Ref. Ares(2023)4562768 - 30/06/2023

ExtremeXP

Experiment Driven and user Experience Oriented Analytics for Extremely Precise Outcomes and Decisions

D6.1: Use case requirements

Project Acronym/Title	Experiment Driven and user Experience Oriented Analytics for Extremely Precise Outcomes and Decisions
Grant Agreement No.:	101093164
Call	HORIZON-CL4-2022-DATA-01
Project duration	36 months | 01 January 2023 – 31 December 2025
Deliverable title	Use case requirements
Deliverable reference	D6.1
Version	1.0
WP	6
Delivery Date	30/06/2023
Dissemination level	PU - Public
Deliverable lead	I2CAT
Authors	Maxime Compastié (I2CAT), Borja Tornos (IDEKO), Laurent Drouglazet (ADS), Kevin Larnier (CS), Sonu Preetam (I2CAT), Konstantinos Mavromatis (MOBY), Giorgos Giannopoulos (ARC), Eleni Zarogianni (ICOM), Vasileios Theodorou (ICOM), Thanassis Mantes (ARC), George Papastefanatos (ARC)
Reviewers	Amela Karahasanovic (SINTEF), Ilias Gerostathopoulos (VUA)
Abstract	To demonstrate the advancements of the experiment-driven analytics paradigm, the ExtremeXP project features five different use cases tackling five different application domains: (i) crisis-management, (ii) cybersecurity, (iii) public safety, (iv) transportation and (v) manufacturing. This deliverable elicits, analyses, and consolidates the different requirements framing the preparation of use case pilots and defines their technical design. For each use case, this deliverable will cover aspects pertaining to the selection of adequate datasets, domain modelling, variability points identification, the specification of experiment models, the elicitation of user intents, and the determination of the technical settings for evaluation.
Keywords	Requirements elicitations, use case, pilot architecture, validation environment, experiment modelling.

Dissemination Level	
PU	Public, fully open
Type	
R	Document, report (excluding the periodic and final reports)

Co-funded by the European Union

Version History

Version	Date	Owner	Author(s)	Changes to previous version
0.1	2023-04-11	I2CAT	Maxime Compastié, i2CAT	Outline
0.2	2023-05-09	I2CAT	Maxime Compastié, I2CAT	Abstract
0.3	2023-05-15	IDEKO	Borja Tornos, IDEKO	Description of Use case 5 context in Section III
0.4	2023-05-17	ADS	Laurent Drouglazet, ADS	Description of Use case 3 context in Section III
0.5	2023-05-18	CS	Kevin Larnier, CS	Description of Use case 1 context in Section III
0.6	2023-05-18	I2CAT	Sonu Preetam, i2CAT	Description of Use case 2 context in Section III
0.7	2023-05-19	MOBY	Konstantinos Mavromatis, MOBY	Description of Use case 4 context in Section III
0.8	2023-05-22	I2CAT	Maxime Compastié, I2CAT	Executive summary
0.9	2023-06-15	ICOM	Eleni Zarogianni, ICOM	Review UC contributions
0.10	2023-06-19	I2CAT	Maxime Compastié, i2CAT	Update and closure Section 2
0.11	2023-06-25	I2CAT	Maxime Compastié, i2CAT	Update and closure of section 3,4,5; for review
0.99	2023-06-29	ALL	All Contributors	Application of the remarks from reviewers
1.0	2023-06-29	ARC	Giorgos Giannopoulos, George Papastefanatos, ARC	Final version for submission

Table of content

List of Figures

Co-funded by
the European Union

List of Tables

Co-funded by
the European Union

List of Acronyms

Abbreviation	Meaning
ABAC	Attribute-based Access Control
AI	Artificial Intelligence
AMR WB	Adaptive Multi-Rate Wideband
AOI	Area of Interest
API	Application Programming Interface
AR	Augmented Reality
ASCII	American Standard Code for Information Interchange
AWS	Amazon Web Services
BCs	Boundary Conditions
CAP	Complex analytics process
CNC	Computer Numerical Control
CPU	Central Processing Unit
CSV	Comma Separated Values
DBSCAN	Density-based spatial clustering of applications with noise
DEM	Digital Elevation Models
DSM	Digital Surface Model
DoW	Description of Work
DSM	Digital Surface Model
ELK	Elasticsearch, Logstash, and Kibana
ESRI	Environmental Systems Research Institute
ETL	Extract-Transform-Load
FHD	Full High Definition
GIS	Geographic Information System
GPU	Graphical Processing Unit
GTFS	General Transit Feed Specification
HR	High Resolution
IDS	Intrusion Detection Systems
IOC	Indicators of compromise
IP	Internet Protocol address
JSON	JavaScript Object Notation
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GPU	Graphics processing unit
KPI	Key Performance Indicator
LAN	Local Area Network
LAS	Laser
LiDAR HD	Light Detection and Ranging Hight Dimensional data
ML	Machine Learning
MSSP	Managed Security service Provider

POI	Point-of-interest
PPDR	Public Protection and Disaster Relief
RAM	Random Access Memory,
RTMP	Real-Time Messaging Protocol
RTP	Real-time Transport Protocol
RTSP	Real-Time Transmission Protocol
SD	Standard Definition
SIEM	Security Information and Events Manager
SOC	Security Operations Centre
TCP	Transmission Control Protocol
TIFF	Tagged Image File Format
TRL	Technology Readiness Levels
TTP	Tactics, Techniques and Procedures
UC	Use Case
VPN	Virtual Private Network
VTK	Visualization Toolkit
WP	Work package

Co-funded by
the European Union

Executive Summary

This deliverable reports on the activities of Task 6.1 “Requirements of use cases: intents, data analytics workflows, data sources, datasets", which is part of the WP6 “Use Cases and Validation”. The key objective of this deliverable is to define and analyse the requirements spanning from the ExtremeXP Use Cases (UC) and define how these will impact the conception and development of the ExtremeXP framework.

The five UCs identified within the ExtremeXP project are the following:

1. Improvement of flash flood forecasting thanks to the use of AI
2. Increased Cybersecurity situation awareness for efficient threat mitigation
3. Situational intelligence and decision making for PPDR
4. Flexible transportation analysis and visualization
5. Failure prevention for manufacturing industry

Based on the UC descriptions provided by the UC owners from the ExtremeXP consortium, we elicit the requirements framing the technical design of the different pilots in charge of validating ExtremeXP advancements and interpret them as requirements for the design of the ExtremeXP framework. Specifically, each UC is analysed to identify its overall goals and specify the related experiment, including the datasets utilised, the variability points of its experiments, the user objectives, technical settings, as well as key performance indicators (KPIs) that are to be achieved.

This deliverable provides the basis for the design of several components of the framework that is contemplated in WP2 “*Conceptual foundations for experiment-driven analytics over extreme data*” and provides the input to the design of the UC pilots in task T6.2 “Implementation of UC pilots”. Finally, this deliverable also serves as a baseline for the requirements of the validation of the ExtremeXP framework in task T6.4 “*Validation of ExtremeXP framework in the UCs*”.

1 Introduction

1.1 Context and goals

Deliverable 6.1 “UC requirements” analyses the ExtremeXP UCs with the scope of eliciting their overall goals and requirements, contextualizing, and informing the preparation of the ExtremeXP framework, and providing a starting point for the design of the technical demonstrators serving evaluation purposes in the project. The deliverable investigates the requirements stemming from five realistic industrial UCs in five different verticals: natural disaster prevention, cybersecurity, public safety and crisis, transportation, and failure prevention in manufacturing.

ExtremeXP introduces the concept of **experiment-driven analytics**, by placing the end user at the centre of complex analytics processes (CAPs). In particular, it proposes a human-in-the-loop experimentation approach, where experimental data and outcomes are presented, visualized, and explained to end-users allowing, at the same time, **user feedback** to be taken into account in the form of preferences and constraints (e.g., in performance expectations, resource usage, processing time and options for model selection). This facilitates the execution of complex analytics processes in decision support systems, where the system’s output gradually improves, meeting both system-level quality metrics (e.g., latency, accuracy, precision, specificity, anonymity) and task-and-user-related metrics (e.g., usefulness, correctness).

Research in ExtremeXP spans from modelling complex analytics (Objective 1) to automated data management (Objective 2), scenario- and constraint-driven machine learning (Objective 3), experiment-driven optimization (Objective 4), visualization and explainability methods (Objective 5), and data access and knowledge management (Objective 6). All these ambitious research threads come together in the development of the **ExtremeXP framework**, along with a collection of tools, methods, models, and software for complex experiment-driven analytics (Objective 7).

The framework is motivated by and will be validated on the ExtremeXP UCs described in this deliverable. In particular, the design of the framework will be directly informed by the UC requirements related to experimentation parameters, user involvement, metrics, deployment settings and overall constraints; and the expected benefits brought by the application of the framework (such as improved accuracy, precision, fit-for-purposeness, trustworthiness, etc.) will be demonstrated by applying the framework on realistic scenarios from the five UCs of ExtremeXP.

Given the overall concept and objectives of ExtremeXP overviewed above, the **objectives** of this deliverable are the following:

- Detail the different UCs, and identify technical barriers that can be overcome with the experiment-driven analytics;
- Prescribe the experiments, the support infrastructure, and the properties of extreme data to be involved in the different UCs in order to guide the design of the ExtremeXP framework;
- Elicit and describe the functional requirements, both related to the demonstrator of each UC and to the different components of the ExtremeXP framework.

1.2 Target audience

The primary audience of this document consists of members of the ExtremeXP consortium, who will participate in the design and development of the ExtremeXP UC pilots, as well as the design and implementation of several components comprising the experiment-driven analytics framework. Additionally, this document might be of wider interest to extended communities of academics and practitioners in data science, as well as to representatives of the several verticals covered by the five different UCs, i.e., public safety and crisis management, transportation, manufacturing, and cybersecurity, interested in the application of complex analytics processes in these domains.

1.3 Relation to other work in the project

Results from D6.1 will contribute to future work in the project as follows:

- The UC requirements will be exploited towards the development of the modelling languages and the ExtremeXP reference architecture to be reported in D2.1 "*Initial architecture, languages and models for complex experiment-driven analytics*" and D2.2 "*Final architecture, traceability, and trustworthiness for complex experiment-driven analytics*".
- The individual requirements stemming from the dataset usage and acquisition methodology, the data analytics pipelines and the experiment definitions will inform and shape the work in the technical tasks and work packages of the project. In particular, they will influence the work to be reported in deliverable D3.1 "*Data selection, integration, and simulation services*", D3.2 "*Constraint-aware ML and analytics services*", D4.1 "*Visualization and explainability services*", and D4.2 "*Human in the loop services – User intents, profiles, feedback*" by impacting the design of the prepared components.
- The requirements issued from the definition of the runtime environment from the different UCs will inform and constrain the implementation of ExtremeXP core services, detailed in D5.1 "*Core framework services – Monitoring, Planning, Scheduling*" and D5.2 "*Core framework services - Data and knowledge management*", and will have a more direct impact on the specification the pilots' testbed environments used in the validation presented in D6.3 "*Pilot demonstrators and validation report*".

1.4 Structure of the document

This deliverable is structured as follows.

- Section 2 provides a review of ExtremeXP concepts and details the methodology to obtain the UC requirements.
- Section 3 thoroughly describes the five different UCs of the project to identify requirements framing them and issues a set of criteria and technical requirements for the design of the UC demonstrators.
- Section 4 capitalises on the analysis of Section 3 to include several functional requirements over the ExtremeXP framework, covering the specificities of the several UCs.
- Finally, Section 5 summarises the content of this document and identifies some of the next steps.

2 Requirement elicitation methodology

In this section, we detail the methodology adopted to elicit the requirements from the UCs that will inform the design of both the UC pilot demonstrators and the ExtremeXP framework architecture. In the first subsection, we describe the general objectives of the project and its ancillary notions to pinpoint under which aspects the UCs should be scrutinised. We then detail the methodology leveraged for requirements extraction at the level of each UC.

2.1 ExtremeXP Concepts and Objectives

ExtremeXP endeavours to foster the generation of accurate, precise, fit-for-purpose and trustworthy data-driven insights, by proposing and instrumenting a novel paradigm called experiment-driven analytics. As defined in ExtremeXP's description of action,

"ExtremeXP proposes a new paradigm for data analytics, which we call experimentation-driven analytics. The main contribution is that it puts the end user, i.e., requirements, preferences, constraints, interpretation, explanations, feedback, and decision-making, at the centre of complex analytics processes (from data discovery to novel interactions), proposing a human-in-the-loop, associated experimentation approach for gaining knowledge and making decisions from data with varying and extreme characteristics […]."

To that extent, the notion of *experiment* occupies a central role in the framework design: it consists of the introduction of variants in several steps in the acquisition and integration of datasets, in the execution of processing and in the visualisation and explanation of the obtained results. The instantiation of variants is elicited from users' intent and controlled via interactive user feedback as illustrated in Figure 2.1-1.

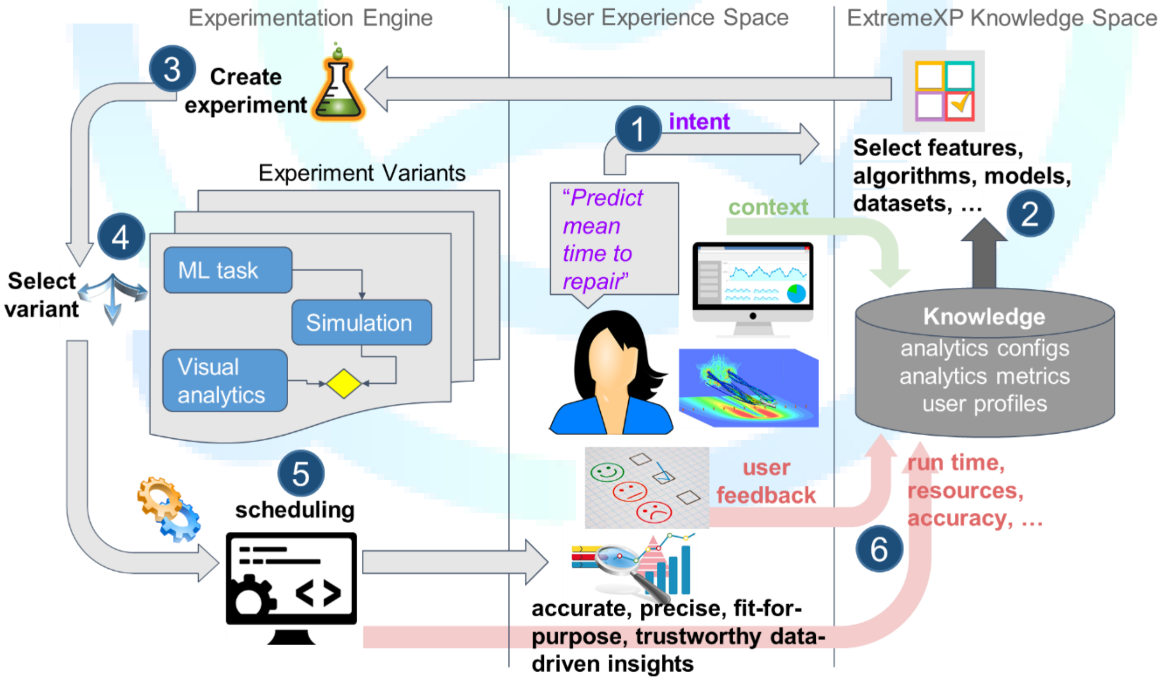


Figure 2.1-1: Experiment workflow in ExtremeXP framework

To implement this vision, ExtremeXP identifies five core research & development directions stated as follows:

- **Analysis-aware Data integration:** Investigate and implement the methodologies to select adequate datasets, cleanse and transform them to mitigate their quality issues in experiment workflows
- **User-driven AutoML:** Research and design tools to instrument simulations to augment data involvement in ML pipeline, design ML algorithms supporting the selection of models based on constraints expressed by the user, to permit the continual learning of the model selection strategy and to optimally deploy ML-pipelines.
- **Transparent & Interactive Decision Making:** Provide a visualization and explainability layer over the ExtremeXP framework and contemplate the usage of Augmented Reality therein.
- **Extreme Data & Knowledge Management:** Control and supervise the lifecycle of the knowledge assets obtained from the experiments and control the acquisition of the involved datasets.
- **User-driven Optimisation of Complex Analytics:** Explore different ways that the user's intents, preferences, and constraints are captured and used in the optimization of complex analytics, including gamification methods.

The several advancements are catered around seven objectives. We review each of them to understand under which aspects each UC should be specified.

Objective 1: Specification and semantics for modelling complex user-experience-driven analytics

This objective will deliver a modelling framework associated to a reference architecture to conduct experiment-driven analytics. The modelling aspects cover the variability in complex analytics, the data serving the evaluation of system properties and user feedback, user requirements and preferences, trade-offs between quality of service and experience metrics.

From a UC perspective, this suggests we should analyse them to extract the different steps (i) constituting the experiments and the data processing, and (ii) the definition of the several parameters inducing variants of processing tasks and model parameterisation. The existence of any pre-existing data-processing workflows should be indicated to position with more precision the expected value from the ExtremeXP project.

Objective 2: Automated and scalable data management for complex analytics workflows

The second objective will investigate and develop the tools, concepts, and methods for analysis-aware data integration. This process accounts for data obtained from existing heterogenous data sources with varying quality, but also generated datasets from simulations.

Consequently, the project should obtain the description of the data to be acquired and the specification of the datasets to assess their properties. These descriptions would assist in comprehending how to evaluate the quality of the datasets and how to augment them if required (e.g., via simulation). Finally, estimating the data velocity would serve to estimate how data processing workflows should be parallelised and scaled up.

Objective 3: Scenario-driven and opportunistic machine learning (ML)

This objective contributes toward the conception of methods for automated ML (AutoML) with user constraints. In detail, this objective contemplates the design and development of AutoML mechanisms performing the scenario-based algorithms and model selection through user-issued constraints, the elaboration of a continuous learning for model selection, and the performance improvement of models though automated feature engineering and data augmentation.

To materialise this objective, for each UC, the different processing of the data should be disclosed so that the research partners (i) identify and model the user constraints related to data processing, (ii) understand for which processing task AutoML should be implemented and (iii) understand how feature augmentation and data augmentation should be introduced in the data processing pipelines.

Objective 4: User-experience and experiment-driven optimization of complex analytics

The fourth objective covers the preparation of the experimentation engine and the support for user involvement in experiment-driven analytics. To meet this end, the objective prescribes the design of the ExtremeXP framework architecture, the development of the experimentation engine and the profiling of the framework user and context.

Consequently, the studied UCs should disclose the different steps of the data processing to identify the several workflow variants that will be executed by the experimentation engine and to profile the several classes of actors interacting with the framework.

Objective 5: Transparent decision making with interactive visualisation methods

The fifth objective tackles the preparation of the explainability-oriented user interaction toolset and the interactive visualisation support. To that extent, this objective (i) explores the visualisation techniques including augmented reality and visual analytics; (ii) prepares user interface guidelines; (iii) develops the interactive and progressive visual analytics adapted to extreme data analytics; and (iv) develops the explanation methods for the decision-making process of the user.

The steps composing the experiment workflows, including the visualisation modalities and the expected interactions for decision-making, should thus be specified for each UC to elicit requirements on visualisation techniques to implement, justify user interface guidelines, and frame the need for explainability of the decision-making process.

Objective 6: Extreme data access control and knowledge management

The sixth objective focuses on the privacy and security aspects to be included in the holistic data and knowledge management. More specifically, this work in this objective aims to (i) adapt the ABAC model to the peculiarity of analytics on extreme data, (ii) develop distributed ledger and smart contract-based methods to ensure the non-reputability of granted access requests and (iii) improve the provenance aspect of the knowledge and data involved in ExtremeXP experiments.

To that end, the UCs should detail which resources (e.g., knowledge, datasets) of the experiments should have their access controlled, and under which modalities, and also provide information/ways so that their provenance can be verified.

Objective 7: Test and validation framework and application on different impactful real-life UCs

The last objective oversees developing the five pilot demonstrators of the project via their deployment in their relevant environments to validate ExtremeXP's contributions, including the modelling framework and reference architecture. This comprises the integration of the several components composing the ExtremeXP framework, their applications in the environment of the five UCs for quantitative validation of the obtained benefits, and the transfer of reusable knowledge assets.

Therefore, all UC descriptions should make the expected benefits introduced by the ExtremeXP approach explicit in the UC verticals. They should also provide a set of KPI to quantify the benefits, and the settings of the evaluation environments: the specification of the current testbeds and the presence of an existing set of frameworks already supporting the execution of data analytics jobs, if applicable.

2.2 Methodology for UC analysis

The identification of requirements was based on several steps, as illustrated in Figure 2.2-1 and detailed below.

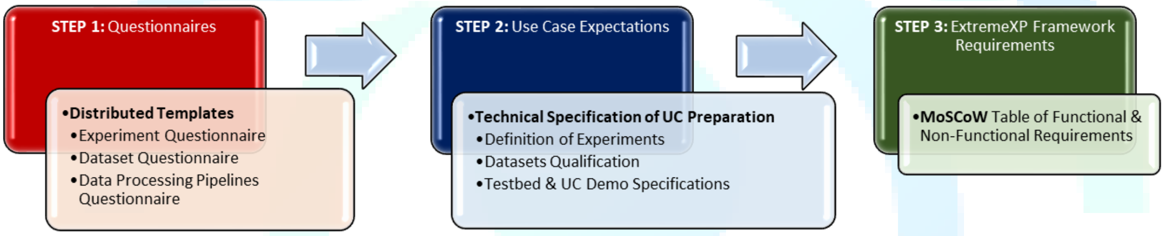

STEP 1: Questionnaires
•Distributed Templates
•Experiment Questionnaire
•Dataset Questionnaire
•Data Processing Pipelines Questionnaire
STEP 2: Use Case Expectations
•Technical Specification of UC Preparation
•Definition of Experiments
•Datasets Qualification
•Testbed & UC Demo Specifications
STEP 3: ExtremeXP Framework Requirements
•MoSCoW Table of Functional & Non-Functional Requirements

Figure 2.2-1: Steps of UC analysis methodology

As a first step, WP and task leaders have jointly prepared templates, including questions pertaining to the needs and research areas of interest of each specific WP.

- A first experiment questionnaire tackled several topics involved in the preparation of the experiments and the deployment, the modelling of the data analytics workflow, the definition of class of users and the access control to the data.
- A second dataset questionnaire focused on the specification of the dataset, the modalities of data integration, the feature augmentation, and the current involvement of ML algorithms.
- A last integration questionnaire focused on the existing data processing pipelines to identify integration challenges.

Those questionnaires were submitted to the representatives of the UC providers for completion.

The information gathered has served the planning of technical activities by revealing several technical aspects pertinent to the design, experimentation, and evaluation settings.

The recipients of these questionnaires were, in priority, the organisation owning the UC of the project. This includes the CS for the flash flood forecasting UCs (UC1), I2CAT and ITU for the cybersecurity UC (UC2), ADS for situational intelligence UC (UC3), MOBY for the transportation UC (UC4) and IDEKO for the failure prevention UC (UC5). Technical experts from the ExtremeXP consortium were also invited to share their relevant experience from involvement in UCs from other research projects they have participated in, with the incentive to exemplify aspects to be investigated according to their technical contributions in the project. Each questionnaire was submitted by WP leaders for completion during a two-week period and were iterated based on feedback for another week. The leader of task T6.1 iterated on the answers of these questionnaires to obtain additional information as per the technical specification of the UC preparation, especially as per the definition of the experiments, the qualification of the datasets, the specification of the testbed and the technical specification of each demonstrator.

Building on these initial questionnaires, the UC partners were asked to provide detailed, structured descriptions of their use cases, with the aim to further facilitate the requirements elicitation process and clarify the objectives of the UC demonstrators, and their expectations over the project. These descriptions are presented in Section 3.

The requirements, elicited from the aforementioned material, were interpreted as functional requirements for different research directions, proposed by ExtremeXP framework and previously described. Namely:

- Analysis-aware Data integration,
- User-driven AutoML,
- Transparent & Interactive Decision Making,
- Extreme Data & Knowledge Management,
- User-driven Optimisation of Complex Analytics.

The above features will serve as taxonomy to classify the requirements of the framework in Section 4.

The different requirements are collected in tables and prioritised according to the MoSCoW Model [1]. This model relies on the following 4 levels of priority:

- **MUST** have**:** These requirements are deemed necessary to the achievement the ExtremeXP Framework objectives. They must be designed, implemented, and verified to meet the project objective.
- **SHOULD** have**:** The requirements labelled in this category are important to meet the project’s end, but not strictly mandatory. Their implementation as part of the framework is recommended.
- **COULD** have**:** These requirements contribute to the project’s objective and are desirable, but not mandatory. They can be considered for the implementation with an extra effort.
- **WON'T** have**:** These requirements have been concurred to be least critical for the project execution. They can be deemed to be safely ignored during the timeframe of the project execution and potentially considered for future activities.

The several requirements introduced by this document reflect the current vision of the UC demonstrators and the ExtremeXP framework at the time of edition of this deliverable. These might be subject to minor re-evaluation, as the development of scientific and technical enablers of the ExtremeXP framework might provide new opportunities to the UCs providers.

3 UC descriptions

This section details the description of the five UCs, which will serve as a basis for the evaluation of the ExtremeXP framework. Each UC is described and analysed to issue a set of functional and technical requirements that will frame the preparation of the demonstrators. The stated information has been obtained following the methodology described in Section 2.2.

Each subsection of this section focuses on one UC, and follows a similar structure, as described below:

- **Subsection 3.X.1** provides the main incentives and technical barriers of the verticals considered by the UC and known architectural considerations.
- **Subsection 3.X.2** elaborates on the *experiment* related to the UC. An experiment is comprised of: (i) different actors (ii) a number of variability points differentiating the execution of data processing pipelines, (iii) the related *experiment model*, (iv) the types of *user intents* expected to be expressed during the experiment definition and the requirements related to visualisation.
- **Subsection 3.X.3** covers the specification of datasets to be utilized within the context of the UC and, if found adequate, the methodology to acquire them. This subsection also tackles the possible restriction to the access of the datasets.
- **Subsection 3.X.4** technically describes the evaluation environments that will serve for the deployment and the evaluation of the demonstrator.
- **Subsection 3.X.5** consolidates the expected benefits to be introduced by ExtremeXP in the UC usages.
- **Subsection 3.X.6** finally lists a set of technical requirements, known at the time of writing of this deliverable, which will frame the design of the UC demonstrator.

3.1 Definitions

The following terms are important as background for this deliverable but also for the project as a whole.

1. **Complex analytics process** (CAP): A CAP in ExtremeXP is a general concept capturing the process perspective of analysis that extreme data is undergone, from the discovery of datasets, to the interaction of users with analytics results. A CAP is a composition of data processing tasks, which can take the form of machine learning (ML) training and model serving pipelines, simulations, data analytics, and extract-transform-load (ETL) pipelines, to name a few concrete examples. Crucially, we assume that (i) a CAP contains variability points (i.e., different options regarding its configurations, algorithms, or even input data and sub-processes), (ii) these variability points can be set at runtime by selecting a concrete variant for each variability point, and (iii) selecting variants leads to different values on quality metrics (such as latency, accuracy, precision, specificity, anonymity, as well as user acceptance and usefulness).
2. **ExtremeXP experiment**: An experiment in ExtremeXP consists of several trials. In each trial, variants are selected for each variability point in a CAP and the quality metric results are computed. The overall goal of an experiment is to identify the dependency between variants and the quality metrics, so that the execution of a CAP can be optimized.
3. **User intent**: A user intent expresses the objective of users in interaction with ExtremeXP framework, from the perspective of their own business context. From the perspective of the framework, these objectives translate as knowledge requirements to be satisfied with properly defined ExtremeXP experiments.

3.2 UC1: Improvement of flash flood forecasting thanks to the use of AI (CS)

3.2.1 Context description and objectives

A flash flood is a rapid flooding, most often caused by extremely heavy rainfall, often resulting in significant destructions of goods or even fatalities. Climate change and the growing urbanization and land artificialization have significantly contributed to increased number of flash floods in urban areas.

Hydrological models are currently utilized by flooding alert services to simulate the evolution of water flows, fluxes, and water storages over time. These models require, however, vast amounts of data, both in terms of volume and diversity, making it difficult for in-situ infrastructures to generate and/or collect them.

In this regard, AI-related methodologies could help alleviate issues pertaining to the integration and cross-analyses of several data sources, and also significantly improve prediction rates for flash flood-related scenarios.

This UC aims to improve the accuracy of flash flood forecasting, either through the development of an AI model that will hybridise the physical model to reduce its need of input data (in volume, diversity, and resolution) or AI methods to facilitate the definition of the inputs of the hydrodynamic model. For instance, the following methods can be considered:

- Computation of the building footprints from a DSM (Digital Surface Model) using an AI model.
- Computation of the hydraulic structures (weirs, levees, etc.) footprints from a DSM.
- Upscaling a coarse DEM (Digital Elevation Model) to obtain a HR (High Resolution) DEM.

In the context of ExtremeXP, this UC will seek to implement a **CAP** featuring **accuracy**, **scalability**, and **precision**. This UC will include a mixture of steps and methodologies in data preparation, data cleansing and physics-based analytics. One critical aspect also relates to the **dedicated visualisation insights,** as flash flood predictions are useful for numerous stakeholders, each of them having their own objectives, competences, and background. Therefore, the visualisation of results must be personalised and interactive.

3.2.2 Experiment definition

The experiments will be split into two phases.

The first phase consists of designing, calibrating, and validating the 2D hydrodynamic (physical) model and will serve either as input to or in the validation of our ML experiments. This phase will be conducted offline (with respect to the ExtremeXP framework) on the METIS platform of CS GROUP[1]. This phase requires a lot of work to transform and adapt the input data such as data pertaining to topography, buildings and streets footprint, land cover etc.) that will be performed using mainly GIS (Geographical Information System) software and Python libraries. Then, the transformed input datasets will be provided to the ExtremeXP project.

The hydrologic workflow for this setup is depicted in Figure 3.2.2-1.

[1] https://www.csgroup.eu/en/offerings-solutions/data-intelligence/metis/

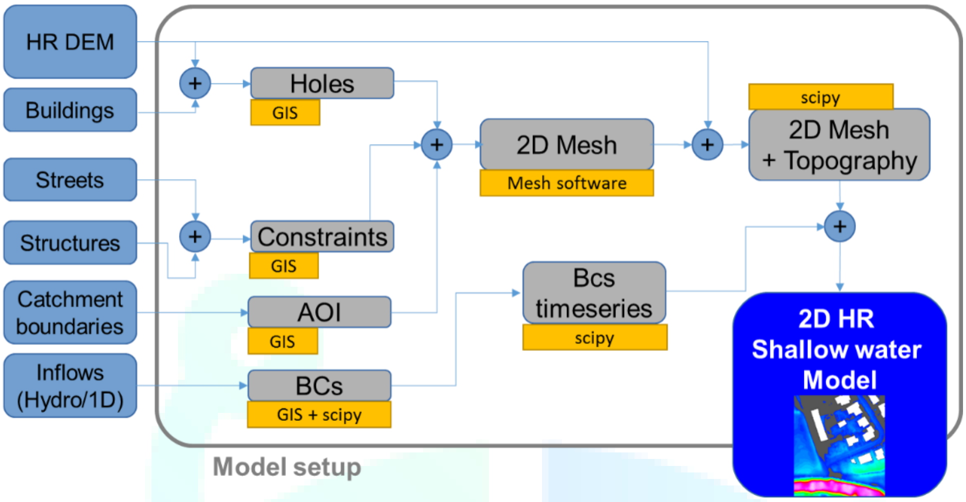


Figure 3.2.2-1: Workflow of the hydrodynamic model

In Figure 3.2.4-1, The following treatments are conducted:

- High resolution DEM and buildings footprints are combined to define the high positions and the places on the ground mapping, where water cannot flow.
- Streets and structures determine the main flow constraints that influence the direction of the flow.
- Data for the mesh are restricted to the Area of Interest (AOI), meaning the polygon that encloses the watershed (i.e., the area that collects the rainfall, which flows to the river or an outlet in the drainage network).
- The location of the Boundary Conditions (BCs), meaning the inflow conditions at liquid frontiers that are also from the intersection between the AOI polygon and the large-scale river network. Finally, the 2D mesh (elementary surfaces of processing) and the Boundary conditions are formatted in the hydrodynamic model inputs format.

During the validation phase the following workflow will be used to compute metrics:

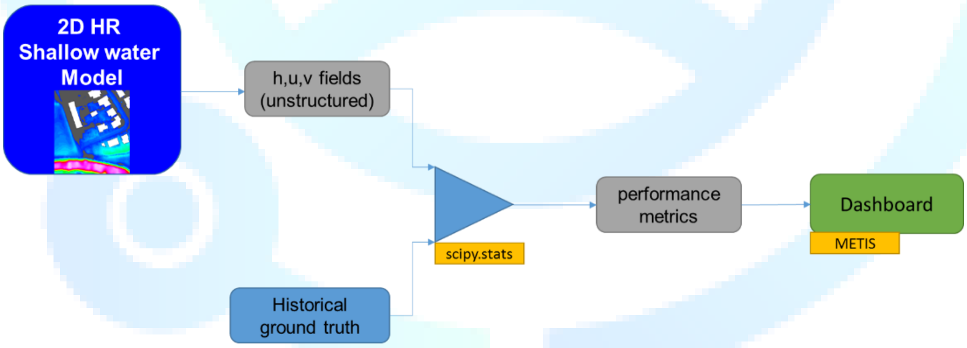


Figure 3.2.2-2: Validation phase of the workflow

The validation of the hydrologic model is done by comparing and analysing the differences with respect to the historical records.

Finally, the following workflow for the visualisation will be used:

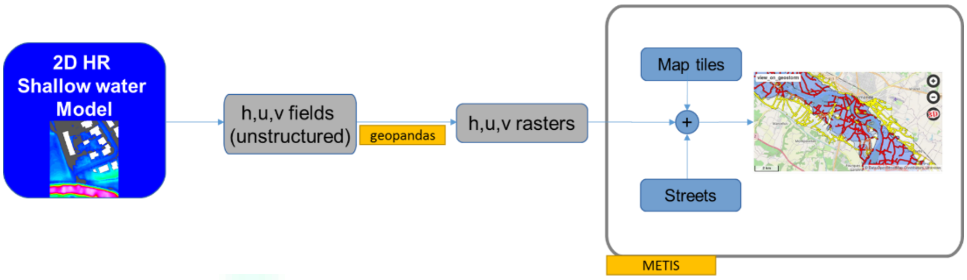


Figure 3.2.2-3: Workflow for visualization

The visualisation of the results (i.e., model outputs, historical records, difference between them) is the main way of validation, and can be completed by more specific analysis if required (when visual difference are not self-explaining)

Once the reference hydrodynamic model is up and running, two kinds of experiments will be conducted:

- Development of AI methods that facilitate the preparation of the input dataset for the hydrodynamic model from the raw input datasets,
- Design and benchmark of an AI model to replace the hydrodynamic model.

These two experiments aim to provide tools for easy and as automatic as possible forecast of urban flash floods.

Considering the first experiment, AI model dedicated to nowcast or forecast of floods are present in the literature. For instance, the following model architecture can be found in Hofmann and Schûttrumpf, 2021 [2]:

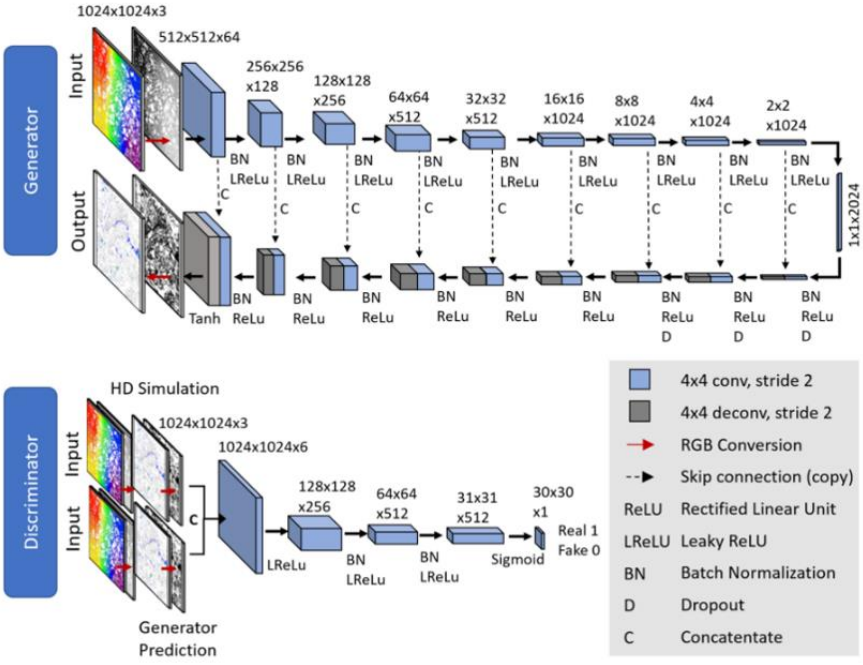


Figure 2. floodGAN architecture and process flow of the training.

Figure 3.2.2-4: FloodGAN architecture and process flow of the training

In [2], Hofmann and Schûttrumpfd trained an AI model with the architecture shown in Figure 3.2.2-4 on the results of a hydrodynamic model and assessed the performance and accuracy of this AI model compared to the hydrodynamic model. They found that the AI model is up to 10^6 times faster than the hydrodynamic model and promising in terms of accuracy and generalizability. Thus, their experiment is similar to what we want to achieve in this UC.

Moreover, other models exist in the literature. For instance, Physics-informed Neural Networks (PINNS) [4] or combinations of multiple machine learning models [5] have shown promising results for the prediction of flash flood events. Such models will be reviewed and benchmarked to select the most relevant AI architecture for our UC.

The goal of this experiment is to provide a good balance between model accuracy and accessibility (in terms of availability and processing requirements) of the input datasets. The targeted accuracy will be discussed in detail with the City Council of Nîmes, in order to establish a list of metrics and qualitative criteria. As the City Council of Nîmes has the best field expertise it is important that they participate and give us insight of the critical flow areas and phenomena that need to be accurately simulated or forecasted.

Next, the ExtremeXP framework will be used to test the impact of some variability points, such as modification of the topography (for instance the creation of levees, storage areas, etc.) or any other modification in the area that could be envisioned by the City Council of Nîmes to mitigate the impact of flooding events. In practice, this will consist in altering the topography (DEM) either for the creation of levees (increase of the topography) or the storage areas (decrease of the topography).

The impact of such modifications will be assessed by comparing the flooded areas and water heights before and after modifications, with the obvious target of reducing the damages in some critical locations (main roads, etc.). Here, we envision dedicated metrics, such as the number of locations where the water height after the modifications is less than a specific threshold or the ratio between the flooded areas after and before the modifications.

The targeted actors of this UC are either the City Council of Nîmes or engineering offices involved in the mitigation of flood damages. These actors have limited knowledge in hydrodynamic modelling and data analytics in general. Hence the interactions of these users in the ExtremeXP framework must be as easy as possible. Hence this UC must provide synthetic definitions for the input data as well as dedicated visualization tools (for capturing flood maps, timeseries plots of the predicted vs true water height at in-situ station, etc.).

3.2.3 Exploited datasets & data acquisition methodology

In order to model the flooding events in the city of Nîmes, two types of datasets are required by the 2D hydrodynamic model.

- The first type concerns the static datasets (Basin contour, Digital Elevation Model, buildings and streets footprints, Land Cover, underground sewer, and rainfall network plans).
- The second type concerns the dynamic datasets (rainfall, inflow timeseries at liquid frontiers of the domain, in-situ data for validation). These datasets are then used to build up the 2D hydrodynamic model and to perform simulations. After that some of these inputs will be used by the ML scenarios.

Regarding static datasets, most of them are open data and acquired through direct download from the provider's website and will be pre-processed internally at CS GROUP to ensure consistency with other static datasets. However, some datasets provided by the City Council of Nîmes (such as the

Area of Interest, sewer, and rainfall network plans) are not public. These datasets have been or will be given to CS GROUP via email attachments or download links and are or will be located in CSGROUP's premises.

The static datasets currently considered are:

- City of Nimes Basin contour (Area of Interest), provided by the City Council of Nîmes
  - Data: geo-referenced polygon
  - Format: ESRI Shapefile
- LiDAR HD (Light Detection and Ranging Hight Dimensional data)
  - Data: Digital Service Model (Lidar Point Cloud),
  - Variables: x, y, z
  - Format LAS
- RGE_ALTI (Referentiel Grande Echelle Altimétrie, https://geoservices.ign.fr/rgealti)
  - Data: Digital Elevation Model
  - Variable: z (x, y)
  - Format: GeoTIFF
- Corine Land Cover
  - Variables: land type
  - Format: GeoTIFF
- BD TOPO (Base de Données Topographique, https ://geoservices.ign.fr/bdtopo)
  - Variables: buildings polygons, streets polylines
  - Format: ESRI Shapefile
- OpenStreetMap (via osmnx tool, https://github.com/gboeing/osmnx)
  - Variables: buildings polygons, streets polylines
  - Format: geopackage

With regards to the dynamic datasets have been provided by the City Council of Nîmes via email attachments or download links. During the experiments in the ExtremeXP framework, similar datasets (e.g., satellite rainfall data) could be tested and hence acquired via the API of the providers. Finally, the results provided by the hydrodynamic model will be used for the calibration and validation of the AI forecast model.

The dynamic datasets currently considered are:

- Radar Rainfall grids, provided by the City Council of Nîmes
  - Variables: 5min cumulated rainfall depth
  - Format: ESRI ASCII raster
- Inflow discharge at liquid frontiers of the domain, provided by the City Council of Nîmes
  - Variables: time, discharge
  - Format: CSV or similar
- In-situ data, provided by the City Council of Nîmes
  - Variables: time, height
  - Format: CSV or similar
- Hydrodynamic model results
  - Variables: time, water height (x, y), velocity vector (x, y)
  - Format VTK or GeoTIFF

These datasets are of best quality in terms of spatial and temporal resolutions. As stated before, some similar datasets that are publicly available will be tested and the quality of these datasets will be assessed trough the results obtained in the simulation.

Regarding the data acquisition modalities, the static datasets are downloaded in a single batch. The dynamic datasets provided by the City Council of Nîmes are also downloaded by batch. However similar datasets (for instance for replacing the Radar Rainfall grids with Satellite Rainfall grids) could be downloaded using API that allow streaming the data during the simulation.

3.2.4 Experiment infrastructure & testbed

The hydrodynamic model will be set-up (raw data preprocessing), calibrated and validated in the METIS platform on the cloud infrastructure of CS GROUP using Python libraries and GIS software (QGIS) [3].

METIS is a web-based platform compatible with all cloud providers. Based entirely on open-source technologies, it is highly configurable, thanks to a wide range of optional components that enable its configuration based on user's requirements. The METIS platform focuses on the integration of services, such as federated data, and is highly modular. It relies on lower-level services layers linked to the hardware and application infrastructure. Thanks to its capacity for customisation, METIS is able to address the needs of a wide range of users, from developers to business users, and perform a range of tasks, from prototypes to mass processing. Finally, it allows all levels of integration of third-party applications and/or data exchanges based on standards.

The demonstrator for the experiments will be designed and tested either on the METIS platform on the cloud infrastructure of CS GROUP or on a cluster of GPUs in CS GROUP. Estimations of the computing resources will be soon estimated on the METIS platform.
The testing and validation of the user-driven experiments (modifying variability points, etc.) will be performed using the ExtremeXP framework services.

3.2.5 Benefits introduced by ExtremeXP

This UC will benefit the "Extreme Data and Knowledge Management" research direction of ExtremeXP, in the context of performing benchmarks of the analytics workflow, by utilizing different data sources. For instance, one could leverage the impact of the resolution of the topography dataset in terms of accuracy using this concept (useful for assessing the feasibility of using global open topography datasets).

The concept of AutoML will also be of major importance for this UC. This will help end-users develop and test various sets of models, input datasets and variability points, and facilitate the execution of their experiments. Finally, the ExtremeXP framework along with the technical requirements of this UC will provide a solution with automated data and results analytics pipelines that will strongly facilitate the decision-making process for the end users.

The following table lists the retained UC KPIs:

KPI Id	Name	Target
Kpi1-1	Loss of accuracy for the hybrid (AI/physical) model compared to a full physical model	< 5%
KPI1-2	Reduction of the volume of data required as input for the hybrid model for a loss of accuracy of less than 5% (compared to physical model with all input data available)	> 50%

KPI1-3	Reduction of input data resolution for the hybrid model for a loss of accuracy of less than 5% (compared to physical model with high resolution data available)	> 50%
KPI1-4	Reduction of the processing time for the hybrid model for a loss of accuracy of less than 5% (compared to physical model with all input data available)	> 50%

Table 3.2.5-1: List of KPIs for the UC 1

3.2.6 Demonstrator technical requirements

In the following table, we list the requirements to be expected on the demonstrator of the first UC. We identify each requirement with a specific identifier, provide a name and short description, as well as its priority level using the MoSCoW model.

Req. Id	Name	Description	Priority (MoSCoW)
UC1-1	***GIS visualization tools***	*The operator should be able to visualise flood maps produced by the hydrodynamic or AI forecast model onto a basemap (Google Maps, OpenStreeMaps, etc.) representing the Area of Interest.*	Should
UC1-2	***End-user definition of custom metrics***	*Each end user must be able to define and use own custom metrics to analyse the simulation/forecast results. Common metrics exist (RMSE, MAE, etc.) but they are less used than custom metrics.*	Must
UC1-3	***Datasets groups***	*In order to benchmark similar datasets with different resolutions and accuracy, the datasets could be organized in groups.*	Could
UC1-4	***Dataset alteration as variant***	*The dataset (e.g., the topography map) can be a25ltered to introduce a new variant of the experiment*	Should
UC1-5	***AI model as variant***	*The AI model to be used by the data analytics task is a variability point of the experiment.*	Must
UC1-6	***METIS integration***	*The ExtremeXP framework supports CS's METIS platform as a data processing runtime environment.*	Should
UC1-7	***Ancillary tools in data analytics***	*The visualisation method reflects the accuracy of the forecasting obtained from different AI models.*	Must
UC1-8	***Visualisation of model accuracy***	*The visualisation should reflect the difference between several models, specifically the output of the model prediction, the historical data, and the historical records.*	Should
UC1-9	***Visualisation of models' prediction difference***	*The visualisation reflects the difference between several models, specifically the output of the model prediction, the historical data, and the historical records.*	Must
UC1-10	***HR-DEM & Building***	*High resolution DEM and building footprint datasets can be combined.*	Must

	integration		
UC1-11	***Streets & structure datasets integration***	*Street & data set structured can be combined.*	Must
UC1-12	***2D Mesh and water series integration***	*2D Mesh and water series can be combined*	Must
UC1-13	***User Interaction in experiments for variability points***	*The user interface of the demonstrator allows to define the variability points (selection of the model and introduce variance in the dataset).*	Could
UC1-14	***User Interaction in experiments for scenario execution***	*The user interface of the demonstrator permit to control scenarios execution (workflow launch).*	Could
UC1-15	***User Interaction in experiments and analysis results***	*The user interface of the demonstrator permits the user to access results analysis (through visualisation).*	Could
UC1-16	***Formats of the static dataset to be integrated***	*The data analytics in the demonstrator of UC 1 support the following data format: ESRI Shapefile, Format LAS, GeoTIFF and Geopackage*	Must
UC1-17	***Formats of the dynamic dataset to be integrated***	*The data analytics in the demonstrator of UC 1 support the following data format: ESRI, ASCII raster, VTK, Geotiff and CSV*	Should
UC1-18	***Dataset acquisition via streaming***	*The demonstrator of UC 1 can interface with the API to acquire by streaming by connecting to the API of the provider.*	Could
UC1-19	***Dataset acquisition via batch***	*The demonstrator of UC 1 can process datasets stored on the experiment environment.*	Must
UC1-20	***Private datasets***	*Private datasets are to be persisted on CSGROUP's on-premises infrastructure, and not be transferred to another infrastructure during the data processing execution.*	Must

Table 3.2.6-1: List of requirements for the UC 1

3.3 UC2: Increased Cybersecurity situation awareness for efficient threat mitigation (i2CAT & ITU)

3.3.1 Context description and objectives

Through cybersecurity situation awareness, this UC aims to provide solutions to the fundamental challenges faced by the cybersecurity research community through the simulation of real-world attack scenarios. The perpetual evolution of the attackers to utilise advanced techniques with refined codes and to perform targeted penetration strategies towards internal and external access points of organizations make continuous enhancement of mitigation technologies essential. The challenges of the Security Information and Events Manager (SIEM) technologies and Intrusion Detection systems (IDS) for threat detection outweighs the advancement of attacks. Threat classification has become more accessible with the use of Artificial Intelligence, but there are still concerns regarding transparency and usability. The analysis of threat actors and classification of threats based on their behaviour is currently seeking to detect the evolution of such attacks. The main challenges of these mitigation technologies that the UC currently focuses on are as below:

(I) **Limitations of Security Data:** Challenges of the cyber threat datasets are often due to the continuous advancement of the threat actors' technological capabilities resulting in datasets that are either outdated, missing, limited to time constraints or unable to provide significant visibility of the diverse threats.

(II) **Automation drawbacks:** Even a single host can generate millions of alerts. Automation through machine learning and deep learning methods has received active interest through years of detecting suspicious alerts. These methods have proven to be useful on large logs and datasets but are only efficient for a subset of threats or system resources.

(III) **Lack of security contextualization:** The solo analysis of the indicators of compromise (IOC) like the IP addresses, file hashes, etc. are easily changeable by the attackers as per the Pyramid of Pain [6]. But the attacker's footprint for performing a cyberattack is something that is difficult to change. Most of the efforts to provide context and relationships of attacker's footprint require deep knowledge and expertise in security or currently rely on an external Managed Security service Provider (MSSP).

(IV) **Limitations of current security tools:** The adoption of security information management system (SIEM) technologies has provided increased visibility to security teams for threat detection. But they are expensive and require onerous efforts requiring security knowledge and expertise to identify the missed alarms, false positives, and negatives. So, the sole use of these technologies is not efficient for a reliable security system.

(V) **Criticality of time:** As it is understood that the evasion of security would imperil an organization's financial security, and reputation and compromise its astute properties, time to manage such risk is highly crucial. The above challenges in current SIEM and Threat monitoring systems affect the ***time-dependent measurement metrics,*** such as the mean time to detect, system uptime, time to resolve, etc. putting critical business systems at risk.

The objective of the UC is to design a reliable and robust SIEM demonstrator that effectively manages the above challenges to implement multimodal threat detection and classification feature model trained with cybersecurity expert skills. To achieve this, the main objectives of the UC are the following:

1. Detect Threat in a real-time environment. In this case we have access to the real traffic in the UPC (Universitat Politècnica de Catalunya) and a visibility of logs from more than 3K resources of an information system.
2. Explore the Cyber Kill Chain concept as the ensemble of different non legitimate activities (further explained below).
3. Threat detection exploiting other techniques than signatures-based and coping with the current threat classification.

The behaviour-based classification uses the concept of TTP where the attacker behaviour and tools are implied as the most valuable indicators of threat. The classification of this behaviour utilizes the publicly accessible threat knowledge base, called MITRE ATT&CK framework, through the emulation of the Cyber Kill Chain concept. The kill chain concept addresses the stages of a cyberattack from a high-level perspective where each stage serves a purpose, and the threat actor must follow step by step in order for the attack to be successful. Such that by breaking the kill chain of a cyber-attack, its impact can be reduced. To understand the cyber kill chain, we can use the example of kill chain of a crypto ransomware, where it encrypts the files or systems of the victims and asks for a ransom to retrieve their access. The ransomware performs certain set of actions otherwise known as techniques and procedures to attack the victim's systems and files.

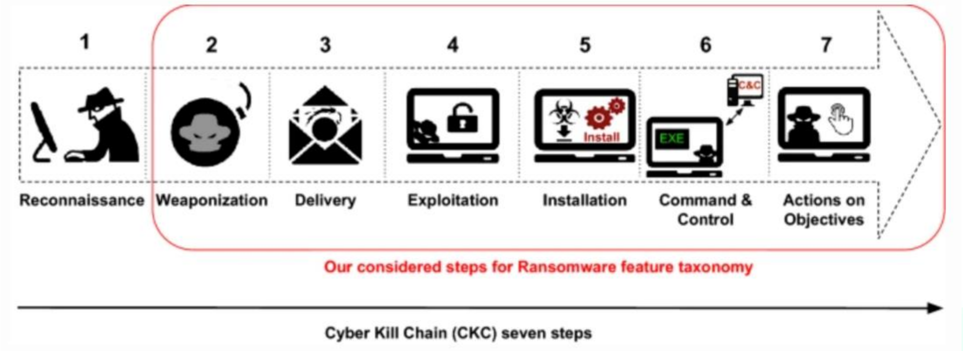


Figure 3.3.1-1: Cyber Kill Chain for Ransomware Attack

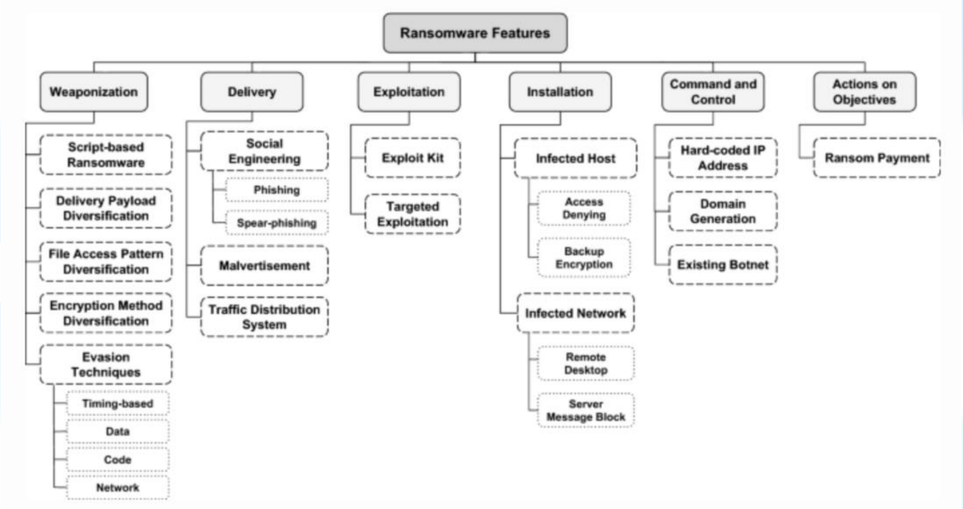


Figure 3.3.1-2: Actions taken by a Ransomware Attack at each step

The cyber kill chain created for such a crypto-ransomware is depicted in Figure 3.3.2-1 [13] and its course of actions is depicted in Figure 3.3.2-2 [13]. Although this depiction differs in providing the granularity of MITRE ATT&CK stages, it gives an idea of the attack steps taken by the ransomware to reach its goals in terms of cyber kill chain.

The MITRE ATT&CK framework provides granular details of the attack stages and refines the objectives, behaviours, processes, actions, and strategies used by threat actors into ***Tactics, Techniques***, and ***Procedures.*** Although these frameworks reduce the semantic gap between the low-level alerts and high-level kill chain stages, they fail to build relationships between attack stages for the attacks. The UC investigates the following methods to provide context and correlation between the cyberattacks in terms of:

a. Development and visualization of a complete kill chain of a cyber-attack

b. Identify causality between the co-occurrences of the attack techniques used by the threat actor [8]

c. Generating relationships between the cyber kill chains.

The correlation mechanism for understanding and implementing such relationships between the different attack techniques and threat actors plays an important role in determining known, unobserved, or unknown attacks.

In the context of ExtremeXP, this design aims to utilize its complex data analytics, experimentation engine, and visualization insights to enhance the decision-making process of the security analyst. The demonstrator aims to achieve the following (i) Training the platform to classify the threat as per the expertise of professionals in our Security Operations Centre (SOC) Team. (ii) Providing risk and time measurement metrics to the threats based on their classification (iii) Detecting and visualizing the threat occurrences and relationships. (iv) Facilitating timely decision making of the analyst.

3.3.2 Experiment definition

In this UC, the main objective of the demonstrator is to conduct an experiment addressing the following objectives:

- Onboarding tools and systems and Log sources.
- Threat modelling.
- Performance Evaluation.
- Threat detection.
- Threat Intelligence.

The main pipelines for the UC experimentation have the following steps that need to be developed from scratch:

1. Generate an attack on the sandbox environment (simulator).
2. Collect forensic information of the attack and merge it with UPC-ITU real-time data.
3. Model each step of the attack as per their techniques. The techniques here are the detailed threat actions within the context of an attacker's goal at each step.
4. Validate the model using public datasets or incidents gathered.

As per the architecture of our UC in Figure 3.3.2-1, at first, the user's and entity behaviour logs from different data sources in the UPC environment are recorded in the ELK (Elastic, Logstash, Kibana) [15] platform of the ITU. This data is integrated into the ELK platform of I2CAT's AI4Cyber environment. The logs are processed in the AI engine, which facilitates the classification, inference validation, explainability proxy, and evaluation metrics for the threat logs. These are aligned to the threat knowledge base and sharing environment to provide accessibility to wider range of knowledge for the threat logs. The resulting data is evaluated as per its severity, risk, or impact to generate the alerts. This evaluation is performed by a security expert or researcher for false positives or missed alerts. The alerts are utilized by the SOC team in two modes as below.

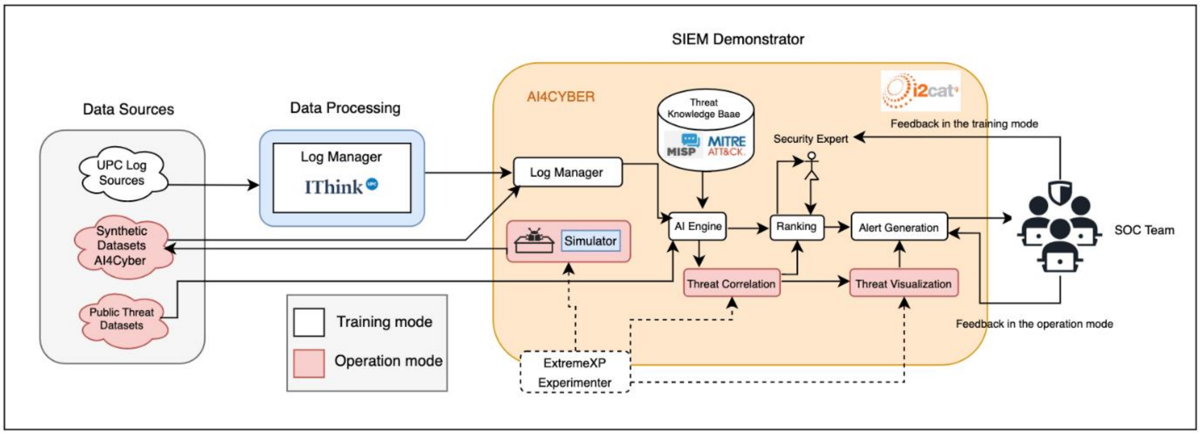


Figure 3.3.2-1: Architecture for the Cybersecurity UC

In the **training mode**, the *SOC team or a security operator* is expected to provide feedback for the alert generated. A security researcher is expected to train the platform to gain experience in detection and classification and the best practices to be enacted when responding to heterogeneous cybersecurity incidents. This user is expected to know the threat to effectively train the system and how the techniques are expected to correlate into a complete procedure (attack scenario).

In the **operation mode**, the gained experience regarding the detection of threat occurrences and their qualification are exploited, and countermeasures that would reactively and proactively mitigate the threat are initiated. Also, we can expect an attacker to conduct the attack scenarios in the sandbox environment that will generate activity logs (Synthetic dataset), later to be detected by the demonstrator in the log manager in operation mode. The regular users of the information system are to conduct benign activities, producing logs and not revealing any sort of technique execution. The SOC team should be informed through a dedicated visualization about the encountered techniques correlation.

The workflow of the above steps is presented as below:

- **Security log extraction:** The extraction of the logs will be conducted on UPC's premises, based on the available resources, and will also be collected by ITU. Each resource is expected to have its own technical challenges and tooling, requiring specific log extractors. For instance, in order to extract logs from a network appliance, we expect to interface with Microsoft Sysmon for Windows operating systems and ingest these logs into our ELK stack. The logs will be integrated to the I2CAT's Elastic Platform, using Virtual Private Network (VPN) security. This data will be used for training the AI engine. The public datasets are used for the validation purposes.

- **Techniques instance identification and classification:** The Logstash in the ELK stack can ingest data from a multitude of sources simultaneously and transform it. Those transformed data is then sent to the Elasticsearch engine for its easy and fast search and analytics capabilities and then Kibana in the ELK stack [15] can be used for its flexible visualization of all the data. The ML model in the AI engine allows to determine if a behaviour is legitimate or non-legitimate for the time window (specifically behaviour aligned to a threat). From a retrieved set of logs, we want to run a set of trained ML models to conduct the recognition of a used technique. In the literature, LightGBM has been identified as the best in terms of overall accuracy for classification purposes [10]. The model to employ will vary on the techniques to be recognized and the data source to be exploited. The result of these classifiers is expected to be a probability expressed as a float ranging between 0 and 1, and a set of properties qualifying the source of this activity (e.g., IP address, name, and version of the communication protocol used ...). The asset at AI4Cyber testbed analyses the information contained in the logs, aligns it as per MITRE ATT&CK framework data sources, and provides a high-level view of the ATT&CK Matrix, reflecting the color-coded tactics and techniques that can be monitored with the current information of the system. Techniques will be aligned on the MITRE ATT&CK referential as per the Kill chain concept. The asset also provides a detailed view of the data analysed on each individual Elasticsearch Index. It will allow the Security Operations Centre (SOC) teams to assess the degree of visibility they have over several Tactics, Techniques, and Procedures (TTPs) from real-world threat scenarios. The techniques to be recognized and the implementation will be selected from a threat catalogue.

- **Procedure recognition:** If a set of techniques is detected, we would like to correlate them, to link their instantiation to an attack scenario modelled as an attack graph. This graph will technically detail how an attacker technically conducts an attack. One challenge will be to recognise the concordance of several threat techniques to a common root cause and temporality. The asset on explainability proxy intends to provide a better insight of how the threat behaves. This explainability proxy is based on annotated explanations using the MITRE taxonomy building a ground truth knowledge base. This system covers the biggest open challenges of cybersecurity by providing transferability through transfer learning, contextualization through risk assessment of augmented threat data as per the operator context and trust aware interfaces through reduction of cost for false positives and false negatives.

- **Threat actors' correlation:** For enhanced detection and threat visualization, our research would be on the threat correlation through attack graphs, clustering based on sophisticated algorithms and as per their kill chain to detect the unobserved attack scenarios. The correlation will be using the APIs from the threat intelligence sharing platforms of MISP, OpenCTI and MITRE. This component can enable threat intelligence gathering, implementation of algorithms and evaluation of the security capabilities. The several attack graphs in different temporality are needed to be correlated to assist the operator to achieve a high level cyber situational awareness [14]. The final objective is to detect the similarity between a current attack graph from one already encountered in the past. In practice, this similarity will permit us to identify if a procedure has already been identified in the past. In that case, this will (i) reveal similarities in the activity of a threat actor and (ii) enact anticipation of the next techniques used in the procedures. The visualization of the similarity is formulated through a similarity score (a float) reflecting how much an attack diverges from another [9] (e.g., Pearson Correlation Similarity).

During the operation phase, several steps of the workflow detailed above are expected to be instantiated with synthetic data.

- **Techniques instance identification and classification:** the techniques recognition model is trained with synthetic dataset reflecting techniques exploitation and augmented with contextual data. A first prior step will be the establishment of a multimodal dataset.
- **Threat actors' correlation:** The model for threat correlation will be manually trained from the information stored from ITU's ticketing system. We would like to address this process as an ExtremeXP experiment, where the user can train this correlation process based on this know-how.

- **Threat visualization:** The selective visualization of the alerts based on their malicious activity needs to be aligned with a kill chain, show prevalent techniques and also the less noticeable ones, and align to different threat actor kill chains.

The definition of the variability points is expected to be manual work conducted by an experienced analyst. The pertinence of a set of values of variability points shall be reflected in the detectability of threats: diminution of the false positive and false negative of the detected threat. The impact on threat correlation is yet to be defined. The selection of the threat detection model & workflow is a clear variability point. Our main objective is to select a detection method delivering the most precise result on a threat occurrence but remains explainable enough to be correlated into an attack graph [7].

3.3.3 Exploited datasets & data acquisition methodology

The different logs allow to generate multimodal time windows allowing to profile the behaviour of an entity in a specific time range. The real logs from ITU are integrated to our laboratory infrastructure using VPNsec for our research purposes. The format is Bro logs that transforms the ASCII logs, and we store them using Elastic search (open-source version). We will be utilizing a Side-to-side ELK platform and a VPNSec with ITU infrastructure to transmit the logs. This will be facilitating the AI engine for the classification of threats. The public datasets will be initially used for the validation of the AI Engine. We will be simulating different application-based attacks in our AI4Cyber Testbed and generating a multi-model dataset to feed to our ELK platform for validation. Depending upon the type of data source in the dataset, the set of columns for the logs will vary as shown in the table of Pages 7-9 in the document Guide to Cyber Threat Information Sharing.

Currently, at UPC, approximately 168,937,836 logs from different applications and network infrastructures are generated daily, and there is no single repository where all the logs converge. So, ITU's focus will be on the repository where most application logs are stored in the ELK service. The data sources of the UPC environment are Firewall (traffic), Firewall (layer 7), Switch, Routers, Load Balancer (F5), Virtualization (Cloud), Kubernetes, Servers, Desktops/Laptops, Antivirus, SIEM, Databases (back ends) and Web Servers (front ends). The size of the data is currently less than 50GB and is maintained for 14 days to 3 years depending on the data source. For accessing the logs of the applications, ITU will be using the ELK API to query the database directly or through the Kibana graphical interface. The formats extracted are in the standard JSON format.

As the diagram below explains, Splunk system at ITU receives all the logs expected to generate alarms. On Splunk, ITU applies rules selected by I2CAT for the attack techniques from the MITRE ATT&CK framework.

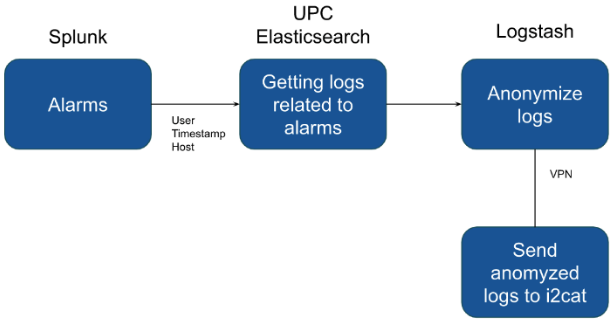
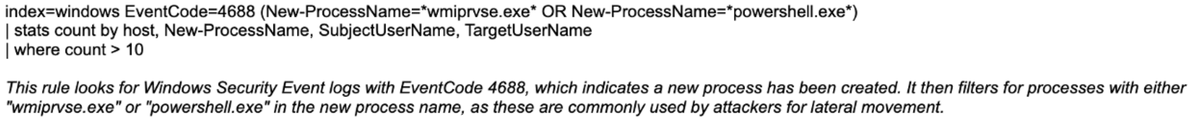


Figure 3.3.3-1: Data acquisition from ITU

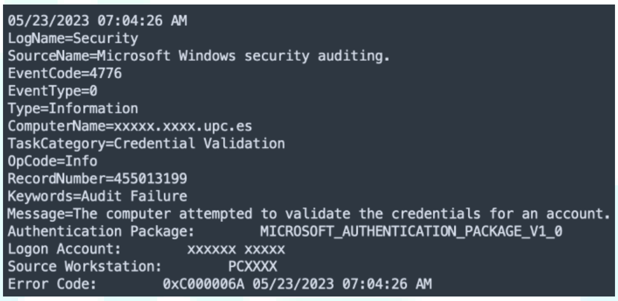
When a suspicious alarm is raised, ITU takes the user, timestamp, and host to search for all the logs related in the ELK service, anonymise them by a Logstash plugin and then sent to I2CAT's ELK platform through a VPN service. An example of the Splunk rule provided to ITU is as specified as below:



```
index=windows EventCode=4688 (New-ProcessName="*wmiprvse.exe* OR New-ProcessName="*powershell.exe*")
| stats count by host, New-ProcessName, SubjectUserName, TargetUserName
| where count > 10
```

This rule looks for Windows Security Event logs with EventCode 4688, which indicates a new process has been created. It then filters for processes with either "wmiprvse.exe" or "powershell.exe" in the new process name, as these are commonly used by attackers for lateral movement.

Figure 3.3.3-2: Example of Splunk rule



```
05/23/2023 07:04:26 AM
LogName=Security
SourceName=Microsoft Windows security auditing.
EventCode=4776
EventType=0
Type=Information
ComputerName=xxxxx.xxxx.upc.es
TaskCategory=Credential Validation
OpCode=Info
RecordNumber=455013199
Keywords=Audit Failure
Message=The computer attempted to validate the credentials for an account.
Authentication Package:        MICROSOFT_AUTHENTICATION_PACKAGE_V1_0
Logon Account:        xxxxxx xxxxx
Source Workstation:        PCXXXX
Error Code:        0xC000006A 05/23/2023 07:04:26 AM
```

Figure 3.3.3-3: Raw anonymised logs from ITU for Bruteforce Attack

The limitation of such logs remains in raw format and there is low standardization of data leading to complexity in selecting the fields of the dataset for its utilization.

3.3.3.1 Synthetic Dataset

The dataset obtained from the simulation of attacks in the AI4Cyber testbed at I2CAT will be injected to the ELK platform to validate the accuracy of the models, the evaluation metrics and identification the threat behaviour.

3.3.3.2 Public Dataset

For the operational purposes before constructing a synthetic dataset, the UC will utilize public datasets for the analysis of threats. Public datasets are mainly used in this UC to validate SIEM demonstrator's accuracy to understand the behaviour in comparison to the real-time data from UPC and explore other attacks. Currently the below-mentioned datasets have been identified for our initial validation:

- Cybersecurity datasets. Canadian Institute of Cybersecurity have large number of datasets that are widely used in security testing by industries and universities. CIC-MalMem-2022, CIC-Evasive-PDFMal2022 (Stores malware attacks)
- Malware captures by Stratosphere Lab provides datasets on recent prevalent malwares.
- USTC-TFC2016: Stores traffic datasets.
- DOHBrw-2020: Stores DNS over HTTPS.

These datasets have been used for many research publications for the analysis of threats and their size is expected to be less than 50 GB. These datasets will be used to provide variability to the logs and to validate the efficiency of the SIEM demonstrator. The datasets will be injected into the AI Engine for further processing.

3.3.4 Experiment infrastructure & testbed

The demonstrator will be implemented and evaluated in the AI4Cyber testbed provided by i2CAT. The lab is controlled with OpenStack and Terraform technologies, which allow us to deploy flexible and scalable environments in an agile way. It comprises of a complete stack of tools for project development and management in the field of cybersecurity and artificial intelligence, including user behaviour analysis, threat analysis, threat profiling and modelling. The processing infrastructure comprises of 104 CPU processing cores, 20Tb of disk, 512 GB RAM, GPU NVIDIA Tesla V100 5.120 cores. The evaluation will be conducted by using synthetic datasets prepared by exploiting pen-testing tools. According to the selected threats for evaluation, the malevolent activity will be generated by simulating several attacks in an isolated environment, in the AI4Cyber testbed, or retrieved from open repositories. Eventually, the UC will be confronted against production log data retrieved from UPC infrastructures (e.g.., NetFlow logs from network equipment, system logs from 3000 workstations and servers) by ITU to validate the practicality of the demonstrator in real-world scenarios. We will be using the VPN to integrate the tasks between ITU and i2CAT. We target the collection of 100 resources into our dataset and retrieve the logs for several days.

3.3.5 Benefits introduced by ExtremeXP

The shared situational awareness through the collective experience, knowledge, and analytic capabilities of the ExtremeXP framework, the defence capabilities of the SIEM demonstrator can be enhanced. This demonstrator will exploit the ExtremeXP framework to increase the recognition of cyber threats by featuring efficient and accurate AI on extreme data, contributing to the emergency management of information systems in several verticals to generate threat intelligence in the information layer. The demonstrator will feature the three pillars of ExtremeXP: (i) Extreme Data and Knowledge Management, to collect the data from distributed and heterogeneous resources across and information systems, (ii) User-driven AutoML, to exploit the adequate AI models to recognise threat techniques and draw attack scenarios from service logs and metrics, and (iii) Transparent & Interactive Decision Making, to learn the best analytics practices in incident detection from the experience of an operator. Also, the experiments showcased in the UC might be deployed in other UCs but also in other SOC domains such as critical infrastructures, SMEs etc.

The list of KPIs is highlighted in the table below where the first KPI will determine 10 different attack techniques from the threat classification mechanism. The second KPI is to reduce the false positives or negatives in the alerts from the achieved classification of threats and determination of techniques. The third KPI is the mean time to act on the classification of alerts traditionally versus the usage of classification of alerts by threat analysis and technique classification is reduced by 30 minutes. This KPI is achieved using the model prediction as well as the correlation of behaviour from the logs with the threat techniques to create attack patterns and prioritization of the risk associated. The operator can have exact visibility of the cyber threat with its context which is necessary to take faster decisions.

KPI Id	Name	Target
KPi2-1	Detection of multimodal threat techniques referenced by the MITRE Att&ck framework	10
KPI2-2	False positives and negatives on threat techniques classification	<10%
KPI2-3	Mean time to classification compared to traditional human techniques	<30’

Table 3.3.5-1: List of KPIs for the UC 2

3.3.6 Demonstrator technical requirements

In the following table, we list the requirements to be expected on the demonstrator of the second UC. We identify each requirement with a specific identifier, propose a name and a description, and priories its execution using the MoSCoW model.

Req. Id	Name	Description	Priority (MoSCoW)
UC2-1	***Threat method selection***	*The security operator should be able to select different ML algorithm as threat detection methods.*	Must
UC2-2	***Threat correlation and feedback***	*The security operator is informed with a representation qualifying the detected threat and be able to confirm its existence.*	Must
UC2-3	***Threat Visualization***	*The security operator is able make proper decision and observations through the graphical and technique-based knowledge representation of the threat.*	Could
UC2-4	***Kill chain Visualisation***	*In UC2 demonstrator, the detected and correlated threat techniques are visualisable as a kill chains.*	Must
UC2-5	***MITRE Att&ck matrix involvement***	*In UC2, the detected threats can be visualisable as a set of correlated set of TTP in the MITRE Att&ck matrix.*	Should
UC2-6	***Attack simulations in sandbox environment***	*In UC2 demonstrator, a sandbox environment permits simulating attacks, monitor the techniques.*	Must
UC2-7	***Forensic information collection***	*In UC2 demonstrator, forensic information can be collected to generate datasets from simulated attacks.*	Must
UC2-8	***Attack phase modelling***	*The demonstrator detects and models the different attack techniques from a set of data via the AI engine.*	Must
UC2-9	***Threat analysis***	*The AI engine analyses the logs to detect and classify threat techniques, provide explainability context.*	Must
UC2-10	***Threat detection feedback***	*In UC2 demonstrator, the security operator can provide feedback as per the risk, severity, and impact of a detected threat (alert).*	Must
UC2-11	***User feedback in training mode***	*In training mode, the UC2 demonstrator permits a human operator to collect feedback on the detected techniques.*	Must
UC2-12	***Experience usage operation mode***	*In operation mode, the UC2 demonstrator informs a human operator about the detected threat techniques and correlation achieved without asking for feedback.*	Must
UC2-13	***Data collection for Windows endpoint***	*The UC2 endpoint supports logs generated by Microsoft Sysmons.*	Could
UC2-14	***Demonstrator accessibility and Security***	*The UC2 demonstrator is only accessible through a VPN, requiring a VPN connection for data ingestion.*	Should
UC2-15	***Multiple data ingestion***	*The UC2 demonstrator is able to ingest logs data from multiple sources simultaneously.*	Must

UC2-16	***Threat sharing platform involvement***	*Threat intelligence sharing platforms, such as MISP, OpenCTI and MITRE, will interface with the UC2 demonstrator for the correlation process.*	Should
UC2-17	***Network equipment data sources***	*The network equipment serving as data sources for UC2 comprises of Firewall (traffic), Firewall (layer 7), Switch, Routers, Load Balancer (F5) from UPC infrastructure.*	Could
UC2-18	***Application data sources***	*The application serving as data sources for UC2 comprises of* Antivirus, SIEM, Databases (back ends) and Web Servers (front ends) *from UPC information system.*	Could
UC2-19	***Infrastructure Data sources***	*The infrastructure data source for UC2 comprises K8S cluster, server, desktop, and laptop endpoints from UPC's information system.*	Could
UC2-20	***Threat alarms***	*The UC2 demonstrator interfaces with Splunk to obtain alarm on potential threat and initiate the data analytics task.*	Should
UC2-21	***Dataset Selection***	*The UC2 demonstrator has two modes, one is a training mode and the other one is an operation mode where the selection and utilization of different datasets and their features take place.*	Must

Table 3.3.6-1: List of requirements for the UC 2

3.4 UC3: Situational intelligence and decision making for PPDR (Public Protection and Disaster Relief) (ADS)

3.4.1 Context description and objectives

In the third UC, the scenario focuses on the significant role of extreme analytics coupled with data personalisation and presentation for the detection, characterization, and operation of fire scenarios in an urban area, via an enhanced situational awareness and critical human operators' decision making.

We will explore ExtremeXP capabilities, such as the experience-centred and assisted decision making, trustworthy and explainable AI, extremely precise prediction, in building strong situational awareness, as well as a precise prediction of operational evolutions that first responders will be able to share through their Mission Critical collaboration platform. This platform allows on-field users form different organisations to communicate with talk groups, or video groups, or messages with the control centre. All data collected from sensors are available on the control centre application on a map or in different menus (type of user, location, status, videos, pictures, etc.).
The ExtremeXP framework will be applied to demonstrate the added value that combined complex analytics, personalization, assisted decision making, and mission critical services could provide almost real-time in the scope of improving operational situational awareness and decision making, based on the mission, organisation, role, location, and context of the first responders.

- **Detection phase:** Currently operators survey multiple screens displaying videos from on-field cameras. When an operator detects an issue, such as a fire or smoke, he/she would have to continuously monitor the situation on the camera, identify the location, choose users on the field and alert pertinent users. An analysis of all the videos with a form detection capability will send off alerts to the command control with all relevant data, such as pictures of the detected fire, location, type of object/vehicle in fire of detected, fire progress, etc. to allow the operator to validate the alert.

- **Decision making & actuation triggering phase:** In case of a validated alert, the operator would have firstly to detect all pertinent procedures (forest fire, vehicle/truck in fire, etc.) and secondly choose the corresponding categories of users on the command-and-control map. This choice entails a series of actions: search, and selection of the closest users in each selected category and secondly a status check of the availability of each user. If a user is available, he has to contact by phone or message each user and check if the user can reply to his request. In case of negative answer, the operator has to select other users of the same category closer to the alert location and contact this user in order to choose another unit. In case of positive answer, the operator sends required tasks to the user.

Within the ExtremeXP's framework, the main objective is to decrease the duration of the decision-making process, by providing the operator with a better view of the current situation, thus enhancing situational awareness and facilitate the actuation:

- Depending on the type of alert, the system will display a selection of actions to be launched on the control centre by the operator. Elected, on-field users will be displayed on the control centre map to facilitate the appointment of selected users as available and close to the alert location and allocation of corresponding tasks.
- Next the operator will have to validate the appointment, by clicking the validation or propose another choice on "another choice menu"
- Then the system will directly send a message to all selected users with a request for an answer. If the user responds to the request, all pertinent data will be sent to his/her smartphone.
- If the user rejects the order, a message is sent to the closest available user of the same category for a new selection.

In this UC, we identify two classes of users: the operator, who is in charge of processing alerts received in the control center and makes decisions, and on-field users in charge of responding to those decisions.

3.4.2 Experiment definition

The threat of fire is a significant concern for people all around the world. Traditional methods of detecting fire often rely on sensors that use fire parameters to identify the presence of flames. Unfortunately, these sensors can take some time to detect the fire, which can be problematic as the flames can quickly grow out of control and cause significant property damage. To address this issue, a more efficient system for detecting indoor and outdoor fires should be developed, using video surveillance cameras. In recent years, computerized monitoring methods have become increasingly popular for identifying fires. However, accurately determining whether a region is actually on fire remains a challenge due to the changing colour of flames from red-yellow to nearly white.

The system will contain the following modules:

- **Detector:** Detecting and localizing fire.
- **Recommender:** Recognize the context and disseminate information until the appropriate first responder unit.
- **Optimizer**: Optimize first responder flow and response time.

The experiments will be split into three separate phases.

The first phase will consist of effectively detecting the fire and pre-event situations as smoke with object recognition algorithm by considering **spatial and temporal complexity**.

The second phase will consist of context recognition for situation as forest fire, car fire, or narrow street fire by representing content in **visual relationship** as Spatial-Temporal Graph Neural Network and suggest the appropriate first responder unit.

The last phase will consist of managing and optimizing first responder, unit interventions by formulating as **resource allocation** problem with availability and localization constraint related to real world road constraint.

3.4.2.1 Early-stage fire detection and localization

The main objective of fire detection is to immediately identify the presence of fire and signal an alarm to take swift action. As stated in |10], it is crucial to detect fire as quickly as possible to prevent any serious damage to property or lives. Various effective methods for detecting fire are available, but detection speed is of utmost importance. Therefore, it is essential to develop a reliable system using the available resources that can detect fire instantly and alert authorities promptly to prevent any potential hazards. The challenges of accurately detecting fire and sounding alarms urgently need to be addressed to allow for sufficient time for evacuation and extinguishing the fire.

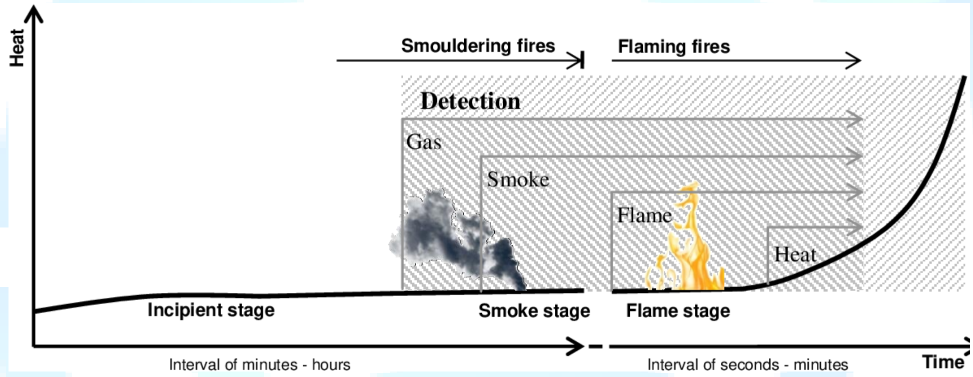


Figure 3.4.2-1: Survey of fire detection in vehicles, from [10]

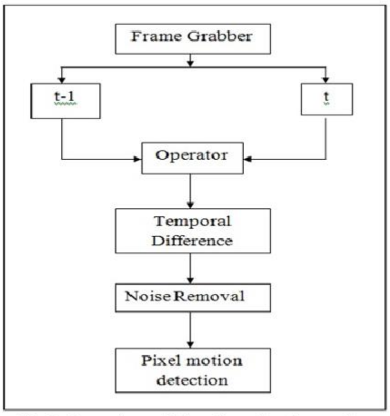


Fig 2: Overview of Fire detection by motion

Figure 3.4.2-2: Fire detection by motion, from [11]

This innovative system proposed by [11] uses three key features of fire, including colour, motion, and shape, which are enhanced with fuzzy logic for added accuracy. The main goal of this system is to minimize false positives in fire detection and improve the effectiveness of the model in both indoor and outdoor settings.

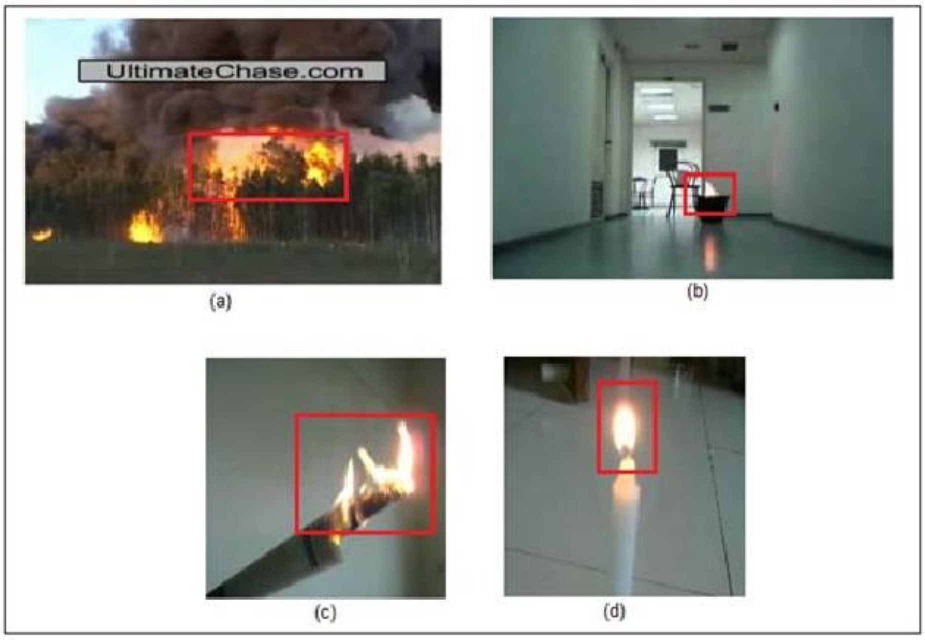
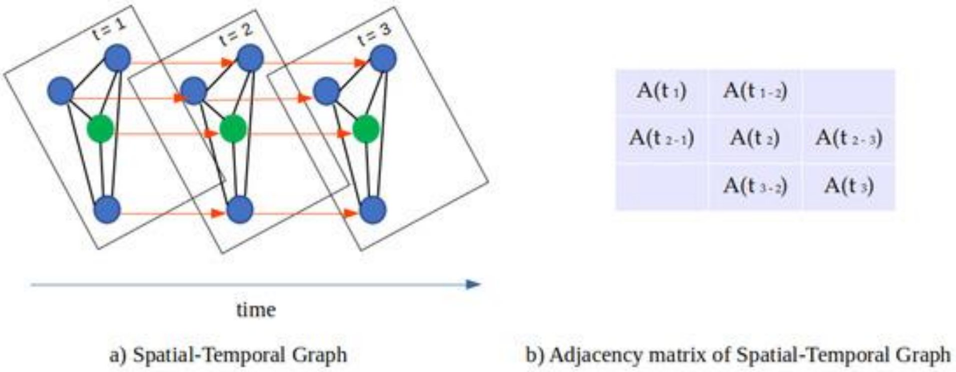


Figure 3.4.2-3: Fire detection by video surveillance, from [11]

Figure 3.4.2-3 (a) through (d) images are taken from our in-house datasets and depict the detection of fire.

3.4.2.2 Recognize the context and disseminate information until the appropriate first responder unit.

When it comes to recommendation systems, it is important to integrate side information along with user-item interactions (in this context, the operator) in order to provide personalized recommendations and improve overall performance. As suggested by [12], this can be achieved by treating recommendation problems as a link prediction task in a bipartite graph between user and item nodes, which are labelled with rating information on edges. By incorporating context, such as the circumstances surrounding the interaction between users and items, recommendation systems can better cater for user preferences and opinions.


t = 1
t = 2
t = 3
time
a) Spatial-Temporal Graph
b) Adjacency matrix of Spatial-Temporal Graph

$A(t_1)$	$A(t_{1-2})$	
$A(t_{2-1})$	$A(t_2)$	$A(t_{2-3})$
	$A(t_{3-2})$	$A(t_3)$

Figure 3.4.2-4: Spatial-Temporal Attention for Multi-Sources Time Series Data, from [12]

In order to understand the relationships between nodes over time, we connect each node with its neighbours sequentially for each time step. By doing this, we can create a spatial-temporal graph that connects nodes from previous, current, and future time steps. This approach provides insight into how nodes interact with each other over time.

3.4.2.3 Optimize first responder flow and response time.

A common way to frame a scheduling and resource allocation problem is by outlining the various activities and resources involved. This can be illustrated using a generic diagram that highlights the relationships between the two.

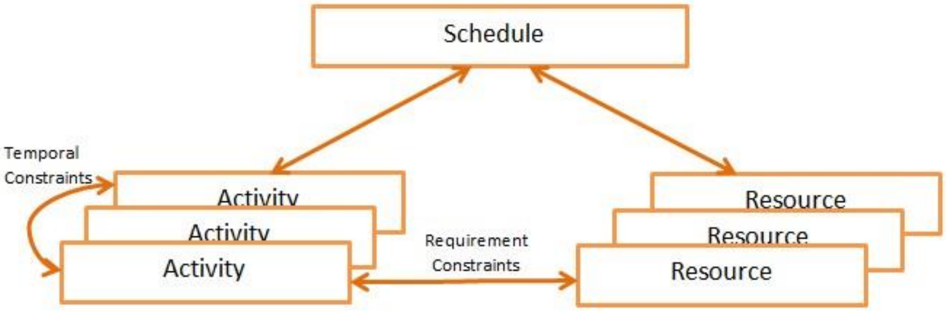


Figure 3.4.2-5 Scheduling and resource allocation with activities

In the realm of resource management, it is important to keep in mind that resources have limitations, and those limitations can change over time. For instance, an individual may only be available for a certain number of hours during the day. Resources can also be classified as recoverable, like humans, or consumable, like money or gasoline. In order to properly allocate resources, it is necessary to consider the various requirements of each activity and ensure that the proper constraints are in place. This high-level schema provides a useful framework for decision modelling, allowing us to solve a wide range of practical problems.

3.4.3 Exploited datasets & data acquisition methodology

The datasets will be available through an API on WebSocket or simple RTSP/RTMP. The user will register to dedicated data services and receive real-time data. The API will allow requests for users list (characters lists – Mb) and users characteristics (Status (characters kb), location (characters kb). The API will also have access to Users camera display (same request for fixed cameras). The video streaming is encoded in H264 and conveyed on RTP. The API will also allow to send messages to the users (characters - kb), pictures (jpg – Mb), real-time videos streams (RTSP H264– Mb/Gb) and requests. The API can receive messages (text - kb), video (RTSP H264– Mb/Gb) and pictures (jpg – Mb). Some users will be simulated and have the capability to move on the map (with pre-generated travels), reply to messages requests (characters - kb), cknowledge orders and move to assigned locations (characters - kb).

3.4.4 Experiment infrastructure & testbed

The testbed will be composed of a server hosting the multimedia communication platform and the user simulation application.

The server requirements are the following:

- 1 vCPU and 1GB RAM

- Bandwidth Rx: Around 500kbps per video stream and per user for SD stream (640*480) 15fps /Around 5Mbps per video stream and per user for FHD stream (1920*1080) 30fps/ Around 50Mbps per video stream and per user for 4K 60fps.
- Bandwidth Tx: Around 500kbps per video stream for SD stream (640*480) 15fps /Around 5Mbps per video stream for FHD stream (1920*1080) 30fps / Around 50Mbps per video stream for 4K stream (3 840*2 160) 60fps.

Tablets or computers will host the command-and-control application and the on-field user can access the system via the application installed on Smartphones or tablets.
The main requirements for tablets and smartphones are as following:

- Android version 9 minimum.
- 4BD RAM.

The evaluation will be conducted in a lab environment, which will simulate the conditions of a real scenario for the UC, as closely as possible. It will include a combination of simulated data and/or pre-registered real-time information for the various flows (e.g., sensors, video flows) to emulate the huge flow of data expected in a real-life scenario, real data from sensors and video flows included in the scenario, and open-source data. Typical dimensioning for the expected streams to be analysed for such a scenario would be:

- Real time video quality: 4K resolution.
- Number of concurrent real time video streams: > 30.
- Real time audio quality: at least AMR WB 23.85 Kbit/s.
- Number of concurrent real time audio streams: >30.
- Data transfers average size: 2 Mbytes.
- Number of concurrent data transfers: >30.
- Number of concurrent short messages: >300.
- Open-source data: event related data, such as social networks or public information.

3.4.5 Benefits introduced by ExtremeXP

The main benefit introduced by ExtremeXP would be the diminution of the decision-making duration via the enhanced situational awareness of the operator, and coordination with the responders. The following KPIs formulate those improvements:

KPI Id	Name	Before	After
KPI3-1	Time to automatically detect critical information and give initial alert.	>10 s	<2s
KPI3-2	Practitioners' reaction time (decision making & actuation triggering) from the moment alert is received	>5min	<1min

Table 3.4.5-1: List of KPIs for the UC 3

The first KPI defined (Time to automatically detect critical information and give initial alert. Before: 10 seconds / After <2 seconds) is to improve the detection time between the incident to the validation of the alert. This detection phase can be used in multiple UCs. For the police organisation this improvement can be applied to gun detection, attack detection... The fire-brigades can use it for fire detection, accident detection...

The second KPI (Practitioners' reaction time (decision making & actuation triggering) from the moment alert is received. Before > 5minutes / After < 1 minute) starts from the validation of the alert to the actions taken

The solution will provide making decision assistance, providing to the operator an overview of the situation, a list of operations and on-field users to imply in the mission.

The proposition can be validated by the operator, orders will automatically be sent to corresponding actors of the mission.

The operator will be able to modify the proposition if needed for a more accurate mission.

3.4.6 Demonstrator technical requirements

In the following table, we list the requirements to be expected on the demonstrator of the modification UC. We identify each requirement with a specific identifier, propose a name and a description, and priories its execution using the MoSCoW model.

Req. Id	Name	Description	Priority (MoSCoW)
UC3-1	***Detection alert***	*A notification stating an anomaly detection is displayed on the control center interface along with pertinent data (type of alert, location, etc.)*	Must
UC3-2	***Protocol selection***	*Depending on the type of alert, a protocol is proposed to the operator displayed on the control center interface.*	Should
UC3-3	***Users' selection***	*The system displays lists of relevant on-field users (selected by distance from the alert location, status, and type of user) on the control centre interface.*	Must
UC3-4	***Scenario replay***	*Users are simulated, with location and status scenario. The scenario will be re-playable.*	Should
UC3-5	***Video Stream as datasets***	*The UC3 demonstrator acquires data from data streaming from video stream.*	Should
UC3-6	***Data visualisation***	*In the UC3 demonstrator, the detected fire and contactable personnels are visualised on a map.*	Should
UC3-7	***Interaction with third party***	*In the UC3 demonstrator, on-field users (first responders) confirm the usage the decision made by an operator in the control center via an interaction with their smartphone.*	Should
UC3-8	***User simulation***	*Some users of the UC3 demonstrators are simulated and their location predefined as travels.*	Must
UC3-9	***Mobiles and tablets support***	*The user interface of the UC3 demonstrator is accessible from smartphones & tablet.*	Should

Table 3.4.6-1: List of requirements for the UC 3

3.5 UC4: Flexible transportation analysis and visualization (MOBY)

3.5.1 Context description and objectives

In the transportation platforms ecosystem, the main stakeholders/decision-makers are transportation authorities, transportation planners and modellers. They need to assess different transportation policies, using a variety of simulation software, advanced econometric modelling, or other data-driven tools.

However, each stakeholder group (transportation planners, modellers, policy makers) has different requirements with respect to the granularity/level of analysis and corresponding outcome precision and quality. There is no "one-size fits all" when it comes to data analytics and visualisations in the transportation domain.

Therefore, ideally, the transportation analysis platform can adapt the

- (i) simulations and data analytics performed in the backend, and
- (ii) the visualisation methods and techniques and learn the above for each user segment.

The transportation UC provides a rich space of data analytics and visualization alternatives that can be employed by different users according to their needs. For instance, users may have the need to evaluate a transportation policy before applying it or exploring potential future scenarios and corresponding policies/interventions.

This UC will identify several users' needs to improve decision making in scenarios such as:

- (i) Introduction of zero-emission zones for freight transportation in urban areas;
- (ii) Introduction of different level of autonomous vehicles in the transportation network;
- (iii) Investigation of the relationship between different levels of remote work, impact on transportation and residential choice.

The present UC will explore the possible improvement for providing data-driven estimation, introduction of data pipelines for the calibration of models and adaptive visualisations that fit the purpose and needs of the various users and actors of the transport platforms. This objective will cope with the different data availability and granularity by combining with smart and adaptive data usages and re-usage by delivering customised policy analysis and decision support. Additionally, each case will rely on receiving data from live sources.

The following ExtremeXP artefacts will be applied:

- (i) Interactive and adaptive visualization of extreme data, for GIS visualizations of both reported (GPS traces) and predicted (simulation results) data of passenger or freight traffic flows.
- (ii) User-driven Optimization of Complex Analytics, for estimation and application of econometric and other sub-modules on the datasets, generating the necessary results of the simulation platform along with latent classification of users, attitudes, and travel behaviour.
- (iii) User Feedback via Serious Games, to perform stated preference experiment data collection using serious gaming techniques to ease the burden of reporting on the respondent.
- (iv) Transparent and Interactive Decision Making, for scalable explanation of results based on customer/audience (from experienced modellers to non-technical policy makers), including the use of AR techniques.

3.5.2 Experiment definition

UC 4 aims at improving the decision-making process for transport planners and enabling better insights for forecasts into the future. The approach is twofold, firstly through the provision of an improved annotated dataset containing the habits and travel patterns of a given study area, and secondly through the utilization of a multilayer modelling approach. The outcome of the modelling process serves as the testbed for running various scenarios and evaluating the results through meaningful statistics and visualisations.

When it comes to data collection, this is done through Moby's mobile application, MobyApp (see Figure 3.5.2-1, MobyApp's Architecture), which allows end-users to voluntarily share their daily trips and preferences, along with their precise locations throughout a period of time (typically 2-4 weeks). Their trips are computed by MobyApp's ML engine (i.e., Pythia) and then presented to the users for verification. Effectively, once users have verified their trips, what is made available is a thorough mapping of their travel patterns for the given study area and time period.

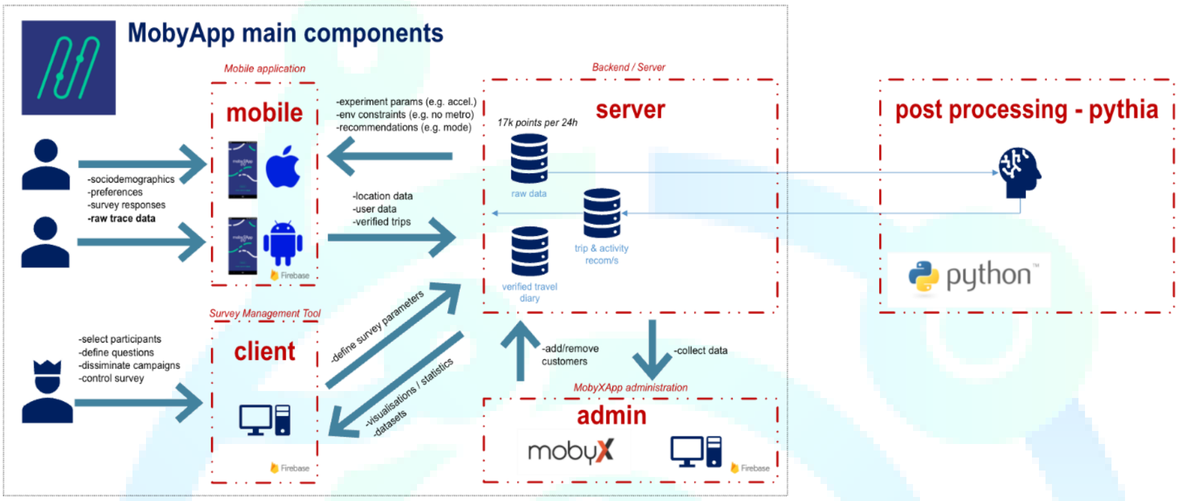


Figure 3.5.2-1: MobyApp architecture

During data collection (i.e., MobyApp) the main actors are the user (or passenger), who is using the app, and secondarily the organisation/planner/researcher who is collecting the data. The main goals are, thus, mainly focusing on how to make the data collection process easier and more effective for the user, thus, resulting to a more accurate and detailed dataset in the end.

To that end, the effort is on the ***minimization of manual data entry***, by producing more accurate prediction of the user's activity, both for stationary and moving segments. Once identified by Pythia, all segments that are presented to the user need to be as close to reality as possible, to enable the user to have limited tampering with their trip diary, so as to be enticed to only fine-tune some missing details, if any.

Additionally, ***efficient data collection with reduced battery consumption, and the need for adaptive sensing***, will allow for less dissatisfaction by the users, who may often see their mobile phone battery depleted due to the continuous stream of data from their phone. The current granularity is set to every 5 seconds, so there is a question of how much data is enough data to produce accurate predictions for their trip diaries.

Real-time information of current/recent trips and activity is an essential milestone for allowing users to view their trips while on the go and will in turn enable the addition of a multitude of features. At the moment, location data for each user is batched into chunks and processed at the end of the day.

Lastly, ***predictive accuracy that increases per user on each subsequent use of the app*** will enable the user to receive more tailored recommendation about their whereabouts and their selected modes of transport. Patterns across the entire userbase can be used for a specific individual, harnessing the power of machine learning, and calibrating per individual. For example, a user may have multiple verified trips by bike, therefore the produced predictions thereon may be deliberately skewed towards predicting more trips by bike than on other users.

Co-funded by
the European Union

Regarding variability points for the data collection by MobyApp, t*here is* ***currently one algorithmic logic for converting GPS traces into meaningful trip diaries****.* This means that regardless of the application area of each survey, the governing logic is the same, whichever thresholds or assumptions used may not perform equally across regions. For example, bike may be overestimated in countries like Cyprus where biking is not preferred, while it may be underestimated in countries like the Netherlands, where cycling is much more common. These methods are applied to different geographic locations; thus, users exhibit significantly different travel behavior depending on the location where the survey is taking place. Currently, any existing variability points are resolved during the design of the algorithm, where a set of possible parameters is tried out per case and decided upon per case. For instance, when using Density-based spatial clustering of applications with noise (DBSCAN) per each user per each day trip, the parameters are selected depending on the produced silhouette score, after clustering has taken place using a range of different parameters. Another consideration would be to develop a close-to-real-time estimation of trips, instead of using batches of data. For example, the estimation by Pythia could only consider the past few hours of tracking or could be triggered only when a user stands idle. A real-time estimation of trips is also a possible approach, and it would require a drastic re-vamp of the ML module, to perhaps transition to a stop-detection based algorithm or other real-time methods.

Once the data from MobyApp is transformed into meaningful trips by applying the above-described process, it is passed into HARMONY MS to be used for modelling purposes. Harmony-MS has been the outcome of the HARMONY H2020 project and constitutes a TRL 7 product which aims at satisfying different objectives corresponding to the different goals of actors of transport modelling sphere. It is a tool which aims at streamlining the modelling process for the different individuals included in the modelling circle.

There have been 3 main actors defined as also seen in Figure 3.5.2-2 (Harmony MS architecture)

- **Transport planners** are the users who are interacting via the graphical user interface of the HMS and perform comparative analysis of transportation cases, interventions, and policies.
- **Transport modelers** are the users specifying the models of the available transportation cases, interventions, and policies that transport planners can compare. They are also interacting with the integrators to ensure the validity of the comparisons.
- **Integrators (Components developers)** ensure that the different platform components (transport models and simulators) at different levels (strategic, tactical, and operational) are correctly integrated to the HMS. The transportation cases, interventions, and policies that can be compared via the HMS, and hence the capabilities of the platform available to transport modelers and ultimately transport planners, depend on the components that are integrated to it. Integrators are the actors that are extending the capabilities of the HMS.

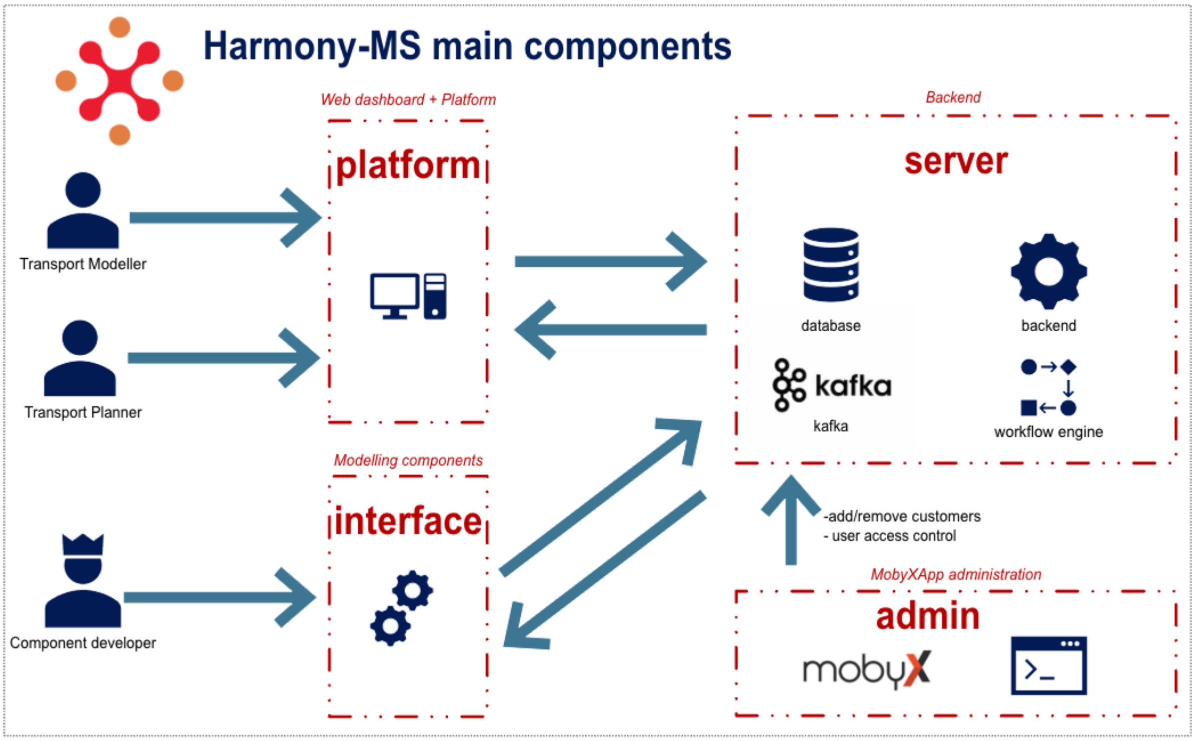


Figure 3.5.2-2: HarmonyMS architecture

- Main Extension points for Harmony MS can be categorised into three major groups: **Transferability of models and data** means that whenever a city needs to test an intervention, but has no data, via Harmony MS they can use data other cities to extrapolate the impact of this intervention to the city in question.
- **Adaptiveness** stands for enabling the flexibility to select the needed tools, methods, and outputs. A city should be able to select the best model out of a pool of models that have the same purpose. In addition, they also need to select the most appropriate set of KPIs and related visualisations for the presentation of results.
- By **streamlining**, Harmony MS aims to provide a seamless environment for the users by adaptive error messaging. User-related actions and personalisation are also options to consider since different cases come with different requirements.

Variability points for Harmony MS refer to the different datasets and different data needs across cases, as matching data needs to the available data is sometimes a significantly hard task.
Different algorithms should be able to be selected, depending on the available models and the given scope.

A set of possible visualizations must be available, depending on the output of the models and the needs of the user.

There are various internal transaction surfaces, where endpoints can be created for facilitating third party access to data streams and internal endpoints. The main steps of the process are the following:

1. Pre-processing
   a. The MobyApp collects raw GPS data and feeds it to the backend in batches (not in real-time currently).

b. Then an ML module (I.e., Pythia) utilises machine learning techniques to identify segments at which the user is stationary or moving.

2. Analytics

a. For the moving segments: identification of the mode of transport and the type of public transport line is being used (via GTFS).

b. For the stationary segments: identification of the type of activity (via POIs) or whether the user was at a Public Transport stop (GTFS)

c. For the modelling tools on Harmony-MS, selection from pre-existing models, creation of models according to available data, pairing between inputs and outputs of modelling modules.

3. Visualization workflows

a. On the mobile app: visualizations on map and on timeline (this currently built-in on the mobile applications).

b. On the client web app of MobyApp: KPIs, metrics, statistics, maps.

c. On Harmony MS: a selection of more advanced KPIs.

3.5.3 Exploited datasets & data acquisition methodology

Datasets that are collected by MobyApp require caution since user locations are considered sensitive data according to GDPR. The reason being that in some cases, a GPS trace can be uniquely pinpointed to a specific individual. For Harmony MS, the datasets are primarily secondary data used for modelling purposes.

For the ExtremeXP project, the available data coming from the data collection pipeline are:

a. Collected raw GPS data that are fed to the backend in batches (currently not implemented in real-time)

b. Annotation from users from their verified trips

On the Harmony MS front the datasets are:

a. Sample models from past case studies.

b. Pipelines and related association between datasets and models.

c. Datasets used as inputs to the different models..

d. Datasets - expected outputs of the different models

The steps of the data acquisition process is summarised in the illustration below (Figure 3.5.3-1):

Collect data	The MobyApp collects raw GPS data and feeds it to the backend in batches (not in real-time currently)
A module utilizes machine learning techniques to identify segments for each user	For the moving segments: we identify the mode of transport and which public transport line is being used (via cross-checking with GTFS1 data) For the stationary segments: we identify the type of activity (via cross-checking with OSM2 POIs3)
User action	The user verifies their trips
On the mobile app	visualisation on map and on timeline
On the client web app	KPIs, metrics, statistics, maps
Final dataset from MobyApp	All travel diaries of all users
Input for Harmony MS	travel diaries, open datasets, existing models
User action	Connect model output to model inputs Configure models
Run scenarios	
Visualizing results	

Figure 3.5.3-1: Data acquisition process

3.5.4 Experiment infrastructure & testbed

Both products of MobyX are hosted on AWS, using EC2 cloud services, and a docker-enabled, microservice infrastructure. There is a production server running, which is customer-facing, and also a development server that could be utilised for the purposes of ExtremeXP. Communication is end-to-end encrypted, and access could potentially be granted to partners on the development server for exploring the different alterations that are explored within the present UC.

3.5.5 Benefits introduced by ExtremeXP

The ExtremeXP framework will introduce the following benefits for the transportation sector:

- Quantify the trade-off regarding how much information the end-user can pass into the app and how much data is at minimum needed in order to produce meaningful predictions for each user’s travel diary (KPI4-1, KPI4-2).
- Enhance user participation through producing better, tailored recommendations, and potentially incorporate Serious Games in the data collection process (KPI4-1).
- Bolster adaptive visualization of extreme data, both during data collection (MobyApp) and also for simulation results to enable a clearer and more real-time overview of the ongoing survey, as well as a preview of intermediate results (KPI4-3, KPI4-4).
- Employ User-driven Optimization of Complex Analytics, for estimation and application of econometric and other sub-modules on the datasets, generating the necessary results of the

simulation platform along with latent classification of users, attitudes, and travel behaviour (KPI4-4, KPI4-5).

- Contribute to interactive decision making, in relation to data availability and through incorporating relevant KPIs that fit the needs of the stakeholder/user of the modelling suite, through examining diverse scenarios (zero-emission zones, remote work, autonomous vehicles, etc.)(KPI4-5).

The following table lists the UC KPIs:

KPI Id	Name	Target
KPI4-1	Reduce manual entry in the travel diary of the user by better identifying modes, trips, and activities	30% less unresolved trips
KPI4-2	Reduce battery consumption through implementing adaptive sensing for collecting only necessary data	30% battery consumption reduction
KPI4-3	Visualization modules based on live data feeds	> 3
KPI4-4	MS platform usability reported by customers/stakeholders (self-reported satisfaction)	> 80%
KPI4-5	Alternative simulations and analytics processes for each policy or scenario	> 5

Table 3.5.5-1: List of KPIs for UC 4

3.5.6 Demonstrator technical requirements

In the following table, we list the requirements to be expected on the demonstrator of the fourth UC. We identify each requirement with a specific identifier, propose a name and a description, and prioritize its execution using the MoSCoW model.

Req. Id	Name	Description	Priority (MoSCoW)
UC4-1	***Adaptive Sensing***	*Efficient data collection with reduced battery consumption*	Must
UC4-2	***Real-time predictions***	*Adapt ML module to predict information of current/recent trips and activity in real-time*	Should
UC4-3	***Minimize manual data entry***	*Produce more accurate predictions of the user's activity both for stationary and moving segments*	Must
UC4-4	***Personalised data collection***	*Predictive accuracy that increases per user on each subsequent use of the app*	Could
UC4-5	***Transferable models***	*Make HMS models reusable across different regions*	Should
UC4-6	***Adaptive modelling***	*Select tools, methods and outputs depending on needs/requirements*	Should
UC4-7	***Quantify data trade-offs***	*Determine what amount of data from the end-user is required at minimum before predictions become less accurate*	Must

UC4-8	***Enable adaptive visualizations of extreme data***	*Both during data collections (MobyApp) and also for simulation results (Harmony MS)*	Must
UC4-9	***Computation off-load***	*Cloud infrastructure, specifically AWS EC2, are to provide computing resources on-demand.*	Should

Table 3.5.6-1: List of requirements for the UC 4

3.6 UC5: Failure prevention for manufacturing industry (IDEKO)

3.6.1 Context description and objectives

Maximizing machine uptime in an industrial plant is crucial for attaining high overall system availability and maintaining competitiveness. The industry 4.0 paradigm has introduced novel techniques and methods to achieve this objective, including predictive maintenance, lifetime assessment, and critical machine element diagnosis. Manufacturing industries are motivated to prevent production line stoppages and are willing to explore and invest in innovative solutions that minimize or eliminate such stoppages. Advanced machines come equipped with highly accurate embedded sensors that generate crucial "health" data, which can be utilized to diagnose and fulfil the afore-mentioned goals.

The key is to have precise information and avoid erroneous warnings that lead to a decrease in the production capacity. In this context, this UC objective is to develop and maintain a dashboard for the technical service team that will allow them to take decisions over incipient and imminent failures of the critical components of the machine, such as heads, balls screws or axis.

Figure 3.6.1-1: Machine components: Heads balls screws and axis

This UC plans to create a simulated fleet of machines that will incorporate self-diagnostic cycles. Using data gathered from the testbed, IDEKO will initiate failure scenarios to test the accuracy of the implemented solution. Additionally, IDEKO will simulate changes in machine behaviour scenarios, such as sudden variations from normal operation (e.g., due to altered manufactured part references or replacement of a complete component). This is aimed at detecting any AI model deviations that require retraining based on simulated data.

Simulated data refers to the process of using real acquired data from a physical machine and feeding it as input to a script that generates data outputs mimicking the behaviour of an actual machine. Thus, we can control the execution of the data, for example to stop whenever we want, etc. While the term "simulated" is used to describe this process, it is important to note that the data being generated are based on real-world data that have been collected from a functioning machine from the IDEKO’s Digital Grinding Innovation Hub. The purpose of this simulation is to replicate the machine's behaviour and generate data outputs as if it were a real operational machine.

Figure 3.6.1-2: Digital Griding Innovation hub

To acquire the data, an edge device known as the Savvy Smart Box is installed inside the electric board of every machine, within the shop floor environment. The primary function of the Smart Box is to collect data from various sources, and it can be readily configured to obtain data, by configuring different connectors for distinct machine Computer Numerical Control (CNC) systems.

Figure 3.6.1.3 shows a simplified architecture of the Smart Box, which comprises of three key components:

1. Computer Numerical Control (CNC) libraries that facilitate the reading of diverse machine controls,
2. interoperability modules that provide access to machine data, and
3. Docker containers for deploying solutions.

The containers housed within the box provide an excellent option for deploying various ExtremeXP modules of the project, or launching connectors that connect with project modules. The box comes equipped with over 50 pre-established containers as standard, but introducing new containers of different types into the box is feasible if the project demands it and can be easily achieved.

Furthermore, the Smart Box also enables the transmission of machine data to a private company cloud storage.

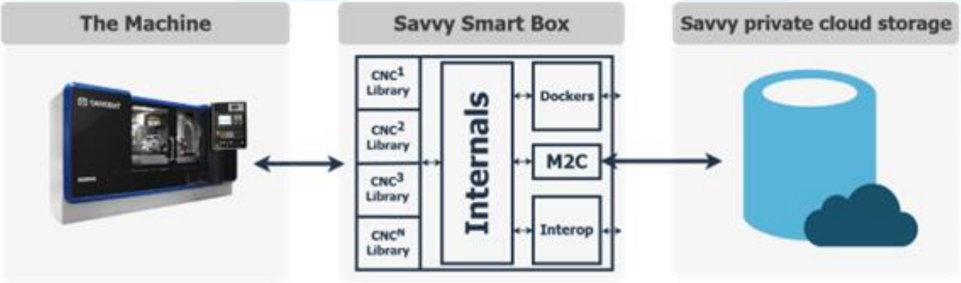


Figure 3.6.1-3: High level Savvy Smart Box architecture

For this UC, IDEKO will use the complete ExtremeXP Experimentation Engine with the spotlight on:

- The **Analysis-aware Data Integration** module to make the proper user data selection based on the stakeholder needs.
- The **User-driven Optimization of Complex Analysis**, and particularly, the **User Profiling** submodule.
- The **User-driven AutoML** and the algorithm and model selection features.

- The **Transparent & Interactive Decision-Making** module to bring the users the rationale behind the systems decisions.

3.6.2 Experiment definition

To define an experiment, it is crucial to consider several factors, such as the actors covering the domain modeling, identifying the variability points, and eliciting user intents or visualization needs.

With respect to the main **actors** for the domain modellings, we have identified the following:

- the *machine operator*, who is responsible for the proper functioning of the machine, including programming the machining process, monitoring the piece, and possessing sufficient knowledge to assess the correctness of a given process. It is important to point out that the operator is the final user of the development of the data analyst.
- the *data analyst,* who is the individual responsible for accurately modeling the algorithm and works closely with the machine operator.

They collaborate to ensure that the algorithm is properly designed and optimized for the specific machine and process. The data analyst utilizes their expertise to analyze the data generated by the machine and make informed decisions regarding the algorithm's configuration and parameters. By working synergistically with the machine operator, they can fine-tune the algorithm and ensure that it effectively addresses the specific needs and challenges of the machine operations.
The operator can visually observe through applications that the machining process did not function properly, and then contacts the data analyst, who downloads high-frequency data and analyzes potential causes. The data analyst communicates the findings to the operator, who can either adjust the machine parameters to prevent recurrence or the data analyst modifies their anomaly detection algorithm to avoid detecting those false positives.

Due to machines having diverse configurations, there is a high variability of data in the machine tool industry, therefore setting the **variability points** is a manual process that is carried out for each case or each analysis. The ML models that are chosen and adjusted depend on the type of problem that has to be addressed, whether it is a supervised or unsupervised problem, and a classification problem, regression, or clustering. The parameters of each of these models can be fixed manually or define a hyperparameter fit to find the combination of parameter values that best fit the data.

Regarding the **experiment** itself, data will be captured from machines located in IDEKO's Digital Grinding Innovation Hub, encompassing both high and low-frequency data. It is important to note that data capture will be performed for different machine configurations, and additionally, the captured data will undergo preprocessing using Python data analysis libraries. The data generated by the machine will serve as input to the ExtremeXP framework.

Since IDEKO currently lacks an automated mode for detecting anomalies that could lead to future breakdowns of specific machine components, the experiment for this UC will be to develop different AI models for anomaly detection for different machine configurations and test its performance. Failure scenarios will be triggered to test the precision of the implemented solution and machine behavior scenario changes will be also provoked to detect ML model deviations requiring retraining. The performance will be supervised and validated by both the machine operator and the data analyst and used as user feedback to boost the continuous improvement of the ExtremeXP framework knowledge.

Regarding **user intents**, the user itself is the machine operator, whose intentions are as follows:

- Rapid notification upon detecting a potential machine anomaly on critical components such as heads, balls screws or axis.
- Visualization of time series data with a high volume of data points in a clear, fast, and smooth manner.
- Visualization of the historical progression of fingerprint results along with their corresponding ranges.

Finally, concerning the **visualization**, it is highly dependent on the specific data being presented, and it is common to represent time series data such as vibrations, power readings, and more. These time series often consist of a large number of data points. Therefore, it is crucial to utilize a rendering library that can effectively handle and display high volumes of data. The chosen library should possess the capability to efficiently render graphs with large datasets to ensure accurate and smooth visual representation.

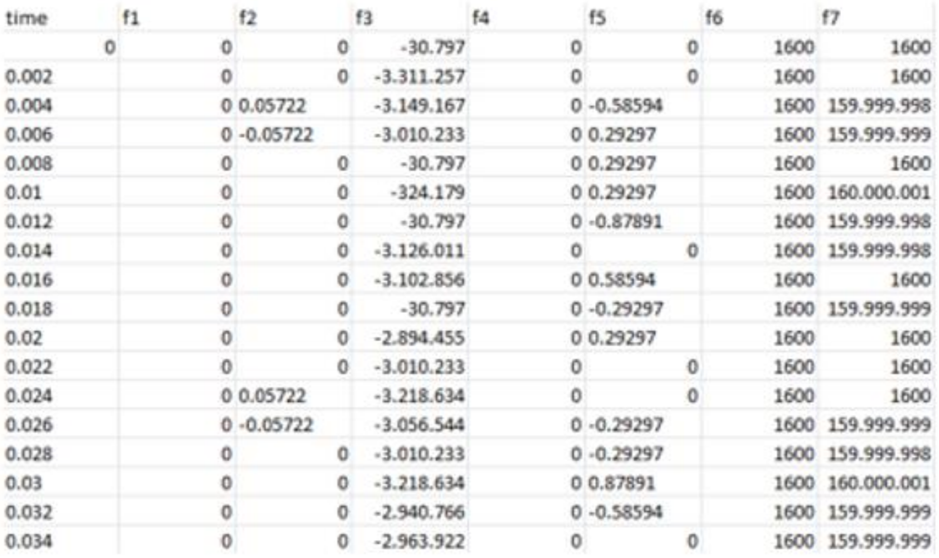
3.6.3 Exploited datasets & data acquisition methodology

Machines generate two types of data:

- high-frequency data resulting from self-diagnostic cycles and providing arbitrary metrics on several component on the machines, and
- low-frequency real-time data, providing information about the machines' performance.

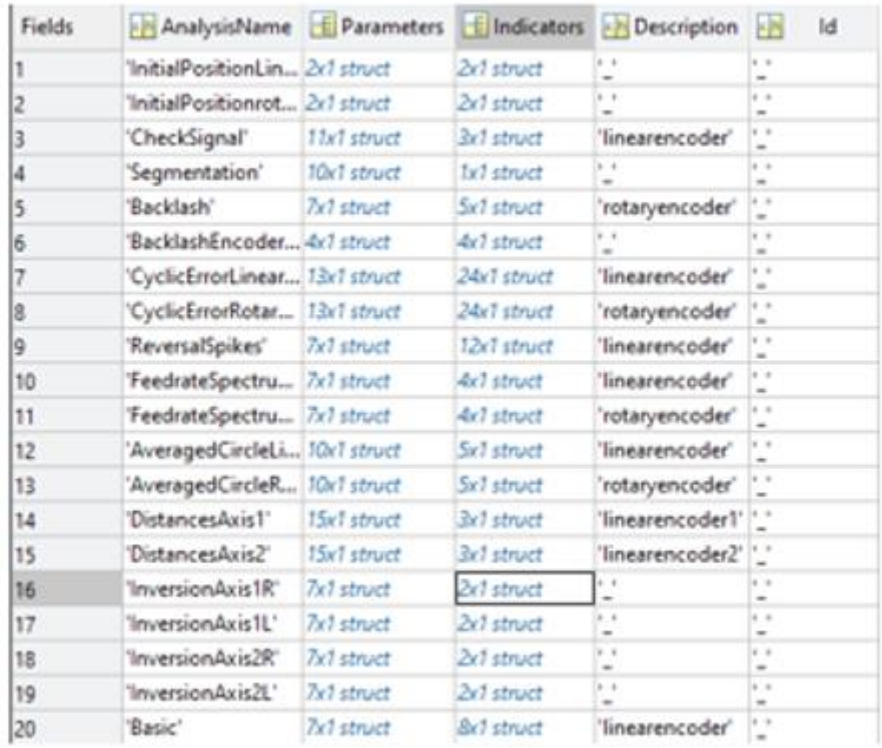
On the one hand, high frequency data is generated periodically by the machine's self-diagnostic cycles. This data provides a detailed insight into the machine's performance and information about the state of health of the machine and its components. Snapshots of high-frequency data are presented in Figures 3.6.3-1 and 3.6.3-2.

The data is collected at a rapid pace when the machine performs a series of predefined, non-destructive movements, without machining, and it is often used to identify any issues or faults in the machine. During these movements, high-frequency signals are captured until the entire process is completed. The aim of the analysis of these data is usually to know the current state of the machines and detect anomalies or changes of scenario in the behavior of these machines by applying, for example, different statistical or artificial intelligence techniques.



time	f1	f2	f3	f4	f5	f6	f7
0	0	0	-30.797	0	0	1600	1600
0.002	0	0	-3.311.257	0	0	1600	1600
0.004	0	0.05722	-3.149.167	0	-0.58594	1600	159.999.998
0.006	0	-0.05722	-3.010.233	0	0.29297	1600	159.999.999
0.008	0	0	-30.797	0	0.29297	1600	1600
0.01	0	0	-324.179	0	0.29297	1600	160.000.001
0.012	0	0	-30.797	0	-0.87891	1600	159.999.998
0.014	0	0	-3.126.011	0	0	1600	159.999.998
0.016	0	0	-3.102.856	0	0.58594	1600	1600
0.018	0	0	-30.797	0	-0.29297	1600	159.999.999
0.02	0	0	-2.894.455	0	0.29297	1600	1600
0.022	0	0	-3.010.233	0	0	1600	1600
0.024	0	0.05722	-3.218.634	0	0	1600	1600
0.026	0	-0.05722	-3.056.544	0	-0.29297	1600	159.999.999
0.028	0	0	-3.010.233	0	-0.29297	1600	159.999.998
0.03	0	0	-3.218.634	0	0.87891	1600	160.000.001
0.032	0	0	-2.940.766	0	-0.58594	1600	159.999.999
0.034	0	0	-2.963.922	0	0	1600	159.999.999

Figure 3.6.3-1: High frequency data example (csv format)



Fields	AnalysisName	Parameters	Indicators	Description	Id
1	'InitialPositionLin...	2x1 struct	2x1 struct	'_'	'_'
2	'InitialPositionrot...	2x1 struct	2x1 struct	'_'	'_'
3	'CheckSignal'	11x1 struct	3x1 struct	'linearencoder'	'_'
4	'Segmentation'	10x1 struct	1x1 struct	'_'	'_'
5	'Backlash'	7x1 struct	5x1 struct	'rotaryencoder'	'_'
6	'BacklashEncoder...	4x1 struct	4x1 struct	'_'	'_'
7	'CyclicErrorLinear...	13x1 struct	24x1 struct	'linearencoder'	'_'
8	'CyclicErrorRotar...	13x1 struct	24x1 struct	'rotaryencoder'	'_'
9	'ReversalSpikes'	7x1 struct	12x1 struct	'linearencoder'	'_'
10	'FeedrateSpectru...	7x1 struct	4x1 struct	'linearencoder'	'_'
11	'FeedrateSpectru...	7x1 struct	4x1 struct	'rotaryencoder'	'_'
12	'AveragedCircleLi...	10x1 struct	5x1 struct	'linearencoder'	'_'
13	'AveragedCircleR...	10x1 struct	5x1 struct	'rotaryencoder'	'_'
14	'DistancesAxis1'	15x1 struct	3x1 struct	'linearencoder1'	'_'
15	'DistancesAxis2'	15x1 struct	3x1 struct	'linearencoder2'	'_'
16	'InversionAxis1R'	7x1 struct	2x1 struct	'_'	'_'
17	'InversionAxis1L'	7x1 struct	2x1 struct	'_'	'_'
18	'InversionAxis2R'	7x1 struct	2x1 struct	'_'	'_'
19	'InversionAxis2L'	7x1 struct	2x1 struct	'_'	'_'
20	'Basic'	7x1 struct	8x1 struct	'linearencoder'	'_'

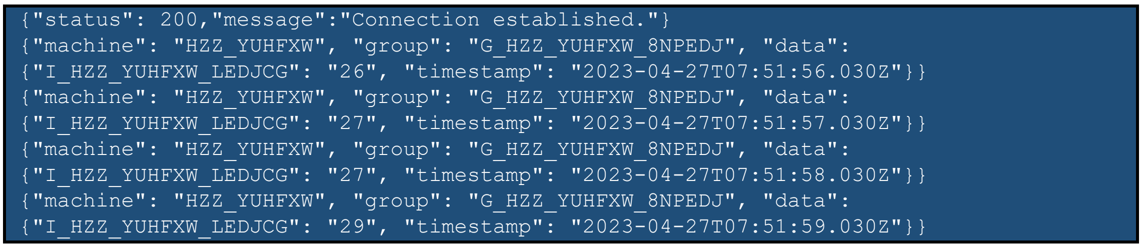
Figure 3.6.3-2: High frequency data example (mat format)

On the other hand, low-frequency real-time data are continuously generated by the machine during normal operation. Data are collected at a slower pace, resulting in a steady stream of information regarding the machine's performance. This allows, on the one hand, to monitor times the status and operation of the machines at all times, as well as the production process, keeping the machine under constant control at any given moment. On the other hand, these data allow operators to monitor their status and adjust, as necessary.

In this way, it is possible to automatically monitor the behavior of the machines and their components, such as spindles and axes. Historical machine data allows for different analyses to be carried out by applying the necessary techniques in each one of them, such as machine downtime analysis, process analysis or vibration analysis. The results obtained can be used to know the uptime and the intensity of machine use, as well as to define normal behavior in different machining processes.

As stated on the following figure, low-frequency real-time data is a constant stream textual data in JSON format following the specified structure:

- machine: Machine identifier
- group: Indicator group identifier
- data: Machine indicators data
- timestamp: Data timestamp



```
{"status": 200,"message":"Connection established."}
{"machine": "HZZ_YUHFXW", "group": "G_HZZ_YUHFXW_8NPEDJ", "data":
{"I_HZZ_YUHFXW_LEDJCG": "26", "timestamp": "2023-04-27T07:51:56.030Z"}}
{"machine": "HZZ_YUHFXW", "group": "G_HZZ_YUHFXW_8NPEDJ", "data":
{"I_HZZ_YUHFXW_LEDJCG": "27", "timestamp": "2023-04-27T07:51:57.030Z"}}
{"machine": "HZZ_YUHFXW", "group": "G_HZZ_YUHFXW_8NPEDJ", "data":
{"I_HZZ_YUHFXW_LEDJCG": "27", "timestamp": "2023-04-27T07:51:58.030Z"}}
{"machine": "HZZ_YUHFXW", "group": "G_HZZ_YUHFXW_8NPEDJ", "data":
{"I_HZZ_YUHFXW_LEDJCG": "29", "timestamp": "2023-04-27T07:51:59.030Z"}}
```

Figure 3.6.3-3: Low frequency data stream example

A snapshot of low-frequency data is presented in Figure 3.6.3-3.
IDEKO's cloud platform provides a section for indicator reading verification, allowing users to access a list of the latest values of these indicators, as well as their historical evolution. This feature enables users to assess the quality of the data before consuming the APIs by, for instance, checking if the speeds or temperatures indicators include logical values.

Both types of data are important for ensuring that machines operate efficiently and effectively: high-frequency data for a detailed view of the machine's performance and low-frequency data to have information of the machine's performance at all times.

These data can be accessed using two different APIs, either through a secure cloud API that requires authentication with a username and password to ensures that only authorized personnel can access the data, or an unsecured local API that does not have internet access.

The restrictions for both types are as follows:

Cloud API	Local API
Auth	No auth
Does not work on a browser	Works on a browser
Accessible through internet	Accessible through local network only
With historical data	Without historical data
Real time streaming	Real time streaming

Table 3.6.3-1: Cloud API vs Local API

Although both APIs offer the possibility of viewing and streaming data in real time, the main difference between local API and cloud API is that there is no historical data query in the first one. Unlike the cloud, storage is limited and not scalable, so it is not possible to have persistent data and store historical data. In contrast, the cloud is scalable and has a very large storage capacity.

Although the local API currently has restricted access due to its localized nature, it is possible to open it up and create an exclusive access for the project. However, this decision needs to be made with careful consideration of the potential security risks involved. Opening up the local API would essentially provide an entry point into the system. Therefore, before any such decision is made, it is essential to evaluate the risks and implement appropriate security measures to ensure that the system remains secure.

In summary and as can be seen in Figure 3.6.3-4, we have two main data sources, high-frequency data, and low-frequency data, both generated by the machine. High-frequency data can be in ‘. mat’ and ‘.csv’ files, while low-frequency data consists of textual data in JSON format served through APIs. Additionally, from the high-frequency data and by applying certain mathematical algorithms in Python, we generate fingerprints, which are internally referred to as AR files, saved in .mat format.

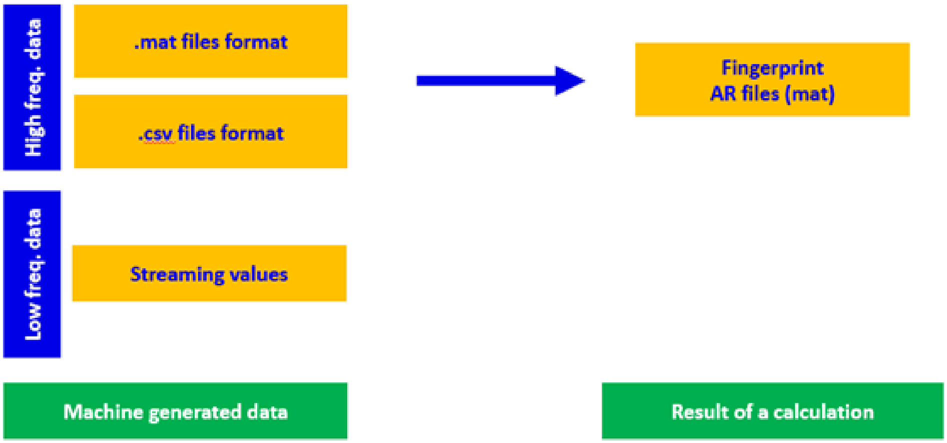


Figure 3.6.3-4: IDEKOS’s data sources

3.6.4 Experiment infrastructure & testbed

Since the requirements for each of our customers are different, we have an infrastructure that can be adapted to most of our potential customers. This infrastructure is built around an IoT device, Savvy SmartBox. The device mentioned in the first section.

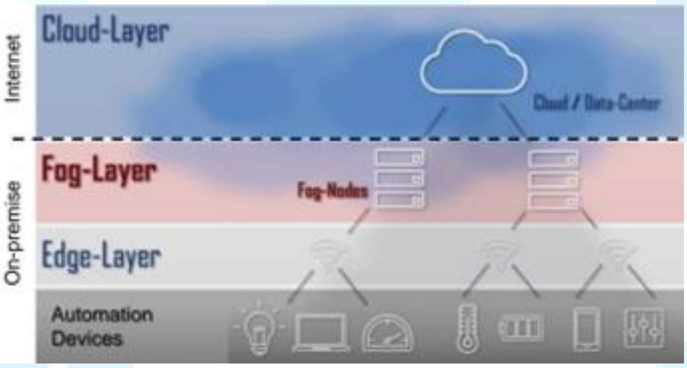
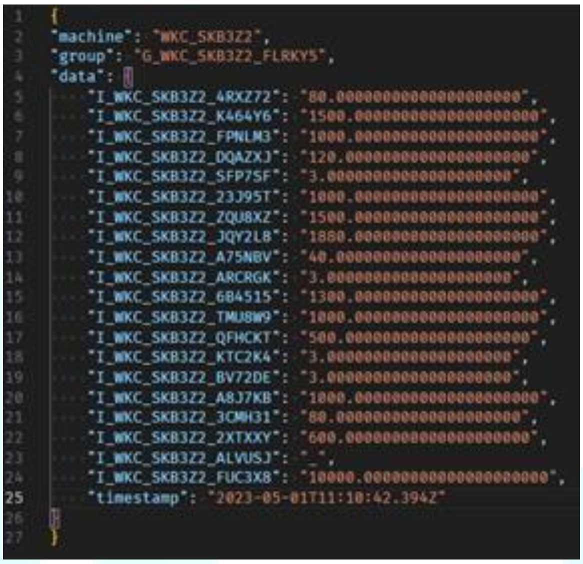


Figure 3.6.4-1: Cloud. Fog, edge

As stated on Figure 3.6.4-1, three types of infrastructure can be built for the consumption of the data obtained by the machine:

3.6.4.1 Edge infrastructure

The collected data is stored in the Savvy Smart Box. The processing of this data can be done directly on the device. For this purpose, the device has interoperability mechanisms to provide the data such as Modbus TCP, MTConnect, and Rest Streaming API. In this case, we will use the Rest Streaming API. This API allows to continuously obtain the indicators that are configured in it. This mechanism returns a value of the set of indicators every second with the current value. This data is given in JSON format.



```
{
"machine": "WKC_SKB3Z2",
"group": "G_WKC_SKB3Z2_FLRKY5",
"data": [
    "I_WKC_SKB3Z2_4RXZ72": "80.00000000000000000000",
    "I_WKC_SKB3Z2_K464Y6": "1500.00000000000000000000",
    "I_WKC_SKB3Z2_FPNLM3": "1000.00000000000000000000",
    "I_WKC_SKB3Z2_DQAZXJ": "120.00000000000000000000",
    "I_WKC_SKB3Z2_SFP7SF": "3.00000000000000000000",
    "I_WKC_SKB3Z2_23J95T": "1000.00000000000000000000",
    "I_WKC_SKB3Z2_ZQU8XZ": "1500.00000000000000000000",
    "I_WKC_SKB3Z2_JQY2L8": "1880.00000000000000000000",
    "I_WKC_SKB3Z2_A75NBV": "40.00000000000000000000",
    "I_WKC_SKB3Z2_ARCRGK": "3.00000000000000000000",
    "I_WKC_SKB3Z2_6B4515": "1300.00000000000000000000",
    "I_WKC_SKB3Z2_TMU8W9": "1000.00000000000000000000",
    "I_WKC_SKB3Z2_QFHCKT": "500.00000000000000000000",
    "I_WKC_SKB3Z2_KTC2K4": "3.00000000000000000000",
    "I_WKC_SKB3Z2_BV72DE": "3.00000000000000000000",
    "I_WKC_SKB3Z2_A8J7KB": "1000.00000000000000000000",
    "I_WKC_SKB3Z2_3CMH31": "80.00000000000000000000",
    "I_WKC_SKB3Z2_2XTXXY": "600.00000000000000000000",
    "I_WKC_SKB3Z2_ALVUSJ": "_",
    "I_WKC_SKB3Z2_FUC3X8": "10000.00000000000000000000",
    "timestamp": "2023-05-01T11:10:42.394Z"
]
}
```

Figure 3.6.4-2: API response example

Once the data has been retrieved via the Rest Streaming API, multiple Docker containers can be launched on the device itself, either in development or production, for processing. This is where any clustering or transformation could be applied, if necessary. And it would also allow the development of any utility that can be deployed in a Docker container. Although it should be borne in mind that the capabilities of the utilities to be developed will be limited by the capabilities of the device in question, CPU, memory, etc., it is important to point out that this solution can only be accessed via a LAN connection to the device and it features a 1.83GHz Quad Core CPU, 8GB RAM, and a 240GB SSD.

3.6.4.2 Fog infrastructure

If the capabilities of the IoT devices themselves are not enough, or we need to deal with data from multiple devices. We have two Linux-based servers at IDEKO; each one being an independent work environment, one for development utilities and the other for production utilities. They have greater capabilities than the edge device but are still not as powerful as using cloud tools. However, they are still at our premises. The development server features an Intel(R) Xeon(R) CPU E5-2640 0 @ 2.50GHz, 8GB RAM and 110GB SSD and the production server features an Intel(R) Xeon(R) Gold 5215 CPU @ 2.50GHz, 12GB RAM and 110GB SSD.

For this solution, the device and the servers must be in the same network and to be able to access them you must connect to this network.

3.6.4.3 Cloud infrastructure

When the capacities required by the utility are very large or there is no restriction related to data governance, the cloud structure is used. Tools such as AWS EC2 instances can be used to process these data, making this solution accessible from anywhere with an internet connection.

3.6.4.4 Hybrid infrastructure

It should be noted that the solutions mentioned are not exclusive. A hybrid infrastructure can be carried out. For example, the edge device gets and pre-processes the data. Filling in the missing data, eliminating duplicates, or filtering the information (obtaining the data only when a condition is met). The servers that are in the fog infrastructure collect the pre-processed data from multiple devices and group it, transform it, reduce it to be able to send it to the cloud. Where, finally, multiple ML mechanisms could be applied to detect many anomalies such as tool breakage, wear in bearings, etc.

3.6.5 Benefits introduced by ExtremeXP

The ExtremeXP framework will introduce the following benefits for the manufacturing sector:

- **Increased Automation:** The system can automate routine tasks and optimize business processes, reducing the need for manual intervention and improving overall efficiency.
- **Enhanced Predictive Maintenance**: The system can use AI models to monitor equipment performance and detect potential issues before they occur, reducing downtime and maintenance costs.
- **Improved Scalability:** The ability to deploy AI models can help businesses scale their operations more efficiently, improving overall business agility and competitiveness.

The following table lists the retained UC KPIs:

KPI Id	Name	Before	After
KPI5-1	Unpredicted downtimes over a simulated production line	0.01%	0.001%
KPI5-2	Imminent failures detection over a simulated production line	87%	90%
KPI5-3	Average time from anomaly occurrence to detection	1h	5min
KPI5-4	Average time from anomaly detection to response compared to traditional human techniques	30min	10min
KPI5-5	Number of data points while ensuring a smooth and manipulable visualization for the user	30000	50000

Table 3.6.5-1: List of KPIs for the UC 5

3.6.6 Demonstrator technical requirements

In the following table, we list the requirements for the demonstrator of the fifth UC. We identify each requirement with a specific identifier, propose a name and a description, and prioritise its execution using the MoSCoW model.

Req. Id	Name	Description	Priority (MoSCoW)
UC5-1	***Anomaly detection algorithm selection***	*Machine operator must select the most suitable model for anomaly detection.*	Must
UC5-2	***IA model result comparison***	*The data analyst should compare the results obtained by the AI model with the feedback of the machine operator in order to measure the detection capacity of the algorithm.*	Should
UC5-3	***Algorithm effectiveness***	*The machine operator must be able to evaluate whether the result is correct or not.*	Must
UC5-4	***Fleet simulation***	*The demonstrator of UC5 permits the simulation of a fleet of machines to be tested.*	Should
UC5-5	***CNC library***	*The UC5 demonstrator can interface with CNC*	Should

	support	*library in Smart Boxes to read the values from the machines controls (sensors).*	
UC5-6	***Variability point***	*In UC5 demonstrator, the variability points regard the selection of the AI models for prediction and its parameters.*	Must
UC5-7	***Python library***	*The data processing environment supports compatibility with python data analysis libraries to conduct data processing.*	Must
UC5-8	***User Interface***	*In UC5 demonstrator, the user is informed via rapid notification of an anomaly on critical components.*	Should
UC5-9	***Time series visualisation***	*The user can easily explore time series via adequate visualization method.*	Must
UC5-10	***Fingerprint visualisation***	*The user of the demonstrator of UC5 can inspect the historical progression of results' fingerprints.*	Should
UC5-11	***Self-diagnosis data intention***	*In UC5 demonstrator, self-diagnosis data are obtained after self-diagnosis cycle.*	Must
UC5-12	***Low frequency real-time data***	*Low-frequency data is continuously obtained by the demonstrator via streaming.*	Must
UC5-13	***Self-diagnosis file formats.***	*In UC5 demonstrator, self-diagnosis data shall be processed as csv files and MATLAB files.*	Must
UC5-14	***Low frequency real-time data file format.***	*In UC5 demonstrator, low frequency data shall be interpreted as JSON data file.*	Must
UC5-15	***Computation off-load***	*Cloud infrastructure, specifically AWS EC2, are to provide computing resources on-demand.*	Could

Table 3.6.6-1: List of requirements for the UC 5

4 Requirements for ExtremeXP framework

In the previous section, five different UCs have been presented, highlighting their potential and envisioned contribution within the ExtremeXP framework, along with the technical requirements of the associated pilot demonstrators. In this section, we interpret those requirements into a set of design expectations for the framework architecture. These requirements are consolidated into a set of tables presented below. The table provides a bird's eye view of the contribution of the different UC toward specific research lines, qualitatively as per the number of requirements they induce over the ExtremeXP framework.

Research Line	UC1	UC2	UC3	UC4	UC5
Analysis-aware Data integration	3	1	1	5	1
User-driven AutoML	4	4	2	4	5
Transparent & Interactive Decision Making	6	5	2	4	3
Extreme Data & Knowledge Management	1	2	1	1	1
User-driven Optimisation of Complex Analytics	4	5	4	4	4

Table 4-1: Summary of the involvement of the different UCs in ExtremeXP's research lines. The columns quantify the number of requirements for each UC.

4.1 Analysis-aware Data integration

The table below reviews the requirements serving the analysis-aware data integration. It interprets and consolidates the requirements issued by the five UCs on aspects related to the heterogeneity of data sources, simulation involvement, and expectation over parallelisation of the data processing steps.

Req. Id	Name	Description	Priority (MoSCoW)	Related UC
ADC-1	***Adjust trade-offs between system level metrics***	*The analytics workflow configuration will allow adjusting the importance of the different objectives that are selected to be optimized.*	Must	UC1, UC4
ADC-2	***Quantify trade-offs between selected metrics to be optimized***	*The framework will provide quantifications of the trade-offs between the selected metrics to be optimized, between different variants of an experimental workflow.*	Should	UC4
ADC-3	***Execute data preparation processes***	*Data preparation processes, including data selection, preprocessing, integration, and cleaning will be supported by the framework.*	Must	UC1, UC2, UC4
ADC-4	***Analyse various***	*The analytics workflow will support*	Must	UC1,

	data types	*various data types, including timeseries data and geospatial data.*		UC2, UC3, UC4, UC5
ADC-5	***Execute simulation tasks***	*The framework will support the execution of variants (using different variability points configurations) of simulation tasks*	Must	UC2, UC4

Table 4.1-1: Requirements over the Analysis-aware Data integration feature of ExtremeXP framework.

4.2 User-driven AutoML

In this subsection, we review the requirements inferred from the UCs that are leaning towards the application of automated Machine Learning (AutoML). These constraints cover the specifications of the mechanisms supporting AutoML practices within the ExtremeXP framework and the continuous learning process over model selection.

Req. Id	**Name**	**Description**	**Priority (MoSCoW)**	**Related UC**
UDA-1	***Scalability***	*The deployed machines pipeline can scale up and down based on its current workload and computing resource needs.*	Should	UC1, UC5
UDA-2	***User feedback involvement in AutoML process***	*The framework will support the utilization of user feedback as input in AutoML processes, during an experimental workflow.*	Must	UC2
UDA-3	***Execute analytics/ML tasks via AutoML processes, requiring minimum user input***	*The framework will support the execution of variants of analytics/ML tasks by non-expert users with minimum technical knowledge.*	Should	UC1, UC2, UC4, UC5
UDA-4	***Execute forecasting tasks***	*The framework will support the execution of variants (using different variability point configurations) of forecasting tasks via machine learning methods.*	Should	UC1, UC4, UC5
UDA-5	***Execute event detection tasks***	*The framework will support the execution of variants (using different variability point configurations) of event detection tasks on time series via machine learning methods.*	Should	UC2, UC5
UDA-6	***Execute concept drift detection tasks***	*The framework will support the execution of variants (using different variability points*	Could	UC5

		configurations) of concept drift detection tasks on time series via machine learning methods.		
UDA-7	***Execute classification tasks***	*The framework will support the execution of variants (using different variability points configurations) of classification tasks on tabular and geospatial data via machine learning methods.*	Should	UC1, UC2, UC3, UC4
UDA-8	***Execute recommendation tasks***	*The framework will support the execution of variants (using different variability points configurations) of recommendation tasks on tabular and geospatial data via machine learning methods*	Should	UC3, UC4

Table 4.2-1: Requirements over the User-driven AutoML feature of ExtremeXP framework.

4.3 Transparent & Interactive Decision Making

In the following table, we detail the requirements impacting the preparation of the explainability-oriented user interaction toolset and the interactive visualisation. These requirements will constraint the selection of the visualisation techniques, the design of the user interfaces, the conception of visual analytics framework, and define requirements over explanation method to be developed.

Req. Id	**Name**	**Description**	**Priority (MoSCoW)**	**Related UC**
DM-1	***Workflow variant as variability***	*The selection of a workflow variant is a variability point.*	Must	UC2
DM-2	***Obtain map-based visualizations***	*Visualization mechanisms will allow the visualization of data and results on maps.*	Should	UC1, UC3, UC4
DM-3	***Obtain spatial and spatiotemporal data visualizations***	*Visualization mechanisms will allow the visualization of spatial and spatiotemporal data and results.*	Should	UC1, UC4
DM-4	***Obtain timeseries visualizations***	*Visualization mechanisms will allow the visualization of multivariate timeseries data and results.*	Should	UC1, UC2, UC5
DM-5	***Obtain visualizations for simulation stages and results***	*Visualization mechanisms will allow the visualization of simulation stages and simulation results.*	Should	UC1, UC2, UC3, UC4, UC5
DM-6	***Obtain comparative visualizations***	*Visualization mechanisms will allow the comparison and verification of the results between different analytics workflows, executed using*	Must	UC1, UC5

		different variability points configurations.		
DM-7	***Obtain interactive and adaptive visualizations***	*Visualization mechanisms will support user interaction with regard to, for example, data and result selection, visualization types and user feedback provision.*	Must	UC1, UC2, UC4,
DM-8	***Explain the decisions of an analytics/ML model***	*The framework will provide explanations on the decisions of an analytics/ML model.*	Should	UC2

Table 4.3-1: Requirements over Transparent & Interactive Decision Making of ExtremeXP framework.

4.4 Extreme Data & Knowledge Management

The Extreme data and knowledge management feature encompasses the access control to datasets and knowledge artefacts involved in the experiments, the non-repudiability of the made decision and the attestation of the provenance of the involved dataset. In the following table, we list the requirements extracted from the UCs related to this feature.

Req. Id	Name	Description	Priority (MoSCoW)	Related UC
KM-1	***Authentication***	*The access to data source is controlled via authentication.*	Must	UC1, UC5
KM-2	***Attestation***	*The consumed dataset needs to be attested.*	Could	UC3
KM-3	***Context-wise authorisation to dataset***	*The access to a specific data is controlled via attribute-based access controlled.*	Could	UC2
KM-4	***Non-repudiation***	*The access and consumption of a dataset is imputed via a NFT (Non-Fungible Token) token.*	Could	UC4
KM-5	***Access control to knowledge data***	*The knowledge acquire from the experiment can be exposed and made accessible via an access mitigated via access control.*	Could	UC2

Table 4.4-1: Requirements over Extreme Data & Knowledge Management of ExtremeXP framework.

4.5 User-driven Optimization of Complex Analytics

In this subsection, we detail the requirements impacting the interaction between the complex experiment-driven analytics and the user. The consolidated requirements will cover some aspects of the experiment engine, the expression of user intent and the profiling of the user and its context.

Req. Id	Name	Description	Priority (MoSCoW)	Related UC
CA-1	***Configure experiment's***	*The framework will support the selection of variability points (alt.*	Must	UC1, UC2, UC3, UC4,

	variability points	*experiment configurations, hyperparameters) and the manual or semi-automatic exploration of their space, for an analytics experiment.*		UC5
CA-2	***Assess the impact of variability points***	*The framework will support the assessment of the impact of different variability points configurations of the experiment on the performance of the analytics task.*	Must	UC1, UC2, UC3, UC4, UC5
CA-3	***User feedback***	*The usage of a workflow variant should be affected by the feedback of the user.*	Must	UC2
CA-4	***Optimize system level metrics***	*Several system level metrics, including accuracy, precision, recall, mean prediction error, scalability, will be available as objectives to be optimized in the experimentation workflow*	Should	UC1, UC2, UC3, UC4, UC5
CA-5	***Finding optimal variant in a semi-automated way***	*The framework will offer the possibility to identify the best performing analytics variant using different search and optimization algorithms (genetic algorithms, local search, multi-armed bandits, etc.)*	Must	UC1, UC2, UC3, UC4, UC5

Table 4.5-1: Requirements over User-driven Optimisation of Complex Analytics of ExtremeXP framework.

5 Conclusion

This document presented a first approach to the definition of the ExtremeXP UCs. To that extent, we have analysed the objectives of the project to identify which aspects the proposed UCs should be further analysed and justified. We detailed the strategy to collect those requirements with partners providing the different research and technology results of ExtremeXP.

Following this strategy, a thorough technical analysis of the identified UCs has been conducted. We have reviewed the context of each UC and the core objectives of each pilot. We have described and defined the different experiments to be conducted with the use of extreme data, by identifying the compounding of the analytics workflows, the involvement of variation points and the several classes of actors they involve. We have detailed the properties of the dataset to be consumed, and where applicable, the specific methodologies to acquire and access them. The environment and testbed serving the deployment of the UC demonstrator is reviewed in their current state and will be refined for the preparation of the validation framework. We justify the added value conferred by ExtremeXP to each UC, and finally summarise the technical requirements of the demonstrator.

Furthermore, we interpret and consolidate the UC requirements as directives serving the development activities of ExtremeXP’s key features. Each functional requirement is detailed and prioritised according to its criticality for UC owners, serving the preparation of UC demonstrators and serving as a baseline for the investigation and implementation of ExtremeXP’s features. Beyond ExtremeXP, the collected requirements also reflect the typical functionalities and design guideline that can be expected from real-world data analytics platform exploiting extreme data in different domains (manufacturing, mobility, cybersecurity, etc.).

The requirements exposed in this deliverable will be refined in the different technical activities of the project, and the specification of the UC demonstrators, evaluation and testbed specification will be exploited during the implementation of the several demonstrators, their deployment, and the preparation of their validation plans.

References

[1] MoSCoW method, available online at https://en.wikipedia.org/wiki/MoSCoW_method, last accessed in May 2023

[2] J. Hofmann and H. Schüttrumpf, “floodGAN: Using Deep Adversarial Learning to Predict Pluvial Flooding in Real Time,” Water, vol. 13, no. 16, p. 2255, Aug. 2021. https://doi.org/10.3390/w13162255

[3] “Bienvenue sur le projet QGIS !” https://www.qgis.org/fr/site/ (accessed Jun. 23, 2023).

[4] Mahesh, R.B., Leandro, J., Lin, Q. (2022). Physics Informed Neural Network for Spatial-Temporal Flood Forecasting. In: Kolathayar, S., Mondal, A., Chian, S.C. (eds) Climate Change and Water Security. Lecture Notes in Civil Engineering, vol 178. Springer, Singapore. https://doi.org/10.1007/978-981-16-5501-2_7

[5] Hou, J., Zhou, N., Chen, G. et al. Rapid forecasting of urban flood inundation using multiple machine learning models. Nat Hazards108, 2335–2356 (2021). https://doi.org/10.1007/s11069-021-04782-x

[6] Davidjbianco, “Enterprise Detection & Response: The Pyramid of Pain,” Enterprise Detection & Response, Mar. 01, 2013. http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html (Accessed Jun. 20, 2023).

[7] R. Al-Shaer, J. M. Spring, and E. Christou, ‘Learning the Associations of MITRE ATT&CK Adversarial Techniques.’ arXiv, May 12, 2020. Accessed: May 09, 2023. [Online]. Available: http://arxiv.org/abs/2005.01654

[8] W. Wang, B. Tang, C. Zhu, B. Liu, A. Li, and Z. Ding, ‘Clustering Using a Similarity Measure Approach Based on Semantic Analysis of Adversary Behaviors,’ in 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, Hong Kong: IEEE, Jul. 2020, pp. 1–7. https://doi.org/10.1109/DSC50466.2020.9194468

[9] M. Mehmood, R. Amin, M. M. A. Muslam, J. Xie, and H. Aldabbas, "Privilege Escalation Attack Detection and Mitigation in Cloud Using Machine Learning," in IEEE Access, vol. 11, pp. 46561-46576, 2023. https://doi.org/10.1109/ACCESS.2023.3273895

[10] Willstrand, Ola et al. “Fire detection & fire alarm systems in heavy duty vehicles: WP1 – Survey of fire detection in vehicles.” (2015).

[11] Jadhav, Rohini Ravindra and Poonam D. Lambhate. “Enhancing Fire Detection for Indoor and outdoor locations via video surveillance.” (2016).

[12] Lira H, Martí L, Sanchez-Pi N. A Graph Neural Network with Spatio-Temporal Attention for Multi-Sources Time Series Data: An Application to Frost Forecast. Sensors. 2022; 22(4):1486.

[13] T. Dargahi, A. Dehghantanha, P. N. Bahrami, M. Conti, G. Bianchi, and L. Benedetto, ‘A Cyber-Kill-Chain based taxonomy of crypto-ransomware features,’ J Comput Virol Hack Tech, vol. 15, no. 4, pp. 277–305, Dec. 2019. https://doi.org/10.1007/s11416-019-00338-7

[14] L. F. Sikos, ‘Cybersecurity knowledge graphs,’ Knowl Inf Syst, Apr. 2023. https://doi.org/10.1007/s10115-023-01860-3

[15] ‘El Elastic Stack: Elasticsearch, Kibana, Beats y Logstash’, Elastic. https://www.elastic.co/es/elastic-stack (Accessed Jun. 27, 2023).