D5.4 Ontwikkeling van een eengemaakt Big Data-fusieraamwerk voor de exploitatie van rijprestatiegegevens in i-DREAMS.

Interview met Md Rakibul Alam

Naast de experimenten houdt WP5 zich ook bezig met de ontwikkeling van het dataraamwerk dat we in i-DREAMS gebruiken om de gegevens samen te voegen, wat de focus is van dit rapport. D5.4 beschrijft hoe de verschillende aspecten van big data worden beheerd. Verder bevat het een bespreking van verschillende aspecten m.b.t. het delen van data binnen het project en na afloop van het project. Deze keer spraken we met Md Rakibul Alam van de Technische Universiteit München. Hij is de verantwoordelijke auteur van D5.4.

Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

Hallo Rakib, bedankt dat u tijd voor me heeft vrijgemaakt om over D5.4 te praten. Het eerste wat ik zou willen vragen is of u ons kunt uitleggen wat u precies bedoelt als u het heeft over datafusie.

RAKIB: *"In het rapport definiëren wij datafusie als het proces van het verzamelen van gedetecteerde informatie uit verschillende bronnen en het samenvoegen daarvan. De verzamelde informatie van verschillende sensoren wordt gecombineerd om tot betere conclusies en nauwkeurigere resultaten te komen. Met andere woorden, datafusie is het proces van integratie van meerdere gegevensbronnen om meer consistente, nauwkeurige en bruikbare informatie te produceren dan die welke door elke afzonderlijke gegevensbron wordt verstrekt."*

Hoe heeft u de ontwikkeling van een raamwerk om alle verzamelde i-DREAMS DATA samen te voegen aangepakt?

RAKIB: *"In het rapport kunt u lezen dat we eerst een uitgebreid literatuuronderzoek hebben gedaan naar gegevensfusietechnieken in de transportliteratuur en vervolgens de relevante fusietechnieken hebben besproken met betrekking tot de i-DREAMS-data. Dit hielp ons om de state-of-the-art in het onderzoek naar datafusie te begrijpen en ook de aard van de datafusie die nodig was voor de i-DREAMS-data in ons big data raamwerk."*

Kunt u, voordat we op dat raamwerk ingaan, eerst nog eens kort uitleggen wat u precies bedoelt als u het over big data heeft?

RAKIB: *"Eenvoudig gezegd zijn big data grotere, complexere datasets, vooral uit nieuwe gegevensbronnen. Deze datasets zijn zo omvangrijk dat traditionele gegevensverwerkingsinstrumenten en -technieken ze niet aankunnen. Maar deze enorme hoeveelheden gegevens kunnen worden gebruikt voor onderzoek naar zaken die voorheen niet konden onderzocht worden."*

U heeft het over enorme hoeveelheden data. Welke i-DREAMS-data bedoelt u dan precies in deze context?

RAKIB: *"De data die wij in i-DREAMS verzamelen kunnen in twee grote categorieën worden verdeeld: ruwe en verwerkte data die worden verzameld via diverse bronnen, waaronder sensoren (bv. GPS, Mobileye, Gateway, de armband en Cardiowheel), rijsimulatoren, enquêtes en videocamera's. Hoe langer een rit is, hoe groter het volume van de gegevens die met de rit samenhangen. Alles komt samen in de back-office voor later gebruik in het project. De heterogene aard van de i-DREAMS-data vereist dat de data worden samengevoegd voordat ze naar de back-office worden overgebracht. De technieken die worden gebruikt om deze fusie uit te voeren, hangen af van het soort data. En zoals ik al zei, maken we onderscheid tussen ruwe en verwerkte data."*

Kunt u het verschil uitleggen tussen ruwe en verwerkte data?

RAKIB: *"Ik begin met de ruwe data. Enkele voorbeelden zijn gps-coördinaten, Mobileye gebeurtenissen, versnellingen van het voertuig, inter-beat intervallen (IBI's) en reistijd. Deze worden verzameld via de Gateway, Mobileye, de armband of Cardiowheel, afhankelijk van de vervoerswijze. Het is misschien een beetje te technisch om de verschillende technieken uit te leggen die we gebruiken om alles samen te voegen, maar het is belangrijk om te onthouden dat elk type data een specifieke frequentie heeft waarmee rekening moet worden gehouden. Maar er zijn ook andere ruwe gegevens: smartphonegegevens en simulatorgegevens. De smartphonedata worden uiteraard verzameld via de smartphones van de bestuurder en omvatten informatie zoals: datum, tijd, gps-coördinaten, snelheid, versnellingsmetergegevens, gyroscoopgegevens, type activiteit (bv. rijden), schermtoestand en gegevens na de rit, zoals snelheidsbeperkingen. In sommige gevallen hangt de verzamelfrequentie af van de specificaties van de smartphonefabrikant. Simulatordata worden uiteraard verzameld door rijsimulatoren. Hoewel we verschillende soorten simulatoren gebruiken, waren er veel gemeenschappelijke elementen bij de gegevensverzameling. Demografische gegevens van de bestuurder en andere individuele kenmerken worden verzameld aan de hand van vragenlijsten."*

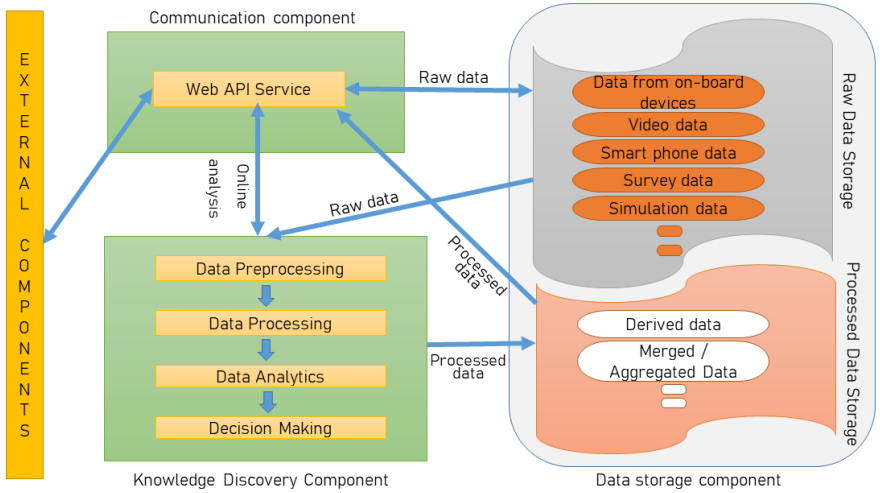
Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

En de verwerkte data?

RAKIB: *"Verwerkte data zijn voorafgaand aan de verwerking al een keer samengevoegd, bv. door de technieken die wij op de ruwe data hebben losgelaten. Soms moeten ze nog steeds geaggregeerd worden voor specifieke data-analysebehoeften in i-DREAMS. Zo kan het nodig zijn data te aggregeren in fijnere tijdsintervallen (bv. 30 seconden) voor analyse in real-time en in langere tijdsintervallen (bv. 2 minuten) voor analyse achteraf. Als zodanig is gegevensaggregatie de laatste datafusietechniek die op de gegevens wordt toegepast. Aggregatiemethoden hangen weer af van het soort gegevens en het specifieke doel ervan in i-DREAMS. Wat belangrijk is om op te merken, is dat voor de identificatie van de fase in de Veiligheids-Tolerantie-Zone (VTZ) in real-time, gegevens van verschillende sensoren worden geaggregeerd met intervallen van 30 seconden (d.w.z. gemiddelde, min, max) en zullen worden gebruikt om een Dynamisch Bayesiaans Netwerk te voeden voor dynamische classificatie op meerdere niveaus. Bovendien worden de gegevens voor de verklarende analyse na de rit geaggregeerd in tijdsintervallen van 2 minuten of meer (d.w.z. op ritniveau) en worden de geaggregeerde gegevens gebruikt in methoden voor gegevensanalyse na de rit (d.w.z. Discrete Keuzemodellen en Structurele Vergelijkingsmodellen)."*

OK, na al het fusiewerk, belanden de gegevens in een back-office. Kunt u daar wat dieper op ingaan?

RAKIB: *"Het i-DREAMS-project heeft een back-office die databeheer mogelijk maakt en toekomstige data-analyse vergemakkelijkt. Het stelt de consortiumpartners in staat gegevens die zijn verzameld tijdens wegproeven, simulaties, enquêtes enz. op te slaan en op te vragen. Bovendien biedt de back-office een manier om specifieke data-analysetaken uit te voeren op de ruwe gegevens. De resultaten van dergelijke analyses worden opgeslagen in de back-end database. Via het back-officesysteem hebben alle consortiumpartners toegang tot die resultaten, volgens een bepaalde toegangsstrategie. De architectuur van de back-office bestaat uit drie componenten: een communicatiecomponent, een gegevensopslagcomponent en een kennisontdekkingscomponent. Van deze componenten voert de kennisontdekkingscomponent de gegevensverwerking en -analyse uit, met inbegrip van gegevensfusie. Deze component moet beschikken over een analytisch kader dat complexe dataverwerking en -analyse van grote datasets die door het project worden verzameld, mogelijk maakt. Het volume en de heterogeniteit van de gegevens vereisen dat het raamwerk efficiënte manieren heeft om met de data om te gaan.*



Figuur 1: i-DREAMS back-office componenten

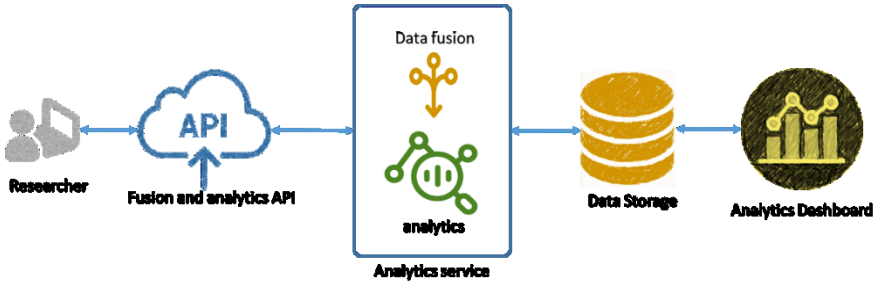
Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

Als ik het goed begrijp, is het belangrijkste onderdeel van de back-office het raamwerk voor gegevensanalyse, dat deel uitmaakt van het onderdeel kennisontdekking. Wat gaat dit raamwerk precies doen?

RAKIB: *"Het focus van dit raamwerk is beperkt tot het laden van verschillende gegevenssets uit de database, het samenvoegen van gegevens om verdere analyse te versnellen, het uitvoeren van de analytische taken en vervolgens het opslaan van de resultaten terug naar de database. Deze resultatensets kunnen vervolgens worden gevisualiseerd in het dashboard dat ook deel uitmaakt van het raamwerk. Gegevensfusie is zeer flexibel georganiseerd. Gegevens kunnen op verschillende manieren en op verschillende niveaus worden samengevoegd. Het is te verwachten dat voor verschillende onderzoeksvragen gespecialiseerde datafusietechnieken nodig zijn die speciaal voor de desbetreffende analyse zijn ontworpen. Daarom zal het implementeren van een specifiek type gegevensfusie niet praktisch zijn in het kader en de uitvoering ervan. Het raamwerk moet veeleer flexibel genoeg zijn om verschillende implementaties van algoritmen voor gegevensfusie te ondersteunen in termen van efficiënte gegevensverwerking."*

U legde uit dat de back-office uit drie componenten bestaat (zie Figuur 1). Hoe zit dat met het analytisch raamwerk?

RAKIB: *"Het raamwerk voor big data-analyse bestaat uit vier componenten. De gegevensopslagcomponent bevat de i-DREAMS-data voor onderzoek. De fusie en analyse API component dient als communicatie interface tussen de onderzoeker en de analyse service component. Via de API roept de onderzoeker een specifieke datafusie- en/of analysetaak op. De antwoorden worden via de API teruggestuurd naar de onderzoeker. De analysecomponent bestaat uit verschillende scripts voor datafusie en -analyse. Op basis van het verzoek van de onderzoekers laadt deze component eerst verschillende datasets uit de gegevensopslag. Vervolgens wordt op die verschillende datasets de juiste datafusietaak uitgevoerd om een samengevoegde dataset te vormen. Deze samengevoegde dataset wordt vervolgens gebruikt in de analysepijplijn. Wanneer de analysepijplijn klaar is met de gegevens, worden de resultaten opgeslagen in de gegevensopslag voor visualisatie en toekomstig gebruik. De dashboardcomponent wordt gebruikt voor de visualisatie van de in de gegevensopslag opgeslagen gegevens met behulp van een aantal visualisatietechnieken.*



Figuur 2: Componenten van het i-DREAMS analytisch raamwerk

Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

Ik kan me voorstellen dat het samenstellen van een dergelijke architectuur veel werk en veel expertise vereist. Wat komt er na afloop van het project terecht van al die inspanningen?

RAKIB: *"Bij transportonderzoek is het aantal Field Operational Tests (FOT) en Naturalistic Driving Studies (NDS) wereldwijd snel toegenomen om een beter inzicht te krijgen in de voordelen van veiligheidssystemen en de factoren die incidenten en ongevallen veroorzaken. Door het uitvoeren van FOT's en NDS'en zijn enorme hoeveelheden gegevens verzameld. Toch ontbreekt in de literatuur nog onderzoek naar het delen van gegevens om inzicht te krijgen in de uitdagingen en vooruitzichten van data-uitwisseling. In i-DREAMS is uitdrukkelijk gekozen voor het Open Access (OA) model voor data-uitwisseling, dat verwijst naar een reeks beginselen en praktijken die toegang tot onderzoeksresultaten mogelijk maken via online verspreiding zonder technische, monetaire of strikte auteursrechtelijke belemmeringen voor gebruikers. In i-DREAMS komen de meeste uitdagingen afgedekt."*

Welke uitdagingen bedoelt u?

RAKIB: *"Globaal gezien zijn er talrijke uitdagingen waarmee rekening moet worden gehouden. Ten eerste maakt de beschikbaarheid en toegankelijkheid van gegevens op zich de gegevens nog niet bruikbaar, aangezien een goede beschrijving van de dataset meestal nodig is om de context en de redenering van de dataverzameling en de kwaliteit van de dataset te begrijpen. Er moeten dus goede metadata beschikbaar zijn, samen met de eigenlijke datasets. Een aanzienlijke inspanning is om dergelijke metadata te produceren. Ten tweede is er de kwestie van de eigendom van de gegevens. Vaak zijn licenties een manier om te voldoen aan bepaalde regels die partners willen opleggen. Maar er is natuurlijk ook de kwestie van distributie van en toegang tot gegevens en gegevensinstrumenten, de aard van het gegevensgebruik dat tijdens en na het project wordt toegestaan, en natuurlijk de kwestie van financiering na afloop van het project voor het hosten van de gegevens en de instrumenten die de gegevens ondersteunen. Sinds de invoering van de GDPR-privacywet medio 2018 moet ook rekening worden gehouden met gegevensbescherming, aangezien NDS-datareeksen privacygevoelige informatie bevatten die kan worden herleid tot de individuele deelnemer, wat juridische gevolgen heeft. Om privacybeperkingen te ondervangen, zijn anonimisering van gegevens door het filteren van gevoelige informatie en aggregatie van gegevens nuttig, wat leidt tot het publiceren van slechts een selectie van gegevenseigenschappen en -waarden. En natuurlijk is er ook behoefte aan ondersteunende en onderzoeksdiensten om hergebruik van gegevens mogelijk te maken. Ondersteunende diensten kunnen bestaan uit documentatie die onderzoekers helpt die niet vertrouwd zijn met het soort gegevens. In het geval van OA kunnen onderzoekers uit een ander vakgebied geïnteresseerd zijn in het gebruik van de gegevens. Onderzoeksdiensten zijn meer gericht op het doen van een deel van het analysewerk, zoals het extraheren van bruikbare datasets voor de hergebruiker van de gegevens of zelfs het uitvoeren van het hele onderzoek zelf en het verstrekken van resultaten. Dat is natuurlijk een heel ideale situatie."*

Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

Hoe worden al deze uitdagingen in i-DREAMS afgedekt?

RAKIB: *"Ondersteunende en onderzoeksdiensten worden momenteel door TUM geleverd tijdens de looptijd van het project, aangezien TUM de back-end toegang en opslag beheert. Er moet nog worden besloten hoe de middelen na afloop van het project zullen worden beheerd. De beschikbaarheid van metadata wordt behandeld in verschillende deliverables, zoals D1.2, dat het Data Management Plan is. Hier is informatie opgenomen over de behandeling van onderzoeksgegevens tijdens en na afloop van het project. Maar meer informatie over alle andere uitdagingen is ook te vinden in de deliverables D5.1, D4.2, D3.4 en D5.3. Het financieringsaspect wordt behandeld in de kaderovereenkomst. Daarin is bepaald dat de industriële partners het onderhoud en de opslag van de data-infrastructuur financieren voor een periode van drie jaar, maar dat de kennisinstellingen het recht hebben de gegevens voor academische doeleinden te gebruiken."*

Bedankt, Rakib. Ik moet zeggen dat ik opnieuw onder de indruk ben van het geleverde werk dat in dit document wordt gepresenteerd.

Edith Donders

i-DREAMS DisCom manager

Dit project is gefinancierd door het onderzoeks- en innovatieprogramma Horizon 2020 van de Europese Unie in het kader van subsidieovereenkomst nr. 814761.

Rapport 5.4 is deel van WP5:
4 fasen, 5 landen experiment
Download het rapport hier

i-DREAMER in de kijker

Md RAKIBUL ALAM

Afgestudeerd als *MSc in Informatica aan de Technische Universiteit München*

Werkzaam bij *The Chair of Transportation Systems Engineering van de Technische Universiteit Münchens sinds 2020*

Gepassioneerd door *Data Engineering & Data Architectuur.*

Taken in i-DREAMS: *Hoofd van data en systeem engineering en ontwikkeling voor de back-office van i-DREAMS. Belast met onderzoek en ondersteunende diensten voor i-DREAMS projectpartners, leiden van Deliverable 4.3. en het assisteren bij andere enz.*