VAN DATA NAAR BRUIKBARE INFORMATIE

LECTORALE REDE | Ander de Keijzer

VAN DATA NAAR BRUIKBARE INFORMATIE

LECTORALE REDE
Ander de Keijzer

INHOUD

INLEIDING

Ik fiets graag, tenminste als het niet regent of al te hard waait. Om boodschappen te gaan doen, of gewoon om even buiten te zijn. Ik woon in Zeist en hoewel dat een zeer groene omgeving is, moet je toch eerst langs de nodige verkeerslichten voordat je ongehinderd en zonder oponthoud kunt fietsen. Als ik voor mijn plezier een stuk ga fietsen, wil ik uiteraard zo snel mogelijk buiten bereik van de verkeerslichten zijn, maar ook als ik verkeerslichten tegenkom, wil ik daar zo min mogelijk voor hoeven wachten. Bij sommige verkeerslichten, maar dat worden er steeds meer, wordt tegenwoordig aangegeven hoe lang het nog duurt voor het verkeerslicht op groen springt. Die aanduiding is niet in seconden, maar veelal aangegeven door een cirkel van lampjes, die een voor een uitgaan. Als alle lampjes uit zijn, mag je als fietser doorrijden (zie ook figuur 1).

figuur 1: wachttijdindicatie voor de fiets.

In theorie is dat natuurlijk een mooie vinding. Als fietser weet je tijdens het wachten waar je aan toe bent, hoe lang je geduld moet hebben en wanneer je weer door mag, maar helaas is dat inderdaad slechts theorie. Zo'n indicatie werkt namelijk alleen maar als de fietser weet hoe het verloop van het uitgaan van de lichtjes zal zijn. Anders gezegd, wanneer door het verloop van de lampjes te voorspellen is wanneer de wachttijd voorbij is. Daarbij hebben we in dit geval een aantal problemen:

- De voortgang van de wachttijd is alleen te bepalen als de fietser het ritme van het uitgaan van de lampjes kent. Dat betekent dat er minimaal enige tijd overheen zal gaan, voordat het interval tussen de verschillende lampjes en de extrapolatie naar de gehele cirkel kan worden gemaakt.
- De extrapolatie zoals genoemd in het vorige punt, kan alleen worden gedaan als de lampjes in een voorspelbaar ritme uitgaan. Dat ritme moet dan bovendien zo zijn, dat door een (willekeurige) persoon in te schatten is hoe zich dat verhoudt tot de gehele wachttijd.

De wachttijd voor de fietser zelf is echter (enigszins) onvoorspelbaar. Die wordt namelijk bepaald door het verkeer op de overige wegen van het kruispunt. Als er veel verkeer is op kruisende wegen, zul je als fietser langer moeten wachten, terwijl je vrijwel direct door kunt als er zich geen verkeer op de andere wegen bevindt. Dit betekent dat de aanduiding voor het wachten per definitie onnauwkeurig zal zijn. Dit zie je ook terug in de wachtindicatie, aangezien de lampjes niet altijd met dezelfde snelheid uitgaan, er soms een aantal lampjes tegelijk uitgaat, of er zelfs soms lampjes weer aangaan in plaats van uit. Het probleem in dit geval zit dus vooral in het feit dat van een onvoorspelbare situatie toch een voorspelling gemaakt wordt, terwijl dat eigenlijk helemaal niet kan.

Kortom, hoewel een goed idee, kan er met behulp van de beschikbare data voor de fietser vaak helemaal geen correcte voorspelling van de te wachten tijd worden gedaan. Uiteraard is dat in dit geval helemaal niet zo'n punt en is die onnauwkeurigheid voor de fietser, hoewel misschien ergerlijk (zeker als het regent), geen onoverkomelijk probleem. Bij andere toepassingen is dit uiteraard heel anders. Denk bijvoorbeeld aan de besturing van zelfrijdende auto's. Daarbij moeten beslissingen direct genomen worden, terwijl de omgeving relatief onvoorspelbaar is. De omgeving kan weliswaar met camera's bekeken worden, maar de acties van de overige bestuurders zijn niet altijd even goed te voorspellen. Die voorspelbaarheid wordt zelfs nog lastiger

als bestuurders de verkeersregels negeren. Als zo'n zelfrijdende auto te laat reageert op overig verkeer zijn de gevolgen waarschijnlijk niet te overzien. De wijze waarop het systeem van een auto omgaat met de data is daarom gelukkig anders dan die bij de wachtindicatie voor fietsers.

Het voorbeeld van de wachtindicatie geeft aan dat het omzetten van data naar bruikbare informatie niet triviaal is. De applicatie, in dit geval de wachttijdindicatie, wordt door velen gebruikt, terwijl de eigenlijke informatievoorziening incorrect is. In plaats van data om te zetten naar informatie, wordt een bepaalde soort informatie (aanwezigheid van verkeer) omgezet in andere data (lampjes die aangeven welk deel van de wachttijd nog rest), terwijl er niet de beoogde informatie (hoe lang duurt het nog voor fietsers weer mogen doorrijden) van is gemaakt. Aan de hand van de onderzoekslijnen van het lectoraat Data Science & ICT geef ik enkele voorbeelden van methoden en technieken in de informatica om met data om te gaan en waar de uitdagingen liggen voor onderzoek op dit gebied.

DATAONTWIKKELINGEN

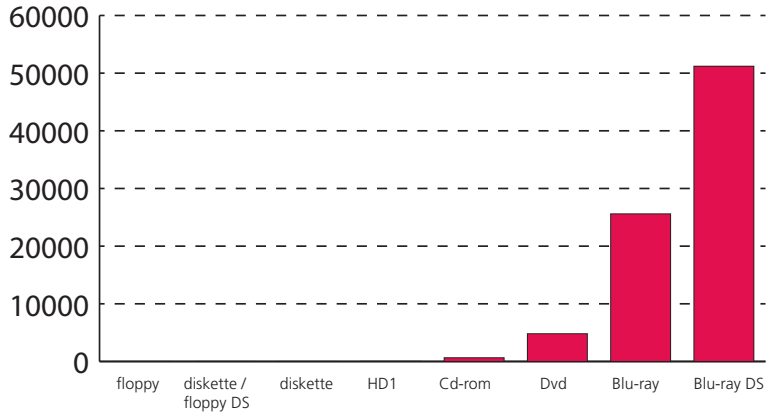
Het feit dat het lectoraat Data Science & ICT bestaansrecht heeft, is mede te danken aan de enorme ontwikkelingen in de ICT. Zonder de miniaturisering van hardware zouden we veel werk nog gewoon met pen en papier doen. In de jaren 40 van de vorige eeuw begon het computertijdperk met de ENIAC[1], de eerste elektronische computer. Toegegeven, velen van ons zouden hem niet als computer herkennen. Het apparaat besloeg een enorme ruimte en kon slechts een heel beperkte hoeveelheid informatie aan. Het was echter een belangrijke stap voor de ontwikkelingen die daarna zouden volgen. Die ontwikkelingen gaan ook steeds sneller. Het heeft vanaf de ENIAC nog ongeveer 40 jaar geduurd voordat computers voor thuisgebruik beschikbaar waren. Nu, weer ongeveer 30 jaar verder, heeft niet alleen het merendeel van de bevolking een computer thuis, maar zitten er computerchips in heel veel apparaten. Telefoons, tablets, afwasmachines, auto's, televisies, allemaal apparaten waar chips in zitten. Als we elke chip als computer zien, heeft een gemiddelde persoon heel wat computers in huis.

Als gevolg van de toename van het aantal computers, is ook de hoeveelheid geproduceerde data enorm toegenomen. Een goed voorbeeld van de toename in grootte is de opslagcapaciteit van de media die we thuis in gebruik hebben. Toen mijn ouders hun eerste computer kochten, een XT 8088, met een snelheid van 5 MHz (die overigens met behulp van het programma speed op te voeren was tot 7 MHz), zaten daar twee disc drives in, een voor floppy's met een opslagcapaciteit van 360 KB en een voor diskettes met een opslagcapaciteit van 720 KB. Beide waren door gebruik te maken van high density varianten te verdubbelen tot respectievelijk 720 KB en 1,44 MB. De harddisk van deze computer was 20 MB groot. Na enige tijd hebben we die vervangen door een harddisk van 40 MB. Dat was voor die tijd zo veel opslagcapaciteit dat ik toen, al was het maar tijdelijk, echt dacht dat we voor de rest van ons leven voldoende opslagcapaciteit hadden.

[1]Electronic Numerical Integrator and Computer

Na de diskettes kwamen de beschrijfbare cd-roms van 650 MB, gevolgd door de dvd waar 4,7 GB op past. De huidige blu-ray schijven zijn 25 GB groot en double sided zelfs 50 GB.

Uitgezet in een grafiek (zie figuur 2), is te zien dat de toename in grootte enorm is. De grootte van de floppy's, diskettes en harddisk van 20 MB zijn zelfs zo klein ten opzichte van de overige opslagmedia, dat deze niet eens in de grafiek te zien zijn.



Figuur 2: opslagcapaciteit van verschillende media.

Een tijdje geleden ging ik drie dagen naar het buitenland en omdat je tegenwoordig erg gewend bent aan internet op de smartphone, had ik een databundel aangeschaft om in het buitenland ook te kunnen internetten. Dat had ik al eens eerder gedaan toen ik anderhalve week wegging en toen ging het om een databundel van 500 MB, waarvan ik ruim de helft over had toen ik terugkwam in Nederland. Ik ging er dan ook vanuit dat ik met die 500 MB ruimschoots voldoende zou hebben voor drie dagen. Bij terugkomst in Nederland bleek echter dat ik vrijwel de gehele databundel had gebruikt.

Nu is 500 MB een nogal abstract begrip en het is lastig om gevoel te krijgen voor hoeveel data dat precies is. Een bekend studieboek over databases is Database Systems: The Complete Book [6]. Dat boek heeft ongeveer 1200 blz. Elke bladzijde heeft ongeveer 45 regels tekst en elke volledige regel tekst bestaat uit 80 karakters. Het totale boek bevat dus ruwweg 1200 * 45 * 80 = 4.320.000 karakters. Als we er voor het gemak van uitgaan dat elk karakter één byte is, is dat voor het totale boek dus iets meer dan 4 MB[2]. De 500 MB data die ik gebruikt heb in drie dagen in het buitenland, is dus 125 keer zo veel. Tijdens mijn verblijf zijn er dan ook ruim 40 boeken aan informatie op een dag verstuurd via mijn telefoon. Ik vind dat voor een apparaat waarmee je eigenlijk geacht wordt te telefoneren nogal veel, vooral als je bedenkt dat ik tijdens die drie dagen slechts één telefoongesprek heb gevoerd. Ik moet wel eerlijk bekennen dat ik ook 12 gesprekken geweigerd of helemaal niet gehoord heb.

[2] Het pdf-bestand van dit boek is overigens 28 MB groot.

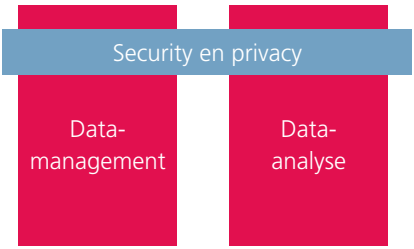
LECTORAAT DATA SCIENCE & ICT

Uit het eerdere voorbeeld van de wachttijdindicatie bij het verkeerslicht blijkt wel dat de data voor de fietser, de lichtjes die een voor een uitgaan, op zich niet zo heel veel zeggen. De betekenis wordt pas duidelijk als je weet dat de totale tijd tot het uitgaan van alle lichtjes aangeeft wanneer het verkeerslicht op groen springt en bovendien kunt inschatten hoelang het duurt voor alle lichtjes uit zijn. Het lectoraat Data Science & ICT richt zich dan ook op het verwerken van data tot informatie, oftewel het betekenis geven aan de data. Bij het verwerken van die data tot informatie en het aanbieden ervan, houden we rekening met de gebruiker. Het begrijpelijk presenteren van de informatie is een essentieel onderdeel van het verwerken van data tot informatie.

Het lectoraat Data Science & ICT is onderdeel van het Expertisecentrum Technische Innovatie van Avans Hogeschool. Gezamenlijk werken de lectoraten van het expertisecentrum aan het overkoepelende thema Resilient City; de duurzame, weerbare, leefbare stad. Onderwerpen die hierbij bekeken worden komen dan ook veelal vanuit de domeinen van de overige lectoraten, *gebouwde omgeving, nieuwe materialen en hun toepassingen, robotisering en sensoring en smart energy.*

Vanuit verschillende opleidingen werken bij elk van de lectoraten - en dus ook bij Data Science & ICT - docent-onderzoekers. Via deze docent-onderzoekers wordt de opgedane kennis uit het lectoraat teruggekoppeld naar het onderwijs. Ook nemen studenten van de betrokken opleidingen in groepsopdrachten, maar ook individueel via stages en afstudeeropdrachten, deel aan de onderzoeken.

De onderzoeken van het lectoraat zijn praktijkgericht van aard. Dat wil zeggen dat in tegenstelling tot theoretisch onderzoek, de resultaten van ons onderzoek (direct) toepasbaar moeten zijn in de beroepspraktijk. Ook moeten de vragen die we beantwoorden via het onderzoek relevant zijn voor de beroepspraktijk. Om die reden zijn er altijd partners uit het bedrijfsleven of de publieke sector betrokken bij de onderzoeken van het lectoraat. Veelal komen de onderzoeksvragen zelfs direct uit de beroepspraktijk en worden die in samenwerking met het lectoraat beantwoord en gevalideerd.



Figuur 3: schematische weergave onderzoekslijnen Data Science & ICT.

ONDERZOEKSLIJNEN

Het lectoraat Data Science & ICT voert onderzoek uit in drie verschillende onderzoekslijnen:

- Datamanagement
- Data-analyse
- Security en privacy

In figuur 3 is een schematisch overzicht gemaakt van deze lijnen. Hieruit blijkt direct dat de eerste twee onderzoekslijnen (in ieder geval min of meer) los van elkaar staan, terwijl de derde lijn over security en privacy dwars op de eerste twee lijnen staat. Met 'los van elkaar staan' bedoel ik dat de technieken die gebruikt worden om de projecten voor deze onderzoekslijnen uit te voeren onafhankelijk van elkaar zijn, terwijl de technieken voor de derde lijn, de security en privacy, verweven zijn met de technieken uit de twee andere onderzoekslijnen.

De onderzoekslijn *Datamanagement* richt zich op het efficiënt opslaan en transporteren van data. Het doel is te zorgen dat de data op zo'n manier wordt opgeslagen dat deze (later) efficiënt kan worden opgevraagd en verwerkt.
De onderzoekslijn *Data-analyse* richt zich op het verwerken van de data. Die verwerking moet zo efficiënt mogelijk gebeuren. Efficiënt is een lastig begrip. Bij de ene toepassing gaat dit vooral om het beperken van de benodigde opslagcapaciteit, terwijl het bij een andere toepassing vooral om snelheid gaat. De gezochte oplossing is dus altijd afhankelijk van de toepassing waar we aan werken. De onderzoekslijn *Security & privacy* houdt zich bezig met het beveiligen van de data, terwijl de bruikbaarheid daardoor niet in het geding komt.

Binnen het lectoraat Data Science & ICT richten we ons niet specifiek op bepaalde technieken, maar zoeken we bij een onderzoek naar de voor dat onderzoek geschikte methode of techniek. In de volgende hoofstukken laten we wel enkele technieken zien om een idee te geven van het soort onderzoek waar we ons binnen het lectoraat mee bezighouden en de uitdagingen die daarbij komen kijken.

DATAMANAGEMENT

Als data geproduceerd is, moet het in veel gevallen worden opgeslagen. Binnen de onderzoekslijn *datamanagement* houden we ons bezig met de efficiënte opslag van data, zodat deze op een later moment eenvoudig en snel kan worden opgevraagd en verwerkt. Een beproefde en veelgebruikte methode is om die data op te slaan in een database. In 1969 bedacht Edgar Codd het relationele model [1, 2]. Dit vormde de basis voor de databasesystemen die we nog steeds dagelijks gebruiken. De wijze waarop in zo'n database gegevens worden opgeslagen is gebaseerd op een tabel. Elke rij in deze tabel bevat een entiteit, terwijl elke kolom een eigenschap of attribuut van die entiteit weergeeft. Tabel 1 geeft een voorbeeld. In deze tabel zijn locaties van Avans weergegeven. Elke regel (entiteit) geeft informatie (de attributen) over een locatie. Een andere soort entiteit slaan we op in een aparte, eigen tabel. In ons voorbeeld zouden we voor Avans de academies op kunnen slaan, zie tabel 2. Ook hier vertegenwoordigt elke regel weer een entiteit, in dit geval dus een academie, en geeft elke kolom een eigenschap of attribuut weer van de entiteit. Een logische vervolgvraag zou kunnen zijn welke academies actief zijn op welke locatie. We moeten hiervoor een relatie leggen tussen de tabel met academies en de tabel met locaties. Van deze relatie komt ook de naam van het relationele model; er worden relaties tussen verschillende tabellen, of entiteiten, gelegd. Ook de relatie tussen twee tabellen geven we weer met een tabel. Veelal door gebruik te maken van het nummeren van de entiteiten in tabellen, in ons voorbeeld aangeduid als id. Als we die nummering gebruiken om relaties te leggen tussen academies en locaties, krijgen we Tabel 3. Deze tabel koppelt per regel één locatie (via locatie_id) aan één academie (via academie_id).

Id	Straat	Nummer	Stad	Postcode
1	Hogeschoollaan	1	Breda	4818CR
2	Lovensdijkstraat	61	Breda	4818AJ
3	Lovensdijkstraat	63	Breda	4818AJ
4	Onderwijsboulevard	215	Den Bosch	5223DE
5	Professor Cobbenhagenlaan	13	Tilburg	5037DA

Tabel 1: Tabel met locaties Avans

Behalve dit relationele model zijn er ook andere modellen om data in databases op te slaan. Elk van deze modellen maakt gebruik van specifieke eigenschappen van de entiteiten om de data zo efficiënt mogelijk op te slaan. Ongeacht het model, wordt wel steeds aan een aantal basisprincipes vastgehouden. Zo wordt de opgeslagen data geacht correct te zijn en zodra het is opgeslagen, is het de verantwoordelijkheid van het databasemanagementsysteem (DBMS) dat de data ook blijft bestaan en niet onbedoeld wordt gewijzigd. Bij het relationele model is er de eis dat elke cel van de tabel een enkelvoudig stuk informatie bevat. Dat wil zeggen dat er slechts één soort data wordt opgeslagen in een cel. Bijvoorbeeld een getal, of een stukje tekst, of een tijdstip. Samengestelde informatie is in een cel niet toegestaan. We kunnen dus, in tegenstelling tot ons voorbeeld over locaties van Avans, wel het hele adres opslaan in één cel, maar het huisnummer wordt dan als onderdeel van de tekst gezien en we kunnen de database niet specifiek vragen om het huisnummer. Een vraag als *geef alle locaties waarvan het huisnummer groter is dan 61* is dan dus niet mogelijk, of in ieder geval niet eenvoudig.

Id	Academie
1	Academie voor Bouw & Infra
2	Academie voor Communicatie en User Experience
3	Academie voor Engineering & ICT
4	Academie voor Industrie & Informatica
5	Academie voor de Technologie van Gezondheid en Milieu

Tabel 2: ETI Academies

id	Academie_id	Locatie_id
1	4	4
2	3	2
3	3	5

Tabel 3: koppeltabel

Voor toepassingen waarbij de computer beslissingen moet nemen en daarover ook informatie moet kunnen opslaan, is de traditionele database niet altijd even geschikt. Een beslissing opslaan in een database zou in dit geval gepresenteerd worden als een zekere beslissing. Onzekerheid over de

keuze kan dan niet (eenvoudig) worden uitgedrukt. Als we de database in ons voorbeeld over locaties vragen waar ik naartoe moet voor de Academie voor Industrie & Informatica, dan is het antwoord eenvoudig: Onderwijsboulevard 215 in Den Bosch. Dit komt omdat er voor deze academie maar een locatie bekend is in de database. Vragen we echter om de locatie van de Academie voor Engineering & ICT dan zijn er twee antwoorden: Lovensdijkstraat 61 in Breda en Professor Cobbenhagenlaan 13 in Tilburg. De betekenis van dit antwoord is ingewikkeld. Moeten we naar een van de twee locaties, of naar beide? De data doet geen uitspraak over dit antwoord. Hooguit zou je als gebruiker van het systeem kunnen afspreken hoe deze data te interpreteren. Het probleem bij deze oplossing is dan weer dat het systeem afhankelijk wordt van afspraken buiten datzelfde systeem om, terwijl onafhankelijk van waar en door wie het systeem gebruikt wordt, de werking en betekenis gelijk zouden moeten zijn.

FLEXIBELER DOOR ONZEKERHEID

Een van de problemen bij de huidige databases is dat informatie in zo'n database per definitie zeker moet zijn. In het eerdere voorbeeld over de locaties en academies van Avans zijn de gegevens in de tabellen de echte locaties. We hebben die informatie in de database gezet, omdat we zeker weten dat dat de locatiegegevens zijn. Dat is ook precies de reden waarom we bij een vraag over wat de locatie van een academie is een onduidelijk antwoord kunnen krijgen; namelijk als de academie verdeeld is over meer dan een locatie. De database combineert dan meerdere tabellen, die elk zekere gegevens bevatten, en door die combinatie ontstaat onzekerheid die niet is op te slaan in een (zekere) database. De database is dan niet gesloten. Dat wil zeggen dat het resultaat van een operatie op de database niet opgeslagen kan worden in deze zelfde database.

Door verschillende universiteiten is onderzocht hoe je onzekere data kunt opslaan in een relationele database en op welke manier deze database dan uitgebreid zou moeten worden met extra functionaliteit [3, 4, 11]. Om onzekere data in een database op te kunnen slaan, moeten we de database aanpassen zodanig dat deze om kan gaan met alternatieven [3]. In plaats van een zekere waarde voor een cel, moet het mogelijk worden om voor elke cel een aantal alternatieve waarden op te geven. Dat aantal kan 0 (de waarde is onbekend of bestaat niet), 1 (het resultaat is zeker, de situatie van de originele database), of meer dan 1 (de inhoud van de cel heeft alternatieve waarden) zijn. Uiteraard moet daarvoor de semantiek van de database worden aangepast. Alle vragen aan de database kunnen dan onzekere antwoorden

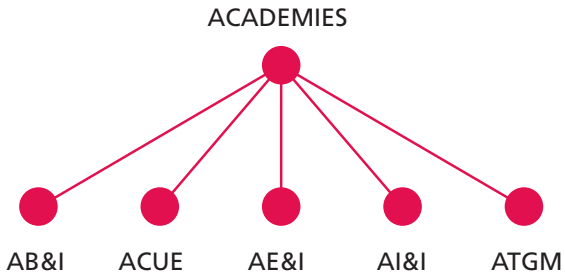
opleveren. Lastiger nog, als combinaties worden gemaakt tussen verschillende tabellen met onzekere data, dan moet het juiste antwoord, bestaande uit combinaties van onzekere waarden, worden opgeleverd. Een voorbeeld van zo'n onzekere database is gegeven in tabel 4. Hierin staat een tabel met het antwoord op de vraag naar welke locatie we moeten voor de academie AE&I. Het resultaat van deze vraag is Lovensdijkstraat 61 in Breda *of* Professor Cobbenhagenlaan 13 in Tilburg. Het feit dat het echte adres slechts één van de twee genoemde adressen kan zijn, en niet beide tegelijk, wordt weergegeven door het teken ||.

Academie	(Straat, Nummer, Stad, Postcode)
AE&I	(Lovensdijkstraat, 61, Breda, 4818AJ) || (Professor Cobbenhagenlaan, 13, Tilburg, 5037DA)

Tabel 4: tabel met onzekere data

ALLES IS EEN TABEL

Niet alle data is even geschikt om in tabelvorm weer te geven of op te slaan. Websites op het internet, in HTML-formaat, zijn bijvoorbeeld gebaseerd op een boomstructuur, vergelijkbaar met een stamboom waarbij de persoon in kwestie bovenaan staat en vervolgens naar beneden steeds de kinderen en kinderen van kinderen worden genoemd. Deze boomstructuren komen in de IT veel voor, maar passen als zodanig niet in het relationele model van Codd. Ook teksten of afbeeldingen zijn niet erg geschikt om zo te worden opgeslagen in een relationele database. Database-onderzoekers zijn er echter goed in om informatie die in eerste instantie misschien niet erg geschikt is voor het relationele model, toch om te zetten in een structuur die wel opgeslagen kan worden in een tabel. In veel gevallen wordt dat gedaan door een geschikte manier van nummering toe te passen op de data. Door elementen in de data te nummeren, kan deze nummering worden gebruikt voor de opslag in een tabel. In figuur 4 is een boomstructuur weergegeven met in figuur 5 de bijbehorende XML en de codering naar het relationele model volgens Grust et al. [9]. Het voordeel van het omzetten van data van een niet-tabel gebaseerde manier naar een tabel, is dat we gebruik kunnen maken van de eerdere ontwikkelingen in het database-vakgebied. De afgelopen jaren is er veel onderzoek gedaan naar het zo efficiënt en snel mogelijk maken van relationele databases. In plaats van nieuwe database-soorten ontwikkelen, is het efficiënter om de bestaande database-systemen te kunnen (her)gebruiken voor andere vormen van data. Na het omzetten van de data naar het relationele model zijn immers alle bewezen technieken weer te gebruiken.



Figuur 4: XML boom

Hoewel dat voor de toepassing van de techniek heel fijn is - we kunnen immers weer gebruikmaken van bewezen technieken en bovendien verschillende soorten data in dezelfde database opslaan - zitten er ook wat nadelen aan deze wijze van data omvormen. Allereerst is de taal waarmee vragen aan de database worden gesteld gebaseerd op het relationele model en daarmee niet altijd even geschikt om een vraag te stellen in een ander model. In het voorbeeld van de boomstructuur is een logische vraag welk punt in de boom heeft een 'ouder' met de waarde *Academies*. Een ouder in een boomstructuur is het aangrenzende punt één plaats boven het huidige punt. In dit specifieke geval voldoen, behalve het punt Academies zelf, alle punten aan deze voorwaarde. Deze vraag is niet direct eenvoudig te stellen in het geval van data in een tabel. Een logische vraag voor de tabel zou wel kunnen zijn: welke rij heeft in de meest rechtse kolom (met de naam *tag*) de waarde *Academies*. Beide modellen, relationeel en boomstructuur, hebben dus hun eigen logische manier van vragen stellen. Het omzetten van de een naar de ander is niet altijd even eenvoudig en al helemaal niet makkelijk te doorgronden voor een gebruiker van het systeem.

pre	post	level	tag
0	5	0	Academies
1	0	1	AB&I
2	1	1	ACUE
3	2	1	AE&I
4	3	1	AI&I
5	4	1	ATGM

```
<Academies>
        <AB&I></AB&I>
        <ACUE></ACUE>
        <AE&I></AE&I>
        <AI&I></AI&I>
        <ATGM></ATGM>
</Academies>
```

Figuur 5: XML-codering en bijbehorende XML-document.

DATA-ANALYSE

De enige reden om data op te slaan, is om het later weer te kunnen gebruiken. Dat is het domein van de tweede onderzoekslijn, data analyse; hoe halen we informatie uit deze opgeslagen data? Er zijn en worden allerlei technieken ontwikkeld om informatie uit data te halen. Veel van dat onderzoek vindt plaats aan universiteiten en levert uiteindelijk vaak een proof of concept op, meestal in de vorm van software. Zo'n proof of concept toont aan dat het ontwikkelde idee ook daadwerkelijk werkt. Vaak is zo'n proof of concept een vereenvoudiging van de werkelijkheid; de gebruikersinteractie is misschien niet zoals je van een doorontwikkeld product verwacht, het product kan wellicht nog slechts een beperkte hoeveelheid data aan, of er worden andere beperkende eisen gesteld bij het gebruik van het product. Kortom het product is nog niet gebruiksklaar. Hoewel er zeker uitzonderingen zijn, zijn deze onderzoeksprototypes niet bedoeld voor direct gebruik in het bedrijfsleven en vaak ook stopt de ontwikkeling zodra het basisprincipe is aangetoond. Werkt het prototype, dan gaan onderzoekers door naar een volgende onderzoeksvraag. In deze onderzoekslijn is het niet zozeer de bedoeling om nieuwe analysetechnieken te ontwikkelen, maar veel meer om bestaande technieken toepasbaar te maken in echte bedrijfssituaties.

DATA-INTEGRATIE

De data die we binnen de data science opslaan en analyseren, komt ergens vandaan. Data science is in dat opzicht altijd afhankelijk van andere domeinen. In het geval van het lectoraat is de samenwerking binnen het expertisecentrum daarom van essentieel belang. Soms komt die data van één bron binnen één domein, soms van meer bronnen binnen één domein, andere keren weer van meerdere bronnen van verschillende domeinen.
Om informatie te verzamelen, is het vaak nodig om data van verschillende bronnen te combineren. Dit is het terrein van de data-integratie. Denk bijvoorbeeld aan de fusie van twee bedrijven, waarbij de klantbestanden moeten worden gecombineerd tot één bestand van het nieuwe, gefuseerde bedrijf. In dat geval worden twee databronnen met dezelfde soort gegevens samengevoegd tot een databestand. Als beide bronnen er qua vorm exact hetzelfde uitzien, is het samenvoegen misschien nog niet zo ingewikkeld. Hooguit zullen er duplicaten ontstaan bij het samenvoegen van deze bronnen, die dan vervolgens moeten worden gefilterd, zodat er van iedere klant nog

slechts een instantie aanwezig is. Als de twee bronnen in vorm van elkaar afwijken wordt het integreren ingewikkelder.

De vorm van de databestanden moet in dit geval eerst op elkaar worden afgestemd. Laten we als voorbeeld weer de locaties van Avans nemen. Stel dat die locaties in twee verschillende databronnen staan opgeslagen, bijvoorbeeld een bron voor Breda en een andere bron voor Den Bosch en Tilburg, zoals tabel 5 en 6. Dan is in dit geval eenvoudig te zien dat de kolommen straat en nummer (Den Bosch en Tilburg) samen dezelfde informatie bevatten als de kolom adres (Breda). Het is dan nog slechts een kwestie van kiezen welke vorm de gecombineerde bron moet hebben en de data van beide bronnen omzetten naar deze vorm. Eventuele duplicaten moeten ook in dit geval natuurlijk worden verwijderd.

Id	Straat	Stad	Postcode
1	Hogeschoollaan 1	Breda	4818CR
2	Lovensdijkstraat 61	Breda	4818AJ
3	Lovensdijkstraat 63	Breda	4818AJ

Tabel 5: locaties in Breda.

Id	Straat	Nummer	Stad	Postcode
1	Onderwijsboulevard	125	Den Bosch	5223DE
2	Professor Cobbenhagenlaan	13	Tilburg	5037DA

Tabel 6: locaties in Den Bosch en Tilburg.

Uit dit voorbeeld blijkt dat kennis nodig is van het domein waarvan de data komt. In dit geval moeten straat en nummer gecombineerd worden, maar elk domein heeft zo zijn eigen logische manieren om data op te slaan. Lastiger nog, vaak zijn er binnen een domein meerdere manieren, die dan dus naar elkaar geconverteerd moeten worden. Voor een mens is die conversie van bekende domeinen goed te doen, maar als deze conversie geautomatiseerd wordt, ligt de beslissing bij de computer. Een uitleg hoe de keuze door de computer tot stand is gekomen, is daarbij belangrijk. Zo kun je bij een verkeerde keuze achteraf nog corrigeren.

PROFIELEN

Een van de mogelijkheden die je hebt met een grote hoeveelheid data, is het groeperen van de opgeslagen entiteiten. Als de opgeslagen entiteiten persoonsgegevens zijn, is het groeperen een manier om profielen te maken. Bedrijven gebruiken deze methode om klanten bijvoorbeeld suggesties te doen voor een volgende aankoop of - in het geval van verzekeringsmaatschappijen - om risico's in te schatten. Bij personen van wie de gegevens overeenkomen, worden dezelfde suggesties gedaan, of wordt het risico gelijk ingeschat. Andersom kunnen acties van een persoon uit zo'n groep gebruikt worden om inschattingen te maken van de overige personen uit dezelfde groep.

De manier om op elkaar gelijkende beschrijvingen te vinden, heet clustering [10]. Hierbij worden alle eigenschappen van de objecten vergeleken. Hoe meer eigenschappen op elkaar lijken, hoe dichter de objecten bij elkaar liggen. Objecten die dicht bij elkaar liggen behoren tot hetzelfde cluster, of in het geval van personen hetzelfde profiel. Natuurlijk komen niet alle eigenschappen overeen. Dat betekent dat sommige eigenschappen van twee objecten op elkaar kunnen lijken, terwijl andere eigenschappen juist verschillen. Beslissen of die twee objecten in zo'n geval tot hetzelfde cluster behoren is niet altijd eenvoudig. Vandaar ook dat in het geval van suggesties op websites op basis van een profiel, het systeem er soms behoorlijk naast kan zitten en je een product als suggestie krijgt, waar je geen interesse in hebt.

PREDICTION

Een ander gebruik van de data is het voorspellen van de toekomst. Uiteraard gaat het hier niet over waarzeggerij, maar over toekomstige waarden van de data, gebaseerd op de verkregen waarden tot nu toe. Simpele vormen van deze voorspellingen zijn bijvoorbeeld de prognose die de energieleverancier stuurt over het energieverbruik voor het jaar, gebaseerd op deze en vorige maanden. Een ander voorbeeld is het voorlopig studieadvies dat studenten krijgen halverwege het eerste jaar, waarin wordt aangegeven of ze op schema liggen om door te mogen met de studie. Uiteraard zijn deze voorbeelden redelijk eenvoudig en kunnen ze veelal met pen en papier door de persoon in kwestie zelf worden nagerekend, maar als er veel en gecompliceerde data aan de berekening ten grondslag ligt, wordt het uitrekenen al snel een stuk ingewikkelder. Net als in het geval van integratie, is het toelichten hoe tot de berekening is gekomen dan van essentieel belang. Niet alleen om het resultaat

voor de gebruiker inzichtelijker te maken, maar ook - zeker bij ingewikkelder en ingrijpender onderwerpen - om de acceptatie bij de gebruiker te vergroten.

STREAMING

Tot nu toe hebben we steeds aangenomen dat data eerst wordt opgeslagen en op een later tijdstip wordt opgevraagd en gebruikt. Dat maakt dat de berekeningen die gedaan worden meestal niet heel tijd-kritisch zijn. Natuurlijk is het prettig als dit zo snel mogelijk gebeurt, maar een seconde meer of minder zal de uitkomst of bruikbaarheid van de resultaten niet beïnvloeden. Dat is anders in situaties waarin gegevens direct gebruikt worden om te bepalen wat er moet gebeuren. Denk bijvoorbeeld aan de routeplanner in de auto. Als er een file ontstaat, moet de routeplanner direct berekenen of de huidige route nog steeds de snelste is, of dat een andere route vanwege de file efficiënter is. In dit geval wordt de file-informatie niet eerst opgeslagen, maar direct gebruikt. Ons wegennet is in termen van data redelijk overzichtelijk, dus in dit voorbeeld is het direct berekenen van een nieuwe route niet zo ingewikkeld, maar in situaties die onvoorspelbaarder zijn en waar meer data bij nodig is, wordt dat anders.

In het kader van de resilient city en het verduurzamen van de energie wordt energie meer en meer lokaal geproduceerd. Het in kaart brengen van de energiebehoefte en de productie daarop afstemmen is dan belangrijk. Uiteraard is die afstemming direct nodig en mag de berekening daarvan niet voor al te veel vertraging zorgen, omdat dit de beschikbaarheid van energie in gevaar zou brengen. Dit directe gebruik van data, zonder tussenkomst van een database, wordt streaming data gebruikt en vergt andere eigenschappen van de applicatie.

FEEDBACK EN UITLEG

Met de grote hoeveelheden data die tegenwoordig beschikbaar zijn, is het noodzakelijk dat veel van de verwerking automatisch wordt afgehandeld. Met de eerder genoemde technieken is veel mogelijk. Helaas kan het gebeuren dat het systeem tot een verkeerde beslissing komt. Op zich is daar niets mis mee. Ook als mensen data verwerken, worden fouten gemaakt. Hopelijk worden die dan (op tijd) gesignaleerd en gecorrigeerd. Het probleem bij automatische verwerking van data is vaak dat personen de informatie die op die manier gegenereerd wordt, zonder meer voor waar aannemen.
Een mooi voorbeeld is een sketch in de comedyserie Little Britain met de inmiddels bekende uitdrukking *computer says no*, of de reclame van een verzekeringsmaatschappij met een paarse krokodil. Allebei voorbeelden

waarbij de data in de computer boven de werkelijkheid wordt geplaatst. De door het systeem aangeleverde informatie wordt voor waar aangenomen en er wordt door de persoon die het systeem gebruikt, niet verder gekeken.

De systemen die ontwikkeld worden voor het automatisch verwerken van data, zouden twee onderdelen moeten faciliteren. Allereerst een feedbackmechanisme waarmee de gebruiker van het systeem aan kan geven dat de werkelijkheid anders is dan aangegeven door het systeem. Op deze manier kan de data door de gebruiker worden aangepast, om het systeem weer in overeenstemming te brengen met de werkelijkheid [7]. Door gebruik te maken van dit feedbackmechanismen zal de kwaliteit van de databron, naarmate er meer feedback wordt gegeven, steeds beter worden [8]. Uiteraard gaan we er hier voor het gemak wel vanuit dat de feedback steeds correct is.

Ten tweede zou het systeem een mechanisme moeten hebben om aan te geven of - beter nog - uit te leggen, hoe en waarom een bepaald resultaat tot stand is gekomen. Een veelbelovende techniek om dit voor elkaar te krijgen is lineage [3, 4]. Lineage van de data geeft aan waar deze data vandaan komt. Door dit consequent bij te houden, is achteraf te herleiden hoe een bepaald resultaat is verkregen. Het nadeel van lineage is dat het erg veel schijfruimte vergt, waardoor het in veel gevallen niet toepasbaar is. Door de grootte van de data is de beschikbare ruimte snel op. Bovendien vraagt het verwerken van al die data relatief veel tijd, waardoor het in sommige applicaties te veel vertraagt.

NOGMAALS ONZEKERHEID

De verwerking van data levert veelal een aantal mogelijke uitkomsten op. Bij profielen voldoet een persoon nooit precies aan de kenmerken van een profiel, maar is het meestal een combinatie van enkele profielen. Met enig geluk is er wel sprake van een overduidelijk zwaartepunt bij een profiel. Meestal wordt dat profiel dan aangehouden.

Bij het voorspellen van toekomstige waarden wordt een verwachting uitgesproken. De realiteit kan anders uitpakken. De mogelijkheid met de grootste kans wordt gepresenteerd als het resultaat. Als we gebruikmaken van de onzekere dataopslag kunnen we die onzekerheid juist gebruiken en bijhouden. Het voordeel van het opslaan van die onzekerheid, is dat als er nieuwe feiten bijkomen alle data nog beschikbaar is en er opnieuw kan worden gekeken wat de meest waarschijnlijke uitkomst is. Zouden we steeds een harde (ja/nee) beslissing nemen zonder onzekerheid, dan kan die nieuwe,

meest waarschijnlijke uitkomst niet correct worden uitgerekend, ook als die nieuwe kennis vooraf geleid zou hebben tot een andere uitkomst.

VISUALISEREN

De eerder beschreven manieren van data opslaan, verwerken en beveiligen zijn belangrijk en het is evident dat dit op een voor de applicatie en gebruiker acceptabele snelheid gebeurt. In sommige gevallen zijn ook andere aspecten van belang, zoals de benodigde opslagruimte. Maar ook al heb je dit alles goed voor elkaar, zonder duidelijke communicatie met de gebruiker kom je nergens. Als de gebruiker, meestal een persoon, de informatie niet snapt, wordt het gewenste doel niet bereikt.

Soms is de eerder genoemde tabelstructuur een goede oplossing om informatie te presenteren. Vooral als het om eenvoudige entiteiten gaat met weinig attributen (en dus kolommen), waarbij bovendien slechts één tabel nodig is. Denk bijvoorbeeld weer aan het voorbeeld met de Avans-locaties. Een tabel met daarin de straat, stad en postcode van een locatie wordt door een gemiddelde gebruiker waarschijnlijk wel begrepen.

Dat wordt een ander verhaal als de informatie complexer is, er relaties tussen entiteiten zijn, of er zelfs helemaal geen entiteiten worden voorgesteld, maar abstracte elementen. Een voorbeeld van dit soort abstracte elementen kunnen sensorwaarden zijn, zoals temperatuurmetingen of verbruikswaarden bij een slimme energiemeter. In die gevallen moet een manier gevonden worden om de informatie zo te presenteren dat deze begrepen wordt door *elke gebruiker*. Het begrip elke gebruiker wil overigens niet zeggen dat het begrepen hoeft te worden door elke persoon. Wel is het noodzakelijk dat alle personen die de applicatie ook daadwerkelijk gebruiken, hiermee overweg kunnen.

SECURITY EN PRIVACY

Meer en meer systemen zijn tegenwoordig aangesloten op het internet. Ook systemen waarvan je je kunt afvragen of dat wel zo noodzakelijk is. Ik kocht een tijdje geleden een nieuwe afwasmachine en tot mijn verbazing kon die worden aangesloten op het internet via wifi. Er werd door de leverancier ook een app geleverd, zodat je ook als je niet in de buurt van de afwasmachine bent de machine kunt aanzetten, of kunt zien of het programma al is afgelopen. Jammer genoeg ondersteunt deze app nog niet het op afstand inladen van de afwasmachine, zodat ik dat noodgedwongen nog steeds zelf moet doen. Daardoor moet je overigens ook echt aanwezig zijn, dus het echte nut van de online-activiteiten van een afwasmachine ontgaan me nog een beetje. Nog interessanter werd het toen ik laatst ook een nieuwe oven kocht, en wat bleek, ook deze heeft de mogelijkheid om aangesloten te worden op het internet. Het is, zeker als je net een apparaat hebt gekocht, natuurlijk grappig om via de app bij te houden wat de voortgang van het apparaat is, maar na verloop van tijd (en dat was in mijn geval vrijwel direct) wordt zo'n functionaliteit toch minder interessant. De verbinding met het internet echter heb ik niet verbroken en dat betekent dat er een onbedoelde mogelijkheid is gecreëerd om digitaal bij mij thuis binnen te komen. Ik vermoed dat mijn huis niet erg interessant is om digitaal in te breken, maar het kan wel. Laten we eerlijk zijn, niets ten nadele van de betreffende fabrikanten, maar ik kan me zo voorstellen dat in dit geval digitale beveiliging van de data en het apparaat niet de allerhoogste prioriteit heeft.

Veel data is privacygevoelig en zelfs als dit niet zo is, is het vaak toch niet de bedoeling dat iedereen de data zo maar kan inzien. Bij een afwasmachine of oven denk je misschien niet direct aan privacygevoelige data, toch zeggen deze apparaten iets over, bijvoorbeeld, de aanwezigheid van bewoners. Als de oven en afwasmachine langere tijd niet gebruikt worden, is de kans groot dat er niemand thuis is, Als we dan ook nog de informatie uit een slimme meter kunnen uitlezen, die immers ook gegevens via het internet verstuurt, dan is de aan- en afwezigheid van bewoners helemaal goed vast te stellen.

De privacygevoeligheid van data is uiteraard veel duidelijker bij patiëntgegevens. Niemand is verbaasd over het feit dat deze gegevens goed beveiligd moeten worden en alleen mogen worden ingezien door bevoegde personen. De vraag is dan echter, wie is bevoegd? De behandelend arts, is

een logisch antwoord op deze vraag. Maar wat als er een noodgeval is en de behandelend arts is niet aanwezig? Dan is het voor de patiënt in kwestie niet alleen prettig, maar soms zelfs van levensbelang als een andere arts ook toegang heeft tot bepaalde informatie. Om de situatie nog lastiger te maken, zullen sommige gegevens alleen door de arts en weer andere gegevens ook door administratief medewerkers mogen worden bekeken. En bij sommige ziekenhuizen zullen delen van de data ook voor onderzoek mogen worden gebruikt. Een manier om te zorgen dat onbevoegden niet bij data kunnen waar ze geen toegang toe zouden moeten hebben, is door deze te versleutelen. Technisch zijn er verschillende mogelijkheden om deze versleuteling (encryptie) te realiseren, maar zelfs het niveau waarop versleuteld wordt, is van belang. Denk daarbij bijvoorbeeld aan versleuteling per rij, met voor elke rij afzonderlijk een sleutel, zodat je per rij kunt aangeven of iemand rechten heeft om de informatie te lezen. In het voorbeeld met de patiënten betekent dit, dat op patiëntniveau toegang gegeven kan worden tot de data. In een tabel zou immers elke rij overeenkomen met één patiënt. Maar als iemand dan data van een patiënt mag bekijken, heeft hij ook meteen toegang tot álle informatie van deze patiënt.

Een voor de hand liggende oplossing om te zorgen dat gebruikers alleen toegang hebben tot de voor hen toegestane informatie per patiënt, is om alle data op het niveau van cellen in de tabel uniek te versleutelen. Dat wil zeggen dat we per cel kunnen aangeven of iemand of wel geen toegang tot de data heeft. Dit niveau van versleutelen geeft de grootst mogelijke flexibiliteit. Elmer Lastdrager heeft in zijn afstudeerwerk [5] een prototype van zo'n systeem gemaakt. Het nadeel van dit systeem is dat het relatief traag is, omdat alle cellen apart van elkaar moeten worden ontsleuteld (decryptie). Dit is niet alleen nodig als de data gelezen moet worden als resultaat van een vraag aan de database, maar ook als de data nodig is om te bepalen of een rij uit de tabel als resultaat gegeven moet worden. Denk daarbij aan de vraag: *Geef alle namen van patiënten die ouder zijn dan 12 jaar*. In dit geval moeten zowel de naam als de leeftijd van de patiënt ontsleuteld zijn. De naam omdat die weergegeven moet worden in het antwoord. De leeftijd omdat die nodig is om te bepalen of een patiënt wel of niet tot de gewenste groep behoort. In tabel 7 staat een fictief resultaat als antwoord op deze vraag. We gaan er in dit voorbeeld gemakshalve vanuit dat het versleutelen van de data heeft plaatsgevonden op rij-niveau. Dat wil zeggen dat iemand bevoegd is om alle gegevens van een patiënt te zien, of helemaal niets van die patiënt mag zien. Blijkbaar heeft de huidige gebruiker van het systeem geen rechten om

de patiënten waarvan de id-velden vermoedelijk 3 en 5 zijn, te zien. In dit simpele voorbeeld kan de waarde van het id-veld van de twee niet getoonde personen nog wel redelijk worden voorspeld, hoewel er geen garantie is dat dit ook klopt. De overige waarden zijn uiteraard volledig onvoorspelbaar. Dat we in het resultaat alleen personen zien die ouder zijn dan 12 jaar is logisch, want dat was immers de vraag aan het systeem. De getoonde personen hoeven echter niet de enigen te zijn die ouder zijn dan 12 jaar. Alle personen die ouder zijn dan 12 jaar, maar waarvoor de huidige gebruiker van het systeem geen rechten heeft om die te lezen, zullen niet worden getoond. Niet alleen omdat de gegevens niet mogen worden getoond, maar ook omdat het leeftijd-veld dat nodig is om te controleren of iemand ouder dan 12 jaar is, niet toegankelijk is voor deze persoon. Afhankelijk van de achterliggende vraag is het gegeven antwoord dus mogelijk zelfs incorrect.

id	naam	adres	leeftijd
1	Jan	Dorpsweg 1	13
2	Marie	Hoofdstraat 4	14
4	Klaas	Stadhuislaan 23	15

Tabel 7: fictief resultaat van vraag naar patiënten ouder dan 12.

Naast de problemen rondom de correctheid van antwoorden, is ook het bijhouden alle sleutels en het registreren van (veranderingen in) het recht tot toegang, een heel administratief proces waar veel in fout kan gaan. Met een goede data-encryptie ben je er nog niet. Zeker zo belangrijk is het dat het systeem werkbaar is en blijft voor de gebruiker. Dat het gebruiksvriendelijk is. Als het systeem, bijvoorbeeld, zo ingewikkeld wordt dat gebruikers het wachtwoord met een sticker op het toetsenbord plakken, dan heeft de beveiliging zelf uiteraard niet zo veel zin meer.

TOT SLOT

De komende jaren wil ik als lector gaan werken aan nieuwe, uitdagende praktijkgerichte data science-oplossingen. Enkele van de technieken en methoden heb ik hiervoor geschetst, maar gezien de snelheid waarmee de ontwikkelingen plaatsvinden, zou het me niets verbazen als veel van wat we nodig hebben nog helemaal niet bekend is.

Ik kom aan het eind van mijn verhaal. Ik zeg 'mijn' verhaal, maar eigenlijk is dit ook het verhaal van het lectoraat. Het lectoraat Data Science & ICT bestaat, naast mijzelf, uit zeven docent-onderzoekers. Ik heb enorm veel zin om samen met hen, docenten en studenten van de verschillende opleidingen en het bedrijfsleven, de projecten uit te voeren en de nieuwe ontwikkelingen niet alleen samen te ontdekken, maar hopelijk ook deels vorm te geven.

REFERENTIES

1. Codd, E.F (1969), *Derivability, Redundancy, and Consistency of Relations Stored in Large Data Banks*, Research Report, IBM.
2. Codd, E.F (1970), A Relational Model of Data for Large Shared Data Banks, https://web.archive.org/web/20070612235326/http:/www.acm.org/classics/nov95/toc.html, Classics. **13** (6): 377–87, visited 8 december 2019
3. M. Mutsuzaki, M. Theobald, A. de Keijzer, J. Widom, P. Agrawal, O. Benjelloun, A. Das Sarma, R. Murthy, and T. Sugihara. Trio-One: Layering Uncertainty and Lineage on a Conventional DBMS. Proc. Third Biennial Conference on Innovative Data Systems Research (CIDR '07), Pacific Grove, California, January 2007. Demonstration description.
4. Das Sarma, O. Benjelloun, A. Halevy, and J. Widom. Working Models for Uncertain Data. Proc. 22nd Intl. Conference on Data Engineering, Atlanta, Georgia, April 2006. Note: much of the material in this conference paper also appears in the journal paper *Representing Uncertain Data: Models, Properties, and Algorithms,*
5. Lastdrager, Elmer (2011) Securing Patient Information in Medical Databases, MSc thesis, University of Twente,
6. Jeffrey Ullman, Hector Garcia-Molina, Jennifer Widom (2001), Database Systems: The Complete Book, Prentice Hall, ISBN:0130319953
7. de Keijzer, A., & van Keulen, M. (2007). User Feedback in Probabilistic Integration. In Second International Workshop on Flexible Database and Information System Technology (FlexDBIST 2007) (pp. 377-381). Los Alamitos: IEEE Computer Society. https://doi.org/10.1109/DEXA.2007.146
8. van Keulen, M., & de Keijzer, A. (2009). Qualitative Effects of Knowledge Rules and User Feedback in Probabilistic Data Integration. VLDB journal, 18(5), 1191-1217. [10.1007/s00778-009-0156-z]. https://doi.org/10.1007/s00778-009-0156-z
9. Grust, T., van Keulen, M., Teubner, J., (2003). Staircase Join: Teach a Relational DBMS to Watch its (Axis) Steps. VLDB Conference, 524-535.
10. Knowledge Discovery in Databases – Part III – Clustering, Heidelberg University, 2017, https://dbs.ifi.uni-heidelberg.de/files/Team/eschubert/lectures/KDDClusterAnalysis17-screen.pdf, visited 8 december 2019
11. Dan Suciu, Nilesh N. Dalvi, Foundations of probabilistic answers to queries. SIGMOD Conference 2005: 963.

INTRODUCTIE LECTORAATSLEDEN

ERCO ARGANTE

Erco Argante is docent Informatica, curriculumcoördinator en voorzitter van het team aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Na zijn studie natuurkunde aan de Radboud Universiteit heeft Erco een tweejarige AIO-opleiding Technische Informatica gedaan, gevolgd door een promotieonderzoek naar parallelle verwerking van collision event data, gegenereerd door de Large Hadron Collider van het CERN in Zwitserland. Hierna heeft hij in verschillende ICT-gerelateerde rollen (team leader, technical coordinator, software architect, system manager) gewerkt bij Ericsson Telecommunicatie. Naast het werk bij Avans doet Erco bij tijd en wijle ICT-consultancy.

Erco's expertise ligt op het gebied van software architecture, solution architecture en enterprise application integration. Daarnaast heeft hij veel ervaring met software engineering en is hij geïnteresseerd in software security en computernetwerken. Recentelijk is Erco veel bezig met machine learning. Hobbymatig heeft hij interesse in aerodynamica.
Erco is betrokken bij het project Race management voor de zonneboot van het Avans Solar Mariteam d.m.v. machine learning.

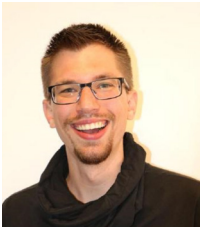
REINIER DICKHOUT

Reinier Dickhout is docent Informatica aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Reinier volgde de Hogere Informatica Opleiding aan Hogeschool West-Brabant. Daarna werkte hij als een van de laatste dienstplichtigen in Nederland een jaar lang als officier bij de Koninklijke Marine op het Centrum voor Automatisering van Wapen- en Commandosystemen.
Na zijn diensttijd ging Reinier aan de slag als Oracle-programmeur en -ontwikkelaar. Hij begon bij IBAS, daarna werkte hij voor detacheerder Yacht. In 2000 koos hij voor een functie met een vaste standplaats in Etten-Leur, waar hij onder meer werkte als IT-manager bij een wereldwijd opererend olie- en gasbedrijf.

In 2013 maakte hij de overstap naar het onderwijs en ging hij bij Avans Hogeschool als docent aan de slag. Hier hield hij zich onder meer bezig met het vraagstuk hoe de opleiding datagerelateerde onderwerpen kan inpassen in het curriculum.

Reinier studeert momenteel af voor de master Business Processes and IT, met de specialisatie Data Science. Zijn expertise ligt op het gebied van databases, bedrijfsprocessen en het verbinden van mensen uit verschillende disciplines.



JEROEN DE HAAS

Jeroen de Haas is docent Informatica aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool. Zijn expertise ligt op het snijvlak van machine learning en statistiek.

Jeroen behaalde zijn mastertitel in statistiek aan de KU Leuven en was daar vervolgens wetenschappelijk onderzoeker. Hij maakte de overstap naar het hbo, waar hij achter de schermen werkte aan de automatische analyse van enquêteresultaten gericht op de onderwijskwaliteit.

Sinds 2016 staat Jeroen voor de klas bij Avans Hogeschool. Hij verzorgt vakken over kunstmatige intelligentie en data science. Daarnaast is hij betrokken bij onderwijsvernieuwing en -ontwikkeling en de totstandkoming van diverse initiatieven op het gebied van data en data science.

Belangrijk voor Jeroen is het leggen van de verbinding met het bedrijfsleven. In samenwerking met het platform Driven by Data verzorgt hij masterclasses voor het mkb en werkt hij samen met studenten aan oplossingen voor vraagstukken rond data.

MARTIN LECLERCQ

Martin Leclercq is docent Visual Design & Research aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Al langere tijd werkt Martin als grafisch ontwerper voor diverse opdrachtgevers. Hij denkt na over de identiteit van ontwikkeltrajecten, verzorgt huisstijlbewaking en realiseert multimedia toepassingen.

De laatste 15 jaar is Martin ook parttime docent bij de opleiding Communication & Multimedia Design (CMD) van Avans Hogeschool. Hij geeft vakken op het gebied van grafisch ontwerpen, webdesign, worldbuilding, data design en beeldretorica. Daarnaast werkt hij mee aan het ontwikkelen en coördineren van meerdere opleidingsmodules. Als toetscommissie- en AR-lid is hij actief betrokken bij de opleiding.

De expertise van Martin ligt bij grafisch ontwerpen, waarbij de focus ligt op het vertalen en het toegankelijk maken van informatie voor multimedia toepassingen. Met zijn brede interesse en een luisterend oor weet hij studenten en opdrachtgevers met diverse achtergronden aan zich te binden.

Het deelnemen aan de kenniskring bij het lectoraat Data Science & ICT geeft Martin ruimte voor onderzoek en kennisverdieping. Ook vindt hij het belangrijk voor kennisuitwisseling met studenten en voor hun participatie bij het derdejaars lab.

ANDER DE KEIJZER

Ander de Keijzer is sinds februari 2019 lector van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie (ETI) van Avans Hogeschool.

Ander de Keijzer studeerde Technische Informatica aan de Universiteit Twente. Al tijdens zijn studie had hij interesse voor zowel onderzoek als onderwijs. Na enkele studentassistentschappen startte hij daarom naast zijn opleiding met de eerstegraads lerarenopleiding.

Na zijn studie begon hij aan een promotie bij de database-groep (tegenwoordig Data Science) aan dezelfde universiteit. Ook gaf hij les op het gebied van informatica aan diverse opleidingen, zoals Werktuigbouwkunde, Technische Geneeskunde en Biomedische Technologie.
Na zijn promotie in 2008 werkte hij achtereenvolgens als docent aan de Universiteit Twente en Hogeschool Windesheim en vervolgens als hoofddocent en waarnemend lector aan Hogeschool Utrecht. Steeds combineerde hij daar onderwijs en onderzoek en altijd in een multidisciplinaire omgeving. Het grootste deel van de projecten waaraan hij werkte waren in de zorgsector.
Binnen het lectoraat onderzoekt Ander hoe organisaties beter gebruik kunnen maken van beschikbare data en hoe ze toegankelijker gemaakt kunnen worden voor de eindgebruiker.

MARTIJN SCHUURMANS

Martijn Schuurmans is docent Informatica aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Hij doceert onder meer over software-architectuur en databases. Grote expertise heeft hij op het gebied van databases waarbij performance en juistheid centraal staan. Machine learning is een ander onderwerp waar Martijn veel kennis over heeft.

Voordat hij als docent startte bij Avans Hogeschool, werkte Martijn bij een IT-dienstleverancier als softwareontwikkelaar voor grootschalige applicaties voor zorgbedrijven. Hij werkte ook als business intelligence engineer. In die functie ontwikkelde hij data warehouses en voerde analyses uit voor onder meer grote verzekeraars, bouw- en overheidsbedrijven. Op dit moment volgt Martijn een master op het gebied van data science.

MAURICE SNOEREN

Maurice Snoeren is docent Technische Informatica aan Avans Hogeschool en kenniskringlid van de lectoraten Smart Energy en Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Sinds mei 2018 werkt Maurice als docent binnen de opleiding Technische Informatica. Tegelijkertijd startte hij als onderzoeker bij het Expertisecentrum Technische Innovatie. Zijn expertise ligt op de terreinen van software, hardware, cybersecurity, blockchain en regelsystemen.

Maurice heeft een elektrotechnische achtergrond en is afgestudeerd aan de Technische Universiteit Eindhoven in 2004. Inmiddels heeft hij meer dan 15 jaar ervaring in het ontwerpen en realiseren van software en elektrotechnische en embedded systemen. Als afdelingsmanager en senior consultant binnen de energiesector was hij verantwoordelijk voor de operatie en cybersecurity van procesautomatiseringssystemen. Hij voerde penetratietesten, vulnerability assessments en audits uit op kritische (procesautomatiserings)systemen voor het verbeteren van de cybersecurity.

Binnen het lectoraat doet Maurice onderzoek op het gebied van nieuwe technologieën, zoals blockchain, artificial intelligence en beveiliging van IoT (Internet of Things) systemen. Daarnaast is hij betrokken bij de volgende projecten:

- Energie Learning Community (ELC)
- Wordt een Smart Grid écht slim met Artificial Intelligence?
- Smart Sensor (digitalisering van gebouwen)
- Energiemeter

JAAP VAN VELDHOVEN

Jaap van Veldhoven is docent Civiele Techniek aan Avans Hogeschool en kenniskringlid van het lectoraat Data Science & ICT bij het Expertisecentrum Technische Innovatie van Avans Hogeschool. Zijn expertise ligt op het gebied van civieltechnische constructies.

Na het afronden van zijn hbo-opleiding Civiele Techniek deed Jaap een half jaar onderzoek aan de universiteit van Johannesburg (Zuid-Afrika) naar labour intensive based civil engineering. Daarna werkte hij eerst als tekenaar en vervolgens als constructeur in de weg- en waterbouw. Naast zijn werk studeerde Jaap nog 6 jaar verder met de opleidingen Pabo, Betonconstructeur HBO+, Geotechniek CGF1 en de masteropleiding Betonconstructeur MSEng. Na 3 jaar werken bij Heijmans aan een aantal grotere water- en infraprojecten, ging Jaap aan de slag als docent Civiele Techniek bij Avans.

De expertise van Jaap ligt bij het ontwerpen van civieltechnische constructies. Vooral op het gebied van geprefabriceerd beton en voorgespannen brugdekken heeft hij veel ervaring. Jaap is altijd op zoek naar verbeteringen, bijvoorbeeld op het gebied van ICT.
Bij de Academie voor Bouw en Infra verzorgt Jaap lessen en trainingen, voltijd en duale coaching, en project-, afstudeer- en stagebegeleiding.

MONIQUE WINTERMANS

Monique Wintermans is Senior Management Assistente bij de lectoraten Data Science & ICT en Smart Energy bij het Expertisecentrum Technische Innovatie van Avans Hogeschool.

Na haar opleiding tot directiesecretaresse heeft Monique de opleiding Travel Trade Management gevolgd aan de Nationale Hogeschool voor Toerisme en Verkeer in Breda. Destijds kon zij niet direct een passende baan in de toeristische sector vinden, waarna zij emplooi vond in de makelaardij. Ruim 10 jaar is zij werkzaam geweest als commercieel medewerkster binnen- en buitendienst bij een kleinschalig makelaarskantoor waar zij verantwoordelijk was voor de gehele commerciële en administratieve ondersteuning van de makelaars. In die tijd heeft zij ook de vastgoed praktijk opleiding afgerond. Hierna had zij een financiële functie bij het familiebedrijf van haar partner, maar na 4 jaar samenwerking wilde zij werk en privé toch graag weer wat meer gescheiden houden. Na een aantal functies bij kleine organisaties ging Monique in 2018 terug naar haar roots als managementassistente, ditmaal bij Avans Hogeschool.

Monique's expertise ligt op het gebied van plannen, regelen en organiseren. Het werken voor een lectoraat zorgt voor een extra uitdaging.

AANTEKENINGEN

COLOFON

Dit is een uitgave van Avans Hogeschool.
Uitgegeven ter gelegenheid van de lectorale rede van Ander de Keijzer

Coördinatie
Diensteenheid Marketing, Communicatie en Studentenzaken,
Avans Hogeschool

Eindredactie
Tekstbureau Geert Braam

Druk & vormgeving
De Studio, powered by Avans & Canon

Contact
Expertisecentrum Technische Innovatie
Info.ETI@avans.nl
088 – 525 9276l

ISBN/EAN:978-90-74611-71-8

avans hogeschool

AVA190391