IN2DREAMS

Contract No. 777596

INtelligent solutions 2ward the Development of Railway Energy and Asset Management Systems in Europe

D5.1: Data Analytics Scenarios

DUE DATE OF DELIVERABLE: 31/05/2018

ACTUAL SUBMISSION DATE: 14/05/2018

Leader/Responsible of this Deliverable: Renzo Canepa - RFI

Reviewed: Y

Document status		
Revision	Date	Description
0.1	16/02/2018	First Issue of the TOC
0.2	20/02/2018	Final TOC and Reference Pictures
0.3	10/03/2018	Conclusion of UNIGE an RFI parts in Sections 1, 2, 3, and 5
0.4	10/03/2018	Conclusion of UNIGE Cross-Scenario 2 description
0.5	16/03/2018	Conclusion of UNIGE Specific-Scenario 3 description
0.6	18/03/2018	Conclusion of UNIGE Specific-Scenario 1 and 4 description
0.7	19/03/2018	Conclusion of RFI Specific-Scenario 2
0.8	10/04/2018	Conclusion of EVOLUTION ENERGIE Specific-Scenario 5
0.9	10/04/2018	Conclusion of KUL Section 2.3
0.10	13/04/2018	Conclusion of UKON Sections 2.2, 3.4, and Cross-Scenario 1 description
0.11	13/04/2018	First Draft for Review
0.12	21/04/2018	Internal Review Completed (KUL+CEFRIEL+RFI)
1	03/05/2018	IN2SMART Partners Review Completed (AST+STRUKTON)
2	14/05/2018	Final Version after TMT approval and Quality check

Project funded from the European Union's Horizon 2020 research and innovation programme		
Dissemination Level		
PU	Public	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Start date of project: 01/09/2017 - Duration: 24 Months

Executive Summary

The general objective of WP5 is to study, design and develop data analytics solutions for knowledge extraction from railway asset data. This objective will be achieved through the following tasks:

- Definition of data analytics scenarios
- Development and demonstration of tools and methodologies aiming at extracting knowledge from data analytics algorithms, and contemporarily making them interpretable in an easier way
- Study and proof-of-concept of metrics and methods/tools to measure the accuracy of analytics algorithms

Task 5.1 is responsible to prepare the work of the other tasks included in WP5 and to define a set of analytics scenarios that will be used as case studies for the development of data analytics algorithms and metrics for their assessment. The scenarios will focus on relevant railway assets whose malfunction and maintenance policies have an impact on the KPIs targeted by the SHIFT2RAIL program. The scenarios will be clearly defined from both the data availability and the analytic perspectives, and the different goals for data analytics tasks and related metrics will be identified. In particular, seven scenarios will be identified:

- Cross-Scenario 1: Visualizations in Control Center
- Cross-Scenario 2: Marketplace of Data and Data Monetization
- Specific-Scenario 1: Track Circuits
- Specific-Scenario 2: Train Delays and Penalties
- Specific-Scenario 3: Restoration Time
- Specific-Scenario 4: Switches
- Specific-Scenario 5: Train Energy Consumption

Two of them are cross-scenario in the sense that they cover, in some way, many aspects of the railway ecosystem while five of them are specific-scenario in the sense that they focus on a single particular aspect.
In the next WP5 deliverable some or all of them will be further develop based on data availability, quality, and quantity.

Abbreviations and Acronyms

Abbreviation	Description
A	Ampere
AMS	Asset Management System
ASTS	Ansaldo STS (IN2SMART Partner)
CBI	Computer Based Interlocking
CEFRIEL	Cefriel (IN2DREAMS WS2 Partner)
COLA	Collaboration Agreement
CSV	Comma Separated Value (file format)
DLR	DLR - Institute of Transportation Systems (IN2SMART Partner)
EU	European Union
EVOLUTION ENERGIE	Evolution Energie (IN2DREAMS Partner)
H2020	Horizon 2020 framework programme
HVAC	Heating, Ventilation and Air-Conditioning
ID	Identifier
IM	Infrastructure Manager
IN2DREAMS	INtelligent solutions 2ward the Development of Railway Energy and Asset Management Systems in Europe (SHIFT2RAIL Recipient)
IN2RAIL	Innovative Intelligent Rail (SHIFT2RAIL Recipient)
IN2SMART	Intelligent Innovative Smart Maintenance of Assets by integRated Technologies (SHIFT2RAIL Recipient)
IP	Innovation Programme
KPI	Key Performance Indicator
KUL	Katholieke Universiteit Leuven (IN2DREAMS WS2 Partner)
ODM	Open Data Management
POC	Proof of Concept
RFI	Rete Ferroviaria Italiana (IN2DREAMS WS2 Partner)
SHIFT2RAIL JU	SHIFT2RAIL Joint Undertaking
STRUKTON	Strukton Rail (IN2SMART Partner)
TCS	Track Circuit System
TD	Technical Demonstrators
TMS	Traffic Management System
TO	Train Operator
TRL	Technology readiness level
UKON	University of Konstanz (IN2DREAMS WS2 Partner)
UNIGE	University of Genoa (IN2DREAMS WS2 Partner)
V	Volt
VA	Visual Analytics
VAC	Voltage in Alternating Current
VDC	Volts of Direct Current
WP	Work Package
WS	Work Stream

Table of Contents

List of Figures

1 Introduction

The general objective of WP5 is to study, design and develop data and visual analytics solutions for knowledge extraction from railway asset data. For this reason the cornerstone of WP5, which is the focus of this deliverable, is Task 5.1 which is in charge of developing scenarios and use-cases in order to accomplish the main WP5 objectives. This deliverable will focus on relevant railway assets whose malfunction and maintenance policies have an impact on the KPIs targeted by the SHIFT2RAIL program. In particular Task 5.1 is responsible for preparing the work of the other tasks included in WP5 by defining a set of analytics scenarios that will be used as case studies for the development of data analytics algorithms and metrics for their assessment. Consequently, T5.1 is crucial to the Work Package and to the whole WS2. The scenarios will be clearly defined from both the data availability and the analytic perspectives, and the different goals for data analytics tasks and their metrics will be identified. One or more scenarios will be used as a basis for the development of the WP5 proof-of-concepts and demonstrators. This task is led by RFI, the Italian Railway Infrastructure Manager, in order to share its view of direct railway stakeholder with the other partners of this WP. Cooperation with WP6 is assured by partner EVOLUTION ENERGIE and coordination with other SHIFT2RAIL recipients is assured by the collaboration with IN2SMART and IN2RAIL.
Seven scenarios will be presented. Two of them are cross-scenario in the sense that they cover, in some way, many aspects of the railway ecosystem while five of them are specific-scenario in the sense that they focus on a single particular aspect. The two cross scenarios are:

- Cross-Scenario 1: Visualizations in Control Center
  This scenario will deal with the problem of improving the information visualization systems of the railway operators with visual analytics. This problem is crucial because of the increasing complexity of the information systems of the railway infrastructure and the advent of new predictive models (including the one developed by the SHIFT2RAIL recipients); the operators need to be able to understand the state of the entire railway infrastructure by just observing a condensed, intuitive, and informative visualization system.
- Cross-Scenario 2: Marketplace of Data and Data Monetization
  This scenario will deal with the problem of helping the actors of the railway ecosystem in having an interpretable and reliable support decision system which can help them in automatically exchanging, evaluating, and monetizing the data collected and stored by their information system. Data monetization is the act of generating measurable economic benefits from available data sources. Typically these benefits accrue as revenue or expense savings. Data monetization leverages data generated through business operations, available exogenous data or content, as well as data associated with individual asset such as that collected with sensors.

The five specific scenarios are:

- Specific-Scenario 1: Track Circuits
  This scenario deals with the problem of building an interpretable and reliable data-driven condition based maintenance system for the track circuit, a simple electrical device used to detect the absence of a train on rail tracks, in order to inform signallers and control relevant signals.
- Specific-Scenario 2: Train Delays and Penalties
  This scenario deals with the problem of building an interpretable and reliable data-driven model for train delays and train delay penalties prediction. This is a critical task for the train management system since it is important to both improve the train circulation and to reduce the delays-related penalties. In this way the train dispatcher can exploit this information and choose a new dispatching solution which minimizes both delay and penalty costs.
- Specific-Scenario 3: Restoration Time

This scenario deals with the problem of building an interpretable and reliable data-driven incident restoration time forecast system. The information outputted by the models can be very useful because it could be used by the traffic management system to reroute trains through safer paths, minimizing the risks of any problem.

- Specific-Scenario 4: Switches
  This scenario deals with the problem of building an interpretable and reliable data-driven condition based maintenance system for the train switches, a mechanical installation enabling railway trains to be guided from one track to another, such as at a railway junction or where a spur or siding branches off.
- Specific-Scenario 5: Train Energy Consumption
  This scenario deals with the problem of building an interpretable and reliable data-driven train power consumption forecast model and use this prediction to compare the results to the real data. This will allow better understanding of the system and better management of energy. The development of the forecasting models are included in the objective of the WP6 of the IN2DREAMS project, which is to develop solutions to assist infrastructure managers and railway operators to select optimal strategies and resources in order to support, in a cost-effective and energy-efficient manner, a variety of railway applications addressing operational requirements and passenger's requests. For example: solutions for operations optimization in line with intelligent train enabling the operation of the rolling stock in an autonomous mode.

The remaining part of the deliverable is organized as follows. Section 2 is dedicated to a brief introduction to Data and Visual Analytics, to the description of the metrics to evaluate the quality of the analytics itself and finally to the general legal requirements and constraints in using these technologies. Section 3 considers and describes the railway ecosystem, its connections with other IN2DREAMS WSs and WPs and SHIFT2RAIL recipients, and the potential impact of Data and Visual Analytics. Section 4 describes in details the IN2DREAMS WP5 scenarios. Section 5 concludes the deliverable.

2 Data and Visual Analytics & Metrics

This section is dedicated to a brief introduction to Data and Visual Analytics, to the description of the metrics to evaluate the quality of the analytics themselves and the general legal requirements and constraints in using these technologies.

2.1 Data Analytics & Metrics

Vast amounts of data are being generated in many fields, especially in the railway one, and data analytics is a tool able to collect, clean, process, analyze, and gaining useful insights from them [4, 25, 26, 62, 85, 88, 97, 230]. The deluge of data is a direct result of advances in technology and the computerization of every aspect of modern life. It is therefore natural to examine whether one can extract concise and possibly actionable insights from the available data for application-specific goals. This is where the task of data analytics comes in. The raw data may be arbitrary, unstructured, or even in a format that is not immediately suitable to be processed by an automated computer program to gain insights. To address this issue, data analysts use a pipeline of processing, where the raw data are collected, cleaned, and transformed into a standardized format. Data analytics applications are often closely connected to one of the two most common types of problems: supervised and unsupervised problems. These problems are important because they are often used as building blocks in a vast number of applications. In a supervised scenario an outcome measurement, usually quantitative or categorical, is provided and we want to predict it based on a set of known features.

Moreover a set of data, in which we observe the outcome and feature measurements for a set of objects, is available. In this context the term supervised indicates the presence of the outcome variable to guide the learning process. In the unsupervised framework, instead, with the same goal we are able to observe only the features and we possess no measurements of the outcome, making the task more complicated. Using this data, we build a prediction model which will enable us to predict the outcome for new unseen objects. A good learner is defined as the one that can be easily interpreted and can accurately predict such an outcome based on some problem specific metrics.

2.1.1 From Data to Data Analytics

Data, in general, can be defined as a piece of digitally information which describes or is produced by a system. Data can be structured or unstructured. Structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and it can be readily searchable with simple queries. On the contrary, data is considered unstructured when it does not fit neatly in a row-column database.
The first step of data analysis is to have a good set of data; otherwise, a data preprocessing step should be applied in order to provide it. Real world data are generally

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- Noisy: containing errors or outliers
- Inconsistent: containing discrepancies in codes or names

The tasks in data preprocessing [55, 72, 78, 83, 101, 196] are then the following ones

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.
- Feature extraction: from the raw data meaningful features should be extracted.

Once data are available it is important to understand what is the task to solve and the best algorithm able to solve it.
There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data. It is popular in data analytics to first consider the learning styles that an algorithm can adopt. There are only a few main learning styles or learning models that an algorithm can have and we will go through them here with few examples of algorithms and problem types that they suit.
This taxonomy or way of organizing data analytics algorithms is useful because it forces to think about the roles of the input data and the model preparation process and select the one that is most appropriate for the problem in order to get the best result.
There are three main different learning styles in machine learning algorithms [4, 183, 185]

- Supervised Learning: input data is called training data and has a known label to describe it (such as faulty/non-faulty asset). A model is prepared through a training process where it is required to make predictions, correcting them when they are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Example problems are classification and regression. Example algorithms include Logistic Regression and Back Propagation Neural Networks.

- Unsupervised Learning: input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may signify to extract general rules. It may happen through a mathematical process to systematically reduce redundancy, or through organizing data by similarity. Example problems are clustering, dimensionality reduction and association rule learning. Example algorithms include: Apriori algorithm and k-Means.
- Semi-supervised Learning Algorithms: input data is a mixture of labeled and unlabeled examples. There is a desired prediction problem but the model must also learn the structures to organize the data. Example problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.

Algorithms are often grouped by similarity in terms of how they work (e.g. tree-based methods and neural network inspired methods) [4]. This is a useful grouping method, but it is not perfect. There are still algorithms that could just as easily fit into multiple categories, like Learning Vector Quantization, that is both a neural network inspired method and an instance-based method. There are also categories that have the same name that describe the problem and the class of algorithms, such as Regression and Clustering. We handle these cases by listing algorithms twice or by selecting the group that is the best fit. In this section we list many of the popular machine learning algorithms grouped in the most intuitive way [4, 183, 185]. The list is not exhaustive in either the groups or the algorithms, but it is representative and will be useful to get an idea of the lay of the land[1].

- Regression Algorithms [115, 135, 183, 185]: regression is concerned with modeling the relationship between variables, and it is iteratively refined by applying a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be misleading because we can use the term regression to refer either to the class of problem or the class of algorithms. The most popular regression algorithms are: Ordinary Least Squares Regression, Linear Regression, Logistic Regression, Stepwise Regression, Multivariate Adaptive Regression Splines, and Locally Estimated Scatterplot Smoothing.
- Instance-based Algorithms [50, 52, 136, 172, 217]: instance-based learning models are decision problems with instances or examples of training data that are deemed important or required to the model. Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances. The most popular instance-based algorithms are: k-Nearest Neighbor, Learning Vector Quantization, Self-Organizing Maps, and Locally Weighted Learning.
- Regularization Algorithms [11, 12, 21, 144, 183, 200, 201, 231]: an extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing. We have listed regularization algorithms separately here because they are popular, powerful and generally simple modifications made to other methods. The most popular regularization algorithms are: Ridge Regression, Least Absolute Shrinkage and Selection Operator, Elastic Net, Least-Angle Regression.
- Decision Tree Algorithms [118, 129, 162, 178]: decision tree methods construct a model of decisions based on actual values of attributes in the data. Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often chosen because of their speed and accuracy. The most popular decision tree algorithms are: Classification and Regression Trees, Iterative Dichotomiser 3, C4.5, C5.0, Chi-squared Automatic Interaction Detection, Decision Stump, M5, and Conditional Decision Trees.

[1]There is a strong bias towards algorithms used for classification and regression, the two most prevalent supervised machine learning problems we will encounter

- Bayesian Algorithms [16, 38, 169, 170]: bayesian methods explicitly apply the Theorem of Bayes for problems such as classification and regression. The most popular Bayesian algorithms are: Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes Averaged One-Dependence Estimators, Bayesian Belief Network and Bayesian Networks.
- Clustering Algorithms [3, 24, 66, 209, 218]: clustering, like regression, describes the class of problems and the class of methods. Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchical. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality. The most popular clustering algorithms are: k-Means, k-Medians, Expectation Maximization, and Hierarchical Clustering.
- Association Rule Learning Algorithms [225]: association rule learning methods extract rules that best explain observed relationships between variables in the data. These rules can discover important and commercially useful associations in large multidimensional datasets that can be exploited by an organization. The most popular association rule learning algorithms are: Apriori algorithm and Eclat algorithm.
- Artificial Neural Network Algorithms [10, 29, 92]: artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching commonly used for regression and classification problems, but they represent an enormous subfield comprised of hundreds of algorithms and variations. Note that we have separated out Deep Learning from neural networks because of the massive growth and popularity in the field. Here we are concerned with the more classical methods. The most popular artificial neural network algorithms are: Perceptron Back-Propagation, Hopfield Networks, and Radial Basis Function Networks.
- Deep Learning Algorithms [79, 90, 119, 147, 180, 194]: Deep Learning methods are a modern update of Artificial Neural Networks that exploit abundant cheap computation. They are concerned with building much larger and more complex neural networks and, as commented on above, many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data. The most popular deep learning algorithms are: Deep Boltzmann Machine, Deep Belief Networks, Convolutional Neural Networks, and Stacked Auto-Encoders.
- Dimensionality Reduction Algorithms [20, 51, 69, 120, 160, 198, 212, 227]: like clustering methods, dimensionality reduction methods seek and exploit the inherent structure in the data, but in this case in an unsupervised manner or order to summarize or describe data using less information. This can be useful to visualize dimensional data or to simplify data which can then be used in a supervised learning method. Many of these methods can be adapted in classification and regression: Principal Component Analysis, Principal Component Regression, Partial Least Squares Regression, Sammon Mapping, Multidimensional Scaling, Projection Pursuit, Linear Discriminant Analysis, Mixture Discriminant Analysis, Quadratic Discriminant Analysis, and Flexible Discriminant Analysis.
- Ensemble Algorithms [34, 37, 179, 224, 229]: ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction. In this context, the effort consists in deciding which types of weak learners to choose and in which way to combine them. This is a very powerful and popular class of techniques: Boosting, Bagging, AdaBoost, Stacked Generalization, Gradient Boosting Machines, Gradient Boosted Regression Trees, and Random Forest.
- Novelty detection [132, 163, 192]: novelty and outlier detection methods address the problem of identifying new or unknown data that a data analytics system has not been trained with and was not previously aware of. Novelty detection is also referred to as one-class classification because it is trained only on the one class of known data. Novelty detection is one of the fundamental problems in a classification system. A data analytics system cannot be trained with all the possible object classes and hence the performance of the model will be poor for those classes that are under-represented in the training set. A good classification system must have the ability to differentiate between known and unknown

objects during testing. For this purpose, different models for novelty detection have been proposed: Support Vector Data Description, Gaussian Data Description, Parzen Window Data Description, Linear Programming Distance Data Description, k-Nearest Neighbor Data Description, k-Nearest Neighbor Outlier Detection, Local Outlier Factor, Local Correlation Integral, Angle-Based Outlier Detection, and Global-Local Outlier Scores from Hierarchies. Novelty detection is a hard problem in machine learning since it depends on the statistics of the already known information. A generally applicable, parameter-free method for novelty detection in a high-dimensional space is not yet known. An important application of novelty detection is the detection of a potential fault whose class may be under-represented in the training set.

Apart from the data analytics learning algorithms, there are two other main tasks to address in real world data analytics problems: feature selection, model selection, and error estimation.
In parallel from predicting the behavior of a system itself, another goal relates to the identification of the most influencing variables. For this purpose, methods like Feature Selection and Ranking may be used. Feature selection is the process of selecting a subset of relevant variables or predictors in the model construction. The central premise when using a feature selection technique is that the data contains many features that are either irrelevant or redundant, and they can thus be removed without incurring in much loss of information [82, 96, 113]. Feature ranking [42, 58, 91, 204, 222] specializes these approaches for ranking the importance of each variable in building a prediction model. In this way, variables that have the same informative content are ranked as equally important. Many feature selection algorithms include variable ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success. The goal of feature selection/ranking is three-fold: (a) improving the prediction performance of the predictive model, (b) providing faster and more cost-effective predictors, and (c) providing a better understanding of the underlying process that generated the data. There are many approaches for addressing this issue, such as: Wrappers and embedded methods, nested subset methods, direct objective optimization and filters for subset selection for feature selection, correlation criteria, single variable forecast, information theoretic ranking criteria and other optimization-based techniques for feature ranking.
Model Selection and Error Estimation [84, 152] try to answer two main simple but fundamental questions in data analytics. How can we select the best performing predictive model? How can we rigorously estimate its generalization error? Among the several methods proposed in literature for Model Selection and Error Estimation, it is possible to identify two main categories: Out-Of-Sample and In-Sample methods [9]. The first one works well in many situations and allows to apply simple statistical techniques in order to estimate the quantities of interest by splitting the data in different sets, each one for different purposes (training, validation and test). Some examples of Out-Of-Sample methods are the well-known k-Fold Cross Validation[112], Leave-One-Out, and Bootstrap [61]. Instead, the In-Sample methods exploit the whole set of available data for training the model, assessing its performance and estimating its generalization error, thanks to the application of rigorous statistical procedures. In-Sample methods can be further divided into two subgroups: the Hypothesis Space-based methods and the Algorithm-based methods [154]. The first subgroup requires to know the hypothesis space from which the algorithm will choose the model. Some examples of these methods are the Vapnik-Chervonenkis Theory [208], (Local) Rademacher Complexity Theory [18, 19, 155, 156] and PAC Bayes Theory [74, 117, 121, 134]. The second subgroup of methods do not require to know in advance the hypothesis space, but they just need to apply the algorithm on a series of modified training sets. Some examples are the Compression Bound [68] and Algorithmic Stability [31, 164].

2.1.2 Data Analytics & Interpretability

If data analytics model performs well, why not just trust the model and ignore why it made a certain decision? *The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-*

world tasks [57].
In predictive modeling, you have to make a trade-off: do we simply want to know what is predicted (i.e. the probability that a switch will brake or a score for the effectiveness of some train dispatching solution) or do we want to know why that prediction was made, possibly paying for the interpretability with a drop in accuracy? In some cases we do not care why a decision was made, only the assurance that the predictive performance was good enough but in other cases, knowing why can help understand more about the problem, the data and why a model might fail. Some models might not need explanations, because they are used in a low risk environment, meaning a mistake has no severe consequences, (e.g. a train prediction system) or the method has already been extensively studied and evaluated. The necessity for interpretability comes from an incompleteness in the problem formalization [57], meaning that for certain problems or tasks it is not enough to get the answer (the what). The model also has to give an explanation about how it came to the answer (the why), because a correct prediction only partially solves the original problem. The following reasons drive the demand for interpretability and explanations [57, 139]:

- Human curiosity and learning. Humans have a mental model of their environment, which gets updated when something unexpected happens.
- Find meaning in the world. We want to reconcile contradictions or inconsistencies between elements of our knowledge structures.
- Data analytics models are taking over real world tasks, that demand safety measurements and testing.
- By default most data analytics models pick up biases from the training data.
- The process of integrating machines and algorithms into our daily lives demands interpretability to increase social acceptance.
- Explanations are used to manage social interactions.
- Only with interpretability data analytics algorithms can be debugged and audited.

If you can ensure that the data analytics model can explain decisions, the following traits can also be checked more easily [57]:

- Fairness: Making sure the predictions are unbiased and not discriminating against protected groups (implicit or explicit). An interpretable model can tell why it decided that a certain person is not worthy of a credit and for a human it becomes easier to judge if the decision was based on a learned demographic (e.g. racial) bias.
- Privacy: Ensuring that sensitive information in the data is protected.
- Reliability or Robustness: Test that small changes in the input don't lead to big changes in the prediction.
- Causality: Check if only causal relationships are picked up, meaning a predicted change in a decision due to arbitrary changes in the input values is also happening in reality.
- Trust: It is easier for humans to trust a system that explains its decisions compared to a black box.

The most straightforward way to get to interpretable data analytics is to use only a subset of algorithms that create interpretable models. Very common model types of this group of interpretable models are:

- Linear models: linear models have been used since a long time by statisticians, computer scientists, and other people tackling quantitative problems. Linear models learn linear (and therefore monotonic) relationships between the features and the target. The linearity of the learned relationship makes the interpretation easy [88].
- Decision trees: tree-based models split the data according to certain cutoff values in the features multiple times. Splitting means that different subsets of the dataset are created, where each instance belongs to one subset. The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes. For predicting the outcome in each leaf node, a simple model is fitted with the instances in these subsets. Trees can be used for classification and regression [88].

- Rule-based models: for example, the RuleFit algorithm [70] fits sparse linear models which include automatically detected interaction effects in the form of binary decision rules.

2.1.3 Metrics

Metrics are some parameters or measures of quantitative assessment used for measurement or comparison in a given context [59, 216, 230]. A metric for all practical purposes is just a variable and it needs to be clearly defined. The number of metrics needs to be kept under control to ensure that the measuring task is achievable. It is thus reasonable to expect that, as the context changes, the metrics would change. Literature has not defined data analytics metrics as such. Data analytics metrics may be defined as a set of measurements which can help in determining the efficacy of a data analytics method or technique or algorithm. They are important to help taking the right decision as choosing the right data analytics technique or algorithm. Each type of designed model will have its own metrics by which it can be assessed, but there may be assessment tools that are independent from the type of model. In many cases, a single metric may not be sufficient to evaluate. In such cases, we might have to look at multiple metrics which can be used to validate one another and maximize the accuracy of the evaluation. Choosing the right metrics for the assessment is of paramount importance. Data analytics metrics generally fall into the categories of accuracy, reliability, and usefulness.
Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that have been provided. There are various measures of accuracy, but all measures of accuracy are dependent on the data that is used. In reality, values might be missing or approximate, or the data might have been changed by multiple processes. Particularly in the phase of exploration and development, we might decide to accept a certain amount of error in the data, especially if the data is fairly uniform in its characteristics. For example, a model that predicts sales for a particular store based on past sales can be strongly correlated and very accurate, even if that store consistently used the wrong accounting method. Therefore, measurements of accuracy must be balanced by assessments of reliability.
Reliability assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless the test data that is supplied. For example, the model that we generated for the store that used the wrong accounting method would not generalize well to other stores, and therefore would not be reliable.
Usefulness includes various metrics that tell us whether the model provides useful information. For example, a data mining model that correlates store location with sales might be both accurate and reliable, but might not be useful, because one cannot generalize that result by adding more stores at the same location. Moreover, it does not answer the fundamental business question of why certain locations have more sales. We might also find out that a model that appears successful in fact is meaningless, because it is based on cross-correlations in the data.
Measuring the effectiveness or usefulness of data analytics approach is not always straightforward. In fact, different metrics could be used for different techniques and also based on the interest level. From an overall business or usefulness perspective, a measure such as Return on Investment (ROI) could be used. ROI examines the difference between what the data analytics technique costs and what the savings or benefits from its use are. Of course, this would be difficult to measure because the return is hard to quantify.

2.2 Visual Analytics & Metrics

The challenges in the IN2DREAMS project comprise a variety of tasks substantially differing in nature: Processing and presenting, analysis and decision making, simulation and interpolation are very different duties whose integration cannot be achieved with conventional solutions. In addition, extracting relevant and meaningful information from heterogeneous data sources is notoriously complex and cumbersome. Researchers

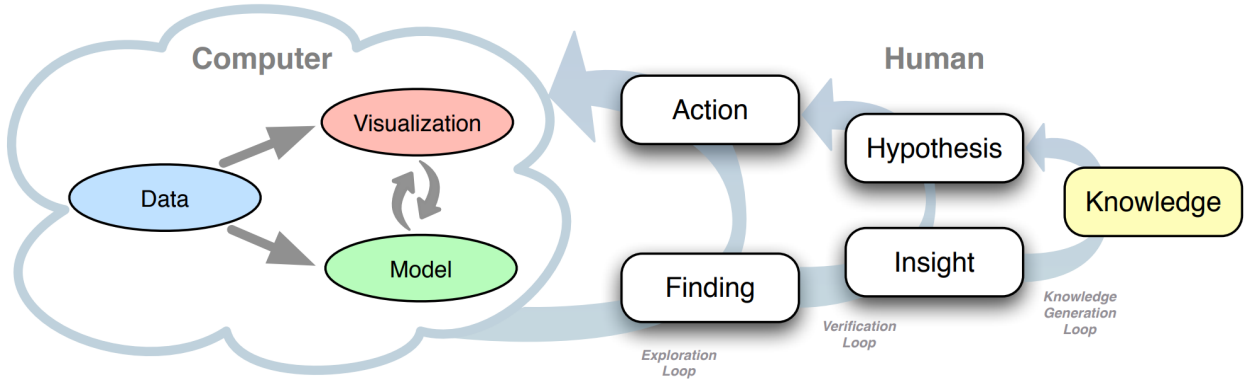

Figure 1: The knowledge generation model describes the relation of data, models, visualization as well as how the human is interfaced with this triple in order to generate knowledge.

have been trying to solve these problems through either automatic data analysis or interactive visualization approaches. However, only the combination of both approaches allows to leverage both the computational power of modern algorithms and machines as well as a user's experience and unmatched ability to perceive and interpret patterns.
Visual Analytics (VA) is an interdisciplinary approach towards complex data analysis scenarios based on this combination of man and machine. VA "combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets", a definition given by Keim et al. as summary of the VisMaster EU research project [105]. Besides direct knowledge generation, following Visual Analytics principles also fosters a user's constructive reflection and correction of conducted analyses, resulting in improvements for processes and models, and ultimately, of decisions taken and knowledge generated by the users.
VA combines multiple research areas and subjects including data management and analysis, spatio-temporal data processing, statistics, human-computer-interaction and visualization [108]. It is intended to allow to derive insights from large, in-homogeneous and ambiguous datasets and enables both to confirm expected results as well as finding unexpected coherence. Users can quickly come to comprehensible, verifiable results and are able to communicate their findings and derived consequences for action efficiently.

The Visual Analytics process has been described and modeled extensively. A refined view is provided by Sacha et al. [176] with the introduction of the Knowledge Generation Model. This model for VA defines and relates human and machine concepts embedded in a three-loop framework [176]. The model is shown in Figure 1 and consists of the VA process model on the left hand side and is related to human knowledge generation process on the right hand side. The model clearly conveys that lower-level processes, which are part of the exploration loop, are guided by higher-level analytic activities, which are part of the verification and knowledge generation loops. A typical analysis flow can be described as follows. Based on previous knowledge, which is always part of the analysis, humans derive hypotheses. Based on a single hypothesis an analyst forms strategies to collect pieces of evidence that change, refine, confirm, or reject the hypothesis and enters the verification loop. The verification loop guides the exploration loop that covers all direct interactions with the system. That means that all actions produce a reaction of a system. Even no reaction of system could be interpreted as a result of previously applied actions. Humans stay in the exploration loop until they find something that can be used for further analysis. Findings can be visual patterns, missing values or model outputs – anything that could be of interest to the analyst. As a next step, humans start to interpret their findings with respect to their problem domain and relate these to their hypothesis. At this stage, findings

have developed into insights and the analyst left the exploration loop. These processes may happen several times until the confidence or trust in these pieces of evidence are high enough to become knowledge in the end (knowledge generation loop).
Applied to the IN2DREAMS scenarios (Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7), the VA process will be used to improve the decision-making process of analysts in complex use cases that require to simultaneously oversee impending consequences of collaboratively made decisions and to communicate these consequences to further affected organizational units. As well, knowledge generated by decision-making results and reviews of the consequences can be used to develop long-term strategies.

2.2.1 Metrics and Uncertainty

The measurement approach discussed in Section 2.2 opens a new area of meta-data analysis on top of exploring the fundamental data sources: Metrics are a set of derived measurements that can express performances of prediction models or the quality of aggregated data that has been transformed, interpolated or sampled for representation.
Considering the choice and handling of such metrics to improve the analysis results is a sensitive endeavour. In Visual Analytics environments, quality metrics play an important role when it comes to dealing with high-dimensional and/or heterogeneous data sets with unkown correlations, and an extensive amount of research has been conducted in the area (see [27, 105] for an overview). Schneidewind et al. [181] show, how pixel-based visualizations can be improved using quality metrics.
As well, metrics can be used to quantify uncertainties within the displayed information: A vital part of the decision-making-process is being informed about the data that is processed and explored. Inherently, most real-world data sources are inaccurate to a certain degree and subject to varying data quality. This uncertainty can be the result of many influences, from plain errors over sensor inaccuracy to information loss through spatial or temporal interpolation. Hunter and Goodchild define uncertainty as “degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations without caution” [93]. Sacha et al. [174, 175] have explored how uncertainty and awareness about it affects a user’s trust in the system she is operating. A wealth of publication is available on how to communicate uncertainties visually [44, 95, 182].
Consequently, effective data analysis in complex scenarios such as the ones solved in the IN2DREAMS project can only be conducted if sources and amount of uncertainties are considered: Algorithmic methods have to incorporate uncertainty information and visual representations and interfaces have to communicate uncertainties to the expert in order to prevent biased decisions made on false grounds.

2.3 General Legal Requirements and Constraints

The following section intends to give a general overview of a few major legal challenges arising from the design and use of data analytics. Deliverable D5.5 will analyze more thoroughly the legal challenges posed. We will firstly present the challenges pertaining to the legal status of data (Section 2.3.1); then we will indicate the legal regime generally having an impact on the design and use of data analytics models (Section 2.3.2).

2.3.1 Rights related to data

Effective and efficient data analytics is predicated on the wide availability of training data. The internal market for data has notably been identified as the “fifth” freedom in EU law[2]. However, different training sets may be subject to diverging sets of legal rules depending on the nature of the data at stake. Different legal

[2] See, e.g., the Estonian Vision Paper on the Free Movement of Data - the Fifth Freedom of the European

frameworks yield different rights and obligations for the parties such as training set providers, data curators and individuals. This section thus outlines the different legal regimes which might be applicable to data.
While literature has generally identified that there is no "ownership" right on data as such[3] and especially not on personal data[4] , it is a common understanding that data are - directly or indirectly - subject to different rights and obligations[5] including sector-specific regulation (such as data in intelligent transport systems). This leads to a legal patchwork and legal uncertainty as to the legal nature of data when data is the object of a transaction[6].
Especially, literature has insufficiently clarified under which conditions the GDPR allows for transactions of personal data on the basis of "consent" of the data subject[7]. The literature has mostly advised against the creation of a data ownership right to clarify the legal status of data and enhance data transactions[8] . However, it has not sufficiently brought solutions as to what legal regime should then apply to data and notably whether it is advisable to enact general regulation or to rather opt for sector-specific solutions with a view to sector-specific challenges and data protection considerations.
(Big) "data" is defined in literature as the "Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value"[9]. The volume, variability and complexity of big data rendered most of the traditional tools and techniques for technical processing inefficient. Furthermore, the extracted value from big data does not really originate from the data themselves but rather from the techniques that are applied to these data to extract value, patterns etc. which may otherwise remain hidden and unknown. However variable the data may be, typically in law data are classified as either personal or non-personal.

Personal and non-personal data In EU law, there is a dichotomy between personal and non-personal data. For example, the General Data Protection Regulation defines "personal data" as simply "any information relating to an identified or identifiable natural person"[10]. The scope of what an "identifiable natural person" means is very board and it includes anyone who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. The GDPR applies only to personal data, i.e. it does not apply to anonymous information[11]. As long as information does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable, the GDPR will not apply. As simple as it may sound, the interpretation of what "anonymous" means has caused a lot of controversy in both academic literature and practice[12]. Furthermore, anonymization itself constitutes further processing of personal data and must, as such, be in accordance with the requirement of compatibility of further processing of personal data.

[3] Zech, "Data as a Tradeable Commodity - Implications for Contract Law"; Drexl, "Designing Competitive Markets for Industrial Data - Between Propertisation and Access"; Farkas, "Data created by the internet of things"; Determann, "No One Owns Data".

[4] Janecek, "Ownership of Personal Data in the Internet of Things".

[5] Drexl, "Designing Competitive Markets for Industrial Data - Between Propertisation and Access".

[6] Graf von Westphalen and Westphalen, "Contracts with Big Data". Sein, "What Rules Should Apply to Smart Consumer Goods?"

[7] Clifford, Graef, and Valcke, "Pre-Formulated Declarations of Data Subject Consent - Citizen-Consumer Empowerment and the Alignment of Data, Consumer and Competition Law Protections".

[8] Wiebe, "Protection of Industrial Data - a New Property Right for the Digital Economy?" Drexl, "Designing Competitive Markets for Industrial Data - Between Propertisation and Access"; "The "Data Producer"s Right?".

[9] Andrea De Mauro, Marco Greco, and Michele Grimaldi, "A Formal Definition of Big Data Based on Its Essential Features".

[10] Article 4(1) GDPR.

[11] Recital 26 GDPR.

[12] In its Opinion 05/2014 on Anonymisation Techniques, the Article 29 Working Party took the view that anonymization is truly achieved when it irreversibly prevents the identification of data subject. Effectively, this means that there should be no feasible way to achieve identification of the subject following the anonymization. Full text of the opinion is available here: `http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`

Unlike the definition of personal data in the GDPR, presently there is no single, uniform legal definition in EU law of what constitutes "non-personal" data. In 2017, as part of its Communication "Building a European Data Economy", the EC announced a Proposal for a Regulation of the European Parliament and of the Council on a framework for the free flow of non-personal data in the EU[13]. Notably, the large policy goal of this proposal is to "ensure the free movement of data other than personal data within the Union by laying down rules relating to data localisation requirements, the availability of data to competent authorities and data porting for professional users."[14] Non-personal data is defined negatively as data other than personal data as interpreted under the GDPR[15].
This proposal is only a piece of the fragmented puzzle of legal frameworks applicable to data across the EU. Thus, for example, such rules may stem from legal frameworks applicable to databases (e.g., Directive 96/9/EC, known as the Database Directive), public data (e.g., the Public Sector Information Directive 2013/37/EU), confidential data (e.g., the Trade Secrets Directive 2016/643).

Fragmented rules on the use of data in the EU The PSI legal framework consists of the Directive on the re-use of public sector information (Directive 2003/98/EC ('PSI Directive') and Directive 2013/37/EU which amends it. Public sector information is defined as information held by the public sector, i.e. government held data. The scope of the directive is re-use of documents collected by public sector bodies in the framework of their public tasks and which are accessible. Public sector bodies with commercial or industrial activities in nature as well as private entities entrusted with fulfilling public sector tasks are excluded from the scope. Reuse is defined as use by persons or legal entities of documents held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced. Exchange of documents between public sector bodies purely in pursuit of their public tasks does not constitute re-use. The directive regulates the release of datasets for reuse of such information, but not the access thereto. Finally, it should also be noted that there is a common misunderstanding that all public information is open data. The harmonized PSI regime is based on the various non-harmonized national regimes of access to information which are often rooted in the freedom of information legislation. In addition to the PSI regime, there are also other legal frameworks that govern open and public data, such as the INSPIRE (Infrastructure for Spatial Information in Europe) directive, the PAEI (Public Access to Environmental Information) directive etc.
The legal framework on copyright protection does not apply as such to data but only to expressions in an original form. However, the control provided by copyright effectively means that one can prohibit the use of the composing data. Furthermore, under copyright law, if a data collection (e.g., a dataset) is the result of an original selection and the curation itself is original, such a dataset may be entitled to copyright protection. In addition to the general copyright protection, in EU law, there is a sui generis right to database protection, which is independent from the traditional copyright regime. This right is given to the producer of a "collection of data" provided there is a substantial investment made in obtaining, verification and presentation of the data. However, investment in the creation of the data themselves does not give rise to protection; it is only extended to the extraction of "substantive parts" of the database and not of individual small amounts of data.
The legal framework on trade secrets established with the Trade Secret Directive is relatively new in EU law. A trade secret is defined as information which meets all of the following requirements: (1) it is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question; (2) it has commercial value because it is secret; and (3) it has been subject to reasonable steps

[13] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a framework for the free flow of non-personal data in the European Union, COM/2017/0495 final - 2017/0228 (COD), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2017%3A495%3AFIN
[14] Article 1 of the Proposal.
[15] Article 3.1 of the Proposal.

under the circumstances, by the person lawfully in control of the information, to keep it secret[16].
As a general conclusion from this overview on the different frameworks applicable to data, it could be said that there is a growingly confusing use of the terminology in the first place. The parlance in the domain of data is becoming more and more property-oriented, whereas what is at stake in essence is not the ownership "of" data but rather the control "over" data which has more or less been dealt with to a certain extent by the existing legal framework, as discussed supra. Moreover, some authors, such as Alain Strowel, argue that if data are already the subject of commercial transactions, it makes little sense to devise a new, property-like right to govern these transactions[17].
The real-world concerns related to data seem to boil down, on the one hand, to possible anti-competitive behavior of players with dominant market positions, and, on the other hand, to possible violations of the fundamental rights and freedoms of citizens as a result of unlawful processing of personal data, especially when it comes to automated decision-making under the GDPR.

2.3.2 Legal regime having an impact on data analytics

In environments characterized by high level of complexity, deterministic computing (explicit specifications and programming) would be hardly possible or extremely costly. In order to deal with that complexity, data analytics systems discover correlations between data: they are based on inductive reasoning - as opposed to deductive reasoning[18] . Machine learning goes a step further as it "refers to the automated detection of meaningful patterns in data"[19] . For the purpose of the analysis of legal challenges, the determination of the actual computing activities at stake is crucial. So are also the concrete environment in which the model takes place, the purpose assigned to it, the entity in charge of the operations, the nature of data at stake and the entities on which such computing activities have an impact.

Practical issues with datasets containing personal data: automated decision-making, the right to an explanation and machine learning It is a general rule that, to the extent, certain data contain information relating to an identified or identifiable natural person, the rules of data protection law apply. There is one specific concern, though, which has recently caused a lot of controversy amongst both academics and practitioners. Thus, with the adoption of the GDPR, questions have been raised as to whether the requirements of Article 22 on automated decision-making actually impede the use of machine learning techniques in the EU.
The provision of Article 22 stipulates that a data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. Furthermore, the GDPR provides that data controllers must provide information to data subjects on the existence of automated decision-making, including profiling, and, at least in the cases of decisions which produce legal effects for them or concern the processing of special categories of personal data under Article 9 GDPR, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject[20].
This provision has been said by many to introduce a general right to an explanation when it comes to machine learning algorithms[21], but there are also dissenters[22]. Such a right to an explanation effectively means that

[16] Article 2(1) Trade Secrets Directive.
[17] See, e.g., https://nexa.polito.it/nexacenterfiles/STROWEL_BigData&Platform.pdf
[18] Understanding machine learning; loyaute des decisions algorithmiques
[19] Shalev-Shwartz and Ben-David, Understanding Machine Learning.
[20] Article 13.2(f) and 14.2(g) GDPR.
[21] Andrew D Selbst, Julia Powles; Meaningful information and the right to explanation, International Data Privacy Law, Volume 7, Issue 4, 1 November 2017, Pages 233-242, https://doi.org/10.1093/idpl/ipx022
[22] Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). International Data Privacy Law, 2017. Available at SSRN:

data controllers may be obliged to provide a clear explanation of how the algorithms they use work ("global explanation") and which factors have contributed to one decision or another being made ("local explanation")[23].
The Article 29 Working Party has recently noted that while the GDPR requires the controller to provide meaningful information about the logic involved, such an "explanation" does not necessarily have to be "a complex explanation of the algorithms used or disclosure of the full algorithm"[24]. However, it must be possible for the data subject to deduce the reasons for a particular decisions based on this explanation. Furthermore, the Article 29 Working Party has provided some guidelines on how this right to information may be effectively implemented. It highlights that "[i]nstead of providing a complex mathematical explanation about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver the information to the data subject"[25]. Examples of such ways provided by the working party include listing: (1) the categories of data that have been or will be used in the profiling or decision-making process; (2) explanation as to why these categories are considered pertinent; (3) how any profile used in the automated decision-making process is built, including any statistics used in the analysis; (4) why this profile is relevant to the automated decision-making process; and (5) how it is used for a decision concerning the data subject[26].
It follows that, to the extent a dataset may contain personal data and to the extent an automated decision making may produce legal consequences for a data subject, data controllers must observe these requirements in both the design of machine learning algorithms and the actual processing activities. However, one can reasonably imagine that while this may be more or less straightforward in less complicated, single-system environments, this will not be the case in "black box" environments of complex systems which continuously interact with each other.

Railway-specific regulation having an impact on data analytics There is no specific regulation of data analytics applicable to industrial data. However the railways and especially railway infrastructure management are subject to sector-specific regulations that can be divided up in two sets of rules: (1) market regulation and (2) technical and safety regulation.

- **Market regulation**

  Applicable legal regime The railway sector has experienced drastic regulatory changes initiated from the European Union since the 1990s in order to open it to competition (see section xx in Deliverable D4.1). Whereas infrastructure management (performed by an infrastructure manager, "IM") may remain a monopolist activity due to its natural monopoly character, the carriage activity (performed by railway undertakings, "RUs") has gradually been liberalized. By doing so, the European law-maker created the market of the use of the infrastructure characterized by the allocation of infrastructure capacity ("train path") from the IM to the RUs upon payment of track access charges. The governance of the IM as well as its relationships towards its customers RUs are heavily regulated in order to ensure non-discrimination (especially vis-à-vis new entrants) and more generally fair treatment. A "regulatory body" was created with the competence to supervise compliance[27] .

`https://ssrn.com/abstract=2903469` or `http://dx.doi.org/10.2139/ssrn.2903469`.

[23] See, e.g., `https://www.kdnuggets.com/2018/03/gdpr-machine-learning-illegal.html`.

[24] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Adopted on 3 October 2017 As last Revised and Adopted on 6 February 2018, 17/EN WP251rev.01, p. 25.

[25] Ibid. 31.

[26] Ibid.

[27] See Directive 2012/34/EU of the European Parliament and of the Council of 21 November 2012 establishing a single European railway area (recast) as consolidated.

In its relations with its RU customers considered as weaker parties deserving legal protection, the IM is subject to general principles of transparency, non-discrimination and fairness[28] . With regard to the identified use cases, article 7b of Directive 2012/34[29] (impartiality of the infrastructure manager in respect of traffic management and maintenance planning") is of particular relevance. "It provides for the obligation falling onto Member States to ensure that "the functions of traffic management and maintenance planning are exercised in a transparent and non-discriminatory manner [...]" (article 7b.1). This provision further reads "as regards traffic management, Member States shall ensure that railway undertakings, in cases of disruption concerning them, have full and timely access to relevant information [...]". Further, article 7b.3 reads "[...] the scheduling of maintenance works shall be carried out by the infrastructure manager in a non-discriminatory way"[30] .

General legal challenges The challenges mostly arise out of the prediction phase when pre-discovered correlations are "applied to new data to generate predictions [and] recommendations [...]??[31]. Discrimination (i.a. against customers) or otherwise inaccurate or unfair treatment may arise from the model as already documented by literature[32] and as expressed in section xx of the present deliverable: it may stem from the input datasets (e.g. historical data) or from the various stages of the operation of the model. These challenges are reinforced by the fact that the operation of the system remains more or less obscure to human scrutiny[33] while the severity varies according to the actual computing process at stake. It is considered more acute in the case of machine learning where the output depends not only on the original input but also on the various interactions of the machine with its environment after its release, making it "a moving target"[34] . Against this background, the situation in our scenarios shares similarities with the use of data analytics by e-commerce operators to profile consumers in various settings[35]. The similarity lies in the fact that the stronger party - the e-commerce operator or the IM - uses data analytics to make decisions impacting the customers deemed weaker party - consumers or RUs. The use of data analytics may lead to unfair, discriminate or otherwise inaccurate treatment without the weaker parties to be able to challenge it.

Specific legal challenges in the railways The railway regulatory framework is functional. Legal obligations shall be complied with by the IM when dealing with the relevant activities, indifferently from the means used for doing so. Reliance on data analytics and machine-learning is not prohibited as such but in doing so the IM shall comply with the regulatory framework. The extent to which the IM decisions are based on data analytics and reciprocally the determination of the segregation of duties between human and machine hereby seems crucial[36] .

While the regulatory framework is aimed to protect RUs as weaker parties, there is no accountability principle in directive 2012/34 so that - pursuant to the general principles of law - the IM does not

[28] See article 10 and 13.1 of Directive 2012/34

[29] This article was recently included by Directive (EU) 2016/2370 of the European Parliament and of the Council of 14 December 2016 amending Directive 2012/34/EU as regards the opening of the market for domestic passenger transport services by rail and the governance of the railway infrastructure. It shall be transposed by the Member States by 25th December 2018 (article 2.1 of Directive 2016/2370).

[30] With regard to traffic management, see also article 16 and 17 of Regulation (EU) No 913/2010 of the European Parliament and of the Council of 22 September 2010 concerning a European rail network for competitive freight as consolidated.

[31] Price, "Regulating Black-Box Medicine".

[32] "ACCOUNTABLE ALGORITHMS | Secondary Sources | Westlaw".

[33] Edwards and Veale, "Enslaving the Algorithm". Desai and Kroll, "Trust But Verify: A Guide to Algorithms and the Law".

[34] Price, "Regulating Black-Box Medicine".

[35] Edwards and Veale, "Enslaving the Algorithm".

[36] (Ambrose) Jones, "The Ironies of Automation Law".

bear the burden of proof of compliance towards the RUs or regulatory body so that the latter need to demonstrate harm. Where the IM decisions are based on data analytics, it may prove difficult by lack of understanding of the process having led to the decision[37] . RUs may however appeal to the regulatory body in case they have been "unfairly treated, discriminated against or in any other way aggrieved"[38] and the regulatory body is legally entitled to make enquiries and to ask for any relevant information to the IM[39] . In addition the regulatory body may "decide on its own initiative on appropriate measures to correct discriminations [...], market distortion and any other undesirable developments [...] in particular with reference to [traffic management and renewal planning of maintenance]"[40] . In a situation where the IM decisions impacting RUs are based on opaque data analytics one may only wonder - by lack of precedent - how the regulatory body would react in order to exert genuine supervision. In cannot be excluded that the regulatory body would consider the opacity of the data analytics system per se as a breach of the general principle of transparency applying to the IM and/or as unfair treatment[41].

- **Safety and technical regulation**

Applicable legal regime The railway sector is subject to technical and safety regulations with a view to (1) ensure technical harmonization and interoperability amongst the operations of the various actors while (2) preventing incidents in the course of railways operations. The sector is characterized by complex regulatory framework: safety and technical regulation are harmonized at the EU level by the Railway Safety and Railway Interoperability directives[42] which are complemented by Technical Specifications for Interoperability ("TSIs") of the European Commission[43] , by national law and internal regulations of the railway actors. Besides, TSIs and other regulations may refer to technical standards with more or less binding effect[44] . Technical and safety regulations are closely related[45] . Generally and for reasons pertaining to both the systemic character of the railways and the need for safety, three layers of certifications and authorizations are required to operate in the railways: technical certification of components as such[46] (1), authorization to place in service railway subsystems (e.g. a train) with a view to their compatibility with the railway system as a whole (interoperability)[47] (2); certification of safety organization: the Safety Management System[48] ("SMS") of the railway professional is certified intuitu personae (3).

[37] Price, "Regulating Black-Box Medicine".
[38] Article 56.1 of Directive 2012/34
[39] Article 56.8 of Directive 2012/34
[40] Article 56.9 of Directive 2012/34
[41] Similarly the French Cour de Cassation recognized in 2012 that the lack of clear information in prioritized referencing is likely to distort the economic behavior of consumers and thereby qualifies as unfair commercial practices within the meaning of French law transposing directive 2005/29, see Cass. Com., 4 déc. 2002, N° de pourvoi: 11-27729 as quoted by Castets-Renard, "Loyauté des traitements et décisions algorithmiques Aspects juridiques".
[42] See Directive (EU) 2016/797 of the European Parliament and of the Council of 11 May 2016 on the interoperability of the rail system within the European Union and Directive (EU) 2016/798 of the European Parliament and of the Council of 11 May 2016 on railway safety
[43] See article 4 of the Railway Interoperability Directive. The railway system is divided between subsystems. The TSIs consist of technical rules applying to a certain subsystem with a view to its compatibility with the rest of the railway system. The subsystems are listed in Annex II of Railway Interoperability Directive: for instance, "operation and traffic management" of "maintenance" are functional subsystems while "infrastructure" and "rolling stock" are structural subsystems.
[44] See article 4.8 of the Railway Interoperability Directive: when TSIs make an "explicit, clearly identified reference to European or international standards or specifications or technical documents [they] shall be regarded as annexes to the TSI concerned and shall become mandatory [...]".
[45] See notably the definition of "interoperability" in Railway Interoperability Directive, article 2.2: "Interoperability means the ability of a rail system to allow the safe and uninterrupted movement of trains which accomplish the required level of performance".
[46] See chapter III of the Railway Interoperability Directive
[47] See chapters IV and V of the Railway Interoperability Directive
[48] See article 9 of the Railway Safety Directive

The safety regulation includes harmonized safety targets and methods for achieving it[49] . Core to safety management is the concept of "risk acceptance criteria or target safety levels"[50] to determine the minimum applicable safety levels. Safety regulation specifically targets the "main actors in the Union rail system"[51] - the infrastructure manager and the railway undertakings - as the ones bearing main responsibility for the safety[52] of the operation of their part of the railway system, in their capacity as railway professionals which shall also include "supply of material and contracting of services [...]"[53] . The railway professionals shall notably "control [the] risks associated [with the safe operation of their part of the railway operation]" and "implement the necessary risk control measures [...][54]. Where a safety risk is identified, the railway professionals shall "take any necessary corrective measure to tackle the safety risk identified"[55] on the basis of return of experience. They shall contractually "oblige the other actors to implement risk control measure" when the latter have "potential impact on the safe operation of the rail system" and they shall "ensure that the contracts [they conclude with contractors] implement risk control measures"[56]. A National Safety Authority (NSA) is established in every member State[57]. Its tasks are generally to grant authorizations of placing onto the market of relevant components and safety authorizations and to supervise compliance of railway professionals with the safety regulatory framework.

Technical certification and standards: technical certification in the railways[58] is based on compliance of a product or process with pre-determined rules and standards on a static basis (rule-based approach). Certification of machine learning systems - where appropriate - may come as a challenge, notably with regard to their core feature of learning loop after release, on the basis of external input. European Union law has already taken note of this challenge to the integration of innovation[59] but it remains to be seen how it can concretely be overcome.

Safety management - control Safety responsibility is assigned to the certified railway professionals (IM and RUs) for their part of the railway system and reciprocally they shall retain control and oversight over safety-related activities, even when delegated to third parties. Where data analytics systems are designed to be used for safety-critical activities, the opacity of the system may come as a challenge for the IM to genuinely retain control and oversight. Core to the organization of safety management is how human operators and machine systems interact.

[49] Common safety indicators, common safety methods and common safety targets, see article 5 to 7 of the Railway Safety Directive as implemented by the European Commission and regularly updated.
[50] Article 7.1 of the Railway Safety Directive
[51] Recital (7) of the Safety Directive
[52] See recital (7) and (8) and article 4 of the Safety Directive
[53] Article 4.1.a of the Railway Safety Directive
[54] Article 4.1.d and 4.3.a of the Railway Safety Directive
[55] Article 4.5.a of the Railway Safety Directive
[56] Article 4.3.c of the Railway Safety Directive
[57] Article 16 of the Railway Safety Directive
[58] As in other transport sectors such as aviation, see Emanuilov, "Autonomous Systems in Aviation: Between Product Liability and Innovation".
[59] See Commission delegated decision (EU) 2017/1474 of 8 June 2017 supplementing Directive (EU) 2016/797 of the European Parliament and of the Council with regard to specific objectives for the drafting, adoption and review of technical specifications for interoperability. Article 3.3 reads "the TSIs shall be reviewed where appropriate to ensure the right balance between rule-based and risk-based approaches". It is supplemented by recital (7) which in substance distinguishes technical rules aimed at (i.) technical compatibility between subsystems which would need still to be based on a rule-based approach and (ii.) those aimed at "specifying functions and performances" which could be subject to risk-based approach. The EC Decision makes an explicit reference to Shift2Rail in recital (8) with a view to adapt the railway technical regulatory framework to innovation.

Safety management - assessment of risk Railway safety is based on assessment of risks and determination of acceptable levels of risks. Where relevant, this means that the design of the data analytics and machine learning systems should come with error assessment[60] in order for the IM to make a well-informed choice pertaining to the acceptance of safety risks. More generally the methods for assessing the risks are likely to be disrupted by the shift of paradigm brought by machine learning.

High standard duty of care and liability The safety management system comes with a high duty of care falling onto the railway professionals; in particular, they bear the responsibility to take measures when safety risks are identified (see above) failing what they may notably be held liable in case of subsequent damage, subject to national law[61]. Against this background, the wealth of predictions produced by data analytics and machine learning systems may have a chilling effect on railway professionals as additional information brings additional obligations.

3 The Considered Railway Ecosystem

This section considers and describes the railway ecosystem, its connections with IN2DREAMS WS2 and WP4, other IN2DREAMS WSs and WPs and SHIFT2RAIL recipients, and the potential impact of Data and Visual Analytics.

3.1 Description of the Considered Railway Ecosystem

To better identify relevant use cases, it is important to understand the considered railways ecosystem. Not all that composes the railways ecosystem may be considered in the scope of our project: our goal is to pave the way towards intelligent asset maintenance, aiming at supporting the deployment, usage, and maintenance of railway assets in a more effective and cost-efficient way. This task should be achieved with the use of data and analytics tools and techniques assessed with railways specific metrics by taking also into account the legal constraints in using these technologies. These tools rely on the historical and real time data collected from the different parts of the railway ecosystem and from exogenous sources collected in the different data lakes of the different railway ecosystem actors.
Consequently, we can consider in scope of our project all the components and processes inside the railways ecosystem that deal with the integration and automation of the asset management. It is important to define that for asset we intend all the physical and non-physical entities that compose the railway infrastructure. Example of classical assets are: switch, crossing, track, catenary, bridge, tunnel, embankments, line sections, and level crossing. Other less classical assets or less easy to define as an asset are trains. A train is an asset only if we consider it as an object that is employing a specific line of the infrastructure preventing others to use it; but it is not an asset per se, as it is not owned by the IM. Another less classical asset is the TMS which is not an asset per se but can be considered an asset if we want to improve its performance and functionalities. Even if the railways ecosystem is composed by a multitude of different actors, talking about the asset management, we have identified three main actors:

- The IM, responsible of the physical infrastructure

[60] From a more general perspective on this matter, see Castets-Renard, "Loyauté des traitements et décisions algorithmiques Aspects juridiques". The authors consider that such a requirement shall be imposed on machine learning, similar to French law applicable to polling institutes.

[61] While extra-contractual liability (ex delicto) is not harmonized and depends upon national law, contractual liability is harmonized at international level in the case of international transport by the International Convention concerning International Carriage by Rail (COTIF 1999 as amended by the Vilnius Protocol)

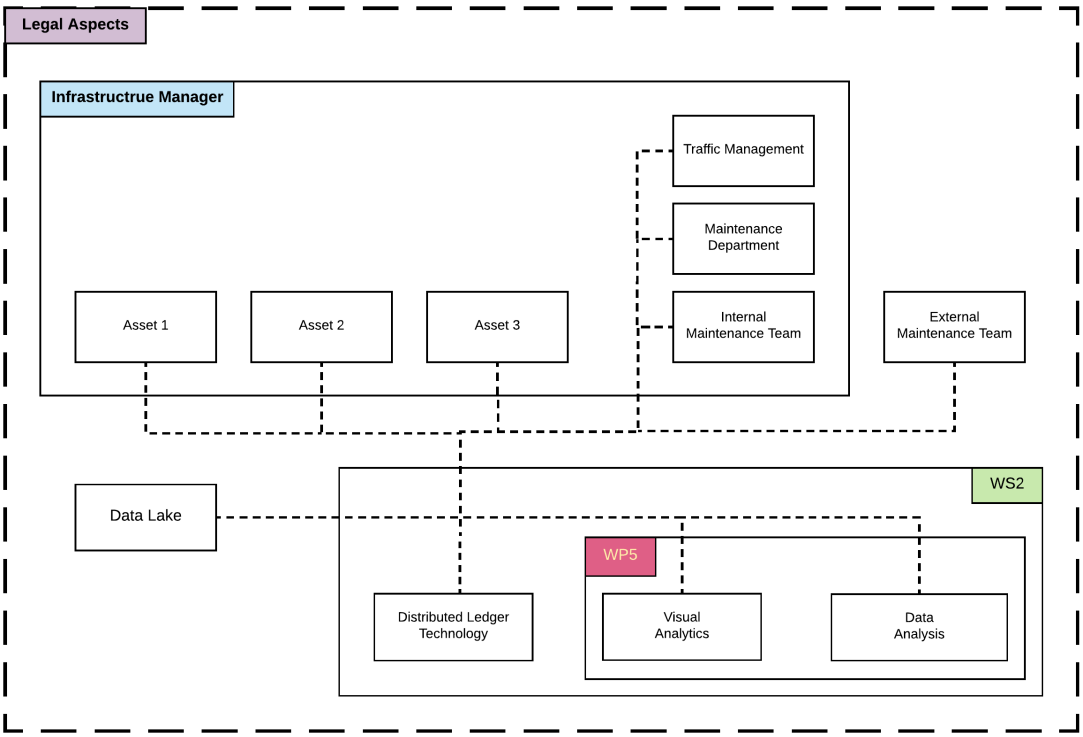

Figure 2: The Identified Railway Ecosystem

- The Suppliers, that provide assets and services to the IM
- The Maintainers, responsible of the maintenance of a specific physical asset on behalf of the IM

A railway network is managed by a company that has the role of IM; managing the network means controlling access to the (and usage of) infrastructure with the TMS, planning and performing maintenance, and upgrading railway lines when needed. The IM makes sure that all the maintenance activity does not disrupt network traffic and ensures reliability and safety. Suppliers design and manufacture the railways components, according to the technical specifications that the IM requires and they also offers services to the IM for the revamping and maintenance of existing structures. The IM schedules maintenance intervention with the goal of keeping a desired standard of service and to preserve the reliability of the network. These tasks are planned according to some criteria chosen by the IM. Maintenance teams, coordinated by the TMS, take care of the maintenance tasks working with the maintenance contractors. Maintenance can be performed both by the IM or by maintenance subcontractors. In Figure 2, we report a simplified representation railway ecosystem and the links between the various actors.
According to Rail Safety and Standards Board (RSSB), research, development and innovation in the railway sector are currently following the technical strategy [81] defined by the following key aspects:

1. Running trains closer together: it reduces the distance between adjacent trains while not reducing the high safety levels, in order to increase the capacity of railway networks and the frequency of trains.
2. Minimal disruption to train services: it reduces as much as possible the number of incidents that cause disruptions to the railway operations. This can be achieved by improving preventive maintenance, reducing corrective one, and implementing on condition and predictive maintenance in order to increase the railway assets reliability. From the operational point of view, intelligent transportation systems able to take into account the state of the railway network, including current traveling trains as well as way-

side assets, and the possibility to simulate and suggest options for decision making will greatly increase the performance in case of disruptions.

3. Efficient passenger flow through stations and trains: a more effective, optimized planning of multi- and railway-modal interfaces will generate added value for both the passengers, which would be able to increase their trust in the railway transportation, and the train operators, which could create new services and traveling possibilities, possibly increasing their revenues. Smarter ticketing and human-centered design will make moving through stations and trains easier and quicker reducing overcrowding at busy stations.
4. More value from data: the growing availability of data and the increasing perception of its value will pave the way towards the research and development of new data-driven services for reducing asset life cycle costs and increasing revenues through added value services to railway customers. In particular, new data collection systems and real-time information will help rail staff to make better decisions and will provide customers with useful and up to date information.
5. Optimum energy use: intelligent distribution, energy storage technologies, green energy and energy usage optimization will deliver more cost effective use of energy on the railway.
6. More space on trains: increase the number of passengers on trains while keeping the same service standards, through large rolling stocks renewal investments.
7. Service timed to the second: knowing the exact location and speed of all trains in real time will improve situational awareness, increase operational flexibility and allow for faster recovery from disruption. Optimized timetable planning and service disruption minimization strategies will increase railway network efficiency. All these advancements will affect positively the reliability perception of customers with respect to railways.
8. Intelligent train: internet of things devices will increase self maintenance of trains, on condition and predictive monitoring, and will enable anomaly detection thanks to distributed intelligence. Intelligent trains will be aware of themselves and their surroundings, knowing where they need to be and when, and able to automatically adjust journeys to meet demand.
9. Personalized customer experience: providing passengers with tailored information and services so that travel by rail becomes a seamless part of their overall journey. Moreover, advanced ticketing, marketing, personalized discounts and advertising, as well as onboard offers, will increase passengers engagement and retention. Freight customers will be provided with value added services, such as advanced goods tracking.
10. Flexible freight: composition of freight trains and good wagons will be improved with higher flexibility, aiming at minimizing carbon footprint, increasing efficiency, time of delivery, and railway network occupations. Trains designed to carry varying loads, combined with better planning and tracking capabilities, will increase flexibility and capacity for freight customers.
11. Low cost railway solutions: reducing the cost of railway solutions through off-the-shelf technologies, relying more and more on data and so reducing asset management and operational costs. Railway lines and trains which are designed, built and operated at low cost will make lightly used lines viable and allow rail to compete for new transport links.
12. Accelerated research, development and technology deployment: implementing agile methodologies in research and development, streamlining bureaucracy and investing in railway research, enabling technologies to be more readily and rapidly integrated into the railway system by creating the environment for increased research and developments investment, technology demonstration and removing barriers to the adoption of new technology.

Our work will focus mainly on points:

- Minimal disruption to train services
- More value from data

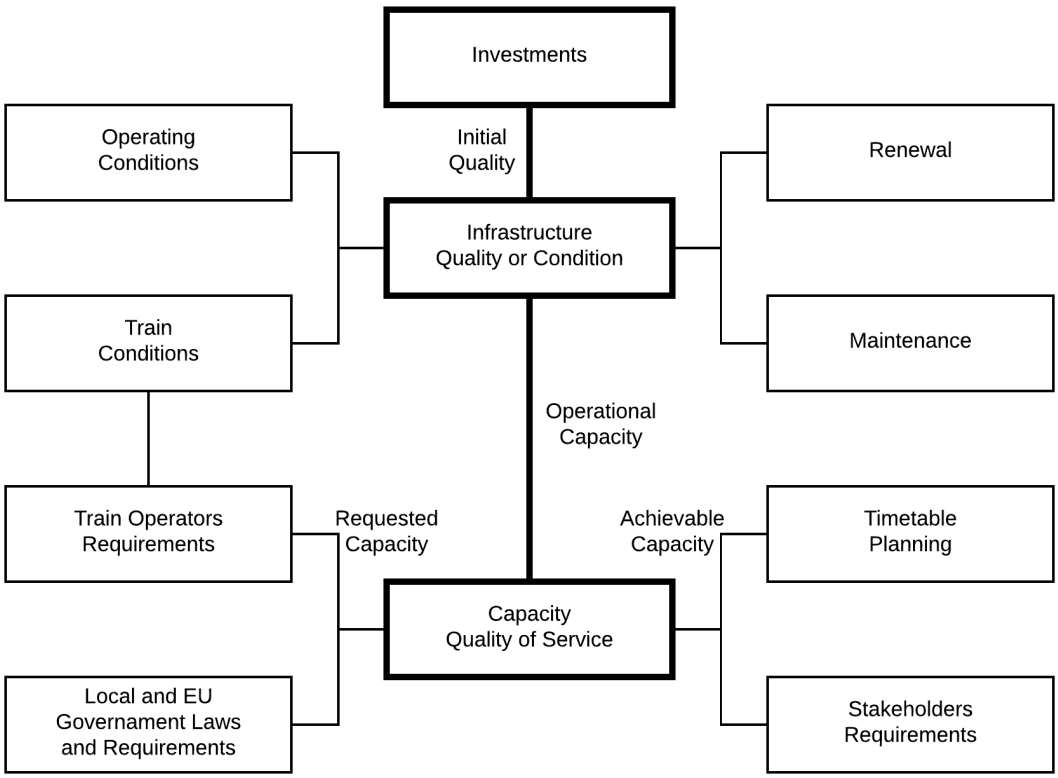

Figure 3: A holistic view of railway infrastructure capacity and its influencing factors

but obviously our work will have impact on the other ones. More in general the final purpose of our work is to increase the railway infrastructure operational capacity. The operational capacity of a given railway infrastructure depends on its technical state or quality and the way it is utilized. Capacity utilization is largely influenced by market requirements, traffic planning, regulations and other operational requirements. An important aspect in railway infrastructure maintenance is the dependence between operational capacity with associated quality of service and infrastructure condition. High operational capacity and expected quality of service is guaranteed when railway infrastructure is in a good state with high quality. Conversely, an increase in capacity or traffic loads leads to rapid quality deterioration of infrastructure and deformation of its components. This consequently leads to higher maintenance and renewal needs and more requests for track possession that eventually reduces the operational capacity. A holistic view of railway infrastructure capacity and its influencing factors is presented in Figure 3 with emphasis on infrastructure quality.

3.2 Connection with other In2Dreams WS and WP and Shift2rail Recipients

The considered railway ecosystem is described in Section 3.1 and depicted in Figure 2. This section is devoted to the description of the connections between the considered railway ecosystem and the IN2DREAMS WP5, other IN2DREAMS WS and WP, and other SHIFT2RAIL recipients. These connections are depicted in Figure 4. More in details, the synergy between IN2DREAMS WP5, other IN2DREAMS WS and WP, and other SHIFT2RAIL recipients is ensured by the following connection:

- the IN2DREAMS WP5 will be in charge of delivering the competences on data and visual analytics with particular reference to the problems of identifying the specific metrics and performance indicator to optimize using data and visual analytics models and techniques. Moreover the delivered data-driven models should be interpretable and their quality must be carefully assessed with the previously mentioned metrics and estimated with state-of-the-art statistical techniques. UNIGE and UKON, IN2DREAMS WS2 contributors will provide competences respectively in data and visual analytics.
- RFI, as IN2DREAMS WS2 contributor, Italian IM, and the only railway actor inside IN2DREAMS WS2, will be in charge to define and provide, together with the other IN2DREAMS partners, the railway specific metrics, the impacts on the railway ecosystem, and the actual historical data about assets and maintenances to be exploited for building and estimating the quality of data and visual models and their impacts on the day to day operations.
- a connection with IN2RAIL, SHIFT2RAIL recipient will be ensured by RFI and UNIGE, IN2RAIL WP9 contributors, that will bring results, lesson learned, the prospected future improvement mentioned in the IN2RAIL WP9 deliverable inside IN2DREAMS. In particular, the experience acquired in IN2RAIL about the impact of data-driven technologies on the next generation of TMS will be taken into account.
- the IN2DREAMS WP4 will be in charge of delivering his competences on distributed ledger technologies with particular reference to their scenario on distributed ledger technologies for data marketplace developed in the IN2DREAMS WP4 (see Deliverable 4.1). This scenario will be the the main link between IN2DREAMS WP4 and WP5 inside the IN2DREAMS WS2. In fact, in data marketplaces, it is necessary to exploit data-driven and visual analytics tools for predicting the value trends of the data exchanged between the different actors of the railway ecosystem. CEFRIEL, IN2DREAMS WS2 contributor, will provide competences on distributed ledger technologies. Moreover there is a strong connection between this scenario and the WP4 selected scenario on Asset Maintenance (D4.1 Section 5.1). In fact the operators need to be aware of the status of the maintenances and this information can be extracted from the blockchain developed in WP4 and vice-versa the blockchain can be fed with the output of the data driven models developed in WP5.
- an important link between IN2DREAMS WS1 and WS2 and other SHIFT2RAIL recipients (with particular reference to IN2SMART and IN2RAIL) will be the exploitation of the data that all the actors of the railway ecosystem are starting to store in their data lakes built based on the new generation big data platform developed and designed by the SHIFT2RAIL recipients in previous and current projects based on the specific requirements of the railway ecosystem. In fact, in the future, these data lakes will contain every data produced by physical assets, maintenance activities, information systems, and exogenous railway-related data sources.
- the connection with the IN2DREAMS WS1 is also ensured by a common scenario on energy consumption forecast developed in the IN2DREAMS WS1 WP6. WP5 will be in charge to assess and estimate the performance of the energy consumption forecast models based on the specific railway related metrics developed in the context of the IN2DREAMS WP5. EVOLUTION ENERGIE, IN2DREAMS WS2 contributor, will be in charge to link IN2DREAMS WS1 WP6 with IN2DREAMS WS2 WP5 by act as abridge between the two WP.
- a strong connection with the SHIFT2RAIL recipient IN2SMART will be ensured by the collaboration with ASTS and STRUKTON, IN2SMART WP8 contributors. ASTS and STRUKTON are suppliers and maintainers for many IMs in Europe. They will provide scenarios and data regarding assets of the railway ecosystem together with their experience in maintenance. ASTS and STRUKTON were also IN2RAIL WP9 contributors so they will help IN2DREAMS in further develop the scenario depicted and developed in IN2RAIL.
- finally KUL, IN2DREAMS WS2 contributor, will be in charge of delivering his competences with respect to the legal requirements and constraints of using all the above mentioned technologies, tools, and techniques in the railway ecosystem.

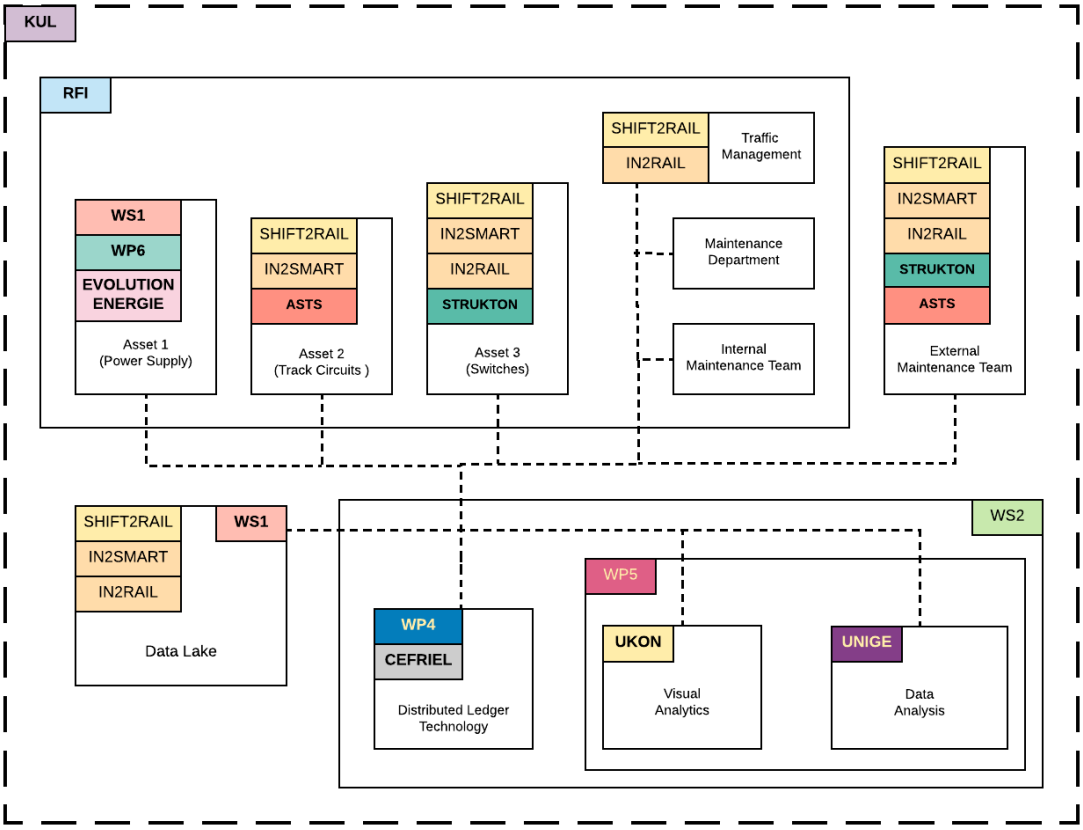

Figure 4: Connection with other In2Dreams WS and WP and Shift2rail Recipients

3.3 Potential Impact of Data Analytics & Metrics

In virtually any city, one is likely to see the results of analytics in the operation of trains that are essential for maintaining the mobility of the metro areas [197]. Furthermore, it is useful for both the public and the industry to realize how significantly public transportation has been a leading pioneer in the rich and extensive historic development of these tools. There are many studies about usage of data analytics in the area of transportation. It applied for Istanbul's automated fare collection system and pricing for BRT-Bus Rapid Transit line planning to obtain better recommendations for consumers with visualization metrics [76]. Using Markov-chain approach, a multi-modal transport network in London was developed with better information clusters for efficiency of transport [63]. City Intelligent Energy Network used statistical data including city economy, construction, population, and different energy parameters to develop the comprehensive model for low-carbon emissions [213].

Using data analytics, several key indicators and metrics for transport ware demonstrated to measure performance of cities [127] and for sustainability [49]. Data analytics has been used to develop architecture for traffic cloud data mining and optimization of strategies and related data processing and network optimization methods [67]. There are other case studies such as big data analytics for safety management [157], an assistant decision-supporting method for urban transportation planning using Global Positioning System data [219], crowdsourcing for intelligent transport system [221], crowd sourcing geo-social network [143], intelligent transport system for predicting drivers behavior [223], realtime monitoring from traffic for TomTom [54] and in Netherlands [207]. Several opportunities in public transport and processing techniques and

their challengers [39] were described in transport domain [36].
There are endless possibilities of data analytics in the transportation and one of the main areas is to look into maintenance aspects for maximum customer experience. It can identify behavior of bottlenecks, maximum loads, variation in traffic, unplanned delay timings, inspection timings and accidents that will impact customer's comfort, business's losses and asset's reputation. Data analytics is used in maintenance to improve the asset's performance, such as scheduling the maintenance windows when there is less traffic, do maintenance tasks at bottlenecks, rescheduling the assets with respect to the amount of traffic by passengers, and enhance overall efficiency that can lead to reduce the operating costs.
The EU specific policies stated that the railway infrastructure managers should focus more and more on reducing operational costs and at the same time increase performances in financial assets and safety[62]. The main component of the system is a commercial computerized maintenance management system that implements linear asset management that allows the dynamic segmentation of the line, permitting to specify attributes and features along the railway lines. This system also allows the definition of linear relationships indicating intersections, parallel or grade crossing of linear assets [47].
There have been studies on the application of data analytics in the area of Railways. It is stated that although data analytics is at nascent stage, for railways, both quality and quantity are going to increase further and several challenges of data analytics are discussed [13]. Innovation teams are discussing on future prospects in applying data analytics in railways[63]. Even the top business leaders of Hitachi were also brainstorming on effective utilization of data analytics in railways[64]. One of the applications of maintenance on railways was carried out by Dutch railways on axle box acceleration measurements with on terabyte of track degradation data for performing adaptive and self-learning mechanisms [149]. Data analytics was applied in Utrecht, Netherlands, to handle the traffic and explain the usage of mobile phones, smart cards and computers to predict the traffic and improve operations accordingly [207]. The maintenance of railways was pointed out on applications by using data analytics by Markov state classification [190]. The metaheuristics can be seen as sophisticated and intuitive methods which mimic natural phenomena and explores the solution within a feasible region in order to achieve specific goals in railway engineering [150]. The implementation of a railway asset monitoring system based upon semantic data models that offer greater capabilities for data integration, extensibility, and compatibility over traditional approaches was carried out for railway asset management [205]. The use of support vector machine technique that effectively utilizes large-scale data and provides valuable tools for operational sustainability was described for alarm prediction in railways [122].
Another system where data analytics is playing an important role is the traffic management system. For instance, in [137] a Fuzzy Petri Net model is developed to estimate train delays based both on expert knowledge and on historical data. In [22] a stochastic model for train delays propagation and forecasts based on directed acyclic graphs is presented. In [166] authors worked on data-driven models for train delays predictions, treating the problem as a time series forecast one. Their system was based on autoregressive integrated moving average and nearest neighbor models, although their work reports the application of their models over a limited set of data from a few trains. In [80, 87, 103, 104] authors developed an intensive research in the context of train delays prediction and propagation by using process mining techniques based on innovative timed event graphs, on historical traffic movement data, and on expert knowledge about railway infrastructure. Finally in [153] a dynamic data-driven train delay prediction system exploits the most recent tools and techniques in the field of time varying big data analysis.
The future of railways with data analytics and the Internet of Things will allow transportation modes to communicate with each other and with the wider environment, paving the way for truly integrated and intermodal transport solutions by Arup Report[65]. Rail travel has to become a safer, cheaper and more efficient

[62] https://www.railwaypro.com/wp/benefits-of-linear-asset-management-applications/
[63] https://www.railsupplygroup.org/wp-content/uploads/2017/09/RSG-Brochure-Jan-2016.pdf
[64] http://www.hitachi.com/IR-e/library/presentation/111207/111207.pdf
[65] http://www.driversofchange.com/projects/future-of-rail-2050/

mean of transport, as well as a source for revenue generation. This is why data analytics solutions have been designed to enhance business and travel experience, surpass the existing silos of systems and processes, drive innovation and build performance[66]. The prospects of using data analytics for railway management on handling large quantities of data was discussed [193]. For integrated maintenance analysis and perspective of innovation in railway sectors, there is a need of complex and centralized big data management to cope up with technological and engineering aspects [146]. Data analytics technologies for railway freight marketing decision by data acquisition, data preprocessing, storage and management using Hadoop was utilized [226]. The characteristics of increasing data volume related to equipment management of high-speed railway to summarize large datasets by using modern management methods to create a more complex situation of data management were demonstrated with emphasis on value and vital [184].
Other examples of using data analytics in the railway field are: condition based maintenance of railway assets [71, 123, 124], automatic visual inspection systems [14, 65], network capacity estimation [32], optimization for energy-efficient railway operations [15], marketing analysis for rail freight transportation [220], usage of ontologies and linked data in railways [145, 206], big data for rail inspection systems [125], complex event processing over train data streams [131], fault diagnosis of vehicle on-board equipment for high speed railways [148, 211, 228] and for conventional ones [28], research on storage and retrieval of large amounts of data for high-speed trains [210], development of an online geospatial safety risk model for railway networks [177], train marshalling optimization through genetic algorithms [168], research on new technologies for the railway ticketing systems [202].

3.4 Potential Impact of Visual Analytics & Metrics

Visual Analytics an approach to data analysis leverages the strengths of the combination of human perception and computational power (see Section 2.2). Analysis processes are designed in close cooperation with domain experts in order to create semi-automatic workflows leaving the operator in full control, while supporting her at the same time with context-sensitive, automatically extracted information. Applying VA approaches supports workflows of operators and analysts in multiple ways. Ultimately, improving the decision making process intrinsically leads to better informed decisions and thus, a reduction in delays managed by TMS operators.
Concerning errors and mistakes, Visual Analytics approaches even put special emphasis on dealing with uncertainties in the data and biases of the user as intrinsic part of the VA process [173]. The data sources relevant to the IN2DREAMS scenarios as well as data sources in railway management environments in general are inherently subject to a variety of uncertainties: Sensors can fail or produce inaccurate values, operators can override settings and parameters or forget to set them at all, and data models carry structural uncertainties [199]. In addition, expectations and other biases of operators and system users are likely to occur in complex management scenarios [100]. Consequently, IN2DREAMS scenarios can benefit from the application of VA principles (e.g. as demonstrated in [48] not only with respect to the specific tasks, but also regarding the trust an operator will put in the system [174].
While situation awareness issues are most important to consider in TMS environments, supporting prioritization and assessment are key tasks for asset management. Technicians have to make timely and sometimes costly decisions as reaction to developing events. In this respect, VA solutions are able to actively aid technicians by displaying relevant metrics. Further, metrics-based predictions can actually externalize the assessment of possible outcomes of a decision, thus helping a user to meet informed decisions and to avoid preoccupations. Current approaches [140, 141] already allow analysts to parameterize predictions for the outcome of unfolding events. Such predictions can even provide multiple results with respect to different

[66] https://www.altran.com/as-content/uploads/sites/7/2017/05/smartinsight_railway_connectivity.pdf

foci. For example, a user could be able to base his decisions on different scenarios provided by the system, such as most time-efficient outcome, most cost-efficient outcome, least schedule-disturbing outcome. Finally, within the IN2DREAMS ecosystem of end users, there are many inter-dependencies between different experts. For example, TMS operators rely on updates from asset managers to know when to expect maintenance operations or to get updates about current technical issues they have to deal with. Vice versa, asset managers coordinate with TMS operators to manually determine optimal maintenance windows. Many more inter-dependencies between users of different areas exist. With VA, shared information spaces can be introduced that automatically bring together such users. This way, knowledge and information transfer between them is increased and the need for manual coordination can be decreased, resulting in less communication overhead, better situational awareness and faster decisions. For example, a train breaks down during transit, and integrated VA processes go into action to present schedule predictions and propose schedule changes to a TMS operator. At the same time, asset management is automatically contacted about the issue, and the operators of each area can start coordinating any necessary action. In a shared view, everyone can update schedules and context information directly visible to others.

4 Scenarios

This section is devoted to the description of the selected scenarios. Seven scenarios will be presented. Two of them are cross-scenarios in the sense that they cover, in some way, many aspects of the railway ecosystem while five of them are specific-scenarios in the sense that they focus on a single particular aspect. For each scenario we will report:

- the responsible partner(s)
- the connections with other scenarios
- the connections with other IN2RAIL WS and WPs and SHIFT2RAIL projects
- the scenario objective(s)
- the scenario description
- the available data and data access policies
- the impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs
- the prospected analytics approaches and methods and metrics

The selected scenarios are the result of many discussions and interviews with RFI, ASTS, and STRUKTON and of a review of the state-of-the-art and current research and industrial trends in the field of data-driven analytics for railways.

4.1 Cross-Scenario 1: Visualizations in Control Center

4.1.1 Summary

Cross-Scenario 1 (CS1) provides visualization techniques for the control room. More specifically, this involves the Traffic Management System (TMS) and the Asset Management System (AMS). The goal is to improve the existing systems and enhance them with state-of-the-art visualization and visual analytics techniques. This shall enable the operators and managers to gain a better overview of the systems as well as understand proposed resolutions and automatic predictions better. This also involves past decisions and available uncertainties. As CS1 targets different systems we divide this scenario into three tasks:

1. Decision Support System for Rail-Conflict Resolution

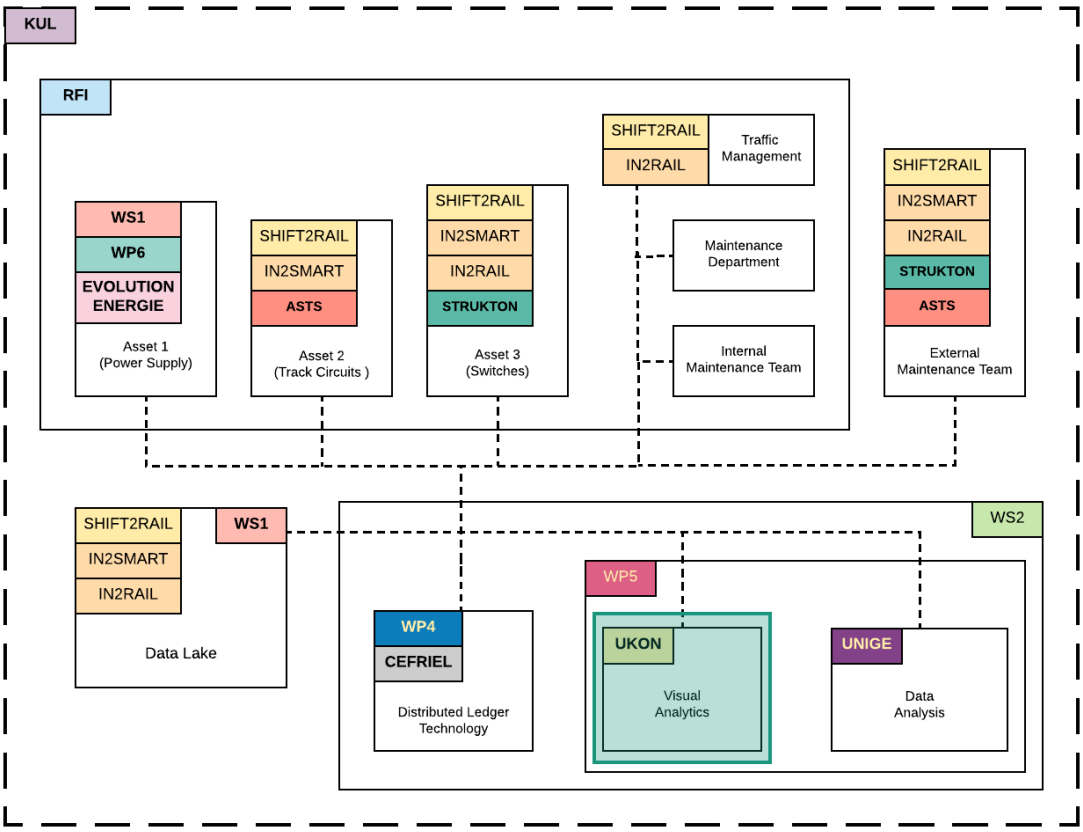

Figure 5: Cross-Scenario 1 Reference Picture (Visualizations in Control Center)

2. Alert Management and Prioritization System for AMS
3. Improving the TMS and Directing the Awareness of the Operator

4.1.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
UKON	IN2DREAMS	WS2	WP5	Responsible Partner	No
RFI	IN2DREAMS	WS2	WP5	Contributor	Yes
UNIGE	IN2DREAMS	WS2	WP5	Advisor	No

4.1.3 Connection to other Scenarios

CS1 is direclty connected to all other scenarios. In the following this will be elaborated for each of the other scenarios.

- Connections to Cross-Scenario 2 (Section 4.2)
  CS2 provides a data marketplace where it is beneficial to analyze and visualize trends regarding the

market. This information will help decision makers to gain better insights into the value of the data and future predictions including uncertainties.

- Specific-Scenario 1: Track Circuits (Section 4.3)
  This scenario is directly connected to Task 2 as the information can be directly included into the AMS. The data and predictions provide insights onto the condition of track circuits allowing predictive maintenance. The visualization also includes information about the uncertainty of the predictions.

- Specific-Scenario 2: Train Delays and Penalties (Section 4.4)
  This scenario is connected to Task 1 and 3. Visualizing the results and predictions of this Specific-Scenario 2 enables the operator of the TMS to make more informed decisions and inspect the impact of the decisions. The prediction also provides its quality in form of uncertainty as well as it reasons about the prediction itself. This enables the user to understand not only how the decision was made by the prediction but also how certain, or good, the prediction is.

- Specific-Scenario 3: Restoration Time (Section 4.5)
  Task 1 and 3 are connected to Specific-Scenario 3 as this scenario provides predictions of the restoration time of an asset. This means that the operator can already plan ahead by using this information and dispatch trains more efficiently. The prediction provides also reasoning as well as quality measures. All of this has to be visualized for the operator in order to make an informed decision.

- Specific-Scenario 4: Switches (Section 4.6)
  Specific-Scenario 4 provides more information of the condition of switches for the asset managers (Task 2). This helps the users to better understand not only the general condition of the switch but also, if faults are present the probable cause. Furthermore the prediction provides reasoning and information about the quality of the prediction.

- Specific-Scenario 5: Train Energy Consumption (Section 4.7)
  This scenario provides data on the power supply of trains. This data can be visualized for operators (Task 3 as well as asset managers (Task 2). The prediction model provides data about the energy consumption. The prediction includes external information such as weather or traffic conditions. The predicted energy consumption can be used to operate trains in a more energy-efficient way.

4.1.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

CS1 is connected to other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects. In the following, we summarize this as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection of CS1 to WP4 is two-fold: first there is a direct connection to visualize properties and metrics of the blockchain helping the maintainers to keep the marketplace efficient and spot problems. Secondly, the analysis of CS2 can be used to identify trends which essentially helps the data producers to estimate the value of the data better and tailor the monetization. IN2DREAMS WP4 focuses its attention on the use of blockchain as software connector in order to empower companies to profit from their data while allowing sustainable data monetization and enhancing trust between actors. They will examine several possible software architectures for the data marketplace, depending on the specific blockchain and smart contracts frameworks adopted. This scenario, in combination with CS2, visualizes the analysis results and other information of the marketplace which essentially enables the actors and maintainers of the marketplace to tailor the data that is being offered, spot trends and anomalies.

Moreover there is a strong connection between this scenario and the WP4 selected scenario on Asset Maintenance (D4.1 Section 5.1). In fact the operators need to be aware of the status of the maintenances and this information can be extracted from the blockchain developed in WP4 and vice-versa the blockchain can be fed with the output of the data driven models developed in WP5.

- Connections with IN2DREAMS WS1, IN2SMART, and other SHIFT2RAIL projects
  Many SHIFT2RAIL projects (e.g., IN2RAIL, IN2DREAMS, IN2STEMPO, and IN2SMART) collect or produce data. The question arises of how to make these data valuable in order to improve asset status nowcasting and forecasting, how to improve the maintenance efficiency, and how to improve the railway capacity. Visual Analytics aims to visualize this data as well as the models and their results in order to give the human a better access to the models and interact with them. This essentially improves the models as the user is able to influence the models with domain specific knowledge that is not inherently available in the data. The access to the models that is provided through visualization ultimately increases the trust in the whole system as the users better understand the models inner workings. Through uncertainty visualization, the users even receive information of how reliable a certain prediction might be which helps them in their decision making.

4.1.5 Scenario Objective(s)

As earlier reported, CS1 is separated into three tasks for a better structure and overview. However, the general objective across all three tasks is to improve the existing visualizations as well as embedding additional information from prediction models allowing the respective users to make more informed decisions and to plan ahead.

Task 1: Decision Support System for Rail-Conflict Resolution The objective for this task is to provide a decision support system that enables the user to make a better informed decision regarding rail-conflicts. This shall include information from predictions (Specific-Scenario 2, 3, 5) as well as past decisions made by the operator.

Task 2: Alert Management and Prioritization System for AMS The objective is to improve the existing alert management system leveraging the existing semantic relations and hierarchical structures of the systems that may generate alarms. In combination with prioritization of systems and types of alarms the system can steer the awareness of the user more efficiently and point him/her to the more imminent failing systems.

Task 3: Improving the TMS and Directing the Awareness of the Operator The current overview screens provide several information simultaneously but, at the same time, produce clutter and are overplotted. The objective is to simplify this visualization system providing context aware information as well as steering the users awareness to the crucial, imminent tasks.

4.1.6 Scenario Description

Task 1: Decision Support System for Rail-Conflict Resolution A rail-conflict can occur when, for example, two trains want to use the same rail-segment at the same time. The basic options that can be taken are to first let train A or train B pass. If the TMS cannot automatically resolve the conflict it will point the user to it. Such an example is shown in Figure 6 where on the center-right part of the screen a yellow target-icon is visible. The visualization shows trains using segments (vertical) over time (left to right). Predictions for the future are made using a rule-based system and the implemented schedule of the trains. When the operator clicks on the icon, another window opens showing more information about the conflict including

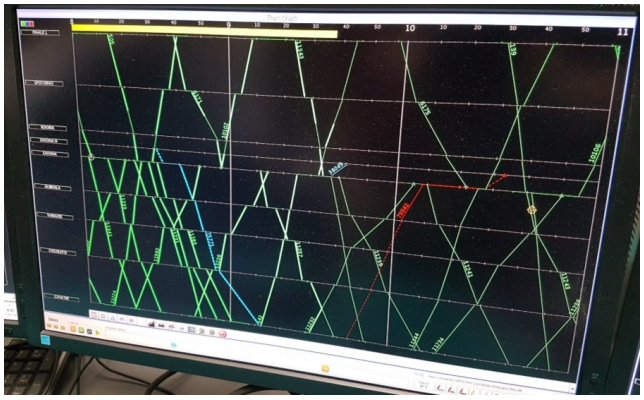

Figure 6: A visualization used by the train operators to display the schedule of trains of past, present and future. A rail conflict is visible at the center-right part of the screen.

the affected trains, the location as well as the optimal resolution based on the rule-based TMS. The operators, however, include more information affecting their decisions. Examples include the type of train, the time of the day and the location as well as other external factors. The goal of this task is to include and visualize that information to provide better and more objective metrics for the operators. This includes a prediction model from Specific-Scenario 2 (Section 4.4) as well as predictions on the energy consumption (Specific-Scenario 5, Section 4.7). Furthermore, as conflicts are often repetitive, information will be displayed on how the operators have decided in the past for the same conflict in similar conditions. All of the previous measures enable the user to make a more informed decision by including multiple metrics. The decision is also less dependent on the operators experience.

Task 2: Alert Management and Prioritization System for AMS The current system does not take any prioritization of alarms into account and requires the asset manager to acknowledge every single alarm. As roughly 15000 alarms occur at one day the asset managers don't bother acknowledging every single alarm as most of them are irrelevant or have a tremendously low importance factor or are even already known to the asset manager. All of this results in a cluttered and highly distracting screen which imposes a high risk of missing an important alarm from a failing system. The systems that are being constantly surveyed have a strict hierarchical structure as well as semantic connections. Furthermore, the systems can be categorized whereas every category contains another hierarchy. For an optimized system the operators must be able to prioritize various categories resulting in prioritized alarms. Additionally, through the modelled semantics the system can automatically detect related alarms and visualize this connection resulting in an aggregated alarm and further providing reasoning onto the root-cause of the alarm which is effectively the failed system. By visualizing this information the asset manager retains the overview of the system continuously as well as is being assisted by additional relevant information enabling her/him to make faster and more informed decisions. The current visualization provides a topological view. This task will also improve the topological view by leveraging the hierarchies of the systems as well as the prioritization. The system will be also improved by using state-of-the-art visualization techniques that guarantee a less distracting environment.

Task 3: Improving the TMS and Directing the Awareness of the Operator The current overview screen of the TMS (Figure 7) shows a topology of the rail network including occupied segments, the position of the trains, as well as other additional information. The overview screen is composed of multiple monitors and, thus, the area is very large making it hard for the user to maintain an overview. One goal is to compress the information visualized on the screen to effectively decrease the screen area. Additional measures include the

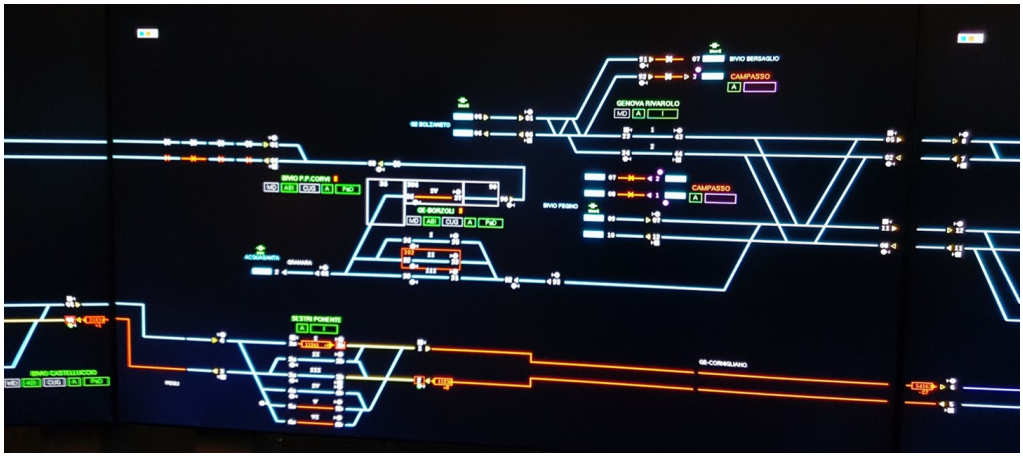

Figure 7: The overview screen of the TMS showing a topology of the rail network.

agnostic screen that accommodates to different contexts and prioritizes the tasks for the user. This does also include the guidance of the users' awareness to direct her/him onto the important displayed information. All of the measures will aim to reduce the clutter of the screen. Predictions, and other information provided by various other scenarios (e.g., Specific-Scenario 2,3,5) should be included into this screen as it provides a basis for the operator. However, it is not necessary to always display this information continuously but only provide it on request or when suggested by the system. The system should adopt to various situations and tasks of the user.

4.1.7 Available Data & Data Access Policies

The current available data is provided by RFI and can be used for Task 1 and 2.

Task 1: Decision Support System for Rail-Conflict Resolution The data for this task contains information on how conflicts in the past were resolved, either automatically or by the operator.

Data Fields	Description
Date	Date of the conflict, the resolution, or change
Time	Time of the conflict, the resolution, or change
Station	The station which is essentially the location of the conflict
Train 1	The identifier of the first train involved
Solution	The solution that was chosen
User	The user that made the resolution; can also be the system when it was automatic
Solution	Further details on the chosen solution for the conflict
Train 2	The identifier of the second train involved
Solution Detail	Detailed information about the solution

Task 2: Alert Management and Prioritization System for AMS The data for this task is a log file of alarms of one day.

Data Fields	Description
Date_Time	Date and time of the alarm
Station	The station where the alarm occurred
System	The system where the alarm occurred
Equipment	The equipment where the alarm occurred
Equipment Detail or Event	Either features detail of the equipment or the event
Status	The status showing the beginning, ending, cancellation and acknowledgement

Data from RFI will be provided anonymously, meaning that no specific asset name will be provided regarding the operational service. All the references to the real data, as object/asset names, will be anonymized in order to avoid privacy and security issues.

4.1.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

Cross-Scenario 1 impacts the SHIFT2RAIL initiative and the IN2DREAMS WS2 WP5 KPIs in multiple ways. The visualization techniques and visual analytics, especially, aim to provide interfaces for the human to data and models. The user can interact with these visualizations to explore the data space and the models to gain a better understanding. This generates more knowledge eventually. The so gained knowledge plus the already available domain and real-world knowledge that is not reflected in the data can therefore be propagated into the models to steer the models and improve them further. Therefore, the domain knowledge by asset managers or train operators might improve the prediction models of the Specific-Scenarios and thereby impacts the stated projects and work streams. In general, the visualization and the improvements of the current system will help the users to maintain overview and increase the trust into the systems as the prediction models already report about their quality of prediction and this is propagated to the user through visualization.

Specifically we expect impact on IP3: cost-efficient and reliable high-capacity infrastructure. This is due to the fact that the improved visualizations in the systems scale better with respect to a growing number of assets. This helps the users to maintain the overview at all times (Task 2) and better assign resources for maintaining the assets. The additionally collected metrics for the assets are analyzed to give the user a better impression of the status of the asset. Predictive maintenance uses algorithms and models to forecast when and how likely an asset will fail in the future. All this information will increase the ability of the asset managers to better plan maintenance.

Regarding the IN2DREAMS WS2 WP5 KPIs we state that CS1 can impact the following scope of the project:

- Impact on specific visualization and VA techniques on the railway ecosystem
  The data and models of the railway ecosystem provide many interesting attributes such as the variability, spatio-temporal data, network-data, as well as uncertainty information and other multidimensional metrics. Visualizing all of this information is challenging but at the same time provides great impact on the efficiency and effectiveness of its users. Interpretable data and visual data-driven models can be better understand to learn what is the real value of the data and what could be the scenario if all the parts of the railway ecosystem would be datified.
- Visual exploration is not only useful for the data space but also for the model- and metric-space. This helps analysts to generate knowledge in finding useful models and metrics for relevant problems. This is not only valuable for the operators of the TMS but also for asset managers as well as for the data marketplace (data-lake). Interacting with the models through visual interactive interfaces allows the user to steer the models and adapt metrics with domain knowledge. This is especially useful in edge cases where only little training-data exists or when knowledge is not datified.

4.1.9 Analytics & Metrics

Exploring spatio-temporal data as provided in the IN2DREAMS context is a well-researched area where many techniques have been developed to solve common tasks. Yet, almost always, specific application scenarios also comprise individual analysis challenges for which novel, customized VA solutions have to be developed. Improving the quality of visual displays both for TCS operators and maintenance technicians requires considering the use-case-intrinsic conditions and requirements. State-of-the-art systems from similar domains such as air traffic [35, 94] and sports analytics [165, 191] illustrate, how common techniques are combined with use-case specific solutions to provide users with both intuitive and powerful, visual-interactive analysis interfaces. Further, examples such as [35] illustrate how VA approaches help increasing the understandability of complex models. Through visual analytics, the user can even feedback knowledge into the models and steer these, for example by observing prediction results in real-time when modifying parameters. This increases the user's awareness and understanding of how the models work and how and why a specific prediction was made by a model. Additionally, the quality of the model can be visualized as uncertainty which ultimately increases a users' trust into the model and the system in general.
Especially concerning a TCS operator's view, visualizations for sequence analysis techniques will play a big role in this scenario, as they have to provide complex train schedule information to the operators while highlighting potential conflicts. Ideally, contextual information is also provided to alleviate the resolution of the conflict or to provide information critical for decision making at the right place, relieving an operator from having to do costly mental context switches between separate systems collecting the necessary information. Train schedules can be seen as sequences running in parallel. Visual Analytics approaches can help in schedule optimization and conflict resolution by automatically precomputing solutions to emerging problems. Different optimization strategies can be regarded, such as cost-efficiency, punctuality or affected passengers. Sequential pattern mining [5] can be applied to learn from recorded data and previous solutions to other problems. Sequential pattern mining in the VA context has already been applied to different domains such as crime analytics [98] or financial data analysis [214].
Finally, identifying and applying suitable metrics and analysis algorithms for supporting the goals of Cross-Scenario 1 is a sensitive task where data sources, types and regulations have to be considered. For example, quality metrics in the visualization domain can be applied to improve data views and interactive VA processes [27, 33]. The advantage of such data-agnostic methods is their universal applicability to heterogeneous data, which can be leveraged in the Cross-Scenario 1 data context.

4.2 Cross-Scenario 2: Marketplace of Data and Data Monetization

4.2.1 Summary

This scenario deals with the problem of helping the actors of the railway ecosystem in having an interpretable and reliable support decision system which can help them in automatically exchanging, evaluating, and monetizing, on the data marketplace developed in the IN2DREAMS WS2 WP4, the information collected and stored in their data-lakes under the legal requirements of the current and prospected EU state members regulations. Data monetization is the act of generating measurable economic benefits from available data sources. Typically these benefits accrue as revenue or expense savings. Data monetization leverages data generated through business operations, available exogenous data or content, as well as data associated with individual asset (i.e. data collected with sensors). In Figure 8, we report a simplified representation of the Cross-Scenario 2 in relation with the railway ecosystem.

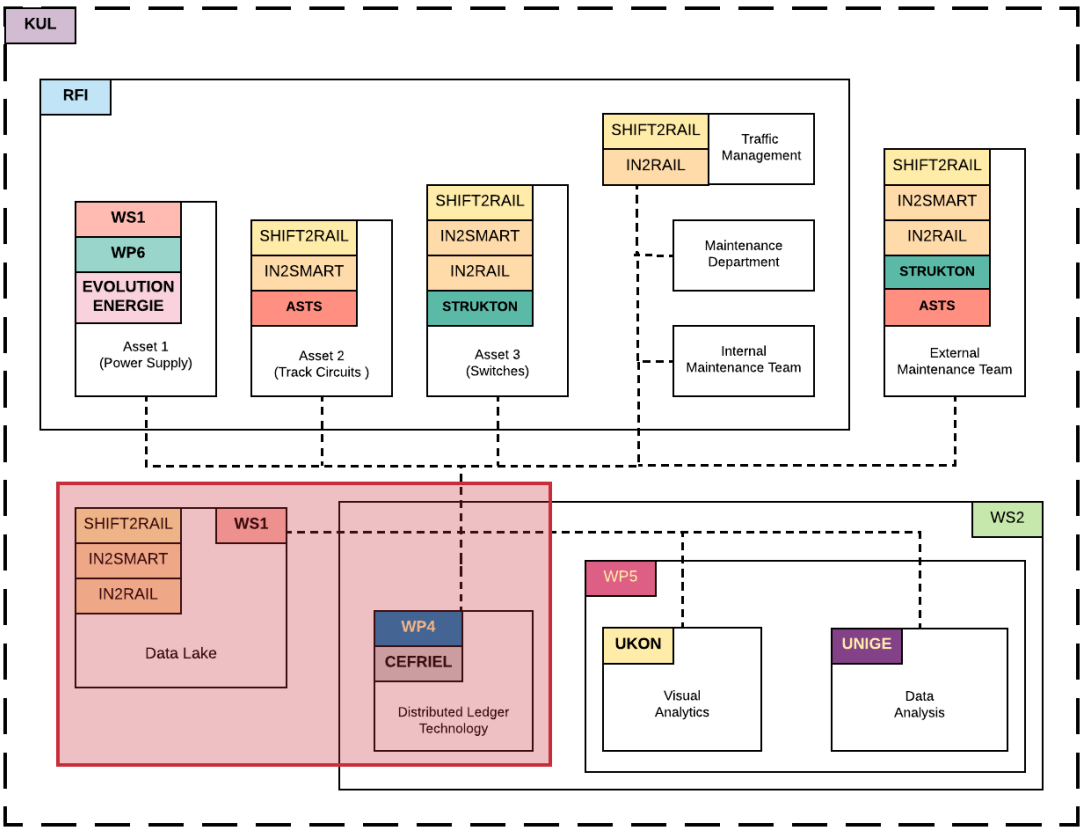

Figure 8: Cross-Scenario 2 Reference Picture (Marketplace of Data and Data Monetization)

4.2.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
UNIGE	IN2DREAMS	WS2	WP5	Responsible Partner	No
CEFRIEL	IN2DREAMS	WS2	WP4 & WP5	Contributor	Potentially
UKON	IN2DREAMS	WS2	WP5	Contributor	No
RFI	IN2DREAMS	WS2	WP5	Contributor	Yes
ASTS	IN2SMART	-	WP8	Advisor	Potentially
STRUKTON	IN2SMART	-	WP8	Advisor	Potentially

4.2.3 Connections with other Scenarios

Cross-Scenario 2, as its name states, is a cross-scenario and for this reason it is connected with all the others more or less directly.

- Connections with Cross-Scenario 1 (Section 4.1)
  The connection with this scenario is rather straightforward in the sense that, once a data marketplace envisioned in the IN2DREAMS WS2 WP4 is available, it is necessary to visualize and analyze the trends

of this market and display this information to the people involved in the strategic decisions (e.g. put some source on the marketplace or start to collect/buy a particular data source). Basically one has to provide all the visualization tools and techniques that are already available in the financial market or in the market of the physical objects adapted to the peculiarities of the railway ecosystem.

- Connections with Specific-Scenario 1, 2, 3, 4, 5 (Sections 4.3, 4.4, 4.5, 4.6, and 4.7)
  Also in this case the connections of Cross-Scenario 2 with all the specific-scenarios is quite intuitive. From one side all these scenarios are characterized by the availability of data about a particular subsystem of the railway ecosystem and these data can be made available, exchanged, sold, and monetized on this marketplace. From the other side the models developed and fed by these data and their prediction can be in turn sold on the very same marketplace.

4.2.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

The connection of Cross-Scenario 2 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection in this case is quite tight since Cross-Scenario 2 is a common scenario between IN2DREAMS WP4 and WP5 inside the WS2. IN2DREAMS WP4 focuses its attention on the use of blockchain as software connector in order to empower companies to profit from their data while allowing sustainable data monetization and enhancing trust between actors. They will examine several possible software architectures for the data marketplace, depending on the specific blockchain and smart contracts frameworks adopted. IN2DREAMS WP5 instead, as described in this deliverable, focuses its attention on the analytics that can further empower the different actors of the marketplace in exploiting the data produced and exchanged on the marketplace.
- Connections with IN2DREAMS WS1, IN2SMART, and other SHIFT2RAIL projects
  A common characteristic of many SHIFT2RAIL projects (e.g. IN2RAIL, IN2DREAMS, IN2STEMPO, and IN2SMART) is that the actors of the railway ecosystem exploit the new generation of big data storage architectures, also called data-lakes, which are designed to collect big, heterogeneous, and continuous flowing sources of information from the railway ecosystem and from the related activities. Once these data have been collected, the question arises of how to make these data valuable in order to improve asset status nowcasting and forecasting, how to improve the maintenance efficiency, and how to improve the railway capacity. Nevertheless, another fundamental question is how to make profit from these data by exchanging these data with other actors. Consequently, two other important questions arise from these new requirements:
  - What are the technical, technological, and legal tools and techniques available to create this marketplace? The answer to this question, raised mainly by IN2SMART WP7 and IN2DREAMS WS1, will be provided in IN2DREAMS WP4.
  - How is it possible to estimate the value of these data and how is it possible to understand if, when, and how it is better to share my data or keep them secret? IN2DREAMS WP5 tries to find an answer to these questions, raised mainly by the IN2SMART WP8 in particular by ASTS and STRUKTON, with visual and data analytics tools and metrics.

4.2.5 Scenario Objective(s)

The main objectives of this scenario are the following ones.

- Envisioning a future in which the data produced inside the railway ecosystem and owned by the different actors can be exchanged under the legal requirements that the actual and prospected regulations

impose and with the adequate support in terms of tools and technologies for monitoring and estimating the value of the different data sources.
- Study the state-of-the-art data-driven technologies and techniques able to support these decisions and potentially also automatize them by plugging machine learning models inside the smart contracts itself.

4.2.6 Scenario Description

The Data Marketplace for Monetization and Servitization Use Case of IN2DREAMS WS2 WP4 refers to the data sources ecosystem for asset management which have been and will be defined by many SHIFT2RAIL projects (e.g. IN2RAIL, IN2SMART, IN2DREAMS, and IN2STEMPO). As output of the IN2DREAMS WS2 WP4, a system for the exchange of data sources will be designed. This system may play a crucial role in the EU Railways Ecosystem, especially considering the Open Market scenario envisioned by the EU. In this use case, actors will be IMs, suppliers, or any other entity that is involved in data creation and data utilization in this field. The system will have to consider the legal aspects of exchanging data sources, since there may be relevant issues related to the ”ownership of data”. The system will collect all the data sources produced and used in the ecosystem by all actors and would have the main goal to allow their business exploitation. The railway ecosystem is composed by a heterogeneous group of actors that produce data during their business activity or consume data as an input for their products and services. The main challenges are: the lack of trust between the users in the network, the quality of the data and their analytics and the lack for a sustainable monetization model. Therefore, IN2DREAMS WS2 WP4 envisions a digital marketplace where

- The actors involved in the process can manage and control their data without the need of intermediary third party or centralized repository.
- The exchange of the data is regulated by adopting open standards and could be improved by adopting smart contracts that will enable data monetization of the data exchange. Thus, we are enabling the monetization for the exchange of data in digital ecosystem and enabling the automation of governance logics in the digital ecosystem.
- All parties involved in the exchange have access to the same data, this will lead to acceleration of data acquisition/sharing, and improving the quality of data and data analytics.
- The adoption of smart contracts could tackle the problem of managing the marketplace dynamically. Data will be decentralized and will have dynamic value based on the usage or other predefined characteristics. Unlike current data exchange markets, this solution requires no trusted third-parties.
- Data producers and customers can cooperate together to build up a network of data-based value transfer.

Once this digital marketplace will be available it will produce by itself data regarding how, how much, and at what value the data have been exchanged in this marketplace. This is the link where the IN2DREAMS WS2 WP4 is connected with the IN2DREAMS WS2 WP5. In the IN2DREAMS WS2 WP5 we will use this data for understanding the behavior of the market for the final purpose of automatically estimating the value of a particular data source and its possible future value trends in order to automatize the data exchange and evaluation inside the smart contracts. This specific case follows the more general problem of Data monetization. Data monetization is the act of generating measurable economic benefits from available data sources. Typically these benefits accrue as revenue or expense savings. Data monetization leverages data generated through business operations, available exogenous data or content, as well as data associated with individual asset such as that collected with sensors. A fundamental scope of the analysis of these data is to show to the operators the potential benefits of the datification which is the modern technological trend turning many aspects of the industries into computerized data and transforming this information into new forms of information of value. The operators, in this case, can belong both to the management but also to the labor: having a more direct way of understanding the value of collecting the data can change both the mentality

of the managers but also the awareness of the operators about the value of the data that they produce by filling forms and online reports. Having a clear picture of the data generated, exchanged, exploited, and their value can improve and enforce a datification of the railway ecosystem which is still a quite non-automated ecosystem. Datification is not yet an hot topic in the railway ecosystem principally because of the difficulties of understanding the value of these data: a marketplace would allow to better understand the value of the data also from the economic perspective and would enforce a datification not just for research or industrial purposes but also for its intrinsic economic value. Moreover, if from one side all these data about a particular subsystem of the railway ecosystem are available, exchanged, sold, and monetized on the other side this marketplace means that people are building analytics from them and the models developed and fed by these data and their prediction can be in turn sold on the very same marketplace. This would create a sort of parallel new business of data.

4.2.7 Available Data & Data Access Policies

At the current stage we do not foreseen the availability of any real data since the data marketplace do not exist, hence we will keep the scenario in the TRL 1 and TRL 2. We will keep working on the scenario in order to understand if it is possible to retrieve some data in order to be able to realize some POC or advance the scenario to TRL 3, TRL 4, and TRL 5.

4.2.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Cross-Scenario 2 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  For sure this scenario has a great impact on the IP3, in fact the ability to exchange data between the actors will improve the ability of the single actors to refine their process with the purpose to improve the capacity. To make an example, the capacity of an IM to obtain data from the TOs can dramatically improve his ability to understand the performance, the age, the weight, and the number of passengers of the trains that are running over the network. Another examples are the external maintenance that would have the ability to access the detailed information from the IM and the TO about the stress of particular components (e.g. switches and level crossing) and act preventively before faults happen during low railway traffic conditions in order to impact less on the circulation. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs we can state that the Cross-Scenario 2 can have an impact into the following scope of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario surely goes into the direction of better spread the datification mentality which is not yet well understood by the railway ecosystem principally because of the difficulties of understanding the value of these data. A marketplace would allow to better understand the value of the data also from the economic perspective and would enforce a datification not just for research or industrial purposes but also for its intrinsic economic value. Interpretable data and visual data-driven models can better understand what is the real value of the data and what could be the scenario if all the parts of the railway ecosystem would be datified.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  As any financial or physical market, finding metrics is a key problem and our market scenario is not an exception. Finding metrics able to understand when it is better to share or not or when it is better to buy or not a particular source of data without risking to spread key information without the adequate

remunerations is a key problem in any industry and especially in the railway one, which is always more liberalized. These metrics developed in this scenario would help to better control and understand this evolution.

4.2.9 Analytics & Metrics

In order to have an idea of the analytics and metrics that we can exploit to reach the goals of the Cross-Scenario 2 we can review the state-of-the-art on the subjects related to it.
Data analytics techniques have been widely used in the field of financial time series predictions [102, 116, 167]. Analizing market trends is a challenging task due to its high volatility and noisy environment. Many factors influence the performance of a market including political events, general economic conditions, and actors' expectations. Although stocks and futures traders have relied heavily upon various types of intelligent systems to make trading decisions, the performance has been a disappointment [2]. Many attempts have been made to predict the markets trends, ranging from traditional time series approaches to artificial intelligence techniques, such as fuzzy systems and artificial neural network methodologies [1]. However, the main drawback with black-box techniques is the tremendous difficulty in interpreting the results. They do not provide an insight into the nature of the interactions between the technical indicators and the market trends fluctuations. Thus, there is a need to develop methodologies that provide an increased understanding of market processes even with specific designed metrics [41, 56, 203].
Contemporarily, also visual analytics covers business and market applications [106]. The markets with its thousands of different stocks, bonds, futures, commodities, market indices and currencies generates a lot of data every second, which accumulates to high data volumes throughout the years. The main challenge in this area lies in analyzing the data under multiple perspectives and assumptions to understand historical and current situations, and then monitor the market to forecast trends and to identify recurring situations. Visual analytics applications can help analysts obtaining insights and understandings into previous stock market development, as well as supporting the decision making progress by monitoring the stock market in real-time in order to take necessary actions for a competitive advantage, with powerful means that reach far beyond the numeric technical chart analysis indicators or traditional line charts. One popular application in this field is the well-known Smartmoney [215], which gives an instant visual overview of the development of the stock market in particular sectors for a user-definable time frame. A new application in this field is the FinDEx system [107], which allows a visual comparison of the performance of a fund to the whole market for all possible time intervals at one glance.

4.3 Specific-Scenario 1: Track Circuits

4.3.1 Summary

This scenario deals with the problem of building an interpretable and reliable data-driven condition based maintenance system for the track circuit, a simple electrical device used to detect the absence of a train on rail tracks, used to inform signallers and control relevant signals. In Figure 9, we report a simplified representation of the Specific-Scenario 1 in relation to the railway ecosystem.

4.3.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

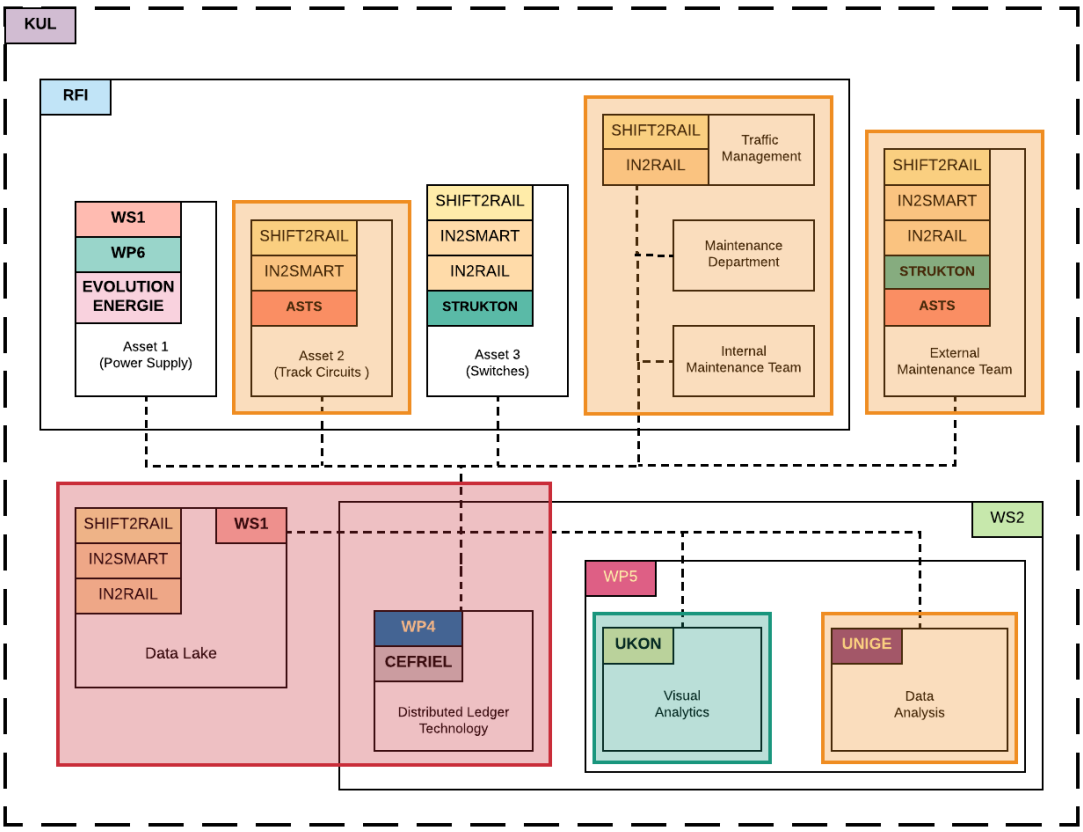

Figure 9: Specific-Scenario 1 Reference Picture (Track Circuits)

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
UNIGE	IN2DREAMS	WS2	WP5	Responsible Partner	No
ASTS	IN2SMART	-	WP8	Advisor	Yes

4.3.3 Connection with other Scenarios

Specific-Scenario 1, as its name states, is a specific-scenario and for this reason it is mainly connected with the cross-scenarios.

- Connections with Cross-Scenario 1 (Section 4.1)
  Specific-Scenario 1 will provide to Cross-Scenario 1 other useful information which will be exploited from the operators (mainly from maintainers) by means of visualization. In particular, the possibility to understand track circuits conditions and consequently the possibility to maintain them before breaks and recognize false occupations is fundamental for limiting disruptions. Moreover, together with prediction results, Cross-Scenario 1 has to show also the quality of the prediction and the reason of the prediction itself. Starting with results from interpretable models developed in Specific-Scenario 1, this aim will be pursued developing a nice and intuitive graphical representation that can help the operator in deciding through his personal experience weather or not to rely on the prediction.
- Connections with Cross-Scenario 2 (Section 4.2)
  The connection of Cross-Scenario 2 is rather straightforward, in fact the models developed with the

data provided by ASTS and the associated predictions can be sold on the data marketplace as the data itself. Moreover, the data marketplace can be a source of additional data that can be exploited for improving the quality of the model. The results of Specific-Scenario 1 can give a hint on what could be a new source of information that could help improving the quality of the models.

- Connections with Specific-Scenarios 2, 3, 4, and 5 (Sections 4.4, 4.5, 4.6, and 4.7)
  The connection of Specific-Scenario 1 with Specific-Scenarios 2, 3, 4, and 5 is quite tight. Specific-Scenarios 4 and 5 deal with the problem of predicting a possible malfunction. Specific-Scenario 3 deals with the problem of predicting the restoration time from a maintenance or a malfunctions. Specific-Scenarios 2 deals with the train delay prediction. Then, the scope of the different scenarios is complementary. We envision a future where predictive models of a malfunction will be exploited together with the one which forecasts the restoration time of an asset and the train delays in a way to optimize the train circulation for improving the cost efficiency, the reliability, and the capacity of the infrastructure.

4.3.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

The connection of Specific-Scenario 1 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection here is well depicted in Figure 9, in fact the data that we will use in this scenario are stored in the data lakes of ASTS. Then, these data can be shared in the marketplace envisioned in the IN2DREAMS WS2 WP4 with a twofold objective: on one side, the monetization of this information by selling it to other actors of the railway ecosystem, on the other, the enrichment of the available data by buying some other detailed information from other actors (e.g. exact usage of the switch with data of the TO about tons) in order to improve the quality of the prediction. Moreover, the models and the predictions themselves can be sold to other actors as a service.
  There is also a strong connection between this scenario and the WP4 selected scenario on Asset Maintenance (D4.1 Section 5.1). In fact the operators need to be aware of the status of the maintenances and this information can be extracted from the blockchain developed in WP4 and vice-versa the blockchain can be fed with the output of the data driven models developed in WP5.
- Connections with IN2SMART and other SHIFT2RAIL projects
  This use case has been designed inside WP8 Dynamic Railway Information Management System Data Mining and Predictive Analytics in the framework of the IN2SMART Project. Like all the other WP8 case-studies, this one concerns relevant assets, whose malfunction and maintenance policies have an impact on the KPIs targeted by the SHIFT2RAIL program, and not included into the dynamic modelling activities of IN2RAIL. The objective of WP8 is to identify, design and study the most suitable Data mining and Big Data analytics approaches, for extracting information and knowledge from data in the context of railway systems. More in details, this scenario encompasses three different activities, outlined inside different WP8 tasks: the first activity is pursued inside Task 8.1 (”Automatic Detection of Anomalies”), the second one inside Task 8.2 (”Process Mining”), and the last one inside Task 8.3 (”Predictive Models of Decaying Infrastructures”). From these activities, the main connection with IN2DREAMS WP5 is related to the anomaly detection activity. It is also worth to note that in the framework of IN2SMART, results coming from this scenario inside IN2DREAMS WP5 (as well from IN2SMART Task 8.1 and Task 8.3) will be exploited in the IN2SMART Task 9.2 (”Design of a generic framework for decision support in maintenance and interventions planning”) of WP9 and the IN2SMART Task 9.4 (”Business cases, prototype development and in-lab demonstration”) where there will be a prototype on ”Track circuits: false track occupancies mitigation” that can take advantage of the results of the IN2SMART WP8 and

IN2DREAMS WS2 WP5 by developing a decision-support system able to plan the Track circuits maintenance and calibration.

4.3.5 Scenario Objective(s)

The main objectives of this scenario are the following ones.

- Develop interpretable and gray-box models of the track circuits behaviour (using historical data collected by ASTS) in order to predict disruption or false occupations.
- Develop metrics and KPIs able to discern between situations in which our models can be effectively be applied in real operations and situations in which we have to leave the choice to the operators because the developed models are not reliable enough (e.g. due to the quality or the quantity of the historical data).
- The same KPIs developed for testing the model developed in IN2DREAMS WP5 will be exploited to test the quality of the black-box models developed in IN2SMART WP8.

4.3.6 Scenario Description

System Description The Track Circuit System (TCS) is a wayside component of the central automatic control system and it is used to provide both train detection and transmission of digital cab signalling data to enable automatic train protection functions (Figure 10). TCS are composed by an electronic board (TCS board) and a physical equipment (i.e. electrical circuit, cable, relays). Track circuits boards communicate with the Computer-Based Interlocking system (CBI), using a specific protocol, in order to enable trains management functions (train detection and cab signalling). Information from the CBI (e.g. train speed/direction, next carrier transmit frequency, track circuit ID etc.) is then used from the central system, which elaborate information coming from all the sub-systems operating in the line to provide higher level information regarding the behaviour of the whole plant (e.g. train behaviour and movements, alarms, communication issues). Track occupancy is observed from TCS through the measure of the shunt level (a parameter related to the resistance of the physical circuit). In normal conditions the shunt level lies in a defined range of values (should be approximately 150% to 160%) while its value sharply decrease to a nil level when a train enters in the track block.

Problem Statement TCS has been designed to be robust to many different failures and problems, and their performances are durable over time. Despite this, some degradation effects exist and unexpected failures may occur (how historical data demonstrates). One of the main issues related to the track circuits systems described above is the false track occupancy phenomena, that means the track is erroneously considered in occupied state due to some Track Circuit malfunctions or external conditions that may affect the correct behavior: this is the kind of anomaly under investigation in this scenario. The reason of such a choice is due to the relevant impact of this anomaly in the standard workflow, since its consequences include traffic interruption and penalties. False occupancy could be caused by:

- Electronic board malfunction (Critical Error)
- Directional Relays failure
- Physical equipment degradation and external factors influence

While board and relay failures occur suddenly (and there are not parameters that could be monitored to assess board and relays status), failures related to equipment degradation are strictly related to shunt level and variance, which are already monitored from TCS boards. This last category of failures is taken into account for the use case definition. Track Circuits parameters (mainly shunt level and variance) have to remain into a

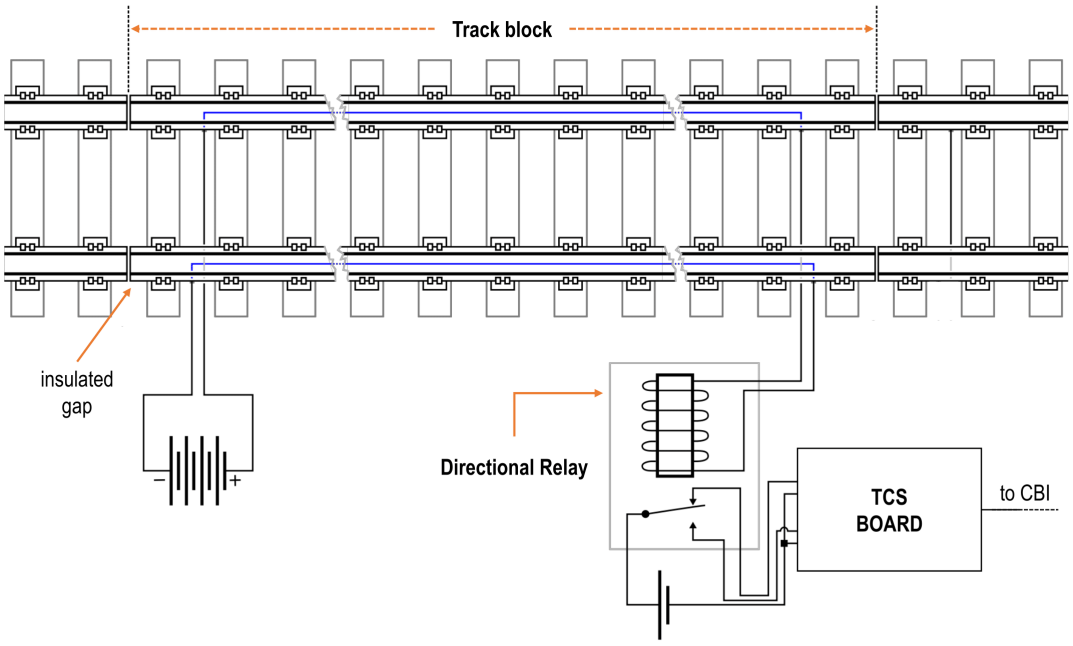

Figure 10: Example of track circuit physical components for a single track block

defined range of values (defined by design), to guarantee the correct behavior of the train detection system. If these parameters go out of the defined range, then a false occupancy event may occur (i.e. due to dirt on track coupling or some dust on the track). This behaviour is shown in Figure 11 for shunt level. Parameters described above are not monitored in an automated fashion and the out-of-range events are managed only once they caused a false occupancy. TCS failures, and then false occupancy occurrences, are observed from TCS board and MicroLok level and then reported from the central system through a set of specific alarms.

Detailed Objectives The objective of the use case is to study and develop data-driven models in order to support and improve TCS maintenance process. Assets involved belong to an urban line in which ASTS developed the signalling system and is in charge of maintenance management. In this scenario, TCS maintenance is so far performed in a corrective fashion. Moreover, some preventive actions are carried out periodically. While corrective maintenance aims to restore functionalities of an asset only once it failed, preventive maintenance consists in a set of scheduled actions in order to avoid asset degradation. However, preventive maintenance is performed mainly due to contract constrains and it is driven by technicians' experience on the field and executed on all the assets of the line without considering their status.
In this context, the general objective of this use case is to develop data-driven models for online and real-time identification and modelling of TCS' status and behaviour, in order to enable diagnostics and prognostics functionalities. In other words, a shift from a standard maintenance approach to a predictive one is pursued. Specifically, two main activities exist: one in the context of anomaly detection (WP8, Task 8.1) and another in the context of predictive modelling (WP8, Task 8.2). Considering the first activity, the goal is the study and the development of a model, based on historical data, for the detection of track circuits anomalous behaviours: TCS' behaviour will be modelled through data of measures of some important parameters of the circuit itself (detailed later) and, when possible, weather data. Considering predictive modelling activity, the objective is the study and the development of a model capable of predicting failures probability with different time horizons, track circuits remaining useful life or failure root causes. Since this second activity is more complex

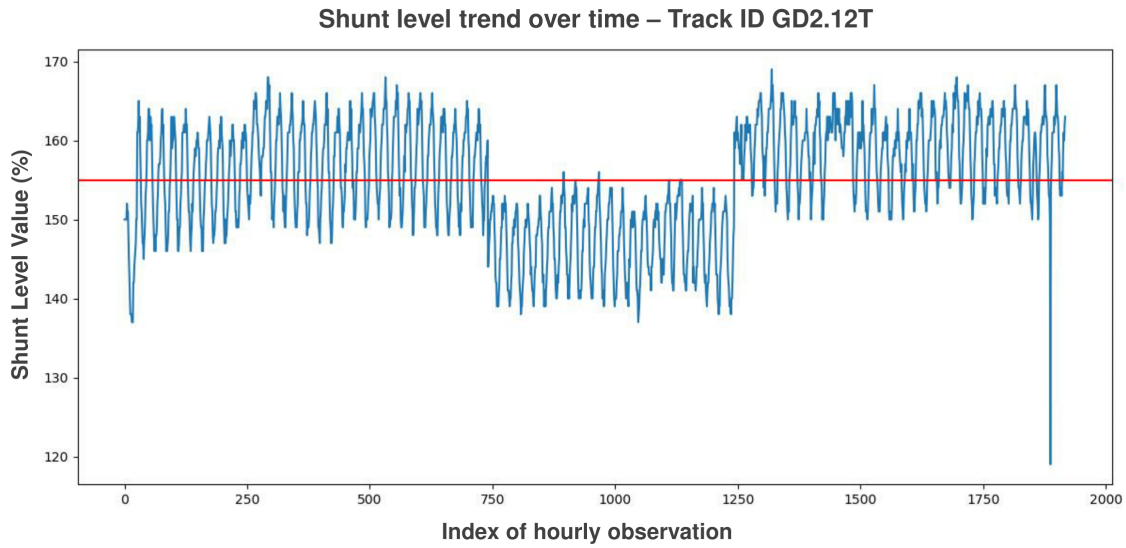

Figure 11: Example of Shunt Level trend over time (3 months observations) for a specific track circuit. The red line represents the ideal lower threshold for TCS correct functioning.

and Task 8.3 is running below with respect to other WP8 tasks, it is possible that the specific goal of this activity will be slightly modified during the remaining project's life depending on data quality and availability. Lastly, it is of course planned to investigate any other interesting failure/malfunction mode of track circuits that will be discussed during the project and for which an evidence in data is conceivable.
Track circuit system has been selected as railway asset to be investigated (both in IN2DREAMS and IN2SMART projects) because its malfunctions and maintenance policies have a high impact on the KPIs targeted by SHIFT2RAIL. Specifically, the work proposed and pursued in this scenario is an essential instrument for SHIFT2RAIL research since it paves the way towards the exploitation of the information content hidden in railway asset data (here represented by TCS) in order to enable Intelligent Asset Management System functionalities. With this aim, transforming data coming from the Railway Integrated Measuring and Monitoring System into actionable knowledge is a key task that must be solved in order to effectively exploit Intelligent Asset Management System potential. Consistent data-driven anomaly detection models, metrics definition and accuracy evaluation will potentially change the shape of the maintenance decision support systems of the future, enabling automated diagnostic and prognostic features for condition-based and predictive maintenance. From a wider perspective, this use case contributes to the definition of the Dynamic Railway Information Management System which aims to define an innovative system for the management, processing and analysis of railway data.

4.3.7 Available Data & Data Access Policies

Available data (mainly coming from track circuits systems and central system) could be grouped depending on their source (as shown in Figure 12):

- **TCS boards parameters.** Collection of measures of most relevant parameters for TCS status monitoring (shunt level, variance, raw signal level and receive level). This data has been acquired from July 2017 (collection is still ongoing) with a frequency of one observation each hour (from January, the acquisition rate has been changed to on observation each 5 minutes). All the fields composing an observation are described in the following table

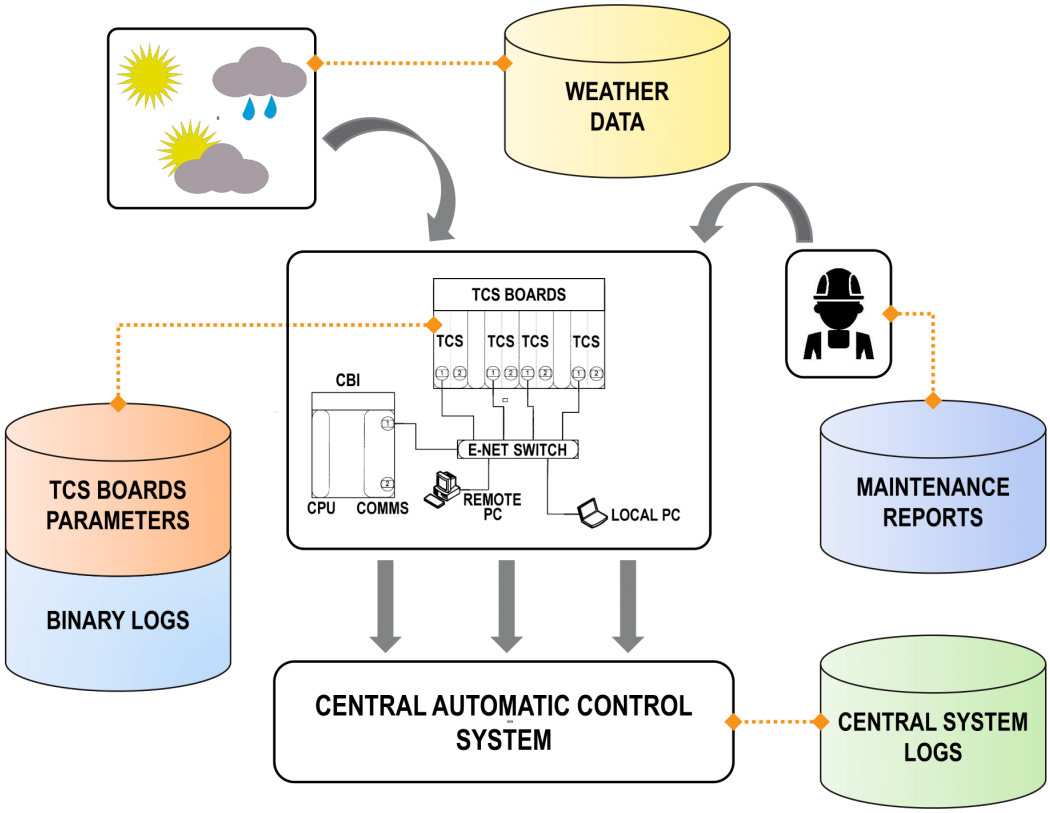

Figure 12: System components interaction and available data sources

Data Fields	Description
Timestamp	Date and time of the observation
Station	Reference station (TCS board rack location)
Track ID	TCS unique identifier
Primary/Backup	Reports to which board (primary or backup) the observation is referred
Shunt Level	TCS status parameter
Variance	TCS status parameter
Raw Signal Level	TCS status parameter
Receive Level	TCS status parameter
Direction	Direction in which the TCS is set

- **Central System logs.** Collection of events and alarm log from the central system. These large set of logs allow to extract information about:
  - Alarms related to TCS (e.g. board failures or occupancy out of sequences etc.)
  - TCS occupancy events (e.g. ”Track A occupied” or ”False occupancy on Track B”)
  - Trains movements events (e.g. ”Train_id XXX moves from Track A to Track B”)
  - Alarms related to other components (e.g. switches, trains, trains doors etc.)
- **Weather data.** Information collected from the closest weather station to the line (multiple weather stations could be exploited if available). These data contain information about the most important weather parameters such as temperature, pressure, humidity, precipitation and so on.

- **Binary logs from boards.** These logs come from TCS boards and contain detailed events about track circuit status, electrical malfunctions and software failures (e.g. ”Config Data Vital CRC Failure”, ”Software Version Critical Failure”)
- **Maintenance reports.** These reports (manually recorded by technical team) contain, for each specific asset, all the related ”Corrective” and ”Scheduled” maintenance activities.

Data from ASTS will be provided anonymously, meaning that no specific asset name will be provided regarding the operational service. All the references to the real data, as object/asset names, will be anonymized in order to avoid privacy and security issues. For what concerns the data shared by ASTS (which is an IN2SMART partner) and IN2DREAMS, refer to the COLA signed between IN2SMART and IN2DREAMS for the details about data exchange policies.

4.3.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Specific-Scenario 1 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  The ability to make the railway infrastructure aware of one of its essential component, the TCS, surely impacts its costs and its capacity. Planning maintenance based on this information can consistently improve the quality of the service provided by the infrastructure by also reducing the costs due to disruptions. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs, we can state that the Specific-Scenario 1 can have an impact into the following scopes of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario is a good use case where to apply the ideas of predictive models which require to be interpreted by an operator. In fact an interpretable model can further improve the ability of the operators to understand the processes of maintenance and optimize them based on the results of these predictions. Moreover interpretable models can easily adapt to take into account the experience of the operators or the previous knowledge about the problem.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  Finding metrics able to understand when our models work well and when they do not is essential to make these data-driven models effective in real world situations. For these reasons, together with ASTS, we will develop specific metrics and KPIs able to detect in what conditions the data driven models perform well and when it is better to exploit the experience of the operators. The metrics developed in this scenario will need to help the operators to better control and understand the TCS decay phenomena and behaviours.

4.3.9 Analytics & Metrics

Concerning use case objectives, the contribution of the work inside IN2DREAMS will mainly focus on anomaly detection in track circuits (with a clear connection to Task 8.1 of IN2SMART WP8) [43, 89, 111, 128, 133, 158]. Anomaly detection is an important problem and it has been researched within many research areas and application domains. The state-of-the-art on this topic is extremely wide, but in the last ten years some exhaustive and comprehensive survey have been published [6, 40, 77]. Moreover, IN2SMART D8.1 contains a brief survey on this subject and introduces the problem declined over railway ecosystem.
Anomaly detection refers to the identification of patterns in data that do not conform to the expected pattern of a given set of data. The definition of unusual pattern (or state) plays a key role in the process and it is

strictly related to systems ”normal” behaviour, which may be defined in advance or learned from data during the process. In this scenario the anomaly is represented by a particular TCS status in which the asset is characterized by some malfunction but still not affected by false track occupancy phenomena. Thus, the proposed anomaly detection system should be able to allow to predict in reasonable advance if a TCS is about to go in a false occupied state or, in other words, to detect the beginning of a degradation pattern. Since no labels are available for TCS status (i.e. normal or degraded/malfunctioning), the underlying analytics scenario could be represented as an anomaly detection task in an unsupervised setting. Moreover, due to the specific and unique behaviour of each a different model will be designed for each TCS. In order to identify unusual patterns, the system will mainly exploit TCS board parameters observations occurred in a specific time-interval. Finally, the potential impact of other variables which could influence TCS behaviour will be investigated (e.g. influence of weather condition on outside track blocks).
More specifically, the contribution of this work will be two-fold: on one side, some metrics in order to evaluate IN2SMART Task 8.1 output model performance will be researched, on the other, a withe box model for degradation status detection will be developed. The output of IN2SMART Task 8.1 activities on this use case is in fact an anomaly detection system which should maximize predictive performance. Thus, a black box unsupervised model (i.e. one-class SVM or k-NN-based models) has been exploited. The evaluation of model performance in an unsupervised setting is a complex activity and thus we aim to investigate which can be the more efficient metrics for the model performance evaluation. Furthermore, considering the model design activity, some common techniques (i.e. Decision Tree or rule-based models) will be exploited and compared. The adoption of a gray-box model for TCS anomaly detection is clearly motivated by the need of a deeper understanding of the problem and a system which could not only alert the technical team about a possible failure but also aid in the failure root causes analysis.

4.4 Specific-Scenario 2: Train Delays and Penalties

4.4.1 Summary

This scenario deals with the problem of building an interpretable and reliable data-driven train delays and train delay penalties prediction model. This is a critical task for the train management system since it is important to both improve train circulation and to reduce delays-related penalties. In this way the train dispatcher can exploit this information and choose a new dispatching solution which minimizes both delay and penalty costs. Moreover we will use the changes in average delays and average penalties due to the new predictive system as a measure of the quality of the penalty system. In Figure 13, we report a simplified representation of the Specific-Scenario 2 in relation with the railway ecosystem.

4.4.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
RFI	IN2DREAMS	WS2	WP5	Responsible Partner	Yes
UNIGE	IN2DREAMS	WS2	WP5	Contributor	No

4.4.3 Connection with other Scenarios

Specific-Scenario 2, as its name states, is a specific-scenario and for this reason it is mainly connected with the cross-scenarios.

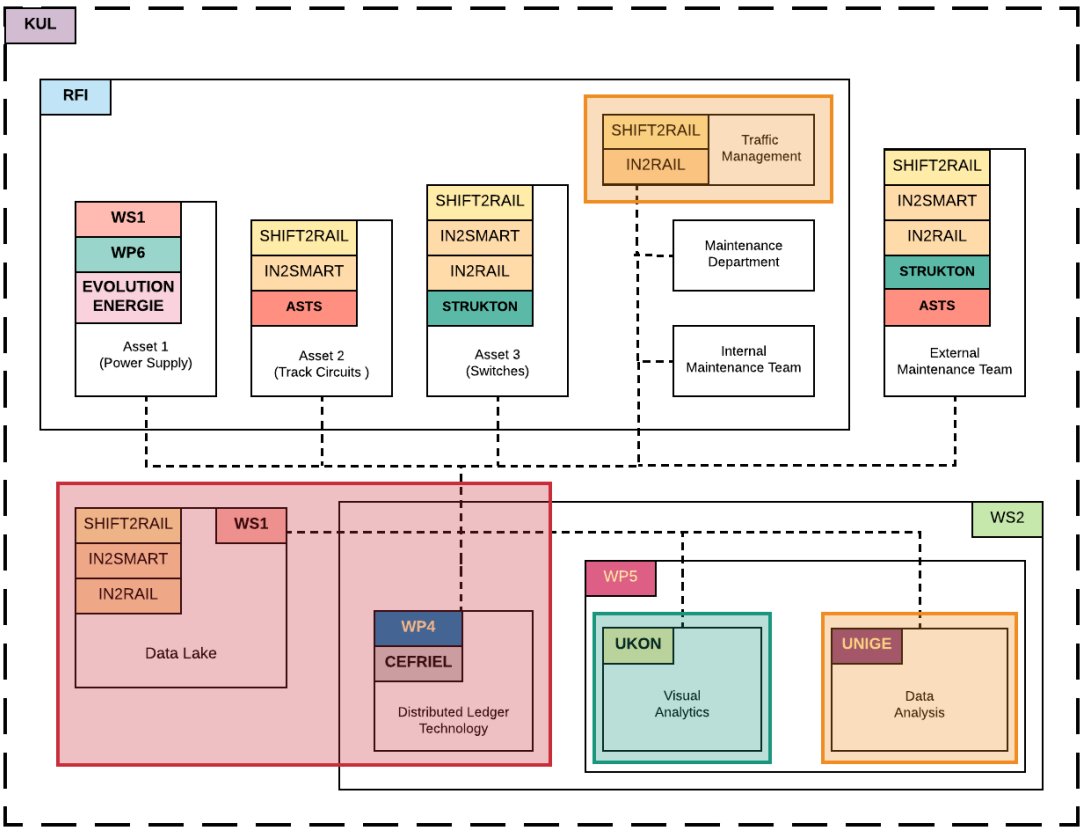

Figure 13: Specific-Scenario 2 Reference Picture (Train Delays and Penalties)

- Connections with Cross-Scenario 1 (Section 4.1)
  Specific-Scenario 2 will provide to Cross-Scenario 1 other useful information which will be exploited from the operators (mainly from dispatchers) by means of visualization. In particular, the possibility to know in advance the delay of a train and the impact of these delays on the penalty costs that RFI has to pay is fundamental to optimize and improve the decision-making process of a dispatcher. Moreover, together with prediction results, Cross-Scenario 1 has to show also the quality of the prediction and the reason of the prediction itself. Starting with results from interpretable models developed in Specific-Scenario 2, this aim will be pursued developing a nice and intuitive graphical representation that can help the operator in deciding through his personal experience weather or not to rely on the prediction. Moreover an interpretable model can surely give to the dispatcher the ability and the evidence of a needs for changes in the timetable planning.
- Connections with Cross-Scenario 2 (Section 4.2)
  The connection of Cross-Scenario 2 is rather straightforward, in fact the models developed with the data provided by RFI and the associated predictions can be sold on the data marketplace as the data itself (for example to the TOs). Moreover, the data marketplace can be a source of additional data that can be exploited for improving the quality of the model. The results of Specific-Scenario 2 can give a hint on what could be a new source of information that could help improving the quality of the models (for example number of pessengers from the TOs).
- Connections with Specific-Scenarios 1, 3, 4, and 5 (Sections 4.3, 4.5, 4.6, and 4.7)
  The connection of Specific-Scenario 2 with Specific-Scenarios 2, 3, 4, and 5 is quite tight. Specific-

Scenarios 1, 4 and 5 deal with the problem of predicting a possible malfunction. Specific-Scenario 3 deals with the problem of predicting the restoration time from a maintenance or a malfunctions. Specific-Scenarios 2 deals with the train delay prediction. Then, the scope of the different scenarios is complementary. We envision a future where predictive models of a malfunction will be exploited together with the one which forecasts the restoration time of an asset and the train delays in a way to optimize the train circulation for improving the cost efficiency, the reliability, and the capacity of the infrastructure while limiting the penalties that the IMs has to pay the delays.

4.4.4 Connection with other IN2DREAMS WSs ans WPs and SHIFT2RAIL Projects

The connection of Specific-Scenario 2 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection here is well depicted in Figure 13, in fact the data that we will use in this scenario are stored in the data lakes of RFI. Then, these data can be shared in the marketplace envisioned in the IN2DREAMS WS2 WP4 with a twofold objective: on one side, the monetization of this information by selling it to other actors of the railway ecosystem, on the other, the enrichment of the available data by buying some other detailed information from other actors (e.g. number of passengers reservations given by the TOs) in order to improve the quality of the prediction. Moreover, the models and the predictions themselves can be sold to other actors as a service.
- Connections with IN2SMART and other SHIFT2RAIL projects
  This use case has been designed inside WP9 of the IN2RAIL Project and also exploited in the WP8 of the IN2SMART Project. This scenario in an extension of the scenario started in IN2RAIL and our scenario builds upon these results, lesson learned, and suggestions coming from IN2RAIL. In particular in IN2RAIL the main question raised by the project was to define a simple and interpretable model of the train delay phenomena which should be not simple constructed based on data but also based on the actual physical system in a gray box approach. In this scenario we will address these requests and suggestions.

4.4.5 Scenario Objective(s)

The main objectives of this scenario are the following ones.

- Develop interpretable and gray-box models that exploit a similar approach to the current in place model by RFI of delay prediction. However, in our model we differentiate the prediction taking in consideration the weather conditions, the typology of the train, the different days of the week, the different hours of day, and the current delay of the train.
- Study the penalty system in case of delayed train and integrate such study in the predictive model.
- Permit the prediction also for trains for which no historical information are provided and/or changes in the timetable.

4.4.6 Scenario Description

This scenario aims at developing a set of interpretable data-driven models able to estimate the delay of a train and to predict it at future times while predicting also the cost of a delay in terms of the penalties that the train management system has to pay. In other words, the main goal is to predict the series of delays that will affect a specific train in all the subsequent “checkpoints” included in its journey, with the highest possible accuracy and, when possible, with an estimate of the forecasting accuracy itself. While predicting the delay

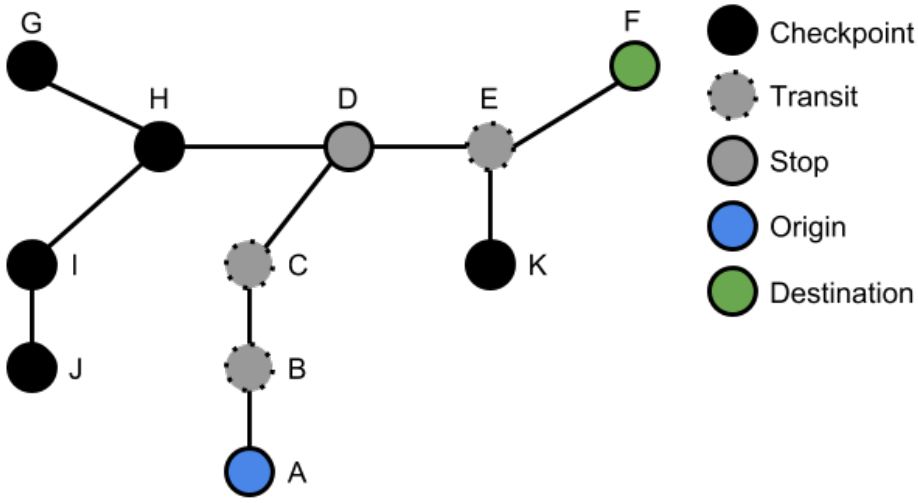

Figure 14: A train itinerary on a railway network

we consider also the penalty cost because making a mistake in predicting the delays it is not just a problem for the TMS but also for the penalties that RFI has to pay. In the rest of this section we firstly describe the problem under exam and the main definitions in order to understand the section. Subsequently, we describe the system currently in-place by RFI to perform the train delay prediction and how it has been improved in the previous IN2RAIL project. Finally, we describe the shortcomings of the two approaches and we show how our novel proposal can improve the previous approaches integrating train delays with penalty estimation.

A railway network can be considered as a graph. Figure 14 describes a railway network where a train follows an itinerary characterized by a station of origin (station A), a station of destination (station F), some stops (station A, D, and F) and some transits (station B, C, and E). We call a checkpoint a station without differentiating between station where the train stops or transits. For any checkpoint, the train is scheduled to arrive at a specified time and to depart at a different specified time, defined in the timetable. Usually, the time references included in the nominal timetable are approximated with a precision of 30 seconds. The difference between the time references included in the nominal timetable and the actual time, either of arrival or of departure is defined as delay. If the delay is greater than 30 seconds, then a train is considered a delayed train. Note that for the origin station there is no arrival time, while for the destination station there is no departure time. We define the dwell time as the difference in seconds between the departure time from a specific checkpoint and the arrival time in the very same checkpoint. Moreover, we define the running time as the amount of time needed to depart from the first of two subsequent checkpoints (i.e. i) and to arrive to the second one (i.e. $i + 1$).

To each train is associated a unique identifier. From this identifier it is possible to retrieve the category of the train. Moreover, each checkpoint is characterized by a unique identifier that can be exploited to retrieve the kind of the railway network. The category of the train and the kind of railway network affect how the penalty cost must be computed. We define also a section equals to the itinerary between two consecutive stations. For instance, section $D - E$ of Figure 14 is a section where the starting point is D and E is the destination station. Note that we are considering also the orientation of the itinerary, for this reason the section $D - E$ is different from the section $E - D$.

In the actual RFI system, for each section, it is defined by RFI a coefficient which estimates the time that can be regained by any train in that section. This coefficient is static, i.e. it does not change for different trains, state of the network, weather conditions, etc. In such system, when predicting a delay, it is assumed that a delayed train is always able to regain in a given section an amount of time equals to the coefficient of that

section. For this reason, when RFI predicts the delay of a train in a subsequent checkpoint it subtracts from the current delay all the coefficients of the sections between the origin station until the considered one.
In the IN2RAIL project has been proposed a different approach with respect to the RFI system to predict the train delays. Such approach is based on a set of data-driven models that, working together, make it possible to perform a regression analysis on the past delay profiles and consequently to predict the future ones. In particular, for each train and for each checkpoint composing its trip, a set of data-driven regression models is built connecting one checkpoint to its successive ones. Specifically, for each train characterized by a specific itinerary of checkpoints, models have to be built for the first checkpoint, for the next one, and so on. Consequently, the total number of models to be built for each train can be calculated as $n * (n-1)/2$ where n is the number of checkpoints visited by the train. These models work together to make it possible to estimate the delays of a particular train during its entire journey. For a single train arriving at (departing from) a specific checkpoint included in its trip, the data-driven regression models take as inputs the sequence of arrival and departure delays affecting that specific train in the current day at its passage/stop in the previous checkpoints (i.e. from origin to the last visited checkpoint) and both the sequence of running times and dwell times for the considered train and output the predicted delay for the specified checkpoint. The models are built exploiting the Random Forest regressor (RF). RF is a machine learning algorithm especially known to be one of the most effective tools for classification purposes, but it can be adapted to solve regression problems.
The above described models present several downsides. The static coefficients of the RFI system do not permit to take into account that the traffic on the railway network can be peculiar in different moment of the days, in weekend and weekdays, if the train is affected by a delays, or because the weather is different. Instead, the IN2RAIL models are specific for each train and it is not possible to predict the delays for trains for which we have not historical data, and, also, the number of generated models may be huge.
For these reasons we propose a different solution where we bring the benefits of the above solutions and we improve them in the scenarios where are not behaving correctly. In our solution we choose to follow a similar approach to the one in place by RFI based on the coefficients to estimate the train delays. However, our idea is to make such coefficients not static but dependant to several variables. Such variables are all correlated variables, similarly to the IN2RAIL model, (e.g. the delay in the previous checkpoints, the delay in the previous days on the same itinerary, the amount of time similar trains spend to traverse the same section etc.) to our variable of interest being the definition of a dynamic coefficient for each section. The goal is to define a methodology able to model the link between the variable of interest with its past values (i.e. its history) and the behaviour of similar trains traversing the same sections in similar scenarios.
Our solution builds a set of interpretable data-driven models to define the coefficients of the sections dynamically. Such coefficients are then exploited to predict the future delays making use of a prediction algorithm similar to the one in place by RFI, where the time a train can re-gain in a given section is defined by the coefficient and subtracted from the current delay.
An example of our methodology is shown in Figure 15, where a train is in the middle of its journey in checkpoint C and the prediction system has to provide predictions about the delay in the checkpoints $D, E, \ldots, F$ composing the train trip. Such train trip is composed of several sections: $C-D, D-E, \ldots, E-F$. Each section exploits a specific data-driven model able to estimate the amount of time that the train is able to re-gain with respect to the delay in checkpoint C. The Get function is exploited to retrieve such time. Moreover, in such case we also update the model (i.e. $Update$ function in the figure) regarding the section $B-C$ already traversed computing the transit time in that section for the considered train.
To be more precise, in the proposed solution we build a set of models to estimate a specific coefficient differentiating between several variables. The variables considered are characteristics of a train and exogenous ones. To sum up, we build one model for each combination of the following variables taken in input:

- A section, where we consider also the orientation. A section from checkpoint A to B is different from the section B to A;

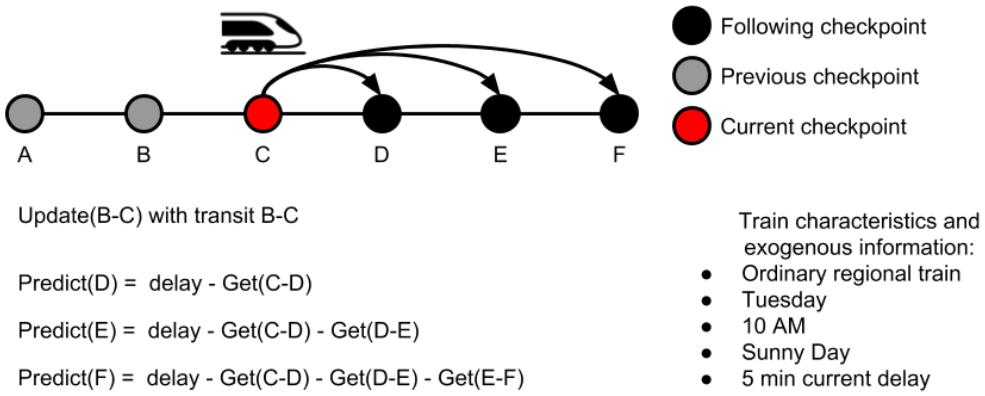

Figure 15: The prediction and update

- The time of the day, we consider the scheduled hour in the considered checkpoint;
- The type of the train (regional, intercity, . . .);
- The current day of the week (Monday, Tuesday, . . .);
- The weather conditions (Sunny, Light Rain, Heavy Rain, Snow);
- A set of intervals representing the last known delay of the train in minutes ($[0, 2]$, $]2, 5]$, $]5, 10]$, . . .).

Moreover, each model gives in output a coefficient specialized for one combination of the above variables. Two trains travelling on a same section and sharing the same values for the above variables will share also the same coefficient. For this reason, this solution is able to provide a prediction also for a new train travelling in the network based on similar trains that have been travelling on the same section in a similar scenario. In addition, such solution permits to handle changes in the timetable considering for each train similar trains that in the past travelled on the same sections.
In addition, to the above considerations, these models take also in account the cost of a wrong prediction by studying the penalty system defined in the RFI contract. The RFI contract defines the cost P of a delayed train as following:

$$P = P_u(M_{GI}C_tC_{cat}C_{del} + M_{NJ}C_{cat}C_{del}) + P_s(S_{GI}P_{sop}C_{sop}) \quad (1)$$

Where C_t is the coefficient of the section. Such value ranges between 0.25 to 3. It mainly differentiates between the types of the section: fundamental, complementary or a node section. C_cat is the category coefficient. It can assume value 1 for scheduled trains, 0.5 for trains operated in operational management, 0.25 for other types of scheduled services. C_{del} is the coefficient of multiplication that changes considering the average and maximum delay registered in a checkpoint where the train is stopping for the passengers or the delay in the final checkpoint for freight trains. M_{GI} is the amount of minutes where the responsibility is of the management of the infrastructure. Whereas, M_{NJ} is the amount of unjustified minutes that will be attributed to the responsibility of management of the infrastructure.
Finally, it is necessary to add some considerations about this formula. The part of the formula regarding the suppressed trains (i.e. $P_s(S_{GI}P_{sop}C_{sop})$ is not relevant in our scenario because in our solution we want to predict the delays of the trains that have actually traversed their itinerary. In addition, RFI experts have communicated during the creation of the models that we can assume the total responsibility of the minutes of the delays is in charge of the infrastructure manager (M_{GI}) and all the minutes of delays are justified. For this reasons, the above formula can be simplified removing the part of M_{NJ} and suppressed trains and we obtain the following formula:

$$P = M_{GI}C_tC_{cat}C_{del} \quad (2)$$

Where we set also P_u equals to 1.00 (one) Euro/minute.

4.4.7 Available Data & Data Access Policies

For what concern the Railway Information Systems, from every TMS, it is possible to retrieve:

- Data about train movements, with time and position references (e.g. timestamps at station arrivals and station unique IDs);
- Theoretical timetables, including planning of exceptional train movements, cancellations, etc.

Taking the RFI TMS databases as an example, the following table shows the data that is recorded every time a train passes through a station or a checkpoint.

Field Name	Datatype	Description
TRAIN DATE	datetime	the date expressed in day, month, and year (formatted as ”dd/mm/yy”) of the record
TRANSPORT NUM	string	the unique identifier of the train
PIC PLACE CODE	string	the unique identifier of the station
PLACE DSC	string	the name of the station
ARRIVAL DATE	datetime	the date of arrival of the train in the station (formatted as ”dd/MM/yy”)
ARRIVAL TIME	datetime	arrival time of the considered train (formatted as ”HH:mm:ss”)
ARRIVAL DIFF	float	the delay of arrival expressed in seconds
DEPARTURE DATE	datetime	the date of departing of the train in the station (formatted as ”dd/MM/yy”)
DEPARTURE TIME	datetime	departing time of the considered train (formatted as ”HH:mm:ss”)
DEPARTURE DIFF	float	the delay of departing expressed in seconds
PASSAGE TYPE	char	the type of station checkpoint from the train point of view. There are four possible values: O (Origin), D (Destination), T (Transit), and F (Stop)

From the management principles document it is explained how to use the train identifier to retrieve the category of the train (note that this applies only to RFI IM). Different categories have different coefficients to be used during the computation of the penalty system. Based on the train identifier the category of a train can be retrieved from the following table.

Train categories	From	To
Ordinary and extraordinary regional trains	103000	103999
	109000	109999
Ordinary and extraordinary regional trains belonging to the market and universal service	100000	100999
	101700	108999
	110000	112999
	120000	125499
Freight trains	125500	127999
	137800	137999
	139000	199999

In addition it is defined:

- the coefficient of the type of the section. It differentiate between the types of the section: fundamental, complementary or a node section;
- the category coefficient. Differentiating between scheduled trains, trains operated in operational management, and other types of scheduled services;
- the coefficient of multiplication that changes considering the average and maximum delay registered in a checkpoint where the train is stopping for the passengers or the delay in the final checkpoint for freight trains.

In particular, the coefficient of the section C_t can be retrieved from the following table.

C_t: **the coefficient of the section**	
High Speed Network	3
Fast Lines (DD)	3
Nodes	2.5
Lines of Connection with Europe	2
Central-north National Corridors	1
Central-south National Corridors	1
Other Lines	1
Secondary Complementary Network	0.5
Complementary Network of Goods Routes	0.5

The coefficient relative to the category of the train C_{cat} follows the following rule.

C_{cat}: **the coefficient for the category of train**	
Train scheduled on time	1
Train operated in operational management	0.5
Other types of scheduled services	0.25

Finally, the coefficient of delay C_{del} must be retrieved from the following three tables based on the category of the train.

C_{del} **for ordinary and extraordinary regional trains**						
AVG / MAX	≤ 5	≤ 15	≤ 30	≤ 60	≤ 120	> 120
≤ 5	0.25	0.5	0.75	1	1.25	1.5
≤ 15	-	1	1.25	1.5	2	2.5
≤ 30	-	-	1.5	1.75	2.25	2.75
≤ 60	-	-	-	2	2.5	3
≤ 120	-	-	-	-	3	3.5
> 120	-	-	-	-	-	4

C_{del} **for ordinary and extraordinary regional trains belonging to the market and universal service**						
AVG / MAX	≤ 5	≤ 15	≤ 30	≤ 60	≤ 120	> 120
≤ 5	0.25	0.25	0.5	1	1.5	2
≤ 15	-	0.5	0.75	1.25	1.75	2.25
≤ 30	-	-	1.25	1.5	2	2.5
≤ 60	-	-	-	2	2.5	3
≤ 120	-	-	-	-	3	3.5
> 120	-	-	-	-	-	4

C_{del} **for freight trains**	
≤ 5	0.25
≤ 15	0.25
≤ 30	0.5
≤ 60	1
≤ 120	1.25
≤ 180	1.5
> 180	2

External Data sources. Other useful exogenous information to be correlated with train delays could be retrieved from external databases, like for example:

- Information about tourists presence in an area. Unfortunately, these data are often not available because of privacy/confidentiality issues;
- Information about the number of passengers on each train. This information is often available only for trains with seat reservations;
- Information about weather conditions from weather stations in the area. For instance, by looking for the closest weather station to the railway station/line.

Consequently, weather historical data will be included in the predictive models developed for this scenario by taking advantage of the Italian national weather service databases, which are publicly accessible.

4.4.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Specific-Scenario 2 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  The ability to make the railway infrastructure aware of a reliable an interpretable model of the train delays and penalties costs has surely an impact in terms of costs-savings and improved capacity. Planning an overtaking or a precedence based on these forecast can improve the capacity of the network preventing clog and unnecessary high penalty costs. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.
- Impacts on the IP2: Advanced Traffic Management and Control Systems
  This scenario also impacts the IP2 because we can make the TMS aware of the future train delays and the motivations can improve the ability of the dispatchers to handle maintenance or repairs. Consequently TD2.9 is surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs we can state that the Specific-Scenario 2 can have an impact into the following scopes of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario is a good use case where to apply the ideas of predictive models which require to be interpreted by an operator. In fact an interpretable model can further improve the ability of the dispatchers to understand the recurrent delays and their cause and optimize the scheduling or the timetable planning based on the results of these predictions. An interpretable model, for example, can suggest that in particular days of the week or with particular weather conditions certain type of delays are more or less frequent. Moreover interpretable models can easily adapt to take into account the experience of the dispatchers or the previous knowledge about the problem.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  Finding metrics able to understand when our models work well and when they do not is essential to make these data-driven models effective in real would situations. For these reasons, together with RFI, we will develop specific metrics and KPIs able to detect in what conditions the data driven models performs well and when it is better to exploit the experience of the operators. The metrics developed in this scenario will need to help the dispatchers to better control and understand the train delays and associated penalties costs.

4.4.9 Analytics & Metrics

In the context of train delay prediction we already introduced several models in Section 4.4.6 (i.e. the one in place by RFI and the one proposed in the IN2RAIL project) because are the two models most related to ours. Such models present several downsides. The static coefficients of the RFI system do not permit to take into account that the traffic on the railway network can be peculiar in different moment of the days, in weekend and weekdays, if the train is affected by a delays, or because the weather is different. Instead, the IN2RAIL model is specific for each train and it is not possible to predict the delays for trains for which we have not historical data and, moreover, the number of generated models may be huge.
Nevertheless, a large literature covering this problem [23, 64, 86] already exists. Such models consider the railway network parameters and the information that can be retrieved from the classical railway information systems as the only data sources (e.g. the TMS for records about train movements). However, in a railway network it is common to have changes on the timetable and to handle trains for witch no historical information are available. For this reason, in this scenario we build a simplified model that it is able to forecast train delays also in such cases based on historical information about similar trains. Moreover, other works consider different ways to construct the models, for instance making use of fuzzy Petri net (FPN) model for estimating train delays [138]. The FPN model has been used to simulate traffic processes and train movements in a railway system. However, when there were no historical data on train delays, they make use of expert knowledge to define rules along with the timetable and infrastructure data. For this reason the model is not automatically adapting to previously unseen trains or changing in the timetables. Moreover, Barta et al. [17] studied a method to forecast the delay propagation in railway networks by developing a Markov-chain based model, which was based on the examination of a large set of historical data and can discover the probability of absorbing or propagating delays. Also in this case the model is based only on historical information for each train being not able in case were no historical data on train delays is available.
Contrary to all the above described works we study and consider also the penalty costs. This is of paramount importance in order to minimize the penalty costs that must be payed. In addition, many other exogenous factors affect railway operations, such as passenger flows, strikes, celebrations and similar public events. Thus, this scenario concentrates on predicting train delays by means of a set of data-driven models integrating data about past train movements and weather conditions.

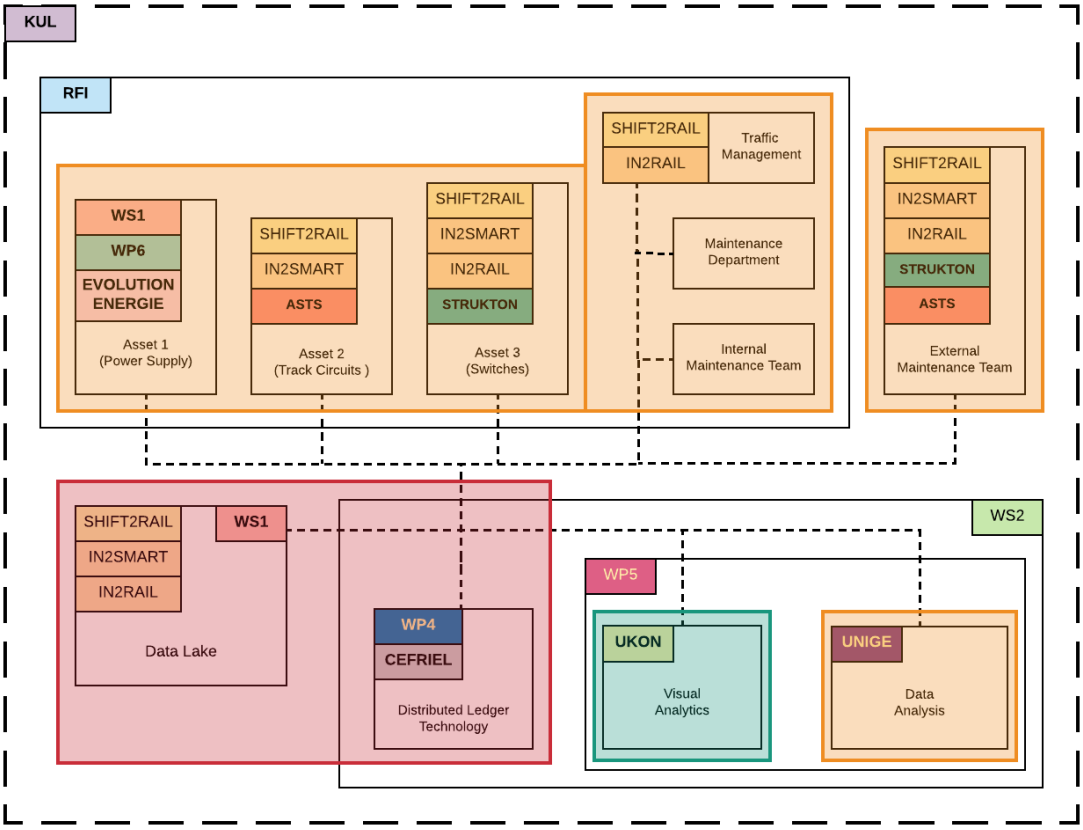

Figure 16: Specific-Scenario 3 Reference Picture (Restoration Time)

4.5 Specific-Scenario 3: Restoration Time

4.5.1 Summary

This scenario deals with the problem of building an interpretable and reliable data-driven incident restoration time forecast system. The information outputted by the models can be very useful because it could be used by the traffic management system to reroute trains through safer paths, minimizing the risks of any problem or delay.

4.5.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
UNIGE	IN2DREAMS	WS2	WP5	Responsible Partner	No
RFI	IN2DREAMS	WS2	WP5	Contributor	Yes
STRUKTON	IN2SMART	-	WP8	Advisor	Yes

4.5.3 Connection with other Scenarios

Specific-Scenario 3, as its name states, is a specific-scenario and for this reason it is mainly connected with the cross-scenarios.

- Connections with Cross-Scenario 1 (Section 4.1)
  Specific-Scenario 3 will provide to Cross-Scenario 1 other useful information to visualize to the operators, principally the dispatcher. In particular, the restoration time is quite useful, from the dispatcher perspective, since knowing in advance with high accuracy when a track section will be free and then again available for train circulation is fundamental and can deeply change the effectiveness of the dispatcher's work. Moreover, together with the prediction itself (obtained thanks to the interpretable models developed in Specific-Scenario 3), Cross-Scenario 1 has to show also the quality of the prediction and the reason behind it, with a nice and intuitive graphical representation that can help the operator in deciding to or not to rely on the prediction on its experience.
- Connections with Cross-Scenario 2 (Section 4.2)
  The connection of Cross-Scenario 2 is rather straightforward, in fact the models and the prediction of the models developed with the data provided by RFI and STRUKTON can be sold on the data marketplace as the data itself. Moreover the data marketplace can be a source of additional data that can be exploited for improving the quality of the model. The results of Specific-Scenario 3 can give an hint on what could be a new source of information that could help improving the quality of the models.
- Connections with Specific-Scenarios 1, 2, 4, and 5 (Sections 4.3, 4.4, 4.6, and 4.7)
  The connection of Specific-Scenario 3 with Specific-Scenarios 1, 2, 4, and 5 is quite tight. Specific-Scenarios 1, 4, and 5 deal with the problem of predicting a possible malfunction and Specific-Scenario 3 deals with the problem of predicting the restoration time from a maintenance or a malfunction. Specific-Scenarios 2 deals with the train delay prediction. Then the scope of the different scenarios is complementary. We envision a future where predictive models of a malfunction will be exploited together with the one which forecasts the restoration time of an asset and the train delays in a way to optimize trains circulation. This will lead to an improvement of the cost efficiency, the reliability, and the capacity of the infrastructure.

4.5.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

The connection of Specific-Scenario 3 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection here is well depicted in Figure 16, in fact the data that we will use in this scenario are stored in the data lakes of STRUKTON and RFI. Then these data can be shared in the marketplace envisioned in the IN2DREAMS WS2 WP4 in order, from one side, to monetize this information by selling it to other actors of the railway ecosystem and, from the other side, these data could be enriched by buying some other detailed information from other actors (e.g. restoration time of other maintainers for assets that STRUKTON or RFI do not maintain often) in order to improve the quality of the prediction. Moreover, the models and the predictions themselves can be sold to other actors as a service.
  There is also a strong connection between this scenario and the WP4 selected scenario on Asset Maintenance (D4.1 Section 5.1). In fact the operators need to be aware of the status of the maintenances and this information can be extracted from the blockchain developed in WP4 and vice-versa the blockchain can be fed with the output of the data driven models developed in WP5.
- Connections with IN2DREAMS WS1, IN2SMART, and other SHIFT2RAIL projects
  In this scenario the connection with IN2DREAMS WS1 is not so strong but still present, in fact power

supply assets under study belong to the maintained assets belonging to the data provided by RFI and STRUKTON. Instead, the connection to other SHIFT2RAIL projects, in particular IN2RAIL and IN2SMART, is very strong. In fact, STRUKTON, one of the data provider of this scenario, belongs to the IN2SMART project and also to IN2RAIL. This scenario in an extension of the scenario started in IN2RAIL and our scenario builds upon these results, lesson learned, and suggestions coming from IN2RAIL.

4.5.5 Scenario Objective(s)

The main objectives of this scenario are the following ones.

- Develop interpretable and gray-box models of the restoration time phenomena for the purpose of predicting the restoration time of planned and urgent maintenances. This model will be based on historical data collected in two different scenarios: the Italian railway infrastructure (data provided by RFI) and the Dutch railway infrastructure (data provided by STRUKTON).
- Develop metrics and KPIs able to determine the situation in which our models can be effectively be applied in real operations and when we have to leave the choice to the operators because the developed models are not reliable enough because of the quality or the quantity of the historical data.

4.5.6 Scenario Description

Every time an infrastructure asset is affected by a failure/problem, it is clear that this will affect not only the single asset functional behavior, but also the normal execution of railway operations. The functional behavior of railway infrastructure assets degrades for many different reasons: age, extreme weather conditions, heavy loads, and the like. For example, the influence of snow on switches is critical, in particularly when switch heating is not functioning properly. Even worse is the case of wind in combination with snowfall, when the assets belonging to a specific area can be significantly affected. Additionally, problems can be introduced unknowingly by performing maintenance actions, for example by a simple human error or as a reaction of the system to changes made on an object. For instance, some maintenance activities (e.g. tamping or ballast dumping) performed close to a switch can change the track geometry, then other parts of this asset must be adjusted to the new situation. For this reason in this scenario we will investigate the problem of estimating and predicting the time to restoration for different assets and different failures and malfunctions. In other words, the objective of this analysis is to estimate the time to restoration for future (planned) and urgent maintenance actions by looking at the past maintenance reports, correlated to the different assets and different types of malfunctions. The predictive models that will be designed will be able to exploit the knowledge enclosed into maintenance reports in order to predict the time needed to complete a maintenance action over an asset in order to restore its functional status. This information will help the TMS in properly assessing the availability of the network, for example by estimating the time at which a section block including a malfunctioning asset will become available again.
This work has been started in the WP9 of IN2RAIL with data proved by STRUKTON but the obtained results were not good enough to be employed directly during operations. One open question and output of the WP9 of IN2RAIL was the suggestion to further develop the work done in IN2RAIL by:

- improving the data cleaning phase, since data were pretty noisy, in order to improve the quality of the predictive models, given that IN2RAIL results were quite promising and actually the data-driven models seemed to be able to detect the reasonable and important sources of information that can be used to predict the restoration time.
- comparing and combining the results with human made estimates in a sort-of grey box approach if this human made estimates can be made and recorded in an historical database.

- using white box and interpretable models for reducing the complexity of the problem to group subsets of similar maintenance tasks and build different models for different problems; in this way the data would be of smaller cardinality and less affected by noise.

For this reason in this scenario we will go in this direction by both following the suggestions and hints given by the IN2RAIL work to further develop the models on the STRUKTON data, and we will employ new data coming from RFI again about the restoration time of planned and urgent maintenances.
The final scope will be to develop white box and interpretable model assessed with custom KPIs developed in collaboration with RFI and STRUKTON.

4.5.7 Available Data & Data Access Policies

For the implementation of this scenario STRUKTON and RFI made available a series of data sources.
For what concerns STRUKTON they made available two main datasets from 2014 to 2017:

- **Weather condition:** data retrieved from the Royal Netherlands Meteorological Institute (KNMI).

Field Code	Field Description
DD	Wind direction averaged over the last 10 minutes of the last hour
FH	Hourly average wind speed
FF	Wind speed averaged over the last 10 minutes of the last hour
FX	Highest gust over the last hour
T	Temperature at 1.50 m altitude during observation
T10N	Minimum temperature at 10 cm height in the last 6 hours
TD	Dew point temperature at 1.50 m altitude during observation
SQ	Duration of sunshine per hour, calculated from global radiation
Q	Global radiation per hour compartment
DR	Duration of precipitation per hour compartment
RH	Hourly precipitation
P	Air pressure converted to sea level, during observation
VV	Horizontal view during observation
N	Overcast during observation
YOU	Relative humidity at 1.50 m altitude during observation
WW	Weather code observation mode
IX	Weather code indicator for automatic station
M	Fog in the previous hour and / or during the observation
R	Rain in the previous hour and / or during the observation
S	Snow in the previous hour and / or during the observation
O	Thunderstorm in the previous hour and / or during the observation
Y	Ice formation in the previous hour and / or during the observation

- **Maintenance/repair actions:** historical datasets regarding the maintenance activities (including their duration) will be provided by STRUKTON. This data is collected by STRUKTON but commissioned by ProRail (Dutch rail infrastructure manager). Data is originally stored in a Maintenance Management System. In addition to the malfunctions, extra information can be added to better quantify the malfunctions in relation to occurred delays or affected trains. This information could be provided by Asset Manager, ProRail.

Field Code	Field Description
Priority-Code	A numerical code that indicates the urgency of the failure
Geo-Code	The code of the Geographical location of the failed asset
Km-Location	Better localization of the failure location on the track
Failure-Type	Generic classification of the (main) problem
Object-Type	The main type of the asset
ObjectID	Unique ID of the asset which failed
Reference-number-SR	Reference ID (internal Database number)
Technical-Department	Department responsible for fixing the failure
DT-notification	Date and Time when failure was reported
DT-Mechanic-informed	Date and Time when the repair team was informed about failure
DT-Mechanic-on-location	Date and Time when the repair team was arrived on the site / location
DT-Starting-repair	Date and Time when the repair team started fixing the problem
DT-function-restored	Date and Time when the failure was fixed and function of the asset restored
DT-Repair-wanted	Date and Time of the problem to be fixed as notified by the train operator
Year	The Year when the problem was reported
Month	The month when the problem was reported
ResponceTime	Responce time in minutes
RepairTime	Repair time in minutes
Function-restorationTime	Total repair-time in minutes
Part-Code	Failed part code / Failed part number
Part-description	Standardized description of the failed part
Action-carried-out-code-description	Description of the coded action (number) taken in order to solve the problem
Action-carried-out-code	Standardized number of the action taken
Failure-cause-main-group	Main group of the failure cause
Failure-cause-code	Standardized number of the fail-cause
Failure-cause-Description	Failure cause
GEO-Shape-Length	Length of the track section involved
X(Long.)-Begin	Longitude position of the beginning of the track section involved
Y(Lat.)-Begin	Latitude position of the beginning of the track section involved
X(Long.)-Mid	Longitude position of the center of the track section involved
Y(Lat.)-Mid	Latitude position of the center of the track section involved
X(Long.)-End	Longitude position of the end of the track section involved
Y(Lat.)-End	Latitude position of the end of the track section involved
Role	Role(s) of the mechanic(s) involved
MechanicID	ID of the mechanic(s) involved
Mechanic-sequence nr	Sequence number of the mechanic(s) involved

For what concerns RFI they made available one dataset containing all the maintenance activities between 2015 and 2018 of a particular section of the railway network (north west of Italy). Each maintenance activity is described by the following information.

Field Code	Field Description
Codice	Identifier of the maintenance
Num. Prg.	Flag to indicate if the maintenance was scheduled or not
Cod. Progetto	Code of the maintenance project
Tipo	Type of the maintenance
Stato	State of the maintenance
Localita' Inizio	Start location of the maintenance
Incl/ Escl	Flag of inclusion or exclusion of the start location from the maintenance
Localita' Fine	End location of the maintenance
Incl/ Escl	Flag of inclusion or exclusion of the end location from the maintenance
Bin.	Track involved
Data Ora Inizio Prog	Timestamp start maintenance (scheduled)
Data Ora Fine Prog	Timestamp end maintenance (scheduled)
Durata Prog	Length maintenance (scheduled)
Data Ora Inizio Reale	Timestamp start maintenance (actual)
Data Ora Fine Reale	Timestamp end maintenance (actual)
Durata Reale	Length maintenance (actual)
Note	Note of the operators

Moreover historical data about weather conditions and forecasts are publicly available from the Italian weather services and contain the following information.

Min/Max/Average Hourly temperature
Min/Max/Average Hourly relative humidity
Min/Max/Average Hourly wind direction
Min/Max/Average Hourly wind intensity
Min/Max/Average Hourly rain level
Min/Max/Average Hourly pressure
Min/Max/Average Hourly solar radiation

Data from RFI and STRUKTON will be provided anonymously, meaning that no specific asset name will be provided regarding the operational service. All the references to the real data, as object/asset names, will be anonymized in order to avoid privacy and security issues. For what concerns the data shared by STRUKTON (which is an IN2SMART partner) and IN2DREAMS, refer to the COLA signed between IN2SMART and IN2DREAMS for the details about data exchange policies.

4.5.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Specific-Scenario 3 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  The ability to make the railway infrastructure aware of the time needed for its restoration can surely impact its costs and its capacity. Planning maintenance based on the repair time and rescheduling the trains based on the forecasted repair time can consistently improve the quality of the service provided by the infrastructure by also reducing the costs due to the penalties that the infrastructure has to pay based on unplanned or badly handled disruptions. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.

- Impacts on the IP2: Advanced Traffic Management and Control Systems
  This scenario also impacts the IP2 because we can make the TMS aware of the restoration time and then improve the ability of the dispatchers to handle maintenances or repairs. Consequently TD2.9 is surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs we can state that the Specific-Scenario 2 can have an impact into the following scopes of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario is a good use case where to apply the ideas of predictive models which require to be interpreted by an operator. In fact an interpretable model can further improve the ability of the operators to understand the processes of maintenance and optimize them based on the results of these predictions. An interpretable model, for example, can suggest that in particular days of the week or with particular weather conditions it is better to not plan maintenance. Moreover interpretable models can easily adapt to take into account the experience of the operators or the previous knowledge about the problem.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  Finding metrics able to understand when our models work well and when they do not is essential to make these data-driven models effective in real would situations. For these reasons, together with RFI and STRUKTON, we will develop specific metrics and KPIs able to detect in what conditions the data driven models performs well (e.g. a subset of the possible maintenance or only in planned maintenance) and when it is better to exploit the experience of the operators (e.g. in the maintenance that are not very frequent and then not enough historical records are present). The metrics developed in this scenario will need to help the operators to better control and understand the restoration time phenomenon.

4.5.9 Analytics & Metrics

This scenario is the typical field of application of the predictive analytics. Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts in order to make predictions about future or otherwise unknown events [60, 151]. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown, whether it is in the past, present or future. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions. In future industrial systems, the value of predictive analytics is to predict and prevent potential issues to achieve near-zero break-down or to predict the time needed to restore from an issue [186, 188].
In our specific case the purpose of the scenario is to try to predict the time needed for restoring the railway network form a fault or from a planned maintenance. The capacity of predicting this time has a dramatic impact in the ability to restore from a disruption [53, 75, 99, 114, 130]. Our specific target will be to develop white box models (rule or decision tree based [70, 129, 178, 225]) for the purpose to build an interpretable model of the phenomena that can be easily modified to take into account the experience of the operators of STRUKTON and RFI or the background knowledge of the problem in a gray-box fashion [45, 46, 159, 161]. Moreover, together with STRUKTON and RFI, we will develop specific KPIs and metrics for testing the quality of the derived models based on the way in which STRUKTON and RFI operate.

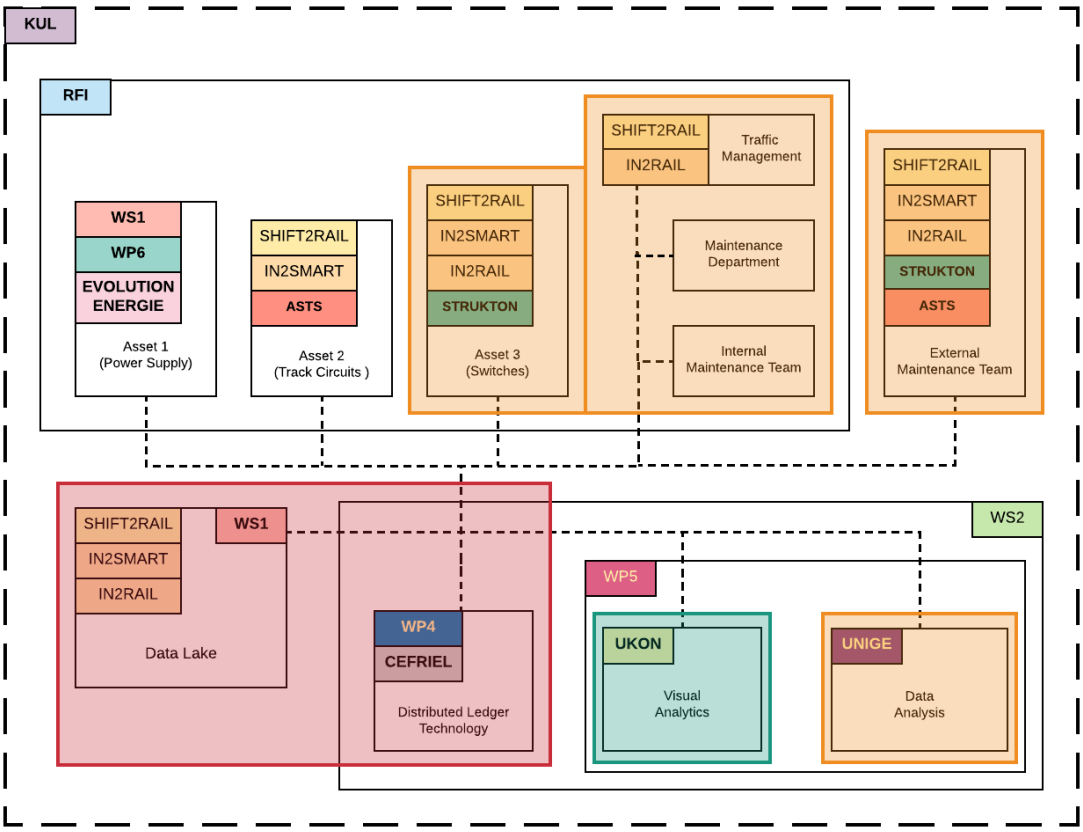

Figure 17: Cross-Scenario 4 Reference Picture (Switches)

4.6 Specific-Scenario 4: Switches

4.6.1 Summary

This scenario deals with the problem of building an interpretable and reliable data-driven condition based maintenance system for the train switches, a mechanical installation enabling railway trains to be guided from one track to another, such as at a railway junction or where a spur or siding branches off.

4.6.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
UNIGE	IN2DREAMS	WS2	WP5	Responsible Partner	No
STRUKTON	IN2SMART	-	WP8	Advisor	Prospected
DLR	IN2SMART	-	WP8	Advisor	Prospected

4.6.3 Connection with other Scenarios

Specific-Scenario 4, as its name states, is a specific-scenario and for this reason it is mainly connected with the cross-scenarios.

- Connections with Cross-Scenario 1 (Section 4.1)
  Specific-Scenario 4 will provide to Cross-Scenario 1 other useful information to visualize to the operators, principally the maintainers. In particular the possibility to understand the switches conditions and consequently the possibility to maintain them before breaks is fundamental for limiting disruptions. Moreover, together with the prediction itself, Cross-Scenario 1 has to show also the quality of the prediction and the reason of the prediction, thanks to the interpretable models developed in Specific-Scenario 4, with a nice and intuitive graphical representation that can help the operator in deciding to or not to rely on the prediction on its experience.
- Connections with Cross-Scenario 2 (Section 4.2)
  The connection of Cross-Scenario 2 is rather straightforward, in fact the models and the prediction of the models developed with the data provided by STRUKTON can be sold on the data marketplace as the data itself. Moreover the data marketplace can be a source of additional data that can be exploited for improving the quality of the model. The results of Specific-Scenario 4 can give a hint on what could be a new source of information that could help improving the quality of the models.
- Connections with Specific-Scenarios 1, 2, 3, and 5 (Sections 4.3, 4.4, 4.5, and 4.7)
  The connection of Specific-Scenario 4 with Specific-Scenarios 1, 2, 3, and 5 is quite tight. Specific-Scenarios 1 and 5 deal with the problem of predicting a possible malfunction. Specific-Scenario 3 deals with the problem of predicting the restoration time from a maintenance or a malfunction. Specific-Scenarios 1 deals with the train delay prediction. Then the scope of the different scenarios is complementary. We envision a future where predictive models of a malfunction will be exploited together with the one which forecasts the restoration time of an asset and the train delays in a way to optimize the train circulation for improving the cost efficiency, the reliability, and the capacity of the infrastructure.

4.6.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

The connection of Specific-Scenario 4 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection here is well depicted in Figure 17, in fact the data that we will use in this scenario are stored in the data lakes of STRUKTON. Then these data can be shared in the marketplace envisioned in the IN2DREAMS WS2 WP4 in order, from one side, to monetize this information by selling it to other actors of the railway ecosystem and, from the other side, these data could be enriched by buying some other detailed information from other actors (e.g. exact usage of the switch with data of the TOs about tons) in order to improve the quality of the prediction. Moreover, the models and the predictions themselves can be sold to other actors as a service.
  There is also a strong connection between this scenario and the WP4 selected scenario on Asset Maintenance (D4.1 Section 5.1). In fact the operators need to be aware of the status of the maintenances and this information can be extracted from the blockchain developed in WP4 and vice-versa the blockchain can be fed with the output of the data driven models developed in WP5.
- Connections with IN2SMART and other SHIFT2RAIL projects
  This use case has been designed inside WP8 Dynamic Railway Information Management System Data Mining and Predictive Analytics in the framework of the IN2SMART as a continuation of the work started

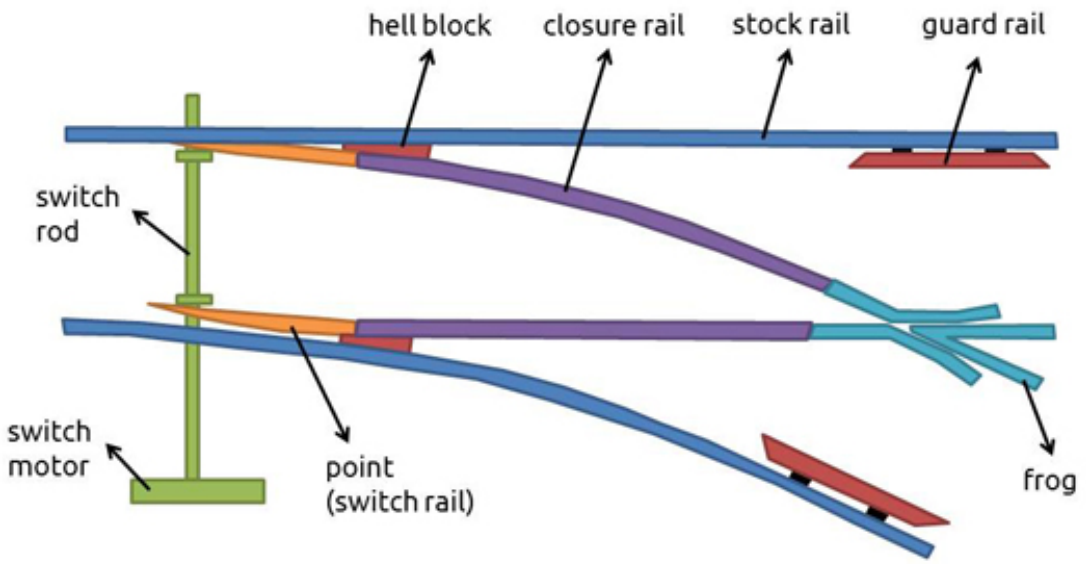

Figure 18: Switch Components

in IN2RAIL WP9. In fact, STRUKTON, one of the data providers of this scenario, belongs to the IN2SMART project and also to IN2RAIL. This scenario in an extension of the scenario started in IN2RAIL and our scenario builds upon these results, lesson learned, and suggestions coming from IN2RAIL.

4.6.5 Scenario Objective(s)

The main objectives of this scenario are the following ones.

- Develop interpretable models of the switches failure predictability with the data provided by STRUKTON able to detect the drift in the behavior of the switches by taking into account heterogeneous sources of information regarding the status, the maintenances, and the operating conditions of the switches. An interpretable model is essential in this case for understanding the biases in the data and reduce the false alarms.
- Develop metrics and KPIs able to determine the situation in which our models can be effectively applied in real operations and when we have to leave the choice to the operators because the developed models are not reliable enough because of the quality or the quantity of the historical data.

4.6.6 Scenario Description

A railroad switch, turnout, or (set of) points is a mechanical installation enabling railway trains to be guided from one track to another, such as at a railway junction or where a spur or siding branches off. Switches are part of the infrastructure of a railway network, and they can be classified as specialized Track Equipment. Railroad switches can be of many types and are composed of many sub-components. The main components of a switch, depicted in Figure 18, are:

- Points (switch rails or point blades) are the movable rails which guide the wheels towards either the straight or the diverging track. They are tapered on most switches, but on stub switches they have square ends.

- Stock rails are the running rails immediately alongside of the switch rails against which the switch rails lay when in closed position. The stock rails are otherwise ordinary rails that are machined, drilled, and bent as required to suit the design of the turnout switch and the individual switch point rails.
- Frog is a component placed where one rail crosses another, and refers to the crossing point of two rails.
- Closure rails are the straight or curved rails that are positioned in between the heel of switch and the toe of frog.
- Guard rail (check rail) is a short piece of rail placed alongside the main (stock) rail opposite the frog. These exist to ensure that the wheels follow the appropriate flangeway through the frog and that the train does not derail.
- Heel block assemblies are units placed at the heel of the switch that provide a splice with the contiguous closure rail and a location for the switch point rail to pivot at a fixed spread distance from the stock rail.
- Switch point rail stops act as spacers between the switch point rail and the stock rail. Stops laterally support the switch point from flexing laterally under a lateral wheel load and thereby possibly exposing the open end of switch point rail to head-on contact from the next wheel.
- A switch operating device moves switch rails. Switch rails can be thrown (moved) from one orientation to another by either a hand-operated (manual) switch stand or a mechanically or electro-mechanically (power-operated) switch machine. In both cases, the operating devices are positioned at the beginning of the turnout opposite the switch-connecting rods near the point of the switch rails.

Switches have a very important role in the context of the railways infrastructures. In fact, the impairment of the latter may potentially lead to a variety of problems, including infrastructure service disruptions and delays on the trains. TMS should be aware not just of whenever a switch is available or not but also when it will be not available in the future in order to provide this information to maintainers and dispatchers. For this reason, TMS has to take advantage of the predictive and interpretable models of the future status of the switches and the accuracy related to these data-driven models developed by this scenario. Indeed, thanks to these tools, scheduling and maintenance interventions will be performed in a proper way, in order to improve the asset condition status before it eventually fails. The necessity for this research has the same origins as the research done on asset status as described in IN2RAIL D9.3 by DLR and STRUKTON, which is also partly described in this section. The information regarding the power consumption of the switch engine during the movements of the switch blades will be provided by STRUKTON. STRUKTON already analyzed this data in IN2RAIL and shown that, inside those data, the asset status information is actually presents. A large variety of behavior conditions and types of malfunctions of the switches has been highlighted by the significant variances reported in the results related to the functional states. Using graphs which describe the behavior of the power consumption combined with the knowledge of experienced engineers allowed to visually identified and distinguished between normal and abnormal behaviors. Obviously, the analysis of several hundreds of switches is not a task than can be easily performed by experts. Besides, minor and systematic changes over time are not often detected by a human expert. In this context, advanced data analytic techniques may help to automatize the process, detect drifts and predict the future status of the switches. While recognizing failing movements or a not functioning switch may be simply performed based on single criterion thresholds, the real challenge is represented by the distinction between normal behavior (switch in good condition) and abnormal behavior (switch in degradation) and the prediction of future behavior if the switch is still functioning (i.e. switch is locked at the end of the switch movement). In fact, maintenance activities and weather influences, together with other factors have a big impact on switches behavior and quantify this relation represents a potential challenge in determining anomalies or drift in the switch engine power consumption. These factors may strongly influence the power consumption while not indicating a negative change of switch condition. Consequently, single threshold criteria such comparing an area under a curve with the reference one which represent a useful and simple example of strategy implemented to analyze the time series of the current graph in an automatic way, embeds the difficulties of determining the reference graph and the risk to not

understand what is really happening to the switch. Moreover, an operator reporting the malfunction of a switch usually fills a database that will result sparse, thus creating a difficulty in the switch status prediction based on the power consumption. The same may happen with the corresponding failure data obtained from the maintenance team. Since the operators typically report only non-functioning switches, the functioning switches are not reported and not consequently labelled for a successful supervised learning approach for forecasting purposes. This research therefore also involves the determination of the correct labels, even if no incident is reported (e.g. a stone which blocks the movement).

4.6.7 Available Data & Data Access Policies

For the implementation of this scenario STRUKTON made plans to make available a series of data sources:

- Switch Monitoring data: they are presented in STRUKTON databases, which includes data from many (thousands) switches related to many years. This monitoring system has its own thresholds, managed by maintenance engineers or switch experts, which are also an input for the analysis. Not all switches are maintained by STRUKTON so we will not have access to all these switches (probably data about hundred switches will be available).
- Recorded Incidents: historical datasets regarding the recorded failures on the same switches/points in the same timeframe as the monitoring data. Because of win/loosing contracts, these records may not be complete for the whole history.
- Maintenance actions data: historical datasets regarding the maintenance activities executed in the same timeframe on the same switches. Because of win/loosing contracts, these records may not be complete for the whole history.
- Usage/Load data: an additional dataset, the load data is generated and provided by ProRail, the Asset Manager. This data is mainly related to the usage conditions of the assets, and includes information about the number of trains that passed over the switch, their weight, etc. Unfortunately, this data is aggregated at year level and in few cases at month level.
- Weather condition: data retrieved from the Royal Netherlands Meteorological Institute (KNMI). Unfortunately, the weather stations can be pretty far from the location of the switches so the conditions recorded may not be well represented for a switch. Local weather in Netherlands can strongly deviate.

The details of the data format and fields are not yet available due to the fact that STRUKTON is still investigating internally if to go forward in this direction or not.
Data from STRUKTON, in case STRUKTON will decide to go forward with this scenario, will be provided anonymously, meaning that no specific asset name will be provided regarding the operational service. All the references to the real data, as object/asset names will be anonymized in order to avoid privacy and security issues. We refer to the COLA signed between IN2SMART and IN2DREAMS for the details about data exchange policies.

4.6.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Specific-Scenario 4 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  The ability to make the railway infrastructure aware of one of its essential component, the switches, surely impacts its costs and its capacity. Planning maintenance based on this information can consistently improve the quality of the service provided by the infrastructure by also reducing the costs due to disruptions. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs, we can state that the Specific-Scenario 1 can have an impact into the following scopes of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario is a good use case where to apply the ideas of predictive models which require to be interpreted by an operator. In fact an interpretable model can further improve the ability of the operators to understand the processes of maintenance and optimize them based on the results of these predictions. Moreover interpretable models can easily adapt to take into account the experience of the operators or the previous knowledge about the problem.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  Finding metrics able to understand when our models work well and when they do not is essential to make these data-driven models effective in real would situations. For these reasons, together with STRUKTON, we will develop specific metrics and KPIs able to detect in what conditions the data driven models perform well and when it is better to exploit the experience of the operators. The metrics developed in this scenario will need to help the operators to better control and understand the switches decay phenomena and behaviours.

4.6.9 Analytics & Metrics

Most of the rail infrastructure managers will indicate switches as very critical assets for their operation, because whenever the availability of a switch is compromised, it introduces numerous problems leading to unavailability of the train path and resulting in train delays, significant disturbances in the operation, increased fuel costs, crew expenses, maintenance and repair costs, and generally in a negative impact on reputation and revenues. In order to prevent this kind of undesirable situation and events, seeing the problems developing and being able to anticipate before they get critical would provide significant benefits for train operations in order to prevent problems before they occur. In case a maintenance crew is engaged in order to fix a growing problem in a switch, more information about the possible problem is necessary in order to successfully perform right maintenance and to fix the real problem or the cause of it.
Many approaches are available in literature for the purpose of predicting the switch failure predictability [73, 109, 126, 149, 171, 189, 195]. These approaches can be basically divided in two families based on the operating conditions:

- A supervised learning strategy based on black-box algorithms is adopted when high quality training data sets are available. This operating condition is quite rare since real data are often quite messy and, especially for switch monitoring based on the power consumption, the available databases are sparse for reported incidents (an operator is reporting the malfunction of a switch) and corresponding failure data (the switch was checked by a maintenance). Moreover black-box algorithms are often not used in real world railway conditions.
- An unsupervised approach is adopted when there is a lack of an extensive database of reported incidents, as discussed before, to capture anomalies which are not captured by the reported incidents and to identify the most relevant characteristics of the current graph for analysis.

In our work we will try to use a combination of unsupervised techniques, for cleaning the data, and white-box based supervised techniques for predicting the failure predictability of the switches which can be safely applied because of their interpretability. Moreover we will develop metrics and KPIs able to determine the situation in which our models (and the black box models already developed in STRUKTON) can be effectively applied in real operations and when we have to leave the choice to the operators because the developed models are not reliable enough because of the quality or the quantity of the historical data.

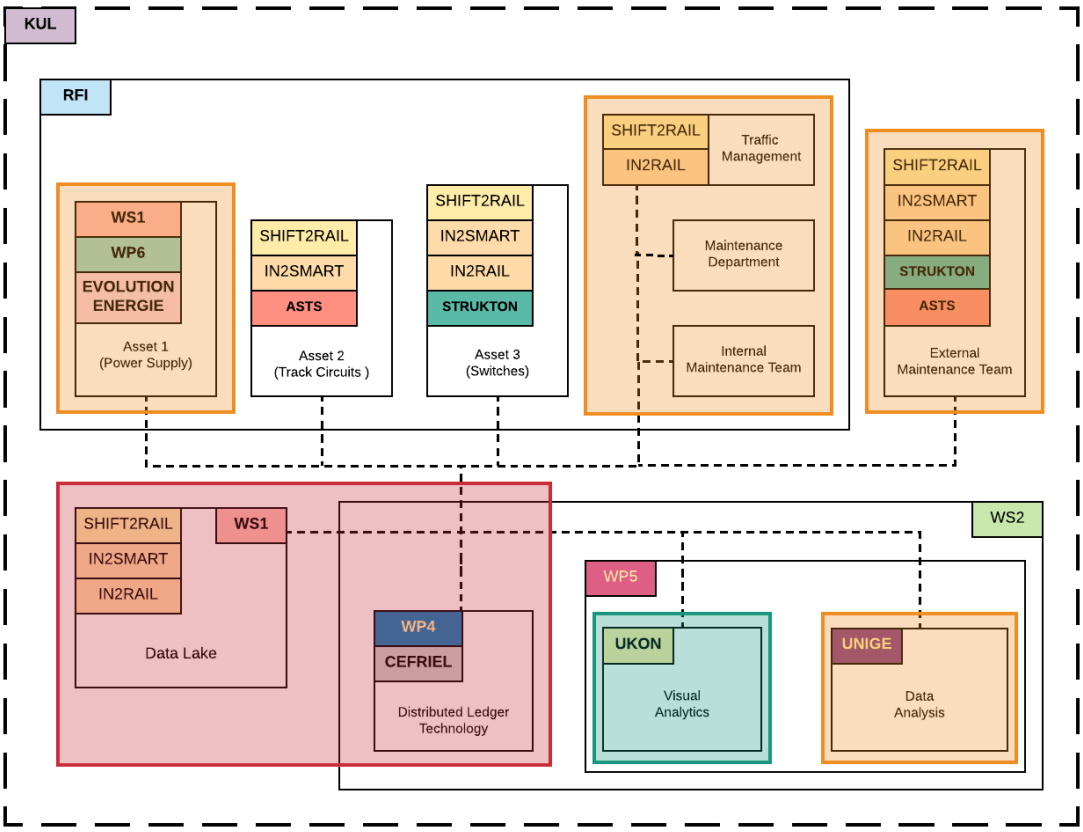

Figure 19: Specific-Scenario 5 Reference Picture (Train Energy Consumption)

4.7 Specific-Scenario 5: Train Energy Consumption

4.7.1 Summary

This specific scenario deals with another asset of train operation: power supply of trains. More specifically, it deals with the measurement, simulation and modelling of data and forecasts related to energy consumption that is done in WP6 of the WS1 of the IN2DREAMS project. The evaluation of those models based on commonly defined KPIs is done in WP5, as well as the possibility to develop a white-box model for energy consumption forecasting based on the available data.

4.7.2 Responsible partner(s)

The following table lists the partners involved in the scenario, their connection with other projects, and their roles.

Partner	SHIFT2RAIL	WS	WP	Role	Data Provider
EE	IN2DREAMS	WS2/WS1	WP5/WP6	Responsible Partner	Yes
UNIGE	IN2DREAMS	WS2	WP5	Advisor	No
ISKRATEL	IN2DREAMS	WS1	WP6	Advisor	No

4.7.3 Connection with other Scenarios

Specific-Scenario 5, as its name states, is a specific-scenario and for this reason it is mainly connected with the cross-scenarios.

- Connections with Cross-Scenario 1 (Section 4.1)
  Specific-Scenario 5 will provide to Cross-Scenario 1 other useful information to visualize to the operators, mainly to the supervisors and the maintainers, but also to the global infrastructure manager or the railway operator to see energy consumption. In particular, it will give them the possibility to understand the link between the energy consumption and other parameters such as weather or traffic conditions, so there can be a choice to operate the lines in the least energy-consuming way. Moreover, together with the prediction itself, Cross-Scenario 1 has to show also the quality of the prediction, based on the KPIs evaluated in Specific-Scenario 5, with a clear and intuitive graphical representation to help the operator in deciding to or not to rely on the prediction on its experience.
- Connections with Cross-Scenario 2 (Section 4.2)
  The link between this specific scenario and Cross-Scenario 2 is rather clear, since the raw data acquired, the models and also the results and predictions of the models can be sold on the data marketplace. Moreover the data marketplace can be a source of additional data that can be exploited for improving the quality of the model. The results of Specific-Scenario 5 can give a hint on what could be a new source of information that could help improving the quality of the models.
- Connections with Specific-Scenarios 1, 2, 3, and 4 (Sections 4.3, 4.4, 4.5, and 4.6)
  The connection of Specific-Scenario 5 with Specific-Scenarios 1, 2, 3, and 4 is quite tight. Specific-Scenarios 1 and 4 deal with the problem of predicting a possible malfunction. Specific-Scenario 3 deals with the problem of predicting the restoration time from a maintenance or a malfunction. Specific-Scenarios 2 deals with the train delay prediction. Then the scope of the different scenarios is complementary. We envision a future where predictive models of a malfunction will be exploited together with the one which forecasts the restoration time of an asset and the train delays in a way to optimize the train circulation for improving the cost efficiency, the reliability, and the capacity of the infrastructure.

4.7.4 Connection with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects

The connection of Specific-Scenario 5 with other IN2DREAMS WSs and WPs and SHIFT2RAIL Projects can be summarized as follows.

- Connections with IN2DREAMS WS2 WP4
  The connection here is well depicted in Figure 19: in fact the data that we will use in this scenario are stored in the data lakes developed and set up in the scope of the IN2RAIL project. Then these data can be shared in the marketplace envisioned in the IN2DREAMS WS2 WP4 in order, from one side, to monetize this information by selling it to other actors of the railway ecosystem and, from the other side, these data could be enriched by buying some other detailed information from other actors (e.g. global energy consumption with data from the energy supplier) in order to improve the quality of the prediction. Moreover, the models and the predictions themselves can be sold to other actors as a service.
- Connections with IN2DREAMS WS1 WP6
  The connection with WS1, and particularly with WP6 of the IN2DREAMS project is strong, since the input of this scenario is the result of the energy consumption prediction model developed in this WP. In addition, the definition of the KPIs and methodology to evaluate the models and their results will be done in collaboration between WP5 (WS2) and WP6 (WS1).

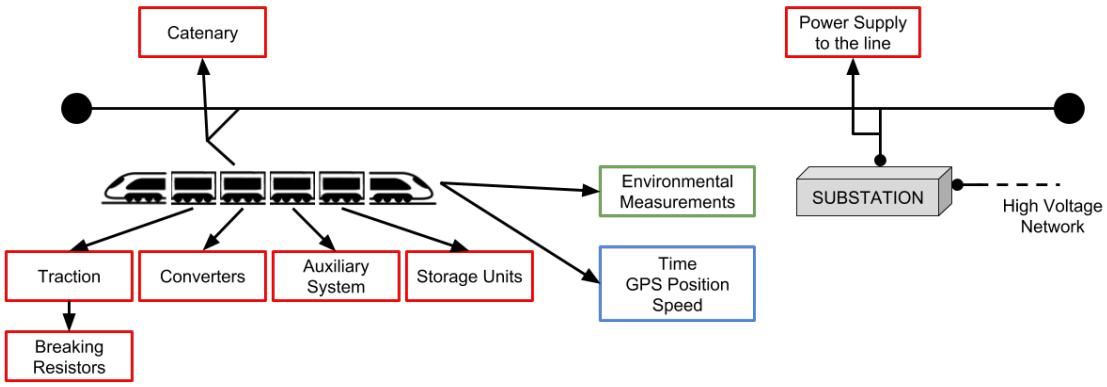

Figure 20: Schematic representation of the monitoring of energy flows

- Connections with IN2RAIL and other SHIFT2RAIL projects
  The use case from which the data has been acquired is the IN2RAIL project, from their experimentation on the Reims urban lines. The connection to this project is also important, because the data already accessible is hosted in the ODM platform developed in the scope of the IN2RAIL project. The present scenario in an extension of the scenario and use case started in IN2RAIL, and it builds upon these results, lesson learned, and suggestions coming from IN2RAIL.

4.7.5 Scenario Objective(s)

The main objectives of this scenario are the following ones:

- Since the focus of WP5 is model evaluation, one of the objective is to develop metrics and KPIs in collaboration with IN2DREAMS WS1 WP6 in order to evaluate the energy consumption forecasting models developed in WP6, and test their quality. These models are built based on the data obtained through the IN2RAIL project, but also from the data management platform developed in WP3 and through the communication platform developed in WP2 in IN2DREAMS project.
- The other focus of WP5 is the development of interpretable models, so if possible, based on the same available data from the use case, another objective of this scenario would be the development of a white-box type model.

4.7.6 Scenario Description

Because power supply is part of the assets necessary for the whole infrastructure to operate, it is important to understand the parameters that have an influence on energy consumption in order to prevent failure and other events impacting negatively the behavior of the trains or other assets. Besides, due to the global concern in a better and optimized energy consumption to reduce the environmental impact, also the management of energy is an aspect to take into consideration.
The scenario of energy consumption in trains for this WP is based on the use case developed in the scope of the IN2RAIL project. Its objective was to support sustainable development of railway infrastructure, by monitoring and analyzing energy flows for the energy system, the stations and the rolling stock. Therefore, sensors have been installed and measure data from the Reims urban lines with data coming from both a train and a power substation. Measurements and data from the same use case are also going to be part of the IN2DREAMS WP2 task, which deals with the communication platform providing data for the use cases defined into the IN2DREAMS project. Using the data available from both these elements as shown in Figure

20 (taken from the deliverable D6.1 of the WP6 of the IN2DREAMS project), WS1 WP6 of the IN2DREAMS project will develop two kinds of prediction models:

- Forecasting energy demand on the train: the models developed for the prediction of on-train energy load are based on data analytics of the information acquired inside the train, including features such as geolocation. The level of modeling will be from one to few minutes time horizon.
- Forecasting energy demand of the substation: the models developed for the prediction of substation charging are based on data analytics of the information acquired from a substation, considered as a stationary node. The level of modeling will be of one hour to one day time horizon.

Those models and their results will be evaluated by a set of KPIs defined by IN2DREAMS WS1 WP6 and WS2 WP5 of the same project. The conclusions of the reliability study will be exposed by WP5 in the deliverables to come. In addition, there is a possibility for WP5 to develop an interpretable model of the power consumption based on the same set of data. This opportunity, integrated into cross scenario 1 about data visualization, would allow the operators to better understand the parameters influencing the train and substation energy consumption, and then guide their decisions toward a more energy-efficient management of power supply. Indeed, it would allow the operator to make decisions based on the parameters that have the greatest influence, or on the choices made previously based on a similar set of conditions (day and time of the week, weather or traffic condition, etc). Also, by adding this set of data into the cross scenario 2, the data market would be enriched, and some insight would be gained by being able to cross many data from various devices and sources (described for example in the other specific scenarios). It could also help managing common maintenance activities for several assets for both the rolling stock and the infrastructure.

4.7.7 Available Data & Data Access Policies

From the IN2RAIL project, a set of data is already available:

- 3 years of measurement of data from one train (moving object) with different kind of data acquired at a frequency of 1 second (table from deliverable D6.2 of the WP6 of the IN2DREAMS project):

Quantity	Channel	Unit
Complete time and date		seconds
Longitude position of the train	LONGITUDE	degrees
Latitude position of the train	LATITUDE	degrees
Train speed	SPEED	km/s
Overhead line or ground power supply	U_CTPP	Boolean
Total supplied voltage	U_CAT	V
Total measured current	I_CAT	A
Current from two traction units	I_TCU1, I_TCU2	A
Current from auxiliary converter	I_CVS	A
3-phase voltage of air conditioning (HVAC)	U_HVAC_AC	V
3-phase current from two HVAC sensors	I_HVAC_AC1, I_HVAC_AC2	A
Rheostat currents for 4 resistors	I_RHEO_11, I_RHEO_12, I_RHEO_21, I_RHEO_22	A
External temperature	TA_EXT	°C
Internal temperature	TA_INT	°C
CO2 level inside the train	CO2	ppm

- 4 months of measurement of data from one substation (stationary node) with different kind of data acquired at a frequency of 1 second:

Quantity	Unit
Busbar voltage	V
Injection current to the depot	A
Injection current to the single line	A
Injection current to the double line	A

Some of the data is available as a CSV file containing non filtered / processed data (mainly for historical data), and also on the ODM platform (for real-time data) developed in the scope of the IN2RAIL project and further for the IN2DREAMS project in order to be accessible in the Analytics platform. If weather data should be needed as a parameter for the model, an external provider will be used.

4.7.8 Impacts on the SHIFT2RAIL and IN2DREAMS WS2 WP5 KPIs

With respect to the specific SHIFT2RAIL KPIs we can state that the Specific-Scenario 5 can have an impact into the following IP.

- Impacts on the IP3: cost-efficient and reliable high-capacity Infrastructure
  The ability to make the railway infrastructure aware of one of its essential asset, power supply and energy consumption, surely impacts its costs, its capacity and its reliability. Planning maintenance based on this information can consistently improve the quality of the service provided by the infrastructure by also reducing the costs due to disruptions. In addition, being able to manage energy by knowing the parameters impacting the consumption can help improving energy efficiency. Consequently TD3.6, TD3.7, and TD3.8 are surely impacted by this scenario.

Furthermore, with respect to the specific IN2DREAMS WS2 WP5 KPIs, we can state that the Specific-Scenario 5 can have an impact into the following scopes of the project.

- Impacts on the study and development of interpretable data-driven models
  This scenario is a good use case where to apply the ideas of predictive models which require to be interpreted by an operator. In fact an interpretable model can further improve the ability of the operators to understand the processes of energy consumption and optimize them based on the results of these predictions. Moreover interpretable models can easily adapt to take into account the experience of the operators or the previous knowledge about the problem.
- Impacts of the study and development of railway specific metrics to validate the data driven models
  Finding metrics able to understand when our models work well and when they do not is essential to make these data-driven models effective in real world situations. For these reasons, together with ISKRATEL from WS1 WP6 of the IN2DREAMS project, we will develop specific metrics and KPIs able to detect in what conditions the data driven models perform well and when it is better to exploit the experience of the operators. The metrics developed in this scenario will help the operators to better control and understand the energy consumption of the urban lines.

4.7.9 Analytics & Metrics

Energy is an important asset for the railway management system. Besides, the optimization of energy consumption for trains can have positive impacts on both the cost and the infrastructure and operation management. That is the reason why many studies have tried several methods such as simulated annealing or other stochastic iteration algorithms, neural networks or genetic algorithms to forecast electricity load or optimize energy consumption taking into account the main constraints of train operation (traffic and schedule demands or operational restrictions for instance) [7, 8, 30, 110, 142, 187].

Regarding the forecasting models that will be developed in WP6 of the IN2DREAMS project, several mathematical approaches of machine learning have been envisioned as described in D6.1: linear or support vector regression, K nearest neighbours (k-NN), random forest and incremental decision tree, or also stay point detection. These methods and algorithms, used with the Qminer tool, for the forecast of energy consumption can be applied to the railway ecosystem by taking into account its specificities. Besides, if those analyses are applied to the constant data flow coming from both the static parts (substations) and the rolling stock (moving train), it can yield to accurate forecasting results than can be compared thanks to relevant KPIs for model evaluation. The knowledge gained from energy consumption forecast by black-box models, which results are evaluated by relevant KPIs, can help the operators to make the best decisions for an optimized power supply and reduce the risk of failure, as well as a better management of costs induced by energy consumption. The KPIs will be defined in common between the teams of WP5 and WP6 of the IN2DREAMS project, considering the results of the models developed by the IN2DREAMS WS1 WP6.
Besides, the possibility of the development of a white-box model to predict energy consumption can allow the operator to have a better understanding of the parameters influencing the behavior under given circumstances, and also guide them toward the best decisions. This option is considered by WP5, based on the available data from IN2RAIL but also from the WP2 of the IN2DREAMS project.

5 Conclusions

The general objective of WP5 is to study, design and develop data and visual analytics solutions for knowledge extraction from railway asset data. For this reason the cornerstone of the WP5 is Task 5.1, which is in charge of developing scenarios in order to accomplish the main WP5 objectives. This deliverable focused on relevant railway assets whose malfunction and maintenance policies have an impact on the KPIs targeted by the SHIFT2RAIL program. The scenarios have been clearly defined in terms of responsibility, data availability, analytic perspectives, and goals for data analytics algorithms and metrics. Seven scenarios have been presented. Two of them are cross-scenarios in the sense that they cover, in some way, many aspects of the railway ecosystem while five of them are specific-scenarios in the sense that they focus on a single particular aspect. Some scenarios will be used in the next WP5 tasks as a basis for the development of the WP5 POC and demonstrators. This task has been lead by RFI, the Italian Infrastructure Manager, in order to share its view of direct railway stakeholders with the other partners of this WP. Cooperation with IN2DREAMS WS1 WP6 is assured by partner EVOLUTION ENERGIE and the coordination with other SHIFT2RAIL recipients is assured by the collaboration with IN2SMART and IN2RAIL.

References

[1] A. Abraham, B. Nath, and P. K. Mahanti. Hybrid intelligent systems for stock market analysis. In *International Conference on Computational Science*, 2001.

[2] Y. S. Abu-Mostafa and A. F. Atiya. Introduction to financial forecasting. *Applied Intelligence*, 6(3):205–213, 1996.

[3] C. C. Aggarwal and C. K. Reddy. *Data clustering: algorithms and applications*. CRC press, 2013.

[4] C. D. Aggarwal. *Data Mining - The textbook*. Springer, 2015.

[5] R. Agrawal and R. Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, 1995.

[6] S. Agrawal and J. Agrawal. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713, 2015.

[7] T. Albrecht. Reducing power peaks and energy consumption in rail transit systems by simultaneous train running time control. *WIT Transactions on State of the Art in Science and Engineering*, 39:3–12, 2010.

[8] H. K. Alfares and M. Nazeeruddin. Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34, 2002.

[9] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1390–1406, 2012.

[10] M. A. Arbib. *The handbook of brain theory and neural networks*. MIT press, 2003.

[11] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in neural information processing systems*, 2007.

[12] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[13] N. Attoh-Okine. Big data challenges in railway engineering. In *IEEE International Conference on Big Data*, 2014.

[14] C. Aytekin, Y. Rezaeitabar, S. Dogru, and I. Ulusoy. Railway fastener inspection by real-time machine vision. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(7):1101–1107, 2015.

[15] Y. Bai, T. K. Ho, B. Mao, Y. Ding, and S. Chen. Energy-efficient locomotive operation for chinese mainline railways by fuzzy predictive control. *IEEE Transactions on Intelligent Transportation Systems*, 15(3):938–948, 2014.

[16] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[17] J. Barta, A. E. Rizzoli, M. Salani, and L. M. Gambardella. Statistical modelling of delays in a rail freight transportation network. In *Proceedings of the Winter Simulation Conference*, 2012.

[18] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

[19] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[20] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[21] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[22] A. Berger, A. Gebhardt, M. Müller-Hannemann, and M. Ostrowski. Stochastic delay prediction in large train networks. In *OASIcs-OpenAccess Series in Informatics*, 2011.

[23] A. Berger, A. Gebhardt, M. Müller-Hannemann, and M. Ostrowski. Stochastic delay prediction in large train networks. In *OASIcs-OpenAccess Series in Informatics*, 2011.

[24] P. Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, 2006.

[25] M. J. Berry and G. Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[26] M. R. Berthold, C. Borgelt, F. Hoppner, and F. Klawonn. *Guide to Intelligent Data Analysis*. Springer, 2010.

[27] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[28] Z. Bin and X. Wensheng. An improved algorithm for high speed train's maintenance data mining based on mapreduce. In *International Conference on Cloud Computing and Big Data*, 2015.

[29] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[30] Y. V. Bocharnikov, A. M. Tobias, and C. Roberts. Reduction of train and net energy consumption using genetic algorithms for trajectory optimisation. In *IET Conference on Railway Traction Systems*, 2010.

[31] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[32] S. A. Branishtov, Y. A. Vershinin, D. A. Tumchenok, and A. M. Shirvanyan. Graph methods for estimation of railway capacity. In *IEEE International Conference on Intelligent Transportation Systems*, 2014.

[33] R. Brath. Metrics for effective information visualization. In *IEEE Symposium on Information Visualization*, 1997.

[34] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[35] J. Buchmüller, H. Janetzko, G. Andrienko, N. Andrienko, G. Fuchs, and D. A. Keim. Visual Analytics for Exploring Local Impact of Air Traffic. In *Eurographics Conference on Visualization (EuroVis 2015)*, 2015.

[36] S. Buckley and D. Lightman. Ready or not, big data is coming to a city (transportation agency) near you. In *Annual Meeting Transportation Research Board*, 2015.

[37] P. Bühlmann. *Bagging, boosting and ensemble methods*. Springer, 2012.

[38] S. Chakraborty. *Bayesian machine learning*. University of Florida, 2005.

[39] A. A. Chandio, N. Tziritas, and C. Z. Xu. Big-data processing techniques and their challenges in transport domain. *ZTE Communications*, 1(10), 2015.

[40] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys*, 41(3):15, 2009.

[41] P. C. Chang, C. Y. Fan, and J. L. Lin. Trend discovery in financial time series data using a case based fuzzy decision tree. *Expert Systems with Applications*, 38(5):6070–6080, 2011.

[42] Y. W. Chang and C. J. Lin. Feature ranking using linear svm. In *Causation and Prediction Challenge*, 2008.

[43] J. Chen, C. Roberts, and P. Weston. Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems. *Control Engineering Practice*, 16(5):585–596, 2008.

[44] M. Chen, J. Heinrich, J. Kennedy, A. Kerren, F. Schreiber, S. Simon, C. Stolte, C. Vehlow, M. Westenberg, and B. Wong. Uncertainty Visualization. In *Biological Data Visualization (Dagstuhl Seminar 12372)*, 2013.

[45] A. Coraddu, L. Oneto, F. Baldi, and D. Anguita. Ship efficiency forecast based on sensors data collection: Improving numerical models through data analytics. In *OCEANS 2015-Genova*, 2015.

[46] A. Coraddu, L. Oneto, F. Baldi, and D. Anguita. Vessels fuel consumption forecast and trim optimisation: A data analytics perspective. *Ocean Engineering*, 130:351–370, 2017.

[47] F. Cores, N. Caceres, F. G. Benitez, S. Escriba, and N. Jimenez-Redondo. A logical framework and integrated architecture for the rail maintenance automation. In *European Transport Conference 2013Association for European Transport*, 2013.

[48] C. D. Correa, Y. H. Chan, and K. L. Ma. A framework for uncertainty-aware visual analytics. In *Visual Analytics Science and Technology*, 2009.

[49] C. D. Cottrill and S. Derrible. Leveraging big data for the development of transport sustainability indicators. *Journal of Urban Technology*, 22(1):45–64, 2015.

[50] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[51] P. Cunningham. Dimension reduction. In *Machine learning techniques for multimedia*, 2008.

[52] W. Daelemans and A. Van den Bosch. *Memory-based language processing*. Cambridge University Press, 2005.

[53] T. Darmanin, C. Lim, and H. Gan. Public railway disruption recovery planning: a new recovery strategy for metro train melbourne. In *Asia Pacific Industrial Engineering and Management Systems Conference*, 2010.

[54] E. De Romph. Using big data in transport modelling. In *Data & Modelling Magazine*, 2013.

[55] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[56] D. Delen, C. Kuzey, and A. Uyar. Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, 40(10):3970–3983, 2013.

[57] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. In *arXiv preprint arXiv:1702.08608*, 2017.

[58] W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik. Comparison of feature ranking methods based on information entropy. In *IEEE International Joint Conference on Neural Networks*, 2004.

[59] M. H. Dunham. *Data mining: Introductory and advanced topics*. Pearson Education India, 2006.

[60] W. Eckerson. Extending the value of your data warehousing investment. In *The Data Warehouse Institute*, 2007.

[61] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[62] C. K. Enders. *Applied Missing Data Analysis*. The Guilford Press, 2010.

[63] M. Faizrahnemoon, A. Schlote, L. Maggi, E. Crisostomi, and R. Shorten. A big-data model for multi-modal public transportation with application to macroscopic control and optimisation. *International Journal of Control*, 88(11):2354–2368, 2015.

[64] W. Fang, S. Yang, and X. Yao. A survey on problem models and solution approaches to rescheduling in railway networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):2997–3016, 2015.

[65] H. Feng, Z. Jiang, F. Xie, P. Yang, J. Shi, and L. Chen. Automatic fastener classification and defect detection in vision-based railway inspection systems. *IEEE Transactions on Instrumentation and Measurement*, 63(4):877–888, 2014.

[66] Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta. A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1):176–190, 2008.

[67] J. Fiosina, M. Fiosins, and J. Müller. Big data processing and mining for the future ict-based smart transportation management system. *Jurnal Teknologi (Sciences & Engineering)*, 62(1):33–40, 2013.

[68] S. Floyd and M. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[69] I. K. Fodor. *A survey of dimension reduction techniques*. Lawrence Livermore National Lab., CA (US), 2002.

[70] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[71] E. Fumeo, L. Oneto, and D. Anguita. Condition based maintenance in railway transportation systems based on big data streaming analysis. *The INNS Big Data conference*, 2015.

[72] T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. W. Paton. Data wrangling for big data: Challenges and opportunities. In *International Conference on Extending Database Technology*, 2016.

[73] M. Garcia, P. Fausto, C. Roberts, and A. M. Tobias. Railway point mechanisms: condition monitoring and fault detection. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 224(1):35–44, 2010.

[74] P. Germain, A. Lacasse, M. Laviolette, A. ahd Marchand, and Roy J. F. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Jorunal of Machine Learning Research*, 16(4):787–860, 2015.

[75] N. Ghaemi and R. M. P. Goverde. Review of railway disruption management practice and literature. In *International conference on Railway Operations Modelling and Analysis*, 2015.

[76] I. Gokasar and K. Simsek. *Using Big Data For Analysis and Improvement of Public Transportation Systems in Istanbul*. Academy of Science and Engineering, USA, 2014.

[77] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4), 2016.

[78] D. Goldston. Big data: Data wrangling. *Nature News*, 455(7209):15–15, 2008.

[79] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. MIT press Cambridge, 2016.

[80] R. M. P. Goverde. A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C: Emerging Technologies*, 18(3):269–287, 2010.

[81] Rail Supply Group and Rail Delivery Group. Rail technical strategy – capability delivery plan. In *Rail Safety & Standards Board*, 2017.

[82] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[83] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[84] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

[85] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[86] I. A. Hansen, R. M. Goverde, and D. J. van der Meer. Online train delay recognition and running time prediction. In *IEEE International Conference on Intelligent Transportation Systems*, 2010.

[87] I. A. Hansen, R. M. P. Goverde, and D. J. Van Der Meer. Online train delay recognition and running time prediction. In *IEEE International Conference on Intelligent Transportation Systems*, 2010.

[88] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[89] R. J. Hill, D. C. Carpenter, and T. Tasar. Railway track admittance, earth-leakage effects and track circuit operation. In *IEEE/ASME Joint Railroad Conference, 1989. Proceedings*, 1989.

[90] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[91] S. J. Hong. Use of contextual information for feature ranking and discretization. *IEEE transactions on knowledge and data engineering*, 9(5):718–730, 1997.

[92] G. B. Huang, Q. Y. Zhu, and C. K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[93] G. J. Hunter and M. F. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. In *Papers from the Annual Conference-Urban and Regional Information Systems Association*, 1993.

[94] C. Hurter, B. Tissoires, and S. Conversy. Fromdady: Spreading aircraft trajectories across views to support iterative queries. *IEEE transactions on visualization and computer graphics*, 15(6):1017–1024, 2009.

[95] D. Jäckle, H. Senaratne, J. Buchmüller, and D. A. Keim. Integrated Spatial Uncertainty Visualization using Off-screen Aggregation. In *EuroVis Workshop on Visual Analytics (EuroVA)*, 2015.

[96] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158, 1997.

[97] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Pres, 2011.

[98] W. Jentner, G. Ellis, F. Stoffel, D. Sacha, and D. A. Keim. A Visual Analytics Approach for Crime Signature Generation and Exploration. In *IEEE VIS 2016 Workshop - The Event Event: Temporal & Sequential Event Analysis*, 2016.

[99] J. Jespersen-Groth, D. Potthoff, J. Clausen, D. Huisman, L. Kroon, G. Maróti, and M. N. Nielsen. Disruption management in passenger railway transportation. In *Robust and online large-scale optimization*, 2009.

[100] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[101] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[102] Y. Kara, M. A. Boyacioglu, and O. K. Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

[103] P. Kecman. *Models for predictive railway traffic management (PhD Thesis)*. TU Delft, Delft University of Technology, 2014.

[104] P. Kecman and R. M. P. Goverde. Online data-driven adaptive prediction of train event times. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):465–474, 2015.

[105] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010.

[106] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual data mining*, 2008.

[107] D. A. Keim, T. Nietzschmann, N. Schelwies, J. Schneidewind, T. Schreck, and H. Ziegler. A spectral visualization system for analyzing financial time series data. In *Eurographics/IEEE TCVG Symposium on Visualization*, 2006.

[108] D. A. Keim and J. Thomas. Scope and Challenges of Visual Analytics, 2007. Tutorial at IEEE Visualization.

[109] I. A. Khouy, P. O. Larsson-Kraik, A. Nissen, J. Lundberg, and U. Kumar. Geometrical degradation of railway turnouts: A case study from a swedish heavy haul railroad. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 228(6):611–619, 2014.

[110] K. Kim and S. I. Chien. Optimal train operation for minimum energy consumption considering track alignment, speed limit, and schedule adherence. *Journal of Transportation Engineering*, 137(9), 2011.

[111] Y. K. Kim, J. H. Baek, and J. Y. Park. Analysis of tuning unit characteristic for track circuit maintenance efficiency. *Journal of the Korea Academia-Industrial cooperation Society*, 10(12):3594–3599, 2009.

[112] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, 1995.

[113] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[114] L. Kroon and D. Huisman. Algorithmic support for railway disruption management. In *Transitions Towards Sustainable Mobility*, 2011.

[115] M. H. Kutner, C. Nachtsheim, and J. Neter. *Applied linear regression models*. McGraw-Hill/Irwin, 2004.

[116] R. K. Lai, C. Y. Fan, W. H. Huang, and P. C. Chang. Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, 36(2):3761–3773, 2009.

[117] J. Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6:273–306, 2005.

[118] D. T. Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[119] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[120] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[121] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[122] H. Li, D. Parikh, Q. He, B. Qian, Z. Li, D. Fang, and A. Hampapur. Improving rail network velocity: A machine learning approach to predictive maintenance. *Transportation Research Part C: Emerging Technologies*, 45:17–26, 2014.

[123] H. Li, D. Parikh, Q. He, B. Qian, Z. Li, D. Fang, and A. Hampapur. Improving rail network velocity: A machine learning approach to predictive maintenance. *Transportation Research Part C: Emerging Technologies*, 45:17–26, 2014.

[124] H. Li, B. Qian, D. Parikh, and A. Hampapur. Alarm prediction in large-scale sensor networks - a case study in railroad. *IEEE International Conference on Big Data*, 2013.

[125] Q. Li, Z. Zhong, Z. Liang, and Y. Liang. Rail inspection meets big data: Methods and trends. In *International Conference on Network-Based Information Systems*, pages 302–308, 2015.

[126] Q. Li, Z. Zhong, Z. Liang, and Y. Liang. Rail inspection meets big data: methods and trends. In *International Conference on Network-Based Information Systems*, 2015.

[127] R. Li, A. Kido, and S. Wang. Evaluation index development for intelligent transportation system in smart community based on big data. *Advances in Mechanical Engineering*, 7(2):541651, 2015.

[128] Z. Lin-Hai, W. Jian-Ping, and R. Yi-Kui. Fault diagnosis for track circuit using aok-tfrs and aga. *Control Engineering Practice*, 20(12):1270–1280, 2012.

[129] R. Lior. *Data mining with decision trees: theory and applications*. World scientific, 2014.

[130] I. Louwerse and D. Huisman. Adjusting a railway timetable in case of partial or complete blockades. *European Journal of Operational Research*, 235(3):583–593, 2014.

[131] M. Ma, P. Wang, C. H. Chu, and L. Liu. Efficient multipattern event processing over high-speed train data streams. *IEEE Internet of Things Journal*, 2(4):295–309, 2015.

[132] M. Markou and S. Singh. Novelty detection: a review. *Signal processing*, 83(12):2481–2497, 2003.

[133] F. P. G. Márquez, F. Schmid, and J. C. Collado. A reliability centered approach to remote condition monitoring. a railway points case study. *Reliability Engineering & System Safety*, 80(1):33–40, 2003.

[134] D. A. McAllester. Some pac-bayesian theorems. In *Computational learning theory*, 1998.

[135] S. Menard. *Applied logistic regression analysis*. SAGE publications, 2018.

[136] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[137] S. Milinković, M. Marković, S. Vesković, M. Ivić, and N. Pavlović. A fuzzy petri net model to estimate train delays. *Simulation Modelling Practice and Theory*, 33:144–157, 2013.

[138] S. Milinković, M. Marković, S. Vesković, M. Ivić, and N. Pavlović. A fuzzy petri net model to estimate train delays. *Simulation Modelling Practice and Theory*, 33:144–157, 2013.

[139] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. In *arXiv preprint arXiv:1706.07269*, 2017.

[140] S. Mittelstädt, D. Spretke, D. Thom, D. Jäckle, A. Karsten, and D. A. Keim. Situational awareness for critical infrastructures and decision support. In *NATO STO IST-116 Symposium on Visual Analytics*, 2013.

[141] S. Mittelstädt, X. Wang, T. Eaglin, D. Thom, D. A. Keim, W. Tolone, and W. Ribarsky. An Integrated In-Situ Approach to Impacts from Natural Disasters on Critical Infrastructures. In *IEEE Annual Hawaii International Conference on System Sciences*, 2015.

[142] M. Miyatake and H. Ko. Optimization of train speed profile for minimum energy consumption. *IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING*, 5:263–269, 2010.

[143] A. Z. Mohamed. A review on crowd sourcing geo-social related big data approaches as solution to transportation problem. In *Applied Mechanics and Materials*, volume 663, pages 622–626, 2014.

[144] V. A. Morozov and M. Stessin. *Regularization methods for ill-posed problems*. CRC press Boca Raton, FL:, 1993.

[145] C. Morris, J. Easton, and C. Roberts. Applications of linked data in the rail domain. In *IEEE International Conference on Big Data*, 2014.

[146] R. Nappi. Integrated maintenance: analysis and perspective of innovation in railway sector. *arXiv preprint arXiv:1404.7560*, 2014.

[147] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International conference on machine learning*, 2011.

[148] K. Noori and K. Jenab. Fuzzy reliability-based traction control model for intelligent transportation systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(1):229–234, 2013.

[149] A. Nunez, J. Hendriks, Z. Li, B. De Schutter, and R. Dollevoet. Facilitating maintenance decisions on the dutch railways using big data: The aba case study. In *IEEE International Conference on Big Data*, 2014.

[150] S. G. Nunez and N. Attoh-Okine. Metaheuristics in big data: An approach to railway engineering. In *IEEE International Conference on Big Data*, 2014.

[151] C. Nyce. Predictive analytics white paper. In *American Institute for CPCU. Insurance Institute of America*, 2007.

[152] L. Oneto. Model selection and error estimation without the agonizing pain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018.

[153] L. Oneto, E. Fumeo, C. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and Anguita. D. Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. *IEEE Transactions on Systems, Man and Cybernetics: Systems*, 47(10):2754–2767, 2017.

[154] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*, 45(9):1913–1926, 2015.

[155] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters*, 43(2):567–602, 2015.

[156] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

[157] C. E. Otero, M. Rossi, A. Peter, and R. Haber. Determining human-perceived level of safety in transportation systems using big data analytics. In *Proceedings on the International Conference on Internet Computing*, 2014.

[158] L. Oukhellou, A. Debiolles, T. Denœux, and P. Aknin. Fault diagnosis in railway track circuits using dempster–shafer classifier fusion. *Engineering Applications of Artificial Intelligence*, 23(1):117–128, 2010.

[159] Y. Oussar and G. Dreyfus. How to be a gray box: dynamic semi-physical modeling. *Neural networks*, 14(9):1161–1172, 2001.

[160] D. L. Padmaja and B. Vishnuvardhan. Comparative study of feature subset selection methods for dimensionality reduction on scientific data. In *IEEE International Conference on Advanced Computing*, 2016.

[161] R. K. Pearson and M. Pottmann. Gray-box identification of block-oriented nonlinear models. *Journal of Process Control*, 10(4):301–315, 2000.

[162] B. Pfahringer, G. Holmes, and R. Kirkby. New options for hoeffding trees. In *Australasian Joint Conference on Artificial Intelligence*, 2007.

[163] M. A. F. Pimentel, C. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[164] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

[165] T. Polk, J. Yang, Y. Hu, and Y. Zhao. Tennivis: Visualization for tennis match analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2339–2348, 2014.

[166] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah. Improving arrival time prediction of thailand's passenger trains using historical travel times. In *International Joint Conference on Computer Science and Software Engineering*, 2014.

[167] B. Qian and K. Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, 2007.

[168] Y. Qingyang and Y. Xiaoyun. Scheduling optimization model and algorithm design for two-way marshalling train. In *International Conference on Intelligent Transportation, Big Data and Smart City*, 2015.

[169] C. E. Rasmussen. *Gaussian processes in machine learning*. Springer, 2004.

[170] Christian Robert. *Machine learning, a probabilistic perspective*. Taylor & Francis, 2014.

[171] C. Roberts, H. P. B. Dassanayake, N. Lehrasab, and C. J. Goodman. Distributed quantitative and qualitative fault diagnosis: railway junction case study. *Control Engineering Practice*, 10(4):419–429, 2002.

[172] S. J. Russell, . Norvig, P, J. F. Canny, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*. Prentice hall Upper Saddle River, 2003.

[173] D. Sacha, I. Boesecke, J. Fuchs, and D. A. Keim. Analytic Behavior and Trust Building in Visual Analytics. In *Eurographics Conference on Visualization (EuroVis)*, 2016.

[174] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology)*, 22(01):240–249, 2016.

[175] D. Sacha, H. Senaratne, B. C. Kwon, and D. A. Keim. Uncertainty Propagation and Trust Building in Visual Analytics. *IEEE VIS 2014 - Provenance for Sensemaking Workshop*, 2014.

[176] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge Generation Model for Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visual Analytics Science and Technology 2014)*, 20(12):1604 – 1613, 2014.

[177] J. Sadler, D. Griffin, A. Gilchrist, J. Austin, O. Kit, and J. Heavisides. Geosrm - online geospatial safety risk model for the gb rail network. *IET Intelligent Transport Systems*, 10(1):17–24, 2016.

[178] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[179] R. E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms, Adaptive Computation and Machine Learning Series*. The MIT Press, 2012.

[180] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[181] J. Schneidewind, M. Sips, and D. A. Keim. An automated approach for the optimization of pixel-based visualizations. *Information Visualization*, 6(1):75–88, 2007.

[182] H. Senaratne, S. Mittelstädt, C. Jacob, and T. Schreck. Uncertainty Visualization for Crisis Management in Smart Grid Environments. In *International Conference on Geographic Information Science (GIScience 2014) - Workshop on Visually Supported Reasoning with Uncertainty*, 2014.

[183] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[184] Y. Q. Shao, R. K. Liu, F. T. Wang, and M. D. Chen. Research on big data management for high-speed railway equipment. In *Applied Mechanics and Materials*, 2014.

[185] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[186] G. Shmueli and O. R. Koppius. Predictive analytics in information systems research. *Mis Quarterly*, 35(3):553–572, 2011.

[187] C. Sicre, P. Cucala, A. Fernández, J. A. Jimenez, I. Ribera, and A. Serrano. A method to optimise train energy consumption combining manual energy efficient driving and scheduling. *WIT Transactions on The Built Environment*, 114:549–560, 2010.

[188] E. Siegel. Predictive analytics. In *Hoboken: Wiley*, 2013.

[189] J. A. Silmon and C. Roberts. Improving railway switch system reliability with innovative condition monitoring algorithms. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 224(4):293–302, 2010.

[190] B. Y. Song, Y. Zhong, R. K. Liu, and F. T. Wang. Railway maintenance analysis based on big data and condition classification. In *Advanced Materials Research*, 2014.

[191] M. Stein, H. Janetzko, T. Breitkreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. Couzin, and D. A. Keim. Director's Cut: Analysis and Annotation of Soccer Matches. *IEEE Computer Graphics and Applications*, 36(5):50–60, 2016.

[192] L. Swersky, H. O. Marques, J. Sander, R. J.G.B. Campello, and A. Zimek. On the evaluation of outlier detection and one-class classification methods. In *IEEE International Conference on Data Science and Advanced Analytics*, 2016.

[193] M. Tanaka. Prospective study on the potential of big data. *Quarterly Report of RTRI*, 56(1):5–9, 2015.

[194] J. Tang, C. Deng, and G. B. Huang. Extreme learning machine for multilayer perceptron. *IEEE transactions on neural networks and learning systems*, 27(4):809–821, 2016.

[195] H. Tao and Y. Zhao. Intelligent fault prediction of railway switch based on improved least squares support vector machine. *Metallurgical and Mining Industry*, 7(10):69–75, 2015.

[196] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. Data wrangling: The challenging yourney from the wild to the lake. In *Conference on Innovative Data Systems Research*, 2015.

[197] A. Thaduri, D. Galar, and U. Kumar. Railway assets: A potential domain for big data analytics. *Procedia Computer Science*, 53:457–467, 2015.

[198] K. Thangavel and A. Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, 9(1):1–12, 2009.

[199] D. P. Thunnissen. Uncertainty classification for the design and development of complex systems. In *Annual predictive methods conference*, 2003.

[200] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[201] A. N. Tikhonov and V. Y. Arsenin. *Methods for solving ill-posed problems*. John Wiley and Sons, Inc, 1977.

[202] Ying ting Zhu, Fu zhang Wang, Xing hua Shan, and Xiao yan Lv. K-medoids clustering based on mapreduce and optimal search of medoids. In *International Conference on Computer Science Education*, 2014.

[203] S. Tiwari, R. Pandit, and V. Richhariya. Predicting future trends in stock market by decision tree rough-set based hybrid system with hhmm. *International Journal of Electronics and Computer Science Engineering*, 1(3), 2010.

[204] L. Toloşi and T. Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[205] J. Tutcher. Ontology-driven data integration for railway asset monitoring applications. In *IEEE International Conference on Big Data*, 2014.

[206] J. Tutcher. Ontology-driven data integration for railway asset monitoring applications. In *IEEE International Conference on Big Data*, 2014.

[207] N. Van Oort. Big data opportunities in public transport: Enhancing public transport by itcs. In *IT-TRANS*, 2014.

[208] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

[209] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[210] B. Wang, F. Li, X. Hei, W. Ma, and L. Yu. Research on storage and retrieval method of mass data for high-speed train. In *International Conference on Computational Intelligence and Security*, 2015.

[211] F. Wang, T. h. Xu, Y. Zhao, and Y. r. Huang. Prior LDA and SVM based fault diagnosis of vehicle on-board equipment for high speed railway. In *IEEE 18th International Conference on Intelligent Transportation Systems*, 2015.

[212] J. Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer, 2011.

[213] M. Wang, J. Wang, and F. Tian. City intelligent energy and transportation network policy based on the big data analysis. *Procedia Computer Science*, 32:85–92, 2014.

[214] F. Wanner, W. Jentner, T. Schreck, A. Stoffel, L. Sharalieva, and D. A. Keim. Integrated visual analysis of patterns in time series and text data - Workflow and application to financial data analysis. *Information Visualization*, 2015.

[215] M. Wattenberg. Visualizing the stock market. In *CHI'99 extended abstracts on Human factors in computing systems*, 1999.

[216] G. M. Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004.

[217] D Randall Wilson and Tony R Martinez. Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286, 2000.

[218] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[219] X. Xu and W. Dou. An assistant decision-supporting method for urban transportation planning over big traffic data. In *International Conference on Human Centered Computing*, 2014.

[220] Z. Xueyan and G. Depeng. Application of big data technology in marketing decisions for railway freight. *The International Conference of Logistics Engineering and Management*, 2014.

[221] K. F. Yan. Using crowdsourcing to establish the big data of the intelligent transportation system. In *Advanced Materials Research*, 2013.

[222] H. Yoon, K. Yang, and C. Shahabi. Feature subset selection and feature ranking for multivariate time series. *IEEE transactions on knowledge and data engineering*, 17(9):1186–1198, 2005.

[223] H. J. Yu, Z. G. Wang, X. Y. Liu, and D. Hu. A big data application in intelligent transport systems. In *Applied Mechanics and Materials*, volume 734, pages 365–368, 2015.

[224] C. Zhang and Y. Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[225] C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

[226] X. Zhang and D. Gong. Application of big data technology in marketing decisions for railway freight. In *ICLEM 2014: System Planning, Supply Chain Management, and Safety*, 2014.

[227] Y. Zhang and Z. H. Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14, 2010.

[228] Y. Zhao, T. h. Xu, and W. Hai-feng. Text mining based fault diagnosis of vehicle on-board equipment for high speed railway. In *17th IEEE International Conference on Intelligent Transportation Systems*, 2014.

[229] Z. H. Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[230] X. Zhu. *Knowledge Discovery and Data Mining: Challenges and Realities*. Igi Global, 2007.

[231] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.