DELIVERABLE

D2.4 – Lessons Learned through cross-pilot analysis

Project Acronym	TT
Project Title	Transforming Transport
Grant Agreement number	731932
Call and topic identifier	ICT-15-2016
Funding Scheme	Innovation Action (IA)
Project duration	31 Months [1 January 2017 – 31 July 2019]
Coordinator	Rodrigo Castiñeira (INDRA)
Website	www.transformingtransport.eu
Project Acronym	TT

Document fiche	
Authors:	Prof. Andrés Monzón [UPM], Pablo Vázquez [UPM], Alessandra Boggio-Marzet [UPM]
Internal reviewers:	[INSIGHT] [AIA]
Work Package:	WP2
Task:	T2.2
Nature:	Report
Dissemination:	Public

Document History			
Version	Date	Contributor(s)	Description
1.0	4/06/2019	UPM (TRANSyT)	Draft document for submission to Reviewers
1.1	19/06/2019	INSIGHT & AIA	P2P Internal Review
1.2	25/06/2019	UPM (TRANSyT)	Final document for submission to EC

Keywords:	KPI, Assessment Framework, Performance Targets, KPI Category, KPI Variation, Lessons Learned, Cross-pilot analysis, Impact, Usability
Abstract:	This deliverable reports on the work performed in WP2/T2.2 "Pilots Requirements Analysis & Lessons Learned with respect to the final assessment of the Performance Indicators established by each one of the pilots within the Transforming Transport project. As well, it is aimed at extracting some lessons that could be replicated in future similar actions, once their usability is depicted here.

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. This document may contain material, which is the copyright of certain TT consortium parties, and may not be reproduced or copied without permission. All TT consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the TT consortium as a whole, nor a certain party of the TT consortium, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

ACKNOWLEDGEMENT

This document is a deliverable of Transforming Transport project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 731932.

Table of Contents

List of Figures

List of Tables

Definitions, Acronyms and Abbreviations

Acronym	Title
AM	Asset Management
CO_2	Carbon Dioxide
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CV	Connected Vehicles
DAS	Distributed Acoustic Sensing
DL	Deliverable Leader
DoA	Description of Action / Degree of Achievement
Dx	Deliverable (where x defines the deliverable identification number e.g. D1.1.1)
EC	Energy Consumption
EEA	European Environment Agency
EF	Economic
EQ	Environmental Quality
EU	European Union
GHG	Green House Gas
IU	Integrated Urban Mobility
KPI	Key Performance Indicator
NLP	Natural Language Processing
MSx	Project Milestone (where x defines a project milestone e.g. MS3)
Mx	Month (where x defines a project month e.g. M10)
NOx	Nitrogen Oxides
OE	Operational Efficiency
PC	Project Coordinator
PLH	Ports as Intelligent Logistics Hubs
PM	Particulate Matter
PPM	Partner Project Manager
PT	Priority Topic
PU	Public
R	Report
RI	Proactive Rail Infrastructure
RNN	Recurrent Neural Networks
SA	Smart Airport Turnaround
SF	Safety
SH	Smart Highways
SN	Dynamic Supply Networks

TBD	To be defined
TL	Task Leader
TMC	Traffic Management Center
TT	Transforming Transport
Tx	Task (where x defines the task identification number e.g. T1.1)
UI	User Interface
WPx	Work Package (where x defines the WP identification number e.g. WP1)
WPL	Work Package Leader
WPS	Work Package Structure

Executive Summary

The overall goal of the Transforming Transport project (TT) is to demonstrate in a realistic, measurable, and replicable way, the transformative effects that Big Data will have on the mobility and logistics sector. The project has been designed to validate Big Data as capable of reshaping transport processes and services significantly increasing operational efficiency, improving customer experience, and fostering new business models. All of this has been achieved by demonstrating, evaluating and validating, in real operational scenarios (Pilots) belonging to different transportation domains, the capability of Big Data innovations to develop more efficient solutions.

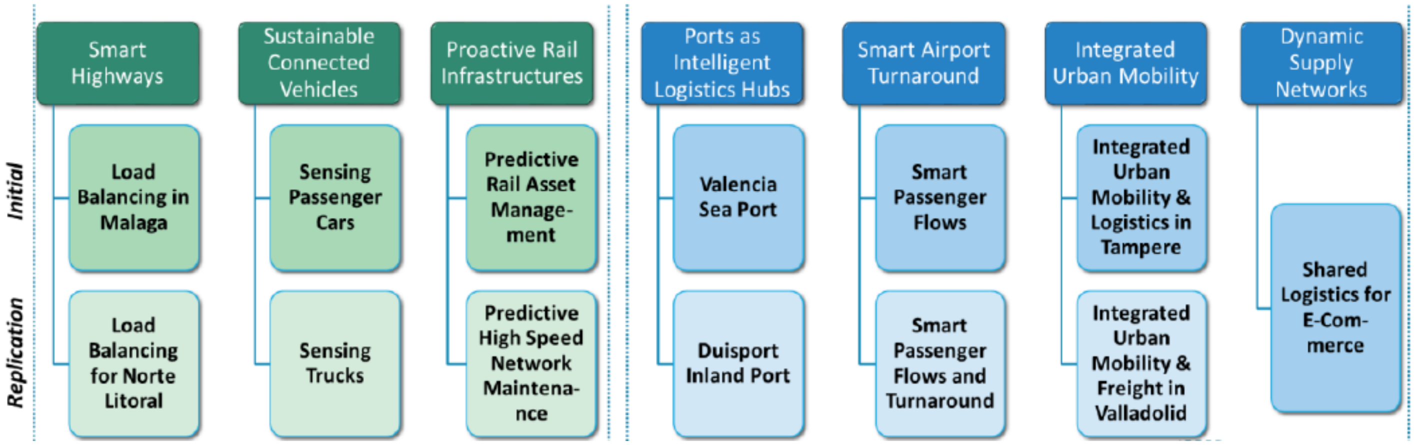
This deliverable reports on the work performed **in WP2/T2.2 "Pilots Requirements Analysis and Lessons Learned".** The goal of this task is to develop the conclusions deriving from the framework that has performed the technical and economic evaluation of the innovations tested in the 13 project pilots and to extract all the positive activities and good or bad practices to bear in mind for further post-project activities. The deliverable mainly covers all the non-technical lessons learned of the **Transforming Transport** project. It also includes a comparison between all the lessons learned by the different Pilots all along the project timespan and the exchange and learning among them.

1 Introduction

The overall goal of TT is to demonstrate in a realistic, measurable, and replicable way the transformative effects that Big Data will have on the mobility and logistics sector. The project is designed to validate Big Data as capable of reshaping transport processes and services significantly increasing operational efficiency, improving customer experience, and fostering new business models. All of this by demonstrating, evaluating and validating, in real operational scenarios (Pilots) belonging to seven different transportation domains, the capability of Big Data innovations to develop more efficient solutions. To achieve this objective, the project is based on a consolidated methodological approach characterized by three main activities:

- Defining both global and pilot domain performance targets
- Testing the innovations in real operation conditions.
- Evaluation and validation of the potential impact of Big Data innovations.

Figure 1. Initial and Replication pilots for each of the seven transport domains


Smart Highways
Sustainable Connected Vehicles
Proactive Rail Infrastructures
Ports as Intelligent Logistics Hubs
Smart Airport Turnaround
Integrated Urban Mobility
Dynamic Supply Networks
Initial
Replication
Load Balancing in Malaga
Load Balancing for Norte Litoral
Sensing Passenger Cars
Sensing Trucks
Predictive Rail Asset Manage-ment
Predictive High Speed Network Maintena-nce
Valencia Sea Port
Duisport Inland Port
Smart Passenger Flows
Smart Passenger Flows and Turnaround
Integrated Urban Mobility & Logistics in Tampere
Integrated Urban Mobility & Freight in Valladolid
Shared Logistics for E-Com-merce

The main objective of task T2.2. – Pilots Requirements Analysis and Lessons Learnt was to: “assess the potential and the requirements for the pilots to effectively demonstrate the transformative potential of Big Data on Mobility and Logistics”. Additional T 2.2 aims were:

- To perform a thorough intra-domain and cross-pilot analysis on the use of Big Data technologies.
- To offer an analysis of the applicability of TT Big Data solutions in post-project replications.

T 2.2 has delivered three reports:

- D 2.2: "Pilot Requirements Analysis"; Development of the evaluation framework including the definition of Priority Topics and KPI selection.
- **D 2.4** & 2.5: "Lessons learned"; execution of the final assessment and extraction of Big Data use recommendations and lessons learnt.

The Development of the Evaluation Framework was a key outcome of Task 2.2, and consequently, the task had significant relevance for the overall impact of the project. The establishment of at least one KPI that will evaluate the performance variation achieved in every objective of the Pilot cases. Furthermore, through the definition of units and data sources, Task 2.2 set the context which outlines the type of data that had to be gathered during the tests implemented throughout the project.

Table 1 presents each Pilot's WP, code and denomination, as well as the two Pilots (initial and replication) which, when necessary, will be coded by adding 1 or 2 to the Pilot Domain Code.

Table 1. Assessment code and denomination for TT Pilots

Pilot Domains	Code	*Initial pilot*	*Replication pilot*
WP4 Smart Highways	SH	Load balancing in Malaga	Load Balancing for Norte Litoral
WP5 Connected Vehicles	CV	Sensing passenger Cars	Sensing Trucks
WP6 Proactive Rail Infrastructures	RI	Predictive Rail Asset Management	Predictive High-Speed Network Maintenance
WP7 Ports as Intelligent Logistic Hubs	PLH	Valencia Sea Port	Duisport Inland Port
WP8 Smart Airport Turnaround	SA	Smart Passenger Flows	Smart Passenger Flows and Turnaround
WP9 Integrated Urban Mobility	IU	Integrated Urban Mobility and Logistics in Tampere	Integrated Urban Mobility and Freight in Valladolid
WP10 Dynamic Supply Networks	SN	Shared Logistics for E-Commerce Pilot	-

In the following sections, the main lessons learned gathered are presented by Pilot, Transport Domain and regarding the Transportation Sector as a whole.

Section 2 describes lessons learned at pilot level concerning only each one of the thirteen pilots individually. This section also aims at summarizing the common findings between Pilots within the different seven domains. **Section 3** reflects all the learnings and the reuse possibilities that pilots have gained from each other across the project, not only from the perspective of the Domain partner but also from all the domains involved and from the project as a whole. **Section 4** intends to collect the most shared lessons learned derived from the pilots' experience along with the Transforming Transport project. Finally, **Section 5** summarizes what is the impact of the

Big Data technologies on each main Assessment Category (Operational Efficiency, Asset Management, Environmental Quality, Energy Consumption and Safety).

Topics such as Methodology Implemented, Data Visualization, Data Analytics and Data Management among others are considered of main relevance when it comes to developing a Big Data project. As well the impacts of Big Data technology on the Transportation sector are summarized.

Table 2 below synthesizes all the fields and dimensions in which any lesson learned has been encountered:

Table 2. Tabular summary of the lessons learned across the identified dimensions

Identified Dimension	Lessons Learned
Profitable end-users	• **Costumers**, **Drivers**, **Passengers**, **Consumers** as those who experience personally the quality of the service. • **Managers** and **Personal Staff** as those who benefit directly from the pilot results.
Methodology	• **End-users** should be involved from the very **beginning** and collaborate all along the different stages of the project. • Constant involvement of working group partners in the development of the project and in the document submissions. • Visits of the **pilot site** and **weekly calls** among partners permits a realistic understanding of the existing situation and an efficient follow-up of the systems. • "**Initial** and **Replication**" terminology not much accurate, since both pilots within the same Transport Domain have worked hand in hand. • **KPI measurement** and definition has resulted in a very challenging task. Thus, addressing only most relevant KPIs is the best option, avoiding distractions from the main ambitions of each pilot. • The support from a dedicated **KPI Team** is essential for the definition and continuous follow-up of the KPIs, as well as acting as main channel of communication and coordination with each pilot.
Data Visualization	**Dashboard** design for data visualization and real-time control, considered as a "*success key factor*" in TT, should: • Show information hierarchically. • Be intuitive, simple and clean.

	• Only display critical and enough well-validated events. • Provide dynamic customization. • Implicate human factors to understand the different issues, selecting the right customer or user to do it. • Enable the users to recognize critical trends easier using colours above and below certain thresholds. • Permit decision makers to use them for strategic planning purposes.
Data Management	• A **Data Management Plan** and a calendar of Quality control tasks should be a must in further projects. • To identify **valuable data sources** that support the understanding of the different transport domains and to define the structure, data format and the different variables since the beginning. • To plan sufficient time at project start for **data refinement** and fine-tuning of data collection before blindly starting any task, since cleansing process is estimated in 80% of total work of Data Analytics Team.
Data Analytics	• Among the most universally accepted principles of analytics is "**Garbage in – Garbage out**", which means that if poor-quality data enter the system, no matter how trendy the software for the analysis can be, the output value is expected to be of poor quality too. • To overcome this, check and cope with missing data, data accuracy, data timelines, different time-zones (clocks), etc. is a must. Also is assigning "**data owners**" that understand data and its field (domain) being able to be in care of data quality. • Using **Deep Learning** and Neural Networks helps to make more efficient development and engineering. • **Time series models** can be successfully approached by traditional **machine learning techniques.** • It is useful to **keep historical non-reproducible data** and, when possible, in **raw format**. Possible errors or improvements in the code that not allow to rebuild processed data if the original data has been deleted.
Data Processing Architectures	• The use of **Big Data platforms**, that have shown their capacity to generate valuable knowledge and new insights for each one of the stakeholders involved in the pilots.

	• The use of **scalable data storage and data processing architecture** is needed as the data volumes are going to be high when the systems start the regular data collection. • **Processing of subsets of data** (e.g. using all of the records for a particular asset) was greatly simplified as the subset could be created, and then processing could be applied to only the relevant data, avoiding time losses by analysing the entire dataset. • **Data lake and cloud environment** were identified to be a good solution for the management of big data to enable data sharing and communication.
Pilot Innovations	• The development of **Dashboards**, predictive **algorithms** and **analytics techniques** have been one of the most rehearsed fields within the project thanks to the Big Data technologies.
Data Protection, Engineering and Standards	• The work under this project has been impacted by the new European regulation in the matter of data protection, EU **General Data Protection Regulation** (GDPR), where prior preparatory actions were taken. • To plan and develop well-proved **disaster recovery systems** in order to be safe for unexpected fails and lose of information. • Using as much as possible **open source technologies** to process and analyse data, as well as in dashboard developments. Since there are thousands of volunteers generating algorithms that everyone can use, contribute using and generating open sources solutions will help to improve this community.
Big Data Impact on Transport Assessment Categories	Big Data technologies have demonstrated its utility in improving the: • **Operational Efficiency** (OE): operating cost, time savings, the average commercial speed, accuracy of the estimated time of arrival, availability of assets, customer satisfaction with the service… • **Asset Management** (AM): number of maintenance interventions, accuracy of failure prediction… • **Environmental Quality & Energy Consumption** (EQ & EC): amount of NOx, CO_2 and PM and GHG emissions. • **Safety** (SF): drivers' perception of safer driving conditions, number of track-side activities…

2 Lessons Learned at Pilot Level

In this section, the objective is to depict all the lessons learned concerning only each one of the thirteen pilots individually. Some of the following topics have been included depending on their relevance for the pilot: Profitable End-Users, Data Visualization and User Interaction, Data Analytics, Data Processing Architectures, Data Protection and Innovative Technologies. This section also aims at summarizing the common findings between Pilots within the different seven domains.[1]

It's worth mentioning that this section summarizes and pulls together the lessons learned elaborated in more detail in the pilot-specific *deliverables Dx.5*, such as to serve as a single point of reference. To this aim, only the most relevant findings are displayed here. For a deeper understanding and insight of pilot-specific discoveries, please refer to the aforementioned documents.

2.1 Smart Highways

The **initial pilot's scenario** is the 100-km long Highway of **Ausol**, a **toll Highway with barriers** that belongs to the AP-7 Highway, which is part of the European route E15, Costa del Sol. This highly-congested semi-urban corridor connects the cities of **Málaga,** Estepona and Guadiaro in the South of Spain. The Highway Concession runs parallel to the free alternative – a dual carriageway with two lanes per direction and grade or signalized intersections at some points along the road. The Platform deployed at Concessions does address the following three main pilot objectives: Understanding **mobility patterns**, **optimizing** AUSOL Highway **operation and maintenance and** guarantee **safer roads**.

The **Norte-Litoral** Highway (**replica pilot**) runs along the northwest coast of Portugal, from Oporto to Caminha, near the Spanish border, with a branch towards the interior between Viana do Castelo and Ponte da Lima. The highway is 109 km long, of which 72 km are refurbished and 47 km new construction. Norte-Litoral adds some additional challenges, since it runs under free-flow infrastructure model (vehicles don't stop to pay), where some indicators are more relevant than in the initial pilot (toll-barrier infrastructure model). Objectives are the same as for both.

1.1. [1] This exchange does not apply for the case of the Supply Chain Network Pilot, since it is a unique pilot whose actions have not been replicated or followed by another pilot pertaining to its same Domain.

Since both Pilots belonging to the *Smart Highways* Domain have made much the same and worked at the same time, in this section only distinct points are going to be presented.

The objective of the pilot was to create an advanced analytical framework to follow the same phase execution pipeline (in a simplified way) by the one proposed by the CRISP-DM (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) methodology. This **methodology** aims to serve as a guide to anyone who, within the requirements of the project, needs to include advanced analytical techniques and consequently needs support in the development and implementation of **predictive models** of advanced analytics, available resources, reuse of tools and environments, offered and supported by the Transportation sector and **estimates of costs and time** for each of the phases of the project.

One of the most important points has been the **collaboration between both pilots**, since in the initial pilot has proceeded to define the common methodology so that in the next phase, the replication pilot was put into practice. The collaboration carried out by both pilots has been based on the following: Firstly, the **reference of steps** to follow and set of good practices that can serve as a basis for the development of advanced analytical models. Secondly, the **automation of resources** to minimize the time in each of these phases and / or steps. And thirdly, the **reuse of algorithms and models** implemented and validated by different teams in the transport sector under the same specification and operationalization mode.

• Malaga Toll Motorway (Spain)

Profitable End-Users

Within the context of the domain of Smart Highways, end users fall into two possible profiles: on one side **drivers and the general public** and on the other hand, the **concession's staff**.

1) **Drivers feedback**: the most important indicator to assess highway efficiency and management are drivers. A distinction must be made between the ones that use the Highway regularly, the ones who use it occasionally and the ones that don't use the highway at all. Collecting information about the thinking, the preferences and dislikes of drivers, is a very strenuous activity. It has been demonstrated the usefulness of making the **surveys** on site at the **toll stations** taking advantage of the stop people are forced to do at toll stations. The main problem is that surveys are performed at long intervals and, once performed, the sample is not representative enough, since many people fall into one of the two last typologies mentioned before whose feedback is not pretty much relevant.

To try to improve the process of gaining the driver perception of the infrastructure, developing a **mobile app** has turned out to be a great solution. The pilot has created an application that allows capturing driving behaviour from built-in accelerometers and GPS in the user's mobile, to capture driver's use of Ausol's corridor, per road segment. The main advantages are that its use is not limited only to drivers of Ausol Highway, it can be used at any time by the user and it can gather much more information than the toll stations surveys. The main downside is that its usage depends exclusively on the driver's will. This problem has been overcome throughout rewards to drivers offering 1-month off for the toll fee of the highway. Thanks to this modern approach have been possible to extract many conclusions concerning the quality of the highway service, such as the driver's safety perception, the physical conditions of the road which have a direct effect on the comfortability of travellers.

The main barrier for the usage of the tolled alternative is obviously the fact of having to pay for it. Nevertheless, this kind of techniques helps them to improve the driver's experience, directly related to safety (road accidents) and traffic flows (congestion) on the road. Once the drivers' awareness had been gained, hopefully it will help to change the mind of those reluctant drivers not fully convinced of using a tolled highway.

2) **Concession staff feedback**: the second group of end-user groups much of the staff of the tolled highway concession. From the management level to operational teams, the tool and the dashboard can have a direct use from these people. The following structure has turned out to be useful to enhance the understanding and management of the tolled-highway operations:

- A Traffic Management Center team (TMC) monitors the Dashboard tool being able to manage in real-time the alarms received and coordinate the actions to be taken by getting in touch with those departments the alarm or incident concerns.
- A toll station management team for monitoring and the control of the alarm management system which are features of interest for the team in order to increase the efficiency and easiness in the activities of stations configuration, toll station personnel planning, etc.
- An operation team, being aware of both, the current state of the traffic flows and the short-term predictions ahead, can provide a better understanding of users' behavior and mobility patterns along the highway that could results in a more efficient and profitable operational strategy, raising issues not previously identified.
- Maintenance team, responsible for the maintenance tasks but also useful for reporting on the real-time location of other maintenance teams on the field, relevant situations or incidents previously identified by the system.

A key point for success is **team coordination** at all levels since the beginning of the project. Weekly follow-up sessions have been held with the end users for collecting the feedback from the teams.

Data Analytics

A complete methodology has been developed during this project, from which the following lessons were learned:

- **Traffic descriptive models**: A detailed data analysis has been carried out on every segment of the highway to calculate traffic profiles grouped by segment, working/not-working day (day type) and month. **First finding is that there might be not enough data at all clusters, so grouping clusters is necessary work** – although never optimal.
- **Incidents Model**: the quantity of data directly derived from incidents or accidents has not been large enough to be considered as optimal and enough to build a consistent model. In this specific case this is due to the low ratio of accidents in this kind of highways. Interestingly, after running the models the number of **incidents on the road is not inversely proportional to the Service Level** (ratio between the number of vehicles driving and the nominal capacity of the road segment). Instead, it was proved that incidents happened the most when the road service level is at its best rate and when it's somewhat downgraded, but not enough to consider it risky. This fact supports the idea that accidents at the wheel are related to the driver's perception of speed and traffic intensity changes, more than the commonly accepted idea of a direct relationship between lower service levels and the accident rate.

Data Protection

The pilot is implemented with high availability, **disaster recovery**, and data backup services from Azure, keeping all services up and running, in a resilient and reliable environment. The platform allows the road concession user to own and control their data with built-in privacy controls with some of the industry's strictest privacy standards in order to reduce system vulnerabilities and protect all the information based on assets of the highway.

Other action taken, in this case from a more detailed technical point, is, within the Bluetooth beacon sub-pilot, all the Bluetooth MAC addresses broadcasted from the vehicles (either the hands-free device of the vehicle or the smartphones of the drivers and passengers) have been trimmed to keep only the information coded into the MAC about the vehicle manufacturer and model, leaving out any other information related to the individual device or vehicle.

Innovative technologies

The value emerged from the perspective of business in this deliverable is closely related to the agreement established between INDRA and CINTRA. This agreement establishes the conditions, to be accepted for both parties, under which, INDRA and CINTRA will exploit all the data and tools beyond the scope of the project, in a commercial relationship either CINTRA as a customer of the service or as co-provider of the tool to third parties. CINTRA will take advantage of the systems and tools developed by INDRA and it will have preferential treatment in the licensing. This agreement can be seen as one case of success of the EC in its effort to foster innovation and innovative collaborations through its H2020 Initiative and beyond.

Regarding the innovative technologies that have helped to fulfil the commitments established at the start of the project, here are the most outstanding ones:

Predictive traffic flow algorithm: The main target for the traffic flow algorithm is to provide a short-time traffic forecast (maximum 120 min approach) with the aim of adapting the operation of the highway according to traffic conditions. It means that, based on that traffic forecast, road operators could be able to adapt the **toll station configuration** adding extra lanes or more efficient payment modes to satisfy the current traffic demand, minimize the impact on the road traffic throughout **lane closures** and provide efficient information to road users thanks to signals on **messaging panels**.

Predictive road incident algorithm: the system provides to the user, the accident probability level based on traffic flow data and external meteorological systems, in order to facilitate proactive decision making in real time whenever there is any accident probability.

HORUS Dashboard: for the visualization of information already processed in an illustrative way to be able to see at a glance the sum of relevant information that will help the user make good strategic decisions.

Animal encroachment – Low-cost network of infrared detectors: deployed over a 2 km-long section, previously identified as a high risk of animal intrusion spot in AUSOL Highway. The system should be able to notify that animals or human beings are moving in the surroundings of the segment under monitoring and that the risk of intrusion exists. The information, collected by the beacons, won't be ready to use before project's deadline due to issues in the integration process but, once positive results demonstrated, we will study if the product has market demand.

Bluetooth receptors network: the system can track all devices Bluetooth MAC (trunked for privacy) turned on along the corridor. This tracking could gather key non-personal information from the users and vehicles such as **travel times, traffic density, origin-destination matrixes, user loyalty analysis and user segmentation.**

• Free-Flow Motorway in Norte Litoral (Portugal)

Data Management

A DAS (distributed acoustic sensing equipment over optic fibre) has been integrated under real traffic conditions for the **accident detection** and control of **animal intrusions** into the road area. The amount of data processed with this technology is around 1TB, which must be cleaned and processed as fast as possible to send the TCC only relevant events related to traffic flow and abnormal events like vehicles stopped on the hard shoulder or guardrail impact.

Innovative Technologies

DAS (Distributed acoustic sensing): the key element of the Replication Pilot is the DAS technology applied to the road environment. Detection of vehicles, the direction of the driving, defects in the pavement, those all are possible use cases where the utilization of DAS technology can suppose an important breakthrough in digitizing linear infrastructures. This device is connected directly to the existing optic fiber buried along the road shoulder at Norte-Litoral Highway. This facilitates the monitoring of on-road traffic as well as off-road surroundings, by sending warning messages to the control center when non-standard behaviours take place.

Further investigation must be done as a post-project activity to demonstrate this technology fast and accurate enough to gather all information about traffic events in real time – a “must do” before designing large scale experiments at other Highway concessions.

2.2 Sustainable Connected Vehicles

Sustainable Connected Cars has been carried out by SoFLEET, Autoaid and Answare. This pilot has focused on cars that belong to different sorts of companies. These companies are interested in achieving efficient management of their fleets. Thanks to telematic dongles installed on the cars and a Big Data architecture, the partners applied techniques and algorithms to offer a decision support system to achieve predictive maintenance of a fleet, monitoring and promote eco-friendly driver behaviours and identification of traffic jams. The dongles gather valuable information from the vehicle such as location, speed, accelerations and fault codes (DTCs). Big Data is able to deal with this huge amount of information and provide functionalities such as data collection, predictive analysis and data visualization.

Sustainable Connected Trucks has been carried out by PTV, JDR, TomTom and Fraunhofer IGD. The main objective for this pilot has been the enhancement of planning and optimization systems for fleet managers. In order to achieve this goal, it was necessary to assess traffic flow for trucks journeys and to detect and analyse logistics hotspots such as terminals, toll stations and ferry stations. Large amounts of Big Data processing, specifically related to truck fleets all over Europe, were necessary for this task. Additionally, the use of satellite images as an extra data source on different planning stages in the context of applications for truck fleet managers was incorporated in this pilot to detect not only the current state of a location but also its changes over time.

Within this domain, the cars and trucks sectors were approached with the idea to replicate the Cars pilot with the Trucks pilot. For this purpose, there was an ongoing exchange between both pilots, e.g. related to the technical architecture, data sources and the definition and measurement of KPIs to find replication possibilities. However, Pilot learned that a **replication** in the proper sense was **not possible**. The main reason for this is the fact both pilots had quite **different requirements**: while the focus of the Cars pilot was on being informed about e.g. tasks needed for reducing emissions to save fuel or related to maintenance, the trucks pilots had to face requirements resulting from the day-to-day business in the logistics sector which is driven by efficiency improvement to save costs.

Thus, an integrated replication would also be difficult in future projects and actions. Yet, a better replication could be possible by e.g. focussing on **common geographical areas** to benefit from common data sources and insights related to the area. Another opportunity to improve the replication activities could be given by a more detailed exchange about data **analytics techniques** and architectures to identify common approaches to use.

• Sustainable Passenger Cars (Spain & France)

Profitable End-Users

Two types of end-users have been fundamental for the pilot's development and further actions:

- **Drivers**: drivers are enabled to access a mobile phone application to which dedicated tips are issued. This way, throughout the observance of the tips by the drivers, these can get the improvement of **fuel consumption** and better **car maintenance**.
- **Fleet managers**: fleet managers can control through **dashboards** the cars with more emissions and problems. In the Cars pilot SoFLEET, as the main end user, has manifested its positive feedback about the usefulness (e.g. statistics dashboard and graphics for tips evolution).

The results of this pilot will highly improve technological competences and competitiveness in the field of predictions and analytics of vehicle diagnostics data, driver behaviour, traffic and driving risks and vehicles locations data. Combing this information will lead to new valuable processes. This sub-project will allow the development of high added value for new services related to the car sector and more precisely with the connected-vehicle sector.

Pilot Innovations

Emission Reduction System: more and more vehicles are connected to the Cloud, as new cars are already prepared for connectivity when they get out from factories and old cars can be adapted using a wide range of devices. This raw data provides a high potential value for analysis and processing; even more, when it is combined with additional data sources like road maps, weather information or brand specifications. AnswareTech has been working on processing and aggregating all this information to provide valuable services available at the online dashboard and mobile applications.

The main objectives of the services aimed at reducing fuel and operational costs and increasing security and fuel performance by monitoring drivers' behaviour and educating them at improving the way they drive. Services such as pattern behaviours, driver monitoring and provision of advice have been proved and validated through a pilot with 20k connected vehicles.

The following features developed in the pilot can be highlighted:

- Easy exploration of all vehicle trips related information such as route, type of route, traffic jams, weather, driver behaviour.
- 15 factors for monitoring driving performance of every trip. From fine grain (trip) to a wider overview (weeks or months).
- Thanks to Pilot's fuel consumption algorithm developed with advanced Machine Learning techniques, one can find out and focus on the most potential fuel reduction vehicles and trip types.
- An intelligent notification system that provides customized advice to drivers according to their behaviour.
- Better decisions by a data-driven approach.
- Integrated API to retrieve the information from different backend systems.

As for the technological novelty aspects of this innovation, we mention the predictive and descriptive analysis of readily available in-vehicle data coming from the cloud for connected car service providers and the use of the machine algorithms for estimating fuel consumption of vehicles.

Sustainable Connected Trucks (The Netherlands & Germany)

Profitable End-Users

In the course of the project, two groups of end users have been identified:

The first end-user, the **logistics service provider**, focused on the transportation process along the chosen European corridors. These corridors were chosen based on the feedback of the end-user to assure the usefulness of the pilot results. Therefore, the big data technologies were applied with a practical relevance especially for the end-user, by looking specifically at the geographical field the end-user is working in. The results provided an overview and a basic understanding of the day-to-day processes, especially as the different data provided by the end-user were combined and assessed with a holistic approach. This provided new and useful insights to the end-user. The usefulness of the on-time arrival approach that was chosen as the second demo-case was highlighted by the end-user as this approach helps to improve the planning of arrival times and therefore helps the logistics service provider to better plan the truck fleet and the general capacity utilization.

The second group of end-users are **providers using satellite images** for validation purposes in the field of traffic-related analyses. Combining these two kinds of data – satellite images and truck traffic data – turned out to be very useful as satellite images serve as additional data sources that have never been set into context to truck traffic data. Furthermore, satellite images are available worldwide and therefore allow for large-scale analyses. This additional information was used to validate results, especially in pre-processing steps before the actual traffic planning phase. One example where this is very useful is model training for moderate or low volume uncongested traffic or parking stops in cities. It's worth mentioning that the images that were not provided on a real-time basis but rather once a day or every couple of days at specific times of the day. Yet, the end-users still gave positive feedback as use cases could be developed beyond the real-time view.

Suggestions for further analysis are covering larger areas to continue the validation activities. As well, visualization techniques and analysis of emissions or long-term applications such as landslides and other changes in the environment are worth it to be addressed following the pilot findings.

Data Visualisation and User Interaction

In the replication pilot, one important aspect was the visualisation of **GPS traces** of truck routes and activities and logistics hotspots. These visualizations were perceived as very useful by the end-user, specifically because the **raw data** could be understood and interpreted better. The visualizations itself contained geographical maps for the routes and pie charts showing the time consumption of different activities such as loading or resting. In the development process, we learned that not every available information has to be shown, but a **useful aggregation** is important. In the available data sources, many different activities and traces for a very large European area were available for example. Showing all of them would have been very complex and confusing. Therefore, we **only showed the most relevant activities** and focused on different geographical corridors.

Data Management

In the trucks pilot, one of the research questions was to **identify valuable data sources** that support the understanding of the trucks transport domain. Therefore, many different data and data sources were part of the pilot, namely satellite images, floating car data, handling activities, maps, routing data etc. These data sources differed in terms of format (e.g. images versus tables), timely availability (e.g. one picture a day versus one trace every minute) or geographical spreading (e.g. one hot spot versus one corridor). One of the first things pilots learned was to **abandon the idea of a holistic technical integration** of all data sources. Data can also provide valuable insights when considered separately to some extent. Concerning the visualization, it was important to develop good use cases and to define the right data for them. Therefore, only **useful data** were used and further processed and could finally reduce the complexity and increase the understandability. Some specific lessons learned in our case were the following:

- **Floating car data** do often contain timely gaps or incorrect positioning. That's why it's important to not fully trust the raw data but to backfire the meaningfulness.
- **Time zones** can differ from data source to data source which was the case for satellite images versus truck data. Before comparing and validating data, we had to obtain the correct time zones for both.
- "**Good" data quality** is required to process data. However, the definition of "good" differs for data sources and that's why the quality criteria must be defined separately. A sufficiently small-time interval for two data sources, for example, can be one day for satellite images and some minutes for GPS traces.

2.3 Predictive Rail Asset Management

The two pilots for this programme of work utilise infrastructure owners' data, namely Network Rail and ADIF for the lead and replication pilot respectively. Each of these Infrastructure owners has differing business goals and ambitions but can be generalised as improving operational efficiency, customer experience, and new business models. The infrastructure owners for both pilots reside in two different physical locations: **UK** (initial) and **Spain** (replica). The key difference between the locations, other than geographic location itself, is the type of railway line each operator is using. The Network Rail railway line of interest is classed as '**mainline'**, whereas the ADIF railway line is categorised as a '**high-speed**' line. Both pilots strive to provide functionality that predicts the failure of mainline railway assets with sufficient foresight that preventative maintenance can be scheduled and performed. Improving the maintenance of rail assets consistently and predictably improves the safety case and other sub goals of the project such as cost efficiency and minimising disruption.

The **pilots worked independently** through the duration of the project and were unable to collaborate on the work due to the **significant differences** in the types of railways being analysed, i.e. **Mainline Rail** (UK Pilot) and **High-Speed** Lines (Spain Pilot).

However, a relevant lesson learned is that for the mutual assets that both pilots have worked on, such as points; it would be useful to have some **common aims and objectives** that could be achieved and measure through the different approaches taken by the pilots. During the timeline of the project, the knowledge transfer would be a key aspect of collaborating which would help understand the complexities and any potential bottlenecks by learning and assistance from the other pilot. Having a common set of KPI in both pilots would have also allowed for a better comparison of the results and a better understanding of the difference in the approaches.

• Mainland Rail in UK

Profitable End-Users

The end user for the use cases in the pilot has been the **maintenance engineers** and **asset maintenance business analysts**.

The demonstrators for several use cases were presented to the users, which found the prototypes intuitive and easy to use. The demos were rated as good overall user experience and relatively easier to navigate on different pages. The performance was highlighted as something that could be improved with additional features that could be added to the environments: **Manual Labelling** (a labelling tool for different asset states), **overlaying other data** such as delayed maintenance activities, line renewals and passenger data, incremental **Learning tool** and **Confidence levels** on the screen.

Data Visualisation and User Interaction

Data visualisation is one of the **key analysis methods for extracting knowledge** regarding the behaviour of the different phenomena of interest e.g. the wire height, stagger, thickness and dynamic forces of the OLE (Overhead Line Equipment)-pantograph interface, and the collating, cleansing and enhancement of datasets in the Points classifier. As these phenomena vary in space and time, are numerous and cover large geographical areas and long timespans, there is a need to visualise the data at different levels of information aggregation and with filtering functionality. Further, to aid the complete understanding of the behaviour of the phenomena, the visualisations should overlay information from different datasets, like OLE and point swing measurements, detected anomalies, incidents and maintenance works.

Data Analytics

The data analyses for studying the feasibility of Big Data for the predictive maintenance of these use cases were performed on **historical data** originally recorded for underpinning the general maintenance and operation procedures.

Many of the challenges when performing the analyses with regards to OLE were to do with the fact that these data are sparse in time and also that the signals believed to have the strongest relation to the phenomena of interest were missing. Further, incidents and maintenance works information that can be used as ground truth is unstructured and not co-registered with the measurement data. All these challenges are well understood, specifically the high cost of generating dense data and machine-readable information vs. the value of such data and information. The following general lessons learned can be stated as:

- For enabling predictive maintenance, **data collection** must be specifically planned in terms of the signals to be recorded and the adequate sampling rate for modelling the target phenomena.

- **Data labelling** is a time and resource consuming process and requires access to a variety of information sources so adequate tools are needed for the accurate and efficient labelling by the domain experts.
- There is **rich information hidden in unstructured text reports** and tools are needed for automating the extraction of this information and encoding it in machine-readable formats.

Additionally, **Python 3.6** was chosen as the base technology for analysis. Python excelled at providing pre-written libraries for various functionalities (particularly for graphical results) and this greatly helped with the analysis of the large datasets used. However, python is **not the tool of choice when considering processing speed**, and therefore a distributed computation framework was required to make analysis practical.

Pilot Innovations

Unsupervised anomaly detection: the unsupervised anomaly detection algorithm was created for the Network Rail OLE measurement data of height stagger and wire thickness. This enabled the use of high-volume measurement data coming from the passenger trains.

Autoregressive Wire Thickness decay model: the autoregressive wire thickness decay model was created to aid the estimation of the remaining useful life of that contact wire asset.

NLP based text algorithm: Natural language processing (NLP) based text processing algorithm was developed for extracting location, asset and failure mode information from incident reports for enabling further analysis for the OLE incidents data.

Fault Classifier algorithm: The fault classifier algorithm was developed to automatically detect the false alarms and anomalies in point machines. The primary step for this was to organise, collate, cleanse, and enhancement of the datasets available. The complexity of the data used has previously been a hindrance when trying to link records together to create meaningful insights. This algorithm performed complex linking between disparate datasets and did this with a scale of data that had not previously been examined.

Precipitation Compensation algorithm: The Precipitation Compensation algorithm (PCA) was developed to reduce the number of false alarms raised by the track circuits with the effect of weather. The PCA algorithm takes into account the amount of precipitation and based on the effect on the current levels, compensates the track circuit current in-order to identify that the alarm was due to weather.

Feature engineering for medium/long term degradations: Temporal effects are important to take into account for the analyses on the track-train interface. The anomalies do not appear suddenly and are more of a long-term process. To address this problem, feature engineering was used by performing time series to build clusters that gather the change over time of the factors studied. This allowed a significant improvement in modelling results.

Smart Heat-map: Most of the data engineering was done at the lowest level possible. But it is impracticable for the supervisor to stay at this level. Using different zoom levels, the information can be aggregated or recomputed for the chosen level, allowing the user to have only the right amount of information. The originality of the approach relies on the fact that the aggregation at each level is completely controlled by the algorithms that have been developed for the project.

• High-Speed Rail in Malaga (Spain)

Profitable End-Users

The main advantage that this pilot presents is the involvement of the end users due to their participation as partners, implementing in different levels the usability of big data technology, which is directly proportional to their participation in the project and the data quality provided.

Adif: as the main infrastructure manager and end-user of the project, Adif has actively participated from the early beginning of the project, and results obtained by the developed predictive application are promising. Therefore, Adif has planned, and it is already in process, to implement the application developed in the main maintenance operation centre of the high-speed railway Cordoba-Malaga, in Antequera. The possibility of **upgrading maintenance schedule system from preventive and corrective maintenance to predictive maintenance** is the main objective of the project. In case that real trials are satisfactory (as it is expected), the possible installation of the predictive model would add an important value to the company, not only in an economical way, but also as pioneers at employing big data technology to optimise railway maintenance. Additionally, in order to improve the management and prediction process, some **advice** for future actions are:

- Involvement of the maintenance operators from the beginning of the full project.
- Automatize the process for collecting data (Seneca Train optimization).
- Include new data that is currently being collected as OLE (Overhead Line Equipment) or ultrasounds.

Ferrovial: Ferrovial clearly understands that the use and management of data through Big Data technologies adds value to the company. Ferrovial together with Ci3 were determined to calculate the theoretical activities that would have been performed when following the predictive model so that the theoretical planning obtained could be compared with the real activities performed each month. This comparison has shown that with the time it would be possible to reduce the number of assets to be maintained, to optimize the travelling distances to the assets and reduce vehicles fuel consumption, having a direct impact on the reduction of greenhouse emissions generated in the maintenance of railways infrastructures. Besides, the participation along the whole project understanding the predictive tool and its visualization has reinforced the awareness of the importance to gather data for its later exploitation.

Thales: due to its late incorporation as point-machine maintenance operator to the project, the development of the results obtained for this use case is not as satisfactory as the results obtained for others. However, important progress for developing predictive maintenance in point-machines by using big data technologies has been achieved. Although great effort has been made, predictions obtained for the use case related to point machines are not as useful as expected due to the **high efficiency** that **point machines in high-speed railways** already have. Some suggestions about how to improve the model in order to upgrade preventive maintenance to predictive maintenance for future projects are:

- Re-feed after each fault detected and maintenance tasks accomplished the model, to update the initial stage from model continues predicting faults.
- Collect and analyse the behaviour of other techniques affecting the point machine as mechanical maintenance.
- Determine if would be a mechanical or electrical fault. To decrease the cost of maintenance, the prediction should indicate in which engine the fault is going to appear. If not, the maintenance team would have to check all the engines, increasing that the maintenance costs.

Data Analytics

Some of the most important aspects that this pilot had to face were the data analytics, specifically everything related to **classification and selection of the most important variables** that affects tracks and point-machines condition. Due to this, the main conclusion reached is that a good algorithm (in this case related to railway failure prediction) is really difficult to develop (and sometimes impossible) unless good data is available, which means that a great amount of data availability is not enough if it has not enough quality, so good data is more important than anything else. Additionally, once the algorithm has been developed, a low coefficient of adjustment is not always a bad result, because it allows discovering opportunities to improve the cycle of collection or provisioning of information.

Data Processing Architectures

The main achievement related to data processing architectures is the utilisation of big data platforms like **Sofia2**, which have shown their capacity to generate valuable knowledge and new insights for each of the stakeholders involved in the pilots. In addition, the study of the different possibilities to obtain and optimise the accuracy of the predictions in different aspects of railway have given different options and points of view about the algorithms to use, and results obtained by the employment of the different algorithms differ in function of the aspect that wants to be upgraded. This means that some algorithms give better results for railway **management** while others give better results for **maintenance**.

Pilot Innovations

Among the innovative actions achieved in the different fields that this pilot covers, must be highlighted:

Big Data Innovation: Big Data Technology, Techniques and Algorithms used for the presented project are based on the **SOFIA2** Platform, a platform that brings an open source toolset for the Big Data exploit, based on the main standards and tools like Hadoop. Using standard technologies, the project evolutions are ensured without the need to change or modify the basic architecture to adapting to new communication mechanisms.

Predicted Models Development: the final goal that Predictive High-Speed Network Maintenance pilot wants to achieve is the development of a tool that predicts where maintenance activities are needed in order to optimise maintenance and management activities. The predictive models focus on three elements:

1. **Prediction of the track profile degradation**.
   Different sources from track profile degradation variables are used in order to collect the most important information that affects the track structure. These variables come from the analysis that the different maintenance companies perform, highlighting dynamic inspection, geometric inspection and maintenance task. Two different severity thresholds are obtained from this analysis, and tracks maintenance management will experiment with an important optimisation in its schedule depending on the criticality of the different measures obtained.
2. **Prediction of the degradation of point machines**.
   Different sources from point machine degradation variables are used in order to collect the most important information that affects the point-machines structure. These variables come from the analysis that the different maintenance companies perform, highlighting movement's time, maintenance task and characteristic data.
3. **Optimization of railway operation** in the Rail Traffic Management System.
   By using the predictions obtained in the points described above, new data will be available to modify and optimise the railway maintenance and traffic management.

As the **main conclusion** of the Pilot, the evolution from traditional business management to a more predictive or prescriptive one, requires a deep cultural and technological change that must be undertaken without delay to ensure the competitiveness of companies.

2.4 Ports as Intelligent Logistic Hubs

The Port Authority of Valencia (PAV), also known as Valenciaport (**initial pilot**), is the public body responsible for running and managing three state-owned ports along an 80km stretch of the Mediterranean coast in Eastern Spain. The PAV is responsible for managing the ports of Valencia, Sagunto, and Gandia in line with the model implemented in the Spanish state owned port system, in which the port authority provides the areas and infrastructures that support port activity, whilst the private sector is responsible for carrying out operations and providing the equipment and services using the aforementioned infrastructure. Valenciaport is Spain's leading Mediterranean port in terms of containerised commercial traffic thanks to its dynamic area of influence and an extensive network connecting to major world ports. In specific terms, Valenciaport handled over 71 million tonnes in 2016, which represents an increase of 1.71% over the previous year and constitutes the highest ever throughput figure for the Port Authority of Valencia. In turn, container traffic, which represents the largest share of its throughput, grew by 2.32% to 4.72 million TEUs, thanks to good import-export and transit figures. The objective is to make the most of the data available for the three cases considered in the port: The **optimization of port operations**: improve transport and logistic operations by using all the available data to create models for forecasting and algorithms for optimization; The introduction of **predictive maintenance strategies**; and the creation of a predictive dashboard that gathers the entire information available (IoT platform, accesses, AIS, environmental, TOS) and provides useful indicators and their trends to end-users.

The Duisburger Hafen AG, duisport (**replica pilot**), owns and manages the Port of Duisburg. With a total handling of 3.7 million TEU3, duisport is the world's largest inland port. For this port and logistics location, the duisport Group offers full service packages in the area of infra- and supra-structure, including relocation management. In addition, the subsidiaries also provide logistics services, such as the development and optimization of transport and logistics chains, rail freight services, building management and packaging logistics. The duisport pilot will demonstrate the use of big data solutions for the proactive management of bi-modal terminal operations as well as for predictive maintenance of terminal equipment. This pilot will thereby assess in how far solutions developed in the Valencia pilot may be replicated and reused for the more challenging setting of the duisport inland port. Compared to Valencia port, the added complexity in duisport stems from the fact that the port is situated **in the middle of a large city** (with close to ½ million inhabitants) and at the center of Germany's largest metropolitan area, the Rhine-Ruhr metropolitan region (with close to 10 million inhabitants). This means that duisport has a multitude of roads, tracks and water ways that serve as entry and exit points for containers to and from the actual terminals and ports. In addition, roads need to be shared with many other cars within the metropolitan area.

Collaboration among the pilots has been an interesting approach to understand how different terminals work. However, their specific **logistic operative is quite different**: Duisport focuses primarily on train transportation whereas Valencia Port highlights vessel and truck transportation. To some extent, both pilots complement each other to provide an overall view of multi-modal transportation in port terminals. Even with different operatives, we found several **points in common**. Firstly, the different discussions for KPIs definition helped to find out common ground in the objectives for the **Terminal Productivity Cockpit**. Specifically, pilots found that the operative timing, train TT and truck TT where a common challenge to solve. Pilots discovered that in both terminals this sort of indicator is relevant from the operative point of view. Consequently, we focus our data gathering approaches and analytics scenarios in the **time series** area.

In the early stages of the project, pilots decided to change the initial equipment to consider for fulfilling the predictive maintenance scenario. In order to select which equipment to consider, both pilots met with the maintenance staff involved. As a result, **twistlock spreaders** were the main cause of failure. This **cross-evaluation task** has been helpful for the future post-project replication, as it is expected that similar issues are happening in other terminals.

Valencia Outer Port (Spain)

Profitable End-Users

The goals of the Pilot have focused on understanding the terminal processes and the maintenance tasks involved in such context. For that reason, the end user involvement and feedback come primarily from the **Noatum Terminal** end users, as they are the ones who could benefit directly from the pilot results. Next the end-user involvement is detailed according to the scenario they participate in.

In the case of the Terminal Productivity Cockpit, main end users are the **head of the operations department and one project manager**: in their daily work, one of their responsibilities is to monitor the different KPIs that describe the business processes in the yard terminal. Therefore, according to the data pipeline of the pilot, their role is to understand the visualizations of the data in order to transform insights into business decisions. Both of them have actively participated in the early stages of the pilot, defining requirements and relevant KPIs, how to measure them, providing detailed explanations regarding the data sources, how a terminal works and the underlying processes. During the project, they have actively engaged in monthly meetings in order to review the pilot progress and suggest improvements. Next, it is briefly presented some initial feedback regarding usefulness and ease of use of the Terminal Productivity Cockpit:

- The **dashboard** of the Terminal Productivity Cockpit is a remarkable improvement in terms of visual usability throughout interactive graphical charts. They soon requested the possibility to create custom charts from similar data.
- **Predictive models** usually provide a value as a result, but not how “accurate” are such values in the current context. Statistical indicators, such as RMSE, are not intuitive to understand in a business environment and they usually hide the error trend. They point out to report this accuracy taking into account the specific business context.

Regarding the predictive maintenance scenario, the main end user involved was the **head of the maintenance department:** he is in charge of supervising the maintenance operations and providing the maintenance KPIs on a monthly basis. From the early stages of the pilot, he participated in meetings related to the predictive maintenance scenario and the deployment of the infrastructure. Specifically, he recommended establishing the goal of the scenario in monitoring data for understanding why spreaders failed. He also provided access and detailed explanations regarding the maintenance reports' data source.

The pilot also involved **two technicians** who usually perform assistances of the faulty spreaders on a daily basis. Their main role in the pilot is to check the information provided by the predictive maintenance models. They will validate if the generated alerts and indicators are consistent with the assistance they perform.

Data Visualisation and User Interaction

Data visualization has been an essential point to develop in the pilot as the stakeholders requested a high usability degree. From the technological point of view, we follow two approaches: 1) The development of personalized web responsive component and charts using Javascript frameworks and 2) the use of a business intelligence-oriented solution for the definition of custom visualization. The first approach provides the flexibility to implement rich interactions and seamless integrations with the underlying predictive models. The latter was included as several stakeholders pointed out the possibility to adapt the visualization to their own preferences. During the pilot, we combined both concerns in the context of the terminal productivity cockpit. Next, we present four visualization approaches we fulfilled in the pilot implementation and their corresponding lessons learned.

Predictive metrics: one of the most interesting points of the pilot was to provide predictive metrics using ML models combined with the latest historical data. These metrics address mainly operative information as the time to fulfil an order, the classification of containers in the yard, the assigned resources (cranes, truck) in a specific period and so on. Briefly, historical data of each metric is used to create a predictive model. Then, data is gathered in real-time from current systems and the predictive model provides a trend for each metric for the next hour, shift, day or week. One request of the stakeholders was to understand how accurate the model was. This is a tricky request as statistical indicators are not trivial to understand without a mathematical background. Therefore, our approach was to provide visualization on how the model performed in the past. In other words, perform prediction in a point of time in which we already know the actual value of the indicator.

Optimization of real-time operations: these visualizations focus on showing information in real-time from the terminal in order to monitor and, optimize, the yard operations. The main difference with other visualization is that they use a data stream directly from the infrastructure and perform fast calculations of the indicators. We have applied this visualization in the context of two user interactions. For example, new data of each crane (mainly GPS position and current operation status) is received on a one-second basis and the visualization is automatically refreshed every five seconds. This visualization is particularly useful for yard planners as they get on a quick view of the resources currently available and the specific jobs they are performing.

Another optimization goal of the pilot was to show the best possible order of processing of containers in a yard block to reduce the time trucks should wait in the terminal to pick up a container. This best order is calculated using a genetic algorithm that is executed on demand by the yard planner. The algorithm takes as input the current trucks waiting for delivery/pick up a container and minimizes total delay over an expected threshold. An interesting remark of end-users is that they were not interested in knowing the exact amount of delay, but to check easily if some trucks have a huge delay in reacting accordingly. Taking into account such need, the visualization proposed to meet their expectations.

Predictive maintenance: regarding the predictive maintenance visualization, their main goal is to show historical information of the assistance required by the STS cranes. This information is currently available in excel-like reports and the specific part to be fixed is introduced using natural language or acronyms. Therefore, it is not easy to get indicators as the number of maintenance assistance associated with a twistlock failure. To create a suitable visualization, first, we aggregated the information of maintenance assistance and we automatically labelled them using a “bag-of-words” algorithm and the classification of the maintenance staff. Maintenance staff found this visualization useful not only for reporting, but also to understand better the reasons for the failure.

Another relevant visualization of this scenario is to understand when a **twistlock** fails, i.e. when it requires maintenance assistance. The pilot gets information regarding the different timestamps for each twistlock open/close movement. In this context, the overall idea is to highlight anomalies: if a twistlock open/close timing is not aligned with the rest of the twistlocks or if the time of the movement is longer/shorter. Such anomalies point out to an issue like damages in the container, a harmful operation or an inner mechanical issue. From a technical point of view, this visualization is challenging due to the amount of point to be sent to the web interface. Our approach was to draw a single point if the timing of twislocks was the same (being the red twistlock the reference) and then, include only the anomalies. This visualization received positive feedback from the maintenance staff, as they used it indirectly as an alert system when the timing of any twistlock is not regular.

Data Analytics

For data analytics, the pilot considered two different scenarios with a different set of techniques:

Port and Terminal Productivity Cockpit

In the context of the terminal productivity cockpit, the pilot selected 77 indicators related to the performance of terminal operations. The main objective from an analytics perspective is to predict the value of such indicators at different time aggregation levels. The indicators raw data is available for every hour and we generate a prediction for the next hour, shift, day and week. As training data, Pilot uses mainly information of the year 2017 and we use the year 2018 as test data. The indicators prediction is considered as a time-series forecasting problem, in which past values of the indicators influence actual values.

After performing some feature engineering techniques, the pilot tested several techniques using the additional data includes. One of the best techniques in terms of accuracy and computational time required was the use of **Random Forests** (RF). The best advantage of RF is that they are **less sensitive to hyperparameter** tuning than Support Vector Machines and Neural Networks, for example, and, in general, they are faster to train. The main disadvantage is that they are not intended for time series analysis. In that case, Pilot had to perform a feature engineering step that consists in generate lagged variables. In the context of time-series forecasting, there are methods like Auto-Regressive Integrated Moving Average (**ARIMA**) that are more suited for this sort of problems. The main disadvantage of ARIMA models is parameter tuning, which requires a trial and error procedure and is difficult to identify the values of its parameters using autocorrelation and partial autocorrelation plots.

Other methods that are designed for time series analysis tested in the pilot were Recurrent Neural Networks (**RNN**): a type of Neural Networks in which the connections between the different artificial neurons of the network are circular. However, there are studies that indicate that this type of networks is not capable of capturing long-term dependencies. For this purpose, new architectures have been developed, such as the Long Short-Term Memory Networks (LSTM), which are capable of learning long-term temporary dependencies.

Predictive Maintenance

A baseline approximate survival analysis model has been explored based on a dataset consisting of the registered maintenance operations of a single crane. Said dataset contains the timestamp of the request, start and end of each operation. It also includes manually filled-in descriptions of the causes of the failures and other discarded data. The failure type can be inferred from the free-text field, but it should ideally be categorized into discrete codes for ease of analysis. In the current implementation, the dataset is analyzed both as a whole and isolating those incidents that mention the twistlocks. This dataset contains information about 7381 maintenance operations on 27 cranes in the yard from 2014/01/10 to 2018/12/08. Of these, 1852 are related to the twistlocks. By subtracting the operation end time from the request time, the operational lifetime between the failures is obtained, allowing the construction of survival models. The nonparametric Kaplan-Meier estimator is chosen to obtain the empirical conditional time-to-failure (TTF) function. Other parametric functions were also tested, namely Weibull, exponential, log-logistic and log-normal, but their fit quality is deemed insufficient and therefore are discarded. The a priori median lifetime for the whole dataset is of 26.85 hours, while for only twistlock related failures it is 132.04 hours. However, it has been proved that the observed risk is greater in the early region of the lifetime but soon loses most of its slope. This means machines that survive early failures are more likely to survive for far longer. Additional data (temperature, humidity, rain, TEU weight, etc.) would help characterize the condition of the machines. The main objective is developing a set of observables that can aid the **prediction of machine failure** states, trying out different techniques that can be applied in a production environment.

Data Processing Architectures

One of the main challenges of the pilot was to deploy an infrastructure scalable enough to deal with the data coming from the terminal equipment: cranes (RTG, STS) and trucks. Data is generated on a second basis and sent to a central server using a wireless infrastructure deployed inside the cranes/trucks. The data ratio, which ranges from 50 to 200 messages/sec, is not the only challenge. Due to the harsh working conditions in the terminal, network connectivity issues are quite common. In that scenario, messages are temporally stored until the network is available again, but generating peaks of messages to be sent. We register peaks of more than 10.000 msg/sec but currently the infrastructure scales up to 500 msg/sec because a delay of several minutes in such scenarios is admissible. Next, we summarize the main components that support such infrastructure and why they were selected.

- **Data broker**: this broker implements a scalable solution to support the data streaming challenge. The pilot has selected **Apache Kafka** to implement this component and currently it is deployed in a two-node cluster (4 cores and 32GB of memory per node) on the CSP Spain terminal premises. Using the broker, we have a set of topics for supporting each stage of our data processing pipeline: raw, formatted, pre-processed and processed. In other words, Kafka role is two-fold: establish a scalable entry point to the system and coordinate the data flow in the processing pipeline. Additionally, Kafka has been a useful tool for supporting the historical storage of raw data at the entry point of the data pipeline.
- **Distributed event processor:** this component performs live transformation over the messages received. One interesting lesson learned from CEP is that rules could be modified without restarting the process entirely. So potentially the data processing could be modified in live.
- **Historical Database**: a **Cassandra** cluster has been set up to be the historical database of all the data from the terminal equipment. Cassandra has been a great choice for storing time-series data providing simultaneously responsive reading and writing times. An interesting point is that Cassandra compacts the tables periodically. This a critical feature,

as we usually received a lot of attributes without changes between two messages. So, in a database without compaction of repetitive values, our current data flow will lead to huge storage requirements.

- **Metrics/Indicators Database**: Cassandra response time is enough to support analytical tasks. But because of the amount of data, the response time of the queries required by the cockpit are far to being optimal (30 – 60 seconds). Additionally, later versions of Cassandra are schema-oriented, so it is not a good database system to store schema-flexible data.

To avoid a lot of network traffic between components, our initial approach was to combine the broker and the historical database in the same cluster. However, this quickly generated huge partitions and memory errors in the cluster. As a solution, Pilot stored data on daily partitions per each equipment to solve this issue. A lesson learned is that the physical design of the data stores in Big Data environments is a critical task. As it has been said previously commented due to the design of the infrastructure, frequently the network communication between the cranes and the central server is lost. This issue generates a randomly huge amount of messages to be sent. This increment puts a lot of stress in the broker but the real bottleneck was the competition of Cassandra and Kafka for resources, specifically for writing data to the hard disks and memory. The solution was to define a bigger cluster with Cassandra nodes isolated from the Kafka ones.

Another infrastructure challenge worth mentioning, was to gather the information for the **twistlock open/close movements**. The initial approach was to setup an encoder to understand the movements of these locking devices. Initial tests on a controlled environment were promising as we detected relationships between the delay of the open/close state and the risk of failure. Therefore, Orbita tested three different designs to install the encoders in the crane to gather live data. However, any device installed to obtain measurements close to the twistlocks was damaged due to the impacts and the harsh working conditions of the crane.

As a workaround, we used the spreader PLC that receives the open-close order from the crane cabinet and checks the status of the twistlocks (open or closed) getting the time lapse since the twistlock starts a movement until the change of state is detected by the PLC with a precision of 10 ms.

Data Management

It is worth to mention that in the port domain there are common interoperability mechanisms, for instance for sharing customs documentation. However, in the context of business operations, each organization has its own information systems, usually proprietary and without easy interoperability mechanisms. Integration in our pilot was supported in a high-level of abstraction, i.e. pilot only integrate data related to business indicators and KPIs. Our approach was to use an indicators database flexible enough to deal with such different data sources. This option avoids complex ETL processes required to integrate data in relational databases. Of course, quality is compromised, mainly because of missing values and timestamps correlation, but it was enough for the pilot requirements.

Pilot Innovations

The pilot has introduced some clear improvement from the use of Big Data technologies. The most interesting advantage is in the context of **historical reports**. One of the pilot results has been to create an integrated data view from the whole set of historical data. Now this view can be queried from a single point, so it is easier for them to generate reports or export data to excel for further analysis. This result has improved their daily reporting and monitoring tasks.

Regarding **crane monitoring** the pilot is now able to monitor up to 150 machines and to store up to three months of data. This is a huge improvement, since the previous system was able to handle around two weeks of data. Another interesting point is that the data infrastructure has been designed to be scalable in the future. Therefore, this pilot is the starting point to define a potential solution to terminal equipment monitoring of the whole terminal. Thanks to Big Data technologies, Pilot monitors the equipment that fulfils such processes and could implement analytical applications on top of such fine-grained data. This is especially relevant for maintenance operations.

Port & Terminal Productivity Cockpit:

TPC is a web-based tool for supporting a common set of shared and specific metrics to evaluate the logistics process from the port access gate to the port terminal. The Pilot uses Big Data architecture to integrate information from several stakeholders involved in the delivery of the containers and create KPIs. The defined KPIs provide benchmarking capabilities (e.g. related to costs and performance) that may indicate different levels of competitiveness. The analytics models have been developed using machine learning techniques and the whole set of the historical data available. Therefore, these models provide insights about current operations, thus helping stakeholders to understand how to improve the related KPIs. Main features of the solution are:

- A real-time dashboard to display relevant information for decision-makers and, hence, improve current resources planning. This dashboard is supported by a user-friendly interface
- Improvement of the planning and execution of port and terminal operations using historical information
- Optimization of asset utilization, specifically reducing maintenance work. A maintenance request implies a minimum crane downtime of 30 minutes. Knowing beforehand such scenario will reduce costs and assign a most suitable resource before failure.

- Increment of productivity and efficiency, translated to moves per hour increase and cost per move reduction.

Duisport Inland Port (Germany)

Profitable End-Users

The identified major stakeholders for the Duisport pilot are (1) the **terminal managers**, which manage the operative business at the terminal and (2) the **maintenance managers**, which schedule and manage the maintenance sessions of the terminal equipment.

Based on demonstrations and structured interview sessions involving the terminal manager one key point was raised: given the **amount and diversity of data available** for the TPC Terminal Productivity Cockpit), the manager felt overloaded by the amount of information displayed in it. So far it displays every relevant information available. However, the terminal manager suggested only providing information that could indicate a problem and its root cause. Addressing such **data overload** (in the presence of the availability of big data) thus can be considered an important lesson learned.

Data Visualisation and User Interaction

The TPC offers a balanced combination of ad-hoc visible information, such as traffic lights, and interactive selections, such as drop-down boxes and mouse over information. Based on these two values the dashboard calculates the adjusted days to next maintenance. To enable the user, to get a quick impression about the health of the devices, traffic lights indicate specific values. Finally, even though the terminal only covers a small geographic area, the terminal operators preferred visualization of information also using maps, as this allows them to better match virtual data with the real world. By way of example, the container movements in the terminal were visualized using heatmaps.

Data Analytics

Predictive Terminal Process Analytics: Pilot uses RNNs as base learners for predictive analytics solutions. The main reason being that RNNs can handle arbitrary length sequences of input data. Thus, a single RNN can be employed to make predictions for business processes that have an arbitrary length in terms of process activities. In contrast, other prediction techniques (such as random forests or multi-layer perceptrons) either require training a prediction model for each of the checkpoints or they require the special encoding of the input data to train a single model. However, these encodings entail information loss and thus may limit prediction performance. The pilot uses RNNs with Long Short-Term Memory (LSTM) cells as they better capture long-term dependencies in the data. However, RNNs also face specific challenges when used for predictive process monitoring. Even though the data that is fed into an RNN is sequential, i.e., a sequence of events, these events represent the execution of business processes which may include loops and parallel regions. Such non-sequential control flows can make a prediction with RNNs more difficult, as RNNs were conceived for natural language processing, which is sequential by nature.

To address these difficulties, Pilot employs the following two solutions[2]. First, instead of incrementally predicting the next process event until we reach the final event and thus the process outcome, we directly predict the process outcome. Thereby, we avoid the problem RNNs may have in predicting the next process activity when process execution entails loops with many repeated activities. Second, we encode parallel process activities by embedding the branch information as an additional attribute of the respective process activity. Thereby, the pilot addresses the problem that parallel process activities can make the prediction task more difficult.

Finally, it should be mentioned that RNNs (as a deep learning technique) provide an additional benefit when applying in practice. Based on our empirical evidence, RNNs work well even without an extensive hyper-parametrisation[3].

Predictive Equipment Maintenance Support: the first step in data analysis for Predictive Equipment Maintenance Support is to clean the data of anomalies. Anomalies are mostly caused by sensor malfunctions and specific values that indicate that there is no value at this point. Another important problem to solve was timing. The different machines which collect the data were configured with different time zones, although they all are physically located in the same time zone. To analyse the data with regards to our predictive maintenance aims, we defined factors that hold more information about the problem. The predictive maintenance main problem is the **time of replacement of the twistlocks**, a wear part that is used to transport containers with a crane. This is an important lesson learned to bear in mind for further actions in Ports.

Data Processing Architectures

The data collected at Duisport is sent every minute to an SAP HANA database at Smart Data Innovation Lab (SDIL) in Karlsruhe. The data consists of four tables, one table stores the **state data** that is collected every five seconds, and it consists mainly of sensor values. Another table stores **statistic** about every container transport the crane executes and another stores

1.2. [2] A. Metzger and A. Neubauer, “Considering non-sequential control flows for process prediction with recurrent neural networks,” in 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2018), Prague, Czech Republic, August 29-31, 2018, T. Bures and L. Angelis, Eds. IEEE Computer Society, 2018, pp. 268–272.

1.3. [3] A. Neubauer, “Predictive Process Monitoring for Transport Processes with Recurrent 2018 Neural Networks”, Master thesis, University of Duisburg-Essen, 2018.

operational data. The last table stores all faults and errors that occur, this was important for the development of their PMML model. One part of the data processing happens through views in the SAP HANA database. The other part is realized through SQL queries sent from Apama and Mashzone Next Gen. Apama periodically requests the most recent rows of data and feeds them to the Zementis plugin, which holds the PMML model pilot developed. The prediction made by our model, is sent to Mashzone Next Gen along with the row that was used for this prediction. Mashzone Next Gen requests data from different views based on user input. It further processes this data in data feeds. For example, the days to next maintenance are calculated by querying the average number of moves per day from the SAP HANA database and combining them with the number of moves a pair of twistlocks can endure. This is done in a data feed inside Mashzone Next Gen.

Pilot Innovations

Terminal Productivity Cockpit (TPC): analysing port logistics in a smart manner can improve port operations to save both time and money. TransformingTransport is piloting the concept of the Terminal Productivity Cockpit (TPC), which exploits advanced data processing and predictive analytics to facilitate proactive decision-making and process adaptation, ultimately achieving more efficient port operations. In particular, the TPC leverages cutting-edge predictive business process monitoring solutions, i.e., real-time predictive Big Data analytics for terminal processes based on deep learning technology. The TPC features a web-based dashboard that presents data it has accumulated and processed to decision makers. The dashboard was continuously improved by considering feedback from several user studies. It displays the current state of the terminal, the predictions from the predictive models and several KPIs computed from live data for comprehensive decision-making support.

2.5 Smart Airport Turnaround

By means of the use of big data, pilots are transferring two dogmas to the real world to gain benefits in terms of operational efficiency and saving costs: **Pro-active disruption management** by analysis of historical data and comparison with actual real time data. This allows the prediction of unfavourable situations and the early deployment of counter-measures to avoid them; **Optimizing operations** across all involved domains through processing available data ac-cording to business goals. In general, Athens Airport (**initial pilot)** focusses on the full passenger process, analysing and describing the passenger behaviours in order to anticipate the number of resources required to manage the expected volume of passengers and predicting when passenger process might affect the aircraft departure times. Malpensa Airport (**replication** pilot) is going to monitor the total turnaround process, analysing and predicting ETAs, predict turnaround processes and aircraft delays, and making a link with passenger process only at the boarding gate.

The main idea of the initial and the replication pilot is to leverage big data analysis to boost capacity by more efficiently using resources at airports and to find new business opportunities in areas with passenger presence. One of the major success factors of the domain, for both the initial as well as the replication pilot, has been the **collaborative approach** from the early beginning on. The aim of this collaboration was to determine the real business need as the driver for activities regarding **data analytics** and **model development**. End-user and partners jointly identify the areas to be improved and subsequently they discussed and agreed on the possible solutions to address the needs.

As the team had experiences through earlier data analytics activities, they expected a considerably high effort for data selection and data cleaning. Thus, high focus was put on the thorough selection of an effective database and the isolation of accurate and completeness of the data to build the models on. Therefore, the team performed a **weekly assessment of the data gathered**. This allowed the execution of measures when identifying noticeable deviations amongst the predictions and to quickly fix any inconsistencies by applying immediate remedial maintenance.

• Smart Passenger Flows in Athens Airport (Greece)

Profitable End-Users

In the next lines, the different actively involved end-users that has been part of the pilot describing the contribution to the project are shown.

Corporate Information & Market Intelligence Unit: its roles include airline route development, customer profile and satisfaction, commercial development, corporate quality, corporate image and tourism. Its engagement is about sharing information about the typology of passengers and their behaviours that are complementing the surveys they do to the passengers.

Airport Security Screening: to ensure the efficient and effective security screening of passengers and ensure passenger processing service levels (SLAs) by third-party contractors. Its engagement is about sharing information about the optimum allocation of security personnel that will allow the efficient processing of passengers while balancing passenger service levels and cost of required security personnel employed.

Retail Services: to optimize the usage of retail space at the Airport and maximize benefits for all stakeholders i.e. Airport, concessionaires and passengers. This end-user shares information about the segmentation of passengers and their behaviours to provide passenger-related insights both to airport's retail services management and to concessionaires. This will enable the development of actionable strategies with the aim to maximize non-aeronautical revenues. Speaking generally, the usability of the retail dashboard design is good and the results of the predictions of passengers in security lines is very similar to reality.

Airport Operations- Aviation Scheduling and Planning function: to ensure the safety and efficiency of aircraft operations and promote passenger experience levels. It shares information about the transfer passenger's arrival and departure information to calculate minimum connection times to optimise aircraft positions. Also provides insights to Airlines about late show passengers at Gates.

Pilot Innovations

Passenger Flow Descriptive Algorithm: this algorithm has allowed the Pilot to understand the different variables and its relationship in passenger's behaviours. In business terms, the main market benefits are the attraction of new customers to optimize their airport resources, reducing operational costs and therefore increasing revenues, and improving the passenger experience.
The passenger flow descriptive algorithms have provided us with the prediction of passengers estimated arrival time, provision of models to understand the internal flows and movement across airport terminals, has enhanced the allocation of airport resources according to the expected demand and allows more direct communication among Airport and Airlines. In analytics and visualization, innovation has allowed us to manage global statistics to show predictive and real arrival of passengers to the airport, integrate all this information in dashboards and cockpits and the use of KPIs for measuring the performance of airport operations.

Operation Management Predictive Optimization Module: this module has exploited predictive analytics with Passenger Flow data obtained in real time from airport and airline systems in order to facilitate proactive decision making in real time whenever there is any disruption over the initial plan. The operation management predictive optimization module provides us with a prediction of passengers estimated arrival time, provision of models to understand the internal flows and movement across airport terminals and has enhanced the allocation of airport resources according to the expected demand.

• Smart Passenger Turnaround in Malpensa Airport (Milan, Italy)

Profitable End-Users

The main end-user for Malpensa Airport Pilot is **SEA**. As well, the **market (clients)** itself has been considered as the most important end-user for the well-being of a business.

Once the TT solution will be proven effective and fully validated, SEA will be interested in applying the project results in the perspective of its airport-system optimization, integration of information and improvement of airport activities flow. With the purpose of translating the KPIs selected in possible improvements of the operational efficiency, thus making the results of the analysis as close as possible to the airport reality, it is necessary to **extend the range of target values by selecting specific variables** to be applied to the TT algorithms. In this way a sort of "operational mitigation" will be implemented. To reduce the complexity of the model, the Partners have agreed on the side factors the models need to take into consideration by prioritizing the most relevant factors for the Airport operations:

- Consider some a/c stands limitations (i.e. type of aircraft allowed in a particular stand);
- Schengen/Non-Schengen Area
- Weather limitations in case of fog – snow – ice – heavy rain

From an airline point of view, **digitization** not only bears an enormous potential for quantifiable **cost savings**, but also for qualitative benefits which are not directly quantifiable, but which need to be considered in the overall perspective of smooth operations and customer loyalty. Any delay, misconnect or technical problem spreads through **social media** more quickly than the airline can comment on it. While this constitutes a factor for additional pressure in case of disruptions, it is also a chance to advertise, without any additional cost, the own achievements in reliability and punctuality by the most credible voice on the **market**: the **end customer itself**. As soon as the more accurate times in the system will be used for planning purposes and other systems will consequently be linked to the new planning to determine and minimize downstream impacts, the non-quantifiable impacts will arrive at the customer and manifest increased passenger satisfaction and rising positive airline reputation.

Data Visualisation and User Interaction

Malpensa pilot has used data visualization in different approaches:

- Visualization of the predictive algorithms results for understanding and evaluation.
- Real-time visualization of the predictions for demo/live purposes.
- Providing production result for end users.

For providing results for production (shadow system in this case) the approach has been a seamless integration in existing user interfaces, so the users can consume Big Data results in the way they are used to. So, in the very same format the data from current systems was presented to users, new columns in the current shadows system are presenting the data produced by the data-driven algorithms. It is important to have not tried to change the way the users are working to avoid changing working procedures or introduce disturbance to end users.

Since Malpensa pilot predictions work in real time as flights arrive and depart from the airport, a nice way to show live how the system works has been to produce a map in which the real flights are presented and the user can pick a flight and see the algorithm's results in real time.

Data Analytics

Malpensa pilot has used some descriptive analytics, mainly for the proper understanding of the existing data and the business process and, as the main contribution, has developed **predictive analytics** leveraging **machine learning**.

Descriptive analytics have been a building block for assessing the quality of the data, detecting any error or mistake in the data management components and presenting the existing performance of the different systems; taxi times real distribution compared to current tables in use, or differences among scheduled times and actual times are samples of the last. The pilot has relayed in well know analytics packages in **R** and **Python** to prepare the descriptive analytics. It has been key to exploit the **descriptive phase at the beginning of the project** to understand the data and be prepared for more challenging predictive analytics. The predictive analytics have been based on training models with machine learning algorithms (particularly gradient boost). It has been key the amount of data available, and along the project new ground truth has been collected to increase the size of training datasets enhancing the performance of the new models. Feature selection techniques have proven to be too effective and models have increased performance after the process. The availability of a framework to use the models in real time was important for the pilot too, and the use of **H2O** framework has been a great success.

Data Processing Architectures

Malpensa pilot most challenging data processing comes from:

- The training on models with very large historical datasets (data-at-rest).
- The real-time processing of the surveillance messages of the airborne flights heading to the airport (data-in-motion).

For the data-at-rest processing the pilot relies on in **Hadoop** (Cloudera distribution) and in **Hive** for preparing the dataset (filtering, merging, etc…). For the data-in-motion processing the pilots use Apache **Flume**, Apache **Kafka** and **python**. The use of Apache Kafka has decoupled the data distribution from the processing making the pilot have great stability.

Data Management

The main challenge in Malpensa pilot about Data Management has been **making the predictions produced in the Big Data platform available** to the shadow system. To this aim, two approaches have been used:

- For the continuous predictions (i.e. ETA) a stream (TCP socket) has been opened from the shadow system to the big data cluster.
- For the request-reply predictions (i.e. taxi times) several HTTP based APIS has been developed.

In both cases the setup of a VPN was requested as a first basic security measure. The stability of the stream has proven to be more difficult to maintain, due to network timeouts and failures to reconnect from time to time. A good practice has been a shadow system's weekly feedback form showing what data was received, to allow checking the completeness and help to debug the connections.

Pilot Innovations

ETA analysis and prediction algorithms: this is a system allowing the analysis of recorded big data to derive a predictive model that can deliver more accurate expected landing times for a current situation. For operational airport planning systems and airline operations centres it is essential to have early predictions for their flights to arrive at their destination. This depends on flight plan, routing, weather and traffic situation. The predictive model is capable of comparing historical data with current data to calculate a prediction for a touchdown.

Taxi times analysis and prediction algorithms: this is a system allowing the analysis of recorded big data to derive a predictive model that can deliver more accurate expected taxi times for a current situation. For operational airport planning systems and airline operations centres it is essential to have early predictions for all flights to arrive at the gate at the destination. This depends on the actual time of arrival, taxi routing, gate location, construction sites, weather and traffic situation. The predictive model is capable of comparing historical data with current data to calculate a prediction for the estimated in-block time.

Boarding times analysis and prediction algorithms: this is a system allowing the analysis of recorded big data to derive a predictive model that can deliver more accurate expected boarding times for a current situation. For operational airport planning systems and airline operations centres it is essential to have early predictions for the duration of the boarding time of all flights at the departure location. This depends on gate location, aircraft type, type of flight (more business or leisure related), number of passengers and airline boarding policies. The predictive model is capable of comparing historical data with current data to calculate a prediction for the estimated boarding time for a given flight.

2.6 Integrated Urban Mobility

The **initial pilot** is located in the city of **Tampere** in Finland. Tampere is the third largest city in Finland, and the largest inland centre in the Nordic countries. Tampere has a population of 231 853 inhabitants1 and about half a million inhabitants in the Tampere Region. The pilot aims to improve the situational awareness regarding traffic context with big data methods, and by providing tools to the traffic management centre. The second objective of the pilot is to provide solutions for urban logistics, taking into account the new parking policy, which will reduce the number of parking places in the city centre considerably. The following partners are involved in the Tampere pilot: VTT, Infotripla, Mattersoft, Taipale Telematics and the City of Tampere.

The **replication** pilot is located in the city of **Valladolid** in Spain, the capital of the Castile and León region. The main objective of the pilot is to generate a traffic model for specific areas in the city where freight transport has more impact, to analyse different freight delivery scenarios, and to create a planning tool for delivery fleets. The following partners are involved in the Valladolid pilot: CARTIF, PTV, TomTom, LINCE and the City of Valladolid. The replication pilot hence is more concentrated on modelling of traffic using big data analytics in order to support decisions made by the City council in order to keep urban logistics viable. The initial pilot uses real-time data sources to support urban traffic management centres, drivers and travellers in their daily task. For this purpose, the pilots make use of the data, which is available.

Again, one of the main barriers for a collaborative was the different approaches of both pilots: **Tampere** Pilot was focused on developing tools of up-to-date information on traffic provided to Traffic Management Centre and citizens in real-time. On the other hand, **Valladolid** Pilot is focused on traffic simulations models that need reliable knowledge based on updated traffic and logistic data, but not in real-time. Despite this, some lessons have been learned from each other along with the project.

Following actions done in Tampere, Valladolid is analysing the viability to acquire an app or similar that permit them to know in real time the use of loading/unloading parking places, mainly those located in the city centre. Following Tampere pilot activities within TT Project, Valladolid City Council also showed interest in the analysis of the images of the traffic cameras from Valladolid, in order to examine congestion or to discriminate the type of vehicles. In this specific case the quality of the image was not good enough, so no more actions took place. Again, the **updating of the technology** is a must for well-proved replications.

The work performed in the Valladolid pilot results in tools for the city council to select the parking places but requires data regarding freight transport in the city. Currently, there is **no data available** regarding the amount, frequency and geographical distribution of freight delivery in the city centre of Tampere. The parking application would be able to collect statistics regarding the use of the service, and hence could provide input to a model, similar to the micro-level model developed in Valladolid.

As the main lesson learned here, **open data availability** should be a priority in order to take advantage of the current technology to improve the efficiency and accuracy of the models, which should be trained and run with as many information as possible.

For Tampere, the work on traffic simulation can be used to optimise the location of the parking places for the urban freight application. However, at this moment there are large infrastructure **works ongoing** in the city centre, such as tramline construction, which will have a large impact on the future road network and traffic demand.

An important lesson is that piloting tasks should be based on these inconveniencies from the start in order to design **contingency or alternative plans** in case these hurdles should be faced.

• Integrated Urban Mobility and Logistics in Tampere (Finland)

Profitable End-Users

Traffic expert at the city of Tampere and the operator at the urban Traffic Management Centre: the operator at the TMC was interviewed prior and after the pilot. The tools, developed in the project, were demonstrated and made available for them. Through the tools, they get better information on the status of the traffic in the city, and the tools support them in decision making. The tools give a quick overview and alert on the general traffic status in the city and on changes in the traffic. This information can in the future also be shared with other stakeholders than the authorities. Also, the dashboard was especially seen useful for mid-size cities, which do not have their own Traffic Management Centre. The tool is a much cheaper solution than a fully equipped TMC.

The TMC operator gave some ideas for further development, mainly related to the combination of information coming from different sources (e.g. information on disturbances with nearby traffic camera views).

Logistic operator: the freight delivery company Niinivirta was very motivated to contribute to the development of the system and to test the delivery parking system. They recognise a clear need for this kind of low entry and operator equal service. Real-time parking reservation plays also an important role by the goods delivery customers of Taipale and is seen as a very useful feature in day to day work. Also, the custom geofencing property is seen as a very useful tool for tracking, e.g. loading and unloading events in a terminal area.

Data Visualisation and User Interaction

The Tampere pilot developed tools for **automation of the information sent to travellers**. Only critical validated events are sent to travellers. Transmission of erroneous data and information overflow have a negative impact on service acceptance by the travellers. Automation of the transmission of validated events reduces the workload of the TMC operator and frees resources for the efficient management of these events. It also allows informing the travellers outside of the working hours of the urban TMC.

Data Analytics

The quality of the data has to be taken into account in the analysis of the data. This includes:

- **Missing data**: the frequency of the data delivered may be higher than specified in order to identify traffic disturbances. If there are gaps in the information provided, analysis is not possible anymore. Regarding missing event information data should be available during all the workflow of the event, being reported not only at the beginning but also at the end of the disruption.

- **Data accuracy**: the resolution of the data may be too low for analysis purposes. Also, camera pictures may only be available in low-resolution formats, in order to avoid data privacy issues.
- **Timeliness of the data**: delays in information networks can have multiple reasons, including network component overload or failure or communication failures. Timeliness becomes more important if the systems are more automated
- **Open data**: data which is made available as open data may be of lower quality than data which is made available through bilateral contracts. Reasons are, in addition to commercial aspects, to reduce the amount of traffic, e.g. there may be restrictions on the frequency or extent of data requests. The resolution of the quality may be lower, e.g. for cameras to avoid privacy and bandwidth issues. Only processed data may be made available.
- **Data from social media**: only a low portion of social media contains location data, and this data then mainly comes from commercial sources (e.g. other traffic information providers, event organisers). It is very hard to find good keywords for extracting social media, as all words have multiple meanings or can be used in a different context.

Data Management

In order to maintain the quality of the data, a **data owner** is needed which is responsible for data management and maintenance. Tasks include assuring that all data sources are working (e.g. all traffic loops collect information), that all interfaces and servers are working and all material is up to date, and representing the real world. For instance, changes in the traffic environment or in public transport routes, timetables or bus stops are reflected in the data.

Access to data

Vehicle data: floating car data, produced by the vehicle, contain valuable information for traffic management: from the position and speed information sent by the vehicles disturbances in traffic can be rapidly detected, e.g. traffic jams. A major problem related to floating car data is the privacy of the data, as in-vehicle data (especially the combination of location and time) is to be considered as personal data[4].

Traffic cameras*:* as the accuracy of traffic cameras is increasing, it becomes easier to identify individual traffic participants, e.g. through the identification of vehicle license plate or identification of pedestrian through face recognition. Cameras in the street network should be installed so, or the images modified so, that they are not aimed at residential buildings. Due to

[4] McCarthy, M Seidl, S Mohan, J Hopkin, A Stevens, F Ognissanto, Access to In-vehicle Data and Resources, European Commission, May 2017

privacy concerns authorities are reluctant to make data available. Video is extremely hard to get and camera data are only provided in low resolution so that vehicle license plates and pedestrians cannot be identified.

Open Data License*:* several data sources used in the Tampere pilot are made available as Open Data. The terms of service allow free use and distribution of data, provided that the producer of the data and the link to the license is provided with the data. An issue is on how to apply the attribution for real-time messages, which may have a limited length (e.g. twitter feeds). The preferred solution is to add the attribution to each message, but this may require a rather large part of the message payload. The selected solution is therefore to include the attribution with the license information and link clearly with the service information, e.g. a website home page and Twitter feed home page, and in the service documentation.

Other Non-technical Lessons Learned

Policies: Big Data can support the city in its policies. Through Big Data authorities can get more information on transport and support the authorities to assess potential measures to achieve its goals, e.g. to reduce emissions. In Tampere, the parking policy is to reduce the amount of on-street parking significantly. A solution has been developed in the Tampere pilot to mitigate the impact of this parking space reduction on goods delivery in the city centre. However, as reducing parking space is a very sensitive matter in local politics, there is a **long time between policy development and the introduction of active measures**.

Pilot Innovations

A fluency model for the city of Tampere, a city of about 200 000 inhabitants, has been developed, with information on the fluency in major streets in the city centre and on the road network around the city. There exist 4 main innovations:

1) A **Dashboard** with the traffic fluency status, the latest information sent, and views of the traffic at major critical points have been developed. The dashboard allows having the traffic manager a view of the traffic in the city at one glance. The fluency model uses a wide amount of traffic sensors, including loop sensors at traffic lights and the permanent traffic sensors at the national road network. The existing fluency model is remarkably improved by the use of floating car data (FCD) from the city buses. Compared to the models provided by global players such as HERE and Google, the model relies on more accurate data, such as traffic loops and counters. The dashboard for the TMC operator allows the pilot to have a view of the traffic status in a single view and have a view of all traffic cameras. The optimal **area of application** is the traffic management for small and medium-size cities.
2) **Approach for informing the public in real-time events through social media:** the method has been developed to automatically inform travellers on critical events in the road network using Twitter. The information is brought to the travellers in real-time, without

the need to install a specific app (other than the social media app). In the traditional approach, the urban TMC operator assesses all events, prior to sending out information to the public and other organisations. To ease the work of the operator and to provide travellers with more real-time support, a process which transmits automatically Twitter messages has been developed. The following information is sent automatically: access control restrictions to the tunnel, unexpected traffic events, such as accidents, reported by the ITM Finland or the Urban TMC. Also, information about roadworks or traffic fluency. Taking the service into use by the end user only requires subscribing to the twitter feed (and assuring that messages are shown on the phone). No installation of a specific app is required.

3) **Method for reserving parking places for urban freight:** the application, which is built on top of the parking management system of the city, contains information on parking places. Drivers and logistic operators can reserve parking spaces through the app. Parking wardens can verify through the parking management system the validity of the reservation.
4) **Methods for detection of traffic disturbances from open data:** during the project different innovations have been made regarding the detection of traffic disturbances, mainly from open data. The following methods have been developed: detection of traffic jam from traffic cameras, detection of traffic-related messages from social media (VTT), analysis of nearby traffic light loop detectors, analysis of bus delay between bus stops, analysis of floating car data for local events, analysis of acceleration data from vehicles and estimation of the possibility of disturbances based on directed acyclic graph model.

• Integrated Urban Mobility and Freight in Valladolid (Spain)

Profitable End-Users

Four main groups of actors have been identified as end-users in this pilot: mobility manager, e.g. city council, logistic operators, traffic modellers and mobility consultants.

Involved with **micro-simulation dashboard** there are two main groups: mobility manager and logistic operators. In both cases the objective of the tasks is to give support to them in the decision-making process.

For the **mobility manager**, e.g. Valladolid City Council, the main advantages of traffic micro-simulation models are:

- For traffic managers and planners, having these models lets them analyse and simulate in advance changes in the places reserved for loading/unloading activities, the extension of these zones or the type of zone (regular zone, strategic zone… which is related to the days and times that these zones are operative and reserved for freight delivery). Also, these models let them simulate the impact of pedestrianization of certain streets, as loading/unloading zones have a different regulation if they are in these pedestrian-only streets.

- Also, for traffic managers and planners, this kind of models are crucial to show politicians that decision-making process is backed up in data and objective criteria. This means that changes in mobility policies are firmly supported in analyses and the results and consequences are studied in advance.

For the **logistic operator**, e.g. Grupo Lince, the main advantage is related to the planning tool developed. From their point of view, this tool is an important support to optimize routes and delivery times and easily integrate into their work system. Nevertheless, drivers have reported an issue associated with the use of reserved places. Since the work schedule (from 8 a.m. to 2 p.m.) includes time slots of maximum affluence and occupation of these stops, sometimes it is difficult to stop in that places, so they have to look for other areas to stop. Moreover, from their point of view, they consider very interesting the possibility to know in real-time the occupancy rate (i.e. free spaces available) of reserved places for loading/unloading and/or to be able to reserve them in advance (which would make easier to optimize the routes).

Besides their feedback, end-users have also provided following suggestions to be taken into account for future versions of the developed tools:

- Improve the traffic micro-simulation model through the analysis of more logistic data, mainly **increasing the number of logistic vehicles monitored**.
- Improve the delivery planning tool taking into account, not only the addresses, but also the timeslots of final customer requirement.

Main actors as end-users of the **macroscopic transport modelling** results could be for instance city council (as mobility manager), PTV group (traffic modellers, software provider) or other mobility consultants. Traffic modellers for commercial customers can use, analyse and assess the OD-matrices for different modelling purposes.

Initial feedback from traffic modellers side can be given in terms of an overall assessment of the macroscopic approach. In that sense the pilot concluded that OD matrices from cellphone data have good quality and are adequate to improve current traffic models (based on a household survey) – depending on the purpose and regional focus. Initial efforts have to be spent in order to define the modelling purpose and needed data extent for analysing traffic demand in and around urban areas. Further research work is needed to investigate data sources from different regions and with specific relation to truck fleets which can be used then for models by applying the Big Data analytics approach developed and used in the frame of this project.

Data Analytics

Within this topic two main issues had been addressed:

1. **GPS track analysis** to discover behaviour in logistic routes (within traffic micro-simulation model)
2. **Data quality assessment** for mobile phone-derived data (within macroscopic approach)

Microscopic approach

GPS data analysis to discover behaviour in logistic routes, specifically where this kind of vehicles stops and for how much time, is a promising and valuable procedure within Knowledge Discovery from Data (KDD) process.

In the beginning, within our pilot, drivers indicated if their stop was due to a service, excluding from the GPS tracks other kinds of stops (traffic jams, refuelling, etc.) The issue came when we had to analyse more GPS tracks without this ground truth. Then, a review of the literature gave us some clues, most of them based on how to **detect data indicating 0 speed**. It is feasible when the granularity of the data is low, but in our case (where GPS data came from third platforms out of our scope) it didn't always happen. In the end, an ad-hoc stop detection algorithm was implemented which detects most of the service's stop. Specifically, when the duration is similar to the duration of a red traffic light.

As the main lesson learned within this issue points out that to know the real/reliable behaviour of logistic vehicles in the city centre it is needed have a monitoring device that allows measuring, not only longitude/latitude, but also real speed and/or other features as: engine condition, door opening, etc. Of course, a low frequency of monitoring is also necessary.

Macroscopic approach

The use of movement data coming from mobile phones for macroscopic traffic modelling turned out to be more complex than initially expected. The OD-Matrices had to be analysed in detail in order to assess the usability for different traffic modelling use cases. In the preparation phase before data provision it is highly recommended to define the area, the zoning, the daytime relation and the activities to be covered very carefully. The process of pre-processing the mobile phone data respectively the transition into OD-Matrices is quite complex and therefore the quality procedures in order to assess the usability of the usability are as well rather time-consuming. However, by applying automated processes and standardized parameters this process could be accelerated.

Pilot Innovations

1) Traffic micro-simulation models in city centres

CARTIF is the proprietary of these services that include as a technological novelty the definition of a common framework of how to generate traffic micro-simulation model based on data analytics, including loading/unloading activities. Freight delivery within the city has a great impact on traffic flow and is responsible for traffic congestion to a great extent. Central areas of the city face a certain number of situations every day:

- Misuse of parking areas by private vehicles, so that they are not free for freight delivery vehicles as they should.
- Not enough load and unload areas to cover current needs.
- Not enough control and surveillance to ensure adequate use of load and unload areas.

In some cases, new regulations that ensure efficient operation of last mile delivery, that does not hinder the city traffic flow, are required. In order to come up with adequate recommendations that can lead to such regulations, a thorough analysis of current traffic and urban freight delivery activities must be performed, so that all the knowledge extracted from data can be used by the traffic simulation models to be developed. Generating these traffic models for particular areas in the city where freight transport has more impact and analysing different freight delivery scenarios to support decision-making process are the main features on this service "Traffic micro-simulation model in city centres". As main results, different traffic micro-simulation models will be developed, as well as a **dashboard** to show valuable insights from Data Analytics tasks.

2.7 Shared Logistics for E-Commerce (Athens)

E-commerce consists one of the high priority sectors in European Economy that has grown to 530 billion euros in 2016 and was expected to grow at 603 billion euros in 2017. The sector presents a steadily growth of 14% in 2017 that differs strongly per country with around 57% of European Internet users shopping online (E-commerce Europe, 2017). On top of that, the omnichannel growth has created new challenges for retailers. According to (E-commerce Europe, 2017), the key challenges for e-retailers relate to delivery and payment options. Speed of delivery is important, but it is not everything. Shoppers want better control over how, when and where their goods are delivered. They want flexibility over delivery times and locations. They want to be kept in the loop, so tracked goods services will continue to grow. The main ambition of this pilot is to provide: a) a roadmap for applying big data analytics in order to tackle specific requirements in the e-commerce logistics, and b) empirical evidence about the impact that big data analytics could have on e-commerce logistics. To this direction, after extensive user requirements elicitation and analysis a set of pilot objectives have been identified and translated to five application scenarios in the city of **Athens** that have been presented in detail at *Deliverables 10.1* and *10.2.*

Profitable End-Users

The main pilot 3PL (**Third Party Logistics**) end-user states that through Big Data technologies and business analytics, the distribution processes can be deeper analysed and decisions can be adopted based on the customers (depositors) needs. The significant results derived from the project help them to reduce cost and better organize our procedures. Also, they will be able to organize the distribution planning better taking into consideration seasonal trends, potential patterns and specific problems that arise. Through forecasting tools, they will re-engineer their processes and create new business models.

Through Big Data technologies and the deep analysis of data they see in more depth the problematic points and we better forecast problematic situations. By identifying weaknesses and by improving processes and time, our customers will be satisfied for: a) Response time, b) Immediate information, c) Satisfied recipient and d) Improvement of customer's service.

LOGIKA, as the end user, is satisfied with the usefulness of the pilot solutions, as these solutions help them to improve our delivery process and reduce daily operating costs. The vehicle routing use case helps to organize better their fleet (capacity per vehicle, number of vehicles used, distance, time) resulting in a reduction of cost expenses. The number of vehicles used per day will be minimised. The ability to better control cancelled or returned orders will result in reduced costs and improved customer satisfaction.

Data Visualisation and User Interaction

1) **Real-time updates assist in better understanding of the results:** Data Visualization in some scenarios has been deployed using **Moriarty Dashboard** which is a tool developed by ITAINNOVA. Moriarty Dashboard provides a fast and easy way to develop customized user interfaces that connect with the Big Data infrastructure. However, the **static nature of the graphs** was identified as a bottleneck, when presenting the results to end users. The end users wanted a mo**re dynamic interface** in which they could adjust configurations and see the results, refreshing/updating them in real time, while allowing getting the desired visuals with the minimum interaction.

2) **Straightforward presentation and high level of interaction are appreciated by end users:** several different visualisations were developed, including *word clouds, interactive correlation cycles* and *descriptive analytics reports*. In more detail, each one of these visualisations was focusing on complementary goals, namely word clouds, interactive correlation cycles and descriptive analytics reports.

The main lessons learned regarding the deployed interface, based on users' feedback, is that it is critical to ensure: **Straightforward presentation of the key outcomes / results**. i.e. describe the main identified problems to the users in a clear and simple way. **High level of interaction** instead of having a static presentation of descriptive analytics outcomes.

3) **Achieving dynamically generated colour contrast with colour lightness:** one approach to remedy this issue was to compute the value (a.k.a. lightness or tone) of the category colour. This value denotes the colour brightness. When the colour is brighter than a specified threshold, black text is used for the label. When the colour value is less than the

specified threshold, white text is used for the label. This approach can serve as a lesson to anyone who wants to work with dynamically generated colours in their graphs.

4) **ML and OR optimization algorithms to serve hub selection and inventory routing:** the hub selection module emphasized the importance of providing a clear visualization of the demand. For this purpose, **OpenStreetMaps** were used to display the demand. The analytics and algorithms were implemented in **Python** programming language and were incorporated in the Moriarty Dashboard for ease of use. As a result, Pilot realized how important it is for decision makers to know exactly the demand pattern for online demand. According to user feedback the most convenient way for this is the map view and the ability to provide an interactive service where you can alter parameters and examine their effect.

5) **Experimental evaluation with user feedback from two grocery retailers:** for the last mile component, the end users and our data providers helped to realize the importance of the geography of the demand as well as the timing of them. For example, weekdays and time windows play an important role in performing last mile deliveries. Time windows that are frequently asked cause demand imbalance and as a result, more vehicles are needed during peak hours that remain idle on hours with no large volume of deliveries. The simulations showed that a collaborative distribution between retailers would become a win-win situation.

6) **Open Street Map for rapid development:** the visualization of the location optimization results in was an important part since a map view of the logistics network and locations was necessary for the end user. One lesson learned was that the free and open approach of OSM is a perfect baseline for designing smart and efficient map tools that do not require in-depth software development. Even the generation of heat maps was possible and can be used by any interested company or organization.

7) **Heatmaps for location optimization:** the concept of using heatmaps to show how stable certain candidate locations are, over a given number of optimizations or network sizes, was raised during the project. This was a new visualization approach that could easily be created and provide lots of valuable information to potential end-users.

Data Analytics

1) **Careful aggregation for better comprehension:** Moriarty Analytics (Kajal Forecasting and Kajal Routing) were used during the implementation. These tools provided a set of algorithms which allowed for fast information extraction in all use cases. The use of different algorithms led us choosing the best option in any situation depending on the input data, transparently to the user. Several criteria were defined to group deliveries and showed several graphs with aggregated data: **Range of dates**, to evaluate historical data, **Locations** grouped in five different levels, to search and identify areas of most

movements, **Depositor**, to study them one by one, **Product category**, **Unit type**, **Order type, Customer code**, to study groups separately.

2) **Location optimization with multiple logistics networks and open data:** one of the main challenges was to align the network and shipment data from multiple companies as well as add open data to it, in order to enrich the optimization results. It was shown that the optimization algorithm structure was suitable also for this newly created scenario and that no general adaptations needed to be made. This heavily refers to the data quality and, even more, the data scale of the geographically based data. However, the project showed that adding new and third-party data to existing optimization approaches heavily impacted the results and created increased benefit.

3) **Carefully combine multiple data sources to prove or disprove a given hypothesis:** a major challenge was the multiplicity of data sources. The pilot had to combine data sources of different types from different partners. The hub selection module was based on demand data of an online retailer and a brick-and-mortar- retailer. The pilot used ML clustering techniques to analyse the data and provide insights into the optimal location for installing hubs. Data cleansing and transforming procedures were preceded to create the structured input that ML algorithms need. A complex pipeline of data cleaning, data transforming, algorithm running, and output visualization had to be designed to serve this goal.

4) **More efficient classification using predefined clusters and 2-level classification:** some of the project's scenarios are comprised of text analysis, sentiment analysis, and text classification based on several different ML-based classification algorithms. The main lesson learned here includes the identified challenge to distinguish between problems reported by the end-users. It was proven that to facilitate better and more efficiently the classification of reviews to specific problems (and thus to identify the most important-and-often reported ones) it is suggested either to: **a)** use a pre-defined, small number of classes, i.e., a few high-level problems. **b)** use both low-level problems and high-level problem categories in a two-level classification approach. The first option leads to higher classification accuracy, but on the other hand some more low-level detailed problems could not always be included in the results.

Data Processing Architectures

1) **Early adaption of technologies pays off despite the problems:** Big data infrastructure was deployed by ITAINNOVA to be used in an elastic way. At the beginning of the project, a basic big data infrastructure was deployed. During the project's lifetime, more data were uploaded, and more demanding services were developed. As a result, the **infrastructure configuration had to be more precise**. Therefore, IT resources were increased and adapted to the developed services' needs. Mainly, memory resources and CPU were the

most common requirements while data size was growing, and more calculation power was needed. Some of the technologies evolved and matured during the project development. As a result, the scenarios had also been pilots in finding bug fixes for these kinds of solutions as various problems were emerged and addressed during that time. This was a big lesson learnt and should serve as a warning to early technology adopters.

2) **Geolocate using different services and levels of detail:** different ways and services for geolocating addresses and postcodes were used. Online services such as Gisgraphy, Nominatim and Google maps were employed to pinpoint addresses on the map. A great challenge was that the addresses were manually inputted by users, and as such, contained inconsistent or incorrect values. For example, different formats, different language, data from other fields, etc. Eventually, there was a compromise between lower location accuracy and higher data volume as postcodes were used to geolocate the problematic addresses. The trade-off was eventually justified as the analysis of the results proved to be concise. The main lesson learned is that if exact geolocation does not work, geolocate at postcode level instead of eliminating problematic entries.

Data Management

1) **Guidelines and data ID cards helped with data quality:** maintaining the data management of identification cards helped us ensure that the processed data was of high quality. The guidelines served as a good starting point for meeting and preserving the data quality KPIs. More specifically the completeness, uniqueness, timeliness, validity, accuracy, consistency was measured against LOGIKA's datasets and was used to produce more meaningful data. For example, duplicates were removed or merged depending on the level of duplicity, invalid or missing fields such as postcodes were filled in or their rows were ignored, inconsistencies in products were remedied by grouping them into categories. Overall, aiming for high-quality data translates to aiming for high-quality analytics results.

2) **Minimize system restore after a catastrophic failure:** one of the greatest challenges was handling catastrophic system failures. One lesson learned is also to request for backups to be available from the data providers. In such cases, the data providers will simply provide their data again. This will eliminate the costs associated with export time and computational power.

3) **Open data from national access points can help in the future:** while especially Greek open data sources where quite scarce during the project with the ongoing deployment and enrichment of the national access points for transportation data this can be a nice opportunity that should be deeply evaluated in future research projects.

Data Protection, Engineering, Standards

1) **Faster integration of connectors and data providers through standardization:** the technical approach of sharing data using an infrastructure of lightweight company side connectors and standardized central data processing has proven to be suitable for this kind of project where different data sources are used in combination to generate more valuable output. The importance of collaboration between all partners in early project stages to gain a mutual common understanding of the data pipeline has proved quite valuable during the integration phase of the project.

Pilot Innovations

1) **Descriptive Analytics for Distribution Patterns Identification:** descriptive analytics for distribution patterns identification offer to the potential users: a) interaction with data through the selection of various criteria/dimensions, and b) detailed analysis of data behaviour in 3PL partners through a set of appropriate visuals, and deduction of patterns and trends. The criteria/dimensions finally selected are an important outcome since they consider the specificities of the 3PL company and e-commerce sector and at the same time are generic enough.

2) **Forecasting Analysis:** it comprises the following innovations: (i) use of several forecasting algorithms and detection of the best one depending on the used input data, (ii) interaction with big data infrastructure almost in real time. For the 3PL end-user, the forecasting analysis consists of an important step in the **digitalization** process and coordination with its depositors/customers. It has the ability to track all types of orders but especially outbounds that consists the biggest part of distribution volume and recognize patterns that can lead them to derive new business opportunities related to the new vehicles’ dimensions, inverse logistics and new consortium agreements with depositors.

3) **Location Optimisation Demonstrator:** the web service is a demonstrator of the possibilities of big data-based location optimization. Its innovative combination of multiple data sources from different companies, combined with open data, allows for more matured insights and optimization results for shared logistics networks that would not have been possible otherwise. It can be easily accessed by an end-user and configured or updated to a new scenario background. In addition to the output of the essential optimization KPIs, the resulting networks can be visualized in a map and also heatmaps can be generated.

4) **Geographical Segmentation Leveraging Customer Orders Behaviours:** a data mining-based approach utilizing text mining and clustering techniques developed that introduces a new type of geographical segmentation based on the customer ordering behaviours and patterns. This approach can be used to design marketing campaigns and bundled

promotions that considers both customers buying preferences and delivery locations. Innovative dynamic pricing models for deliveries and products promotions can be designed for addressing the challenges of last-mile logistics and smoothing the delivery demand in various areas. It can be targeted at 3PL companies, courier companies and any company that deals with the transportation of goods and supply chain solutions in the last mile.

5) **Hub Selection:** defined as the supermarket that serves online orders (picking and delivering). Usually only some of the facilities of a retailer are selected to serve online orders. This is due to the fact that not only supermarkets can support the operations due to size or other restrictions and also the online channel in e-grocery is significantly smaller than the physical although the trend is increasing rapidly. The hub selection module is a decision support module integrated as a service and provided in the TT platform. It enables the end user and decision maker to visualize the demand on the map and provides helpful descriptive analytics.

6) **Route Optimisation:** after selecting which supermarkets become hubs to serve online demand a problem that arises is the replenishment of specific products only available online (not in store) from a central warehouse. The problem faced can be modelled as an Inventory Routing Problem characterized by great complexity. For the purpose of scalability, Pilot designed and implemented a new algorithm, able to identify which route, when and what levels of inventory quantities should be kept in the selected hubs in order to achieve viable inventory costs, with respect to the limited supermarkets capacities.

3 End-user Survey

3.1 Aim and Background

The aim of the TT end-user survey was to complement the lessons learned collected in the various TT pilots (and expressed in deliverables Dx.5 and the other sections of this deliverable) with a more structured and comparable approach across all pilots. To this end, we shared a structured survey questionnaire with all end-users involved in the TT pilots. End-user here mean actual transport system operators or transport managers.

Acknowledging that the TT pilot solutions (cockpits, dashboards, demonstrators etc.) are still prototypes, we asked the end-users to assume that when answering the questions that the TT pilot solutions would be fully developed, tested and deployed for real business operations at your workplace.

The questionnaire is based on the Technology Acceptance Model (TAM), a widely used model for information systems assessment[5]. The benefits of TAM include the model's simple use, the model's low complexity, empirically validated measurement scales as well as the robustness of

[5] F. Davis, „Perceived usefulness, perceived ease of use, and user acceptance of information technology“. MIS quarterly, MIS Quarterly 13(3): 319-340, 1989.

the model and its outcomes[6]. TAM was successfully used to evaluate software prototypes and applications in different application areas, including transport[7,8,9,10].

TAM aims to determine four main indicators:

- **Perceived usefulness (PU)**, which gives "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989).
- **Perceived ease-of-use (EOU)**, which gives "the degree to which a person believes that using a particular system would be free from effort" (Davis 1989).
- **Attitude toward Using (ATU)**, which gives the degree to which a person has a positive or negative attitude towards using the solution.
- **Behavioural Intention to Use (ITU)**, which gives the degree to which a person has formulated conscious plans to perform or not perform some specified future action.

3.2 Questionnaire

Responses to the questionnaire questions (listed below) below were to be provided along a 5-poinz Likert-scale, which also offered a “not applicable” option:

Strongly agree – agree –neutral – disagree – strongly disagree – not applicable.

[6] W. R. King, K. He: A meta-analysis of the technology acceptance model. Information & Management 43(6): 740-755, 2006.

[7] M. Morris, A. Dillon, „How User Perceptions Influence Software Use“. IEEE Software 14(4): 58-65, 1997.

[8] O. Laitenberger, H. Dreyer, „Evaluating the usefulness and the ease of use of a web-based inspection data collection tool”, Fifth International Software Metrics Symposium, 1998.

[9] A. Metzger, P. Schmidt, C. Reinartz, and K. Pohl, “Management operativer Logistikprozesse mit Future-Internet-Leitständen: Erfahrungen aus dem LoFIP-Projekt (Industrietransfer-Beitrag),” in Software Engineering 2014, Fachtagung des GI-Fachbereichs Softwaretechnik, February 25-28, 2014, Kiel, Germany, ser. LNI. GI, 2014.

[10] A. Braun, O. Bock, „Bewertung der Anwendbarkeit und Nützlichkeit der Modelle von Future-Internet-Logistiksystemen und der föderierten Future-Internet-Leitstände“, Projektdeliverable D-5.1/5.2, LoFIP-Projekt, 2014

Perceived Usefulness

1. I found the TT pilot solution useful
2. The TT pilot solution helped me gain new insights
3. Using the TT pilot solution would allow me manage operations better (for example, detect more problems)
4. The data visualizations of the TT pilot solution would help me manage operations better
5. The data analytics (descriptive or predictive) of the TT pilot solution would help me manage operations better
6. Using the TT pilot solution would allow me to manage operations faster (for example, detect problems in less time)
7. The data visualizations of the TT pilot solution would help me manage operations faster
8. The data analytics (descriptive or predictive) of the TT pilot solution would help me manage operations faster

Perceived Ease of Use

9. I found the TT pilot solution easy to use
10. It was clear to me how to interact with the TT pilot solution
11. It was easy for me to find the information needed in the TT pilot solution
12. The data visualizations in the TT pilot solution are understandable
13. The data analytics outcomes of the TT pilot solution are understandable
14. It would be easy for me to become skillful at using the TT pilot solution

Attitude toward Using

15. I like the idea of using the TT pilot solution
16. I believe it is a good idea to use the data visualizations of the TT pilot solution to manage operations
17. I believe it is a good idea to use the data analytics of the TT pilot solution to manage operations

Intention to Use

18. I would use the TT pilot solution frequently if available at my workplace
19. I would prefer using the TT pilot solution to other forms for data analysis

Comments

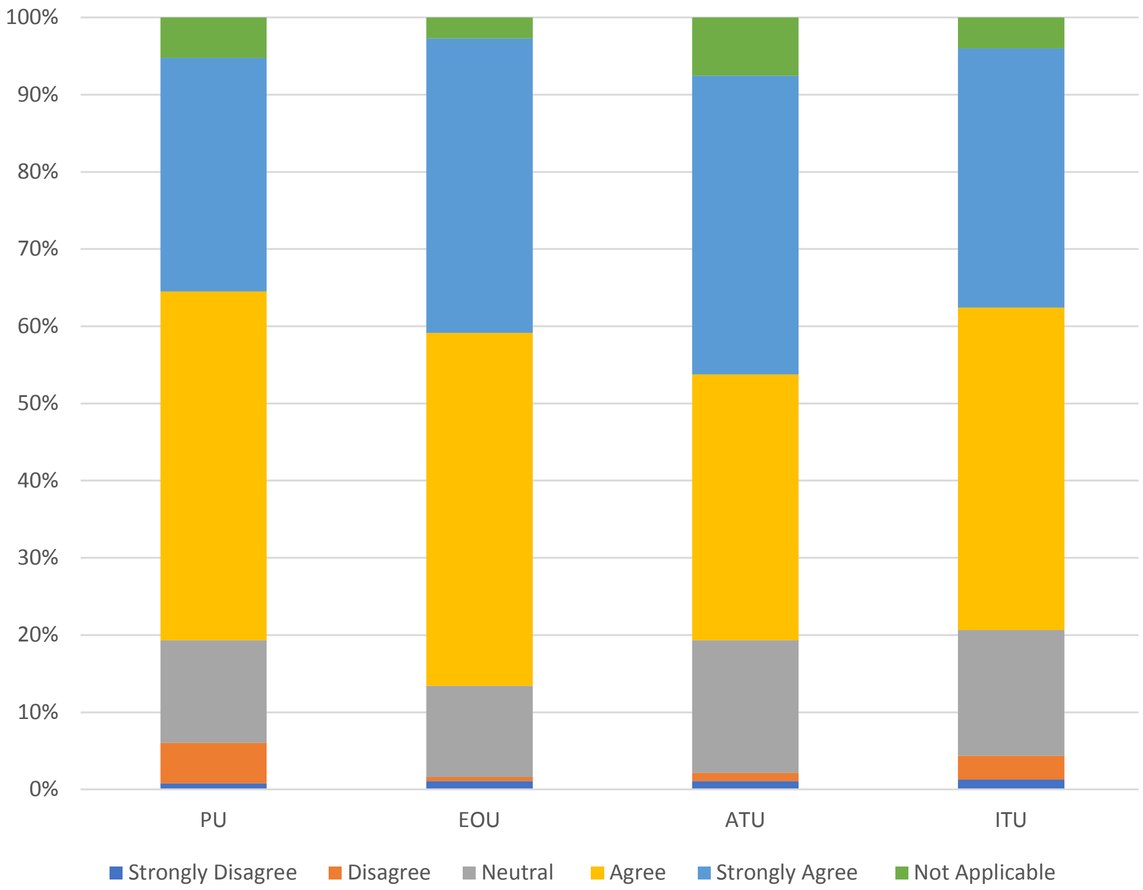
3.3 Results

Overall, we collected 31 responses from the TT end-users in all the 13 pilots, with the following breakdown in the number of responses:

• Roads (WP4):	6
• Vehicles (WP5):	4
• Rail (WP6):	3
• Ports (WP7):	7
• Airports (WP8):	2
• Smart Cities (WP9):	6
• Logistics (WP10):	3

The below diagrams visualize the responses on two levels. First, per each of the four main indicators, second for each of the questions individually.

Figure 2: Responses per TAM indicator



What is evident in Figure 2 is the very positive assessment of the TT pilots. A significant majority of respondents agreed or fully agreed that the pilot solutions are useful and easy to use.

In cases, where respondents disagreed, the main reason was that the end-users found it hard to fully assess the use of the TT pilot solutions given the fact that they were prototypes and not yet fully developed products.

Figure 3: Responses per survey question

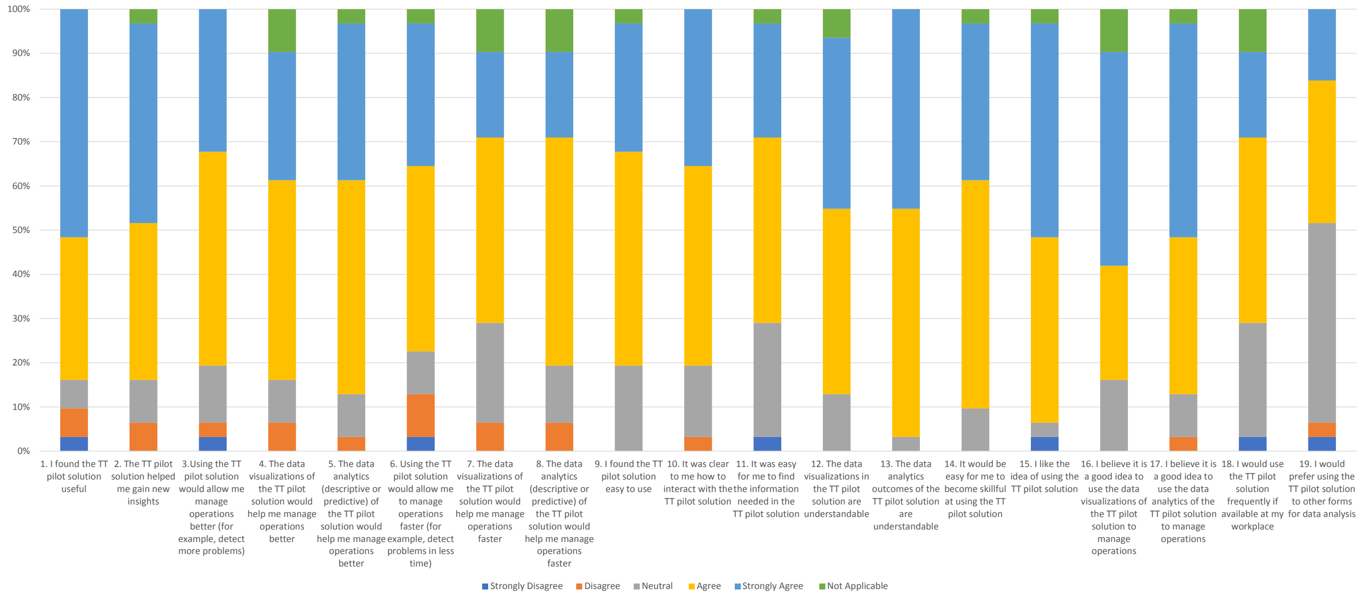


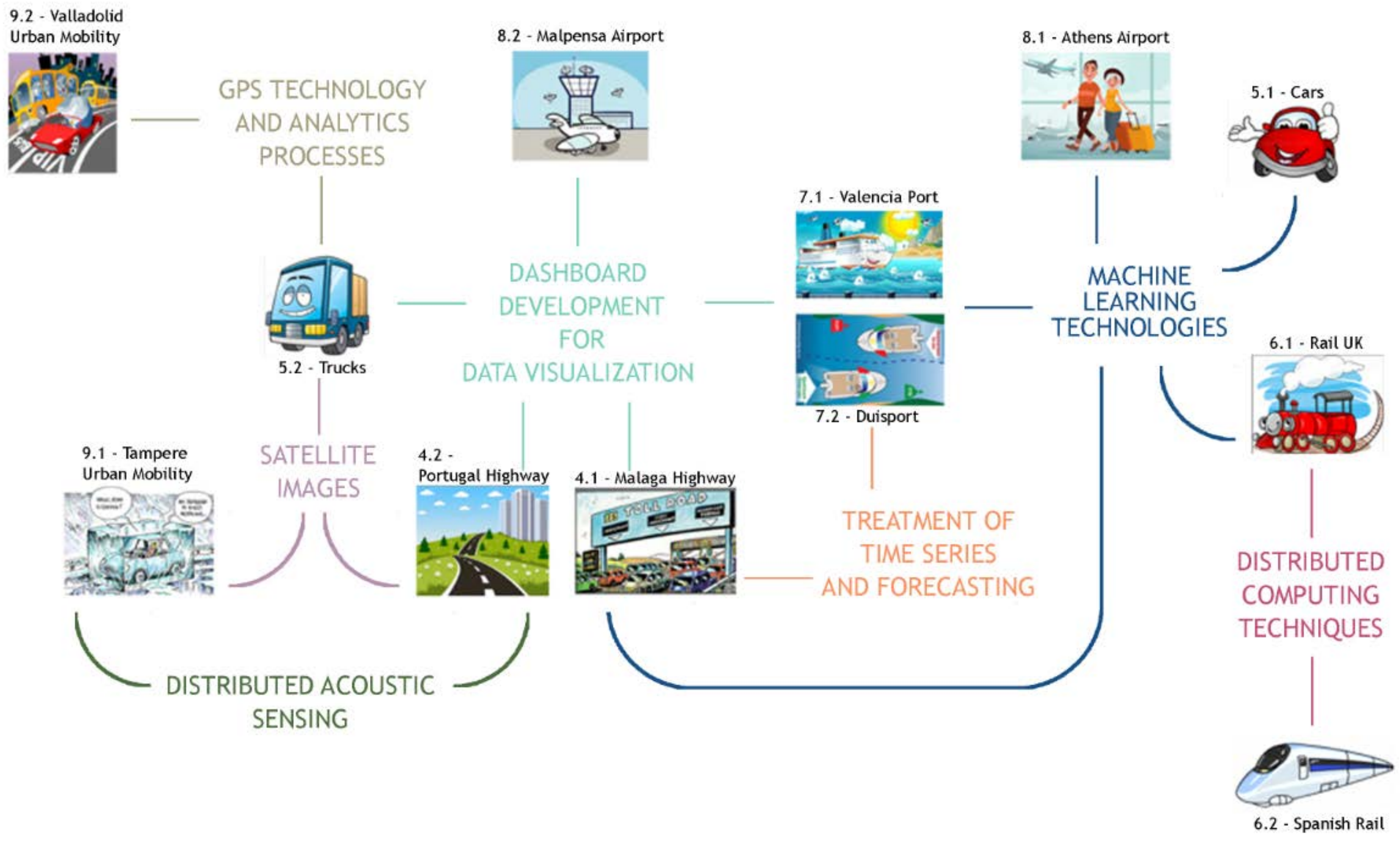
Figure 3 shows the results on the level of the individual questions. Especially for what concerns the analytics and visualization features of the pilot solutions, feedback was very positive. Again, the disagreement to some of the questions (e.g., Q6 on managing operations faster) was due to the prototype nature of the TT solutions.

4 Mutual Learning among Pilots

The following lines are intended to display all the learnings and the reuse possibilities that pilots have gained from each other across the project, not only from the perspective of the Domain partner but also from all the domains involved and from the project as a whole. A very rich learning among pilots of all sorts has been encountered.

Figure 4 displays the mutual learning and sharing among different transport domains in an illustrative way:

Figure 4. Reuse and Learning among Transport Domains


9.2 - Valladolid Urban Mobility
GPS TECHNOLOGY AND ANALYTICS PROCESSES
8.2 - Malpensa Airport
8.1 - Athens Airport
5.1 - Cars
7.1 - Valencia Port
DASHBOARD DEVELOPMENT FOR DATA VISUALIZATION
MACHINE LEARNING TECHNOLOGIES
5.2 - Trucks
7.2 - Duisport
6.1 - Rail UK
9.1 - Tampere Urban Mobility
SATELLITE IMAGES
4.2 - Portugal Highway
4.1 - Malaga Highway
TREATMENT OF TIME SERIES AND FORECASTING
DISTRIBUTED COMPUTING TECHNIQUES
DISTRIBUTED ACOUSTIC SENSING
6.2 - Spanish Rail

4.1 Smart Highways

The sharing of information from other Pilot Domains and the contacts made during the events has been productive for CINTRA. From conversations with Fraunhofer (partner involved in the Pilot Domain of Connected Trucks) and considering its works in the processing of **still satellite images** coming from different platforms for the monitoring and tracking of trucks on the road, CINTRA has known another technology not considered before in house but with great potential to the Business.

The technology has been evaluated internally by technical teams at CINTRA in collaboration with other players belonging to the sector. Fruit of this study, several cases of use with high interest for the Business were defined, becoming into real initiatives already ongoing.

4.2 Connected Vehicles

As for the Cars pilot, it has evaluated some different technologies that other pilots were using (e.g. related to **machine learning** techniques) to see if they could be applied to the pilot.

Referring to the Trucks pilot, it learned from other pilots and pilot domains how to better develop **dashboards**. One aspect was the visualization of data, e.g. geographically. Another aspect was the inclusion of **performance indicators**. So basically, the pilot learned where to set the focus and how to better convey the key message of the dashboard.

4.3 Proactive Rail Infrastructures

Some of the assets that were part of the pilot such as **OLE** (Overhead Line Equipment) could be explored and worked on further by the other piloting activities as a post-project replication. The work carried out so far could provide sufficient knowledge and baseline to understand and potentially build on the results.

The application of **distributed computing** systems was also greatly facilitated by the breadth of knowledge available across other domains. Many of the problems relating to the **size and complexity of the dataset** were solved by utilising solutions that have been publicised by other engineers in their work across different pilots.

4.4 Ports as Intelligent Logistic Hubs

The pilots reused and learned **analytic approaches** applied by other pilots. Specifically, a lot of pilots were dealing with **time series** and the **forecasting** of values from such a series. This a common type of data in the transport domain, so pilots found **alternative algorithms** to their initial proposal. For instance, several pilots presented interesting results using **LSTM networks** (Long Short-Term Memory) and for that reason, pilots applied this approach.

It was also found quite interesting the **data visualization** examples provided by other pilot demonstrations. It was close to a brainstorming session on how to improve the usability of the data in our specific pilot. Thanks to these demonstrations, pilots also discover interesting technologies, such as Bokeh or Plotly, which finally were not applied to the pilot, because of the technological stack was already defined. However, they will be considered in future Big Data projects.

4.5 Smart Airport Turnaround

The good news is, as the airport scheduled services are **replicating every day** again and again thousandfold worldwide since decades, there is a tremendous amount of data and experience that the industry can rely on to pro-actively overcome many of the disturbing events. The clue is to predict the **probability of occurrence** through analysis of current data with historical data and the impact it had on operations. Given these facts, the opportunity to apply today's technologies of big data analytics to these operational challenges across all transport domains is obvious.

The solutions proposed in the project are specific to the onsite scenarios in Athens and Malpensa airports. However, the **models developed are open for data of any other airport** and can be tweaked with the expert knowledge of the local operators to obtain an adapted solution. With that, the actors of the aviation domain pilots in Transforming Transport were able to develop a replicable solution which can be used beyond the scope of the research project.

4.6 Integrated Urban Mobility

During the pilot, the use of **sensors** which were deployed in other domains, especially the **Highway domain**, were assessed for use in Tampere. The Distributed Acoustic Sensing (**DAS**) method applied in "Norte Litoral" Pilot could provide more information on disturbances in traffic. However, since there is more non-traffic related to environmental noise, the signal to noise ratio is worse. The use of DAS for urban traffic requires more research for more specific use cases.

The use of **satellite imaging** as potential input for the situational awareness was also assessed. A potential use could be to get a view of the status of roadworks on the traffic. Currently, the accuracy of free satellite imaging is not accurate enough. Another issue that it is difficult to address is the view from the air to the traffic on the ground. An alternative solution, which is taken into use, is the use of **drones** to get an inventory of local roadworks.

Within Valladolid pilot main interactions with other pilots had been related to **GPS data**, and how to manage them in the correct way. Technically there had been no more interactions with other pilots, because most of them were focused on maintenance and real-time analysis, but not in simulations.

In the course of the project collaborations with the "Trucks" Pilot for sharing comparable data sources have been investigated. Also, from Trucks pilot fleet data of heavy trucks, different **data sources** have been analysed. In a first step data from toll providers and other telematics providers have been analysed with regard to the usability in urban areas – also for the purpose of **traffic modelling**. In "Trucks" routing and planning for the commercial truck fleet are core objectives, so the requirements for the data (coverage, content etc.) are different. In the end it turned out that these data coming from various areas in Europe are rather different in terms of data quality indicators like coverage, density, transport mode, vehicle type etc. For instance, some data from **telematics providers** are quite good in central Europe (NL, GE, FR), however the coverage and density in Spain especially in rural areas not related to the main urban areas remain rather low. However, these findings have been very important in the frame of the Big Data analytics processes and form the basis for further activities in this field.

4.7 Dynamic Supply Networks

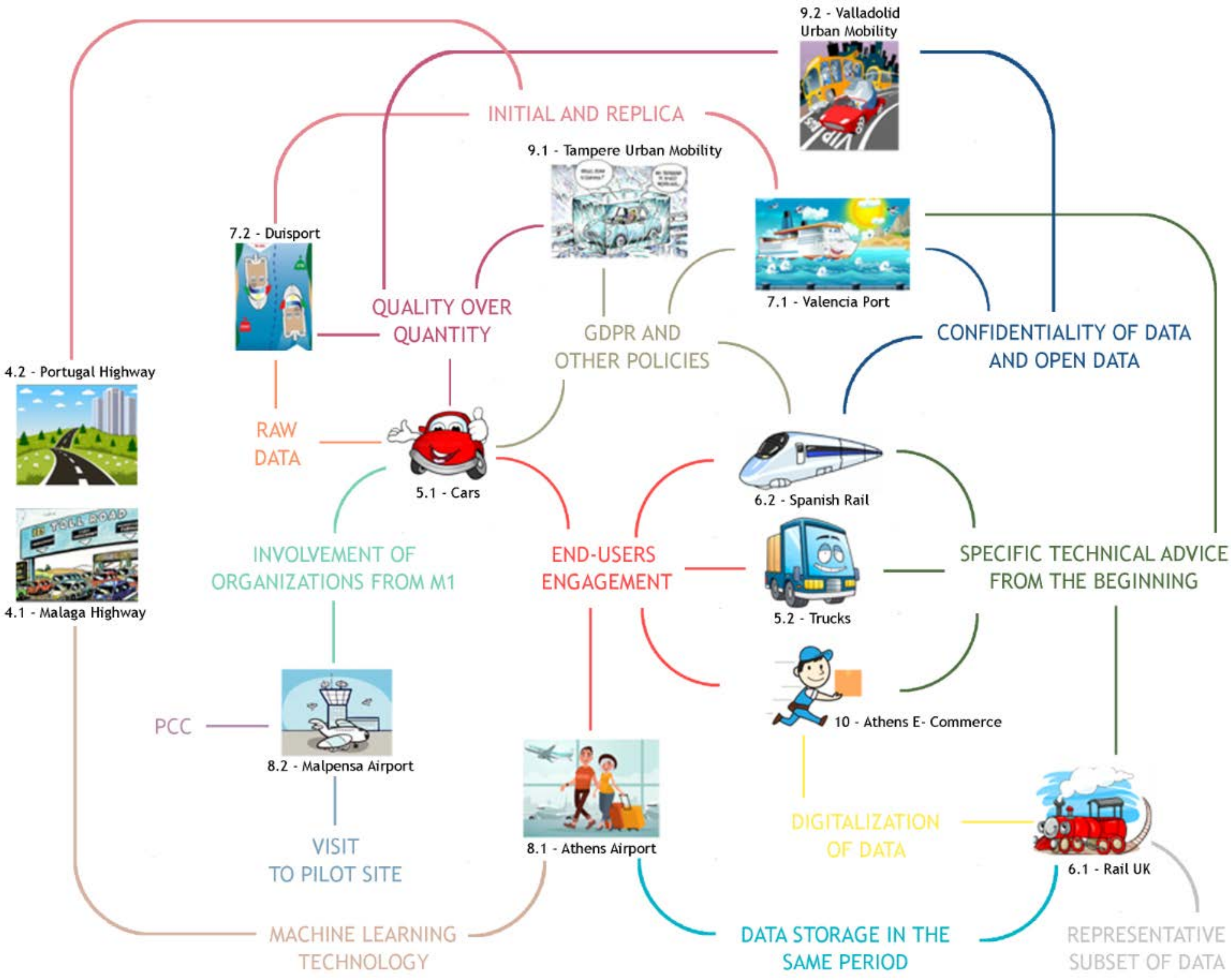
This pilot has no replication pilot. So instead of discussing cross-pilot findings like the other pilots, it has discussed some interesting similarities among the routing optimization and location optimization use cases addressed in different scenarios of the pilot and replicate the techniques in the future. For example, end user LOGIKA, wishes to refine its **delivery process** in Attica, the most populated area of Greece, where the majority of its customers are based, to reduce the daily operating costs raised by vehicles use. To investigate the delivery process, the pilot used of LOGIKA's routing dataset and associated them with the list of LOGIKA vehicles, taking into account the capacity and the routing zone that each vehicle is serving. As pilot noticed it is very important for **decreasing the total distance travelled** and **avoid "half-empty" vehicles**, to optimize the routing process and more specifically to control the number of vehicles used per day over all zones and the number of orders per route.

Dynamic Supply Networks have become a major theme for research, innovation and commercial development, given the **vast amount of e-commerce sales**. The focus here remains on the potential impact of the Pilot services on logistics processes, in terms of **operational efficiency** and **customer experience**.

5 Common Lessons Learned for Big Data implementation on further transport projects

In this section, the intention is to collect the most shared lessons learned derived from the pilots' experience along with the Transforming Transport project. Topics such as Methodology Implemented, Data Visualization, Data Analytics and Data Management are considered of main relevance when it comes to developing a Big Data project. In the following lines, a general overview of these findings is given to be considered in further actions. Figure 5 displays most common lessons learned from different transport domains in a schematized way:

Figure 5. Most Common Lessons Learned among all Transport Domains



5.1 Methodology

Here are presented most relevant lessons learned related to the procedures initially planned and which have turned out to be not as efficient or precise as expected. Some issues concerning initial terminology of the pilots, KPI Measurement process and end-user involvement are shown below.

1) The differentiation between **Initial and Replication** that was proposed at the very beginning of the project was not much effective, since both pilots within the same transport Domain have ended up working practically at the same time. The main reason for this is that deadlines for delivering documents or reporting results have been identical for all the pilots involved in the project. The idea was to take advantage of the previous work carried out by the Initial Pilot so that the Replication could start their tasks on a well-proven basis. As it has been pointed out in this document, the reliable replication of the tasks performed by the first Pilot has turned out kind of utopic, since different interests and objectives were addressed by the pilots.

As a matter of fact, whereas the terminology Initial and Replication could shed light on some sort of difference in the range or the category between pilots of the same domain, it has been proved that all of them have finally worked simultaneously according to their preferences. Thus, the delay planned for the start of the Replication works has been perceived unnecessary since it was eager to start work to carry on with the investigation and fulfil the deadlines. What it's worth mentioning here is that oftentimes even the Replication has offered interesting solutions to the Initial one, in an attempt of improving the methodology. This reveals that start piloting activities in parallel offers many ample opportunities for exchange and interaction among pilots.

2) **KPI measurement** has resulted in a very challenging task for all pilots. At the start of the project, the team responsible for suggesting the most suitable indicators for each one of the pilots namely "KPI Team", made a proposal of a set KPIs whose measurement was assumed by the pilots at the early stages of the project. Finally, it has been proved that addressing **too many KPIs may distract** from the main ambitions of each respective pilot. As well, establishing fair KPI baselines was tricky (historical data required to compute baselines; real-world moves on, so should baseline; time periods of baselines vs. actual affected by seasonality), so targets have been pulled out from the evaluation framework in a strict sense, although have been partially taken into account.

For this reason, several Reviewers suggested focussing only on the most relevant KPIs which have obviously led to the most interesting results from a business perspective. Also, targets have been fairly useful to follow up on the performance of the indicators even though the preliminary predictions were not much accurate. As a conclusion, KPI definition is a strenuous task but once the most important ambitions are addressed, these indicators are the true key to validate the usefulness of outcomes.

3) **End-user involvement** only in Stage 3 (final phase of the project) has been perceived too late by the different pilots. All of them agree on this issue, stating that at least relevant end-users should be involved from the very beginning of the project, in order to fix the most suitable methodology that addressed the main interest of all parts. Their participation at early stages is a key point to focus on the main needs from the user perspective, which when all is said and done, has the final word.

In order to calibrate models and technologies, end-users must be present from the start and take part in the definition of the processes and measurements to be addressed.

5.2 Visualization Techniques

As the project concludes, one of the most useful and profitable techniques being considered as a "*success key factor*" has been the **Dashboard** for data visualization and real-time control. The dashboard is a flexible Human-Machine Interface (HMI) designed to help operators on day-to-day monitoring, where pilots have shared their knowledge to gain the most valuable insights from these tools. According to Pilots comments, many conclusions with respect to the optimization of the Dashboard can be extracted:

1) Despite being an excellent tool to really see what is happening around the pilot, Dashboard should not be exhaustive in relation with the amount of information displayed, which can lead to cognitive overload due to information overflow. There are three main requisites: the **information** must be **shown hierarchically** from top to bottom interface, enabling making summaries with the most relevant details. Secondly, widgets must be **intuitive, simple and "clean"** for the user and allow for quick handling to easily grasp the information shown. Thirdly, Dashboard should **only display critical and enough well-validated events**, in order to avoid overloading the interface with superfluous warnings and focus the attention on the most important ones.

2) Static User Interfaces (UI) may be limiting. Providing **dynamic customization** of UI from simple multi-option dropdowns to more complex interchangeable requests could boost the efficiency of the analysis adapting itself to a specific user and operator wishes.
3) Visualisation helps to take decisions, with synthetic and clear results. Implications of the **human factors** team were found useful to understand these aspects. Moreover, getting them involved in the early stages of the project also helped to get a better perspective of the demonstrator.
4) It is relevant to address the **right customer or user** who is going to work with the visualized data. In the day-to-day business, there is often not enough time to only look at visualizations without an explicit added value. Yet, if the dashboard also serves as a decision-making tool, e.g. to plan routes or has another technical implementation, it

provides more added value. Another group to be approached could be **decision makers** who can use these dashboards for rather strategic planning purposes.

The goal of data visualisation is to make the data easily understandable and usable by the operators. To accomplish this, visualizations beyond just showing the quantitative data in big tables must be developed, thereby enabling the users to intuitively make a qualitative assessment of quantitative data. The terminal operators must be sure, that the **data is up-to-date** and current data. However, only knowing the current state is not sufficient for the operator. Also, the **date and time** of the last critical event were perceived important, in order to allow the operator to visualize/search for anomalies around the fault in historical data, and not only rely on the prediction algorithm. To enable the user to recognize critical trends more easily, spaces above and below certain thresholds are recommended to be **coloured**.

As it turns out, Dashboards are an excellent mean that allows gaining a real clear perception of the current status of the activities. Nevertheless, excessive overload in the presentation of the results can be risky for a good understanding of the actual and relevant situation.

5.3 Analytics

Another important topic in a Big Data project are analytics techniques. Many conclusions can be derived from pilots' experience:

1) **Data Quality**: Among the most universally accepted principles of analytics is "**Garbage in – Garbage out**", which refers to the quality of the data in the training models. It means that if poor-quality data enter the system, no matter how trendy the software for the analysis can be, the output value is expected to be of poor quality too. To overcome this, check and cope with missing data, data accuracy, data timelines, different time-zones (clocks), etc. is a must. Also is assigning "data owners" that understand data and its field (domain) being able to be in care of data quality.

2) Using **Deep Learning** and Neural Networks helps to make more efficient development and engineering. It has been proved to work well even without extensive hyper-parametrization, provided that enough good quality data is available. This means, the time- and resource-consuming step of extensive experimentation with hyper-parameters may be skipped, contributing to more efficient development and deployment process of big data applications.

3) **Data accuracy**: Operators benefit from information about data accuracy. It results in improved decision making and helps to determine when to trust a prediction, existing a trade-off between earliness and accuracy. Augmenting the quality of data (live or

predicted) with confidence intervals, error ranges or reliability estimates allows operators to acquaint themselves with the most actual situation.

4) **Time series models** can be successfully approached by traditional **machine learning techniques**. It has been verified that, machine learning techniques and Arima models are quite similar in short-term predictions, while the former tend to be more accurate as the time to be predicted gets longer. Not only predictive models are useful to improve a process, but it is also necessary to have **teams with enough experience** to select the most suitable alternative (descriptive or predictive). Additionally, machine learning techniques and Arima models are quite similar in short-term predictions, while the former tend to be more accurate as the time to be predicted gets longer. Another lesson learned is that external variables are easily included in the modelization.

5) **Historical data**: Regarding data analytics, Pilots found it useful to **keep historical non-reproducible data** and, when possible, in **raw format**. There are several reasons that support that way of proceeding, like possible errors or improvements in the code that not allow to rebuild processed data if the original data has been deleted. If substitutes raw data with processed data and there are no possible turning back mechanisms, important information can be missed in ulterior processing stages. The drawback for maintaining unprocessed raw data could be the increase of the storage capacity. Raw historic data can also be used for training in machine learning algorithms. The main idea is to keep the information (understanding here the concept of information as to the quantity of information according to the information theory -e.g. data can be lossless compressed-), since some bits of previous not treated information can be very important for future analyses.

5.4 Data Processing Architectures

The most important valuable lesson learned has been the use of **Big Data platforms**, that have shown their capacity to generate valuable knowledge and new insights for each one of the stakeholders involved in the pilots.

1) The use of **scalable data storage and data processing architecture** is needed as the data volumes are going to be high when the passenger trains start the regular data collection. Sometimes is necessary the processing of data into two categories; the first one was for values that required the **entire dataset** to be calculated, and the second was for values that could be calculated using a subset of the total data. Significant implementation issues were initially discovered when the entire dataset was needed, as the large quantity of data coupled with the chosen analysis technologies (**python**) meant that **calculation times could run into many days**. By way of example, the implementation of **Apache Spark**

processing allowed this calculation to be reduced by a factor of 40, which made analysis much more practical. This processing framework also allowed the data to be searched through to create subsets of records based on various filter criteria. Therefore, **processing of subsets of data** (e.g. using all of the records for a particular asset) was greatly simplified as the subset could be created using the Spark infrastructure, and then processing could be applied to only the relevant data. **Data lake and cloud environment** were identified to be a good solution for the management of big data to enable data sharing and communication. Due to the size of data being large, it is important to create partitions in the dataset. The data is split in a number of smaller packets with each packet assigned a specific key. The merge can occur for each packet so to ensure efficiency, the choice of a good key is crucial to create packets as small as possible while allowing the merge to occur.

2) **Changes in the code** are relatively frequent and should be implemented and put into service as soon as possible. Assuring a Continuous Integration and Continuous Deployment implies an organizational architecture to which the use of **microservices** can help. Some benefits of an architectural microservice approach include modularization, reusability and individual deployment, what could contribute to optimize the time to market constraints. This help alongside with the use of a non-relational database, in the maintenance of code yielded to processed data generation from raw data, as modifications and bug corrections were rapidly implemented and deployed. The use of a non-relational database in combination with the microservices allowed us to improve flexibility (e.g. adding new fields to the collections were quickly implemented) while retaining the performance of a shared database.

5.5 Data Management

The access to the data sources has turned out quite more complicated than expected due to the following reasons: Firstly because of the number of **different sources** and data production & storage systems. Secondly because of the **access characteristics of data sources**: from the technical point of view, some of these sources and systems didn't have the optimal flexibility. For future actions or projects involving BIG DATA technology application, it would be very useful, before starting, to ensure a comprehensive preliminary study of all the data sources and systems at disposal. Thirdly, the **data quality** has not been enough for many datasets. Here are some lessons learned:

1) **Management Plan**: A Data Quality Management Plan and a calendar of Quality Control tasks would be a must in further projects. Thus, **data cleansing** has taken a **high**

percentage of the effort in the project just for these activities, estimated in 80% of total work of Data Analytics Team.

2) Concerning **Real-Time analysis**, tools have been in many cases implemented **not as pure real-time but as near-real-time** system adapting the reaction time of the tools to the more lagged data producing process. This is an important lesson because expecting pure real-time systems is nowadays far from being easy due to ageing ITS in several cases. This technology should be updated for further replications mainly concerning Big Data projects, in order to take advantage of the new technologies to the extent.

3) **Storage structures**: In order to provide services at real-time, extra storage is required (what it should be considered in the dimensioning phase of the system). That means that special care must be taken in defining **optimized structures** derived from the raw data that allow **lower latency** to process data. Additionally, in the case of databases, it is important to define appropriate indexes, reaching a compromise between the speed of writing in the database and reading from it. It has also been found that non-relational databases are more appropriate than classic relational databases, for evolving systems. Relational databases are more restrictive in their structure and do not allow rapid changes, offering advantages such as flexible schemas and better scaling (e.g. when new dataset are added and more fields in a table -or collection- are necessary, the addition is much easier in a non-relational database).
4) **Profitable sources**: One of the research questions was to **identify valuable data sources** that support the understanding of the different transport domains. Therefore, many different data and data sources were part of the pilots. These data sources differed in terms of format, timely availability or geographical spreading (for Pilots with large areas of action). One of the first things that many pilots learned was to **abandon the idea of a holistic technical integration** of all data sources. Data can also provide valuable insights when considered separately to some extent. Concerning the visualization, it was important to develop good use cases and to define the right data for them. Therefore, only **useful data** were used and further processed and could finally reduce the complexity and increase the understandability.

5) The management of data required in many cases two approaches depending on whether processes required the use of the **raw datasets** or whether they required datasets produced from the results of processing. Raw data were stored in file structures which were accessible to all workers. The **parallelisation of computations** was then organised such that each process task would use a different file to other processes, resulting in a mitigation of file access conflicts. Results were then stored in a variety of data structures that are capable of both receiving data very quickly from multiple sources, as well as very

fast search and retrieval times for records. These two levels of information were very useful to understand the problems to study. But the most important was to have access to people who really know the data. **Data standardisation** has not always been completed.

6) Defining the **structure** and the data **format**, as well as the different **variables** and how they were going to be obtained has affected the whole project. For these reasons, among the conclusions and lessons learned related to data management highlights the necessity to understand the possibility of not disposing of every source of data when it is required, so lack of some data might be possible when the model is about to be displayed. This happens because of the **high amount of data sources from too many providers**, which means that not all needed data might be available on time for real-time predictions.

7) **Data availability** and fit for purpose: having data available at day 1 of the project does not mean it is fit for purpose (enough to answer the addressed business or operational needs) since technical access (interfaces) and organizational access (ownership) may require time to resolve. Because of this, firstly data analytics and visualization goals must be defined and then determine which data and how to access it or vice versa.

8) **Data quality and Integration**: in order to achieve enough data quality, time and effort required for data integration and refinement of data collection have been estimated at around 80% by some pilots. Oftentimes no control over the quality of data from third parties (incl. open data). If characteristics such as **Quality** and **Periodicity** are not good enough, the collaboration of the business teams in the imputation of data or the completion of it is needed to ensure a higher quality of the model.

Because of this, there must be planned sufficient time at project start for **data refinement** and fine-tuning of data collection before blindly starting any task.

5.6 Data Protection, Engineering or Standards

1) **Regulations**: From the regulatory perspective of data confidentiality and personal data protection, the work under this project has been impacted by the new European regulation in the matter of data protection, EU **General Data Protection Regulation** (GDPR), with an effective date on May 25th 2018. In order to achieve compliance with new legislation, prior preparatory actions were taken. To this aim, different mechanisms have been adopted such as the Bluetooth MAC delivering trimmed addresses.

2) **Recovery systems**: Every organization must plan and develop well-proved disaster recovery systems in order to be safe for unexpected fails and lose of information.

3) **Data reliability**: usually, one has to bear in mind that raw data is not always correct or error-free. Ways for outliers' detection, even the simplest ones like establishing a deviation threshold from an average, can be very useful to discard erroneous data.

4) **Latency in data**: it is convenient to register not only the timestamp when the data was produced, but also when data arrived at the system. That way, it can be checked if there is a delay problem with the data provision and to an asset if the succeeding process can be considered real-time or not.

5) **Data volume**: Counters in the reception of data and free disk space are needed to indicate the occupancy degree of the system and to foresee when space scalability is necessary.

6) **Health status**: It is convenient, apart from the registration of the physical parameters of the cluster, such as memory, disk space or CPU speed, to have a dashboard with counters for the previously mentioned topics (percentage of data discard, delay in data and data throughput). Through these indicators, an operator could detect some problems straightforward by simple inspection.

7) **Cloud infrastructure**: it can be more convenient and time-saving to use a Cloud Big Data Platform. Its benefits are avoiding administrative tasks and being provided with more flexible/scalable (computation/budget by demand). Also, high computational capacity is required in punctual situations.

8) A general recommendation is using as much as possible **open source technologies** to process and analyse data, as well as in dashboard developments. Since there are thousands of volunteers generating algorithms that everyone can use, contribute using and generating open sources solutions will help to improve this community.

Next, Table 2 reunites the level of commitment of the pilots along with the project with regard to the topics described above, where **1** = main focus, **2** = addressed but not mainly, **3** = addressed marginally and **4** = not addressed.[11]

1.4. [11] S. Zillner, E. Curry, A. Metzger, R. Seidl (Eds.), "European big data value strategic research and innovation agenda (SRIA)," Version 4.0, October, 2017

Table 3. Map of the scope of the pilots to the BDVA reference model. SRIA (2017)

	WP 4		WP 5		WP 6		WP 7		WP 8		WP 9		WP 10
Technical priorities and challenges	**Ausol**	**Norte Litoral**	**Cars**	**Trucks**	**UK Rail**	**Spanish Rail**	**Valencia**	**Duisport**	**Athens**	**Malpensa**	**Tampere**	**Valladolid**	**Supply Networks**
Data Management													
Semantic Annotation of unstructured and semi- structured data	2	2	3	3	3	3	3	3	3	3	4	4	3
Semantic interoperability	3	3	3	4	3	3	3	3	3	3	4	4	3
Data quality	3	3	4	4	2	2	4	4	4	4	4	4	4
Data lifecycle management and data governance	4	4	4	4	4	4	4	4	3	3	3	3	3
Integration of data and business processes	3	2	3	4	4	4	4	4	4	4	4	4	4
Data-as-a service	4	4	4	4	4	4	4	4	4	4	4	4	3
Distributed trust infrastructures for data management	4	4	4	4	4	4	4	4	4	4	4	4	4
Data Processing Architectures													
Heterogeneity	4	4	4	4	4	4	4	4	4	4	3	3	4
Scalability	3	3	3	3	3	3	3	3	3	3	3	3	3
Processing of data-in-motion and data-at-rest	4	4	4	4	4	4	4	4	4	4	4	4	4
Decentralizatrion	4	4	4	4	4	4	4	4	4	4	4	4	4
Performance	4	4	4	4	4	4	4	4	4	4	4	4	4
Novel architectures for enabling new types of big data workloads	3	3	4	4	4	4	4	4	4	4	4	4	4
Introduction of new hardware capabilities	4	4	4	3	4	4	4	4	4	4	4	4	3

	WP 4		WP 5		WP 6		WP 7		WP 8		WP 9		WP 10
Technical priorities and challenges	**Ausol**	**Norte Litoral**	**Cars**	**Trucks**	**UK Rail**	**Spanish Rail**	**Valencia**	**Duisport**	**Athens**	**Malpensa**	**Tampere**	**Valladolid**	**Supply Networks**
Data Analytics													
Semantic and knowledge-based analysis	3	2	3	3	2	2	3	3	2	2	2	2	2
Content validation	4	4	4	4	3	3	4	4	3	3	4	4	4
Analytics frameworks & processing	2	3	3	3	3	3	3	3	3	3	3	3	3
Advanced business analytics and intelligence	3	2	2	1	1	1	2	2	3	3	2	2	2
Predictive and prescriptive analytics	1	1	1	2	1	1	1	1	2	2	1	1	1
High Performance Data Analytics (HPDA)	2	2	2	2	1	1	2	2	2	2	3	3	2
Data analytics and Artificial Intelligence	4	4	4	3	4	4	4	4	4	4	4	4	3
Data Protection													
Generic and easy to use data protection approaches	4	4	4	4	4	4	4	4	4	4	4	4	4
Robust Data privacy (incl. multi-party computation)	4	4	4	4	4	4	4	4	4	4	4	4	4
Risk based approaches	4	4	4	4	4	4	4	4	4	4	4	4	4
Data Visualisation and User Interaction													
Visual data discovery	3	3	3	2	2	2	3	3	3	3	3	3	3
Interactive visual analytics of multiple scale data	2	2	3	2	2	2	2	2	3	3	2	2	2
Collaborative, intuitive and interactive visual interfaces	2	2	2	2	2	2	2	2	3	3	2	2	2
Interactive visual data exploration and querying in a multi-device context	2	2	2	2	2	2	2	2	3	3	2	2	2

5.7 Other non-technical Lessons Learned

1) One of the key lessons was the **involvement of working group partners in the document submissions**. The individual use case owners should be involved in contributing to the respective documents throughout the length of the project in each deliverable, instead of the pilot leader putting the document together. This would help build more meaningful and accurate content for the documents. The pilot leader should agree to an action plan for the duration of the project with the working group partners at the start by setting some key milestones. The group can then measure their performance against the milestones. This would also help the pilot leader in putting the contributions together and generally running the project smoothly.

2) **Partners contribution**: Regarding the data management and the different stakeholders, the **involvement of the partners** in the development of the project must be constant. The involvement of the data providers is crucial to be ensured since the beginning of the project, not only for the duration of the project, but especially to guarantee the continuity of the project afterwards. The end-user might comprise many different departments of a company, where each of them has different needs and objectives. It is essential to carefully identify, assess and prioritize the real end-user needs so that the limited effort can be almost entirely applied to what they expect to be improved. Considering the nature of an innovation project, where the result of the development is highly unknown, it is quite easy to fall into a contradictory and challenging situation where end-users request certainty on what they will receive in exchange of the data or effort provided.

3) **Tangible and realistic outcomes** have to be set as a target in order to encourage data providers' participation. Big data algorithms obtained during the project have to be trained and adjusted when the solution is completely installed and in use. That's why strong collaboration basis has to be previously established, keeping such interest of the end-user on the solution afterwards. Also, the contribution of business experts is highly recommended so that specific knowledge is applied during the development of business cases.

4) Full involvement and **full cooperation** proved to be the key to the project's success. Each partner, in fact, has been involved in a continuous dialogue that guarantees, in parallel, continuous surveillance of the various steps. Another essential key for the Project's success is the **visit of the Pilot site**, enriched with the presence of the operators engaged in their everyday activities: this permits a realistic dialogue (questions and related answers) and an understanding of the existing systems. Another winning tip is the **weekly**

calls. Live conversations prove to be particularly effective and guarantee continuous, effective and efficient full coordination and total control of the situation.

5) As main lessons learned it could be said that a focused **and effective solution is preferable to an ambitious** but wide scope one due to the higher probability of loss of performance.

5.8 Support from Dedicated KPI Team

As responsible for the Task 2.2: “Pilot Requirements Analysis and Lessons Learned”, the KPI Team has carried out the development of the evaluation framework of the Transforming Transport project including the definition of the Key Performance Indicators (KPI) that finally were selected by the Pilots. This section is to present the usefulness perceived by the pilots around the KPI Team collaboration on the definition of the KPI and the corresponding coordination of the periodic reports on their values and solving the specific questions from the pilots that could arise in the measurement process affecting the figures directly.

The support from a dedicated KPI team and a project-wide homogenous KPI measurement process have been perceived very positive. The participation of an external team in the definition and continuous **follow-up of the KPIs** is an important point that it is worth to be commented on. The KPI team not only proposed KPIs for the pilots, but also helped in the establishment of most suitable **formulas** and the way they could be measured. This proposal has also been a good **baseline** to understand the aims and objective of the use case and how they could be linked to the measure of success in applying big data technologies. A constant channel of communication with such a team is basic in order to keep up to date during the works, for achieving good performance of the **coordination tasks**.

As it is expected and due to the complexity and different “subprojects” carried out, pilots have made some recommendations regarding future similar collaborations:

1) **More collaborative work in the definition of the KPIs** in a first phase (all the pilot representatives, and the KPI team, should talk to each other to find out what KPIs can be representative and valid for most of the pilots. A more global communication channel is better perceived than a more oriented dialogue between KPI team and pilot leaders.
2) Put the effort on a **deeper understanding of the content of the pilots’ activities** to provide better content-related feedback to specific KPIs. A next step could even include to come up with suggestions for further pilot KPIs based on the feedback of all pilots to create a more common frame of KPI measurement.
3) The **number of meetings** to discuss the KPI’s could be increased to have a clearer understanding of the updates to the KPI during the length of the project.
4) **Support** and assistance regarding **calculating KPI’s** with a higher number of inputs.

5) **Selecting the categories** for the KPI's could be a **joint activity** between the KPI team and Pilot leaders to further develop these categories by considering the concrete ambitions of each Pilot.

Due to the wide variety in the focus of the pilots, the different environments and the **different transport domains**, developing a common approach for the **definition**, the **measurement** and the **analysis** of the **KPIs is very challenging**. Challenges are directly related to the success of the pilots and not by changes in the environment (e.g. changes in policies related to transport vehicles, the availability of new roads or large infrastructure works). Baselines may be hard to define, as they may require setting up additional measurement infrastructure, which is not yet available.

However, the pilots agree that having a dedicated KPI team helped to clearly define how success and value added by the work could be measured since the project start. Particularly interesting in such a big consortium, is to scale these measurements to all domains and to aggregate all results coherently. Without a dedicated team, this task would have been otherwise very difficult to achieve given the different team's background and topics we have worked on.

6 Big Data findings and impacts by Assessment Category for improving the performance of transport activities

Within the following lines, a conclusion about what is the impact of the Big Data technologies on each main Assessment Category is addressed. According to this review, further projects using Big Data in the field of Transport will be able to start from a well-founded baseline to develop their works, when it's time to design the most suitable performance indicators and methodologies.

To this aim and throughout the performance of the **Transversal KPIs**, the benefits of Big data on the Transportation Sector for five Assessment Categories (Operational Efficiency, Asset Management, Environmental Quality, Energy Consumption and Safety) are depicted here. The conclusions will start from the less relevant (Safety) to the most impacted (Operational Efficiency) according to TT project results.

6.1 Operational Efficiency Category (OE)

Operational Efficiency is defined as the ability to deliver products and services cost-effectively without sacrificing quality. Enhancing efficiency in the transportation system is a priority topic for the EC appearing reflected as so in multiple transport related EC publications. "The future prosperity of our continent will depend on the ability of all of its regions to remain fully and competitively integrated in the world economy. Efficient transport is vital in making this happen" (White paper on transportation, EC 2011). In TT project the introduction of Big Data technologies is expected to contribute to the achievement of a more efficient provision of transport services in all the tested domains as reflected in one of the goals of the aforementioned White paper: using transport and infrastructure more efficiently through the use of improved traffic management and information systems.

Among the 136 KPI that were finally selected for evaluating each Pilots' performance, up to **74** belonged to the Operational Efficiency Category, evidencing its huge relevance for all the partners, since its importance it is largely proved from a business perspective.

In this sense, it can be extracted several conclusions and lessons learned from the results of all these KPIs for each one of the pilots. Obviously, Big Data Implementation is directly related to the increment of the number of sources displayed along the Pilot's area of influence and the high efficiency of the processing systems for analysing the retrieved data. Bering in mind this overall updating, all pilots have taken advantage of these technologies to enhance more Pilot-specific problems, not only gathering data itself but also to extract added value according to their interests and preferences.

- For **Highways**, Big Data technologies have demonstrated its utility in improving the time savings to users, the average commercial speed, the number of informative messages to users or the adaptation of the lane to the actual circumstances.
- For **Connected Vehicles**, it has been possible to improve the operating cost, the travel time savings, the average commercial speed in peak and off-peak hour. As for the pilot Cars, the number of harsh accelerations and decelerations of the vehicles has also been improved. In the case of the Pilot Trucks, the accuracy of the estimated time of arrival has been clearly improved.
- As for **Rail**, reduction in operating cost related with three different use cases (Overhead Electric Line, Points and Tamping) as well as the increased availability of assets have been achieved in the case UK Rail. The Spanish Rail has also got good results in reducing the train lateral accelerations of the car body.
- In the case of **Ports**, it has improved the average truck turnaround time in the terminal area, the average equipment idle time and in general the average time per container transaction. As for Duisport, the number of trains (instead of trucks) leaving the terminal on time has been enhanced.
- For **Airports**, Athens (passenger-oriented) has improved the number of security lanes capacity and availability, the security staff hours and the accuracy of their predictions through the machine learning technologies. As for Malpensa (aircraft-oriented), has enhanced the predictions of taxi-in and taxi-out time, the estimated time of arrivals and the boarding times or boarding gate allocation predictability.
- As for **Urban Mobility**, Tampere has improved customer satisfaction with the service, reflected in the increasing amount of subscribers to the city centre region automated tweets. On the other hand, Valladolid has reduced the time used by freight vehicles in the city centre for driving and parking and the daily number of stops per vehicles and day.
- Finally, **E-Commerce** Pilot has incremented the time savings per delivery, decreased the walking distance from customer to the collection points and the distance from the hub to the customer, the average of daily deliveries and the number of vehicles used per day. As well, it has enhanced the forecast accuracy of problematic deliveries such as returns.

6.2 Asset Management Category (AM)

Asset management is a systematic process of deploying, operating, maintaining, upgrading, and disposing of assets cost-effectively. Infrastructure asset management is the combination of management, financial, economic, engineering, and other practices applied to physical assets with the objective of providing the required level of service in the most cost-effective manner. It includes the management of the entire lifecycle of physical and infrastructure assets. Operation and maintenance of assets in a constrained budget environment require a prioritization scheme.

Big Data technologies offer the potential to optimise the management of the usage and the maintenance of transport assets.

Asset Management is the second most applied category by pilots with up to **31** KPIs. But in this case, not all of them have included this category to evaluate their project performance:

- **Highways** Pilots have reduced the number of maintenance interventions, incremented the traffic volume and diminished the minutes of the queue in the most important toll station for the case of the Spanish Pilot in Malaga.
- As for the **Connected Vehicles**, only for the case of Cars Pilot, the number of messages related to breakdowns have considerably decreased, which reflects better maintenance of the vehicles.
- For **Rail**, there has been a reduction in the number of interventions and in the number of alerts per circulation. As well the percentage of preventative maintenance activities compared with corrective ones has increased. The accuracy of failure prediction has also improved. Concretely for Spanish Rail, the distance covered by the machinery to the working area has been optimized too.
- For **Ports**, Valencia has reduced the average cost, breakdowns and downtime per monitored equipment and augmented the mean container handlings between failure like Duisport. This one has also incremented the proportion of productive time of the operation time of a crane.
- As for **Urban Mobility**, Pilots have augmented the number of freight delivery places covered by the urban logistics applications and the traffic camera installations for traffic management in the case of Tampere.

6.3 Environmental Quality Category (EQ)

Europe’s goals toward sustainability are widely acknowledged. The EU transport system is currently not sustainable, and in many respects moving away from sustainability rather than towards it. The European Environment Agency highlights particularly the sector's growing CO2 emissions that threaten the EU meeting its target under the Kyoto protocol. It also points to additional efforts that are needed to reach existing air quality targets, points to a large proportion of the population that is exposed to annoying or harmful noise levels and highlights many more environmental impacts from transport in the EU. The Commission has been working in cooperation with the Council to prepare the ground for making European transport policy sustainable.

Air pollution has been one of Europe's main political concerns since the late 1970s. European Union policy on air quality aims to develop and implement appropriate instruments to improve air quality. The control of emissions from mobile sources, improving fuel quality and promoting and integrating environmental protection requirements into the transport and energy sector are part of these aims. In this regard the Commission establishes environmental quality standards so as to limit the concentrations of certain chemical substances that pose a significant risk to the environment or to human health. The implementation of the piloted solutions should prove useful in the achievement of a more environmentally respectful transportation sector.

As for this Category, only **9** KPIs have been measured to prove the efficiency of Big Data technologies in relation to the reduction of the pollution derived from the Transportation Sector. Although results have been quite hopeful, more efforts should be made in future projects to address this kind of problems that affect all of us the same.

- Only Pilots such as **Highways** (Malaga), **Connected Vehicles** (Cars), **Airports** (Athens and Malpensa) and **E-commerce** have addressed this kind of indicators. As a result, big reductions in NOx, CO_2 and PM emissions have been achieved.

6.4 Energy Consumption Category (EC)

Linked with the previous item, but with a special focus on resource efficiency for the transportation sector, the European Commission's 2012 Energy Efficiency Directive establishes a set of binding measures to help the EU reach its 20% energy efficiency target by 2020. Under the Directive, all EU countries are required to use energy more efficiently at all stages of the energy chain, from production to final consumption. This objective will also be monitored in the different pilots allowing to assess the improvements in efficiency regarding energy consumption achieved by better system management derived from the use of Big Data technologies.

Very related to the former one, this category has been evaluated throughout only **7** KPIs, also showing good results.

- Pilots such as **Highways** (Malaga) has considerably reduced the GHG emissions at toll stations waiting times, **Connected Vehicles** (Cars) has decreased the vehicle energy consumption and the GHG emissions throughout real-time messages to users. **Airports** (Malpensa) has improved fuel consumption due to better taxi fuel loaded. **Urban Mobility** (Valladolid) and **E-Commerce** Pilots diminished their GHG emissions throughout fuel consumption savings per delivery.

6.5 Safety Category (SF)

Safety in transportation is another of EU's priorities for the sector. More specifically road safety, which was addressed in the last Valletta declaration (2017) as a commitment across the EU. Last year 25,500 people lost their lives on EU roads in 2016, 600 fewer than in 2015 and 6,000 fewer than in 2010. A further 135,000 people were seriously injured on the road according to

Commission's estimates. (EC, 2017). The likely improvements in this regard through the processing of large amounts of data (mostly on road traffic but not exclusively) will also be a category assessed by the framework.

Albeit this is also a very important category to be analysed, it has turned out to be one of the least interesting from the point of view of the pilots, in order to enhance their production systems. Only **8 KPIs** have been designed to demonstrate the Big Data impact on Transport safety:

- **Highways** Pilots have evaluated this category, obtaining improvements in the drivers' perception of safer driving conditions. As well, UK **Rail** Pilot has reduced the number of track-side activities for each one of their three use cases displayed above (OLE, Points and Tamping).

6.6 Economy Category (SF)

An economic analysis of the impact of the implementation of Big Data solutions in diverse transport sector domains is at the core of the TT project. The economic impact of the piloted technologies will be addressed, inferring their outcomes in terms of competitiveness, market impact and business improvement for operators.

It is worth mentioning that the Economic Category above is limited to the internal economic assessment of pilots. Bearing this in mind no cross-pilot analysis is feasible, since interests vary among all the pilots. Nevertheless, for Pilots addressing this KPI Category, outcomes show positive improvements in the reduction of operation, maintenance and environmental costs.

7 Conclusions

This section summarizes the main findings and some recommendations for future activities. Best practices should pave the way for further actions once usefulness of Big Data in the transport sector have been demonstrated along TT developments. The conclusions are structured in three sections.

1) Best practices on Big Data processes

Data Visualization: One of the most useful and profitable techniques has been the **Dashboard** for data visualization and real-time control. According to Pilots comments, many outputs related to the optimization of the Dashboard can be extracted: Dashboard should not be exhaustive in relation with the amount of information displayed, which can lead to cognitive overload due to information overflow. The **information** must be **shown hierarchically** from top to bottom interface, enabling making summaries with the most relevant details. Secondly, widgets must be **intuitive, simple and "clean"** for the user and it should allow quick handling to easily grasp the information shown and **only display critical and enough well-validated events**.

The information should also provide **dynamic customization** of the user interface, which would boost the efficiency of the analysis adapting itself to specific users and operators. The validation team of the dashboard should be formed by dedicated personnel. This group of users should include **operators** and also **decision makers,** who can use these dashboards for their current work and for strategic planning purposes, respectively.

Analytics: Among the most universally accepted principles of analytics is "**Garbage in – Garbage out**", which refers to the quality of the data in the training models. To overcome this risk, check and cope with missing data, data accuracy, data timelines, different time-zones (clocks), etc. is a must. To assign "data owners" that understand data and its field (domain) is necessary to care of data quality. Operators benefit from information about **data accuracy**, which results in improved decision making and helps to determine when to trust a prediction, existing a trade-off between earliness and accuracy. Augmenting the quality of data (live or predicted) with confidence intervals, error ranges or reliability estimates allows operators to acquaint themselves with the most actual situation.

On the other hand, using **Deep Learning** and Neural Networks helps to make more efficient development and engineering and **Time series models** can be successfully approached by traditional **machine learning techniques**. Predictive models are useful to improve a process, but it is also necessary to have **teams with enough experience** to select the most suitable alternative (descriptive or predictive). Regarding data analytics, pilots found it useful to **keep historical non-reproducible data** and, when possible, in **raw format**, in order to mend possible errors or make improvements in the code that not allow to rebuild processed data if the original data has been deleted.

Data Processing Architectures: The most important valuable lesson learned has been the use of **Big Data platforms**. The use of **scalable data storage and data processing architecture** is needed as the data volumes are growing when data collection starts. Sometimes it is necessary the processing of data into two categories; the first one refers to values that required the **entire dataset** to be calculated, and the second is for values that could be calculated using a subset of the total data. **Processing of subsets of data** (e.g. using all of the records for a particular asset) was greatly simplified as the subset could be created, and then processing could be applied to only the relevant data. **Data lake and cloud environment** were identified to be a good solution for the management of big data to enable data sharing and communication. When the size of data became large, it is important to create partitions in the dataset.

Data Management: The access to the data sources has turned out quite more complicated than expected due to the following reasons. Firstly, because of the number of **different sources** and data production & storage systems. Secondly, because of the **access characteristics of data sources**: From the technical point of view, some of these sources and systems didn't have the optimal flexibility. For future actions or projects involving BIG DATA technology application, it would be very useful, before starting, to ensure a comprehensive preliminary study of all the data sources and systems at disposal. Thirdly, the **data quality** has not been enough for many datasets.

Some lessons learned are: A Data Quality **Management Plan** and a calendar of **Quality Control** tasks would be a must in further projects, since **data cleansing** has taken a **high percentage of the effort in the project just for these activities, estimated in 80% of total work of Data Analytics Team.** Concerning **Real-Time analysis**, tools have been in many cases implemented **not as pure real-time but as near-real-time** system, adapting the reaction time of the tools to the more lagged data producing process due to ageing ITS in several cases. In order to provide services in real-time, extra storage is required (what should be considered in the dimensioning phase of the system). That means that special care must be taken in defining **optimized structures** derived from the raw data that allow **lower latency** to process data.

One of the research questions was to **identify valuable data sources** that support the understanding of the different transport domains. These data sources differed in terms of format, timely availability or geographical spreading (for Pilots with large areas of action). One of the first things that many pilots learned was to **abandon the idea of a holistic technical integration** of all data sources. Data can also provide valuable insights when considered separately. The management of data required in many cases two approaches depending on whether processes required the use of the **raw datasets** or whether they required datasets produced from the **results of processing**. The **parallelisation of computations** was organised such that each process task would use a different file to other processes, resulting in a mitigation of file access conflicts. Defining the **structure** and the data **format**, as well as the different **variables** and how they were going to be obtained has affected the whole project.

For these reasons, among the conclusions and lessons learned related to data management, we can highlight the necessity of understanding that it is not possible the availability of every source of data when it is required, so lack of some data might be possible. This happens because of the **high amount of data sources from too many providers**, which means that not all needed data might be available on time for real-time predictions. If characteristics such as **Quality** and **Periodicity** are not good enough, it is needed the collaboration of the application teams in the imputation of data, or their completion, to ensure a higher quality of the model. Because of this, there must be planned sufficient time at project start for **data refinement** and fine-tuning of data collection before blindly starting any task.

Data Protection, Engineering or Standards: The work under this project has been impacted by the new European regulation in the matter of data protection, EU **General Data Protection Regulation** (GDPR). In order to achieve compliance with new legislation prior preparatory actions were taken. Every organization must plan and develop well-proved **disaster recovery systems** in order to be safe for unexpected fails and lose of information. Ways for outliers' detection can be very useful to discard erroneous data. It is convenient to register not only the timestamp when the data was produced, but also when data arrived at the system. That way, it can be checked if there is a delay problem with the data provision and to an asset if the succeeding process can be considered real-time or not.

Finally, a general recommendation is using as much as possible **open source technologies** to process and analyse data, as well as in dashboard developments. Since there are thousands of volunteers generating algorithms that everyone can use, contribute using and generating open sources solutions will help to improve this community.

2) Methods for measuring Big Data impacts in transportation

End-users & Partners Contribution: End-user involvement only in the final phase of the project has been perceived too late by the different pilots. All of them agree on this issue, stating that at least relevant end-users should be involved from the very beginning of the project, in order to fix the most suitable methodology that addressed the main interest of all parts. Their participation at early stages is a key point to focus on the main needs from the user perspective, which when all is said and done, has the final word.

One of the key lessons was the **involvement of working group partners in the document submissions**. The individual use case owners should be involved in contributing to the respective documents throughout the length of the project in each deliverable, instead of the pilot leader putting the document together. This would help build more meaningful and accurate content for the documents.

Regarding the data management and the different stakeholders, the **involvement of the partners** in the development of the project must be constant. It is essential to carefully identify, assess and prioritize the real end-user needs so that the limited effort can be almost entirely applied to what they expect to be improved. Full involvement and **full cooperation** proved to be the key to the project's success. Another essential key for the Project's success is the **visit of the Pilot site**, enriched with the presence of the operators engaged in their everyday activities. Another winning tip is the **weekly calls** among partners, providing effective and efficient full coordination and total control of the situation.

Profitable end-users: According to pilot comments, the most important indicator to assess pilot efficiency and management are **users or customers'** feedback, since they are who really test and experience personally the quality of the infrastructure and services. **Drivers** feedback, in the case of Highways, Connected Vehicles or Urban Mobility Domains, is the most useful information in order to acquaint with the good performance of roads and the vehicles. **Passengers** in the case of Rail or Airport Domains, **consumers** in the case of E-commerce Domain or the **Terminal Managers and Maintenance Managers** in the case of Port Domain play the most important role when it comes to evaluate the overall efficiency of the pilot. In general, a very useful end-user is the **Personal Staff**, as they are the ones who benefit directly from the pilot results. Moreover, a well-designed structure of this group with sufficient knowledge on the field is essential for a good performance.

Pilot Innovations: Along with the document several innovations have been highlighted. As for Highways most outstanding innovation is the **DAS** (distributed acoustic sensing). Cars Pilot has developed an **emission reduction system** and Trucks Pilot outlines the usefulness of their **satellite image system**. In general, main innovations are related to the different **Dashboards** designed for real-time conditions monitoring (e.g. Terminal Cockpit in the case of Ports Domain) and the development of accurate **algorithms** to predict future events (e.g. Rail, Ports, Airports…). As well, Urban Mobility and E-Commerce Domains highlight the development of **analytics techniques** to extract information from the **social media**, really useful to gain insights of the current tendencies.

KPI Issues: **KPI measurement** has resulted in a **very challenging task** for all pilots. At the start of the project, the team responsible for suggesting the most suitable indicators for each one of the pilots namely "KPI Team", made a proposal of a set KPIs whose measurement was assumed by the pilots at the early stages of the project. Finally, it has been proved that addressing **too many KPIs may distract** from the main ambitions of each respective pilot. As well, establishing fair KPI baselines was tricky (historical data required to compute baselines; real-world moves on, so should baseline; time periods of baselines vs. actual affected by seasonality), so targets have been pulled out from the evaluation framework in a strict sense, although have been partially taken into account. For this reason, several Reviewers suggested focusing only on the **most relevant KPIs** which have obviously led to the most interesting results from a business perspective. As a conclusion, KPI definition is a strenuous task but once the most important ambitions are addressed, these indicators are the true key to validate the usefulness of outcomes.

The support from a dedicated KPI team and a project-wide homogenous KPI measurement process have been perceived very positive. The participation of an external team in the definition and continuous **follow-up of the KPIs** has been an important point that it is worth to be highlighted. A constant channel of communication with such a team has been basic in order to keep up to date during the works, for achieving good performance of the **coordination tasks**.

Nevertheless, pilots have made some recommendations regarding future similar collaborations: **More collaborative work in the definition of the KPIs** in a first phase (all the pilot representatives, and the KPI team, should talk to each other to find out what KPIs can be representative and valid for most of the pilots. Put the effort on a **deeper understanding of the content of the pilots' activities** to provide better content-related feedback to specific KPIs. The **number of meetings** to discuss the KPI's could be increased to have a clearer understanding of the updates to the KPI during the length of the project. **Selecting the categories** for the KPI's could be a **joint activity** between the KPI team and Pilot leaders to further develop these categories by considering the concrete ambitions of each Pilot.

However, the pilots agree that having a dedicated KPI team helped to clearly define how success and value added by the work could be measured since the project start. Particularly interesting in such a big consortium, is to scale these measurements to all domains and to aggregate all results coherently. Without a dedicated team, this task would have been otherwise very difficult to achieve given the different team's background and topics we have worked on.

3) Improving performance of different transport modes

With respect to the Operational Efficiency Category (the most representative in TT) Big Data technologies have demonstrated its utility for all the pilots. In the case of the **highways** pilots TT has produced time savings to users, increased average commercial speed, increased the number of informative messages to users and enabled new extra lanes when demand exceeds capacity in toll stations by real-time monitoring.

Vehicles pilots have improved the operating cost, the travel time savings, the average commercial speed in peak and off-peak hour. As for the pilot Cars, the number of harsh accelerations and decelerations of the vehicles has also been improved. In the case of the Pilot Trucks, the accuracy of the estimated time of arrival has been enhanced.

In the **rail** domain, reduction in operating cost related with three different use cases (Overhead Electric Line, Points and Tamping) as well as the increased availability of assets have been achieved for the UK Rail pilot. The Spanish Rail has also revealed good results in reducing the train lateral accelerations of the car body.

In the case of **ports**, it has improved the average truck turnaround time in the terminal area, the average equipment idle time and in general the average time per container transaction. As for Duisport, the number of trains (instead of trucks) leaving the terminal on time has been augmented.

Athens **airport** (passenger-oriented) has improved the number of security lanes capacity and availability, the security staff hours and the accuracy of their predictions through the machine learning technologies. Malpensa airport (aircraft-oriented) has enhanced the predictions of taxi-in and taxi-out time, the estimated time of arrivals and the boarding times or boarding gate allocation predictability.

For pilots related to the **urban mobility**, Tampere has improved customer satisfaction with the service, reflected in the increasing amount of subscribers to the city centre region automated tweets. On the other hand, Valladolid has reduced the time used by freight vehicles in the city centre for driving and parking and the daily number of stops per vehicles and day.

Finally, **e-commerce** pilot has incremented the time savings per delivery, decreased the walking distance from customer to the collection points and the distance from the hub to the customer, the average of daily deliveries and the number of vehicles used per day. As well, it has enhanced the forecast accuracy of problematic deliveries such as returns.

The second most successful results correspond to the Asset Management Category. **Highways** have reduced the number of maintenance interventions, incremented the traffic volume and diminished the minutes of the queue in the most important toll station for the case of the Spanish Pilot in Malaga.

In the case of the **connected vehicles** pilots, only for the case of Cars, the number of messages related to breakdowns have considerably decreased, which reflects better maintenance of the vehicles.

Rail domain has reduced the number of interventions and in the number of alerts per circulation. As well, the percentage of preventative maintenance activities compared with corrective ones has increased. The accuracy of failure prediction has also improved. Concretely for Spanish Rail, the distance covered by the machinery to the working area has been optimized too.

Valencia **port** has reduced the average cost, breakdowns and downtime per monitored equipment and augmented the mean container handlings between failure like Duisport. This one has also incremented the proportion of productive time of the operation time of a crane.

Pilots concerning **urban mobility** have augmented the number of freight delivery places covered by the urban logistics applications and the traffic camera installations for traffic management in the case of Tampere.

Finally, Categories such as **Environmental Quality**, **Energy Consumption** or **Safety** have been much less analysed and more efforts should be put on these concepts in the future, since hopeful results have also been reported. Big reductions in NOx, CO_2 and PM emissions have been achieved in some pilots. **Highways** (Malaga) has considerably reduced the GHG emissions at toll stations waiting times, **Connected Vehicles** (Cars) has decreased the vehicle energy consumption and the GHG emissions throughout real-time messages to users. **Airports** (Malpensa) has improved fuel consumption due to better taxi fuel loaded. **Urban Mobility** (Valladolid) and **E-Commerce** pilots diminished their GHG emissions throughout fuel consumption savings per delivery.

Related to Safety, **Highways** Pilots have improved drivers' perception of safer driving conditions. In its turn, UK **Rail** Pilot has reduced the number of track-side activities for each one of their three use cases (OLE, Points and Tamping), reducing the risk of accidents.

As it has been widely demonstrated, Big Data technologies are the key for future development of the Transportation sector, where outstanding results have been achieved all along the different modes and fields. Future actions should be based on many lessons learned in TT. However, more attention on environmental aspects should be drawn in order to reduce negative externalities of transport activities. Nevertheless, results displayed here are quite encouraging to further research.