



## ICARUS:

**“Aviation-driven Data Value Chain for Diversified Global and Local Operations”**

### **D2.2 – Intuitive Analytics Algorithms and Data Policy Framework**

<b>Workpackage:</b>	WP2 – ICARUS Big Data Framework Consolidation		
<b>Authors:</b>	Suite5, SILO, UCY, CINECA, CELLOCK		
<b>Status:</b>	Final	<b>Classification:</b>	Public
<b>Date:</b>	22/01/2019	<b>Version:</b>	1.00

#### **Disclaimer:**




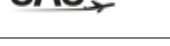



The ICARUS project is co-funded by the Horizon 2020 Programme of the European Union. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

© Copyright in this document remains vested with the ICARUS Partners.

## ICARUS Project Profile

<b>Grant Agreement No.:</b>	780792
<b>Acronym:</b>	ICARUS
<b>Title:</b>	Aviation-driven Data Value Chain for Diversified Global and Local Operations
<b>URL:</b>	<a href="http://www.icarus2020.aero">http://www.icarus2020.aero</a>
<b>Start Date:</b>	01/01/2018
<b>Duration:</b>	36 months

## Partners

	UBITECH (UBITECH)	Greece
	ENGINEERING - INGEGNERIA INFORMATICA SPA (ENG)	Italy
	PACE Aerospace Engineering and Information Technology GmbH (PACE)	Germany
	SUITE5 DATA INTELLIGENCE SOLUTIONS LIMITED (SUITE5)	Cyprus
	UNIVERSITY OF CYPRUS (UCY)	Cyprus
	CINECA CONSORZIO INTERUNIVERSITARIO (CINECA)	Italy
	OAG Aviation Worldwide LTD (OAG)	United Kingdom
	SingularLOGIC S.A. (SILO)	Greece
	ISTITUTO PER L'INTERSCAMBIO ISI SCIENTIFICO (ISI)	Italy
	CELLOCK LTD (CELLOCK)	Cyprus
	ATHENS INTERNATIONAL AIRPORT S.A (AIA)	Greece
	TXT e-solutions SpA (TXT) – 3 <sup>rd</sup> party of PACE	Italy

## Document History

Version	Date	Author (Partner)	Remarks
0.10	06/09/2018	Fenareti Lampathaki (Suite5)	Initial Table of Contents
0.20	21/09/2018	Fenareti Lampathaki (Suite5)	Initial draft incorporating sections 1, 2.1, 3.1, 3.2
0.21	25/09/2018	Nikos Papagiannopoulos (AIA)	Contribution to Section 4.4
0.22	27/09/2018	Fenareti Lampathaki (Suite5)	Creation of templates for Annex II and Sections 3.2-3.4
0.23	18/10/2018	Giorgio Pedrazzi (CINECA)	Contribution to Section 3.3
0.24	29/10/2018	Dimosthenis Stefanidis, George Pallis (UCY)	Contribution to Section 3.5 and Annex II
0.25	29/10/2018	Marios Zacharias (SingularLogic)	Contribution to Sections 4 and 5
0.26	31/10/2018	Matt Colling (OAG)	Contribution to Section 4.4
0.27	01/11/2018	Haris Zacharatos, Neofytos Vlotomas, Yiannis Diellas (CELLOCK)	Contribution to Annex II
0.30	02/11/2018	Fenareti Lampathaki (Suite5)	Updated draft including revisions in Sections 2 and 3
0.31	07/11/2018	Giorgio Pedrazzi (CINECA)	Updated contribution to Section 3.3
0.32	12/11/2018	Haris Zacharatos, Neofytos Vlotomas, Yiannis Diellas (CELLOCK)	Updated contribution to Annex II
0.33	15/11/2018	Marios Zacharias (SingularLogic)	Updated contribution to Sections 4 and 5
0.40	30/11/2018	Fenareti Lampathaki (Suite5)	Updated draft including revisions in Sections 4 and 5
0.50	05/12/2018	Fenareti Lampathaki (Suite5)	Updated full draft circulated for internal review
0.60	20/12/2018	Fenareti Lampathaki (Suite5)	Updated version addressing feedback received during the plenary meeting and the internal review process by ISI (on 07/12/2018), UBITECH (on 11/12/2018) and CELLOCK (on 13/12/2018)
0.70	21/01/2019	Fenareti Lampathaki (Suite5)	Updated version addressing feedback received by PACE (on 10/01/2018), AIA (on 14/01/2018) and OAG (on 21/01/2019)
1.00	22/01/2019	Fenareti Lampathaki (Suite5), Dimitrios Alexandrou (UBITECH)	Final version for submission to the EC

## Executive Summary

The ICARUS Deliverable D2.2, entitled “Intuitive Analytics Algorithms and Data Policy Framework”, documents in detail the performed work and the produced results of the ICARUS Task T2.3 “Deep Learning and Prescriptive Analytics Algorithms” and Task T2.4 “Data Policy and Assets Brokerage Frameworks”. The scope of the current deliverable can be described along the following axes:

*Axis I.* An in-depth review of the data analysis state-of-play was performed as a first step towards elaborating the theoretical foundations for the novel data analytics that will be made available through ICARUS. Specifically, the analysis commences with a literature review of the most recent advancements in machine learning, from 2014 onwards, and their contribution in modern big data analytics and applications so as to outline the full potential of data analysis. Subsequently, data analytics are put into the aviation perspective, through (a) studying scientific literature surveys on machine learning for data analytics in aviation, and (b) exploring additional methods and aviation applications grouped under the four key elements that move through airports, as identified by the IATA NEXTT initiative: aircrafts, passengers, baggage and cargo. As highlighted by NEXTT, air transport entails the complete journey from home to end destination. In this context, data analytics trends and potential benefits throughout the journey of such four elements are examined, covering a broad spectrum of applications and stakeholder interactions and identifying cases and approaches that are of interest to ICARUS.

*Axis II.* The technical and algorithmic perspectives of data analysis are examined in detail in order to understand how data-savvy decision-making can be realised in the complex and frequently changing aviation environment. Real-world data frequently manifest heterogeneity, noise, incomplete attributes and several other characteristics that make data analysis impossible to be examined as a monolithic process. The 5 key steps that are typical to any data analytics approach are in line with the ICARUS methodology in D1.2 and are briefly described, including: (I) Data Ingestion, (II) Data Cleansing and Transformation, (III) Dimensionality Reduction, (IV) Data Analysis and (V) Visualisation. Emphasis is given on the fourth step, i.e. the core data analysis process, by selecting and studying the most important algorithms based on three criteria: a) applicability to aviation specific tasks, b) proven robustness in the research community throughout the years, and c) implementations in commonly used software frameworks/ libraries for data analysis. The selected algorithms are grouped under three types, based on their applicability in Descriptive, Predictive or Prescriptive Analytics, and are presented along 3 axes: Basic Analytics, Machine Learning and Deep Learning. For each of the selected algorithms (9 in Basic Analytics, 17 in Machine Learning and 5 in Deep Learning), a detailed presentation is provided, along with established variations and application examples in the aviation domain. As a final step of this analysis, concrete examples of use cases that map some very early demonstrator scenarios to the aforementioned algorithms are presented, in order to provide early insights into the ICARUS stakeholders’ perspective during the data analysis process.

*Axis III.* Data sharing practices and complexities are examined in detail and reported as a first essential step towards defining and implementing the ICARUS Data Policy and Assets Brokerage Framework. In this context, an in-depth landscape analysis of data sharing practices, incentives and challenges is

performed based on an extensive review of the state-of-play. First, the broader spectrum of data sharing practices is studied, starting from the open data paradigm and moving on to non-open data sharing/ trading approaches and considerations. In this regard, several attributes of data and data sharing agreements are discussed, including IPR, licenses, pricing, trust and security, privacy and protection, ownership and liability. Frameworks that attempt to model the data sharing barriers and drivers and address their challenges are thoroughly studied and discussed. Data marketplaces are also examined, along with the disruptive application of Distributed Ledger Technologies in this context, which bring enforced transparency, rigorous provenance and data democratisation. Then, the aviation-specific challenges and potential benefits of data sharing are discussed through: (a) identifying and studying the major data sharing initiatives in aviation and (b) reporting on the data sharing practices of the project's industry partners. This holistic analysis concludes in the identification of the key considerations for data sharing in aviation.

*Axis IV.* Following the above detailed landscape analysis, the ICARUS Data Policy and Assets Brokerage Framework is defined. The Framework is designed on top of three core entities, namely the Data Asset, the Policy and the Contract and two supporting entities, namely Attributes and Terms, with the latter being classified into Prohibition, Permission and Obligation. The ICARUS Data Policy and Assets Brokerage Framework aims at formalising all data attributes and qualities that affect, or are in any way relevant to, the ways in which data assets can be shared / traded and handled subsequently to their acquisition. Therefore, concrete examples of Attributes and Terms are provided to facilitate its implementation and application on real life cases. Furthermore, workflows that capture the basic provider-consumer interactions are defined and demonstrate how ICARUS envisions to enable the creation of structured, machine-processable data contracts for the aviation industry, whilst maintaining the data owner in control of the provided data. The ultimate goal of the Framework is essentially to link data providers and data consumers at all levels of the data value chain in the aviation industry, through secure and trustworthy data trading agreements. Towards this goal, several challenges have already been identified which will help steer subsequent work towards successfully addressing them.

The results of the current deliverable have been produced hand-in-hand with the developments in deliverables D2.1 ("Data Management and Value Enrichment Methods") and D3.1 ("ICARUS Architecture, APIs Specifications and Technical and User Requirements") providing them with insights stemming from the preliminary data analytics needs and outlining the data sharing requirements presented in this document. Work under the tasks of this Deliverable, namely T2.3 and T2.4, will continue to advance since feedback from the rest of the ICARUS Work Packages will be effectively collected (especially from the development activities and the external MVP validation activities) and will be reported in detail in Deliverable D2.3 ("Updated ICARUS Data Management, Analytics and Data Policy Methods"), as both tasks remain active until M18.

## Table of Contents

<b>1</b>	<b>Introduction.....</b>	<b>9</b>
1.1	Purpose .....	9
1.2	Methodological Approach .....	10
1.3	Relation to other ICARUS Results.....	10
1.4	Structure .....	11
<b>2</b>	<b>Data Analytics in Aviation .....</b>	<b>13</b>
2.1	Data Analytics Background .....	13
2.2	Data Analytics in Aviation Landscape.....	16
2.2.1	Analytics in the Aircraft Journey .....	18
2.2.2	Analytics in the Passenger Journey .....	24
2.2.3	Analytics in the Baggage Journey .....	26
2.2.4	Analytics in the Cargo Journey .....	27
2.3	Key Considerations for Data Analytics in Aviation .....	30
<b>3</b>	<b>ICARUS Data Analytics.....</b>	<b>33</b>
3.1	Purpose .....	33
3.2	ICARUS Data Analytics Approach .....	33
3.3	Axis I: Basic Analytics .....	36
3.3.1	Summary Statistics .....	37
3.3.2	Hypothesis Testing .....	37
3.3.3	Sampling .....	38
3.3.4	Pearson’s and Spearman’s Correlation .....	39
3.3.5	Linear Regression methods .....	39
3.3.6	Logistic regression .....	40
3.3.7	Principal component analysis.....	41
3.3.8	Features selection .....	41
3.3.9	Autoregressive Integrated Moving Average (ARIMA).....	42
3.4	Axis II: Machine Learning Algorithms .....	43
3.4.1	Self-Organising Map .....	44
3.4.2	K-Means.....	45
3.4.3	Streaming K-Means .....	46
3.4.4	DBSCAN .....	46
3.4.5	Gaussian mixture models .....	47
3.4.6	Apriori.....	48
3.4.7	Collaborative filtering.....	49
3.4.8	Content-based filtering .....	50
3.4.9	Support Vector Machines.....	50
3.4.10	Classification and Regression Tree .....	51
3.4.11	Random Forest .....	52
3.4.12	Gradient Boosting Machine.....	53
3.4.13	K-NN .....	54
3.4.14	Naïve Bayes .....	54
3.4.15	Multilayer Perceptron (MLP).....	55
3.4.16	Adaptive Neuro-Fuzzy Inference System (ANFIS) .....	56
3.4.17	Genetic Algorithms (GA).....	56
3.5	Axis III: Deep Learning.....	57

3.5.1	Deep Feedforward Networks (DFFN) .....	58
3.5.2	Convolutional Neural Networks (CNN) .....	58
3.5.3	Recurrent Neural Networks (RNN) .....	59
3.5.4	Deep Autoencoders.....	61
3.5.5	Deep Q-Networks (DQN).....	62
<b>3.6</b>	<b>Data Analytics Perspectives for the ICARUS Demonstrators.....</b>	<b>63</b>
3.6.1	Demonstrator 1: AIA .....	64
3.6.2	Demonstrator 2: PACE.....	64
3.6.3	Demonstrator 3: ISI .....	65
3.6.4	Demonstrator 4: CELLOCK .....	66
<b>4</b>	<b>Data Sharing in Aviation .....</b>	<b>67</b>
4.1	Background .....	67
4.2	Data Marketplaces.....	72
4.3	Data Sharing Motivation and Initiatives in Aviation.....	76
4.4	Current Data Sharing Agreements from ICARUS Stakeholders .....	82
4.5	Key Considerations for Data Sharing in Aviation.....	83
<b>5</b>	<b>ICARUS Data Policy and Assets Brokerage Framework .....</b>	<b>88</b>
5.1	Purpose .....	88
5.2	ICARUS Data Sharing Model .....	88
5.3	ICARUS Data Sharing Workflow .....	91
5.4	Assets Brokerage Challenges .....	93
<b>6</b>	<b>Conclusions &amp; Next Steps .....</b>	<b>98</b>
<b>Annex I: References .....</b>		<b>101</b>
<b>Annex II: Aviation Data Analytics State-of-Play - Indicative papers' in-depth analysis</b>		<b>107</b>

## List of Figures

Figure 1-1: ICARUS Data Analytics and Sharing – Method of work .....	10
Figure 1-2: Relation to other ICARUS Work Packages .....	11
Figure 2-1: A brief history of Data Science (Capgemini 2014) .....	13
Figure 2-2: Gate-to-Gate Flight Procedures List (Canino-Rodríguez et al. 2015) .....	19
Figure 2-3: Causes of delay (Comendador, Valdés, and Sanz 2016) .....	20
Figure 2-4: Influence Factors of Baggage Demand (Cheng, Gao, and Zhang 2015) .....	27
Figure 2-5: A landscape of air cargo operations (Feng, Li, and Shen 2015).....	28
Figure 2-6: Air cargo journey operations state of the art (Feng, Li, and Shen 2015) .....	30
Figure 3-1: Typical Data Analytics Workflow .....	34
Figure 3-2: Relation of the proposed methodology to descriptive, predictive and prescriptive analytics. ....	36
Figure 4-1: Privacy & Protection Attributes of Data Sharing Agreements (Grabus and Greenberg 2017) .....	70
Figure 4-2: Licensing and Rights Management Initiatives Classification (Koko et al. 2018) .....	72
Figure 5-1: ICARUS Data Sharing Model - High-Level View .....	89
Figure 5-2: Basic Data Trading Workflow.....	92

## List of Tables

Table 2-1: Latest machine learning literature review papers .....	14
Table 2-2: Related reviews on Machine Learning Data Analytics in Aviation domain .....	16
Table 3-1: Basic Analytics Algorithms .....	37
Table 3-2: Machine Learning Algorithms .....	44
Table 3-3: Deep Learning Algorithms.....	57
Table 3-4: Scenario use cases for the AIA demonstrator.....	64
Table 3-5: Scenario 1 use cases for the PACE demonstrator .....	65
Table 3-6: Scenario 2 use cases for the PACE demonstrator .....	65
Table 3-7: Scenario 1 use cases for the CELLOCK demonstrator .....	66
Table 3-8: Scenario 2 use cases for the CELLOCK demonstrator .....	66
Table 5-1: ICARUS Data Brokerage Framework: Attributes Definition .....	89
Table 5-2: ICARUS Data Brokerage Framework: Terms Definition .....	90



# 1 Introduction

---

## 1.1 Purpose

ICARUS aims at building a novel data value chain in the aviation-related sectors, acting as multiplier of the “combined” data value that can be accrued, shared and traded. Using technologies that are currently on the rise (e.g. big data analytics, deep learning, semantic data enrichment, and blockchain powered data sharing), ICARUS will address critical barriers for the adoption of Big Data in the aviation industry (i.e. data fragmentation, data provenance, data licensing and ownership), and will enable aviation-related big data scenarios for EU-based companies, organisations and scientists, through a multi-sided platform that will allow exploration, curation, integration and deep analysis of original, synthesized and derivative data characterized by different velocity, variety and volume in a trusted and fair manner.

In this context, the ICARUS Deliverable D2.2 “Intuitive Analytics Algorithms and Data Policy Framework” aims at elaborating on the theoretical foundations for: (a) novel data analytics and (b) a fair and trusted data sharing approach, that will drive the implementation of the corresponding core and advanced data services bundles in the ICARUS platform. D2.2 is released in the scope of the WP2 “ICARUS Big Data Framework Consolidation” activities and practically documents the preliminary outcomes of Tasks T2.3 “Deep Learning and Prescriptive Analytics Algorithms” and T2.4 “Data Policy and Assets Brokerage Frameworks”. In particular, in accordance with the ICARUS Description of Action that dictates the objectives of: (a) suggesting the necessary algorithms for knowledge extraction, business intelligence and usage analytics deriving from big cross-sectorial data, using advanced deep learning and prescriptive analytics patterns, and (b) defining the Data policy and Business Brokerage methods”, the scope of this deliverable is:

- To study and understand the underlying state-of-play regarding data analytics and data sharing in general, and in aviation in particular, from a research / academic and market perspective.
- To get insights into the status quo and the inner workings of data analytics and data sharing agreements in the ICARUS demonstrators.
- To identify the machine learning and deep learning algorithms that are most relevant for the aviation industry and the demonstrators’ needs.
- To elaborate on data sharing and secure information exchange aspects.

As both tasks T2.3 and T2.4 remain active in the following months, D2.2 contributes the first iteration of the data analytics and sharing approaches that will be constantly updated to follow the project’s activities. Such initial outcomes are expected to be refined, enriched and finalized in the middle of the second year of the ICARUS project implementation (on M18) in alignment with the project development activities, as well as the experiences gained during the development activities and the feedback acquired from the early demonstration activities and the external MVP validation activities.

## 1.2 Methodological Approach

In order to conceptualize the ICARUS Data Analytics and Data Sharing Frameworks, the following 2-axis approach was adopted as depicted in Figure 1-1:

- **Axis I: Data Analytics** (addressing the Task T2.3 activities) that included an in-depth study of state-of-the art data analytics approaches in aviation and of the readily available algorithms in different software libraries. The ICARUS Data Analytics approach was defined, ensuring its direct relation to the descriptive, predictive and prescriptive analytics classification. The most relevant algorithms for ICARUS were identified along the (often cross-cutting) Basic – Machine Learning – Deep Learning classification, that led to a preliminary brainstorming and mapping of the demonstrators’ scenarios ideas to specific algorithms.
- **Axis II: Data Sharing** (addressing the Task T2.4 activities) that studied the underlying state-of-play and the current demonstrators’ bilateral agreements for sharing their private data. The ICARUS data sharing approach in terms of data sharing model and basic workflow was elaborated and iteratively discussed.

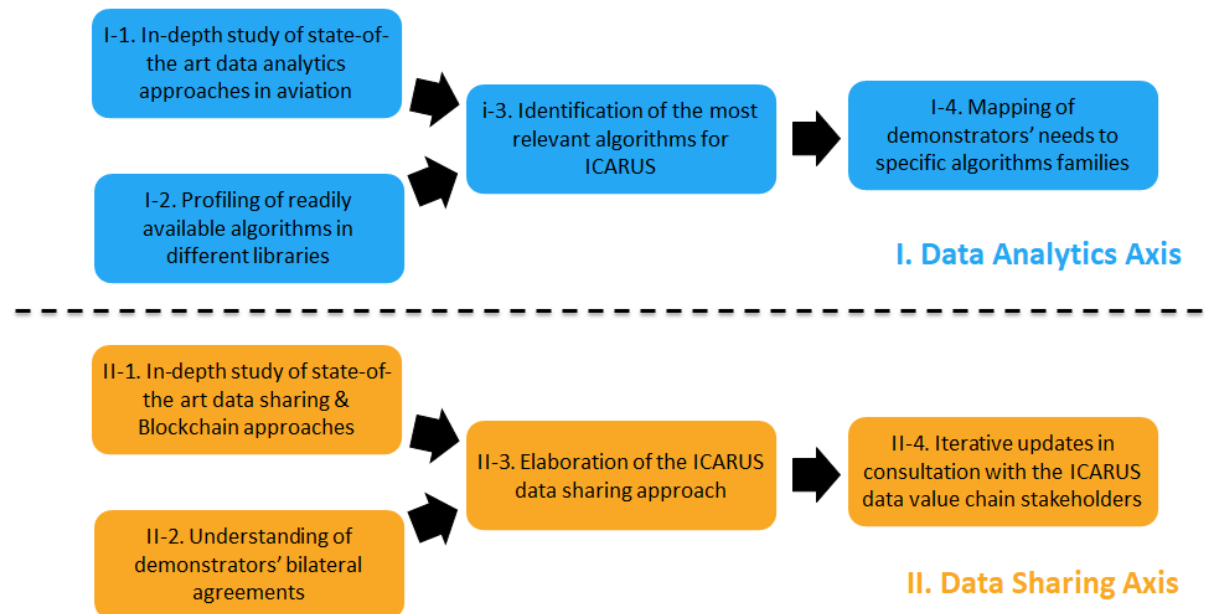


Figure 1-1: ICARUS Data Analytics and Sharing – Method of work

Further updates to be performed on Axis I and II will be reported in the final D2.3 (Updated ICARUS Data Management, Analytics and Data Policy Methods).

## 1.3 Relation to other ICARUS Results

As depicted in Figure 1-2, D2.2 is released in the scope of the WP2 “ICARUS Big Data Framework Consolidation” activities as the initial outcome of Tasks T2.3 “Deep Learning and Prescriptive Analytics Algorithms” and T2.4 “Data Policy and Assets Brokerage Frameworks” that have been brainstormed and elaborated in tight collaboration with the rest of the WP2 tasks (namely T2.1 “Data Collection,

Provenance and Safeguarding Methods” and T2.2 “Data Curation, Harmonisation and Linking Frameworks”).

D2.2 and WP2, in general, are strongly dependent on the outcomes of WP1 “ICARUS Data Value Chain Elaboration” with regard to the ICARUS methodology and the ICARUS Minimum Viable Product (MVP). Although the initial results were documented in D1.2 and were available since the D2.2 preparation activities started, the ongoing external validation activities for the MVP have drastically influenced the definition of the data analytics and especially the data sharing aspects.

Overall the approach specified in D2.2 will be applied, refined and further elaborated during the design and development of the ICARUS platform in WP3 and WP4, respectively. Finally, the experimentation with different analytics algorithms and the application of the data sharing framework in the ICARUS demonstrators (in WP5) will be performed taking into account the specifications, challenges and considerations as reported in this document.

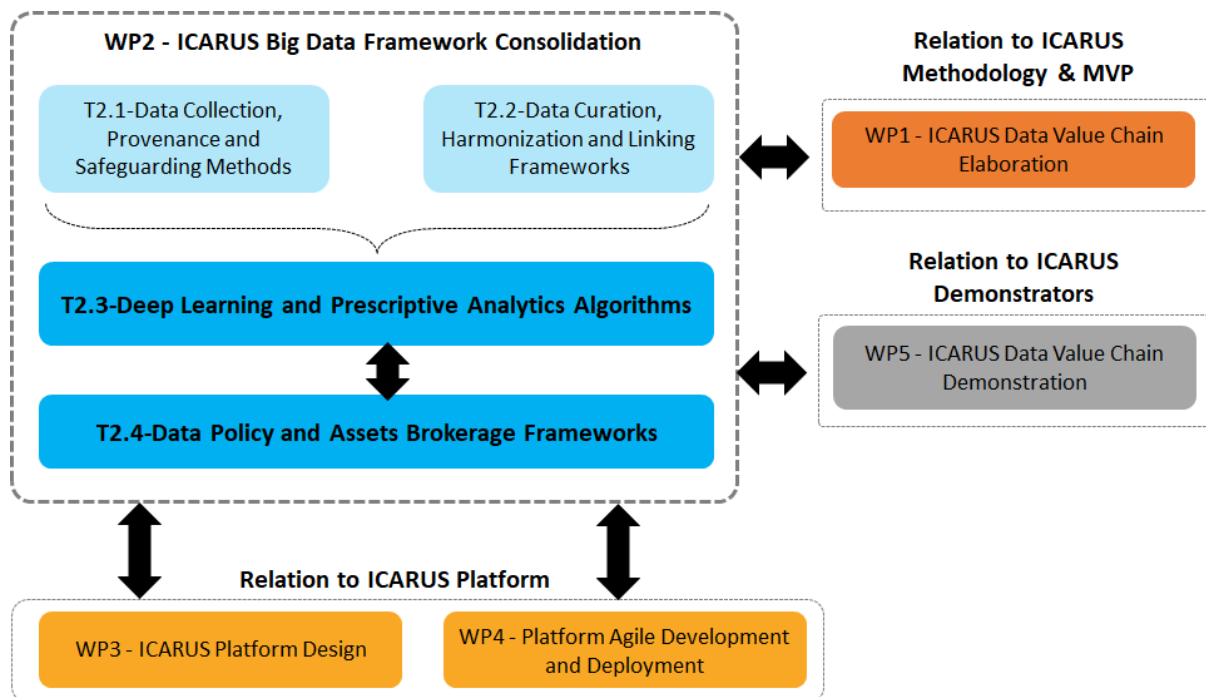


Figure 1-2: Relation to other ICARUS Work Packages

## 1.4 Structure

The structure of the document is as follows:

- Section 2 investigates the current state-of-the art in Data Analytics for aviation, in terms of background and how analytics are applied in the aircraft journey, the passenger journey, the baggage journey and the cargo journey, and concludes with a set of key considerations for ICARUS.
- In Section 3, the ICARUS Data Analytics approach is defined along 3 axes and the different algorithms that fall within the scope of such axes are presented in detail and accompanied by concrete examples on how they could be applied in the ICARUS demonstrators.

- Section 4 sets the scope of data sharing and analyses the relevant state-of-play in aviation, taking into account generic-purpose data marketplaces, but also specific data sharing initiatives in aviation and the underlying data sharing agreements that are put in place by the ICARUS stakeholders.
- Section 5 describes the core of the ICARUS Data Policy and Assets Brokerage Framework, sets the basic data trading workflows and identifies a number of challenges that need to be addressed as the ICARUS project progresses.
- In Section 6, the conclusions deriving from the work performed and documented in the deliverable at hand, as well as directions and recommendations for the next steps are reported.
- Annex I lists the references included in the present deliverable.
- Annex II presents in detail the analysis of selected academic literature for the different areas investigated in Data Analytics (i.e. aircraft journey, the passenger journey, the baggage journey and the cargo journey).

## 2 Data Analytics in Aviation

### 2.1 Data Analytics Background

In the emerging era of Big Data, the analysis of information is motivated by an exponentially increasing volume and availability of different types of data, as well as the desire to create data-driven computer applications that can aid human beings in making complex decisions. To this end, more sophisticated techniques have been developed capable of extracting useful knowledge from data, while the computational power and storage have steadily improved. The convergence of these trends is fuelling a rapidly growing amount of attention on the notion of big data analytics, both in academic and business communities.

**Data analytics** is an arbitrary collection of computational methods and algorithms used to examine data sets and harvest meaningful insights from them. The term data analytics is often seen together with, or even used interchangeably with, the term Data Science. The term Data Science was used in 2001 by William S. Cleveland (Cleveland 2014) to describe a new discipline, closely related to computer science and contemporary work in data mining. As shown in Figure 2-1, Data Science actually incorporates various disciplines, including computer science, statistics, and mathematics.

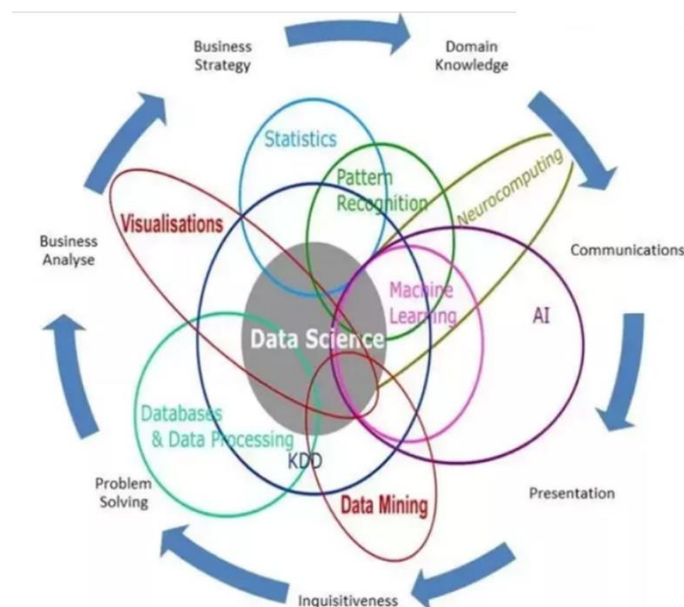


Figure 2-1: A brief history of Data Science – adapted from (Capgemini 2014)

A commonly used distinction between the two terms is that Data Science is responsible for asking questions, whereas Data Analytics provide the processes and techniques that provide answers to questions. In this sense, Data Analytics can be viewed from three major perspectives: descriptive, predictive and prescriptive (Deka 2016). In few words, descriptive (or diagnostic) analytics refer to methods that attempt to describe raw data and extract some form of useful information interpretable by humans. In a way, its purpose is to describe the past (what has happened and why), and as such, it is closely related to Data Mining (Han, Kamber, and Pei 2011). Predictive analytics, on the other hand,

aims to forecast the future and make predictions based on discovered patterns in the given dataset. It originates from AI (Artificial Intelligence) theories and aims to unravel future events or trends. The last category, prescriptive analytics, is a relatively new field that goes beyond descriptive and predictive analytics by recommending particular courses of action that lead towards a solution. Prescriptive analytics use a combination of computational intelligence techniques, tools and procedures, applied against input from different data sets, in order to take advantage of predictions and provide useful pieces of advice.

When data analytics are to examine large and varied data sets, then big data technologies come into play that not only support the ability to collect large amounts, but also to understand and take advantage of their full value, offering high performance, speed and efficiency. The strategy followed to achieve all these is called big data analytics.

Big data analytics involve a number of disciplines, including statistics, data mining, natural language processing, signal processing, pattern recognition, optimisation methods and visualisation approaches. Nevertheless, the primary toolset employed by big data analytics to uncover hidden patterns, predict future trends and assist in decision making derives from the field of machine learning (ML). Machine learning is a branch of artificial intelligence that uses algorithms and mathematical models to parse data, learn from it, and then make relevant predictions or provide useful insights. The ability of ML algorithms to generalise from examples and automatically improve with experience, without the need to be explicitly programmed, make them an essential asset to any modern computer system.

The following table presents a list of latest literature review papers, from 2014 onwards, concerning machine learning techniques and their contribution in modern big data analytics and applications.

**Table 2-1: Latest machine learning literature review papers**

Reference	Title	Contribution	Limitations
Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. <i>IEEE transactions on emerging topics in computing</i> , 2(3), 267-279.	A survey of clustering algorithms for big data: Taxonomy and empirical analysis.	<ul style="list-style-type: none"> <li>• The paper thoroughly reviews ML clustering algorithms for big data and proposes a categorizing framework based on their typical properties.</li> <li>• The paper recommends one efficient algorithm from each category.</li> <li>• The analysis provides a comparison of the recommended algorithms on real big data, evaluating validity, stability, runtime and scalability.</li> </ul>	<ul style="list-style-type: none"> <li>• Ensembles of clustering algorithms are not considered during the comparative study.</li> </ul>
Långkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised	A review of unsupervised feature learning and deep learning	<ul style="list-style-type: none"> <li>• The paper examines and compares deep learning and unsupervised feature learning techniques in a number of time-series related applications.</li> </ul>	<ul style="list-style-type: none"> <li>• The review analysis is somewhat restricted by comparing the methods on some common time-oriented problems and</li> </ul>

Reference	Title	Contribution	Limitations
feature learning and deep learning for time-series modeling. <i>Pattern Recognition Letters</i> , 42, 11-24.	for time-series modeling	<ul style="list-style-type: none"> <li>It highlights the challenges present in time-series data and proposes improvements on learning algorithms.</li> </ul>	benchmark data sets, and not on real-world data.
Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. <i>Journal of Big data</i> , 2(1), 21.	Big data analytics: a survey	<ul style="list-style-type: none"> <li>A thorough review of studies on data analytics from traditional data mining to big data mining is presented in this paper.</li> <li>Frameworks and platforms for big data analytics are also described.</li> <li>Open issues on computational resources, quality of results, security and privacy are discussed.</li> </ul>	<ul style="list-style-type: none"> <li>No comparison is made in terms of efficiency or accuracy of results.</li> </ul>
Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. <i>Big Data Research</i> , 2(3), 87-93.	Efficient machine learning for big data: A review	<ul style="list-style-type: none"> <li>The authors review the theoretical and experimental data-modelling literature for large-scale machine learning techniques, including deep learning.</li> </ul>	<ul style="list-style-type: none"> <li>The value of the paper's insights is quite low.</li> <li>There is no evaluation and comparison of the different approaches presented in the paper.</li> </ul>
Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i> .	Multimodal machine learning: A survey and taxonomy	<ul style="list-style-type: none"> <li>The paper studies in depth the topic of multimodal fusion and surveys recent advances in related machine learning algorithms and methods.</li> <li>It recommends a common taxonomy based on a number of technical challenges.</li> <li>The issue of data misalignment is highlighted.</li> </ul>	<ul style="list-style-type: none"> <li>There is no evaluation and comparison of the different approaches presented in the paper in terms of efficiency.</li> </ul>

The study of the aforementioned review papers offers a brief overview of the current research trends and hot topics related to big data analytics and machine learning. A general conclusion is that the traditional machine learning algorithms cannot easily cope with big data characteristics and thus, they need to scale up and be modified accordingly (e.g. deep learning networks) or become more adapted to the new challenges (e.g. via ensemble methods or dimensionality reduction techniques).

## 2.2 Data Analytics in Aviation Landscape

The purpose of data analytics in aviation is to examine the vast amount of data generated daily and provide useful information to airlines, airports, and other aviation stakeholders so that they can improve their operational planning and execution, as well as any related products and services.

Nowadays, aviation data derive from diverse sources, and usually lack standardisation, uniformity and fault controls required for reliable integration in a common analytics platform. A mere examination of the relative literature shows an evolving domain that takes data collection and mining very seriously, but faces a real obstacle when it attempts to scale up and combine data sources. In order to cope with this challenge, many participants in the aviation industry have implemented their own, isolated solutions, which may improve particular processes but fail to capture the whole picture.

Today, with the use of big-data analytics and technologies, the accumulation and processing of massive data sets has become easier, while, at the same time, the increasing application of machine learning techniques has led to advanced predictive analytics, extracting trends and insights from heterogeneous data sources able to address modern operational needs and requirements.

It is, therefore, essential for this project to examine the state-of-play in aviation analytics, and highlight important contributions and challenges. The research analysis commences with a review on scientific literature surveys on machine learning data analytics in aviation, presented in the table below (Table 2-2).

**Table 2-2: Related reviews on Machine Learning Data Analytics in Aviation domain**

Reference	Title	Contribution	Limitations
Maheshwari, A., Davendralingam, N., & DeLaurentis, D. A. (2018). A Comparative Study of Machine Learning Techniques for Aviation Applications. In <i>2018 Aviation Technology, Integration, and Operations Conference</i> (p. 3980).	A Comparative Study of Machine Learning Techniques for Aviation Applications	<ul style="list-style-type: none"> <li>It is a survey on machine learning techniques applied on the Aviation domain.</li> <li>The analysis provides a comparison of popular algorithms using an air travel demand modelling problem.</li> <li>Uses publicly available data.</li> </ul>	<ul style="list-style-type: none"> <li>The scope of this survey is limited in one particular example, that of air travel demand.</li> </ul>
Ariyawansa, C. M., & Aponso, A. C. (2016, May). Review on state of art data mining and machine learning techniques for intelligent Airport systems. In	Review on state of art data mining and machine learning techniques for intelligent Airport systems	<ul style="list-style-type: none"> <li>Examines and compares data mining techniques that can be applied on flight delay prediction, passenger profiling, passenger segmentation and association rule mining.</li> </ul>	<ul style="list-style-type: none"> <li>Provides selected data mining methods that could be integrated in an intelligent airport system. So, the review analysis is somewhat restricted by this assumption.</li> </ul>



Reference	Title	Contribution	Limitations
<i>Information Management (ICIM), 2016 2nd International Conference on</i> (pp. 134-138). IEEE.			<ul style="list-style-type: none"> <li>Insights and key-takeaways from this survey are of limited extent.</li> </ul>
Gavrilovski, A., Jimenez, H., Mavris, D. N., Rao, A. H., Shin, S., Hwang, I., & Marais, K. (2016). Challenges and Opportunities in Flight Data Mining: A Review of the State of the Art. In <i>AIAA Infotech@Aerospace</i> (p. 0923).	Challenges and Opportunities in Flight Data Mining: A Review of the State of the Art.	<ul style="list-style-type: none"> <li>It reviews data mining techniques for aviation operational safety and quality.</li> <li>It tackles the problem of time series clustering.</li> </ul>	<ul style="list-style-type: none"> <li>Focuses mainly on anomaly detection tasks for rotorcraft and fixed-wing general aviation.</li> <li>There is no evaluation and comparison of the different approaches presented in the paper.</li> </ul>

The presented review papers were not easy to find, but it is evident from their publication year that the application of data analytics and machine learning techniques on the aviation domain has only recently attracted a lot of interest in the research community. All these works analyse and compare mainly machine learning methods on aviation related challenges and data sets. However, most of the review papers have a narrow scope and base their comparative analysis on one or two specific tasks. The work of Ariyawansa & Aponso in (Ariyawansa & Aponso, 2016) constitutes an exception, since it tackles five different challenges from a general airport's perspective.

From an algorithmic point of view, artificial neural networks (ANNs), ensemble methods (e.g. random forest), and lately, deep learning methods seem to be the most effective choices when dealing with aviation industry data. Text mining approaches are also gaining popularity in analysing accident reports to improve our understanding of the safety-related incidents. The most commonly used aviation data sets include flight tracking data, airport operations, weather conditions, airline and passenger information, aircraft logs and air safety reports.

While Table 2-2 provides insights gained from broader data analytics studies in the aviation domain, in the next sub-sections, additional methods and aviation applications are presented, grouped under the four key elements that move through airports, as identified by the IATA NEXTT initiative: aircrafts, passengers, baggage and cargo. As highlighted by NEXTT, air transport is not just about the flight, it's about the complete journey from home to end destination. Each of the subsequent subsections therefore examines the data analytics trends and potential throughout the journey of the four elements, thus covering a broad spectrum of applications and stakeholder interactions.

It should be clarified that the scope is not to present data analytics methods from a technical perspective (as this will be provided in Section 3), but to explore the state-of-play in applying data analysis through concrete aviation-specific examples. In the highly complex air travel value chain,

though, even specific use cases cannot be isolated completely, so the grouping under the four categories serves as a means to facilitate the review process and its presentation, but the boundaries among the areas are, at least, blurry. Indicatively, fuel consumption is a very popular and actively researched topic, which is highly correlated with aircraft taxi-out time and hence flight delay prediction, which is in turn by itself another important topic. Going a step further, delays affect passenger experience and may reflect badly on their perception of the airport and/or airline. Passenger experience is, of course, also a very important topic with its own studies and approaches.

It should also be stressed that several of these problems have proposed solutions stemming from the Operations Research domain, e.g. optimisation approaches, which will not be discussed here. Operations Research is a partially overlapping discipline with Data Analytics in the sense that there are data problems in which optimisation techniques from Operations Research have to be applied to detect best fitting structures (under suitable constraints) in the underlying data. Furthermore, Operations Research is often based on model formulations for which some model parameters might be unknown or even unobservable, in which cases model parameters may be obtained through the application of Data Analytics.

#### 2.2.1 Analytics in the Aircraft Journey

The aircraft journey concept is broad and includes pre-flight, en route and post-flight operations. Moreover, even though aircraft manufacturing processes and aircraft software are not in scope, there is an immense volume of data being generated from aircraft systems and sensors during flights that can be leveraged to extract insights for various processes relevant to the many tasks and involved stakeholders in the aircraft journey. Indicatively, the forthcoming Airbus A380-1000 will be equipped with 10,000 sensors in each wing. The current A350 model has a total of close to 6,000 sensors across the entire plane and generates 2.5 Tb of data per day, while the newer model – expected to take to the skies in 2020 – will capture more than triple that amount. Submerged in this massive amount of data, data analysts are presented with the opportunity to extract useful information through the application of data analysis. Even if the data from the aircraft sensors are currently only available to manufacturers for analysis, the potential advantages are nevertheless immense for all stakeholders and various other data sources are also available.

The aircraft journey phases, as shown in Figure 2-2, encompass processes that are directly or indirectly relevant to various aviation stakeholders. Aircraft systems are numerous, have complex functionalities and interactions and are the result of an ongoing research and development process, which also includes various data analysis procedures. However, as stated above, for the purposes of the current deliverable, such systems (e.g. airborne navigation systems, trajectory guidance and conflict detection-resolution systems, aircraft environment information management systems), as well as functionalities related to air traffic control, are out of scope. The aircraft journey will instead be examined from the perspectives of the involved stakeholders and not from the technical challenges related to the core aircraft operations. The phases more relevant to ICARUS for the aircraft journey

are thus the pre- and post-flight ones, whereas the en route phases will be mostly examined from the passengers' perspective in the next sub-section.

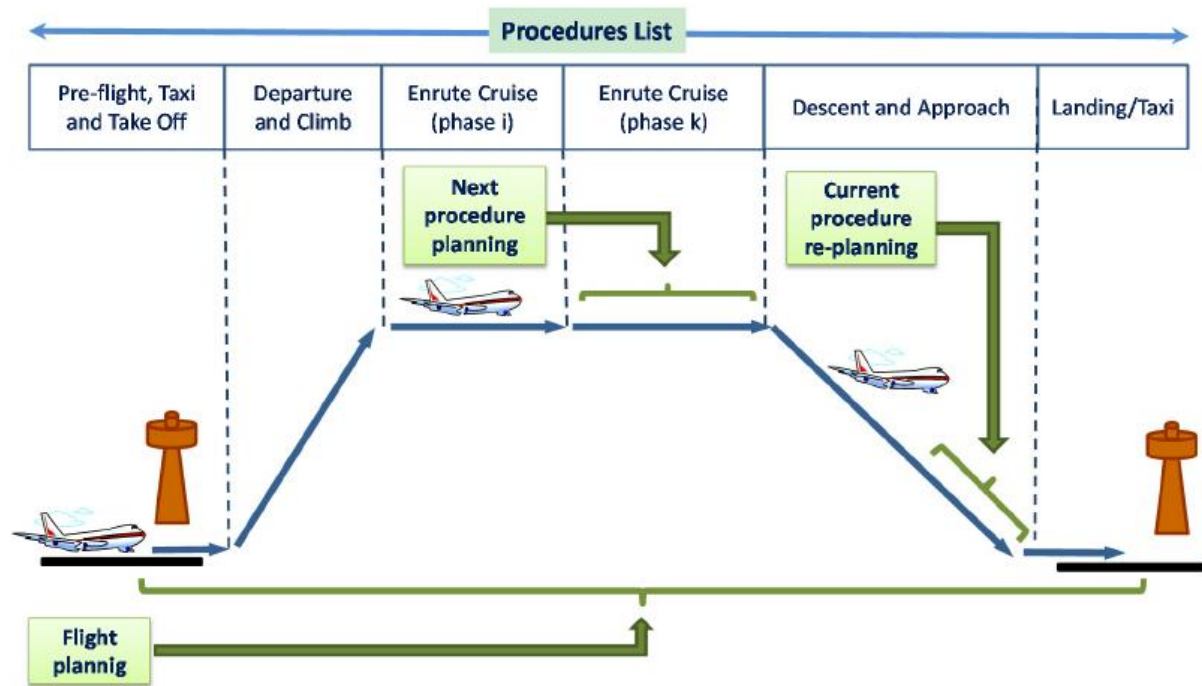


Figure 2-2: Gate-to-Gate Flight Procedures List (Canino-Rodríguez et al. 2015)

Commercial aviation is a complex transportation system that has to orchestrate a sophisticated origin-destination matrix and ensure flight safety and smooth operations for and by all involved stakeholders, including but not limited to, passengers, airports, airlines, pilots, cargo handling companies, ground handlers etc. The most prominent research problem related to the aircraft journey which affects all involved parties, seems to be the ***prediction of flight delays***. (Comendador, Valdés, and Sanz 2016) identify 9 main causes of delay in E-TMA processes. Each of them has different performance drivers and indicators and different approaches are used for their prediction.

Cause of Delay	Definition / Explanation
Reactionary	Delays due to the late arrival of aircraft delayed during its previous leg operation, late arrival of a connecting flight, passengers or load, and late arrival of crew members, expected from another flight
ATC / ATFCM	Delays due to ATC/ATFCM management: standard demand/capacity problems, reduced capacity caused by industrial action or staff shortage, equipment failure, weather, military exercise or extraordinary demand due to capacity reduction in neighbouring area (noise abatement, night curfew, special flights). It also includes restrictions related to air traffic services, start up and push back
Passenger and Baggage Processes	Delays due to inefficiencies and failures during passengers and baggage processes: check-in reopened for late passengers, check-in not completed by flight closure time, errors with passenger or baggage details, booking errors (overselling), discrepancies or missing checked in passengers during boarding, late or incorrect order given to catering supplier, late or incorrectly sorted baggage.
Cargo (including mail)	Delays due to inefficiencies and failures related to cargo processes: late or incorrect documentation for booked cargo, late delivery of booked cargo to airport/aircraft, acceptance of cargo after deadline, repackaging and/or re-labelling of booked cargo, booked load in excess of saleable load capacity (weight or volume), cargo reloading or off-load
Weather	Delays due to weather conditions below operating limits. It includes removal of ice, snow, water, and sand from airport (runway, taxiways, apron), and ground handling impaired by adverse weather conditions (high winds, heavy rain, blizzards, monsoons etc.).
Airport Facilities and Operations	Delays due to disruptions or problems related to airport facilities (parking stands, ramp congestion, lighting, buildings, gate limitations, etc.) and operations (check-in, security, immigration, customs, health, boarding, etc.). It also includes operational restrictions such as airport and/or runway closed due to obstruction industrial action, staff shortage, weather, political unrest, noise abatement, night curfew, special flights.
Technical and Aircraft Equipment	Delays due to failures or problems related to technical and aircraft equipment: aircraft defects, late release from scheduled maintenance, special checks and/or additional works beyond normal maintenance schedule, lack of spares, lack of and/or breakdown of specialist equipment required for defect rectification, aircraft change for technical reasons (e.g. a prolonged technical delay), scheduled cabin configuration adjustments, aircraft damage during operations (bird or lightning strike, turbulence, heavy or overweight landing, collisions during taxiing and ground operations).
Airline Operations	Delays due to inefficiencies and failures during airline operations: late completion of or change to flight plan, late alteration to fuel or payload, late crew boarding or departure procedures, flight deck shortage or special request, extraordinary captain requests for security checks outside mandatory requirements.
Handling	Delays due to inefficiencies or failures during aircraft and ramp handling processes: late or inaccurate aircraft documentation, problems regarding loading/unloading, servicing, cleaning, fuelling/defueling and catering.

**Figure 2-3: Causes of delay** (Comendador, Valdés, and Sanz 2016)

(Sternberg et al. 2017) provide an extensive review of flight delay prediction methods and identify grouping parameters that help to better understand their strengths, weaknesses and intended usage. In reality, delays can be induced by different sources and affect airports, airlines, en route airspace or an ensemble of them, but in order to apply data analysis, a simplified system is usually assumed in order to identify how a specific actor is affected and how a specific type of delay can be predicted. It is observed that machine learning approaches are still less common compared to operations research solutions, but there has been a significant increase in their usage in the last decade. Classification, forecasting and even recommender systems (Zonglei, Jiandong, and Guansheng 2008) have been also employed in this direction.

Indicatively, (Rebollo and Balakrishnan 2014) propose a new class of models for predicting air traffic delays based on random forests using both spatial and temporal delay states, whereas (Ayhan, Costas, and Samet 2018) use weather data, air traffic, and airport data along the potential flight path in order to devise a novel Estimated Time of Arrival (ETA) system for commercial flights. The system feeds these data to various regression models and a Recurrent Neural Network (RNN) and the obtained results are reported as superior to the ones provided by EUROCONTROL. (Khanmohammadi et al. 2014) present an approach based on a fuzzy inference system used in the context of a fuzzy DSS to sequence arrivals in JFK airport. (Mathur, Nagao, and Ng 2013) used 105 features to fit two Naïve Bayes models, one Gaussian for the 14 continuous features and one multivariate multinomial for the 89 categorical features, to create a classifier for determining whether a flight will be on time.

The features used in delay forecasting generally vary in literature and depend on the type of delay being examined, the selected stakeholder perspective and, subsequently, the data availability. (Sternberg et al. 2017) identify the following commonly used data categories and indicative features:

- Temporal: season, month, day of week, time of day
- Planning: flight plan, airline schedule, airport schedule
- Weather: visibility, ceiling, convective weather, surface weather
- Spatial: airport, city, region
- Operations: capacity, demand
- System state: prior delay levels, operational conditions
- Features: airline status, airport infrastructure, aircraft model, aircraft occupancy, fares, frequency

Another popular research topic, interlinked with delay times in air transport, is accurate **taxi-out time prediction**, which is a significant precondition for improving the operability of the departure process at an airport, as well as reducing congestion and excessive emission of greenhouse gases. Taxi-out time is usually defined as the time spent by a flight between its actual off-block time (AOBT) and actual take-off time (ATOT). (Lee, Malik, and Jung 2016) review and compare machine learning algorithms for taxi-out prediction using real-world traffic data and find that random forests and linear regression give the most accurate results. (Lian et al. 2018) focus on congested airports and evaluate a Generalized Linear Model, a Softmax Regression Model, and an Artificial Neural Network method and two improved Support Vector Regression (SVR) approaches based on swarm intelligence

algorithm optimisation, which include Particle Swarm Optimisation (PSO) and Firefly Algorithm. The last two, which are developed in the context of that work, are proven to achieve a better predictive performance when dealing with abnormal taxi-out time states than the other three more conventional methods.

Taxi-out time is highly correlated with the **Aircraft Landing** problem (Zipeng and Yanyang 2018), which is studied mainly from the perspective of airports who are the responsible stakeholders to provide this planning. In this regard, the problem has two distinct directions: the single runway problem (SRP), i.e. the scheduling of aircraft landing on the unique available runway, and the multi-runway problem (MRP) which requires runway assignment additionally to sequencing (M. Ahmed et al. 2018) propose a population-based meta-heuristic approach based on a genetic algorithm for optimal **aircraft sequencing and runway configuration**. (Hancerliogullari 2013) perform a deep analysis on the combined **arrival-departure aircraft sequencing** and **reactive scheduling** problems on multiple runways based on various aircraft data, including its operational type (i.e. arrival or departure), weight-class (i.e. heavy, large, or small), priority (aircraft tardiness penalty), ready time, target time, deadline, and separation times. The reactive part refers to the fact that air traffic systems frequently encounter various disruptions due to unexpected events such as inclement weather, aircraft failures or personnel shortages, hence such data are also important to properly model the situation. The sequencing problem is part of the commonly encountered **Ground Movement Problem**, i.e. the allocation of efficient routes to taxiing aircrafts, which becomes increasingly important as air traffic levels continue to increase. If taxiways cannot be reliably traversed quickly, aircraft can miss valuable assigned slots at the runway or waste fuel waiting for other aircraft to clear. (Brownlee et al. 2018) propose a fuzzy approach to tackle the inherent uncertainty in these situations and, more specifically, an adaptive Mamdani fuzzy rule-based system to estimate taxi time and relevant uncertainty. They also adapted the currently used QPPTW algorithm to use fuzzy time estimates and evaluated their model on simulated data for taxi movements at Manchester airport, finding its predictions to be more accurate. Other relevant topics here include **aircraft turnaround operations** (Schmidt 2017), with particularly rich literature for congested airports that operate at or close to their capacity limit and **gate assignment** (Bouras et al. 2014) for which genetic algorithms have been and continue to be widely used (Bolat 2001; Gu and Chung 1999; Hu and Di Paolo 2007; Xu and Cai 2019).

Another extremely important research topic, from both academic and industry perspectives is the **short-term and long-term air traffic forecasting**, which is horizontal to all problems discussed above, affecting all of them directly and being indirectly affected through the satisfaction levels and feedback of involved stakeholders. (Nai et al. 2017) propose a hybrid air traffic forecasting model based on empirical mode decomposition (EMD) and seasonal auto regressive integrated moving average (SARIMA). Monthly air cargo and passenger flow data from 2006 to 2014 (from the Civil Aviation Administration of China - CAAC) are used to test the model's forecasting performance in comparison to other known forecasting methods. (Bao, Xiong, and Hu 2012) propose an ensemble EMD-based SVM model to tackle the nonlinearity and irregularity along with implicit seasonality and trend in the context of short-term air passenger traffic forecasting. Air traffic forecasting based on **predicted**



**passenger flows** is an active research field with numerous custom approaches and ensemble models, including combination of Grey Theory and ANN, SVMs and Decision Trees (Ke-Wu 2009; Laik, Choy, and Sen 2014; Yan and Zhang 2010). The investigated issues in this domain are clearly related to enhanced **demand prediction** for all involved stakeholders. (S. C. Chen et al. 2012) use back-propagation neural networks to forecast air passenger and cargo demand combined and individually. (Sulistiyowati et al. 2018) also examine both cargo and passenger demand together and propose a hybrid method that utilises first a linear modeling with time series regression model (TSR) and Autoregressive Integrated Moving Average with Exogenous Factor (ARIMAX) and in a second phase, analyse the error of the linear model using machine learning methods such as Neural Network (NN) and Support Vector Regression (SVR) to capture nonlinear patterns. They evaluate four combinations of these methods at three international airports in Indonesia, namely, Soekarno Hatta, I Gusti Ngurah Rai, and Juanda Airport and find that hybrid ARIMAX-NN and TSR-NN give the most accurate results. The **Hub Location Problem (HLP)** has strong connections to the demand forecast and capacity planning and is also very popular in academic literature. Hubs are facilities that work as consolidation, connecting, and switching points for flows between stipulated origins and destinations, therefore HLP has obvious applications in airlines and airport networks (Farahani et al. 2013). (Wu, Zhang, and Wei 2018) examine how regional hub airports can enhance the network performance in terms of reduced total network cost and congestion at the hub airport.

Apart from mining trends in air traffic flows for traffic forecasting, interesting applications include mining air traffic flows to identify patterns. (Murça et al. 2018) have developed a data analytics framework for air traffic flow characterisation from flight tracking data. The framework performs spatiotemporal **pattern recognition in flight trajectory data** using a dual machine learning approach: A density-based trajectory clustering algorithm (Density-Based Spatial Clustering of Applications with Noise - DBSCAN) and a multi-way classifier based on Random Forests. The framework is tested on terminal area traffic flows of three multi-airport (metropolitan) systems of the global air transportation system: New York, Hong Kong and Sao Paulo.

The performed state of the art review highlighted the significance of problems related to airport capacity and efficient aircraft sequencing towards predictable and, when possible, minimum delays in air travel. However, there are numerous other aspects of the broader aircraft journey that benefit from data analytics processes. Environmental challenges, including **noise abatement** and **fuel consumption minimisation** (which are also relevant to taxi-out times and operations), and assessment of airports' **energy sustainability** are some indicative examples (Alba and Manana 2016; Glenn, Panarat, and Graham 2018; Lluís 2018; Salah 2014). (Matthews et al. 2013) leverage scalable data mining algorithms to detect anomalous safety events and highlight the types of anomalies that can be identified through the proposed process. Traditional **Business Intelligence** applications are of-course also commonly applied by aviation stakeholders (e.g. data mining applied by airlines for cost optimisation, airport situation revenue per flight, cost per seat (Akpınar and Karabacak 2017) etc.), but will not be extensively studied in the context of this deliverable as they are more generic. However, many of the above discussed problems are related with BI aspects. As an example, (Tang, Yan, and

Chen 2008) examine how improved demand forecast can lead to optimised passenger, cargo, and combined flight scheduling, which in turn can increase efficiency, operability and revenues for airlines and airports, while lowering the costs, fares and passenger tickets.

### 2.2.2 Analytics in the Passenger Journey

Global passenger traffic is expected to double by 2035, reaching 6 billion passengers, so it's no surprise that airport technology is one of the fastest developing aspects of the entire industry. But despite the surge in innovation, very few of the changes being made focus on improving the passenger experience. Ultimately, the industry needs to re-think its current customer service strategies in line with what the passenger expects. The service industry is moving quickly to personalize every step of the process, but the aviation sector is a step-change behind. There is currently an abundance of data in various forms and from various sources that can be leveraged through advanced data analytics: from thermal condition sensors in airport terminals and biometric passenger processing (Patel 2018), to social media conversations and hand-written feedback forms being digitised, to directly acquired passenger data and inferred behaviours and more. Research in the field shows immense potential in “bringing the digital passenger to life”.

One of the main issues from the passenger's perspective is the minimisation of waiting times in airport queues. According to the 2017 Global Passenger Survey conducted by IATA (IATA 2017), most passengers expect emerging technologies to minimise time uncertainties in air travel and make them more in charge of their travel experience. This is gradually being achieved through the provision of off-airport capabilities, e.g. the online check-in, as well as automated processes in airports, such as the self bag drop services available in many airports. Data analytics cannot directly provide such services, but they are leveraged for **exploratory data analysis of waiting times (check-in, security control, baggage check-in, immigration counter)** in order to extract insights on how they could be avoided or accurately predicted. Data mining on passenger-movements from the time passenger arrives at the airport until they depart helps at predicting passenger flows and avoiding bottlenecks, improving both airports' and airlines' capacity management. (Sankaranarayanan, Agarwal, and Rathod 2016) employ big data visual analytics techniques towards this goal, whereas (Chiti, Fantacci, and Rizzo 2018) propose an integrated and flexible service platform suitable for airport operations management which integrates a queueing analytical model for queue time prediction. (Guo, Grushka-Cockayne, and De Reyck 2018) develop a predictive system that generates quantile forecasts of transfer passengers' connection times and also produces quantile forecasts for the number of passengers arriving at the immigration and security areas. The system applies various regression models, including regression trees, quantile regression forests and linear regression on data from the Business Objective Search System (flight information data), the Baggage Daily Download (records of every piece of connecting baggage through Heathrow on the previous day), and the Conformance data sets (records of boarding pass scans) for Heathrow airport.

(Rahaman, Hamilton, and Salim 2017) focus on another type of queue and propose a methodology and algorithms on **minimizing taxi and waiting passenger queues** under different contextual factors,



including time, taxi trips, passengers and weather. For travellers that ride the metro or train to get to the airport, another study performing **micro-analysis on travel behaviour patterns** sheds light on how to optimize efficiency to better serve the demand (Cardell-Oliver et al. 2017). It is important to understand that the passenger journey is not limited to the time spent in airports and aircrafts but includes off-airport activities as well. Leveraging data from the steps prior even to the passenger journey beginning, (Kim and Shin 2016) try to **forecast short-term air passenger demand** from search engine query logs and successfully model the identified short-term fluctuations in the air passenger demand regression model. Although data used in this paper are only part of the passenger journey, the performed work is clearly linked to the demand prediction problem which was extensively discussed in the previous section and it can help airports forecast more accurately their short-term passenger demand. Apart from waiting times, ambient parameters are also important in providing a satisfactory traveling experience. With the availability of thermal sensors in airports, it is now easier to explore which are the **optimum ambient (e.g. thermal) conditions** for passengers, as shown in (Kotopoulos and Nikolopoulou 2016). The aforementioned studies provide insights as to how the overall passenger experience can be improved, but also ultimately help improve passenger opinion for the corresponding airports and their facilities.

Understanding passenger opinions and behaviours is critical for many stakeholders in the air transport value chain for reasons beyond waiting times prediction and flight safety. (Balcombe, Fraser, and Harris 2009) applied Bayesian methods on data obtained through an online choice experiment to examine consumer choices with respect to the **bundle of services on offer when deciding to purchase a flight**. The data included socio-economic information (e.g. age, gender etc.) and offer features such as ticket price, seat dimensions and on-board entertainment options. Results showed that passengers are willing to pay a relatively large amount for enhanced service quality. Frequent flyer programs are loyalty programs designed to enable (mainly) business travellers to have a smoother travel experience, but at the same time incentivize them to choose specific airlines, even if there are cheaper competing carriers. **Data mining on frequent flyer programs data**, together with other airline data can help predict loyalty, rate and are used in airline decision support systems (Akpınar and Karabacak 2017; Gössling et al. 2017). **Flight ticket pricing** is also a popular application field for predictive analytics in the aviation industry. (S. Chen et al. 2015) investigates airline passenger behaviour based on three types of travel data, provided by a Chinese airline company: passenger name record (PNR), share of wallet (SOW) and web trends. PNR data analysis aims to identify passengers with influential power in their social interactions. SOW data analysis is used to identify potential high-value travellers and suggest corresponding marketing **segmentation** and promotion strategies. Passenger's webtrends information (e.g. mobile number, membership number, web browsing records) are used to determine the passengers' web and mobile usage levels. (Zheng et al. 2016) propose a deep learning approach to **passenger profiling**, using a pythagorean fuzzy deep Boltzmann machine (PFDBM), biogeography-based optimisation (PFDBM-BBO), a Gaussian mixture model (PFDBM-G) and hybrid gradient and enhanced BBO learning (PFDBM-EBO). They use a variety of data sources that provide heterogeneous data, indicatively including PNR information, passenger's flight history from the Aviation

Administration (collected from different airlines), travel statistics from other transportation methods, such as railway and marine, statistics from the Tourism Administration (collected from travel agencies) and preprocessed telecommunication behaviour records. Such profiling methods and other predictive analytics techniques are applied to help airlines offer personalized advertising services and provide personalised *in-flight entertainment*, thereby both increasing ancillary revenues and improving customer satisfaction (Akpinar and Karabacak 2017; Elena 2018).

### 2.2.3 Analytics in the Baggage Journey

The growth in the number of passengers worldwide exerts considerable pressure on the systems and baggage processes of the industry. According to the Baggage Report for year 2018 compiled by SITA (CITA 2018), mishandling rate has dropped by more than 70% since 2007. With the increase to more than 4 billion passengers in 2017, airlines have succeeded in reducing the amount of incorrectly issued bags. In fact, 5.57 suitcases per thousand passengers is the lowest level ever recorded. Despite this improvement, according to the report, improperly managed baggage cost the industry an estimated \$ 2.3 billion in 2017, so there is a considerable margin for cost reduction that favors airline investments in the tracking of luggage from end to end.

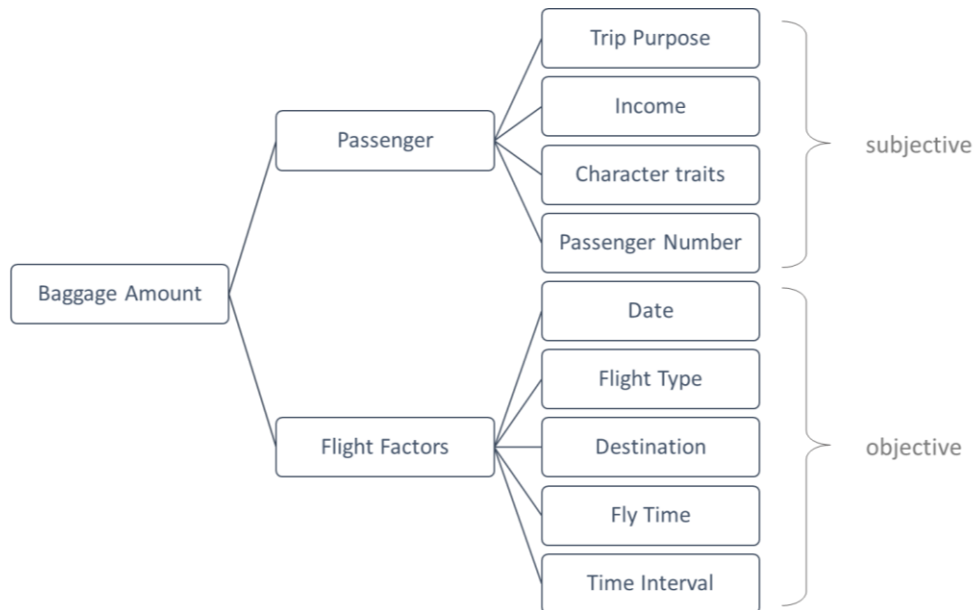
Thanks to the state-of-the-art indoor tracking technologies like RFID, Bluetooth, and Wi-Fi, tracking of object movements from one symbolic location to another within the indoor spaces is possible. However, the resulting tracking data can be massive in volume. Analyzing these large volumes of tracking data can reveal interesting patterns that can provide opportunities for different types of location-based services, security, indoor navigation, identifying problems in the system, and finally service improvements.

The European project BagTrack<sup>1</sup> aims to **identify the patterns that are responsible for baggage mishandling** towards improving the worldwide aviation baggage handling quality.

Baggage mishandling data are highly imbalanced, posing challenges in the data preprocessing steps of model development. (T. Ahmed, Calders, and Pedersen 2015) highlight this fact and propose a detailed methodology for mining risk factors from RFID baggage tracking data that will help baggage management vulnerabilities. In this work, data are first fragmented based on various important factors and then different classifiers (specifically Decision Trees, Naïve Bayes, KNN, Linear Regression, Logistic Regression and SVM) are used to classify the baggage records under two categories: mishandled and properly handled.

---

<sup>1</sup> <http://daisy.aau.dk/bagtrack>



**Figure 2-4: Influence Factors of Baggage Demand** (Cheng, Gao, and Zhang 2015)

(Zeinaly, De Schutter, and Hellendoorn 2015) study how the **integrated control of baggage handling systems** can be improved through a predictive model that includes **routing, line balancing** and empty-cart management. (Cheng, Gao, and Zhang 2015) try to model the baggage demand prediction problem in isolation of the overall air travel demand and identify several high-level influence factors for **baggage demand prediction** (Figure 2-4), for which weights are assigned using grey relational analysis. For the prediction, two models are used and compared, specifically a back-propagation Neural Network and a multiple regression model, with the latter showing superior performance.

The baggage journey is closely linked to the aircraft and passenger journeys that were presented above and is affected by many of the problems discussed there, e.g. demand prediction and waiting times in airport queues (for baggage drop-off). This sub-section therefore briefly discussed only problems pertaining specifically to the baggage journey.

#### 2.2.4 Analytics in the Cargo Journey

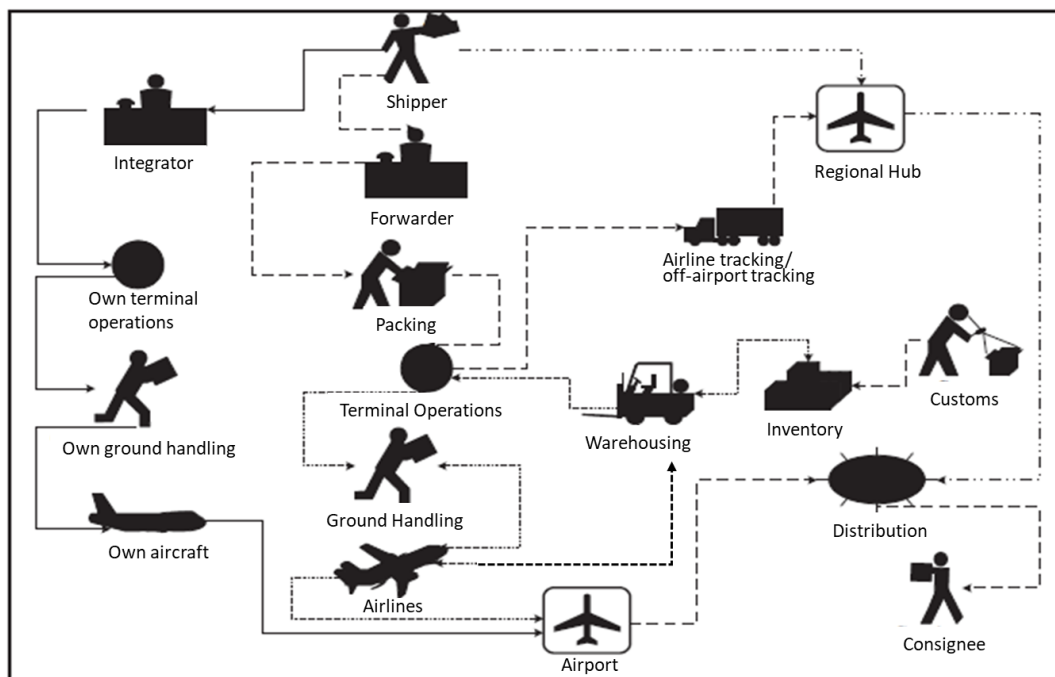
Air transport is vital for manufactures trade (particularly cross border) and air cargo is essential to many facets of modern life, including moving perishable goods across countries/ continents. It is vital for the pharmaceutical industry (particularly vaccines), but also for the transportation of live animals, electronic devices, and various e-commerce needs. Indicative of the air cargo importance is the fact that IATA has an interest group of around 80 members, named Cargo iQ, with the mission of creating and implementing quality standards to improve shipment control and processes for the air cargo industry, worldwide. IATA forecasts a rise in cargo carried to 62.5 million tonnes in 2018 (+4.5% on the 59.9 million tonnes in 2017) representing less than 1% of world trade by volume, but over 35% by value. The value of goods carried by airlines is expected to exceed \$6.2 trillion in 2018, representing 7.4% of world GDP. The growth of international trade and e-commerce, combined with companies' incentives to adopt inventory reduction and also keep lead times shorter, makes demand forecasting

in the framework of the air cargo industry a very important problem affecting numerous stakeholders. Speed is not the only reason air transport of cargo is chosen; reliability and predictability are also crucial factors for shippers. In this direction, the reliability of air cargo forecasting affects directly the reliability of the whole Industry. However, air cargo transport is a very complex process that encompasses many operations performed by various stakeholders (as shown in Figure 2-5), hence has many coordination requirements and inherent uncertainty (Feng, Li, and Shen 2015).

The main entities in a physical air cargo supply chain are the following three:

- Shippers, who are the customers of the air cargo services offered by Airlines.
- Freight forwarders, who are responsible for the physical movement (generally ground transportation) of the cargo, to and from the shippers' facilities.
- Airlines, the stakeholders that actually carry the cargo by air.

However, most Airlines outsource the management of their air cargo operations to specialized companies called Air Cargo Handling Companies (ACHC). In practice, Airlines also outsource the air cargo demand forecasting, so ACHCs are responsible to predict the tonnages of cargo to handle per month and per customer (airline) and make the necessary resource planning.



**Figure 2-5: A landscape of air cargo operations** (Feng, Li, and Shen 2015)

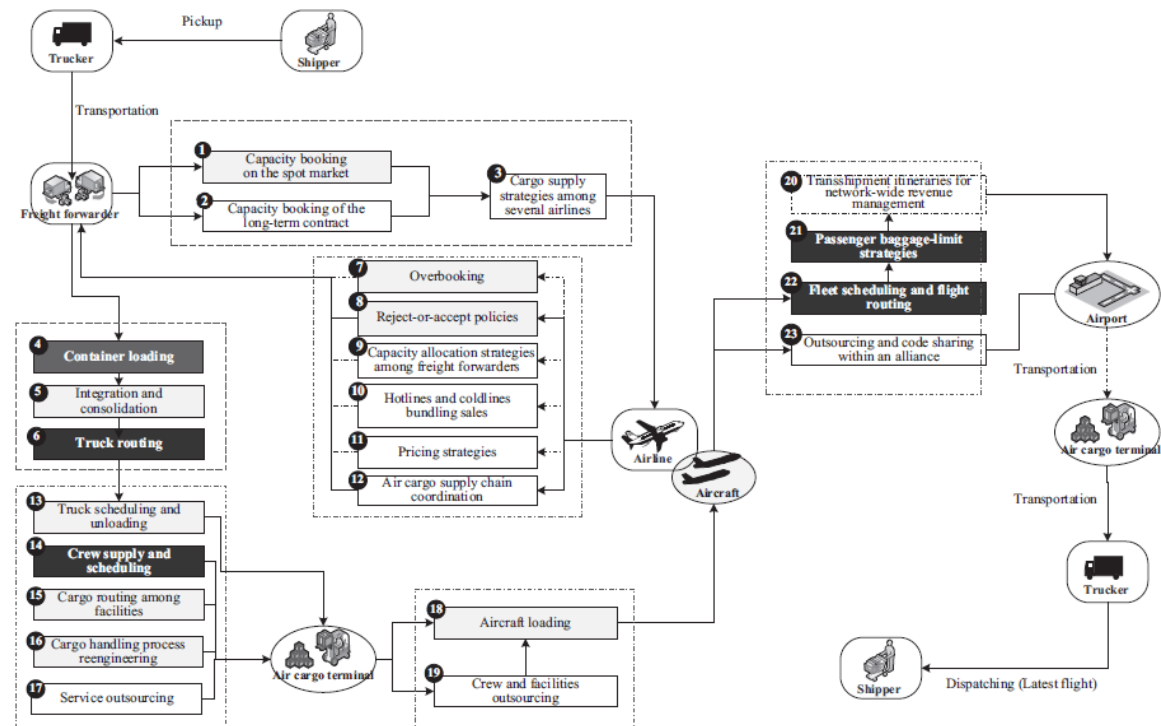
Cargo capacity is by itself a complex concept that causes uncertainty in the complete process, as it depends on the container types used, called unit load devices (ULDs), which are further specified by multiple dimensions, such as pivot weight, pivot volume, type, and centre of gravity. Capacity availability forecast has much higher uncertainty than passenger transport, also because ACHCs and freight forwarders need to plan the capacity booking for air cargo many months ahead of the actual transport. Furthermore, there is an imbalanced capacity supply for different routes and difficulty in implementing overbooking. **Cargo capacity and demand forecasting** is therefore a very challenging, actively researched, problem. Proposed solutions include approaches based on exponential

smoothing, time series analysis, econometric modelling, trend analysis and seasonality indices. These mostly focus on the long-run demand forecast and adopt a broader macroeconomic perspective, therefore such models include features related to population, employment rates, incomes per capita, GDP, GNP, economic growth rates, etc (Magaña, Mansouri, and Spiegler 2017). There are also short-term air cargo demand forecast approaches that employ techniques closer to data analytics. The aircraft journey section referred to two such approaches that leverage Artificial Neural Networks. (Baxter and Srisaeng 2018) is another approach where Artificial Neural Networks are used to predict Australia's annual export air cargo demand. (Chou, Liang, and Han 2013) propose a Fuzzy Regression Forecasting Model (FRFM) to forecast demand by examining present international air cargo market, leveraging fuzzy regression analysis to reduce the residual caused by uncertainty.

Accuracy of demand forecast is a decisive factor for the smoothness of the complete workflow, but there are many more operations that need to be planned and coordinated throughout the cargo journey, e.g. ***cargo scheduling for optimal pick up and drop off times***, which could include real-time data being captured by connected devices and forwarded to Warehouse Management Systems, Load Control Systems and even handling by AR (Augmented Reality) devices. Ground transportation vehicles could then adjust speed to optimize cargo arrival time at the airport facilities. ***Shipment health*** is also monitored throughout the flight, enabling event detection processes to trigger mitigation actions upon early warnings (NEXTT 2018). Indicatively, (Vijay et al. 2018) employ image processing and various classifiers to develop an air cargo monitoring system. Data mining techniques are used to ***measure the performance of freight forwarders*** and data-driven decision analysis is leveraged for improved ***cargo routing and scheduling, manpower requirements planning and personnel shift*** and ***truck scheduling*** (Feng, Li, and Shen 2015). From a different perspective, (Zhu, Yang, and He 2015) propose a custom co-clustering based dual prediction for ***cargo pricing optimisation***, which uses multi-label learning and multi-way clustering model to predict both the optimal price for the bid stage and the outcome of the transaction (win rate) in the decision stage of cargo pricing definition, given a pair of origin and destination. It should be clarified here that in the spot market of air cargo, airlines use dynamic pricing. Because of the hub-and-spoke structure used by most airlines, a dominant player or leader commonly exists for each regional market. Thus, dynamic pricing is often a leader–follower game.

The cargo journey is multi-faceted and comprises many interacting operations which can be either studied in isolation or in the real-life complex environment. (Feng, Li, and Shen 2015) provides a comprehensive diagram of the journey phases (as depicted in Figure 2-6) along with the research issues related to each step. Each issue may have different perspectives and approaches depending on the acting stakeholder, i.e. the stakeholder who identifies and models the problem.

It is instrumental to ensure that a holistic view of the cargo journey is provided at a high-level, however it is not in the scope of the current deliverable to delve into the details of each step in the process. Deliverable D2.3 will re-evaluate the use cases that need to be examined in ICARUS and, if needed, provide a state of the art review for additional cases of the ones discussed here.



**Figure 2-6: Air cargo journey operations state of the art (Feng, Li, and Shen 2015)**

As a closing remark of the four journey sections (2.2.1-2.2.4), it should be stated that in the context of the current state-of-the-art review, numerous papers were collected and studied, resulting in the extraction of the presented insights. Annex I provides indicative examples of the way information was initially extracted from each paper.

## 2.3 Key Considerations for Data Analytics in Aviation

The previous sections provide some preliminary insights regarding the role of data analytics in various aviation-related use cases and highlight the potential benefits across the complete value chain. The decision to adopt such processes raises also certain considerations that will be briefly discussed here.

### Data Availability

As revealed by the term data analytics, data constitute the most significant part of the process. The infamous “Garbage in – Garbage out” principle reflects the need to feed any data analytics process with appropriate data. The word appropriate is hard to define in this context, as it is not limited to the data content. Data structures, data veracity, data quality and, most of all, data availability and completeness need to be examined prior to any data analytics application. The data also significantly influence the selection of the most suitable algorithm and their pre-processing is in many cases the most demanding part of the whole process.

Data availability can also be seen from the data sharing perspective, i.e. the need to acquire data from another stakeholder, which will be discussed in detail in Section 4.

#### Type of Problem

First of all, it should be stressed that not all problems can or should be expressed as data analytics problems. Regarding machine learning in particular, whose advantages were individually examined and discussed as it is a very powerful data analysis method, it should be stressed that approaches that fall under the operations research category may be more suitable for certain problems.

Having concluded that a given problem can benefit from a data analytics solution, the next important consideration is how to select the most appropriate one or how to combine some models in a hybrid solution. Some insights were provided in this section based on the state of play for selected use cases and some more technical details and considerations will be provided in Section 3. Moreover, certain high-level guidelines as to the factors that need to be considered can be found, which may steer the user/ analyst towards some high-level directions. However, identifying the right option is a multifaceted problem and no predefined answers can be apriori provided.

#### Stakeholder Perspective

As shown in section 2.2, the aviation value chain is complex and use cases are highly interconnected and assumptions and simplifications are often required in order to examine a certain problem in isolation. Even then, each problem can be seen from diverse perspectives depending on the stakeholder who acts as the decision-maker. As in most business to business transactions, the objective of each involved party may differ, in this case resulting in a different formalisation and modelling of the same underlying processes. This fact may significantly impact not only the specified objective of the data analytics process and possible constraints, but also the selection of the most appropriate algorithm.

#### Technical Issues

Although this is not a technical section, there are inherent technical issues related to the application of a data analytics strategy that cannot be neglected. Computational needs, especially in a big data context, should be taken into consideration and appropriate infrastructures should be foreseen when developing a data analytics solution. Furthermore, other aspects such as the integration of data sources and extensibility of the adopted solution can also pose significant technical challenges along the way.

#### Data Analytics Mentality

The data analytics concept in an Industry perspective should not be treated as a must-have because it is a trend hype, especially in a complex environment like the one formed by the interacting aviation stakeholders. This would create unrealistic expectations and would push involved parties to develop rushed solutions, hindering the adoption of a truly beneficial data analytics strategy. Biases hidden in the data, datasets that fail to represent properly the underlying situation and overlooked correlations

are almost certain to appear in the first steps of a data analytics model development. A successful data analytics approach requires time and the will to experiment, evaluation and refine the developed models. Data analytics should be therefore be considered part of the core business processes.



## 3 ICARUS Data Analytics

---

### 3.1 Purpose

The highly competitive business environment in the aviation industry can be a clear example of a real-world big data pool. Structured and unstructured information is exchanged every second and every minute world-wide, regarding flight statuses, passenger control, in-flight entertainment and airport management. In such a frequently changing environment, future predictions and decision-making assistance could be vital, whereas simple visualisation of historical perspectives would be of little use. Leading enterprises in the aviation domain have already started gathering and exploiting big data using the latest technologies, while at the same time they are making significant investments in data science personnel. A recent example is Airbus and its Skywise data platform<sup>2</sup>, where many airlines and selected suppliers deposit and share their data in an innovating e-maintenance eco-system.

Having all these in mind, the purpose of data analytics in ICARUS is to offer multiple ways of producing, handling and monitoring data-driven knowledge to support decision-making in an aviation-related framework. In that context, the ICARUS platform should be able to provide a sufficient number of descriptive, predictive and prescriptive methods that are widely accepted by the research community, on the one hand, and efficiently applied in the aviation industry, on the other hand.

In particular, the ICARUS users should be able to perform a basic statistical analysis of historical data in order to evaluate the data, understand the past and examine how prior behaviours could affect the future. This descriptive analysis could be also assisted by data mining techniques, like clustering or feature correlation, to produce additional insights. The next step contains predictive machine learning algorithms that could be employed to forecast future trends and events, or fill in information that is missing. Deep learning methods could also be applied on dynamically updated data sets for further predictive analytics. Finally, on a more advanced level, prescriptive analytics should provide comprehensible solutions and recommend actions in the form of combined techniques and algorithms. Optimisation and decision-making tasks fall under this category.

ICARUS aims to facilitate all the aforementioned capabilities of data analytics considering widely accepted methods and algorithms already applied successfully in the aviation industry. In addition, ICARUS plans to provide pretested combinations of techniques as complete, domain-specific, methodologies that encompass different data sources from different providers (e.g. ICARUS pilot partners), as well as appropriately adapted algorithms, verified to produce explicit outcomes.

### 3.2 ICARUS Data Analytics Approach

Within the ICARUS context, an enormous volume of real-time data is expected to be generated on a daily basis. Data coming from multiple sources will be collected and curated, before being fed to a

---

<sup>2</sup> <https://www.airbus.com/aircraft/support-services/skywise.html>

data analytics process, where, not only the typical machine learning techniques, but also deep learning algorithms will be employed to infer knowledge and useful insights for the aviation-related stakeholders.

When speaking of data analytics, it needs to be noted that there are many processes involved besides the actual analysis of the information (as mentioned in D1.2, too). Real-world data can usually be characterised by heterogeneity, noise, incomplete attributes and several other features that hinder a proper analysis. For this reason, it makes sense to pre-process the data after collecting them to increase their value. This procedure may involve several stages, from data ingestion and cleansing to data transformation and dimensionality reduction, up until the actual data analysis and the visualisation of the results.

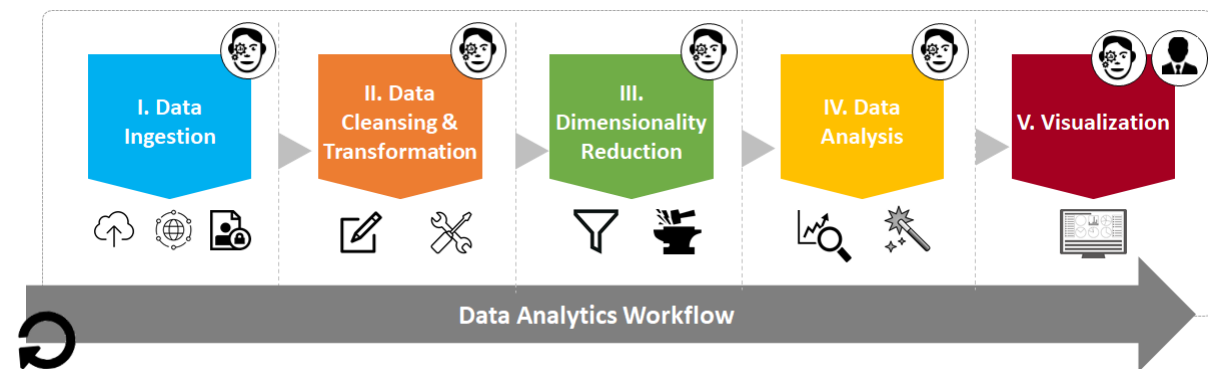


Figure 3-1: Typical Data Analytics Workflow (in line with ICARUS D1.2)

#### Step 1: Data Ingestion

The process of importing, transferring, and loading data from a variety of sources is called data ingestion. Data can be obtained in real-time (streaming data) or in batches, and are usually stored in a database or be available for immediate use (see D2.1 for more information).

#### Step 2: Data Cleansing and Transformation

This pre-processing step is the most valuable one since it determines the efficiency and accuracy of the analysis. Through data cleansing any corrupt, incomplete or erroneous data can be detected and amended or removed. This aims to increase the dataset quality, minimizing computational and resource costs (see D2.1 for more information).

#### Step 3: Dimensionality Reduction

The purpose of this optional step is to reduce the size of the dataset in order to achieve computational efficiency and quality increase. There are many techniques available for dimensionality reduction, such as sampling, feature selection methods, sketching techniques for stream of data, and transformation related algorithms for feature extraction.

#### Step 4: Data Analysis

Data analysis is the process of systematically evaluating the data and discovering useful information using statistical or machine learning techniques. Data analysis can also be viewed as asking questions about what happened, what is happening, and what will happen, with the answers being extracted

from data examination and processing. In essence, such a data analysis encapsulates an iterative cycle of feature engineering, model fitting and model evaluation until the data analyst is satisfied with the outcomes.

#### Step 5: Visualisation

The visual representation of the results of data analysis, in the form of charts, tables, line graphs, column charts, and many other forms, offers a comprehensive picture of the produced insights and predictions, making also explicit the trends and patterns inherent in the data.

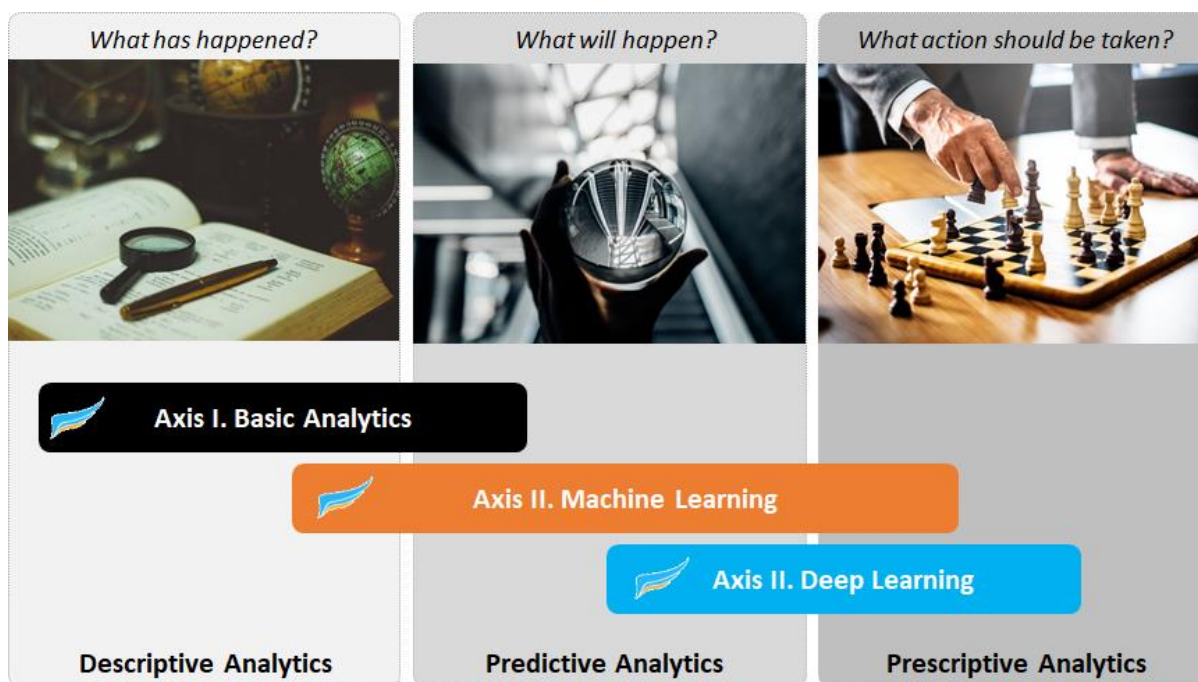
The ICARUS data analytics approach aims to address efficiently all of the aforementioned steps. For the analysis phase in particular, ICARUS will explore several methods and algorithms that cover most of the complex aspects encountered in the aviation industry. These algorithms were selected with the following criteria in mind: a) to adhere to the ICARUS platform requirements and be applicable to aviation specific tasks, b) to have proven their ability and robustness in the research community through the years, and c) to have been implemented in a commonly used software framework or library.

For the last criterion, several popular software frameworks and libraries have been considered. More specifically:

- Spark MLlib (<https://spark.apache.org/mllib/>)
- Scikit-learn (<https://github.com/databricks/spark-sklearn>)
- Tensorflow (<https://www.tensorflow.org/>)
- Keras (<https://keras.io/>)
- H2O (<https://h2o.ai/>)
- Deeplearning4j (<https://deeplearning4j.org/>)
- BigDL (<https://bigdl-project.github.io/>)
- PyTorch (<https://pytorch.org/>)
- Caffe, Caffe2 (<http://caffe.berkeleyvision.org/>, <https://caffe2.ai>)
- Apache Mxnet (<https://mxnet.apache.org/>)
- Microsoft CNTK (<https://www.microsoft.com/en-us/cognitive-toolkit/>)
- Custom implementations in Python, Java or R

The analytical study that follows relies on three axes, as shown in Figure 3-2.

The first axis describes methods mostly utilized when one wants to examine and understand the data at hand, and extract simple predictive insights. Consequently, most of these methods derive from the field of statistics. The second axis covers the field of machine learning, presenting the most representative algorithms for descriptive, predictive, as well as prescriptive analytics. Lastly, the third axis includes the most recent deep learning techniques, which are able to handle big data efficiently and result into useful predictions or important conclusions.



**Figure 3-2: Relation of the proposed methodology to descriptive, predictive and prescriptive analytics.**

By examining the figure, it is noticeable that the three axes can be partially overlapping in scope. This is true in a sense that a number of algorithms belonging to one axis could be applied on the same tasks that the algorithms of another axis could also tackle. However, the appropriate selection of an algorithm is actually depending on the nature of the problem at hand, the parameters involved and the way each approach leads to a solution. It is, therefore, required for ICARUS to provide a variety of techniques for the same analysis, so that the analyst should be able to employ the best candidate for each task.

### 3.3 Axis I: Basic Analytics

Basic analytics is a rather vague term to describe a set of algorithms and statistical methods commonly employed by an analyst as an initial step in order to examine and understand a dataset. In that sense, most of these approaches belong to the descriptive or diagnostic analytics category.

They are usually applied on raw data or content in order to visualize what the data is about, which input features are correlated or why something has occurred (test a hypothesis). This frequently provides an analyst with a quick and quite accurate overview of the information available, which actually determines the next steps of the data analysis and the appropriate predictive or prescriptive algorithms to apply.

The following table summarizes the suggested methods assigning each one to its respective algorithmic family. All of these methods will be presented in detail in the lines that follow.

**Table 3-1: Basic Analytics Algorithms**

Section	Algorithm Family	Algorithm Name
3.3.1	Statistical Analysis	Summary Statistics (mean, std, etc.)
3.3.2	Statistical Analysis	Hypothesis Testing
3.3.3	Statistical Analysis	Sampling
3.3.4	Feature Correlation	Pearson's and Spearman's Correlation
3.3.5	Feature Correlation, Regression Analysis	Linear Regression methods
3.3.6	Feature Correlation, Regression Analysis, Classification	Logistic Regression
3.3.7	Dimensionality reduction, Feature extraction	Principal component analysis (PCA)
3.3.8	Dimensionality reduction	Feature Selection
3.3.9	Time series prediction	Autoregressive Integrated Moving Average (ARIMA)

### 3.3.1 Summary Statistics

Algorithm Family	Statistical Analysis
Description	Summary Statistics (or Descriptive Statistics) are useful to quickly describe a set of observations and explore numerical variables as simply as possible. In other words, they provide descriptive information about the dataset available. Summary Statistics can include measures of location (Mean, Median, Mode, Minimum Value, Maximum Value, etc.) and measures of dispersion or spread (Range, Standard Deviation, etc.). In the case of categorical variables, a Frequency distribution is recommendable for exploratory analysis. Knowing such basic statistics regarding each attribute of a data set makes it easier to fill in missing values, smooth noisy values, and spot outliers.
Typical Applications	Summary statistics are used during the data validation step to explore a data set. Data validation is a decisional procedure ending with the acceptance or refusal of data attributes. The decisional procedure is usually based on rules expressing the acceptable combinations of values. Rules are applied to data. If data satisfy the rules, which means that the combination expressed by the rules is not violated, data are considered valid for final usage.
Application examples in the aviation domain	Summary statistics can be employed on air traffic or airport data in order to provide some initial information and reveal particular trends in aviation data.
Popular Libraries / Software tools	Custom implementation in Python (scipy.stats), R (base).

### 3.3.2 Hypothesis Testing

Algorithm Family	Statistical Analysis
Description	Hypothesis testing is a method to determine the probability that a given hypothesis is true. It consists of: <ul style="list-style-type: none"> <li>• formulation of the null hypothesis (observations are the result of pure chance);</li> <li>• identification of a statistic that can assess the truth of the null hypothesis;</li> <li>• computation of the P-value (the probability of the statistic under the null hypothesis, the smaller the P-value, the stronger the evidence against the null hypothesis);</li> </ul>

	<ul style="list-style-type: none"> <li>comparison of the p-value to a significance value alpha.</li> </ul>
<b>Typical Applications</b>	Hypothesis testing refers to statistical procedures used to accept or reject a statistical hypothesis, usually involving two datasets and the relationship between them. If the result is statistically significant, then the statistical hypothesis can be accepted else it should be rejected.
<b>Application examples in the aviation domain</b>	Hypothesis testing can be used to compare the on-time performance of flights (means comparison) among airports. Hypothesis testing is also applicable for most of the algorithms to assess the statistical significance of the results.
<b>Popular Libraries / Software tools</b>	Custom implementation in Python (scipy.stats), R (stats, specific packages).
<b>References</b>	<p>Stuart A., Ord K., Arnold S. (1999), <i>Kendall's Advanced Theory of Statistics: Volume 2A—Classical Inference &amp; the Linear Model (Arnold)</i></p> <p>Kennedy, Q., Taylor, J. L., Reade, G., &amp; Yesavage, J. A. (2010). Age and expertise effects in aviation decision making and flight control in a flight simulator. <i>Aviation, space, and environmental medicine</i>, 81(5), 489-497.</p>

### 3.3.3 Sampling

<b>Algorithm Family</b>	Statistical Analysis
<b>Description</b>	
Sampling is the process of selecting a representative sample from a target population and collecting data to infer something about the population as a whole. There are two main types of sampling: probability and non-probability sampling. The difference between the two types is whether the sampling selection involves randomisation. Randomisation occurs when all members of the sampling frame have an equal opportunity of selection. The most basic type of sampling is Simple Random Sampling where sample members are selected randomly and purely by chance.	
<b>Typical Applications</b>	Frequently used in market research and statistical surveys.
<b>Application examples in the aviation domain</b>	Sampling technique can be used to conduct analysis on passengers' opinions on airport and airline services.
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Stratified random sample: The population is split into groups. The overall sample consists of some members from every group. The members from each group are selected randomly.</li> <li>Cluster random sample: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.</li> <li>Systematic random sample: uses a specific system to select members.</li> <li>Non-probability sampling: there are several sampling methods of this category, like Quota, Convenience, Snowball, etc. The disadvantages of these methods are that it is impossible to know how well they are representing the population.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python (pandas), R (base, sampling).
<b>References</b>	Lohr, S. L. (1999). <i>Sampling: Design and Analysis</i> . Duxbury Press. Pacific Grove, CA.

	<a href="https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Consumer-research/Departing-passenger-survey/Sampling-methodology/">https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Consumer-research/Departing-passenger-survey/Sampling-methodology/</a> accessed on line 05/11/2018
--	---

### 3.3.4 Pearson's and Spearman's Correlation

<b>Algorithm Family</b>	Feature Correlation
<b>Description</b>	
<p>A correlation coefficient measures the extent to which two variables tend to change together. The coefficient describes both the strength and the direction (negative or positive) of the relationship between the two variables.</p> <p><b>Pearson's product moment correlation</b>, or simply Pearson correlation, is a measure that ranges from -1 to 1, with -1 indicating a perfect negative linear relationship and 1 a perfect positive linear relationship between the two variables. A zero value is an indication of no linear relationship.</p> <p><b>Spearman's rank-order correlation</b> is also a measure that summarizes the strength and direction of a relationship using the same range as Pearson correlation. However, in Spearman's case, the correlation coefficient is based on the ranking (ordering) of each variable's values and not on the raw data values.</p>	
<b>Typical Applications</b>	Both methods are commonly employed to evaluate the mutual association between two sets of data and can be applied as a first step to better understand the underlying relationships and subsequently build better statistical models.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• Provided a dataset with measurements of an airplane's environmental impact for multiple flights as well as weather conditions, length of trip, altitude, it can be assessed whether there is any correlation between the environmental impact data with any of the other features.</li> <li>• Connecting flights could be used as a ranking variable and be evaluated against any other variable related to the flight (e.g. the number of passengers at each connecting point).</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• Kendall rank correlation</li> <li>• Point-Biserial correlation</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementation in Python (scipy.stats), R (base).
<b>References</b>	<p>Kendall, M. G., &amp; Gibbons, J. D. (1990). <i>Rank Correlation Methods</i> (5th ed.). London: Edward Arnold.</p> <p>Tsamboulas, D. A., &amp; Nikoleris, A. (2008). Passengers' willingness to pay for airport ground access time savings. <i>Transportation Research Part A: Policy and Practice</i>, 42(10), 1274-1282.</p>

### 3.3.5 Linear Regression methods

<b>Algorithm Family</b>	Feature Correlation, Regression Analysis
<b>Description</b>	
<p>Linear Regression is one of the most widely known variable modelling techniques. The dependent variable is continuous while the independent variables (one or more) can be continuous or discrete. Linear regression makes the assumption that any changes in the dependent variable can be modelled as a monotonic linear function of the independent variables. Ordinary Least Squares (OLS) is a method for estimating the unknown</p>	



parameters with the goal of minimising the sum of the squares of the differences between the observed responses (values of the variable being predicted) in the given dataset and those predicted by a linear function of a set of explanatory variables.	
<b>Typical Applications</b>	Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Regression methods can be used to predict the aircraft performance parameters as accurately as possible, using the data given in the Aircraft Flight Manual (AFM) to reduce fuel consumption.</li> <li>Regression can also be applied to flight delay and taxi-out time prediction.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference.</li> <li>Polynomial Regression is a regression equation when the power of independent variable is more than 1. In this non-linear regression technique, the best fit line is a curve that fits into the data points.</li> <li>Stepwise Regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.</li> <li>Ridge Regression is a technique used when the data suffers from multicollinearity. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.</li> <li>Lasso (Least Absolute Shrinkage and Selection Operator) also penalises the absolute size of the regression coefficients. In addition, it can reduce the variability and improving the accuracy of linear regression models</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (stats, glm, glmnet, MASS, biglasso).
<b>References</b>	<p><i>Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley.</i></p> <p>Harju, T, (2017). Derivation of Aircraft Performance Parameters Applying Machine Learning Principles. <i>Master Thesis, Aalto University School of Engineering.</i></p>

### 3.3.6 Logistic regression

<b>Algorithm Family</b>	Feature Correlation, Regression Analysis, Classification
<b>Description</b>	
Logistic regression is a variant of Linear regression when the dependent variable is binary or dichotomous. If the dependent variable has more than two possible values, it is possible to use its extension Multinomial Logistic Regression.	
<b>Typical Applications</b>	Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Logistic Regression can be used to predict evaluate the customer satisfaction and quality of services of an airline.</li> <li>Multinomial Logistic Regression Model can be used for Predicting Flight Arrival &amp; Delay using Flight On-time performance and weather data.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Ordered Logistic regression</li> </ul>



<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (glm, MASS).
<b>References</b>	<p>Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". <i>J Roy Stat Soc B</i>. 20 (2): 215–242.</p> <p>De La Vina, L., &amp; Ford, J. (2001). Logistic regression analysis of cruise vacation market potential: Demographic and trip attribute perception factors. <i>Journal of Travel Research</i>, 39(4), 406-410.</p>

### 3.3.7 Principal component analysis

<b>Algorithm Family</b>	Feature extraction, Dimensionality reduction
<b>Description</b>	
The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set, while retaining the variability present in the data set. This is achieved by transforming the original variables to a new set, the principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the original variability.	
<b>Typical Applications</b>	PCA is mostly used for data exploration and as a pre-processing step of predictive modelling. The first principal components can be represented in a 2D or 3D visualisation for better understanding of the dataset.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Large datasets are increasingly common and are often difficult to interpret in aviation domain. PCA is able to reduce the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. For example, applying PCA to airline quality profiles can provide us with a two-dimensional solution that clearly exposes each airline's competitive positioning in the field and the competitive groups that are formed.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Correspondence Analysis (CA), for categorical data</li> <li>Independent Component Analysis (ICA)</li> <li>Independent PCA (iPCA),</li> <li>Sparse PCA (sPCA)</li> <li>Sparse independent PCA (siPCA)</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (FactoMineR, stats).
<b>References</b>	<p>Pearson, K. (1901). Principal components analysis. In <i>The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science</i>.</p> <p>Jolliffe, I. T., &amp; Cadima, J. (2016). Principal component analysis: a review and recent developments. <i>Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences</i>.</p> <p>Gursoy, D., Chen, M. H., &amp; Kim, H. J. (2005). The US airlines relative positioning based on attributes of service quality. <i>Tourism management</i>, 26(1), 57-67.</p>

### 3.3.8 Features selection

<b>Algorithm Family</b>	Dimensionality reduction
<b>Description</b>	

<p>The number of variables (features) of a data set can be reduced using the results of the application of different filter methods. In classification and regression tasks, it is possible to test strength of the correlation or connection between the target variable and all the other variables. The variables more related to the target variables can be kept in the training step. Different statistics or mathematical formulas can be used for this task depending on the nature of variables (numeric or categorical). Examples of these techniques are Gain Ratio, Chi Squared, R Squared, Singular Value Decomposition, etc.</p> <p>Filter methods can rank individual features or evaluate whole feature subsets. Feature subset generation for multivariate filters depends on the search strategy. While there are many search strategies, there are four usual starting points for feature subset generation: 1) forward selection, 2) backward elimination, 3) bidirectional selection, and 4) heuristic feature subset selection.</p>	
<b>Typical Applications</b>	Feature selection (FS) methods are used in data pre-processing to achieve effective data reduction. This is useful for improving the model accuracy and avoiding the “curse of dimensionality”. Typical examples can be found in the fields of cryptanalysis and bioinformatics, where there are many properties (input variables) to be compared and evaluated.
<b>Application examples in the aviation domain</b>	In general, a feature selection technique is utilized in cases where the available dataset is characterized by high dimensionality and the input should remain untransformed. In the aviation domain, such datasets could be the operational data of a flight or an airport.
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• Univariate selection</li> <li>• Recursive Feature Elimination (RFE)</li> <li>• Feature importance</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (caret).
<b>References</b>	<p>Jović, A., Brkić, K., &amp; Bogunović, N. (2015). A review of feature selection methods with applications. In <i>2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings</i> (pp. 1200–1205). Institute of Electrical and Electronics Engineers Inc.</p> <p>Lukacova, A., Babic, F., &amp; Paralic, J. (2014, January). Building the prediction model from the aviation incident data. In <i>Applied Machine Intelligence and Informatics (SAMII), 2014 IEEE 12th International Symposium on</i> (pp. 365-369). IEEE.</p>

### 3.3.9 Autoregressive Integrated Moving Average (ARIMA)

<b>Algorithm Family</b>	Time Series Prediction
<b>Description</b>	
<p>The autoregressive integrated moving average (ARIMA), also known as the Box-Jenkins methodology, is a very popular statistical model for time series analysis and prediction. ARIMA models apply usually to stationary data (where there is no seasonality, trend, etc.) and can be estimated using the Box-Jenkins approach.</p>	
<b>Typical Applications</b>	Common forecasting applications such as weather predictions, stock market or tourism forecasting.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• An ARIMA model could be used to perform time series analysis in a number of aviation applications. For example, common forecasting usages include air passenger traffic, aircraft orders and deliveries, airlines’ seat sales, baggage carousel planning, or flight delay estimation.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• Seasonal ARIMA, when a seasonal effect is expected in the series.</li> </ul>

	<ul style="list-style-type: none"> <li>Fractional ARIMA, when a long-range dependence is suspected in the series.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python (statsmodel package), Spark (spark-timeseries), R (stats, forecast).
<b>References</b>	<p>Asteriou, Dimitros; Hall, Stephen G. (2011). "ARIMA Models and the Box–Jenkins Methodology". <i>Applied Econometrics</i> (Second ed.). Palgrave MacMillan. pp. 265–286.</p> <p>Pitfield, D. E. (2007). Ryanair's impact on airline market share from the London area airports: a time series analysis. <i>Journal of Transport Economics and Policy (JTEP)</i>, 41(1), 75-92.</p> <p>Yoon, S. W., &amp; Jeong, S. J. (2015). An alternative methodology for planning baggage carousel capacity expansion: A case study of Incheon International Airport. <i>Journal of Air Transport Management</i>, 42, 63–74.</p> <p>Cheng, J. (2015). Estimation of flight delay using weighted Spline combined with ARIMA model. In <i>Proceedings of 2014 IEEE 7th International Conference on Advanced Infocomm Technology, IEEE/ICAIT 2014</i> (pp. 8–20). Institute of Electrical and Electronics Engineers Inc.</p>

### 3.4 Axis II: Machine Learning Algorithms

The algorithms of this section belong to the broad field of Machine Learning and tackle more complex and resource intensive tasks. Their primary concern is to detect useful patterns within the given datasets, form comprehensive and informative rules and predict future trends or behaviours. For these reasons, most of them can be eligible to tackle problems of the whole spectrum of data analytics.

A common characteristic shared by all machine learning algorithms is that they require some form of training before being able to perform efficiently. In other words, they need to learn a mapping from inputs to outputs. The training of the algorithms takes usually three forms that are known as supervised, unsupervised and reinforcement learning. Supervised learning requires a sample of the data to be labelled in some useful way, so that the system is aware of the output during training and can thus optimize its performance. On the contrary, unsupervised learning requires no labels in order to produce an output. This is achieved by identifying similarities or patterns in the data and when a new input is inserted, the system reacts according to the defined similarity measures in order to cluster or classify it. Reinforcement learning (RL), on the other hand, is mostly concerned with how software agents behave within a specific environment and follows a totally different approach. RL algorithms try to find the best strategy of actions for an agent in order to earn the maximum reward based on feedback from the environment. Common RL applications involve robotic movement, self-driving cars, gaming, and other goal-oriented tasks.

Provided with proper training from historical observations, machine learning algorithms exhibit the ability to generalise in any given problem. This means that they are capable of adapting automatically to new, previously unseen data, without the need to be explicitly programmed to do so, which makes machine learning algorithms particularly suitable for the dynamic environment of the aviation industry.

In this context, a number of representative machine learning algorithms were selected based on the following criteria:

- to be relevant to the demonstrators' needs
- to have proven their ability and robustness in aviation-related tasks
- to have been implemented in a commonly used software framework or library

An introductory summary of the selected techniques is presented in the table below. The detailed descriptions follow.

**Table 3-2: Machine Learning Algorithms**

Section	Algorithm Family	Algorithm Name
3.4.1	Clustering, Dimensionality Reduction	Self-Organising Map (SOM)
3.4.2	Clustering, Anomaly Detection	K-means
3.4.3	Clustering, Anomaly Detection	Streaming K-means
3.4.4	Clustering, Anomaly Detection	DBSCAN
3.4.5	Clustering	Gaussian Mixture models
3.4.6	Association Rules	Apriori
3.4.7	Recommendation Systems	Collaborative Filtering (CF)
3.4.8	Recommendation Systems	Content-based Filtering (CBF)
3.4.9	Classification, Regression, Outlier detection	Support Vector Machines (SVM)
3.4.10	Classification, Regression	Classification and Regression Tree (CART)
3.4.11	Classification, Regression, Outlier detection	Random Forest (RF)
0	Classification, Regression	Gradient Boosting Machines (GBM)
3.4.13	Classification, Regression, Outlier detection	K-NN
3.4.14	Classification	Naïve Bayes (NB)
3.4.15	Classification, Regression, Time Series Prediction	Multi-Layer Perceptron (MLP)
3.4.16	Classification, Regression, Time Series Prediction	Adaptive Neuro-Fuzzy Inference System (ANFIS)
3.4.17	Optimisation	Genetic Algorithms (GA)

#### 3.4.1 Self-Organising Map

<b>Algorithm Family</b>	Clustering, Dimensionality Reduction
<b>Description</b>	
A self-organizing map (SOM), also known as Kohonen network, is a popular neural network typically utilized for clustering and dimensionality reduction purposes. SOMs apply competitive (unsupervised) learning to produce a low-dimensional (usually 2-D) map of the input space, while trying to preserve its topological properties.	
<b>Typical Applications</b>	High-dimensional data exploration, text clustering, weather prediction, marketing, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• A common application of SOM networks is in the grouping of consumer transactions, which results in a mapping of similar consumer behaviour that can be easily visualized and function as a compass for a future marketing</li> </ul>

	<p>strategy. This could be applied to Duty-free shop transactions, along with flights information, and form clusters of similar customer mentality and behaviour, which would then operate as predictive models for personalized promotional offers.</p> <ul style="list-style-type: none"> <li>• Create an input space using passenger demographics, flight data and destination details to suggest new routes and destinations for specific target groups. The SOM network will use a competitive learning technique to assemble clusters of passengers and destinations based on their characteristics.</li> <li>• Use SOM for diagnostics on aircraft engine condition data.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• <i>Growing self-organizing map (GSOM)</i> is a growing variant of SOM in the sense that it starts with a minimal number of nodes and grows new ones based on a heuristic method.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python (SomPy), R (Kohonen, class).
<b>Considerations</b>	Data should be represented as vectors.
<b>References</b>	<p>Kohonen, T. (2012). <i>Self-organisation and associative memory</i> (Vol. 8). Springer Science &amp; Business Media.</p> <p>Cumming, S. (1993). Neural networks for monitoring of engine condition data. <i>Neural Computing &amp; Applications</i>, 1(1), 96-102.</p>

### 3.4.2 K-Means

<b>Algorithm Family</b>	Clustering, Anomaly Detection
<b>Description</b>	<p>The K-Means algorithm clusters data in a chosen number of groups of equal variances, by minimizing a criterion known as the inertia or within-cluster sum-of-squares. The algorithm requires specifying a priori the desired number of clusters. It scales well when the number of observations is quite large.</p>
<b>Typical Applications</b>	K-means (and its variations) is one of the most commonly used clustering algorithms: they have been successfully used in various topics, including market segmentation, computer vision, geostatistics, astronomy and agriculture. It is often used as a pre-processing (initialisation) step for other algorithms, for example to find a starting configuration.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• Detect and characterize anomalies in large sets of high-dimensional symbol sequences that arise from recordings of switch sensors in the cockpits of commercial airliners. The approach taken uses unsupervised clustering of sequences using the normalized length of the longest common subsequence as a similarity measure, followed by detailed outlier analysis to detect anomalies.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• K-medoids</li> <li>• k-means++</li> <li>• k-means   (parallelized version of k-means++)</li> <li>• Mini Batch K-Means</li> <li>• Bisecting K-Means</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (stats, sparklyr, cluster).

<b>Considerations</b>	<p>K-means can only handle numerical values and requires a large set of data.</p> <p>K-means also assumes that the clusters are of spherical shape and have roughly equal number of samples.</p>
<b>References</b>	<p>Kaufman, L., &amp; Rousseeuw, P. J. (1990). <i>Finding Groups in Data: An Introduction to Cluster Analysis</i> (Wiley Series in Probability and Statistics).</p> <p>Budalakoti, S., Srivastava, A. N., &amp; Otey, M. E. (2009). Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. <i>IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews</i>, 39(1), 101–113.</p>

### 3.4.3 Streaming K-Means

<b>Algorithm Family</b>	Clustering, Anomaly Detection
<b>Description</b>	<p>Streaming algorithms were introduced in order to analyse data arriving through a real-time stream. Streaming k-Means is a streaming clustering algorithm that forms clusters in a dynamic way and updates them when new data arrive.</p>
<b>Typical Applications</b>	Streaming k-Means is used to cluster data that arrive continuously such as telephone records, multimedia data, financial transactions, sensor measurements, etc.
<b>Application examples in the aviation domain</b>	While traditional clustering algorithms merely considered static data, today's applications and research issues in data mining have to deal with continuous, possibly infinite streams of data, arriving at high velocity. Air traffic data, click streams, surveillance data, sensor measurements, customer profile data are only some examples of these daily-increasing applications in the aviation domain.
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• BIRCH builds a hierarchical data structure to incrementally cluster the incoming points using the available memory and minimizing the amount of I/O required.</li> <li>• COBWEB is an incremental clustering technique that keeps a hierarchical clustering model in the form of a classification tree.</li> <li>• C2ICM builds a flat partitioning clustering structure by selecting some objects as cluster seeds and a non-seed is assigned to the seed that provides the highest coverage.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementation in Spark (MLlib), R (stream).
<b>References</b>	<p>Guha, S.; Mishra, N.; Motwani, R.; O'Callaghan, L. (2000). "Clustering Data Streams". <i>Proceedings of the Annual Symposium on Foundations of Computer Science</i>.</p> <p>Ailon, N., Jaiswal, R., &amp; Monteleoni, C. (2009). Streaming k-means approximation. In <i>Advances in neural information processing systems</i> (pp. 10-18).</p>

### 3.4.4 DBSCAN

<b>Algorithm Family</b>	Clustering, Anomaly detection
-------------------------	-------------------------------

Description	
<p>The DBSCAN algorithm is a density-based clustering algorithm. It views clusters as areas of high density separated by areas of low density. Due to this generic approach, clusters found by DBSCAN can be of any shape, as opposed to k-means which assumes that clusters are convex shaped. The data points in regions of low point density are typically considered noise or outliers. The algorithm does not require the number of clusters to be pre-specified.</p>	
<b>Typical Applications</b>	<p>Density-based clustering are particularly suitable for applications where clusters cannot be well described as distinct groups of low within-cluster dissimilarity. This is the case in spatial data (e.g. satellite images) where clusters of points in an area may be formed due to natural structures such as rivers, seismic faults, etc.</p>
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• DBSCAN shows promising results when applied to the characterisation of traffic flows based on recorded radar tracks.</li> <li>• DBSCAN can be used to detect abnormal flights based on Flight Data Recorder (FDR) data. Results from cluster analysis are provided to domain experts to verify operational significance of such anomalies and associated safety hazards.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• HDBSCAN: Hierarchical DBSCAN with simplified hierarchy extraction.</li> <li>• OPTICS/OPTICSXi: Ordering points to identify the clustering structure clustering algorithms.</li> </ul>
<b>Popular Libraries / Software tools</b>	<p>Custom implementations in Python (scikit-learn), R (dbscan).</p>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Datasets with altering or varying densities cannot be handled well.</li> <li>• Fails to identify clusters if dataset is too sparse.</li> <li>• Sensitive to clustering parameters and to sampling.</li> </ul>
<b>References</b>	<p>Ester, M., Kriegel, H. P., Sander, J., &amp; Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. <i>Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining</i>.</p> <p>Basora, L., Morio, J., Mailhot, C. (2017). A Trajectory Clustering Framework to Analyse Air Traffic Flows. <i>SIDs 2017, 7th SESAR Innovation Days</i>, Belgrade, Serbia.</p> <p>Lishuai, L. (2013). Anomaly detection in airline routine operations using flight data recorder data, <i>Thesis (Ph. D.)-Massachusetts Institute of Technology</i>.</p>

### 3.4.5 Gaussian mixture models

<b>Algorithm Family</b>	Clustering
Description	
<p>A typical application of Gaussian Mixture Models (GMMs) is clustering. The observations are assigned to a cluster by maximizing the posterior probability that a data point belongs to the cluster itself. Like other clustering methods, the number of desired clusters must be specified before training the model. Other parameters for GMMs are component covariance structure, initial conditions and regularisation parameter.</p>	
<b>Typical Applications</b>	<p>GMMs are useful when there is need of a probabilistic classification of the observations. The probability of belonging to each cluster is calculated and a classification is achieved by assigning each observation to the most likely cluster.</p>



<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Based on Gaussian mixture models (GMM), speed, flow and occupancy are used together in the cluster analysis of traffic flow data. Compared with other clustering and sorting techniques, as a structural model, the GMM is suitable for various kinds of traffic flow parameters. Clustering analysis and pattern recognition can also be used to cluster and classify dynamic traffic flow patterns for freeway on-ramp and off-ramp weaving sections as well as for other facilities or things involving the concept of level of service, such as airports, parking lots, intersections, interrupted-flow pedestrian facilities, etc.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python (scikit-learn), Spark (MLlib), R (mclust).
<b>Considerations</b>	Works well when patterns follow a gaussian distribution
<b>References</b>	<p>A.P. Dempster, N.M. Laird, and D.B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM algorithm", <i>Journal of the Royal Statistical Society, Series B</i>, vol. 39, 1:1-38</p> <p>Sun, L., Zhang, H., Gao, R., Chen, L. (2011). Gaussian mixture models for clustering and classifying traffic flow in real-time for traffic operation and management, <i>Journal of Southeast University (English Edition)</i> 27(2):174-179</p>

#### 3.4.6 Apriori

<b>Algorithm Family</b>	Association Rules
<b>Description</b>	<p>Apriori is an algorithm to discover association rules. It proceeds in two steps:</p> <ul style="list-style-type: none"> <li>generation of all frequent item sets above the minimum support;</li> <li>generation of the association rules above a minimum confidence.</li> </ul>
<b>Typical Applications</b>	<p>Association rules are used for:</p> <ul style="list-style-type: none"> <li>Basket data analysis: to plan product placement in a storefront, to run a marketing campaign or to design a business catalogue,</li> <li>Web usage mining and intrusion detection: to security threats and network performance issues.</li> <li>Bioinformatics: to discover to intrinsic associations between gene annotations and expression data.</li> </ul>
<b>Application examples in the aviation domain</b>	<p>Association Rules allow the mapping of passenger behaviour in different moments of flight experience, such as:</p> <ul style="list-style-type: none"> <li>purchases of goods at the airport and during the flight.</li> <li>purchases of In-Flight-Entertainment services (i.e. movies and music).</li> </ul> <p>Additionally, quality rules can be discovered in safety-related datasets, like the aviation safety incident reports.</p>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li><i>The FP-Growth Algorithm</i> is an efficient and scalable method for mining the complete set of frequent patterns. Han proved that this method outperforms other popular methods for mining frequent patterns, e.g. the Apriori Algorithm and the TreeProjection.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementation in Spark (MLlib), Python (apyori), R (arules, arulesviz).

<b>Considerations</b>	<p>The algorithm can become very slow when there are many alternatives to analyse, due to its time and space complexity.</p> <p>Suffers from efficiency issues.</p>
<b>References</b>	<p>Agrawal, R., &amp; Srikant, R. (1994). Fast algorithms for mining association rules. <i>In 94 Proceedings of the 20th International Conference on Very Large Data Bases</i>.</p> <p>Han, J., Pei, J., &amp; Yin, Y. (2000). Mining frequent patterns without candidate generation. <i>In Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00</i>.</p> <p>Nazeri, Z., &amp; Bloedorn, E. (2004). Exploiting available domain knowledge to improve mining aviation safety and network security data. <i>The MITRE Corporation, McLean, Virginia, 22102</i>.</p>

### 3.4.7 Collaborative filtering

<b>Algorithm Family</b>	Recommendation systems
<b>Description</b>	<p>One of the predictive processes of Recommendation Systems is Collaborative filtering (CF). Collaborative filtering draws inferences about the relationship between user preferences and items or services, by mimicking user-to-user recommendations.</p>
<b>Typical Applications</b>	Collaborative filtering is used in recommendation systems. They seek to predict the "rating" or "preference" a user would give to an item.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Recommendation system can be used into an end-to-end solution that enhances and unifies all services that a passenger requires before, during and after a flight, such as a car rental, taxi bookings, retail purchases and entertainment. It also increases revenue for the airlines and elongate customer retention "long before" and "long after" a flight. The Recommendation System provides personalised information and recommendations and offers value - for - money services to passengers for the entire journey.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Hybrid approaches that combine both collaborative filtering and content-based filtering.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementation in Python, Spark (MLlib), R (rrecsys).
<b>Considerations</b>	Collaborative filtering suffers from the "cold start" effect. This implies the requirement of a large dataset of active users who have rated a product before, in order to make accurate predictions. Thus, the algorithm becomes impractical for new products which still have no ratings.
<b>References</b>	<p>Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. <i>Internet Computing, IEEE, 7(1), 76-80</i>.</p> <p>Ricci, F., Rokach, L., Shapira, B., and Kantor, P.B. (2011). <i>Introduction to Recommender Systems Handbook</i>. Springer.</p> <p>Frank, C. P., &amp; Deveraux, M. N. (2015). A consumer-centric methodology for selecting architectures against multiple objectives and its application to the in-flight catering and entertainment system. <i>In 6th European Conference for Aerospace Sciences</i>.</p>

### 3.4.8 Content-based filtering

<b>Algorithm Family</b>	Recommendation systems
<b>Description</b>	
Content-based filtering (CBF) methods, also known as cognitive filtering methods, base their recommendations on a comparison between the content of an item (e.g. description) and a user profile. The content of an item is represented as a set of terms, typically keywords, and a user profile is built up by analysing the terms of items the user has already viewed or liked. A content-based recommender system makes automatic predictions (filtering) on new items a user might prefer.	
<b>Typical Applications</b>	Movies recommendations based on content (e.g. Netflix)
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>A content-based filtering system could be employed to offer passenger recommendations on various products and services related to aviation that also have textual content or description. These can be duty-free products, hotels, in-flight movies, or meal suggestions. As with any recommendation system, the personalised services provided can eventually increase passenger satisfaction.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Hybrid approaches, that combine both collaborative filtering and content-based filtering.</li> <li>Context-based filtering, that considers additional context information in order to provide a recommendation.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python, Spark (MLlib), R (rrecsys).
<b>Considerations</b>	The method is sensitive to the keywords selected to represent the content.
<b>References</b>	<p>Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. <i>Artificial intelligence review</i>, 13(5-6), 393-408.</p> <p>Basilico, J., &amp; Hofmann, T. (2004, July). Unifying collaborative and content-based filtering. In <i>Proceedings of the twenty-first international conference on Machine learning</i> (p. 9). ACM.</p> <p>Liu, H., Hu, J., &amp; Rauterberg, M. (2015). Follow your heart: Heart rate controlled music recommendation for low stress air travel. <i>Interaction Studies</i>, 16(2), 303-339.</p>

### 3.4.9 Support Vector Machines

<b>Algorithm Family</b>	Classification, Regression, Outlier detection
<b>Description</b>	
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. SVMs are effective in high dimensional spaces even if the number of dimensions is greater than the number of samples. They use a subset of training points in the decision function (called support vectors), so they are also memory efficient. Different Kernel functions can be specified for the decision function. The disadvantages of SVMs include overfitting if the number of features is much greater than the number of samples and they don't provide probability estimates.	
<b>Typical Applications</b>	<ul style="list-style-type: none"> <li>SVMs have been used successfully in many real-world problems: text categorisation, image classification, bioinformatics (Protein classification, Cancer classification) and, hand-written character recognition.</li> </ul>

<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Flight delays are frequent all over the world and they are estimated to have an annual cost of several tens of billion dollars. The prediction of flight delays is a primary issue for airlines and travellers. SVM can be used for this task.</li> <li>To improve airport ground transport operations SVM can forecast queue contexts (i.e. taxis are waiting for passengers, passengers are waiting for taxis, both are waiting for each other, none is waiting). In an airport, this is a challenging problem due to the presence of different contextual factors i.e., time, weather, taxi trips, flight arrivals and many more.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Least squares support vector machines (LS-SVM).</li> <li>One-class SVM for outlier detection.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (e1071, caret).
<b>Considerations</b>	<p>SVMs require a lot of parameter fine tuning for each task.</p> <p>They do not provide the probability of each class membership.</p> <p>They can be susceptible to over-fitting.</p>
<b>References</b>	<p>Cortes, C., &amp; Vapnik, V. (1995). Support-Vector Networks. <i>Machine Learning</i>.</p> <p>Belcastro, L., Marozzo, F., Talia, D., &amp; Trunfio, P. (2016). Using Scalable Data Mining for Predicting Flight Delays. <i>ACM Transactions on Intelligent Systems and Technology</i>.</p> <p>Rahaman, M. S., Hamilton, M., &amp; Salim, F. D. (2017). Predicting Imbalanced Taxi and Passenger Queue Contexts in Airport Predicting Imbalanced Taxi and Passenger. <i>Pacific Asia Conference on Information Systems</i>.</p>

### 3.4.10 Classification and Regression Tree

<b>Algorithm Family</b>	Classification, Regression
<b>Description</b>	
<p>The Classification and Regression Tree (CART) analysis is a term used to refer to models that can predict a categorical target variable (Classification tree) or a numeric one (Regression Tree).</p> <p>CART algorithms create models that predict the value of a target variable by learning simple decision rules inferred from the data features. Trees used for regression and trees used for classification have some similarities, but also some differences, such as the procedure used to determine where to split the input space, since they utilize a “divide and conquer” technique.</p>	
<b>Typical Applications</b>	CART applications include classification and regression problems when the focus is on explanation of the prediction rather than performance.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Regression tree has been used to develop a system to provide real-time information about transfer passengers’ journeys through the airport. This information is important for the airport to best assist the passengers, the airlines, and their employees. Although simple in nature, Regression Tree provides accurate predictions and easy interpretations.</li> <li>Airlines habitually overbook flights based on the expectation that some fraction of booked passengers will not show for each flight. Accurate forecasts of the expected number of no-shows for each flight can increase airline revenue by reducing the number of spoiled seats and the number of involuntary denied boardings at the departure gate. Decision Tree like C4.5</li> </ul>

	can predict cabin-level no-show rates using specific data on the individual travellers booked on each flight.
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>ID3, C4.5, CHAID, MARS</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (rpart, caret, C50).
<b>Considerations</b>	Decision trees are very sensitive to the input space. Even a slight change may cause large changes in a tree's outcome. So, being also a nonparametric technique, this makes them not so efficient to make any generalisation on the underlying phenomenon based on the results observed.
<b>References</b>	<p>Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth &amp; Brooks/Cole Advanced Books &amp; Software.</p> <p>Xiaoja, G., Grushka-Cockayne, Y. and De Reyck, B. (2018) Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning. <i>Harvard Business School Working Paper, No. 19-040</i>. (Under Review.)</p> <p>Lawrence, R. D., Hong, S. J., &amp; Cherrier, J. (2003). Passenger-based predictive modeling of airline no-show rates. In <i>Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03</i> (p. 397). New York, New York, USA: ACM Press.</p>

### 3.4.11 Random Forest

<b>Algorithm Family</b>	Classification, Regression, Outlier detection
<b>Description</b>	Random forests are an ensemble learning method for classification and regression. They operate by creating a multitude of decision trees during the training phase. The class of a new observation is computed as the mode of the classes (classification) or as the mean prediction (regression) of the individual trees generated by the model. They were introduced as an enhancement of the generalisation capability of single decision trees.
<b>Typical Applications</b>	Random forest applications include classification and regression problems when the focus is on performance of the model rather than on explanatory description.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Predict flight delays. Flight delays are frequent all over the world and they have an estimated annual cost of several tens of billion dollars. This scenario makes the prediction of flight delays a primary issue for airlines and travellers.</li> <li>To improve airport ground transport operations Random Forest can predict queue contexts indicating the states of queues (i.e. taxis are waiting for passengers, passengers are waiting for taxis, both are waiting for each other, none is waiting). In an airport this is a challenging problem due to the presence of different contextual factors i.e., time, weather, taxi trips, flight arrivals and many more.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>Random Survival Forests, Multivariate Random Forests, Enriched Random Forests, Quantile Regression Forests are variants of the original method.</li> <li>Isolation Forest is one efficient way of performing outlier detection in high-dimensional datasets using Random Forests.</li> </ul>

<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (randomforest, caret).
<b>References</b>	<p>Ho, T. K. (1998). The random subspace method for constructing decision forests. <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i>.</p> <p>Liu, F. T., Ting, K. M., &amp; Zhou, Z. H. (2008). Isolation forest. In <i>Proceedings - IEEE International Conference on Data Mining, ICDM</i> (pp. 413–422).</p> <p>Rahaman, M. S., Hamilton, M., &amp; Salim, F. D. (2017). Predicting Imbalanced Taxi and Passenger Queue Contexts in <i>Pacific Asia Conference on Information Systems</i>.</p>

### 3.4.12 Gradient Boosting Machine

<b>Algorithm Family</b>	Classification, Regression
<b>Description</b>	<p>Gradient Boosting Machines (GBMs) are an ensemble learning method for regression and classification problems. It builds a prediction model in the form of an ensemble of weak prediction models, usually decision trees. Like in other boosting methods the model is built in a stage-wise way, permitting optimisation of an arbitrary differentiable loss function.</p>
<b>Typical Applications</b>	<p>One of the important properties of GBMs is the possibility of building sparse models. This property is desirable in several practical cases, for example, when the predictor data comes from a very high dimensional distribution whilst containing very little, sparsely distributed information.</p>
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Low-visibility conditions at airports can lead to capacity reductions and therefore to delays or cancelations of arriving and departing flights. Accurate visibility forecasts are required to keep the airport capacity as high as possible. The nowcasts are generated with tree-based statistical models “boosting” based on highly-resolved meteorological observations at the airport. Short computation times ensure the instantaneous generation of predictions.</li> <li>Predicting aircraft trajectories is a key element in the detection and resolution of air traffic conflicts. Predictions with Gradient Boosting Machine can improve the root mean square error on the predicted descent length up to 24 %, when compared with the baseline BADA method (a baseline method relying on the Eurocontrol Base of Aircraft Data).</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>XGBoost, LightGBM, Catboost</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (xgboost, caret).
<b>References</b>	<p>Breiman, L. (1997). Arcing the edge. <i>Statistics</i>.</p> <p>Dietz, S. J., Kneringer, P., Mayr, G. J., Zeileis, A. (2017). Forecasting Low-Visibility Procedure States with Tree-Based Statistical Methods, <i>Working Papers 2017-22, Faculty of Economics and Statistics, University of Innsbruck</i>.</p> <p>Alligier, R., Gianazza, D., Durand, N (2016). Predicting Aircraft Descent Length with Machine Learning. <i>ICRAT 2016, 7th International Conference on Research in Air Transportation</i>, Philadelphia, United States.</p>

## 3.4.13 K-NN

<b>Algorithm Family</b>	Classification, Regression
<b>Description</b>	
<p>K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its K neighbours. The case being assigned to the class is most common amongst its K nearest neighbours measured by a distance function.</p> <p>These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbour. At times, choosing K turns out to be a challenge while performing K-NN modelling.</p>	
<b>Typical Applications</b>	It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>K-NN algorithm can be used to predict airline delays.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), R (class).
<b>Considerations</b>	<p>As a “lazy learner” algorithm, i.e. it does not actually learn from data, K-NN is more appropriate for datasets of low volume and complexity.</p> <p>K-NN is also strongly affected by the curse of dimensionality, i.e. performance decreases with the increase of input attributes.</p>
<b>References</b>	<p>Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". <i>The American Statistician</i>. 46 (3): 175–185.</p> <p>Choi, S., Kim, Y. J., Briceno, S., &amp; Mavris, D. (2017, September). Cost-sensitive prediction of airline delays using machine learning. In <i>Digital Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th</i> (pp. 1-8). IEEE.</p>

## 3.4.14 Naïve Bayes

<b>Algorithm Family</b>	Classification
<b>Description</b>	
<p>Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.</p>	
<b>Typical Applications</b>	<p>Naive Bayes was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorisation, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.</p>
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Time Performance (OTP) is an important aspect for flight service user and provider. OTP is one of factors that affect positive or negative assessment of flight service. To overcome the problem, it requires a departure on time prediction system. Naïve Bayes Classification can be applied to create on</li> </ul>



	time prediction system that can be used by the airlines to prepare more for the possibilities that can be happened in the future.
<b>Popular Libraries / Software tools</b>	Custom implementations in H2O, Python (scikit-learn), Spark (MLlib), R (naivebayes, caret).
<b>Considerations</b>	Naïve Bayes classifier main disadvantage is that it makes strong assumptions on the shape of data distribution, based on its feature independence principle. This makes Naïve Bayes not very suitable when data have continuous features.
<b>References</b>	Russell, Stuart; Norvig, Peter (2003) [1995]. <i>Artificial Intelligence: A Modern Approach (2nd ed.)</i> . Prentice Hall.  Andi Nugroho, A., Fahmi, R. (2017) On-Time Flight Departure Prediction System Using Naive Bayes Classification Method (Case Study: XYZ Airline). <i>International Journal of Computer Trends and Technology (IJCTT) V53(1):4-10</i> .

### 3.4.15 Multilayer Perceptron (MLP)

<b>Algorithm Family</b>	Classification, Regression, Time Series Prediction
<b>Description</b>	<p>The multi-layer Perceptron (MLP) is the most common representative of the Artificial Neural Networks (ANNs) and belongs to the category of feedforward neural networks. It consists of more than one layer of nodes, (in contrast to the single layer perceptron) and employs a supervised learning technique, known as backpropagation, for training. The ability of MLPs to solve non-linear problems makes them eligible as <i>universal approximators</i>.</p> <p>MLPs were very popular during the '80s especially in the fields of speech processing and image recognition, but this interest degraded gradually due to the appearance of faster and simpler algorithms (e.g. SVMs). Today, the enthusiasm towards multi-layer networks has returned as an aftereffect of the advancements in <i>deep learning</i> technology.</p>
<b>Typical Applications</b>	Numerous and various applications, such as image recognition, speech processing, currency exchange rates and stock prices prediction, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>An MLP can be utilized in several applications of the aviation industry, especially in quality control and operations management: Marketing and control of airline seat allocation, scheduling and planning of air flights, aircraft trajectory control and collision avoidance, personnel management, etc.</li> </ul>
<b>Popular Libraries / Software tools</b>	Python (theano, scikit-learn, keras, caffe), Spark (MLlib), H2O (deep-learning), R (monmlp), tensorflow.
<b>Considerations</b>	MLPs cannot guarantee the finding of a global minimum (i.e. best model of a function). Their training approach may often get stuck in a local minimum.
<b>References</b>	<p>Pal, S. K., &amp; Mitra, S. (1992). Multilayer Perceptron, Fuzzy Sets, Classification. Widrow, B., Rumelhart, D. E., &amp; Lehr, M. A. (1994). Neural networks: applications in industry, business and science. <i>Communications of the ACM</i>, 37(3), 93-106.</p> <p>Kaidi, R., Lazaar, M., &amp; Ettaouil, M. (2014). Neural network apply to predict aircraft trajectory for conflict resolution. <i>In 2014 9th International Conference on Intelligent Systems: Theories and Applications, SITA 2014. IEEE Computer Society</i>.</p>

### 3.4.16 Adaptive Neuro-Fuzzy Inference System (ANFIS)

<b>Algorithm Family</b>	Classification, Regression, Time Series Prediction
<b>Description</b>	
An adaptive neuro-fuzzy inference system (ANFIS) is a special kind of artificial neural network which integrates a fuzzy inference engine. This technique facilitates the main principles of both neural networks and fuzzy logic, as well as their advantages. The outcome of the learning procedure is a system capable of approximating nonlinear functions using a data-driven set of fuzzy IF–THEN rules.	
<b>Typical Applications</b>	The unparalleled generalisation ability of ANFIS has made it a very popular algorithm in research studies, especially for diagnosis and fault detection, optimisation and control, time series prediction, image processing and more.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>ANFIS can be employed to tackle advanced tasks, such as machine fault diagnosis, gas emissions predictions, collision detection, or passenger demand.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Spark, Python (anfis), R (anfis), tensorflow (tensorANFIS).
<b>Considerations</b>	Despite ANFIS's wide acceptance as a universal approximator, it suffers from limitations related to the curse of dimensionality and computational expense.
<b>References</b>	<p>Jang, J. S. (1993). ANFIS: adaptive-network-based fuzzy inference system. <i>IEEE transactions on systems, man, and cybernetics</i>, 23(3), 665-685.</p> <p>Srisaeng, P., Baxter, G., &amp; Wild, G. (2015). An adaptive neuro-fuzzy inference system for modelling Australia's regional airline passenger demand. <i>International Journal of Sustainable Aviation</i>, 1(4), 348-374.</p>

### 3.4.17 Genetic Algorithms (GA)

<b>Algorithm Family</b>	Optimisation
<b>Description</b>	
<p>Genetic algorithms are a set of metaheuristic techniques that belong to the field of evolutionary computation. As such, they do not belong to the machine learning family, but they are quite close relatives and, in many methodologies, they can be seen working together.</p> <p>Genetic algorithms attempt to simulate Darwin's theory of natural evolution in such a way that the most capable (fittest) individuals from a population of solutions are selected for reproduction and the creation of the next, more fitting according to a fitness function, generation. This iterating process eventually leads to a globally optimal solution.</p>	
<b>Typical Applications</b>	Common optimisation problems such as vehicle routing, time scheduling, searching strategies, filtering and signal processing, encrypting and decrypting.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>The advantages of GA can be exploited in several aviation-related applications, such as the reduction of airspace congestion or the optimisation of airport traffic planning, airline slot allocation and gate scheduling.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li><i>Adaptive genetic algorithms</i>, that change their parameters dynamically</li> <li><i>Hybrid genetic algorithms</i>, where GA are combined with a more conventional optimisation method, such as gradient descent.</li> </ul>
<b>Popular Libraries / Software tools</b>	Custom implementations in Python (DEAP), Spark (pyspark/DEAP), R (GA).

<b>Considerations</b>	Genetic algorithms do not scale well with complexity and dynamic data sets.
<b>References</b>	<p>Mitchell, M. (1998). <i>An introduction to genetic algorithms</i>. MIT press.</p> <p>Daniel, D., Oussedik, S., &amp; Stephane, P. (2005, March). Airspace congestion smoothing by multi-objective genetic algorithm. In <i>Proceedings of the 2005 ACM symposium on Applied computing</i> (pp. 907-912). ACM.</p> <p>Ferreira, D. M., Rosa, L. P., Ribeiro, V. F., de Barros Vidal, F., &amp; Weigang, L. (2014, September). Genetic algorithms and game theory for airport departure decision making: GeDMAN and CoDMAN. In <i>International Conference on Knowledge Management in Organisations</i> (pp. 3-14). Springer, Cham.</p>

### 3.5 Axis III: Deep Learning

In recent years, there is a lot of research and investment in deep learning and thus there has been a gigantic increase in the number of deep learning architectures. Deep learning architectures refer to artificial neural networks that have a large number of layers, enabling them to automatically learn complex features at multiple levels of abstraction. In contrast to task-specific algorithms, deep learning architectures are based on learning data representations. They can be used for supervised, semi-supervised or unsupervised learning and the output value depends on the activation function. Theoretically, these models can outperform typical machine learning models as they can be fitted with tremendous amounts of data and can automatically learn and use more complex features. However, they have high time and space complexity, demanding exponential computational time due to their deep architecture.

This section aims to identify and provide the state-of-the-art deep learning architectures in order to utilise them for the discovery and communication of meaningful knowledge and new patterns that were unattainable or hidden in the previous isolated data structures. Furthermore, the proposed deep learning architectures will allow the extraction of patterns and the execution of “what-if” simulation scenarios based on the existing data that is available. The proposed deep learning architectures were selected based on the following criteria:

- to be relevant for the aviation industry and the demonstrators’ needs
- to have proven their ability and robustness
- to have been implemented in a commonly used software framework or library

A summary of these architectures is presented in the following table, along with the respective section where their detailed description is provided:

**Table 3-3: Deep Learning Algorithms**

Section	Algorithm Family	Algorithm Name
3.5.1	Classification, Regression, Deep Learning	Deep Feedforward Networks (DFFN)
3.5.2	Classification, Regression, Deep Learning	Convolutional Neural Networks (CNN)
3.5.3	Classification, Regression, Time Series Prediction, Deep Learning	Recurrent Neural Networks (RNN)
3.3.53.5.4	Dimensionality Reduction, Clustering, Data visualisation, Feature Learning, Deep Learning	Deep Autoencoders
3.5.5	Reinforcement Learning, Deep Learning	Deep Q-Networks (DQN)

### 3.5.1 Deep Feedforward Networks (DFFN)

<b>Algorithm Family</b>	Classification, Regression, Deep Learning
<b>Description</b>	
<p>Deep feedforward networks, also called feed forward neural networks, or multilayer perceptrons (MLPs), are deep neural networks where the connections between the nodes do not form a cycle. In particular, this class of networks consists of multiple layers of computational units and the goal is to approximate a non-linear function (e.g. sigmoid function). Furthermore, the information in these models moves only forward, from the input nodes to the output nodes without any feedback connections.</p> <p>Deep feedforward neural networks are suitable for tabular datasets and they are primarily used for classification or regression prediction problems. Furthermore, they support for multivariate input, multivariate output and learning complex functional relationships. Overall, they are very flexible and can be used generally to learn a mapping from inputs to outputs. This flexibility allows them to be applied to other types of data like image data, text data, etc.</p>	
<b>Typical Applications</b>	Speech recognition, image recognition, machine translation, timeseries forecasting etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• Provided a dataset with the monthly demand of the stock of aircraft spare parts, it can be predicted the demand of the spare parts in order to minimize inventory cost and reduce the risk of stock-out.</li> <li>• Provided a dataset with meteorological data like temperature, dew point temperature, relative humidity, pressure, wind speed and direction, it can be predicted the horizontal visibility during fog over the airport.</li> </ul>
<b>Popular Libraries / Software tools</b>	Tensorflow, Keras, BigDL, PyTorch, Caffe, Caffe2, Mxnet, Microsoft CNTK, H2o, Deeplearning4j
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Data should be represented as vectors</li> <li>• Require much more data than traditional Machine Learning algorithms</li> <li>• Computationally expensive</li> <li>• Setting the number of hidden neurons too low may result in underfitting while setting the value too high may result in overfitting</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>• Rosenblatt, Frank. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. No. VG-1196-G-8. CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961</li> <li>• Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.</li> </ul>

### 3.5.2 Convolutional Neural Networks (CNN)

<b>Algorithm Family</b>	Classification, Regression, Deep Learning
<b>Description</b>	
<p>Convolutional neural networks (CNNs) are deep feed-forward artificial neural networks, composed of one or more convolutional layers with fully connected layers on top. In particular, they consist of an input and an output layer as well as multiple hidden layers. The hidden layers of a convolution neural network consist of convolutional layers, pooling layers, fully connected layers and normalisation layers.</p>	

CNNs are suitable for processing data that have grid-like topology (e.g. 2-D grid of pixels from image data) and spatial relationship. Furthermore, CNNs are easier to train than other deep feed-forward neural networks as they have fewer parameters to estimate and they can learn the useful features without depending on prior knowledge or human effort in feature design. Additionally, they support for multivariate input, multivariate output and they are primarily used for classification or regression prediction problems. Finally, CNNs can be used on data that has a spatial relationship and they achieve state-of-the-art results on problems such as image and text classification.

<b>Typical Applications</b>	Image and video processing, natural language processing, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Provided a dataset with aircraft turbine gearbox fault data and frequency data of vibration signals, the gearbox health conditions can be predicted.</li> </ul>
<b>Popular Libraries / Software tools</b>	Tensorflow, Keras, BigDL, PyTorch, Caffe, Caffe2, Mxnet, Microsoft CNTK, Deeplearning4j
<b>Considerations</b>	<ul style="list-style-type: none"> <li>Data should be represented as vectors</li> <li>Require much more data than traditional Machine Learning algorithms</li> <li>Computationally expensive</li> <li>Setting the number of hidden neurons too low may result in underfitting while setting the value too high may result in overfitting</li> <li>Do not work very well with temporal sequences</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>Lippmann, Richard P. "Review of neural networks for speech recognition." Neural computation 1, no. 1 (1989): 1-38.</li> <li>Waibel, Alexander, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. "Phoneme recognition using time-delay neural networks." In Readings in speech recognition, pp. 393-404. 1990.</li> <li>Denker, John S., W. R. Gardner, Hans Peter Graf, Donnie Henderson, Richard E. Howard, W. Hubbard, Lawrence D. Jackel, Henry S. Baird, and Isabelle Guyon. "Neural network recognizer for hand-written zip code digits." In Advances in neural information processing systems, pp. 323-331. 1989.</li> <li>LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86, no. 11 (1998): 2278-2324.</li> </ul>

### 3.5.3 Recurrent Neural Networks (RNN)

<b>Algorithm Family</b>	Classification, Regression, Time Series Prediction, Deep Learning
<b>Description</b>	<p>Deep recurrent neural networks (RNNs) are a family of neural networks where connections between nodes form a directed graph. They consist of a sequence of multiple nonlinear layers. Deep RNNs can use their internal states (memory) to process multiple time-scales representations of inputs.</p> <p>Deep RNNs support for multivariate input, multivariate output and they are primarily used for classification or regression prediction problems. Finally, they are suitable for temporal data processing and they achieve state-of-the-art results on problems like natural language processing.</p>
<b>Typical Applications</b>	Language translation, natural language processing, timeseries forecasting, anomaly detection, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Provided a dataset with multivariate, time-series aircraft's flight data, several anomalies and trends that reduce safety margins can be predicted.</li> <li>Provided a dataset which contains nodes (cities), hubs (airports) and the traffic (airline passengers) that receives and sends each airport, the hub</li> </ul>

	<p>(airport) locations can be chosen from among these nodes to act as switching points in case of uncapacitated hubs.</p> <ul style="list-style-type: none"> <li>• Provided a dataset with multivariate, time-series aircraft's flight data, and health patient data (sick with an infectious disease or not) to predict the spreading of the particular disease.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• <b>Long short-term memory (LSTM) network:</b> A very popular variation of the RNNs that was developed to deal with the exploding and vanishing gradient problems of the traditional RNNs.</li> <li>• <b>Gated recurrent units (GRU) network:</b> A variation of LSTM that have fewer output gates than LSTM. Additionally, they have fewer parameters than LSTM and thus they are faster to train.</li> <li>• <b>Bidirectional RNN (BRNN) / Bidirectional LSTM (BLSTM):</b> The network splits the neurons of a regular RNN (or LSTM) into two directions, one for forward states (forward in time) and another for backward states (backwards in time) in order the output layer to get information from past and future states.</li> <li>• <b>Deep Stacked RNN/LSTM:</b> The network consists of a stacked composition of multiple non-linear recurrent (or LSTM) hidden layers. The state computation proceeds by following the hierarchical network organisation, from the lowest layer to the highest one. Specifically, at each time step <math>t</math> the first recurrent layer in the network is fed by the external input while each successive layer is fed by the activation of the previous one.</li> </ul>
<b>Popular Libraries / Software tools</b>	<ul style="list-style-type: none"> <li>• <b>Simple RNN, LSTM and GRU:</b> Tensorflow, Keras, BigDL, PyTorch, Caffe, Caffe2, Mxnet, Microsoft CNTK, Deeplearning4j</li> <li>• <b>Bidirectional RNN/LSTM:</b> Tensorflow, Keras, BigDL, PyTorch, Mxnet, Microsoft CNTK, Deeplearning4j</li> <li>• <b>Deep Stacked RNN/LSTM:</b> Custom implementations in Tensorflow, Keras, BigDL, PyTorch, Mxnet, Microsoft CNTK, Deeplearning4j</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Data should be represented as vectors</li> <li>• Require much more data than traditional Machine Learning algorithms</li> <li>• Computationally expensive</li> <li>• Setting the number of hidden neurons too low may result in underfitting while setting the value too high may result in overfitting</li> <li>• Do not work very well with image and tabular data</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>• Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." <i>nature</i> 323, no. 6088 (1986): 533.</li> <li>• Elman, Jeffrey L. "Finding structure in time." <i>Cognitive science</i> 14, no. 2 (1990): 179-211.</li> <li>• Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." <i>Neural computation</i> 9, no. 8 (1997): 1735-1780.</li> <li>• Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." <i>arXiv preprint arXiv:1406.1078</i> (2014).</li> <li>• Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." <i>arXiv preprint arXiv:1412.3555</i> (2014).</li> <li>• Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." <i>IEEE Transactions on Signal Processing</i> 45, no. 11 (1997): 2673-2681.</li> </ul>

	<ul style="list-style-type: none"> <li>Pascanu, R., Gulcehre, C., Cho, K., &amp; Bengio, Y. (2014). How to construct deep recurrent neural networks. In Proceedings of the Second International Conference on Learning Representations (ICLR 2014).</li> </ul>
--	--

#### 3.5.4 Deep Autoencoders

<b>Algorithm Family</b>	Dimensionality Reduction, Clustering, Data visualisation, Feature Learning, Deep Learning
<b>Description</b>	
Deep Autoencoders are artificial neural networks which are trained to learn a representation of the original data in an unsupervised manner and they are suitable for dimensionality reduction and clustering tasks. They have two major parts, the encoder and the decoder. The encoder compresses the input data to lower dimensional features, and then the decoder uncompresses the features into something that closely matches the original data.	
<b>Typical Applications</b>	High-dimensional data exploration, text clustering, image clustering, clustering sensor data, machine translation, anomaly detection, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>Provided a dataset with the trajectory data of an aircraft, the irregular air traffic patterns (anomalous flights) can be detected.</li> <li>Provided a dataset with flight data of aircrafts, the potential safety anomalies of aircrafts can be detected as they occur.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li><b>Deep Embedded Clustering (DEC):</b> A variation which is suitable for clustering and dimensionality reduction tasks. DEC learns a mapping from the data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective. In particular, it uses a deep stacked autoencoder (SAE) to initialize the parameters and then it iterates between computing an auxiliary target distribution and minimizing the KL divergence loss. In general, it offers significant improvements over several other clustering methods in terms of accuracy, running time and hyperparameters robustness.</li> <li><b>Improved Deep Embedded Clustering (IDEC):</b> A variation of DEC that manipulates feature space in order to scatter data points using a clustering loss as guidance. Additionally, an under-complete autoencoder is included to the network structure.</li> <li><b>Variational Deep Embedding (VaDE):</b> A unsupervised generative clustering approach which use a variational autoencoder (VAE). In particular, VaDE models the data generative procedure with a Gaussian Mixture Model (GMM) and a deep neural network (DNN). VaDE significantly outperforms the state-of-the-art clustering methods on several benchmarks and datasets.</li> </ul>
<b>Popular Libraries / Software tools</b>	<ul style="list-style-type: none"> <li><b>Deep Autoencoders:</b> Custom implementations in Tensorflow, Keras, BigDL, PyTorch, Caffe, Mxnet, Microsoft CNTK, Deeplearning4j</li> <li><b>DEC and IDEC:</b> Custom implementations in Tensorflow, Keras, Mxnet and Caffe</li> <li><b>VaDE:</b> Custom implementations in Keras</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>Data should be represented as vectors</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." nature 323, no. 6088 (1986): 533.</li> <li>Hong, Chaoqun, et al. "Multimodal deep autoencoder for human pose recovery." IEEE Transactions on Image Processing 24.12 (2015): 5659-5670.</li> </ul>



	<ul style="list-style-type: none"> <li>• Chicco, Davide, Peter Sadowski, and Pierre Baldi. "Deep autoencoder neural networks for gene ontology annotation predictions." Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 2014.</li> <li>• Xie, Junyuan, Ross Girshick, and Ali Farhadi. "Unsupervised deep embedding for clustering analysis." In International conference on machine learning, pp. 478-487. 2016.</li> <li>• Guo, Xifeng, Long Gao, Xinwang Liu, and Jianping Yin. "Improved deep embedded clustering with local structure preservation." In International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 1753-1759. 2017.</li> <li>• Jiang, Zhuxi, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. "Variational deep embedding: An unsupervised and generative approach to clustering." arXiv preprint arXiv:1611.05148 (2016).</li> </ul>
--	---

### 3.5.5 Deep Q-Networks (DQN)

<b>Algorithm Family</b>	Reinforcement Learning, Deep Learning
<b>Description</b>	<p>Deep Q-Network (DQN) is a deep reinforcement learning neural network which learns control policies directly from high-dimensional input using reinforcement learning. DQN utilizes a variant of Q-learning whose output is a value function estimating future rewards and a neural network architecture (e.g. convolutional networks, LSTMs, etc.). In particular, the Q-Learning variant combines stochastic minibatch up-dates with experience replay memory to ease the training of deep networks.</p>
<b>Typical Applications</b>	Robotics, gaming, task scheduling, resource management, stock trading, etc.
<b>Application examples in the aviation domain</b>	<ul style="list-style-type: none"> <li>• Provided a dataset with multivariate, time-series aircraft's flight data, several anomalies and trends that reduce safety margins can be predicted.</li> <li>• Provided a dataset with aircraft's flight geolocation, red signal for emergency landing, nodes (cities) and hubs (airports), the most appropriate hub for emergency landing can be predicted.</li> <li>• Provided a dataset with multivariate, time-series aircraft's flight data, the taxi-out time for each flight can be predicted.</li> </ul>
<b>Established Variations</b>	<ul style="list-style-type: none"> <li>• <b>Double DQN</b> decouples the selection from the evaluation phases and reduces the overestimations of the action values of the DQN.</li> <li>• <b>Dueling DQN</b> separates the representation of state value function and the state-dependent action advantage function. It consists of two streams that represent the value and advantage functions, while sharing a special aggregating layer to produce an estimate of the state-action value function Q.</li> <li>• <b>Continuous DQN</b> uses a variant of Q-learning that can be used in continuous control domains. Furthermore, it combines the variation of Q-learning algorithm with learned models so as to accelerate learning while preserving the benefits of model-free RL.</li> <li>• <b>Deep Deterministic Policy Gradient (DDPG)</b> relies on the actor-critic architecture with two elements, actor and critic. An actor is used to decide the best action for a specific state whereas critic is used for evaluating the policy function estimated by the actor according to the temporal difference (TD) error.</li> <li>• <b>Deep State-Action-Reward-State-Action (SARSA)</b> is very similar to DQN. The main difference is that SARSA learns the Q-value based on the action performed by the current policy instead of the greedy policy.</li> </ul>

<b>Popular Libraries / Software tools</b>	<ul style="list-style-type: none"> <li>• <b>DQN, Double DQN, Dueling DQN and DDPG:</b> Custom implementations in Tensorflow and Keras</li> <li>• <b>Continuous DQN and Deep SARSA:</b> Custom implementations in Keras</li> </ul>
<b>Considerations</b>	<ul style="list-style-type: none"> <li>• Data should be represented as vectors</li> <li>• Requires enormous amount of training data</li> <li>• Balancing exploration (try out non-optimal actions to explore the environment) versus exploitation (exploit the optimal action in order to make useful progress)</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>• Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., &amp; Riedmiller, M. (2013). Playing atari with deep reinforcement learning.</li> <li>• Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... &amp; Petersen, S. (2015). Human-level control through deep reinforcement learning. <i>Nature</i>, 518(7540), 529-533.</li> <li>• Lillicrap, Timothy P., Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. "Continuous control with deep reinforcement learning." <i>arXiv preprint arXiv:1509.02971</i> (2015).</li> <li>• Zhao, Dongbin, Haitao Wang, Kun Shao, and Yuanheng Zhu. "Deep reinforcement learning with experience replay based on SARSA." In <i>Computational Intelligence (SSCI), 2016 IEEE Symposium Series on</i>, pp. 1-6. IEEE, 2016.</li> <li>• Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep Reinforcement Learning with Double Q-Learning." In <i>AAAI</i>, vol. 2, p. 5. 2016.</li> <li>• Wang, Ziyu, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. "Dueling network architectures for deep reinforcement learning." <i>arXiv preprint arXiv:1511.06581</i> (2015).</li> <li>• Gu, Shixiang, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. "Continuous deep q-learning with model-based acceleration." In <i>International Conference on Machine Learning</i>, pp. 2829-2838. 2016.</li> </ul>

### 3.6 Data Analytics Perspectives for the ICARUS Demonstrators

During the ICARUS project, four (4) real-life demonstrator cases will be realised (see related data sets in D1.1). ICARUS will provide the appropriate conditions and tools for basic and advanced data analytics in the context of separate scenarios derived from such demonstrators, with the aim to reveal the actual potential of the ICARUS platform.

The following sections present a preliminary approach of a mapping between the very early demonstrators scenarios and the potentially applicable algorithms, given the brief, early descriptions of scenarios as presented by the pilot partners in the ICARUS plenary meetings. For each scenario, a number of indicative use case examples is provided, in an attempt to correlate the scenario objectives with computational tasks commonly encountered in literature of the aviation domain analytics. A group of algorithms is then proposed in accordance with each task, in order to address it in the most efficient way. It should be noted, however, that in the most complex cases, a combination of algorithms and techniques will be required for better and more accurate results.

### 3.6.1 Demonstrator 1: AIA

The Athens International Airport (AIA) is responsible for the first demonstrator, which involves an early usage scenario. Such a scenario is inspired by the airport’s transitional stage, upgrading from Level 1 to Level 2 in terms of capacity. It is, thus, devoted to airport capacity optimisation, in respect to airside capacity, airport infrastructure and runway operations. This is a scenario of high complexity and depth, possibly involving several subtasks. Some indicative subtasks are presented in the following table in the form of use case examples:

**Table 3-4: Scenario use cases for the AIA demonstrator**

Scenario 1: Optimize Airport Capacity (Airside Capacity, Airport Infrastructure, Runway Operations Capacity)		
Use case examples	Algorithmic Task(s)	Algorithm(s)
Delay modelling	Clustering	SOM, k-means, DBSCAN
Flight delay prediction	Regression Analysis (predict amount of delay)	Linear regression, SVM, Random Forest, MLP, ANFIS, DFFN
	Classification (predict whether flight will be delayed or not)	SVM, Random Forest, k-NN, NB, MLP, ANFIS, DFFN
Taxi time prediction	Regression Analysis (predict taxi time amount)	Linear regression, Random Forest, MLP, ANFIS, DFFN
Air travel demand	Classification	CART, SVM, Random Forest, MLP, ANFIS, DFFN
	Regression Analysis	CART, SVM, Random Forest, MLP, ANFIS, DFFN
	Time Series Prediction	ARIMA, ANFIS, RNN
Airline Slot Allocation	Optimisation	Genetic Algorithms
Flight Gate scheduling	Optimisation	Genetic Algorithms
Air traffic forecasting	Regression Analysis	CART, SVM, Random Forest, MLP, ANFIS, DFFN

### 3.6.2 Demonstrator 2: PACE

This demonstrator is run by PACE and includes two preliminary scenarios. The first scenario focuses on the analysis of pollution data and aircraft emissions. Typical use case examples in this field involve the statistical evaluation of weather data and pollution data and the prediction of aircraft performance in relation to the environmental impact.

**Table 3-5: Scenario 1 use cases for the PACE demonstrator**

Scenario 1: Pollution Data Analysis		
Use case examples	Algorithmic Task(s)	Algorithm(s)
Weather data evaluation	Statistical Analysis	Simple statistics, Hypothesis testing, sampling
Pollution data evaluation	Regression Analysis (feature correlation, predictions)	Simple statistics, Hypothesis testing, sampling
Flight performance prediction	Regression Analysis (feature correlation, predictions)	SVM, Random Forest, ANFIS, DFFN

The second scenario aims to analyse pollution data on a larger scale, that of a massive route network. Typical use case examples in this field involve the statistical evaluation of weather data and pollution data and the prediction of aircraft performance in relation to the environmental impact, as depicted in table 3-6.

**Table 3-6: Scenario 2 use cases for the PACE demonstrator**

Scenario 2: Massive Route Network Analysis and Evaluation		
Use case examples	Algorithmic Task(s)	Algorithm(s)
Weather data and Pollution Data evaluation	Statistical Analysis	Simple statistics, Hypothesis testing, sampling
Modelling route network	Clustering	SOM, k-means, DBSCAN
Route performance prediction	Regression Analysis (feature correlation, predictions)	SVM, Random Forest, MLP, ANFIS, DFFN

### 3.6.3 Demonstrator 3: ISI

The ISI demonstrator is set in the context of computational epidemiology and is focused on improving the state-of-the-art modelling of global-scale infectious disease spreading. In studying such phenomena, human mobility data, and in particular airline transportation data, are of fundamental importance in order to develop realistic models, capable of making predictions, and evaluate the efficacy of response strategies. The ISI simulation framework makes use of custom algorithms and methods<sup>3</sup> that are of minor interest for the aviation industry, and there are no current plans on using advanced data analytics on the ICARUS platform for this demonstrator. However, the diverse data sets employed by the respective scenarios could potentially take advantage of some basic analytics tools

<sup>3</sup> <https://www.gleamproject.org/publications>

in order to examine, align and pre-process them. Moreover, some output results of this demonstrator could be made available through the ICARUS platform for the benefit of other stakeholders.

### 3.6.4 Demonstrator 4: CELLOCK

The last demonstrator is devoted to the passenger experience. CELLOCK has undertaken this task and has suggested two preliminary scenarios. The first scenario is about enhancing the passenger experience pre-flight, post-flight, as well as in-flight. The related use cases are presented in the table below.

**Table 3-7: Scenario 1 use cases for the CELLOCK demonstrator**

Scenario 1: Enhance pre, in and post flight Passenger Experience			
Use case examples		Algorithmic Task(s)	Algorithm(s)
Passenger modelling	profile	Clustering	SOM, k-means, Gaussian mixtures
Passenger prediction	profile	Classification	SVM, NB, Decision Tree, k-NN, MLP, ANFIS, DFFN
Queues prediction/ Wait time forecasting		Regression Analysis	CART, SVM, Random Forest, MLP, ANFIS, DFFN
		Time Series Prediction	ARIMA, ANFIS, RNN
Enhance passenger experience with special offers	passenger with special	Recommendation Systems	CF, CBF, Hybrid
		Association Rules	Apriori
		Optimisation	Genetic Algorithms

The second scenario aims to offer personalised content to the passenger by analysing the passenger's preferences and behaviour, as shown in the following table.

**Table 3-8: Scenario 2 use cases for the CELLOCK demonstrator**

Scenario 2: Offer personalised content (entertainment)			
Use case examples		Algorithmic Task(s)	Algorithm(s)
Passenger profile segmentation and modelling		Clustering	SOM, k-means, Gaussian mixtures
Context-aware passenger prediction	passenger profile	Classification	SVM, NB, Decision Tree, Random forest, k-NN, MLP, ANFIS, DFFN
Recommend personalized content		Recommendation Systems	CF, CBF, Hybrid
		Association Rules	Apriori

## 4 Data Sharing in Aviation

---

### 4.1 Background

The abundance of data created each day and the advancements in the data analytics area constitute an evident motivation to adopt data-sharing practices towards enabling more informed decision-making and more innovative problem-solving.

An important concomitant to data sharing has been the open data paradigm, particularly through efforts to concretely define openness and articulate its principles. The following properties, prerequisites for published data to be considered open, also serve as hints for the implications involved in making non-open data available to interested third parties (OpenDataCharter 2018; WorldBank 2018).

1. Open data are open by default, i.e. both legally (placed in the public domain or under liberal terms of use with minimal restrictions) and technically (non-discriminatory, publicly available and accessible on a public server, without password or firewall restrictions and published in machine processable and non-proprietary electronic formats). Data should be license-free, i.e. not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.
2. Open data are timely and comprehensive, i.e. made available as quickly as possible to preserve the value of the data.
3. Open data are accessible and usable, i.e. made available to the widest range of users for the widest range of purposes and also comparable and interoperable.
4. Open data are complete and primary, i.e. all data not subject to valid privacy, security or privilege limitations should be made available and also with the highest possible granularity level, not in aggregate or modified forms.

The aforementioned principles are often seen in the context of open government data, where openness serves as a transparency driver, promoting improved governance and increased citizen participation. Open data are credited for bringing benefits to various domains (Bertagnolli et al. 2017), and are gaining popularity, although their adoption is not always straightforward or even possible, depending on the domain and the sensitivity of data involved. Even if incentives to keep data private for personal professional advancement are gradually reduced based on the evident advantages of collaborating in more open environments (Olson and Downey 2013), legal and privacy concerns cannot be neglected even in the research domain. Indicatively, scepticism is very common when personal information may be revealed, as in the case of health data, where numerous data sharing barriers have been identified and even grouped under six wide categories: technical, motivational, economic, political, legal and ethical (Van Panhuis et al. 2014). These barrier categories are directly generalisable and applicable to other domains.

However, when significant advancements and benefits for the general public are expected, people are incentivised to overcome the data sharing barriers and devise solutions which foresee data exchange

under string terms among selected trusted parties (Aitken et al. 2016). Restrictions imposed by privacy, sensitivity and confidentiality exist in a wide spectrum of cases and not all data can or should be made broadly available, so data sharing does not refer only to the open data sphere. However, even though openness is gaining momentum regarding data (and also software code and algorithms) bringing further advancements in terms of new enablers and innovative services, sharing of sensitive data is not showing equally significant advancements. Privacy, legal and organisational policies and even infrastructure limitations may hinder data sharing and, thus, the benefits that stem from it. Data sharing agreements are neither new nor rare in Industry and the same is true for the discussions regarding their various aspects, drivers and barriers, but holistic frameworks to tackle these challenges are not yet established.

In this direction, the International Data Spaces (also known as Industrial Data Space) proposed by the IDS Association<sup>4</sup> (2019) aims at enabling a “network of trusted data” built on the core principles of sovereignty, protection of confidence, decentralization and security. It can be pragmatically viewed as a virtual data space leveraging existing standards and technologies, as well as accepted governance models, to facilitate the secure exchange and easy linkage of data in a trusted business ecosystem. In practice, with the help of its Reference Architecture Model (RAMI 4.0), it conceptually supports the establishment of secure data supply chains from the lowest layer (i.e. the data source) to the highest layer (i.e. data use), while at the same time ensuring data sovereignty for the data owners.

The NSF Spoke project “A Licensing Model and Ecosystem for Data Sharing”, led by researchers at Massachusetts Institute of Technology (MIT), Drexel University’s Metadata Research Center, and Brown University also aims to provide such a data sharing framework, aspiring to address the various data sharing challenges, at least, as stated, for the 80% of the cases - a statement that can only be interpreted as an admission of the underlying problem’s complexity. The project’s aspired contribution is three-fold:

1. A licensing model to facilitate data sharing between different organisations
2. A prototype data sharing software platform (ShareDB)
3. Relevant metadata that will accompany the datasets shared under the different licenses, making them easily searchable and interpretable

The foreseen contributions point to two very important axes of data sharing that are also very relevant to ICARUS:

1. A model that can express important data and data sharing attributes currently included in data trading contracts across domains for diverse data coming from and acquired by diverse stakeholders
2. A software platform capable of handling the data sharing, enforcing policies, ensuring quality and consistency of data and data contracts

As for the metadata, it is a common approach in this type of application to leverage metadata as a link between the model and its implementation. The number of metadata properties that will be required

---

<sup>4</sup> <https://www.internationaldataspaces.org/>



and their level of sophistication depends on the overall scope of the data sharing system to be developed. When metadata properties can be arranged in a set of predefined value groups, efficient and easier to implement solutions can be expected. Dataverse, an open source web application to share, preserve, cite, explore, and analyse research data has designed colour-coded data tags as a means of specifying security and access requirements for sensitive data that are instantly apprehended (visually) when used in commercial platforms (Bar-Sinai, Sweeney, and Crosas 2016).

In order to reduce the complexity of all the involved regulations and policies into a set of tags, a simultaneously broad and deep understanding of the domains and the data is required. The process usually starts with the collection and analysis of many data sharing agreements/contracts.

(Grabus and Greenberg 2017) have performed an analysis on 26 data sharing agreements from industry, academia and government and have identified the following six high-level, possibly partially overlapping, aspects of data licenses that affect data sharing:

- **General:** attributes relating to the project and the agreement itself, e.g. description of data, definition of terms
- **Privacy & Protection:** the protection of sensitive information and security
- **Access:** the definition of who can make contact with the data and how this can be done, including approved software and hardware
- **Responsibility:** legal, financial, ownership and rights management pertaining to the data, e.g. indemnity clause and establishment of data ownership
- **Compliance:** ensuring fulfilment of agreement terms, e.g. third-party compliance with contract
- **Data Handling:** specifics of permissible interactions with the data

It should be clarified that the term data license in this context refers to a broader concept than the one examined in D1.1 Section 4.5.1. where several predefined data licenses, such as Creative Commons, CDLA and Open Data Commons, were examined. Here, the aim is not to discuss the standardised data licenses, but instead to attempt to cover the complete spectrum of possible ad-hoc data sharing contracts, hence the features that are studied span more both horizontally and vertically. For each of the aforementioned aspects more fine-grained attributes are therefore gathered, identified and defined. Indicatively, for the “Privacy & Protection” category, the attributes depicted in the following figure are identified.

<b>Privacy &amp; Protection</b>		
<i>Sensitive Information</i>		
<i>Regulations</i>	<i>Preparing data</i>	<i>Access</i>
<ul style="list-style-type: none"> <li>• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)</li> <li>• Compliance with federal/state/international data protection laws and regulations</li> </ul>	<ul style="list-style-type: none"> <li>• Identification of confidential/special categories of information (e.g., pii, proprietary)</li> <li>• Individual identifiers removed/anonymized prior to transfer</li> </ul>	<ul style="list-style-type: none"> <li>• Who has access to pii/confidential data</li> <li>• Who has access to proprietary information</li> </ul>
<i>Privacy</i>	<i>Avoiding re-identification</i>	<i>Exceptions</i>
<ul style="list-style-type: none"> <li>• Anonymization of data</li> <li>• Confidentiality and safeguarding of PII/sensitive data</li> <li>• Removal/nondisclosure of company/personnel identification in materials and publications</li> <li>• No contact with data subjects</li> </ul>	<ul style="list-style-type: none"> <li>• No direct/indirect re-identification</li> <li>• Statistical cell size (how many people, in aggregated form, can be released in groups)</li> <li>• Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify)</li> </ul>	<ul style="list-style-type: none"> <li>• Exceptions to confidentiality</li> <li>• Conditions of proprietary information disclosure</li> <li>• Conditions of pii disclosure (who, what, and for what purpose?)</li> <li>• Limitations on obligations if data becomes public</li> <li>• Limitations on obligations if data is already known prior to agreement</li> <li>• Limitations on obligations if data given by 3<sup>rd</sup> party without restriction</li> </ul>
<i>Security</i>		
<ul style="list-style-type: none"> <li>• Sharing non-confidential data</li> <li>• Password protection/authentication of files</li> <li>• Encryption</li> <li>• Security training for involved personnel</li> <li>• Establishing infrastructure to safeguard confidential data</li> </ul>		

**Figure 4-1: Privacy & Protection Attributes of Data Sharing Agreements** (Grabus and Greenberg 2017)

It becomes evident that an exhaustive list of attributes that address all aspects of any potential data sharing scenario is impossible to create and even then, the appropriate values for all these attributes would have to be specified. To avoid this complexity, as well as the implications stemming from the inherent domain-dependence of certain aspects, (Truong et al. 2012) propose an extensible model to be created through a process centred on a combination of community- and people-centric collaboration, enabling the domain experts to extend/adapt the model to the data agreements of the domain. Their model is designed to capture contractual terms for data contracts and represent them in a form that can be reasoned upon automatically. In order to identify the fundamental elements of data sharing contracts, they have also conducted an analysis of existing data contracts. Their study concludes on the following five distinct categories:

- **Data rights**, e.g. derivation, collection, reproduction, attribution, non-commercial use.
- **Quality of data (QoD)**
- **Regulatory Compliance**
- **Pricing**
- **Control and relationship**, i.e. an indication of the geographical regions where the underlying terms are applied. As an indicative example of why this is very important and may have legal implications, the value “Austria” in certain contexts should be interpreted as a sub element of European Union (EU).

There are obvious links between these and the six categories from (Grabus and Greenberg 2017), and even the health related categories (Van Panhuis et al. 2014). As it can be expected, the term categories above appear with variations in titles and/or expected values, in various other studies in literature, with each one focusing on the attributes most important for the specific underlying data sharing problem.

Indicatively, regarding pricing, (Cao et al. 2016) identify the following prevalent practices:

- **Payment on package delivering (API handle):** data are split into separated packages (e.g., messages or images) and consumers are charged every time they successfully receive a number of packages from the marketplace.
- **Payment on data size:** consumers are charged on the size of received data, e.g. per MB, GB, etc.
- **Payment on time of subscription:** consumers are then charged on total time of each moment they subscribed. This model is appropriate for streaming data where the data is generated in a duration.
- **Payment on data unit:** providers split their data in different data units and set up the basic unit charge fee for each. Consumers will pay one time and get the data until reaching the limitation of unit.
- **Payment on plan (fixed payment on a period):** consumers subscribe to use data in a subscription period (e.g., a week or a month) and only pay one time for this period with or without maximum limitation of received data.
- **Free usage:** consumers can use these services at no charge from providers. There are some reasons to offer a service for free such as: (i) the data comes from the government and the consumer is a public authority funded by tax money. This case is usually constrained by a data contract. (ii) a person or an organisation can also provide the free data as a social responsibility because the generating data fee is supported by the other organisation or the government.

These are fixed payment plans, but pricing may also refer to more dynamic schemes computed over the available data assets. (Heckman et al. 2015) have collected a set of attributes based on which they construct machine learning regression models that help establish a correlation between data attributes and the price of a given dataset. These attributes, taken directly from this work, are as follows:

- Value-based parameters (value of data to the consumer):
  - The value of the data in terms of saving in time, effort, or money
  - The ROI for the customer
  - Risk exposure, i.e. inducing higher costs for data cleansed of personally identifiable information and privacy violations
  - Data exclusivity
  - Level of ownership, ranging from transfer of ownership to allowing use for a fixed time to allowing limited use for a specific purpose
- Qualitative parameters (attributes or meta-attributes of the dataset):
  - Age of the data

- Credibility of the data
- Accuracy of the data elements
- Quality of the data
- Format and level of structure of the data
- Fixed and marginal cost parameters (directly measurable cost):
  - Cost of collecting the data
  - Cost of data storage, bandwidth, and other operational costs
  - Cost of data-as-a-service offerings – add-on services to process the data, computing resources for the data, analytic reports, or aggregation on the data
  - Delivery cadence – one-time, batch, or continuous basis

From this perspective, several attributes examined to define the appropriate pricing strategy depend on the other term categories, such as the data rights regime and the data quality. The latter was discussed in D1.1, whereas the rights regime, tied to the data licensing, is by itself a very complex issue, as can be shown by the following 18 Licensing and Rights Management Initiatives, presented below grouped under six overlapping categories:

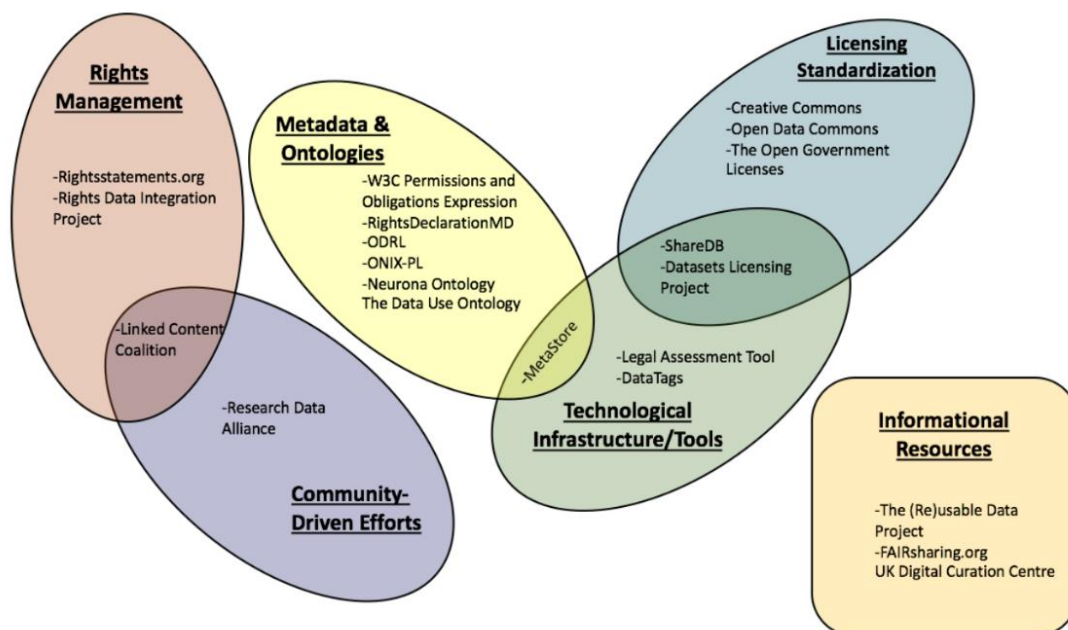


Figure 4-2: Licensing and Rights Management Initiatives Classification (Koko et al. 2018)

## 4.2 Data Marketplaces

Apart from the studies that aim to capture specific perspectives of the generic data sharing problem, other, more targeted works attempt to address specific challenges that manifest in a given domain and examine all perspectives related to one type of data and/or one particular domain.

Indicatively, (Balint and Truong 2017) introduce a new extensible platform for enabling contract-aware IoT dataspace services, which supports data contract specification and IoT data flow monitoring based on established data contracts, whereas (Sakr 2018) study the challenges related to spatial data

sharing and propose a model and query algorithms for this type of platform. The European project AEGIS has defined a conceptual policy and brokerage framework, which covers several aspects related to data assets rights, data quality, policies and pricing models (AEGIS 2017). The presented framework aims to cover numerous trading workflows for various virtual assets, including datasets and data-as-a-service, but also spans to microservices, algorithms and analytics reports, hence it is inherently more generic.

The motivation behind the aforementioned studies and initiatives largely stems from the need to create data contracts in non-textual, machine-processable form, which could serve as automation enablers for data marketplaces. A data marketplace in this context is a multi-sided platform, where a digital intermediary connects data providers, data purchasers, and other complementary technology providers (HBR 2006). Data marketplaces are a relatively new type of supplier, but constitute a promising growing market in the EU (Carnelley et al. 2016). The current marketplace landscape includes generic purpose platforms provided by large cloud computing companies (e.g. Microsoft Azure and Amazon), open data platforms with market features and vertical marketplaces, i.e. within-industry targeted to specific types of data or domains of application. Data abundance and integration with cloud-based data analysis and visualisation solutions have bolstered these marketplaces as data sharing enablers across a variety of domains and applications. Therefore, the problem of data sharing cannot be seen independently of data marketplaces.

(Koutroumpis, Leiponen, and Thomas 2017), in their attempt to develop a conceptual market design framework to examine possible forms of governance for trading data, including their benefits and deficiencies, identify four main types of data marketplaces, depending on the way buyers and sellers are matched, each of which manifests different qualities in terms of design and data typology:

1. One-to-One: These regard a bilateral relationship, common in how data brokers operate, with negotiated terms of exchange. They have low liquidity, as locating trading partners in an environment with secretive transactions can be difficult. Transaction costs are expected to be high, provenance and boundaries are clear, rules are strong with invasive monitoring. Data traded in these marketplaces are of high value and high confidentiality.
2. One-to-Many: These dispersal marketplaces refer to cases when a single seller transacts with many buyers for the same data, e.g. data being distributed through APIs to many interested buyers. Usually terms of exchange are standardised, due to the prohibitive costs associated with individual, per buyer, negotiations. Therefore, transaction costs tend to be low, while provenance and boundaries are unclear and enforced rules are weak with minimal monitoring. Data are usually of low value and confidentiality.
3. Many-to-One: These marketplaces are characterised by the harvesting of data of many users by one service provider, common in cases when users get “free” access to a service in exchange for their data. Liquidity is obviously high, transaction costs low and provenance and boundaries unclear. The underlying rules are weak, with minimal monitoring, and data are of low value (when examined individually) and low confidentiality. This marketplace type is vulnerable to repugnance concerns such as norms related to privacy. As an example, the General Data Protection Directive

(GDPR) gives users a “right to be forgotten”, which could inflate the costs associated to monetising user data.

4. Many-to-Many: These refer to multilateral trading platforms upon which anybody (or at least a large number of registered users) can upload and maintain datasets. They can be centralised or decentralised and there are varying licensing models – standardized or negotiated – to regulate data trading. Data confidentiality and value also vary, transaction costs are usually low, at least in non-big data applications, and liquidity is high. These marketplaces often emphasise on data discoverability and other facilitation activities, including online payment. In their most common form, these platforms do not have ownership of the data, but only act as intermediaries that facilitate transactions. This last characteristic is why other attempts to formalise data trading contracts also aim to limit the liability of data and data-as-a-service (DaaS) providers in case of failure of the provided data (Truong et al. 2012).

The last category provides high flexibility in terms of involved stakeholders and data trading options and has sparked the development of machine processable data contracts. In this context, there is an emerging need to develop contract engines that provide querying and validation mechanisms for access to and usage rights of data assets, as well as the status of the agreements being performed. However, the majority of existing marketplaces are far from this level of automation and lack an adequately expressive, but not prohibitively -for implementation- complex, information model (Vu et al. 2012). Apart from overcoming numerous of the aforementioned data sharing barriers, the model is required for usability purposes as well. Current data marketplaces require manual search from prospective buyers and do not allow on-demand data integration or query optimisation to be performed. The latter is extremely important in decentralised many-to-many marketplaces that may offer the same or similar data assets or DaaS services under different terms and/or data quality. Furthermore, without designing and implementing this model, data information and DaaS engineering cannot be tied, therefore making it possible for inconsistencies to manifest between what is described in data descriptions (HTML documents) and what is actually offered through the DaaS (Vu et al. 2012).

Another important aspect hindering the adoption of data marketplaces in several domains operating with non-open data is the difficulty in establishing rigorous provenance through verifiable information for the data being sold, i.e. the questionable robustness of the existing rights regime (Vu et al. 2012). Towards this goal, distributed ledger technologies (DLTs) are now being leveraged in the design of the decentralised multilateral platforms. The advantages of DLTs in this context are numerous and widely accepted. Among others, they address the need to enforce transparency and data democratisation (Özyılmaz, Doğan, and Yurdakul 2018), they eliminate the single-point of failure problem of centralised systems, they enable wider participation in the data markets previously monopolised by technology giants. One important reason behind the technology’s adoption in such applications is the fact that, at least one of the DLTs and probably the most popular, the blockchain, facilitates the creation and application of cryptocurrency models, hence economic incentives are easier to create, thus making it a perfect match for marketplaces. But the adoption of DLTs brings also several domain-, data- and

application-dependent advantages in data sharing platforms, as can be seen by the numerous initiatives in the domain of DLTs and their application in data trading context. Indicatively:

- **IOTA**<sup>5</sup> is an open-source distributed ledger designed to enable feeless microtransactions over IoT data and allow machine-ensured data integrity. IOTA uses the Tangle DLT technology that solves certain blockchain inefficiencies in terms of transaction times and scalability, and has already launched its data marketplace.
- **Wibson**<sup>6</sup>, currently preparing to enter the market, is a decentralized blockchain-run marketplace that will allow members of the public to both manage and profit from securely and anonymously selling their personal data that are validated for accuracy.
- **Hu-manity.co**<sup>7</sup> narrows down this concept to only health related data and allows users to sell their patient data to big pharma directly by cutting the middleman.
- **Quorum**<sup>8</sup> is an Ethereum-based distributed ledger protocol with transaction/contract privacy and new consensus mechanisms, which can support directed transfer of private data to network participants.
- **Datapace**<sup>9</sup> offers a marketplace for IoT data which are stored encrypted anonymised.
- **Repux**<sup>10</sup> helps businesses sell anonymised data to developers who use them to implement machine learning models and may in turn sell their AI-enhanced applications to interested businesses, all through the platforms' cryptocurrency, the Repux tokens.
- **Corda**<sup>11</sup> is a blockchain and smart contract platform that allows parties to transact directly, with value and realise complex agreements of any asset type, applicable across industries including finance, supply chain and healthcare.
- The **Datum**<sup>12</sup> network allows anyone to store structured data securely in a decentralized way on a smart contract blockchain, optionally enabling selling and buying data using the DAT smart token.
- The **ContractVault**<sup>13</sup> platform offers various Smart Templates and legally-enforceable smart contracts, making the creation, management and integration of contractual processes on the blockchain simple.
- **Streamr**<sup>14</sup> offers a marketplace for real time data, leveraging blockchain and Ethereum-based smart contracts for security critical operations like data transfers.

---

<sup>5</sup> <https://www.iota.org/>

<sup>6</sup> <https://wibson.org/>

<sup>7</sup> <https://hu-manity.co/>

<sup>8</sup> <https://quorum.org/>

<sup>9</sup> <https://www.datapace.io/>

<sup>10</sup> <https://repux.io>

<sup>11</sup> <https://www.corda.net>

<sup>12</sup> <https://datum.org>

<sup>13</sup> <https://www.contractvault.io>

<sup>14</sup> <https://marketplace.streamr.com>



- The **Enigma Data Marketplace**<sup>15</sup> contract was recently announced on the Ethereum testnet. The underlying protocol, Enigma, solves privacy and scalability issues for any blockchain, but is particularly relevant for data marketplaces.

Most of these platforms build upon the need to redefine, innovate and ensure the core principles in data sharing, namely trust, ownership, fairness in value creation and data profits, all of which are inherent in the way DLTs are designed, i.e. offering transparency of state and execution and assurance. The popularity of DLTs is pushing researchers to devise ways to overcome certain inherent limitations so as to further broaden their potential. As an example, privacy and auditability can be required in certain industry-level data trading agreements. (Bhaskaran et al. 2018) propose a new way of creating double-blind data contracts over blockchain which can ensure anonymity of relationships between providers and consumers, with dynamic consent and data access control. (Park et al. 2018) focus on the potential scepticism against peer-to-peer systems and the way this could pose a risk for significant data trading agreements if the identity of both parties is not ensured, especially in the IoT data landscape where there tend to be numerous small-scale providers than a small number of well-known entities. They have developed a review blockchain-based mechanism that enforces the usage of specific metadata to ensure data integrity and prevent malicious behaviour by members of the network. (Biryukov, Alex, Dmitry Khovratovich 2018) examine ways on how to bring the traditional KYC (Know Your Customer) procedures in the new DLT landscape. The proposed solution addresses the need to limit the set of users allowed to participate in a smart contract based on some dynamically defined criteria. (Özyilmaz, Doğan, and Yurdakul 2018) also provide ways to eliminate unreliable data providers in their marketplace that attempts to address the scalability issues of employing blockchain based solutions in encrypted IoT data sharing, by relaxing the requirement for time-critical operations. The previous and current sub-sections aimed to provide a comprehensive summary of the main data sharing practices and considerations, but the field is too broad to attempt an exhaustive analysis. It is important to highlight that data sharing is an advancing multi-faceted field, actively being researched in the theoretical sphere and bearing strong links to the flourishing data marketplace economy. However, the implications and requirements of creating data sharing systems, even conceptually, are important, diverse and challenging to address and are better addressed when put into the specific data, domain and application context. The next sub-section will delve into the specificities of the aviation industry and its data sharing practices, current and future.

### 4.3 Data Sharing Motivation and Initiatives in Aviation

Data sharing is not a new concept for the aviation industry. GAIN, the Global Aviation Information Network, aiming to promote and facilitate the voluntary collection and sharing of safety information by and among users in the international aviation community to improve safety (GAIN 2003), was proposed by the Federal Aviation Administration (FAA) in 1996. Since then, numerous multi-airline

---

<sup>15</sup> <https://enigma.co/>

and multi-national data sharing programs and initiatives which involve centralising airline flight data storage have been established, the most important of which are presented below:

- **FDX<sup>16</sup>**: Flight Data eXchange (FDX) is an aggregated de-identified database of FDA/ FOQA type events that allows to identify commercial flight safety issues for a wide variety of safety topics. It covers many types of aircraft, across a global database and enables flight operations and safety departments to proactively identify safety hazards. FDX has data from more than 100 airports, which makes up about 500 runways, thus allowing operators to accomplish over 50 different types of runway specific safety analysis and hazard identification utilising industry data, fine tuning their own operations to minimise risk.

Currently, numerous different events are displayed by location including Ground Proximity Warning System (GPWS/ TAWS) locations, Traffic Collision and Avoidance System (TCAS) events, Windshear warnings, Unstable approaches, Go-arounds, and high tailwind landing events.

- **ASIAS<sup>17,18</sup>**: The Federal Aviation Administration (FAA) and the aviation industry developed the Aviation Safety Information Analysis and Sharing (ASIAS) program to promote an open exchange of safety information. ASIAS has drawn together a wide variety of safety data and information sources across Government and industry, including voluntarily provided safety data. ASIAS incorporates voluntarily provided safety data from operators that represent 99 percent of U.S. air carrier operations in the National Airspace System (NAS). ASIAS continues to pioneer advanced analytical capabilities to provide safety teams with enhanced insight into these operations. ASIAS stakeholders currently include, 45 commercial air carriers, 83 general aviation operators and many other from Industry and Government. ASIAS has established metrics that enable Commercial Aviation Safety Team (CAST) to monitor and evaluate the effectiveness of deployed safety mitigations. In recent years, CAST has evolved to a proactive approach that focuses on detecting risk and implementing mitigation strategies before accidents or serious incidents occur.

The ASIAS data repository continues to expand, incorporating a wide variety of public and proprietary data sources. Each source provides information from different parts of the NAS.

- **GADM<sup>19,20</sup>**: GADM is the IATA Global Aviation Data Management programme and platform. Over 90% of IATA member carriers have agreed to participate in this programme which currently receives data from more than 470 organisations. The participation in the programme allows data providers to access aggregated and de-identified reports on safety metrics and trends, including analyses on accidents and incidents, operational reports and ground damage reports. The IATA Ground Damage Database (GDDB) is a key initiative supporting the IATA Global Ground Operations activities. Key stakeholders from the Air Transport Industry collectively identified ground damage data fields that can and should be reported into the GDDB consistently amongst all participants, as well as the parameters for how it is to be reported.

---

<sup>16</sup> <https://www.iata.org/services/statistics/gadm/Pages/fdx.aspx>

<sup>17</sup> [https://www.faa.gov/news/fact\\_sheets/news\\_story.cfm?newsId=23036&omniRss=fact\\_sheetsAoc&cid=103\\_F\\_S](https://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=23036&omniRss=fact_sheetsAoc&cid=103_F_S)

<sup>18</sup> <https://portal.asias.aero/web/guest>

<sup>19</sup> <https://www.iata.org/services/statistics/gadm/Pages/index.aspx>

<sup>20</sup> <https://www.iata.org/services/statistics/gadm/Pages/gddb.aspx>

- **STEADES<sup>21</sup>**: STEADESTM is IATA's aviation safety incident data management and analysis program and one of the data sources of the Global Aviation Data Management (GADM). With over 200 members, the STEADES database of de-identified airline incident reports is the world's largest, offering a secure environment for airlines to pool safety information for global benchmarking and analysis needs.
- **Skybrary<sup>22</sup>**: SKYbrary is an electronic repository of safety knowledge related to flight operations, air traffic management (ATM) and aviation safety in general. It also is a portal that enables users to access the safety data made available on the websites of a variety of aviation organisations. SKYbrary was initiated by EUROCONTROL in partnership with the International Civil Aviation Organisation, Flight Safety Foundation, the U.K. Flight Safety Committee and the European Strategic Safety Initiative. SKYbrary's objective is to become a single point of reference for aviation safety knowledge by making universally available and accessible the safety knowledge accumulated by various aviation organisations, entities and initiatives. The SKYbrary knowledgebase is a dynamic enterprise that has taken several years to develop.
- **Data4Safety<sup>23</sup>**: Data4Safety (also known as D4S) is a data collection and analysis programme that aims to support the goal to ensure the highest common level of safety and environmental protection for the European aviation system. The programme aims to collect all data that can help the management of safety risks at European level, including safety reports (or occurrences), flight data (i.e. data generated by the aircraft via the Flight Data Recorders), surveillance data (air traffic data), weather data - but those are only a few from a much longer list. On the 31st of March 2017, key actors from the aviation sector agreed to join in a co-operative partnership the Data4Safety programme initiated by EASA. Participants were: EasyJet, British Airways, Iberia, Deutsche Lufthansa, Ryanair, Airbus, the Boeing Company, the European Cockpit Association (ECA), the Spanish Aviation Safety and Security Agency (AESA), Direction de la Sécurité de l'aviation civile (DSAC France), the Irish Aviation Authority (IAA), the United Kingdom Civil Aviation Authority (UK CAA), the European Aviation Safety Agency (EASA) (Airlines, Aircraft Manufacturers, National Aviation Authorities and Pilot Unions).
- **A-CDM<sup>24,25</sup>**: The European Airport Collaborative Decision Making, also known as A-CDM, was based on the American concept of Collaborative Decision Making that was introduced in January 1998 to cope with heavy capacity reductions due mainly to en route or airport bad weather conditions. Delays during ground delay programs were reduced by 15 percent during the experimental period. In early 2000, trials were conducted at several major European airports to study and develop a CDM concept for Europe. This led to the creation of the Airport CDM Task Force under the EATM Airport Throughput Division (APT) to guide the Airport Operations Team (AOT) in A-CDM issues and undertake specific work. The decision making by the A-CDM Partners

---

<sup>21</sup> <https://www.iata.org/services/statistics/gadm/steades/Pages/index.aspx>

<sup>22</sup> <https://flightsafety.org/resource/skybrary/>

<sup>23</sup> <https://www.easa.europa.eu/newsroom-and-events/news/data4safety-partnership-data-driven-aviation-safety-analysis-europe>

<sup>24</sup> <https://infrastructuremagazine.com.au/2018/05/31/data-sharing-system-to-benefit-aviation-industry/>

<sup>25</sup> [https://www.skybrary.aero/index.php/Airport\\_Collaborative\\_Decision\\_Making\\_\(A-CDM\)](https://www.skybrary.aero/index.php/Airport_Collaborative_Decision_Making_(A-CDM))

is facilitated by the sharing of accurate and timely information and by adapted procedures, mechanisms and tools. The main A-CDM Partners are the Airport Operator, Aircraft Operators, Ground Handlers, De-icing companies, the Air Navigation Service Provider (ATC), the Network Manager and support services (Police, Customs and Immigration etc).

- **SkyFusion<sup>26</sup>**: SkyFusion is an outcome of the strategic partnership between IATA and HARRIS. It was developed to help core ATM stakeholders, namely airlines, ANSPs and airports, to easily overcome the limitations of today's systems and effectively meet the challenges ahead by providing them with the ability to connect, communicate, share data, and make collaborative decisions in real-time. SkyFusion is a SWIM-configured platform that features data exchange and CDM tools, including data exchange and a chat hub that enables stakeholders to communicate real time in querying the information on the platform, promoting situational awareness and real-time alignment of involved stakeholders.

Finally, the European Aviation Safety Agency (EASA) and the International Air Transport Association (IATA) announced in 2014 an agreement on sharing of safety information and joint analysis of safety trends<sup>27</sup>. These analyses primarily will be based on the information derived from the Safety Assessment of Foreign Aircraft (SAFA) program, and the IATA Operational Safety Audit (IOSA). IATA is also developing an ACID (Air Cargo Incident Database<sup>28</sup>), as part of StB Cargo<sup>29</sup> program. This database of de-identified airline incident reports will offer a secure environment for airlines and ground handlers to pool their safety and operations information, supporting a proactive data-driven approach for advanced trend analysis, predictive risk mitigation and improvement programs.

The primary goal of such initiatives is to ensure safety in the air travel, which in turn requires the optimisation of a wide range of operations, e.g. the way Airline Operations Centres build schedules, plan flight routings and fuel uplift and ensuring passenger connections, the way ANSPs organise and manage the airspace over a country with Air Traffic Services, the way Military Operations Centres plan their missions, block airspace to conduct training operations and fulfil national security tasks. Information to be shared in these scenarios includes aeronautical data, flight trajectories, aerodrome operations, historical and current meteorological data, air traffic flow information, surveillance data (from radar, satellite navigation systems, aircraft datalinks, etc.), capacity and demand data (actual and foreseen).

The aforementioned programs and initiatives may, as stated, be mainly focused on the core ATM stakeholders and the facilitation of data sharing and the effort to agree on common standards among them, however the launch of such Aviation Data Exchange programmes has also opened the door to data sharing with trusted third parties (DataScience.aero 2018). The term third party in the aviation industry may refer to core domain stakeholders, due to the fact that the industry comprises numerous diverse stakeholders who operate in extreme proximity, but, flight safety aside, are in a different business (Mindtree 2018). For example, airlines are in the transportation business whereas airports

---

<sup>26</sup> [https://www.iata.org/services/safety-flight-operations/Documents/SkyFusion\\_Brochure\\_2017.pdf](https://www.iata.org/services/safety-flight-operations/Documents/SkyFusion_Brochure_2017.pdf)

<sup>27</sup> <https://www.internationalairportreview.com/news/16788/iata-easa-reach-agreement-on-information-sharing/>

<sup>28</sup> <https://www.iata.org/whatwedo/cargo/Documents/cargo-strategy.pdf>

<sup>29</sup> <https://www.iata.org/whatwedo/cargo/stb/Pages/index.aspx>

are in the real estate business, hence their efficiency is measured in completely different KPIs and their revenues are increased through different activities. Considering that in many airports a lot of the provided services are in fact offered by different stakeholders, e.g. parking spaces in the AIA, the data fragmentation becomes more evident. In the current data-driven era, stakeholders are starting to realise the potential benefits from data sharing and ad-hoc collaborations are spawning across the Industry. EasyJet has partnered with Gatwick Airport on the latest update for its mobile app, which is designed to make the departure and arrival processes as simple as possible (FTE 2015b). Dublin airport has partnered with Aer Lingus on a data sharing pilot (FTE 2015a), expected to significantly improve both their operations and the overall passenger journey experience. The **Skywise**<sup>30</sup> data platform, launched by Airbus in 2017 in collaboration with Palantir Technologies, is another larger-scale initiative towards improved data sharing and it aspires to become the reference platform for core aviation stakeholders to improve their operational performance.

The advantages of data sharing in the aviation industry are numerous and multi-faceted and bolster the development of further larger-scale collaborations. IATA, Eurocontrol, other core stakeholders of the aviation industry but also stakeholders of the broader spectrum, all report on the expected advantages of collaborations built on shared data and insights (FAA 2012).

Indicatively, easier real-time communication among airlines, ANSPs and airports will significantly contribute to:

- increased airline safety,
- greater situational awareness to resolve en route and terminal constraints and manage airspace more efficiently,
- enhanced usage of resources and demand-capacity balancing
- improved on time performance and reduced holding costs
- reduced pilot flights and duty times
- optimisation of ATC staffing levels
- increased participation in the decision-making process
- integration with a-CDM initiatives
- development of a SWIM-compliant air traffic network

For each of the above generic benefits, concrete examples are easily devised. Indicatively, increased airline safety can be accomplished by airlines by:

- Gathering additional information of an event that an airline has not yet experienced to early identify an emerging trend or assess the effectiveness of potential corrective action
- Gathering information of an encountered event to identify other airlines that have experienced the same or similar problems to obtain information on the problem characteristics and their experience with corrective actions
- Creating an alerting system for airlines regarding occurrence of events and enable comparison of experiences regarding frequency of events and severity of incidents

---

<sup>30</sup> <https://www.airbus.com/aircraft/support-services/skywise.html>

- Gathering information on an operational area where an airline has limited or no experience, e.g. a new aircraft or a new airport

Thus, the airlines would be able to:

- Take corrective actions that are less costly and timelier than they would otherwise be, thus leading to less flight delays and cancellations due to early recognition of maintenance problems, reduce incidents like flap overspeed and hard landings
- Save staff time by reducing the need to make individual inquiries to gather information
- Eliminate the element of chance in sharing info through ad-hoc networking

The importance of data sharing in the aviation domain is underlined, especially in the safety context, by the limited data availability in some cases, e.g. the number of safety incidents. Traditionally, data sharing systems in aviation belong to three main types, depending on the time the sharing is realised compared to when the event actually took place and the level of processing of the data being shared have undergone, i.e. whether the data are (almost) raw or whether data have been processed and what is shared is the result of the analysis: near-real time event sharing systems, periodic aggregation and analysis systems and lessons learned and corrective action systems. Technological advancements can gradually empower truly real time data sharing systems and data fusion from various sources, thus enabling a 360-degree view of the passenger journey and inspiring innovative services that go beyond security and safety. Indeed, examples not directly relevant to safety are also becoming increasingly appealing to aviation stakeholders looking to enhance their offered products and services (FTE 2015a; Mindtree 2018). Indicatively:

- Airline-Airport data sharing can speed up check-ins. The airport knows how many passengers have arrived, but the airline only knows how many haven't checked in. If this simple information is shared between the two, an airport can quickly estimate how many extra check-in counters will be required to process all the passengers and allocate the right numbers. Moreover, the airport knows that the airline's passenger is in the terminal when the passenger goes through security. So, by sharing that information with the airline, when they are ready to close the gate but know they're missing a number of passengers, that information can help the airline make an informed decision on whether to wait or go ahead and close the doors.
- Enhanced forecasting of expected time to be spent by passengers in waiting queues can also be accomplished by using a variety of techniques, including heat sensing and video analytics using computer vision. Sharing such information can thereby enable airlines to trigger appropriate passenger notifications.
- Better passenger movement, service and revenue through communication of airline information to the airports. Indicatively, an airline knows the number of children to board a flight, or the number of people of a given nationality, well in advance and could pass along this information to the airport to help provision facilities for families with children or provide targeted shopping offers for specific segments and nationalities. Wheelchairs availability, baggage assistance and golf carts to move passengers with connecting flights over large distances between gates could also be provided in a dynamic, more targeted, manner.

The potential benefits from establishing broader and more efficient data sharing practices among aviation stakeholders have attracted recently research interest. The European Project SafeClouds<sup>31</sup> aims to combine various datasets from diverse users, such as airlines and ANSPs, in a novel data mining approach for aviation safety. The project aims to address challenging aspects of data analysis in the domain, including fusion of proprietary confidential data and performing benchmarking among competitive stakeholders whilst ensuring that specific parts of the data will not be explicitly shared or inferred, e.g. through applying multiparty computation techniques. Datascience.aero<sup>32</sup> on the other hand aspires to provide data analysis use cases that will inspire stakeholders to stimulate joint initiatives towards aviation data knowledge discovery, which can be expected to foster broader data sharing partnerships. From a more academic perspective, (Reis, Rocha, and Castro 2018) propose an airline marketplace, modelled as a multi-agent system with an automated negotiation mechanism, where airlines can announce availability of resources (aircraft or aircraft and crew) for lease and other airlines can go there to contract resources to fill gaps in the operation, typically due to disruptions and/or an unexpected increase on the operation. In practice though, the data sharing reality in the industry is nowhere near this level of automation, as shown in the next section.

#### 4.4 Current Data Sharing Agreements from ICARUS Stakeholders

The previous sections provided a landscape analysis of data sharing practices, initiatives, incentives and challenges, based on an extensive review of several sources. Insights from the projects' industry partners were leveraged to ensure that the performed analysis examines all aspects that are significant for the domain, but the spectrum was deliberately kept broader, so that findings would bring new ideas and unforeseen challenges into consideration. Having ensured this, the next step is to report the data sharing practices of the project's industry partners, which provide a level of detail that cannot be obtained from external sources. Each of the four industry partners was asked to provide templates and samples of the documents currently used in data trading agreements/contracts in order to understand what are the IPR, legal, technical etc. aspects of these agreements that are relevant and important in the domain and how they are addressed in practice now. The provided documents cannot be reported as-is in the deliverable. Instead, their core aspects were extracted and are presented below.

##### **Athens International Airport (AIA)**

Athens International Airport does not have a formal data sharing agreement document, i.e. a data contract. Instead, in these situations a standard non-disclosure agreement (NDA) is used, which specifies the following information:

- The date as of which the NDA is binding
- The parties that are involved in the agreement

---

<sup>31</sup> <https://ec.europa.eu/inea/en/horizon-2020/projects/h2020-transport/aviation/safecLOUDS.eu>

<sup>32</sup> <https://datascience.aero/discovering-hidden-knowledge-aviation-data/>



- The purpose of the meetings and exchange of correspondence, information and documentation that are foreseen under the agreement
- The definition of what constitutes confidential information according to the agreement
- The allowed actions for the receiving party
- The ownership of the information to be shared
- Other legal obligations of the two parties, including indemnification

### **OAG**

OAG, due to the nature of its core business that involves data sharing contracts, has very targeted template documents for such situations. Specifically, these include:

- A document specifying the standard general terms and conditions of any agreement. This document holds information regarding:
  - Products and/or services involved in the agreement
  - Payment information and obligations
  - Ownership, Restrictions on Use, Licensing, law compliance and Confidentiality clauses
  - Termination clause
  - Limitation of liability and indemnification
  - Other common contractual terms, such as Governing Law and Force Majeure.
- A document listing the permitted usage of the provided by OAG data/products. The list has 12 predefined options and it is obligatory that one of them is chosen for the agreement.
- A set of terms that are relevant to specific OAG products.

The remaining Industry stakeholders of the project, namely PACE and CELLOCK, do not have legal documents for usage in a data sharing context as they have not acted as data providers in the past.

## **4.5 Key Considerations for Data Sharing in Aviation**

The previous sections provided a comprehensive, yet not exhaustive, review of data sharing aspects both general and targeting the aviation industry, which showed that there is growing interest in the data sharing field and, at the same time, numerous challenges that have yet to be addressed. The problem is multi-faceted and in many cases domain dependent or even data-dependent, in the sense that each piece of data may be inherently different in terms of privacy, sensitivity and legal implications.

On one hand, efforts in the broader data sharing area largely aim to enable the automated data trading through data marketplaces capable of enforcing contractual terms and regulating the process in a way that makes it trustworthy by both sellers and buyers. This ambitious goal pushes towards developing metadata models aiming to capture the diverse aspects of data sharing agreements, including licensing, pricing, ownership, provenance, allowed usage and indemnification. The attempts to design and implement such models, in turn help identify and highlight complexities involved, thus

contributing towards deeper understanding of the underlying issues and gradually enabling the provision of solutions.

On the other hand, data sharing initiatives among the core aviation stakeholders, although common, are mainly either ad-hoc one-to-one collaborations or large initiatives usually related to safety incidents. One-to-one agreements typically involve lengthy discussions and custom memoranda of understanding, NDAs and bailee agreements. The limited availability of templates and/or standard documents for data sharing agreements, presented in Section 4.4 indicates that the core aviation industry is not in sync with the advancements in the field, although as established earlier in the review, this is not due to limited expectations or lack of interest from the stakeholders' perspective. As explained, from the four data marketplace types discussed in Section 4.2, the last one, i.e. the many-to-many, is the one that shows the most promising potential to foster broader data sharing that will lead to innovative services creation. This is especially true in domains that, like aviation, do not comprise data selling businesses, but stakeholders that potentially have payable data as a side product of their work. Hence, they would profit from a data sharing enabler, but are not inclined or capable of developing their own system.

There are numerous challenging aspects when it comes to fostering a data sharing mentality and transforming the current data-siloed situation to a more open collaborative environment. The key aspects and considerations, revealed through the performed landscape analysis and through insights provided directly by the domain stakeholders of the ICARUS consortium are as follows:

#### **Key Data Sharing Consideration I: Data privacy and sensitivity**

Data privacy and confidentiality should not be sacrificed under any circumstances, regardless of the potential benefits. Companies operating in the EU are, since May 2018, bound by the General Data Protection Regulation (GDPR), which enhances personal data privacy, strengthens citizens' rights to gain information about the use of their personal data and thus has forced companies to review their personal data handling processes and inform people of the data that are being collected and the ways they are used. Sensitive data in the aviation industry mostly refer to passengers' personal data. Formally, passenger data comprises the Advance Passenger Information (API) and the Passenger Name Records (PNR). These data can be a useful tool for governments' border control or security processing as it can help them pre-identify travellers and patterns. The concept of a PNR (IATA 2010) was first introduced by airlines that needed to exchange reservation information in case passengers required flights of multiple airlines to reach their destination ("interlining"). For this purpose, IATA and ATA defined standards for interline messaging of PNR and other data through the "ATA/IATA Reservations Interline Message Procedures - Passenger" (AIRIMP). IATA states that personal data should be transferable and secure (IATA 2016), but the personal data processed must be minimal and proportionate: the PNR Directive refers to the EU Charter of Fundamental Rights, underlining the proportionality and data minimisation principles, as does the GDPR (Regulation EU 2016/679 GDPR). IATA has even issued guidelines (IATA 2018) regarding the explanatory memorandum that should be added to e-tickets, whereas individual stakeholders have published relevant information addressed to

flight passengers so as to disclose their private data collection and usage policies and reassure them that personal information is respected, e.g. (Finavia 2018).

When it comes to sharing personal data outside the above PNR context, e.g. to facilitate the provision of innovative services as the ones discussed earlier in the chapter, GDPR is stringent and prohibitive. Usage of anonymous tokens that link various reservations to each other is the industry's recommendation in cases where the personal information does not need to be shared. As an example, an OTA may not want to share customer contact information with an airline, but if the flight is cancelled the airline has no way to reach out to the passenger. In this situation, the OTA could generate an anonymous token that is passed to the airline and serves as the verification of passenger identity. The airline would not need the passenger's phone number, but rather could trigger a flight alert based on that token and have the actual notification generated through the OTA's servers (FTE 2015c).

When such a solution is not possible, i.e. when analysis on data is required, the most common approach is anonymisation of the data, a process that eliminates all personal information, making the involved people impossible to identify (de-identification). Ensuring that all identifiable information has been removed is a challenging task and the ways to achieve it have been discussed in the ICARUS Deliverables D1.1 and D2.1. Anonymised data are entirely excluded from the GDPR, as they were never personal data to begin with, or have undergone such extensive technical anonymisation that no natural person is identifiable through them (Ramsay 2018). It is also worth mentioning that differential privacy and multi-party computation techniques are being explored to achieve privacy-preserving analysis on private data from various providers (Pettai and Laud 2015).

Data sensitivity is not limited to personal data, but may also refer to data that are critical for the company's strategy, therefore deciding which datasets to share is not a straightforward process. Finally, it should be highlighted that there are conflicts dependent on the type of the stakeholders involved in a potential data sharing agreement, as not all stakeholder types are legally allowed to exchange data.

### **Key Data Sharing Consideration II: Trust**

Although the data marketplace type being discussed applies the many-to-many paradigm, this does not preclude controlled membership. In fact, the aviation domain has very strong KYC (Know-Your-Customer) requirements, so for Industry stakeholders to trust such an initiative, participants should be limited to well-known organisations and businesses operating in the core aviation industry or have explicit and clear connection to it. The same is true for solutions based on blockchain, where approaches with controlled network membership also exist.

### **Key Data Sharing Consideration III: Security**

Cybersecurity has become an elevated risk for the aviation industry, as today's cyber adversaries are more persistent, trained, skilled and technologically savvy than before, according to Palo Alto reports

(PaloAlto 2017). The highly sensitive nature of flight systems, passenger data, increase of Wi-Fi connectivity, PCI-DSS compliance and IoT revolution could potentially introduce cyber security vulnerabilities that affect the business. The data breach at Cathay Pacific and Hong Kong Dragon Airlines, which exposed the sensitive personal and financial details of up to 9.4 million customers, is an alarming event in commercial aviation, forcing the review and update of airlines' data security practices, and their handling of events when they happen (CNBC 2018). In this environment, participation in any cloud data marketplace requires strong security guarantees.

Measures that can be taken mainly concern the implementation design and specification of the devised solution and include password management, firewalls and database access controls. Apart from infrastructure security, two important aspects that ensure higher security by design for the collaborating stakeholders, is (a) the inclusion in the system of carefully selected, in terms of content and granularity, and pre-processed in some cases data (e.g. striped of any sensitive personal information) and (b) the encryption of the data prior to leaving the company premises as discussed in D2.1, as well.

#### **Key Data Sharing Consideration IV: Data IPR, licensing and ownership**

Data licensing issues have been discussed earlier in the current section but also in D1.1. The legal implications pertain to data IPR frameworks, or lack thereof, and definition of ownership not only of data but also ownership of data-products. When such products combine data from different datasets under different licenses, compatibility checking may be difficult even as a manual process. Aspects like the algorithmic principles that should underpin the combination of data sharing agreements and the representation and ensured continuity of data provenance are decisive for successful data contracts and yet the relevant mechanisms are not yet developed (Koutroumpis, Leiponen, and Thomas 2017). Data IPR is a contentious issue that needs to be carefully defined and adjudicated. It may be difficult to do this using a blanket agreement, and there may arise needs for more tailored contracts and protections. Contract management thus needs to account for two contradicting forces: the requirement to homogenise the process under a common framework to increase efficiency and the demand for customised rights, terms and conditions and negotiation mechanisms.

#### **Key Data Sharing Consideration V: Technical Issues and Barriers**

The lack of a common data model, at least for the most commonly encountered data in aviation, is a significant technical barrier hindering the creation of a wider data sharing system. Data browsing, exploration and discoverability are among the principal functionalities that the envisioned system should provide. Otherwise, the process would lapse to ad-hoc inquiries for data availability and acquisition. But in order to implement the required mechanisms to handle the inherent data heterogeneity, an underlying model (as presented in the ICARUS Deliverable D2.1) needs to be created to help link data coming from different providers and in different formats, identify connections among data with similar content but different structures, etc. For example, airlines may have different

approaches to documenting a similar incident within their organisation, which in turn translates to different data structures and formats, but also different content. Since a common set of standard operating procedures and reporting has not been accomplished, a data and metadata model to facilitate the data sharing processes in the context of the system being discussed, will need to be developed and, for the ICARUS purposes, is presented in D2.1. It is worth mentioning that the availability of such a model could also boost the development of services, e.g. data analysis and visualisations, that take advantage of the facilitated data fusion.

The quality of the data is another significant aspect that will determine the stakeholders' incentives to participate. Data quality, especially in the Big Data spectrum, is difficult even to define, as it spans along numerous dimensions: accuracy, completeness, consistency, redundancy, readability, accessibility and trustworthiness (Batini et al. 2015). Hence, designing and implementing mechanisms to provide guarantees regarding the quality of the provided data that go beyond the accompanying textual descriptions and assurances is a challenging task.

Other technical considerations include the handling of high dimensionality data, the actual size of data, the encryption requirements etc.

## 5 ICARUS Data Policy and Assets Brokerage Framework

---

### 5.1 Purpose

The ICARUS Data Policy and Assets Brokerage Framework has a dual role, as revealed by its name:

- (I) To formalise all data attributes and qualities that affect, or are in any way relevant to, the ways in which data assets can be shared / traded and handled subsequently to their acquisition. This involves licenses, IPR, characterisation of sensitivity and privacy risk levels, but also more generic metadata regarding data content and structure. Here there is a clear link to the data model described in D2.1, as there are certain properties that although not directly related to the policy part, play a role in understanding relevant limitations and potential risks and benefits of a data asset being shared through the ICARUS system.
- (II) To enable the creation of structured, machine-processable data contracts for the aviation industry. This entails the expression of contractual terms pertaining to data trading agreements into an appropriate machine-processable language. Furthermore, the framework will foresee all possible interactions of stakeholders in aviation data sharing scenarios and will define the system's expected behaviour in this context.

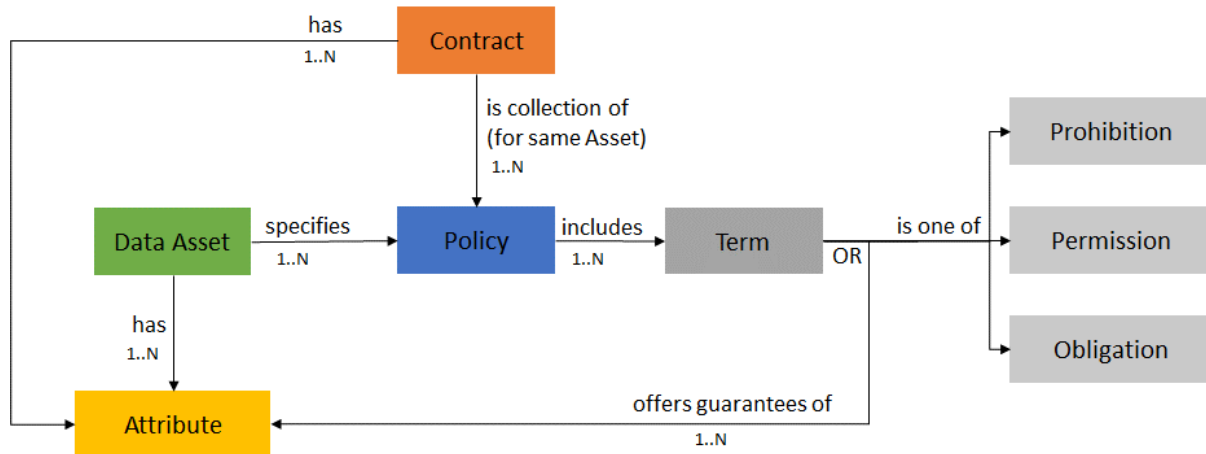
### 5.2 ICARUS Data Sharing Model

Work in the context of WP2 activities was performed with close collaboration to WP3 activities, as there are obvious interdependencies between them. Specifically the Data Policy and Business Brokerage Framework has clear links to the architecture decisions that need to be made, as it will drive the design and development of the project's data sharing mechanisms. Therefore, the current section builds on top of (and partially shows) some technical decisions that have been made and will be properly reported in D3.1. More specifically:

- Datasets provided by the ICARUS stakeholders will not leave their premises unencrypted, unless they comprise only public information. This has emerged as a clear stakeholder requirement also in the context of the WP1 MVP validation activities. A hybrid approach will be followed in this regard: Assuming tabular data (or data that can be easily structured as such, which is the case at least in the first version of the system according also to the collected requirements), most columns will be encrypted, leaving only specific columns that hold spatiotemporal information unencrypted. These columns will be used to enable efficient data browsing and selection without revealing any sensitive/proprietary information prior to an acquisition.
- ICARUS will adopt a DLT-based solution for the data brokerage, leveraging the benefits of smart contracts discussed in section 4.2.
- Data entry to the ICARUS platform will require the provision of certain metadata by the data provider.

In order to present how the proposed approach fulfils the dual role of the Framework, the section will first present the defined data and data sharing attributes that complement the Sharing Metadata of the ICARUS Metadata Schema presented in D2.1, and then the basic data acquisition workflows, showing how the data consumer interacts with the ICARUS engine and, through it, with the data provider.

Based on the initially identified requirements, the Framework will be built on top of three core entities, namely the Data Asset, the Policy and the Contract and two supporting entities, namely Attributes and Terms, with the latter being specified as one of Prohibition, Permission and Obligation:



**Figure 5-1: ICARUS Data Sharing Model - High-Level View**

As Data Asset, ICARUS defines a specific dataset from a data provider. A Data Asset, at least in the Framework's first version, corresponds to a single file which will either be already in or be easily formatted in a tabular form, i.e. comprising rows and columns. There is no separate entity to express the concept of DaaS, as these will be offered through sharing agreements that foresee updates and not through real-time data streams.

A Policy is the way all legal, IPR, license, quality etc. terms are expressed. Each Data Asset specifies a number of Policies which control how it can be shared and accessed. A Policy comprises a group of terms and/or attribute guarantees. Terms are specific prohibitions, permissions or obligations stemming from the above-mentioned aspects, whereas Attributes are expressions of certain facts and/or qualities, e.g. the date a Data Asset was created.

Finally, contracts represent the official data sharing agreements between a data provider and a data consumer in regard to one single Data Asset under specific Policies.

The following two tables provide high-level examples of Attribute and Term instantiations which have been designed in alignment with the broader ICARUS metadata schema presented in D2.1.

**Table 5-1: ICARUS Data Brokerage Framework: Attributes Definition**

Attribute Subject	Attribute
Contract	• temporal validity
	• spatial validity & coverage
	• validation date



Attribute Subject	Attribute
	<ul style="list-style-type: none"> <li>liability</li> </ul>
	<ul style="list-style-type: none"> <li>involved provider</li> </ul>
	<ul style="list-style-type: none"> <li>involved consumer</li> </ul>
	<ul style="list-style-type: none"> <li>termination clause</li> </ul>

Table 5-2: ICARUS Data Brokerage Framework: Terms Definition

Policy Category	Terms	Exemplary Definitions/ Values
Pricing	cost calculation scheme	per row   fixed for dataset   query dependent (number)
	amount	
	currency	euro   cryptocurrency units
	tax scheme	% over amount prepay   postpay   dynamic
Access	software	custom security -oriented software constraints
	hardware	custom security-oriented hardware constraints
	requestor's organisation	Athens International Airport
	requestor's organization type	Airport
	request origin country	Greece
	requestor's location	identification for geo-restricted access
	duration of retention	fixed start & end dates   interval
Responsibility	duration of use	fixed start & end dates   interval
	versioning & updates	included   included until   no
	ownership	publisher   creator   consumer
	addressed to	individual   group
Rights & Usage	liability & indemnification	custom clauses
	license & copyright notice	CC   custom   ...
	derivation	modify (Y N)   excerpt (Y N)
	attribution	annotate (Y N)   aggregate (Y N)
	reproduction	required
	distribution	allowed   prohibited
	target purpose	allowed   prohibited
	target industry	scientific usage
	online storage	limited to aviation (select from predefined)
Quality	re-context	not allowed
	accuracy	allowed
	completeness	%
	consistency	guaranteed for coverage X
Privacy & Protection	credibility	checked
	accessibility & online availability	estimated %
	privacy & sensitivity compliance	guaranteed
	liability	levels   disclaimers   guarantees
Privacy & Protection	applicable law	custom clause
		state conforming to regulations, laws

The presented terms and attributes, as well as the entities and relationships among them, correspond to the first definition of the ICARUS Data Policy Framework. As explained in Section 4, a balance needs to be achieved between expressivity and applicability / efficiency, therefore the initially adopted

model is simplified when compared to the complete set of considerations and options of data trading in aviation. It is believed though that the performed assumptions do not harm the future extension and refinement of the Framework. The final translation and mapping of the real-life contractual terms into the above defined concepts will certainly reveal challenges and, if needed, the above modelling will be revised accordingly and documented in D2.3.

### **5.3 ICARUS Data Sharing Workflow**

The second part of the ICARUS Data Policy and Assets Brokerage Framework is related to all data brokerage aspects, hence requires the definition of workflows that capture the basic provider-consumer interactions. The simplest version of a data trading workflow, which is assumed to be completed smoothly, is presented in three phases in Figure 5-2.

#### **Phase I: Data Assets Exploration**

The workflow is initiated by an ICARUS user performing a query to search for data. The query is constructed by selecting which fields should be available in the returned results (selection options based on the underlying predefined data model) and under which conditions. The conditions are defined as filters on the datasets' unencrypted fields that are used to specify (mainly) the spatiotemporal bounds of the query.

The system then performs the query and identifies the matching data from the ICARUS database. These are either individual datasets that match the whole query or appropriate combinations of datasets. Dataset policies are checked to ensure that the requesting user is eligible to purchase the datasets in the results. As an indicative example, it may be in a dataset's policy that a user representing an airline cannot have access to the data under any circumstances. After this filtering is performed, the user is presented with the query results as a list of alternatives from which to choose. Whether data excerpts/samples will be included in the presented options or only the datasets' metadata and pricing schemes, is a technical choice that does not affect the overall Framework.

#### **Phase II: Smart Contract Drafting**

Assuming the easiest possible choice in each step of this basic workflow, the user who performed the query then explores the results, reviews the terms and pricing scheme options and selects one dataset that matches the query. At this point, a request to purchase is issued and the data provider is notified. Again assuming the smoothest flow, the data provider approves the request and, through the system, creates a smart contract that defines the dataset policies and the trading terms. The smart contract is in the ledger, but not yet validated.

#### **Phase III: Smart Contract Validation**

The requesting user, i.e. from this point on the data consumer, reviews the contract and accepts all defined policies. The status of the contract in the ledger changes to reflect that it has been accepted by both parties and is only pending the respective payment in order to become effective. The data consumer proceeds to pay and when the system is notified, the smart contract changes status to valid. It should be clarified that the validation date of the smart contract may differ from the date that the contract specifies that the data will be available.

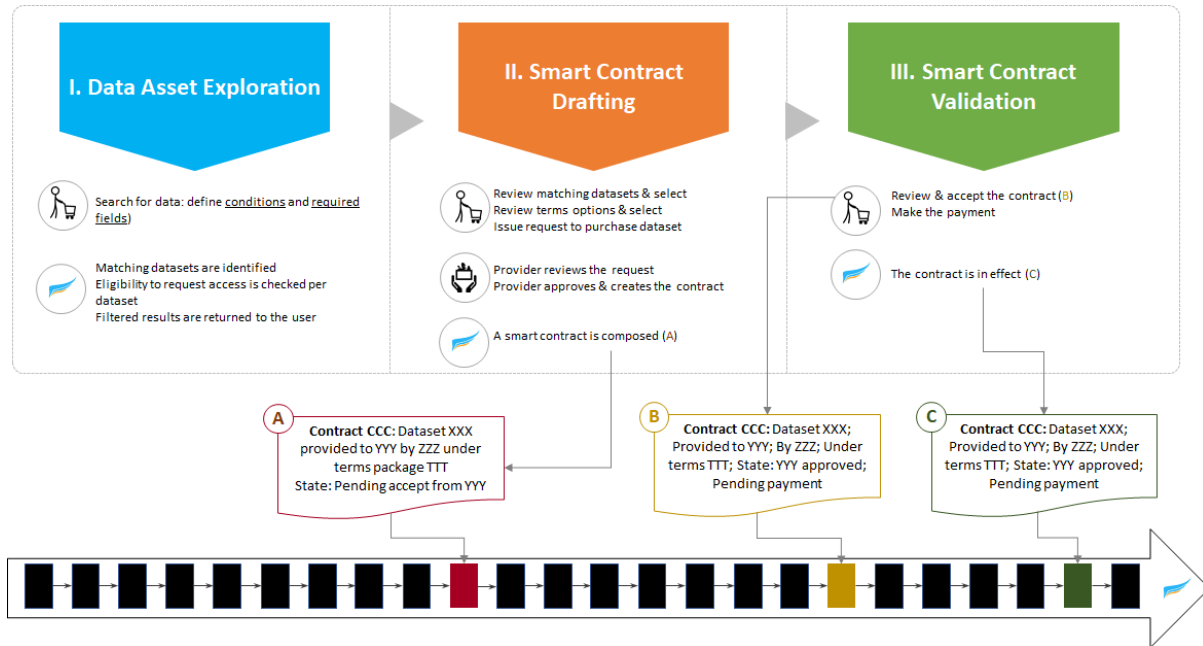


Figure 5-2: Basic Data Trading Workflow

There are also certain well anticipated alternatives, or deviations, to the above workflow, discussed briefly below.

**Deviation 1:** The data provider decides not to provide the dataset to the requesting user.

ICARUS will not enforce any automated data trading and allows the data owners to be in full control of their data. If, for any reason, a data provider does not wish to proceed to any transaction, or even negotiation of such, with a requesting user for a specific dataset (or in general), then the request is dropped, and the requesting user is notified accordingly.

**Deviation 2:** The requesting user is interested in purchasing more than one datasets through a single query.

This is actually expected to be a very common workflow in ICARUS, as the querying mechanism will enable the selection of fields, irrespectively of whether these are found in the same dataset, hence it is very likely that users will search for data that need to be gathered from different datasets. When a user chooses a query result that combines multiple datasets, then for each of them a separate request is issued and the process spawns parallel processes, equal in number with the number of different datasets. Each of the processes starts with a data provider being notified of the request and finishes

with the smart contract terms being accepted by both the consumer and the provider, after which the payments are performed. The number of created smart contracts matches the number of datasets and each process is independently writing on the ledger. It should be stressed here that it may be the case that the requesting user wants to purchase all datasets included in the returned result or none. The definition of such restrictions, if needed, is foreseen by the Framework and their realization is not expected to pose any technical challenges.

**Deviation 3:** The data consumer and the data provider do not agree in the first definition of the smart contract's terms and want to negotiate.

Intuitively, this is also expected to be a frequent requirement when forming data trading agreements. Although some dataset policies will not be negotiable, other terms related to the pricing scheme for example may be open for discussion. Whether this will be performed through a flexible smart contract updating mechanism available to negotiating parties, or done externally to the ledger through the main ICARUS platform, is a technical decision that will be examined and reported in the context of WP3.

## 5.4 Assets Brokerage Challenges

The ICARUS policy and brokerage framework will set the foundations of the ICARUS platform that will link data providers and data consumers at all levels of the data value chain in the aviation industry, through secure and trustworthy data trading agreements. In order to achieve this, it is first of all crucial to incentivise stakeholder participation, which entails (a) eliminating scepticism regarding trust, security and confidentiality, (b) facilitating data trading and reducing the required effort, both time- and resource-wise, and (c) ensuring that all potentially worrying aspects, from a technical, legal, and conceptual perspective, have been at least foreseen and, to the extent possible, resolved. The current section presents some of the challenges that have already been identified by the consortium and constitute drivers of further work to be performed.

### **Asset Brokerage Challenge I: Lack of commonly agreed upon data and metadata model**

Data assets and data and metadata attributes should be defined in a way that is both flexible but also meaningful for the domain so that users can intuitively browse and query the available datasets in order to identify which ones, or which subsets and combinations, they are interested in. There is one IATA initiative currently underway<sup>33</sup>, however it is both a work in progress (hence cannot be fully leveraged at this point), but also too broad for the data sharing context that ICARUS operates in. The data model should be expressive enough to accommodate all informational needs of the platform to ensure robustness and integrity, intuitive enough in the way metadata purposed for data browsing are structured and, at the same time, scalable (in terms of how it can be queried and reasoned upon) and extensible (as the closed world assumption cannot be applied here).

---

<sup>33</sup> <https://www.iata.org/whatwedo/passenger/Pages/industry-data-model.aspx>

Although ensuring the conformance to one common model will be a time-consuming and difficult process, it is an indisputable requirement of a data sharing system. Beyond the increased data exploration and discoverability, provision of high-quality enhanced data, i.e. represented in a manner that facilitates data integration and visualization, has emerged as a clear market need (Stahl, Schomm, and Vossen 2018).

Work is already being performed to identify the data assets relevant to ICARUS, their structure, types, characteristics and modelling requirements in terms of content and metadata spanning from generic descriptive attributes to rights, licenses, ownership etc, i.e. regarding all aspects described in Section 5.2 and potentially some additional from Section 4, which could be deemed meaningful to include, considering also the overall approach and the results presented in D2.1.

### **Asset Brokerage Challenge II: Data Policy Language**

A robust framework for smart contract creation needs to address both the data heterogeneity (not only in terms of content as discussed above, but also regarding metadata that affect the way they can be handled in a data trading process) and the required expressivity and formality of a binding data sharing agreement that imposes specific terms for IPR, pricing, quality, legal issues etc. In order to do this, an appropriate language that expresses and formalises such attributes is required. Literature shows that the language decision involves finding the right balance in the trade-off between generic usage for all data types and modalities on one hand, and expressive price and IPR functions, rich metadata models and detailed data sharing contracts on the other hand. Essentially this decision also involves a choice between reusing an existing language or designing a custom metadata schema (Sakr 2018).

There are several Rights Expression Languages (RELs), i.e. machine interpretable languages that convey the rights and restrictions associated with a particular asset, that are used in this context, either in metadata definitions or as part of Digital Rights Management (DRM) systems (García and Mercé 2005).

The Creative Commons Rights Expression Language (ccREL (W3C 2008) ) is the standard recommended by Creative Commons for machine-readable expression of copyright licensing terms and related information but does not scale to CC+ scenarios (Alton 2009) and it is not concerned with access and usage control (García and Mercé 2005). ODRL is an XML vocabulary to express rights, rules, and conditions including permissions, prohibitions, obligations and assertions (W3C 2018a, 2018b). Motion Picture Experts Group, Standard 21, Part 5 (MPEG-21/5) REL provides industry with the means to create expressions that can then be associated with audio/video and other content to express what the consumer can do with that content (Trondheim 2003). Picture Licensing Universal System (PLUS) is a rights language developed by the PLUS Coalition in order to simplify licensing images. RightsML (IPTC n.d.) builds on ODRL, extending and refining it, to meet the specific needs of the media industry (García and Mercé 2005). Adobe Content Manager is a fully implemented REL, used only in the Adobe Reader product for protected files, which has a small vocabulary but covers the basics of printing, copying, lending, and text-to-speech (Hess 2016).

RDFa (W3C 2010) also provides attributes to define the license and creator information of online resources, e.g. documents or images, found in websites and the SKOS vocabulary includes a concept for rights statement. (García and Mercé 2005) have developed a copyright ontology, CopyrightOnto, geared towards the development of copyright-aware DRM systems. XACML, a well-known access control policy language encoded in XML enables the use of arbitrary attributes in different types of policies, including privacy policies, and has therefore been used in this context (Yang, Li, and Niu 2015). Finally, as DaaS services are emerging, rights expression languages to address their complexities are required and recently work is being performed in this direction as well, e.g. through ODRL-S (Gangadharan et al. 2007) which extends ODRL by implementing clauses of service licensing.

RELs were not examined in Section 4, as literature review there did not highlight them as a significant factor in the design phases of a data sharing framework, in general or in aviation. The language selection is not a decisive aspect when collecting the requirements of the framework, i.e. when the legal terms, the IPR attributes, the ownership definition and other metadata properties to be included. It is, however, a significant step that follows these specifications in the future work, as the language will need to provide the formalisations and also the appropriate expressivity, flexibility and stringiness to ensure the framework's robustness and the stakeholders' ability to use it and to trust its mechanisms.

#### **Asset Brokerage Challenge III: Data Pricing**

The envisioned ICARUS approach regarding data pricing is very flexible, foreseeing both predetermined prices (per asset, per line etc.) but also enforcement of custom prices which may depend on the provision of updates for pre-paid assets for specific time periods. The data pricing and payment options will be primarily designed and developed to address the stakeholders' requirements, however there are additional aspects to examine, such as the ICARUS commission for each data contract. Furthermore, the ICARUS model for pricing policies should foresee the need to extend and customise certain features in the future, if needed. Therefore, an intuitive, flexible and, of-course, fair to the members, strategy need to be devised in collaboration with the ICARUS exploitation activities in WP7.

The decision of the payment system to be applied, specifically whether it will be based on money or virtual currency and what are the advantages and disadvantages in each approach, is also relevant in this context.

#### **Asset Brokerage Challenge IV: Provision means of data**

The way the purchased data will be served to the users is another challenging aspect of the process, especially considering that some datasets may be provided under agreements that include updates. Datasets that are updated on a regular basis, e.g. information on daily flights, are very common in the aviation industry and the way such data should be delivered is more than a technical decision, it is at the core of the business brokerage framework. Furthermore, apart from the option to download,

offering pre-configured cloud computing environments where data can be directly used or feeding data via APIs to be integrated into customer enterprise applications are options that need to be considered (Sakr 2018), so as to succeed in innovating the complete data sharing value chain in aviation.

#### **Asset Brokerage Challenge V: License compatibility analysis**

As also established by the extensive literature study, data licensing and IPR management is a complex issue and, depending on the granularity level and the scope in which it is examined, it imposes several challenges, both technical and theoretical/ legal. Data assets provided through the envisioned ICARUS platform cannot be expected to share the same licenses, restrictions and permissions and therefore conflicts may arise when a data consumer wishes to combine datasets in order to create, and potentially publish, data-products (e.g. new enhanced datasets, analysis results etc). Situations where license incompatibilities may forbid the realisation of a data consumer's plans are numerous, making copyright clearance a challenging problem. Devising mapping and compatibility rules (Villata and Gandon 2012) is not a straightforward process and a balance will need to be found between solutions that require pure manual checking by the users (which could prove demotivating) and automated compliance verification, which could prove difficult for the domain stakeholders to trust. It needs to be noted that the literature study performed in the context of this deliverable did not find any such production-ready solutions anyway.

**Liability** and **accountability** are also relevant here as, apart from the license checking, ensuring that all parties involved in a data trading agreement will honour its terms is an almost impossible task. To this end, concrete legal agreements should be used to detail the responsibilities of all parties involved to underpin ethical commitments in data transfer, storage and sharing.

#### **Asset Brokerage Challenge VI: Legal standing of smart contracts**

##### **Enforceability**

Smart contracts are a recent advancement and therefore enforceability of their legal code has not yet been affirmed by legal authorities (Wright 2016). Concerns have been raised regarding smart contracts' inability to handle ambiguity and the difficulty for programmers to plan for every possible contingency (Glidden 2018). Translating all real-world data needed to adjudicate these situations cannot be realistically expected, but at least certain aspects relevant to the formation and enforcement of legal contracts via smart contracts will need to be clearly defined. (Raskin 2016) state that innovative technology does not necessitate innovative jurisprudence, and traditional legal analysis can help craft simple rules as a framework for this complex phenomenon (Raskin 2016). A white paper on smart contract enforceability by the Smart Contracts Alliance Initiative of the Chamber of Digital Commerce, a trade association representing the digital asset and blockchain industry, provides an extensive study of the various complications that may arise in the context of blockchains (e.g. identification of parties, modification of terms, choice of law, enforcement, liability, etc.) , the enforceability of smart contracts under existing law and also examines some issues surrounding the



perfection of a security issue in blockchain (Chamber of Digital Commerce 2018). It becomes thus evident that work in the area is progressing as the interest around smart contracts rises, along with their usage.

Until the legal aspects of adopting smart contracts in data sharing agreements are clarified, risk averse parties may be reluctant to enter such agreements, therefore this is a crucial challenge to be addressed. It should be stressed, however, that ICARUS does not propose a fully automated process of smart contract creation, negotiation and validation; instead, human intervention is always required, from both the seller and the buyer parties.

#### Compliance with GDPR

The blockchain has inherent contradictions to certain GDPR principles, such as rectification and removal, but it also strongly conforms with the technical data protection principles according to the GDPR, as it has proven structure security. The biggest conflict between the blockchain and the GDPR is the blockchain's immutability: Although its biggest strengths originate from this immutability and having immutable objects is in line with some of the GDPR's purposes (integrity, security and transparency), it results in the data subject losing the retroactive control over their personal data. The GDPR assesses these principles as absolute but does not discuss if alternative usage would provide the most security for the individual (Ramsay 2018).

The extent to which the GDPR is applicable in combination with a DLT-based solution will be further assessed in the course of the project, as the ICARUS solution will gradually become more tangible and the exact aspects to examine will be identified. However, this already seems to be achievable, based on current literature findings and taking into account that the ICARUS platform will not be required to handle personal data, but de-identified, anonymized data or non-personal data according to the preliminary data assets collection documented in D1.1.

#### Security

Security analysis is becoming a significant area in smart contract development environments, expected to bolster their adoption as they will increase stakeholders' trust in them. The adoption of an appropriate strategy is therefore an important aspect that needs to be studied. Existing solutions can be potentially, e.g. the Mythrill Platform<sup>34</sup> that aims to raise the baseline security level of all smart contracts deployed on the Ethereum blockchain.

The aforementioned points do not cover all the challenging aspects identified, but present the ones considered to pose major considerations regarding the first steps in refining and realising the ICARUS policy and brokerage framework. Other aspects, like applying conditions related to the cancellation of a contract (e.g. the allowed time interval between acceptance of terms and payment), clarifying the technical implications of the hybrid data encryption approach and providing appropriate contract negotiation mechanisms, will be also examined and taken into consideration in the subsequent design and implementation phases.

---

<sup>34</sup> <https://mythrill.ai/>

## 6 Conclusions & Next Steps

---

The present deliverable documents the produced results of the activities performed in the first iteration of Task T2.3 “Deep Learning and Prescriptive Analytics Algorithms” and Task T2.4 “Data Policy and Assets Brokerage Frameworks”. In this regard, the deliverable is structured on two distinct axes, each linked to one of the two aforementioned tasks, as explained in the methodological approach of the deliverable in Section 1.2.

The first axis, namely the “Data Analytics” axis, involves the work presented in sections 2 and 3 of the current report and is related with the definition and design of all the intelligence generation functions to be integrated in the ICARUS platform, as defined by the T2.3 objectives. In this context, an in-depth review of the data analysis state-of-play, from a scientific and business perspective, was effectively performed. The review commenced with the collection and study of the most recent advancements in machine learning, from 2014 onwards, and their contribution in modern big data analytics and applications, a process that helped outline the full potential of employing data analytics techniques. Subsequently, scientific literature surveys on machine learning for data analytics targeting specifically the aviation domain were studied in order to understand the current status of data analytics in the evolving aviation landscape. Having set the theoretical foundations for a more in-depth analysis, the next step was to discover, collect and investigate additional, more specific, methods and aviation applications. These were grouped under the four key elements that move through airports, as identified by the IATA NEXTT initiative: aircrafts, passengers, baggage and cargo to help maintain the business perspective as well whilst performing the state of play analysis. For each of these four elements, their complete “journey” cycle was studied and real-life use cases, problems and data analytics solutions were collected and studied. For each of the identified scientific approaches, detailed information was initially extracted (e.g. data requirements, relation to ICARUS, availability of data and/or algorithms used etc.) and then combined to produce insightful reports regarding the major issues in each element’s journey, as well as commonly employed data analytics algorithms to address them, along with potentially relevant data and data sources.

As a next step, the technical and algorithmic perspectives of data analysis were examined in detail. Inherent data characteristics in the modern aviation landscape, including heterogeneity and noise, render data analysis a complex, multi-step process. Therefore, five distinct steps were defined and presented that comprise the ICARUS data analysis approach in line with the ICARUS Deliverable D1.2 “The ICARUS Methodology and MVP”: (I) Data Ingestion, (II) Data Cleansing and Transformation, (III) Dimensionality Reduction, (IV) Data Analysis and (V) Visualisation. The current deliverable placed its focus mainly in the fourth step to help identify the most suitable algorithms for knowledge extraction, business intelligence and analytics in ICARUS. Towards this goal, three criteria were defined for the algorithm selection process: a) to adhere to the ICARUS platform requirements and be applicable to aviation specific tasks, b) to have proven their ability and robustness in the research community through the years, and c) to have been implemented in a commonly used software framework or library. For the last criterion, 12 popular software frameworks and libraries have been considered. The selected algorithms were then grouped under three categories, based on their applicability in

Descriptive, Predictive or Prescriptive Analytics, and were presented along 3 axes: Basic Analytics, Machine Learning and Deep Learning. In total 9 algorithms were studied and described in detail for Basic Analytics, 17 for Machine Learning and 5 for Deep Learning. For each of the algorithms, a predefined template was filled to extract specific information, including well known variations and specific application examples of the algorithm in the aviation domain. As a final step of this analysis, concrete examples of use cases that map some very early demonstrator scenarios to the aforementioned algorithms were presented in order to provide early insights into the ICARUS stakeholders' perspective in the data analysis process. For each scenario, a number of indicative use case examples is provided, in an attempt to match the scenario objectives with computational tasks commonly encountered in literature of the aviation domain analytics. A group of algorithms is then proposed, according to each task, that can cope with it in the most efficient way.

The second axis of the current deliverable is the “Data Sharing” axis. It involves the work presented in sections 4 and 5 of the current report and is related with the activities towards delivery of the ICARUS Data Policy and Assets Brokerage Framework that will be used to facilitate the data sharing and trading features that will be offered by the platform to link data providers and data consumers at all levels of the data value chain, as defined by the T2.4 objectives.

In this regard, the first step was to study and understand the underlying state-of-play regarding data sharing in general, and in aviation in particular, from a research / academic and market perspective. An in-depth landscape analysis of data sharing practices, incentives and challenges was performed based on an extensive literature review. First, the broader spectrum of data sharing practices is studied, starting from the open data paradigm and moving on to non-open data sharing/ trading approaches and considerations. In this regard, several attributes of data and data sharing agreements were identified and discussed, including IPR, licenses, pricing, trust and security, privacy and protection, regulatory compliance, legal responsibility, data quality, ownership and data access rules. Furthermore, frameworks and initiatives that attempt to model the data sharing barriers and drivers and address their challenges were explored. It was shown the motivation behind many such studies and initiatives largely stems from the need to create data contracts in non-textual, machine-processable form, which could serve as automation enablers for data marketplaces. Data marketplaces were also studied and the reasons hindering their adoption were examined. The emergence of Distributed Ledger Technologies as enablers for innovative data marketplaces was studied and several examples were discussed to showcase how DLTs can enforce transparency, rigorous provenance and data democratisation. Then, the aviation specific challenges and potential benefits of data sharing were identified, first through collecting and studying the major data sharing initiatives in aviation and then through reporting on the data sharing practices of the project's industry partners so as to get insights into the status quo and the inner workings of data sharing agreements in the ICARUS demonstrators, which also served as representative cases for the aviation domain in general.

Finally, having concluded the detailed landscape analysis of data sharing, the first version of the ICARUS Data Policy and Assets Brokerage Framework was defined. The Framework was based on top

of three core entities, namely the Data Asset, the Policy and the Contract and two supporting entities, namely Attributes and Terms, with the latter being specified as one of Prohibition, Permission and Obligation. Concrete initial term and attribute examples are provided that will be used for the formalisation of all data features and qualities that affect, or are in any way relevant to, the ways in which data assets can be shared / traded and handled subsequently to their acquisition. The foreseen policies involve licenses, IPR, characterisation of sensitivity and privacy risk levels, but also more generic metadata regarding data content and structure. One of the main defined goals of the ICARUS Data Policy and Assets Brokerage Framework is to enable the creation of structured, machine-processable data contracts for the aviation industry. Towards this end, workflows that capture the basic provider-consumer interactions were defined, which show how ICARUS envisions to enable the creation of structured, machine-processable data contracts for the aviation industry, whilst maintaining the data owner in control of the provided data. The simplest version of a data trading workflow, as foreseen by the Framework, comprises three distinct phases: data assets exploration, smart contract drafting and smart contract validation, which were described in detail. Furthermore, possible deviations from the basic workflow were identified and addressed.

As the final goal of the Framework is to link data providers and data consumers at all levels of the data value chain in the aviation industry, through secure and trustworthy data trading agreements, the final step along the “Data Sharing” axis was to identify the challenges in its realisation and adoption. Lack of commonly agreed upon data and metadata models, selections of an appropriate data policy language, data pricing models, selection of provision means of data, license compatibility analysis , accountability and liability, legal standing of smart contracts, including enforceability, GDPR compliance and security, are the main identified challenges and will serve as input to subsequent activities in the current Work Package as well as all other related Work Packages in order to help steer subsequent work towards successfully addressing them.

Since D2.2 was prepared in parallel with D2.1 and D3.1 to ensure the results’ consistency, in the forthcoming steps, the outcomes of D2.2 will feed the detailed specification and design tasks in WP3 and will also be leveraged by the ongoing activities in WP2. Further feedback gathered by aviation stakeholders (internally in the consortium and through the MVP external validation activities) will be addressed in the final version of the “Intuitive Analytics Algorithms and Data Policy Framework” that will be documented in D2.3 “Updated ICARUS Data Management, Analytics and Data Policy Methods”, that will be released on M18 of the ICARUS project implementation.

## Annex I: References

---

- AEGIS. 2017. *D2.1 Semantic Representations and Data Policy and Business Mediator Conventions*. AEGIS-D2.1-Semantic-Representations-and-Data-Policy-and-Business-Mediator-Conventions-v1.0.pdf.
- Ahmed, Md et al. 2018. "A Cooperative Co-Evolutionary Optimisation Model for Best-Fit Aircraft Sequence and Feasible Runway Configuration in a Multi-Runway Airport." *Aerospace*.
- Ahmed, Tanvir, Toon Calders, and Torben Bach Pedersen. 2015. "Mining Risk Factors in RFID Baggage Tracking Data." In *Proceedings - IEEE International Conference on Mobile Data Management*,.
- Aitken, Mhairi et al. 2016. "Public Responses to the Sharing and Linkage of Health Data for Research Purposes: A Systematic Review and Thematic Synthesis of Qualitative Studies." *BMC Medical Ethics*.
- Akpınar, Musab Talha, and Muhammed Emin Karabacak. 2017. "Data Mining Applications in Civil Aviation Sector: State-of-Art Review." In *CEUR Workshop Proceedings*,.
- Alba, Sergio Ortega, and Mario Manana. 2016. "Energy Research in Airports: A Review." *Energies*.
- Alton, Roland. 2009. "Semantic Copyright Expression Options." *SlideShare*. <https://www.slideshare.net/rasos/semantic-copyright-expression-options>.
- Ayhan, Samet, Pablo Costas, and Hanan Samet. 2018. "Predicting Estimated Time of Arrival for Commercial Flights." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*,.
- Balcombe, Kelvin, Iain Fraser, and Liam Harris. 2009. "Consumer Willingness to Pay for In-Flight Service and Comfort Levels: A Choice Experiment." *Journal of Air Transport Management*.
- Balint, Florin Bogdan, and Hong Linh Truong. 2017. "On Supporting Contract-Aware IoT Dataspace Services." In *Proceedings - 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2017*,.
- Bao, Yukun, Tao Xiong, and Zhongyi Hu. 2012. "Forecasting Air Passenger Traffic by Support Vector Machines with Ensemble Empirical Mode Decomposition and Slope-Based Method." *Discrete Dynamics in Nature and Society*.
- Bar-Sinai, Michael, Latanya Sweeney, and Merce Crosas. 2016. "DataTags, Data Handling Policy Spaces and the Tags Language." In *2016 IEEE Security and Privacy Workshops (SPW)*,.
- Batini, Carlo, Anisa Rula, Monica Scannapieco, and Gianluigi Viscusi. 2015. "From Data Quality to Big Data Quality." *Journal of Database Management*.
- Baxter, G, and P Srisaeng. 2018. "The Use of an Artificial Neural Network to Predict Australia's Export Air Cargo Demand." *International Journal for Traffic and Transport Engineering* 8(1): 15–30. [http://ijtte.com/uploads/2018-02-11/b2ddbaec-be4b-b57fijtte.2018.8\(1\)02.pdf](http://ijtte.com/uploads/2018-02-11/b2ddbaec-be4b-b57fijtte.2018.8(1)02.pdf) (December 20, 2018).
- Bertagnolli, Monica M. et al. 2017. "Advantages of a Truly Open-Access Data-Sharing Model." *New England Journal of Medicine*.
- Bhaskaran, Kumar et al. 2018. "Double-Blind Consent-Driven Data Sharing on Blockchain." In *Proceedings - 2018 IEEE International Conference on Cloud Engineering, IC2E 2018*,.
- Biryukov, Alex, Dmitry Khovratovich, and Sergei Tikhomirov. 2018. "Privacy-Preserving KYC on Ethereum." In *Proceedings of the 1st ERCIM Blockchain Workshop 2018, Reports of the European Society for Socially Embedded Technologies*,.
- Bolat, A. 2001. "Models and a Genetic Algorithm for Static Aircraft-Gate Assignment Problem." *Journal of the Operational Research Society*.
- Bouras, Abdelghani, Mageed A. Ghaleb, Umar S. Suryahatmaja, and Ahmed M. Salem. 2014. "The Airport Gate Assignment Problem: A Survey." *Scientific World Journal*.
- Brownlee, Alexander E.I. et al. 2018. "A Fuzzy Approach to Addressing Uncertainty in Airport Ground Movement Optimisation." *Transportation Research Part C: Emerging Technologies*.
- Canino-Rodríguez, José M. et al. 2015. "Human Computer Interactions in Next-Generation of Aircraft

- Smart Navigation Management Systems: Task Analysis and Architecture under an Agent-Oriented Methodological Approach." *Sensors (Switzerland)*.
- Cao, T. et al. 2016. "MARSA: A Marketplace for Realtime Human Sensing Data." *ACM Trans. Internet Technol.*
- Capgemini. 2014. "Impcat of Big Data on Analytics." *SlideShare*: 6. [https://www.slideshare.net/slideshow/embed\\_code/36866068?utm\\_content=buffer701ec&utm\\_medium=social&utm\\_source=twitter.com&utm\\_campaign=buffer](https://www.slideshare.net/slideshow/embed_code/36866068?utm_content=buffer701ec&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer).
- Cardell-Oliver, Rachel et al. 2017. "Travel Behaviour Patterns – Micro Analysis." <https://research-repository.uwa.edu.au/en/publications/travel-behaviour-patterns-micro-analysis> (December 20, 2018).
- Carnelley, P et al. 2016. *Europe's Data Marketplaces - Current Status and Future Perspectives*.
- Chamber of Digital Commerce, Smart Contracts Alliance. 2018. *Smart Contracts – Is the Law Ready*.
- Chen, Shu Chuan, Shih Yao Kuo, Kuo Wei Chang, and Yi Ting Wang. 2012. "Improving the Forecasting Accuracy of Air Passenger and Air Cargo Demand: The Application of Back-Propagation Neural Networks." *Transportation Planning and Technology*.
- Chen, Sien et al. 2015. "Understanding Airline Passenger Behavior through PNR, SOW and Webtrends Data Analysis." In *Proceedings - 2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015*.
- Cheng, Shaowu, Qian Gao, and Yapping Zhang. 2015. "Comparative Study on Forecasting Method of Departure Flight Baggage Demand." In *2014 IEEE Chinese Guidance, Navigation and Control Conference, CGNCC 2014*.
- Chiti, Francesco, Romano Fantacci, and Andrea Rizzo. 2018. "An Integrated Software Platform for Airport Queues Prediction with Application to Resources Management." *Journal of Air Transport Management*.
- Chou, Tsung Yu, Gin Shuh Liang, and Tzeu Chen Han. 2013. "Application of Fuzzy Regression on Air Cargo Volume Forecast." *Quality and Quantity*.
- CITA. 2018. *Baggage Report 2018*. <https://www.sita.aero/resources/type/surveys-reports/baggage-report-2018>.
- Cleveland, William S. 2014. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *Statistical Analysis and Data Mining*.
- CNBC. 2018. "Hong Kong Privacy Watchdog to Investigate Cathay Pacific over Massive Data Breach." <https://www.cnn.com/2018/11/06/hong-kong-watchdog-to-investigate-cathay-pacific-over-data-breach.html>.
- Comendador, Victor Fernando Gomez, Rosa María Arnaldo Valdés, and Álvaro Rodríguez Sanz. 2016. "Reduction of Uncertainty Propagation in the Airport Operations Network | Request PDF." In *XII Congreso de Ingeniería Del Transporte*, Valencia. [https://www.researchgate.net/publication/312523969\\_Reduction\\_of\\_uncertainty\\_propagation\\_in\\_the\\_Airport\\_Operations\\_Network](https://www.researchgate.net/publication/312523969_Reduction_of_uncertainty_propagation_in_the_Airport_Operations_Network) (December 21, 2018).
- DataScience.aero. 2018. "Aviation Data Exchange Programmes Worldwide."
- Deka, Ganesh Chandra. 2016. "Big Data Predictive and Prescriptive Analytics." In *Handbook of Research on Cloud Infrastructures for Big Data Analytics*, IGI Global, 30–55.
- Elena, Hawk. 2018. "Dynamic Airline In-Flight Entertainment Systems Using Predictive Analysis." Hochschule Anhalt.
- FAA. 2012. "Performance Success Stories - Data Sharing Helps Airlines Reduce Delays." <https://www.faa.gov/nextgen/snapshots/stories/?slide=9>.
- Farahani, Reza Zanjirani, Masoud Hekmatfar, Alireza Boloori Arabani, and Ehsan Nikbakhsh. 2013. "Hub Location Problems: A Review of Models, Classification, Solution Techniques, and Applications." *Computers and Industrial Engineering*.
- Feng, Bo, Yanzhi Li, and Zuo Jun Max Shen. 2015. "Air Cargo Operations: Literature Review and Comparison with Practices." *Transportation Research Part C: Emerging Technologies*.



- Finavia. 2018. "Finavia Personal Data Collection Notification." <https://www.finavia.fi/en/newsroom/2018/airport-data-privacy-what-kinds-personal-data-does-finavia-collect-and-how-it-used>.
- FTE. 2015a. "Data Sharing Transformation Driving Kerb to Gate Improvements at Dublin Airport." <https://www.futuretravelexperience.com/2015/04/data-sharing-transformation-driving-kerb-gate-improvements-dublin-airport/>.
- . 2015b. "EasyJet and Gatwick Launch Mobile Host to Simplify Airport Experience." <https://www.futuretravelexperience.com/2015/04/easyjet-and-gatwick-launch-mobile-host-for-iphone/>.
- . 2015c. "IATA Should Lead Efforts to Develop Travel Industry-Wide Data Sharing Standards Concludes FTE Collaboration Forum." <https://www.futuretravelexperience.com/2015/10/iata-should-lead-efforts-to-develop-travel-industry-wide-data-sharing-standards/>.
- GAIN. 2003. *Guide to Automated Airline Safety Information Sharing Systems*.
- Gangadharan, G. et al. 2007. "Service License Composition and Compatibility Analysis." In *Service-Oriented Computing - ICSOC 2007, Fifth International Conference, Vienna, Austria, September 17-20, 2007*.
- García, Roberto, and Jaime Delgado Mercé. 2005. Department of Technologies "A Semantic Web Approach to Digital Rights Management."
- Glenn, Baxter, Srisaeng Panarat, and Wild Graham. 2018. "An Assessment of Airport Sustainability, Part 2—Energy Management at Copenhagen Airport." *Resources* 7(2): 32.
- Glidden, Andrew. 2018. "Should Smart Contracts Be Legally-Enforceable?" *Blockchain at Berkeley*. <https://blockchainatberkeley.blog/should-smart-contracts-be-legally-enforceable-599b69f73aea>.
- Gössling, Stefan, Frank Fichert, Peter Forsyth, and Hans Martin Niemeier. 2017. "Subsidies in Aviation." *Sustainability (Switzerland)*.
- Grabus, Sam, and Jane Greenberg. 2017. "Toward a Metadata Framework for Sharing Sensitive and Closed Data: An Analysis of Data Sharing Agreement Attributes." In Springer, Cham, 300–311. [http://link.springer.com/10.1007/978-3-319-70863-8\\_29](http://link.springer.com/10.1007/978-3-319-70863-8_29) (December 20, 2018).
- Gu, Yu, and Christopher A. Chung. 1999. "Genetic Algorithm Approach to Aircraft Gate Reassignment Problem." *Journal of Transportation Engineering*.
- Guo, Xiaojia, Yael Grushka-Cockayne, and Bert De Reyck. 2018. "Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning." *SSRN Electronic Journal*. <https://www.ssrn.com/abstract=3245609> (December 20, 2018).
- Han, Jiawei., Micheline. Kamber, and Jian. Pei. 2011. *Data Mining : Concepts and Techniques*. Elsevier Science.
- Hancerliogullari, Gulsah. 2013. ProQuest Dissertations and Theses "Approximate Algorithms for the Combined Arrival-Departure Aircraft Sequencing and Reactive Scheduling Problems on Multiple Runways."
- HBR. 2006. "Strategies for Two-Sided Markets." <https://hbr.org/2006/10/strategies-for-two-sided-markets>.
- Heckman, Judd Randolph et al. 2015. "A Pricing Model for Data Markets." <https://www.ideals.illinois.edu/handle/2142/73449> (December 20, 2018).
- Hess, Demian. 2016. "Introducing Four Rights Expression Languages: CcREL, PLUS, MPEG 21/5 and ODRL." *LinkedIn*.
- Hu, X. B., and E. Di Paolo. 2007. "An Efficient Genetic Algorithm with Uniform Crossover for the Multi-Objective Airport Gate Assignment Problem." In *2007 IEEE Congress on Evolutionary Computation, CEC 2007*.
- IATA. 2010. *Guidelines on Passenger Name Record (PNR) Data*. [https://www.iata.org/iata/passenger-data-toolkit/assets/doc\\_library/04-pnr/New Doc 9944 1st Edition PNR.pdf](https://www.iata.org/iata/passenger-data-toolkit/assets/doc_library/04-pnr/New Doc 9944 1st Edition PNR.pdf).
- . 2016. "Passenger Data Must Be Transferable and Secure."



- <https://airlines.iata.org/analysis/passenger-data-must-be-transferable-and-secure>.
- . 2018. “Data Protection and Privacy Notice Implementation.” <https://www.iata.org/whatwedo/workgroups/Documents/ADS-AB-2018-03v4-GDPR-General-Notice.pdf>.
- IPTC. “RightsML.” IPTC. <http://dev.iptc.org/RightsML>.
- Ke-Wu, Yan. 2009. “Study on the Forecast of Air Passenger Flow Based on SVM Regression Algorithm.” In *First International Workshop on Database Technology and Applications*, IEEE, 325–28.
- Khanmohammadi, Sina, Chun An Chou, Harold W. Lewis, and Doug Elias. 2014. “A Systems Approach for Scheduling Aircraft Landings in JFK Airport.” In *IEEE International Conference on Fuzzy Systems*,.
- Kim, Seongdo, and Do Hyoung Shin. 2016. “Forecasting Short-Term Air Passenger Demand Using Big Data from Search Engine Queries.” *Automation in Construction*.
- Koko, Famien et al. 2018. “Differential Privacy in Licensing Model and Ecosystem for Data Sharing.” NSF. <https://cci.drexel.edu/mrc/wp-content/uploads/2018/03/RDABerlinPoster.pdf>.
- Kotopoulos, Alkis, and Marialena Nikolopoulou. 2016. “Thermal Comfort Conditions in Airport Terminals: Indoor or Transition Spaces?” *Building and Environment*.
- Koutroumpis, Pantelis, Aija Leiponen, and Llewellyn Thomas. 2017. ETLA Working Papers *The (Unfulfilled) Potential of Data Marketplaces*.
- Laik, Ma Nang, Murphy Choy, and Prabir Sen. 2014. “Predicting Airline Passenger Load: A Case Study.” In *Proceedings - 16th IEEE Conference on Business Informatics, CBI 2014*,.
- Lee, Hanbong, Waqar Malik, and Yoon C. Jung. 2016. “Taxi-Out Time Prediction for Departures at Charlotte Airport Using Machine Learning Techniques.” In *16th AIAA Aviation Technology, Integration, and Operations Conference*, , 3910.
- Lian, Guan et al. 2018. “Predicting Taxi-out Time at Congested Airports with Optimization-Based Support Vector Regression Methods.” *Mathematical Problems in Engineering*.
- Lluís, Palma García. 2018. “Trajectory Optimization for Noise Abatement Arrival Procedures. Case Study at Barcelona Airport.” Universitat Politècnica de Catalunya.
- Magaña, Uriel, S. Afshin Mansouri, and Virginia L.M. Spiegler. 2017. “Improving Demand Forecasting in the Air Cargo Handling Industry: A Case Study.” *International Journal of Logistics Research and Applications*.
- Mathur, Arjun, Aaron Nagao, and Kenny Ng. 2013. *Predicting Flight On-Time Performance*.
- Matthews, Bryan et al. 2013. “Discovering Anomalous Aviation Safety Events Using Scalable Data Mining Algorithms.” *Journal of Aerospace Information Systems*.
- Mindtree. 2018. “Data Sharing Between Airlines and Airports.”
- Murça, Mayara Condé Rocha, R. John Hansman, Lishuai Li, and Pan Ren. 2018. “Flight Trajectory Data Analytics for Characterization of Air Traffic Flows: A Comparative Analysis of Terminal Area Operations between New York, Hong Kong and Sao Paulo.” *Transportation Research Part C: Emerging Technologies*.
- Nai, Wei, Lu Liu, Shaoyin Wang, and Decun Dong. 2017. “An EMD–SARIMA-Based Modeling Approach for Air Traffic Forecasting.” *Algorithms* 10(4): 139. <http://www.mdpi.com/1999-4893/10/4/139> (December 20, 2018).
- NEXTT. 2018. “NEXTT THE CARGO JOURNEY.” IATA. <https://nextt.iata.org/?page=main&data=cargo> (November 27, 2018).
- Olson, Steve, and Autumn S Downey. 2013. *System Sharing Clinical Research Data : Workshop Summary*.
- OpenDataCharter. 2018. “Open Data Charter.” <https://opendatacharter.net/principles/>.
- Özyılmaz, Kazım Rifat, Mehmet Doğan, and Arda Yurdakul. 2018. “IDMoB: IoT Data Marketplace on Blockchain.” <http://arxiv.org/abs/1810.00349> (December 21, 2018).
- PaloAlto. 2017. *Next-Generation Security for Aviation Environments*.

- Van Panhuis, Willem G. et al. 2014. "A Systematic Review of Barriers to Data Sharing in Public Health." *BMC Public Health*.
- Park, Ji-Sun et al. 2018. "Smart Contract-Based Review System for an IoT Data Marketplace." *Sensors (Basel, Switzerland)* 18(10). <http://www.ncbi.nlm.nih.gov/pubmed/30360413> (December 21, 2018).
- Patel, Vishra. 2018. "Airport Passenger Processing Technology: A Biometric Airport Journey." EMBRY-RIDDLE Aeronautical University. <https://commons.erau.edu/edt/385> (December 20, 2018).
- Pettai, Martin, and Peeter Laud. 2015. "Combining Differential Privacy and Secure Multiparty Computation." In *Proceedings of the 31st Annual Computer Security Applications Conference on - ACSAC 2015*,.
- Rahaman, Mohammad Saiedur, Margaret Hamilton, and Flora D Salim. 2017. "Predicting Imbalanced Taxi and Passenger Queue Contexts in Airport Predicting Imbalanced Taxi and Passenger." In *Pacific Asia Conference on Information Systems*,.
- Ramsay, Sebastian. 2018. "The General Data Protection Regulation vs. The Blockchain: A Legal Study on the Compatibility between Blockchain Technology and the GDPR."
- Raskin, Max. 2016. SSRN *The Law and Legality of Smart Contracts*.
- Rebollo, Juan Jose, and Hamsa Balakrishnan. 2014. "Characterization and Prediction of Air Traffic Delays." *Transportation Research Part C: Emerging Technologies*.
- Reis, Luis, Ana Paula Rocha, and Antonio J. M. Castro. 2018. "An Electronic Marketplace for Airlines." In Springer, Cham, 60–71. [http://link.springer.com/10.1007/978-3-319-94779-2\\_6](http://link.springer.com/10.1007/978-3-319-94779-2_6) (December 20, 2018).
- Sakr, Mahmoud. 2018. "A Data Model and Algorithms for a Spatial Data Marketplace." *International Journal of Geographical Information Science* 32(11): 2140–68. <https://www.tandfonline.com/doi/full/10.1080/13658816.2018.1484124> (December 20, 2018).
- Salah, Khadi. 2014. "Environmental Impact Reduction of Commercial Aircraft around Airports. Less Noise and Less Fuel Consumption." *European Transport Research Review*.
- Sankaranarayanan, Hari Bhaskar, Gaurav Agarwal, and Viral Rathod. 2016. "An Exploratory Data Analysis of Airport Wait Times Using Big Data Visualisation Techniques." In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions, CSITSS 2016*,.
- Schmidt, Michael. 2017. "A Review of Aircraft Turnaround Operations and Simulations." *Progress in Aerospace Sciences*.
- Stahl, Florian, Fabian Schomm, and Gottfried Vossen. 2018. *The Data Marketplace Survey Revisited*.
- Sternberg, Alice, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. 2017. "A Review on Flight Delay Prediction." <http://arxiv.org/abs/1703.06118> (December 20, 2018).
- Sulistyowati, Ratna et al. 2018. "Hybrid Forecasting Model to Predict Air Passenger and Cargo in Indonesia." In *2018 International Conference on Information and Communications Technology, ICOIACT 2018*,.
- Tang, Ching Hui, Shangyao Yan, and Yu Hsuan Chen. 2008. "An Integrated Model and Solution Algorithms for Passenger, Cargo, and Combi Flight Scheduling." *Transportation Research Part E: Logistics and Transportation Review*.
- Trondheim. 2003. "MPEG Working Group Completes MPEG21-5 Rights Expression Language (REL)." *Trondheim*. <http://xml.coverpages.org/MPEG21-5-Trondheim.html>.
- Truong, Hong Linh et al. 2012. "Data Contracts for Cloud-Based Data Marketplaces." *International Journal of Computational Science and Engineering*.
- Vijay, Rathinamala et al. 2018. "Air Cargo Monitoring: A Robust Tamper Detection and Reliable Communication System." In *2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES)*, IEEE, 1–4. <https://ieeexplore.ieee.org/document/8442087/> (December 20, 2018).
- Villata, Serena, and Fabien Gandon. 2012. "Towards Licenses Compatibility and Composition in the

- Web of Data.” In *CEUR Workshop Proceedings*,.
- Vu, Quang Hieu et al. 2012. “DEMOS: A Description Model for Data-as-a-Service.” In *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*,.
- W3C. 2008. “CcREL: The Creative Commons Rights Expression Language.” W3C.
- . 2010. “RDFa in XHTML: Syntax and Processing 1.1.” W3C. <https://www.w3.org/MarkUp/2010/ED-rdfa-syntax-20100113/>.
- . 2018a. “ODRL Information Model 2.2.” W3C. <https://www.w3.org/TR/odrl-model/>.
- . 2018b. “ODRL Vocabulary & Expression 2.2.” W3C. <https://www.w3.org/TR/odrl-vocab/>.
- WorldBank. 2018. “Open Data Essentials.” <http://opendatatoolkit.worldbank.org/en/essentials.html>.
- Wright, Aaron. 2016. “Conceptualizing Smart Contracts.” *SlideShare*.
- Wu, Weiwei, Haoyu Zhang, and Wenbin Wei. 2018. “Optimal Design of Hub-and-Spoke Networks with Access to Regional Hub Airports: A Case for the Chinese Regional Airport System.” *Transportmetrica A: Transport Science*.
- Xu, Ran, and Kaiquan Cai. 2019. “Solving Airport Gate Assignment Problem Using an Improved Genetic Algorithm with Dynamic Topology.” In *Advances in Intelligent Systems and Computing*,.
- Yan, Zhang, and Jun Zhang. 2010. “A Hybrid Model of Neural Network and Grey Theory for Air Traffic Passenger Volume Forecasting.” *Key Engineering Materials* 439: 818–22.
- Yang, Ji Jiang, Jian Qiang Li, and Yu Niu. 2015. “A Hybrid Solution for Privacy Preserving Medical Data Sharing in the Cloud Environment.” *Future Generation Computer Systems*.
- Zeinaly, Yashar, Bart De Schutter, and Hans Hellendoorn. 2015. “An Integrated Model Predictive Scheme for Baggage-Handling Systems: Routing, Line Balancing, and Empty-Cart Management.” *IEEE Transactions on Control Systems Technology*.
- Zheng, Yu-Jun, Wei-Guo Sheng, Xing-Ming Sun, and Sheng-Yong Chen. 2016. “Airline Passenger Profiling Based on Fuzzy Deep Machine Learning.” *IEEE Transactions on Neural Networks and Learning Systems*.
- Zhu, Yada, Hongxia Yang, and Jingrui He. 2015. “Co-Clustering Based Dual Prediction for Cargo Pricing Optimization.” In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, New York, New York, USA: ACM Press, 1583–92. <http://dl.acm.org/citation.cfm?doid=2783258.2783337> (December 20, 2018).
- Zipeng, Li, and Wang Yanyang. 2018. “A Review for Aircraft Landing Problem” ed. Dan Zhang. *MATEC Web of Conferences* 179: 03016. <https://www.matec-conferences.org/10.1051/mateconf/201817903016> (December 20, 2018).
- Zonglei, Lu, Wang Jiandong, and Zheng Guansheng. 2008. “A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning.” In *Proceedings - 2008 International Symposium on Knowledge Acquisition and Modeling, KAM 2008*,.
- IATA. 2017. 2017 GLOBAL PASSENGER SURVEY. <https://www.iata.org/publications/store/Documents/GPS-2017-Highlights-report.pdf>.

## Annex II: Aviation Data Analytics State-of-Play - Indicative papers' in-depth analysis

### 1. Approximate algorithms for the combined arrival-departure aircraft sequencing and reactive scheduling problems on multiple runways

**Link**

<https://search.proquest.com/openview/bcd294d62734f672d6d53b32273c821a/1?pq-origsite=gscholar&cbl=18750&diss=y>

**Year**

2013

**Published/Appeared in**

OLD DOMINION UNIVERSITY

**Main topic**

Aircraft Sequencing Problem (ASP): determine the assignment of each aircraft to a runway, the appropriate sequence of aircraft on each runway, and their departing or landing times. Also addresses the Aircraft Reactive Scheduling Problem (ARSP) as air traffic systems frequently encounter various disruptions due to unexpected events such as inclement weather, aircraft failures or personnel shortages.

**Algorithms used (could be also interesting models, methods)**

- 1) Adapted Apparent Tardiness Cost with Separation and Ready Times (AATCSR),
- 2) the Earliest Ready Time (ERT),
- 3) Fast Priority Index (FPI),
- 4) Simulated Annealing (SA),
- 5) Metaheuristic for Randomized Priority Search (Meta-RaPS)
- 6) Do-Nothing and Left-Shift are the repair strategies for the flight cancellations
- 7) RepairBySlack, RepairByEDD, InsertDelayed algorithms to repair the schedule for flight delays
- 8) RepairByTWST and InsertNew are for the arrival of new aircraft.
- 9) TWST Algorithm
- 10) SA-Re Algorithm

**Important assumptions / limitations**

It is assumed that each runway can accommodate at most one aircraft at any time, that runways are reliable, and that they operate independently. **The problem can then be modeled as an identical parallel machine scheduling problem** with the runways being machines and the aircraft being jobs that have ready times (release times), target times (due dates), deadlines, tardiness penalties (weights), and sequence-dependent separation (setup) times.

**Types of Data Used**

Each aircraft is characterized by its operational type (i.e, arrival or departure), weight-class (i.e, heavy, large, or small), priority (aircraft tardiness penalty), ready time, target time, deadline, and separation times.

**Format of Data Used**

text

**Data Language**

n/a

**Features Used**

- Airline Schedule
- Flight Disruption information
  - o Flight delay (Yes/No)
    - Number of Delay
    - Delay flight index
    - Time of delay
  - o Flight Cancellation
    - Number of Cancellation
    - Cancel flight index
  - o Unexpected flight
    - Number of unexpected flight
    - Unexpected flight index

- Ready time
- Target time
- Weight

**Includes real-time data**

Yes

**Includes historic data**

No

**Includes real-time analysis**

Yes

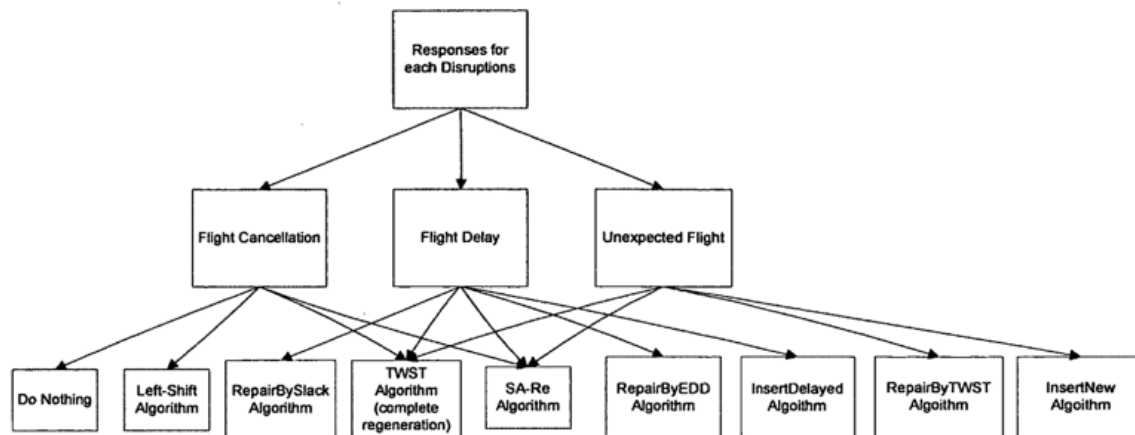
**Mentioned software**

Algorithm implementation in C, Minitab 15.1 is used for analysis

**Main outcome/ conclusion**

The following figure illustrates the reactive scheduling strategies for corresponding disruptive event. The proposed responses for each disruption are considered at every decision point, and then, the best (i.e., the one with the minimum objective function value) reactive sequencing policy is identified from several candidates.

- a) A complete regeneration method, TWST Algorithm, which treats flight cancellation, delay and unexpected arrivals simultaneously.
- b) The SA-Re Algorithm is a hybrid-metaheuristic, which gets the initial solution from the TWST Algorithm and then applies simulated annealing algorithm.
- c) Do-Nothing and Left-Shift are the repair strategies which are considered specifically for the flight cancellation disruption.
- d) Partial repair strategies, RepairBySlack (Reschedule the affected flights by minimum slack time), RepairByEDD (Reschedule the affected flights by earliest deadline) and InsertDelayed Algorithms (Insert the delayed flights to the best position), are proposed to repair the schedule after the delays in particular.
- e) RepairByTWST and InsertNew Algorithms are the repair strategies for the unexpected flight arrival disruption.


**Research- /Industry-oriented**

Industry

**Addressed to**

Airports, Airlines

**Relevance to ICARUS**

High

**Relevant ICARUS pilot**

AIA, PACE

**Relevant data are available in ICARUS**

Yes

**Relevant Journey**

Aircraft

**Reference**

Hancerliogullari, G. (2013). Approximate algorithms for the combined arrival-departure aircraft sequencing and reactive scheduling problems on multiple runways. Old Dominion University.

## 2. An integrated model and solution algorithms for passenger, cargo, and combination flight scheduling

<b>Link</b>
<a href="https://www.sciencedirect.com/science/article/pii/S1366554508000185">https://www.sciencedirect.com/science/article/pii/S1366554508000185</a>
<b>Year</b>
2007
<b>Published/Appeared in</b>
Transportation Research Part E 44 (2008) 1004–1024
<b>Main topic</b>
An integrated scheduling model that combines passenger, cargo and combi flight scheduling.
<b>Algorithms used (could be also interesting models, methods)</b>
Lagrangian relaxation technique , sub-gradient method, four self-developed upper bound heuristics, flow decomposition algorithm,
<b>Important assumptions / limitations</b>
The scope of this paper is confined to pure fleet routing and flight scheduling. Although the scheduling process is closely related to the aircraft maintenance and the crew scheduling processes, these processes are usually separated, to facilitate problem solving. In practice, maintenance and crew constraints are rather flexible, due to the use of stand-by crews and a progressive maintenance policy, thus excluded in the modeling. In addition to the three aforementioned major elements, there are several operating constraints that need to be considered, including specifically the number of available airplanes in each fleet, the quota for each airport/airport pair, and the airplane's capacity, respectively. As well, the same flight leg in the passenger-fleet-flow and the combi-fleet-flow networks can be served at most once. Similarly, the same flight leg in the cargo-fleet-flow and the combi-fleet-flow networks can be served at most once.
<b>Types of Data Used</b>
1) The fleet-flow time–space networks <ol style="list-style-type: none"> <li>1. Flight leg - flight connecting two different airports.</li> <li>2. Ground– indicates the holding or the overnight stay of an aircraft at an airport in a time window.</li> <li>3. Cycle - shows the continuity between two consecutive planning periods. It connects the end of one period to the beginning of the next period, for each airport. The arc cost is the cost of holding an airplane overnight.</li> </ol>
2) The passenger-flow time–space networks <ol style="list-style-type: none"> <li>1. Passenger delivery - represents the transportation of passengers from one airport to another on a flight leg.</li> <li>2. Passenger holding - passengers staying at an airport in a time window.</li> <li>3. Passenger demand - the actual service demand for an Origin Destination(OD) pair. It connects the arrival station to the departure station of the corresponding network OD pair.</li> </ol>
3) The cargo-flow time–space networks <ol style="list-style-type: none"> <li>1. Cargo delivery - represents the transportation of cargo from one station to another on a flight leg.</li> <li>2. Cargo holding - indicates that the cargos are held at an airport in a time window.</li> <li>3. Cargo demand - shows the service demand for the OD–time-pair that would actually be served in the network.</li> </ol>
<b>Format of Data Used</b>
text

<b>Data Language</b>
n/a
<b>Includes real-time data</b>
No
<b>Includes historic data</b>
No
<b>Includes real-time analysis</b>
No
<b>Experimentation/ Validation Dataset</b>
based on data obtained from a major Taiwan airline's operations in Asia during 2002. There were 12 and 10 cities served by passenger and cargo services, respectively. Three types of aircraft were used, including five B767-300 passenger aircraft (226 seats each), 8 B747-400 combi aircraft (272 seats and 35 metric tons each), and 6 MD-11F cargo aircraft (80 metric tons each).
<b>Mentioned software</b>
C , CPLEX 8.1
<b>Main outcome/ conclusion</b>
The practical values of the integrated model have several points: (1) direct and systematic integration of flight sources, (2) effective management of passenger–cargo relationships, (3) speeding up scheduling process and enhancing cooperation among different processes, and (4) providing a systematic computerized tool, all of which are helpful for airlines' operations.
<b>Research- /Industry-oriented</b>
Research
<b>Addressed to</b>
Airlines
<b>Relevance to ICARUS</b>
High
<b>Relevant data are available in ICARUS</b>
Partially
<b>Relevant Journey</b>
Aircraft, Passenger, Cargo
<b>Reference</b>
Tang, C. H., Yan, S., & Chen, Y. H. (2008). An integrated model and solution algorithms for passenger, cargo, and combi flight scheduling. <i>Transportation Research Part E: Logistics and Transportation Review</i> , 44(6), 1004-1024.

### 3. A Review on Flight Delay Prediction

<b>Link</b>
<a href="https://arxiv.org/pdf/1703.06118.pdf">https://arxiv.org/pdf/1703.06118.pdf</a>
<b>Year</b>
November 6, 2017
<b>Published/Appeared in</b>
CEFET/RJ
<b>Main topic</b>
A thorough literature review of approaches used to build flight delay prediction models
<b>Algorithms used (could be also interesting models, methods)</b>
<ul style="list-style-type: none"> <li>- Machine learning algorithms <ul style="list-style-type: none"> <li>o k-Nearest Neighbor</li> <li>o neural networks, reinforcement learning</li> <li>o SVM</li> <li>o fuzzy logic</li> </ul> </li> </ul>



- random forests.
- Operational Research
  - Simulations
  - Queueing Models
- Network Representation
  - Graph Approaches (Direct acyclic graphs)
  - Bayesian network
- Probabilistic Models
  - Conditional probability
  - Survival model
  - expectation-maximization model combined with genetic algorithms
- Statistical analysis
  - Regressions
  - Econometric models
  - Tests
  - Correlation analysis

#### Types of Data Used

- Weather data
- Airline Data
- Airport Data

#### Data Language

Region (Asia , Brazil, Europe, US)

#### Features Used

- Planning features
  - Flight Plan
  - Airline schedule
  - Airport Schedule
- Temporal features
  - Season
  - Month
  - Day of the week
  - Time of day
- Weather
  - Visibility
  - Ceiling
  - Convective weather
  - Surface weather
- Spatial
  - Airport
  - City
  - Region
- Operations
  - Capacity
  - Demand
- General
  - Airline Status
  - Airport Infrastructure
  - Aircraft Model
  - Aircraft Occupancy
  - Fares
  - Frequency
  - Prior delay levels
  - Operational conditions

#### Includes real-time data

YES

#### Includes historic data

YES

#### Includes real-time analysis

YES

#### Experimentation/ Validation Dataset

- The United States Department of Transportation
- The Federal Aviation Administration
- The Bureau of Transportation Statistics
- Eurocontrol
- National Oceanic and Atmospheric Administration of the United States
- The Weather Company (US)

#### Main outcome/ conclusion

In the light of the domain-problem classification, this timeline showed a dominance of delay propagation and root delay over cancellation analysis. Researchers used to focus on statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the use of machine learning and data management is increasing significantly. This clearly characterizes a Data Science trend.

#### Research- /Industry-oriented

Research

#### Addressed to

Airports

#### Relevance to ICARUS

High

#### Relevant ICARUS pilot

AIA

#### Relevant Journey

Aircraft, Passenger

#### Reference

Sternberg, A., Soares, J., Carvalho, D., & Ogasawara, E. (2017). A Review on Flight Delay Prediction. arXiv preprint arXiv:1703.06118.

## 4. Anomaly Detection in Aircraft Data using Recurrent Neural Networks (RNN)

#### Link

<https://ieeexplore.ieee.org/abstract/document/7486356>

#### Year

2016

#### Published/Appeared in

2016 Integrated Communications Navigation and Surveillance (ICNS)

#### Main topic

The application of Recurrent Neural Networks (RNN) for effectively detecting anomalies in flight data. Recurrent Neural Networks with Long Short Term Memory cells (RNN LSTM) and Recurrent Neural Networks with Gated Recurrent units (RNN GRU) are capable of handling multivariate sequential time-series data without dimensionality reduction, and can detect anomalies in latent features. RNNs can also be implemented for real-time anomaly detection on the flight deck.

#### Algorithms used (could be also interesting models, methods)

Recurrent Neural Networks (RNN), Long Term Short Term Memory (LSTM), Gated Recurrent Units (GRU), Exceedance Detection, Multiple Kernel Based Anomaly Detection (MKAD), Clustering Based Anomaly Detection (ClusterAD)

#### Important assumptions / limitations

Need for Dimensionality Reduction,

#### Types of Data Used

21 continuous and discrete variables are recorded at the sampling rate of 2 Hz, which include aircraft state and automation state parameters for the approach phase of flight. The dataset included a total of 500 flights of which 485 are normal flights and 15 are anomalous flights.

**Format of Data Used**

text

**Features Used**

21 continuous and discrete variables are recorded at the sampling rate of 2 Hz, which include aircraft state and automation state parameters for the approach phase of flight.

**Includes real-time data**

No

**Includes historic data**

Yes

**Includes real-time analysis**

Yes

**Experimentation/ Validation Dataset**

The dataset included a total of 500 flights of which 485 are normal flights and 15 are anomalous flights. Eleven canonical anomalies from were introduced into the data set. The abbreviations are used to refer to the anomalies: 1) Very High Airspeed Approach (Rushed and Unstable Approach) (VHSPD) 2) Landing Runway Configuration Change 1 (RW1) 3) Landing Runway Configuration Change 2 (RW2) 4) Auto Land without Full Flaps (Unusual Auto Land Configuration) (FLAP) 5) Auto Land with Single Auto Pilot (Unusual Auto Land Configuration) (1AP) 6) High Energy Approach (Too High or Too Fast or both) (HENG) 7) Recycling of FDIR (FDIR) 8) Influence of Wind (WIND) 9) High Pitch Rate for Short Duration (PTCH) 10) High Airspeed for Short Duration (SHORT) 11) Low Energy Approach (LENG)

**Mentioned software**

X-Plane Simulation, using the X-plane Software Development Kit (XSDK), was configured to run in a Monte Carlo shell. External plugins were developed to manipulate the simulation set-up configuration, the pilot commands, and a Monte Carlo shell.

**Main outcome/ conclusion**

All RNN models are able to detect 8 out of 11 anomalous cases. Anomalous flights with abnormal Pitch for short duration and the second case with a Runway change were not detected by either the MKAD or RNN models.

Normal flights were also part of the test set. Both the RNN and MKAD algorithms successfully classified them as negative. In this way, the RNNs did not exhibit any false positives yielding an ideal precision value equal to 1

**Addressed to**

Airports, Airlines

**Relevance to ICARUS**

Low

**Relevant Journey**

Aircraft

**Reference**

Nanduri, A., & Sherry, L. (2016, April). Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). In Integrated Communications Navigation and Surveillance (ICNS), 2016 (pp. 5C2-1). IEEE.

## 5. Current Icing Potential: Algorithm Description and Comparison with Aircraft Observations

**Link**

<https://journals.ametsoc.org/doi/abs/10.1175/JAM2246.1>

**Year**

2005

**Published/Appeared in**

Journal of Applied Meteorology

**Main topic**

The “current icing potential” (CIP) algorithm combines satellite, radar, surface, lightning, and pilot- report observations with model output to create a detailed three-dimensional hourly diagnosis of the potential for the existence of icing and supercooled large droplets. It uses a physically based situational approach that is derived from basic and applied cloud physics, combined with forecaster and onboard flight experience from field programs. Both fuzzy logic and decision-tree logic are applied in this context. The CIP algorithm, its individual components, and the logic behind them are described.

**Algorithms used (could be also interesting models, methods)**

Current Icing Potential (CIP)

**Important assumptions / limitations**
**Types of Data Used**

To maximize the value of multiple data sources and to represent better the hybrid nature of icing conditions, CIP merges satellite, surface, radar, lightning, and PIREP observations with model forecasts of T, RH, SLW, and vertical velocity and then uses fuzzy- logic and decision-tree techniques to determine the likelihood of icing and SLD at each location. CIP determines icing and SLD potentials in a step- wise fashion (see Fig. 1 for a conceptual diagram and Fig. 2 for a flowchart of the process) and will be described in this way. In step 1, the datasets are placed onto a common grid. In step 2, the 3D locations of clouds and precipitation are found using satellite, surface, and radar observations. In step 3, fuzzy-logic membership functions are applied to icing-related fields to create interest maps. In step 4, the physical icing situation is determined by using a decision tree. In step 5, the initial icing and SLD potentials are calculated by situationally combining interest maps from basic fields (e.g., T, RH). In step 6, the final icing potential is calculated by increasing or decreasing the initial icing potential using the vertical velocity, SLW, and PIREP interest maps.

**Format of Data Used**

text

**Features Used**

Cloudiness, Cloud-top height, Cloud-base height, precipitation presence and precipitation type, Temperature, Cloud-top temperature, Relative humidity, Vertical velocity, Explicit supercooled liquid water predictions, Pilot reports

**Includes real-time data**

Yes

**Includes historic data**

No

**Includes real-time analysis**

Yes

**Main outcome/ conclusion**

CIP combines satellite, surface, radar, lightning, and pilot observations with model output to provide a detailed three-dimensional hourly diagnosis of the potential for icing and SLD. It uses a physically based situational approach derived from basic and applied cloud physics principles, combined with forecaster and onboard flight experience from field programs. CIP uses a conservative approach to the determination of the locations of clouds and precipitation and in its depiction of the potential for icing and SLD, showing the worst possible conditions that are likely to exist within a given portion of airspace (3D grid volume).

CIP provides users with accurate, high-resolution depictions of icing and SLD potential, allowing them to make route-specific decisions that can help aircraft to avoid icing, including that associated with SLD. The use of the operational CIP in combination with high-resolution diagnoses and forecasts of convection, turbulence, ceiling and visibility, flight-level winds, and so on in an easy-to-use graphical form may soon allow pilots and dispatchers to choose their routes and altitudes appropriately to allow for more efficient and, most important, safer flights

**Addressed to**

Airports, Airlines

**Relevance to ICARUS**

Medium

**Relevant data are available in ICARUS**

Yes

**Relevant Journey**

Aircraft

## Reference

Bernstein, B. C., McDonough, F., Politovich, M. K., Brown, B. G., Ratvasky, T. P., Miller, D. R., ... & Cuning, G. (2005). Current icing potential: Algorithm description and comparison with aircraft observations. *Journal of Applied Meteorology*, 44(7), 969-986.

## 6. Efficient Active Set Algorithms for Solving Constrained Least Squares Problems in Aircraft Control Allocation

### Link

<https://ieeexplore.ieee.org/abstract/document/1184694>

### Year

2002

### Published/Appeared in

Proceedings of the 41st IEEE Conference on Decision and Control, 2002.

### Main topic

In aircraft control, control allocation can be used to distribute the total control effort among the actuators when the number of actuators exceeds the number of controlled variables. The control allocation problem is often posed as a constrained least squares problem to incorporate the actuator position and rate limits. In this paper we investigate the use of classical active set methods for control allocation. We develop active set algorithms that always find the optimal control distribution and show by simulation that the timing requirements are in the same range as for two previously proposed solvers.

### Algorithms used (could be also interesting models, methods)

SLS Sequential least squares  
MLS Minimal least squares  
WLS Weighted least squares  
WLS2 WLS with a maximum of  $N = 2$  iterations  
RPI Redistributed pseudoinverse  
FXP Fixed-point iteration algorithm

### Types of Data Used

Aircraft data, consisting of the control effectiveness matrix and the position and rate limits. The commanded virtual control trajectory corresponds to the helical path. It contains 85 samples, each consisting of the commanded aerodynamic moment coefficients.

### Format of Data Used

text

### Includes real-time data

No

### Includes historic data

No

### Includes real-time analysis

No

### Mentioned software

MATLAB

### Main outcome/ conclusion

All algorithms produce solutions which satisfy the actuator position and rate constraints. By construction, SLS and MLS both generate the exact solution to the control allocation problem formulated as the sequential least squares problem.

WLS only solves an approximation of the original problem, but with the weight 7 set to 1000 as in the simulations, WLS comes very close to recovering the true optimal solution.

WLS2 uses at most two iterations, corresponding to at most two changes per sample in the set of active actuator constraints. For the feasible trajectory, WLS2 almost exactly recovers the optimal solution, while for the infeasible trajectory, the solution quality is somewhat degraded. In most sampling instants, WLS (without any restriction on the number of iterations) finds the optimum in only one or two iterations. In those instants where WLS needs three or more iterations, WLS2 only finds a suboptimal solution, but can be thought of as retrieving the correct active set a few sampling instants later. Similar restrictions on the number of iterations could also be introduced in SLS and MLS.

The RPI performance seems difficult to predict. The general flaw with RPI is the heuristic rule it is based on which claims that it is optimal to saturate all control surfaces which violate their bounds in some iteration. Note that RPI yields the highest average number of actuator saturations in all cases.

In all of the simulated cases, FXP generates rather poor, although continuous, solutions to the control allocation problem.

#### **Research- /Industry-oriented**

Research

#### **Addressed to**

Airports

#### **Relevance to ICARUS**

Low

#### **Relevant data are available in ICARUS**

No

#### **Relevant Journey**

Aircraft

#### **Reference**

Harkegard, O. (2002, December). Efficient active set algorithms for solving constrained least squares problems in aircraft control allocation. In Decision and Control, 2002, Proceedings of the 41st IEEE Conference on (Vol. 2, pp. 1295-1300). IEEE.

## **7. Integrated recovery of aircraft and passengers after airline operation disruption based on a GRASP algorithm**

#### **Link**

<https://www.sciencedirect.com/science/article/pii/S1366554516000028>

#### **Year**

2016

#### **Published/Appeared in**

Elsevier, Transportation Research Part E: Logistics and Transportation Review Volume 87

#### **Main topic**

Considers the integrated recovery of both aircraft routing and passengers. It is shown that the integrated recovery of flights and passengers can decrease both the recovery cost and the number of disrupted passengers.

#### **Algorithms used (could be also interesting models, methods)**

GRASP algorithm, separate recovery method (SRM), passenger reassignment algorithm (PRA), minimum cost path algorithm.

#### **Important assumptions / limitations**

Minimum turnaround time is assumed to be 40 min, the maximum delay time limitation is 4h, and the recovery period varies from 7:00 AM to 12:00 AM. Crew members' availability.

#### **Format of Data Used**

text

#### **Features Used**

Synthetic data on passengers, including the number of passengers on each flight and the cost for each passenger refunded.

- Delay cost per passenger in flight
- Cost per passenger re-assigned from itinerary
- Flight Delay time
- Number of aircraft
- Maximum delay time
- Number of passengers in flight
- Number of passengers in itinerary
- Number of seats in aircraft
- Minimum time for passenger connection between flights
- Number of flights of itinerary
- Schedule connection time

**Includes real-time data**

No

**Includes historic data**

No

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

Disruption scenarios (instances) were randomly generated as follows.

- 2–10 aircraft are grounded throughout the day.
- 2–5 aircraft are delayed for a period of time ranging from 1 h to 4 h.
- A 40-min minimal turnaround time is assumed at all airports.
- Delay cost for each passenger is CNY 0.1/min.
- The number of the passengers in each planned aircraft is 80% of its capacity.
- Reassignment cost for each passenger is determined according to the difference between the actual arrival time of the new itinerary and the arrival time of the scheduled itinerary. The rate is assumed to be CNY 0.15/min.
- The cost for each passenger refunded refers to the average ticket price of the itinerary.
- The recovery period is one day, that is, from 7:00 AM to 12:00 AM.

**Main outcome/ conclusion**

The solution from heuristic offers a great improvement over the other three solution types. The comparison between SRM and heuristic reveals that heuristic can achieve 10.7% fewer passengers refunded as well as almost 10% lower total cost. The comparison between the airline heuristic and this heuristic shows sharp differences. The average number of disrupted passengers derived from the airline heuristic is much greater than that provided by this heuristic. In addition, solution values from our heuristic dominate in the distribution of delay, reassigned and refunded costs with a reduction of approximately 65%.

**Research- /Industry-oriented**

Research

**Addressed to**

Airports, Airlines

**Relevance to ICARUS**

Low

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Aircraft, Passenger

**Reference**

Hu, Y., Song, Y., Zhao, K., & Xu, B. (2016). Integrated recovery of aircraft and passengers after airline operation disruption based on a GRASP algorithm. *Transportation research part E: logistics and transportation review*, 87, 97-112.



## 8. A Multi-Aircraft Model for Conflict Detection and Resolution Algorithm Evaluation

<b>Link</b>
<a href="https://hybridge.nlr.nl/documents/D1.3%20Version1.3%20(18%20Feb%202004).pdf">https://hybridge.nlr.nl/documents/D1.3%20Version1.3%20(18%20Feb%202004).pdf</a>
<b>Year</b>
2004
<b>Published/Appeared in</b>
HYBRIDGE, IST-2001-32460
<b>Main topic</b>
A method for modelling the evolution of multiple flights from the point of view of an air traffic controller
<b>Important assumptions / limitations</b>
1. Spoilers, flaps, etc. are not used as inputs. Their effect on the aerodynamic parameters is considered. 2. The angle of attack and side-slip angles are small. The flight path angle and the bank angle are treated as inputs rather than states. 3. Ground speed is given by a simple addition of wind speed to airspeed. 4. A 3D FMS is used. Aircraft do not correct for a long track errors. 5. During final approach and landing, no attempt is made to track a glide path. As a consequence, the aircraft generally tend to miss the beginning of the runway longitudinally. 6. The same controller is used to track straight paths and turns. 7. No nominal weather data is provided to the model. The nominal wind is zero and all wind is treated as a stochastic perturbation. 8. Aircraft turning at a way point ignore the measured wind and compute the point where they should start their turn as though the wind was zero. 9. The model ignores the effect of weather phenomena other than wind speed (e.g. fluctuations in humidity).
<b>Types of Data Used</b>
The multi-aircraft model contains several parameters, such as the masses of aircraft, their aerodynamic coefficients, the gains of the controllers used to model the FMS, the variance and spatio-temporal correlation of the wind, etc. Typical values for many of these parameters (aircraft masses, aerodynamic coefficients) are obtained from the Base of Aircraft Data (BADA) database. The wind statistics are estimated from publicly available weather data. Monte-Carlo simulations of our multi- aircraft model driven by random wind with the computed statistics, compare the results to earlier studies on the deviation of aircraft from their flight plan, and tune the values of the FMS gains to get the results of the simulations to match the conclusions of these studies.
<b>Format of Data Used</b>
text
<b>Includes real-time data</b>
Yes
<b>Includes historic data</b>
No
<b>Includes real-time analysis</b>
Yes
<b>Mentioned software</b>
Monte-Carlo
<b>Main outcome/ conclusion</b>
Develop a model that does not necessarily reproduce exactly the systems used in commercial aircraft, but adequately simulates their behavior from the point of view of an Air Traffic Controller (ATC), while maintaining a workable degree of simplicity. The purpose of the model is to be used primarily as a basis for numerical experiments for the evaluation of conflict detection and resolution algorithms. The model also allows to include uncertainty about some of the actions of the air traffic controller, for example the time at which they order an aircraft to begin its final descent. Finally, the model also allows to capture parametric uncertainty, for example, uncertainty about the initial mass of the aircraft. The long-term aim is to enable evaluation of algorithms that estimate such parameters, based for example on adaptive control and system identification methods.
<b>Research- /Industry-oriented</b>
Research
<b>Addressed to</b>
Airports
<b>Relevance to ICARUS</b>
Low
<b>Relevant ICARUS pilot</b>

**Relevant data are available in ICARUS**

Yes

**Relevant Journey**

Aircraft

**Reference**

Glover, W., & Lygeros, J. (2004). A multi-aircraft model for conflict detection and resolution algorithm evaluation. HYBRIDGE Deliverable D, 1, 3.

## 9. Airline Passenger Profiling Based on Fuzzy Deep Machine Learning

**Link**
<https://ieeexplore.ieee.org/abstract/document/7577870>
**Year**

2016

**Published/Appeared in**

IEEE Transactions on Neural Networks and Learning Systems (Volume: 28, Issue: 12, Dec. 2017)

**Main topic**

A deep learning approach to passenger profiling.

**Algorithms used (could be also interesting models, methods)**

Pythagorean fuzzy deep Boltzmann machine (PFDBM), biogeography-based optimization (PFDBM-BBO), Gaussian mixture model (PFDBM-G), hybrid gradient and enhanced BBO learning (PFDBM-EBO).

**Types of Data Used**

1) The passenger name record, including the identity information of the passenger and booking information of the flight. 2) Travel statistics of the passenger's flight history from the Aviation Administration (collected from different airlines). 3) Travel statistics from other transportation methods, such as railway and marine. 4) Travel statistics from the Tourism Administration (collected from travel agencies). 5) Criminal records from the Public Security Department. 6) Educational records from the Education Department. 7) Tax records from the Tax Department. 8) Reported information from banks and the Housing Administration. 9) Consumption records or patterns from large retailers (the detailed records have typically been preprocessed, summarized, and/or mined by retailers' systems). 10) (Preprocessed) telecommunication behavior records or patterns from telecom operators. 11) (Authenticated and preprocessed) Internet behavior records or patterns from Internet operators.

**Format of Data Used**

Text

**Data Language**

n/a

**Features Used**

Based on their data types, the input features can be divided into the following four classes. 1. Static binary-valued features, which can be input to the PFDBM directly. 2. Static real-valued features, which are first transformed into binary values using a Gaussian-Bernoulli RBM (GRBM) and then input to the PFDBM. GRBM is typically effective for image pixel binary classification. 3. Static labeled features, which are first transformed into real values using membership functions of fuzzy sets defined on the domain of the labels, and then transformed into binary values using GRBM. 4. Dynamic (temporal) features, which are first transformed into real values using an additional temporal filter, and then transformed into binary values using GRBM.

**Includes real-time data**

Yes (for real life application)

**Includes historic data**

Yes

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

Air China dataset

### Main outcome/ conclusion

The key component of the DNN is a novel PFDBM model whose governing parameters are expressed based on PFSs. We propose for the PFDBM a hybrid learning algorithm that employs an evolutionary metaheuristic for facilitating exploration and a gradient-based method for enhancing exploitation. Experimental results show that the proposed DNN with evolutionary learning exhibits competitive performance on the training data sets.

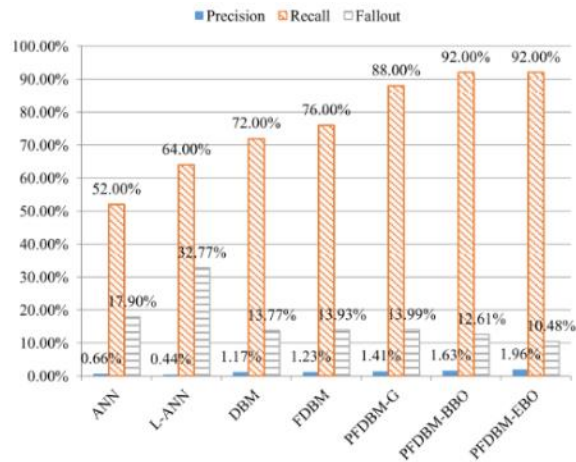


Fig. 8. Precision, recall, and fallout results of the seven models for individual passenger profiling.

### Research- /Industry-oriented

Research

Addressed to

Airports, Airlines

Relevance to ICARUS

High

Relevant ICARUS pilot

AIA, CELLOCK

Relevant data are available in ICARUS

No

Relevant Journey

Passenger

Reference

Zheng, Y. J., Sheng, W. G., Sun, X. M., & Chen, S. Y. (2017). Airline passenger profiling based on fuzzy deep machine learning. IEEE transactions on neural networks and learning systems, 28(12), 2911-2923.

## 10. Predicting Imbalanced Taxi and Passenger Queue Contexts in Airport

Link

<https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1109&context=pacis2017>

Year

2017

Published/Appeared in

Pacific Asia Conference on Information Systems PACIS 2017 Proceedings

Main topic

Employs various sampling techniques and machine learning algorithms to predict the imbalanced queue contexts related to taxi and passenger.

Algorithms used (could be also interesting models, methods)

QCD algorithm, Pearson's Correlation Coefficient, Oversampling (OS), Undersampling (US), Joint Sampling (JS), No Sampling (NS), Naïve Bayes (NB), decision tree (J48), random forest (RF), decision table (DT), PART decision rule, support vector machine (SVM), k-nearest neighbor (k-NN),

#### Important assumptions / limitations

For the sake of simplicity, the study is limited to using three datasets. However, the integration of more contextual datasets (e.g. traffic condition data, public transport usage data) together with the three datasets could provide more insights about the problem domain.

#### Types of Data Used

- i) the taxi trip data,
- ii) the JFK airport passenger wait times data and
- iii) the JFK weather condition data.

#### Format of Data Used

Text

#### Data Language

n/a

#### Features Used

Taxi Trip	Passenger Wait Time	Weather	Queue Context
<ul style="list-style-type: none"> <li>- Medallion</li> <li>- Pickup date / Time</li> <li>- Drop off date/Time</li> <li>- Pickup Latitude / Longitude</li> <li>- Drop-off Latitude / Longitude</li> <li>- Trip Distance</li> <li>- Rate Type</li> <li>- Payment Type</li> <li>- Tip Amount</li> <li>- Passenger Count</li> </ul>	<ul style="list-style-type: none"> <li>- Terminal</li> <li>- No of Flights</li> <li>- No of Passengers</li> <li>- No of booths</li> <li>- Average wait time</li> </ul>	<ul style="list-style-type: none"> <li>- Temperature</li> <li>- Dew Point</li> <li>- Humidity</li> <li>- Wind speed</li> <li>- Precipitation</li> <li>- Condition</li> <li>- Event</li> </ul>	<ul style="list-style-type: none"> <li>- Day / Hour</li> <li>- Total Pax</li> <li>- Total flights</li> <li>- Total booths</li> <li>- Pax wait time</li> <li>- Taxi queue wait time</li> <li>- Pax Pickup Frequency/Rate</li> <li>- Pax Drop-Off Frequency/Rate</li> <li>- Temperature</li> <li>- Humidity</li> <li>- Wind speed</li> <li>- Precipitation</li> <li>- Dewpoint</li> </ul>

#### Includes real-time data

No

#### Includes historic data

Yes

#### Includes real-time analysis

No

#### Experimentation/ Validation Dataset

8 three real world datasets based on the JFK international airport in New York City.

#### Main outcome/ conclusion

The proposed framework is able to identify the most effective prediction techniques for the queue context prediction problem and analyze their performance from two different points of views: the taxi drivers and the passengers. The SVM and the Random Forest are the two most effective predictor algorithms identified by the proposed framework. The experimental results using the framework show the effectiveness of the approach. The SVM performs better from the taxi drivers' point of view while Random Forest shows better results from the point of view of the airport passengers to predict different queue contexts in a given future time stamp.

#### Research- /Industry-oriented

Research

#### Addressed to

Airports

#### Relevance to ICARUS

Medium

**Relevant ICARUS pilot**

AIA, CELLOCK

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger

**Reference**

Rahaman, M. S., Hamilton, M., & Salim, F. D. (2017). Predicting Imbalanced Taxi and Passenger Queue Contexts in Airport. In Proc. of the Pacific Asia Conf. on Info. Systems (PACIS). <https://doi.org/pacis2017/172>.

## 11. Understanding airline passenger behavior through PNR, SOW and webtrends data analysis

**Link**
<https://ieeexplore.ieee.org/abstract/document/7184897>
**Year**

2015

**Published/Appeared in**

2015 IEEE First International Conference on Big Data Computing Service and Applications

**Main topic**

Investigates airline passenger behavior by analyzing three types of travel data: passenger name record (PNR), share of wallet (SOW) and web trends. The PNR data analysis will help the airline company to identify who are influential passengers in their social circles. With SOW data analysis, this study identifies who are potential high-value travelers, and suggest corresponding marketing segmentation and promotion strategies based on different SOW level. Passenger's webtrends information includes mobile number, membership number, identity number, and other web browsing records. Connecting these webtrends data with other information sources, this study provides an overview and insights on individual passenger's website and mobile usage.

**Algorithms used (could be also interesting models, methods)**

PNR Social Network Development

- Take every passenger as a node, if two passengers have travelled together, then there is a connection
- Add two directed connections to every pair (assume passenger A and passenger B who travelled together before are a pair), every connection strength is determined by the following factor: the proportion of "go together" times to single travel times  
For example: Passenger A (a corporate executive) travelled 100 times in total. There are 5 times that he went with B (executive's mother), so the connection strength directed from A to B is 0.05. However, B only got 5 trips in all her life, then the connection strength from B to A is 1.
- Count and show the detail information of passengers; such as gender, age, travel times etc.
- Once the network is developed, the featured relationship and value of passengers can be determined.

SOW Analysis in PNR Social Network

Share of wallet (SOW), a marketing term referring to the amount of the customer's total spending that a business captures in the products and services that it offers. SOW here means a ratio of tickets purchase amount from an airline company to passenger's total travel times.

**WEBTRENDS ANALYSIS**

Passenger's webtrends information includes mobile number, membership number, identity number, and other web browsing records.

1) Import WebTrends data into hdfs by flume. 2) Use pig to preprocess the raw data, includes: identification of ID number, mobile phone number, member 325 card number, user account; statistics and analysis of necessary customer information; map of process steps name and event types. 3) Data integration of SVC customer account and PNR customer account 4) The treated data can be used to a) Link the views of recognised WebTrends users and system users, do the business statistics, and store the processed data into hdfs/HBase/MySQL b) Extract present data, form HFile, and mount it on Hbase.

**Important assumptions / limitations**

This study only has the purchase data from a Chinese airline company, the passenger's overall airline consumption is not available.

**Types of Data Used**

Passenger Demographics

**Format of Data Used**

text

**Data Language**
**Features Used**

WebTrend Analysis	PNR Analysis
Users web data (browsing etc)	Electronic Ticket Data Record
Electronic Ticket	Customer ID
Customer ID	Customer preferences
Contact information	Customer Value
	Check-in records
	First Class/Business Class
	Booking type (Own, Traveller, Call)

**Includes real-time data**

No

**Includes historic data**

No

**Includes real-time analysis**

No

**Main outcome/ conclusion**

The three types of travel data (PNR, SOW and webtrends) will be joined by the user's identity with the addition of user name and the order number, then be imported into the big data platform. Each type of data can reflect the characteristics of the user's behavior. The webtrends data can reflect the user's network behavior when they attempt to book an air ticket online, such as mining the users' preferred access path, stay time in the website, page view, etc. The PNR can reflect the characteristics of the user's travel, such as how often the user travels, which place the user travel most, etc. The SOW can reflect user's loyalty to the airline.

**Research- /Industry-oriented**

Research

**Addressed to**

Airlines

**Relevance to ICARUS**

Medium

**Relevant ICARUS pilot**

Cellock

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger

**Reference**

Chen, S., Zhu, J., Xie, Q., Huang, W., & Huang, Y. (2015, March). Understanding airline passenger behavior through PNR, SOW and webtrends data analysis. In Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on (pp. 323-328). IEEE.

## 12. Travel Behaviour Patterns – Micro Analysis

<b>Link</b>
<a href="http://www.patrec.uwa.edu.au/__data/assets/pdf_file/0005/3085376/Travel-Behaviour-Patters-Project-4.2-Final-Technical-Report.pdf">http://www.patrec.uwa.edu.au/__data/assets/pdf_file/0005/3085376/Travel-Behaviour-Patters-Project-4.2-Final-Technical-Report.pdf</a>
<b>Year</b>
November 2017
<b>Published/Appeared in</b>
Planning and Transport Research Centre, University of Western Australia
<b>Main topic</b>
The aim is to develop a system for querying, analysis and data mining, to support a knowledge discovery process centred on passengers, hubs, and journeys.
<b>Algorithms used (could be also interesting models, methods)</b>
CLARA (k-means hierarchical method that clusters around medoids rather than means and is optimized for large datasets)
non-smooth NMF method (non-negative matrix factorization)
<b>Types of Data Used</b>
Card ID Number
Location of Tag-On (Stop ID Number)
Location of Tag-Off (Stop ID Number)
Time of Tag-On
Time of Tag-Off
Type of transaction (standard, concession, senior)
<b>Format of Data Used</b>
Text
<b>Data Language</b>
n/a
<b>Features Used</b>
Based on their data types, the input features can be divided into the following four classes. 1. Static binary-valued features, which can be input to the PFDBM directly. 2. Static real-valued features, which are first transformed into binary values using a Gaussian-Bernoulli RBM (GRBM) and then input to the PFDBM. GRBM is typically effective for image pixel binary classification 3. Static labeled features, which are first transformed into real values using membership functions of fuzzy sets defined on the domain of the labels, and then transformed into binary values using GRBM. 4. Dynamic (temporal) features, which are first transformed into real values using an additional temporal filter, and then transformed into binary values using GRBM.
<b>Includes real-time data</b>
Yes
<b>Includes historic data</b>
Yes
<b>Includes real-time analysis</b>
No
<b>Experimentation/ Validation Dataset</b>
9 TransPerth & SmartRider ticketing logs
<b>Research- /Industry-oriented</b>
Research
<b>Addressed to</b>
Airports
<b>Relevance to ICARUS</b>
High
<b>Relevant ICARUS pilot</b>
AIA
<b>Relevant data are available in ICARUS</b>
No
<b>Relevant Journey</b>
Passenger
<b>Reference</b>

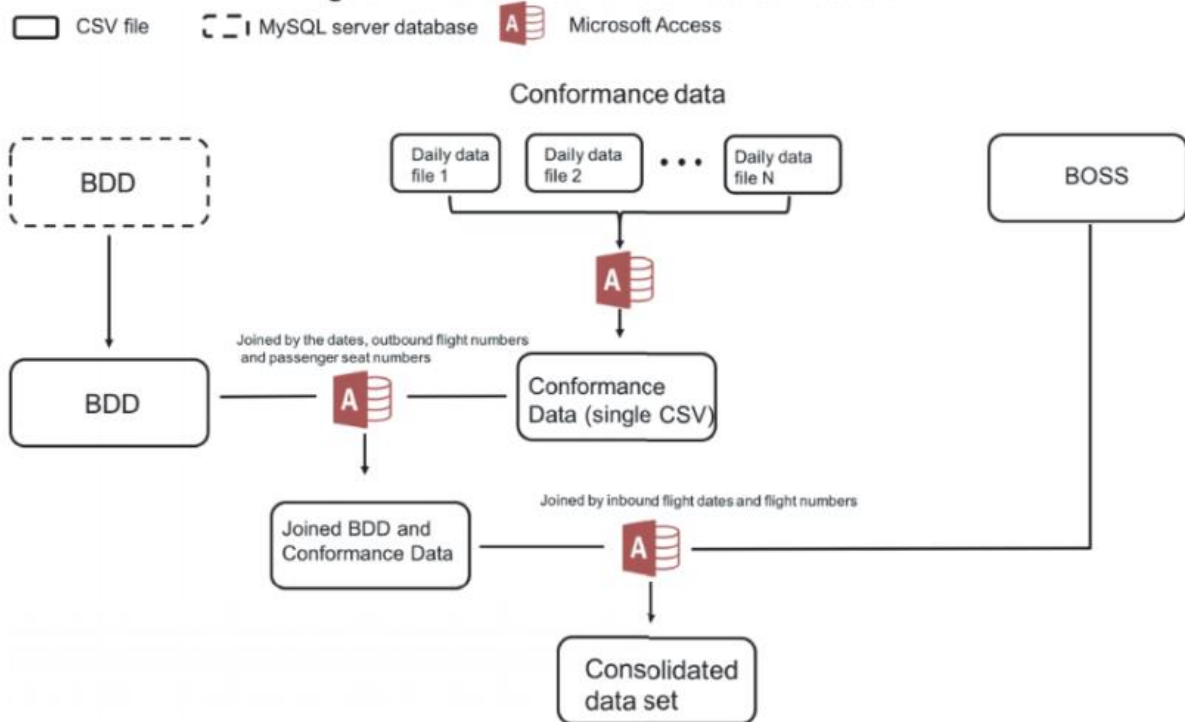


Cardell-Oliver, R., Povey, T., Dokuchaeva, L., Li, J., Lilburne, J., & Biermann, S. (2017). Travel Behaviour Patterns – Micro Analysis. Planning and Transport Research Centre (PATREC).

### 13. Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning

<b>Link</b>	<a href="https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3245609">https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3245609</a>
<b>Year</b>	2018
<b>Published/Appeared in</b>	SSRN
<b>Main topic</b>	A predictive system that generates quantile forecasts of transfer passengers' connection times. Sampling from the distribution of individual passengers' connection times, the system also produces quantile forecasts for the number of passengers arriving at the immigration and security areas.
<b>Algorithms used (could be also interesting models, methods)</b>	regression tree method, linear regression, quantile regression, quantile regression forest, and gradient boosting machine
<b>Important assumptions / limitations</b>	
<b>Types of Data Used</b>	The Business Objective Search System (BOSS), the Baggage Daily Download (BDD), and the Conformance data sets. The BOSS data contains all flight information data. The BDD data records every piece of connecting baggage through Heathrow on the previous day. Each of the records also contains passenger information, such as passengers' arriving and connecting flight number. The Conformance data, provided on a daily basis, stores the records of passengers' boarding pass scans at the conformance desks.
<b>Format of Data Used</b>	
<b>Data Language</b>	n/a

**Figure 2 Consolidation of the Historical Data Set**



Features Used					
Aircraft Body (wide/Narrow)	Runway No	Pax travel class	Conform time passenger arrives	lb punct	
Aircraft Type	Schedule time	lb terminal	Conform code	lb hour	
Pax capacity	Stand number	lb Stand type	Terminal Number	Perceived delta	
Total Pax	Inbound date	Ob stand type	lb region	lb load	
Pax transfers	Flight Number	Pax seat number	Ob region	Ob load	

**Includes real-time data**

Yes

**Includes historic data**

Yes

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

Boss , BDD,

**Mentioned software**

Python GUI scripting interface, Azure Machine Learning platform.

**Algorithm availability**
**Main outcome/ conclusion**

Among all six models shown in Table below, the regression tree has the lowest average pinball loss (2.73) and is best at three of the five quantiles. Not surprisingly, the naïve model performs the worst with an average pinball loss of 3.38. The quantile regression forest and the gradient boosting machine, which are often considered as advanced machine learning methods with high accuracies, perform worse than the second-best model, the quantile regression. The weak performance of these two advanced machine learning methods is likely due to the fact that only seven variables were used to train the models. In terms of point forecasting, the regression tree model has the lowest MAE. The difference between the MAE of our regression tree model and the second-best model, however, is not significant. While our regression tree approach is not significantly more accurate in generating point forecasts of passengers' connection times, it can be easily applied to forecast the number of passengers arriving at immigration and security.

**Table 1 Accuracy of Forecasts on Connection Times in the Test Set**

	MAE	Pinball losses					
		$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	average
Naïve model	10.20	0.96	3.54	5.10	4.86	2.44	3.38
Linear regression	8.52	1.18	3.14	4.26	4.31	2.15	3.01
Quantile regression	8.18	<b>0.78*</b>	2.81	4.09	<b>3.99</b>	2.03	2.74
Quantile regression forest	8.24	0.81	2.84	4.12	4.01	2.06	2.77
Gradient Boosting Machine	8.38	0.81	2.90	4.19	4.07	2.07	2.81
Regression tree	<b>8.16</b>	0.80	<b>2.77*</b>	<b>4.08</b>	4.01	<b>1.98*</b>	<b>2.73*</b>

The symbol \* indicates significance at the 0.1% level.

Values in bold indicate the lowest errors.

**Table Accuracy of Forecasts on Connection Times in the test set**

Passenger experience has been improved through reduced queuing, as capacity and resourcing along the journey more closely match with the dynamic demand. In addition, the managers are able to identify passengers who are at risk of missing their connection, and work with Heathrow and airline teams to assist and expedite their journeys. An accuracy test of the expected passenger flows has been conducted over July and August 2017. The MAE of the forecasts generated from the new predictive system is 38.7, which is 21% lower in comparison to Heathrow's previous system that only generates static predictions the day before operation.

**Research- /Industry-oriented**

Industry

**Addressed to**

Airports

**Relevance to ICARUS**

Medium

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger, Aircraft

**Reference**

Guo, X., Grushka-Cockayne, Y., & De Reyck, B. (2018). Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning. Available at SSRN 3245609.

## 14. Airport Passenger Processing Technology: A Biometric Airport Journey

**Link**
<https://commons.erau.edu/cgi/viewcontent.cgi?article=1384&context=edt>
**Year**

2018

**Published/Appeared in**

Thesis at the Embry-Riddle Aeronautical University

**Main topic**

Analyzes the current passenger processing technology implemented at airports around the world and their associated challenges that passengers face. A new passenger processing technology called a biometric single token identification (ID) is presented as a solution to help alleviate current issues. By using a medium-sized international airport as a case study, the results show that a single token ID is beneficial to the time it takes to process a passenger. Furthermore, it demonstrates that implementation of a single token ID with self-service technology can provide enhanced passenger travel experience, improving operational process efficiency, all while ensuring safety and security

**Algorithms used (could be also interesting models, methods)**

Blockchain technologies, Merkle Hash Tree, SHA-256, Facial Recognition Technology (FRT)

**Important assumptions / limitations**

Biometrics are subjected to varieties of attack. Some potential attacks, along with potential defenses are listed in the following Table by (Stallings 2015)

Attacks	Definition	Examples	Typical Defenses
Client Attack	Adversary attempts to achieve user authentication without access to the remote host	False match	Large entropy; Limited attempts
Host Attack	Directed at the user file at the host where biometrics codes are stored	Template Theft	Capture device authentication ; challenge response
Eavesdropping, theft and copying	Attempts to learn password by attach that involves the physical proximity of the user	Spoofing Biometric	Copy detection at capturing device and authentication
Replay	Repeats a previously capture user response	Replay stolen biometric template response	Copy detection at capturing device and authentication via challenge-response protocol
Trojan Horse	An application or physical device acting as authentic application or device	Installation of rogue client or capture device	Authentication of client within trusted security perimeter
Denial of Service	Disable a user authentication by flooding the service with attempts	Lockout by multiple failed authentication	Multifactor with token.

**Lack of Revocability.** Biometrics have permanent association with every individual. If a system is compromised and the biometric credentials are leaked, then revocability of biometric data is impossible. In this case, once a user's biometric has already been entered into a system, then ability to change or recompute an account with

new or update biometric is not possible. In cases where a user loses a hand or finger or even suffers from biometric theft, then the biometric can be revoked or cancelled, but cannot be replace or substituted.

**Format of Data Used**

Text, Images

**Data Language**

n/a

**Features Used**

Biometrics is a general technical term used for body measurements. There are two types of biometrics: physical and behavioral. Physical biometrics include iris, fingerprints, hand, retinal, face recognition, and DNA. Behavioral biometrics include gait, voice, keystroke, and signature (BioMetrica, 2018) (Agrawal, 2017). Successful application of biometrics relies on the combination of two or more of these approaches to obtain a considerably strong security system. For passengers, highly applied biometrics processing technology includes facial, iris and finger print recognition. Its characteristics varies depending on the type (Facial, Fingerprint, and Iris) with different type of features (Facial patters, Fingertip patterns, Iris patterns)

**Includes real-time data**

No

**Includes historic data**

Yes

**Includes real-time analysis**

No

**Mentioned software**

Simio simulation environment

**Main outcome/ conclusion**

In average, the waiting time in line for security has the potential to be reduced significantly. The simulation models reported a 92.33% decrease, from 47.98 minutes to 3.68 minutes needed for Token ID passengers. The number of passengers waiting in the security line given by the Token ID processing is decreased by 97.86% from an average of 12 passengers compared to 54 passengers

**Research- /Industry-oriented**

Research

**Addressed to**

Airports

**Relevance to ICARUS**

Medium

**Relevant ICARUS pilot**

Airport

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger, Aircraft

**Reference**

Patel, V. (2018). Airport Passenger Processing Technology: A Biometric Airport Journey.

## 15. Understanding Door-to-Door Travel Times from Opportunistically Collected Mobile Phone Records. A Case Study of Spanish Airports

**Link**

[https://www.sesarju.eu/sites/default/files/documents/sid/2017/SIDs\\_2017\\_paper\\_37.pdf](https://www.sesarju.eu/sites/default/files/documents/sid/2017/SIDs_2017_paper_37.pdf)

**Year**

2017

**Published/Appeared in**

SINGLE EUROPEAN SKY ATM RESEARCH (SESAR) Innovation Days, 28th – 30th November 2017

**Main topic**

A methodology for the measurement of door-to-door travel times based on the analysis of opportunistically collected data generated by personal mobile devices. Anonymized mobile phone records are combined with data from the Google Maps Directions API to reconstruct the different legs of the trip and estimate the travel times. One of the high-level goals defined in the Flightpath 2050 report is that, by 2050, “90% of travelers within Europe are able to complete their journey, door-to-door within 4 hours”, extending the concept of “time spent in the airports” to a wider and multimodal concept that includes all the stages of the passengers’ travel from their origins to their destinations.

#### **Algorithms used (could be also interesting models, methods)**

Stay Point Detection Algorithm (SPD), Google Maps Directions API

#### **Important assumptions / limitations**

The study focuses on the passengers flying to Madrid through direct flights from other Spanish airports in July 2016.

An average smartphone user who has his mobile phone switched on and with a data connection enabled typically produces a register at least every 30 to 60 minutes, which provides a reasonably good resolution for the analysis of the user’s mobility patterns.

#### **Types of Data Used**

Mobile phone data (Call Detail Record (CDR)) containing an anonymized identifier of the user together with the time when an interaction with the network occurred and the cell tower to which the user was connected at that moment.

Google Maps Directions API to retrieve the different travel options for the different trips (or trip legs) detected for a user, in order to estimate the selected transport mode(s), the chosen route(s), and the associated travel times.

Demand Data Repository (DDR2) to obtain the average flight duration and the number of flights for each Spanish domestic route with destination the Adolfo Suárez Madrid-Barajas Airport during July 2016.

#### **Format of Data Used**

Text, Images

#### **Data Language**

n/a

#### **Features Used**

Call Detail Record (CDR) is a data record produced every time a mobile phone interacts with the network through a voice call, a text message or an Internet data connection. Each of the records available for this study contains an anonymized identifier of the user together with the time when an interaction with the network occurred and the cell tower to which the user was connected at that particular moment. The registers do not provide the exact location of the users, but the location of the tower to which they are connected, which typically provides an accuracy of around 100-200 meters in urban environments and up to a few kilometers in rural areas, where the mobile network is less dense. To refine the estimation of the user position inside each of these areas, the cell plan has been integrated with a layer of land use information, which assigns users to different areas in a cell with a probability that

#### **Includes real-time data**

No

#### **Includes historic data**

Yes

#### **Includes real-time analysis**

No

#### **Experimentation/ Validation Dataset**

The project has access to a dataset of anonymized mobile phone records provided by Orange Spain; therefore the data analysis methodology has been developed and tested in the context of a case study focused on Spanish domestic flights arriving in the Madrid-Barajas airport, in July 2016.

#### **Main outcome/ conclusion**

The case study presented in this paper shows that mobile phone data can be a useful source of fine-grained passenger information. By analyzing the registers produced by mobile phone users, it is possible to obtain valuable insights into door- to-door travel times, which are very difficult to measure by using more conventional methods.

The mobile phone registers generated by the user during his airport access/egress trip can be used to produce information about the transport mode used and the routes followed by airport users, e.g. by using map matching techniques.

Data from DDR was used to obtain an estimation of the average flight duration from each origin airport to the Madrid airport. However, as mobile phone data provides data about individual trips, it may be possible to extract a sample of passengers for each flight. This would add another level of disaggregation to the door-to-door study, allowing us to extract mobility patterns at different times of the day

**Research- /Industry-oriented**

Research

**Relevance to ICARUS**

Medium

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger

**Reference**

García-Albertos, P., Ros, O. G. C., Herranz, R., & Ciruelos, C. Understanding Door-to-Door Travel Times from Opportunistically Collected Mobile Phone Records.

## 16. Consumer willingness to pay for in-flight service and comfort levels: A choice experiment

**Link**
<https://www.sciencedirect.com/science/article/abs/pii/S0969699708001762>
**Year**

2009

**Published/Appeared in**

Journal of Air Transport Management, Volume 15, Issue 5

**Main topic**

An on-line choice experiment to examine consumer choices with respect to the bundle of services on offer when deciding to purchase a flight. With these data we use the Bayesian methods to estimate a mixed logit specification. Our results reveal that in principle passengers are willing to pay a relatively large amount for enhanced service quality.

**Algorithms used (could be also interesting models, methods)**

multinomial logit (MNL) specification, mixed logit (ML) specification, Bayesian methods, log-normal distribution, MCMC sampler

**Important assumptions / limitations**

The socio-economic variables included in the analysis are Age, Income, Gender and Education. All variables have been included as dummy variables. In the case of Age and Income the data have been divided at the sample mean; 34 years old and £35,000 per annum.

**Types of Data Used**

Price is the price of the ticket, Pitch and Width are seat specification characteristics, Bar is a drinks service, Amenity Pack and Screen are on-board entertainment options, and Sandwich and No Meal are the food options.

**Format of Data Used**

Text

**Data Language**

n/a

**Features Used**

The attributes and levels that was used in the experiment are the following :

Attribute	Units	Status Quo Levels	Additional Levels
Seat Pitch	Inches	28	31,34
Seat Width	Inches	17	18.5
In-Flight Meal	Level	Hot Meal	None, Sandwich
Inflight Entertainment	Level	Standard	Standard plus amenity pack, plus own screen

Complementary inflight drinks	Level	None	Complementary bar service
Ticket Prices	Euro	300	285, 325, 400, 500

**Includes real-time data**

No

**Includes historic data**

No

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

The sample consisted of 568 useable responses, comprising 56% males with the vast majority of respondents being UK nationals (almost 90%). Next, in our sample 63% of respondents had no children, and 32% were single. The distribution of income was relatively evenly distributed over the sample, with 26% earning below £25,000, 16% earning over £50,000 and a sample average of £35,000. Finally, the average age of respondent was 34 years old whereas for the UK it was 39 in 2007.

**Main outcome/ conclusion**

A CE has been deployed to examine consumer WTP for onboard/in-flight service provision and level of comfort. Our results indicate that product differentiation and avoidance of the intra-EU competition are viable strategies to deal with the significantly increased competition within the LCC 3-h range. This survey has addressed which attribute levels are perceived to be valuable by consumers and has provided WTPs for specific levels of these values. In addition, we have been able to identify which type of customer is WTP for which specific type of service provision.

Overall the results indicate that a revised provision of additional on-board comfort and service levels yield a net WTP of approximately €120. This may appear to be a rather large amount, but it needs to be remembered that price differences of this magnitude already exist in the market for trips of this type when comparing different airlines. Thus, it would appear that there is scope for CAs to consider the overall quality of their on-board service provision and not simply follow the approach adopted by the LCCs and operate a no-frills service. Indeed, the WTP plus the WTA estimates found for the reduction in food service provision indicate that following the LCC no-frills strategy need not be the only business model to pursue.

**Research- /Industry-oriented**

Research

**Relevance to ICARUS**

Medium

**Relevant ICARUS pilot**

CELLOCK

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger

**Reference**

Balcombe, K., Fraser, I., & Harris, L. (2009). Consumer willingness to pay for in-flight service and comfort levels: A choice experiment. *Journal of Air Transport Management*, 15(5), 221-226.



## 17. Path2Go: Context-Aware Services for Mobile Real-Time MultiModal Traveler Information

### Link

<https://ieeexplore.ieee.org/abstract/document/6083071>

### Year

2011

### Published/Appeared in

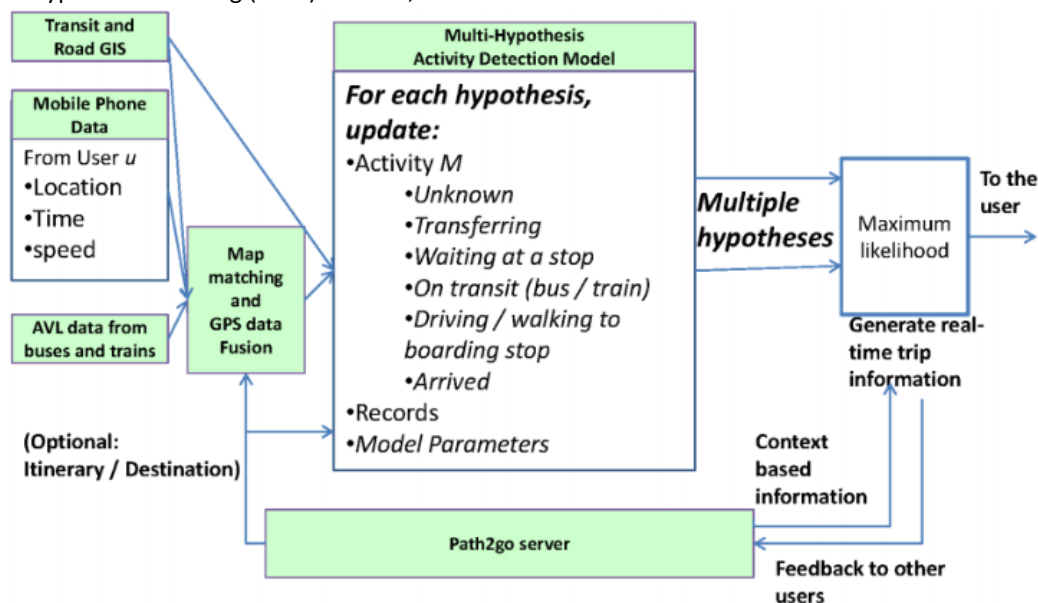
2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)

### Main topic

To improve the accuracy and relevance of the mobile traveler information, context awareness has become an active research topic. In this paper, we describe algorithms and services provided by Path2Go, a multimodal traveler information system. It incorporates real-time traffic, transit and parking information into an integrated system to provide the travelers with reliable, multimodal information via a real-time multimodal trip planner, a real-time traveler information web tool and a mobile multimodal real-time information application.

### Algorithms used (could be also interesting models, methods)

multiple-hypothesis tracking (MHT) method, Markov Chain method



### Types of Data Used

GPS data from transit vehicles, GPS data from smart phones,

### Format of Data Used

Text

### Data Language

n/a

### Features Used

The following features are used :

- Transit and Road GIS
- Mobile phone data
  - Location
  - Time
  - Speed
- AVL data from busses and trains
- Activity
  - Unknown
  - Transferring
  - Waiting at a stop
  - On Transit (bus/train)
  - Driving/waling to boarding stop
  - Arrived

**Includes real-time data**

Yes

**Includes historic data**

No

**Includes real-time analysis**

Yes

**Experimentation/ Validation Dataset**

As part of a field operational test (FOT), users were recruited from the San Francisco Bay Area during August 2010 to mid-November 2010 to test the application and provide survey feedbacks for the effectiveness of the real-time multimodal information. During the project period over 1000 registered mobile users were recruited.

**Main outcome/ conclusion**

The activity detection model employed had several features that helped to achieve reliable activity detection performance, including a multi-hypothesis Bayesian model to address the uncertainties caused by the lack of data in detection; a GPS fusion algorithm that matches travelers to buses or trains so that a user with a priori itinerary or destination can be matched to a transit route faster and more reliably; and several services enabled by the activity detection, such as the variable-frequency communication and need-based GPS use.

System testing carried out by the development team showed a 92% correct detection rate for 109 multimodal trips made. There was an independent evaluation of the application, results of which showed that the mode detection algorithm successfully identified majority of the test scenarios.

**Research- /Industry-oriented**

Research

**Relevance to ICARUS**

Low

**Relevant ICARUS pilot**

CELLOCK

**Relevant data are available in ICARUS**

No

**Relevant Journey**

Passenger

**Reference**

Zhang, L., Gupta, S. D., Li, J. Q., Zhou, K., & Zhang, W. B. (2011, October). Path2Go: Context-aware services for mobile real-time multimodal traveler information. In Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on (pp. 174-179). IEEE.

## 18. Comparative Study on Forecasting Method of Departure Flight Baggage Demand

**Link**
<https://ieeexplore.ieee.org/abstract/document/7007431>
**Year**

2014

**Published/Appeared in**

IEEE Chinese Guidance, Navigation and Control Conference

**Main topic**

Passengers baggage estimation.

**Algorithms used (could be also interesting models, methods)**

Feed forward Neural Network, Multivariate Linear Regression Model

**Important assumptions / limitations**

-

**Types of Data Used**

Flight data

**Format of Data Used**

Text

**Data Language**

N/A

**Features Used**

passenger's boarding number, flight type (domestic or international), flight date, flight time interval, flying time, flight destination, number of luggage per passenger

**Includes real-time data**

Yes (for real-life application)

**Includes historic data**

Yes

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

Flight data of an international airport passenger terminal in 2012 May, a total of 3,050 sets of records composed of boarding number, flight date, flight type, time interval, flight time, Checked Baggage number.

**Mentioned software**

No particular tool, implementation in MATLAB.

**Algorithm availability**

-

**Main outcome/ conclusion**

Multivariate Linear Regression outperform Feed forward Neural Network. Overall results are good.

**Research- /Industry-oriented**

Research

**Addressed to**

Airlines, airports

**Relevance to ICARUS**

High

**Relevant ICARUS pilot**

-

**Relevant data are available in ICARUS**

No

**Notes/Comments**

-

**Relevant Journey**

Baggage

**Reference**

Cheng, S., Gao, Q., & Zhang, Y. (2014, August). Comparative study on forecasting method of departure flight baggage demand. In Guidance, Navigation and Control Conference (CGNCC), 2014 IEEE Chinese (pp. 1600-1605). IEEE.

## 19. Mining Risk Factors in RFID Baggage Tracking Data

**Link**

<https://ieeexplore.ieee.org/abstract/document/7264327>

**Year**

2015

**Published/Appeared in**

16th IEEE International Conference on Mobile Data Management

**Main topic**

Estimate the score of a bag being mishandled.

**Algorithms used (could be also interesting models, methods)**

Resampling techniques to deal with imbalance data e.g. random oversampling, random undersampling, Synthetic Minority Over-sampling Technique (SMOTE).

Classifiers: Decision Tree, Naive Bayes classifier, KNN classifier, Linear regression, Logistics regression, Support vector machine

**Important assumptions / limitations**

-
<b>Types of Data Used</b>
RFID baggage tracking data
<b>Format of Data Used</b>
Text
<b>Data Language</b>
N/A
<b>Features Used</b>
FromAirport, ToAirport, IsTransit (whether it is a transit bag), Weekday, FlightTimeHour, DurationBeforeFlight (Available time for the bag to catch the flight), IsLongerStayFound (If any stay duration between readers at the FromAirport is longer than expected), DelayInArrival, TotalBagInThatHour (Total number of bags read), BagStatus ('OK' or 'Mishandled')
<b>Includes real-time data</b>
Yes (for real-life application)
<b>Includes historic data</b>
Yes
<b>Includes real-time analysis</b>
No
<b>Experimentation/ Validation Dataset</b>
RFID-based baggage tracking data collected from 13 different airports with a total of 124 RFID readers deployed. There are 874,347 records collected for the period from January 1, 2012 until December 2, 2013.
<b>Mentioned software</b>
N/A
<b>Algorithm availability</b>
-
<b>Main outcome/ conclusion</b>
Decision Tree outperform the other classifiers. Re-balancing the data set by under-sampling helps to achieve a better predictive model for the longer transit bag. The proposed method can identify the risky objects very accurately when they approach the bottleneck locations on their paths and can significantly reduce the operation cost.
<b>Research- /Industry-oriented</b>
Research
<b>Addressed to</b>
Airlines, airports, passengers
<b>Relevance to ICARUS</b>
High
<b>Relevant ICARUS pilot</b>
-
<b>Relevant data are available in ICARUS</b>
No
<b>Notes/Comments</b>
-
<b>Relevant Journey</b>
Baggage
<b>Reference</b>
Ahmed, T., Calders, T., & Pedersen, T. B. (2015, June). Mining risk factors in RFID baggage tracking data. In Mobile Data Management (MDM), 2015 16th IEEE International Conference on (Vol. 1, pp. 235-242). IEEE.

## 20. Co-Clustering based Dual Prediction for Cargo Pricing Optimization

<b>Link</b>
<a href="https://dl.acm.org/citation.cfm?id=2783337">https://dl.acm.org/citation.cfm?id=2783337</a>
<b>Year</b>
2015
<b>Published/Appeared in</b>
Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
<b>Main topic</b>
Air cargo pricing optimisation.
<b>Algorithms used (could be also interesting models, methods)</b>
COCOA (developed by the paper authors and first described there)
Hierarchical clustering
Hierarchical Logistic Regression
<b>Important assumptions / limitations</b>
-
<b>Types of Data Used</b>
Cargo origination and destinations, historical bidding prices and bidding stages (win or loss), number of cargo pieces, cargo weight, cargo volume, lead time and customer size, etc.
<b>Format of Data Used</b>
Text
<b>Data Language</b>
N/A
<b>Features Used</b>
Same as data
<b>Includes real-time data</b>
No
<b>Includes historic data</b>
Yes
<b>Includes real-time analysis</b>
No
<b>Experimentation/ Validation Dataset</b>
(a)synthetic data, (b) data from worldwide cargo company challenge
<b>Mentioned software</b>
N/A
<b>Algorithm availability</b>
-
<b>Main outcome/ conclusion</b>
The proposed algorithm can improve revenue significantly.
<b>Research- /Industry-oriented</b>
Research
<b>Addressed to</b>
Airlines, ACHCs
<b>Relevance to ICARUS</b>
Low
<b>Relevant ICARUS pilot</b>
-
<b>Relevant data are available in ICARUS</b>
No
<b>Notes/Comments</b>
-
<b>Relevant Journey</b>
Cargo
<b>Reference</b>
Zhu, Y., Yang, H., & He, J. (2015, August). Co-clustering based dual prediction for cargo pricing optimization. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1583-1592). ACM.

## 21. The Use of an Artificial Neural Network to Predict Australia's Export Air Cargo Demand

### Link

[https://www.researchgate.net/publication/324809092\\_The\\_use\\_of\\_an\\_artificial\\_neural\\_network\\_to\\_predict\\_Australia's\\_export\\_air\\_cargo\\_demand](https://www.researchgate.net/publication/324809092_The_use_of_an_artificial_neural_network_to_predict_Australia's_export_air_cargo_demand)

### Year

2018

### Published/Appeared in

International Journal for Traffic and Transport Engineering (IJTTE)

### Main topic

Air Cargo Demand Forecast

### Algorithms used (could be also interesting models, methods)

Artificial Neural Networks

### Important assumptions / limitations

-

### Types of Data Used

Annual data from 1993 to 2016 including world real merchandise exports, world population growth, world jet fuel prices, world air cargo yields (proxy for air cargo costs), outbound flights from Australia, and Australian/United States dollar exchange rate

### Format of Data Used

Text

### Data Language

N/A

### Features Used

Same as data and two dummy variables, which controlled for the strong cyclical fluctuations in air cargo demand in 2003 and 2015

### Includes real-time data

No

### Includes historic data

Yes

### Includes real-time analysis

No

### Experimentation/ Validation Dataset

Real annual data from 1993 until 2016

### Mentioned software

N/A

### Algorithm availability

-

### Main outcome/ conclusion

The results were good and proved that the demand for Australia's export air cargo products is dependent upon the growth in the world population, which increases the size of the potential market, transportation costs and the development of world merchandise trade volumes.

### Research- /Industry-oriented

Industry

### Addressed to

ACHCs

### Relevance to ICARUS

Low

### Relevant ICARUS pilot

-

### Relevant data are available in ICARUS

No

### Notes/Comments

-

### Relevant Journey

Cargo

### Reference

Baxter, G., & Srisaeng, P. (2018). The Use of an Artificial Neural Network to Predict Australia's Export Air Cargo Demand. *International Journal for Traffic and Transport Engineering*, 8(1).

## 22. An EMD–SARIMA-Based Modeling Approach for Air Traffic Forecasting

<b>Link</b>	<a href="https://www.mdpi.com/1999-4893/10/4/139">https://www.mdpi.com/1999-4893/10/4/139</a>
<b>Year</b>	2017
<b>Published/Appeared in</b>	Algorithms, Multidisciplinary Digital Publishing Institute (MDPI)
<b>Main topic</b>	Air traffic forecasting.
<b>Algorithms used (could be also interesting models, methods)</b>	Hybrid EMD (Empirical Mode Decomposition)-SARIMA (Seasonal autoregressive integrated moving average, used for statistical time series forecasting)
<b>Important assumptions / limitations</b>	-
<b>Types of Data Used</b>	cargo and passenger flow data, both domestic and international
<b>Format of Data Used</b>	Text
<b>Data Language</b>	N/A
<b>Features Used</b>	N/A
<b>Includes real-time data</b>	No
<b>Includes historic data</b>	Yes
<b>Includes real-time analysis</b>	No
<b>Experimentation/ Validation Dataset</b>	Time series data gathered from the official website of the Civil Aviation Administration of China (CAAC), with monthly cargo and passenger flow data, both domestic and international, ranging from January 2006 to July 2014, 103 months in total
<b>Mentioned software</b>	N/A
<b>Algorithm availability</b>	-
<b>Main outcome/ conclusion</b>	The EMD–SARIMA framework can improve the forecasting accuracy to a great level under the four different evaluated cases: domestic cargo, domestic passenger, international cargo, and international passenger.
<b>Research- /Industry-oriented</b>	Research
<b>Addressed to</b>	Air cargo companies
<b>Relevance to ICARUS</b>	Low
<b>Relevant ICARUS pilot</b>	-
<b>Relevant data are available in ICARUS</b>	No
<b>Relevant Journey</b>	



Aircraft

**Reference**

Nai, W., Liu, L., Wang, S., & Dong, D. (2017). An EMD–SARIMA-Based Modeling Approach for Air Traffic Forecasting. *Algorithms*, 10(4), 139.

### 23. A fuzzy approach to addressing uncertainty in Airport Ground Movement optimisation

**Link**

<https://www.sciencedirect.com/science/article/pii/S0968090X18305461>

**Year**

2018

**Published/Appeared in**

Algorithms, Multidisciplinary Digital Publishing Institute (MDPI)

**Main topic**

Route allocation to taxiing aircrafts under uncertainty.

**Algorithms used (could be also interesting models, methods)**

Adaptive Mamdani fuzzy rule based system

**Important assumptions / limitations**

-

**Types of Data Used**

Flight movement data: 1767 tracks for the time period 5–12 November 2013 for aircrafts with an altitude of zero within 5 km of Manchester airport's centre

**Format of Data Used**

Text

**Data Language**

N/A

**Features Used**

N/A

**Includes real-time data**

No

**Includes historic data**

Yes

**Includes real-time analysis**

No

**Experimentation/ Validation Dataset**

Same as data

**Mentioned software**

N/A

**Algorithm availability**

-

**Main outcome/ conclusion**

The proposed approach produces routes that are more robust, reducing delays due to uncertain taxi times by 10–20% over the original QPPTW.

**Research- /Industry-oriented**

Research

**Addressed to**

Airports

**Relevance to ICARUS**

Medium

**Relevant ICARUS pilot**

AIA

**Relevant data are available in ICARUS**

(Yes)

**Notes/Comments**

Aircraft

**Reference**

Brownlee, A. E., Weiszer, M., Chen, J., Ravizza, S., Woodward, J. R., & Burke, E. K. (2018). A fuzzy approach to addressing uncertainty in Airport Ground Movement optimisation. *Transportation Research Part C: Emerging Technologies*, 92, 150-175.