# D5.4 Development of a unified Big Data fusion framework for exploiting driving performance data in i-DREAMS.

*Interview with Md Rakibul Alam*

Besides the experiments, WP5 also deals with the development of the data framework that we use in i-DREAMS to fuse the data, which is the scope of this deliverable. D5.4 describes how the different aspects of big data are managed. Furthermore, it includes a discussion on several data sharing aspects within the project and after the project end. This time we talked with Md Rakibul Alam from the Technical University of Munich. He is the responsible author of D5.4.

**Hello Rakib, thank you for making time for me to talk about D5.4. The first thing I would like to ask is if you could explain to us what it is exactly that you mean when you talk about data fusion.**

RAKIB: *"In the deliverable we define data fusion as the process of collecting sensed information from several sources and integrating those together. The collected information using several sensors is combined to reach a better inference and more accurate results. In other words, data fusion is the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source."*

**How did you approach the development of a framework to fuse al the collected i-DREAMS DATA?**

RAKIB: *"In the deliverable you can read that we first did an extensive literature review of data fusion techniques in transport literature and then we discussed the relevant fusion techniques in the case of the i-DREAMS data. This helped us to understand the state-of-the-art in data fusion research and also the nature of data fusion that was needed for the i-DREAMS data in our big data framework."*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**Before we go into that framework, can you first briefly explain again what you exactly mean when you talk about big data?**

RAKIB: *"Simply said, big data are larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing tools and techniques can't manage them. But these massive volumes of data can be used to research issues you wouldn't have been able to tackle before."*

**You talk about massive volumes of data. Which i-DREAMS data are you referring to then in this context?**

RAKIB: *"The data we collect in i-DREAMS can be divided into two broad categories: raw and processed data[1] which are collected from various sources including sensors (for example GPS, Mobileye, Gateway, wearable and Cardiowheel), driving simulators, surveys and video cameras. The longer a single trip is, the higher the volume of the data associated with the trip becomes. Everything comes together in the back-office for later use in the project. The heterogeneous nature of i-DREAMS data requires data fusion prior to transfer to the back-office. The techniques used to perform this data fusion, depend on the type of data. And like I said before, we distinguish raw and processed data."*

**Can you explain the difference between raw and processed data?**

RAKIB: *"I will start with the raw data. Some examples are GPS coordinates, Mobileye events, vehicle acceleration, inter-beat intervals (IBI's) and travel time. These are collected via the Gateway, Mobileye, the wearable or Cardiowheel, depending on the transport mode. Explaining the different techniques that we use to fuse everything, might be a bit too technical, but it is important to remember that each type of data has a specific frequency to take into account. But there are other raw data as well: smartphone data and simulator data. The smartphone data are of course collected via the driver's smartphones and they include information such as: date, time, GPS coordinates, speed, accelerometer data, gyroscope data, type of activity (e.g. driving), screen state and post-trip data such as speed limits. In some cases, the frequency of collection depends on the specifications of the smartphone manufacturer. Simulator data are collected by driving simulators of course. Although we use different types of simulators, there were many common elements among them in terms of data collection. Driver demographics and other individual characteristics are collected using questionnaires."*

[1] More info in deliverable 4.3 (WP4)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**And the processed data?**

RAKIB: *"Processed data have already been fused once before processing, for example by the techniques we have unleashed on the raw data. However, they may still need aggregation for specific data analysis needs in i-DREAMS. For example, data may need to be aggregated in finer time intervals (e.g. 30 seconds) for real time analysis and in courser time intervals (e.g. 2 minutes) for post-trip analysis. As such, data aggregation is the final data fusion technique that is applied on the data. Aggregation methods again depend on the type of data and their specific purpose in i-DREAMS. What is important to note is that for the identification of the Safety Tolerance Zone (STZ) level in real-time, data collected from different sensors are aggregated at 30-second intervals (i.e. mean, min, max, average) and will be used to feed a Dynamic Bayesian Network for multi-level dynamic classification. In addition, for post-trip explanatory analysis of data, they will be aggregated in 2-minute time intervals or higher (i.e. trip level) and the aggregated data will be used in post-trip data analysis methods (i.e. Discrete Choice Models and Structural Equation Models)."*

**OK, after all the fusion work, data end up in a back-office. Can you elaborate a bit on that?**

RAKIB: *"The i-DREAMS project has a back-office which enables data management and facilitates future data analysis. It empowers consortium partners to store and retrieve data collected through road trials, simulations, surveys etc. In addition, the back-office provides a way to perform specific data analysis tasks on the raw data. The results of such analysis are stored in the back-end database. The back-office system allows all consortium partners to access those results, following a certain access strategy. The architecture of the back-office consists of three components: a communication component, a data storage component and a knowledge discovery component. Among these components, the knowledge discovery component does the data processing and analysis tasks including data fusion. This component needs to have an analytics framework that allows complex data processing and analysis of large datasets which is being collected by the project. The volume and heterogeneity of the data require the framework to have efficient ways to deal with the data."*

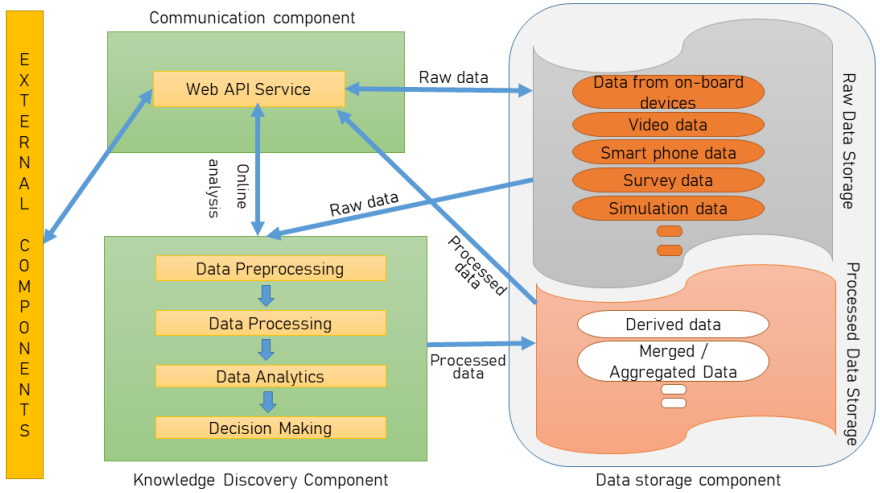


*Figure 1: i-DREAMS back-office components*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**If I understand correctly, then the most important part of the back-office is the data analytics framework, which is part of the knowledge discovery component. What will this framework do exactly?**

RAKIB: *"The scope of this framework is limited to loading different data sets from the database, fusing data to accelerate further analysis, performing the analytic tasks and then storing the results back to the database. These result sets then can be visualized in the analytics dashboard which is also part of the framework. Data fusion is organized in a very flexible way. Data can be fused together in different ways and in different levels. It is anticipated that different research questions will need specialized data fusion techniques designed particularly for the respective analysis to be performed. Therefore, implementing a specific type of data fusion won't be practical in the framework and its implementation. Rather, it is necessary to have the flexibility in the framework that can support different data fusion algorithm implementations in terms of efficient handling of the data."*

**You explained that the back-office consisted of three components (see Figure 1). What about the analytics framework?**

RAKIB: *"The big data analytics framework has four components. The data storage component contains the i-DREAMS data for research. The fusion and analytics API component serves as communication interface between the researcher and the analytics service component. Via the API, the researcher calls for a specific data fusion and/or analytics task. Responses are sent back to the researcher via the API. The analytics service component is comprised of various data fusion and analytics scripts. Based on the researchers' request this component first loads different datasets from the data storage. Then the appropriate data fusion task is performed on those different datasets to form a fused dataset. This fused dataset is then used in the analytics pipeline. Finally, when the analytics pipeline is finished working on the data, the results are stored back to the data storage for visualization and future usage. The dashboard component is used for visualization of the data stored in the data storage using a number of visualization techniques.*

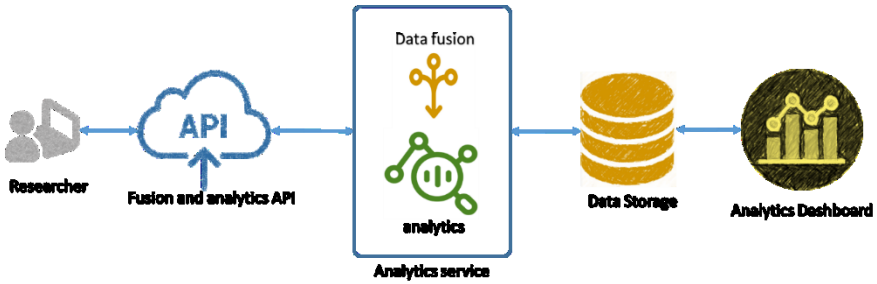


*Figure 2: Components of the i-DREAMS analytics framework*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**I can imagine that assembling such an architecture requires a lot of work and a lot of expertise. What will become of all those efforts after the project end?**

RAKIB: *"Transportation research has seen a fast growth in the number of Field Operational Tests (FOT) and Naturalistic Driving Studies (NDS) performed across the world to better understand the benefits of safety systems and the factors that cause the occurrence of incidents and accidents. Immense volumes of data have been collected through conducting FOT's and NDS's. Nevertheless, literature is still missing the data sharing related research to help understand the challenges and prospects of data sharing. In i-DREAMS, the explicit choice was made to go for the Open Access (OA) data sharing model, which refers to a set of principles and practices that enable access to research outputs through online distribution without any technical, monetary or strict copyright access barriers for users. In i-DREAMS, most of the challenges are covered."*

**Which challenges are you referring to?**

RAKIB: *"In general, there are numerous challenges to take into account. Firstly, the availability and accessibility of data itself does not make the data usable since a proper description of the dataset is usually required to understand the context and reasoning of data collection and quality of the dataset. So, proper metadata need to be available along with the actual datasets. A substantial effort is required to produce such metadata. Secondly, there is the matter of data ownership. Oftentimes, licensing is a way to deal with particular rules that partners may want to impose. But of course, there is also the matter of distribution and access to data and data tools, the nature of data usage that would be allowed during and after the project and of course there is the matter of post-project funding to host the data and tools supporting the data. Since the introduction of the GDPR privacy law somewhere mid 2018, there is also the matter of data protection to take into account, since NDS data sets contain privacy sensitive information that can be traced back to the individual participant, which has legal consequences. To overcome privacy constraints, data anonymization by filtering of sensitive information and aggregation of data are useful which lead to only publishing a selection of data properties and values. And of course, there is also the need for support and research services in order for re-using data to become a fact. Support services can comprise of documentation that helps researchers who are not familiar with the type of data. In the case of OA, researchers from a different field might be interested in using the data. Research services are more targeted towards doing part of the analysis work, such as extracting usable datasets for the data re-user or even perform the whole research itself and provide results. That is a very ideal situation to have that in place of course."*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**How are all of these challenges covered in i-DREAMS?**

RAKIB: *"Support and research services are currently provided by TUM within the project lifetime as TUM is managing the back-end access and storage. It is to be decided how the resources are going to be managed after the project ends. The availability of metadata is covered in several deliverables, such as D1.2, which is the Data Management Plan. Here the information on handling of research data during and after the end of the project is included. But more information about all the other challenges can also be found in deliverables D5.1, D4.2, D3.4 and D5.3. The funding aspect is covered in the framework agreement. There it is stipulated that the industrial partners will finance the maintenance and storage of the data infrastructure for a duration of three years, but knowledge institutions have the right to use the data for academic purposes."*

Thanks a lot, Rakib. I must say, I am again impressed with the work that is presented in this deliverable.

Edith Donders

i-DREAMS DisCom manager

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

**Deliverable 5.4 is part of WP5:**

***4-stage, 5-country experiment***

Download the report here

# i-DREAMER in the spotlight

## Md RAKIBUL ALAM

Graduated as *MSc in Informatics from Technical University of Munich*

Employed at *The Chair of Transportation Systems Engineering of the Technical University of Munich since 2020*

Passionate about *Data Engineering & Data Architecture.*

Tasks in i-DREAMS: *Lead of data and systems engineering and development for i-DREAMS' back-office. In charge of providing research and support services for i-DREAMS project partners, leading Deliverable 4.3 and assisting in others etc.*