D1.1 Technical Resources and Problem definition

Deliverable 1.1

Dispatcher3

Grant:	**886461**
Call:	**H2020-CS2-CFP10-2019-01**
Topic:	**JTI-CS2-2019-CfP10-SYS-01-16**
Consortium coordinator:	**University of Westminster**
Dissemination level:	**Public**
Edition date:	**16 December 2020**
Edition:	**01.01**

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Luis Delgado / University of Westminster	Project member	16 December 2020
Paolino De Falco / University of Westminster	Project member	16 December 2020
Damir Valput / Innaxis	Project member	16 December 2020
Jovana Kuljanin / Universitat Politècnica de Catalunya	Project member	16 December 2020

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Luis Delgado / University of Westminster	Project member	16 December 2020
Paolino De Falco / University of Westminster	Project member	16 December 2020
Gérald Gurtner / University of Westminster	Project member	16 December 2020
Damir Valput / Innaxis	Project member	16 December 2020
Ralph Schultz / PACE	Project member	16 December 2020
Hector Fornes / Vueling Airlines	Project member	16 December 2020

Approved for submission to the CSJU

Beneficiary	Position	Date
University of Westminster	Project coordinator	16 December 2020

Document History

Edition	Date	Status	Author	Justification
01.00	01 December 2020	Release	Dispatcher3 Consortium	New document for review by Topic Manager
01.01	16 December 2020	Release	Dispatcher3 Consortium	Small modifications considering Topic Manager comments

Dispatcher3

INNOVATIVE PROCESSING FOR FLIGHT PRACTICES

This deliverable is part of a project that has received funding from the Clean Sky Joint Undertaking under grant agreement No 886461 under European Union's Horizon 2020 research and innovation programme.

Abstract

This deliverable starts with the proposal of Dispatcher3 and incorporates the development produced during the first five months of the project: activities on different workpackages, interaction with Topic Manager and Project Officer, and input received during the first Advisory Board meeting and follow up consultations.

This deliverable presents the definition of Dispatcher3 concept and methodology. It includes the high level the requirements of the prototype, preliminary data requirements, preliminary technical infrastructure requirements, preliminary data processing and analytic techniques identification and a preliminary definition of scenarios.

The deliverable aims at defining the view of the consortium on the project at these early stages, incorporating the feedback obtained from the Advisory Board and highlighting the further activities required to define some of the aspects of the project.

Table of Contents

EUROPEAN UNION

List of figures

List of tables

Executive summary

Dispatcher3 is a CleanSky2 Innovative Action. This project started in June 2020 and it is 30 months long. The Topic Manager is Thales AVS France SAS. Dispatcher3 aims at developing a software prototype for the acquisition and preparation of historical flight data in order to give support to the optimisation of future flights providing predictive capabilities and advice to relevant stakeholders.

The primary objective of Dispatcher3 is to develop a prototype for the acquisition and preparation of historical flight data in order to give support on the optimisation of future flights providing predictive capabilities and advice to relevant stakeholders (e.g., dispatchers and pilots). This will be done considering airline preferences and impact of flight missions on the overall airline objectives.

This deliverable presents the definition of the problem with the concept and methodology that will be followed during the project. It also includes the high level the requirements of the prototype, data requirements, preliminary identification of technical infrastructure requirements and data processing and analytic techniques, and a preliminary identification of scenarios. The deliverable starts with the proposal of Dispatcher3 but incorporates the development produced during the first five months of the project: activities on different workpackages, interaction with Topic Manager and Project Officer, and very relevantly, the input received during the first Advisory Board meeting, follow up bilateral discussions with different members of the Advisory Board, and a consultation. These activities have significantly supported the definition of the concept of operation and project scope described in this document.

The deliverable aims at defining the view of the consortium on the project at these early stages, while highlighting the feedback obtained from the Advisory Board and the further activities required to define some of the aspects of the project.

Dispatcher is organised in three layers:

- Data infrastructure: BeSt by DataBeacon, a multi-sided, data storage and processing platform. BeSt provides private environments, secure dataframes, machine learning development frameworks and a scalable cloud computing and storage infrastructure.
- Predictive capabilities, provided by two distinct modules:
  - Data acquisition and preparation, which a first phase of data wrangling and a second step of descriptive analytics.
  - Predictive models, consisting on target variable labelling and feature engineering, and training, test and validation of machine learning models.
- Advice capabilities: producing specific advice to targeted roles (dispatcher and pilots).

In the current European operational environment, the dispatching process of generating flight plans is highly automated. However, the dispatching process understood as the management of the fleet on the day of operation has increased the relevance of longer lookahead decision making process. Identifying potential disruptions early in the day might provide possibilities to plan for solutions beyond adjustment of flight plans (e.g., aircraft swapping). Finally, independently on the automation, dispatchers preparing the flight plans might still manually intervene to adjust and modify solutions when non-nominal situations arise (e.g., avoiding a turbulence region by using a different flight level or an ATFM regulation by re-routing).

EUROPEAN UNION

Dispatcher3 will focus on providing support to the processes performed by the dispatcher a few hours prior the flight and to provide advice to the flight crew. However, the project may explore how to support some of these fleet-wide activities performed at pre-departure at different decision periods (e.g., from the day before operations to estimate the likelihood of disturbances in given flights, to hours or minutes before the flight to identify the expected variance between planned and executed flight once the flight plan has been fixed to support the duty manager on fleet management decisions). Independent roles-oriented models will be developed. Dispatcher3 will help the dispatcher to identify non-nominal situations and the crew to better understand their operations.

High-level requirements for the prototype have been identified along with a preliminary identification of required datasets. The first identification of data processing and analytic techniques and scenarios highlight the need to advance on the data acquisition so that a data-driven approach to the methods and use cases selection can be performed.

1 Introduction

Dispatcher3 is a CleanSky2 Innovative Action. This project started in June 2020 and it is 30 months long. The Topic Manager is Thales AVS France SAS. Dispatcher3 aims at developing a software prototype for the acquisition and preparation of historical flight data in order to give support to the optimisation of future flights providing predictive capabilities and advice to relevant stakeholders.

1.1 Dispatcher3 objectives

The primary objective of Dispatcher3 is to develop a **prototype** for the acquisition and preparation of historical flight data in order to give support on the **optimisation** of future flights providing **predictive capabilities** and **advice to relevant stakeholders** (e.g., dispatchers and pilots). This will be done considering airline preferences and impact of flight missions on the overall airline objectives.

Dispatcher3 focuses on supporting the **activities prior departure**: dispatching, understood in this case as the broad flight planning from the day prior operations to the flight plan definition and selection, and advise pilot on how to operate the flight. **Pilot3** project, a Clean Sky 2 Innovation Action granted to members of Dispatcher3 consortium and starting 1st November 2019, in its turn, will tackle the continuous monitoring of the flight during its **execution**, suggesting trajectories modifications and with an understanding of the implication of these alternatives on the different objectives defined by the airlines policies, can be carried out to improve the performance of the flight. Both projects have significant synergies including the collection of flight policies, computation of key performance indicators (KPIs) and the use of the same data infrastructure (BeSt). The coordination between both initiatives contributes to one of the objectives of this project which is to obtain an **improved pilot understanding of the flight and their actions on all the flight processes**: from before the flight plan to its completion.

After discussing with the Advisory Board, as presented in Section 2, Dispatcher3 will focus on providing support to **tactical planners**, **duty managers**, **dispatchers** and **pilots** in the time-frame ranging from the **day prior the flight to minutes before off block**.

In particular, Dispatcher3 will:

- Identify **airliners policies** in terms of expected **targeted KPIs** that need to be estimated and monitored, and understand the different **flight management policies**. This will be done considering the impact of a given flight on the whole airline network.
- Develop **software tools** with **data engineering capabilities** to clean, synchronise and merge **past flight data**, including flight plans and actual flight execution data (e.g., quick access recorders – QAR), along with **other relevant operational data**, such as operational environmental data (e.g., air traffic flow management -ATFM- regulations) or weather.

EUROPEAN UNION

- **Prepare the data gathered** so that data analytic techniques defined in Dispatcher3 can be applied, but also to remain generic enough to facilitate the exploitation of data for other future objectives in a **data infrastructure** capable of scaling up at an acceptable cost
- Use advanced data analytic techniques (machine learning (ML) algorithms) to provide **predictive capabilities** that can be used by dispatchers to estimate **KPIs** for new alternative flight plans at different time-frames.
- Identify the precursors that impact the prediction of the KPIs in order to **gain understanding** on the (internal or external) **events** which explain the variances between planned flight plans and their executed realisation.
- Develop an **advice generator module for planning activities** to process the outcome of the predictive engine and transform it into actionable indications considering airline policies.
- Produce **advice for pilots** on the execution of the selected flight plan. This will consider the expected variability between planned and executed flight plan, the precursors for these variances and the airline's flight policies. The advice will provide the pilots with a **better understanding of the stakes of the current mission and how to manage it during the flight** in close cooperation with **Pilot3.**
- **Validate the prototype with internal and external activities** considering a broad stakeholder community thanks to the capabilities of the members of the consortium and the definition of an Advisory Board.

As expected, there are differences between planned flights and their execution. These variations are due to internal and external events (e.g., holdings due to congestion at arrival airport, shorter routes than planned). Different flight plan alternatives can be considered by experienced dispatchers in order to select one which best captures the airline's policies. However, the variations of a single flight plan might have a limited impact on the overall experienced airline's performance and fleet management actions can be considered (e.g., aircraft swapping) to minimise the impact of disruptions in the network. These processes, however, rely on individual expertise and automatisation, and lacks of the benefit of systematically considering historical performances of flights on same routes under similar conditions. For example, predicting variations between block times or fuel consumed. Tactically, pilots might lack an understanding of the changes ahead and therefore are not provided with specific advice on how to operate a given flight considering the impact of the current operational environmental conditions, such as weather, air traffic congestion, time of the day, etc.

Flight operations generate a large set of data from different sources: from planned activities, such as flight plan, forecast weather at the moment of dispatching the flight or expected airspace and airport congestion, to actual realisations, such as flight performance data (QAR), actual weather or holding times. The scope of Dispatcher3 is to **consider all these data** in order to produce predictions on the outcome of individual flight plans on the different airline's KPIs, which could be **used to generate the operational flight plan** considering the expected trade-offs involved and, which could **provide advice to pilots** on how to operate the given flight considering the precursors of the different variances expected.

One of the main objectives of Dispatcher3 is therefore to **improve these dispatching and flight operating processes by providing an infrastructure able to leverage on historical data and machine learning techniques** to systematically estimate the variability between planned and executed flight plans, providing expected results of flight plans and advice to the flight planning processes and pilots. This tool will help create more informative flight plans better aligned with the airline operator policies. Dispatcher3 tool will also enable airlines to find a suitable solution to fly as efficiently as possible within the known constraints, to ensure the robustness of the airline network against disturbances and environmental conditions (e.g., adverse meteorological conditions, network capacity) and the airline's network-wide impact (changes request in the planning or pilot behaviour). Dispatcher3 has as objective to lead to a **more robust network and a better operational outcome for the European airlines**.

The infrastructure developed and the type of data processing performed in the project will also provide information that could **strategically** be used by airlines to improve the scheduling process (e.g., identifying conditions in which routes require more or less buffers) and even flight policies (e.g., in which conditions advise their pilots to recover delay) as part of the engineering back-office activities.

1.2 Deliverable purpose and intended reader

The purpose of this deliverable is to present the definition of the problem which is considered by Dispatcher3. It presents the concept and methodology that will be followed during the project.

This deliverable captures at a high level the requirements of the prototype, data requirements, preliminary technical infrastructure requirements, a review and identification of preliminary data processing and analytic techniques and a definition of potential scenarios. The deliverable starts with the proposal of Dispatcher3 but incorporates the development produced during the first 5 months of the project: activities on different workpackages, interaction with Topic Manager and Project Officer, and very relevantly, the input received from the Advisory Board.

The Advisory Board of Dispatcher3 is composed of three airlines, the network manager, a consultancy company with expertise on airline operations, a centre for training of dispatchers, and an independent consultant. The first Advisory Board meeting was carried out on the 9th October 2020 when initial feedback was obtained on airlines operations and the scope of Dispatcher3. To further gather insight on current airline operations and challenges in the European context for different business models and type of operation, follow up bilateral meetings and consultations were conducted during the final weeks of October 2020 with various members of the Advisory Board, and with the operational department of the consortium partner Vueling. Finally, a consultation was conducted to verify that the approach selected for Dispatcher3 was aligned with the views of the Advisory Board.

The deliverable aims at defining the view of the consortium on the project scope at this early stages, while highlighting the feedback obtained from the Advisory Board and the further activities required to define some of the aspects of the project. It triggers the technical activities of the remaining part of the project. A particularity of Dispatcher3 is that it is a data-driven project. This means that some of the functionalities and testing scenarios will need to be adjusted and refined once datasets are acquired and processed. This will be part of the agile methodology followed in the project.

The intended reader of the deliverable is the Topic Manager and Project Officer in order to validate that the concept and approach defined are adequate; and broader audience with an interest in the topic. This deliverable aims at improving the dissemination of the project objectives, approach and activities.

1.3 Deliverable structure

The deliverable is structured in 8 sections:

- Section 2 presents the problem definition and scope formulation. This section includes the Dispatcher3 concept and methodology that will be used to develop the prototype. Note that some of the methodologies will be further developed as part of WP3 – Domain driven data engineering and analytic techniques (reported in D3.1 – Data engineering and analytic techniques report); the prototype details will be further considered in the prototype development (WP4 – Prototype development); and the validation approach including developmental methodologies will be captured by the activities of WP5 – Prototype verification and validation.

- The detailed requirements of the modules that form Dispatcher3 will be part of WP4 – Prototype development, but a first high-level set of requirements are provided in Section 3.

- Section 4 includes the preliminary data requirements considering the potential use of the data and approaches towards data availability. Note that the acquisition of the data and their further definition will be performed in WP2.

- Section 5 contains preliminary technical infrastructure requirements describing in more details the data infrastructure platform that will be used for Dispatcher3 (BeSt by DataBeacon).

- Section 6 contains a review of preliminary data processing and analytic techniques. These techniques will be further defined and selected as part of WP3 – Domain driven data engineering and analytic techniques. These techniques will be implemented as part of WP4 – Prototype development.

- Section 7 defines the preliminary scenarios considered in Dispatcher3, these already consider input from the Advisory Board. However, as Dispatcher3 is driven by the data available and processed, this will be reviewed and refined as part of WP5 – Prototype verification and validation considering the datasets identified in WP2 – Data collection and management.

- Section 8 contains next steps and look ahead, being very relevant for this deliverable which triggers the technical activities of the project.

The document closes with references and acronyms.

2 Problem definition and scope formulation

2.1 Current concept of operations and motivation

Different airlines have different processes as part of their flights operations, however they can be generalised in the activities summarised in **Table *1***.

Table 1. Flight management activities

Activity	Time when activity performed	Objectives and consideration
Flight policy (standard definition)	-	• These policies reflect the business objectives of the airlines and, • Their main objective is to indicate how the different operations should be performed in the airline in order to achieve these objectives • Example of these policies could be the nominal cost index to be used for flight planning.
Network planning and scheduling	Months before operation	• Definition of routes (origin-destination). • Definition of schedules (SOBT, SIBT). • Business and operational aspects to be considered such as: historical block times, demand, operational constraints, availability of strategic scheduling slots at airports which require them, network robustness.
Aircraft assignment	15 – 7 days before operations	• Assignment of specific aircraft registration to individual flights. • This process needs to consider operational aspects such as aircraft and crew availability, maintenance, network robustness.
Day of operation planning	1 day before operations	• These activities aim at generating a flight operation plan. • It can be considered the first activities of the flight dispatching as first flight plans could be considered. • The main objective is to identify potential network issues and disruptions to prepare preventive measures such as aircraft tail swapping or crew reassignment.

EUROPEAN UNION

Activity	Time when activity performed	Objectives and consideration
First flight plan generation	10 – 9 hours before operation	• In this phase, a first fully operational flight plan is generated considering available environmental data (e.g., weather forecast, RAD). • This process can be more or less automatised and in some airlines parameters such as the lateral route will be fixed at this moment. • The outcome of these analysis might produce actualisations to the flight operation plan.
Updated flight plan generation	4 – 3 hours before operation	• The flight plans are updated considering updated environmental data. • The number of flight plans that are generated might vary per airline from a continuous automatic update and generation of flight plan each time information is updated to specific time milestones.
Last flight plan	3 – 0.5 hours before operation	• If new relevant information becomes available or arises an update on the flight plan could be generated. • However, submission of new flight plans 30 minutes before operation is seldom done to avoid penalisation as late filler by the network manager.
Flight monitoring	During period between AOBT–AIBT	• Monitoring of estimated arrival times and flight operations to consider impact on network (e.g., further rotations).
Post-operation analysis	Days after operations	• Statistical analysis to improve operations (e.g., cost index to be used, fuel for holdings, etc.) which will close the loop by providing update on some of the flight policies.

The dispatching process understood as the generation and submission of flight plan is being **highly automatised** in Europe with few staff in charge of the supervision of these activities. This differs from the North American market where dispatchers are assigned with more responsibilities and airlines tend to have a large number of dispatchers working on the day of operations. There are several reasons from this:

- The **operational environment is very constraint** (e.g., routes available are limited between origin and destination pairs, regulation imposes fuel loads), and **flights are relatively short**.
- This means that **potential changes on the flight plan are limited** (reduced different number of routes and small impact on delay recovery by modifying parameters such as cost index).

This lead to the use of **flight plan generators** (e.g., Lufthansa LIDO software) which are generally fed with many constraints and pre-defined optimisation parameters (CI estimated by the back-office for specific routes), and a reduced number of dispatchers to generate all the flight plans for the airline.

However, as indicated, the dispatching process understood as the management of the fleet on the day of operation has increased the relevance of longer lookahead decision making process. Identifying potential disruptions early in the day might provide possibilities to plan for solutions beyond adjustment of flight plans (e.g., aircraft swapping). Finally, independently on the automation, dispatchers preparing the flight plans might still manually intervene to adjust and modify solutions when non-nominal situations arise (e.g., avoiding a turbulence region by using a different flight level or an ATFM regulation by re-routing).

Figure 1 depicts the different phases from scheduling to operation of a given flight identifying both the processes and the roles that are involved. As indicated, there are a set of roles in charge of the different activities, which can be summarised as in **Table 2**

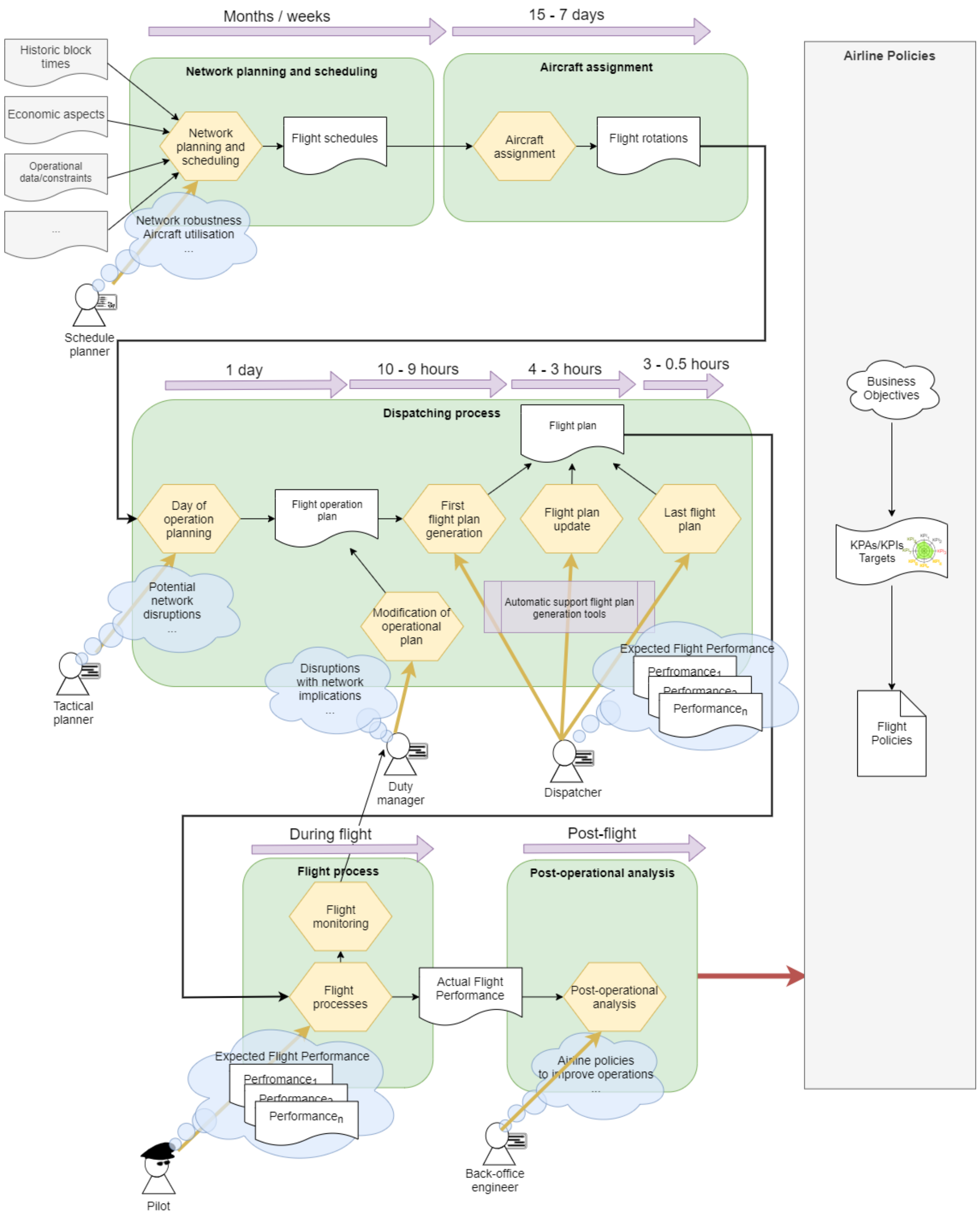

Figure 1. Flight operation concept

Table 2. Different roles involved on flight management activities

Role	Main activities and consideration
Schedule planner	• In charge of the definition of the airline network.
Tactical planner	• In charge of the analysis of the network situation prior the day of operations in order to perform the day of operation planning. • Historical information on probabilities of disruptions and information from the Network Manager can be considered at this stage.
Duty manager	• Responsible for the management of the dispatching activities during the day of operation but from a *network perspective*. • The role does not focus on the generation or adjustment of specific flight plans, but rather on the assessment of delay and disruption in the network. • The type of actions that will be performed are aimed at preventing the propagation of disruption and costs through the network (e.g., with aircraft swapping, consideration of crew reassignment). • It is in charge of executing and updating the operation plan.
Dispatcher	• This is the specific role of generating and verifying the flight plans. • These activities are highly automatised and dispatchers will intervene to adjust some operational aspects, for example, modifying flight levels to avoid a turbulence region, adjusting fuel loading to consider fuel tankering, etc. • Dispatchers might estimate the expected performance of different flight plan options considering the airline policies. • Moreover, they could consider the expected variance between planned flights and their actual execution in order to select the most adequate flight plan.

Role	Main activities and consideration
Pilots	• Pilots are in charge of the tactical execution of the flight. • They are provided with information from the dispatchers. It has been identified that a pre-processing of the dispatching information might be useful in order to help them identify the main challenges of their operations. • Different airlines would allow different degrees of autonomy on the operation of the flight to their crew. They can then react to disruptions considering tactical operational information and using a trajectory optimisation or prediction system, such as the flight management system (FMS) or more advanced systems embedded in electronic flight bags (EFB) or tablets. • Pilots might consider alternatives (e.g., modification of their cost index, request of directs in routes), based on the airline flight policies and on the expected results of their actions on the overall flight performance. • They also use their expertise on their operations to foresee the likelihood of different outcomes such as experiencing holdings at arrival at the end of the day for a given airport or route. However, these decisions are not based on the systematic analysis of previous data and in some cases rely on anecdotal pilot experience.
Back-office	• This role is in charge of conduct the post-operational analysis of the flight to identify areas and situations to improve the operations. • This will translate into updated flight policies (e.g., nominal cost indexes to be used, fuel considered for taxi in and out).

Dispatchers3 aims at improving these activities by providing an infrastructure able to leverage on historical data and machine learning techniques to systematically estimate the variability between planned and executed flight plans. It will contribute to support the different roles previously described.

2.2 Dispatcher3 concept

As shown in **Figure 2**, the supporting layer of Dispatcher3 is the **data infrastructure** (BeSt platform by DataBeacon and developed by Innaxis (see Section 2.4.1 and and Section 5 for more details)). This infrastructure will store, prepare and process historical actual operational, environment and flight data, along historical planned data. Then, machine learning techniques will be used to provide predicting and advisory capabilities which will be used before the operation of the flight. As depicted in the Figure, Dispatcher3 focuses on the pre-departure activities (even is support to post-flight analysis could be provided), while **Pilot3**, a Clean Sky 2 project awarded to members of the **Dispatcher3** consortium, focuses on the execution phase of the flight and on improving the pilot decision making process.

EUROPEAN UNION

As described in the previous section there are a set of roles which participate on the dispatching processes at different time-frames. As indicated the project proposal, Dispatcher3 will focus on the processes which are performed by the dispatcher a few hours prior the flight and it will provide advice to the flight crew. However, the project may explore how to support the processes of other roles performed at pre-departure at different decision periods (e.g., from the day before operations to estimate the likelihood of disturbances in given flights, to hours or minutes before the flight to identify the expected variance between planned and executed flight once the flight plan has been fixed to support the duty manager on fleet management decisions). Therefore, independent **roles-oriented models** will be developed (considering these distinct requirements). It is crucial that these models only consider the information available at the moment the prediction is given. For example, if the system provides support on the day before of operations by the tactical planner, the delay due to ATFM regulations will need to be estimated based on the available information; however, minutes before the flight, when the dispatcher is assessing the flight plan, this information will be already available. This means that at different time moments, some of the information might be the outcome of the predictive model (e.g., probability of having a regulation) or the input to the system (e.g., delay assigned by ATFM).

Dispatcher3 will therefore collect and analyse planned historical datasets for different time references and specific analysis, predictions and advice will be aimed at being produced for each role. In particular Dispatcher3 could support the different roles as indicated in **Table *3***.

Table 3. Potential support of Dispatcher3 for the different identified roles.

Role	Main activities and consideration
Schedule planner[3]	• Out of scope of Dispatcher3 but the project will create the infrastructure needed to store and process planned and actual historical flight and operational environmental data. This will allow strategic decisions to be further developed based on these data, e.g., modifying airline flight policies. • Dispatcher3 could provide advice on which flights, and in which conditions, are more prone to variance between schedules and execution blocks.
Tactical planner[2]	• Identify which flight plans are more likely to be disrupted. • Estimate already block times, fuel usage and impact on reactionary delay. • Support the estimation of benefit of alternatives such as aircraft swapping, crew rotations, etc.
Duty manager[2]	• This position might benefit from enhanced predictive capabilities, not aimed at improving a given flight, but at identifying which flights might suffer from disruptions in the network with a few hours of look-ahead. • The goal is to highlight, identify which flights will be prone to have disruptions and propagate them through the network.

Role	Main activities and consideration
Dispatcher[1]	• Providing enhanced metrics on the result of the flight might be useful, but in most cases, it won't be able to use the information as the actions that can be performed are very limited (e.g., cost index tend to be fixed strategically by the airlines and not modified when generating the flight plans). • There are some particular instances when these enhanced capabilities might be useful: assessment of different flight plans when avoiding areas with turbulence or with ATFM regulations, estimating the fuel required for tankering activities, expected holding times in non-nominal conditions. • Identify the precursors of the different variations between planning and execution in order to highlight the factors influencing these variabilities.
Pilot[1]	• Crews will appreciate having a better understanding on the rationale between some of the decisions performed at dispatching (e.g., fuel on-board for holdings). • It would be beneficial to have an indication of the variances that they can expect during their flight and follow up rotations. • Dispatcher3 could provide information on the expected variance between the flight plan and the execution while indicating the precursors for these changes and advice on some flight operations (such as the possibility to recover some delay in the air).
Back-office analyst[3]	• Dispatcher3 will set up an infrastructure which enables the analysis of past flights to better identify situations and operations which could be optimised (e.g., selecting different baseline cost indexes).

[1]Main scope of Dispatcher3: predictive capabilities based on advanced machine learning and advice generator modules will be created, with a focus on flight analysis.
[2]May be considered in Dispatcher3: predictive capabilities for flights but with greater focus on the network, and identification of disruptions.
[3]Out of scope of Dispatcher3: will benefit from Dispatcher3 infrastructure and capabilities.

EUROPEAN UNION

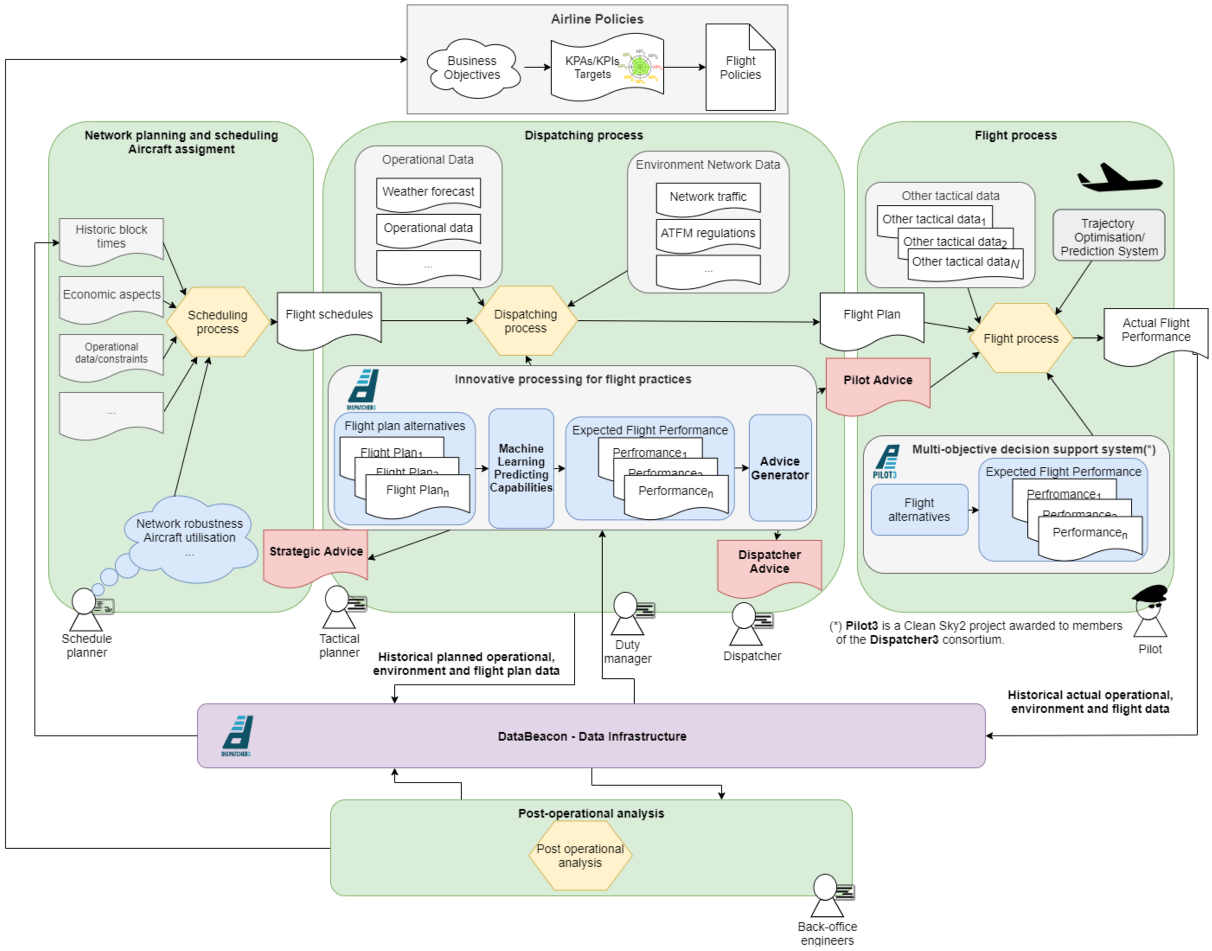

Figure 2. Dispatcher3 concept

As presented in **Table 3** these different roles might benefit from Dispatcher3, however, their needs are different. The schedule planner and the back-office analyst are out of scope of Dispatcher3 but the data infrastructure will be useful for their activities and analysis. Dispatchers and pilots focus on a specific flight and will benefit from the estimation of discrepancies between planned and executed flights. This is the target of Dispatcher3 (as presented in the proposal). The roles of tactical planner and duty manager are relevant in the current management of flight operations in the European market but their role focuses on monitoring and assessing the network operations of the airline. Therefore, even if individual flight predictions are relevant and might provide useful insight, they are more interested on the consideration of the whole fleet and the remaining of flights to be operated in the day. This means that the type of predictions that they are interested might diverge from the analysis performed for dispatchers and pilots. This might require the usage of specific algorithms which are model driven with input from the predictive machine learning models (e.g., simulation of network of flights to estimate reactionary delay considering estimated delays). Dispatcher3 may consider the requirements of these users to provide with useful predictions (computed with a longer look-ahead horizon) and considering the identification of flights which might disrupt the network rather than specific parameters for individual flights.

In general, Dispatcher3 **starts with the users**, the system will take the **current plan flight information** (first input to the system) and compute certain **target variables or KPIs** (second input to Dispatcher3).

Those two inputs, together with the historical datasets, will all be the inputs needed for Dispatcher3 to make a prediction of the KPIs for that particular flight (prediction capabilities). These KPIs might diverge from the planned ones and the precursors of these differences will be identified. The users can directly consider this predictive information or engage the advice generator, which will provide advice during the flight plan design and selection process. The advice generator will also produce information to the pilot about how to operate the flight, e.g., 'no need to increase the cost index as there is a high probability of having a flight shorter than planned due to route short-cuts'. These suggestions will be qualitative by nature and designed to digest the probabilistic outcome of the predictors into actionable indications. Note that the same prediction might produce different advice depending on the airline policies.

Besides airline datasets, Dispatcher3 will **explore the possibility of incorporating information that currently lies on the network side** (e.g., ANSP data). This will allow the project to quantify the potential benefit of sharing/accessing these datasets producing benefit for the airlines, and for the airspace managers as less variability between planned and executed plans will be expected improving the overall system uncertainty.

2.3 Target indicators to predict

Through consultation with the Advisory Board, a set of indicators are identified as candidate to be predicted by Dispatcher3 (KPIs), namely (more relevant highlighted in bold):

- **Time deviations**: block times being the most relevant one but could be disaggregated on its components: taxi times, flight time, etc., and
- **Fuel deviations**

Other directly useful variables identified are:

- **Holdings**,
- **Outcome of fuel tankering**, i.e., estimation of fuel remaining after flight for fuel tankering,
- **ATFM regulations,**
- **Impact of reactionary delay,** i.e., propagation of delay thought the network,
- **Taxi times,**
- **Taxi fuel,**
- Estimated time of departure,
- Turbulence indicator,
- Arrival procedures,
- Number of tactical adjustments to CI performed by crew (useful for longer routes in some airlines), and
- Environmental indicators

A set of research questions to be answered by Dispatcher3 have been identified for the different time-frames where predictions will be produced targeted at specific roles as described in **Table 5**.

EUROPEAN UNION

Table 4. Research questions for Dispatcher3 at different time-frames.

Time-frame	Targeted role	Research Question/Target indicators
Day prior operations (D-1)	Tactical planner	• Identify flights potentially affected by disruptions (delay). • Identify congestion in network impacting flights.
Hours prior the flight (- 10/9 H)	Duty manager	• Flights potentially affected by disruptions (delay). • Congestion in network impacting flights. • Time deviations. • Fuel deviations. • Taxi times and fuel. • Impact on reactionary delay.
Few hours prior flight (- 4/3 H)	Dispatcher	• Holdings (non-nominal conditions). • Fuel tinkering. • Time deviations. • Fuel deviations. • Taxi times and fuel.
	Duty manager	• Impact on reactionary delay.
Before push-back (30')	Pilot	• Advice on what to expect during the flight.

Finally, it has been identified that non-nominal conditions are more relevant as flight policies tend to focus on general flight management and support might be required by the different roles when non-nominal situations arise. See Section 7 for more details on potential scenarios to model.

2.4 Dispatcher3 approach

In order to achieve the concept of operations described in the previous section, Dispatcher3 organises the work on **three layers**, as depicted in **Figure 3: data infrastructure**; **predictive capabilities**; and **advice capabilities**

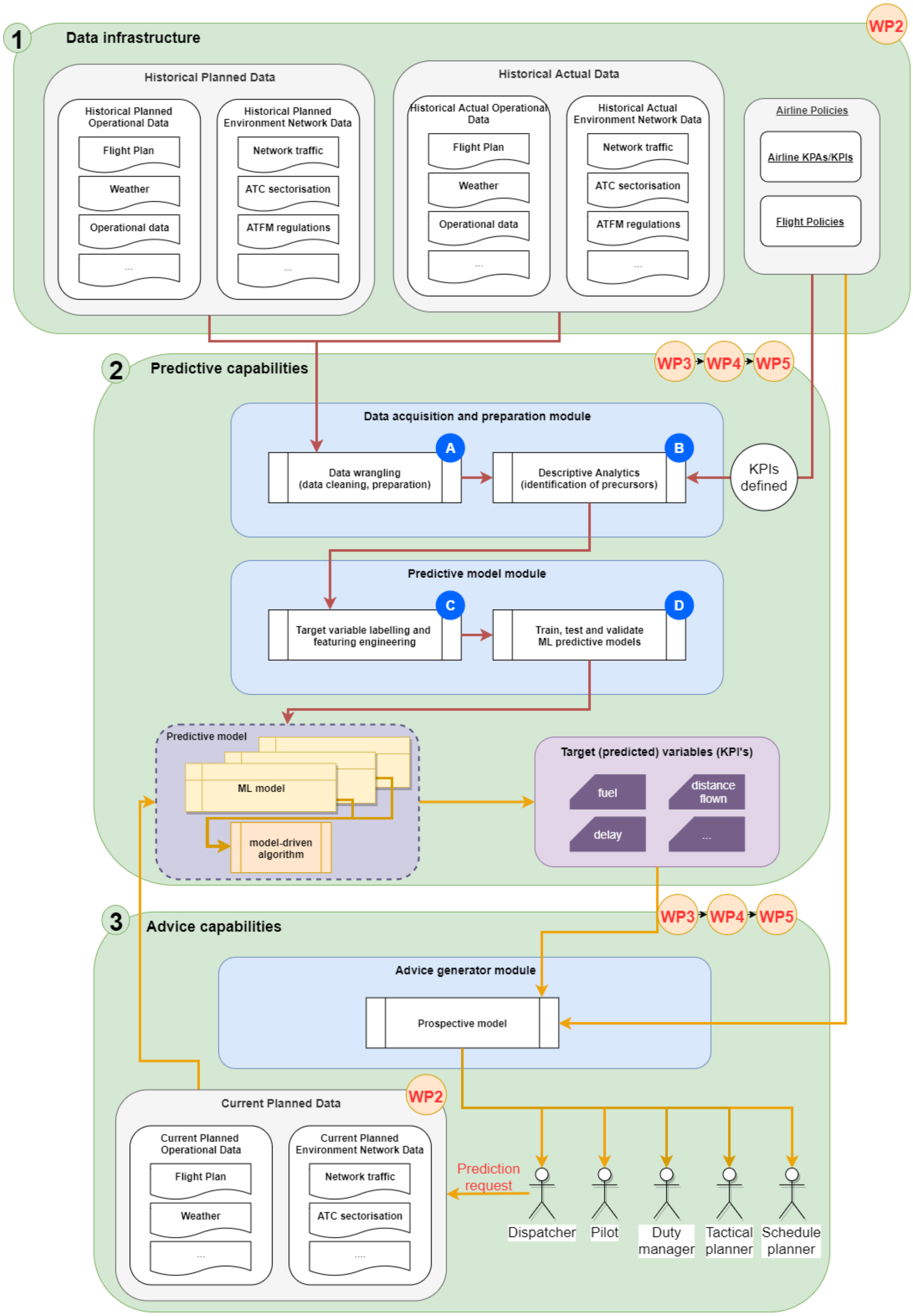

Figure 3. Dispatcher3 prototype architecture

EUROPEAN UNION

2.4.1 Data infrastructure

The data workflow and algorithm development will be powered by **DataBeacon** (https://databeacon.aero). More detailed information on the data infrastructure is provided in Section 5.

DataBeacon is a **multi-sided, open source, data storage and processing platform** focussing on data ownership, confidentiality and data protection. DataBeacon leverages a full-stack AI environment that brings together operational experts with machine learning engineers. The key elements of DataBeacon are:

- **Private environments** that hold data but are not accessible by applications or analysts directly.
- **Secure Data Frames (SDFs)** are de-identified dataframes, private and sensitive fields are substituted by secure cryptographic hash functions to allow identification across different data sets.
- **The full-stack AI environment** is an on-demand cloud computing architecture for analysis and applications. After secure login, a complete data science development environment launches, including toolsets such as Python, Hadoop, Spark and the Anaconda ecosystem. Secure access to a Jupyter notebook allows analysts to work remotely with protected data – no data leaves the cloud. Additionally, a GitLab repository serves as a collaborative and monitoring tool.

Different data sources will be required in Dispatcher3 as reported in Section 4. These data will be prepared and merged in SDFs inside DataBeacon. This will provide, via a negotiated definition between analysts and data providers, a unique and standardised dataset for the different case studies considered (user-oriented). By defining and particularising the SDFs, we ensure that the training and the input data is adequate for the ML algorithms.

An iterative process is proposed to select the features that will describe the KPIs determined by the airline. This will enable analysts to work closely with aviation experts to extract the relevant variables for the machine learning problems, ensuring that both a clear understanding of the training data and learning models and domain knowledge are considered. This methodology will not only allow us to extract and engineer the most influential variables to support the flight plan decision-making optimization from several prospective users while maximising the performance of the trained ML algorithms, but it will be generic enough to provide a data infrastructure with potential to apply machine learning models for other activities beyond Dispatcher3.

2.4.2 Predictive capabilities

The predictive capabilities of Dispatcher3 will be developed in two different modules:

- **Data acquisition and preparation**, composed of two activities:
  - **Data wrangling (preparation and cleaning)**, which focuses on the acquisition of the data and their incorporation into the data lake. Once it is acquired it needs to be cleaned and prepared so that it can be used for the data analytics.
  - **Descriptive analytics**, using data mining techniques to extract the KPIs that will be used as target variable (variables to predict using ML models). It will also focus on identifying the precursors (e.g., based on variables correlation and aviation experts) that will be used to explain the prediction (e.g., having or not a sort cut in the

operation, being regulated by ATFM, crossing an area with a SIGMET weather) and at the end into an advice for the end user.

- **Predictive model**, which consists of:
  - **Target variable labelling and feature engineering:** Supervised machine learning algorithms work by training models based on a set of labelled data. The dataset will be annotated following the KPIs defined as the result of the descriptive analytics. Also, the selected precursors must be engineered from the raw data, calculating the variables necessary.
  - **Train, test and validate ML predictive model:** These activities consist on the actual training of the model which will provide the predictive capabilities.

Details on the different data processing and analytic techniques are provided in Section 6. In this section a summary of main characteristics of each phase is provided.

2.4.2.1 Data acquisition and preparation

Data wrangling (data preparation and cleaning)

This first step focuses on accessing the data and executing the data preparation tasks to obtain a minimum readable dataset that enables basic data analytics. Most of the data sources are not perfectly clean and well-documented, therefore these activities are required in order to understand the structure and idiosyncrasies of the data we are dealing with. Data wrangling will, therefore, comprise the processes of acquiring, cleaning and storing individually each of the data sources involved in the project. Dispatcher3 aims at automating this data preparation process as much as possible in order to facilitate the further acquisition of datasets and the usability of the tool for different cases than the one tested in the project.

Descriptive analytics

Descriptive analysis is the first approach to extract patterns and relations between variables using statistical inferring. The main objectives of these processes are to have a general view of the distribution of the data: extracting common patterns, outliers, and identifying unseen correlations between the vast number of variables that compose our dataset.

2.4.2.2 Predictive model

Target variable labelling and features engineering

The objective of the target variable labelling and features engineering is to create the dataset required for the training and test of the machine learning models. Once the KPIs have been extracted from the descriptive analysis and the list of precursors have been designed following experts' recommendations, these will be translated into the dataset. KPIs are the target variables to predict and a list of potential precursors will be used as input features to feed the models.

Train, test and validate predictive model

Analysing historical data, predictive analytics uses machine learning techniques, such as data mining and predictive modelling, to not only predict a probability of a known outcome, but to also discover hidden patterns and arrive at new conclusions about a known problem. The proposed machine learning problems in Dispatcher3 can be solved using **supervised learning strategies**. These techniques aim to learn a function that maps an **input X** (features matrix) to an **output y** (labels vector), based on

EUROPEAN UNION

examples (X,y) (features, label pairs) mapped as **y = f(X)**. For further information on this, please refer to Section 6 on data processing and analytics techniques.

As previously indicated, some model-driven algorithms might be required to estimate complex indicators, such as KPIs which require the consideration of the airline's network.

2.4.3 Advice capabilities

The activities of data acquisition and preparation module are used to prepare the dataset required to train different machine learning models. The activities of the predictive model module are more specific to the questions that we want Dispatcher3 to answer and in particular to which target variables we want to predict and what are the historical inputs (features) that better describe them. The role of the advice generator is to produce specific advice to stakeholders based on the subset of predicted KPIs. This module is therefore linked to the user policies. With the same predictions different advice could be generated based on the airlines' preferences (e.g., suggest to the pilot to increase the cost index in order to recover delay, or not). The final customers of the prospective (prescriptive) model are the different roles identified in Section 2.3.

The goal of the advice generator module is therefore to collect all the information from the predictive analytics obtained, including information about the quality of the prediction (accuracy, precision, recall, etc.) and build a **decision framework**. It is particularly relevant to consider the uncertainty on the predictions, which might vary at different time-frames and with different uncertain inputs, to produce robust decisions for the users. Probabilistic approaches to quantifying uncertainty forecast, if deemed appropriate or feasible, will be explored. This decision framework should aid all the final customers to understand the predictions produced, stressing the importance of the ML interpretation and providing a hands-on advance model interoperation. The goal is to avoid the "**information overflow**", particularly considering the probabilistic nature of the information generated. Visual analytics will help to mitigate the potential overflow of information and mitigate potential **impact of information automation** on the system. The advice generator will require to consider accuracy versus interpretability trade-offs, and guarantee that both the dispatcher and the pilot understand the predictive analytics provided and are equipped to understand the probabilistic nature of the information.

The advice generator delivers the prediction to different users which need to be synchronised in their understanding of the results that the systems are delivering, despite the fact that they may receive different details at different times:

- The pilot will receive the results from the prescriptive analytics (advice generator) right before departure. At this point, some things may be fixed like for instance an ATFM regulation. By its nature the advice will be qualitative and linked to the precursors and the policies which underpin these decisions.
- The dispatcher will receive the results before the flight is dispatched (a few hours before the flight).
- The duty manager may receive results on the flights affected by deviations from planning from a few hours before the flight to several hours prior the schedule.
- The tactical planner may focus on information which affect the planning of the fleet and the operations one day ahead.

- Strategically, the output from the system might be relevant, as the airline may want to take longer term decisions about, for instance, their schedules, route structure, or even flight policies (such as fuel uplift or delay recovery). This is out of scope of the project, but Dispatcher3 sets the infrastructure for future inclusion of these concepts. These decisions will therefore impact the scheduling and potentially airlines policies.

The **usability** of the Dispatcher3 concepts is critical for the project team. The solutions proposed will meet usability requirements specified by the users. Dynamic dashboards capable of crafting visual stories to explain complex concepts will be particularly useful in distilling and sharing the information with peers. Maintaining information and communicating visually complex concepts enable certain rationale to be presented showing the relationship between variables and meaningful metrics. The different techniques depend not only on the structure of the data or type of datasets, but on the actual values. Storytelling is an inherently collaborative activity with different user types. Furthermore, little work has been done on collaboration in (big) data visualisation within the aviation field.

2.4.4 Verification and validation

Two type of verification and validation activities will be carried out during the duration of Dispatcher3: internal and external. The internal activities involve the members of the consortium and the Topic Manager. They will be performed during the development of the prototype. Once a first version is available, external validation of the tool will be done with the input from airlines from the Advisory Board and in a dedicated workshop. In some cases, sensitivity analysis of the output of the predictions with respect to variations in the input will be performed to capture the effect of uncertainty at an input level.

2.4.5 Interaction with stakeholders

Dispatcher3 will develop independent user-oriented models providing specific predictive and advisory capabilities. In order to ensure that these requirements are properly considered, the project will rely on continuous interaction with experts and stakeholders beyond the industrial partners of the consortium (**PACE** and **Vueling**). This will be achieved with continuous interaction with a broader stakeholder community (airlines, network manager, consultancy), via the **Advisory Board**.

If required, site visits are planned to discuss technical details on the implementation of the prototype. The Advisory Board has been consulted (with a workshop, follow up bilateral meetings and a consultation) to develop the main concepts presented in this deliverable. Particular effort will be allocated to gather information on how to provide advice based on the predictions considering airlines policies when required in the project evolution.

As discussed in Section 2.4.4, validation activities will be carried out internally in the consortium but also presenting the prototype to airlines in a one-to-one interaction and as part of a workshop once the first prototype is created. The Advisory Board will be open to incorporate new members if deemed necessary.

Finally, Dispatcher3 will focus on specific case studies which will be defined around specific origin and destination routes. The selection of these case will be performed at the beginning of the project in cooperation with the Topic Manager and driven by Vueling needs and Skeyes and PACE advise (see Section 0 for more details on the current identification of preliminary scenarios). These will also be consulted with the Advisory Board to ensure that different relevant operational environments are captured during the project development.

EUROPEAN UNION

2.5 Dispatcher3 technical activities

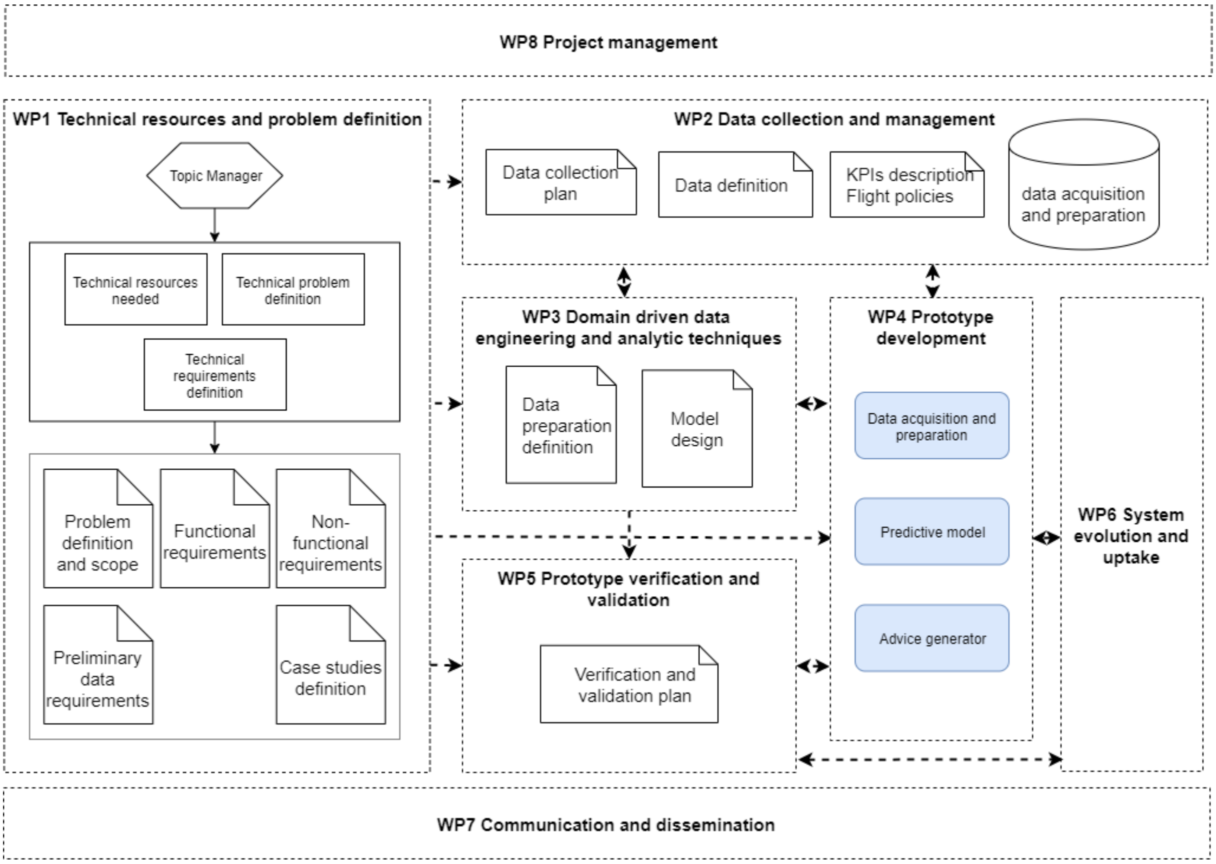

Figure 4. Dispatcher3 workpackage structure overview

Figure 4 presents the high-level workpackage structure of Disptachter3. WP1 (and D1.1) incorporate the first technical activities with a formal definition of the problem tackled, its scope and case studies. These activities have been done seeking input from the Advisory Board and the Topic Manager.

Firstly, WP1 defines the high-level functional and non-functional requirements (see Section 3). The preliminary data and data infrastructure requirements are defined in Section 4 and Section 5 and will be further developed as part of WP2. WP2 will finalise the data definition and their collection plan, and develop the technical activities of data acquisition including the deployment of the data infrastructure.

Section 6 identifies a preliminary set of data processing and analytic techniques. This is the trigger of the activities of WP3 where the data engineering techniques will the technically be defined. This includes the description of the technical solutions to clean, pre-process and prepare the raw data to be used; the identification of techniques required for the processing of data to create the training dataset for the machine learning models; the identification and description of the machine learning techniques that will be used to generate the predictive models; and the scoping of techniques available for advice generation. The prototype will be developed as part of the activities of WP4 grouped in three different sub-modules: data acquisition and preparation, predictive models and advice generator.

A preliminary indication of the type of scenarios that will be considered are presented in Section 7. WP5 will define the verification and validation approach and refine the scenarios and case studies to be evaluated. Input from external stakeholders will be considered during these validation activities.

The final technical workpackage is WP6, which will identify the changes required between the first and the final version of the model, while identifying the future development required for the potential industrialisation of Dispatcher3.

3 High-level prototype requirements

This section comprises the high-level prototype requirements for Dispatcher3. Note that the detailed requirements for the different modules will be developed and captured as part of the activities performed in WP3 – Domain driven data engineering and analytic techniques, WP4 – Model development; and validation requirements will be captured in WP5 – Model verification and validation.

The section includes information on the structure of the requirements (Section 3.1), functional requirements (Section 3.2) and non-functional requirements (Section 3.3).

3.1 Requirements structure

All the requirements of Dispatcher3 prototype will be documented according to the following format (**Table 5**):

Table 5. Format of requirements

Identifier	Version:	<x.y>	SW release:	<a.b>	Priority:	{1,2,3,4,5}
Description	<text>					
Rationale	<text>					
Validation	<text>					
Origin	<text>					

- Identifier: D3-{FR,NFR,DR}-{SYS,DAP,PM,AG,HMI,INT}-nnn
    - D3: Dispatcher3 (fixed tag)
    - FR: Functional Requirement – NFR: Non-functional requirement – DR: Domain requirement
    - SYS: High level system – DAP: Data Acquisition and Preparation – PM: Predictive Model – AG: Advice Generator – HMI: Human Machine Interface – INT: integration
    - nnn: 3 digit number initially increasing in steps of 10
- Version: identifies the version number of the requirement.
- SW release: identifies the software (SW) version of the release in which the requirements should be implemented.
- Priority: 1 for highest priority, 5 for lowest priority for not-binding provisions.

- Description: usage of "shall" for mandatory requirements and "may" for not-binding provisions.

Note as the validation activities will be defined in WP5, the validation information is not provided for these preliminary requirements.

3.2 Functional requirements

Table 6. D3-FR-SYS-010

D3-FR-SYS-010	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** provide a data infrastructure which enables cleaning, synchronisation and merging of past flights data along with other relevant operational data.		
Rationale	The user will be able to access a platform were operations with past flights data, including flight plans and actual flight execution data along with other relevant operational data, such as operational environmental data or weather can be implemented. The platform will provide (partially) automated methods to ingest, clean and prepare aviation and other data for analysis using machine learning techniques.		
Origin	D1.1		

Table 7. D3-FR-SYS-020

D3-FR-SYS-020	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** provide predictive capabilities based on advanced data analytic techniques for the role of dispatcher a few hours before the flight.		
Rationale	Based on advanced data analytic techniques, the dispatcher shall be able to run predictions to estimate targeted indicators for flight plans. Such as: • Holdings (non-nominal conditions) • Fuel tankering • Time deviations • Fuel deviations • Taxi times and fuel		
Origin	D1.1		

Table 8. D3-FR-SYS-030

D3-FR-SYS-030	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** support the understanding of the causes leading to possible variations between planned and executed flight plans.		
Rationale	The user shall be able to identify the precursors that impact the prediction of selected indicators and the variances between planned flight plans and their executed realisation.		
Origin	D1.1		

Table 9. D3-FR-SYS-040

D3-FR-SYS-040	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** provide an advice generator module which transforms the outcome of the predictive engine into advice according to airline policies for the role of pilot before push-back.		
Rationale	The pilot will receive actionable indications. For example, he will be informed about the possibility to recover some delay in the air.		
Origin	D1.1		

Table 10. D3-FR-SYS-050

D3-FR-SYS-050	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** provide an advice generator module which transforms the outcome of the predictive engine into advice according to airline policies for the role of dispatcher a few hours prior the flight.		
Rationale	The dispatcher will receive actionable indications.		
Origin	D1.1		

Table 11. D3-FR-SYS-060

D3-FR-SYS-060	Version: 1.0	SW release: 1.0	Priority: 2
Description	Dispatcher3 **may** provide predictive capabilities for the role of tactical planner on the day prior-operations.		
Rationale	The tactical planner may use this functionality. Specifically, they may be able to: • Identify flights potentially affected by disruptions (delay) • Identify congestion in network impacting flights		
Origin	D1.1		

Table 12. D3-FR-SYS-070

D3-FR-SYS-070	Version: 1.0	SW release: 1.0	Priority: 3
Description	Dispatcher3 **may** provide predictive capabilities for the role of duty manager many hours prior to the flight		
Rationale	The duty manager may use this functionality. Specifically, they may be able to assess: • Flights potentially affected by disruptions (delay) • Congestion in network impacting flights • Time deviations • Fuel deviations • Taxi times and fuel • Impact on reactionary delay		
Origin	D1.1		

Table 13. D3-FR-SYS-080

D3-FR-SYS-080	Version: 1.0	SW release: 1.0	Priority: 4
Description	Dispatcher3 **may** provide predictive capabilities for the role of duty manager few hours prior to the flight		
Rationale	The duty manager may use this functionality. Specifically, they may be able to assess: • Flights potentially affected by disruptions (delay) • Congestion in network impacting flights • Time deviations • Fuel deviations • Taxi times and fuel • Impact on reactionary delay		
Origin	D1.1		

EUROPEAN UNION

Table 14. D3-FR-SYS-090

D3-FR-SYS-080	Version: 1.0	SW release: 1.0	Priority: 5
Description	Dispatcher3 **may** automatically assess different flight plans for the role of dispatcher a few hours before the flight		
Rationale	The dispatcher may use this functionality to avoid areas with turbulence or with ATFM regulations		
Origin	D1.1		

D3-FR-SYS-020	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 **shall** provide predictive capabilities based on advanced data analytic techniques for the role of dispatcher a few hours before the flight.		
Rationale	Based on advanced data analytic techniques, the dispatcher shall be able to run predictions to estimate targeted indicators for flight plans. Such as: • Holdings (non-nominal conditions) • Fuel tankering • Time deviations • Fuel deviations • Taxi times and fuel		
Origin	D1.1		

3.3 Non-functional requirements

Table 15. D3-NFR-SYS-010

D3-NFR-SYS-010	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **shall** provide a cloud-based data storage platform.		
Rationale	The user shall be able to securely and efficiently store the data from various sources in a data lake.		
Origin	D1.1		

Table 16. D3-NFR-SYS-020

D3-NFR-SYS-020	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **shall** provide a generic data infrastructure which might be applied for other projects/goals.		
Rationale	The data infrastructure shall be able to ensure that the gathered data can be used not only for Dispatcher3 but also for future applications.		
Origin	D1.1		

Table 17. D3-NFR-SYS-030

D3-NFR-SYS-030	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **shall** provide adequate functionalities to protect the privacy of the stored data.		
Rationale	The prototype shall ensure that confidential data is de-identified using cryptographic techniques and in such a way, the privacy of the data stored in the platform is ensure.		
Origin	D1.1		

Table 18. D3-NFR-SYS-040

D3-NFR-SYS-040	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **shall** provide adequate functionalities to protect the privacy of the data while processing it.		
Rationale	The prototype shall ensure that the datasets can be securely merged and thus used in various computational tasks, whilst ensuring that the confidentiality and privacy of the data sets is not breached.		
Origin	D1.1		

Table 19. D3-NFR-SYS-050

D3-NFR-SYS-050	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **shall** provide a secure data science development environment.		
Rationale	The prototype shall provide secure sandboxed environments in which developers and analysts can work in a secure and user-friendly way.		
Origin	D1.1		

Table 20. D3-NFR-SYS-060

D3-NFR-SYS-060	Version: 1.0	SW release: 1.0	Priority: -
Description	Dispatcher3 prototype **may** provide data visualisation tools and interactive dashboards.		
Rationale	The user may be able to easily interpret and analyse the results of predictive models.		
Origin	D1.1		

EUROPEAN UNION

4 Preliminary data requirements

The following section contains a description of the identified data sets to support Dispatcher3, including a brief description summarising the main properties, the type of data (actual or planned), along with information on their purpose, and on data availability.

Data are crucial for Dispatcher3 as the project focuses on the processing and analysis of different data sources. Some datasets are owned by airlines and relatively ready to be ingested by the system, while others are out of the airlines domain and therefore their potential use might be restricted due to data availability and completeness. In some instances, the limitation of not having a specific dataset might be overcome with some estimation of the relevant parameters by the mean of other modelling or data preparation activities. Dispatcher3 will develop a prototype, and thus, in some cases, the examples presented will be considered as illustrative and in a production environment further datasets and data pipeline might be required. Therefore, note that the specific case studies and analysis performed might be adjusted as a function of the data availability.

Finally, as presented below, the main source of data will be data from Vueling airlines (as member of the consortium).

This section presents a preliminary identification of datasets as they will be fully identified and acquired as part of the activities of WP2 – Data collection and management.

4.1 Data management

An accurate, precise and efficient data management is essential to the successful development and implementation of the project, as explained in Section 0.

One of the goals of Dispatcher3 is to collect and prepare flight data in order to apply machine learning techniques. Therefore, different data sources will be required. Preliminary data requirements are identified in this section; while WP2 is a dedicated WP for data collection and management. As part of WP2, a full data collection plan and formal data definition will be provided (D2.1 – Data definition and processing report (due M10 (March 2021)).

Particular focus will be given to the different aspects of the data management and processing including:

- **Data management** for providing data to the machine learning procedures, which usually requires parsing, cleaning and merging datasets from a large variety of heterogeneous sources.
- The **data infrastructure** definition capable of ingesting the data at the required level of volume, variety and velocity, and capable of providing adequate computing power for the machine learning procedures. A first identification of these data infrastructure requirements and capabilities are provided in Section 0.

EUROPEAN UNION

- **Data Analytics** is the core practice of data science wherein machine learning techniques are applied to problems of interest. **User engagement** is important for enabling human operators to actually extract and analyse the relationships between observed variables, synthesised models and metrics of interest. **Data protection** is a fundamental factor to be addressed in practice to make sure that massive datasets are only used by authorised actors and for authorised purposes.

4.2 Data processing

As presented in Section 6 there are different activities that are required to process and prepare the data in order to be ingested by the training algorithms of the machine learning techniques used in the predictive models. Feature engineering pipeline which deals with continuous time-series data will be defined. In particular, Dispatcher3 will provide:

- parsing and error correction of raw datasets,
- data cleaning,
- standardisation of data inputs,
- statistical processing which might require model-driven approaches for data reconstruction and preparation, and
- de-identification of fields, as required, to ensure that raw values are not stored in the data infrastructure.

After these processes, the datasets will be ready to be used in particular machine learning problems, which will include the tasks of:

1. descriptive analytics to statistically describe the data and identify potential precursors,
2. feature engineering and data labelling (preparation of the training dataset),
3. training of machine learning models to provide predictive capabilities, and
4. validation of the trained models.

4.3 Datasets identified

Dispatcher3's consortium has already identified three different data categories that will be considered during the project. The full definition of datasets will be part of WP2 – Data collection and management and reported in D2.1 – Data definition and processing report:

- **Airline data:** including flight plans, actual flight executions, airline policies, etc. In order to estimate the impact of different flight plans on the airlines' indicators, a full representation of the airline situation will be required.
- **Operational environment:** including information on the operational conditions of the ATM network (e.g., Route availability, ATFM regulations, demand, sectorisation) and meteorology.
- **Other data sources**: besides the previously indicated for additional data processing.

Note that in some cases, the actual historical data should be considered (e.g., actual flight plan executed). However, in most instances, even if useful to gain insight on how the operational environment evolves, Dispatcher3 is concerned about the planned information at the moment of performing the different actions (e.g., planned demand at arrival airport at the moment of finalising the flight plan) to predict the outcome of the flight plans (actual flight performance experienced).

Table 21 presents the preliminary datasets identified with considerations on their need for the project and availability. The preference for Dispatcher3 will be to use real data from relevant stakeholders, however, as previously indicated it might be the case that some datasets are not available but the data could be estimated enabling the creation of the prototype pipelines and processes. For this reason, the table contains a suggestion on how some datasets could be estimated. This will be further explored in WP2. Note that the focus of Dispatcher3 is to predict the operational indicators as a function of the data available at the moment of making the prediction. This means that most of the data required is composed of estimations and forecast (e.g., forecast weather available at the moment of triggering Dispatcher3). However, the actual realisation (e.g., actual weather or actual congestion) could be useful for the analysis and understanding of the datasets and the deviations between planned and realised flight plans.

Table 21. Preliminary identified data sources required in Dispatcher3

Data	Type (Actual or Planned)	Required or beneficial	Specific purpose in the project	Provider	Could be estimated?
			Airline data		
Computerised flight plans (CFP)	Actual	Required	• Estimate difference between planned and actual. • Have a pool of potential 'alternatives' to flights.	Vueling	Could be estimated from historical trajectories but with limitations.
Actual flight trajectory executed (FDM data)	Actual	Required	• To estimate actual performance of flights.	Vueling	From trajectory integration models.
Flight policies	-	Required	• Identify dispatcher role/actions. • Identify KPIs relevant to predict	Vueling / AB	-
Airline schedules	Planned	Required	• Analysis of the impact on the demand	Vueling	Historical datasets (e.g., from trajectories).
Passenger itineraries	Actual	Beneficial	• Estimate the impact of different flight plans on the airlines' KPIs	Vueling	Passenger models could be used.

Data	Type (Actual or Planned)	Required or beneficial	Specific purpose in the project	Provider	Could be estimated?
Operational costs	Actual	Beneficial	• To calculate/estimate cost of disruptions and to provide better advice to dispatcher/pilot	Vueling	Models on costs of delay.
			Operational environment		
Weather forecast (temperature, wind)	Planned	Required	• Estimate operational parameters at moment of decision. • Estimate difference between forecasted and actual weather.	ECMWF, METAR	-
Actual weather encountered (temperature, wind)	Actual	Beneficial			It could be derived from FDM, e.g., by analysing ground and air speed.
ATFM network status (regulations)	Planned	Beneficial	• Status of network to predict variations between planned/actual due to congestion.	Network manager	-
	Actual	Required			
Route availability (RAD, NOTAM)	Planned	Beneficial	• Impact of route availability on performance/flight plans	Vueling / Network manager / skeyes	-
Other network operations (e.g., military, special events, etc.)	Planned	Beneficial	• Analysis of the impact of military or special events on possible disruptions	Network manager / skeyes	-

Data	Type (Actual or Planned)	Required or beneficial	Specific purpose in the project	Provider	Could be estimated?
Demand and capacity at airports	Planned	Required	• Impact of airport demand on performance	Network manager / skeyes	From planned demand at airports.
	Actual	Beneficial			-
Demand and capacity at airspace	Planned	Required	• Impact of airspace demand on performance	Network manager / skeyes	From planned demand at airspace.
	Actual	Beneficial			-
Sectorisation	Planned	Beneficial	• Impact of airspace sectorisation/usage on performance	Network manager	-
Radar trajectories	Actual	Beneficial	• Predict demand on ATM resources	Skeyes	-
Traffic trajectories (ADS-B)	Actual	Beneficial	• Possibility to use the trajectory integration engine from trajectory datasets	ADS-B providers	-
ATC commands	Actual	Beneficial	• Analysis of impact on ATC	Skeyes	-

EUROPEAN UNION

Data	Type (Actual or Planned)	Required or beneficial	Specific purpose in the project	Provider	Could be estimated?
CRCO charges	Actual	Beneficial	• Impact of en-route charges on route selection	Network manager	-
Air traffic statistics (e.g., eCODA, ATFCM and STATFOR)	Actual	Beneficial	• Statistical information on the air traffic delay in Europe for generating realistic scenarios to simulate	Network manager	-
Other					
Aircraft performance models (BADA)	-	-	• Possibility to reproduce the geometric, kinematic and kinetic aspects of aircrafts' behaviour over the entire operation flight envelope (trajectory simulation)	Consortium members	-

EUROPEAN UNION

5 Preliminary technical infrastructure requirements

An accurate, precise and efficient data management is essential to the successful development and implementation of the project. Dispatcher3 partners have more than 30 years of combined experience in working with aircraft performance, meteorological, air traffic, passengers, airline and economic/financial data. Each partner of the Dispatcher3 consortium has extensive experience in handling vast amounts of data of this sort.

The machine learning algorithms and the software implementation of those algorithms are a key element in designing AI applications. However, most of the effort in these projects tend to be spent on data wrangling, acquiring and cleaning data, creating data pipelines and ensuring the right governance model to guarantee the security of the datasets. Additionally, counting with a well-integrated software solution that runs on a scalable cloud computing accelerates the full lifecycle of AI projects, whether they are in their initial inception phase or in the last stages of deployment. Building AI on a solid platform is therefore an essential foundation of an AI application or project.

Dispatcher3 will for rely on the platform **BeSt** (Beacon Stack), developed in the last several years and managed by the company **DataBeacon**. Through the usage of this platform the first and second non-functional prototype requirement (see D3-NFR-SYS-010 in **Table *15*** and D3-NFR-SYS-020 in **Table *16*** of Section 3.3) will be satisfied. Dispatcher3 aims at progressing beyond the current state of the art on the following areas that are **not covered by existing data infrastructures**:

1. **Enhancement of data management pipelines for aviation**, designing automated and robust flows which encompass tasks such as data ingestion, cleaning, processing, and merging with other sources (including ATM and meteorological sources) so that they are fully prepared for analysis using ML techniques in a development environment;
2. Taking advantage of the **flight plan historical data,** assessing the deviations between planned and actual flight conditions to compare the expected KPIs with the actual KPIs and to detect unexpected disturbances during the flight;
3. Preparing the **infrastructure for novel cases studies**, ready for applying machine learning and deep learning techniques to perform predictive analytics over data available at planning stage so to predict their evolution once the flight is being operated;
4. Analysis of the impact of **external (to the airliners) factors**, such as weather, traffic congestion or sectorisation constraints, which might have had an impact on the expected-actual KPIs deviations;
5. **Researching airline policies applications** during the actual flight execution and **labelling the flight data with the expected KPIs** to examine their impact while recommending alternative flight plans leveraging on ML techniques;

EUROPEAN UNION

6. Providing a complete and reliable **prototype**, automating the data gathering and model execution to obtain the predictions in a near production-ready environment.

5.1 BeSt: Data storage and AI platform

BeSt is a scalable, secure, on-demand multi-side computing (MSP) and data storage platform that allows fast deployment of AI applications in aviation. It securely fuses datasets and runs computations over private, confidential data that are isolated from the rest of the platform.

In an MSP participants are usually both data providers and consumers of analytic services, that interact through a platform using secure common exploitation of data, improving their performance among various aspects of their business. These interactions are funded over an open, participative IT infrastructure as well as a global governance model. The goal is to consummate matches among users and facilitate the exchange of data and applications, thereby enabling value creation for all participants.

BeSt relies on three computing layers to ensure security, scalability and flexibility for a variety of big data and AI applications for aviation. Data protection is a key topic of interest for every industry and aviation data records are heavily protected by airlines and airports. Considering this, traditional analysis requires de-identification techniques, which has important limitations for AI applications, impeding data fusion. DataBeacon overcomes those challenges allowing applications to securely exploit these datasets.

BeSt has been used in previous initiatives leading to an extensive data catalogue such as aircraft performance data from several airlines complemented by radar data from several providers. In the past, the platform has hosted and processed data from several providers like ENAIRE, Austro Control, LFV, Eurocontrol, SITA, Airports Council International, IATA, EASA, Air France, METSAFE, Lufthansa, Aeroport De Paris, Frankfurt Airport, Flight Aware, Open Sky and parsed data from other public providers, like meteorological data. Besides Dispatcher3, BeSt is also being considered for other applications with airlines, air navigation service providers and airports.

5.1.1 Advantages and benefits of BeSt

Some advantages and benefits of using BeSt are:

5.1.1.1 Data collection with a Data Protection Agreement

In most cases data providers do not have the capabilities or resources to collect and prepare the data to be ingested into third party systems. For instance, airlines' Safety Manage System usually analyse data from Flight Data Monitoring (FDM) using proprietary software with limited exportation. Similarly, data are usually used for operational purposes (e.g., air traffic data) and stored under legal requirements, but not shared. This limits the progress of AI applications in aviation. DataBeacon provides a framework to overcome these data collection limitations. Privacy concerns and legal requirements are addressed through a governance agreement with the data owners called Data Protection Agreement (DPA). The DPA provides the general terms and requirements to securely upload and store the data into DataBeacon alongside a few Annexes that specify and limit the usage of the data in particular analyses or applications. In general, a DPA is signed once and updated while Annexes are signed any time a new application is under development.

EUROPEAN UNION

5.1.1.2 Data protection and integrated data processing

Secure Data Fusion

The main advantage of MSP is the combination of data from different sources to generate richer datasets that can power AI applications. The main blockage for this fusion is the privacy restrictions of the different datasets.

In order to overcome this privacy issue BeSt platform incorporates the **Secure Data Fusion (SDF)** technology. For instance, airlines usually share their flight data only in aggregated manner or without identifiable fields to ensure privacy of crew operations. However, unidentifiable datasets cannot be matched with external datasets limiting the context of the data set; and with a limited context datasets lack variety and hinder the potential of AI algorithms as learning from them is limited. The **Secure Data Fusion technology uses cryptography to hide private identifiable data** but still allows the data to be identifiable across different datasets solving this issue**.**

Private environments

BeSt uses a multi-layer architecture for data storage developed to preserve data privacy, protection and accessibility:

- From the data provider point of view, the first layer consists of a series of **private local nodes** one for each data provider, which **collect and store raw, identified data.** Data is pre-processed directly from input sources, and it is then protected and de-identified, as required, and pushed to the next layer.
- The second layer consists of several **storage and processing nodes**; one for each data provider. Data stored in this layer is de-identified, cleaned and standardised, hosted on isolated private clouds environments.

These two layers belong to each data provider and together compose a **private environment**. Private environments scale easily on the cloud, so that when new data provider joins BeSt, the private environment can easily be replicated and configured.

To make data accessible and useful to developers, BeSt implements **secure sandboxed environments in which analytics take place**. It consists of a **high availability, on-demand cloud computing platform**. After secure login, a complete data science development environment is launched, which includes popular data science tool sets such as Python, Hadoop, Spark and the Anaconda ecosystem. A secure access to a Jupyter Notebook allows analysts to work remotely with the data while **no data leaves BeSt platform.** Additionally, a GitLab repository instance serves as a collaborative and monitoring tool, enabling code versioning and continuous integration pipelines.

Private cloud environments hold data but are not accessible by analysts directly from the secure sandboxed environments. Instead, data is filtered and consumed by the Secure Data Fusion (SDF) technology. This allows enriching each isolated data set by combining multiple sources of data while respecting the privacy of the data owners

Despite the complexity of the platform, it can directly be accessed with a secure SSL connection on any compatible internet browser.

Figure 5 shows the overview of these layers in the platform.

EUROPEAN UNION

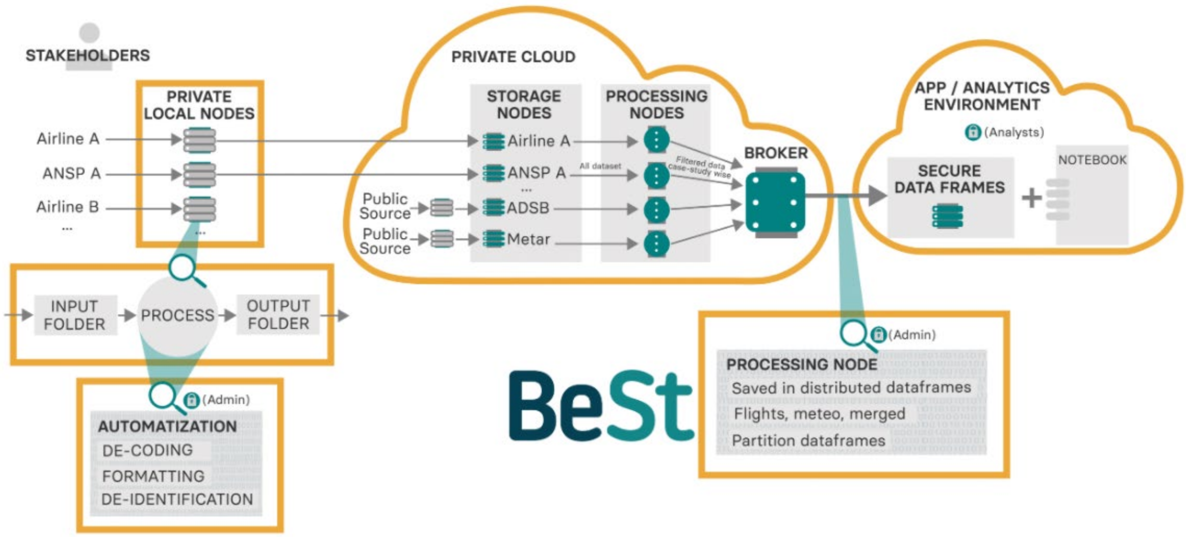

Figure 5. BeSt: Overview

5.1.1.3 Data visualisation

BeSt offers integrated support for Tableau interactive dashboards that provides an easy way to develop and deploy visual and easy understanding of analytics models. This will be exploited in case of developing the non-functional prototype requirement on providing data visualisation tools (D3-NFR-SYS-060, see **Table *20***).

5.1.1.4 Cloud computing

The flexibility of the cloud environment allows the platform to adapt dynamically to the demand, and cloud services comply with the highest standards on security, both physical and virtual.

Additionally, a global governance model is embedded into the platform with several layers of security and responsibility. The global governance model defines who and when has access to which resources and gives the data owners the mechanisms to:

- Join or leave the platform at any time, removing any trace of their data.
- Approve or reject new case studies or applications.
- Monitor data access and usage.

To ensure confidentiality, privacy and non-disclosure of the data, data owners and consortium members will follow, as required, the **DataBeacon's global governance model** (**Figure 6**), the model that consist of a data protection agreement (PDA) with general terms and a series of Annexes describing particular usages of such data, e.g., scenarios. In particular data shall maintain the confidentiality of any information that may, in any manner, violate the commercial secrecy of any particular before any use by the consortium, leaving only such data necessary for the analyses and modelling.

The activities within the project will not involve the collection and/or processing of any kind of 'sensitive data'. The parties understand that health, sexual lifestyle, ethnicity, political opinion,

EUROPEAN UNION

religious or philosophical conviction data can be identified as 'sensitive data', and are completely out of scope of the activities planned.

Since most of, if not all, the data needed by Dispatcher3 is subject to strict confidentiality constraints, at this proposal stage the Dispatcher3 project is not in the position to participate in the Pilot on Open Research Data in Horizon 2020. This decision will be immediately revised in case of changes in the disclosure conditions of the input data or in case of specific agreements with the data owners, and the appropriate Data Management Plan will be formulated.

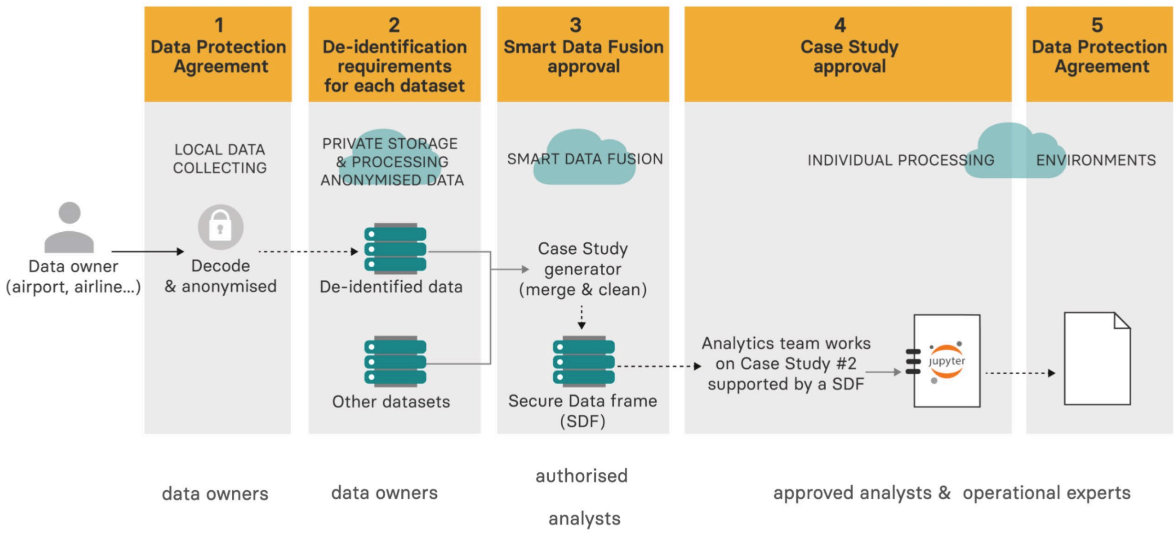

Figure 6. Governance model in BeSt

5.1.2 Data interfaces

In BeSt, there are several interfaces available for accessing data: Data Catalog (AWSGlue [1]), Athena intended principally for relational databases and users familiar with SQL queries, and data lakes intended to work with non-relational databases (AWS data wrangler, parquet and partitioning). All the data interfaces are going to be well documented and the instructions provided to the analysts working with data in Dispatcher3.

The characteristics of BeSt described in this section concerning governance model, data privacy, development environments, etc. are crucial for satisfying the non-functional requirements of the prototype (principally non-functional requirements P3-NFR-SYS-030, P3-NFR-SYS-040 and P3-NFR-SYS-050 and P3-NFR-SYS-050 see **Table *17***, **Table *18*** and **Table *19***).

If required by the project, in addition to the BeSt platform, UoW has the capability to host a database built on one of the well-known standards in the field (MySQL) which would be accessible only through a secured connection which will pass through a dedicated VPN from the user to the University of Westminster's server. All the consortium members will be required to register and access the data through the VPN connection. This will be considered if some network models are required to complement the datasets.

6 Preliminary data processing and analytic techniques identification

Dispatcher3 will predominantly rely on methodologies and tools from the area of data science. Data science is a set of disciplines that combines computer hardware, algorithms and statistics that, instead of describing a system through hypothesis testing and experimentation, tackles problems strictly from the data perspective. Machine learning is one of the most commonly relied on tools by data scientists and it refers to the ability to detect and extract the most likely patterns from data, thereby turning **raw data into knowledge**.

6.1 Data acquisition and preparation

The data acquisition and preparation is composed of two steps: data wrangling and descriptive analytics. Dispatcher3 aims at standardising these processes so that the datasets might be used for different functionalities beyond the ones studied in this project.

6.1.1 Data wrangling

This step focuses on accessing the data and executing the needed data preparation tasks to obtain a minimum readable dataset that enables basic data analytics. **Data wrangling** comprises the processes of acquiring, cleaning and storing individually each of the data sources involved in the project. Some data sources might require **parsing** and error correcting modules to read the information; for instance, those services hosted by third-party APIs such as public data sources, being necessary to parse, collect and save this information.

Once the different data sources are collected, the next step is data cleaning, which implies correcting wrongly decoded or missing values from the stored information. This will require a **data cleaning module** that checks that no data field contains erroneous values, applying corrections such as dropping or substituting the values if possible. Within this cleaning process it is also important to **standardise** the input datasets as much as possible. This understands adapting similar columns with similar properties under a common standard format, which highly improves the accessibility in terms of resources consumption. For example, establishing a shared agreement on measurement units used for a certain magnitude, such as *speed* or *heading*, and transforming all these values to the same unit (e.g., *knots* and *radians).* These activities contribute to the future fast loading of the data processes, avoiding to standardise the whole dataset every time a user wants to consume it.

In case data owners want to protect certain information, it is mandatory to perform a **de-identification** for those columns before the data has been ingested, ensuring that raw values of such sensitive columns are never stored in the platform. Therefore, sensitive columns will be always de-identified and never accessible by analysts, since they were hashed during data collection phase. The output obtained from this process will be a set of isolated, cleaned and de-identified data sources saved in

EUROPEAN UNION

the platform, ready to be inspected and merged with each other to run predictive analytics. This intermediate form of data is what is called in BeSt the **Secure DataFrame (SDF)**, as presented in Section 5.1.1.2.

Dispatcher3 aims at automatising this data preparation process as much as possible in order to facilitate the further acquisition of datasets and the usability of the tool for different cases that the one tested in the project.

As described above, these data cleaning and preparation activities will be supported by **statistical processing** (data driven). However, if needed, some **trajectories might be reconstructed** using a **model-driven approach** relying on aircraft trajectory prediction models. The combination of data- and model-driven methodologies will exploit the capabilities and synergies of both, allowing for the reconstruction of missing data points, contributing to the identification of outliers, smoothing noise datasets and filling up gaps. For this, UPC DYNAMO tool will be used, which is an aircraft trajectory prediction and optimisation engine capable to rapidly compute trajectories using realistic and accurate weather and aircraft performance data [5].

6.1.2 Descriptive analytics

When all the data are already ingested, cleaned and prepared in the BeSt platform, they will be considered ready to be used by the analysts. Descriptive analysis is the first approach to extract patterns and relations between variables using statistical inferring. The main objectives of these processes is to have a general view of the distribution of the data: extracting common patterns, outliers, and identifying unseen correlations between the vast amount of variables that compose our dataset.

In order to properly represent the impact of these variables, it is necessary to define several **Key Performance Indicators** (KPIs). Using the descriptive analysis results, a list of potential **precursors** that might influence the KPIs performance will be extracted. The precursors are features that have been considered to have high influence over selected KPIs, mostly determined by airline experts. For instance, an example of a KPI could be the probability of changing the planned route of a flight. Understanding this route as a sequence of waypoints between origin and destination airports, we would like to predict this KPI using machine learning techniques. To do so, we would perform a descriptive analysis to explore which are the most probable factors that influence route modification. This complete process is also called **precursors analysis**. These precursors will influence the variable that is predicted and in this example they could be, for example, the aircraft type, the weather present in sectors involved (e.g., rain, thunderstorms, dust), air traffic congestion near arrival and departure airports, fuel loaded into the aircraft, etc. Taking into account aviation experts experience, we would determine a *suitable list of precursors* that might impact our KPI metrics.

6.2 Predictive models

Once the datasets have been prepared and the indicators to be considered defined, the different machine learning models can be created. For this three distinct activities are required:

1. Target variable labelling and feature engineering, which objective is to prepare the datasets for the training of the machine learning models.
2. Machine learning models training, which perform the tuning of parameters of the machine learning models to improve their predictive capabilities.

3. Validation and testing of the trained machine learning models.

6.2.1 Target variable labelling and feature engineering

The objective of the target variable labelling and features engineering is to create the dataset required for the training and test of the machine learning models.

Once the KPIs have been extracted from the descriptive analysis and the list of precursors has been designed following experts' recommendations, these will be translated into the dataset for training machine learning models. In our machine learning problem, KPIs are the target variables that we wish to predict. Note that some of those KPIs could be defined as the difference between the obtained KPI in the planned and executed flight.

However, **most of these KPIs will not be directly available in the data collected but will need to be computed or estimated instead**. Hence, we will have to perform a **data labelling** process to identify which flights were affected by the different precursors and to compute the KPIs in our historical data. This way, we approach the problem from a **supervised learning** point of view, labelling all the target variables for each observation. Normally, data labelling task can be hard to automate due to the high amount of resources needed to check if the indicators (and precursors) are being accomplished or not.

In the same way, the precursors list obtained from the descriptive analysis will be the initial set of features to train our model. Probably, not all of them will be relevant for the model, so it is possible that this list has to be updated later on using engineered features. We understand as **feature engineering** the process of using domain knowledge about the data at disposal to create features useful to machine learning models to deliver accurate forecasts.

Normally feature engineering is an iterative process in which we continuously look for new features that better improve the performance of machine learning models, retrain the model and evaluate its performance. Some of these features will be **static**, such as airport names, dates, call sign, etc., but others will be **dynamic**, such as speed or heading values. Since machine learning model cannot handle a complete **time series** as an input, we will have to **sample dynamic features** at specific time points. The sampling can be performed with respect to time, altitude or distance, among others. It could be possible to describe features that group some of these time series into qualitative indicators. Additionally, it is important to provide the model only with the data that are available at the prediction point, thus we need to evaluate the availability of the features depending on the user and the imposed time-frame boundaries (driven by the user-oriented model).

During this process, datasets not currently available to the airlines can be tested in order to evaluate their potential benefit for the airlines in solving these kind of problems. An example of such a dataset could be information currently available at ANSPs.

6.2.2 Machine learning models training

This section describes some of the machine learning techniques considered in Dispatcher3. The problems we aim to solve are mostly of supervised nature, which means the data observations need to be labelled with the target variable that we aim to forecast. The target variable will vary depending on the particular use case or forecasting horizon. Additionally, while we intend to predominantly work with supervised machine learning algorithms, some unsupervised techniques might be considered as needed to analyse the input data and support feature engineering processes. Finally, as indicated in

EUROPEAN UNION

Section 2.2, some models might rely on simulation which uses as input the estimations of trained machine learning models, e.g., to estimate reactionary delay.

Complex problems with highly dimensional feature space usually require more advanced solving approaches than the ones relying on linear machine learning methodologies (e.g., lasso, ridge or logistic regression). With the onset of big data, more innovative and advanced machine learning models and methodologies have proven to be more useful in learning complex non-linear relationships between the input and output variables (in case of supervised learning) and thus delivering superior performance (most popular being various instantiations of artificial neural networks).

6.2.2.1 Overview of machine learning candidate models and techniques

The first step in the development of machine learning models is the definition of the problem in terms of machine learning (e.g., identifying the problem as classification or regression) and selection of the suitable predictive model, depending on the selected KPIs that we want to predict. Afterwards, the models are **trained and tested**. Note that Dispatcher3 will not produce a machine learning model but a set of models (one per KPI) and that the data inputs might vary as a function of the user and time-frame of the prediction. For those purposes, a number of models will be considered and their suitability will be assessed with respect to each particular indicator (KPI).

In continuation we present an overview of the machine learning techniques we consider could be useful for the problems we plan to solve in Dispatcher3. In addition to the techniques mentioned here, we plan to explore other approaches as the project advances and other models or techniques emerge as suitable candidates for a particular problem. This will be further explored as part of WP3 – Domain driven data engineering and analytic techniques.

Support vector machines with kernels are often a good solution with often very good performance [2]. They essentially calculate distance between data observations, and then find a decision boundary that maximises the distance between the closest members of separate classes, i.e., construct a **maximum margin separator**. This helps them generalise well. As an example, an SVM with a linear kernel is similar to logistic regression. They are also quite robust to overfitting and thus especially suitable for high-dimensional spaces. However, when dealing with a large amount of data they do not scale well. Learning times are almost unfeasible for large datasets because they are very difficult to parallelise and tuning them can be challenging due to the importance of selecting the right kernel.

On the other hand, **decision trees** [6] are a fast, rule based family of algorithms, and one of the simplest yet most successful forms of machine learning. They perform well on large datasets and can learn complex functions. Additionally, unlike many other models they provide high level of interpretability due to their rule based nature. However, they are highly unstable with a high risk of overfitting and creating biased trees, and they are nowadays often inferior in their performance by other more advanced and novel methodologies. More often than single trees, ensemble version are used, such as **random forest** [13] that ensembles a number of (weaker) decision tree learners thus delivering a better performing, more accurate model.

Artificial neural networks (ANNs) can approximate any function regardless of its linearity [4]. They are often the best solution for complex and abstract problems, because they are highly tuneable by configuring different architectures of neural networks. Although understanding each architecture requires vast knowledge of deep learning solutions, they are relatively easy to implement and use thanks to diverse **libraries** available (e.g., scikit, tensorflow, keras, theano, pytorch). Additionally, a number of different **ANN architectures are readily available** in open-source packages and can be used as it is or further adapted to the problem at hand. However, in many cases simpler solutions (random forest, SVM, etc.) are more suitable and solve the problem just as well. They also require high volumes of training data and cases, and they operate as a "black box" model, i.e., they are very difficult to interpret. Finally, they are computationally expensive, demanding a large GPU infrastructure or even a cluster.

In addition to more classical feedforward neural networks, innovative and advanced neural network architectures that are often more suitable for solving particular problems will be considered. **Long-short term memory (LSTM)** neural networks are a version of **recurrent neural networks (RNNs)** that rely on long-short term memory blocks to remember past information, i.e., to remember data samples that were "seen" by the model over a longer periods of time [12]. That makes them suitable for forecasting problems where it is very important and beneficial to take into account how some variables evolve over a longer time period, thus giving excellent results in time series forecasting. As it is the case with neural networks, they require a lot of training data and very complex feature engineering process. Another model that might be considered are **convolutional neural networks (CNNs)** [14] and their various versions (e.g., **temporal convolutional networks** for time inputs). They rely on the concept of the mathematical operation of convolution that enables them to capture spatial and temporal dependencies in the output data. A number of architectures of CNNs are available which makes their implementation more feasible: LeNet, AlexNet, VGGNet, ResNet, etc. Lastly, **autoencoders** could be considered if needed, e.g., for the tasks of dimensionality reduction, feature learning, and others [3]. They are feedforward, non-recurrent neural networks similar that learn how to copy its input to its output. Having an internal hidden layers enables them to code the input (encoder), and then reconstruct approximately the original input at the output (decoder), thus learning a representation for some dataset.

A very good compromise between performance and computational complexity of a model is often achieved by relying on **ensemble machine learning** methodologies. The idea of ensemble learning is to select a collection (an ensemble) of hypotheses from the hypothesis space and combine their predictions. Ensemble methods compose multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction [15]. They accept high dimensionality training sets for both the number of features and the number of samples. They are a very powerful class of techniques and they are very popular because they often outperform more complex and computationally expensive deep learning algorithms. In particular two families of ensemble algorithms are used: bagging and boosting frameworks [7].

Bagging ensemble methods use aggregations of multiple decision trees as weak learners. They can learn non-linear functions and present good performance on large datasets. In particular, they address class imbalance very well [8]. The most used bagging algorithm are random forests. **Boosting** ensembles are among the best, off-the-shelf, supervised learning methods available in terms of accuracy prediction performance. They are modest in memory and runtime requirements. There is no need to apply feature transformations for the algorithm to perform well. And can handle a mix of binary, categorical and continuous features. On the other hand, they are difficult to interpret, require careful tuning of hyper-parameters, and they are not useful when the feature space is very sparse. There exist two main boosting algorithms adaptive boosting and gradient boosting.

Adaptive boosting (AdaBoost) is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified datasets are higher [11]. Then the instances are adjusted such that subsequent classifiers focus more on difficult cases. **Gradient boosting machines (GBM)**, for example the LightGBM algorithm [9], is similar to AdaBoost but it does not modify the sample distribution. On the other hand, each weak learner trains on the remaining errors of the strong learner. At each iteration the errors are computed and using a gradient descent optimisation they try to minimise the overall error of the strong learner.

Table 22 presents an overview of the different machine learning models that could be considered in Disptacher3, note that different techniques might be assessed for differ

Table 22. Overview of considered machine learning models and techniques

Model/ technique	Short description, main characteristics	Advantages	Shortcomings	References
Artificial (feedforward) neural networks (ANNs; Multilayer perceptron)	Perceptron-based structures capable of learning complex and non-linear functions, suitable for solving very for complex and abstract problems.	• Easy to implement • Highly tuneable • state-of-the-art results	• Difficulty to interpret the features of the architecture • Require high volumes of training data • Require high computational capacity	[4]
Decision trees and their ensembles (e.g., random forests)	Decision trees are a rule based algorithm adopting a greedy divide-and-conquer strategy of always testing the most important attribute first. Random forest is a meta estimator that ensembles a number of decision tree learners.	• Address class imbalance very well • Fast with few parameters to tune • Interpretable • Able to learn complex functions • ensembles such as random forests able to improve the performance significantly • able to handle categorical features	• Risk of overfitting • Difficulty to tune the high number of hyper-parameters with possible of requirement of high demanding grid search • interpretability of ensemble can be questioned	[6,13]
Support vector machines (SVM)	Kernel-based machines that construct maximum margin separators between data observations.	• generalise well • robust to overfitting and suitable to high-dimensional feature spaces • can provide non-linear solutions	• do not scale well to larger datasets • sensitive to the kernel used and thus trickier to tune	[2]

Model/ technique	Short description, main characteristics	Advantages	Shortcomings	References
Long short-term memory (LSTM) network	A recurrent neural network suitable for using on time series data where dependencies over longer periods of time are relevant.	• capable of automatically extracting features from past events • performing well on time series • high precision and great for memorising evolution of features over longer period of time (time dependencies)	• complex process of data preparation and feature engineering • computationally very intensive • non-interpretable • requires a lot of data for training	[12]
Convolutional neural networks (CNN)	A type of neural network composed of multiple building blocks such as convolution layers, pooling layers, and fully connected layers, capable of learning complex spatial (and even time) dependencies in the input data.	• high precision and great for learning complex dependencies in the input data • capable of automatically and adaptively learning spatial hierarchies of features	• computationally very intensive • non-interpretable • requires a lot of data for training	[14]
HDBSCAN	A hierarchical clustering algorithm.	• fast and makes fewer assumptions on the distribution of the underlying data • gives clusters of arbitrary shapes	• very high time complexity and thus computationally intensive • very hyperparameter-sensitive and sensitive to outliers	[10]

EUROPEAN UNION

Model/ technique	Short description, main characteristics	Advantages	Shortcomings	References
Autoencoder	Neural networks that aim to copy their inputs to their outputs by compressing the input into a latent-space representation, and then reconstructing the output from this representation.	• great for data denoising and dimensionality reduction • can learn data projections more interesting than some more basic techniques (e.g., PCA)	• data-specific: restricted utility to data that is highly similar to its training data • not as efficient as more advanced techniques (e.g., generative adversarial networks)	[3]
Ensemble learning	A methodology that, instead of relying on one prediction from a weaker learner, collects (ensembles) a set of independent predictions from many learners. Most popular methods are bagging and boosting. This methodology yields various meta-learners (meta-estimators).	• high performance with low computational needs, often performing more complex data-hungry models • can learn non-linear functions and perform well on larger datasets • address class imbalance well	• difficult to interpret (ensembling weaker learners that individually might be interpretable produces a complex model that loses on interpretability) • require careful tuning of hyperparameters	[7, 8, 15]
Adaptive boosting (AdaBoost)	Ensemble machine learning procedure which combine several weak learners to form a single strong learner	• easy to read and interpret the algorithm • less susceptible to overfitting than other algorithms • computationally friendly and stronger than weaker learners (decision trees, SVMs, etc.)	• highly sensitive to noisy data and outliers – dataset should be cleaned up from outliers • extremely difficult to scale up	[11]

Model/ technique	Short description, main characteristics	Advantages	Shortcomings	References
Gradient boosting machines (GBM); specifically: LightGBM	GMB is an ensemble machine learning procedure which sequentially adds predictors to the ensemble and follows the sequence in correcting preceding predictors to arrive at an accurate predictor at the end of the procedure	• LightGMB uses histogram based algorithm, which makes it fast and efficient • LightGBM is low on computational needs (memory friendly) • high accuracy that often cannot be beat • performs well on larger datasets • very flexible – can optimise on different loss functions • Easy to read and interpret the algorithm	• prone to overfitting (cross-validation is helpful there); very dependent on outliers • GMB in general can be memory intensive (not LightGBM) • less interpretable than single learners	[9]

6.2.2.2 Model validation and testing

The models will be **validated** using cutting-edge techniques such as Grid Search or Random Search could be used. The **cross-validation** dataset will be used to perform this hyper-parameters selection process. **K-Fold** cross validation is one of the different strategies that can be followed for this. The data is divided in k subsets and the holdout method is repeated k times. Each time, one of the k subsets is used as the cross-validation dataset and the other k-1 subsets are used as the train. This methodology **reduces bias** as most of the data is used for fitting.

Lastly, once the model is trained, its performance will be assessed using a **test dataset** and the importance of the features will be measured. The objective is to assess the **importance of the precursors extracted from the descriptive analysis for produced forecasts**. These results will be finally validated with aviation experts. Standard metrics, such as prediction accuracy, precision or recall, and tools such as confusion matrix are going to be used to display the results and analyse how well the models are performing.

EUROPEAN UNION

7 Preliminary scenarios definition

Dispatcher3 is a data-driven project. This means that specific scenarios and case studies will vary as a function of the data availability; and the available data sources and their analysis will steer the creation of the scenarios and case studies. For example, it might not be possible to analyse the impact of major ATFM disruptions, such as ATC strikes, if the datasets available do not cover significant volume of historical observations of these type of disruptions; or the analysis of the data might indicate situations suitable or interesting for the use of machine learning techniques. Therefore, the specific use cases will be continuously reviewed alongside the data management and descriptive analysis processes. This will be managed by the activities of WP5 – Prototype verification and validation following an Agile methodology.

Besides this consideration, a set of preliminary scenarios and situations, which are more relevant for their operations, have been identified via specific consultation with the industrial partner from the consortium, Vueling and the Advisory Board. Further consultation will be carried out along the project to ensure fine tuning of the use cases in order to explore the full potential of the tool. Moreover, specific use cases might be developed considering the interest of individual members of the Advisory Board as part of the validation activities. This will be thoroughly elaborated in D5.1 – Verification and validation plan. The set of initial scenarios presented in **Table 23** have been developed in order to ensure that Dispatcher3 can provide the supportive information at different process levels considering different roles and their respective time-frames of interest.

From the three roles involved in the flight management process (the tactical planner, the duty manager and the dispatcher), as presented in Section 2 and captured by the requirements in Section 3, the project will focus on providing support to the dispatcher, with predictions and advice, and in addition, to the pilot, with qualitative and informative advice.

Specific predictions could be provided for the tactical planner and the duty manager roles. Support for the role of schedule planner and post-operational analysis (back office) are out of the scope of the project, but the outcome of Dispatcher3 can be particularly useful for analysing the flights that are systematically prone to variations and could be optimised (e.g., some route are persistently getting directs, some routes are having longer taxi-in time than expected, etc.).

Some of the characteristics of the scenarios for which Dispatcher3 is relevant include:

- Major disruptions in network (e.g., ATCOs strike, severe meteorological conditions, volcanic eruption, ATFM regulations etc.)
- Update of weather conditions (e.g., the use of updated weather forecast indicating strong head or tail wind in specific flight phase)
- Delay at destination TMA
- Different time of the day when disruption occurs (e.g., in the morning, evening, etc.)

In addition to these, there are situations which are more relevant for post operational analysis as their results may improve the operations of specific flights. For instance, the flights operating at the airport with an increase in the variability in taxi-out times may benefit from cost index adjustments performed at back office.

The members of the Advisory Board agreed that the consideration of non-nominal situations are more relevant (e.g., snow conditions in central Europe), as operational parameters, such as taxi times, might be affected in unexpected ways.

Table 23 presents an overview of a preliminary set of the type of scenarios considering different roles and different time-frames in which Dispatcher3 could potentially provide support. As mentioned, while these scenarios have been identified, in consultation with Advisory Board, as situations of particular interest where a tool such as Dispatcher3 could be very useful, it is important to stress that the scenarios implemented will be driven by available data. Note that some of the indicators highlighted in the different scenarios complement the identified in Section 2.3 and the research questions compiled in **Table *4***.

EUROPEAN UNION

Table 23. Preliminary scenarios considering roles and time-frames

Name of scenario	Short description	Type of route	Example of routes (routes considered)	Time frame	Example of indicators to estimate	Potential support to
Network very disrupted	Major disruption is expected in the network impacting portion of the flights (e.g., ATCOs strike, severe meteorological conditions, volcanic eruption, ATFM regulations etc.)	• Intra-ECAC • Long haul route	Airline flight plans for the entire network one day before the operation	24 h prior operation	• Identify flight plans that are more likely to be disrupted • Block times (time deviations) • Fuel deviations • Reactionary delay	• Tactical planner
		• Intra-ECAC • Long haul route		10h to 9h prior SOBT	• Identify flight plans that are more likely to be disrupted • Block time (time deviations) • Fuel deviations • Reactionary delay	• Duty manager
		• Intra-ECAC • Long haul route	Particular flight impacted by network disruption	4h to 3h prior SOBT	• Block time (time deviations) • Fuel deviations	• Dispatcher

Update weather forecast	X hours prior to departure, a weather forecast is updated indicating strong (head/tail) wind impacting the particular flight	• Intra-ECAC	• BCN- LGW • BCN-OSL • BCN-LPA • BCN-FUE	4h to 3h prior SOBT	• Block time (time deviations) • Fuel deviations • Fuel tankering	• Dispatcher • Pilot
Disruptive events with respect to meteorolo gical conditions	X hours prior to departure, the information about the specific weather (e.g., fog, snow, etc.) at arrival airport is available	• Intra-ECAC	• BCN-BRU • BCN-ZRH • BCN-AMS	4h to 3h prior SOBT	• Block time (time deviations) • Fuel deviations • Taxi-in time • Holding time	• Dispatcher • Pilot
Turbulence in the cruise phase	X hours prior to departure, a weather forecast is updated indicating severe turbulence in the cruise phase	• Intra-ECAC	• BCN-AMS	4h to 3h prior SOBT	• Block time (time deviations) • Fuel deviation • Turbulence indicator	• Dispatcher • Pilot
Dense TMA	Delay at destination TMA known X hours prior to departure impacting the particular flight	• Intra-ECAC	• BCN- LGW	4h to 3h prior SOBT	• Block time (time deviations) • Time deviation • Holdings • Taxi-in time • Arrival procedures	• Dispatcher • Pilot

EUROPEAN UNION

Particular flight affected by ATFM regulations	X hours prior departure, the information on new ATFM regulation is issued impacting the particular flight, while the network is not particularly congested	• Intra-ECAC	• BCN- AMS	4h to 3h prior SOBT	• Block time (time deviations) • Flight time deviation • Fuel deviation	• Dispatcher • Pilot

8 Next steps and look ahead

This deliverable has presented the high-level requirements and approach to the development of Dispatcher3. This deliverable is based on the proposal but incorporates the development produced during the first five month of the project. The interaction with the Advisory Board has been paramount to gain understanding on the airline flight management processes and to adjust the scope of the project.

The deliverable incorporates the seeds of the different technical activities which have already started on the different project tasks, in particular:

- The data definition identified in Section 4 is being further developed as part of WP2 – Data collection and management. The data definition and processing report (D2.1) will incorporate these considerations (due March 2021). The consortium is working on the identification of the datasets and their collection from different sources (e.g., Vueling, ADS-B provider) to incorporate them in the data processing platform.

- These activities will be performed in parallel to the scope and selection of the data acquisition and preparation techniques (WP3), reaching by March 2021 the milestone MS3 – Domain driven data engineering techniques identified, and the further identification of predictive and advice generator techniques. The outcome of WP3 will feed into the development of the models in WP4 and it will be reported in D3.1 – Data engineering and analytic techniques report (due May 2021) achieving milestone MS5 – Domain driven analytic techniques identified.

- The first models for the data acquisition and preparation module, the predictive model and the advice generator modules are being developed as part of WP4, even if the current focus is on the data acquisition.

- As indicated in Section 9, the use cases that will be implemented will depend on data availability and driven by the first analysis of the datasets. However, the verification and validation plan is already being developed as part of WP5 activities and it will be reported in D5.1 – Verification and validation plan (due May 2021).

- Dissemination activities are being performed and D7.1 – Project communication, dissemination and exploitation report has been delivered in November 2020.

Once the data driven techniques for data processing are defined (WP3) and, parallel to the acquisition of data (WP2), focus will shift towards the development of the prototype with the objective of producing a first prototype of Dispatcher3 by M17 of the project (October 2021), which will be reported in D4.1 – Technical documentation first release and D4.2 – Prototype package (first release). At that moment, the external validation of the system will start. A workshop with the Advisory Board

EUROPEAN UNION

and open to other airlines and stakeholders will be held to present the tool, validate the functionalities and obtain further feedback.

9 References

1. AWS Glue: Simple, scalable, and serverless data preparation. Available at: https://aws.amazon.com/glue/?whats-new-cards (accessed 01.12.2020)
2. Cristianini N. and Shawe-Taylor, J., 2000, An introduction to support vector machines and other kernel-based learning methods
3. D. Bank and Noam Koenigstein and Raja Giryes. Autoencoders. ArXiv, 2020. Volume 003.05991.
4. Da Silva, I.N., Spatti, D.H., Flauzino, R.A. and Liboni, L.H.B., 2017, Artificial neural networks
5. Dalmau, R., Melgosa, M., Vilardaga, S. and Prats, X., 2018, A Fast and Flexible Aircraft Trajectory Predictor and Optimiser for ATM Research Applications. Proceedings of the 8th International Conference on Research in Air Transportation (ICRAT)
6. Decision Trees. Scikit-learn implementation. Available at: https://scikit-learn.org/stable/modules/tree.html (accessed 01.12.2020).
7. Dietterich, T.G., 2000,An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization
8. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F., 2011, A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches
9. Ke, G, Meng, Q, Finley, T., Wang, T. and Chen, W., 2017, Lightgbm: A highly efficient gradient boosting decision tree
10. McInnes, Leland & Healy, John & Astels, Steve. (2017). hdbscan: Hierarchical density based clustering. The Journal of Open Source Software. 2. 10.21105/joss.00205.
11. Peter L. Bartlett and Mikhail Traskin. AdaBoost is Consistent. Journal of Machine Learning Research, volume 8, number 78, pages: 2347-2368. Year: 2007.
12. Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation 1997 9:8, 1735-1780.
13. Tony Jiu. Understanding Random Forest: How the algorithm works and why is it so effective? Year: 2019. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed 01.12.2020)
14. Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
15. Zhang, C. and Ma, Y., 2012, Ensemble machine learning: methods and applications

EUROPEAN UNION

10 Acronyms

AB: Advisory Board

ADS-B: Automatic Dependent Surveillance – Broadcast

AI: Artificial Intelligence

AIBT: Actual In-Block Time

ANN: Artificial Neural Network

ANSP: Air Navigation Service Provider

AOBT: Actual Off-Block Time

ATC: Air Traffic Control

ATCO: Air Traffic Control Officer

ATFCM: Air Traffic Flow and Capacity Management

ATFM: Air Traffic Flow Management

ATM: Air Traffic Management

AWS: Amazon Web Services

BADA: Base of Aircraft Data

CFP: Computerised Flight Pans

CNN: Convolutional Neural Network

CODA: Central Office for Delay Analysis

CRCO: Central Route Charges Office

CSJU: Clean Sky 2 Joint Undertaking

DX.Y: Deliverable number (X=workpackage, Y=deliverable numbering within workpackage)

EFB: Electronic Flight Bag

FDM: Flight Data Monitoring

GBM: Gradient Boosting Machine

H2020: Horizon 2020 research programme

HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise

HMI: Human Machine Interface

IA: Innovation Action

ICRAT: International Conference for Research in Air Transportation

INX: Short name of Dispatcher3 partner: Fundación Instituto de Investigación Innaxis

KPI: Key Performance Indicator

LSTM: Long-Short Term Memory

METAR: Meteorological Terminal Aviation Routine Weather Report/Meteorological Aerodrome Report

ML: Machine Learning

MS: Milestone

NOTAM: Notice to Airmen

PACE: Short name of Dispatcher3 partner: PACE Aerospace Engineering and Information Technology GmbH

PCA: Principal Component Analysis

PDA: Protection Data Agreement

QAR: Quick Access Recorder

RAD: Route Availability Document

RNN: Recurrent Neural Network

SDF: Secure Dataframea

SIBT: Scheduled In-Block Time

SIGMET: Significant Meteorological Information

SOBT: Scheduled Off-Block Time

SQL: Structured Query Language

STATFOR: Statistics and Forecast EUROCONTROL

SVM: Support Vector Machine

TMA: Terminal Manoeuvring Area

TRL: Technology Readiness Level

UoW: Short name of Dispatcher3 coordinator: University of Westminster

UPC: Short name of Dispatcher3 partner: Universitat Politècnica de Catalunya

WP: Workpackage

-END OF DOCUMENT-