sd4m

Smart Data for Mobility

Unlocking new data value chains of multimodal mobility services

06
→ Introduction

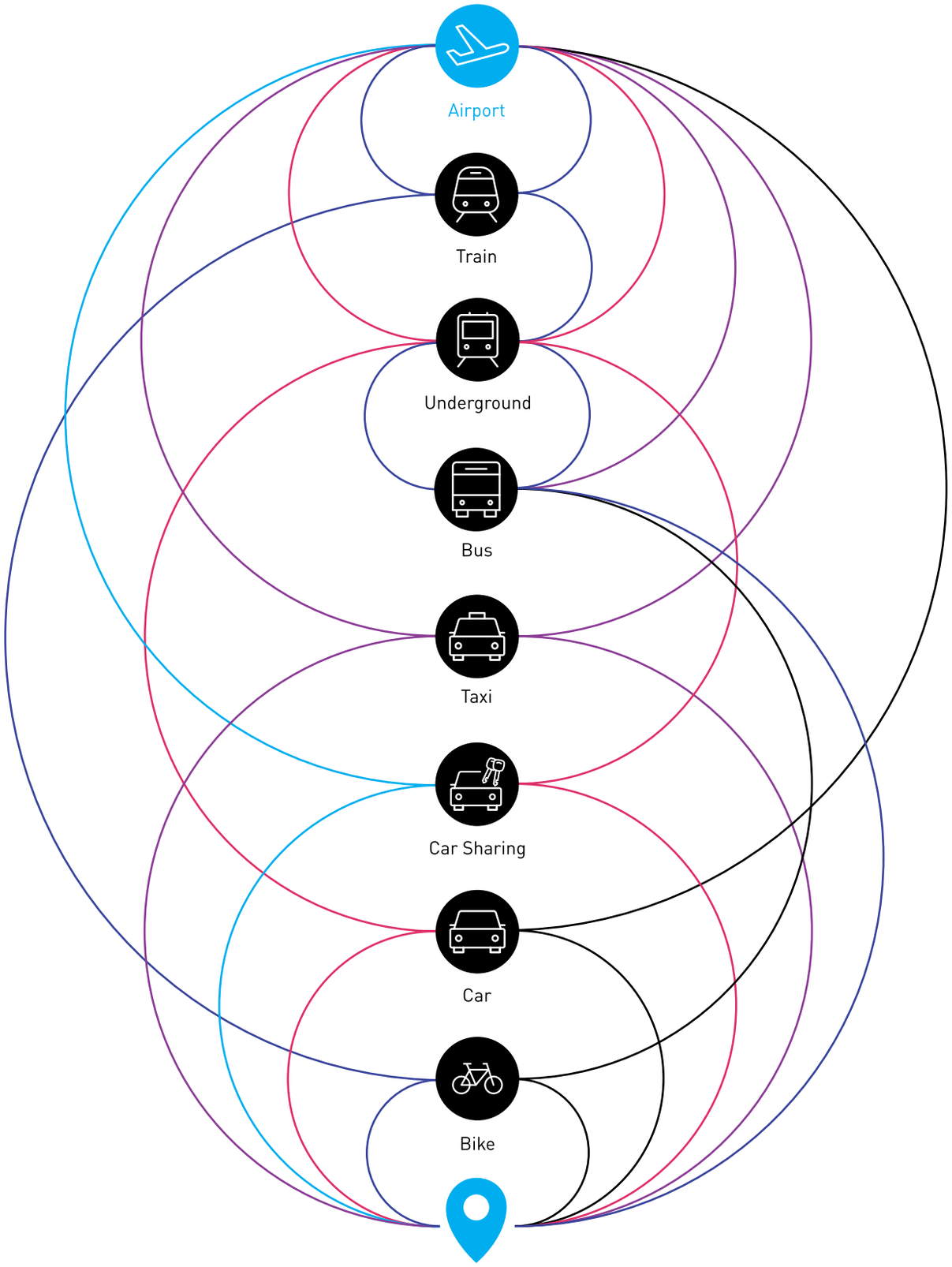
07
→ Untapped, unstructured and dynamic mobility data holds significant value

08
→ Unlocking new mobility Data Value Chains

11
→ Tackling unstructured and dynamic data: The SD4M Platform

13
→ Leveling the playing field, opening up avenues of innovation in mobility

14
→ Imprint & Contact



Introduction

Every day, millions of users access transport-planning applications on their devices (smart-phones, tablets, computers etc.) for everything from train timetables to routing options to ticket purchasing. Undoubtedly, mobility is an area in which technological advances have created immense value for customers. Despite improvements in the overall transport experience, however, a consider-able gap remains between the pre-journey (booking, planning, routing etc.), and the actual journey experience. There are no mobility information services today which are able to inform customers about changes and events anytime and anywhere on demand. However, the emergence of mobility relevant data resources from various channels enables the development of new effective and inno-vative data value chains (DVCs) for multi-modal services, which would significantly contribute to a sustainable improvement of the overall travel experience.

While the pre-journey, or trip-planning, experience has already improved dramatically, all too often actual travel is disrupted or compromised by real-time events, drastically reducing the value of any planning. Indeed, most mobility applications can tackle the pre-journey challenges quite competently - offering multi-modal (train, public transport, air travel, bike rental, car sharing, etc.) options for scheduling and routing - but then offer only little assistance once the customer has embarked on the actual journey. In essence, current state-of-the-art DVCs are by design limited, insofar as they are unable to adequately utilize real-time and dynamic data - much of which is, of course, unstructured. Thus, vast quantities of pertinent minute-to-minute social media data are left out of the mobility equation. To substantially enhance consumers' transport experience and un-leash the full potential of mobility data, the mobility DVC needs to be equipped to fully capture the potential of real-time and dynamic data.

Untapped, unstructured and dynamic mobility data holds significant value

The good news: current shortcomings in the mobility DVC are not fundamentally attributable to a shortage of real-time information. On the contrary, the difficulties lie in the nature and characteristics of relevant data sources at each stage of the mobility experience. In the mobility data ecosystem there are four different types of data, each origin-ating from a broad range of sources.

→ Data type 1: Structured transport data, including not only infrastructure data, i.e. local streets, highways, public transport locations, etc., but also structured data such as timetables, connections and routing options.

→ Data type 2: Relevant data regarding traffic flows, encompassing static data such as the numbers of registered vehicles in a certain area, but also dynamic and real-time information such as traffic jams, road accidents, radio cell data and auto-positioning data (GPS etc.).

→ Data type 3: Interconnecting contextual data for transport-relevant events. This includes, to name a few examples, information about weather conditions, public and school holidays, events such as concerts or football games, protests, fairs and local sights.

→ Data type 4: Highly unstructured, up-to-the-minute data about traffic flows, consisting of information provided by other travelers. This ranges from textual social media data like Twitter, Facebook and blogs to the official channels of transport providers (RSS feeds, etc.), as well as those of public authorities such as police, rescue services, etc.

Current mobility solutions build solely upon the first two types of data (planning & routing) - which are both static and dynamic, but always, to some extent, structured. The processing, streamlining and integration of structured data is on its own not an easy task, but mobility application developers have put their attention primarily on data visualization and user experience. Viewing this resource allocation through the lens of the DVC concept, however, the economic value creation by downstream activities is often undermined, because dynamic and unstructured data is not taken into account. Clearly, the extraction and integration of data from a much broader scope of data sources (data type 3 & 4), such as the unstructured textual data from social media activity, promises significant economic value creation.

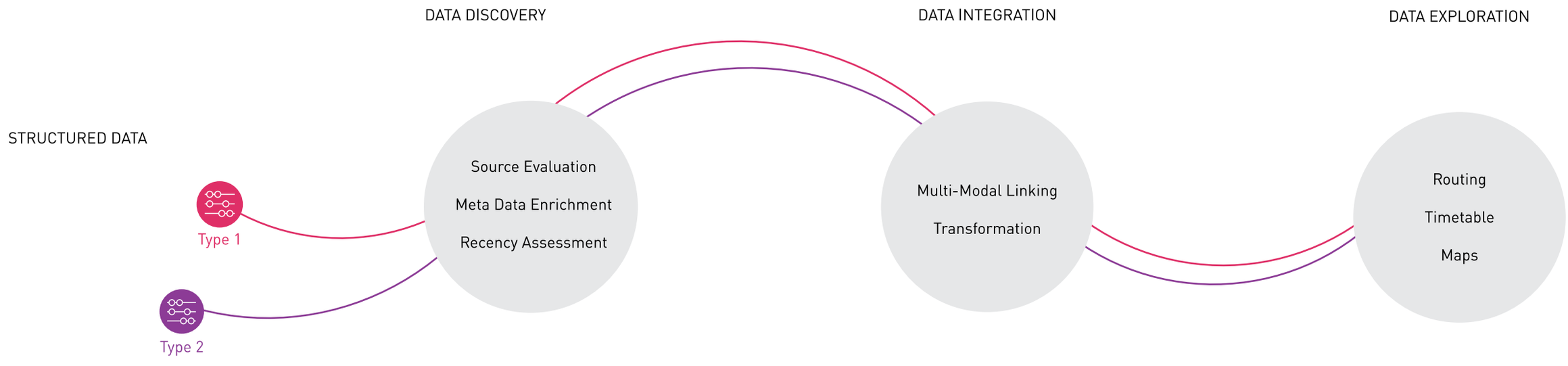
Unlocking new mobility Data Value Chains

What is a data value chain?
The data value chain (DVC) refers to all those activities and resources which support the applicability and value creation of data within a commercial context. DVC encompasses at least three steps: (1) Data Discovery, which includes data collection, selection, preparation and organization, (2) Data Integration, which describes the standardization and interoperability of various data sources and (3) Data Exploitation, which sums up the process of analysis and visualization as the prerequisites for decision-making. The concept of the value chain was originally proposed by Michael Porter. In recent years, scholars and experts have increasingly expanded this concept to the DVC, finding useful both its treatment of data as a resource and its allowing for quick assessment of value-generation and interdependencies at various steps.

The mobility DVC refers to all those types of activities and resources whose application supports value creation through data. Data source selection, preparation and organization are the essential first steps. The subsequent integration of data from disparate sources, thereby ensuring their semantic interoperability, is itself a tedious task - even when only structured data is considered. Naturally, complexity skyrockets when unstructured data is added into the equation. In order to save engineering and technology resources (and sometimes due to a lack of such resources), recent approaches have focused exclusively on structured data sources, resulting in an artificial narrowing of the data funnel, thereby simplifying data integration and exploitation.

This "shortcut" in the primary selection of sources of the mobility DVC hasn't, of course, had purely negative implications. Until now, mobility applications have focused on the creation of multi-modal concepts and have substantially improved the "look & feel" for users. However, by limiting the scope of data sources, even the broadest integration efforts can only predict the transport experience to the degree possible using exclusively structured and non-dynamic information. Consequently, the use-cases are, in actuality, limited to pre-journey planning and (to some degree) routing. Since the interoperability of unstructured and structured data in a coherent mobility application has not yet been achieved, the holy grail of mobility - seamless planning and real-time adaptation and execution - has up to now remained out-of-reach.

Clearly, the fundamental challenge is posed by unstructured and dynamic textual data. Any approach to unstructured textual data must be simultaneously adaptive across varied domains and robust to the challenging dimensions of big data (volume, velocity, variety, veracity). Systems, methods or strategies developed for general purposes, or for a specific domain, are often not directly transferable to other domains. Not surprisingly, domain adaptation has become a central research topic for big data text analytics. Creating an innovative framework for the mobility-specific analysis of unstructured data generates new value creation opportunities for downstream activities, such as data analysis and visualization.

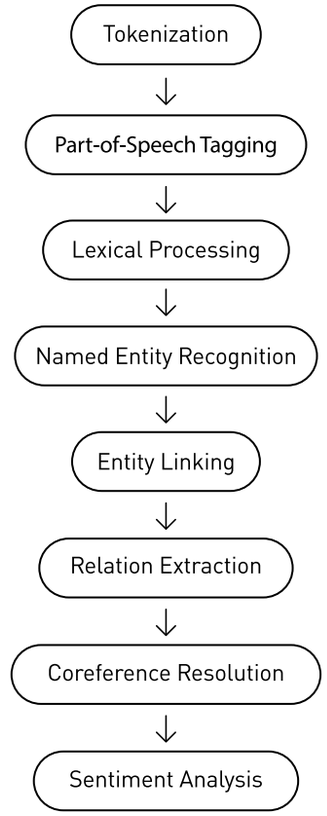

DATA DISCOVERY
DATA INTEGRATION
DATA EXPLORATION
STRUCTURED DATA
Type 1
Type 2
Source Evaluation
Meta Data Enrichment
Recency Assessment
Multi-Modal Linking
Transformation
Routing
Timetable
Maps
DATA DISCOVERY
DATA INTEGRATION
DATA EXPLORATION
STRUCTURED DATA
Type 1
Type 2
Type 3
UNSTRUCTURED DATA
Type 4
Source Evaluation
Meta Data Enrichment
Recency Assessment
Big Data Text Analytics
Outlier Detection
Multi-Modal Linking
Transformation
Analysis
Explorative Data Analysis
Hypothesis generation
Discover
Visualisation
Comprehension
Comparison
Aggreation
Decision Making
Planning
Forecasting

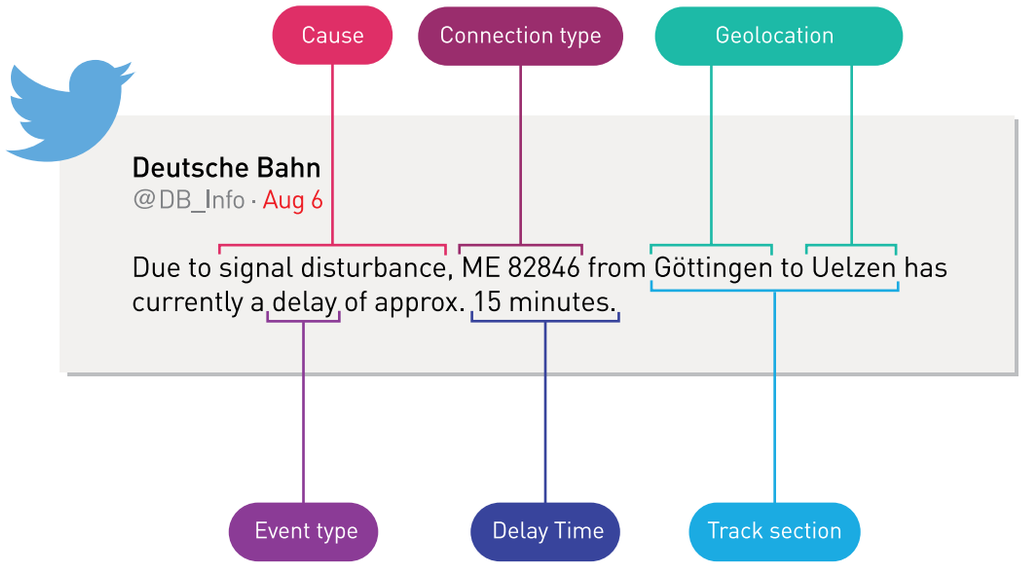
Tackling unstructured and dynamic data: The SD4M Platform

The SD4M Big Data Analytics Platform is the first comprehensive attempt to tackle the integration of static and dynamic, structured and unstructured mobility-related data, and to substantially expand the current data value chains of mobility.

Unstructured data - twitter posts, rss feeds and other social media updates - are primarily textual snippets which evoke meaning, convey information, and create context for the reader. In order to achieve interoperability between structured and unstructured data in a computational setting, unstructured textual sources require further processing and analysis in order for relevant structured information, such as semantic concepts (person, organisation, location, date and time, etc.), facts, events or opinions to be identified and extracted. The SD4M Big Data Analytics Platform builds upon a highly sophisticated stack of big text data analytics components, which are utilized to map and relate the relevant extracted information in the context of the transport sequence. To perform in a real-time environment, the platform must automatically recognize, establish and analyze those relations between the relevant entities or concepts that are salient to the user's needs in any given unstructured text. In light of the volume and velocity of unstructured text to be processed, the under-lying text analytics technologies should not only be domain-adaptive, but also equally robust and scalable, in order to provide for a reliable real-time data stream.

What is Big Data Text Analytics?
Big Data Text Analytics broadly refers to technologies and methods in computational linguistics and computer science used for the automatic detection and analysis of relevant information within large volumes of dynamic, real-time unstructured textual content. Machine learning and statistical methods are often employed for text analytics tasks. In the literature, text analytics is also regarded as a synonym for 1) text mining or 2) information and knowledge discovery from text. Major subtasks are 1) linguistic analysis; 2) named entity recognition and linking these entities to objects in the real word; 3) coreference resolution, 4) relation extraction; and 5) opinion and sentiment analysis. It enables the integration of live feeds from social media (such as tweets, Facebook posts, etc.) into a framework of more structured data.





Leveling the playing field, opening up avenues of innovation in mobility

SD4M provides the first open real-time big data analytics platform for mobility-related data and thus levels the playing field for application developers seeking to exploit such data for new and enhanced applications. In order to enable the full utilization of available data streams for widespread transport use, SD4M will expand current mobility data value chains by adding dynamic structured and unstructured data, thereby providing a broader, more stable, more reliable and fully open data platform for mobility services.

Given the fragmentation of data sources in the field of mobility - with transport providers, municipalities and other related entities being decidedly protective of their own data - as well as the absence of adequate technology, efforts in the last years have focused solely on the integration of structured mobility information. Now, recent technological advances in big text data analytics allow for scalable integration of large volumes of real-time, dynamic unstructured data into the mobility-specific data value chain, forming the basis for promising new value-creation downstream. This undoubtedly signals exciting new prospects, since transport providers can themselves utilize the API to improve their services, while mobility applications can likewise fine-tune their routing using real-time information. The successful integration of these vastly different types of data into mobility apps and services, at every step of the way, is an innovation of inestimable value, granting the consumer access to an improved up-to-the-minute transport experience - one which utilizes the full range and power of data.

Smart Data for Mobility
Imprint & Contact

Contact Data Value Chains in Mobility

Dr. Feiyu Xu
feiyu@dfki.de
+49 (30) 23895-1812

Dr. habil. Feiyu Xu is Senior Researcher and Head of the Research Group Text Analytics in the Language Technology Lab at the German Research Center of Artificial Intelligence. She is technical coordinator of the SD4M project. She has extensive experience in multilingual information systems, information extraction, text mining, big data analytics, business intelligence, question answering and mobile applications of NLP technologies. In 2012, Feiyu Xu has won the Google Focused Research Award for Natural Language Understanding. In 2014, Feiyu Xu was honored as DFKI Research Fellow.

Authors

Paul von Bünau (idalab GmbH)
Holmer Hemsen (DFKI GmbH)
Leonhard Hennig (DFKI GmbH)
Denis Herth (PS-Team GmbH & Co. KG)
Michael Merz (PS-Team GmbH & Co. KG)
Niels Reinhard (idalab GmbH)
Sven Schmeier (DFKI GmbH)
Christine Schwarz (]init[AG)
Philippe Thomas (DFKI GmbH)
Hans Uszkoreit (DFKI GmbH)
Dirk Wieczorek (]init[AG)
Feiyu Xu (DFKI GmbH)

SD4M Projekt Coordinator

Ingo Schwarzer
ingo.schwarzer@deutschebahn.com
+49 (30) 297-16370

DB Systel GmbH
Marktstr. 8
10317 Berlin

Dr. Paul von Bünau
paul.buenau@idalab.de
+49 (30) 814 513-14

Dr. Paul von Bünau is Managing Director of idalab, a Berlin-based agency for data science founded in 2003, specialising in machine learning, AI, mathematical modeling and data strategy. He has advised a wide range of clients from banking, mobility, e-commerce, pharma and the digital sector. A mathematician and computer scientist by training, Paul von Bünau holds a Ph.D. in statistical data analysis from TU Berlin.

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

www.sd4m.net