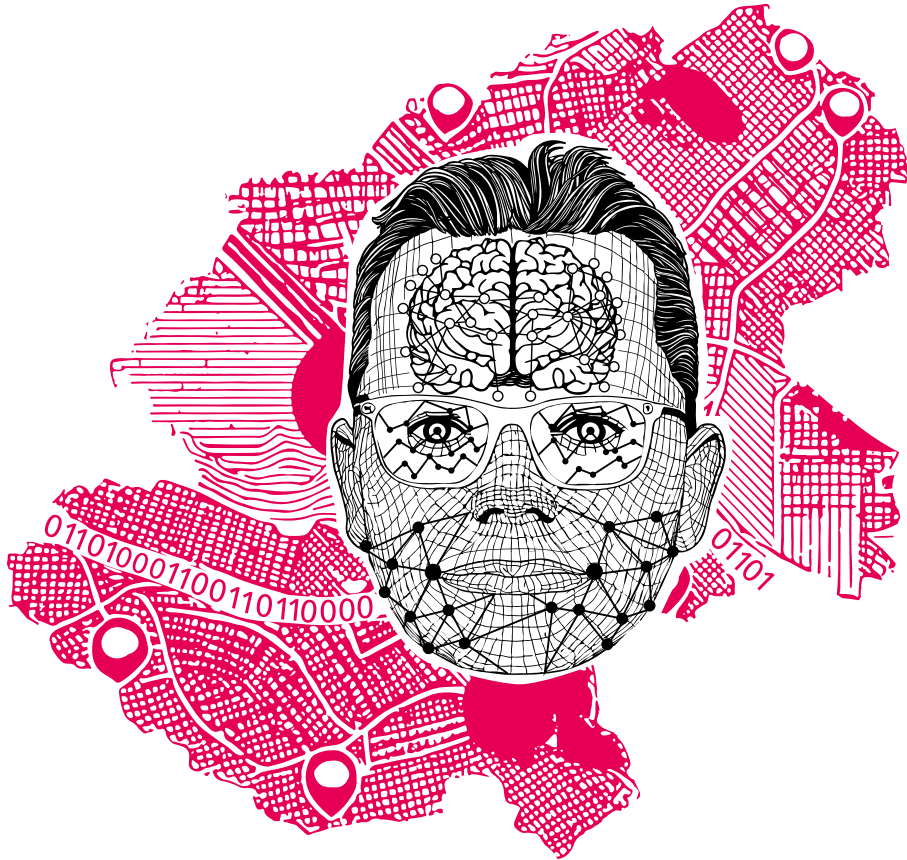


APPLIED DATA SCIENCE & AI: HOE VERDER?



Lectoraat Applied Data Science & AI
Dr. ir. Erwin Folmer
21 november 2024

OPEN UP NEW HORIZONS.

HAN UNIVERSITY
OF APPLIED SCIENCES

COLOFON

HAN University of Applied Science, Academie IT en Mediadesign

Lectoraat Applied Data Science & AI (ADSAI)

<https://www.han.nl/onderzoek/lectoraten/lectoraat-applied-data-science-and-ai/>

LECTOR

dr.ir. Erwin Folmer

Erwin.folmer@han.nl

VORMGEVING

HAN Studio MC, Roswitha Teerink

AFBEELDINGEN

Vera Lange & B302

Uitgave HAN University of Applied Sciences Press, Arnhem, 21-11-2024

@Alles uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotografie, microfilm, geluidsband of op welke andere wijze dan ook, zonder voorafgaande toestemming van de auteur, mits er zorgvuldig wordt verwezen naar de auteur.

APPLIED DATA SCIENCE & AI: HOE VERDER?

Dr.ir. Erwin Folmer - Lector Applied Data Science & AI
21 november 2024

INHOUD

INTRODUCTIE	7
DEEL 1: AANLEIDING EN CONTEXT	10
1.1 Data Science & AI bij de HAN	11
1.1.1 Data Science en AI toegepast in onderwijs	11
1.1.2 HAN Academies & (HBO) Onderwijs	13
1.1.3 Het onderzoeksperspectief	13
1.1.4 Ambitie HAN	14
1.2 Ambitie Nederland	15
1.3 'Applied'	16
DEEL 2: DATA SCIENCE & AI - HET FUNDAMENT	20
2.1 Data & AI	20
2.1.1 Wat is Data?	20
2.1.2 Wat is AI?	21
2.1.3 Symbolic AI	24
2.1.4 Semantiek & Standaarden	25
2.1.5 Linked Data	26
2.1.6 *Data*	29
2.1.7 Data Governance	31
2.2 Data Science	33
2.2.1 Wat is Data Science?	33
2.2.2 Het CRISP-DM model	35
2.3 Machine learning	38
2.3.1 Algoritmes en modellen	39
2.3.2 De algoritmes uitgelegd	42
2.3.3 Evalueren met metrics	51
2.4 Specifieke vormen van AI	55
2.5 Hybrid AI - Neuro-Symbolic AI	60
DEEL 3: DATA SCIENCE & AI - DE PRAKTIJK	64
3.1 TNO - Data Essentie	64
3.1.1 Data standaarden	64
3.1.2 Data governance	65

3.1.3	Data Spaces	65
3.1.4	Stoppen met data delen	66
3.1.5	Lessen geleerd bij TNO	66
3.2	Kadaster Data Science Team	66
3.2.1	Data: De Kadaster Knowledge Graph	66
3.2.2	Data: Toepassingen op de Kadaster Knowledge Graph	70
3.2.3	Data: Lock-Unlock: Lock de data & Unlock het potentieel	72
3.2.4	AI: Voorspelmodellen	74
3.2.5	AI: Detecteren	78
3.2.5	Lessen geleerd bij het Data Science Team	82
3.3	HAN Lectoraat ADSAI	82
3.3.1	Predictive maintenance: CHANGE	82
3.3.2	Voorspellen: Future Factory	83
3.3.3	Lessen geleerd bij HAN ADSAI	85
3.4	AI voor Fun	85
3.4.1	Poging 1 - 't Perceeltje	85
3.4.2	Poging 2 - HANZY PAI	86
	DEEL 4: HET LECTORAAT ADSAI	90
4.1	Federatief Data Delen	91
4.2	Large Language Models & Knowledge Graphs	94
4.3	Responsible & explainable AI	97
4.4	(Machine/Deep learning met) Exacte AI	98
	SAMENVATTING	104
	De uitdagingen	104
	Uitdaging 1: Ruis op de lijn.	104
	Uitdaging 2: Een Data Scientist heeft 80% verloren tijd.	104
	Uitdaging 3: Het gebruik van domme data.	105
	De focus van ADSAI	105
	Tot slot	107
	DANKWOORD	112

```
- exercise_lin_reg_ M * - visualisation-exe_ M * iris.csv x
exam-main-template.tex
32 \usepackage{hyperref}
31 % This packages creates links in the pdf-document.
30
29 % for footers
28 \usepackage{fancyhdr}
27
26 % use this folder for images
25 \graphicspath{ {./img/} }
24
23 % define some shortcuts
22 \newcommand{\N}{\mathbb{N}}
21 \newcommand{\parallelism}{\mathbb{I}\!/ \mkern -5mu \!/}
20
19 \usepackage{shortlabels}[enuniten]
18
17 % Define the 'exerc' counter for exercises
16 \newcounter{exerc}
15
14 % Custom Exercise Environment
13 \newenvironment{exercice}[1][{}]{%
12 \refstepcounter{exerc}%
11 \par\medskip
10 \noindent\textbf{Exercice-}\theexerc. (#1)\! } \rfamily}{\v
9
8
7 \pagestyle{fancy}
6 \fancyhf{} % Clear all header and footer fields
5 \fancyfoot[LO,CE]{DATEXP01 TOETS-01}
4 \fancyfoot[RO,CE]{\thepage}
3
2 % Redefine the plain page style
1 \fancypagestyle{plain}{
56 \fancyhf{}
1 \fancyfoot[LO,CE]{DATEXP01 TOETS-01}
2 \fancyfoot[RO,CE]{\thepage}
3 }
4
5
6 \title{exam UDS1 Data Exploration}
7 \date{20 September 2024}
8 \begin{document}
9
10 \maketitle
11
12 \input{exercice1.tex}
13 \input{exercice_lin_reg_sept_24.tex}
14 % \input{clustering_extra_her.tex}
15 \input{visualisation-exercice.tex}
16 \end{document}
17
18 \input{exercice1.tex}
19 \input{exercice2.tex}
20 \input{exercice3.tex}
```

```
exam-main-template.tex
30
29 % for footers
28 \usepackage{fancyhdr}
27
26 % use this folder for images
25 \graphicspath{ {./img/} }
24
23 % define some shortcuts
22 \newcommand{\N}{\mathbb{N}}
21 \newcommand{\parallelism}{\mathbb{I}\!/ \mkern -5mu \!/}
20
19 \usepackage{shortlabels}[enuniten]
18
17 % Define the 'exerc' counter for exercises
16 \newcounter{exerc}
15
14 % Custom Exercise Environment
13 \newenvironment{exercice}[1][{}]{%
12 \refstepcounter{exerc}%
11 \par\medskip
10 \noindent\textbf{Exercice-}\theexerc. (#1)\! } \rfamily}{\v
9
8
7 \pagestyle{fancy}
6 \fancyhf{} % Clear all header and footer fields
5 \fancyfoot[LO,CE]{DATEXP01 TOETS-01}
4 \fancyfoot[RO,CE]{\thepage}
3
2 % Redefine the plain page style
1 \fancypagestyle{plain}{
56 \fancyhf{}
1 \fancyfoot[LO,CE]{DATEXP01 TOETS-01}
2 \fancyfoot[RO,CE]{\thepage}
3 }
4
5
6 \title{exam UDS1 Data Exploration}
7 \date{20 September 2024}
8 \begin{document}
9
10 \maketitle
11
12 \input{exercice1.tex}
13 \input{exercice_lin_reg_sept_24.tex}
14 % \input{clustering_extra_her.tex}
15 \input{visualisation-exercice.tex}
16 \end{document}
```

INTRODUCTIE

Deze publicatie is geschreven als basis voor de intrede als Lector Applied Data Science & AI. De benoeming als lector gaat hand in hand met het opstarten van het nieuwe lectoraat (met dezelfde naam) aan de HAN.

Mede daarom is deze publicatie bedoeld om zowel de inhoudelijke context als de richting en ambitie van het nieuwe lectoraat neer te zetten. De inhoud bestaat uit vier delen:

Deel 1: De aanleiding & context

Deel 2: Data Science & AI - Het fundament

Deel 3: Data Science & AI - De praktijk

Deel 4: Het lectoraat Applied Data Science & AI (ADSAI)

De publicatie wordt afgesloten met een samenvatting & woord van dank.



HA



DEEL 1: AANLEIDING EN CONTEXT

Er gaat geen dag voorbij of iets met Data of AI (Artificial Intelligence) komt in het nieuws. Van een datalek tot de nieuwste ontwikkelingen rond AI. Die ontwikkelingen gaan in een razend tempo.

Dit lijkt een grote tegenstelling ten opzichte van de ontwikkelingen in de datawereld die vaak juist tergend langzaam gaan. Denk aan EDI en XML-berichtenverkeer; oude technologie die nog steeds gangbaar is. Of de zeer langzame groei van concepten zoals Linked Data. De eerste standaarden voor Linked Data zoals RDF stammen uit 1999 (RDF 1.0 als W3C recommendation). Linked Data is inmiddels niet meer weg te denken uit het IT-landschap, maar is nog altijd geen mainstream IT.

In AI-land gaat het anders. Mogelijk doordat het getriggerd wordt door de Big Tech die verwachten hier veel geld mee te kunnen verdienen. Hoe disruptief AI ook wordt neergezet, en hoe groot de kansen ook zijn: relatief zien we nog weinig toepassingen binnen de overheid en het bedrijfsleven. Tuurlijk zien we de mooie voorbeelden van een chatbot, een gegenereerde tekst, een door AI gemaakt plaatje, of een interessant voorspelmodel op basis van historische data, maar de meeste disruptieve toepassingen moeten nog komen.

We zien veel experimenten, maar die leiden lang niet altijd tot daadwerkelijke implementaties. Veel organisaties zijn risicomijdend en dat is best te begrijpen. Kennis & ethische kwesties spelen daarbij een rol, maar evenzo belangrijk zijn vertrouwen in de aanpak en technologie, en de betrouwbaarheid van de uitkomst.

Aan de andere kant zien we ook veel toepassingen die we al jarenlang gebruiken, maar nu gelabeld worden als AI.

Denk daarbij aan modellen voor personeelsplanning en het voorspellen van economische trends. Veelal statische aanpakken die nu ook onder de noemer AI vallen.

De belangrijkste les die we eigenlijk allemaal al kennen, maar soms toch weer onder de aandacht moeten brengen: AI drijft op data. Om toepassingen van AI te realiseren die implementeerbaar zijn, hebben we goede beschikbaarheid van data nodig. AI en data gaan hand in hand. Daarom focussen we ons niet alleen op de AI (of breder de Data Science toepassingen), maar ook op het datafundament dat we nodig hebben voor Data Science & AI-toepassingen. De risico's en beperkingen van AI hebben veelal een belangrijke relatie met aspecten van de gebruikte data.

En dan komen we in de wereld van Big Data, Data Lakes, Data Mesh, Data Spaces; een wereld van business en techniek, veel hypes, veel leveranciersbelangen; waar niet alle ontwikkelingen in de loop der jaren even succesvol zijn gebleken.

In deze publicatie pakken we de onderwerpen Data (Science) en AI samen op, maar zijn we wel sterk gefocust op het realiseren van toepassingen. Want dat past bij de HAN.

1.1 DATA SCIENCE & AI BIJ DE HAN

De HAN is een grote onderwijsorganisatie, primair gericht op HBO-onderwijs, maar de laatste jaren is dat flink uitgebreid met andere onderwijsvormen en ook onderzoek in lectoraten. Data Science & AI komt dan ook op verschillende plekken terug.

1.1.1 Data Science en AI toegepast in onderwijs

Binnen de HAN zijn principes opgesteld die ondersteunen bij de keuzes rond het inzetten van AI, met name in het onderwijs. Interessant daarbij is hoe AI zeer breed gedefinieerd wordt, inclusief andere Data Science en statistisch benaderingen (zie kader). De principes zijn verder zeer generiek van aard, waardoor de impact van de keuze van een extreem brede definitie op AI gering is.

Definitie Kunstmatige Intelligentie (AI): AI bestaat als vakgebied al sinds de jaren 50. De ontwikkelingen van AI zijn met horten en stoten verlopen, waardoor de definitie en het doel van AI ook aan veranderingen onderhevig zijn geweest. Een veel gebruikte en recente definitie is die van de Europese Commissie: 'artificiële-intelligentiesysteem' (AI-systeem): software die is ontwikkeld [door mensen] aan de hand van een of meer van de technieken en benaderingen die zijn opgenomen in de [onderstaande] lijst en die voor een bepaalde reeks door mensen gedefinieerde doelstellingen output kan genereren, zoals inhoud, voorspellingen, aanbevelingen of beslissingen die van invloed zijn op de omgeving waarmee wordt geïnterageerd; a. Benaderingen voor machinaal leren, waaronder gecontroleerd, ongecontroleerd en versterkend leren, met behulp van een brede waaier aan methoden, waaronder diep leren ('deep learning'). b. Op logica en op kennis gebaseerde benaderingen, waaronder kennisrepresentatie, inductief (logisch) programmeren, kennisbanken, inferentie- en deductiemachines, (symbolisch) redeneren en expertsystemen. c. Statistische benaderingen, Bayesiaanse schattings-, zoek- en optimalisatiemethoden. Volgens deze definitie van AI is er altijd sprake van menselijke invloed op AI: zowel in de creatie, de software, die immers door mensen geschreven wordt, als in de door mensen gedefinieerde doelstellingen. De klasse van AI-systemen die als output 'inhoud' geven (zoals afbeeldingen, tekst, muziek) noemen we generatieve AI (ook wel 'gen AI').

Bron: Framework gebruik AI binnen de HAN.

https://www.han.nl/onderwijsondersteuning/leren-werken-met-ict/artificial-intelligence/Framework_gebruik_AI_binnen_de_HAN.pdf

Kaders en principes zijn belangrijk, maar evenzo belangrijk is de praktische vertaling naar de onderwijssituatie. Daarvoor is in 2023 het experiment Sandbox AI gestart. Met de Sandbox AI maken we het mogelijk om voor een groep collega's te verkennen hoe docenten generatieve AI in de onderwijspraktijk verantwoord

kunnen inzetten. De AI-Sandbox is een leernetwerk en heeft tot doel om samen kennis over AI en toepassingen in de onderwijspraktijk te ontwikkelen. De resultaten van de eerste fase zijn gedocumenteerd¹.

1.1.2 HAN Academies & (HBO) Onderwijs

In verschillende opleidingen bij verschillende academies komt Data Science & AI terug in het onderwijsaanbod. Bij de opleidingen Toegepaste Biowetenschappen en Chemie (ATBC) is het werken met grote datasets en het uitvoeren van data-analyse onderdeel van het onderwijsaanbod. Bij de opleiding Master Ontwerpen Van Eigentijds Leren (MOVE) vanuit academie Educatie (AE) is aandacht voor de toepassing van AI in het onderwijs, zoals de impact van ChatGPT. Ook bij de opleidingen van de Academie Engineering en Automotive (AEA) en Financieel Economisch Management (FEM) zijn flarden van Data Science & AI beschikbaar in het onderwijs.

De Academie IT en Mediadesign (AIM) lijkt op voorhand met de HBO-ICT opleiding de landingsplek voor de inbedding van Data Science & AI in het curriculum; het Data Solutions Development (DSD) uitstroomprofiel bij HBO-ICT komt daar nu het dichtst in de buurt. De studenten kunnen hier al in aanraking komen met machine learning. Het HBO-ICT curriculum wordt in 2024-2025 hervormd, en de nadrukkelijke wens is uitgesproken om meer Data Science & AI in dit curriculum op te nemen, bij voorkeur middels een eigen uitstroomprofiel.

Vanuit AIM (in nauwe samenwerking met andere academies) is in 2023 de deeltijdopleiding MADS² (Master Applied Data Science) gestart, volledig gericht op het onderwerp Data Science & AI.

1.1.3 Het onderzoeksperspectief

Bovengenoemde academies hebben in hun onderzoekstaken ook aandacht voor Data Science & AI. Specifieke aandacht verdient daarbij het lectoraat LEAN (AEA). LEAN is het kenniscentrum voor het slim organiseren en continu verbeteren van productontwikkeling, smart produceren, veranderen en leren, en data speelt natuurlijk een belangrijker rol.

1 <https://www.han.nl/onderwijsondersteuning/leren-werken-met-ict/artificial-intelligence/han-ai-sandbox/>

2 <https://www.han.nl/opleidingen/master/applied-data-science/deeltijd/>

Ook bijzondere aandacht verdient het lectoraat Leren met ICT (AE) en het daaraan gekoppelde iXperium Centre of Expertise Leren met ICT in Nijmegen. In dit iXperium³ ontdekken we met collega's van aangesloten partners mogelijkheden van ICT-rijk onderwijs, ontwikkelen we kennis en bieden we begeleiding bij de implementatie. Uiteraard gaat dit breder dan alleen AI, maar AI neemt hier wel een prominente plek in.

Bij deze academies (AEA, AE, maar ook ATBC, FEM) staan specifieke toepassingsdomeinen centraal, en is Data Science en AI een hulpmiddel, of anders geformuleerd 'een technisch noodzakelijk kwaad'. Bij AIM is dat omgekeerd; daar staat de technologie centraal en proberen we dat breed toe te passen over verschillende toepassingsdomeinen heen. Het lectoraat Data & Knowledge Engineering (DKE) is ook sterk gelieerd aan het onderwerp Data Science & AI, maar is vooral gericht op het onderliggende data fundament, de kennisextractie en -representatie, en de taalaspecten, waarbij de Large Language Models (LLMs) een typische AI-toepassing zijn.

1.1.4 Ambitie HAN

De ambitie van de HAN is opgenomen in het koersbeeld: 'Voor een slimme, schone en sociale wereld van morgen.' Een expliciet geformuleerde doelstelling in het koersbeeld is daarbij: 'Studenten en medewerkers worden digi- en datavaardig.' Tevens worden hiermee de drie zwaartepunten Slim (Smart Region), Schoon (Sustainable Energy & Environment), Sociaal (Fair Health) geïntroduceerd, als focusgebieden beschreven in de HAN Agenda. De zwaartepunten hebben elk thema's gedefinieerd. Voor Slim zijn dat onder andere de thema's Digital Twinning en AI. Oftewel een expliciete relatie met het onderwerp Data Science & AI. Bij de andere zwaartepunten is dat meer impliciet; bijvoorbeeld voor Schoon is Data Science & AI bij uitstek geschikt om vraagstukken rond de energietransitie mee te tackelen.

Daarnaast werkt de HAN met academieplannen; waarin de ambitie en focus vertaald worden naar een tactisch/operationeel plan voor de academie, en met de Kennis en Innovatie Agenda (KIA). In de KIA wordt de nadruk gelegd op het belang van de Key Enabling Technologies (KET) en Key Enabling Methodologies (KEM); in het IT-domein zijn AI en Data Science de nummers 1 en 2 op de lijst van KETs⁴.

³ <https://www.ixperium.nl/ixperiums/ixperium-nijmegen/>

⁴ <https://www.nwo.nl/en/key-enabling-technologies>

Het moge duidelijk zijn dat de HAN ambitie heeft op het gebied van Data Science & AI, en dat de contouren geschetst zijn om hier invulling aan te geven. Het werd in het onderwijs concreet met de lancering van de deeltijd Master opleiding MADS, een belangrijke eerste mijlpaal. Als tweede mijlpaal kan de lancering van het nieuwe lectoraat Applied Data Science & AI⁵ (ADSAI) beschouwd worden. De opleiding MADS en het lectoraat ADSAI zijn nauw met elkaar verbonden; het lectoraat zal de komende jaren een nadrukkelijke rol gaan spelen in het verder uitbouwen van de MADS opleiding.

1.2 AMBITIE NEDERLAND

De ambitie van de HAN sluit naadloos aan bij de ambities van Nederland, verwoord in de Nationale Technologie Strategie (NTS)⁶. De NTS prioriteert 10 sleuteltechnologieën die: 1. een grote bijdrage leveren aan ons verdienvermogen, 2. cruciaal zijn voor maatschappelijke uitdagingen, 3. belangrijk zijn voor de nationale veiligheid en 4. Nederlands technologisch leiderschap mogelijk maken. AI en Data Science zijn daarbij aangewezen als 1 van de 10 topprioriteiten voor Nederland. Uitvoering van de plannen wordt gedaan onder leiding van de Topsector ICT, die in de plannen zeven sleuteltechnologieën heeft vastgesteld: 1. Artificial Intelligence, 2. Data Science, data analytics and Data Spaces, 3. Cyber security technologies, 4. Software technologies and computing, 5. Digital Connectivity Technologies, 6. Digital Twinning and Immersive technologies, 7. Neuromorphic technologies. De nummers 1, 2, 4 en 7 hebben een duidelijke relatie met Data Science en AI.

Onder de digitale sleuteltechnologie Data Science, Data Analytics & Data Spaces vallen alle aspecten rond het werken met data: verzamelen, beheren, ontsluiten, delen, verwerken en analyseren van data.

⁵ <https://www.han.nl/onderzoek/lectoraten/lectoraat-applied-data-science-and-ai/>

⁶ <https://www.rijksoverheid.nl/documenten/beleidsnotas/2024/01/19/de-nationale-technologiestrategie>

Deze digitale sleuteltechnologie (Data Science, Data Analytics & Data Spaces) wordt in de NTS prioritair benoemd voor de aanpak van belangrijke economische en maatschappelijke uitdagingen⁷.

Verdere invulling en afspraken zijn gemaakt in de KIA Digitalisering⁸ en de KIC (Kennis- en Innovatieconvenant⁹). Hierin zijn drie soorten innovatie richtingen voor de sleuteltechnologie AI benoemd¹⁰: 1. Technische innovatie in AI, 2. Maatschappelijke en bedrijfsmatige innovatie met AI, 3. Inzichten in voorwaarden voor het gebruik van AI (vaak samengebracht onder de noemer ELSA - ethical, legal & social aspects).

De ambitie van de HAN sluit vooral goed aan bij nummer 2: Maatschappelijke en bedrijfsmatige innovatie met AI. Dit zullen we in de volgende paragraaf toelichten.

1.3 'APPLIED'

Met een focus op een (Key Enabling) Technologie ligt ook een onderzoeksfocus op de loer die meer fundamenteel academisch op de technologie gericht is. Dat past misschien bij een universiteit (als het überhaupt nog realistisch is om hier als Nederland een rol van betekenis te spelen in dit internationale geweld), maar zeker niet bij de HAN. De naam van het lectoraat bevat daarom bewust de term 'Applied' om te benadrukken dat het uiteindelijk gaat om waardecreatie, of het nu economisch of maatschappelijk is. Die wordt gerealiseerd door toepassingen op basis van data en Data Science technieken zoals AI. De focus ligt dus niet op het onderzoek doen naar fundamentele nieuwe technologische hoogstandjes, maar op de state of the art technologie toepasbaar te maken.

Zoals we bij Data vaak over de V's van Data spreken (Volume, Variety, Veracity, Velocity), kunnen we dat bij Data Science en AI doen met de B's: Betrouwbaarheid, Betaalbaarheid en Beheersbaarheid¹¹.

Onderzoek naar de 3 B's zullen een boost geven aan de toepassingen met Data Science & AI en daarmee aan de waardecreatie. Een mooi startpunt.

7 <https://topsector-ict.nl/nieuws/waarom-datatechnologie-%C3%A9%C3%A9n-van-de-prioriteiten-is-van-de-bv-nederland>

8 <https://kia-digitalisering.nl/>

9 <https://topsector-ict.nl/nieuws/het-nieuwe-kennis-en-innovatieconvenant-2024-2027>

10 <https://hollandhightech.nl/hoe-we-helpen/kia-sleuteltechnologieen>

11 <https://www.computable.nl/2024/04/18/de-3-bs-van-de-ai-wereld/>



SPARQL → Query

LINKED DATA
SEMANTIC
Web

Select ?Huisnummer
 ?huis

```

a kkg:Huis ;
kkg:tuin ?tuinmz ;
kkg:metHuisnummer ?huisnummer ;
kkg:inWoonplaats/rdb:label ?huisnummer ;
?huisnummer desc (?tuinmz)

```

KNOWLEDGE GRAPH



RDF
RDF
SKOS

GRAPHS

DEEL 2: DATA SCIENCE & AI - HET FUNDAMENT

Hal Varian, hoofdeconoom van Google, noemde in 2012 de Data Scientist het meest sexy beroep van het moment¹². Het is misschien belangrijk om te benadrukken dat het hier gaat om de gewildheid van Data Scientists op de arbeidsmarkt in plaats van de uiterlijke kenmerken van de beroepsgroep (al zit het daarmee ook wel goed bij ADSAI). Ruim een decennium na de uitspraak van Hal Varian is Data Science nog steeds hot en is daar AI bijgekomen.

Er wordt veel gesproken over Data Science en AI maar in dit deel proberen we op een toegankelijke manier toch iets dieper de materie in te duiken. We zullen onder andere uitleggen wat het verschil is tussen data en informatie, wat het CRISP-DM model is, wat het verschil is tussen logistische en lineaire regressie, en tussen machine learning en deep learning. Na het lezen van dit deel zul je beter begrijpen waar we de hele dag mee bezig zijn op het lectoraat en wat de uitdagingen en kansen zijn van ons werk.

2.1 DATA & AI

2.1.1 Wat is Data?

De DIKW-piramide¹³ beschrijft hoe ruwe gegevens kunnen worden omgezet in waardevolle en relevante inzichten wanneer data effectief beheerd en gebruikt worden. De piramide, afgebeeld in Figuur 1, maakt onderscheid tussen Data, Informatie, Kennis en Wijsheid; die samen de afkorting DIKW vormen.

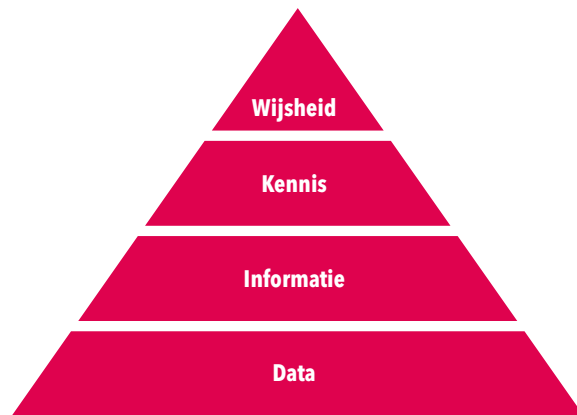
Data vormt de onderste laag van de piramide en bestaat uit ruwe, ongeorganiseerde feiten en cijfers zonder enige context of betekenis. Denk hierbij bijvoorbeeld aan een lijst met getallen of losse woorden zonder enig verband. Het toevoegen van een context en structureren zorgt ervoor dat de data wordt omgezet in informatie. Informatie beantwoordt vragen zoals 'wie', 'wat', 'waar', en 'wanneer'.

¹² <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

¹³ https://en.wikipedia.org/wiki/DIKW_pyramid

Het heeft meer waarde dan data omdat het ons iets vertelt, zoals een grafiek van de onderhoudskosten van een vervoersbedrijf of de frequentie van bepaalde defecten in een vloot van voertuigen. De volgende stap in de piramide is kennis. Kennis ontstaat wanneer informatie wordt geanalyseerd, geïnterpreteerd en begrepen binnen een bepaalde context. Het impliceert een begrip van hoe verschillende stukken informatie met elkaar in verband staan en hoe ze kunnen worden toegepast. Kennis kan bijvoorbeeld inzicht bieden in wat de onderliggende oorzaken van een bepaald type mankement aan een vrachtwagen is. Bovenaan de piramide bevindt zich wijsheid. Wijsheid staat voor een diepgaand begrip en de toepassing van kennis in besluitvorming en actie. Het is het vermogen om de juiste keuzes te maken in complexe situaties, bijvoorbeeld door niet alleen te begrijpen waarom bepaalde defecten aan vrachtwagens ontstaan, maar ook hoe die kennis kan worden gebruikt om beslissingen te nemen over het onderhoud van deze voertuigen.

In essentie toont de DIKW-piramide aan hoe data kan worden omgezet in wijsheid door middel van een proces van het toevoegen van context, begrip en toepassing, waarbij elke stap in de piramide een toenemende waarde en betekenis toevoegt.



Figuur 1 - De DIKW-piramide

2.1.2 Wat is AI?

In de wereld van digitale technologie is AI een fascinerend onderzoeksveld dat steeds meer in de schijnwerpers staat. Waar Data Science zich richt op het analyseren en interpreteren van gegevens binnen een organisatie, beschrijft AI een

digitale technologie die menselijke intelligentie nabootst. IBM¹⁴ geeft de volgende treffende definitie van AI:

“Kunstmatige intelligentie, of AI, is de technologie die computers en machines in staat stelt menselijke intelligentie en probleemoplossende capaciteiten te simuleren.”

Een veel gebruikte toepassing van AI is het voorspellen van zaken op basis van een model met behulp van trainingsdata. Een andere vorm die snel in opmars is, is generatieve AI, waarmee nieuwe content (zoals tekst, afbeeldingen, video, code) gecreëerd kan worden. Sinds de introductie van ChatGPT heeft de term AI enorm aan populariteit gewonnen. Zo stijgt het aantal zoekacties op Google met de termen ‘AI’ en ‘ChatGPT’ significant na november 2022, de maand van ChatGPT’s officiële introductie (zie Figuur 2). Gezien de lange geschiedenis van de term is het toch opmerkelijk om te zien dat de term AI zo trending kan worden. De term is namelijk stokoud voor de begrippen van de digitale wereld.



Figuur 2 - De populariteit op Google van de zoektermen AI & ChatGPT in de periode 2020-2024 in Nederland¹⁵

14 <https://www.ibm.com/topics/artificial-intelligence>

15 <https://trends.google.com/trends/explore?date=today%205-y&geo=NL&q=ai,chatgpt&hl=nl>

De term AI werd voor het eerst gebruikt in 1955 tijdens de voorbereidingen voor een zes weken durend congres, dat een jaar later werd gehouden. Het onderwerp van dat congres was de ontwikkeling van AI. Bij dat congres kwamen twintig wiskundigen bij elkaar om te discussiëren over het nabootsen van menselijke intelligentie met machines¹⁶. De nu zo populaire term is dus bedacht toen een computer nog niet eens in een gemiddelde slaapkamer paste, kijk maar naar de IBM 305 RAMAC (zie Figuur 3) die dat jaar vol trots door IBM werd gepresenteerd.

Na het congres in 1956 was het optimisme groot, de wetenschappers hadden hoge verwachtingen van de nieuwe technologie. Zo bestond het idee dat computers binnen tien jaar beter zouden kunnen schaken dan mensen. De werkelijkheid bleek weerbarstiger en de ontwikkeling van de techniek liep vertraging op. Zoals we nu weten lijkt dit inherent aan het uitvoeren van IT-projecten. Uiteindelijk werd in werkelijkheid de heersende wereldkampioen Garri Kasparov pas in 1997 verslagen door AI.

Toch is tien jaar na de conferentie de eerst echt werkende vorm van AI geïntroduceerd: expert systemen. Deze systemen zijn gebaseerd op de domeinkennis van experts en bestaan uit grote beslisbomen met veel alsdan regels. Gebruikers van de systemen moeten vragen beantwoorden en op basis daarvan ontvangen ze een advies. De systemen zijn nuttig maar hebben verschillende beperkingen. Zo is het moeilijk om de kennis van de experts te vergaren en is het onderhoud van de systemen een uitdaging¹⁷.

16 <https://spectrum.ieee.org/dartmouth-ai-workshop>

17 Koolstra, S., De Veer, B., & Veltman, T. (2021). *Dit is kunstmatige intelligentie: Een introductie in de technologie die ons leven steeds meer bepaalt*. Van Haren.



Figuur 3 - De IBM 305 RAMAC computer ¹⁸

Na de introductie van de expert systemen neemt de aandacht voor AI af. Daarmee begint halverwege de jaren 70 de eerste AI-winter waarin er weinig aandacht is voor de ontwikkeling en het gebruik van AI. In de jaren 80 herleeft kortstondig de aandacht voor AI en groeit het gebruik van de expert systems weer, om vervolgens weer net zo hard in te storten¹⁹. Vervolgens breekt dan ook de tweede AI-winter aan totdat rond het jaar 2000 de rekenkracht en de beschikbaarheid van data verbetert. Hierdoor kunnen meer geavanceerde modellen gebruikt worden die automatisch leren van reeds beschikbare data. Deze modellen zijn makkelijker te onderhouden en zijn niet zo afhankelijk van experts²⁰.

2.1.3 Symbolic AI

Symbolic AI is een onderdeel van een breder scala aan AI-technieken. Symbolic AI verschilt van andere AI-technieken, zoals machine learning en deep learning, doordat het geen enorme hoeveelheden trainingsdata vereist. In plaats daarvan is Symbolic AI gebaseerd op kennisrepresentatie en redenering. Symbolen en concepten (semantiek) staan centraal in plaats van data als getallen.

18 <https://digitalmuseum.org/011015239966/22-0-ibm-modell-305-ramac>

19 https://en.wikipedia.org/wiki/Expert_system

20 <https://www.techtarget.com/searchenterpriseai/definition/AI-winter>

Symbolic AI-algoritmes werken door symbolen te verwerken, die objecten of concepten in de wereld vertegenwoordigen, en hun relaties. De belangrijkste aanpak in Symbolic AI is het gebruik van op logica gebaseerde code, waarbij regels en axioma's worden gebruikt om gevolgtrekkingen en deducties te maken. Bijvoorbeeld de eerdergenoemde beslisbomen (bedrijfsregels, als-dan regels) die al in de expert systemen gebruikt werden. Symbolic AI gebruikt formele talen om kennis te representeren en redenering mogelijk te maken. Op dit moment zijn Semantic Web (Linked Data) standaarden met RDFS (RDF Schema), OWL (Web Ontology Language) en SHACL (Shapes Constraint Language) een veelgebruikte aanpak en taal voor het vastleggen van kennis.

De kracht van de formele kennisrepresentatie maakt het uitermate geschikt voor toepassing in domeinen waar accuraatheid van groot belang is en waar kennis goed gedefinieerd kan worden in logische regels.

De voordelen van Symbolic AI zijn²¹:

Interpreteerbaarheid: Symbolic AI zorgt voor transparantie in het redeneringsproces, waardoor het gemakkelijker wordt om te begrijpen hoe een systeem tot een conclusie is gekomen.

Kennisrepresentatie: Symbolic AI kan complexe kennis op een formele en gestructureerde manier vastleggen, waardoor deze eenvoudig gebruikt kan worden in applicaties.

Flexibiliteit: Symbolic AI is zeer flexibel en kan worden aangepast aan verschillende domeinen door de kennisrepresentatie aan te passen of uit te breiden.

2.1.4 Semantiek & Standaarden

Data heeft betekenis en context. Data gebruiken zonder kennis te hebben over de betekenis, dus zonder het op te werken naar informatie en kennis, is gelijk aan het gebruiken van een medicijn zonder de bijsluiter te lezen. Ongewenst en gevaarlijk. Dus we willen die betekenis vastleggen, bij voorkeur in standaarden, zodat we eenduidig gebruik krijgen van de vastgestelde betekenis, en daarmee semantische interoperabiliteit.

²¹ <https://www.datacamp.com/blog/what-is-symbolic-ai>

Een complicerende factor daarbij is dat de betekenis van mensen, dingen, gebeurtenissen, et cetera, niet constant is. Die betekenis kan variëren. Zo kan iets meerdere benamingen hebben. Zoals bij 's-Hertogenbosch, Den Bosch en Oeteldonk, waarbij het om dezelfde stad gaat (in een andere context). Of dezelfde benaming kan gebruikt worden binnen verschillende contexten met een verschillende betekenis. Zoals bij Bastille, wat een monument, een fort of een gevangenis kan zijn. Dit kan bij onjuist of slordig gebruik tot verwarring leiden. Mensen zijn gewend om contextuele factoren mee te nemen bij het toekennen van betekenis aan informatie. Voor machines geldt dit niet. Om computers toch in staat te stellen de juiste betekenis toe te kennen, is het aanbieden van relevante context van groot belang. Ook andere vormen van metadata, zoals informatie over de herkomst (provenance) van data is van groot belang om data op te werken naar informatie en kennis.

Dit speelt zeer prominent in verschillende toepassingsdomeinen (sectoren). Deze spreken andere talen (begrippen) en gebruiken dan ook andere semantische standaarden, waardoor semantische interoperabiliteit een uitdaging is.

Daar komt Linked Data om de hoek kijken. Linked Data is een techniek om met machine-leesbare context om te gaan, deze te genereren en te interpreteren. Het biedt dan ook de mogelijkheid om met meerdere contexten (en betekenissen) van data om te gaan.

2.1.5 Linked Data

Linked Data is een manier om gestructureerde data te publiceren zodanig dat data met elkaar verbonden kan worden, en semantiek en context als onderdeel van de data kunnen worden opgenomen. Het is gebaseerd op de fundamenteën van het World Wide Web en is in grote mate gestandaardiseerd met open W3C (World Wide Web Consortium) standaarden, zoals RDF en SPARQL.

De openheid van deze standaarden is een cruciaal en uniek aspect van Linked Data om interoperabiliteit en leveranciersafhankelijkheid te bewerkstelligen. Vandaar dat de Linked Data (W3C) standaarden opgenomen zijn op de lijsten van het Forum Standaardisatie, om ze te stimuleren voor het gebruik in de (semi-) publieke sector, of zelfs te verplichten²².

²² <https://www.forumstandaardisatie.nl/open-standaarden>

De basis van Linked Data is een data modellering standaard genaamd Resource Description Framework (RDF), waarmee data in zogenaamde triples (subject-predicate-object structuur) kan worden vastgelegd. In aanvulling op RDF kan RDFS (RDF Schema) gebruikt worden. Met behulp van RDFS kunnen 'klassen' van resources aangemaakt worden en ook beperkingen worden gelegd op de verschillende relaties die mogelijk zijn tussen instanties van deze klassen. Inmiddels is er een flink aantal RDF-standaarddefinities beschikbaar van resource- en relatietypes. Deze zijn vastgelegd in zogenoemde vocabulaires. Voorbeelden van veelgebruikte vocabulaires zijn SKOS (Simple Knowledge Organization System), voor het opstellen van begrippenkaders, gegevenswoordenboeken, taxonomieën en thesauri, en FOAF (Friend of a Friend); een vocabulaire dat gebruik maakt van RDF om personen te beschrijven, hun relaties met andere personen en voorwerpen, en hun interacties. Complexere vocabulaires zijn ontologieën, en deze kunnen bijvoorbeeld gemaakt worden op basis van de OWL (Web Ontology Language) standaard.

Kenmerkende principes van Linked Data zijn:

- Alle data wordt vastgelegd in triples (subject-predicate-object). Deze triples vormen grafen die visueel lijken op ketens en netwerken van datawolken.
- Alle informatie wordt belicht vanuit een bepaalde invalshoek (viewpoints). Voor een volledig beeld van de situatie kunnen viewpoints gecombineerd worden tot een overzichtelijk en samenhangend verhaal.
- Linked Data werkt met een 'open world assumption'. Er kan altijd meer data beschikbaar komen en gebrek aan data betekent niet dat iets niet waar kan zijn. Een antwoord op een vraag kan dan 'misschien' of 'onbekend' zijn.
- In de wereld van Linked Data kan iedereen een gegeven verrijken met eigen informatie. Zo komen verschillende perspectieven bij elkaar. Deze eigenschap wordt ook wel de AAA-slogan genoemd, *Anybody can say Anything about Any topic*.
- Door gebruik te maken van bestaande vocabulaires wordt de interoperabiliteit tussen gegevens vergroot, waardoor data uit verschillende bronnen makkelijker met elkaar gecombineerd kunnen worden. De verschillen tussen data uit verschillende bronnen zijn makkelijker te overbruggen met Linked Data, omdat meer dezelfde 'taal' gesproken wordt.

Ook kunnen datasilo's uit hun isolement worden gehaald en data beter hergebruikt worden zonder dat data onnodig gekopieerd wordt. Er kan gelinkt worden naar één leidende bron.

- Met Linked Data is het mogelijk om een grote hoeveelheid en verscheidenheid aan data met elkaar in verband te brengen. Daarbij maakt het niet uit of het nu kaartmateriaal is, illustraties, informatie op een webpagina, of gegevens uit een database.
- Met Linked Data krijgen gegevens context. Een context die digitaal te verwerken is: de gegevens vertellen een verhaal. Hierdoor kunnen enorme hoeveelheden gegevens met elkaar worden verbonden en verwerkt.
- Met Linked Data is verregaande data discovery mogelijk. Je kunt nieuwe data vinden, waarvan je het bestaan vooraf niet wist, maar die gelinkt is aan de databron waarmee je je zoekvraag begon. Zoekacties worden daarmee gericht.
- Databronnen kunnen op verschillende locaties staan en toch met één SPARQL query bevraagd worden, door gebruik te maken van federated queries.

In feite proberen we met Linked Data te werken aan rijke data, zodat in de toepassingen informatie, kennis en wijsheid kan worden verkregen. Op het gebied van open data is het 5-sterren model²³ een veel gebruikte manier om de mate van herbruikbaarheid van data aan te geven:

1 ster: de informatie is beschikbaar op het internet (met open licentie), in welk formaat dan ook.

2 sterren: De informatie is online beschikbaar in een gestructureerd formaat, dat geschikt is voor automatisch hergebruik (zoals Excel in plaats van een plaatje van een tabel).

3 sterren: De informatie is online beschikbaar in een open bestandsformaat (zoals CSV in plaats van Excel).

4 sterren: De informatie is online beschikbaar conform de standaarden (o.a. RDF) en aanpak (o.a. gebruik van URIs) van Linked Data / W3C. Zo kunnen er eenvoudig relaties tussen dataobjecten worden aangebracht.

5 sterren: Al het bovenstaande, en bovendien wordt er naar data van anderen verwezen voor meer context van de data.

²³ <https://5stardata.info/en/>

De eerste drie sterren betreffen open data en vanaf de vierde en vijfde ster wordt gesproken over Linked (Open) Data. Je zou kunnen zeggen bij vier sterren is de data 'linkable' (URLs, RDF, et cetera) en bij vijf sterren ook daadwerkelijk gelinkt. Uiteraard geldt dit model net zo goed voor niet-open data, maar dan zou het Internet vervangen kunnen worden door bijvoorbeeld Intranet, en zal de licentie de beperkingen op de data moeten beschrijven.

Linked Data (ook onder de noemer Semantisch Web en Web 3.0) is in de basis een visie op het werken met data, met daarbij open standaarden die het fundament bieden om aan deze visie invulling te geven. Toch wordt Linked Data regelmatig als 'technisch' beschouwd, mogelijk door de complexiteit en de nadruk op standaarden. Vermoedelijk is dit ook een reden dat steeds meer de term Knowledge Graphs wordt gebruikt. In feite is 5 sterren Linked Data een Knowledge Graph.

Knowledge Graphs bieden rijke (met betekenis en context) verbonden data. Dit zou dus de ideale basis kunnen worden van slimme Data Science & AI-toepassingen. Nu is de praktijk dat we vooral arme/domme data (data zonder betekenis en context) als input nemen voor AI-toepassingen. We proberen dan domheid te compenseren door maar meer en meer data toe te voegen. Het toepassen van Knowledge Graphs als fundament voor Data Science & AI-toepassing biedt veel potentie. Maar laten we eerst maar eens in meer detail naar Data Science & AI gaan kijken.

2.1.6 *Data*

Big Data was in het begin van de 21e eeuw een flinke hype. De meeste aandacht ging op dat moment uit naar het (big) volume aspect van data. In de praktijk bleek volume wel oplosbaar met geld, en vaak niet de grootste uitdaging. Een grotere uitdaging was en is de variëteit van data, wat alles van doen heeft met semantiek en context van data. De afgelopen jaren is daar 'velocity' bij gekomen, de snelheid waarmee data verandert. Dat speelt vooral bij sensordata, maar ook bij basisregistraties is er een grotere noodzaak om met actuele data te kunnen werken.

De eerder genoemde datastandaarden zijn een middel, dat onder architectuur ingezet kan worden om interoperabiliteit te bereiken (de fameuze driehoek

architectuur + standaarden = interoperabiliteit). Waarbij architectuurstijlen mode-objecten lijken te zijn; elke paar jaar hebben we weer een nieuwe hippe architectuurstijl. Afhankelijk van de leeftijd kennen we nog de Enterprise Service Bus (ESB), of de Service Oriented Architectures (SOA). Meer recent hebben we de Data Mesh architectuur en de Data Fabric; waarbij Data Mesh meer een strategie is, gericht op de organisatorische aspecten van data, terwijl Data Fabric neergezet wordt als technische oplossing. Beide lijken alweer enigszins op hun retour te zijn. Er lijkt meer aandacht voor Event Driven Architectures of Event Sourcing te komen, die meer gericht zijn op het vastleggen en communiceren van de gebeurtenissen die leiden tot een verandering in de dataset.

Onder andere de aandacht voor Data Mesh architectuur heeft eraan bijgedragen dat er meer aandacht is voor data governance. Een belangrijk en onderschat onderwerp; misschien wel de belangrijkste drempel voor het delen van data. Data is voor veel organisaties een 'asset'; oftewel een waardevol bezit. Daar ben je zuinig op, en als je het gaat delen wil je daar afspraken over maken zodat anderen er ook zuinig mee zijn. Doen ze dat niet; dan wil je als eigenaar kunnen ingrijpen.

En dat is het uitgangspunt van Data Spaces. Een Data Space is gericht is op het veilig, vertrouwd en 'soverein' delen van data binnen domeinen en over domeinen heen. Met soeverein wordt dan bedoeld dat het recht doet aan de eigenaar van de data. Ook interoperabiliteit tussen verschillende domeinen/sectoren is een belangrijk streven. Uitgangspunt daarbij is het gebruik van open standaarden om onafhankelijk te zijn van specifieke ICT-leveranciers (leveranciersneutraal).

Dit lijkt ook op SOLID. Het concept van Sir Tim Berners-Lee om het Web te repareren²⁴, en in feite het eigenaarschap van data terug te leggen daar waar het hoort: bij de burger. Bij SOLID is dat in eerste instantie gericht om de macht van de Facebooks & Googles van deze wereld te doorbreken, maar in Vlaanderen zien we de eerste implementaties ook in andere trajecten waarbij data van personen gedeeld wordt met bedrijven en overheid. Het concept van SOLID is ook sterk op (open) standaarden gericht, net zoals zijn eerdere initiatieven (het Web en Linked Data).

24 <https://www.nytimes.com/2019/11/24/opinion/world-wide-web.html>

Data Spaces hebben een oorsprong in de Duitse automobielenindustrie (industrie 4.0), maar zijn geadopteerd door Europa, mogelijk ook vanuit de gedachte om met Europese standaarden minder afhankelijk te worden van de Amerikaanse internetreuzen. Deze Europese standaarden voor onafhankelijke en gecontroleerde datadeling worden ontwikkeld onder de noemer International Data Spaces (IDS). De architectuur is vastgelegd in de IDS Reference Architecture Model (RAM)²⁵. Om de driehoek naar interoperabiliteit compleet te maken, zijn dan nog de standaarden nodig. De eerste standaard is het Dataspace protocol²⁶, en heeft inmiddels een open source referentie-applicatie²⁷. Hiermee begint Data Spaces dat jaren een concept is geweest, nu richting een implementeerbare oplossing te bewegen.

In Nederland werkt de overheid onder de noemer Federatief Datastelsel (FDS) aan een soortgelijk doel van data interoperabiliteit. Hier ligt wat meer nadruk op het 'federatieve'; de data bij de bron houden en daar bevragen in plaats van data centraal opslaan of rondpompen tussen overheidsorganisaties. Redelijke kans dat in de realisatie van het FDS de IDS standaarden als middel worden ingezet. In principe kan het FDS dan gezien worden als een Data Space in het domein van de overheid.

Maar hoe zit het met dan met Linked Data? Het is de verwachting dat semantische interoperabiliteit in Data Spaces met Linked Data (bijvoorbeeld domein vocabulaires) wordt ingevuld.

2.1.7 Data Governance

Toen (begin 21e eeuw) de semantische standaarden in opmars kwamen, was de kwaliteit van die standaarden een uitdaging. Daar is een proefschrift over geschreven²⁸. Die kwaliteit werd beïnvloed door de manier waarop de standaard werd ontwikkeld en beheerd: de governance. De semantische standaarden hadden als kenmerk dat ze buiten de formele standaardisatie organisaties (zoals NEN, CEN, ISO) werden ontwikkeld, maar dichterbij het domein/sector zelf. Soms was de branche-organisatie betrokken, regelmatig werd er een nieuwe organisatie opgetuigd voor het in beheer nemen van de standaarden voor de specifieke domeintoepassing.

25 <https://internationaldataspaces.org/offers/reference-architecture/>

26 <https://internationaldataspaces.org/offers/dataspace-protocol/>

27 <https://projects.eclipse.org/projects/technology.dataspace-protocol-base>

28 <https://research.utwente.nl/en/publications/quality-of-semantic-standards>

Dat was aanleiding om een werkgroep te starten voor het delen van kennis, en heeft geleid tot de BOMOS (Beheer en Ontwikkelmodel voor Open Standaarden) publicatie; in eerste instantie een collectie van best practices hoe governance van semantische standaarden in verschillende domeinen was georganiseerd. De kern van BOMOS wordt gevormd door het activiteiten model: activiteiten die ingevuld moeten worden om een standaard te kunnen ontwikkelen en beheren.



Figuur 4 - BOMOS activiteiten diagram²⁹

BOMOS kan inmiddels zelf als standaard worden beschouwd voor het beheer van standaarden. Maar lijkt ook prima bruikbaar voor onder andere data governance, het beheer van datastelsels, of open source software.

²⁹ <https://gitdocumentatie.logius.nl/publicatie/bomos/fundament/>

2.2 DATA SCIENCE

2.2.1 Wat is Data Science?

Data Science kan het best worden omschreven als de wetenschappelijke discipline die zich bezighoudt met alle aspecten van data. Een abstracte definitie van Data Science is³⁰:

“Data Science is de studie van de generaliseerbare extractie van kennis uit data.”

Typische werkzaamheden van een Data Scientist omvatten het verkrijgen, opslaan, opschonen, analyseren en visualiseren van data, evenals het doen van voorspellingen op basis van deze data³¹. Deze voorspellingen worden onder andere gegenereerd met behulp van AI, waarbij menselijke intelligentie wordt nagebootst via machine learning-algoritmes.

Een concretere definitie, waarin ook de relatie tussen Data Science en AI in terug komt, is van IBM³²:

“Data Science combineert wiskunde en statistiek, gespecialiseerde programmering, geavanceerde analyses, kunstmatige intelligentie (AI) en machine learning met specifieke vakinhoudelijke expertise om bruikbare inzichten bloot te leggen die verborgen liggen in de data van een organisatie.”

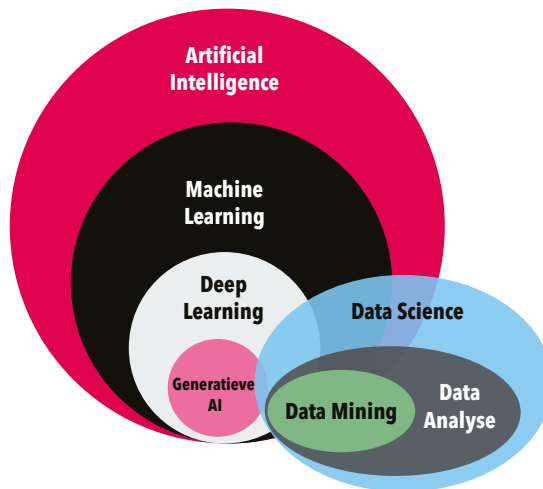
30 <https://dl.acm.org/doi/10.1145/2500499>

31 <https://doi.org/10.3390/electronics10030318>

32 <https://www.ibm.com/topics/data-science>

Nauw verwante termen aan Data Science zijn Data Analyse en Data Mining, met als belangrijkste verschil de scope van de term: Data Mining is het proces van het ontdekken van patronen en inzichten in een (grote) dataset³³. Data Analyse gaat verder in het verkrijgen van inzichten die betrekking hebben op de huidige stand van zaken. Data Science gaat weer verder om ook toekomstige situaties te voorspellen en te sturen.

Figuur 5 schetst een globaal schematisch overzicht van hoe AI en Data Science zich tot elkaar verhouden.

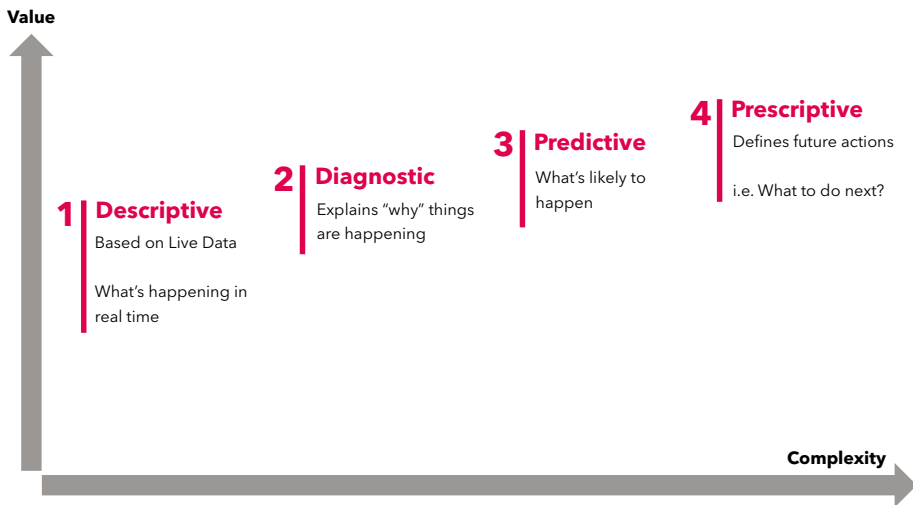


Figuur 5 - Verhoudingen vakgebieden AI en Data Science³⁴

De typen analyses worden veelal ingedeeld in vier categorieën: descriptive, diagnostic, predictive, en prescriptive (zie Figuur 6). Descriptive analyse gebruikt data om te beschrijven wat er precies is gebeurd, terwijl diagnostic analyse uitlegt waarom iets gebeurt. Beide analyses worden doorgaans uitgevoerd door een Data Analyst en vallen onder Business Intelligence (BI). Een Data Scientist gaat verder door toekomstige gevallen te voorspellen op basis van data, ook wel predictive analyse genoemd. Bij dit type analyse is vaak een voorspelmodel het eindproduct. Zo'n model verwerkt de data volgens een vast patroon en produceert zo een voorspelling. Prescriptive analyse bouwt daarop verder door voor te schrijven wat een persoon of organisatie in de toekomst zou moeten doen.

33 <https://www.ibm.com/topics/data-mining>

34 Aangepast op basis van <https://doi.org/10.3390/electronics10030318>



Figuur 6 - De vier type data analyses³⁵

Data Science is een veelzijdige discipline omdat het verschillende vakgebieden combineert, zoals wiskunde, computer science, statistiek met specifieke domeinkennis (zoals gezondheidszorg, automotive of landbouw). Daarnaast zijn de werkzaamheden ook zeer divers. Berucht is daarnaast de 80-20 regel³⁶: een Data Scientist houdt zich in de praktijk meer bezig met het verkrijgen en opschonen van data (80% Data Engineering) en spendeert maar 20% van de tijd aan de 'Science'; de analyse uitdaging. Vandaar ook de opkomst van de rol van Data Engineer, een all-rounder in het verzamelen, verwerken en opslaan van data.

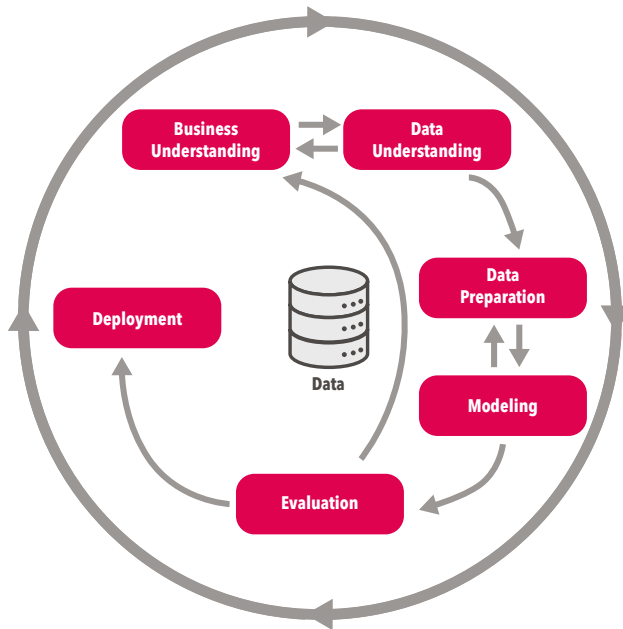
2.2.2 Het CRISP-DM model

Het meest populaire procesmodel binnen het Data Science vakgebied is, met afstand, het Cross-Industry Standard Process for Data Mining (CRISP-DM) model. Het voordeel van het model is dat het universeel inzetbaar is omdat het zich niet op een bepaalde industrie, datatype of voorspelling focust. Het proces bestaat uit de volgende zes fases die meerdere keren doorlopen kunnen worden, afhankelijk van de kwaliteit van de voorspellingen.

35 <https://medium.com/co-learning-lounge/types-of-data-analytics-descriptive-diagnostic-predictive-prescriptive-922654ce8f8f>

36 https://www.researchgate.net/publication/339550606_Playing_the_whole_game_A_data_collection_and_analysis_exercise_with_Google_Calendar

De verschillende fases zijn: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, en Deployment (zie Figuur 7). In de volgende sectie zullen we het gehele model³⁷ en alle stappen doorlopen.



Figuur 7 - Het CRISP-DM model

Business Understanding

Deze fase draait om het verkrijgen van een begrip van de zakelijke context en doelstellingen van het project. Allereerst moeten de behoeften van de organisatie begrepen worden om deze te kunnen vertalen naar specifieke doelstellingen van het voorspelmodel. Deze fase kan pas succesvol afgerond worden als er concrete en relevante doelen voor het project zijn geformuleerd.

Data Understanding

In deze fase wordt de focus gelegd op het verzamelen en verkennen van de beschikbare data. Het doel van de fase is om de eigenschappen van de beschikbare data te ontdekken, inclusief patronen, kwaliteitsproblemen en mogelijke uitdagingen.

³⁷ <https://doi.org/10.1016/j.procs.2021.01.199>

De data understanding fase staat in het teken van een Exploratory Data Analysis (EDA). Tijdens een EDA worden statistische technieken, code en visualisaties gebruikt om de belangrijkste kenmerken van een dataset te onderzoeken. EDA helpt bij het identificeren van patronen, trends, uitschieters en ontbrekende waarden, wat cruciaal is voor het begrijpen van de data voordat verdere analyses of modellering kunnen worden uitgevoerd. EDA is belangrijk om de data zo goed mogelijk te kunnen gebruiken maar ook om de bruikbaarheid te beoordelen.

Stel je bijvoorbeeld voor dat je het ras van een hond wil voorspellen op basis van een foto. Als alleen de foto's van labradors een watermerk bevatten van een bepaalde fotograaf zal het algoritme dat watermerk gebruiken om labradors te onderscheiden. Dat is natuurlijk niet wenselijk omdat deze watermerken in de praktijk niet op de foto's staan. Om zulke fouten te voorkomen is een EDA van groot belang.

Data Preparation

Tijdens de data preparation wordt de ruwe data voorbereid voor analyse en het trainen van algoritmes. Dit omvat het opschonen van de data door het invullen of verwijderen van ontbrekende waarden en het verminderen van ruis of meetfouten. Ook kan het nodig zijn om de data te transformeren, bijvoorbeeld door normalisatie of aggregatie toe te passen. Het doel is om een goed gestructureerde dataset te creëren die geschikt is voor het trainen van een algoritme.

Modeling

In de modellering fase worden vaak meerdere modellen toegepast om patronen en trends in de data te identificeren. Verschillende modellen kunnen worden getraind op de data om voorspellingen te doen. Een model wordt gekozen op basis van onder andere de complexiteit van de data, het type taak, vereiste snelheid en natuurlijk de nauwkeurigheid van de voorspellingen.

Evaluation

De evaluatiefase richt zich op het beoordelen van de modellen die zijn ontwikkeld tijdens de modeling fase. Het doel is om te controleren of de modellen voldoen aan de oorspronkelijke zakelijke doelstellingen en om hun prestaties te beoordelen op basis van objectieve criteria. Deze evaluatie helpt bij het selecteren van het meest geschikte model voor verdere inzet en geeft een indicatie van hoe

goed het model zal presenteren in de praktijk. Dit is belangrijk om te weten voor de eigenaren en gebruikers van het model, zij weten zo hoe goed ze kunnen vertrouwen op het model.

Deployment

Tenslotte, in de deployment fase, worden de resultaten en inzichten van het proces geïmplementeerd in de operationele omgeving van de organisatie. Dit kan variëren van het presenteren van rapporten en visualisaties tot het integreren van voorspellende modellen in operationele systemen. Het doel is om de waarde van de analyses en modellen te benutten en bij te dragen aan betere besluitvorming en operationele efficiëntie.

2.3 MACHINE LEARNING

Je hebt waarschijnlijk al het vermoeden dat AI tegenwoordig niet meer regel voor regel geprogrammeerd wordt, zoals het geval was bij de expert systemen. Het zou immers een onmogelijke uitdaging zijn om elk antwoord van ChatGPT van tevoren te bedenken en te programmeren. Onder de motorkap van moderne AI draaien daarom diverse machine learning (ML) algoritmes. Deze algoritmes kunnen menselijke taken overnemen, zoals die van een vertaler of chatoperator, maar ze kunnen ook de kans op een bepaalde ziekte voorspellen op basis van een DNA-profiel.

ML-algoritmes worden getraind met behulp van datasets en leren hierdoor om complexe taken uit te voeren zonder expliciete instructies voor elke mogelijke situatie. Machine learning kan ook taken uitvoeren die tot dusver niet door mensen uitgevoerd werden. In essentie gaat het om het ontdekken van patronen in grote hoeveelheden data. Deze patronen worden gebruikt om voorspellingen te doen of beslissingen te nemen. De definitie van machine learning is daarom samen te vatten in de volgende zin³⁸:

“Machine learning is een vakgebied dat computeralgoritmes laat leren zonder dat je ze expliciet hoeft te programmeren.”

38 <https://doi.org/10.1147/rd.33.0210>

2.3.1 Algoritmes en modellen

Een ML-algoritme is een set regels of instructies die zijn ontworpen om een bepaald type probleem op te lossen. Voorbeelden zijn lineaire regressie, beslissingsbomen en neurale netwerken. Een model in machine learning is daarentegen de specifieke representatie die is geleerd van data door een algoritme toe te passen. Het is de uitkomst van het trainingsproces en bevat patronen die uit de data zijn geleerd³⁹.

De overgang van een algoritme naar een model is een cruciaal proces in machine learning. Het begint met het selecteren van een geschikt algoritme op basis van de aard van het probleem en de beschikbare data. Vervolgens leert het algoritme van de data via een proces dat training wordt genoemd. Dit leren omvat het aanpassen van de parameters van het algoritme totdat het nauwkeurige voorspellingen of beslissingen kan maken. Het resultaat van dit getrainde algoritme is een model, dat vervolgens kan worden gebruikt om voorspellingen te doen op basis van nieuwe, ongeziene data.

Algoritmes zijn logisch op te delen aan de hand van de manier van leren: supervised, unsupervised en reinforcement. Het verschil tussen deze methodes is hoe en of de algoritmes feedback ontvangen tijdens het trainen. Bij supervised learning leert het algoritme van de labels gekoppeld aan de dataset om zo voor nieuwe data ook deze labels te kunnen voorspellen. Bij unsupervised learning zijn er geen labels en is het doel vaak clusteren, compressie van de data of het leren van associatie regels. In het geval van reinforcement learning is er vooraf geen dataset aanwezig, de computer leert dan om te gaan met zijn omgeving door middel van beloningen. Dit zullen we nog wat nader toelichten.

Supervised learning

Bij supervised learning wordt het model getraind met behulp van een gelabelde dataset. Het model leert onder de supervisie van een expert, de persoon die de labels van de dataset gemaakt heeft. Als je een spamfilter maakt zijn de labels bijvoorbeeld 'spam' of 'geen spam'. Elke invoer (input) in de dataset heeft een bijbehorend label (output). Het model leert door de patronen tussen de input en output te herkennen.

³⁹ <https://medium.com/the-modern-scientist/understanding-the-difference-between-algorithms-and-models-in-machine-learning-71ebacd207fa>

Sommige supervised modellen worden getraind door elke keer voorspellingen te maken op basis van nieuwe voorbeelden. Het model ontvangt dan feedback op de voorspelling en weet zo hoeveel de voorspelling afwijkt van de werkelijkheid, zie het als een soort geavanceerde versie van het spelletje 'warmer kouder'. In het begin zijn de voorspellingen natuurlijk nog niet goed en moet het model nog veel leren. Na een tijdje zal het model patronen in de data ontdekken waarop het zijn voorspellingen kan baseren. Dan begrijpt het bijvoorbeeld dat woorden als 'klik op de link', 'geld verdienen', 'winnen' en 'iPhone' een indicatie van spam zijn.

Supervised learning is op te delen in twee verschillende type voorspellingen: classificatie en regressie. Classificatie gaat om het voorspellen van de categorie van een datapunt, het onderscheiden van spam en normale mail is daarom ook een voorbeeld van classificatie. Regressie daarentegen voorspelt een continue waarde (een getal), bijvoorbeeld de prijs van een huis.

Unsupervised learning

Bij unsupervised learning zijn er geen labels, en is het doel om verborgen patronen of structuren in de data te ontdekken zonder vooraf gedefinieerde antwoorden. Een veelvoorkomende vorm van unsupervised learning is het leren van associatieregels, hierbij wordt in een grote database van transacties gekeken welke items (relatief) vaak samen voorkomen. Supermarkten gebruiken associatieregels om te ontdekken welke producten samen gekocht worden. Deze informatie kan dan gebruikt worden om die producten dichterbij elkaar neer te zetten, zo zie je dat nootjes en wijn vaak naast elkaar in de schappen staan. Het voordeel van unsupervised learning is dat het ook onverwachte patronen kan vinden die wij als mens over het hoofd zouden zien. Zo vond Thomas Blischok, een manager van een retail consultancy groep, dat tussen 17:00 en 19:00 veel supermarktklanten bier en luiers kochten. Het bleek dat vooral jonge vaders rond die tijd langskwamen om boodschappen te doen⁴⁰. Behalve associatieregels zijn er ook nog andere vormen van unsupervised learning zoals clusteren, dimensiereductie en anomaliedetectie.

Reinforcement learning

Reinforcement learning is gebaseerd op een systeem van beloningen en straffen. Het model, vaak een agent genoemd, leert door interactie met een omgeving en het uitvoeren van acties. Voor elke actie ontvangt de agent feedback in de

⁴⁰ <https://www.dssresources.com/newsletters/66.php>

vorm van beloningen of straffen, en de agent probeert de totale beloning te maximaliseren. Een klassiek voorbeeld is een zelflerende robot die door een doolhof moet navigeren. De robot ontvangt beloningen voor elke stap dichterbij de uitgang en straffen voor verkeerde stappen. Op deze manier leert de robot na talloze herhalingen uiteindelijk de snelste weg naar de uitgang. Deze manier van leren zorgt ervoor dat er op voorhand geen dataset beschikbaar hoeft te zijn, de robot leert door interacties met de omgeving.

Baker (2020)⁴¹ beschrijft een experiment waarin robots getraind worden om tikkertje te spelen tegen elkaar. In eerste instantie rennen de poppetjes maar wat willekeurig rond maar al snel leren ze dat het handig is om achter de andere speler aan te rennen. Na verloop van tijd gaan de spelers obstakels gebruiken om zichzelf op te sluiten, zo kan de tikker niet bij hen komen. Uiteindelijk vinden de verstoppers elke keer nieuwe technieken waarop de tikker na verloop van tijd een oplossing bedenkt. Het hele proces heeft veel weg van Darwins evolutietheorie omdat de robots zich kunnen blijven aanpassen aan een nieuwe omgeving⁴².

Overfitting & underfitting

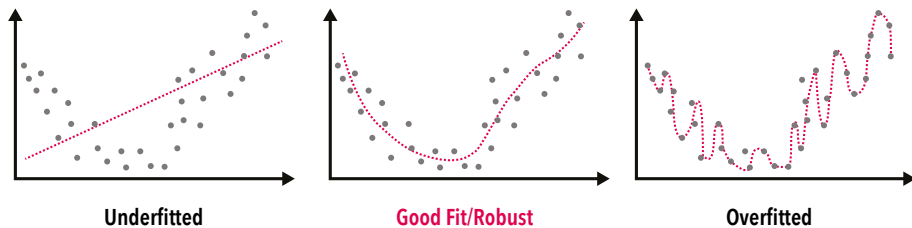
Een risico tijdens het trainen van voornamelijk supervised algoritmes is overfitting en underfitting. Beide fenomenen treden op tijdens het trainen van machine learning-modellen en hebben invloed op de prestaties en generalisatiecapaciteit van het model. In beide gevallen is de mate waarin het model kan leren van de dataset van belang, zie Figuur 8 voor een visuele weergave.

Overfitting: dit fenomeen treedt op wanneer een model te complex wordt gemaakt ten opzichte van de hoeveelheid beschikbare trainingsdata. Het model past zich zodanig nauwkeurig aan de trainingsdata aan dat het ook de ruis en de toevallige variaties in de data oppikt. Hierdoor kan het model heel goed presteren op de trainingsdata, maar slecht op nieuwe, niet eerder geziene data. Het model heeft te veel geleerd van de specifieke voorbeelden in de trainingsdata en kan het niet goed generaliseren naar andere data.

41 <https://arxiv.org/pdf/1909.07528.pdf>

42 <https://youtu.be/kopoLzvh5jY?si=h-0ltakG6xIAZsYN>

Underfitting: het tegenovergestelde van overfitting en treedt op wanneer een model te eenvoudig is om de complexiteit van de data vast te leggen, het leert te weinig van de data. Het model kan de onderliggende patronen in de data niet goed vastleggen en presteert daarom slecht, zowel op de trainingsdata als op nieuwe data.



Figuur 8 - Weergave van underfitting en overfitting, de gestippelde lijn geeft de voorspelling van het machine-learning algoritme weer

Aan de hand van een aantal eigenschappen van de taak wordt bepaald welke algoritmes geschikt zijn. Het doel van de taak (bv. voorspellen of classificeren) bepaalt het type algoritme (unsupervised, supervised, reinforcement learning) dat geschikt is. Daarnaast wordt de keuze voor het algoritme bepaald door andere eigenschappen, onder andere de complexiteit van het model (belangrijk voor het vermijden van underfitting en overfitting), de uitlegbaarheid van de voorspellingen, en de snelheid van het trainen en het maken van een voorspelling. Nadat de geschikte algoritmes zijn geïdentificeerd, kunnen de verschillende algoritmes getest worden, waarna het beste algoritme gekozen kan worden. Binnen het gebruik van het algoritme gebruik je verschillende parameters voor het tunen van het model. Ook hier worden meerdere modellen getest om uit eindelijk het best passende model bij het algoritme voor de taak te selecteren.

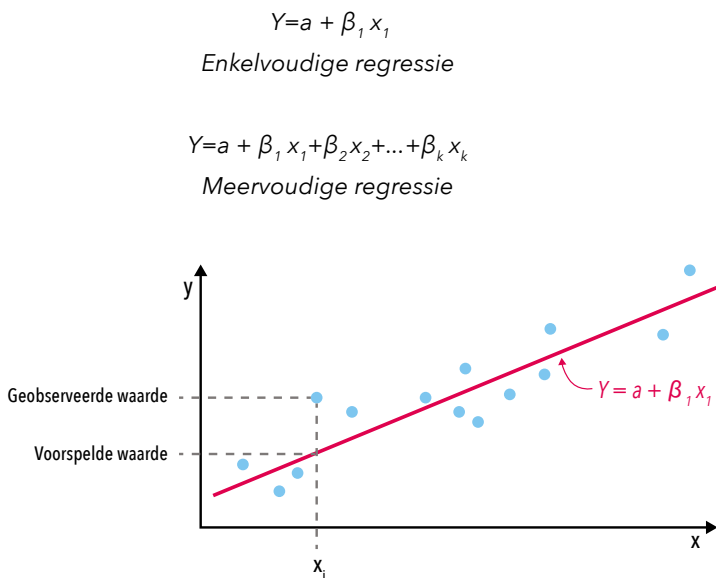
2.3.2 De algoritmes uitgelegd

Door de grote hoeveelheid en variatie in algoritmes is het onmogelijk om in dit hoofdstuk alle varianten in detail te beschrijven. Om toch een beeld te geven van hoe ML-algoritmes werken zullen we in dit hoofdstuk de meest voorkomende varianten behandelen. De meeste modellen zijn geschikt voor een specifieke taak zoals classificatie of clustering. Sommige algoritmes kunnen echter, met een aantal kleine aanpassingen, voor meerdere doeleinden ingezet worden.

Lineair & logistische regressie (supervised - regressie & classificatie)

Lineaire regressie is een bekende methode die wordt gebruikt om de relatie tussen twee of meer variabelen te begrijpen en te voorspellen. Het doel is om de rechte lijn, de lineair, te vinden die het beste past bij de relatie tussen de variabelen (zie Figuur 10). Voor het bepalen van de lineair wordt meestal de kleinste-kwadratenmethode gebruikt, welke beter bekend staat onder de Engelse naam: the least squares method. Deze methode kijkt naar het kwadraat van de afstand tussen de lineair en de datapunten, het is de bedoeling om de som van alle kwadraten te minimaliseren.

Na het bepalen van de lineair kan de regressie gebruikt worden om voorspellingen te maken. Dat gebeurt met onderstaande formules waar \hat{Y} de voorspelde waarde is, a een constante, x de waarde van de variabele en β de lineair (ook wel het gewicht genoemd). Het gewicht kan je zien als de hellingshoek van de lineair, het gewicht staat voor het verband tussen de onafhankelijke en afhankelijke variabele. Er kunnen ook meer variabelen aan de vergelijking toegevoegd worden met elk hun eigen gewichten en waardes, dit wordt een meervoudige regressie genoemd.



Figuur 9 - Illustratie van enkelvoudige regressie, de lineair is de roze lijn in het voorbeeld

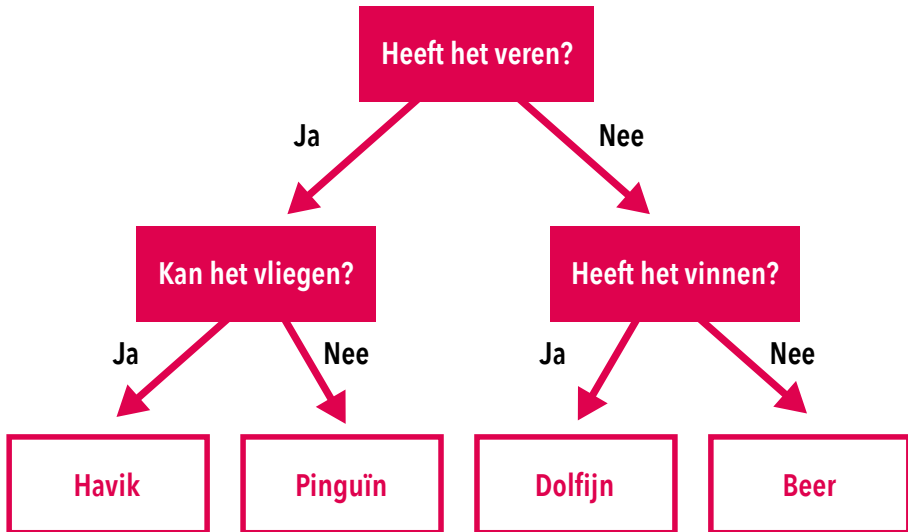
Een andere vorm van regressie is de logistische regressie. Deze vorm voorspelt de kans op een bepaalde uitkomst. Hiermee kan bijvoorbeeld de kans op een ziekte voorspeld worden op basis van verschillende factoren zoals roken, leeftijd en gewicht. In plaats van een rechte lijn zoals bij lineaire regressie, gebruikt logistische regressie een kromme lijn die een kans representeert (bijvoorbeeld de kans dat iemand ziek is). Deze kromme verandert geleidelijk en blijft altijd tussen 0 en 1, de kans op een bepaalde uitkomst kan namelijk niet boven de 100% liggen of negatief zijn. Als de uitkomst van de lineaire regressie hoog is dan kan er bijvoorbeeld een vervolgonderzoek uitgevoerd worden om te bepalen of iemand inderdaad ziek is.

Beslisbomen & random forests (supervised - regressie & classificatie)

Het concept van beslisbomen is veelal bekend: je begint bovenaan de boom en na het beantwoorden van de vragen daal je steeds verder af tot je aankomt bij je keuze. In machine learning worden ook beslisbomen gebruikt maar dan om uit te komen bij de voorspelling in de vorm van een categorie of numerieke waarde. Het concept is eenvoudig maar daardoor niet minder effectief.

In essentie heeft de ML-beslisboom dezelfde structuur als een expert systeem, in beide gevallen worden als-dan regels gebruikt om uit te komen bij een voorspelling of advies. Echter ontbreekt het zelflerende aspect bij de expert systemen, de beslisbomen daarentegen leren zelf van de dataset en worden niet regel voor regel geprogrammeerd. Het zelflerende vermogen maakt het mogelijk om met een paar regels code een volledig model met duizenden als-dan regels te trainen. Deze regels zorgen ervoor dat het model een nauwkeurige voorspelling kan maken. De beslisbomen kunnen gecombineerd worden tot een random forest. Dit zijn combinaties van een groot aantal beslisbomen waarvan de voorspellingen gecombineerd worden tot een enkele voorspelling. Dit maakt de methode veel betrouwbaarder, robuuster en minder gevoelig voor overfitting, je kan het zien als een soort 'wisdom of the crowd'⁴³.

43 <https://doi.org/10.1023/a:1010933404324>



Figuur 10 - Een eenvoudige beslisboom

Naïeve Bayes (supervised - classificatie)

Naïeve Bayes wordt gebruikt om voorspellingen te doen op basis van kansberekeningen. Het model berekent hoe waarschijnlijk het is dat een bepaalde gebeurtenis plaatsvindt, gegeven bepaalde kenmerken. Het model gaat ervan uit dat de aanwezigheid van een bepaald kenmerk in een categorie onafhankelijk is van de aanwezigheid van andere kenmerken in dezelfde categorie. Dit is een vereenvoudiging van de werkelijkheid, vandaar de naam Naïeve Bayes. Het voordeel van de vereenvoudiging is dat het de berekeningen simpeler en sneller maakt. Dat maakt dat er geen dure hardware nodig is voor het maken van een grote hoeveelheid voorspellingen.

Stel je nu voor dat je een systeem wilt bouwen dat films aanbeveelt op basis van genres en andere kenmerken. Naïeve Bayes kan helpen voorspellen of iemand een bepaalde film leuk zal vinden op basis van eerdere voorkeuren. Het model kijkt naar de kenmerken van films die een gebruiker in het verleden leuk vond, zoals genre, acteurs, of regisseur, en berekent hoe vaak deze kenmerken voorkomen in de films die de gebruiker heeft beoordeeld.

Ondanks zijn eenvoud, werkt Naïeve Bayes bijzonder goed voor veel toepassingen, zoals spamfiltering, sentimentanalyse, en andere taken waarbij je moet kiezen tussen verschillende categorieën. De goede prestaties in combinatie met de eenvoud van het model zorgen ervoor dat Naïeve Bayes een populaire keuze is.

K-means (unsupervised - clustering)

K-means is een methode voor het clusteren van data. Het verdeelt een dataset in een vooraf bepaald aantal groepen, de clusters. Dat gebeurt door gegevenspunten zo in te delen dat punten in dezelfde groep zo veel mogelijk op elkaar lijken, terwijl punten van verschillende groepen juist zo verschillend mogelijk zijn.

Het algoritme werkt iteratief: eerst worden willekeurige punten als voorlopige centra van de clusters gekozen. Vervolgens worden alle andere punten toegewezen aan het dichtstbijzijnde clustercentrum. Daarna worden opnieuw de centra berekend maar nu is het clustercentrum het gemiddelde van alle punten in het cluster. Dit proces van toewijzen en berekenen van de centra wordt herhaald totdat de centra van de clusters niet meer significant veranderen of het maximaal aantal iteraties bereikt is. Zo helpt K-means bij het ontdekken van groepen en andere patronen in de data.

DBSCAN (unsupervised - clustering)

DBSCAN is een clustermethode die verschilt van K-means omdat het aantal clusters niet vooraf opgegeven hoeft te worden. In plaats daarvan groepeert DBSCAN-punten op basis van hun dichtheid: punten die dicht bij elkaar liggen vormen samen een cluster, terwijl punten die geïsoleerd liggen worden gemarkeerd als ruis. Dit maakt DBSCAN bijzonder geschikt voor datasets met clusters van ongelijke vorm en grootte, en voor het identificeren van ruis of outliers. Het algoritme begint met een willekeurig punt en breidt een cluster uit door punten toe te voegen die binnen een bepaalde afstand liggen en voldoen aan de minimale dichtheidseis. Dit proces gaat door totdat alle punten zijn verwerkt.

Het identificeren van ruis en outliers is nuttig voor het herkennen van ongewone observaties of gebeurtenissen. Het herkennen van deze ongewone datapunten heet anomaly detection en wordt veel gebruikt om fraude op te sporen. Als je

een dataset hebt met historische transacties en daarop het DBSCAN-algoritme uitvoert zal een frauduleuze transactie buiten de clusters vallen. Zo'n transactie heeft dan bijvoorbeeld in een ander land plaatsgevonden of heeft een extreem hoog bedrag. Dit helpt bedrijven om ongewoonheden op te sporen en misbruik te voorkomen door bijvoorbeeld je creditcard te blokkeren.

Association mining (unsupervised)

Association mining is een techniek die helpt verborgen patronen en regels in gegevens te ontdekken door te analyseren hoe vaak verschillende items samen voorkomen. Deze techniek zoekt naar combinaties van items die vaak gezamenlijk verschijnen in transacties en identificeert daarmee verbanden die anders onopgemerkt zouden blijven.

Stel je voor dat een supermarktketen wil weten welke producten vaak samen worden gekocht. Association mining begint met het analyseren van alle aankopen die klanten doen en telt hoe vaak elk product samen met andere producten wordt gekocht. Vervolgens identificeert het algoritme de meest voorkomende combinaties en zet deze om in regels. Bijvoorbeeld, als blijkt dat klanten die pannenkoekenbeslag kopen ook vaak ijs kopen, dan vormt het een regel die deze associatie beschrijft, zoals 'Als pannenkoekenbeslag, dan ook ijs.'

De kracht van association mining ligt in het feit dat het patronen automatisch ontdekt zonder vooraf gedefinieerde hypothesen, waardoor organisaties nieuwe inzichten kunnen verkrijgen. Dit helpt bij het optimaliseren van productplaatsingen, het ontwikkelen van gerichte marketingstrategieën, en het verhogen van klanttevredenheid door beter in te spelen op koopgedrag. Deze techniek wordt breed toegepast, van aanbevelingssystemen tot voorraadbeheer, en biedt waardevolle inzichten voor datagedreven beslissingen.

Gradient Boosting Machines (supervised - regressie & classificatie)

Gradient Boosting Machines (GBM) worden gebruikt voor zowel classificatie- als regressietaken. Deze algoritmes, waaronder populaire varianten zoals XGBoost, LightGBM en CatBoost, bouwen een sterk voorspellingsmodel door meerdere zwakke modellen (meestal beslisbomen) achtereenvolgens te trainen. In elk iteratiestap leert het model van de fouten van de vorige bomen door gefocust te trainen op de data die eerder verkeerd werd voorspeld. Deze aanpak maakt GBM zeer nauwkeurig en efficiënt, vooral bij gestructureerde data.

XGBoost staat bekend om zijn hoge snelheid en prestaties, mede door technieken als parallelle verwerking en een slimme aanpak voor het omgaan met ontbrekende waarden. LightGBM biedt nog hogere efficiëntie door het gebruik van histogram-gebaseerde splitsingen, wat resulteert in een sneller trainingsproces en lager geheugengebruik. CatBoost onderscheidt zich door een geavanceerde verwerking van categorische variabelen, wat handmatig pre-processen overbodig maakt. Deze algoritmes worden vaak gebruikt in machine learning-wedstrijden en industriële toepassingen omdat ze heel simpel te configureren zijn, gecombineerd met een hoge nauwkeurigheid, robuustheid en mogelijkheid om complexe patronen in data te ontdekken.

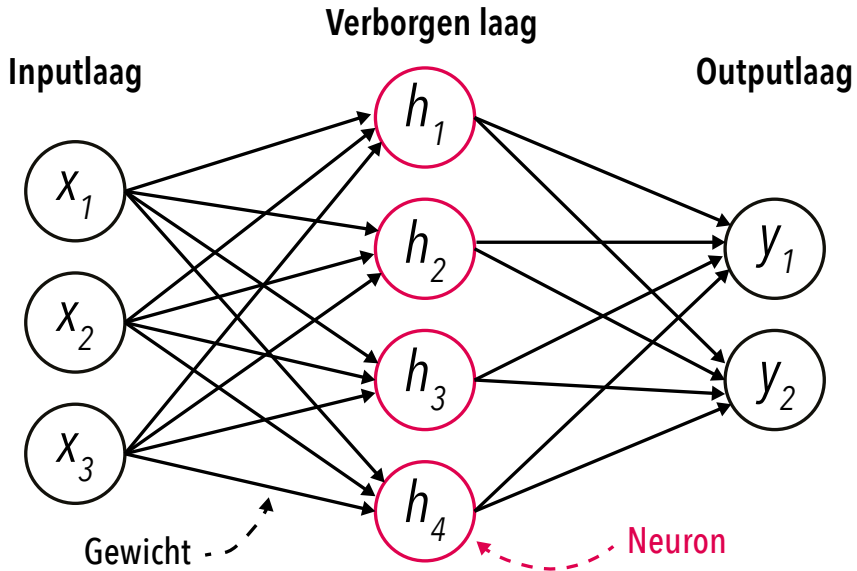
Neurale netwerken (supervised - regressie & classificatie)

Neurale netwerken (NN) zijn een veelgebruikte methode binnen de ontwikkeling van AI, die sterk geïnspireerd is door de werking van het menselijke brein. Door de structuur van het brein na te bootsen, kunnen neurale netwerken complexe patronen en relaties in data leren en herkennen. Dit maakt ze geschikt voor een breed scala aan toepassingen, zoals spraakherkenning, beeldherkenning en het spelen van spellen op menselijk niveau.

De netwerken bestaan uit verschillende type lagen: een inputlaag, de verborgen lagen en de outputlaag (zie Figuur 11). Elke laag bevat knooppunten, die neuronen worden genoemd, en deze zijn verbonden met de neuronen in de lagen ervoor en erna. Elk neuron ontvangt input, voert berekeningen uit en stuurt de resultaten door naar de volgende laag. Een typisch neural network heeft een enkele input- en outputlaag, daartussenin kunnen zich meerdere verborgen lagen bevinden.

De inputlaag van het neurale netwerk ontvangt de data, dat kan van alles zijn zoals een afbeelding, geluidsopname, sensormetingen of andere data. Deze gegevens worden verwerkt en doorgegeven aan de latere zogenaamde verborgen lagen, waar elk neuron gewichten en biaswaarden gebruikt om berekeningen uit te voeren. Het resultaat van de berekeningen wordt elke keer doorgegeven aan de volgende laag, totdat het uiteindelijk bij de outputlaag komt waar ook de uiteindelijke voorspelling uitkomt. Tijdens het trainingsproces wordt het netwerk herhaaldelijk blootgesteld aan voorbeelddata (zoals afbeeldingen van cijfers met hun bijbehorende labels). Het netwerk past zijn gewichten aan op basis van de fouten in zijn voorspellingen, een proces dat bekend staat als backpropagation.

Door deze aanpassingen leert het netwerk steeds beter hoe het voorspellingen kan doen op basis van de patronen in de data.



Figuur 11 - Een voorstelling van de werking van een kunstmatig neurale netwerk⁴⁴

Het bijzondere van een neurale netwerk is dat zelfs een ondiep neurale netwerk van twee lagen diep, in theorie, elke functie die je maar verzint kan benaderen tot op elk niveau van nauwkeurigheid. In de praktijk blijkt dat veel trainingstijd te kosten en is het sterk afhankelijk van de data, maar het blijft bijzonder dat neurale netwerken zo universeel zijn.

De term deep learning is gebaseerd op de diepte (het aantal lagen) van neurale netwerken. Diepere netwerken (tot wel honderden lagen diep) zijn complexer, maar vooral makkelijker te trainen dan de netwerken die daarvoor gebruikt werden. Daardoor kunnen ze nauwkeurigere voorspellingen maken.

De complexiteit van de netwerken betekent dat er veel rekenkracht nodig is voor het trainingsproces en het gebruik in de praktijk. Recentelijke technologische ontwikkelingen hebben ervoor gezorgd dat de benodigde rekenkracht beter beschikbaar is waardoor de populariteit van deep learning algoritmes enorm gegroeid is.

⁴⁴ <https://udlbook.github.io/udlbook/>

Een Convolutional Neural Network (CNN) is een specifiek type neuraal netwerk dat speciaal ontworpen is om visuele patronen te herkennen. Hierdoor is het bijzonder effectief in taken zoals objectdetectie, gezichtsherkenning en classificatie van beelden. Waar CNNs ontworpen zijn voor afbeeldingen, zijn RNNs (Recurrent Neural Network) ontworpen voor taal. Tegenwoordig worden CNNs ook voor andere niet visuele toepassingen gebruikt. Zo kan een CCN gebruikt worden voor spraakherkenning, de spraak (een 1-dimensionaal signaal) wordt dan omgezet naar een afbeelding (een 2-dimensionaal signaal) en dient als input voor een CNN. De CNN kan dan bijvoorbeeld de spraak van personen herkennen. In de praktijk worden NNs, CNNs en RNNs voor allerlei toepassingen gebruikt, met soms verrassend goede resultaten.

In plaats van volledig verbonden lagen, maken CNNs gebruik van convolutielagen die fungeren als filters om kenmerken zoals randen, hoeken en texturen in de inputdata te identificeren. Een convolutie is dus eigenlijk een klein filtertje dat speurt naar een specifiek patroon. Vergelijk het met het zoeken van een bekende in een mensenmassa, dan zoek je vaak naar het patroon of de kleur van iemands kleding. De filters kunnen door het netwerk hergebruikt worden op verschillende plekken in de afbeelding om essentiële visuele informatie te detecteren, waardoor het eenvoudiger wordt om complexe patronen in beelden te leren. Dit maakt CNNs zeer krachtig voor beeldgerelateerde toepassingen, waar traditionele neurale netwerken vaak tekortschieten. Door de werking van de filters en poolinglagen kunnen CNNs efficiënter omgaan met grote hoeveelheden data en zijn ze minder gevoelig voor de positie van objecten binnen een afbeelding.

Transformers (supervised - tekst)

Transformer-taalmodellen zijn een speciaal type neuraal netwerk dat ontworpen is om effectief met tekst om te gaan. Wat hen uniek maakt, is hun vermogen om langeafstandsrelaties in tekst te ontdekken, in tegenstelling tot eerdere modellen die vooral letten op de nabijheid van woorden. Een woord dat aan het begin van een zin of zelfs een paragraaf staat, kan invloed hebben op de betekenis van een woord dat veel later in de tekst voorkomt. Dit vermogen is essentieel voor taken zoals het correct toepassen van enkelvoud en meervoud of het begrijpen van complexe zinsconstructies. Dankzij de zogenaamde 'attention'-laag kunnen transformers zich 'concentreren' op belangrijke woorden en hun relaties met andere woorden in de tekst, waardoor ze de context beter begrijpen

en goed omgaan met subtiele nuances zoals cynisme, humor en ambiguïteit. De aandachtmechanismen zorgen voor meer relevante en coherente output doordat ze de focus leggen op de belangrijkste delen van de tekst.

Transformers zijn uniek doordat ze uitsluitend gebruik maken van 'attention' tijdens het verwerken van tekst, zonder de noodzaak van andere architecturen zoals convoluties of recursies. De verwerking binnen transformer-taalmodellen houdt in dat de inputtekst wordt omgezet in wiskundige representaties die zowel de context als de betekenis van de woorden bevatten. Dit gebeurt door embeddings, numerieke representaties van woorden langs honderden dimensies van betekenis, gecombineerd met het attention-mechanisme dat bepaalt welke woorden belangrijk zijn in relatie tot andere woorden. Deze interne representaties worden vervolgens gebruikt om outputtekst te genereren, bijvoorbeeld bij het vertalen van zinnen of bij tekstgeneratie.

Een van de grote voordelen van transformers is hun schaalbaarheid: ze kunnen efficiënt worden getraind op grote hoeveelheden data en hun prestaties verbeteren naarmate ze meer gegevens en complexere patronen leren. De training van transformers gebeurt vaak door woorden uit een zin weg te laten en het model te laten voorspellen welke woorden er ontbreken, een proces dat eenvoudig te automatiseren is omdat het juiste antwoord altijd in de oorspronkelijke tekst te vinden is. Dit eenvoudige trainingsprincipe heeft geleid tot de ontwikkeling van geavanceerde taalmodellen zoals GPT (Generative Pre-trained Transformer) en BERT, die veel worden toegepast in bijvoorbeeld chatbots, automatische vertalingen en tekstgeneratie.

De introductie van transformers heeft een revolutie teweeggebracht in de wereld van generatieve AI, door de brede toepassing van taalmodellen mogelijk te maken in allerlei gebieden voor taken die eerder moeilijk waren met traditionele AI-methoden. Hierdoor vormen transformers een belangrijke pijler van de hedendaagse AI-technologieën.

2.3.3 Evalueren met metrics

Voor een specifieke taak is het vaak lastig om op voorhand te bepalen welk model (algoritme) het beste zal presteren. Sommige modellen kunnen sowieso niet gebruikt worden voor een bepaald type taak, zo is het niet mogelijk om lineaire

regressie te gebruiken voor classificatie. Voor andere modellen is de kans groot dat ze niet bruikbaar zijn omdat ze in vergelijkbare gevallen niet hoog scoren. Zo zou je bijvoorbeeld geen beslisboom inzetten om geschreven tekst te herkennen. Na het identificeren van de geschikte modellen worden deze met elkaar vergeleken. Daarvoor worden ze eerst getraind met een deel van de beschikbare data: de trainingsset. Na het trainen moeten de modellen op een ander deel van de data voorspellingen doen. De kwaliteit van de voorspellingen van de verschillende modellen wordt met elkaar vergeleken om zo het beste model te kunnen kiezen. De kwaliteit van de voorspellingen wordt behaald aan de hand van kwantitatieve evaluaties, zogenaamde metrics en scores.

Classificatie

Voor classificatiemodellen worden verschillende veelgebruikte metrics berekend op basis van de verhoudingen in een confusion matrix (Figuur 12). Een confusion matrix bestaat uit vier verschillende getallen die een indicatie geven over de voorspellingen van het model ten opzichte van de werkelijke categorieën.

De true positives (TP) en true negatives (TN) zijn correcte voorspellingen waarbij, respectievelijk, een positief geval ook als zodanig voorspeld wordt of er een negatieve voorspelling gedaan wordt voor een negatief geval. False positives (FP), ook wel type I fouten genoemd, zijn gevallen waarin het model onterecht een positieve uitkomst voorspelt. Er wordt bijvoorbeeld voorspeld dat een patiënt ziek is terwijl deze in werkelijkheid gezond is. False negatives (FN), ook wel type II fouten genoemd, zijn gevallen waarin het model onterecht een negatieve uitkomst voorspelt. Dat komt overeen met voorspellen van geen ziekte bij een zieke patiënt.

		Actual Values	
		Positive (1)	Negative (0)
Predicted values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Figuur 12 - Een confusion matrix voor het evalueren van een classificatiemodel⁴⁵

De meest bekende metrics voor classificatie modellen zijn nauwkeurigheid, precisie, recall en de F1-score. Voor alle metrics geldt dat een hogere waarde staat voor betere kwaliteit van de voorspellingen. Nauwkeurigheid is het percentage correcte voorspellingen ten opzichte van het totaal aantal voorspellingen. Precisie daarentegen meet het aandeel correcte positieve voorspellingen van alle positieve voorspellingen. Precisie zegt dus iets over de kans dat een positieve voorspelling in werkelijkheid ook een positief geval is. Recall meet het aandeel correcte positieve voorspellingen van alle werkelijke positieve gevallen, een lage recall betekent dus dat een model veel positieve gevallen gemist heeft. Afhankelijk van de toepassing wordt er meer waarde gehecht aan een hoge recall of precisie. Een hoge recall is bijvoorbeeld belangrijk bij het ontdekken van storingen in een vliegtuig, er mag dan geen enkel geval gemist worden. Zowel recall als precisie werden tijdens de coronapandemie in 2020 plotseling van groot belang voor een breed publiek. Ze beantwoorden namelijk vragen zoals: hoe groot is de kans dat je onnodig in quarantaine gaat na een positieve test, en hoe groot is de kans dat ik toch besmet ben na een negatieve test?

⁴⁵ <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

Voor veel toepassingen zijn beide scores van belang, in dat geval is de F1-score een goede keuze. Deze score is een harmonisch gemiddelde van precisie en recall, de exacte formule voor de F1-score kan je vinden in het onderstaande overzicht. De F1-score ligt altijd tussen 0 en 1, waarbij een hogere score dus staat voor betere prestaties.

Of de waarde van een metric goed is hangt af van de taak en de data en is daarom altijd relatief, het is dan ook niet mogelijk om te zeggen welke absolute waarden goed en slecht zijn. De metrics moeten altijd gebruikt worden voor de vergelijking van de prestaties van algoritmes met dezelfde taak en data.

$$\text{Nauwkeurigheid} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precisie} = \frac{TP+TN}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \times \text{Precisie} \times \text{Recall}}{\text{Precisie} + \text{Recall}}$$

Regressie

Voor regressiemodellen moeten andere metrics gebruik worden omdat er continue waarden voorspeld worden in plaats van categorieën. De meest gebruikte metrics voor regressiemodellen zijn de mean absolute error (MAE), mean squared error (MSE), en R-kwadraat.

De MAE meet de gemiddelde absolute afwijkingen tussen voorspelde en werkelijke waarden. Er wordt gekozen voor absolute waarden omdat het in veel gevallen niet zoveel uitmaakt of een voorspelling hoger of lager is dan de werkelijke waarde. In de eerste plaats gaat het vaak over de grootte van de fout.

De MSE lijkt op de MAE en staat voor de gemiddelde kwadratische afwijkingen tussen de voorspelling en de werkelijke waardes. Door het gebruik van de kwadratische afwijkingen hebben grote verschillen meer invloed op de score dan kleine verschillen. Doordat het kwadraat van de fout altijd positief is gaat het hier ook over de grootte van de fout. Voor zowel de MAE en de MSE is een lagere score beter, als alle gevallen perfect voorspelt zijn zullen beide scores op 0 uitkomen. Dit gebeurt echter maar zelden in de praktijk. Een andere veelgebruikte score is R-kwadraat, het geeft aan hoeveel van de variantie in de afhankelijke variabele wordt verklaard door het model. De mogelijke waardes liggen tussen de 0 en de 1, waarbij juist weer hogere waardes goed zijn.

Ook bij MAE, MSE en R-kwadraat is er geen absolute goede of foute waarde. De metrics moeten alleen gebruikt worden om algoritmes te vergelijken. Behalve de genoemde metrics zijn er nog veel meer varianten beschikbaar. De gebruikte metrics hangen sterk samen met de taak en het domein waarin het voorspelmodel opereert. Zo zijn er speciale metrics die gebruikt worden voor afbeeldingen en tekst. Toch hebben alle metrics hetzelfde doel: de prestaties van voorspelmodellen vergelijken.

2.4 SPECIFIEKE VORMEN VAN AI

Uiteraard hebben we in dit deel slechts een beperkt beeld van Data & AI kunnen neerzetten. In deze laatste paragraaf zullen we nog een aantal specifieke vormen van AI benoemen die nu (of in de nabije toekomst) in de spotlights (komen te) staan.

Natural Language Processing (NLP)

Natural Language Processing (NLP) combineert statistische technieken met machine learning technieken. Met inzet van NLP halen we slim en automatisch informatie uit omvangrijke en ongestructureerde tekstuele data. We halen bijvoorbeeld eenvoudig kernwoorden uit een tekst.

Large Language Models (LLMs)

Large Language Models (LLMs) zijn in de basis minder 'slim' dan NLP-technieken omdat ze geen echt begrip van taal hebben; ze begrijpen geen betekenis zoals mensen dat doen. In plaats daarvan zijn Large Language Models gebaseerd op

het toepassen van deep learning op enorme hoeveelheden trainingsdata, in combinatie met een transformerarchitectuur. Deze architectuur stelt het model in staat om patronen in de data te herkennen en het volgende woord in een zin te voorspellen, wat de kern is van hoe deze modellen functioneren. ChatGPT is het aansprekende voorbeeld van het toepassen van een Large Language Model om teksten en antwoorden te genereren.

Om de output van Large Language Models te verbeteren, kan gebruik worden gemaakt van Retrieval-Augmented Generation (RAG). Deze methode voegt extra context toe aan het Large Language Model door relevante informatie uit externe bronnen op te halen en te integreren in de gegenereerde tekst. Hierdoor kan de kwaliteit van de gegenereerde antwoorden aanzienlijk worden verhoogd, omdat het model niet alleen vertrouwt op zijn trainingsdata, maar ook op actuele en specifieke informatie die relevant is voor de vraag of taak.

Generatieve AI

Generatieve AI is de inzet van AI-technologie om nieuwe content te creëren, zoals tekst, afbeeldingen, muziek, en meer. In het geval van tekst gebruikt de generatieve AI applicatie (zoals ChatGPT) vaak Large Language Models (LLMs) in combinatie met Natural Language Processing (NLP) om op basis van natuurlijke taal (ook wel prompt (engineering) genoemd) coherente en contextuele teksten te produceren. Maar de mogelijkheden van generatieve AI reiken verder dan alleen taal; het kan ook visuele content, zoals afbeeldingen en video's, muzikale composities, en softwarecode genereren. Door gebruik te maken van verschillende neurale netwerken, zoals transformers, kan generatieve AI creatieve output leveren die steeds moeilijker te onderscheiden is van menselijke creaties.

Quantum Machine Learning (QML)

Zoals de term al aangeeft wordt met Quantum Machine Learning (QML) de ideeën uit Quantum Computing en Machine Learning gecombineerd. Modellen in QML maken gebruik van concepten uit de Quantum Mechanica als superpositie om data te verwerken. Dat zorgt ervoor dat deze modellen unieke dingen kunnen die niet mogelijk zijn op een klassieke computer. Daarnaast ligt er een grote kans bij het reduceren van de trainingstijd ten opzichte van klassieke machine learning modellen. Hiermee zouden zowel meer grootschalige toepassingen mogelijk worden, maar ook meer real-time toepassingen.

In ruime zin is er onderscheid te maken tussen twee verschillende vormen van quantum computers: quantum annealing chips en gated chips. De eerste vorm is met name geschikt voor het oplossen van klassieke optimalisatieproblemen, maar kan dat beter dan een klassieke computer. Daarvoor gebruikt het een proces genaamd quantum annealing, waarbij een quantum model door verschillende suboptimale oplossingen voor het klassieke probleem kan 'tunnelen'. Het tweede type maakt gebruik van logische operaties op qubits, quantum versies van bits. Deze vorm van quantum computing is algemener, en maakt gebruik van verstrengeling en superpositie om problemen op een manier op te lossen die niet mogelijk is op een klassieke computer. Denk bijvoorbeeld aan het razendsnel ontbinden van een getal in priemgetallen. Toepassingen op gated quantum chips zijn momenteel nog gelimiteerd, omdat het moeilijk is om de quantum deeltjes lang genoeg geïsoleerd te houden van hun omgeving om gecompliceerde berekeningen te kunnen doen.

Neuromorphic Computing

Een grote uitdaging bij het gebruik van AI is het energieverbruik. Volgens een rapport in The Economic Times kan één jaar ChatGPT genoeg energie verbruiken om Nieuw-Zeeland drie maanden van stroom te voorzien en Nigeria vier maanden. Er wordt ook gesteld dat een enkele ChatGPT-zoekopdracht 2,9 wattuur verbruikt, bijna 10 keer meer dan een Google-zoekopdracht (0,3 wattuur)⁴⁶. En dan hebben we het nog niet over het stroomverbruik van het trainen van modellen. Daartegenover is het menselijk brein een zeer efficiënte computer. Er zijn inschattingen dat ons brein ongeveer 20 watt verbruikt aan energie en daarvoor krijg je een neuraal netwerk met ongeveer 86 miljard neuronen, dat in staat is om een exaFlop (dat is een 1 met 18 nullen) operaties aan berekeningen per seconde kan doen⁴⁷. Met Neuromorphic Computing wordt getracht een stap verder te zetten in het nabootsen van het menselijke brein met als doel om AI chips bouwen die energie-efficiënter zijn dan hedendaagse chips.

Een voorbeeld van een essentieel verschil tussen het brein en computers is dat bij een computer het geheugen gescheiden is van de processor. Dat betekent dat een computer eerst de relevante informatie moet ophalen uit de opslag en naar de processor moet transporteren, alvorens er berekeningen mee te kunnen doen. De snelheid waarmee die informatie ingelezen kan worden, is gelimiteerd. Daarentegen bestaat het menselijk brein uit een verstrengeld netwerk van

46 <http://timesofindia.indiatimes.com/articleshow/111382705.cms>

47 https://doi.org/10.1007/978-981-16-8892-8_28



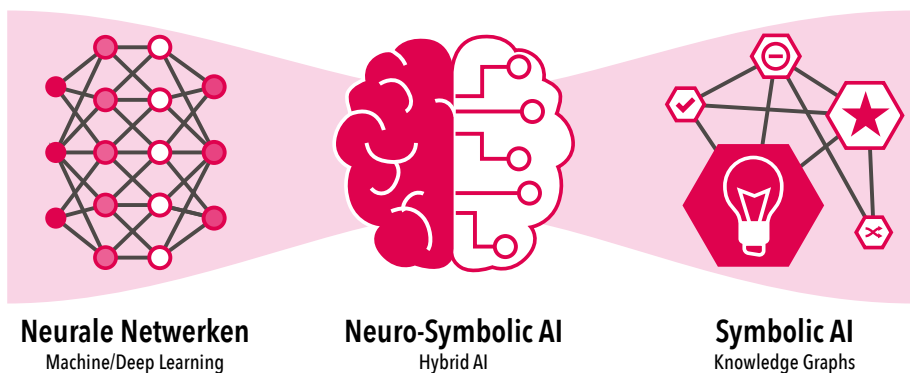


neuronen en synapsen, en wordt informatie lokaal opgeslagen op de plekken waar het ook verwerkt wordt⁴⁸. Neuromorphic computing probeert chips te maken die meer van dit soort kenmerken van het brein nabootsen, waardoor ze krachtiger en efficiënter kunnen worden.

2.5 HYBRID AI - NEURO-SYMBOLIC AI

Elke vorm van AI heeft zijn voor- en nadelen. Het basisidee van Hybrid AI is eenvoudig; het samensmelten van verschillende vormen van AI om de voordelen van bepaalde AI-vormen maximaal te benutten om de nadelen van andere bepaalde AI-vormen te vermijden. Specifiek gaat het daarbij om de voordelen van machine learning (neurale netwerken) te combineren met wereld van regels en logica (Symbolic AI); vandaar ook de concretere term Neuro-Symbolic AI.

In andere woorden: we willen data in al haar kracht kunnen gebruiken; door zowel te kunnen leren van (veel) data, maar daarbij ook alle kennis die we vastgelegd hebben kunnen gebruiken.

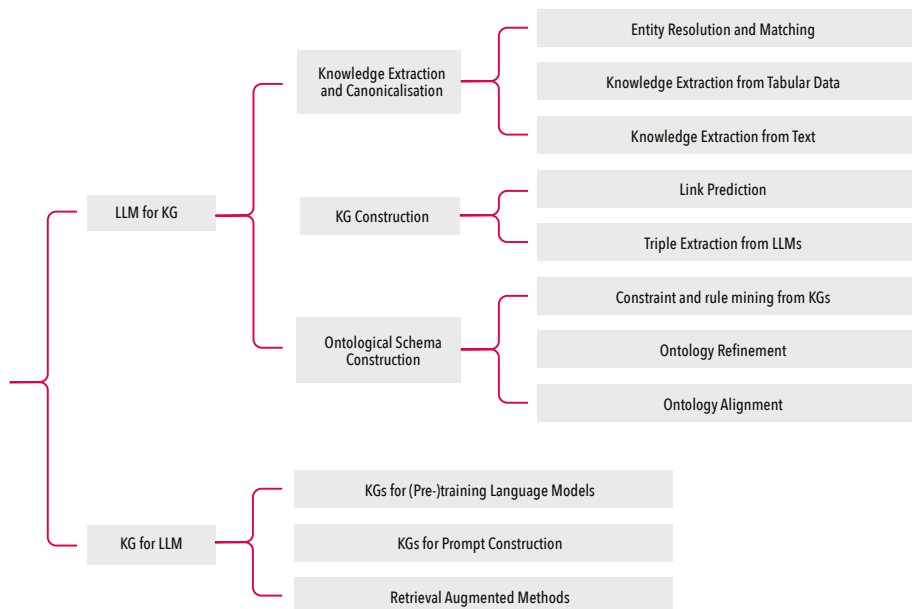


Figuur 13 - Hybrid AI (Neuro-Symbolic AI) waar Neurale Netwerken en Symbolic AI samenkomen

Dit is het beste van twee werelden, en gaat de basis worden van veel grootschalige AI implementaties waar zekerheid en betrouwbaarheid een belangrijk aspect zijn; en dat is nagenoeg voor alle situaties waarbij de maatschappelijke of economische impact groot is.

48 <https://doi.org/10.1038/s42254-020-0208-2>

De combinatie van Neuro-Symbolic AI staat op dit moment concreet in de spotlight met het combineren van Large Language Models (LLMs - Neuro) en Knowledge Graphs (KGs - Symbolic). De kracht van het verbinden werkt beide kanten uit (zie Figuur 14); LLMs kunnen gebruikt worden voor het extraheren van kennis uit data, voor het creëren of aanvullen van een Knowledge Graph of het kennismodel (ontologie). Omgekeerd kunnen KGs gebruikt worden voor het trainen van LLMs of voor het maken van prompts of het inzetten van RAG.



Figuur 14 - De combinatie van LLMs en KGs⁴⁹

⁴⁹ <https://drops.dagstuhl.de/storage/08tgdk/tgdk-vol001/tgdk-vol001-issue001/TGDK.1.1.2/TGDK.1.1.2.pdf>





DEEL 3: DATA SCIENCE & AI - DE PRAKTIJK

De praktijk beschrijf ik aan de hand van drie voorbeeldorganisaties. TNO, Kadaster en de HAN. Niet toevalligerwijs de organisaties waar ik een bijdrage aan lever, of heb mogen leveren.

3.1 TNO - DATA ESSENTIE

3.1.1 Data standaarden

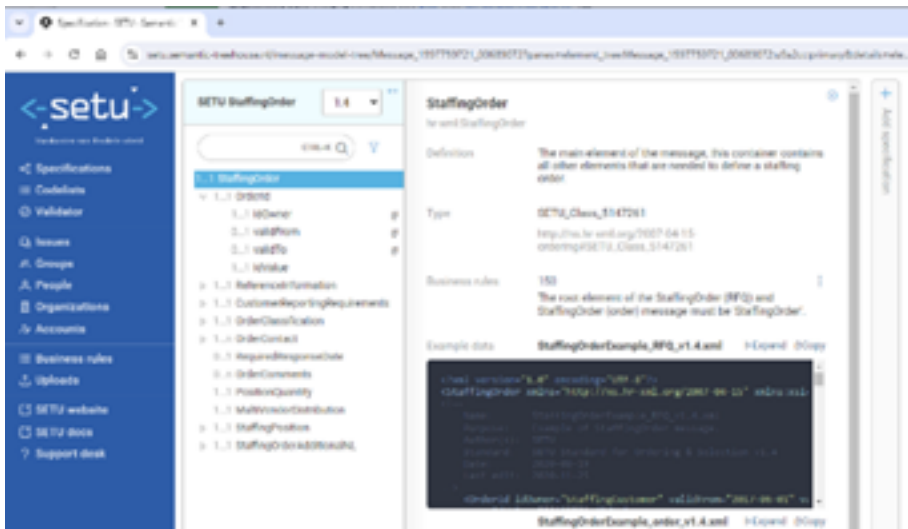
Bij TNO begonnen we rond het jaar 2000 onder de noemer e-business, het belang van interoperabiliteit (standaarden & architectuur) te benadrukken voor ICT-toepassingen. Met het nodige geluk bleken we (achteraf gezien) precies op het juiste moment een semantische data standaard voor de uitzendbranche ontwikkeld te hebben⁵⁰. Deze sector stond aan het begin van een grote automatiseringsslag (van fax naar IT), waarbij maatwerk automatiseringstrajecten tussen inlener en uitzender erg kostbaar waren. De business case voor standaardisatie was dan ook relatief eenvoudig. Deze aanpak is toegepast in vele verschillende sectoren, waarbij telkens de aanpak en tooling verder geprofessionaliseerd zijn. Een voorbeeld hiervan was het Pressure Cooker concept, waarbij een semantische standaard voor een sector in een week ontwikkeld kon worden.

De technologie is door de jaren heen flink veranderd, maar de essentie niet: het belang van het vastleggen van de betekenis van informatie (en context) met als doelstelling interoperabiliteit in een samenwerkingsketen. Nu gebruiken we geen (of minder) XML Schema meer, maar maken domein ontologieën (met OWL, SHACL, SKOS, et cetera), applicatie profielen (met JSON schema en JSON-LD) en Open API specificaties.

TNO heeft met het Semantic Treehouse⁵¹ het gehele proces van datastandaarden ontwikkelen en beheren verder geprofessionaliseerd en met tooling ondersteund.

50 www.setu.nl

51 <https://www.semantic-treehouse.nl/>



Figuur 15 - Voorbeeld Semantic Treehouse voor de SETU standaard

3.1.2 Data governance

Al vrij snel werd duidelijk dat semantische standaardisatie in essentie afspraken maken is, polderen en vastleggen. De belangen zijn daarbij groot. Bepaalde keuzes in een standaard kunnen tot hoge kosten leiden voor de ene organisatie, en lage kosten voor de andere. De business case is in de praktijk nooit eerlijk verdeeld. Dus we hebben regels nodig: Governance. We zijn toen ervaringen gaan delen, en dat resulteerde in het BOMOS-raamwerk en best practices voor de governance van datastandaarden (zie ook paragraaf 2.1.4). BOMOS blijkt in de praktijk veel gebruikt te worden, onder andere ook voor het beheer van de standaarden gerelateerd aan de omgevingswet.

3.1.3 Data Spaces

Het succes van het delen van data, het kopiëren van data, en het beschikbaar stellen van open data heeft ook een keerzijde; data gaat zijn eigen leven leiden, veelal tot ongenoegen van de eigenaar (zelfs bij open data). Dat heeft een beweging op gang gebracht waarin de essentie is dat data moet gaan stromen, maar wel op een manier die recht doet aan de eigenaar van de data. In het Engels ook wel data sovereignty genoemd. Dit is de gedachte achter de Data Spaces ontwikkeling, uiteraard daarbij ook gericht op het behalen van technische,

semantische en organisatorische interoperabiliteit op basis van open standaarden. Standaarden als het Dataspaces Protocol, zijn mede door de inbreng van TNO, recent beschikbaar gekomen⁵² (zie ook paragraaf 2.1.3).

3.1.4 Stoppen met data delen

Recent is er ook veel aandacht om data juist niet meer te delen, vanwege die keerzijde dat data zijn eigen leven gaat leiden. Liever geeft een data-eigenaar (tijdelijk) toegang tot de (meta)data, of dat het algoritme naar de data gebracht wordt (in plaats van andersom). Rond data in de zorg is de Personal Health Train daar een voorbeeld van⁵³.

3.1.5 Lessen geleerd bij TNO

Maar hoe de architectuur, de standaarden, en de techniek ook in de tijd veranderen: afspraken over data zijn cruciaal. Mijn belangrijkste les die ik heb geleerd bij TNO.

3.2 KADASTER DATA SCIENCE TEAM

Kadaster is een echte data-organisatie. In feite heeft het niks anders dan data: data is het kernproduct, en daarom heen zijn er taken van inwinning, registratie, publicatie maar altijd gericht op het dataproduct. Voor de buitenwereld komt het Kadaster over als een betrouwbare organisatie, wat ook past bij de taak van het bieden van (rechts)zekerheid. Misschien zelfs een beetje saai, maar saai is het achter de schermen niet; daar vinden top data innovaties plaats. Een spil in het web is het Data Science Team, een virtueel multidisciplinair team dat vooral kortlopende data en AI-innovaties oppakt. We zullen een aantal voorbeelden geven, zowel meer gericht aan de data kant als mooie voorbeelden van AI.

3.2.1 Data: De Kadaster Knowledge Graph

Kadaster heeft een lange historie met het toepassen van Linked Data. Dat is ook niet zo verwonderlijk, want voor Kadaster zijn principes als betekenis en herkomst van de data, zekerheid bieden over data, van cruciaal belang. Daarnaast om maatschappelijke uitdagingen aan te gaan is de data van het Kadaster altijd een schakel in combinatie met andere data, bijvoorbeeld met data van KvK, CBS, RVO, RWS, et cetera. Dus als Kadaster willen we graag de betekenis onderdeel van de data laten zijn, de mogelijkheid bieden om data met elkaar te verbinden (unieke

52 <https://www.tno.nl/nl/technologie-wetenschap/technologieen/international-data-spaces>

53 <https://www.dtls.nl/fair-data/personal-health-train/>

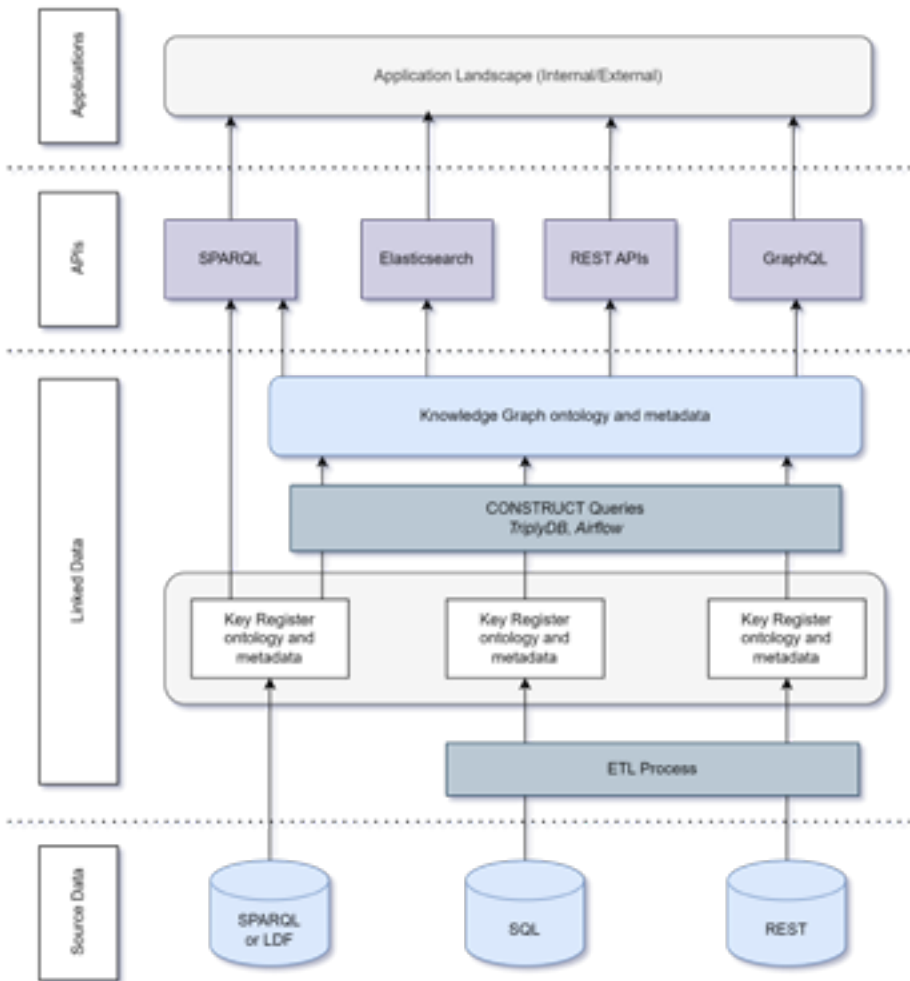
identifiers) en op basis van de open standaarden van het World Wide Web. Linked Data dus.

In verschillende iteraties door de jaren heen is hieraan gewerkt, van eerste stapjes om de Basisregistratie Adressen en Gebouwen (BAG) in RDF (standaard) te publiceren, tot aan het publiceren van wat de Kadaster Knowledge Graph is gaan heten: alle open geo-basisregistraties met elkaar verbonden en bevroagbaar⁵⁴.

Net zoals de visie op het publiceren van Linked Data is de manier waarop we Linked Data maken (de ETL - Extract Transform Load), beschikbaar stellen, en de data modelleringsaanpak in de loop van de tijd veranderd. In de eerste iteratie was de publicatie van een dataset als Linked Data gewoon een ander formaat (yet another format) naast de meer traditionele geospatiale formaten en diensten (bijvoorbeeld WFS). In de basis transformeerden we WFS-data naar een Linked Data publicatie. Omdat er geen formele relaties tussen de individuele datasets beschikbaar waren, moest de gebruiker alsnog zelf de data gaan integreren. Deze 'yet another format' aanpak resulteerde dan ook niet in het doel om de herbruikbaarheid en integraliteit van data te verbeteren. In de volgende iteraties hebben we dan ook meer nadruk gelegd op de integraliteit (verbinden) van de data, wat geleid heeft tot de Kadaster Knowledge Graph.

De architectuur en aanpak van de Kadaster Knowledge Graph kent drie fases. Allereerst, wordt de relationele bron data (bijvoorbeeld de Basisregistratie Adressen en Gebouwen) geladen in een PostgreSQL database nadat een Geography Markup Language (GML) indexeer stap is uitgevoerd. Deze data is dan benaderbaar via een GraphQL API, waarmee het model wordt uitgebreid en gevalideerd om er zeker van te zijn dat we accuraat en efficiënt queries kunnen afvuren. De focus is om de data zo dicht mogelijk bij de bron te laten om accuraatheid en actualiteit te behouden zodat de Linked Data publicatie de bron data reflecteert zonder onnodige transformaties.

54 <https://data.kkg.kadaster.nl/>



Figuur 16 - Architectuur Kadaster Knowledge Graph

Daarna, als tweede nadat het model beschikbaar is, hebben we een Enhancer component ontwikkeld waarmee we de data met vooraf gedefinieerde queries (met tijd en paginering parameters) kunnen bevragen, en als resultaat data in JSON-LD formaat krijgen.

Vervolgens wordt SHACL (Shapes Constraint Language, W3C standaard) gebruikt om zowel het datamodel als de getransformeerde instantie data te valideren, om data integriteit te kunnen garanderen. Het gehele ETL (Extract Transform Load) proces is met minimale code geschreven en wordt beheerd vanuit Apache Airflow. De gevalideerde data wordt geladen in een TriplyDB triplestore, met voor elke dataset een eigen SPARQL endpoint.

Als derde en laatste stap wordt de Kadaster Knowledge Graph gecreëerd door het geïntegreerde data model als laag te leggen bovenop de individuele datasets. We implementeren dat door elke onderliggende dataset te bevragen met een SPARQL CONSTRUCT query waarmee de data conform de mapping wordt getransformeerd van het dataset model naar het SOR (Samenhangende ObjectenRegistratie) model. Deze queries genereren nieuwe data (triples), en worden als nieuwe triples, individuele graphs per feature in het model, geladen in de triplestore. Dan hebben we een Knowledge Graph die vervolgens benaderbaar is met REST APIs, GraphQL, ElasticSearch, en uiteraard (Geo)SPARQL.

Recent is deze architectuur onder de loep genomen om beter aan te kunnen sluiten bij de Enterprise Architectuur van Kadaster, en om de overgang van de Kadaster Knowledge Graph naar een formele dienst van Kadaster te versnellen. Hiervoor is de eerste stap van de ETL aangepast naar een Azure Databricks omgeving. Hier zijn alle relationele databronnen beschikbaar in de Databricks catalogus, met voor elke dataset tabellen die de individuele eigenschappen bevatten. De kolomnamen in deze tabellen worden gerelateerd aan de klassen en eigenschappen in de ontologie van elke dataset. Python scripts worden uitgevoerd in de Databricks Spark omgeving voor de transformatie op basis van RML of R2RML mappings waarin de kolomnamen zijn gerelateerd aan de ontologie. Dit proces is zeer goed schaalbaar, en genereert miljoenen triples in een paar uur door gebruik te maken van de Morph-KGC library. De output triples worden dan geladen in de triplestore (TriplyDB) en zijn dan benaderbaar met onder andere SPARQL.

Speciale aandacht verdienen de linksets, de verbindingen tussen de datasets, die gedefinieerd worden als relationele view en geconverteerd worden met RML, en met SPARQL queries worden de triples gegenereerd waarmee een volledige Kadaster Knowledge Graph beschikbaar komt. Naast de technische aanpassingen in de ETL, is ook het model aangepast naar het nieuwe gestandaardiseerde IMX-Geomodel.

De Kadaster Knowledge Graph bevat 1.7 miljoen triples en wordt op dit moment elk kwartaal geupdate. Ongeveer 30 data stories zijn gemaakt die de Kadaster Knowledge Graph gebruiken voor verschillende toepassingen, zoals ruimtelijke plannen. Technisch gezien staat de Kadaster Knowledge Graph als een huis, maar organisatorisch is er nog werk te verzetten. De governance voor de linksets is bijvoorbeeld nog niet ingeregeld. Afspraken moeten gemaakt worden over eigenaarschap, hoe de links te maken, hoe vaak te updaten, et cetera. Ondanks de noodzaak voor de linksets in grote overheidsontwikkelingen zoals het Federatieve Data Stelsel (FDS), is dit nog niet ingeregeld.

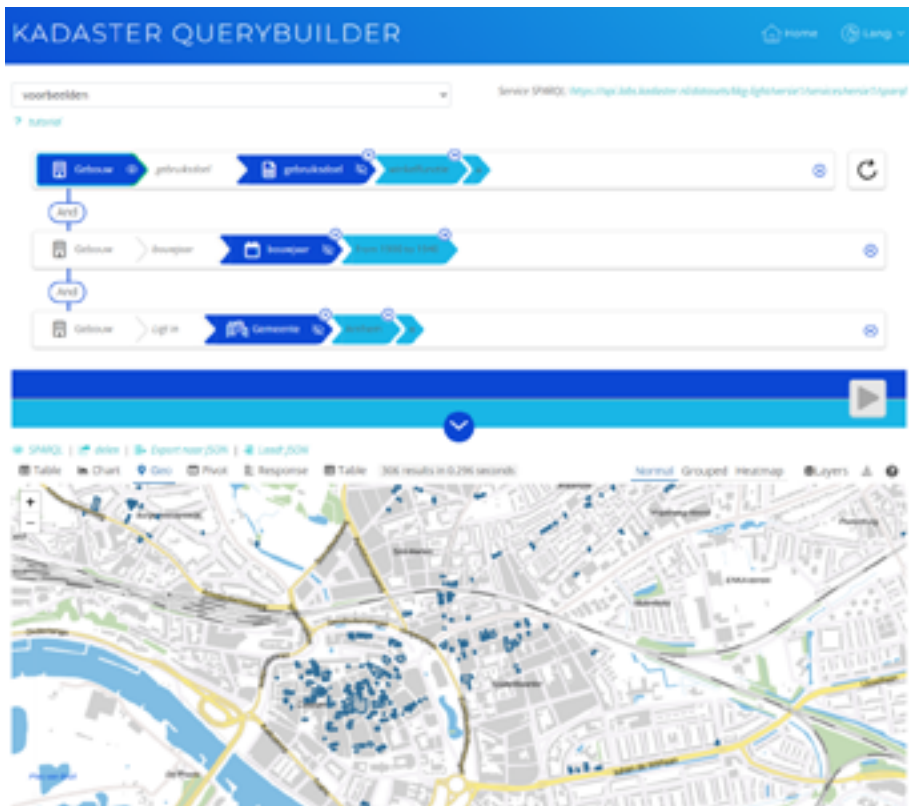
3.2.2 Data: Toepassingen op de Kadaster Knowledge Graph

De Kadaster Knowledge Graph is een huzarenstukje van formaat, maar in feite is het 'slechts' het Datafundament. De proof of the pudding moet komen op twee manieren:

1. Dat andere organisaties dit ook op deze manier gaan doen;
2. De toepassingen.

Het is mooi dat we een rijk datafundament hebben; nog mooier als het nog rijker wordt met data van andere organisaties. Maar uiteindelijk zijn het de toepassingen die de waarde creatie maken.

Een achilleshiel daarbij is het gebruik van SPARQL; het schrijven van een SPARQL query is niet voor iedereen weggelegd; en daarmee is de Kadaster Knowledge Graph maar voor een beperkte groep SPARQL-experts bruikbaar. Gelukkig zijn daar mooie oplossingen voor. Allereerst bestaan er tools waarmee op basis van het datamodel een grafische interface gemaakt kan worden waarmee het maken van een SPARQL query wordt versimpeld tot het klikken in een datamodel (De Kadaster Querybuilder in Figuur 17). De doelgroep wordt hiermee vergroot tot iedereen die niet vies is van een beetje begrip van datamodellen.



Figuur 17 - Grafisch queries maken⁵⁵

Maar dat is nog steeds niet 'iedereen', maar ook daarmee worden we goed geholpen door de ontwikkelingen van ChatGPT. 'Iedereen' (of in ieder geval de grote massa) is het normaal gaan vinden om aan een chatbot in natuurlijke taal vragen te gaan stellen. Naast de bekende teksten, kan hiermee code genereerd worden, en waarom dan niet een SPARQL query? Loki is de experimentele chatbot die dat mogelijk maakt, en getraind is met SPARQL voorbeeld queries. In natuurlijke taal kan de gebruiker aan Loki vragen stellen zoals 'Wat is het bouwjaar van mijn huis?' of 'Hoeveel woningen staan er in mijn straat?'; De chatbot Loki bevat niet zelf antwoorden (in tegenstelling tot ChatGPT), maar kan wel de vraag omzetten naar een SPARQL query, deze live uitvoeren, en het antwoord tonen. Hiermee voorkomen we hallucinaties, of verkeerde en verouderde antwoorden. De query gebruikt de actuele data bij de Kadaster bronnen.

55 <https://labs.kadaster.nl/demonstrators/querybuilder/>



Figuur 18- De chatbot Loki (screenshot)⁵⁶

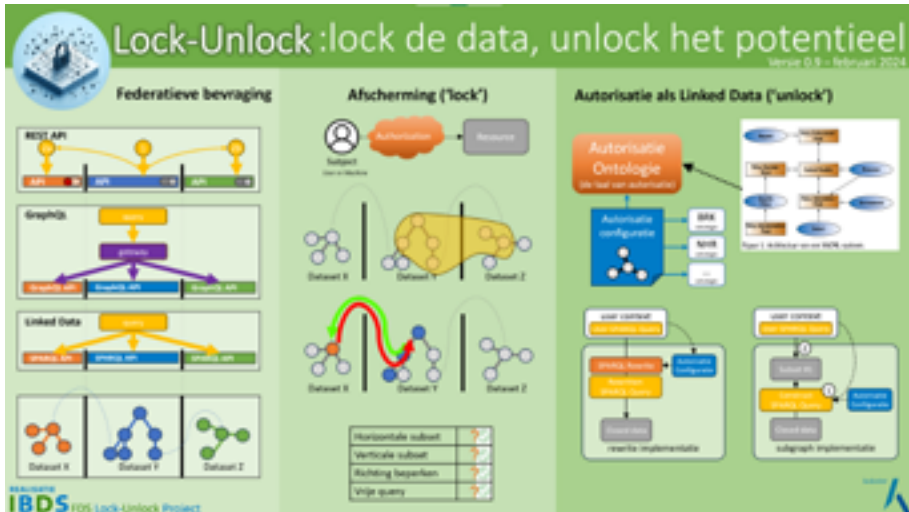
3.2.3 Data: Lock-Unlock: Lock de data & Unlock het potentieel

De Kadaster Knowledge Graph verbindt open geo-data en de Linked Data standaarden werken ook goed op open data. Echter niet alle data k n open zijn. Redenen van privacy en economische waarde kunnen noodzaak geven om data af te schermen. Is deze data dan onder voorwaarden toch te ontsluiten?

In het Lock-Unlock onderzoeksproject⁵⁷ is de doelstelling om het potentieel van data te ‘unlocken’, maar wel met de mogelijkheid tot een ‘lock’ op bepaalde data, of bepaalde combinaties/uitsneden van data. Met andere woorden het integraal verbinden van afgeschermd (Linked) data.

⁵⁶ <https://labs.kadaster.nl/cases/lokiv3>

⁵⁷ <https://labs.kadaster.nl/cases/lockunlock-project>



Figuur 19 - Samenvatting Lock-Unlock

Tijdens het onderzoek kwamen we erachter dat er weinig gestandaardiseerde mogelijkheden voor autorisatie van data in het Linked Data domein zijn, zeker wanneer je binnen een federatief datastelsel werkt. Daarom is er tijdens het project een autorisatie ontologie ontwikkeld (in Linked Data) waarmee verschillende autorisatiepatronen zoals horizontale en verticale subsets gemodelleerd kunnen worden.

Vervolgens hebben we in een proof-of-concept twee implementaties getest die beide gebruikmaken van deze ontologie. Bij de ene implementatie wordt de inkomende SPARQL query op basis van de autorisatie ontologie en configuratie automatisch herschreven naar een query waarin beperkingen zijn toegevoegd. Bij de andere implementatie wordt er op basis van dezelfde gegevens een subgraph gegenereerd met toegestane gegevens waar vervolgens de oorspronkelijke query van de gebruiker op wordt uitgevoerd.

Een van de belangrijkste conclusies van het onderzoek is dat het mogelijk is om fijnmazig autorisatieregels declaratief te modelleren op basis van een autorisatie ontologie voor federatieve bevragingen. We hebben dit aan kunnen tonen in onze

demonstrators wat laat zien dat het mogelijk is om een SPARQL-endpoint op te zetten die alleen informatie teruggeeft waarvoor je de benodigde rechten hebt. Er is nog wel aanvullend onderzoek nodig en de beweging naar een standaard voor de autorisatie ontologie voordat dit standaard toepasbaar is.

3.2.4 AI: Voorspelmodellen

Voorspelmodellen is een veel gebruikte toepassing van AI. Als er trainingsdata beschikbaar zijn, dan is in de regel in relatief weinig tijd een voorspelmodel te maken; Dit is wijd inzetbaar, waardoor er vele voorbeelden zijn.

Bouwjaren voorspellen. De gemeente Amsterdam gebruikte 1005 als default bouwjaar waarde als het bouwjaar onbekend was (de registratie stond lege waardes niet toe). Het bleek goed te doen om voor die panden een betrouwbaarder bouwjaar te voorspellen op basis van data over de omliggende panden. Hiervoor is een random forest model getraind op de verschillende features zoals oppervlakte, bouwjaar en postcode. Wel bleek het organisatorisch (afwijking van het vastgestelde werkproces) uitdagend om een door AI voorspelde waarde op te nemen in de basisregistratie. Maar gelukkig, met enige vertraging zijn de voorspelde waardes nu onderdeel van de Basisregistratie Adressen en Gebouwen.



Figuur 20 - Voorspelde (missende) bouwjaren voor panden in Amsterdam⁵⁸

58 <https://labs.kadaster.nl/demonstrators/bagdemonstrator/>

Bouwlagen voorspellen: Op welke verdieping zit een bepaald adres? Voor dit probleem is een Gradiënt Boosting regressiemodel ontwikkeld op basis van de kenmerken: huisnummer, oppervlakte van het verblijfsobject (vbo), oppervlakte van het pand, aantal vbo's in het pand en de totale oppervlakte van de vbo's in het pand. De labels/annotaties binnen de trainingsdata zijn afkomstig uit appartementsomschrijvingen van het Kadaster Objecten- en Rechtenregistratie Systeem (KOERS).

Voorspellingen zijn gedaan voor verdiepingen tot en met de 19e verdieping. Voor hogere verdiepingen is alles '20+' gelabeld. 53% van de voorspellingen kwamen exact overeen met de annotaties, en in 82% van de gevallen zat de voorspelling er maximaal 2 verdiepingen naast.

Bouwlagen voorspellen: Hoeveel bouwlagen heeft een pand? Er is gekozen voor een XGBoost regressiemodel, dat op basis van geometrische kenmerken zoals hoogte, volume, muuroppervlakte en dak-type uit de 3DBag (Algemeen Hoogtebestand Nederland (AHN)) voorspellingen doet. Andere kenmerken zijn CBS-features uit Kerncijfers Wijken en Buurten 2021, zoals bevolkingsdichtheid, aanwezigheid van horeca in de buurt en het percentage meergezinswoningen. Ten slotte is er informatie gebruikt uit de Basisregistratie Adressen en Gebouwen (BAG), zoals het bouwjaar, aantal aangrenzende gebouwen, aantal hoekpunten van de gebouwgeometrie, gebruiksdoel, aantal vbo's en oppervlakte van de vbo's. De labels binnen de trainingsdata zijn afkomstig uit de gemeenten Amsterdam, Den Haag en Rotterdam, en zijn dus erg gericht op de Randstad.

Voorspellingen zijn gedaan voor elk BAG pand in Nederland. De resultaten zijn opgedeeld in subsets van gebouwen van 1-5 en 6+ verdiepingen. Dit is gedaan om te laten zien dat de kwaliteit van de voorspellingen afneemt voor gebouwen met veel verdiepingen, iets wat we in Nederland niet veel hebben. We zien dat in ongeveer 90,5% van de gevallen het aantal bouwlagen correct wordt voorspeld. Voorheen werden deze voorspellingen puur gebaseerd op geometrie (hoogte gebouw/verwachte verdiepingshoogte) en dit was in minder dan 70% van de gevallen correct.

Het resultaat is ook opgeslagen in een dataset en beschikbaar gemaakt als Linked Data. De bijbehorende data story⁵⁹ bevat een aantal mooie queries om bijvoorbeeld de panden te vinden in een stad met de meeste bouwlagen.

Leerpunt is hierbij dat we met een eenvoudig model uitgevoerd in zeer beperkte tijd al een aardige precisie en recall hebben. Hogere precisie en recall zijn heel goed mogelijk door meer data en/of een andere aanpak te kiezen. Dit leidt ook tot hogere kosten. Maar hoe bepaal je de kosten/baten afweging voor de precisie en recall? Helemaal omdat de baten bij andere organisaties liggen.

Voorspellen van graafschade: Graafschades zijn in Nederland een groot probleem. De Wet informatie-uitwisseling bovengrondse en ondergrondse netten en netwerken (WIBON) heeft als doel het gevaar of economische schade door beschadiging van ondergrondse kabels of leidingen (zoals bijvoorbeeld: water-, elektriciteit-, gas- en telecomleidingen) te voorkomen. In de praktijk is vooral de KLIC-melding bekend die een graver moet doen voordat hij in de grond gaat graven. Er is veel inzicht gecreëerd, maar helaas nog niet minder graafschade, mede ook doordat het aantal graafbewegingen hard gegroeid is. De vraag is simpel: bij het doen van een KLIC-melding, kunnen we dan een inschatting maken van de kans op graafschade. Het antwoord is ook simpel: Ja, dat kunnen we.

Het voorspelmodel bevat een XGBoost machine learning model dat voorspellingen kan doen op basis van een KLIC-melding. In het onderzoek zijn voorspelmodellen getest op basis van XGBoost, LightGBM and CatBoost modellen, waarbij we de modellen geëvalueerd hebben op de accuraatheid, interpreteerbaarheid en efficiency. Het XGBoost model haalt een AUC-ROC score (een alternatief voor F1-score) van 0.829 en een balanced accuracy van 0.947, waarbij de output (voorspelling) in 2 seconden beschikbaar is. Deze gradient boosting models zijn explainable doordat de eigenschappen van de data die het meeste invloed hebben in de voorspellende waarde bekend zijn (in dit geval de aanwezigheid van bomen). Voor het trainen van het model is drie jaar van graafdata en graafschadedata gebruikt aangevuld met onder ander type landgebruik, bodemsoort en bomendichtheid.

⁵⁹ <https://data.labs.kadaster.nl/dst/-/stories/inzichten-bouwlagen>

Een interessante les was de dataverzameling. De aanname was dat locatie-informatie van de kabels en leidingen een belangrijke voorspellende waarde zou hebben. Die locatie-informatie is eigendom van de netbeheerders, die we voor drie gemeentes om toestemming hebben gevraagd; dat ging moeizaam en was een langdurig traject. Ironisch genoeg bleek uiteindelijk die data weinig effect te hebben op het voorspelmodel. De grootste impact bleek het te hebben op de adoptie; zonder locatie-informatie hebben de potentiële gebruikers minder vertrouwen in het model.

Voorspellen van werkaanbod: Het Kadaster probeert het werkaanbod van akten zo nauwkeurig mogelijk te voorspellen, zodat de planning van werkzaamheden hierop afgestemd kan worden. Tot 2023 werd die voorspelling uitgevoerd met Excelmodellen. Het Data Science Team heeft onderzocht of de voorspelling verbeterd kan worden met behulp van AI. 'Kunnen we het aanbod aan aktes accuraat en real-time voorspellen om zo de bemensing goed te plannen?'

De verwerking van aktes is grotendeels geautomatiseerd (ook deels met AI) maar de wat complexere aktes worden altijd nog handmatig verwerkt door juridische medewerkers die hier speciaal voor zijn opgeleid. 'Voorheen konden we vijf dagen vooruit voorspellen met Excel. Met AI kunnen we minimaal vijf weken vooruit voorspellen met een veel lagere afwijking.'

Tijdens het onderzoek zijn eerst de huidige Excelmodellen nagebouwd en zijn de patronen in de data herkend waarna er een zelflerend AI model is ontwikkeld. Er is onderzoek gedaan naar AI-architecturen zoals SARIMAX, LSTM en transformer modellen. Voor de beste resultaten hebben we een neurale netwerk o.b.v. een time2vec en LSTM-architectuur gekozen. Gegeven genoeg data kan het model de trend leren voorspellen zonder handmatige stappen en hierin steeds nauwkeuriger worden. Het toevoegen van marktscenario's maakt het model nog krachtiger. Bijvoorbeeld op basis van externe factoren zoals de rente-ontwikkeling en het BBP wordt de voorspelling automatisch door het model bijgesteld.



Figuur 21 - Voorspelling van werkaanbod voor de komende 30 dagen⁶⁰

3.2.5 AI: Detecteren

Naast de voorgaande voorbeelden gericht op het voorspellen van een fenomeen, wordt AI ook veelvuldig gebruikt voor het detecteren van objecten.

Detecteren van ondergrondse parkeergarages:

Veiligheidsregio's, met name de brandweer, willen graag weten waar de ondergrondse parkeergarages zich bevinden onder andere vanwege de mogelijkheid dat daar elektrische laadpalen staan, die extra risico opleveren bij brand.

Binnen het Kadaster hebben we een dataset samengesteld met behulp van de Basisregistratie Topografie (BRT) en OpenStreetMap om locaties van zowel privé- als publieke parkeergarages in kaart te brengen. Deze dataset, bestaande uit ± 2000 unieke afbeeldingen, combineert panoramabeelden en gerichte uitsneden van parkeergarages. Vervolgens hebben we deze afbeeldingen geannoteerd door de bounding boxes die de ingangen van parkeergarages weergeven, te markeren op Cyclorama streetview afbeeldingen. Dit resulteerde in een uitgebreide set van gekoppelde originele beelden en annotaties.

⁶⁰ <https://labs.kadaster.nl/cases/voorspelmodellen>

Om deze gegevens te gebruiken voor automatische detectie, hebben we een machine learning-algoritme getraind. We kozen voor een YOLO-v8 model, dat is getraind op de samengestelde dataset. Hierbij hebben we gebruik gemaakt van data-augmentatie om de variëteit van de trainingsset te vergroten, door willekeurige aanpassingen zoals helderheidsveranderingen, RGB-verschuivingen en het toevoegen van ruis. Na het trainen werd het model gevalideerd op een aparte set, waardoor het nu in staat is om nauwkeurig de ingangen van parkeergarages te identificeren in diverse afbeeldingen, zoals panoramafoto's.

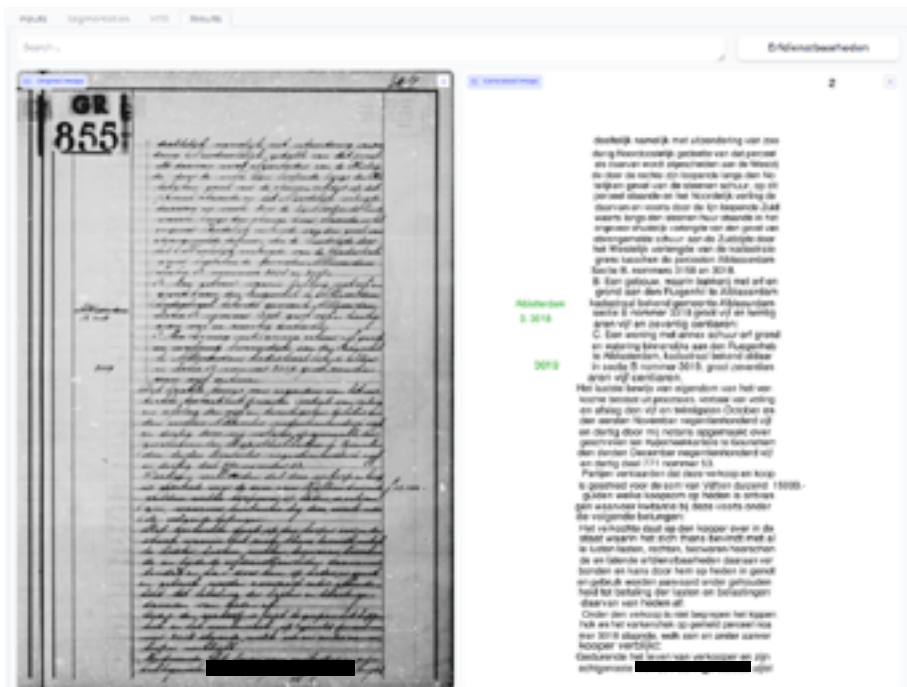


Figuur 22 - Detectie van een ondergrondse parkeergarage op cyclorama

Detectoren; het digitaal doorzoekbaar maken van handgeschreven aktes:

Afgelopen jaar heeft het Kadaster onderzoek gedaan om de handgeschreven aktes uit het archief machine-leesbaar te maken en erfdiensbaarheden daarin automatisch te detecteren. Deze aktes, die dateren van 1838 tot 1950, worden regelmatig gebruikt voor onderzoek, maar zijn momenteel niet doorzoekbaar. Dit maakt het onderzoek tijdrovend, lastig en duur aangezien elke akte handmatig moet worden geopend en gelezen. Het Kadaster beheert naar schatting 5-10 miljoen van deze handgeschreven aktes. Door deze aktes om te zetten naar doorzoekbare tekstbestanden, kunnen onderzoeken efficiënter worden uitgevoerd, wat de kans op fouten verkleint en de operationele efficiëntie verhoogt. Een uitdaging is het gegeven dat Kadaster akten in zwart-wit en in lage kwaliteit opgeslagen zijn. Hierdoor lukt het niet om met off-the-shelf tooling goede transcripties te maken.

Om de aktes digitaal doorzoekbaar te maken is daarom een pipeline van afzonderlijk trainbare AI-modellen gebouwd. Hierin worden de regels gedetecteerd, en wordt per regel de tekst voorspeld. Als laatst wordt de klasse van de tekst voorspeld, zodat er bijvoorbeeld gefilterd kan worden op marginalia en paginanummers. De algoritmes zijn o.a. getraind op Kadaster akten, en begrijpen deze dus optimaal. Er wordt een gemiddelde Character Error Rate gehaald van 5.3%, wat betekent dat 5.3% van de karakters foutief wordt voorspeld door de algoritmes. De pipeline blijkt een zeer waardevolle tool voor de kadaster medewerkers: momenteel worden de modellen getest door de medewerkers die de erfdienstbaarheidsonderzoeken uitvoeren.



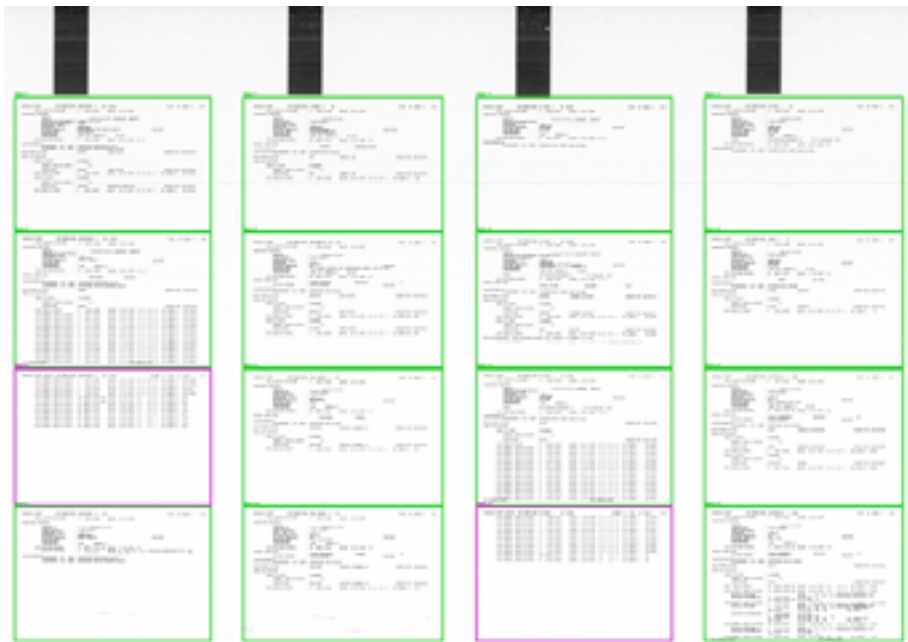
Figuur 23 - Gedigitaliseerde weergave van een handgeschreven akte waarin gezocht kan worden

Detecteren in objectkaarten archief:

Het proces van het extraheren van objectkaarten uit microfilm strips is van groot belang voor het digitaliseren en toegankelijk maken van historische gegevens. Deze microfilms, die vroeger werden gebruikt vanwege de lagere opslagkosten, bevatten rijen en kolommen van objectkaarten die nu gedigitaliseerd moeten worden. Een techniek zoals objectdetectie met behulp van het YOLO-v8 maakt dit

mogelijk. Bij het Kadaster wordt dit model onder andere nog meer ingezet voor de detectie van recreatiewoningen en Top10NL (kaart) objecten. De handmatige annotatie van de microfilm strips, leidde tot een verdeling van de data in training, test, en validatie sets, met respectievelijk 60%, 20%, en 20% van de data.

Bij de extractie stuiten wij op verschillende uitdagingen, zoals het detecteren van objectkaarten die over meerdere pagina's lopen of maar gedeeltelijk op een uitsnede vallen. Dit werd opgelost door het gebruik van overlappende uitsnedes. Na detectie worden overlappende objectkaarten samengevoegd en geordend op basis van rij en kolom. Deze methode resulteert in een nauwkeurigheid van 99% (F1-score) op de validatieset. De uiteindelijke uitsnedes van de objectkaarten kunnen hierna verder worden verwerkt door deze op te nemen in het digitale archief en door de inhoud doorzoekbaar te maken met behulp van OCR (Optical Character Recognition).



Figuur 24 - Detectie van afzonderlijke objectkaarten op microfilms

3.2.5 Lessen geleerd bij het Data Science Team

Les 1: In de praktijk blijkt de data van de basisregistraties veel beter op orde dan de data voor onze ondersteunende processen. In die AI-projecten ontstond vaak vertraging omdat de data niet direct beschikbaar was (de 80-20 regel). Eigenlijk is dat ook wel een beetje goed nieuws voor de basisregistraties.

In het toepassen van AI voor detectiemodellen is ook de uitdaging om goede geannoteerde data te krijgen (voorbeelddata waarop bijvoorbeeld ondergrondse parkeergarages zijn aangegeven). Het is moeilijker om daar budget voor te krijgen terwijl in de praktijk goede (en veel) inputdata leidt tot veel betere resultaten.

Les 2: Er kunnen zeer veel toepassingen ook in brede zin gerealiseerd worden. De (experimentele) toepassingen bij het Kadaster vinden op alle terreinen plaats. Voor het inwinnen van data, voor data kwaliteit, de mogelijkheden van analyse op die data. Maar ook de inzet van AI voor HR/Resource Planning, Finance & Control taken, Marketing, Klant Contact, et cetera.

Les 3: Het hoeft allemaal niet zo complex en duur te zijn; als je team (kennis/kunde/spirit) maar goed is (en data beschikbaar is). Een eerste prototype is veelal met een aantal dagen inzet al mogelijk.

Les 4: Het is lastig om een succesvol AI-experiment te continueren in een productie-omgeving. Dit is vooral een organisatorische uitdaging.

3.3 HAN LECTORAAT ADSAI

Het lectoraat ADSAI (Applied Data Science & AI) is in oprichting, maar toch zijn er al mooie casussen. Bij de HAN werken we bijvoorbeeld aan:

3.3.1 Predictive maintenance: CHANGE

CHANGE is een onderzoeksproject dat zich richt op het gebruik van data voor innovatie met vrachtwagentrailers. Het project focust op de inzet van extra elektrische aandrijving op de trailer en de ontwikkeling van predictive maintenance, met speciale aandacht voor banden en remmen. De keuze voor de trailer is bewust gemaakt, omdat in het verleden de meeste innovaties hebben plaatsgevonden in het trekkend voertuig. Hierdoor ligt de grootste potentie voor verdere optimalisatie en winst nu bij de trailer.

De inzet van elektrische aandrijving op de trailer kan leiden tot aanzienlijke brandstofbesparingen. Om de aandrijving zo efficiënt mogelijk te benutten, is data nodig over onder andere de versnelling, snelheid en het toerental van de vrachtwagen. Momenteel wordt hiervoor data vanuit het trekkende voertuig gebruikt om te bepalen wanneer de elektromotor in de trailer moet worden ingeschakeld of opgeladen. In de praktijk worden trailers echter door een grote verscheidenheid aan voertuigen getrokken, waardoor het telkens verbinden van de trailer met de trekker onpraktisch is. Het doel van het onderzoek is daarom om de inzet van de elektromotor optimaal te benutten met de data afkomstig van (bestaande) sensoren op de trailer zelf, zonder afhankelijk te zijn van gegevens uit het trekkend voertuig.

Het idee achter predictive maintenance is dat door onderhoud te voorspellen, plotselinge defecten onderweg kunnen worden voorkomen. Dit verhoogt de verkeersveiligheid door het verkleinen van de kans op ongelukken en voorkomt dat vrachtwagens onverwachts stil komen te staan, doordat onderhoud tijdig plaatsvindt. Dit maakt het transport betrouwbaarder en bespaart veel tijd en kosten. Hoewel predictive maintenance al langer bestaat in de transportsector, hebben de meeste ontwikkelingen zich tot nu toe gericht op het trekkend voertuig.

Voor de ontwikkeling van predictive maintenance zijn data en machine learning essentieel. Unsupervised machine learning-algoritmes kunnen worden ingezet om afwijkingen van het normale gedrag van een trailer te detecteren, wat kan wijzen op versleten onderdelen. Zo worden restricted Boltzmann machines (RBM) ingezet om de data te comprimeren en vervolgens weer te genereren. Als er veel afwijking is tussen de input en de output van de RBM wordt er gesproken van een afwijking van het normale patroon, ook wel een anomalie genoemd. Verder worden supervised algoritmes ingezet om patronen uit historische (sensor)data te leren. Een long short-term memory algoritme (LSTM) kan bijvoorbeeld variaties in wielsnelheden analyseren om versleten banden te identificeren. Ons lectoraat ADSAI richt zich op het verzamelen, verwerken en analyseren van data om zo effectieve algoritmes voor predictive maintenance te ontwikkelen.

3.3.2 Voorspellen: Future Factory

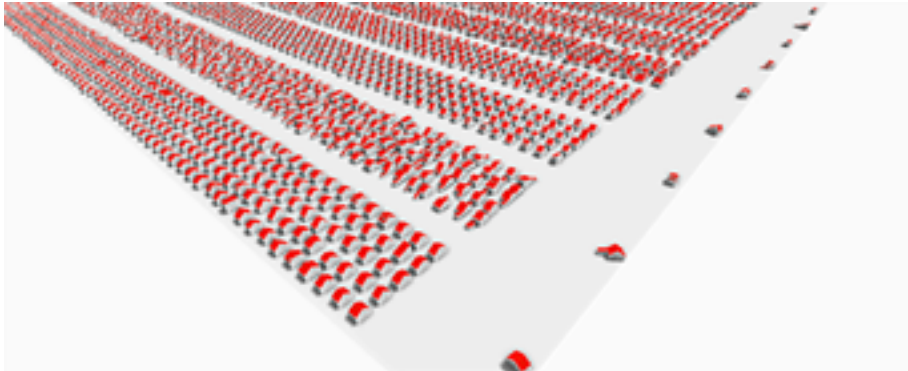
Het Future Factory project richt zich op het opschalen van de verduurzaming van woningen en wooncomplexen in Nederland. Een activiteit is onder andere

het identificeren van geschikte woningen voor renovatie op basis van openbare gegevens. ADSAI ontwikkelt hiervoor in samenwerking met bouwkundige onderzoekers van de Hogeschool Utrecht een prototype. Er wordt een pijplijn ontwikkeld die automatisch geschikte woningen identificeert aan de hand van data uit de 3DBAG en het Kadaster. De 3DBAG is een openbare dataset van 3D-modellen van gebouwen die heel Nederland dekt. In deze pijplijn worden metrics berekend op basis van de 3D-modellen en de 2D-voetafdruk van gebouwen, die bijvoorbeeld informatie geven over de rechthoekigheid van een gebouw. Deze metrics worden gebruikt om gebouwen met elkaar te vergelijken. Naast de specifieke metrics als rechthoekigheid wordt de 2D-voetafdruk ook verwerkt tot een turning functie. Dat is een soort handtekening van de vorm. Deze turning functies kunnen direct met elkaar worden vergeleken, en er kan een afstand berekend worden tussen deze functies. Dit geeft een zeer algemene manier om 2D-voetafdrukken met elkaar te vergelijken.

De verzamelde data wordt gecombineerd met gegevens vanuit het Kadaster, zoals bouwjaar, aantal bouwlagen en gebruiksfunctie. Deze verrijkte dataset kan vervolgens door machine learning-algoritmes worden gebruikt voor uiteenlopende toepassingen. Momenteel worden DBSCAN en K-Means ingezet om clusters van vergelijkbare woningen te detecteren. In Figuur 25 is een overzicht te zien van de resultaten van het clusteralgoritme. Er is gekozen voor een aantal voorbeeldwoningen en de andere woningen worden ingedeeld bij de meest passende voorbeeldwoning. De voorbeeldwoningen staan in de afbeelding vooraan, de andere bijbehorende woningen die veel gelijkenissen vertonen met de voorbeeldwoningen staan daarachter in de groepen. Met de resultaten van de clustering kan de variatie in bouwstijl van woningen in een wijk of stad worden bepaald, wat inzicht biedt in de uniformiteit of diversiteit van de bebouwing. Dit is nuttig voor de bepaling van de geschiktheid van gebieden voor grootschalige renovatieprojecten.

Een andere toepassing is het genereren van labels die woningen categoriseren op basis van hun geschiktheid voor renovatie. Deze labels kunnen vervolgens worden gebruikt om een supervised classificatie-algoritme te trainen, zoals een random forest, neurale netwerk of gradient boosting machine (GBM). Dit getrainde algoritme kan dan de geschiktheid voor renovatie voorspellen voor de gehele Nederlandse woningvoorraad, wat waardevolle inzichten biedt voor beleidsmakers en investeerders.

De ontwikkeling van deze pijplijn maakt het gebruik van de 3D-BAG data toegankelijk voor een breed scala aan toepassingen, van stadsplanning en vastgoedbeheer tot energietransitieprojecten. Zo kan deze technologie bijdragen aan het efficiënter plannen en uitvoeren van renovaties op grote schaal.



Figuur 25 - Overzicht van de verschillende clusters van gebouwen

3.3.3 Lessen geleerd bij HAN ADSAI

In lijn met de les bij Kadaster, zien we hier ook al dat de mogelijkheden van Data Science & AI overal zijn; zeer breed en zeer veel toepassingsmogelijkheden. Te veel kansen en zoveel mogelijkheden; dat is met de huidige kennis-intensieve manier van werken niet schaalbaar in te vullen; alleen met goede kennis is een project uitvoerbaar, en dat is te weinig beschikbaar. De oplossing zit uiteraard in onderwijs (en die handschoen pakken we als HAN graag op), maar ook het eenvoudiger toepasbaar maken van Data Science & AI is essentieel.

3.4 AI VOOR FUN

Werken moet ook leuk zijn. Dus we hebben ook een uitstapje gemaakt om AI in te zetten voor een geheel andere toepassing: het maken van bier. En in plaats van daar alleen over te praten hebben we het ook gerealiseerd; niet 1 maar 2 keer!

3.4.1 Poging 1 - 't Perceeltje

Als je bier wilt maken, heb je een recept nodig, en daar ligt een kans voor het toepassen van AI. Als teamactiviteit van het Kadaster Data Science Team hebben we een experiment gedaan om een bierrecept te voorspellen. Het startpunt zijn

de 201 recepten van de Schotse brouwer Brewdog die open beschikbaar zijn (helaas zijn ze een van de weinige brouwers die dit doen). Daarnaast weten we de scores van deze biertjes op basis van Untappd (community van bierdrinkers) beoordelingen. Dit is de basis waarop we het AI-model trainen, en daarbij een miljoen recepten genereren met daarbij een voorspelling van de potentiële Untappd score. Een potentieel hoog scorend recept hebben we gebrouwen onder de naam 't Perceeltje⁶¹. Ook het etiket en de tekst is met GenAI gemaakt.

Later zijn anderen ook AI gaan inzetten voor bier, en dat heeft geleid tot een publicatie in het toonaangevende tijdschrift Nature⁶². Een gemiste kans voor ons.



Figuur 26 - De etiketten van het 't Perceeltje en de HANZY PAI

3.4.2 Poging 2 - HANZY PAI

De ervaring van 't Perceeltje smaakte naar meer, nu vanuit de HAN. De lancering van het nieuwe lectoraat vormde een mooie aanleiding, daarnaast hebben we vanuit de HAN de mogelijkheid om (MADS) studenten en het café op de HAN campus (Lokaal99) te betrekken. Maar uiteraard willen we dan wel een volgende stap zetten in het gebruik van AI; het idee is om een bier te maken dat aansluit bij de smaak van de HAN drinker.

61 <https://labs.kadaster.nl/demonstrators/data-driven/datagedrevenbier>

62 <https://nos.nl/artikel/2514276-kunstmatige-intelligentie-maakt-belgische-biertjes-lekkerder>

In de app 'Untappd' houden mensen bij wat voor bieren ze waar gedronken hebben en wat voor cijfer ze het bier geven. We hebben data verzameld van de drie HAN-horecalocaties; Lokaal99, de Zalloon en grand café de HANgar. Dit gaf ons een grote database aan biertjes met bijbehorende waardering gedronken door voornamelijk HAN-studenten en medewerkers samen. De bieren hebben we gekoppeld aan een smaakprofiel: een lijst met eigenschappen zoals bitterheid en kleur die het biertje typeren. Met deze data hebben we een voorspellend model gemaakt dat aan de hand van een smaakprofiel kan voorspellen wat voor score de gebruikers van Untappd zouden geven. Toen we eenmaal een goed werkend model hadden, konden we de vraag omdraaien. We zochten naar een nieuw smaakprofiel bij een nog niet bestaand bier dat volgens ons model een hoge score op Untappd zou krijgen. Met een slimme methode genaamd Differential Evolution zochten we in de parameterruimte naar een nieuw smaakprofiel. Daarbij keken we naar kenmerkende smaakprofielen voor bestaande types bier. Zo hielden we rekening met het feit dat het wel een realistisch type bier moet zijn. Het model hebben we uitgebreid door verschillende hopsoorten toe te voegen, en daarbij te zoeken naar hopsoorten die een goed bier opleverden. Het resultaat is de HANZY PAI, een Hazy IPA met mandarijn.





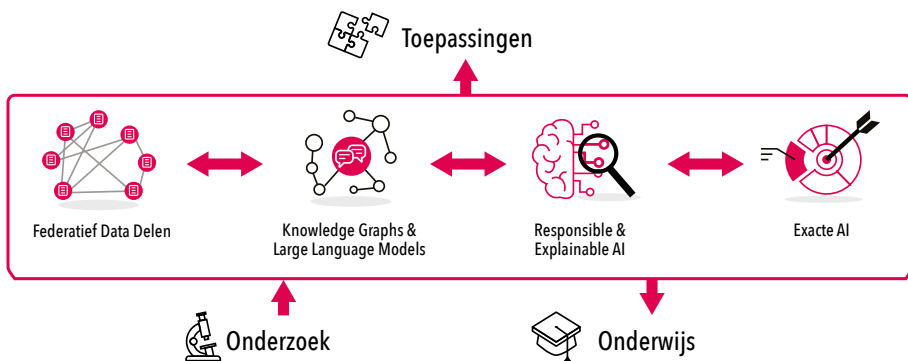
DEEL 4: HET LECTORAAT ADSAI

In de vorige delen hebben we de wereld van Data Science & AI geschetst, en ook prachtige voorbeelden uit de praktijk gegeven. Dit deel zal ingaan op de inhoudelijke focus van het ADSAI- lectoraat voor de komende jaren.

Wij hebben 4 thema's gekozen, en deze zijn:

1. Federatief data delen,
2. Knowledge Graphs & Large Language Models,
3. Responsible & explainable AI en
4. Exacte AI.

Deze 4 thema's zijn aan elkaar gerelateerd: het begint met data delen (en standaarden), en met Knowledge Graphs & Large Language Models kunnen we de volgende stap zetten in het toegankelijk maken van rijkere informatie. Dit vormt ook het datafundament om gewenste stappen, zoals herleidbaarheid naar de bron, te kunnen zetten met responsible en explainable AI. Exacte AI kan daarbij helpen om AI meer verantwoord (responsible) in te zetten. Deze vier thema's gaan we toelichten in de volgende paragrafen.



Figuur 27 - De scope van ADSAI

4.1 FEDERATIEF DATA DELEN

Data organiseren, beheren en delen op een veilige manier

Dit kennisgebied is op te delen in drie onderdelen: data delen, datastandaarden en data governance. Daarbij heeft het vele aanknopingspunten met het lectoraat Data & Knowledge Engineering (DKE).

Data Delen - wat is het?

Data moet stromen om waarde te creëren. In zowel overheid als bedrijfsleven stroomt data beperkt. Dit heeft verschillende oorzaken. Organisaties raken niet graag de controle over de data kwijt. Soms is er sprake van wantrouwen. Data is immers een asset, met waarde, en dient zo ook benaderd te worden. Een belangrijk concept in deze context is 'Data Sovereignty': eerlijk met data omgaan. Dat betekent dat de eigenaar regie houdt over zijn data en deze onder controle en onder de juiste voorwaarden veilig kan delen. Als hulpmiddel kan gebruik gemaakt worden van 'Data Spaces' waarbij organisaties op basis van data contracts data met elkaar delen (zie ook 2.1.3 en 3.1.3).

In de context van de Nederlandse overheid speelt daarbij het Federatief Data Stelsel. Het uitgangsprincipe daarbij is om meer data bij de bron te laten en niet te werken met kopieën. Data wordt waar mogelijk geminimaliseerd en eenvoudig toegankelijk gemaakt via open standaarden.

Toegang tot en controle over data zal daarbij steeds fijnmaziger moeten worden. Mede doordat door de nieuwe datastandaarden (zoals Linked Data) combinaties in data steeds eenvoudiger te maken zijn. Hiermee kan meer waarde gecreëerd worden, maar moeten de risico's opgevangen worden door fijnmaziger toegang, alsook betere monitoring en afscherming.

(Federatieve Machine Learning is in de basis hetzelfde conceptuele idee van federatie (bij de bron); en zien we ook als onderdeel van Federatief Data Delen)

Wat doen wij?

Wij kunnen proof of concepts/demonstrators maken voor het delen van data. Dit doen wij door middel van bijvoorbeeld Data Spaces & Knowledge Graphs, met aandacht voor toegangsbeheer op basis van smart contracts, autorisatie-ontologieën, en andere toepasbare concepten. We werken aan een veilig en eerlijk data ecosysteem.

Data Standaarden - wat is het?

Data Standaard 3.0 is een set regels en richtlijnen die bepalen hoe gegevens moeten worden opgeslagen, gedeeld en gebruikt, zodat verschillende systemen en organisaties gemakkelijk met elkaar kunnen communiceren en samenwerken. Het is als een gemeenschappelijke taal voor gegevens, zodat iedereen dezelfde woorden en zinnen gebruikt en begrijpt wat er wordt bedoeld. In andere woorden: Data Standaard 3.0 zorgt ervoor dat gegevens uniform en begrijpelijk zijn voor iedereen die ze gebruikt, waardoor samenwerking en innovatie worden gestimuleerd.

Iets technischer beschreven: van de traditionele manier van 'berichtenuitwisseling' (via EDI, XML, JSON) conform berichten standaarden bewegen we naar het gebruik van Knowledge Graphs voor data delen. Hierbij zijn de uitwisselingstaal (de ontologie) alsook de eventuele regels aan de uitwisseling vastgelegd, maar hieruit worden geen berichten meer gespecificeerd. Deze nieuwe vorm biedt veel flexibiliteit en toch ook semantische interoperabiliteit, en is daarnaast een invulling van data bij de bron, en dataminimalisatie.

Wat doen wij?

We kunnen het ontwerp maken van een data standaard 3.0, een invulling met nadruk op semantiek (met OWL, SKOS) en regels (met SHACL). Wij kunnen de data standaard 3.0 ontwikkelen voor een specifieke toepassing (bijvoorbeeld Digitaal Product Paspoort (DPP)), en wij onderzoeken de verdere evolutie van datastandaarden. Op die manier creëren we waarde met data en waarborgen we de kwaliteit van datastandaarden.

Data Governance: wat is het?

Data governance is het fundament voor het werken met data. Wij hebben als lectoraat een unieke positie omdat wij historisch ervaring hebben met governance

op data (semantische) standaarden, gebruik makend van het BOMOS-model. Dit model wordt vaak gebruikt bij beheerders van datastandaarden (Geonovum, CROW, Kennisnet, RWS, TNO, et cetera.). Het model is uit te breiden naar governance op data en andere data-gerelateerde artefacten. Daarnaast is hier sprake van een learning community van gebruikers en kennisorganisaties.

Wat doen wij?

Wij leveren een bijdrage aan BOMOS, en de implementatie van BOMOS voor data en standaarden. Wij zullen de BOMOS-community aanjagen. Dat zullen we in nauwe samenwerking doen met Logius en Geonovum, en ook vanuit en met het DEMAND netwerk⁶³.

Wat levert het op?

Het onderwerp Federatief Data Delen is vooral interessant voor spelers binnen het publieke domein, en sectoren/ketens, veelal vertegenwoordigd door branche-organisaties. Om federatief data te kunnen delen heb je afspraken/standaarden nodig. Die standaarden moeten zekerheid bieden en technisch implementeerbaar zijn, waarmee data 'vloeit' en er economische en maatschappelijke waarde ontstaat. Concreet betekent het:

- **Gegevens blijven lokaal:** Elk bedrijf of instelling houdt zijn eigen gegevens bij zich. Ze sturen hun gegevens niet naar een centrale database.
- **Samenwerken zonder delen:** Organisaties kunnen samenwerken en gegevens gebruiken van elkaar zonder de gegevens daadwerkelijk over te dragen.
- **Veiligheid en privacy:** Omdat de gegevens niet naar een centrale plek worden verplaatst, blijven ze beter beschermd en privé. Alleen de resultaten van de gegevensanalyse worden gedeeld.
- **Technologie maakt het mogelijk:** Speciale software en protocollen zorgen ervoor dat de gegevens op een veilige manier kunnen worden geraadpleegd en gebruikt.

Ter illustratie: verschillende ziekenhuizen willen onderzoek doen naar een ziektebeeld. Elk ziekenhuis heeft patiëntgegevens, maar ze willen deze gegevens niet delen vanwege privacy redenen. Met federatief data delen kunnen ze toch samenwerken en analyses uitvoeren op al die gegevens samen, zonder dat de

⁶³ <https://demand.nl/>

gegevens het ziekenhuis verlaten. Zo krijgen ze nuttige inzichten zonder de privacy van patiënten in gevaar te brengen. In essentie gaat het erom dat je kunt samenwerken en informatie kunt gebruiken alsof alles op één plek staat, terwijl dat in werkelijkheid niet zo is.

4.2 LARGE LANGUAGE MODELS & KNOWLEDGE GRAPHS

Van hallucinaties naar feiten

Wat is het?

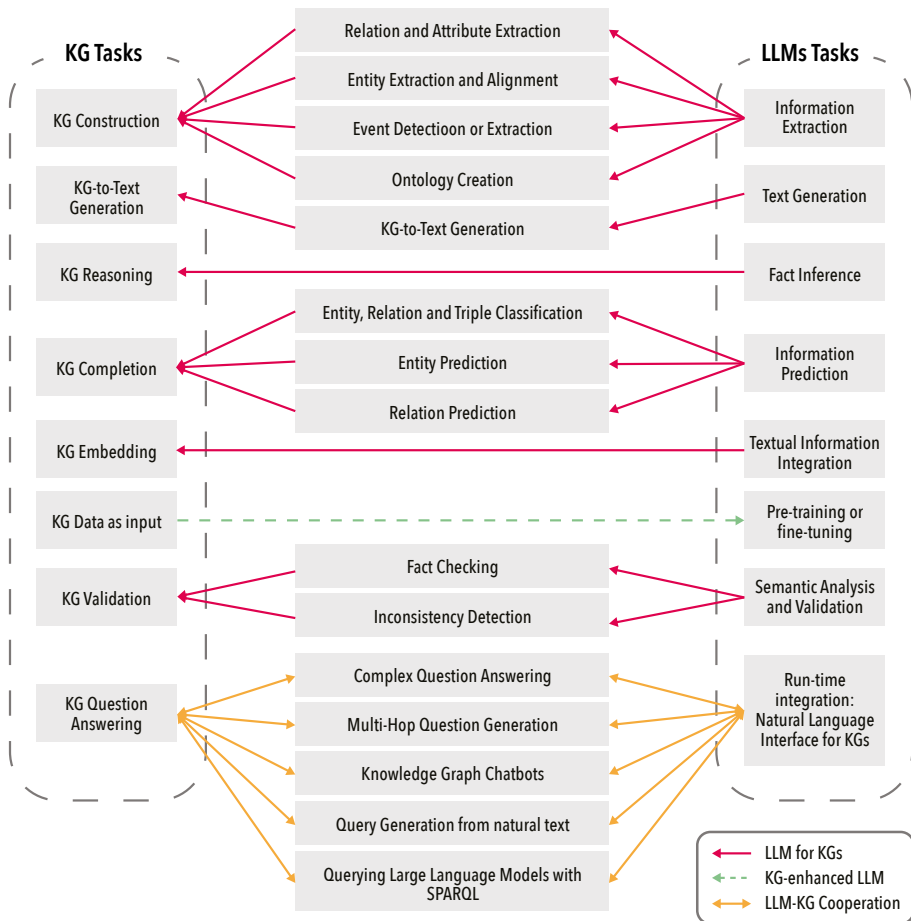
Large Language Models (LLMs) verrassen in vele opzichten. Implementaties zoals ChatGPT zijn laagdrempelig te gebruiken en weten ons vaak te overtuigen met goede antwoorden. Tegelijkertijd geven deze modellen met dezelfde overtuiging verzonnen antwoorden; de zogenaamde hallucinaties. Vaak zijn deze zo subtiel dat ze bijna lijken te kloppen. Dit is een actief probleem waar veel praktische toepassingen tegenaan lopen.

Knowledge Graphs zijn helemaal niet gebruiksvriendelijk, inhoudelijk complex, maar bieden wel feitelijke antwoorden herleidbaar naar de bron, zonder de mogelijkheid van hallucinaties. Niet voor niets dus dat de combinatie van Knowledge Graphs & Large Language Models wordt gezien als een potentieel gouden combinatie (zie ook sectie 2.5). Ze zijn laagdrempelig en voorkomen tegelijkertijd hallucinaties⁶⁴.

Recent onderzoek⁶⁵ heeft vele mogelijkheden geïdentificeerd hoe Knowledge Graphs en Large Language Models elkaar kunnen versterken (zie Figuur 28). Onder andere kunnen Large Language Models ondersteunen bij het (semi) automatisch maken van Knowledge Graphs, of ondersteunen bij het valideren van de Knowledge Graph. Omgekeerd kan de Knowledge Graph gebruikt worden om trainingsdata te genereren voor het Large Language Model. Maar pas echt interessant wordt het als Knowledge Graphs en Large Language Models samen komen in bijvoorbeeld een chatbot op basis van de Knowledge Graph.

⁶⁴ <https://aidanhogan.com/talks/2023-09-22-semantic-2023.pdf>

⁶⁵ <https://arxiv.org/html/2406.08223v2>



Figuur 28 - Relatie Knowledge Graphs en Large Language Models

Wat doen wij?

Elke activiteit uit figuur 28 (midden kolom) is in scope voor ADSAI. Een aantal voorbeeld activiteiten:

- Knowledge Graphs gebruiken als basis van een Large Language Model. Bijvoorbeeld door het Large Language Model in te zetten voor de transformatie van natuurlijke taal naar queries (SPARQL) op de Knowledge Graph;

- Knowledge Graphs creatie. Er zijn manieren om automatisch Knowledge Graphs te genereren van een grote hoeveelheid tekst. Deze Knowledge Graphs zijn een weergave van de globale structuur van een verzameling teksten en kan worden meegestuurd in een poging het antwoord van de Large Language Model te 'gronden' in deze globale structuur;
- Een specifiek Large Language Model trainen op basis van een Knowledge Graph.

Iedereen wil een 'eigen' ChatGPT voor zijn unieke kennisbronnen (variërend van HR-reglement tot technische handleidingen en projectvoorstellen), maar hoe kunnen we de betrouwbaarheid van deze modellen verbeteren, als een Knowledge Graph niet aanwezig is of een stap te ver. Dan kan RAG (Retrieval Augmented Generation) helpen. Hierbij wordt een database aangelegd van eigen, specifieke informatie (bijvoorbeeld een HR-reglement). Op basis van de vraag die een gebruiker stelt aan de Large Language Model wordt snel de relevante context (bijvoorbeeld pagina 8 van het reglement) meegestuurd naar de Large Language Model om een antwoord te geven.

Ook worden binnen het lectoraat andere oplossingen bekeken zoals een JEPA-architectuur (Joint Embedding Predictive Architecture) en Sensory substitution for semantic vectorspaces. Waar de JEPA het Large Language Model wil verbeteren, wil Sensory substitution een manier zijn waardoor mensen betere zintuigen krijgen om de context beter te begrijpen.

Wat levert het op?

Door deze technieken in te zetten lossen we het hallucineren van AI-modellen op en werken wij als het ware samen met de modellen omdat we een betere 'taal' creëren waarmee we het AI-model kunnen bevatten. Voor alle organisaties die betrouwbaarheid van data belangrijk vinden, leveren wij een bijdrage aan AI-systemen die mensen helpen om op een betrouwbare manier informatie te ontsluiten. Ook leren wij organisaties nieuwe manieren van interacteren met AI-systemen. Wij hebben de kennis in huis om te onderzoeken hoe generatieve AI, en in het bijzonder Large Language Models, betrouwbaarder en herleidbaar gemaakt kunnen worden.

4.3 RESPONSIBLE & EXPLAINABLE AI

AI zo uitlegbaar en transparant mogelijk

Wat is het?

AI is 'hot': we horen er iedere dag wel iets over en alle bedrijven en organisaties doen óf moeten er iets mee. We weten goed wat AI is: het nabootsen van de menselijke intelligentie en probleemoplossende capaciteiten met computers. Daarentegen is het niet altijd duidelijk hoe AI werkt omdat de modellen niet perse transparant zijn. Dit betekent dat je niet altijd kan inzien wat een AI-systeem doet en of het immorele beslissingen neemt. Responsible AI zorgt voor een transparante, veilige en ethische inzet van AI. Explainable AI is daarbij essentieel omdat het vertelt hoe AI tot een voorspelling, advies of tekst komt.

Responsible AI betekent dat we AI-systemen bouwen en gebruiken op een ethische en verantwoorde manier. Dit houdt in dat de AI eerlijk, transparant en zonder vooroordelen werkt en dat de privacy van mensen wordt beschermd. Zo schrijft responsible AI voor dat AI-systemen voor werving en selectie niet mogen discrimineren op basis van geslacht, ras of leeftijd. Op deze manier zijn beslissingen eerlijk voor alle kandidaten.

Explainable AI (XAI) richt zich op de uitleg van de voorspellingen van AI. Wanneer je een AI-systeem een besluit laat nemen, wil je soms ook weten hoe het systeem tot dat besluit is gekomen. Bij zogenaamde glass-box modellen is het beslissingsmechanisme begrijpelijk voor mensen en kan een uitleg bij een voorspelling worden gegeven door enkel naar het model te kijken. Voor de meeste populaire modellen is dat echter niet het geval en dan spreken we over black-box modellen. Deze modellen zijn zo complex dat niemand kan uitleggen waarop een voorspelling gebaseerd is. Om hier inzicht in te krijgen zijn XAI-tools nodig.

Met XAI-tools kan nagegaan worden waarom een bepaalde beslissing is gemaakt en wordt die op een begrijpelijke manier aan mensen uitgelegd. Dit wordt bijvoorbeeld gedaan door de inputdata systematisch aan te passen en te observeren welk effect dit heeft op de voorspelling. Hierdoor kan je bepalen waarop het model de voorspelling baseert. Ter illustratie: wanneer AI een lening afwijst, kan XAI uitleggen dat dit komt door een lage kredietscore en te veel bestaande schulden. Hiermee is XAI een middel waarmee je kan aantonen dat je AI responsible is.

Wat doen wij?

Door responsible en explainable AI te combineren, kunnen we AI-systemen bouwen die niet alleen krachtig en efficiënt zijn, maar ook begrijpelijk, eerlijk en betrouwbaar. Wij kunnen de risico's van AI-systemen identificeren en adviseren over het verantwoordelijke gebruik van AI.

Het lectoraat Applied Data Science & AI kan bedrijven en (semi-)overheidsinstellingen adviseren over het implementeren van responsible AI. Wij kunnen organisaties ondersteunen bij de implementatie van XAI door te helpen bij het selecteren, implementeren en interpreteren van XAI-tools.

ADSAI zal vooral een adviserende rol hebben om samen met de organisatie te onderzoeken hoe XAI in specifieke gevallen ingezet kan worden. Daarnaast zullen we ook gezamenlijk de uitkomsten van de tools - de daadwerkelijke uitleg - interpreteren en kritisch beoordelen.

Bijvoorbeeld bij het gebruik van foto's als input data kunnen er heatmaps gemaakt worden waarop te zien is welke delen van de foto zorgen voor positieve of negatieve voorspellingen. Voor input data kunnen bijvoorbeeld SHAP- of LIME-waarden worden berekend. Deze waarden geven inzicht in de invloed van verschillende delen van de data op de voorspelde waarde of nauwkeurigheid. Een andere manier om een voorspelling te verklaren is met counterfactual explanations. Deze geven voorbeelden waarbij een kleine verandering in de inputdata leidt tot een andere voorspelling.

4.4 (MACHINE/DEEP LEARNING MET) EXACTE AI

Wiskunde is het fundament

Wat is het?

AI en machine learning zijn diepgeworteld in de exacte wetenschappen. Praktisch alle modellen zijn gebaseerd op lineaire algebra en calculus. Daarnaast zijn modellen veelal geïnspireerd op andere modellen uit de natuurkunde en biologie. Dit komt ook sterk naar voren in de moderne technieken zoals (quantum) annealing, neuromorphic computing en diepe neurale netwerken (zie 2.4). Het implementeren van dit soort moderne technieken vraagt dan ook veel inhoudelijke kennis: Ze moeten vaak op maat gemaakt worden voor de specifieke toepassing,

en dat vergt dat een praktisch probleem vertaald moet worden naar een uniek wiskundig model.

Ook voor het efficiënt kiezen van het juiste model voor de juiste toepassing is het noodzakelijk de exacte achtergrond voldoende te doorgronden. Een naïeve aanpak om veel grote modellen te trainen is inefficiënt en heeft een hoge klimaatimpact. Veel efficiënter is het om een gelimiteerd aantal modellen te proberen en daarbij niet onnodig diepe netwerken te gebruiken. Voor het kiezen van een geschikt model met een zo klein mogelijke klimaatimpact, is het nodig kennis te hebben over de precieze werking en kosten van een grote hoeveelheid machine learning technieken.

Naast kennis over AI-modellen is domeinkennis vaak instrumenteel aan het behalen van toepasbare resultaten. Beschrijvende modellen die gecombineerd worden met en aangevuld worden door data worden ook wel grey box models genoemd. Een concreet voorbeeld hiervan is een natuurkundig model om bandslijtage bij te houden en te voorspellen wanneer een band aan vervanging toe is, dat daarbij gebruik maakt van verzamelde data om parameters beter te kunnen voorspellen en updaten. De grote kracht van deze modellen ligt in het effectief combineren van machine learning (AI) met exacte berekeningen en onderliggende wetmatigheden.

Wat doen wij?

Als lectoraat ADSAI hebben we een exacte achtergrond met veel kennis in machine learning, AI, wiskunde en natuurkunde. Met deze kennis willen we impact creëren met verschillende toepassingen. Bijvoorbeeld samen met de Hogeschool Utrecht werkten we in Future Factory (zie 3.3.2) aan verduurzaming, door nieuwe en snelle algoritmes te ontwerpen om woningen te vergelijken. Ook in de maakindustrie zetten we moderne technieken als annealing in de planning om de doorloop in fabrieken te verbeteren.

In verschillende projecten zetten we onze natuurkundige en wiskundige kennis in voor het ontwikkelen van grey box models. Bijvoorbeeld voor huurders in het sociale segment maken we inzichtelijk hoe ze kunnen besparen op energieverbruik en kosten. Ook op het gebied van transport ontwikkelen we grey box models om efficiëntie te vergroten en voorspellingen te maken over onderhoud.

Tenslotte gebruiken we onze kennis om in samenwerkingen te adviseren over efficiënte modellen. Daarbij maken we inschattingen welke modellen met minimale klimaatimpact een optimaal resultaat kunnen bereiken, en vermijden we het overmatig trainen van meerdere modellen.

Wat levert het op?

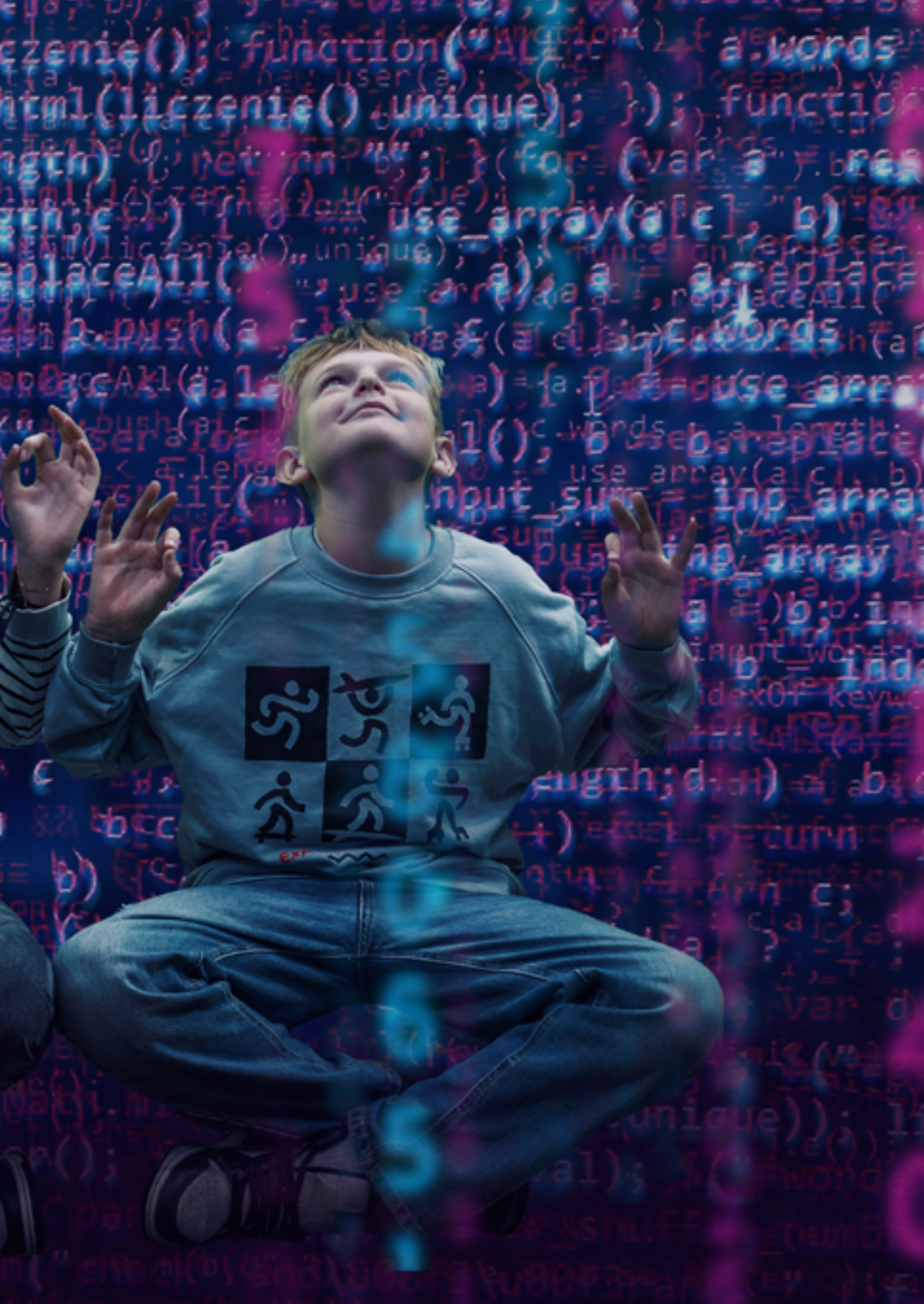
Energiebesparing: Allereerst dragen onze resultaten bij aan de verduurzaming van woningen. We maken inzichtelijk voor bewoners hoe ze kunnen verduurzamen en besparen, en inventariseren welke woningen op een soortgelijke manier gerenoveerd kunnen worden. Ten tweede bereiken we deze resultaten met minimale (klimaat)impact van de gebruikte modellen.

Unieke en zuinige modellen: We ontwikkelen nieuwe algoritmes op basis van onze exacte kennis. Daarbij komen we met unieke oplossingen buiten de gebaande paden. Dit levert efficiëntere methodes op die niet meer middelen gebruiken dan nodig.

Kennis: Door het toepassen van moderne modellen die niet out-of-the-box werken, doen we ervaringen op met het toepassen van deze methodes in de praktijk. De kennis die we daarbij opdoen draagt bij aan het efficiënt en energiezuinig draaien van vervolgprojecten.







SAMENVATTING

Op het gebied van Data & AI leven we in een interessante tijd. De ontwikkelingen gaan in een razend tempo. Als onderzoeker is het een enorme snoepwinkel met prachtige ontwikkelingen waarin het moeilijk kiezen is. Maar zoals in deel 1 geschetst: onze primaire taak is om Data Science & AI-toepassingen mogelijk te maken. In het voorgaande hebben we inhoudelijk onze focusgebieden aangegeven. In dit laatste deel schetsen we drie uitdagingen die grootschalige toepassingen in de weg staan. Vervolgens vatten we de focus van het lectoraat ADSAI samen, en eindigen we met stof tot nadenken.

DE UITDAGINGEN

Uiteraard zijn er vele uitdagingen rond het werken met Data en het toepassen van AI. We schetsen er drie:

Uitdaging 1: Ruis op de lijn.

Er zijn inmiddels meer mensen die over AI praten dan met AI werken, en dat is een probleem. Vooral als men feiten en fabels niet uit elkaar kan houden. Dit resulteert namelijk in heel veel ruis, veel overleg, foute beslissingen, maar heeft ook tot gevolg dat AI-specialisten zich minder in discussies gaan mengen: ze willen graag door met de volgende AI-innovatie. Daardoor ontstaan er twee gescheiden werelden van de mensen die praten over AI en de collega's die werken met AI. Dit is onwenselijk.

We willen een bijdrage leveren aan minder ruis, of een hoger niveau van data & AI geletterdheid in de maatschappij. Onder andere onze publicaties (zoals dit boekje) zouden hier een bijdrage aan kunnen leveren, maar nog belangrijker is de embedding in het onderwijs van de HAN. Met de MADS opleiding, de HBO-ICT opleiding en LLO-modules, zijn we op de goede weg. Maar het streven is dat elke HAN student data & AI geletterd is.

Uitdaging 2: Een Data Scientist heeft 80% verloren tijd.

Eerder hebben we de 80-20 regel beschreven, waarmee gesteld wordt dat een Data Scientist 80% van zijn/haar tijd met Data Engineering (verkrijgen, opschonen,

et cetera) taken bezig is, en slechts 20% met analyse/science taken. Een groot deel van die 80% zou voorkomen kunnen worden en is daarmee verloren tijd, maar nog belangrijker is dat deze tijd kostbaar is: Dit maakt Data Science duur, en daarmee lastig voor MKB-bedrijven.

We willen een bijdrage leveren in het data fundament (bijvoorbeeld door de ontwikkeling van data standaard 3.0, maar ook door het beschikbaar maken van een data lab faciliteit) om die 80% naar beneden te krijgen.

Uitdaging 3: Het gebruik van domme data.

We laten AI-voorspellingen en keuzes maken. We geven AI daarbij data maar om het wat lastiger te maken vertellen we niet alles. Het is alsof je het management van een organisatie een begroting laat zien en vraagt om keuzes te maken. Om het wat spannender te maken voor het management vertel je niet wat de data betekent.

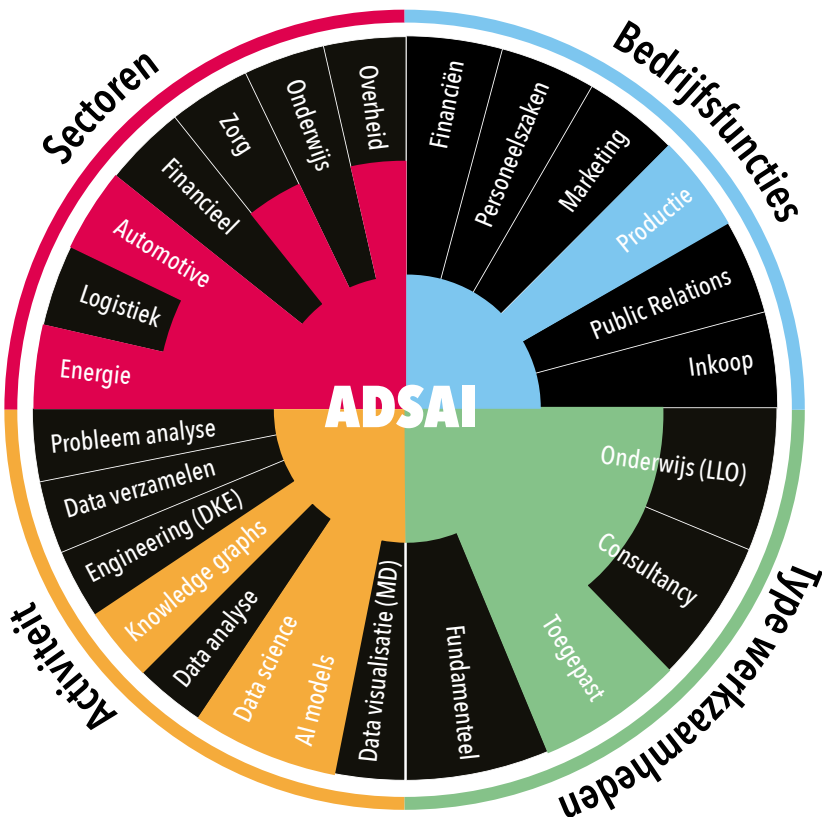
Het gokken ofwel 'hallucineren' van AI staat toepassingen in de weg. Uiteindelijk willen we alleen verantwoord gebruik van AI. Dat heeft meerdere facetten; alle kennis die we hebben willen we ook daadwerkelijk gebruiken. Dit kan met Knowledge Graphs waarin ook alle metadata beschikbaar gesteld kan worden. Maar we willen ook een onderbouwd passend AI-model, zowel in de zin van explainable, responsible met ook oog voor het klimaat: Eerlijke AI.

DE FOCUS VAN ADSAI

Inhoudelijk heeft ADSAI een stevige focus gericht op een rijk semantisch datafundament, met Knowledge Graphs, die als basis dienen voor explainable en responsible AI, waar we de exacte (wiskundige) kant van AI goed willen benutten, gericht op Data Science toepassingen in verschillende sectoren. We kiezen daarbij voor een focus op het primaire 'productie' proces om impact te hebben, ten opzichte van meer ondersteunende bedrijfsfuncties. Die primaire processen kunnen zowel in de industrie (automotive, logistiek) als in de (semi) publieke sector (overheid, zorg, energie) plaatsvinden. De activiteiten die we uitvoeren zullen in hoofdzaak gericht zijn op het concreet inzetten van Data & AI concepten, bijvoorbeeld in grote onderzoeksprojecten of juist in kleine situaties met het regionale MKB. Dat kan eventueel ook in de vorm van kleinschalige consultancy.

Ook willen we de kennis in onderwijsprogramma's onder brengen. Zowel in het reguliere (bachelor/master) onderwijs als ook cursussen in het kader van Leven Lang Ontwikkelen (LLO).

Binnen de HAN kunnen we daarbij goed partnersen. Binnen onze eigen academie (AIM) met het lectoraat DKE (Data & Knowledge Engineering) op het gebied van semantiek (taal) en het lectoraat MD (Media Design) op het gebied van data visualisatie. De andere academies van de HAN zijn de ingangen naar de sectoren, de uitdagingen in de praktijk. Maar nog belangrijker zij bieden de domeinkennis die nodig is om Data & AI technologie in te zetten die waarde toevoegt. De 1+1 =3.



Figuur 29 - Het Lectoraat ADSAI samengevat

TOT SLOT⁶⁶

De eerder beschreven uitdagingen zijn niet het einddoel, maar zijn randvoorwaardelijk om het uiteindelijke doel van waardecreatie met Data Science & AI voor iedereen mogelijk te maken. Geen experimenten, maar serieuze toepassingen; het is de toepassing waar de HAN voor staat.

Tot slot, om het gedachteproces nog wat te prikkelen, nemen we een eenvoudige data-analyse vraag: Hoeveel huizen staan er in Warnsveld?

Die vraag stellen we in september 2024 aan ChatGPT, dan krijg je als antwoord 'volgens de meest recente gegevens van het CBS ongeveer 3700 woningen', verder geen bron of referentie. Het antwoord komt na ongeveer 5 seconden, en het energieverbruik is gelijk aan ongeveer 400 meter rijden met een auto. De hardware die ChatGPT gebruikt is onbetaalbaar.

Stel je dezelfde vraag aan Google dan is het antwoord: 'Wijk 05 Warnsveld heeft in totaal 3.739 geregistreerde woningen' met een referentie naar een niet-authentieke data bron⁶⁷. Overigens het antwoord komt na ongeveer 0.6 seconden, en het energieverbruik is gelijk aan ongeveer 1 meter rijden met een auto. De hardware van Google is zelfs als je het geld zou hebben niet te koop.

Ook kan de vraag aan een Knowledge Graph gesteld worden, bijvoorbeeld de Kadaster Knowledge Graph. De vraag stellen kan met een SPARQL query; het antwoord is dan 3873. Daarbij zien we in de query dat we 'verblijfsobjecten met een woonfunctie' hebben opgehaald uit de authentieke bron: De Basisregistratie Adressen en Gebouwen. Dat ging in ongeveer 0.6 seconde. Knowledge Graphs draaien meestal op eenvoudige goedkope hardware.

Uiteraard kost het meer tijd voor een gebruiker om een SPARQL query te maken en is dit niet voor iedereen weggelegd. Maar ook het maken van SPARQL queries kan via een getrainde AI-chatbot: we stellen dezelfde vraag aan Loki (labs.kadaster.nl): uiteraard krijgen we dan hetzelfde antwoord 3873, uit dezelfde authentieke bron, met zelfs een toelichting en de SPARQL query erbij.

⁶⁶ Noot: Dit voorbeeld is overgenomen en aangepast van: <https://www.youtube.com/watch?v=ww99npDh4cg> en <https://www.poolparty.biz/thank-you/poolparty-summit-2024/>

⁶⁷ <https://kadastralekaart.com/wijken/wijk-05-warnsveld-WK030105>

De tijd en impact is uiteraard iets hoger geworden dan de 'pure SPARQL query', maar komt nog altijd niet in de buurt van de ChatGPT tijd en impact.

Dus of het argument nu is dat je een voorkeur hebt voor een correct antwoord, of dat je een voorkeur hebt dat de officiële authentieke bronnen als data (live bij de bron) worden gebruikt. Of dat je gewoon vindt dat we zuinig met ons klimaat moeten omgaan. De essentie is om vooral een goed data fundament (wat een Knowledge Graph biedt) in te zetten, en daarnaast alleen AI toepassen met een positieve kosten/baten balans (waarde).



HAN_



HANZY PAI

BAYESIAN INFUSED - FOLMER EDITION

6.1% vol

33 cl



HAN_



HANZY PAI

BAYESIAN INFUSED - FOLMER EDITION

6.1% vol

33 cl





DANKWOORD

Ik dank het College van Bestuur van de HAN University of Applied Sciences voor het in mij gestelde vertrouwen. Stijn Hoppenbrouwers en Astrid Hoge wil ik bedanken voor het plaveien van de weg, en nog belangrijker voor het vertrouwen en het warme welkom dat ik heb mogen ontvangen binnen de academie IT & Mediadesign. Een betere start had ik me niet kunnen wensen.

Stijn heeft me een vliegende start gegeven met mooie projecten en mensen. Met collega's, Raoul Grouls, Vincent Coumans, Onno Huijgen, Timo Scheidel en Mohammad Karimi vormen we nu het lectoraat ADSAI. Dank allemaal voor de prettige samenwerking; we gaan er de komende tijd een leuk feestje van maken! En waar zouden we zijn zonder Boukje Postma....in ieder geval niet hier op dit event.

Maar ook mijn (oud-)werkgevers en collega's van TNO, Universiteit Twente en het Kadaster wil ik bedanken. Paul Oude Luttighuis en Jos van Hillegersberg als mijn leermeesters maar ook als zeer prettige collega's. Ik kijk met zeer veel plezier terug op de tijd in Twente, zowel bij TNO als op de UT.

Kadaster noem ik wel eens een verborgen parel. Aan de buitenkant ziet het er misschien uit als een wat saaie organisatie, aan de binnenkant hebben we toppers waarmee we innovatie van topklasse kunnen realiseren. Het Data Science Team is de parel van het Kadaster. Het zijn te veel namen om te noemen, maar in het bijzonder wil ik Charl Vermeer bedanken voor de blanco cheque om het Data Science Team te starten (ok, misschien was de cheque niet blanco, maar zo voelde die wel). Lexi Rowland (voor het krokodilmoment, en wat dingetjes met Linked Data), Anjo Kolk (voor de Slack-avondjes), en Janneke Michielsen aan wie ik het Data Science Team in goede handen kon overdragen.

Ik kon eigenlijk geen afscheid nemen van het Kadaster Data Science Team, en ik ben dankbaar aan zowel de HAN als het Kadaster dat we een constructie hebben gevonden waarbij dat niet nodig is. Sterker nog; we zien de waarde om elkaar de komende jaren te versterken. Bedankt Richard Metz en Frank Tierolff voor de support vanuit het Kadaster, de bijdrage op het event, en de altijd scherpe vragen.

Projectjes lopen bij mij nog wel eens uit de hand...zo ook het schrijven van dit boekje. Het begon met het schrijven van een korte introreede, maar toen kwam ook het idee om een korte beschrijving te maken van AI-concepten. En, ja, dan is dit boekje ook wel een mooie kans om gave projecten in de spotlights te zetten, en ook de plannen van het lectoraat te ontvouwen. Iets kleins werd iets groters... en de tijd was kort. Dit was compleet onmogelijk, zonder de hulp van velen. Het is echt een team-effort geworden. Specifiek woord van dank aan Timo Scheidel, zonder jou was het boekje dunner geweest. Michiel Stornebrink, Janneke Michielsen, Marc van Andel, Wim Florijn, Lexi Rowland, Jiarong Li, Onno Huijgen, Vincent Coumans, Raoul Grouls, Stijn Hoppenbrouwers, Boukje Postma, Elvira Folmer, Sylvie Smeets; allemaal bedankt voor jullie bijdrage. Bijzonder veel dank aan Vera Lange die op het laatst nog de figuren gemaakt heeft. Ik hoop dat lezers het resultaat weten te waarderen.

Bij een mooi boekje hoort dan ook event waar het uitgedeeld kan worden. Ook dat liep iets uit de hand, maar het is wel fantastisch te zien dat er vandaag zoveel mensen aanwezig zijn! Iedereen hartelijk bedankt! Een speciaal woord van dank aan Franka Janssen, Sabine de Vré en Boukje Postma voor de organisatie van dit event.

In de (Linked) Data wereld heb ik vele vrienden gemaakt; ik ben erg blij dat ze ook een bijdrage hebben kunnen leveren aan het event vandaag (Wouter Beek, Rob Wenneker, Richard Zijdeman, Henk van Haaster, Michiel Stornebrink, Linda Oosterheert, Marcel Reuvers), allemaal hartelijk dank!

Blijft nog over mijn familie: De sollicitatiegesprekken combineerde ik met ziekenhuisbezoeken, slechts 4 kilometer van elkaar verwijderd en een wereld van verschil. Ik had het toen niet verwacht, maar het geeft me een enorm gevoel van trots dat mijn beide ouders vandaag hier bij kunnen zijn. Ook mijn broertje en zusje wil ik bedanken, voor te veel om op te noemen.

Lieve Lotte en Jens; ik kan enorm van jullie genieten. Jullie vinden dit gedoe maar gek, en daar hebben jullie best wel gelijk in. Maar dat mag ik niet zeggen...jullie wel! Welk pad jullie ook zullen gaan bewandelen. Ik zal er voor jullie zijn!

Lieve Elvira, we kennen elkaar al even...jouw thesis schreef je op mijn studentenkamer op de campus in Enschede...een paar jaar geleden (tenminste zo voelt het). Daarna zijn er meerdere boekjes gevolgd, deze is echt de laatste. Jij hebt het mogelijk gemaakt dat ik de avonden en weekenden aan dit boekje kon besteden. Maar vooral ook dank dat je het al zo lang met mij weet vol te houden; dat is een prestatie die niet in de buurt komt van het schrijven van alle boekjes samen.

Ik zie er erg naar uit om weer fijn de tijd met jullie door te brengen, of het nou thuis is, met de camper op pad, of de komende nog onbekende vakantiebestemming die Jens gaat kiezen (hopelijk zonder de hulp van AI).

OVER DE AUTEUR



Erwin Folmer (1975) is een ervaren onderzoeker en adviseur op het snijvlak van business en ICT. De afgelopen jaren gaf hij leiding aan het Data Science Team bij het Kadaster en was hij docent en onderzoeker bij de Universiteit Twente.

Eerder werkte hij bij TNO als senior wetenschapper op het gebied van data interoperabiliteit en standaarden. In 2012 promoveerde Erwin op het proefschrift 'Quality of Semantic Standards' bij de Universiteit Twente.

Hij is gespecialiseerd in onder andere Data Science, Linked Data, Semantic Web en AI, en zal zijn functie als lector bij de HAN blijven combineren met een parttime functie bij het Data Science Team van het Kadaster.

OVER HET LECTORAAT

Data Science is een interdisciplinair vakgebied dat zich richt op het verkrijgen van kennis en inzicht uit data én de toepassing hiervan. Dit vakgebied wordt vaak in één adem genoemd met AI dat een verzamelnaam is voor allerlei data-intensieve technologieën, toepassingen en autonome systemen.

Het lectoraat fungeert als aanjager, verbinder en kennisdrager op het gebied van Data Science en AI. Zowel binnen de HAN als expertisecentrum, maar ook in de regio in de samenwerking tussen onderzoek, onderwijs, bedrijfsleven en de overheid. Een kennisnetwerk op het gebied van het delen van data en concepten zoals Knowledge Graphs, Federatie van Data, Data Spaces en AI. Gericht op concrete toepassingen in verschillende domeinen zoals energie, gezondheidszorg en industrie.

<https://www.han.nl/onderzoek/lectoraten/lectorat-applied-data-science-and-ai/>

