LEVERAGING BIG DATA FOR MANAGING TRANSPORT OPERATIONS

Deliverable 1.1 **Understanding and Mapping Big Data in Transport Sector**

Kim Hee[1], Naveed Mushtaq[1], Hevin Özmen[1], Marten Rosselli[1], Roberto V. Zicari[1], Minsung Hong[2], Rajendra Akerkar[2], Sophie Roizard[3], Rémy Russotto[3], Tharsis Teoh[4]

Goethe-University Frankfurt[1], Western Norway Research Institute[2], Confederation of Organizations in Road Transport[3], Panteia B.V.[4]

April 2018

Work Package 1

Project Coordinator

Prof. Dr. Rajendra Akerkar (Western Norway Research Institute)

Horizon 2020 Research and Innovation Programme

MG-8-2-2017 - Big data in Transport: Research opportunities, challenges and limitations

Distribution level	Public (P)
Due date	30/04/2018
Sent to coordinator	08/05/2018
No. of document	D1.1
Title	*Understanding and Mapping Big Data in Transport Sector*
Status & Version	*Final*
Work Package	*1: Setting the stage on big data in transport*
Related Deliverables	*D1.2, D1.3, D3.2*
Leading Partner	*Goethe-University Frankfurt*
Leading Authors	*Hevin Özmen, GUF (Chapter 3, 5)* *Kim Hee, GUF (Chapter 1, 2, 5)* *Minsung Hong, WNRI (Chapter 4)* *Rajendra Akerkar, WNRI (Chapter 4)* *Sophie Roizard, CORTE (Chapter 4)*
Contributors	*Roberto V. Zicari, GUF (Chapter 2, 3, Appendix A-B-C)* *Tharsis Teoh, Panteia (Chapter 2)* *Julien Debussche, B&B (Chapter 2)* *Naveed Mushtaq, GUF* *Marten Rosselli, GUF*
Reviewers	Julien Debussche, B&B Jasmien César, B&B
Keywords	Big Data, Transportation, Opportunities, Challenges

Disclaimer:

This report is part of the LeMO project which has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement number 770038.

The content of this report reflects only the authors' view. The European Commission and Innovation and Networks Executive Agency (INEA) are not responsible for any use that may be made of the information it contains.

Executive summary

European Union's Transport policy's pivotal aim is to strengthen the existing Transport infrastructure, which is crucial to economic development. The improvement in the transport sector should provide efficient logistics of goods, better travel and commuting facilities, and accessibility of the European region.

This report, as part of the first phase of the Leveraging Big Data to Manage Transport Operations (LeMO) project, provides an introduction to big data in the transport sector. It identifies untapped opportunities and challenges and describes numerous data sources.

This report is a part of WP1 which is a cornerstone of the LeMO project. It aims to generate a shared understanding of current big data landscape in transport and identifies a holistic view on opportunities, challenges, and limitations.

The remainder of this report is structured as follows:

Chapter 2 explores the characteristic of big data and highlights the big data challenges in the transport sector. It covers six transportation modes (air, rail, road, urban, water and multi-modal) and two transportation sectors (passenger and freight).

Chapter 3 identifies several opportunities and challenges of big data in transportation, by using: several subject matter expert interviews, nineteen applied cases, and a literature review. It also indicates that the combination of different means and approaches will enhance the opportunities for successful big data services in the transport sector.

Chapter 4 offers an intensive survey of the various data sources, data producers, and service providers. In addition, cartography was modeled to visualize data flows intuitively. Cartography demonstrates where data originated from and where it is flowing to.

Chapter 5 summarizes all findings and provides a conclusion.

Table of contents

List of Figures

List of Tables

Glossary

Abbreviation	Expression
ABS	Anti-skid Braking System
ACID	ACID properties: Atomicity, Consistency, Isolation, and Durability
AFC	Automated Fare Collection
AIS	Automatic Identification System
ANPR	Automatic Number Plate Recognition
APC	Automatic Passenger Counting
API	Application Programming Interface
ARPS	Average Revenue Per Session
AVL	Automatic Vehicle Location
AUTOPILOT	Automated driving Progressed by Internet Of Things
AWS	Amazon Web Services
BDE	BigDataEurope (EU project)
CAPACITY4RAIL	Increasing Capacity 4 Rail networks through enhanced infrastructure and optimised operations
CERTH	Centre for Research and Technology Hellas
CFVD	Cellular Floating Vehicle Data
DATA SIM	DATA science for SIMulating the era of electric vehicles
DBMS	Database management system
ESP	Electronic Stability Program
ETL	Extract Transform Load
EV	Electric Vehicle
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HGV	Heavy Good Vehicles
HTS	Household Travel Surveys
ICT	Information and Communications Technology
IMRS	Intelligent Monitoring And Recording System
IN2RAIL	Innovative Intelligent Rail
INFORM	INstitute of Operations Research and the Management Sciences
INONMAN²HIP	InNovative Energy MANagement System for Cargo SHIP
IoT	Internet of things
ITS	Intelligent Transport System
JSON	JavaScript Object Notation
KNN	K-Nearest Neighbours
LeMO	Leveraging Big Data for Managing Transport Operations
MAPE	Mean Absolution Percentage Error
MDM	Master Data Management
MobiDig	Mobilität digital Hochfranken

MUNIN	Maritime Unmanned Navigation through Intelligence in Networks
NGO	Non-Governmental Organization
NOESIS	NOvel Decision Support tool for Evaluating Strategic Big Data investments in Transport and Intelligent Mobility Services
NYC	New York City
OCR	Optical Character Recognition
OD	Origin-Destination
OPTIMUM	Multi-source Big Data Fusion Driven Proactivity for Intelligent Mobility
ORION	On-Road Integrated Optimization and Navigation System
PFT	Package Flow Technologies
PMQ	Predictive Maintenance and Quality
RF	Random Forest (technique)
RFID	Radio Frequency IDentification
REST	REpresentational State Transfer
SNCF	Société Nationale des Chemins de fer Français
SME	Subject Matter Expert
TT	Transforming Transport
TLC	Taxi & Limousine Commission
UPS	United Parcel Service
URT	Urban Rail Transit
VANET	Vehicular Ad Hoc NETwork
VTOL	Vertical TakeOff and Landing
XML	Extensible Markup Language

1 Introduction

1.1 Abstract

European Union's Transport policy's pivotal aim is to strengthen the existing Transport infrastructure, which is crucial to economic development. The improvement in the transport sector should provide efficient logistics of goods, better travel and commuting facilities, and accessibility of the European region.

This report as part of the first phase of the Leveraging Big Data to Manage Transport Operations (LeMO) project provides an introduction to big data in the transport sector. It identifies untapped opportunities and challenges and visualizes numerous data sources. The authors believe that this report generates a shared understanding of harnessing big data and provides a foundation for phase 2 and 3 of the LeMO project.

This report is a part of work package 1 (WP1) which is a cornerstone of the LeMO project. It aims to generate a shared understanding of current big data landscape in transport and identifies a holistic view on opportunities, challenges, and limitations. The remainder of this report is structured as follows: Chapter 2 explores the characteristic of big data and highlights the big data challenges in the transport sector. Chapter 3 identifies the opportunities and challenges in research, applied cases from governmental, non-governmental and private organizations. Chapter 4 provides an extensive survey of big data sources with the cartography of data flows. Chapter 5 summarizes all findings and provides conclusion. Finally, qualitative interviews with subject matter experts are provided in three Appendices.

1.2 Purpose of the document

Functionality and efficiency of the transport sector significantly rely on data such as sensor-generated data, traffic schedules, flight information and more. Due to the recent developments in Information and Communication Technology (ICT), a large amount of data is generated, collected and processed in the transport sector. This provides untapped opportunities to gain better insight into transportation infrastructure, movement of people and vehicles. However, numerous challenges such as data silos, poor data quality and lack of expertise hinder to seize such opportunities. Thus, the demand to understand the transport sector associated with big data is surging rapidly from the transport stakeholders.

To meet the current needs, the LeMO project aims to provide a comprehensive view that is amplifying opportunities, while diminishing limitations. The LeMO project is comprised of three phases as shown in Figure 1. Phase 1 investigates the role of big data in the transport sector and identifies institutional and governmental issues. Phase 2 explores the societal impact of comprehensive case studies based on the findings of Phase 1. The findings of Phase 1 and 2 will feed into exploring the future direction in Phase 3. The created value from the course of all three phases will be disseminated through various channels in parallel.

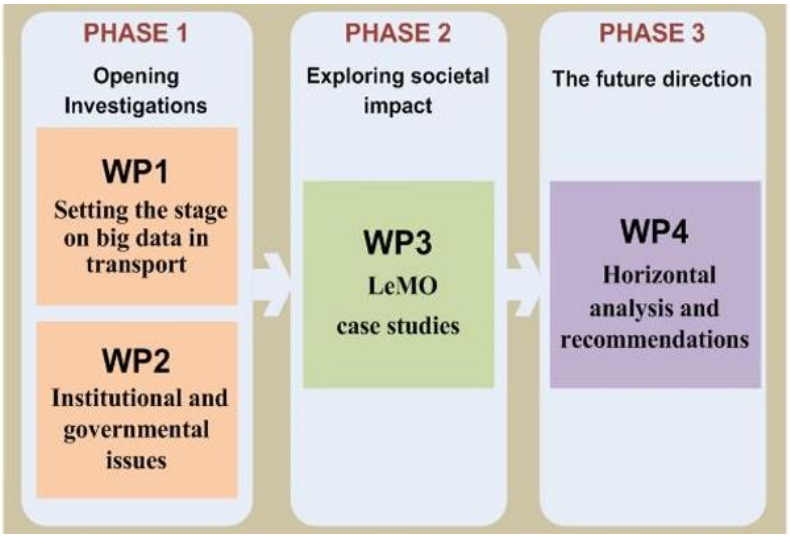

Figure 1: LeMO project - phases and work packages

1.3 Target audience

The target audience of this report comprises of government policy-makers, transport industry, and relevant technology companies. For example, part of this report attempts to help one to make an evidence-driven decision. And the course of the decision may introduce a new policy in the transport sector, which could impact a change in people's lifestyle. Nevertheless, the target audience can be anyone who is interested in examples of big data technologies in general, since it offers a comprehensive overview of harnessing big data.

2 Big Data in Transportation

Numerous data-driven applications are introduced across multiple domains. Amongst them, smart city and transportation, healthcare, social science, energy, manufacturing, and finance are popular in academia and industry. The big data applications in the transport sector, have recently gained considerable attention thanks to the vast amount of data and the improvement of big data technology. Before we go deeper into the application level, it is necessary to understand granular notions and elements comprising the big data application in the transport sector. Thus, this chapter provides the domain knowledge as well as big data and its technologies.

2.1 Understanding transportation

The lay definition of transportation is "an act, process, or instance of transporting or being transported" referring to the movement of people or goods (Merriam-Webster Dictionary, 2018). However, transportation can be viewed in different manners as the result of the interplay of different technologies, behaviours with dependencies and impacts on economic, environmental and geographical aspects (Gennaro et al., 2016; Manheim, 1980). For instance, Figure 2 (Transport Research & Innovation Portal) shows the various transport themes, which are grouped within five transport dimensions: mode, sector, technology, policy, and evaluation. While each theme focuses on specific aspects of transport research, many themes overlap. On the other hand, Figure *3 (Teodorovic and Janic, 2017, p.15)* attempts to categorize the transportation with a focus on two dimensions: mode and sector.

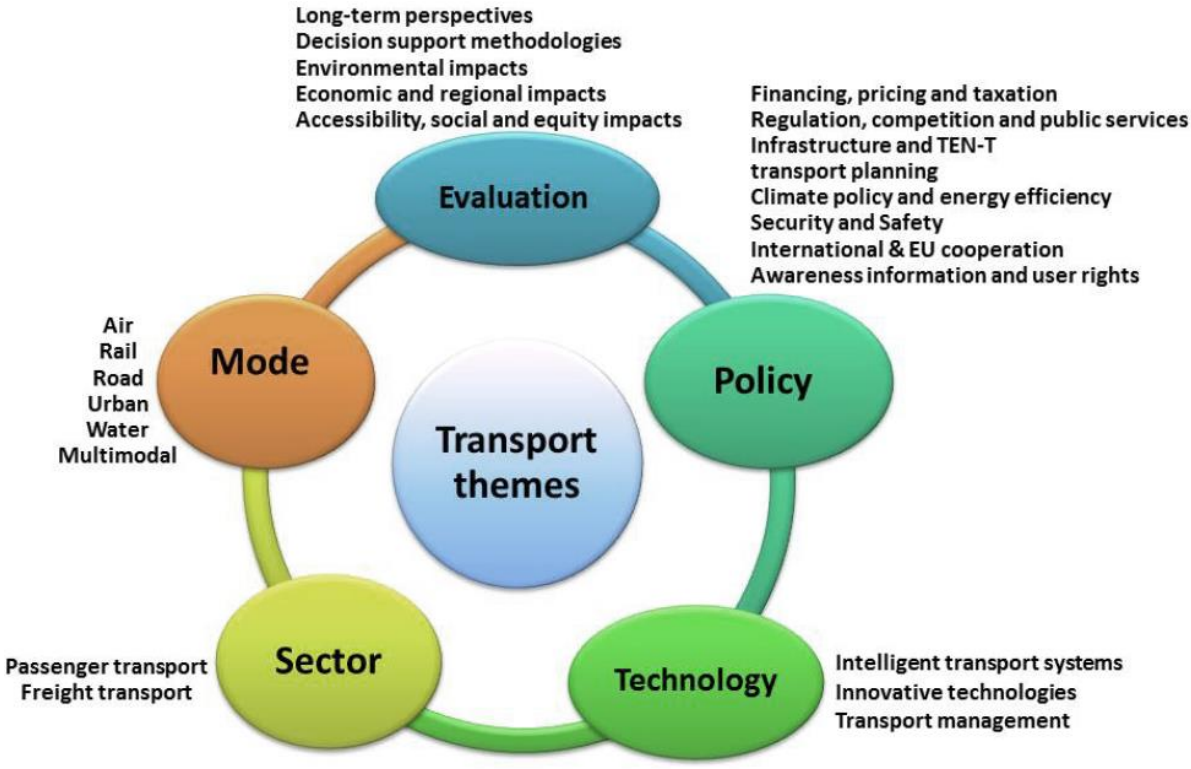

Figure 2: Transport themes within a framework of five transport dimensions

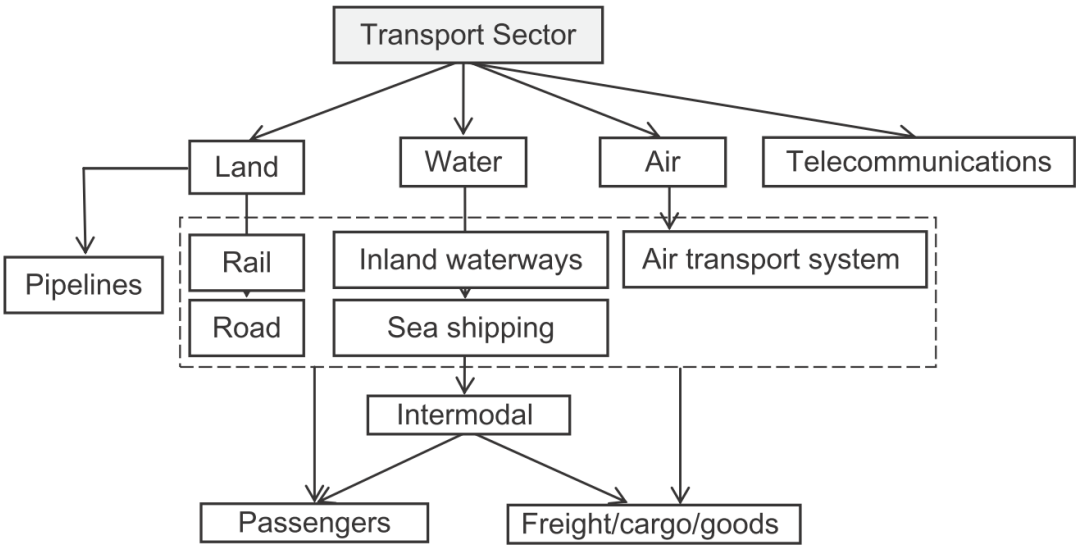

Figure 3: Simplified structure of the transport sector with modes, systems, and sectors

The scope of this report focuses on the physical movement of people and commodities. Hence, this report covers the sectors of passenger transportation and freight transportation but excludes pipelines and telecommunications, which are shown in Figure *3*. Neither of them can transport people or merchandise goods as pipeline can carry only fluid materials and telecommunication can carry radio waves and digital signal. Accordingly, this report covers the mode of transport for the transport of passengers or freights. We subgroup the mode into six categories based on the carrier medium: air, rail, road, urban, water, and a mix of modes which is called multimodal as shown in Figure 2. The examples of carrier medium for each mode are described in Figure *2*. The examples of carrier medium for each mode are described as follows:

- Air: Aircraft, balloon, helicopter
- Rail: locomotive, train, tram,
- Road: car, electric vehicle, tractor, tricycle, trolleybus, truck
- Urban: bicycle, bus, motorcycle
- Water: boat, sailboat, ship, submarine

The goal of transportation services is to satisfy mobility demand for people and freight movement. Mobility demand is determined by the cargo volume to ship or the number of passengers from given origins to destinations during a specific time period. Determinants for transport supply are various transport services in numerous transport modes and the respective system. Thus, stakeholders and components from transport supply including airports, streets, ports, and highways must meet existing transport demand. (Teodorovic & Janic, 2017, p. 15)

2.2 Understanding big data in the transport sector

2.2.1 Motivation

The focus on big data for the transport sector comes from two distinct trends. First, in the past decade, digital technology has permeated many facets - ranging from communications to economic activity, trade, and transportation (Tavasszy & de Jong, 2014). As will be discussed at

length later, these digitized activities have created massive amounts of records – whether as the functional objects or as a by-product – that in some cases have been stored as electronic data. Many have identified these data as potential sources of useful information. So, it is reasonable to ask the question "with all the data that I have access to, what can I do with it?" In short, it is a potential solution seeking a problem. The problem we have identified and will focus on is in the (very wide) transport domain.

The second trend is the realization that existing methods of creating information for the purposes of planning, control, and evaluation, are costly (expensive to carry out, taking a long time, not easily scalable) and very constrained in terms of limited coverage (geographically, sample size and temporally) (Allen et al., 2014). On a small-scale, these challenges are surmountable, but on a large-scale, for example, in urban transport planning or port control, the lack of good timely information can cause long-term negative effects, which are difficult to mitigate. These effects, have been identified as both failures of public policy and market failures, leading to economic losses, and undesirable social and environmental outcomes (Cui et al., 2015).

While it is the purpose of the research project to provide an in-depth analysis of the potential of big data for the transport sector to achieve its aims, we provide a brief description of the aim of "planning, control, and evaluation" in the transport sector (Lindholm & Ballantyne 2016; Marsden & Reardon, 2017). This serves as the basic motivation for the study to promote technology in this direction.

Underlying the planning, control and evaluation activities is the need to meet high sustainability. This concept as applied in transportation is valid for the public and private sector, i.e. those who are managing "public" transportation resources in a governmental role, and those who are managing "private" transportation resources in a commercial role. Briefly, the concept of sustainability is defined as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (WCED, 1987), and is often discussed in an economic, social and environmental framework. In the context of transport, a sustainable transport system has the following general qualities (Wolfram, 2004):

- *Allows the basic access and development needs of individuals, companies, and societies to be met safely and in a manner consistent with human and ecosystem health, and promotes equity within and between successive generations;*
- *Is affordable, operates fairly and efficiently, offers a choice of transport mode, and supports a competitive economy, as well as balanced regional development;*
- *Limits emissions and waste within the planet's ability to absorb them use renewable resources at or below their rates of generation, and, use non-renewable resources at or below the rates of development of renewable substitutes while minimizing the impact on the use of land and the generation of noise.*

Based on this, we can understand the need to expand our understanding of the potential benefits and detriments caused by transport activity. To name a few detriments: transport causes about 25% of the EU’s greenhouse gas emissions (European Environment Agency, 2017), can be attributed to poor health (from air or noise pollution), mortality (caused by accidents)

and other externalities in urban areas (Santos et al., 2010). However, transport is more often, except in some cases, also a "derived demand" that enables work and trade, enables human social activity, and promotes human welfare (Ortuzar & Willumsen, 2011).

What we want to ensure through sustainable transport activity, by means of good planning, control, and evaluation, is that the benefits of transport are enhanced or at least maintained, while the detriments are reduced and eliminated. While such outcomes cannot be guaranteed, techniques and approaches to achieve them through big-data driven analysis hold the potential of improving them.

2.2.2 Big data definition

While big data is ubiquitous terminology today, there is no consensus among scholars and professionals on the definition of big data. According to an online survey in April 2012 polled by Harris Interactive behalf of SAP, among 154 executives at multinational companies varied widely in their understanding of big data (SAP, 2012). Some answered what big data is, while others tried to answer what it does. With 28 percent defining it as the massive growth in transaction data; 24 percent described it as new technologies that address the volume, variety and velocity challenges of big data; 19 percent said big data refers to requirements to store and archive data for regulatory compliance; 18 percent saw big data as the rise in new data sources, such as social media, mobile device, and machine-generated devices; and 11 percent said others.

In 2001, Volume, Variety, and Velocity (also known as three Vs of big data) have been introduced in (Laney, D. 2001). The author introduced the three dimensions based on its challenges in data management. However, the landscape of big data is changing rapidly and this momentum pushes the boundary of the definition of big data consistently. In 2014, about 13 years later, seven Vs of big data are introduced in (Uddin, M. F., & Gupta, N., 2014). This report covers **four Vs of big data with a focus on the transport sector.**

- Volume – This is the most popular and obvious characteristic of big data, as the first word of "big" is indicating this characteristic. The massive volume of data collected from millions of vehicles in Europe, social data, sensor data, etc.
- Variety – The differences within the industry standards, sampling rates, and data types such as video, JSON (JavaScript Object Notation), XML (Extensible Markup Language), pictures, text and more.
- Velocity – The high arrival rate data, for instance, sensor data, weather data, GPS (Global Positioning System) generated data, social media messages, and data generated from vehicle onboard devices and etc.
- Veracity – The potential for missing or erroneous data due to environmental conditions, unreliable data sources, equipment failures, or malicious intent.

2.2.3 Big data challenges related to the four core Vs

In general, big data challenges are laid over two disciplines of data science and big data engineering. In data science disciplines, data analysis pipeline is very complex as it consists of multiple phases including data acquisition, processing, aggregation, and delivery. Furthermore,

the pipeline is recursive and non-linear. Moreover, there are numerous algorithms that need to be selected and hyper-parameters need to be tuned in an ad-hoc manner. On the other hand, in data engineering disciplines, there are numerous big data tools and services available. Thus, designing an optimal architecture is very complex. Figure 4 illustrates example architecture of big data pipeline and numerous tools proposed by Amazon.

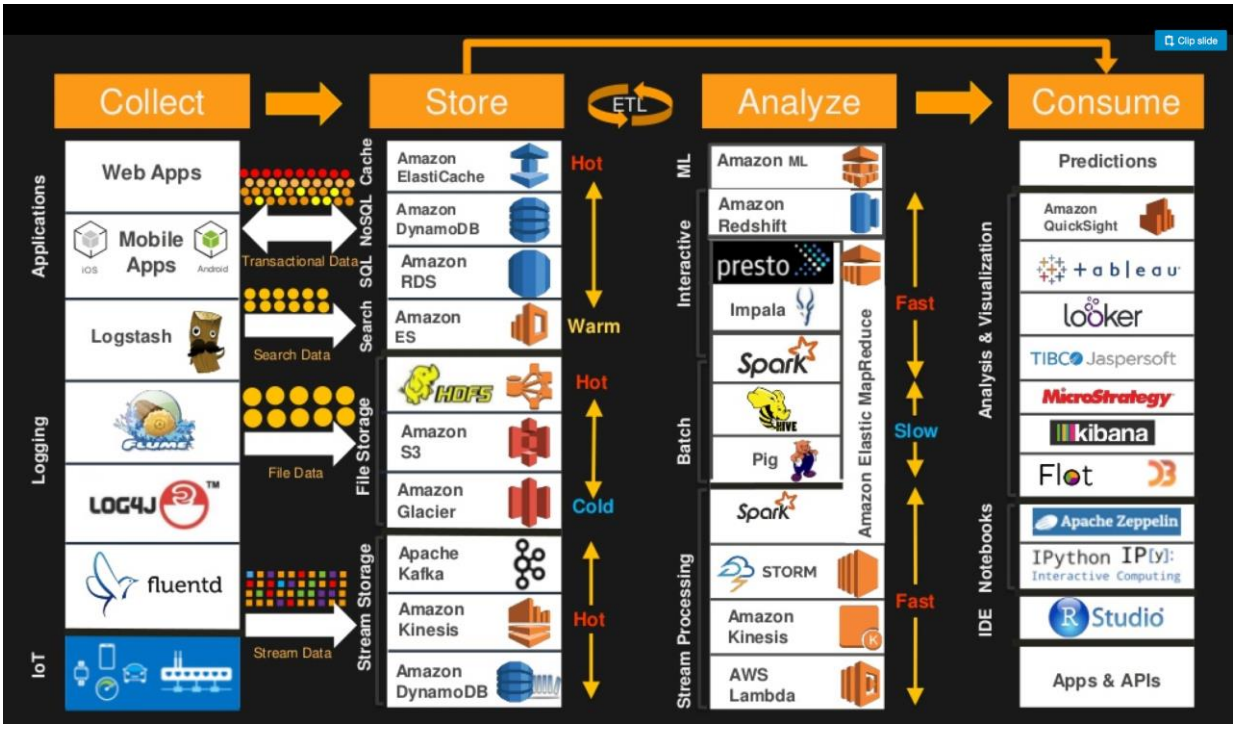

Figure 4: Big data architecture from Amazon (AWS website) selected

When it comes to the challenges in the transport sector, we identify concrete barriers corresponding to the defined four Vs. The first three Vs (volume, variety, and velocity), they are the challenges sitting in big data engineering dimension. The last one veracity is the challenge in data science dimension.

- The first challenge is **volume**. It is hard to store and process the massive amount of data in a conventional database. For example, 480 TB of data were generated by cars in 2013 and it is anticipated to increase to 11.1 PB in 2020 (Chowdhury, 2017).
- The second challenge is **variety of data**. The conventional relational database is not able to store unstructured and heterogeneous data such as signal data generated by a sensor, free text data collected through social media, image data generated by satellite and video data streamed by closed-circuit television.
- The third challenge is **velocity.** The conventional database cannot ingest streaming data generated by multiple sensors in real time. For instance, traffic cameras on highways continuously stream video data and it aims to identify traffic condition and car accidents.
- The last challenge is **veracity**, which means uncertainty and incorrectness. This characteristic emerges because big data is not bounded to the relational database management system (DBMS), which cares Atomicity, Consistency, Isolation, and

Durability (known as ACID properties). Therefore, data embed high potential to be in a low quality.

The last challenge is **the most critical challenge because the low data input quality leads to a false result**, whereas the first three challenges simply cause inconvenience such as service can be slower, or the storage can be out of space. Thus, data quality issues particularly need more attention in this section.

2.2.4 Interviews on data quality

To investigate more closely on Data Quality, GUF team interviewed[1] twenty-one Data Scientists from both industry and academia. The selected interviewees are as follows:

- Jeff Saltz (Associate Professor at Syracuse University)
- Yanpei Chen (Data scientist product-manager at Splunk)
- Manohar Swamynathan (Staff Data Scientist at GE Digital)
- Jonathan Ortiz (Data scientist and knowledge engineer at data.world)
- Anya Rumyantseva (Data Scientist with Pentaho, a Hitachi Group Company)
- Dirk Tassilo Hettich (Tax Technology & Analytics at EY)
- Wolfgang Steitz (Senior Data Scientist at travel audience GmbH)
- Paolo Giudici (Full Professor at University of Pavia)
- Andrei Lopatenko (Director of Engineering for Recruit Institute of Technology)
- Mike Shumpert (Vice President of Analytics at Software AG)
- Romeo Kienzler (Chief Data Scientist in the IBM Watson IoT World Wide team)
- Elena Simperl (Professor of Computer Science at the University of Southampton)
- Mohammed Guller (Glassbeam)
- Natalino Busa (Head of Data Science at Teradata)
- Vikas Rathee (Staff Data Scientist with GE Digital)
- Christopher Schommer (Associate Professor at the University of Luxembourg)
- Slava Akmaev (Senior Vice President and Chief Analytics Officer at Berg, LLC.)
- Jochen Leidner (Director of Research with Thomson Reuters)
- Claudia Perlich (leads the machine learning at Dstillery)
- Richard J Self (Research Fellow at the University of Derby)
- Ritesh Ramesh (Chief Technology Officer at PricewaterhouseCoopers)

The questions asked were the following:

1. How do you ensure data quality?
2. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?

[1] The original interviews have been conducted and published on ODBMS.org, reprinted here with permission of the resource portal.

The detailed interviews are provided in the **Appendix A** and the summary of interviews is given below:

In response to the first question, how to ensure data quality, most interviewees agreed that data quality cannot be ensured, and it is a continuous issue to be addressed. Statistical methods (e.g. univariate analysis, multivariate analysis and outlier detection) can help to ensure the data quality in pre-processing phase. Implementing a holistic and end-to-end process would be more appreciated because data quality can be monitored continuously beyond the isolated task in the data pre-processing phase. Nevertheless, most interviewees pointed out that a validation step with domain experts is inevitable to ensure high data quality.

With respect to the second question, most interviewees came back to the same answer. Applying a statistical method can be useful to evaluate the result but including domain experts is much appreciated to evaluate the obtained insights. For instance, a constant communication with stakeholders and domain experts at the early stage could save tremendous time and resources. And one of the popular statistical evaluation approaches is an out-of-sample method.

3 Opportunities and Challenges of Big Data in Transportation

Utilizing big data in the transport sector is faced with numerous challenges ranging from data generation to service deployment in a scalable system. This chapter identifies opportunities and challenges from three sources with different evidence levels. The hierarchy of evidence is taken from the medical research discipline by the National Health and Medical Research Council (Council et al., 2009). In terms of the level of evidence, the opportunities and challenges from the domain experts are relatively lower than the research papers due to the following reasons: the opinion on a subject matter may differ within experts and it often lacks references.

3.1 Subject matter expert interviews

A number of subject matter experts have been asked to give their opinion on the transport sector associated with big data. Interviews of eight experts[2] in both industry and academia conducted via E-Mail:

- Dr. Frank Wisselink (Managing Consultant Big Data, IoT & Transformation at Detecon)
- Christopher Sciacca (Communications Manager at IBM Research)
- Dr. Gerhard Kress (Director Mobility Data Services at Siemens)
- Stephen Dillon (Senior Architect at Blacksmith Applications)
- Scott Jarr (Co-founder of VoltDB)
- Dr. Alessandra Bagnato (research scientist and project manager at Softeam Group)
- Prof. Jeff Saltz (Associate Professor at Syracuse University)

The detailed interviews are provided in **Appendix B** and the summary of these interviews is given below:

Eight subject matter experts conveyed numerous initiatives and applications in the transport sector associated with big data. The conveyed opportunities of using big data are as follows:

- Automated driving and parking applications
- Internet of Things (IoT) with vehicles and road infrastructures
- Object detection, spatiotemporal track association
- Behaviour analysis and intent prediction

However, drawing actionable insights from big data are not easy. The interviews revealed a few common challenges embedded in the transport sector. Challenges addressed mostly were coherent with technical issues:

- Reducing latency of analytics
- Real-time result delivery
- Deal with a large amount of data
- Transportation planning

[2] The original interviews have been conducted and published on ODBMS.org, reprinted here with permission of the resource portal.

3.2 Applied cases

Nineteen applied cases in (non) government, industry and research projects, and university initiatives are investigated. The major contribution of this section is to identify opportunities and challenges from the project descriptions. Note that some projects lack details because they are under development when this report is written.

The identified opportunities of using big data for transportation can be summarized as follows:

- Automated and unmanned vessels (MUNIN)
- Automated transport planning and management (NOESIS, BDE)
- Automating the entire business delivery chain (ORION)
- Big data analysis for transport policy-making (Strava Metro)
- Passenger travel behaviour analysis (OPTIMUM, DATA SIM)
- Real-time event analysis and prediction for traffic (OPTIMUM)
- Reduction of congestion (Taxi Movement in NYC)
- Self-monitoring infrastructure and maintenance (CAPACITY4RAIL, INONMAN²HIP)
- The reduction of carbon emissions (all)

Compared to the opportunities, not many challenges are addressed through the case studies. One of the reasons is that many projects are in the initial phase. Usually, the challenges are occurring during under the development or after deployment.

- Legal privacy barriers such as data privacy in the EU (EU GDPR)
- Lack of skilled professionals
- Big data challenges as known as the 4V

3.2.1 Non-governmental and government projects and initiatives

A number of Non-Governmental Organizations (NGOs), as well as governmental projects and initiatives, have been put forward in recent years. Among them are various EU-funded initiatives, either concluded or ongoing.

3.2.1.1 Transforming Transport (TT)

The Transforming Transport (TT)[3] project is an EU-funded project (2017-2019)[4] representing a consortium of 47 transport, logistics and information technology stakeholders in Europe. The objective of the project is to **demonstrate the effects of big data on mobility and logistics**. This goal is established by validating the economic and technical viability of big data in transport processes and services. TT covers all six modes and both sectors - passenger and freight.

TT explores seven transportation domains to demonstrate, validate and evaluate the capability for big data applications in real operational scenarios (pilots). The pilots are described as follows:

[3] https://transformingtransport.eu/, Accessed January 22, 2018

[4] http://cordis.europa.eu/project/rcn/206575_de.html, Accessed January 22, 2018

- **Smart highways:** 1,000 GB of data (1 GB per day) are anticipated to be used. The data are traffic information, historical traffic data, meteorology information, OCR (Optical Character Recognition) data, spatial infrastructure, car sensor data, and social network streams. Numerous goals can be achieved in this category: understand mobility patterns in corridors and route choice criteria (forecast traffic flow including external variables such as weather, events etc.); optimize operational highway efficiency (optimizing scheduling and maintenance); increase safety (reduce the number of accidents, prevent and react to accidents).
- **Sustainable vehicle fleets:** 2,500 GB of data (7 GB per day) are anticipated to be used. The data are vehicle, position, speed, brake, black ice sensor, ABS, ESP, fuel level, emergency button, engine status, engine revolutions, tire pressure, and temperature. This category aims to achieve the following objectives: deploy big data infrastructure comprising descriptive and predictive analytics capabilities; enable big data injection; develop generic big data visualization application; develop predicting functionality with maintenance features, machine learning functions, and visualization; reduce emissions; detect traffic jams.
- **Proactive rail infrastructures:** 1,500 GB (2 GB per day) are anticipated to be used. The utilized data are environmental, geospatial, network model, track circuits, axle counters, points heaters, rail temperature monitoring, overhead line tension, rail signalling power, scheduled planning and control data, and track usage. The following objectives can be achieved: verification of quality, accuracy, and provenance of asset data; prediction and prevention of maintenance activities; improvement of track-side asset availability and reliability; increased availability of rail infrastructure for freight and passenger transport; collection, processing and analysis of information for prediction of utilization of rail asset components.
- **Ports as intelligent logistics hubs:** 400,000 GB (110 GB per day) are anticipated to be used. The utilized data are import and export of containerized cargo, general and bulk cargo, tracking and tracing through the whole logistics chain, machines equipped with, AIS (vessel tracking), video streams of trucks and trains (from video gates), and traffic management data. The following objectives can be achieved: improve efficiency in crane movements; apply predictive maintenance models to cranes' spreaders; improve predictive decision making.
- **Efficient air transport:** 3,000 GB (30 GB per day) are anticipated to be used. The utilized data are flights scheduling, flights updates, passenger tracking information across the airport, queuing times for check-in/security/lifts, hand luggage scanning information, shops level of occupation, aircraft sensor data. The following objectives can be achieved: improve the operation (reduce delays, improve baggage processing, increase efficiency in passenger processing stations, and reduce turnaround times); analyze passenger behavior.
- **Multi-modal urban mobility:** 500 GB (10.5 GB per day) are anticipated to be used. The utilized data are real-time probe data from personal, freight and public transport vehicles, traffic light data and sensors, roadside camera images, usage data from public transport. Objectives in this pilot are as follows: improve traffic operations; generate

freight traffic model; improve decision making in freight delivery scenarios; improve efficiency for delivery fleets.

- **Dynamic supply chains:** 1,500 GB (11 GB per day) are anticipated to be used. The type of data is inventories, orders acquired from online retailers, deliveries, distribution information collected by 3PL/courier companies, and customer data. Numerous objectives can be met in this pilot: decrease costs; mitigate carbon emissions; improve customer satisfaction with respect to e-commerce deliveries; increase the precision of future demand estimations; provide alternative shipping methods.

Identified opportunities and challenges

Along with the above-mentioned objectives, the identified opportunities for transport sector are as follows: cross-sector data sharing, open data platforms, and meta-data repositories. The identified barriers are the protection of personal data in the EU (EU GDPR) and the protection of commercial data. Furthermore, the lack of skilled practitioners in the job market is reported as one of the major challenges, as the demand for technical professional exceeds labour supply by far.

3.2.1.2 BigDataEurope (BDE)

The BigDataEurope (BDE)[5] project was an EU-funded project (2015-2017)[6] aiming at integrating big data, software, and communities for addressing Europe's societal challenges. The specific objectives were the development of an adaptable and easily-deployable big data platform for stakeholders to integrate into their business processes and the establishment of a network of stakeholders for key European societal sectors. Finally, BDE aimed at designing, developing, and evaluating a Big Data Aggregator infrastructure meeting usability demands of the various domain stakeholders. The considered societal challenges were:

- Health: heterogeneous data linking and integration, biomedical semantic indexing
- Food and Agriculture: Large-scale distributed data integration
- Energy: real-time monitoring, stream processing, data analytics, decision support
- Transport: streaming sensor network and geospatial data integration
- Climate: real-time monitoring, stream processing, data analytics
- Social sciences: statistical and research data linking and integration
- Security: real-time monitoring, stream processing, data analytics, image data analysis

The outcome of the project was the BDE technology Platform (open source platform designed to capture, manage and process big data) supporting the various dimensions of big data. The pilot in the transport domain "Big Data & Smart, Green and Integrated Transport" addressed congestion as one of the major problems, especially in urban areas. The pilot explored data ingestion, data processing and data storing for the classification of traffic conditions and for traffic predictions.

[5] https://www.big-data-europe.eu/pilot-transport/, Accessed February 1, 2018

[6] http://cordis.europa.eu/project/rcn/194216_en.html, Accessed February 1, 2018

Transport modes are a road (cars, vehicles) in urban areas. The addressed sector was passenger transport. The pilot used the city of Thessaloniki as the test case. Apache Kafka was used as a message broker and allowed the components to communicate asynchronously. Apache Flink was used to processing the stream data, like the taxi data provided by the Transport of the Centre for Research and Technology Hellas (CERTH), and also batch data. Elasticsearch, a document database based on the open source search engine Apache Lucene, was used to store the records after data processing.

Identified opportunities and challenges

The project objectives were to improve mobility-related data collection, utilising real-time data for the provision of accurate mobility services and advanced transport planning. BDE aimed to enable better decisions from the travellers' side as well as improved traffic management at the city level by the respective traffic management authorities.

The pilot tackled some interesting challenges around using sensor data, spatial databases, GPS data, messages from social networks and webcams. The use of mobility data coming from multiple sources presented significant challenges, especially due to the different nature of the datasets both in content and spatiotemporal terms. Data sources in the transport domain are generally characterized by time and spatial dimensions. A further challenge identified was that the data should be collected and processed in real time, as the value of data records decreases quickly in time.

3.2.1.3 NOvel Decision Support tool for Evaluating Strategic Big Data investments in Transport and Intelligent Mobility Services (NOESIS)

The ongoing EU-funded research project NOvel Decision Support tool for Evaluating Strategic Big Data investments in Transport and Intelligent Mobility Services[7] (NOESIS) (2017-2019)[8] aims to identify the critical factors and features leading to successful implementation of Big Data technologies and services in the field of transport and logistics taking into account their socio-economic impact. For this various transportation areas and contexts will be examined for ICT investments and potential. Different use cases will be the basis for the research study.

Identified opportunities and challenges

NOESIS plans to contribute to the holistic understanding of big data impact in the transport sector, not specific to a certain transportation mode. The targeted outcome is the development of a tool for decision support to predict the value generated from these technologies also accounting for the various characteristics of the different transport systems. Due to the distinct similarities of NOESIS and LeMO, both projects aim to collaborate and exchange their findings while they progress. As NOESIS is only in the beginning phase, major challenges have not been identified to the best of our knowledge.

[7] http://noesis-project.eu/, Accessed December 13, 2017

[8] http://cordis.europa.eu/project/rcn/211706_de.html, Accessed December 13, 2017

3.2.1.4 DATA science for SIMulating the era of electric vehicles (DATA SIM)

DATA science for SIMulating the era of electric vehicles (EVs) (DATA SIM)[9] is a completed EU-project (2011-2014)[10] that aimed at providing a detailed methodology for spatial-temporal microsimulation in the domain of human mobility. DATA SIM targeted forecasting the various nation-wide consequences that would accompany a major switch to electric vehicles, especially with the interrelation between mobility and power distribution networks. Included data sources are social data, GPS data, mobile call records, as well as information on energy consumption, land-use, road network, and public transport data. The transport mode covered were road transportation on both passenger and freight transport. The following objectives were set:

- tackle big data challenges such as data storage, integration, management and privacy that data
- include behavioural aspects of travel patterns to understand mobility demand
- understand agent-driven changes in traffic decisions from a behavioural point of view
- overcome scalability issue of big data

Identified opportunities and challenges

The project intended to make use of the mitigating effect of EVs on carbon emissions and on energy consumption. Big data applications enable the analysis of travel behaviour. Addressed challenges in the project can be categorized into data challenges and modelling challenges. Data challenges lied in the complex nature of travel behaviour information, poor data quality, and the complexity of semantic interpretation but also the issues of data integration, data privacy and data storage. In terms of modelling, addressed challenges were the inaptitude of existing model structures for current travel behaviour analysis, as well as the lack of scalability in the models.

3.2.1.5 Multi-source Big Data Fusion Driven Proactivity for Intelligent Mobility (OPTIMUM)

OPTIMUM[11] was an EU-funded project (2015-2018)[12], which aims to improve transportation in urban areas using data from various sensors, systems, and service providers. The result was a scalable, distributed architecture for big data processing, monitoring, and prediction in the transport sector. OPTIMUM is relevant for all modes in passenger and freight transport.

Identified opportunities and challenges

The objectives were advances in transit, freight transportation, and connectivity throughout Europe. OPTIMUM explored opportunities such as travel behaviour analysis, sentiment analysis, predictive analysis, real-time event-based big data processes with potential for maintenance of transportation systems, efficiency, proactive charging for freight transport and vehicle communication integration. Challenges addressed in addition to the four Vs were

[9] https://www.uhasselt.be/datasim, Accessed December 13, 2017
[10] https://cordis.europa.eu/project/rcn/100232_en.html, Accessed December 13, 2017
[11] http://www.optimumproject.eu/, Accessed December 13, 2017
[12] http://cordis.europa.eu/project/rcn/193380_en.html, , Accessed December 13, 2017

handling of real-time data, enabling semantic understanding especially with context-aware forecasting and lastly difficulties decision-making support.

3.2.1.6 Increasing Capacity 4 Rail networks through enhanced infrastructure and optimised operations (CAPACITY4RAIL)

The concluded EU-project CAPACITY4RAIL[13] (2013-2017) aimed at improving railway systems by delivering solutions for track design, freight, operation and capacity and advanced monitoring. Results included technical demonstrations and system-wide guidelines and recommendations for future research and practice.

Identified opportunities

CAPACITY4RAIL assessed big data opportunities for self-monitoring infrastructure and maintenance, especially for rather fragile key components of the infrastructure due to weather conditions, improvements in speed as well as measures for traffic planning. Further opportunities are integrated monitoring, automated coupling, traffic capacity computation and real-time traffic flow analysis.

3.2.1.7 Innovative Intelligent Rail (IN2RAIL)

The ongoing EU-project IN2RAIL[14] (2015-2018)[15] aims at enhancing existing rail capacity, increasing the reliability of rail systems and decreasing costs in the rail sector. Considered sectors are a passenger as well as freight. The targets are addressed with three technical sub-projects smart infrastructure, intelligent mobility management as well as power supply and energy management systems.

Identified opportunities and challenges

The application of big data open up opportunities for integrated asset monitoring, self-diagnostic maintenance, increasing resilience and decreasing energy consumption. The challenges are yet to be explored in the course of the project.

3.2.1.8 oneTRANSPORT

The oneTRANSPORT project[16] (Collaborative R&D, Innovate UK) aims to make transport more accessible and usable by providing an open and scalable platform for multi-modal and multi-system transport integration across local authorities. oneTRANSPORT focuses on integrating transport modes and systems from 4 contiguous counties. Users can use this platform with a subscription fee.

[13] http://capacity4rail.eu/, Accessed December 13, 2017
[14] http://www.in2rail.eu/, Accessed December 13, 2017
[15] http://cordis.europa.eu/project/rcn/193360_de.html, Accessed December 13, 2017
[16] https://onetransport.io/, Accessed December 13, 2017

Identified opportunities

The platform offers real-time data sets within one environment and allows for the collaboration of various stakeholders. The expected impacts are the overall improvement in transport management and traffic operations.

3.2.1.9 Maritime Unmanned Navigation through Intelligence in Networks (MUNIN)

The idea of the EU-project Maritime Unmanned Navigation through Intelligence in Networks (MUNIN)[17] (2012-2015)[18] was to develop an autonomous ship concept (freight transportation), by combining an automated decision system with remote control via a shore-based station. The feasibility of the unmanned vessel was explored at the technical, economic and legal level. Results of MUNIN were the following:

- The technical concept for unmanned cargo vessels for more sustainable maritime transport (Ship is autonomously operated by new systems aboard the vessel, monitoring and controlling performed by the operator on land)
- Assessment of the concept's technical, economic and legal feasibility
- Development of prototypes for subsystems, including both onboard and onshore modules.

Identified opportunities

MUNIN aimed to develop and verify a concept of an autonomous ship with key impacts on European shipping's competitiveness and safety. The economic feasibility in terms of vessel costs, the reduction in fuel consumption as well as the increase in safety was demonstrated. One of the challenges in terms of safety is the threat of cyber-attacks to unmanned automatic vehicles.

3.2.1.10 InNovative Energy MANagement System for Cargo SHIP (INONMAN²HIP)

The objective of the EU-project InNovative Energy MANagement System for Cargo SHIP from 2011 to 2014[19] was the development of smart energy management strategies to reduce onboard CO2 emissions.

Identified opportunities

The project proposed an energy management system to collect data in real time and thus to anticipate and optimise energy requirements for each operational ship configuration. Benefits included

- Life-cycle analysis, risk assessment, cost study to provide an independent assessment of the impact of alternative energy sources for onboard cargo ships
- Optimize energy efficiency for ships for green maritime transport
- Improve efficiency of conventional sources of energy

[17] http://www.unmanned-ship.org/munin/, Accessed January 22, 2018
[18] https://cordis.europa.eu/result/rcn/169600_en.html, Accessed January 22, 2018
[19] https://cordis.europa.eu/result/rcn/185049_en.html, Accessed January 19, 2018

- Consider overall energy production and its management system aboard ships and adopt a holistic approach

3.2.1.11 UN Global Pulse Jakarta – Using Big data analytics for public transport

This project is the ongoing collaboration between UN Global Pulse Jakarta Lab and the Smart City team within the Jakarta City Government. It aims at enhancing transportation planning and improved decision-making via real-time analytics in line with Jakarta's Smart City initiative. The mode is road transport for the passenger sector. The collaboration project analyses real-time bus location data, service demand data, and real-time traffic information in two phases. The first phase focuses on mapping abnormal traffic behaviours and locations and the impact of traffic dynamics on customer demand. The insights from this stage will then be put into action for the bus transport planning in phase two. [20]

The case study uses different data sources to satisfy the project objectives. Firstly, the Jakarta Smart City platform, established in December 2014 as part of the Jakarta Smart City initiative for using big data analytics for improved public transport. The platform aggregates data via government, third parties, and crowd-sourced data with applications such as Qlue and Waze for delivery of real-time data. In addition, the city's Bus Rapid Transit system TransJakarta, founded for improved transport system delivery i.e. reduce commuting time or rush hour traffic, provides data in order to meet the increasing challenges of waiting time, unpredictable service frequency and travel times.

With the collaboration TransJakarta now analyzes real-time TransJakarta data. The data collected is analysed for various insights. GPS bus data is analysed for the identification of problematic regions in terms or traffic inefficiencies and passenger tap-in data is the basis for exploring passenger behaviour (Origin-Destination: OD) statistics and waiting times. TransJakarta's data stems from GPS devices where bus information is updated every five seconds. The data is not openly available. [21]

As initial project results, GPS data from April to June 2016 were processed in order to analyze speed, number of operating vehicles and bottleneck ratios in twelve corridors. From this, problematic lanes and regions in the city could be identified. Decisions that can be derived from this are measures for detection of irregularities and for the reduction of bottlenecks of respective routes. For a better understanding of passenger travel demand, tap-in data from May to June 2016 was analysed for in weekly patterns using data from 422,694 unique passengers. Of these, 48% were included in the OD pair analysis. Various insights were drawn from that, e.g. top destination lines, number of hours between consecutive trips, top OD pairs. Further, anonymized data was the basis for determining average passenger waiting times based on a number of buses as well as the number of passengers. Finally, based on the findings, city

[20] https://www.unglobalpulse.org/projects/improving-transport-planning-with-data-analytics, Accessed February 22, 2018

[21] http://unglobalpulse.org/sites/default/files/Project%20Brief%20-%20Using%20big%20data%20analytics%20for%20imrpoved%20pubic%20transport.pdf, Accessed February 22, 2018

administration used the analysis for introducing targeted measures such as adding more officers, reducing barriers for securing dedicated lanes in more congested areas and adding more vehicles for bus transport on routes, where the project implied the need.

Identified opportunities and challenges

The first phase focuses on mapping abnormal traffic behaviours and locations and the impact of traffic dynamics on customer demand. The insights from this initial stage will then be put into action for Jakarta's bus transport planning in phase two. [22] As the project is currently in progress, no challenges have yet been reported to the best of our knowledge.

3.2.2 Industry and research projects and initiatives

3.2.2.1 ORION (On-Road Integrated Optimization and Navigation system) at UPS

In this section, the usage of big data at UPS is described, specifically with the program *"On-Road Integrated Optimization and Navigation system"* (ORION). With the help of ORION technology UPS drivers are navigated with optimized routes using advanced algorithms and fleet telematics. Thereby, based on online map data tailored to UPS to calculate distances and travel time, drivers are provided the optimal path for package delivery while also integrating customer needs.[23]

Some of the main lessons learned from the development of the ORION are mentioned in an interview with Jack Levis (Senior Director, Industrial Engineering at UPS), published in the ODBMS Industry Watch [24]. The interview is included as an appendix in this report with permission for publication and reveals the wide potential of big data for package delivery but indicated some of the challenges. A summary of opportunities and challenges revealed in the interview is presented below and the interview is provided in the **Appendix C**.

Identified opportunities and challenges

The potential of the ORION project lies in the central challenge of increasing customer demand for personalized services in package delivery. With ORION this challenge was faced based on a proprietary Package Flow Technologies (PFT) data infrastructure using predictive modelling. The required analytics tools are also built in-house combining operations research, IT and business processes. The long-term goal is to comprehensively automate the entire delivery chain (including sorting, loading, moving of vehicles) with the continuous development of ORION. Jack Levis also revealed some critical barriers and success factors when introducing big data applications into transport. It is critical to fully understand the problems addressed, in his case the delivery problem. When handling data, data accuracy is named as a critical inhibiting factor. During data analysis, the typical mistake is not using a decision-centric approach during

[22] https://www.unglobalpulse.org/projects/improving-transport-planning-with-data-analytics, Accessed February 22, 2018

[23] https://pressroom.ups.com/pressroom/ContentDetailsViewer.page?ConceptType=Factsheets&id=1426321616277-282, Accessed February 24, 2018

[24] http://www.odbms.org/blog/2017/08/big-data-at-ups-interview-with-jack-levis/, Accessed February 24, 2018

data analysis to fully grasp big data value specific to the considered domain. The technology behind it should meet the decision requirements. During the implementation of big data projects, the challenge of deployment and change management is presented, which requires complete top-management support from the beginning.

3.2.2.2 Use Case: Analysis of taxi movements in New York City

The analysis of the movement of taxis in NYC, the USA as published by FiveThirtyEight[25] in August 2015 combines data from the NYC Taxi & Limousine Commission (TLC) with Uber trip data (4.5 million Uber pickups in NYC), from April to September 2014. Both datasets were retrieved from the TLC as either csv- or xlsx-file. The data is openly accessible through Github[26]. The analysis aims at assessing road transport in the passenger sector. The directory consists of almost 93 million trips Uber and conventional taxis.

This public dataset allows the user to play with and implement various use cases. For instance, comparison of Uber taxi pickups throughout NYC, it claims that Uber surpasses conventional taxis[27]. In addition, the dataset was used to analyze the potential for using Uber for low-income families, as Uber is currently especially used by high-income households[28]. For this case, it proposes a combination model of using Uber and public transit to reduce car-ownership. While this use case can indicate the potential for reducing emissions while also improving people's access to transportation through car-sharing model, the real potential would increase by decreasing costs for car-sharing while simultaneously improving public transportation infrastructure. Going beyond data for just NYC, Uber itself offers various possibilities and tools to gain better insight into the passenger sector of road transport.

- The open-source **deck.gl Framework[29]** designed for the advanced visualization of large data sets and well suitable for the analysis of Uber data. It offers a complete architecture for reusable JavaScript Layers, visual overlays generally used for maps. The framework is applicable to historical and real-time insights from large, complex data sets.
- The traffic analysis tool **Uber Movement[30]** provides anonymized data from over two billion trips aiming to improve and support urban planning globally. It partly gives insight into Uber's internal demand and usage data. The aggregated dataset help decisions for adaption of infrastructure and more efficient city planning. The data is anonymized to meet privacy guidelines for passengers and Uber drivers. Movement data is freely available. Currently, there is no programmatical access via API to Movement. The data is available for several cities such as Boston, NYC, Manila, Sydney or Bogotá, however service for more cities will be provided gradually. The user can show travel time between

[25] http://data.beta.nyc/dataset/uber-trip-data-foiled-apr-sep-2014, Accessed February 15, 2018
[26] https://github.com/fivethirtyeight/uber-tlc-foil-response, Accessed February 15, 2018
[27] https://fivethirtyeight.com/features/uber-is-serving-new-yorks-outer-boroughs-more-than-taxis-are/, Accessed February 15, 2018
[28] https://fivethirtyeight.com/features/public-transit-should-be-ubers-new-best-friend/, Accessed February 15, 2018
[29] https://uber.github.io/deck.gl/#/, Accessed February 15, 2018
[30] https://movement.uber.com/cities?lang=en-US, Accessed February 15, 2018

any two neighbourhoods in the respective cities. In addition to the map data, charts are provided breaking down travel times by time of day of a weekday. In addition, time series data is openly available in CSV. [31]

Identified opportunities

In general, the datasets provide the substantial potential for assessing models for decreasing gas emissions, meeting transport demand, and also assess socioeconomic aspects of transportation such as analysing data in predominantly low-income neighbourhoods. Opportunities further are using the data to help reduce congestion and improve overall transportation access and efficiency.

3.2.2.3 Strava Metro – Data-Driven Bicycle and Pedestrian Infrastructure Planning

Strava Metro, launched in 2014, is a US-based developer of mobile and online services and applications for athletes (especially running and cycling) worldwide. Their main application uses crowd-sourced data to improve infrastructure for bicycle and running (Musakwa & Selala, 2016). The crowd-sourced data is provided by the users via the Strava App by uploading rides and runs with their smartphone or GPS device and anonymized and aggregated for further analysis. The mobile application records data of each activity for time, date, distance, average speed and route. In addition, users may add textual information. The database holds a trillion GPS points globally increasing by 8 Million activities weekly. The data is marketed to different transportation stakeholders, purchasable for different license options (streets, nodes and origin and destination licenses) and in different formats (Musakwa & Selala, 2016).

Identified opportunities and challenges

Data from the Strava App serves as a basis for identifying athletic patterns and trends for evidence-based infrastructure planning in close relations with transportation planners, policy makers and various additional stakeholders (Musakwa & Selala, 2016). To meet the privacy concerns of their users, all personal user information, and private activities are excluded in the data set. Data is only provided in aggregated version and per-user-view on the data is not possible. [32]

As for the project results, as of November 2017, 125 organizations globally use the toolkit provided by Strava Metro, including departments of transportation in Colorado, Texas, New Hampshire and Copenhagen. Scientific use of the date, with actionable insights for transportation stakeholders in Johannesburg and South Africa, was also done by Musakwa and Selala (2016). In addition to the opportunity of improving infrastructure especially in urban areas, the data also reveals valuable socioeconomic insights. One needs to keep in mind, which the social network usually attracts good-income users and neighbourhoods with just a few GPS

[31] https://techcrunch.com/2017/08/30/uber-movement-traffic-data-finally-makes-it-out-of-beta/, Accessed February 15, 2018

[32] https://support.strava.com/hc/en-us/articles/216918877-What-is-Strava-Metro-, Accessed February 15, 2018

points are in need of improved pedestrian and cycling infrastructure even when the low level of activity might imply lower demand. [33]
The data is collected in English, however, the textual unstructured information such as tags and titles can be in any language, posing a challenge for further analysis for the latter. An additional challenge worth mentioning is also the possible bias during analysis in favour of app-users against the general population which must be considered when deriving decisions from the provided data (Musakwa & Selala, 2016).

3.2.2.4 Railigent – Siemens

Railigent[34] offers digital services for the rail industry as a cloud-based data analytics platform based on Siemens' IoT industrial operating system Mindsphere with the objective of improving rolling stock availability to almost 100% and signal infrastructure of the rail industry. Railigent services are founded on a cloud-based data lake (AWS), where data ingestion is performed in batches or streams and initial analytics models are applied for data validation and augmentation. Analytics Operations are performed either in sandboxes or in the full platform using Python and pySpark. [35]

Identified opportunities
Railigent offers a remote monitoring system capable of transmitting vehicle and infrastructure systems data to respective terminals in real time. The objectives are improvements in train availability, cost-effectiveness, and rail safety.[36] It includes services for remote analytics and maintenance services for prediction and prevention of faults as well as analysis of energy consumption (Brahimi et al., 2017). Additionally, cybersecurity and guidance services are included. Successful applications of Railigent are the cooperation with the Spanish rail organization Renfe, a Russian collaboration as well the application of Railigent for commuter trains in Bangkok integrating metro and infrastructure systems. In 2016, Deutsche Bahn (DB) has begun working with Siemens to incorporate Railigent in various DB entities.[37]

3.2.2.5 Mobilität digital Hochfranken – MobiDig (Digital Mobility)

The project Mobilität digital Hochfranken – MobiDig (Digital Mobility) is developed by a consortium led by the University of Hof and the Technical University in Munich. They collaborate with the counties and the town of Hof (2017-2020). The goal of the project is to provide short-term mobility supply/demand forecasts for optimal operational planning to facilitate resource-efficient public transport operation.

---

[33] https://www.citylab.com/transportation/2017/11/strava-metro-global-heatmap-urban-planning/545174/, Accessed February 17, 2018

[34] https://www.siemens.com/global/en/home/products/mobility/rail-solutions/services/digital-services/railigent.html#QA, Accessed February 17, 2018

[35] E-Mail from Gerhard Kress (responsible for Railigent at Siemens) 2/2/2018)

[36] https://www.siemens.com/global/en/home/products/mobility/rail-solutions/services/digital-services/railigent.html, Accessed February 17, 2018

[37] E-Mail from Gerhard Kress 20/02/2018

The addressed transport sector is public passenger transportation in peripheral regions. The project addresses the demographic change in the peripheral region. This comprises the increasing dependency on public transport in the aging population and at the same time, the reduction of available mobility caused by the shrinking population. Conventional public transport offers no cost-effective solution for this challenge.

The project utilizes data from OpenStreetMap, origins, and destinations for passenger transport (e.g. schools, shops, companies), aerial images, information about existing transport facilities (spatial and temporal), actual usage of transport resources, anonymous movement profiles and weather data. Part of the data is open.

Identified opportunities and challenges

The project will develop a mobility concept for peripheral regions ensuring the economic viability and ubiquitous availability of mobility services through the use of the state-of-the-art concept in passenger mobility and digitalization technologies. Challenges in the project are the acquisition of sufficient data supporting the prediction of transport needs under existing legal and technical constraints, the accuracy of predictions, and the implementation of a flexible transport system making use of these predictions.

3.2.3 University initiatives

3.2.3.1 MIT SENSEable City Lab

The **MIT SENSEable City Lab** is a research initiative for applying real-time analytics and big data methods for urban planning (passenger and freight transportation). Various big data projects in transportation are worth mentioning.

- **Global Mobility Index (2018):** The project shows how people move in 100 cities around the world. The visualization focuses on three aspects of urban mobility (1) congestion levels for real-time traffic-monitoring data, (2) commuting time and (3) estimated the percentage of trips that could be shared if citizens were to wait to share a trip. The objective is to free cities from congestion via car-sharing or bike-sharing usage and public transportation data.
- **Shareable Cities (2017)**: The project provides an analysis of share-ability for different sharing concepts: Based on data on millions of taxi trips in NYC, San Francisco, Singapore, Vienna share-ability curves for each city are computed. The objective is to help planners, transportation companies, and society to shape sustainable global growth[38].
- **Sensing Vehicle: The car as an ambient sensing platform (2017):** The project with VolksWagen Group of America Electronic Research Lab provides an analysis of driver behaviour and urban environment. Project objectives and value created are the

[38] http://senseable.mit.edu/shareable-cities/, Accessed March 22, 2018

reductions in car accidents, decrease stress for drivers and finally to understand how drivers make a decision in road traffic and thereby improve overall road safety.[39]

- **Roboat (2016):** The project provides design and test of autonomous boats in Amsterdam, where units can transport goods and people. The objective is to create temporary floating infrastructures, such as self-assembling bridges and concert stages.[40]
- **lightTRAFFIC (2016)[41]:** The project developed and analyzed slot-based intersections for their potential to replace traditional traffic lights, significantly reducing queues and delays.
- **hubcab (2014)[42]:** The project provided a visualization of 170 million taxi trips in NYC indicating the time and mode of taxi pick-up and drop-off. Project objectives were the identification of zones of condensed pick-up and drop-off activities as well as the analysis of taxi share-ability.

3.2.3.2 Goethe University Frankfurt Data Challenges

The Goethe University Frankfurt Big Data Lab[43] is collaborating with Deutsche Bahn and its internal IT partner DB Systel GmbH. The Deutsche Bahn AG, as one of the few comprehensive mobility suppliers, aims to shape the actual development towards networked, smart and sustainable mobility.

In the context of annual Data Challenges, students and Ph.D. students are encouraged to use the open data provided by Deutsche Bahn to create new business ideas for more efficient transportation. Attendees of the Data Challenge in 2016[44] "*Mobility and the Future*" explored innovative ideas and solutions in the area of passenger transportation to achieve an improved travel experience as an important aspect of future mobility experience.

The Data Challenge in 2018[45] focuses on "*Smart Cities and the design of future urban life"*. Questions addressed are the organization of smart and integrated mobility in different transport modes, the forms of smart logistics and new infrastructural services, various concepts for "mobility on demand" and prototypical demonstration of the potential of autonomous driving. Using the open data pool, attendees are expected to create new ideas and solutions for the questions addressed. Examples of such new ideas are the following reference implementations:

- Bahnhofsbox (picking up goods at the station)
- Cargo Bike (sharing system for bikes)
- Clever shuttle (provides door-to-door mobility using electro cars)

[39] http://senseable.mit.edu/vwsensing/, Accessed March 22, 2018
[40] http://senseable.mit.edu/roboat/, Accessed March 22, 2018
[41] http://senseable.mit.edu/light-traffic/, Accessed March 22, 2018
[42] http://hubcab.org/#12.00/40.7217/-73.9070, Accessed March 22, 2018
[43] www.bigdata.uni-frankfurt.de, Accessed March 22, 2018
[44] http://www.bigdata.uni-frankfurt.de/web-business-data-challenges-ws-2016/, Accessed March 22, 2018
[45] http://www.bigdata.uni-frankfurt.de/web-business-data-challenges-ss-2018/, Accessed March 22, 2018

- ioKi (autonomous car sharing service)

3.2.4 Data characteristics

Common characteristics of big data are identified among nineteen applied cases.

Traffic data: Traffic data or traffic monitoring data can be either historical or live data. The structured data are stored in relational databases, whereas the unstructured data are stored in a data lake such as Hadoop Distributed File System or NoSQL systems. Live data is also known as streaming data and is normally analysed in real-time only, not stored afterwards.

One example of traffic information data is the data of the German national railway company. Most of that data is available online via an open data platform[46]. This platform is used by various apps to display train live schedule data (departure and arrival times) including delays. Public transportation, flight and ship schedules and updates, traffic light data, video gate data at e.g. parking lots or airports, commuting times (car tracking) and information on rail track usage are other examples of this type of data.

Sensor data: Sensor data are used for live monitoring and long-term improvements. They are the most common data source in the transportation domain. Trains, airplanes, cars have multiple sensors installed. For instance, information on the GPS position, speed, fuel level, engine status, tire pressure and the temperature can be controlled by sensors.

This data can be used to trigger live events live or to store the data for later analysis. Sensor data usually are generated and stored as JSON or XML objects. To process these data types efficiently, specific tools and technologies are required.

Tracking data: Tracking data (also known as scanning information) are generated by various means (e.g. cargo, luggage, parcels, buses, taxis, and passengers). Thereby the various objects to be transported are tracked and traced throughout the entire logistics chain.

External Data: External data are seemingly irrelevant datasets. Environmental data, weather data are examples of data in this category. However, the combination of external data with the above-mentioned datasets provides opportunities for generating valuable insights into different external correlations for transportation. For instance, combining transport data with weather data, information on polluting emission or energy consumption data might lead the detection of patterns that allow for a better planning and scheduling of routes. A source of meaningful data can also be the repair reports of the repair shop. Scanning this information with OCR software might make this information available for further analysis.

Social network data: Social network data can also be historic or live. Social media data is used to understand the customers travel behaviour for a reliable analysis of transport demand and to avoid bottlenecks.

[46] http://data.deutschebahn.com/, Accessed February 1, 2018

Crowd-sourced data: Crowd-sourced data if anonymised data from mobile applications. Users of the applications can upload (GPS-)data about their e.g. tours, locations, rides or runs. This centralized collected data can is a valuable data source to explore the movement of people.

3.2.5 Summary

The reported nineteen applied cases are revealing the widespread interest and potential of using big data in the transport sector. Table 1 is a compact view of the nineteen applied cases. Each project is summarised in terms of transport modes, sectors, and data characteristics. Most projects were characterized by a collaboration of various stakeholders. The majority projects (eleven out of nineteen) cover the mode of the road, whereas only three projects cover the mode of air.

Table 1: Summary of applied cases, respective modes, sectors, and data characteristics

Project type	Project name	Mode					Sector		Data Characteristics					
		Air	Rail	Road	Urban	Water	Passenger	Freight	Sensor data	Tracking data	Social network data	Crowd-sourced data	Public Data	3rd Party Data
NGO / government	Transforming Transport	x	x	x			x	x	x	x	x		x	
	BigDataEurope			x			x		x		x		x	
	NOESIS	x	x	x	x	x	x	x	x	x				
	DATA SIM			x			x	x	x		x		x	
	OPTIMUM	x	x	x	x	x	x	x	x					x
	CAPACITY4RAIL		x				x	x	x					
	IN2RAIL		x				x	x	x				x	
	oneTRANSPORT			x			x	x		x			x	
	MUNIN					x		x		x				
	INONMAN²HIP					x		x		x				
	UN Global Pulse Jakarta			x			x		x	x			x	
Industry / research	ORION				x			x						
	Taxi Movement in NYC			x			x		x	x				x
	Strava Metro			x		x	x				x	x		
	Railigent		x				x	x	x					x
	MobiDig			x	x		x		x				x	
University	MIT SENSEable City Lab													
	Goethe University Frankfurt Big Data Lab		x	x	x		x	x		x				x

3.3 Literature review

In recent years, the use of big data in the transport sector has gained considerable attention amongst researchers. This section conducts a comprehensive, but not exhaustive, literature review to investigate its opportunities and challenges.

3.3.1 Opportunities

Numerous experiments are conducted by researchers. The authors regard the novel studies as opportunities due to the following two reasons: often the industry references many research papers for implementing an innovative application. Moreover, many policymakers can take the research papers to propose a data-driven policy. In general, four study types are identified in this report, namely transportation planning, traffic operations, safety, and carbon emissions reduction.

3.3.1.1 Studies for transportation planning

The ***understanding of passenger demand and behaviour*** is crucial for efficient transportation planning. Nakamura et al. analyse the influence of loyalty programs on user travel behaviour based on smart card data and survey results (Nakamura et al., 2016). Various data from fare collection, passenger counters, and mobile devices are utilized for behavioural insights into the movement of urban individual passengers. Of particular interest in this topic, domain is exploring factors that impact a change in travel behaviour such as monetary aspects or analyse factors to promote a change in travel behaviour. (Sánchez-Martínez & Munizaga, 2016)

An important aspect of transport analysis is the ***understanding of OD*** data for analysis of the transport system. In transport literature, OD data are analysed as matrices (Munizaga & Palma, 2012). Munizaga and Palma use smartcard data from Santiago, Chile to develop an estimated multimodal OD matrix for public transport. The model generates estimations for spatial and temporal characteristics of public transportation. These OD matrices are later validated by Pineda et al. (2016) using estimates from a large-scale OD survey to determine passenger flows, loads, trips and transfer stations for Santiago.

Cui et al. (2016) aim at better understanding ***travel demand modelling*** and transport network services using GPS data. For this, data for urban mobility patterns are modelled based on GPS data from operating taxis in Harbin, China for the time period July until September 2013. The patterns capture central traffic features in the region such as travel speed, demand, and direction of travel routes. The model measures interregional road performance based on the indicators travel demand, temporal travel efficiencies and geometric properties of the road connection. Thereby, mismatches between travel demand and transport network services are identified revealing efficiency problems at specific periods of a day as well as problematic regions. The model proposed by Cui et al. can easily be applied to other regions as its effectiveness is shown suitable for spatial and temporal road network analysis based on actual travel demand. Their approach takes advantage of the potential lying in GPS data to monitor individual traffic demand and the possibility to identify poor performance regions in an urban situation, an understanding of the road conditions and interregional travel routes.

For the purposes of **better-understanding passenger flow** in urban areas, Zhiyuan et al. (2017) use big data visualization for passenger flow analysis of urban rail transit (URT) networks for the case of the Shanghai Metro. The analysis is done for different aspects (line, region, turnover capacity, load factor, network) for the purposes of analysing and controlling the distribution and change in passenger flow. A further application for smart transportation management is proposed by Wang et al. (2016) who present a big data approach to support smart transportation for bus transport in the case of Fortaleza, Brazil computing bus travel time and passenger demand.

Linares et al. (2017) introduce an analysis of **different transportation models** based on transport demand and dynamic ridesharing. The authors evaluate various possible uses for on-demand multi-passenger ride-sharing in urban areas, showing some positive impacts for improving travel time and reducing traffic flow.

3.3.1.2 Studies for traffic operations

Sánchez-Martínez and Munizaga (2016) point towards the application of big data ***for real-time monitoring of transit operations and incident management.*** They use vehicle location data from location sensors installed in vehicles. For this, location data can be streamed to data control centres for analysis of driver behaviour. Further, control centres may use data to respond to incidents e.g. by dispatching police or providing emergency maintenance. Data can also be used for ***reduction of waiting times***.

Julio et al. (2016) introduce an approach for **real-time prediction of bus travel speeds** based on machine learning algorithms. Objectives are improving quality and reliability of bus services and increasing effectiveness of traffic control. Various machine learning algorithms, namely Artificial Neural Networks, Support Vector Machines and Bayes Networks, are implemented and compared for the prediction of bus speeds using real-time data on traffic conditions.

Sánchez-Martínez et al. (2016) aim at ***maximizing service performance of public transport services*** for a network of high-frequency bus routes. Based on automatically collected data and simulated modelling, using data from the Boston metropolitan area, traffic performance is evaluated at each route for variable fleet sizes. The model is combined with a greedy algorithm for optimal allocation adjustment. This method is suitable for improving service quality as well as fleet utilization.

The aspect of **improving transport performs** the efficiency of the traffic management system of ship trajectories was addressed by Gan et al. (2018). The proposed trajectory length prediction for intelligent traffic signalling. For their approach, they used historical trajectories data grouped by a clustering algorithm, modelled relationships between variables such as ship speed, freight capacity, weight, maximum power, length, width, and water level and their respective membership using neural networks. Ultimately, their approach is shown to improve traffic control signalling.

Guerreiro et al. (2016) also propose a big data approach for improvement of **performance in traffic services.** They present an Extract Transform Load (ETL) architecture for Intelligent Transportation System (ITS) for a dynamic toll system on highways. The architecture includes predicting functionalities and is capable of processing historical and real-time data.

In the context of traffic operations, big data applications promise various possibilities for facing ***traffic congestion***. Tang and Heinimann (2018) provide a resilience-oriented approach for quantitatively assessing recurrent congestion in urban areas – in contrast to non-recurrent congestion which refers to congestion stemming from incidents. Traffic congestion not only brings along delay and inconvenience but is also associated with road safety issues and negative impacts on greenhouse emissions. The authors use the concept of resilience which refers to a system's capability to react to disturbances and recover and aim to quantify recurring congestion based on traffic patterns on urban streets and freeways. Their approach reveals opportunities to test the approach in other regions. Challenges faced in this approach were located in the vehicle trajectory data used, i.e. its availability, cleanness or the instability found in the sampling rate GPS sensors. Their metric was shown as effective to identify patterns of recurrent congestion.

Wemegah and Zhu (2017) analyse challenges such as congestion in their **evaluation of traffic volume** a count in Nanjing, China with Radio Frequency Identification (RFID) data. They address various challenges revealing peak hours, off-peak traffic volume count, regions of high and low volume traffic and provide tools for congestion analysis.

In urban areas, so-called ***intelligent monitoring and recording system*** (IMRS) as a processing platform for front-end image acquisition and back-end data processing are applied in urban areas. Xia et al. (2016) combine IMRS with HBase for data analysis in contrast to relational databases with the objective to increase performance and efficiency.

Melo et al. (2017) studied the smart city and the traffic optimization. The study demonstrates the positive effects of smart re-routing such as saving travel time and improving road network efficiency. It then evaluates the traffic performance, costs, and environmental conditions in Lisbon, Portugal.

3.3.1.3 Studies for transportation safety

Liu et al. (2016) provide a methodology for ***improving truck safety*** at railway crossings based on local traffic information. They investigate direct and indirect relationships between driver behaviour immediately before truck collisions and combine path analysis with spatial analysis. The authors target a provision of a benchmark for the real-time identification of risky vehicles, the evaluation of control devices located at railroad crossings and ultimately the identification of problematic regions.

Jian et al. (2016) also study the potential of big data for improving **road safety** by investigating high-risk crash regions analysing big data following the Random Forest (RF) technique. RF is based on a collection of decision trees with controlled variations (Breiman, 2001). For their analysis of high-risk crash regions, the Jian et al. identify variables that determine crash risk and evaluate crash risk by collision type and injury level.

Big data applications are also useful for ***fault detection in transportation systems*** (Liu, Li, & Zio, 2017) as shown by Liu and Zio (2018). The authors introduce a method based on a scalable fuzzy support vector, for detecting faults in transportation systems implemented in high-speed trains. The authors calculate fuzzy membership values in their dataset based on K Nearest Neighbours (KNN). The objectives are to overcome data challenges such as imbalanced data on normal or

faulty conditions, data noise, outliers. The proposed method is tested with various benchmark methods on five public datasets and then used for a case study for fault detection in a braking system of high-speed trains. KNN method is used for the identification of data outliers and noise.

Vehicular ad hoc networks (VANETs) pose additional opportunities for addressing safety challenges (Plößl & Federrath, 2008). VANETs represent mobile ad hoc networks (MANETs) attributed to vehicles. Najada and Mahgoub (2016) analyse real-time big data from the Florida Department of Transportation on road accidents for the potential of a system to anticipate and **alert of congestion, and road accidents** in order to ultimately reduce them.

3.3.1.4 Studies for carbon emissions reduction

Gennaro et al. (2016) analyse **big data for supporting low-carbon transport policies** tailored to the road transport in Europe. GPS data from conventional fuel vehicles with onboard location devices were used for their pilot study and algorithm development. It also investigates large-scale mobility statistics and electric vehicle analysis in terms of potential, energy demand, and data on emissions.

Li et al. (2018) analyse opportunities for **reducing carbon emissions**. Individual travel pattern data are evaluated considering urban differences in aspects such as economic development, population, and factors for change in individual passenger travel behaviour. Activity types, energy intensity, and carbon intensity are assessed in a model for 288 cities in China feeding into a scenario development for energy use in Chines urban areas between 2010 and 2050. Li et al. give recommendations based on their analysis, namely, that policies on a national level targeting individual travel behaviour are shown to have a reducing impact on carbon emissions, policies optimizing the environment for vehicle sharing solutions have a stabilizing impact on greenhouse emissions for the short term and promoting electric vehicles are shown to be useful for long-term sustainability.

3.3.2 Challenges

The literature review revealed that the major barriers to big data application in transportation are data silos, data ownership issue, data privacy, and the lack of data quality and standards. One of the most important challenges, however, a lack of expertise of technical knowledge.

According to the study by (Sánchez-Martínez & Munizaga, 2016), ***big data access*** and **big *data utilization*** are the most critical challengesGenerally, the various traffic data is collected by various organizations, ranging from private organizations to public and government organizations. Therefore, full exploitation of available data would require all these different organizations to provide access to their data and share it. **Data ownership** is also mentioned as another critical challenge. Data owners are transportation operators or the agency for Automatic Vehicle Location (AVL) data, Automated Fare Collection (AFC) data or Automatic Passenger Counting (APC) data. Operators or agencies may hesitate to share data concerned that their data might be used by their competitors. Also, when data is owned by a private data collector, data is not publicly available due to the data's monetary value. In various cases, when data is shared with researchers, a non-disclosure agreement is added, inhibiting the full use and

publication of the resultant research. An additional factor, hindering data publication are legal concerns in terms of ***data privacy***, especially in the case of data and origins and destinations of trips, or fare transaction data. , especially in the case of data and origins and destinations of trips, or fare transaction data.

In terms of technical challenges, Sánchez-Martínez and Munizaga conclude that ***poor data quality*** stemming from faults in data sensors and errors in manual processes is one major challenge for data cleaning and therefore for drawing valuable insights from the data. Other identified challenges are the ***lack of standards*** regarding hard- and software, dealing with ***unusual events*** (sports, weather, societal) when analysing average traffic performance, and the acknowledging of ***data bias*** when analysing sample data. Finally, the same authors see the ***lack of expertise*** as one of the major inhibiting factors for exploiting big data opportunities, lacking skills in database design and computer programming.

Giest (2016) also addresses the challenges observed in the analysed initiatives for data-driven policy-making and reduction of carbon emission. It is pointed out that the obstacles for municipalities and government do not lie in the technology, more in the ***complexity of stakeholder number and their interconnectedness***. Projects based on cooperation with other organisations are faced with ***missing cross-departmental and cross-level connection*** hindering ***data exchange and compatibility***, namely due to missing standardisation connected to missing regulations. Further, all initiatives are greatly dependent on organisations from the private sector due to the lack of technical expertise and analytical skills within government and municipalities.

Gennaro et al. (2016) see the major technical challenges of big data application in data structure of the input format and data definition for output to allow for **easy exchange across different systems** (velocity), the **selection of the optimal algorithm for processing high volumes of data** with an acceptable computational burden (volume), the post-processing of data to make them easily understandable but still keep its informational complexity and finally the link between data and the potential **insight** (value).

4 Data Sources

4.1 Sources of big data

Currently Substantial Transport data are generated. This is mainly due to the advancement of Information and Communications Technology. This report identifies four data sources as follows: Route-based data, Vehicle-based data, Traveler-based data and Wide area data.

Route-based data are collected **by sensors** at fixed locations of a path such as a highway or a train. One of the most used sensors is a loop detector. Numerous loop detectors are harnessed for monitoring intersection traffic, as well as detecting incidents, classifying vehicle and re-identifying vehicle.

Vehicle-based data are collected **by mobile devices** or in-vehicle GPSs. Whereas the route-based data are collected at a specific location, vehicle-based data are dynamic such as data of route choice, travel time estimation and more. In particular, connected vehicle technologies enable vehicles to share data in real-time with other vehicles and the transport system.

Traveler-based data are collected **by people**. For example, traffic jams are inferred from one's location and accidents are voluntarily reported by mobile device users.

Wide area data are collected **by sensor networks** to monitor traffic flow. Unmanned aircraft and space-based radar fall into this type of data source. They generate photos and video data.

4.2 Traffic data collection techniques

Traffic data can be obtained by numerous data collection techniques. They are technically diverse and play an important role to classify the type of data sources. Thus, this section takes a closer look at the data producer and understand how data are actually obtained. Choice of a particular data collection methodology depends on the end needs and purposes of the data as well as on available resources. Large-area patterns of traffic flow may not need real-time information, whereas incident alerts, trip plans, and congestion reports would. Also, different sources of data can be combined to enhance information and/or minimise the cost of data collection. Conventional traffic data sources are defined here to include road surface traffic sensors, household travel surveys, floating car surveys and traffic cameras, including automatic number plate recognition (ANPR).

Mobile devices: Mobile, such as in-vehicle GPS and mobile device, offers the largest scope of data of traffic volume, speed and OD flows. The advantage of this generator is inexpensive.

Household travel survey: Household travel surveys (HTS) collects data for a long-term strategic network planning and policy analysis. HTS collects OD trip patterns by transport mode, trip purpose and time of day, and household demographic information to model travel choices.

Traffic camera: Traffic camera is utilized for multiple purposes. It monitors traffic conditions on key arterials, speed and red-light enforcement, adaptive signal control, and incident detection and management. It is an essential element of the Smart Infrastructure/Intelligent ITS and it is implemented on highways around the world. A network of roadside traffic cameras can provide information on vehicle movement.

Traffic sensor: There is two type of traffic sensors: Roadside and Rode. Road-side traffic sensor aims to detect vehicles on the highway. Compared to the road traffic sensor, it provides less intuitive. Examples of road-side traffic sensors are as follows: Acoustic sensors, Microwave radar, Lidar and active infrared sensors, Video image detection, and RFID tags. On the other hand, road-traffic sensor records traffic, classifies vehicle, vehicle speeds and provide real-time information to traffic management systems. It is either embedded in the pavement or laid on the pavement surface. Single sensor record only vehicle movements, whereas sensor networks are able to measure vehicle speeds. Thus, In general, it is installed on a high traffic road. In addition, the installation and repairs of in-pavement road sensors require intrusive roadworks. Examples of road traffic sensor are as follows: Inductive loops, Pneumatic tubes, Piezo-electric sensors, and Vehicle classifiers.

4.3 Traffic data providers

This section provides a comprehensive survey of transportation traffic data providers in a global level. The objective is to have a better understanding of transportation data with real-world use cases. For one who wants to delve deeper into, the authors encourage one to reference LeMO Task 3.2. Table 2 is a compact view of twelve traffic data providers. It shows that a successful transportation service needs to harness numerous data sources and methodologies.

Table 2: Characteristics of traffic data providers

Traffic data provider	Service					Data source				Collection technique			
	navigation	traffic	transport	tracking	fleet management	route-based	vehicle-based	traveler-based	wide area	household travel survey	mobile devices	traffic camera	traffic sensor
HERE	x					x	x						x
TomTom	x					x	x		x		x		x
Google Maps	x						x	x			x		x
INRIX	x	x				x	x		x		x		x
StreetLight			x			x			x		x		x
AirSage				x				x			x		
Logica		x				x			x	x			x
Traffic Network Solutions		x							x				x
Austraffic		x	x						x	x			x
GridTraq				x	x				x		x		
Mercurien				x			x				x		
Optus	x	x					x		x		x		

HERE (www.here.com): HERE provides mobile navigation services, including 3D maps, live traffic, and public transport information. According to HERE, it collects data from more than 80,000 sources over 196 countries and one billion mobile devices are using the service. HERE users NAVTEQ (www.navteq.com) digital maps, which are part of Nokia's Location and Commerce division. The data inflows are depicted in Figure 5 (HERE). The distribution of traffic data indicates the number of users.

United States	16.89%
Brazil	9.25%
Germany	5.11%
Mexico	4.76%
United Kingdom	3.59%

Figure 5: Traffic of data inflow for HERE website by countries[47]

TomTom (www.tomtom.com): TomTom is a Dutch manufacturer of automotive navigation systems. Its products are as follows: in-vehicle navigation devices, in-dashboard navigation and vehicle control services, and navigation software for mobile devices. TomTom also offers a range of navigation services, such as traffic alerts, weather updates, speed camera locations (TomTom PLUS), travel conditions and congestion (TomTom HD Traffic), and fastest route (TomTom IQ Routes). The data underpinning TomTom's traffic data services come from a range of sources including TomTom navigation device GPS data, GSM data, and Traditional road agency data. It also regularly produces regional urban traffic congestion indexes for cities in Australia and New Zealand rep. The data inflows are depicted in Figure 6 (TomTom).

[47] https://www.similarweb.com/website/here.com, Accessed April 15, 2018

Figure 6: Traffic of data inflow for TomTom website by countries[48]

Google Maps (maps.google.com) including **Waze** (www.waze.com): Google provides colour-coded maps of traffic speeds across roads. The data underlying the maps are sourced from third-party data sources and crowd-sourced data. Waze collects crowd-sources traffic information from users and provides real-time traffic information back to users. Waze also allows users to report accidents, police traps or any other hazards. Google purchased Waze in June 2013 when it had over 50 million users worldwide. The data inflows are depicted in Figure 7 (Google Maps) and Figure 8 (Waze).

[48] https://www.similarweb.com/website/tomtom.com, Accessed April 15, 2018

United States	32.36%
France	5.81%
United Kingdom	4.79%
Brazil	4.55%
Germany	4.51%

Figure 7: Traffic of data inflow for Google Maps website by countries[49]

United States	20.63%
Brazil	13.03%
Israel	7.24%
France	6.30%
Slovakia	3.39%

Figure 8: Traffic of data inflow for Waze website by countries[50]

[49] https://www.similarweb.com/website/maps.google.com, Accessed April 15, 2018
[50] https://www.similarweb.com/website/waze.com, Accessed April 15, 2018

INRIX (www.inrix.com): INRIX is one of the leading providers of traffic services in the US and Europe. INRIX offers live traffic information, direction and driver services and flexible developer apps and tools. According to the company website, INRIX currently serves more than 200 customers and industry partners, including the Ford Motor Company, MapQuest, and Microsoft. INRIX's traffic information covers one million miles of roads in North America and one million kilometers in 28 European countries. The Real-time Traffic Flow services are based on road sensor data and mobile devices. Based on the past traffic flows, INRIX predicts future traffic flows per day, season, holidays. It also forecasts weather, accidents and road construction, as well as other events such as school schedules, sports games, and concerts.

StreetLight Data (www.streetlightdata.com): StreetLight Data is a data analytics company, providing readily usable information about transportation and mobility. It uses multiple data sources for its products including Archival anonymised cellular data, Traffic flow data, Archival anonymised GPS data, US census and other demographic data, Geospatial yellow pages, and Open source maps.

AirSage (http://www.airsage.com): AirSage collects mobile device signal data, and transforms into anonymised information. It contains when and where people are moving. It offers the following services: OD trip matrices for specified geographic areas, OD commuting trip matrices for specified geographic areas, and Route-based OD matrices for specified geographic areas. It has important relationships with two major national US wireless carriers.

Logica (www.logica.com.au): The core business model of Logica is developing and implementing integrated transport IT solutions. For example, LogicaCMG developed a Mobile Traffic System in 2004 to monitor traffic speed and provided road authorities with the ability to manage traffic flows and congestion (Steenbruggen et al. 2012). The system was tested in the Noord-Brabant region in the Netherlands and validated against data from floating cars, loop detectors, and number plate surveys. Data are anonymised and aggregated. Anonymised mobile data in California are provided by their core partner AirSage. The services currently appear to be limited to the United States.

Traffic Network Solutions (www.deepbluesensor.com and www.trafficnow.eu): Traffic Network Solutions is a provider of Bluetooth road-side sensor equipment. It offers a virtual traffic control centre service.

Austraffic (www.austraffic.com.au): Austraffic provides traffic and transport data collection services for government and private sector clients. It collects data via OD survey, travel time surveys and delays studies. OD survey utilises state-of-the-art video surveillance technology.

GridTraq (www.gridtraqcentral.com.au): GridTraq is an Australian service provider that provides vehicle tracking and fleet management services. It uses a proprietary GSM/GPS/GPRS (General Packet Radio Service) hardware and NAVTEQ digital.

Mercurien (www.mercurien.com.au): Mercurien is an Australian service provider that captures in-vehicle generated data. It has a partnership with Telstra to develop cloud-based traffic management systems.

Optus (www.optus.com.au): Optus provides Trafficview to its subscribers via SMS alerts and offers alternative routes. It utilizes Cellular Floating Vehicle Data (CFVD) to generate data.

Intelematics Australia (www.intelematics.com.au): Intelematics is a wholly owned member of the RACV Group, focussed primarily on in-vehicle telematics systems, including dash-based information and entertainment.

Telstra (www.telstra.com.au): Telstra provides networking solutions across a range of applications/industries, with particular relevance to fleet management. Telstra also owns the Whereis traffic navigator, through a partnership with Australian Traffic Network. (Telstra announced in 2011 it would partner with Mercurien to develop cloud-based traffic management systems.)

Vodafone (www.vodafone.com.au): Vodafone is the third major mobile telephony network and service provider. It is not clear whether Vodafone is presently active in the mobile device/traffic data space.

Yahoo Weather API: Yahoo Weather API allows us to get current weather information for your location. It makes use of YQL (Query Language) Query a SQL-like language that allows obtaining the meteorological information, the API is exposed like a service REST (Representational State Transfer) and returns the information in a data structure JSON. The data are updated every 2 seconds.

APPmobile: APP for mobile phones developed ad-hoc for the Transforming Transport project with the aim of collecting data about origin and destinations plus travel time. The technology used: REST. Real-time frequency.

CCTV cameras: Last pictures collected by the CCTV cameras deployed on the roads that show the current status of the traffic flows in real time. The technology used: Web service - DATEX II. Real-time frequency.

CCTV – Highway: Videos stored by CCTV cameras deployed on the roads. They show status traffic flows offline. MP4 videos.

Speed Radar: Location of the speed radars (both, fixed and mobile locations) deployed by the Authorities (National, Regional and Local) along the road network. The technology used: Web service - DATEX II. Real-time frequency.

Traffic Events: Traffic events on the road that have an impact on the road traffic, such as road works, accidents, traffic congestions, etc. The technology used: Web service - DATEX II. Real-time frequency.

VariableMessageSign: Information showed on the Variable Message Signs which can have an impact on the road traffic. The technology used: Web service - DATEX II. Real-time frequency.

4.4 Data flows in transportation modes

4.4.1 Cartography

A cartography is selected to visualize the data flows in the transport sector. The data flow of five modes (air, rail, road, water and multimodal) were measured among the USA, UK, Germany, France, and Australia. This includes the average traffic inflow of major transportation-related websites for each mode respectively by countries in the time period August 2017 to January

2018. The authors believe that this analysis provides a comprehensive view of global cross-country data flow as shown in Figure 9 to Figure 13. The raw data used for the visualizations are summarized in Table 3.

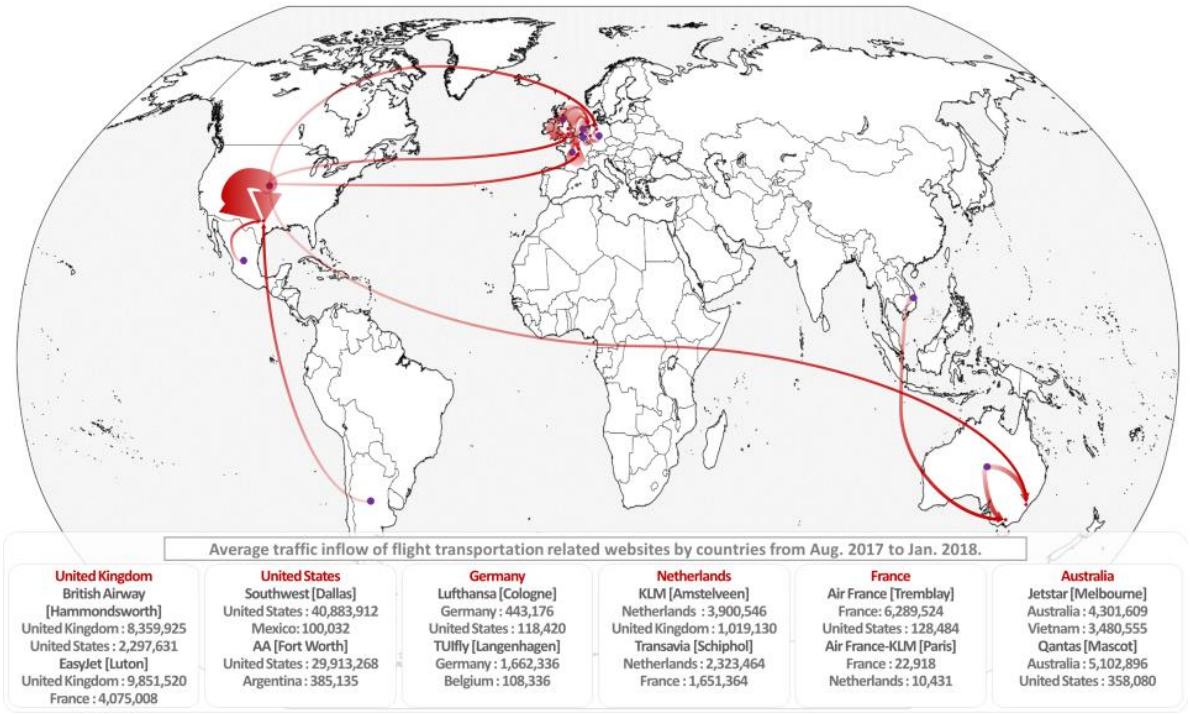

Figure 9: Average traffic inflow of air transportation-related websites by countries

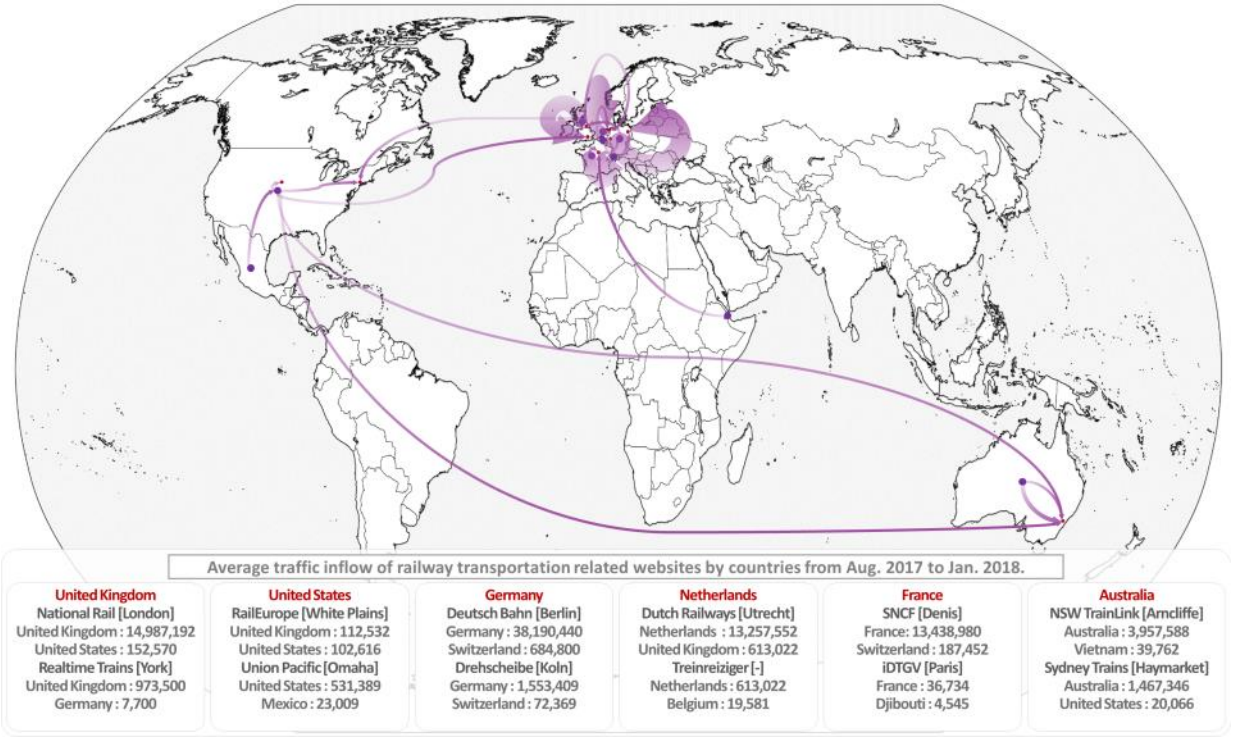

Figure 10: Average traffic inflow of rail transportation-related websites by countries

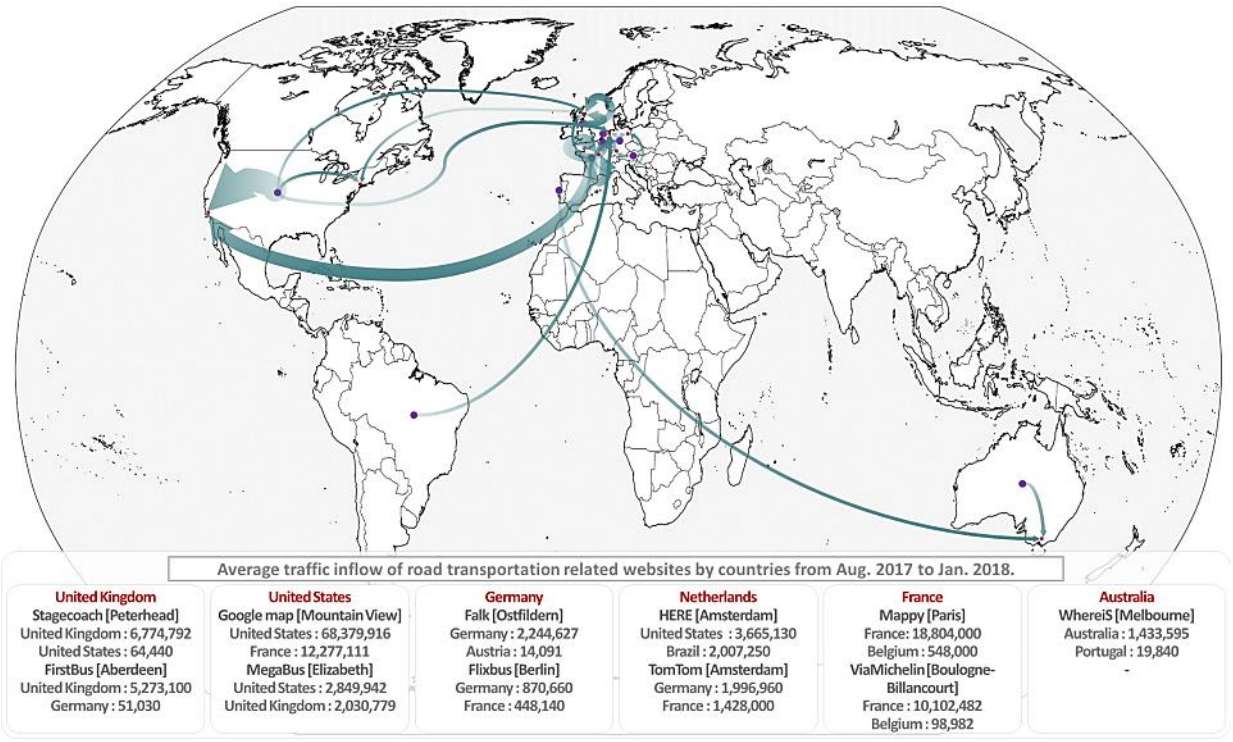

Figure 11: Average traffic inflow of road transportation-related websites by countries

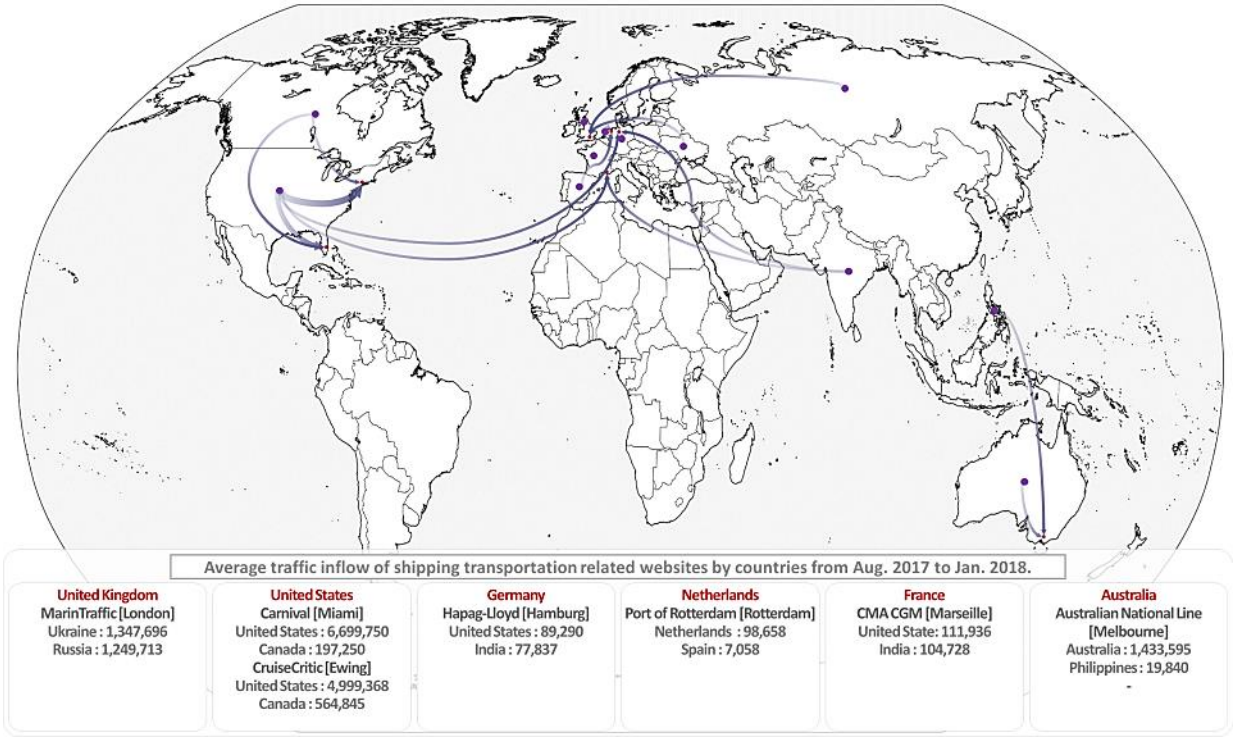

Figure 12: Average traffic inflow of water transportation-related websites by countries

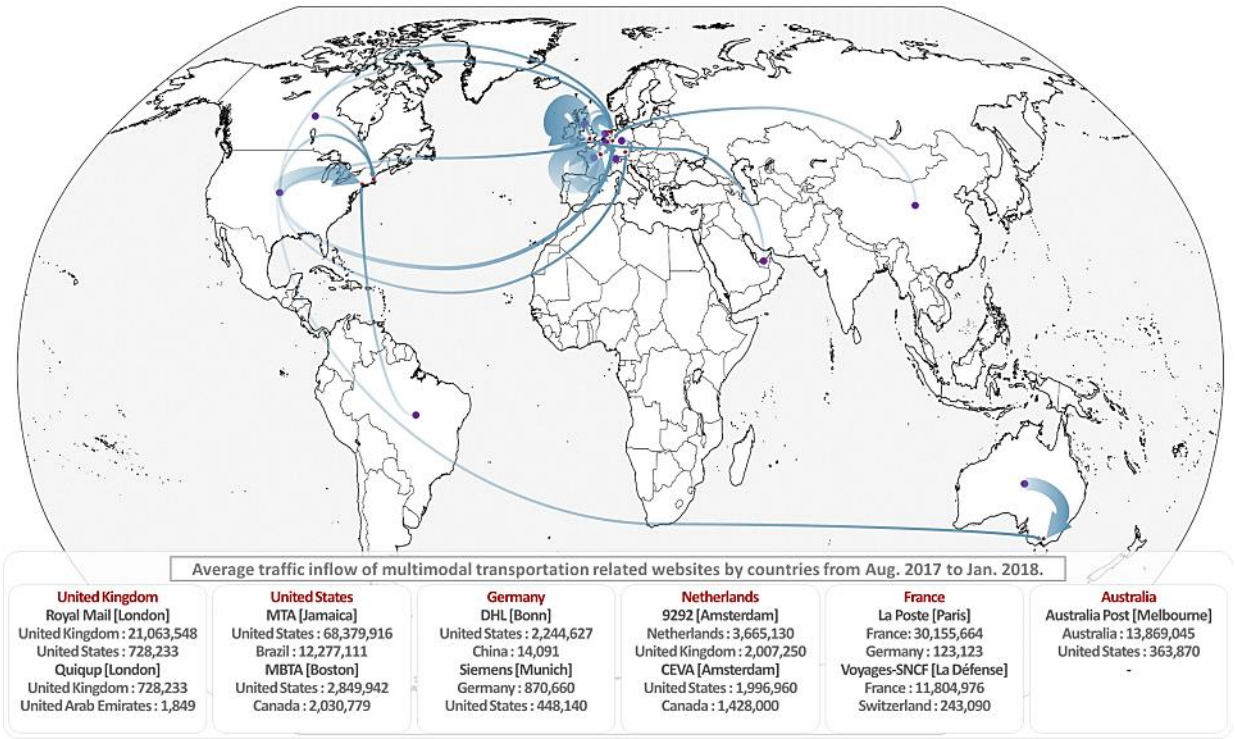

Figure 13: Average traffic inflow of multimodal transportation-related websites by countries

Table 3: Traffic flows of transportation-related websites

No.	Country	Company	Mode	Sector	Rank		Average traffic		Visitors in the last 6 months (2/26/2018)									
					In country	In global	In country	In global	1st	%	Traffic	2nd	%	Traffic	3rd	%	Traffic	Others
1	United Kingdom (UK)	British Airway	Flight	Passenger	199	3,195	8,359,925	15,410,000	UK	54.25	8,359,925	US	14.91	2,297,631	DE	2.16	332,856	28.68
2		easyJet	Flight	Passenger	197	2,226	9,851,520	26,880,000	UK	36.65	9,851,520	FR	15.16	4,075,008	IT	9.46	2,542,848	38.73
3		National Rail	Railway	Passenger	176	4,378	14,987,192	16,060,000	UK	93.32	14,987,192	US	0.95	152,570	DE	0.6	96,360	5.13
4		Realtime Trains	Railway	Passenger	2,000	47,032	973,500	1,000,000	UK	97.35	973,500	DE	0.77	7,700	NL	0.37	3,700	1.51
5		Stagecoach London	Road	Passenger	520	14,049	6,774,792	7,160,000	UK	94.62	6,774,792	US	0.9	64,440	DE	0.48	34,368	4
6		FirstBus	Road	Passenger	680	17,476	5,273,100	5,670,000	UK	93	5,273,100	DE	0.9	51,030	US	0.86	48,762	5.24
7		MarinTraffic	Shipping	-	-	3,291	209,526	17,190,000	UA	7.84	1,347,696	RU	7.27	1,249,713	US	6.63	1,139,697	78.26
8		Royal Mail	Multimodal	Freight	134	2,922	21,063,548	26,290,000	UK	80.12	21,063,548	US	2.77	728,233	IT	1.39	365,431	15.72
9		Quiqup	Multimodal	-	56,274	670,565	63,741	75,460	UK	84.47	63,741	UAE	2.45	1,849	FR	2.25	1,698	10.83
10	United States (US)	Southwest	Flight	Passenger	209	1,195	100,032	41,680,000	US	98.09	40,883,912	MX	0.24	100,032	CA	0.23	95,864	1.44
11		AA	Flight	Passenger	298	1,541	385,135	33,490,000	US	89.32	29,913,268	AR	1.15	385,135	CA	0.99	331,551	8.54
12		RailEurope	Railway	Passenger	21,570	109,260	102,616	768,660	UK	14.64	112,532	US	13.35	102,616	FR	5.78	44,429	66.23
13		Union Pacific	Railway	Freight	24,285	119,453	531,389	639,150	US	83.14	531,389	MX	3.6	23,009	IN	1.35	8,629	11.91
14		Google map	Road	-	236	421	68,379,916	211,310,000	US	32.36	68,379,916	FR	5.81	12,277,111	UK	4.79	10,121,749	57.04
15		MegaBus	Road	Passenger	4,237	11,752	2,849,942	6,010,000	US	47.42	2,849,942	UK	33.79	2,030,779	CA	8.73	524,673	10.06
16		Carnival	Shipping	Passenger	1,133	6,250	6,699,750	7,500,000	US	89.33	6,699,750	CA	2.63	197,250	BR	1.01	75,750	7.03
17		CruiseCritic	Shipping	Passenger	2,432	10,766	4,999,368	6,530,000	US	76.56	4,999,368	CA	8.65	564,845	UK	1.03	67,259	13.76
18		MTA	Multimodal	Passenger	1,110	6,118	14,977,830	16,100,000	US	93.03	14,977,830	BR	0.81	130,410	UK	0.76	122,360	5.4
19		MBTA	Multimodal	Passenger	6,704	37,372	1,902,964	1,960,000	US	97.09	1,902,964	CA	0.36	7,056	UK	0.34	6,664	2.21
20	Germany (DE)	Lufthansa	Flight	Passenger	463	5,407	443,176	1,240,000	DE	35.74	443,176	US	9.55	118,420	IT	5.01	62,124	49.7
21		TUIfly	Flight	Passenger	1,510	32,632	1,662,336	2,220,000	DE	74.88	1,662,336	BE	4.88	108,336	NL	4.59	101,898	15.65
22		Deutsche Bahn	Railway	Passenger	37	1,259	38,190,440	42,800,000	DE	89.23	38,190,440	CH	1.6	684,800	NL	1.11	475,080	8.06
23		Drehscheibe-online	Railway	Passenger	1,134	24,936	1,553,409	1,870,000	DE	83.07	1,553,409	CH	3.87	72,369	AU	3.66	68,442	9.4
24		Falk	Road	Passenger	529	15,544	2,244,627	2,310,000	DE	97.17	2,244,627	AT	0.61	14,091	CH	0.46	10,626	1.76
25		Flixbus	Road	Passenger	2,670	19,676	870,660	4,200,000	DE	20.73	870,660	FR	10.67	448,140	NL	7.81	328,020	60.79

No.	Country	Company	Mode	Sector	Rank		Average traffic		Visitors in the last 6 months (2/26/2018)								
					In country	In global	In country	In global	1st	%	Traffic	2nd	%	Traffic	3rd	%	Traffic	Others
26		Hapag-Lloyd	Shipping	Freight	-	82,025	31,222	729,490	US	12.24	89,290	IN	10.67	77,837	DE	4.28	31,222	72.81
27		DHL	Multimodal	Freight	-	3,239	1,066,148	24,910,000	US	15.97	3,978,127	CN	6.34	1,579,294	IT	5.26	1,310,266	72.43
28		Siemens	Multimodal	-	2,060	7,944	1,312,710	9,870,000	DE	13.3	1,312,710	US	11.76	1,160,712	IN	5.96	588,252	68.98
29	Netherlands (NL)	KLM	Flight	Passenger	153	4,920	3,900,546	12,660,000	NL	30.81	3,900,546	UK	8.05	1,019,130	DE	6.41	811,506	54.73
30		transavia	Flight	Passenger	308	11,813	2,323,464	5,720,000	NL	40.62	2,323,464	FR	28.87	1,651,364	ES	4.1	234,520	26.41
31		Dutch Railways	Railway	Passenger	45	5,626	13,257,552	14,280,000	NL	92.84	13,257,552	UK	1.1	157,080	DE	0.94	134,232	5.12
32		Treinreiziger	Railway	Passenger	1,721	117,156	613,022	668,290	NL	91.73	613,022	BE	2.93	19,581	DE	2.14	14,301	3.2
33		HERE	Road	-	-	2,104	-	21,700,000	US	16.89	3,665,130	BR	9.25	2,007,250	DE	5.11	1,108,870	68.75
34		TomTom	Road	-	-	5,224	1,386,560	11,200,000	DE	17.83	1,996,960	FR	12.75	1,428,000	NL	12.38	1,386,560	57.04
35		Port of Rotterdam	Shipping	-	13,631	421,053	98,658	149,210	NL	66.12	98,658	ES	4.73	7,058	DE	4.55	6,789	24.6
36		9292	Multimodal	Passenger	88	9,521	7,057,212	7,320,000	NL	96.41	7,057,212	UK	0.45	32,940	DE	0.44	32,208	2.7
37		CEVA Logistics	Multimodal	Freight		97,870	30,337	583,410	US	48.42	282,487	CA	9.78	57,057	NL	5.2	30,337	36.6
38	France (FR)	Airfrance	Flight	Passenger	269	8,182	6,289,524	7,740,000	FR	81.26	6,289,524	US	1.66	128,484	NL	1.41	109,134	15.67
39		Air France-KLM	Flight	Passenger	55,935	638,151	22,918	74,240	FR	30.87	22,918	NL	14.05	10,431	BE	12.87	9,555	42.21
40		SNCF	Railway	Passenger	169	5,335	13,438,980	14,760,000	FR	91.05	13,438,980	CH	1.27	187,452	DE	1	147,600	6.68
41		iDTGV	Railway	Passenger	32,529	712,819	63,734	75,380	FR	84.55	63,734	DJ	6.03	4,545	CN	1.25	942	8.17
42		Mappy	Road	-	73	2,472	18,804,000	20,000,000	FR	94.02	18,804,000	BE	2.74	548,000	CH	0.33	66,000	2.91
43		ViaMichelin	Road	-	245	8,301	10,102,482	10,530,000	FR	95.94	10,102,482	BE	0.94	98,982	CH	0.33	34,749	2.79
44		CMA CGM	Shipping	Freight	-	60,018	-	1,060,000	US	10.56	111,936	IN	9.88	104,728	UAE	4.22	44,732	75.34
45		La Poste	Multimodal	Freight	50	1,751	30,155,664	31,570,000	FR	95.52	30,155,664	DE	0.39	123,123	BE	0.35	110,495	3.74
46		Voyages-sncf	Multimodal	Passenger	152	4,322	11,804,976	13,140,000	FR	89.84	11,804,976	CH	1.85	243,090	BE	1.44	189,216	6.87
47	Australia (AU)	Jetstar	Flight	Passenger	130	3,984	4,301,609	13,730,000	AU	31.33	4,301,609	VN	25.35	3,480,555	JP	15.28	2,097,944	28.04
48		Qantas	Flight	Passenger	169	11,344	5,102,896	6,880,000	AU	74.17	5,102,896	US	5.35	368,080	NZ	3.59	246,992	16.89
49		NSW TrainLink	Railway	Passenger	229	19,570	3,957,588	4,230,000	AU	93.56	3,957,588	US	0.94	39,762	NZ	0.79	33,417	4.71
50		Sydney Trains	Railway	Passenger	703	50,839	1,467,346	1,580,000	AU	92.87	1,467,346	US	1.27	20,066	NZ	0.91	14,378	4.95
51		WhereiS	Road		853	59,271	1,433,595	1,550,000	AU	92.49	1,433,595	PT	1.28	19,840	US	1.1	17,050	5.13
52		ANL	Shipping	Freight	52,011	621,829	14,452	66,600	AU	21.7	14,452	PH	14.62	9,737	MY	11.27	7,506	52.41
53		Australia Post	Multimodal	Freight	59	4,931	13,869,045	15,550,000	AU	89.19	13,869,045	US	2.34	363,870	NZ	1.44	223,920	7.03

4.4.2 Road mode

ITS is an integrated application of computer, electronics, communication technologies and management strategies. It aims to increase the safety and efficiency of the road transportation systems.[51] ITS involve infrastructures in the field, vehicles, travellers, road operators, and managers. They are all sources of data and communicating together either using wide-area wireless communication (microwave, satellite, 3G, 4G), in between fixed points, or vehicle-to-vehicle and vehicle-to-infrastructure communications (ETSI, ITS G5 WAVE). Figure 14 represents the architecture and high-level data flows between the communicating subsystems comprising field (orange), vehicle (blue), centres (green) and travellers (yellow). The wide area of applications is also depicted in the figure as white boxes located in the respective subsystem and their relevant data flows, including the following applications:

- traffic demand and capacity management
- information management (real-time)
- public transportation operations
- online trip planning, booking, and payment
- vehicle control assistance and safety systems
- emergency management
- maintenance and construction management
- commercial vehicle operations

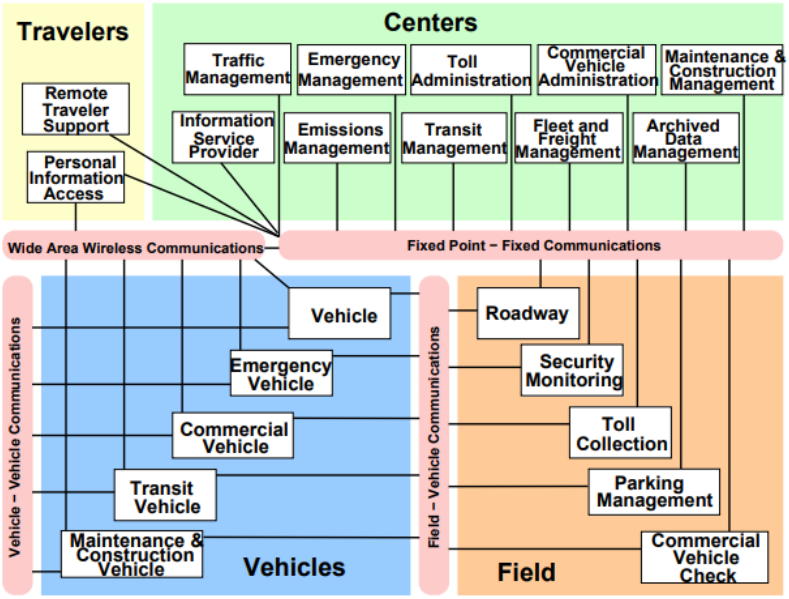

Figure 14: Architecture and subsystems involved in ITS

[51] Systems Transport Engineering Chapter 48, http://nptel.ac.in/courses/105101008/downloads/cete_48.pdf, Accessed April 15, 2018

4.4.3 Urban mode

Urban transport mode is represented by a combined view on public urban transport and rail transport. Figure 15 displays examples of data sources and data flow for urban bus operations and links to ITS. In addition, Data sources, flows and main functions in the public transport sector are depicted in Figure 16. For passenger rail transport, the utilized data sources and applications are visualized in Figure 17.

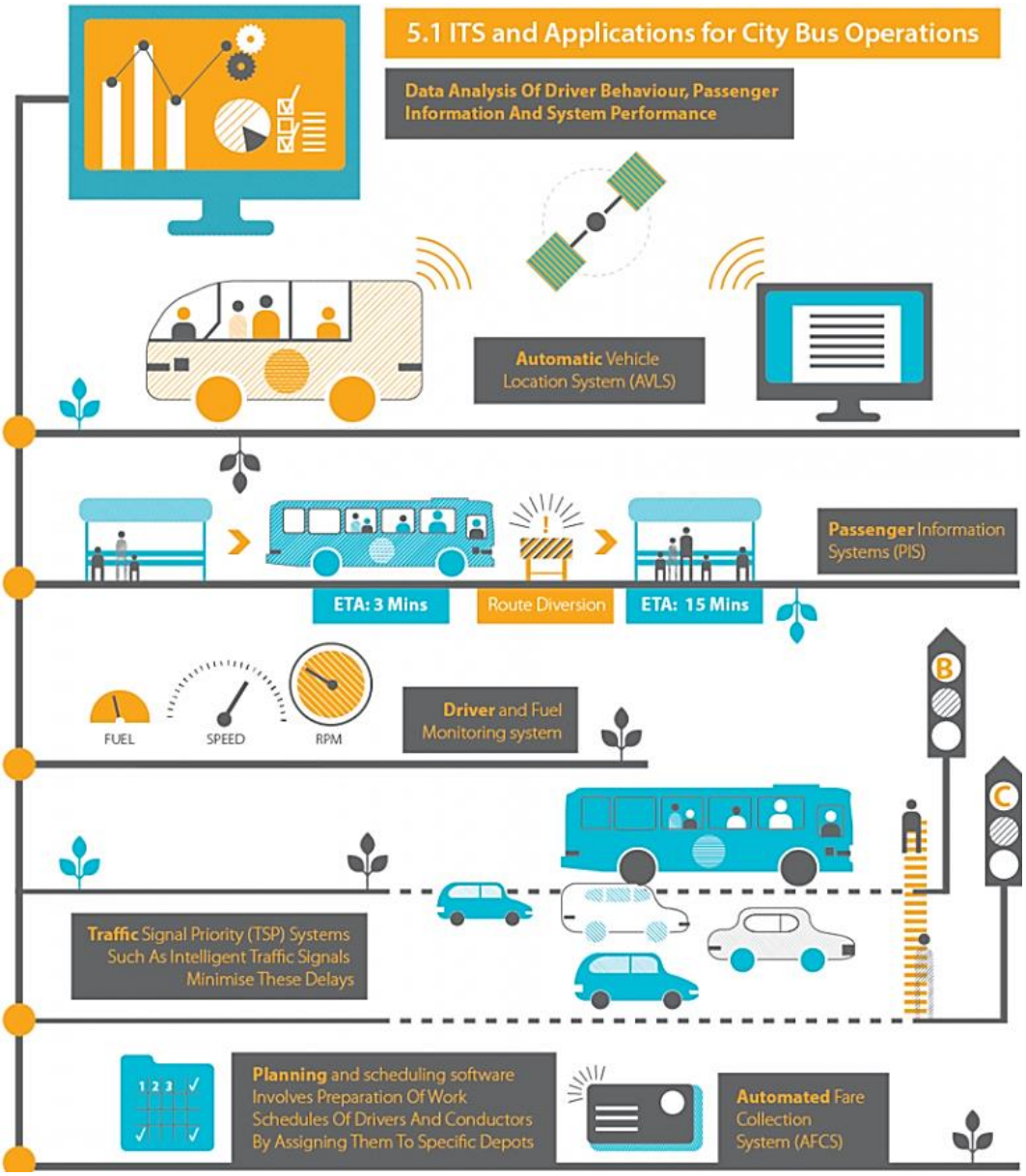

Figure 15: Big Data applications for city bus operations[52]

[52] http://wricitieshub.org/online-publications/chapter-5-intelligent-transportation-systems-city-bus-services

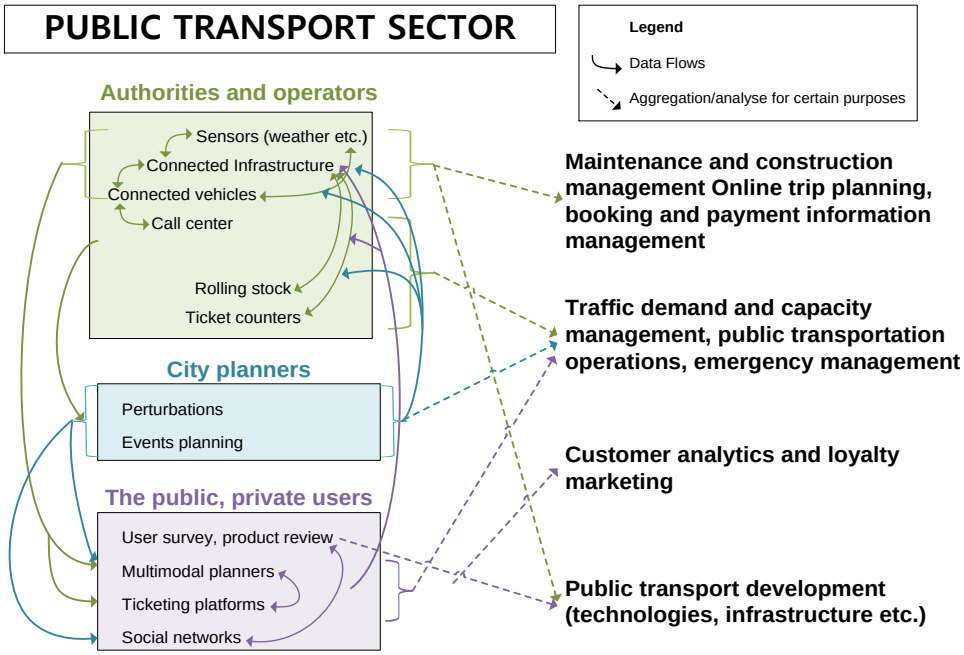

Figure 16: Data sources, flows and main functions in the public transport sector (Source: LeMO visualization)

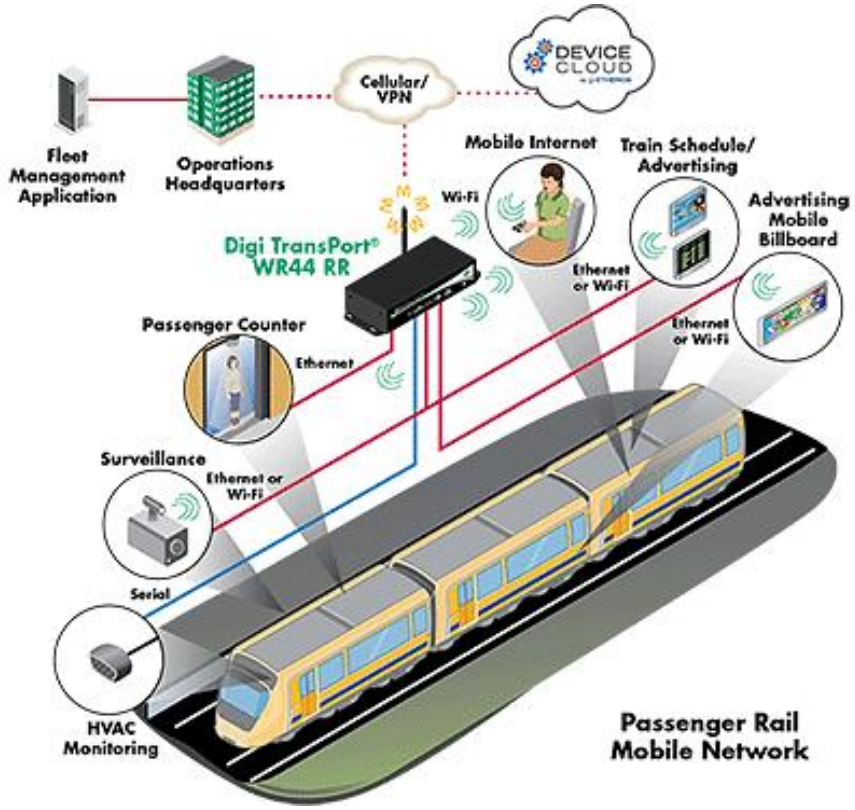

Figure 17: Big Data involved in railway transport operation[53]

[53] https://lightrailnow.wordpress.com/2013/09/03/how-rail-public-transportation-has-been-a-leader-in-the-analytics-and-big-data-revolution/, Accessed April 15, 2018

4.4.4 Rail mode

The sources of data used in the rail sector originate from vehicles, travellers, the infrastructure and fixed centres. In the case of public transport, the focus is set on maintenance and construction management. Considering the high cost of railway infrastructure, predictive modelling of the degradation of railway infrastructure and vehicles allow for optimization of maintenance operations. Modelling uses data inflows from infrastructure, vehicle condition sensors and historical data of the systems.

4.4.5 Air mode

Many airline companies have adopted big data analytics. The airline industry has been a customer experience expert (pre-flight and flight) with its successful loyalty programs for many years. As one of the biggest industries that have access to various kinds of data from multiple sources. The data flows between actors in the airline industry are aggregated in various forms. They are collecting data from everywhere (e.g. weblogs, social media, mobile apps, frequent flyer data, third party data etc.) and hence further requires integration of internal, external, real-time and batch-processed data of structured and unstructured data. Big data helps airlines have a better understanding of the individual passenger, identify patterns in his/her behavior, determine preferences and foresee future requests. By leveraging big data insights, airlines have the ability to make strategic decisions and differentiate themselves in the competitive market. Figure 18 shows data flow between third parties, social media, airline and a frequent flyer in air transport sector.

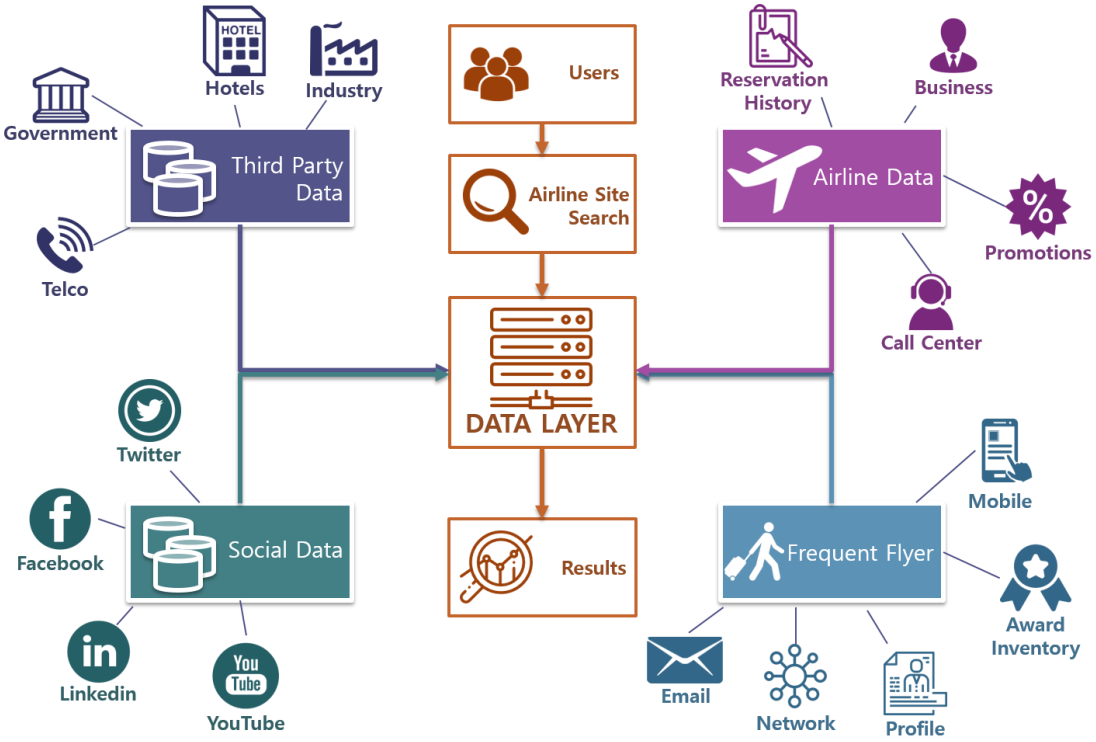

Figure 18: Data flows in the air transport sector (Source: LeMO visualization)

4.4.6 Water mode

In the shipping sector, sensors on the ships combined with sensors at the docks enable effective port resource management. At the dock, information such as the volume of goods to deliver per ship, unloading speed and Estimated Time Arrival of the ships are harnessed to design a dynamic schedule. Notifications are sent to the truck dispatcher for updating arrival time for time and resource management. Additionally, sensors on the ships assist onboard navigation providing data on sea and weather conditions as well as vessel performances. Also, they assist the driver with safety features such as collision avoidance, and collected data is used for predictive maintenance and fleet management.

The data flows between actors in the shipping industry are aggregated in various forms. For example, the National Coastal Administration displays the data flows based on granularity and its use (suppliers and consumers) as depicted in Figure 19. Figure 20 presents a maritime big data platform of Fujitsu for a ship data center and data flows in the shipping industry.

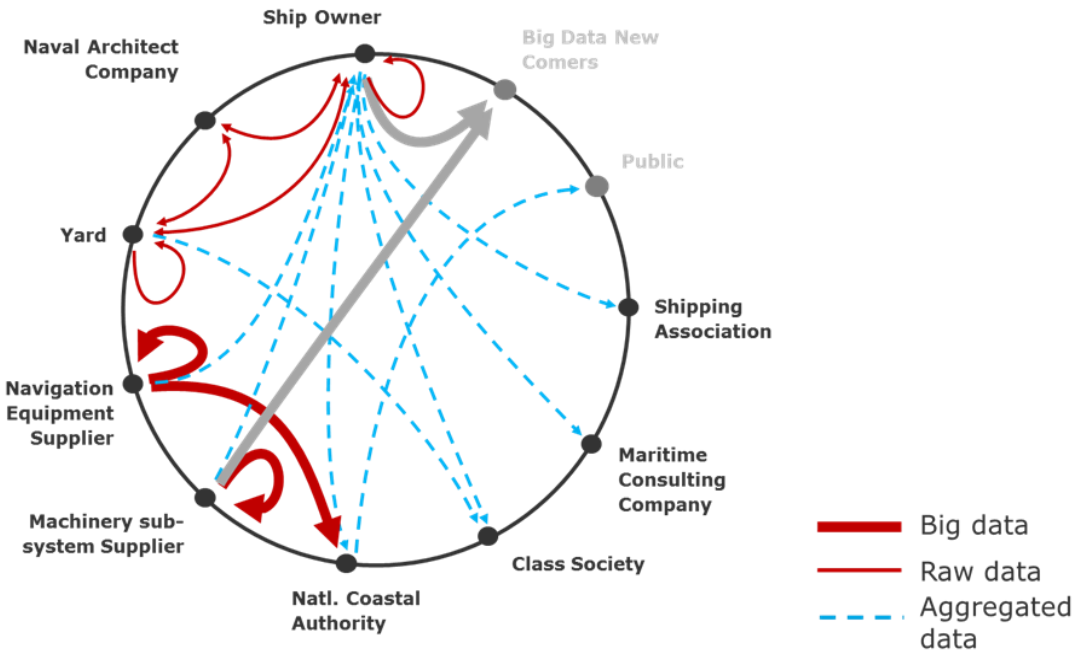

Figure 19: Data flows in the shipping industry (Source: BYTE project – Deliverable 1.1)

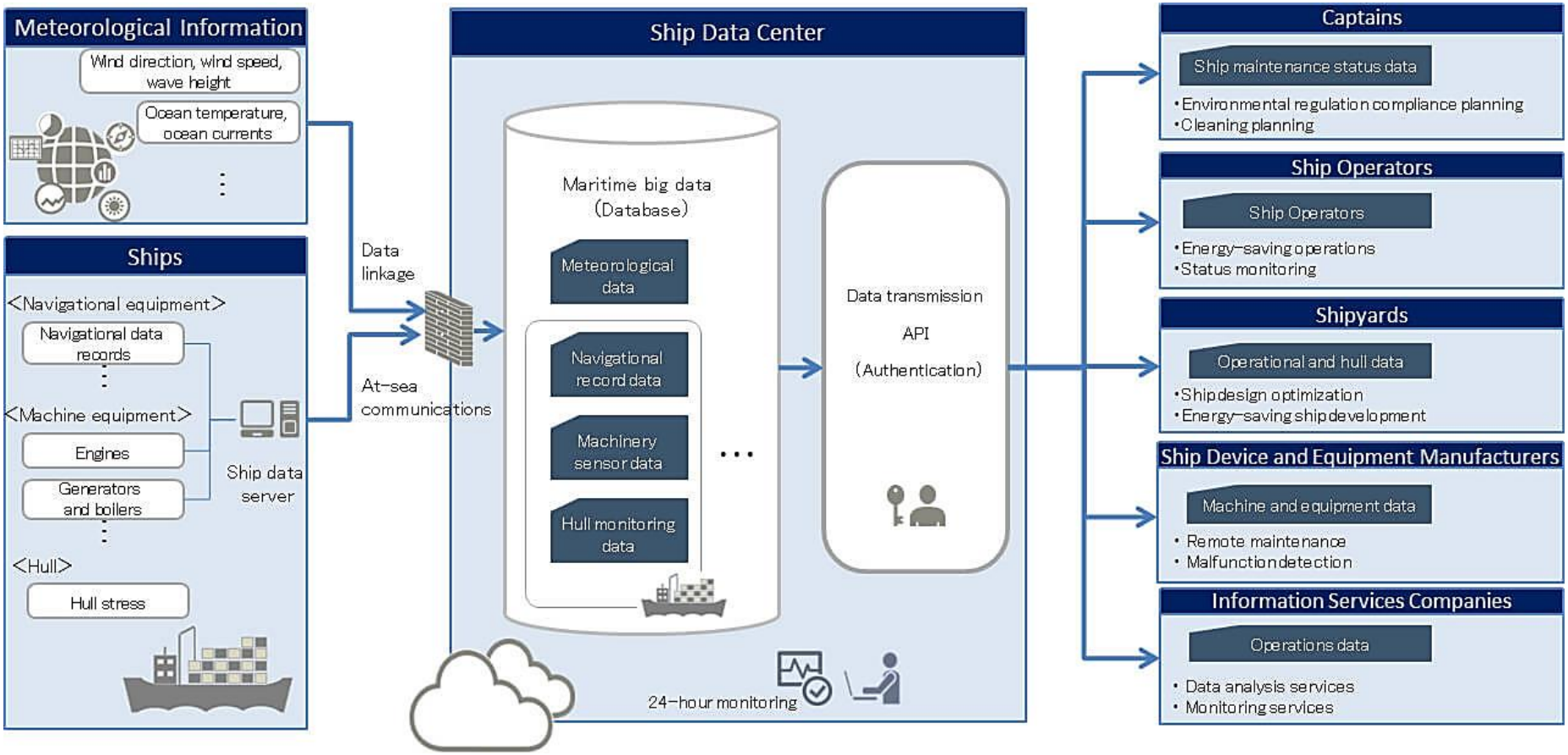

Figure 20: A maritime big data platform of Fujitsu for ship data center[54]

[54] https://www.marineinsight.com/shipping-news/fujitsu-build-industry-first-maritime-big-data-platform/

5 Conclusion

This report delivered two major tasks: (1) Understanding big data in Transport Sector and (2) Cartography of data flows.

The first task was completed in Chapter 2 and 3. Chapter 2 covers a background of the transport sector and big data in transportation. It clarified six modes (air, rail, road, urban, water and multi-modal) and the two sectors (passenger and freight) in the transport sector. Chapter 3 investigated multiple opportunities and challenges of big data in transportation in three steps: the subject matter expert interviews, nineteen applied cases, and then literature review. It indicates that the combination of different means and approaches will enhance the opportunities for successful big data services in the transport sector.

The second task is completed in Chapter 4. It carried an intensive survey of the various data sources, data producers, and service providers. In addition, cartography was modelled to visualize data flows intuitively. Cartography demonstrates where data originated from and where it is flowing to.

References

Allen, J., Ambrosini, C., Browne, M., Patier, D., Routhier, J-L., Woodburn, A (2014), Data Collection for Understanding Urban Goods Movement: Comparison of Collection Methods and Approaches in European Countries. In: Gonzalez-feliu, J., Semet, F., & Routhier, J., Sustainable urban logistics: concepts, methods and information systems. doi.org/10.1007/978-3-642-31788-0

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi.org/10.1023/A:1010933404324

Chopra, S., & Meindl, P. (2010). Supply chain management: Strategy, planning, and operation (4. ed., global ed.). Boston, Mass.: Pearson.

Chowdhury, M. & Apon, A. & Dey, K. & , Editors. (2017). Data Analytics for Intelligent Transportation Systems. Elsevier.

Council, N. H. M. R. (2009). National Health Medical Research Council. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. Canberra.

Cui, J., Dodson, J., & Hall, P. V. (2015). Planning for Urban Freight Transport: An Overview. Transport Reviews. doi.org/10.1080/01441647.2015.1038666

Cui, J., Liu, F., Hu, J., Janssens, D., Wets, G., & Cools, M. (2016). Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast growing Chinese city of Harbin. Neurocomputing, 181, 4–18. doi.org/10.1016/j.neucom.2015.08.100

European Environment Agency. (2017). Greenhouse gas emissions from transport., www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-10. Published 11-07-2017.Accessed 20-07-2017

Gan, S., Liang, S., Li, K., Deng, J., & Cheng, T. (2018). Trajectory Length Prediction for Intelligent Traffic Signaling: A Data-Driven Approach. IEEE Transactions on Intelligent Transportation Systems, 19(2), 426–435. doi.org/10.1109/TITS.2017.2700209

Gennaro, M. de, Paffumi, E., & Martini, G. (2016). Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities. Big Data Research, 6, 11–25. doi.org/10.1016/j.bdr.2016.04.003

Giest, S. (2017). Big data analytics for mitigating carbon emissions in smart cities: Opportunities and challenges. European Planning Studies, 25(6), 941–957. doi.org/10.1080/09654313.2017.1294149

Guerreiro, G., Figueiras, P., Silva, R., Costa, R., & Jardim-Goncalves, R. (2016). An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows. In R. R. Yager (Ed.), 2016 IEEE 8th International Conference on Intelligent Systems: Proceedings (pp. 65–72). Piscataway, NJ: IEEE. doi.org/10.1109/IS.2016.7737393

Jiang, X., Abdel-Aty, M., Hu, J., & Lee, J. (2016). Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. Neurocomputing, 181, 53–63. doi.org/10.1016/j.neucom.2015.08.097

Julio, N., Giesen, R., & Lizana, P. (2016). Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. Research in Transportation Economics, 59, 250–257. doi.org/10.1016/j.retrec.2016.07.019

Laney, D. (2001), 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical report, META Group.

Lee, M.-K., & Yoo, S.-H. (2016). The role of transportation sectors in the Korean national economy: An input-output analysis. Transportation Research Part A: Policy and Practice, 93, 13–22. doi.org/10.1016/j.tra.2016.08.016

Li, P., Zhao, P., & Brand, C. (2018). Future energy use and CO 2 emissions of urban passenger transport in China: A travel behavior and urban form based approach. Applied Energy, 211, 820–842. doi.org/10.1016/j.apenergy.2017.11.022

Linares, M. P., Barceló, J., Carmona, C., & Montero, L. (2017). Analysis and Operational Challenges of Dynamic Ride Sharing Demand Responsive Transportation Models. Transportation Research Procedia, 21, 110–129. doi.org/10.1016/j.trpro.2017.03.082

Lindholm, M., & Ballantyne, E. E. F. (2016). Introducing Elements of Due Diligence in Sustainable Urban Freight Transport Planning. In Transportation Research Procedia. doi.org/10.1016/j.trpro.2016.02.048

Liu, J., & Zio, E. (2018). A scalable fuzzy support vector machine for fault detection in transportation systems. Expert Systems with Applications. Advance online publication. doi.org/10.1016/j.eswa.2018.02.017

Liu, J., Li, Y.-F., & Zio, E. (2017). A SVM framework for fault detection of the braking system in a high speed train. Mechanical Systems and Signal Processing, 87, 401–409. doi.org/10.1016/j.ymssp.2016.10.034

Liu, J., Wang, X., Khattak, A. J., Hu, J., Cui, J. X., & Ma, J. (2016). How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis. Neurocomputing, 181, 38–52. doi.org/10.1016/j.neucom.2015.08.098

Lukoianova, T. & Rubin, V. L. (2014), Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Advances in Classification Research Online.

Manheim, M. L. (1980). Fundamentals of transportation systems analysis (3. pr). MIT Press series in transportations studies: Vol. 4. Cambridge, Mass.: MIT Pr.

Marsden, G., & Reardon, L. (2017). Questions of governance: Rethinking the study of transportation policy. Transportation Research Part A: Policy and Practice. doi.org/10.1016/j.tra.2017.05.008

Melo, S., Macedo, J., & Baptista, P. (2017). Guiding cities to pursue a smart mobility paradigm: An example from vehicle routing guidance and its traffic and operational effects. Research in Transportation Economics, 65, 24–33. doi.org/10.1016/j.retrec.2017.09.007

Merriam-Webster Dictionary. (2018). Definition of "Transportation". Retrieved from www.merriam-webster.com/dictionary/transportation

Möller, D.P.F., & Schroer, B. (2014). Introduction to Transportation Analysis, Modeling and Simulation: Computational Foundations and Multimodal Applications: Springer London. Retrieved from books.google.de/books?id=XUnPBAAAQBAJ

Moraglio, M. (2017). Seeking a (new) ontology for transport history. The Journal of Transport History, 38(1), 3–10. doi.org/10.1177/0022526617709168

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, 24, 9–18. doi.org/10.1016/j.trc.2012.01.007

Najada, H. A., & Mahgoub, I. (2016). Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI): Proceedings : 6-9 December 2016, Athens, Greece (pp. 1–8). Piscataway, NJ: IEEE. doi.org/10.1109/SSCI.2016.7850097

Nakamura, T., Nakamura, N., Schmöcker, J.-D., Uno, N., & Iwamoto, T. (2016). Urban Public Transport Mileage Cards: Analysis of Their Potential with Smart Card Data and an SP Survey. In

Ortuzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport. Modelling Transport. doi.org/10.1002/9781119993308

P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 1st ed. McGraw-Hill Osborne Media (IBM), 2011.

Pineda, C., Schwarz, D., & Godoy, E. (2016). Comparison of passengers' behavior and aggregate demand levels on a subway system using origin-destination surveys and smartcard data. Research in Transportation Economics, 59, 258–267. doi.org/10.1016/j.retrec.2016.07.026

Plößl, K., & Federrath, H. (2008). A Privacy Aware and Efficient Security Infrastructure for Vehicular Ad Hoc Networks.

Sánchez-Martínez, G. E., & Munizaga, M. (2016a). Workshop 5 report: Harnessing big data. Research in Transportation Economics, 59, 236–241. doi.org/10.1016/j.retrec.2016.10.008

Sánchez-Martínez, G. E., Koutsopoulos, H. N., & Wilson, N. H.M. (2016b). Optimal allocation of vehicles to bus routes using automatically collected data and simulation modelling. Research in Transportation Economics, 59, 268–276. doi.org/10.1016/j.retrec.2016.06.003

Santos, G., Behrendt, H., Maconi, L., Shirvani, T., & Teytelboym, A. (2010). Part I: Externalities and economic policies in road transport. Research in Transportation Economics, 28(1), 2–45. doi.org/10.1016/j.retrec.2009.11.002

Tang, J., & Heinimann, H. R. (2018). A resilience-oriented approach for quantitatively assessing recurrent spatial-temporal congestion on urban roads. PloS one, 13(1), e0190616. doi.org/10.1371/journal.pone.0190616

Tavasszy, L. & de Jong, G. (2014), Data availability and model form. In: edited by Tavasszy, L. & de Jong, G., Modelling Freight Transport. Elsevier, Oxford, doi.org/10.1016/B978-0-12-410400-6.00012-4

Teodorovic, D., & Janic, M. (2017). Transportation engineering: Theory, practice, and modeling.

Vlahogianni, E. I., Park, B. B., & van Lint, J.W.C. (2015). Big data in transportation and traffic engineering. Transportation Research Part C: Emerging Technologies, 58, 161. doi.org/10.1016/j.trc.2015.08.006

Wang, Y., Ram, S., Currim, F., Dantas, E., & Saboia, L. A. (2016). A big data approach for smart transportation management on bus network. In Improving the citizens quality of life: IEEE Second International Smart Cities Conference (ISC2 2016) : 12-15 September 2016, Trento - Italy : proceedings (pp. 1–6). Piscataway, NJ: IEEE. doi.org/10.1109/ISC2.2016.7580839

WCED, 1987. Report of the World Commission on Environment and Development. Our Common Future.

Wemegah, T. D., & Zhu, S. (2017). Big data challenges in transportation: A case study of traffic volume count from massive Radio Frequency Identification(RFID) data. In 2017 International Conference on the Frontiers and Advances in Data Science (FADS): 23-25 October 2017, Xian, China : conference proceedings (pp. 58–63). Piscataway, NJ: IEEE. doi.org/10.1109/FADS.2017.8253194

Wolfram, M. (2004) Expert Working Group on Sustainable Urban Transport Plans, Deliverable D4, Cologne: Rupprecht Consult

Xia, Y., Chen, J., Lu, X., Wang, C., & Xu, C. (2016). Big traffic data processing framework for intelligent monitoring and recording systems. Neurocomputing, 181, 139–146. doi.org/10.1016/j.neucom.2015.07.140

Zhiyuan, H., Liang, Z., Ruihua, X., & Feng, Z. (2017). Application of big data visualization in passenger flow analysis of Shanghai Metro network. In 2017 2nd IEEE International Conference on Intelligent Transportation Engineering: ICITE 2017 : September 1-3, 2017, Singapore (pp. 184–188). Piscataway, NJ: IEEE. doi.org/10.1109/ICITE.2017.8056905

Zicari, R. V., Rosselli, M., Ivanov, T., Korfiatis, N., Tolle, K., Niemann, R., & Reichenbach, C. (2016), Big Data Optimization: Recent Developments and Challenges, Springer International Publishing:

Appendix A Interviews on Data Quality

Question 1: How do you ensure data quality?

Answers:

Jeff Saltz: Data quality is a subset of the larger challenge of ensuring that the results of the analysis are accurate or described in an accurate way. This covers the quality of the data, what one did to improve the data quality (ex. remove records with missing data) and the algorithms used (ex. were the analytics appropriate). In addition, it includes ensuring an accurate explanation of the analytics to the client of the analytics. As you can see, I think of data quality is being an integrated aspect of an end-to-end process (i.e., not a "check" done before one releases the results)

Manohar Swamynathan: Looking at basic statistics (central tendency and dispersion) about the data can give good insight into the data quality. You can perform univariate and multivariate analysis to understand the trends and relationship within, between variables. Summarizing the data is a fundamental technique to help you understand the data quality and issues/gaps.

Jonathan Ortiz: The world is a messy place, and, therefore, so is the web and so is data. No matter what you do, there's always going to be dirty data lacking attributes entirely, missing values within attributes, and riddled with inaccuracies. The best way to alleviate this is for all data users to track the provenance of their data and allow for reproducibility of their analysis and models. The open-source software development philosophy will be co-opted by data scientists as more and more of them collaborate on data projects. By storing source data files, scripts, and models on open platforms, data scientists enable reproducibility of their research and allow others to find issues and offer improvements.

Anya Rumyantseva: Quality of data has a significant effect on results and efficiency of machine learning algorithms. Data quality management can involve checking for outliers/inconsistency, fixing missing values, making sure data in columns are within a reasonable range, data is accurate etc. All can be done during the data pre-processing and exploratory analysis stages.

Dirk Tassilo Hettich: Understanding the data at hand by visual inspection. Ideally, browse through the raw data manually since our brain is a super powerful outlier detection apparatus. Do not try to check every value, just get an idea of how the raw data actually looks! Then, looking at the basic statistical moments (e.g. numbers and boxplots) to get a feeling how the data looks like. Once patterns are identified, parsers can be derived that apply certain rules to incoming data in a productive system.

Wolfgang Steitz: It's good practice to start with some exploratory data analysis before jumping to the modeling part. Doing some histograms and some time series is often enough to get a feeling for the data and know about potential gaps in the data, missing values, data ranges, etc. In addition, you should know where the data is coming from and what transformations it went through. Ones you know all this, you can start filling the gaps and cleaning your data. Eventually, there is even another data set you want to take into account. For some model running in production, it's a good idea to automate some data quality checks. These tests could be as

simple as checking if the values are in the correct range or if there are any unexpected missing values. And of course, someone should be automatically notified if things go bad.

Paolo Giudici: For unsupervised problems: checking the contribution of the selected data to between the group's heterogeneity and within the group's homogeneity; For supervised problems: checking the predictive performance of the selected data.

Andrei Lopatenko: Data quality is not enough, it must be automatically checked. In real-world applications, it rarely happens that you get data once. Frequently you get a stream of data. If you build an application for local business, you get a stream of data from the provider. If you build an e-commerce site, then you get regular data updates from merchants, and other data providers. The problem is that you can almost never be sure of data quality. In most cases data are dirty. You have to protect your customers from dirty data. You have to work to discover what problems with data you might have. Frequently problems are not trivial. Sometimes you can see them browsing data directly, frequently toy cannot. For example, in case of local business latitude-longitude coordinates might be wrong because provided has a bad data geocoding system. Sometimes you do not see problems with data immediately, but only after using them for training some models, where errors are accumulated and lead to wrong results and you have traced back what was wrong.

To ensure data quality once I understand what problems may happen, I build data quality monitoring software. At every step of data processing pipelines I embed tests, you may compare them with unit tests for traditional software development which checks the quality of data. They may check the total amount of data, existence or non-existence of certain values, anomalies in data, compare data to data from the previous batch and so on. It required significant error to build data quality tests, but it pays back, they protect from errors in data engineering, data science, incoming data, some system failures, it always pays back. From my experience, almost every company build a set of libraries and code alike to ensure data quality control. We did it in Google, we did it in Apple, and we did it Walmart. In the Recruit Institute of Technology, we work on Big Gorilla tools set, which will include our open source software and references to other open source software which may help companies build data quality pipelines.

Mike Shumpert: On the one hand, one of the basic tenets of “big data” is that you can’t ensure data quality – today’s data is voluminous and messy, and you’d better be prepared to deal with it. As mentioned before, “dealing with it” can simply mean throwing some instances out, but sometimes what you think is an outlier could be the most important information you have.

So if you want to enforce at least some data quality, what can you do? It’s useful to think of data as comprising two main types: transactional or reference. Transactional data is time-based and constantly changing – it typically conveys that something just happened (e.g., customer checkouts), although it can also be continuous data sampled at regular intervals (e.g., sensor data). Reference data changes very slowly and can be thought of as the properties of the object (customer, machine, etc.) at the center of the prediction. Both types of data typically have predictive value: this amount at this location was just spent (transactional) by a platinum-level female customer (reference) – is it fraud? But the two types often come from different sources and can be treated differently in terms of data quality. Transactional data can be filtered or

smoothed to remove transitory outliers, but the problem domain will determine whether or not any such anomalies are noise or real (and thus very important). For example, the $10,000 purchase on a credit card with a typical maximum of $500 is one that deserves further scrutiny, not dismissal. But reference data can be separately cleansed and maintained via Master Data Management (MDM) technology. This ensures there is only one version of the truth with respect to the object at the core of the prediction and prevents nonsensical changes such as a customer moving from gold status to platinum and back again within 30 seconds. Clean reference data can then be merged with transactional data on the fly to ensure accurate predictions.

Using an Internet of Things (IoT) example, consider a predictive model for determining when a machine needs to be serviced. The model will want to leverage all the sensor data available, but it will also likely find useful factors such as the machine type, date of the last service, country of origin, etc. The data stream coming from the sensors usually will not carry with it the reference data and will probably only provide a sensor id. That id can be used to look up relevant machine data and enrich the data stream on the fly with all the features needed for the prediction. One final point on this setup is that you do not want to go back to the original data sources of record for this on-the-fly enrichment of transactional data with reference data. You want the cleansed data from the MDM system, and you want that stored in memory for high-performance retrieval.

Romeo Kienzler: This is again a vote for domain knowledge. I have someone with domain skills assess each data source manually. In addition, I gather statistics on the accepted data sets so some significant changes will raise an alert which – again – has to be validated by a domain expert.

Elena Simperl: It is not possible to “ensure” data quality because you cannot say for sure that there isn’t something wrong with it somewhere. In addition, there is also some research which suggests that compiled data are inherently filled with the (unintentional) bias of the people compiling it. You can attempt to minimise the problems with quality by ensuring that there is full provenance as to the source of the data and err on the side of caution where some part of it is unclassified or possibly erroneous. One of the things we are researching at the moment is how best to leverage the wisdom of the crowd for ensuring the quality of data, known as crowdsourcing. The existence of tools such as Crowdflower makes it easy to organise a crowdsourcing project, and we have had some level of success in image understanding, social media analysis, and data integration Web. However, the best ways of optimising cost, accuracy or time remain to be determined and are different relative to the particular problem or motivation of the crowd one works with.

Mohammed Guller: It is a tough problem. Data quality issues generally occur upstream in the data pipeline. Sometimes the data sources are within the same organization and sometimes data comes from a third-party application. It is relatively easier to fix data quality issues if the source system is within the same organization. Even then, the source may be a legacy application that nobody wants to touch. So you have to assume that data will not be clean and address the data quality issues in your application that processes data. Data scientists use various techniques to address these issues. Again, domain knowledge helps.

Natalino Busa: I tend to rely on the "wisdom of the crowd" by implementing similar analysis using multiple techniques and machine learning algorithms. When the results diverge, I compare the methods to gain an insight into the quality of both data as well as models. This technique works also well to validate the quality of streaming analytics: in this case, the batch historical data can be used to double check the result in streaming mode, providing, for instance, end-of-day or end-of-month reporting for data correction and reconciliation.

Vikas Rathee: Data quality is very important to make sure the analysis is correct and any predictive model we develop using that data is good. Very simply I would do some statistical analysis on the data, create some charts and visualize information. I also will clean data by making some choice at the time of data preparation. This would be part of the feature engineering stage that needs to be done before any modeling can be done.

Christopher Schommer: To keep a data quality is mostly an adaptive process, for example, because provisions of national law may change or because the analytical aims and purposes of the data owner may vary. Therefore, the ensuring of a data quality should be performed regularly, it should be consistent with the law (data privacy aspects and others), and should be commonly performed by a team of experts of different education levels (e.g., data engineers, lawyers, computer scientists, mathematicians).

Jochen Leidner: There are a couple of things: first, make sure you know where the data comes from and what the records actually mean. Is it a static snapshot that was already processed in some way, or does it come from the primary source? Plotting histograms and profiling data in other ways is a good start to find outliners and data gaps that should undergo imputation (filling of data gaps with reasonable fillers). Measuring is key, so doing everything from the inter-annotator agreement of the gold data overtraining, dev-test and test evaluations to human SME output grading consistently pays back the effort.

Claudia Perlich: The sad truth is – you cannot. Much is written about data quality and it is certainly a useful relative concept, but as an absolute goal, it will remain an unachievable ideal (with the irrelevant exception of simulated data). First off, data quality has many dimensions. Secondly – it is inherently relative: the exact data can be quite good for one purpose and terrible for another. Third, data quality is a very different concept for 'raw' event log data vs. aggregated and processed data. Finally, and this is by far the hardest part: you almost never know what you don't know about your data. In the end, all you can do is your best! Skepticism, experience, and some sense of data intuition are the best sources of guidance you will have.

Richard J Self: Data Quality is a fascinating question. It is possible to invest enormous levels of resource into attempting to ensure near perfect data quality and still fail. The critical question should, however, start from the Governance perspective of questions such as:

- What is the overall business Value of the intended analysis?
- How is the Value of the intended insight affected by different levels of data quality (or Veracity)?
- What is the level of Vulnerability to our organisation (or other stakeholders) if the data is not perfectly correct in terms of reputation, or financial consequences?

Once you have answers to those questions and the sensitivities of your project to various levels of data quality, you will then begin to have an idea of just what level of data quality you need to achieve. You will also then have some ideas about what metrics you need to develop and collect, in order to guide your data ingestion and data cleansing and filtering activities.

Ritesh Ramesh: Data Quality is critical. We hear often from many of our clients that ensuring trust in the quality of information used for analysis is a priority. The thresholds and tolerance of data quality can vary across problem domains and industries but nevertheless, data quality and validation processes should be tightly integrated into the data preparation steps.

Data scientists should have full transparency on the profile and quality of datasets that they are working with and have tools at their disposal to remediate fixes with proper governance and procedures as necessary. Emerging data quality technologies are embedding leverages machine learning features to detect proactive data errors and make data quality a business-user-friendly and an intelligent function more than ever it has been for years.

Question 2: How do you evaluate if the insight you obtain from data analytics is "correct" or "good" or "relevant" to the problem domain?

Answers:

Jeff Saltz: With respect to being relevant, this should be addressed by our first topic of discussion – needing domain knowledge. It is the domain expert (either the data scientist or a different person) that is best positioned to determine the relevance of the results. However, evaluating if the analysis is "good" or "correct" is much more difficult, and relates to our previous data quality discussion. It is one thing to try and do "good" analytics, but how does one evaluate if the analytics are "good" or "relevant"? I think this is an area ripe for future research. Today, there are various methods that I (and most others) use. While the actual techniques we use vary based on the data and analytics used, ensuring accurate results ranges from testing new algorithms with known data sets to point sampling results to ensure reasonable outcomes.

Yanpei Chen: Here's a list of things I watch for: Proxy measurement bias. If the data is an accidental or indirect measurement, it may differ from the "real" behavior in some material way. Instrumentation coverage bias. The "visible universe" may differ from the "whole universe" in some systematic way. Analysis confirmation bias. Often the data will generate a signal for "the outcome that you look for". It is important to check whether the signals for other outcomes are stronger. Data quality. If the data contains many NULL values, invalid values, duplicated data, missing data, or if different aspects of the data are not self-consistent, then the weight placed in the analysis should be appropriately moderated and communicated. Confirmation of well-known behavior. The data should reflect behavior that is common and well-known. For example, credit card transaction volumes should peak around well-known times of the year. If not, conclusions drawn from the data should be questioned. My view is that we should always view data and analysis with a healthy amount of skepticism while acknowledging that many real-life decisions need only directional guidance from the data.

Jonathan Ortiz: I think "good" insights are those that are both "relevant" and "correct," and those are the ones you want to shoot for. I always have a baseline for comparison. You can do

this either by experimenting, where you actually run a controlled test between different options and determine empirically which is the preferred outcome (like when A/B testing or using a Multi-armed Bandit algorithm to determine optimal features on a website), or by comparing predictive models to the current ground truth or projected outcomes from current data. Also, solicit feedback about your results early and often by showing your customers, clients, and domain experts. Gather as much feedback as you can throughout the process in order to iterate on the models.

Anya Rumyantseva: I would suggest constantly communicating with other people involved in a project. They can relate insights from data analytics to defined business metrics. For instance, if a developed data science solution decreases shutdown time of a factory from 5% to 4.5%, this is not that exciting for a mathematician. But for the factory owner, it means going bankrupt or not!

Wolfgang Steitz: Presenting results to some domain experts and your customers usually helps. Try to get feedback early in the process to make sure you are working in the right direction and the results are relevant and actionable. Even better, collect expectations first to know how your work will be evaluated later on.

Paolo Giudici: By testing its out-of-sample predictive performance we can check if it is correct. To check its relevance, the insights must be matched with domain knowledge models or consolidated results. What are the typical mistakes done when analyzing data for a large scale data project? Can they be avoided in practice? Forget data quality and exploratory data analysis, rushing to the application of complex models. Forgetting that pre-processing is a key step, and that benchmarking the model versus simpler ones is always a necessary prerequisite.

Andrei Lopatenko: Most frequently companies have some important metrics which describe company business. It might be the average revenue per session, the conversion rate, precision of the search engine etc. And your data insights are as good as they improve these metrics. Assume in an e-commerce company, the main metrics is Average Revenue Per Session (ARPS). And you work on a project of improving extraction of a certain item attribute, for example, from the non-structured text. Questions to ask yourself, will it help to improve ARPS by improving search because it will increase relevance for queries with color intents or faceted queries by color, or by providing better snippets, or by still other means. When one metric does not describe company business and many numbers are needed to understand it. Your data projects might be connected to other metrics. But what's important is to connect your data insight project to metrics which are representative of company business and improvement of these metrics will be as a significant impact to the company business. Such connection makes a good project.

Romeo Kienzler: I'm using the classical statistical performance measures to assess the performance of a model. This is only about the mathematical properties of a model. Then I check with the domain experts on the significance of their problems. Often a statistically significant result is not relevant to the business. E.g. if I tell you that a bearing will break with 95% probability within the next 6 months might not really help the PMQ (Predictive Maintenance and Quality) guys. So the former can be described as "correct" or "good" whereas the latter as "relevant" maybe.

Elena Simperl: The importance of having good enough domain knowledge comes into play in terms of answering the relevance question. Hopefully, a data scientist will have a good knowledge of the domain, but if not then they need to be able to understand what the domain expert believes in terms of relevance to the domain. The correctness or value of the data then comes down to understanding how to evaluate machine learning algorithms in general and using domain knowledge to apply to decide whether the trade-offs are appropriate given the domain.

Mohammed Guller: This is where domain knowledge helps. In the absence of domain knowledge, it is difficult to verify whether the insight obtained from data analytics is correct. A data scientist should be able to explain the insights obtained from data analytics. If you cannot explain it, chances are that it may be just a coincidence. There is an old saying in machine learning, "if you torture data sufficiently, it will confess to almost anything." Another way to evaluate your results is to compare it with the results obtained using a different technique. For example, you can do backtesting on historical data. Alternatively, compare your results with the results obtained using the incumbent technique. It is good to have a baseline against which you can benchmark results obtained using a new technique.

Natalino Busa: Most of the time I interact with domain experts for a first review of the results. Subsequently, I make sure that the model is brought into "action". Relevant insight, in my opinion, can always be assessed by measuring their positive impact on the overall application. Most of the time, as human interaction is part of the loop, the easiest method is to measure the impact of the relevant insight in their digital journey.

Vikas Rathee: Getting Insights is what makes the job of a Data Scientist interesting. In order to make sure the insights are good and relevant we need to continuously ask ourselves what is the problem we are trying to solve and how it will be used. In simpler words, to make improvements in an existing process we will need to understand the process and where the improvement is a requirement or of most value. For predictive modeling cases, we need to ask how the output of the predictive model will be applied and what additional business value can be derived from the output. We also need to convey what does the predictive model output means to avoid incorrect interpretation by non-experts. Once the context around a problem has been defined and we proceed to implement the machine learning solution. The immediate next stage is to verify if the solution will actually work. There are many techniques to measure the accuracy of predictions i.e. testing with historical data samples using techniques like k-fold cross-validation, confusion matrix, r-square, absolute error, MAPE (Mean absolution percentage error), p-value etc. We can choose from among many models which show most promising results. There are also ensemble algorithms which generalize the learning and avoid being overfitted models.

Christopher Schommer: In my understanding, an insight is already a valuable/evaluated information, which has been received after a detailed interpretation and which can be used for any kind of follow-up activities, for example, to relocate the merchandise or to deeper dig in clusters showing a fraudulent behavior. However, it is less opportune to rely only on statistical values: an association rule, which shows a conditional probability of, e.g., 90% or more, maybe an "insight", but if the right-hand side of the rule refers to a plastic bag only (which is to be paid (3 cents), at least in Luxembourg), the discovered pattern might be uninteresting.

Slava Akmaev: In a data-rich domain, evaluation of the insight correctness is done either by applying the mathematical model to new "unseen" data or using cross-validation. This process is more complicated in human biology. As we have learned over the years, a promising cross-validation performance may not be reproducible in subsequent experimental data. The fact of the matter is, in life sciences, laboratory validation of computational insight is mandatory. The community perspective on computational or statistical discovery is generally skeptical until the novel analyte, therapeutic target, or biomarker is validated in additional confirmatory laboratory experiments, pre-clinical trials or human fluid samples.

Jochen Leidner: There is nothing quite as good as asking domain experts to vet samples of the output of a system. While this is time-consuming and needs preparation (to make their input actionable), the closer the expert is to the real end user of the system (e.g. the customer's employees using it day to day), the better.

Claudia Perlich: First of, one should not even have to ask whether the insight is relevant – one should have designed the analysis that led to the insight based on the relevant practical problem one is trying to solve! The answer might be that there is nothing better you can do than status quo. That is still a highly relevant insight! It means that you will NOT have to waste a lot of resources. Taking negative answer into account as 'relevant' – if you are running into this issue of the results of data science not being relevant you are clearly not managing data science correctly. I have commented on this here: What are the greatest inefficiencies data scientists face today?

Let's look at 'correct' next. What exactly does it mean? To me it somewhat narrowly means that it is 'true' given the data: did you do all the due diligence and right methodology to derive something from the data you had? Would somebody answering the same question on the same data come to the same conclusion (replicability)? You did not overfit, you did not pick up a spurious result that is statistically not valid, etc. Of course, you cannot tell this from looking at the insight itself. You need to evaluate the entire process (or trust the person who did the analysis) to make a judgement on the reliability of the insight.

Now to the 'good'. To me good captures the leap from a 'correct' insight on the analyzed dataset to supporting the action ultimately desired. We do not just find insights into data for the sake of it! (well – many data scientists do, but that is a different conversation). Insights more often than not drive decisions. A good insight indeed generalizes beyond the (historical) data into the future. Lack of generalization is not just a matter of overfitting, it is also a matter of good judgement whether there is enough temporal stability in the process to hope that what I found yesterday is still correct tomorrow and maybe next week. Likewise, we often have to make judgement calls when the data we really needed for the insight is simply not available. So we look at a related dataset (this is called transfer learning) and hope that it is similar enough for the generalization to carry over. There is no test for it! Just your gut and experience.

Finally, good also incorporates the notion of correlation vs. causation. Many correlations are 'correct' but few of them are good for the action one is able to make. The (correct) fact that a person who is sick has temperature is 'good' for diagnosis, but NOT good for prevention of infection. At which point we are pretty much back to relevant! So think first about the problem and do good work next!

Richard J Self: The answer to this returns to the Domain Expert question. If you do not have adequate domain expertise in your team, this will be very difficult. Referring back to the USA election, one of the more unofficial pollsters, who got it pretty well right observed that he did it because he actually talked to real people. This is domain expertise and Small Data. All the official polling organisations have developed a total trust in Big Data and Analytics because it can massively reduce the costs of the exercise. But they forget that we all lie unremittingly online.

Ritesh Ramesh: Many people view Analytics and Data science as some magic crystal ball into the future events and don't realize that it is just one of many probable indicators for successful outcomes – If the model predicts that a there's an 80% chance of success, you also need to read it as there's still a 20% chance of failure. To really assess the 'quality' of insights from the model you may start with the below areas.

1) Assess whether the model makes reasonable assumptions on the problem domain and takes into account all the relevant input variables and business context – I was recently reading an article on a U.S. based insurer who implemented an analytics model that looked for number of unfavorable traffic incidents to assess risk on the vehicle driver but they missed out on assigning weights to the severity of the traffic incident. If your model makes wrong contextual assumptions – the outcomes can backfire.

2) Assess whether the model is run on a sufficient sample of datasets. Modern scalable technologies have made executing analytical models on massive amounts of data possible. More data the better although every problem does not need large datasets of the same kind.

3) Assess where extraneous events like macroeconomic events, weather, consumer trends etc. are considered in the model constraints. Use of external data sets with real-time API (Application Programming Interface) based integrations is highly encouraged since it adds more context to the model.

4) Assess the quality of data used as an input to the model. Feeding wrong data to a good analytics model and expecting it to produce the expected outcomes is an unreasonable expectation. The stakes are higher in high regulatory environments where the minimal error in the model might mean millions of dollars of lost revenues or penalties.

Even successful organizations who execute seamlessly in generating insights-struggle to "close the loop" in translating the insights into the field to drive shareholder value. It's always a good practice to pilot the model on a small population, link its insights and actions to key operational and financial metrics, measure the outcomes and then decide whether to improve or discontinue the model.

Appendix B Interviews on Big Data in Transport

Question 1: Are you involved in initiatives (Private, Public) in the area of Transportation where (Big) Data is used? Can you refer us to initiatives in the area of Transportation that you are aware of, where (Big) Data is used?

Answers:

Christopher Sciacca: Yes, we are currently involved in the H2020 project called "Automated driving Progressed by Internet Of Things" (AUTOPILOT). AUTOPILOT brings IoT into the automotive world to transform connected vehicles — moving "things" in the IoT ecosystem — into highly and fully automated vehicles. While using the IoT potential for automated driving, AUTOPILOT also makes data from autonomous cars available to the Internet-of-Things. The AUTOPILOT consortium represents all relevant areas of the IoT ecosystem. Thanks to AUTOPILOT, the IoT eco-system will involve vehicles, road infrastructure and surrounding objects in the IoT, with particular attention to safety-critical aspects of automated driving. AUTOPILOT IoT enabled autonomous driving cars are tested, in real conditions, at six permanent large-scale pilot sites in Finland, France, Italy, the Netherlands, South Korean and Spain. Find out more about the AUTOPILOT pilot sites here. And in a second project, called Up Drive for automated parking and driving will develop robust, general 360° object detection and tracking employing low-level spatiotemporal association, tracking and fusion mechanisms, Accurate metric localization and distributed geometrically consistent mapping in large-scale, semi-structured areas, Representations and mechanisms for efficient and cost-effective long-term data management across devices and scene understanding, starting from detection of semantic features, classification of objects, towards behavior analysis and intent prediction. As a result of this strategy, UP-Drive expects a significant technological progress that will benefit all levels of automation: from driver assistance in the short-term to full automation in the longer term – across a broad range of applications.

Gerhard Kress: Yes, I am involved in three projects: Siemens platform Mindsphere[1], Impact (train timeliness of >99%) on the Spanish rail transport (Railigent)[2], and The Railigent[3]. Basically, our Railigent platform builds on technologies from Mindsphere, enlarged with rail specific elements like data models/semantics, rail specific format translators and of course our applications and data analytics models. The foundation is a data lake in the cloud (Amazon Web Services: AWS) in which we store the data in a loosely coupled format and create the use case specific structures on reading. Data gets ingested in batch or stream, depending on the source and during the data ingest we already apply the first analytics models to validate and augment

[1] https://www.computerwoche.de/a/big-data-meets-heavy-metal,3329549

[2] https://www.siemens.com/innovation/de/home/pictures-of-the-future/digitalisierung-und-software/from-big-data-to-smart-data-heading-for-data-driven-rail-systems.html

[3] https://www.siemens.com/global/en/home/products/mobility/rail-solutions/services/digital-services/railigent.html

the data. For every step in the data lifecycle we use active notifications to move the data to the next stage and as much as it is possible we rely on platform services from AWS to build the applications. Our applications consist out of micro-services which we bundle in a common UI framework. And we have deployed a full CI/CD pipeline based on Jenkins.

Scott Jarr: Yes, I am involved in a company called Cambridge Mobile Telematics. We have developed software and hardware that enables drivers to track their driving behaviour and, in conjunction with insurance companies, make safer roads.

Dr. Alessandra Bagnato: I am not personally involved in the area of Transportation where Big Data is used at this moment but I work on Big Data on the DataBio Project, Unveiling benefits of Big Data technology in raw material production in agriculture, forestry, and fishery for a smart & sustainable data-driven bioeconomy. More details at www.databio.eu/en/ And I am aware of the following initiative: Big Data Magnolia Project at Société nationale des chemins de fer français (SNCF) with SAS.

Jeff Saltz: I am involved in one effort. It is a public effort to improve snow removal. The main source of data is GPS locations of the snow plows (in the city of Syracuse). The goal is to improve the efficiency and/or effectiveness of removing snow (or to improve the public's knowledge of the status of snow removal). Other work that I am aware of Autonomous droves (ex. for package delivery) - how to ensure effective drone navigation, Detecting / predicting where potholes (big holes in the street) will occur, and Travel directions (that avoids traffic, predicts the length of future trips)

Carlo Ratti: Let me answer both questions together. Big Data means a better knowledge of the urban environment. In relation to mobility, let me share an example: At MIT Senseable City Lab, we analyzed all taxi trips that connect the City of New York in a given year. The project-called HubCab– gathered 170 million taxi trips by over 13,000 Medallion taxis in New York City (NYC), with GPS coordinates of all pickup and drop off points and corresponding times. We then created a mathematical model to determine the potential impact of ridesharing applied to such vast database. The project introduced the concept of "share-ability networks" that allows for efficient modelling and optimization of the trip-sharing opportunities. Such an approach could lead to less traffic congestion, reduced operating costs and split fares, and to a less polluted environment.

Question 2: Do you know which Data Sources are typically used in these initiatives?

Answers:

Dr. Frank Wisselink: Insights from the Mobile Network

Christopher Sciacca: There are several papers available which outline the data sets[1]. Many appear to be readily available training data sets of videos and images. In the Up Drive project, they use The Cityscapes dataset is currently the most challenging benchmark considering that

[1]http://up-drive.eu/category/resources/publications/, Accessed April 13, 2018 and http://autopilot-project.eu/autopilot-library/, Accessed April 13, 2018

it consists of video sequences captured in traffic environments from 50 different cities. 5000 images were fully annotated and 25000 images were only coarsely annotated using 19 semantic classes.

Scott Jarr: Yes, it is driver data captured in real-time from either a smartphone application or a small in-car device.

Dr. Alessandra Bagnato: SNCF Open Data

Jeff Saltz: GPS data or location data from mobile devices.

Carlo Ratti: Let me still refer to HubCab. The basis of the HubCab tool is a data set of over 170 million taxi trips of all 13,500 Medallion taxis in NYC in 2011. The data set contains GPS coordinates of all pickup and drop off points and corresponding times. Other sources are data collected from a cellphone, either in opportunistic ways (a-la TomTom) or with special apps (think about Waze). Regarding software, we usually develop our own. In the case of Hub Cab, cartographic data of street shapes were obtained from OpenStreetMap. The streets were cut into over 200,000 street segments of 40m length each with a Python script and the help of the shapely Python library and imported into a MongoDB. Pickup and drop off points were matched to the closest street segments. Street types unlikely to contain taxi drop-offs or pickups, such as footpaths, trunks, service roads, etc. were not used in the matching process. Line widths of yellow and blue street segments on low zoom levels were styled on a logarithmic scale. The pickup and drop off points, represented as dots on the high zoom levels, were generated via an Arcpy script, being placed randomly within a box around a given street segment with the box width again following a logarithmic scale.

GPX (GPS Exchange format) files of the dots were styled using Maperitive, then merged and amended for different zoom levels. The dots and street line files were layered together with MapBox, which is the platform that streams all the map content. The data backend of HubCab runs on a MongoDB, containing all street segments and their coordinates, and all flows between each pair of street segments. The number of all possible street segment pairs is over 40 billion (200,000 times 200,000) per map. Radius selection is dynamic, using MongoDB's $near function to obtain flows from all segments within the radius of the pickup marker to all segments within the radius of the drop off the marker. With nine maps (one for the yearly data, eight for 3-hour time segments on all Fridays/Saturdays) and three selectable radii, there is a total of over one trillion flow combinations that can be explored with HubCab. Communication between MongoDB and the front end is realized via PHP scripts and Javascript+JSONP.

Question 3: Do you know what the main Challenges are in managing data for Transportation?

Answers:

Stephen Dillon: Reducing latency of both the calculation of analytics as well as delivery of the results is a major challenge. Companies are trying to move beyond theoretical to actual working architectures and solutions that provide Edge analytics and only send data Cloud platforms when needed.

Dr. Alessandra Bagnato: Helping Decision making and services for Transportation companies towards billions of travelers

Jeff Saltz: Similar to many domains, there can be a lot of data (ex. second by second GPS data), so the question about how much data to store and how to effectively access the data is key. For example, the GPS data noted in question 1, the vendor only provides minute by minute data (as opposed to second by second data) - some analysis would be better with more detailed data, but storing that amount of data was deemed not worth the expense.

Carlo Ratti: I would say the same issues that one encounters in all Big Data projects.

Question 4: Do you use (or are you aware of) any special software for managing data for Transportation?

Answers:

Stephen Dillon: *In-memory systems such as VoltDB are a specialized form of modern database or New SQL database designed to be able to perform such calculations with low latency. Apache Spark is another technology that allows for calculations and delivery with very low latency. Also; deployable stacks such as Azure Stack that provides the ability for Edge analytics is gaining in popularity.*

Scott Jarr: *From my understanding, it is a fairly common big data stack that includes streaming ingest, analytics, and a NoSQL storage engine.*

Dr. Alessandra Bagnato: *Not personally, but I know that SNCF uses SAS Visual Analytics, SAS Visual Statistics and SAS Office Analytics*[1]*.*

Gerhard Kress: Data analytics happens either in sandboxes when the model is still in development or in the full platform. We use mostly Python and pySpark, but are also using other technologies when needed (e.g. deep learning driven approaches).

In addition to the opportunities and challenges presented in the interviews above, one can also refer to the Q&A with Alan P. Amling (Vice President, UPS Corporate Strategy) on Logistics and 3D printing.

Question 5: 3D printing, also known as additive manufacturing (AM): How does it relate to the business of UPS?

Alan P. Amling: In the early days of 3D printing, when the primary use cases were models and prototypes, we thought the technology was interesting. However, as machine and material quality continued to increase, we saw the potential for using 3D printing for industrial production and the opportunity became compelling. As we looked at how this could play out, we realized that it could be a threat or an opportunity depending on the actions we take today. We acted. UPS is not a manufacturer, so we invested in one, Fast Radius, and integrated their operations into our global logistics network. In Louisville, for instance, we can print until 1 am and get the product anywhere in the U.S. by the next morning. We see 3D Printing as a logistics solution, another tool to help customers eliminate waste, delight customers and optimize their

[1] https://www.sas.com/fr_fr/news/press-releases/france/2015/Fevrier/sncf-gares-connexions-choisit-SAS-pour-son-projet-bigdata-magnolia.html

logistics networks. This is not new to UPS. This is what we've been doing for 111 years. We just have a new tool to do it.

Question 6: What is the On-demand manufacturing strategy of UPS?

Alan P. Amling: The commercial internet has enabled the on-demand consumer and UPS is an integral part of that supply chain. Industry 4.0 will do the same for manufacturing. We see manufacturing becoming more distributed with products/parts produced on-demand, in smaller quantities, more often, closer to the point of use. This will only apply to a small number of parts and products at first but will grow over time as additive manufacturing technologies and advancements in materials increase quality while decreasing time and cost.

By integrating on-demand manufacturing into our global logistics network, we're preparing UPS to be an integral part of the manufacturing ecosystem that will be developing over the next 10 years. We're using a "box-in-a-box" strategy to enable this network. The Fast Radius Additive Micro Factory on our Louisville supply chain campus is a great example. It's also important to note that our strategy is not focused on what we make, it's focused on what we make possible.

- How can we help businesses reduce waste and lower inventory by allowing them to store products in virtual inventory and print on demand? 3D printing holds the potential to invert the traditional supply and demand model, where demand actually comes before supply.
- How can we help our customers create new products customized for each individual? Today we accept "best fit" or "close enough". Tomorrow we'll be fitting everything from artificial knees to hammer handles to the individual instead of the other way around.
- How can we help a small entrepreneur in Cape Town build, test and sell one of their innovations to customers in New York without the high upfront cost and risk of building a mold or tooling for a product that may have to change several times? We call this manufacturing-as-a-service.

Question 7: What are the main advantages of distributed manufacturing?

Alan P. Amling: The big near-term advantage is the ability to meet customer demands while lowering inventory. For our customers, inventory is the devil. If the inventory turns slowly, like you find in critical parts that need to be stored close by for quick repairs, it's a big financial drain for companies. Storing parts in virtual inventory and printing on-demand is a great alternative. We're also seeing great demand for the personalized application, combining what used to be several parts into one 3d printed part, and creating unique and ultra-efficient designs that simply can't be manufactured using traditional methods.

Question 8: Are there any challenges when deploying distributed manufacturing?

Alan P. Amling: There are big challenges to the mainstream adoption of additive manufacturing. Determining what parts make sense to print, going through the certification process, making sure IP is protected, scaling the solution, and making sure you have engineers on staff that are familiar with the new technology are all "speed bumps" preventing progress. Because we were in the market early and identified these issues, it allowed us with our partner's Fast Radius and Carbon to create what we call the Application Launch Program (ALP). We take 2 engineers from 10 non-competing companies in each ALP cohort through a 6-month program with on-site and

field work to knock out each of these issues and get to initial production. The progress we've seen is very encouraging.

Question 9: UPS seems to sit at the intersection of the digital and physical world (digital commerce and physical delivery). Can you please tell us how is on-demand manufacturing allowing UPS to move your position in the supply chain up one notch?

Alan P. Amling: As digital technology shrinks the gap between supply and demand, it's incumbent upon UPS to reach further up and down the supply chain to drive new efficiencies and opportunities for our customers. While this is a new phase in our evolution, it's not our first rodeo. The original messenger service that started the company in 1907 was disrupted by an incredible technology…the telephone. It forced us to think differently and reconfigure our assets to find new ways to deliver value which gave birth to our package delivery business and the rest is history. This changes are part of our 111-year history and I'm excited to see how this next phase unfolds.

Appendix C Interview on Identification of Opportunities and Challenges

Interview with Jack Levis, Senior Director, Industrial Engineering at UPS:

Question 1: Can you give us a background on UPS and some of the challenges that UPS faces?

Answer: The e-Commerce revolution brings with it some interesting challenges. First, from a cost-to-serve standpoint, residential deliveries are less dense in terms of distance between stops and pieces per stop. Second, residential customers want personalization in their delivery experience, which adds cost and other challenges. UPS has met these challenges by utilizing operations technology and advanced analytics.

Question 2: **UPS won the 2016 Franz Edelman prize for its On-Road Integrated Optimization and Navigation system (ORION). You lead the four-year-long development of ORION. ORION completed a deployment in 2016. How did you manage to reduce 100 million miles driven annually and save UPS $300 to $400 million each year?**

UPS has a long history of innovation and constant improvement. The ORION story actually began in the late 1990s when UPS started building our PFT data infrastructure. PFT created predictive models, a "virtual network", and a suite of planning and execution tools. This was the foundation for ORION. PFT deployed in 2003 and reduced 85 million miles driven per year.

ORION was built on this robust foundation. Using the discipline of operations research, ORION created a proprietary route optimization brain. Research into ORION began in 2003, and the first model was field tested in 2008 in Gettysburg, Pennsylvania. The result was that ORION could find ways to serve all customers in a route while at the same time reducing cost.

It does so by systematically analyzing more than 200,000 different ways a route can be run and then presenting the optimal route to a driver. It does so in seconds. The ORION savings of 100 million miles and $300 million to $400 million annually is in addition to the savings from PFT.

Question 3: What data infrastructure and data analytics tools did you use to implement ORION?

Answer: As mentioned above, ORION sits on top of our proprietary PFT technology. The analytics tools are also built in-house by UPS's operations research team. The marriage of operations research, IT and business processes is part of the ORION success story.

Question 4: What were the main challenges and pitfalls you encountered in the project?

Answer: The first challenge was to build the ORION optimization engine (brain) that could not only meet service while reducing cost but do so while thinking like a driver. This meant that ORION needed to balance consistency and optimality. It made no sense to throw things up in the air just to save a penny. To do so, UPS had to re-evaluate business rules, methods, procedures, etc. In essence, UPS redesigned the delivery process.

The second challenge was ensuring the data fed to ORION was "pristine", and maps were a major challenge in this regard. Off-the-shelf maps were not accurate enough. UPS patented a

process for utilizing our "big data" infrastructure to help make maps accurate enough for ORION. For instance, if a speed limit changes or a bridge is out, this edit must be made and updated information sent to ORION. From the time the map is edited, an optimized route can be in a driver's hands in 30 seconds.

The third and largest challenge was changing management. By definition, optimization tools like ORION will require people to change behavior. UPS tested training procedures, new metrics, analysis tools, etc. to ensure the change would take place. Ultimately, the deployment team grew to 700 people. This team impacted tens of thousands of front-line personnel. The deployment team and the front line are the true heroes of the story.

Question 5: What are the three most important lessons learned from this project?

Answer: Never assume you know the answers. The first four years were spent truly understanding the delivery problem. There were many guidelines but few rules. ORION had to turn these guidelines into acceptable algorithms. The team did so by working with drivers and the front line until ORION started thinking like a driver.

Data is always a bottleneck. Initially, the team thought maps that could be purchased would be accurate enough. When the optimizations gave bad answers, the team looked at the algorithm. As it turned out, the problem was the map data accuracy and not the algorithm. The algorithm could not be truly tested until the data inputs were "pristine."

Don't forget deployment and change management. Do not think that "If you build it, they will come." ORION required extensive change management and front-line support. It is important to have support from the top and show that the results are achievable. Without understanding the importance of change management, new programs run the risk of becoming a "flavor of the month."

Question 6: Can you recommend three best practices so that other projects can have a smoother path?

Answer: Focus on decisions. Put effort into areas where a better decision will make an impact. Focus on deployment and data from the beginning. The ORION deployment strategy is now the standard for all operations projects at UPS. Ensure the right data infrastructure is in place with proper data. Utilize appropriate associations and networking for help like the Institute of Operations Research and the Management Sciences (INFORM).

Question 7: What were the key elements that made this project a success?

Answer: Support from senior management to allow the team to continue the research even when failures occurred. Proving that benefits could be achieved. The teams ran 11 different tests in multiple operations to test things like benefit achievement, training, metrics, best practices, etc.

No site could be deployed unless an entrance criterion was met. No site could say deployment was completed unless exit criteria were met. There was constant evaluation of metrics and issues. ORION was built inside a delivery process. Operations do not know they are using such advanced mathematics. They are just doing their job.

Question 8: What are the typical mistakes done when analyzing data for a large scale data project? How can they be avoided in practice?

Answer: We do NOT focus on technology. We focus on decisions. Big data is a how not a what. We care less about Big Data than we care about big insight and big impact.

By focusing on a decision that needs to be made, priorities become clearer. Ensuring decision-makers have the right information to make decisions and then measuring the impact of better decisions helps with the process. It helps to ensure the proper data is in place to make an impact. If an impact can be made from a simple tool, that is a good thing.

Question 9: What are the main barriers to applying machine learning at scale?

Answer: The largest barrier is not focusing on decisions.

Question 10: What is your next project?

Answer: UPS will build out ORION. ORION will begin making suggestions to drivers throughout the day. ORION will also provide navigation to drivers. We are also working on ORION making "dispatch" decisions. ORION will begin deciding which driver should serve customers. In essence, at some point, it will automate the pickup/delivery process. Simultaneously, UPS will provide ORION-like functionality in other areas of the business. There will be a PFT/ORION for Transportation; city-to-city movements. There will be a PFT/ORION for inside the building operations; sorting, loading, moving of vehicles, etc. […] The advances mentioned above along with automated facilities will begin to automate and optimize the network. Today, ORION optimizes a single driver. Tomorrow, ORION will begin to optimize an entire delivery area. UPS has a bold vision to optimize an entire network from end to end.