, on Figure 5 the “ELECTRICITE Exhauster”, and on Figure 6 the “AMPERAGE Exhauster”. The interpretation of results suggests that “Baghouse filter drop pressure” and “drying drum drop pressure” are characterised by small time-intervals, while “ELECTRICITE Exhauster” and “AMPERAGE Exhauster” are characterised by large time-intervals with stable values

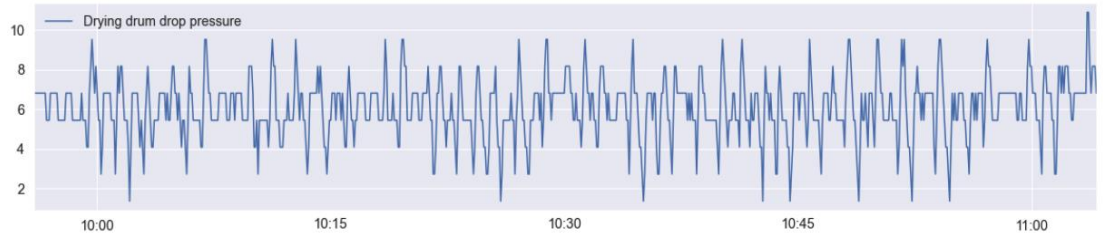

Figure 2: Line plot of Baghouse filter drop pressure

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

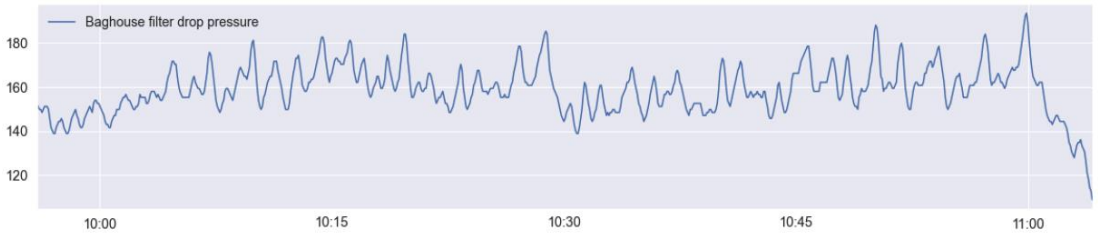

Figure 3: Line plot of Drying drum drop pressure

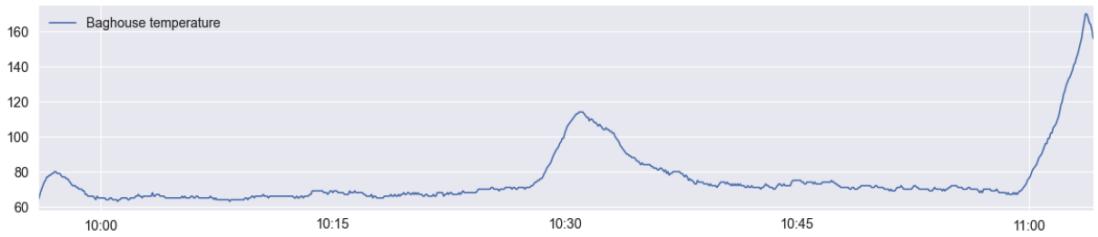

Figure 4: Line plot of Baghouse temperature

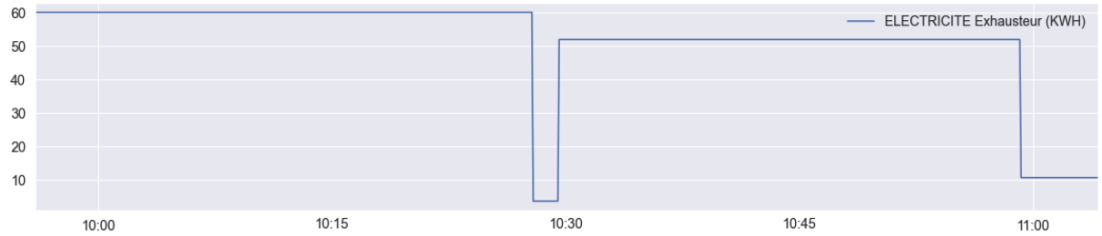

Figure 5: Line plot of ELECRICITE Exhauster (KWH)

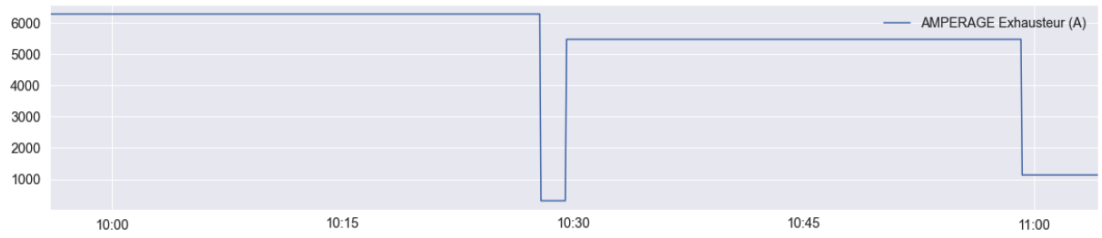

Figure 6: Line plot of AMPERAGE Exhauster (A)

Additionally, Figure 7 and Figure 8 present the line plots of the production flows of SPA0400 and SPB0105, respectively. Clearly, both SPA0400 and SPB0105 are characterized by many time-intervals with stable values. Approximately, all production flow features change their values every 30-60 minutes. Notice that similar conclusions can be taken for the rest of production flow features, and for that reason they were omitted from the data analysis.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Figure 7: Line plot of production flow (SPA0400)

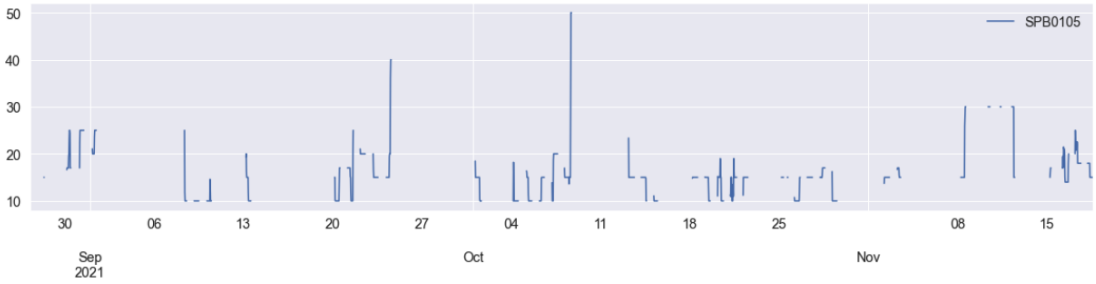

Figure 8: Line plot of dosification setpoint (SPB0105)

Figures 9-15 present the distribution plots and the box plots of the utilized features. The distribution plots are used to study the shape of the distribution, the symmetry possession, the inclination tendency, and the uniformity. Box plots summarize information such as the minimum, first (lower) quartile, median, third (upper) quartile, and maximum in compact form. Additionally, these plots are also utilized to quantify extreme values.

From the distribution plots, some conclusions can be taken:

- The shape of the distribution of Baghouse temperature, is similar to the shape of Normal distribution.
- The distributions of AMPERAGE Exhauster(A) and ELECTRICITY Exhauster' are right skewed.
- The distributions of AMPERAGE Exhauster(A) and ELECTRICITY Exhauster' have similar behaviours.
- The largest value accumulation for each feature i.e., in baghouse filter drop pressure is 110, in ELECTRICITY Exhauster is 70, baghouse temperature is 72 etc.

Additionally, the following conclusions can be made from the box plots:

- The features 'Baghouse temperature', 'Drying drum pressure', 'AMPERAGE Exhauster(A) and ELECTRICITY Exhauster' possess many outliers.
- The boxplot of Baghouse filter drop pressure is comparatively low with respect to the others. This suggests that the overall values have a high level of agreement with each other.
- The boxplots of Drying drum drop pressure, of dosification, and of Baghouse temperature setpoint are comparatively low. This suggests that the overall values have a high level of agreement with each other.
- The boxplots of Baghouse filter drop pressure of production flow are comparatively high. This suggests that the overall values have a low level of agreement with each other on this aspect or sub-aspect.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

- By taking into consideration the previous analysis, for each feature the proper transformation should be applied for later building and training of the Machine Learning model.

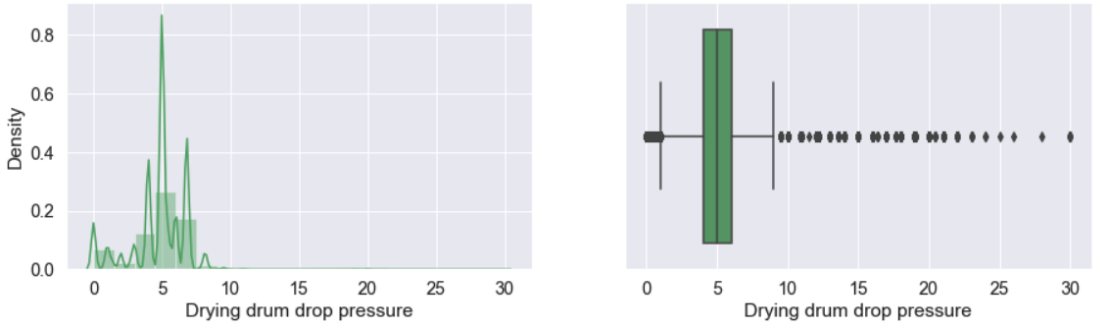

Figure 9: Distribution plot (left) and box plot (right) of Baghouse filter drop pressure

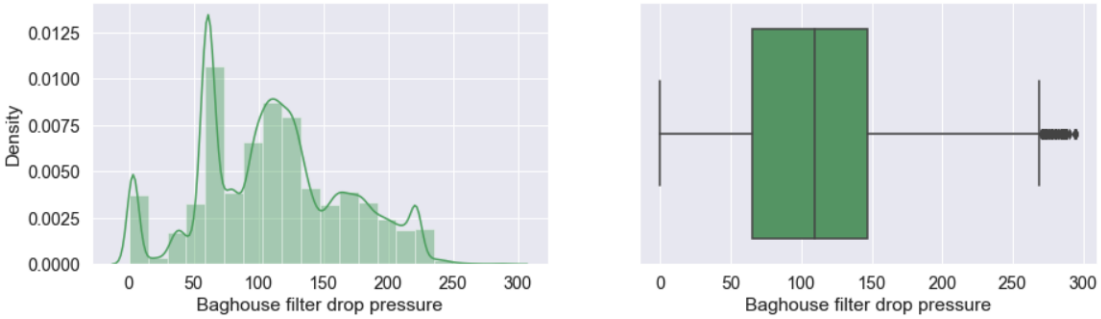

Figure 10: Distribution plot (left) and box plot (right) of Drying drum drop pressure

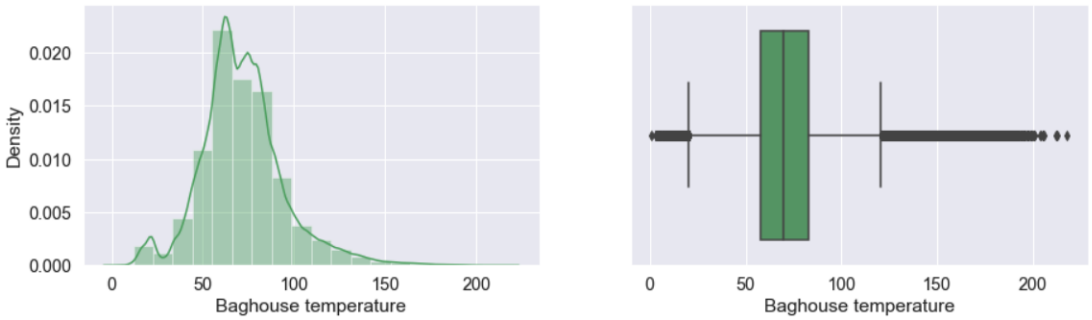

Figure 11: Distribution plot (left) and box plot (right) of Baghouse temperature

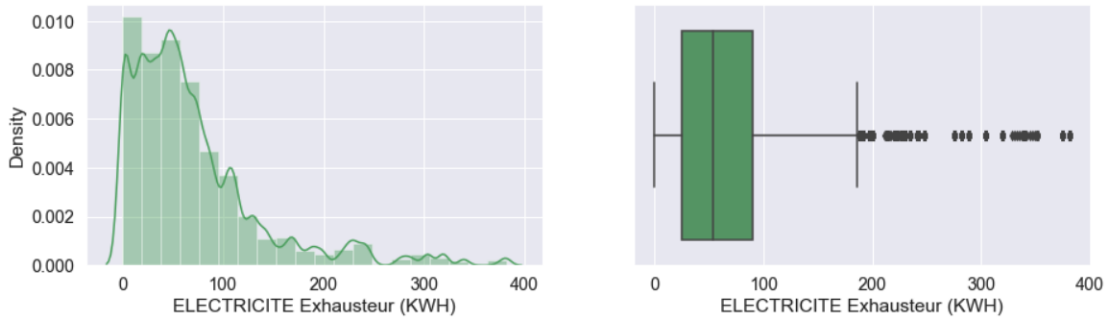

Figure 12: Distribution plot (left) and box plot (right) of ELECRICITE Exhauster (KWH)

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

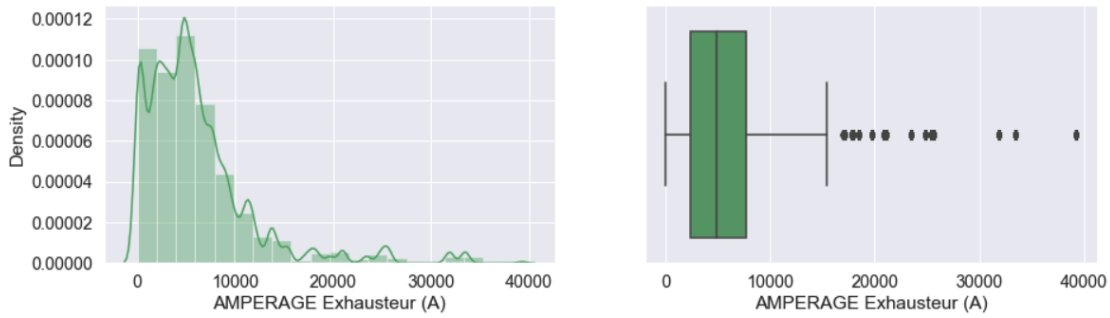

Figure 13: Distribution plot (left) and box plot (right) of AMPERAGE Exhauster (A)

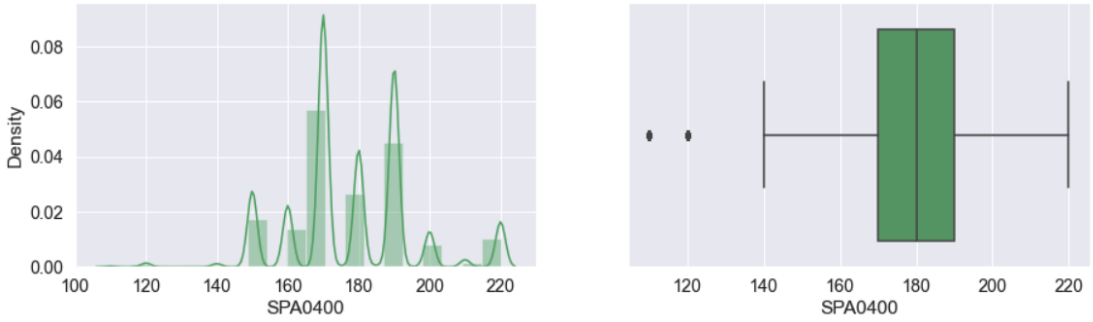

Figure 14: Distribution plot (left) and box plot (right) of production flow (SPA0400)

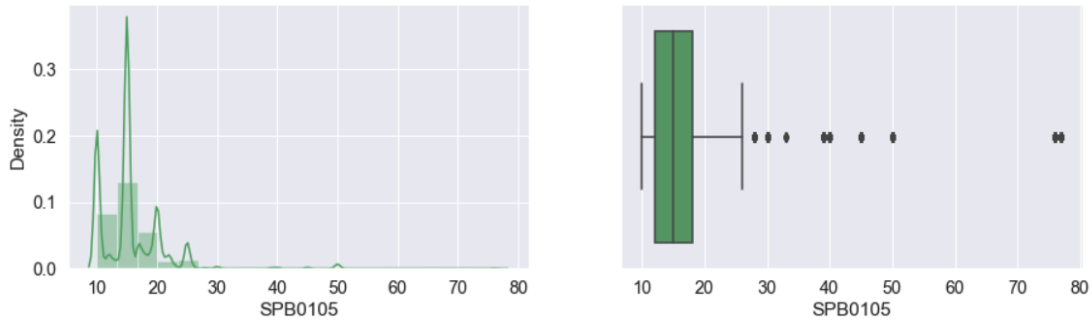

Figure 15: Distribution plot (left) and box plot (right) of dosification setpoint (SPB0105)

Furthermore, the daily box plots (Figures 16-20) of the utilized features were drawn to identify any seasonality. The main results are:

- For all features, the box plot on Saturday presented different median, Q1 and Q3 compared to the box plots on the other weekdays.
- The “Baghouse filters drop pressure”, “Baghouse temperature” and “ELECTRICITE Exhausteur (KWH)” report similar behaviour between Monday and Friday.
- Finally, no daily seasonality was identified for all features.

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

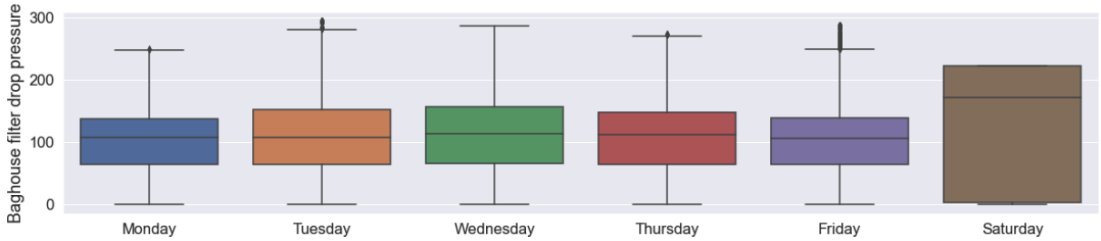

Figure 16: Box plot of daily Baghouse filter drop pressure

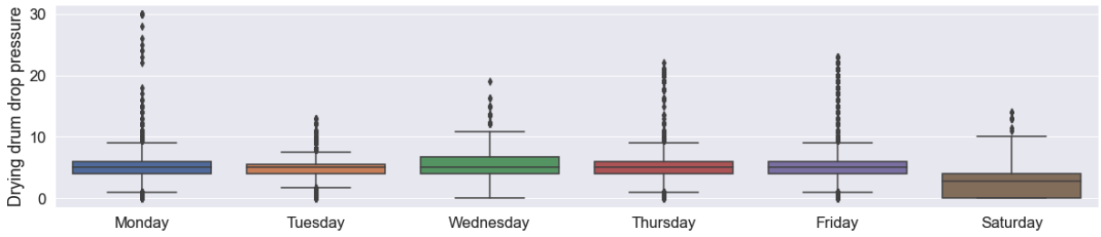

Figure 17: Box plot of daily of Drying drum drop pressure

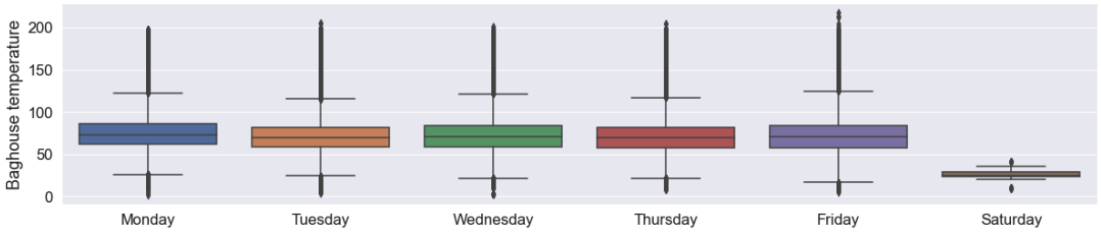

Figure 18: Box plot of daily of Baghouse temperature

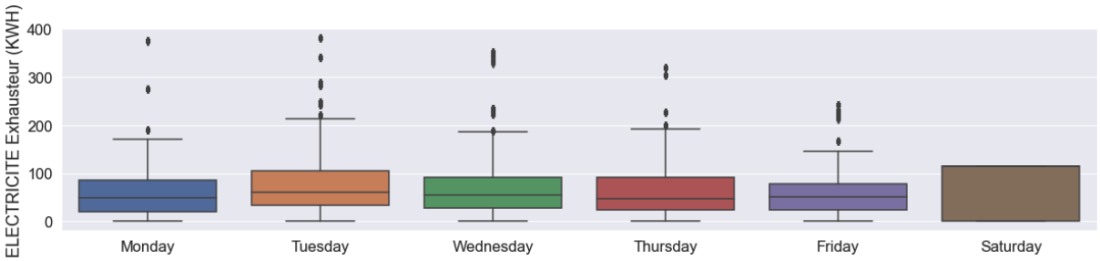

Figure 19: Box plot of daily ELECRICITE Exhauster (KWH)

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

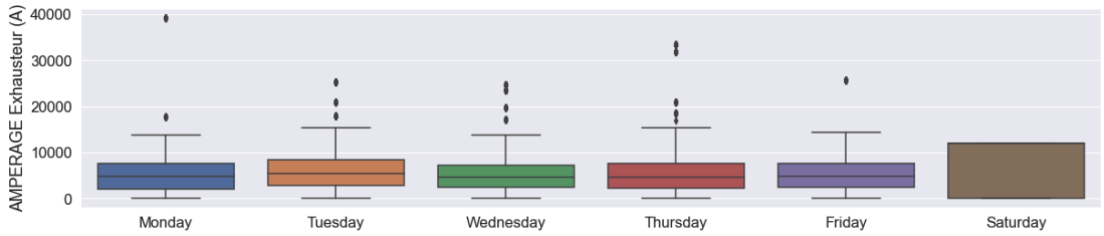

Figure 20: Box plot of daily of AMPERAGE Exhauster (A)

2.1.1.1 Collinearity

A comprehensive correlation analysis was performed in order to identify and examine the dependencies between the relevant features. In D3.1 - "CAPRI final reference architecture", the Pearson correlation matrix was presented to identify any linear dependencies between the features. In the previous analysis, it was stated that only "AMPERAGE Exhauster (A)" and "ELECTRICITY Exhauster" were found to be highly correlated.

However, the major disadvantage of the utilization of Pearson correlation matrix is that no information can be obtained about the non-linear relationship between the features as well as any multi-collinear information. To overcome this, it was used a correlation analysis based on mutual information (MI), which was originally proposed by Ross in 2014[3]. MI is a measure of the relationship between variables, which unlike Pearson correlation, is valid also for non-linear relationships. Nevertheless, for the linear correlation case both MI and Pearson are equivalent. Additionally, the effect of other variables can be removed using a partial correlation. These features make MI a better correlation measure for exploratory analysis of many variable pairs.

Figure 21 presents the MI matrix of all pairs of features included in the data. Clearly, the features "Baghouse filter drop pressure", "Baghouse temperature" and "ELECTRICITE Exhauster" are highly correlated (i.e. correlation > 0.8), as well as the features "AMPERAGE Exhauster" and "ELECTRICITE Exhauster". The grey parts denote that the MI could not be calculated, and therefore, the information related to the non-linear dependencies between the features are limited.

[3] Ross, B. C. (2014). Mutual information between discrete and continuous data sets. PloS one, 9(2), e87357.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

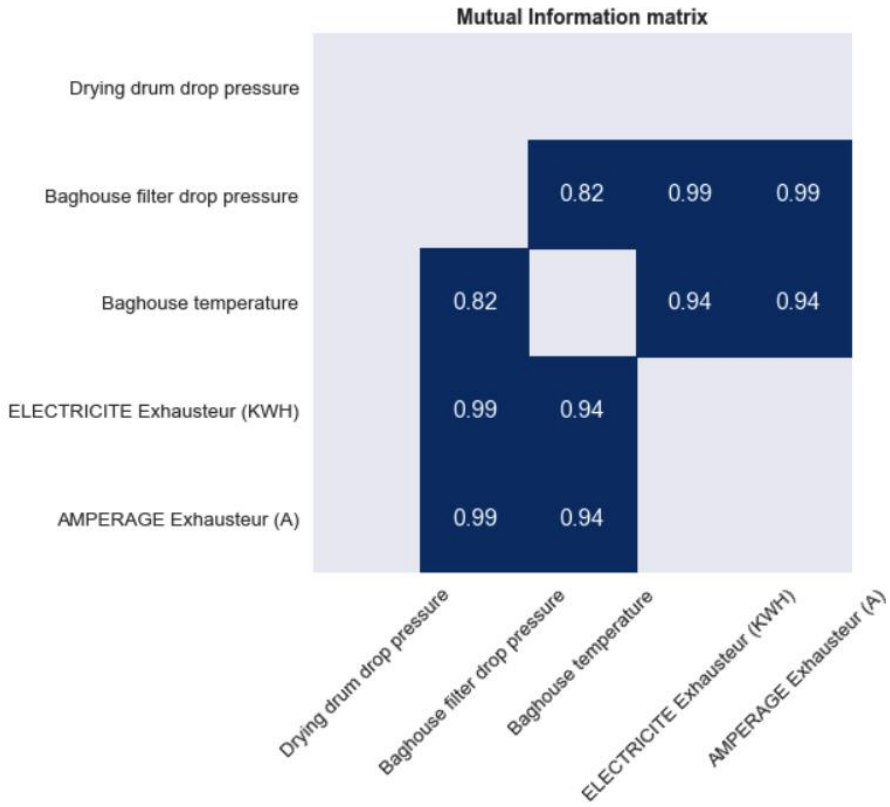

Figure 21: Mutual Information matrix

Finally, to detect multi-collinearity between the features, the Variance Inflation Factor (VIF) was calculated. VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model and it is calculated by:

$$VIF = \frac{1}{1 - R_i^2} = \frac{1}{\text{Tolerance}}$$

where R_i^2 represents the unadjusted coefficient of determination for regressing the i-th independent variable (feature) on the remaining ones.

Table 1 presents the VIF and Tolerance values reporting all features as moderately correlated since all VIF values are in $[1,5)$.

Table 1 VIF values and collinearity status

Feature	VIF	Tolerance	Status
AMPERAGE Exhausteur (A)	3.662024	0.273073	Moderately correlated
ELECTRICITE Exhausteur (KWH)	3.661554	0.273108	Moderately correlated
Baghouse filters drop pressure	1.484779	0.673501	Moderately correlated

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Drying drum drop pressure	1.403375	0.712568	Moderately correlated
Baghouse temperature	1.099836	0.909227	Moderately correlated

2.1.1.2 Next Steps

Based on the presented exploratory data analysis, the next procedures are:

- The integration of information from historical maintenance data.
- The implementation of advanced anomaly detection techniques to identify any abnormal behaviours of the baghouse[4,5,6].
- The application of novel machine learning techniques for predicting the health index of a component of baghouse.

2.1.2 Steel use case

This section contains preliminary exploratory data analysis regarding **Sidenor** steel company. The main goal is to **compare the analysis performed by D2Lab Online** software and the manually conducted analysis.

2.1.2.1 Dataset description

The obtained data consists of two stages:

- The first stage consists of **Measurement Data** (fixed, non-volatile parameters, measured only once);
- The second stage consists of **Time Series** (parameters measured frequently at certain timestamps).

The files containing **secMet** parts in their names contain measurement parameters, while the files containing **cc** in their names contain time series.

2.1.2.2 Brief parameter explanation

The measurement data and its details are presented in Table 2, while the Time series data's details are presented in Table 3:

Table 2 Parameters of Measurement Data

Variable	Description	Unit
heatnumber	Heat ID Number	

[4] Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: A machine learning perspective. Crc Press. ISO 690

[5] Pang, G., Cao, L., & Aggarwal, C. (2021). Deep learning for anomaly detection: Challenges, methods, and opportunities. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 1127-1130).

[6] Yao, D., Shu, X., Cheng, L., & Stolfo, S. J. (2017). Anomaly detection as a service: challenges, advances, and opportunities. Synthesis Lectures on Information Security, Privacy, and Trust, 9(3), 1-173.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

grade	Steel Grade (Sidenor internal nomenclature)	
vacdegasduration	Time in vacuum under 0.13 Kpa	minutes
tipsec	Type of heat in the sequence (N or A is first heat, S is same grade sequence and D is different grade Sequence)	
nsec	Number of this heat in the sequence	
totsec	Total number of heats in sequence	
finStirringDuration	Time from end of vacuum to end of secondary metallurgy	minutes
c	Carbon (final composition)	% * 1000
mn	Manganese (final composition)	% * 1000
si	Silicon (final composition)	% * 1000
s	Sulphur (final composition)	% * 1000
p	Phosphorous (final composition)	% * 1000
ni	Nickel (final composition)	% * 1000
cr	Chromium (final composition)	% * 1000
mo	Molybdenum (final composition)	% * 1000
al	Aluminum (final composition)	% * 1000
cu	Copper (final composition)	% * 1000
sn	Tin (final composition)	% * 1000
sb	Antimony (final composition)	% * 1000
as	Arsenic (final composition)	% * 1000
co	Cobalt (final composition)	% * 1000
pb	Lead (final composition)	% * 1000
ca	Calcium (final composition)	% * 1000
te	Tellurium (final composition)	% * 1000
ti	Titanium (final composition)	% * 1000
v	Vanadium (final composition)	% * 1000
b	Boron (final composition)	% * 1000

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

se	Selenium (final composition)	% * 1000
nb	Niobium (final composition)	% * 1000
n	Nitrogen (final composition)	ppm
h	Hydrogen (final composition)	ppm * 10
bi	Bismuth (final composition)	% * 1000
deoxSi	Ferro Silicon used for deoxidation (at tapping)	Kgr
deoxSiC	Silicon Carbon used for deoxidation (at tapping)	Kgr
deoxSiMn	Silicon Manganese used for deoxidation (at tapping)	Kgr
deoxAl	Aluminum used for deoxidation (at tapping)	Kgr

Table 3 Parameters of Time Series

Variable	**Description**	**Unit**
timestamp	timestamp of the signal values	
steelTempMeniscus	Steel temperature measured at tundish (at the beginning the value is not valid)	ºC
velCast_SX	Casting Speed in the strand X (1..6)	meters / minute * 100
lenght_SX	Length in the strand X (1..6)	meters
mouldWaterFlow_SX	Mould Water flow rate in the strand X (1..6)	liter / minute
mouldWaterTempDiff_SX	Temperature difference in the mold water in the strand X (1..6)	ºC
sprayWaterFlow_Z1_SX	Spray Water Flow Rate in the Zone 1 in the strand X (1..6)	liter / minute
sprayWaterFlow_Z2_SX	Spray Water Flow Rate in the Zone 2 in the strand X (1..6)	liter / minute
sprayWaterFlow_Z3_SX	Spray Water Flow Rate in the Zone 3 in the strand X (1..6)	liter / minute

To analyse data with **D2Lab Online**, one frame is created from both measurement data and time series to suit **D2Lab Online's** wide data format. The results frame consists of **86** columns and **9297** rows.

2.1.2.3 Analysis performed by D2Lab Online

Analysis performed by **D2LabOnline** is displayed in Figure 22.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

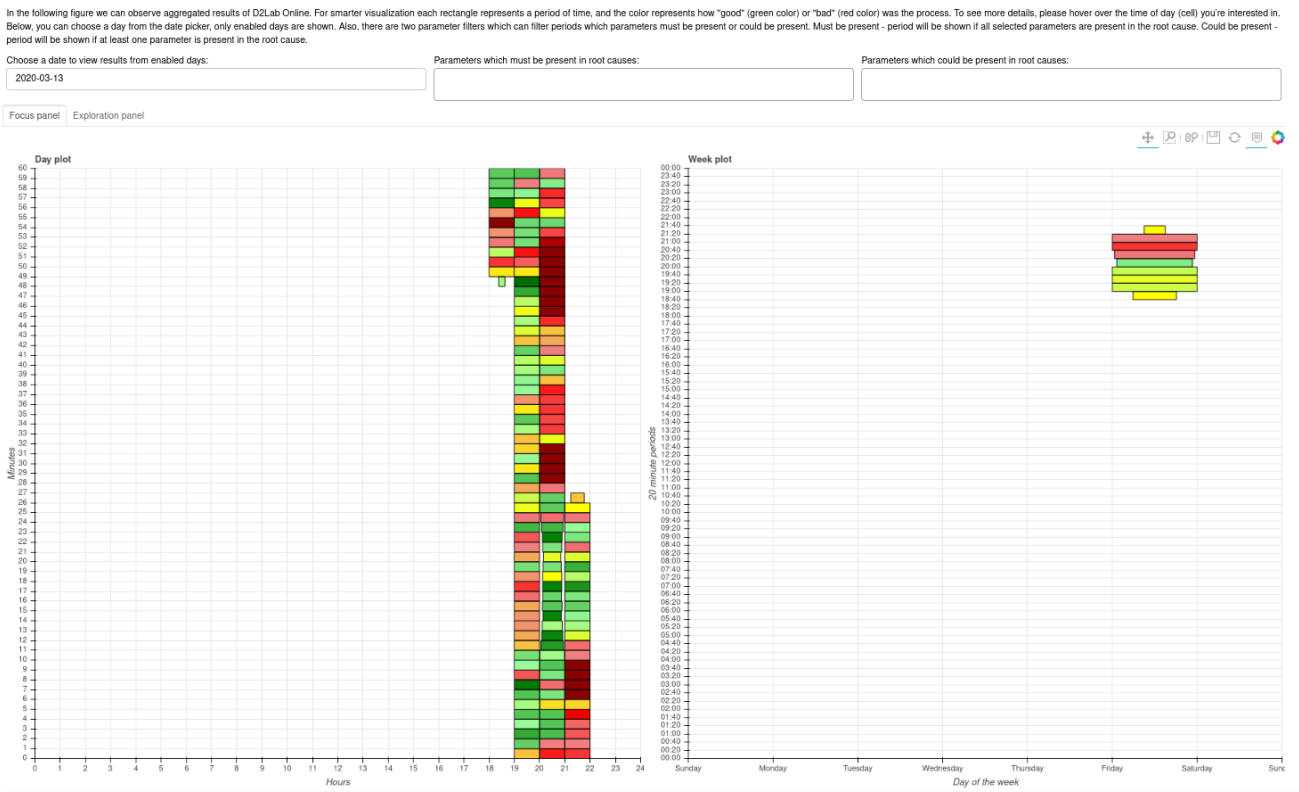

Figure 22: Analysis performed by D2LabOnline

By observing Figure 22, it is clear that only incomplete four hours of data are available, which were recorded on 13th March 2020. The dark green rectangles represent high satisfaction rate, opposite to dark red rectangles that represent low satisfaction rate (zero percent).

The main idea is to observe one “healthy” rectangle, represented with dark green colour (satisfaction rate greater than 90 percent) as opposed to time intervals containing sequential dark red rectangles (a few consecutive dark red rectangles).

The observed dark green rectangle represents a time interval of one minute, from 19:48 to 19:49 (Figure 23). The tooltip is highlighted by the blue rectangle (pops up when the green rectangle is hovered over). The satisfaction rate is high (96.67 percent).

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

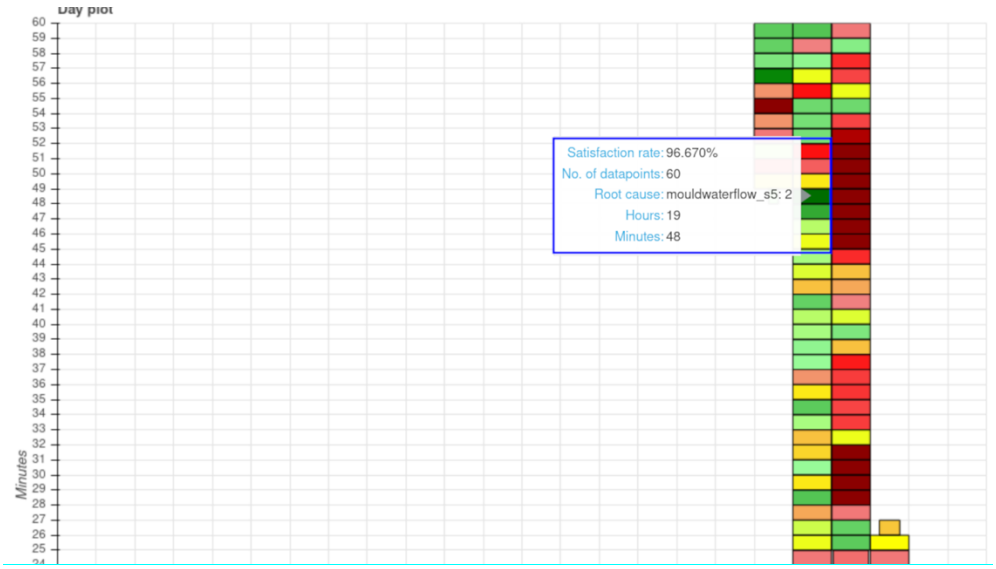

Figure 23: Tooltip representing time interval 19:48 – 19:49

As opposed to highlighted green rectangle represented in Figure 23, Figure 24 represents one dark red rectangle with satisfaction rate of 0 percent. This dark red rectangle also represents one minute interval 21:06 – 21:07.

Figure 24: Tooltip representing time interval 21:06 – 21:07

2.1.2.4 Statistical parameters comparison

Figure 25 displays statistics for all parameters contained within both created frames. The upper frame displays statistics for parameters present in frame marked as four sequential dark red rectangles. The lower frame contains statistics for all parameters in frame marked with dark green rectangle. Due to the number of columns, only part of statistical parameters is represented in the picture below.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

```
df_red.describe()
```

	Unnamed: 0	instance_id	vacdegasduration	nsec	totsec	finStirringDuration	c	mn	si	s	p	ni	cr	mo	al	cu	sn	sb	as
count	241.000000	241.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mean	2622.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	69.714896	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	2502.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	2562.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	2622.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	2682.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	2742.000000	210352.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
df_green.describe()
```

	Unnamed: 0	instance_id	vacdegasduration	nsec	totsec	finStirringDuration	c	mn	si	s	p	ni	cr	mo	al	cu	sn	sb	as
count	61.000000	61.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mean	7322.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	17.752934	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	7292.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	7307.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	7322.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	7337.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	7352.000000	210351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 25: Upper frame contains statistics for red rectangle, lower frame statistics for green rectangle

The main idea is to observe statistical parameters (mean, standard deviation etc.), derived from both created frames separately and compare statistics of corresponding parameters, as shown in Figure 26. For example, statistics of **steelTempMeniscus** parameter of red frame are highlighted with a red rectangle, while statistics of the same parameter in green frame are marked with green rectangle.

```
df_red.describe()
```

steelTempMeniscus	velCast_S1	lenght_S1	mouldWaterFlow_S1	mouldWaterTempDiff_S1	sprayWaterFlow_Z1_S1	sprayWaterFlow_Z2_S1	sprayWaterFlow_Z3_S
241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.0000(
1521.352697	151.410788	1098.269253	2098.863071	6.841621	56.692946	32.435685	21.7925
14.111612	22.376171	1.779191	2.379147	0.169600	6.980114	2.819023	1.4224
1451.000000	111.000000	1094.990000	2093.000000	6.383101	45.000000	27.000000	19.0000(
1525.000000	129.000000	1096.740000	2097.000000	6.759258	48.000000	29.000000	20.0000(
1526.000000	154.000000	1098.460000	2099.000000	6.903934	59.000000	34.000000	23.0000(
1526.000000	174.000000	1099.800000	2100.000000	6.932869	63.000000	35.000000	23.0000(
1528.000000	178.000000	1101.070000	2104.000000	7.077547	64.000000	36.000000	23.0000(

```
df_green.describe()
```

steelTempMeniscus	velCast_S1	lenght_S1	mouldWaterFlow_S1	mouldWaterTempDiff_S1	sprayWaterFlow_Z1_S1	sprayWaterFlow_Z2_S1	sprayWaterFlow_Z3_S1
61.000000	61.000000	61.000000	61.000000	61.000000	61.0	61.000000	61.0
1534.803279	169.524590	961.503230	2099.885246	7.198977	62.0	33.622951	22.0
0.400819	1.246416	0.502514	1.752126	0.036147	0.0	0.488669	0.0
1534.000000	167.000000	960.654000	2097.000000	7.135416	62.0	33.000000	22.0
1535.000000	169.000000	961.079000	2098.000000	7.193285	62.0	33.000000	22.0
1535.000000	170.000000	961.503000	2100.000000	7.193285	62.0	34.000000	22.0
1535.000000	170.000000	961.928000	2101.000000	7.193285	62.0	34.000000	22.0
1535.000000	172.000000	962.352000	2104.000000	7.251154	62.0	34.000000	22.0

Figure 26: Comparing statistics of corresponding parameters

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

2.1.3 Pharma use case

The general purpose of the Pharma dataset processing is to **monitor the quality of pills processing**. The main goal of the process is to perform the **fault detection of the system** and to alert the operator about the malfunction of the system, so that the required corrections are made.

2.1.3.1 Dataset description

The received dataset contains a total number of 20458 datapoints, with sample rate of 1s, with 207 parameters measured in each datapoint. The total size of received dataset is 39.4 Mb. The column DATE is defined as the key index.

The provided dataset contains data for approximately 6 hours, for time periods from 10.1.2020 at 5:38:37 until 10.1.2020 at 11:19:34. For this specific time period, the parameters of Table 4 did not contain any measurements associated, while all other parameters contain data for each datapoint.

Table 4. Parameters for which no data was provided within the 6 hours window

XXXA_BU_ForceControlO	TPSX_B_Modulation_On	PHXA_BDV_AutoAdjOn
STFX_BDV_Sample	XXXX_B_SingleReject	ORSX_BV_BadTablSampled
SCAX_BDU_WeighingOK	Activity	Area
Batch ID	Batch/Trial/Experiment ID	Category
Cell	LIMS Sample ID	Lot ID
Master Recipe ID	Order ID	Sample Nr
Start time	Test Index	Unit

Out of the 207 provided parameters, 129 parameters are marked as important parameters for processing.

According to the documentation provided about the Pharma use case, the processing pipeline was defined as shown on Figure 27.

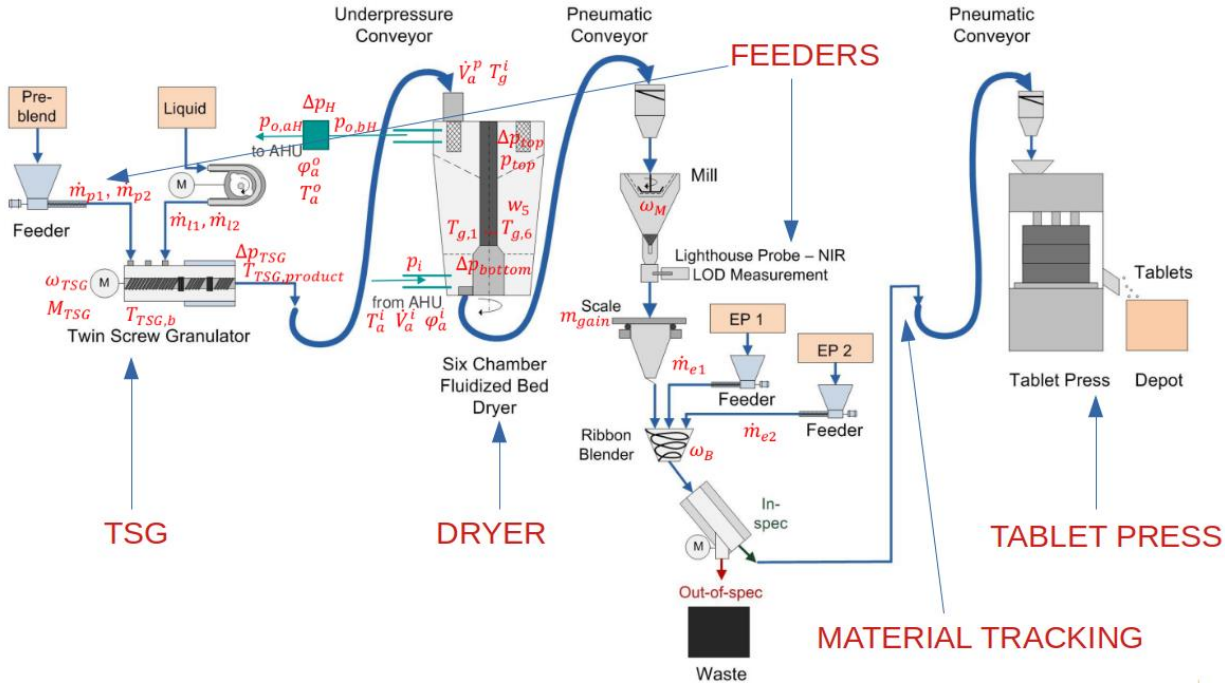

Figure 27: Pharma use case - Processing pipeline

In the specific process, **5 different stages** are defined:

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

- Tablet press
- Material Tracking
- Feeders
- Dryer
- TSG

Also, in order to complete the classification of the parameters, two more stages are defined:

- Others – relevant parameters but not classified as any of upper-mentioned stages
- Non-relevant parameters – other parameters not marked as relevant

The measurement locations for some of the parameters are presented in Figure 27. The same parameters are labelled with green cell in the column “of the flow chart”. According to the visual representation of the relevant parameters and their belonging to particular stages, the assumption is that the order of 5 stages is as presented in the Figure 27.

The main goal of the system is to detect and to prevent the possible appearance of fault scenarios in the system. The general idea is to prevent fault scenarios automatically and to alert the operator as fast as possible by analysing the data in real time using D2Lab.

2.1.3.2 Initial processing idea

The initial idea is to process the data and detect fault scenarios using D2Lab and PCA control chart.

Since there is no information whether the received data contains just processing data and/or all processing stage, the first task is to detect the processing phases in the data.

According to the processing pipeline, as we can see from Figure 27, the first relevant measured parameters are X_feeder_1_massflow (marked as m_{p1}) and XF_TSG_pump2_massflow (marked as m_{l2}). Also, in accordance with the processing pipeline, the last measured parameters are SAMX_SNDV_TotGoodCounter and XXXA_SNDV_TotalBadCntr, which count the number of goodly and badly generated pills. By analysing the changes in the values of these parameters, it was possible to determinate that the process was active between 10.1.2020 at 6:55:19 and 10.1.2020 at 9:44:00, of the received data. For the time period definition, datapoints that fall inside the limits are labelled as active, while the datapoints falling outside the limits are labelled as not active.

As D2Lab processes data on the level of stages, the active period for the received data should be separated on the level of stage. Since characteristic parameters for each stage have values for each datapoint, independently that the stage is active in particular time period or not, the next task is to teach the system to recognize which stage is active in specific time period.

As each dataset contains too many parameters, characteristic parameters for each stage are grouped and separately correlated. Using this approach, the number of parameters to analysed for each stage will be reduced, without losing significant information.

By analysing the processing pipeline and the 5 stages as presented in Figure 27, the following order of stage activation and characteristic parameters for each stage were defined:

1. stage: Feeders -> parameter: X_feeder_1_massflow
2. stage: TSG -> parameter: FC_TSG_pump2_speed
3. stage: Dryer -> parameter: LHP_NDC_CH1_OUTPUT_R.F_CV
4. stage: Material Tracking -> parameters: DT_CELL_INDEX_CELL_(1 to 5)
5. stage: Tablet press -> parameter: SAMX_SNDV_TotGoodCounter
6. stage: Others -> parameter: XFT_volume_flow_air_granulate_transport

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

For those parameters, different modes of operation are expected, therefore it can be assumed that these parameters define whether the specific stage is active or not.

The active state of the process as well as the activation states of each stage for the whole provided dataset are visualized on the figures below. Value 0 means that the state of the stage is inactive, while the value 1 means that the state of the stage is active.

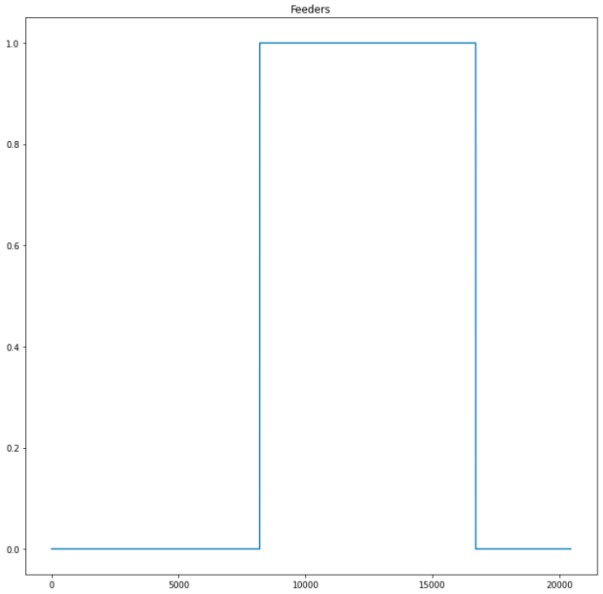

Figure 28: Activation of the Feeders stage in process

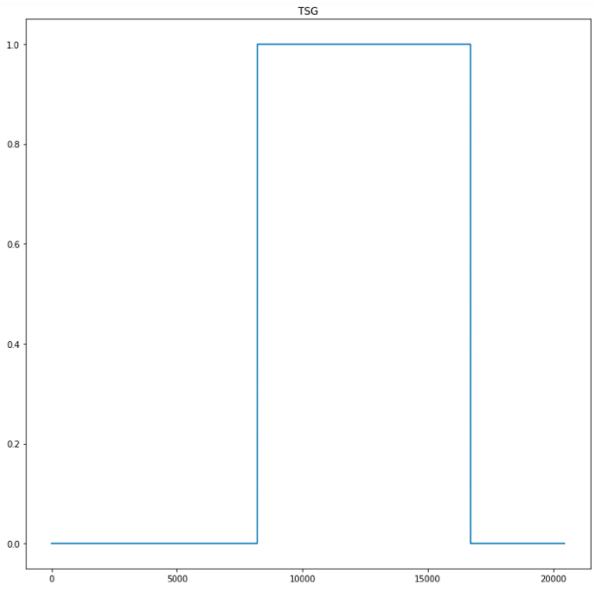

Figure 29: Activation of the TSG stage in process

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

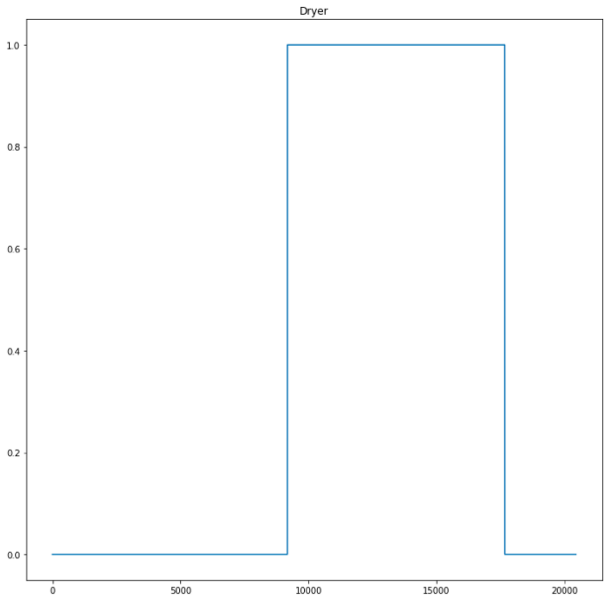

Figure 30: Activation of the Dryer stage in process

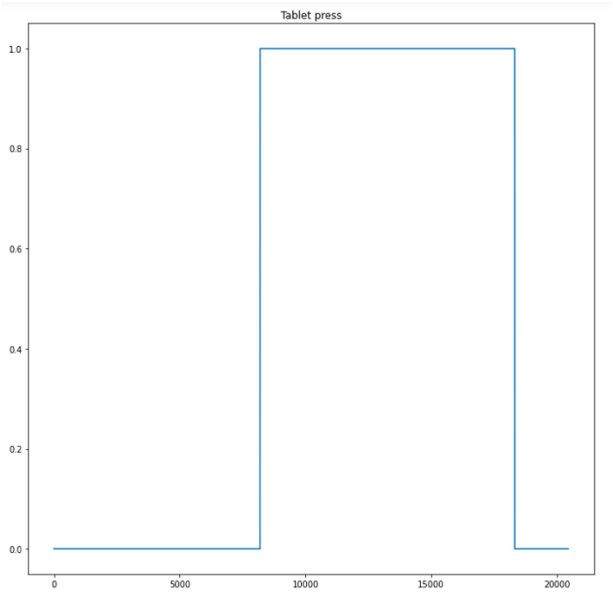

Figure 31:Activation of the Tablet press stage in process

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

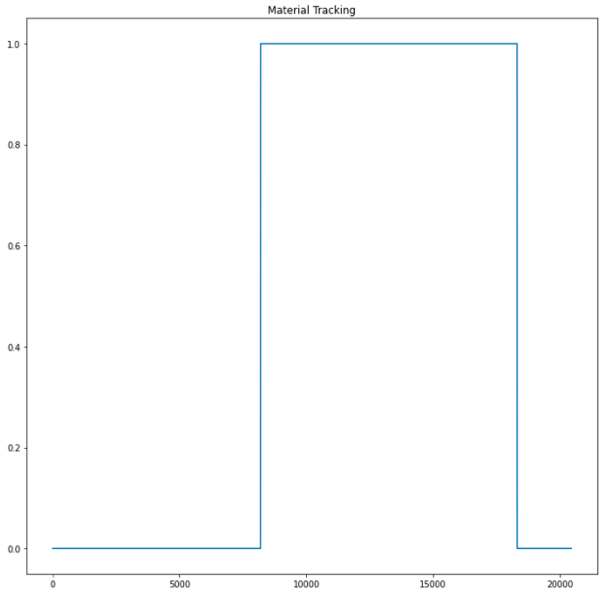

Figure 32: Activation of the Material Tracking stage in process

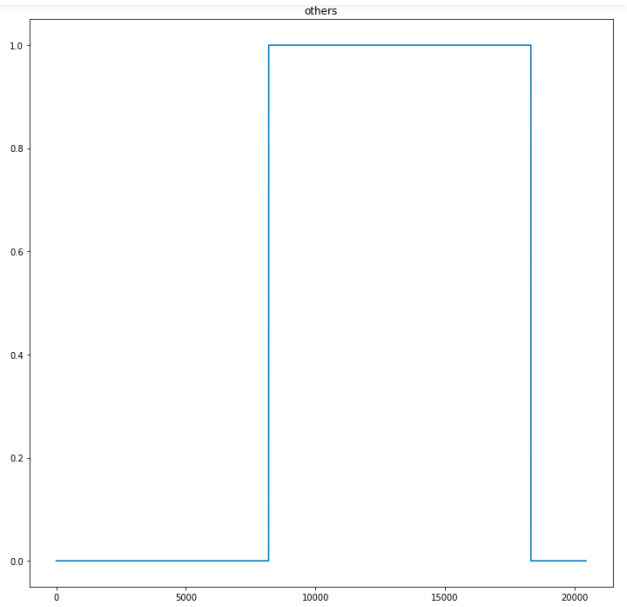

Figure 33: Activation of the Others stage in process

As it can be seen on the figures above, stages are overlapping each other in the process. This means that the activation of the process is consecutive and that multiple stages can be active at the same time. According to the activation times of the stages, additional labels for each stage were added to the data indicating whether the stage is active or not for a particular datapoint.

Therefore, after labelling the data, the data structure is generated as presented in Figure 34.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

	active	Feeders	TSG	Dryer	Tablet press	Material Tracking	others
10000	True	True	True	True	True	True	True
10001	True	True	True	True	True	True	True
10002	True	True	True	True	True	True	True
10003	True	True	True	True	True	True	True
10004	True	True	True	True	True	True	True
...	...	...	...	...	...	...	...
20453	False	False	False	False	False	False	False
20454	False	False	False	False	False	False	False
20455	False	False	False	False	False	False	False
20456	False	False	False	False	False	False	False
20457	False	False	False	False	False	False	False

Figure 34: Generated labeled structure for the data

The value in each cell of the dataframe is True if the stage is active for particular datapoint, else the value of the cell is False.

Using this way of data labelling, it should be possible to **automatically detect the state of the system** in each newly received datapoint in real time. To achieve this, different possibilities for training the data are explored.

Table 5 Results of processing stages detection using different methods

	Logistic regression	SVM	Random forest	Neural network
active	89,6%	89,6%	89,6%	49,5%
Feeders	100%	99,9%	100%	58,4%
TSG	100%	99,9%	99,9%	58,4%
Dryer	99,8%	95%	99,9%	58,5%
Tablet press	89,6%	89,6%	89,6%	49,5%
Material Tracking	89,6%	89,6%	89,6%	49,5%
Others	89,6%	89,6%	89,6%	49,5%
Average success	94%	93,3%	94%	53,3%

As can be seen from Table 5, the best results for the automatic detection of stages are provided by Logistic regression and Random forest methods. Therefore, most certainly one of those two methods should be further tested and used in the actual system.

Additional testing should be done by analysing a larger amount of data, since the presented results are based on the 18000 datapoints for training and on the 2458 datapoints for validation.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

2.1.3.3 Further steps

All described methods should be cross validated with the domain experts in order to proceed with further analysis. Additionally, discussions about further steps with domain experts should be done.

According to the current knowledge about the dataset, the next step to be performed, using the described approach, will be to test the stages' detection methods on larger amount of data, since the current amount of data is definitely not enough for taking valid conclusions. After that, the following step will be to adapt the data for processing inside D2Lab, and to perform the PCA on the generated instances. Ultimately, the final task should be the implementation of a visualization system.

2.2 Open Data

2.2.1 The value of open data

The intrinsic value of open data is that it promotes innovation and progress, ultimately advancing society through job creations, enhanced efficiencies, and economic stimulation[7]. Digital twins of aircraft engines and manufacturing plants can be created to mirror in real time the real-life versions. These digital replicas rely on data (such as weather, usage, activity and history) to predict faults, diagnose issues and target maintenance. If data is openly available, more factors could be inputted to improve the accuracy of the virtual simulation or the recommended adjustments to be made to avoid issues, leading to better efficiency.

Reusing data held by industry companies is particularly advantageous compared with self-collection, saving organisations considerable sums and administrative burdens. The quantity and quality of the available data increases when reusing data, whilst the bias risks associated with self-collected data decreases.

The cross-fertilisation of data across industries allows new value to be unlocked in old data. For instance, a joint report on 'How the Power of Data Will Drive the EU Economy[8]' provides examples of data from mobile telecommunication operators being used by these businesses for internal purposes, by EU statistical offices for official statistics on mobility and demography, as well as by the health industry to control and predict disease outbreaks.

2.2.2 Barriers

Despite the numerous benefits of open data, the **road to achieve open data is not straightforward**. Financial implications are also materially relevant where companies who have invested significant sums in data collection are reluctant to simply give their data away for free. Moreover, companies may fear liability or costs associated with inadvertently disclosing commercially sensitive or confidential information, affecting the competitiveness of their business. As organisations are not afforded foresight of who accesses their data when it is truly open, they may be dissuaded by the notion of their competitors prying on their information. Moreover, there is an enduring perceived public relations benefit in only sharing data for certain campaigns or for targeted reasons (such as for public health campaigns).

If companies opt to make data open, they must prepare the data by cleaning it, disaggregating it, merging it with other datasets, making it readily accessible in a common, interoperable standard. The lack of a preferred or universal method to share data makes this a more cumbersome process.

If privately held data is made open, it often suffers with data coverage bias stemming from an unrepresentative or narrow sample that is used during the data collection. Time, skills and costs are required to ensure that the data validly represents a particular market and a degree of trust is

[7] https://www.cms-lawnow.com/ealerts/2021/06/embracing-open-data-is-now-more-important-than-ever-open-data-note-2-of-2?cc_lang=en

[8] https://datalandscape.eu/sites/default/files/report/EDM_D2.2_First_Report_on_Policy_Conclusions_20.04.2018.pdf

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

required in organisations sharing and using the data. Similarly, time and resource intensive work is also required when manipulating data to extract its value, particularly where raw data is provided in an unusual or incompatible format (i.e. contrary to the above definition of openness). Technicians with appropriate skillsets and knowledge must be deployed for the data's value to be appreciated.

2.2.3 Open data creation in pilots

Our preliminary analysis has shown that the dataset described in Section 2.1.2 (Sidenor pilot) can be of interest for sharing as open data. This data represents the values of parameters used in a particular stage in the steel production process, which is continuously repeated (one heat). Since in each step (heat) the used equipment is degrading (due to harsh condition), this dataset is a very suitable example for inspecting the tool wear process (e.g. methods for detecting tool wear). However, the main problem on sharing the data is the reluctance of sharing their own manufacturing information (some of the reasons are covered in Section 2.2.2). A possible alternative can be the creation of synthetic data, as described in the following section.

2.2.4 Synthetic data

Recent advancements in deep learning and an increase in computational power facilitated the development and early adoption of an emerging anonymization and privacy protection technique: **AI-generated synthetic data**. Synthetic data is **artificial data that is generated based on original customer data**. If well generated, it is highly realistic and statistically representative of the original data and thus suitable to serve as a drop-in replacement for it (e.g., for AI training). Moreover, when generated with appropriate privacy mechanisms – synthetic data is fully anonymous and impossible to re-identify.

Besides creating replica datasets, synthetic data is also capable of augmenting data to reduce bias and to correct imbalances. Research[9] estimates that by 2022, 85% of algorithms will be erroneous due to bias. Bias and discrimination of AI systems are problems that are already being taken seriously, and synthetic data can contribute to mitigate bias with fair synthetic datasets representing the world, not as it is, but as we would like to see it.

Additionally, some types of data are costly to collect, or are rare. For instance, collecting data representing the variety of real-world road events for an autonomous vehicle may be prohibitively expensive. Bank fraud, on the other hand, is an example of a rare event. Collecting sufficient data to develop ML models to predict fraudulent transactions is challenging because fraudulent transactions are rare.

Therefore, generating synthetic data that reflects the important statistical properties of the underlying real-world data can solve these problems. It is inexpensive compared to collecting large datasets and can support AI/deep learning models development or software testing without compromising customer privacy. It's estimated that by 2024, 60% of the data used to develop AI and analytics projects will be synthetically generated (see Figure 35).

[9] https://research.aimultiple.com/synthetic-data/

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

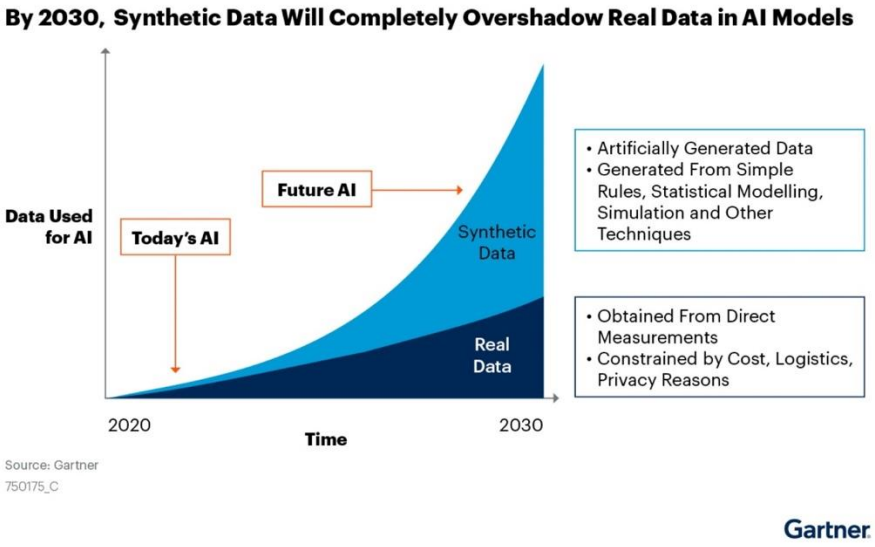

Figure 35: The role of Synthetic data for AI

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

3 CAPRI Open Source Reference Implementations

The Cognitive Automation Platform (CAP) Implementation, already described in D3.1, is an open-source components-based platform aiming to support the entire life data cycle from the sensor to the user/application levels. On the other hand, the platform is able to integrate data coming from external factories, as well as to share data under the data sovereignty principles thanks to the integration of the related vertical structure (Figure 36).

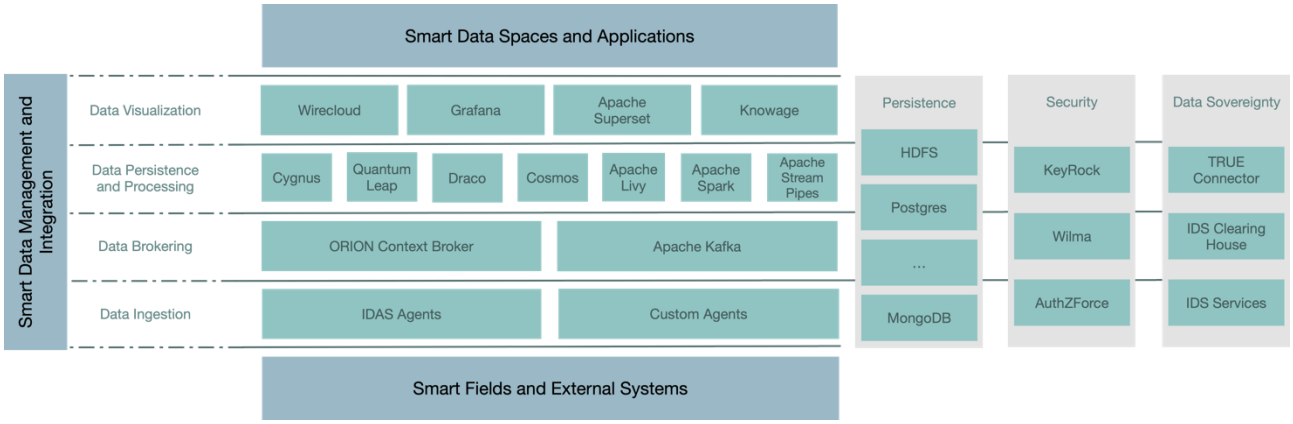

Figure 36: Cognitive Automation Platform Implementation v2

The implementation, taking into account requirements and information coming from the cognitive solutions, has been refined. In particular:

- The Data Brokering layer includes Apache Kafka, as an alternative way to ingest data from the factory (sensors, machines, devices, etc.).
- The Data Visualization layer has been extended to include Apache Superset, in order to design custom dashboards and cockpits to support the data monitoring, visualization and decisions making.

Apache Kafka is a widely used event streaming platform able to:

- Publish (write) and subscribe to (read) streams of events, including continuous import/export of data from other systems.
- Store streams of events durably and reliably for as long as needed.
- Process streams of events as they occur or retrospectively.

The integration of Apache Kafka allows extending the CAP platform, making it able to expand the ingestion layer considering the widespread of the brokering tool. In this way, a specific requirement defined in the Steel domain is covered: existing Kafka connectors can be used to bring data to the CAP platform, thus enabling the system integration of already working solution for data collection.

Apache Superset is an open-source data exploration and visualization platform designed to be visual, intuitive and interactive. It allows users to analyse data using its SQL editor and to easily create charts and dashboards. It has been integrated in the CAP platform since it represents a very modern tool with several features. This makes it much more attractive, reason for which the developers are pursuing it, mainly due to:

- Although still immature, the graphical interface is modern and the interactivity is high, especially when working with different types of datasets.
- The dashboards can be shared among the different users, allowing a more complete collaboration.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

- Data exploration can be shared among different users.
- Superset is much faster in terms of performance than other solutions, ensuring greater effectiveness in terms of data visualization.

The CAP implementation can be deployed in a custom way, taking into account user needs and technical/user constraints. Indeed, the open-source components can be combined to address specific needs in relation to the status and conditions of a single cognitive solution.

After a deep analysis concerning the state of the art of the cognitive solutions, in character with their related descriptions in D3.1, three different scenarios have been defined (Figure 37). Basically, the CAPRI solution and, consequently, the OSS components part of the CAP implementation can be adopted to support the Cognitive Solution, or as an alternative for replicating (or extending) a proprietary solution (or sub-system) in order to enrich the interoperability and improve the openness.

In the next sections, the Cognitive Solutions will be analysed, describing their current situation, constraints and technological limits, aiming to categorize all solutions and create a link with the fitting scenario. For each Cognitive Solution, the CAP implementation building blocks will be hit, describing how they are able to satisfy the requirements of the entire solution (Scenario 1), or partial solution (Scenario 2), or to cover the need for additional functionalities using open-source components (Scenario 3).

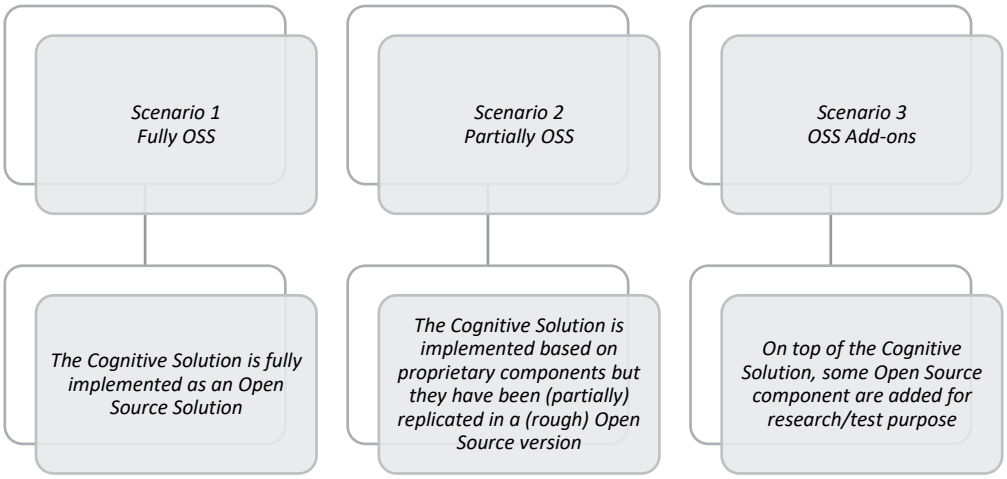

Figure 37: OSS scenarios in Cognitive Solutions

3.1 Asphalt use case

CAS1 – Sensor for bitumen content

CAS1 acquisition and control systems will be implement through c++/CUDA (not OSS), in order to make the most out of the embedded platform. Aravis library (licensed under GNU Lesser General Public License v2.1) will be used for the camera acquisition and openCV and TensorFlow (both licensed under Apache 2.0 License) will be used for data processing. These last three software solutions are open source. As so, a big part of CAS1's system will use some open-source solutions, becoming a part of the Scenario 2, i.e.: it is based partly on proprietary solutions, but it will have some components developed in an OS scenario. However, due to the possible patenting of CAS1, its acquisition and control systems will not be provided in open source. Nevertheless, during the implementation of CAS1 in the asphalt plant, it will feed its output data (time tag and bitumen %) using the OSS FIWARE based platform also, as can be seen in Figure 38.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

CAS2 – Sensor of Amount of Filler

CAS2 Cognitive Solution has not been implemented in an OSS scenario from the beginning. Its core solution is based on 2 different development technologies:

- Part of the cognitive solution is based on proprietary tools developed with the Labview programming and datalogging environment, that are be able to gather and analyse data coming from different sensors located at the asphalt use case production plant.
- The second part of the CAS2 solution will be based on an algorithm programmed using high level structure language like python using an Open Source IDE (Integrated Development Environment). But so far, the programmed code to estimate the filler quantity in the process (main output of the CAS2 solution) is not freely available.

Therefore, CAS2 Cognitive sensor solution is only partially based on OS components as development tools. On its current form, CAS2, partially based on Python programming language, could adapt some of its functionalities, implemented as proprietary Labview tools, to make them open and standard compliant, by using different OS tools.

In a brief way, CAS2 cognitive sensor solution is part of the Scenario 2, that is, it is based partly on proprietary solutions, but it will have some components developed in an OS scenario.

Anyway, on top of the Cognitive Solution CAS2, some Open Source components for data ingestion have been implemented. Additionally, its outputs are part of the data ingestion system using ORION as context broker (so CAS2 is “powered with FIWARE Reference Architecture” (see Figure 38)). A more detailed explanation of this architecture can be found below, in CAC1 subsection, where CAS2 output data feeds, among others, the CAC1 CS data ingestion using FIWARE based Reference Architecture.

CAC1 – Control of the asphalt drum

CAC1 Cognitive Solution has not been implemented in an OSS scenario from its beginning. Its core development solution is based on proprietary MPC (Model Based Predictive Control) libraries available in the MATLAB numerical computing and programming platform within the so-called Model Predictive Control Toolbox. It provides functions, an app, and Simulink blocks for designing and simulating controllers using linear and nonlinear model predictive controls, as well as specifies plant and disturbance models, horizons, constraints, and weights. In addition, the used model has been identified using the System Identification Toolbox, also part of the MATLAB environment, so it is not part of an OS solution.

By default, the CAC1 Cognitive control solution won’t be based or have any OS component within it. It has been developed using MATLAB scripting programming language (similar to C language) and different blocks of the mentioned toolboxes. Also, Simulink blocks have been used for the proper calculations and simulations. The whole CS solution has been compiled using the MATLAB compiler that enables the developer to share MATLAB programs as standalone applications and packages, as well as to deploy MATLAB programs. End users can run these applications using the MATLAB Runtime tool, which is a free tool that contains the libraries needed to run MATLAB applications on a target system without a licensed copy of MATLAB.

It is not expected that the different configuration blocks of CAC1 solution will be replicated implementing OSS components in order to make it open source and standard compliant, due to the nature of the MATLAB proprietary libraries regarding systems identification, MPC libraries and their very specific field of control systems.

Although some of their specifications can be found in other open source libraries, they are scattered throughout not integrated different solutions and they do not seem to be widely used and tested, as much as those based in MATLAB. A lot of additional work would be needed in order to fully integrate and allow compatibility with CAC1 specifications and requirements, which could lead to not fully tested undesired results.

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

As so, CAC1 cognitive solution is part of the Scenario 3, described on Figure 37. On top of the Cognitive Solution CAC1, some Open Source components for data ingestion have been implemented. This way, CAC1 standalone solution does not have any OSS features, just ORION as context broker. Like the rest of Asphalt Cognitive Solutions, CAC1 will be integrated within a FIWARE Reference Architecture which is fully Open Source based.

The OSS features that CAC1 uses are the following (see Figure 38):

- From the asphalt use case plant, all related production data and the data coming from all the new monitoring sensors that have been commissioned and/or developed is sent through a local datalogger using MQTT protocol.
- In a dedicated computer (called “CAPRI Server”, located at CARTIF partner premises), for data ingestion, a dedicated MQTT broker (mosquitto) receives and resends the data to a dedicated IDAS Agent. Specifically, the IoT Agent for JSON (fully OS, based on FIWARE technology) is used as a bridge between MQTT messages coming from the asphalt plant and NGSI (FIWARE API) where the corresponding Orion Context Broker receives all those messages, subscribing to the corresponding context information (current status) of the asphalt production plant.
- From the Orion context broker, a Draco module (fully OS and FIWARE based technology) is an enabler that allows data persistence for managing the context information (asphalt plant data). Based on Apache NiFi, it makes data routing to a dedicated MySQL (also OS based) database server to store all asphalt plant context information (all data received from the plant).
- These data can be accessed by CAC1 to perform its calculations and give the corresponding output variables (mainly temperature setpoints for the aggregates drying process within the asphalt production plant).

Other solutions, like CAS2, CAO1 and CAP1 feed their output data using this OSS FIWARE based platform also. This way, CAC1 can obtain data needed from these solutions using the MySQL database as data source to know asphalt plant context information.

This project receives funding in the European Commission’s Horizon 2020 Research Programme under Grant Agreement Number 870062

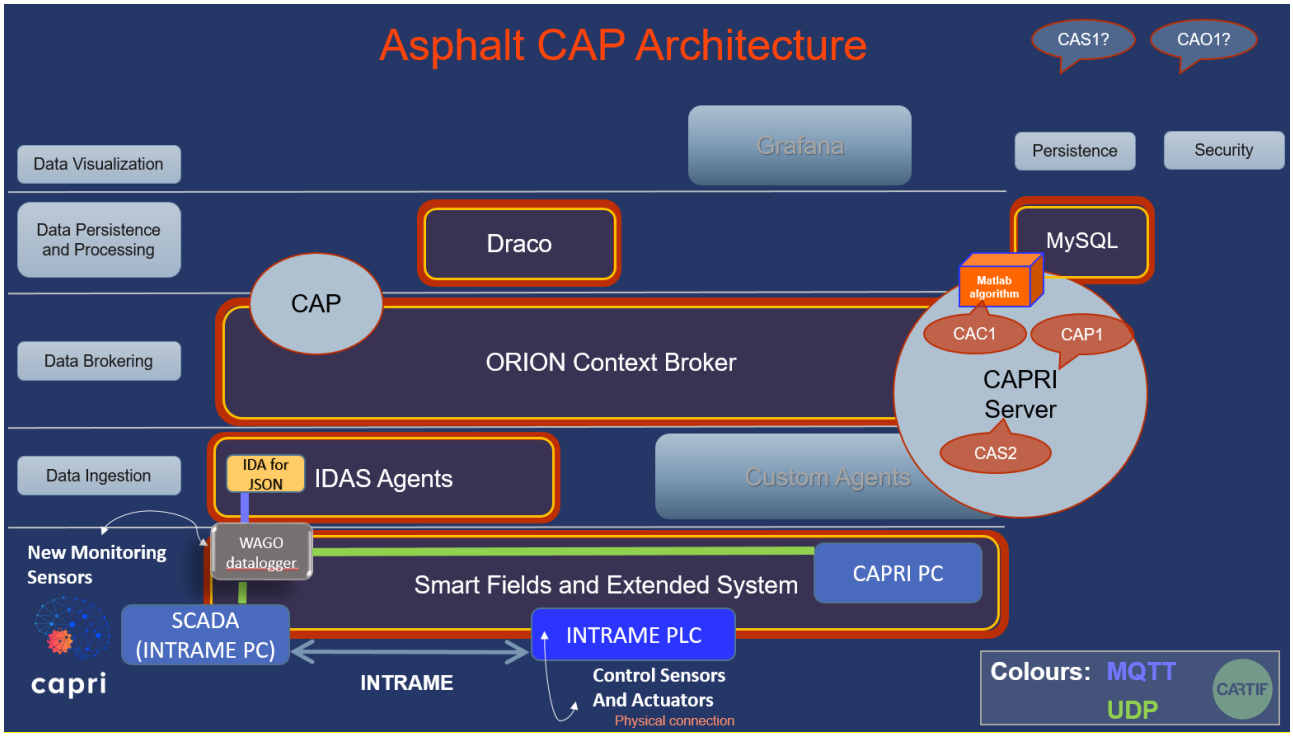

Figure 38: Asphalt CAP Architecture where CAC1 solution (among other asphalt CS solutions) is integrated

CAO1 – Predictive maintenance of the baghouse

The CAO1 solution provides integration for the anomaly detection model and the remaining of useful life of a component of the baghouse. The models are written in python using TensorFlow framework, and will be deployed using TensorFlow Serving, which is an open source API that provides production ready inference through HTTP. The applications that consume and produce data to the CAP middleware are written in python using open-source libraries. The services are wrapped in docker images (Docker is an open-source containerization platform).

The developed services will not be replicated as an OSS. The services will be deployed on CORE's hybrid cloud and will exchange information remotely with CAP middleware through secure communication protocols (HTTPS, SSL, TLS etc).

The components will not be replicated as an OSS. Therefore, additional changes of the components will not be OSS as well.

CAP1 – Planning and control of asphalt production

The CAP1 solution will be based on an algorithm programmed in an open source R programming environment and language. All the modifications and possible improvements made to CAP1 can be implemented in the same open source line.

The cognitive solution CAP1 is part of Scenario 1 described above (see Figure 37), i.e., it is based on solutions developed in an OS scenario. Its outputs are part of the data ingest system using ORION as context broker (so CAS2 is "powered with FIWARE Reference Architecture" (see Figure 38)). CAP1 will obtain the data needed by using the MySQL database as a data source to learn the context information of the asphalt plant.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

To summarize, the asphalt solutions are implemented partially using open source tools, so mostly falling under Scenario 2 described above (see Figure 37), but most are proprietary themselves. It's worth mention that on top of CAC1 and CAP1 some Open Source components for data ingestion are going to be implemented, thus addressing the Scenario 3.

3.2 Steel use case

CSS1 – Steel tracking sensor

CSS1 is a tracking solution based on hardware and software components. The hardware consists of lasers for marking of steel items and cameras for reading out the markers. It has been implemented as a proprietary solution strongly focussed on the SIDENOR.

The replication in open source components is not foreseen.

The results of the product tracking are the basis for the other cognitive sensors, in particular the digital twins (CSO1) and the risk and anomalies soft sensor (CSS5); the latter will be made available as open source.

CSS2 – Steel solidification sensor

CSS2 is a pure software solution that simulates the continuous casting process of steel billets, including the three-dimensional temperature field and the solidification front. It consists of a proprietary C++ library developed by BFI and a wrapper to the Python programming language.

The replication in open source components is not foreseen.

The results of the solidification sensor will feed the risk and anomalies soft sensor (CSS5), whose open source prototype will be made available.

CSS3 – Temperature soft sensor

CSS3 is a pure software solution that interpolates the existing measurements of product surface temperatures in the hot rolling mill to the cooling bed. A proprietary model for the calculation of the cooling curves has been adapted to the products considered in the project and is used for the calculation.

The replication in open source components is not foreseen.

The results of the temperature sensor will feed the risk and anomalies soft sensor (CSS5), whose open source prototype will be made available.

CSS4 – Scale soft sensor

CSS4 is a pure software solution that estimates the amount of secondary scale growing on the surface of steel bars in and after the hot rolling mill. It is based on a proprietary model that has been adapted to the steel grades considered in the project.

The replication in open source components is not foreseen.

The results of the scale sensor will feed the risk and anomalies soft sensor (CSS5), whose open source prototype will be made available.

CSS5 – Risk and anomalies sensor

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

CSS5 is a pure software solution that estimates the processing risk for individual semi-products after the casting and after the hot rolling. It is being implemented in Python, using the well-known Scikit-Learn and Tensorflow open source libraries.

In WP4 we aim to migrate the risk sensor to Apache Spark. At least a prototype will be published as open source.

CSO1 – Digital Twins

CSO1 is a pure software solution, providing Digital Twins for steel (semi-)products. They consist of a persistence layer, an application programming interface (API) and a web-based graphical user interface (GUI). The twins have been implemented using open source components MongoDB[10], NodeJS and Angular.

The Digital Twin solution will be adapted in WP4 for integration with CAPRI's cognitive automation platform (CAP). The exact form remains to be defined, but one aim is to remove the dedicated persistence layer and instead access the data stored in the CAP directly. The usage of open source components for the implementation will not be changed.

The risk and anomalies sensor (CSS5) retrieves its data from the digital twins, and in the future, it will also write back its results to the twins. We plan to publish (a prototype of) the risk sensor as open source.

To summarize, the steel solutions are mostly implemented using open source tools, so falling under Scenario 2 described above (see Figure 37), but most are proprietary themselves. We aim to develop a version of the risk and anomalies sensor, the central cognitive component of the steel use case, as an open source component, thus addressing Scenario 1.

3.3 Pharma use case

CPS1 – Sensor for blend uniformity

A Raman probe has been implemented at the outlet of the twin screw granulator. This probe reads spectrographic data and provides it via an OPC interface. The Python code used to evaluate these spectra will be made open source.

Open-source Python code will be easily understood and able to be used as a base for future enhancements by other actors in the pharma industry.

CPS2 – Sensor for granule quality

Particle size distribution (PSD) data is collected by a Parsum probe at the twin screw granulator outlet. Four moments are required to be calculated from the PSD data. The Python code developed to calculate these moments will be open sourced.

CPS3 – Sensor for product moisture

There are no open-source components planned for this cognitive solution. The model developed to estimate the granule moisture in the fluid bed dryer has been implemented in Matlab/Simulink. Due

[10] The MongoDB license is not generally considered open source, since it imposes certain restrictions on the user regarding the offering of database cloud services. It still shares many features with genuine open source licensed programs, such as availability of the source code and redistributability.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

to a lack of suitable open-source solution alternatives to Simulink, there will not be open-source solutions provided for CPS3.

CPS4 – Sensor for prediction of dissolution

It is planned that some aspects for the dissolution model, based on process settings, will be available as open-source code. The prediction engine is written in Python and will be made available as open-source component.

CPS5 – Sensor for fault detection

There are no open-source components planned for this cognitive solution. The fault detection module is based on D2Lab, which is a proprietary software.

CPC1 – Cognitive Control Concept

The CPC1 control concept, including the process models, has been implemented in Matlab/Simulink. A transfer to open source solutions, especially in terms of the Simulink blocks, is infeasible. Therefore, no open source components will be available from CPC1.

CPO1 – Cognitive Operation Concept

All Python code used to develop this operator solution will be made open source.

CPP1 – Cognitive Planning Concept

The planning solution CPP1 is formulated as a scheduling problem. The code is written in Python and will be made available as open source.

As a general consideration, the pharma industry has built many solutions over the years, based on proprietary hardware, software, and know-how. The CAPRI project offers the ideal opportunity to develop in-part open-source solutions to extend domain knowledge and collaboration opportunities. However, challenges remain, especially around data generated through the manufacture of specific product formulations, many of which fall under existing patent applications. To summarize, the pharma solutions are mostly implemented using open source tools, so falling under Scenario 2 described above (see Figure 37), but most are proprietary themselves.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

4 Data Pipeline and PI – COGNITWIN collaboration

4.1 Introduction

In this section an analysis of the possible collaboration with other SPIRE projects is provided, exploring the potential of the cognitive plants in the Process Industry.

Indeed, the **cognition** can be seen as a **generally applicable method for resolving unknown situations**, which are very dominant in the Process Industry (e.g. due to harsh conditions many previously unforeseen situations should be quickly detected and understood).

Due to the availability of relevant information, the focus is on the possible collaboration with COGNITWIN project[11], where both Nissatech and Sidenor are common partners.

The pipeline for hybridization was selected (creation of hybrid models), since it can be applied in common pilots, and it is a novel processing capability combining data-driven and numerical modelling.

The main argument is that in each of our pilots this method can be very useful.

In this section we provide the description of the element that enables the combination of data driven and numerical models, more specifically the integration of the numerical models into StreamPipes[12] pipelines.

4.2 Data Processor for Numerical models

This Data Processor is used to execute numerical model (physics-based model) developed for the Sidenor pilot, in COGNITWIN project. The goal is to provide additional, previously unknown, information regarding the heat process and ladle usage.

For the model to perform its calculations it needs both acyclic and cyclic data for the heat the calculation is done for. Therefore, the user has the option to connect and configure both Data Streams, one for acyclic data and another for cyclic data, which are going to feed the data to this Data Processor. In addition, the user has to specify the heat IDs for both data types, in order to ensure that the streams are synchronized (*i.e., make sure that the element performs calculations for acyclic and cyclic data that come from the same heat data*) (Figure 39).

[11] https://www.sintef.no/projectweb/cognitwin/
[12] https://streampipes.apache.org/

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

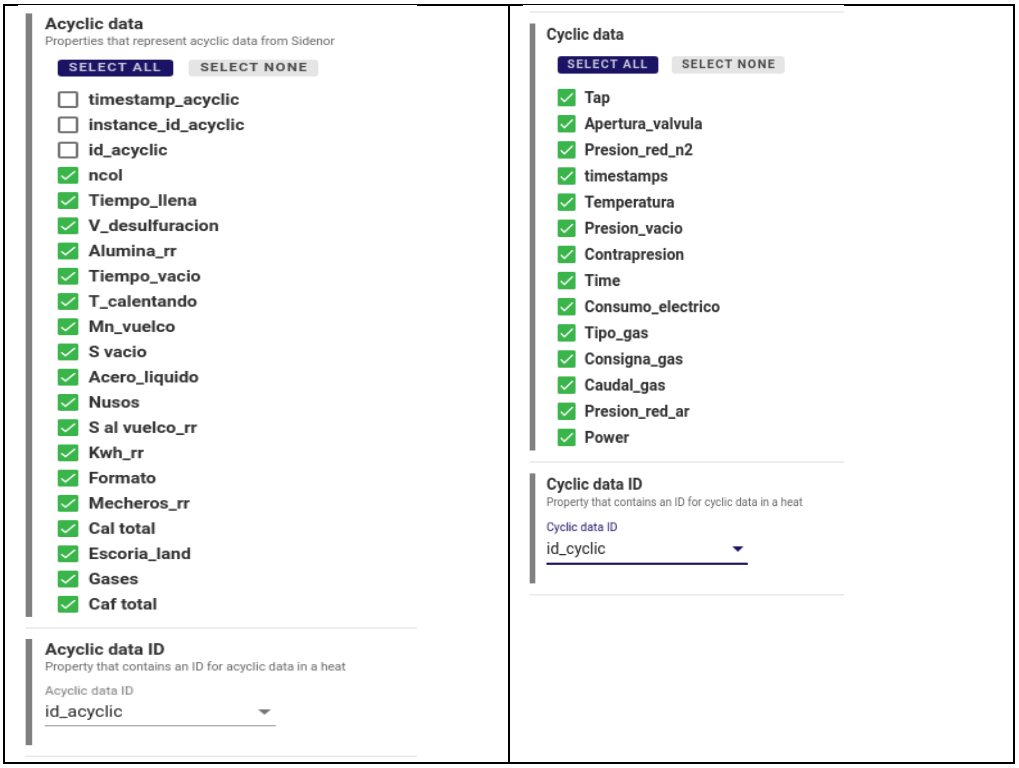

Figure 39: Configuration options of "Numerical Model"

This element requires "ncol", "Acero_liquido", "Cal total", "Alumina_rr" and "Caf total" acyclic properties and "Temperatura", "Time" (*offset in seconds from the initial timestamp*), "Power" (*differential of "Consumo_electrico"*), and "Consumo_electrico" cyclic properties. Other selected properties are being ignored by the model (*not used in calculations*).

Implementation

The numerical model is implemented in Python programming language, which means that this Data Processor cannot execute it on its own, since the StreamPipes Python wrapper is not yet released.

Therefore, we have implemented a REST service that is going to:

- Receive requests for the numerical model execution from this Data Processor. This request contains the necessary data for the calculations.
- Execute the numerical model.
- Respond to this Data Processor with the results of execution.

This means that this Data Processor sends POST request to the service (*2. request*) once it receives data from both Data Streams (*1. data*) and waits for its response with the results. When it receives response from the service (*3. response*), it creates an output event and forwards it to other pipeline elements connected to it (*4. result*). A diagram explaining this communication is shown in Figure 40.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

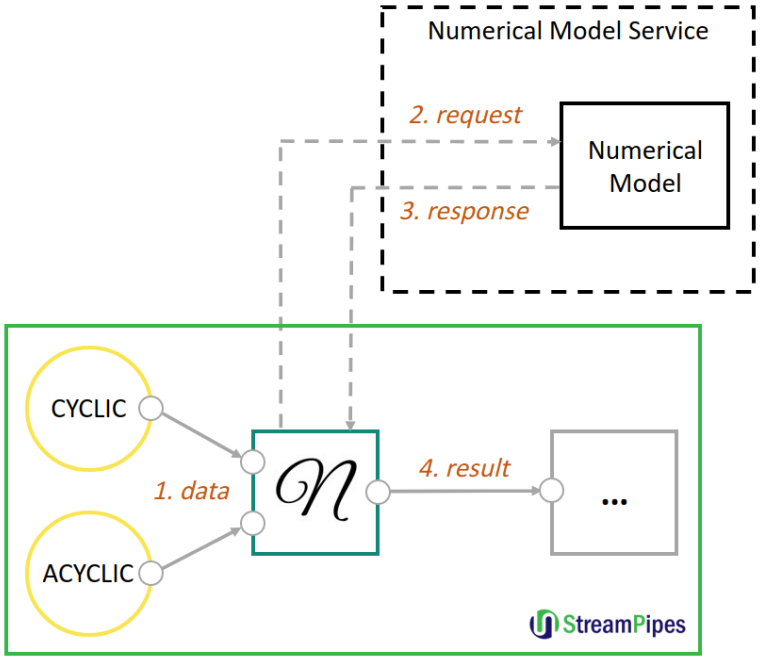

Figure 40: Integration of a numerical model in StreamPipes processing

4.3 Conclusion

Since the presented method is implemented using the StreamPipes open source framework, its integration in CAP is feasible. More precisely, StreamPipes can be seen as a general orchestration framework for an efficient development of complex processing pipelines.

The challenge is that the hybridization approach requires the existence of several models that should be combined (e.g. data-driven models and numerical models) in order to get more precise models. For example, numerical models can be used as simulation tools for the generation of data, which will be used in the data-driven modelling.

The tool wear/degradation scenarios were defined as very suitable for this method and one of the tasks for the future work will be to find such scenarios in existing pilots.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

5 Recommendations and Lessons Learnt for WP4 and WP5

As mentioned, the current document together with D3.2, D3.3, D3.4, D3.5 (all expected at month M24) establishes the formal end of WP3, even if the activities performed in this Work Package will be further enhanced in next months by other WPs. Following the bottom-up approach, WP3 represents the "bottom" (where the Cognitive Solutions are developed at laboratory level as standalone assets), while WP4 and WP5 represent the "up" (where the final platform is implemented and tested, respectively). Hence, it is evident that the achievements of WP3 are the starting point for the next Work Packages and that its lessons learnt must be taken into account for future purposes.

WP3 has collected the requirements for the definition of the CAP Reference Architecture, based on which the Cognitive Automation Platform will be implemented in WP4 and the CSs integrated on top. It is of the responsibility of WP4 to guarantee that the innovative aspects of the Architecture and of the cognitive solutions are correctly exploited, during the implementation phase.

This is the reason why this chapter is entirely dedicated to collect some insights of WP3, relevant for WP4 and WP5, and to gather their experience, identifying success stories and opportunities. Taking into account the implementation of Cognitive Solutions and the process that lead to the definition of the Reference Architecture for the CAP, a number of suggestions have been identified, useful not only for next WPs, but also for future projects dealing with the same topic.

The exercise has been run addressing directly the WP3 partners, involving both demonstrators and technology providers, that participated in the development of the CSs of the three domains, as well as in the definition of the requirements for the Reference Architecture.

In order to have a structured pictured of the information collected, we follow the approach of the SWOT analysis, suitably tailored for CAPRI purposes.

SWOT analysis

The acronym SWOT stands for Strengths, Weakness, Opportunities and Threats and the SWOT analysis is a framework used to evaluate the position of a company/project, in order to develop a strategic planning, taking into account both internal and external factors.

The analysis is based on a simple principle: in order to define a proper strategy, it is fundamental to take into account the following:

- What are the points of excellence, to be pursued and further exploited [**Strengths**]. This aspect is fundamental to make the product/solution attractive for external stakeholders.
- What are the points to be improved, in order to guarantee a high level product/solution [**Weaknesses**]. This aspect is fundamental to define the correct strategy since the weakness's points must be taken into account to be addressed (and hopefully mitigated) in the planning.
- What are the external factors, to be taken into account and from which the company/project may benefit [**Opportunities**]. Brainstorming about it may represent a good way to identify some options to be explored, that otherwise wouldn't be considered.
- What are the external factors, that may harm the company/project [**Threats**]. It is fundamental to identify all of them: even if the company/project can't act directly on them (since as external factors, they don't depend on it), it is possible to mitigate or to prevent them in case of risk.

In the WP3 context, it doesn't make sense to run the SWOT analysis in the "traditional" way, since we are not talking neither of a company nor of an entire project, but simply of a subset of activities useful to perform the next steps. Starting from the pillars on which the SWOT analysis is based, we fit them into CAPRI's purposes, identifying the following topics of discussion:

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

- **Innovative Aspects:** it represents the core of our analysis since it includes the innovative aspects that have been pursued in WP3 both at platform (CAP) and CSs level. The innovative aspects (to be concretised in WP4) are those that make the CAPRI solution unique and interesting for further exploitation and/or external stakeholders. Referring to the SWOT analysis aforementioned, it conceptually corresponds to the Strengths, that is, the points of excellence and the added value that we can provide thanks to the CAPRI solution.
- **Positive Outcomes:** it collects some unexpected events/outcomes that were useful for WP3 activities, since they provided added value to the final solution. Even if not originally foreseen, a "positive outcome" represents a feature to be stressed in terms of exploitation, but also to be taken into account for the development of future projects that might benefit from it, but as "planned" not as "unexpected". They are both points of Strengths for WP3 and Opportunities for WP4/WP5.
- **Occurred issues:** it includes a number of events that occurred during the two years of WP3 and impacted the planned development of the Cognitive Solutions, typically requiring more effort than expected. Even if at first glance, it may assume a negative meaning, this is not necessarily the case, since solving issues often implies acquiring experience and learning lessons for the future. It combines both internal and external factors: in the first case, when the issue is reliant on internal activities; in the other, when it is independent from CAPRI's team. Internal issues are comparable to points of Weakness that have been turned into Opportunities and lessons learnt for WP4 and WP5; external issues are mainly Threads and risks to be controlled.
- **Future Possible Issues:** it indicates possible risks that may occur, impacting the project's results. The risks analysis is part of the management process of any project and in CAPRI, specifically, is driven by WP1. Hence, the starting point has been the list of risks identified by the project coordinator in the Description of Action (DoA), with the objective of detailing for each domain how the specific risk could be shaped. The importance of such analysis is well known and it is fundamental to identify possible ways to mitigate/avoid them. It conceptually corresponds to Threats.

The next paragraphs provide an overview of the analysis run, highlighting the main lessons learnt and recommendations. Many of them are very generic and applicable to the entire project; they are those dealing with the CAP Reference Architecture or related to the management process, including of course the impact of Covid-19 pandemic. Others are more domain related and applicable only in a specific scenario, they are those dealing with the Cognitive Solution implementation/integration.

The key aspects better detailed in next paragraphs are summarised in the table below, to have at a glance the full picture of the analysis results. The arrows represent dependencies and in particular they are very useful to identify whether an issue has been turned into an opportunity.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Table 6 The CAPRI analysis of Innovative, Positive and Negative aspects

Innovative Aspects	Positive Outcomes and Opportunities	Occurred issues	Future Possible Issues
Matlab Compiler to run Matlab for Spark jobs	Agile approach in the implementation phase	Covid-19	Low accuracy in predictive models
Lego like approach	Flexibility in the implementation phase	Several delays (receiving goods, exchange info) / Difficulties in communicating / Initial lack of collaboration between WP3/WP4	Low performance for control and optimisation
Edge-Cloud computation platform	Accurate activities Planning	More effort to recover the delays	Inadequate user acceptance
Both Greenfield/Brownfield development allowed	The CAP layer structure (to integrate only needed features)		Higher implementation cost
[Asphalt] Real time accurate measurement of bitumen content in RAP	Testing of several models to identify the most suitable one	Lack of interaction with manufacturers/ operators	[Steel] CSS1 has low performance
[Steel] Digital Twin integrated in the CAP	Important to collaborate also with the manufacturer/operator		
[Pharma] Real time monitoring of the manufacturing line			

5.1 Innovative Aspects

Since the activities run in WP3 have the twofold objective of developing new Cognitive Solutions and of providing the requirements for the implementation of the CAP, to analyse the innovative aspects it's worth to distinguish between those related to the Platform and those that are strictly connected to the CSs and so, domain specific.

One of the most relevant and innovative aspect at Cognitive Automation Platform level is the **integration of the Matlab Compiler licence** in the CAP, that allows to run job using the Matlab for Spark. The need of exploring this option derives from the large use of Matlab code in the CAPRI Cognitive Solutions implementation, that drove towards the investigation of possible solutions to combine the Open Source principle on which the CAP is based with the benefits of using Matlab as a powerful working tool.

So, what was originally born as a mitigation feature, is now a **key component of the CAP that makes the platform very appealing also outside of CAPRI's borders**, since in many other scenarios and domains Matlab is one of the preferred tools to implement algorithms.

Many other features of the CAP are worth mentioning on this regard. For example, **the LEGO-like approach makes the architecture flexible and adaptable** to the specific needs of the various application domains in process industry. The CAP components in every layer can be combined

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

according with a Lego-like approach: it means that according to the use case and the specific company's specs, the components can be combined in order to provide only the required capabilities.

The Cognitive Automation Platform has been conceived as an **edge-cloud cognitive computing solution**, able to take advantage from both the approaches, according to the use case and to the company's needs: the choice turns to cloud computation in case of high performance requirements and big data (typically, cloud computing services provided by third-party are more efficient when compared to those installed on-premises); conversely, Edge Computing could be helpful in case of data generated locally, inside the manufactory company, requiring a real time processing.

Finally, the possibility to develop it both in a **greenfield and brownfield scenario** makes the CAP an extremely flexible platform that fits very well with the process industry requirements. Currently, although process industry is a very strong sector, it has scarcely approached innovations (plants use less advanced technologies and tools, workers and operators have poor digital background and skills), reason for which there is the urgent need for new solutions deployment (as it is the CAP) in the presence of the existing software to be integrated.

At the cognitive solution level, innovative aspects are identified at domain level.

Asphalt

The implementation of CAS1 – "Sensor for bitumen content" allows an almost **real-time accurate measurement of the bitumen** contained in the Reclaimed Asphalt Pavement (RAP), in the shape of an optical system. CAS1 is conceived as a very innovative solution since there are not any commercial products with comparable capabilities in the market. It has been designed to answer to a well-known issue in the asphalt sector, that is the current lengthy laboratorial process of measurement of bitumen content in RAP. The current measurement process is performed in an external laboratory, requiring human resource's time (around 5,5h for each sample) and the use of several equipment and solvents. Even more, the measurement of bitumen content is made using a small sample of RAP, not being representative of the amount of RAP added to the asphalt mix. All the more, this lengthy process does not provide timely data for the correction of the virgin bitumen content to the asphalt mix planning. CAS1 answers this issue by providing a solution that will be placed in the RAP line and will provide bitumen readings of all RAP added to the asphalt mix in almost real-time, allowing to correct the virgin bitumen content on the asphalt mix planning software and considerably lower the personnel time dedicated to such lengthy process, besides lowering the personnel's health risks by eliminating the usage of dangerous chemical products.

Steel

One of the main results of the Steel use case is the implementation of a **Digital Twin** to represent the entire process of billet casting and bar rolling as a solution **integrated into the CAP**. It means that the bar will be related to its particular history when it was part of a different product (either liquid steel or billet). This relationship is new as in the past the whole steel heat was considered in the analysis, and this meant that many of the measured variables were averaged over a great number of bars.

In order to integrate the Digital Twin into the CAP, some APIs will be soon defined as well as the data model specification. To properly finalise the task, the plan is to describe the solution according to a dedicated semantic model that makes the Digital Twin easily interoperable with the different steel case cognitive solutions (CSS).

Pharma

The implementation of CPC1 – "Cognitive Control Concept" allows the **real-time monitoring of the manufacturing line**, combining the output of some individual physical sensors (CPS1 – "Sensor for

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

blend uniformity", CPS2 – "Sensor for granule quality", CPS3 – "Sensor for product moisture", CPS4 – "Sensor for prediction of dissolution", CPS5 – "Sensor for fault detection"), implemented in CAPRI as well.

The innovative aspect of the proposed concept is the creation and (real-time) use of additional information on the material properties and the process state. By means of the cognitive sensor solutions, quality attributes of intermediates are available in real-time, avoiding their time consuming offline analysis in the laboratory. Further, the existence of this real-time information is used to actively adjust the process settings in real-time. By the incorporation of this real-time data in a quality control concept, the quality of the final product can be maintained while at the same time minimizing the waste material. This minimization of waste is achieved by two means: 1) The process control concept keeps critical quality attributes (captured by CPSx) close to their desired values by automatic adjustment of process settings in the manufacturing line. This approach aims at improving the robustness of the manufacturing process, ultimately reducing the effect of disturbances (e.g., raw material variations) on the quality attributes of the final product; and2) The quality control concept, also relying on information gathered by process sensors and CPSx solutions, discards material that does not conform to the specifications. The developed material tracking model is a valuable tool for identifying and discarding out-of-specification material more precisely when compared to conventional approaches.

The option of retrofitting existing manufacturing lines by the proposed solutions should be highlighted, too. For example, CPS3, the granule moisture prediction cognitive solution, does not need any modifications of the hardware on the manufacturing line, but it uses only available process data for predicting the granule moisture in the fluid bed dryer by means of a suitable dynamic model. This model offers valuable information on the granule properties without the need of installing additional hardware sensors.

5.2 Positive Outcomes

During the implementation of the 19 Cognitive Solutions and the collection of the CAP requirements, a number of positive results have been achieved, both from the technical and the methodological perspectives, besides the main goals already defined in the GANTT.

For instance, the definition of the Reference Architecture in WP3 brought out the complexity of the CAP development, which is based on different layers combining the needs of several use cases. The conclusion reached after the collection of the requirements is that the implementation of the CAP can't be solved in a single iteration: even if the four main tasks of WP4 (T4.1 – T4.4) overlap for most of the time, it is fundamental to proceed step-by-step, starting from the sensor layer, then the control layer, the operation and finally the planning, **adopting an Agile approach**. Such mindset fits very well with **flexibility**, which is another key aspect that came out during the WP3 implementation and proved to be extremely important for reaching the final scope. Developing features step-by-step allows to adjust the requirements defined at the beginning of the project to make them better fit with CAPRI'S purpose, that may change during months even because currently the project is in a research scenario.

The active collaboration between WP3 and WP4 has been particularly fruitful for the Cognitive Solutions' developers since it allowed to **enhance skills and competences in terms of Platform Architecture**, integration methods and user interface. The need of integrating the CSs on the CAP implied an important work on requirements collection and definition, with the contribution of all partners, who got in touch (sometimes for the first time) with "architectural concepts", making them aware of the process that brings to build a platform.

Very often, having at disposal a large amount of information from the pilot represents an obstacle for the technology provider, since it is not always easy to identify which information is relevant for the use case. Hence, it is fundamental to **identify an effective method to investigate the pilot** data and information.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Additionally, matching the technology provider competences with the pilot domain (for instance, choosing one who already worked in a similar use case and is aware of how the production process works in the specific domain) may speed up the activity's development, mainly considering the difficulties in communication due to Covid-19 pandemic's restrictions.

Asphalt

1) In the Asphalt use case, both the **technology providers** (CARTIF and AIMEN) and the **pilot** (EIFFAGE) **are located in the same country**, all Spanish companies. This match came out to be a great advantage from two different perspectives. First, having the same mother tongue speeds up communication and avoids misunderstandings; it's worth to consider that very often not all the workers employed in a company (SMEs but also large enterprises) speak English fluently and it could represent an obstacle to relate with the technology provider. Secondly, due to Covid-19 pandemic's restrictions among countries, travelling in the same country was much easier than going abroad and, apart from the initial strong limitations in 2020, people in the same country were allowed to visit the plant with no constraints.
2) The final version of CAS1 – "Sensor for bitumen content" will no longer be based on multispectral cameras (although its usage was essential for the basic research part of the development of this CS), as though at the beginning of the project. Instead, it will be based on a much simpler and cheaper optical system, which requires less time-consuming data processing. It represents an excellent example of how **testing several different tools/models** from the ones originally in mind can be helpful **to find the optimal results** (in this case, an unexpected optimal result).

Steel

1) As previously mentioned, from negative issues it is often possible to derive positive outcomes. In the case of the Steel plant revamping, the mitigation plan adopted (the restriction of the use case only to "one-billet orders") results in the definition of an easier scenario (that is, easier to be implemented) without losing the cognitive component. The simplification of the use case makes it more flexible and adaptable to other scenarios.

5.3 Occurred issues

One of the most relevant issues that occurred, since the beginning of the project (April 2020), is not surprisingly the **Covid-19 pandemic**, that strongly impacted CAPRI project (as well as any other activities).

In each of the three domains of WP3, the main consequence has been the impossibility for technology providers to go to the pilots' plants to analyse the scenarios and collect details. Even if, very often, remote meetings supported the exchange of information, it is not always possible to replace a physical inspection with a simple online conference, since not everything can be performed remotely. This caused some initial delays in starting the collaborative activities between technology providers and pilots, but since it happened at the very beginning, we are now successfully recovering from it.

Another relevant implication of the Covid-19 pandemic's restrictions was the initial difficulty in communicating among the entire consortium (mainly due to the unexpected need of reorganizing all project's meetings) that has made the strong relationship between WP3 and WP4 not so evident. Mainly from the pilot's perspective, it was not so clear that the development of the CSs at laboratory level was only a preliminary step toward the implementation of a common Cognitive Automation Platform on top of which the CSs could be integrated (following the bottom-up approach).

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

This initial uncertainty caused some delays in the integration activities and some re-work, but WP4 leaders did a great job in elucidating all partners of the process to be followed.

Considering that the Covid-19 pandemic was not something avoidable, the only mitigation action that could and can be performed is to **increase flexibility and the time spent communicating with other partners**, even if remotely. Having reduced the physical contacts with the other partners, clarity and precisions in all "virtual" communications has become fundamental, as well as an early planning, avoiding last minute choices and being sure that everyone is aware in advance of next steps.

At technical level, the initial separation between WP3 and WP4 caused the development of a number of CSs as proprietary or based on licensed software, without taking into account that it is not so straightforward to integrate proprietary solutions into the CAP. However, the **CAP structure based on layers** (to be fully or only partially integrated according to the pilot's needs) and some additional workaround in place allowed to overcome this issue and now, at least the most advanced CSs, are integrated. Additionally, also the use of different programming languages in the three sectors doesn't speed-up (on the contrary, it slows) the integration of the CSs in the CAP.

These apparent issues, that caused additional effort from both sides (pilots and technical providers), provide an added value for further exploitation, since it shows that, with ad-hoc customization, it is possible to integrate also proprietary solutions, developed with different languages (Python and Matlab, mainly).

Another relevant matter, to be taken as lesson learnt for WP4, is the lack of **involvement of the plant manufacturer/operator** in the CSs' development and integration. In WP3, mainly due to the limited possibility to communicate with people and organise meetings, only the user company owner has been involved, with positive outcomes in terms of business requirements, but causing lack of a more operative/technical attitude. To this regard, for the integration and validation activities (WP4 and WP5, respectively), provided that the Covid-19 pandemic's restrictions are definitely over, different professional figures of the plants will be taken into account.

Asphalt

In the Asphalt domain, unfortunately, some delays occurred due to a number of different reasons, mainly related to the effects of the Covid-19 pandemic. For instance, some goods and materials were delivered later than expected and the difficulties in physically reaching the plant also caused some problems in receiving data, as well as an initial mistake made in the data transferring process.

Another reason of delay typical for basic research developments, is the underestimation of the time required to properly acquire the competences and skills needed to get from basic research to physical sensor's development and prototype, as it happened for CAS1.

Steel

The main unexpected issues occurred in the Steel domain is the plant revamping, which caused many delays as it blocked the testing of CSS1 – "Sensor of product tracking" that has many dependencies with other Cognitive Solutions implemented in the same domain, since it feeds them with data. Due to the lack of a certain date for the end of the revamping, it took some months to define a mitigation plan, hoping to be able to recover the original situation as soon as possible. But finally, the **mitigation plan has been adopted,** restricting the use case to the "one-billet orders", in order to be able to collect the data required. The definition of the contingency plan is the result of a delicate analysis to find the balance between what was feasible and the minimum requirements to guarantee the cognitive components of the Steel Solutions.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

Pharma

1) The original planning underestimated the complexity of physically integrating both CPS1 – "Sensor for blend uniformity" and CPS2 – "Sensor for granule quality" at the same time. The physical space (at the twin screw granulator outlet) needed for mounting both hardware sensors in parallel was not sufficient. As so, during the implementation it came out that this approach was not feasible. The **successful workaround** that was adopted consists in the physical integration only of CPS1, while CPS2 has been replaced with a soft sensor based on process data. For operation scenarios that do not need CPS1 information, still the hardware sensor of CPS2 can be installed and used.
2) A second unexpected issue was the damage of CPS1 – "Sensor for blend uniformity". It went under repair but, due to Covid-19 pandemic's restrictions, it took more than expected. However, having **planned all the activities including possible contingencies**, allowed to mitigate the delay with no impact on the other CSs.

5.4 Future Possible Issues

As mentioned, the Future Possible Issues correspond to the risks described in the CAPRI DoA related to WP3, WP4 and WP5, to be discussed at single domain level but also to identify some recommendations/next actions for WP4.

The Future Issues are the following:

- **Low accuracy in predictive models**. This risk is particularly relevant in CSS5 – "Sensor for risks and anomalies" and the mitigation plan is based on the validation and testing of different models to find the most suitable for the use case. The model currently implemented consists of an autoencoder whose reconstruction error we aim to correlate with the occurrence of surface defects, but the approach could not be validated on the limited dataset available. As a next step, a more sophisticated analysis of the autoencoder latent space will be implemented. An additional attempt is the research of further data sources, if necessary, because it is unclear whether all process parameters relevant to the surface quality are really captured. A better understanding of the root causes of surface defects in the available dataset, or the provision of a new dataset with known issues, would also be useful for the algorithm development, which is another action we will pursue.
- **Low performance for control and optimisation.** To avoid it, it's important to act both on models (for instance, reducing complexity) and on the platform (for instance, increasing the computational power), in order to guarantee the right balance between algorithm and machine performance (model accuracy vs execution time). The recommendation is to keep the strong collaboration between "models developers" and "platform engineers", that turned out to be very fruitful in WP3, by organising periodical meetings to collect information and insights. Additionally, the integration process will be based on two different iterations: step one is to include the codes as black boxes into the platform, step two is to be refined them and try to optimise the performance in case it is needed.
- **Inadequate user acceptance.** Even if the risk is restricted to WP5 activities, it's worth to mention that one of the possible mitigation actions is to involve industrial key partners, experts, and end users in the development process, that is something that might start already in WP4. Actually, as highlighted in Sections 5.2 and 5.3, this is a task performed in WP3 and strongly recommended to be pursued also in WP4 and WP5.
- [Only Steel related] **CSS1 has low performance due to harsh conditions.** It is well-known that the issue occurred (due to the plant revamping), but an ad-hoc mitigation plan was put in place, by restricting the use case only to tracking "one-billet orders". The contingency plan, even if it simplifies the use case, is a good solution since it allows to get the most from what it is possible to have.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062

6 CONCLUSION

D3.6, despite being the conclusive deliverable of WP3, didn't provide a full picture of what has been implemented so far (for that, please refer to D3.2, D3.3, D3.4 and D3.5), but it paved the way for future activities to be developed within the end of the project, starting from WP3 achievements. The implementation of the 19 Cognitive Solutions at laboratory level are the basis for further investigation, raising the subject about several possible features to be included in the final solution, or at least to be explored: The **Openness of the CSs' components** that will be integrated in the CAP, the implementation of **"Fast Thinking" cognitive processes** leveraging on the use cases' raw data, the generation of **Open Dataset** from the pilots Datasets. In deeper detail:

- Regarding the **"Openness" of the Cognitive Solutions**, three scenarios have been identified: Fully OSS, Partially OSS, OSS Add-ons. So far, some of the CSs have been implemented as proprietary solutions (CAS1 and CSS1, mainly) and they won't be disclosed with external stakeholders; the large use of Matlab/Simulink (CPS3) and/or proprietary libraries/models (CSS2, CSS3 and CSS4) make difficult the implementation of a solution that can be Openly adopted outside CAPRI's borders. However, many other CSs have been implemented based on Open Source languages (mainly Python) and the code will be available as an open-source component (CAP1, CSS5, CPS1, CPS2, CPO1, CPP1,…). With the integration of the CSs into the CAP, these scenarios will be better outlined and boosted.
- About the implementation of a Cognitive System to simulate the "**Fast Thinking**" processes of humans, the three use cases have been explored and data has been collected from the pilots. A preliminary investigation of possible analysis that can be performed has been conducted, mainly at the sensor and control layers and it will be finalised in next months, with the objective of integrating the PICO platform in the CAPRI CAP and testing a general solution for Process Industry.
- The generation of **Open Dataset** has been explored, underlining its fundamental value but also putting in light the difficulties and barriers typically faced in its creation. Some pilot's datasets have been analysed to identify the most relevant ones for the purpose (for instance, the Steel dataset used in the development of "Fast Thinking" cognitive processes).

Previous achievements have been selected as features to be enhanced to make the CAPRI solution appealing and exploitable outside the project's borders, including research works with the most recent advancements. This approach boosts the Cognitive solutions and processes developed so far, making them interesting also for external stakeholders. However, also the opposite direction has been explored. The objective of the analysis presented in the deliverable about possible collaborations with SPIRE-06 cluster is to explore the potential of other SPIRE projects (COGNITWIN mainly), which are developing solutions that could be applied also in CAPRI's use cases and to be integrated in the CAP. This investigation is still at a preliminary stage and will be intensified in the next months, taking into account the progresses of the other projects.

This project receives funding in the European Commission's Horizon 2020 Research Programme under Grant Agreement Number 870062