Eindrapportage pilot – DCC voor Praktijkgericht onderzoek

Pilotgegevens	
Titel van pilot	**Creating a Data Fabric through Easy-to-use Cloud Computing**
Hogeschool	**Hogeschool Rotterdam [HR]**
Contactpersoon	**Rob van der Willigen**
Emailadres	**r.f.van.der.willigen@hr.nl**
Looptijd	**1 juli tot 8 december 2022**
datum	**8 december 2022**

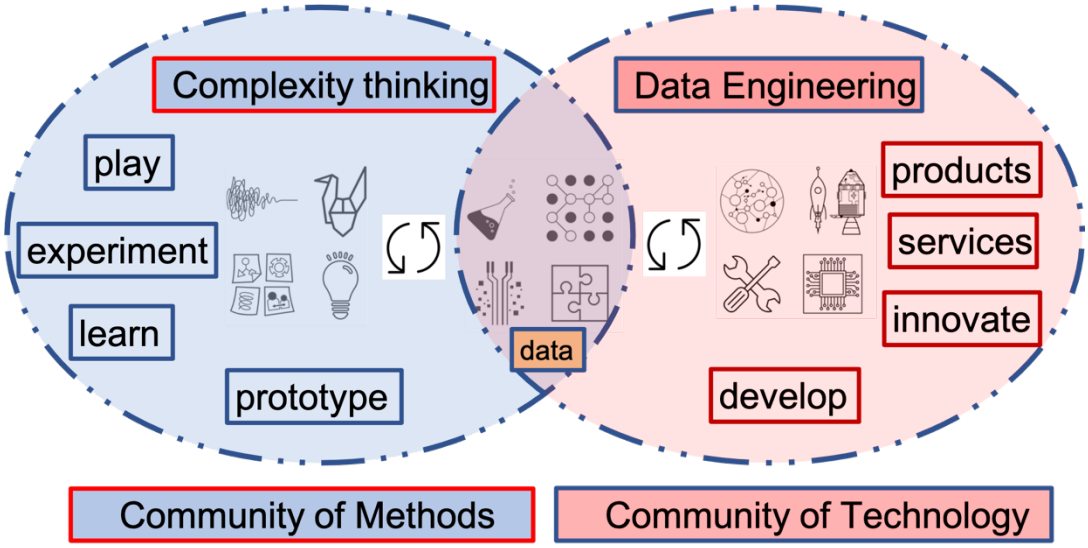
DCC SURF-Pilot -ronde 3- 2022

Creating a Data Fabric through Easy-to-Use Cloud Computing

Opbouwen van kennis & expertise
door *hands-on* disseminatie van AI datatechnologie & data-infrastructuur

Creating a DATA FABRIC

Artwork Adapted from: https://idezo.ch



*Inzetten van **datatechnologie & AI** om te komen tot **Bildung van complexiteitsdenken**, enerzijds, en **het laagdrempelig, verantwoord delen van inzichten voor maatschappelijke vraagstukken die voortvloeien uit het op grote schaal automatisch verzamelen van data**, anderzijds.*

1. Beschrijving van Pilot

Probleem beschrijving

Data Science —DS— gerelateerd praktijkgericht onderzoek binnen het hoger onderwijs wordt bemoeilijkt door twee complicerende factoren:

> **Gebrek aan hands-on kennis over Cloud Computing.**
>
> **Ontbreken aan selfservice toegankelijkheid van data-infrastuur.**

Via het tot stand brengen van een **Data Fabric** kan deze problematiek doorbroken worden (Hollaway et al., 2020). De Data Fabric in deze pilot bestaat uit een Easy-to-Use cloud computing component ***—Jupyter Hub [Lisa-cluster van SURFsara]*** | ***Jupyter Notebook Plug-in [Research Drive]—*** in combinatie met beschikbaar stellen van Data Science tools & knowhow **—HR-brede programma voor AI & Ethiek—**.

Multi-user Jupyter Hubs maken het mogelijk om bachelor studenten **Data Science** (DS) en/of **Computational Thinking** (CT) skills te laten ontwikkelen; zonder zelf iets te moeten downloaden en/of te instaleren. Na te hebben ingelogd op een Jupyter Hub-server, kunnen studenten meteen aan de slag met data zonder gefrustreerd te raken bij het uitzoeken wat te installeren en een te trage computer. Met deze nieuwverworven DS-skills kunnen studenten duurzaam worden ingezet onder supervisie van docentonderzoekers om ruwe researchdata —*aangeleverd door kenniscentra*— te analyseren en/of visualiseren (Beg et al., 2021).

Doel & Data Science Toepassingsdomein

Beoogde doelgroep binnen de hogeschool Rotterdam —**HR**— is docentonderzoeker (inclusief lectoren) en/of bachelor studenten die state-of-the-art Data Science tools willen benutten (zie **Tabel 4.1**).

> **Doel van de pilot is om data-driven onderzoeksinspanningen te faciliteren**
> ***van aan Minoren-onderwijs gekoppelde kenniscentra***
> **door het laagdrempelig introduceren van Data Science tools**
> **als onderdeel van de versterkingsagenda van het HR-brede programma voor AI & ethiek.**

Een overzicht van de deelnemende onderwijseenheden en het aantal betrokken studenten + docent-onderzoekers is weergegeven in **Tabel 1.1**. Het kenniscentrum ***"Creating 010"*** *—Maatschappelijke Digitalisering en Maatschappelijke Transformaties—* heeft deelgenomen via de Minoren **Digital Humans + Data engineering**. Het ***Zorg Tech domein*** *—kenniscentrum* ***"Data Supported Healthcare"****—* heeft deelgenomen als opdrachtgever voor de minor ***Data Engineering***. Het lectoraat Asset Management *—kenniscentrum* ***"Duurzame Havenstad"****— heeft* deelgenomen via een bachelor cursus van de opleiding **Mechanical Engineering**. Tenslotte heeft het kenniscentrum "***Business Innovation***" deelgenomen via de minor **Data-Driven Solutions**. De deelnemende instituten zijn*;* **BSR**: *Instituut Business School Rotterdam*; **CMI**: *Instituut voor Communicatie, Media en Informatietechnologie;* : **EAS**: *Instituut voor Engineering en Applied Science*.

De pilot faciliteert het hands-on toepassen van Natural Language Processing (NLP), Computer Vision (CV) en Statistical Machine Learning [ML] ten behoeve van de analyse & visualisering van researchdata. Hierbij is gebruikgemaakt van ***Computational Thinking (CT) onderwijsmethoden*** in combinatie met populaire open-source ***Data Science tools*** (Beg et al., 2021; Kim & Henke, 2021; Wing, 2006).

Voor disseminatie & kennisverspreiding van Data Science tools + datasets zijn ***GitHub Repositories*** ontwikkeld (zie **Tabel; 2.1**). Deze GitHub Repositories vormen de online component van een HR-brede Data Science Community of Practice (CoP) nodig voor de ontwikkeling van virtuele onderzoeksomgevingen (VREs) waarin docent-onderzoekers/Lectoren & studenten transdisciplinair aan onderzoeksprojecten werken met als verbinden factor research datasets enerzijds, en open-source broncode nodig voor de daadwerkelijke Data Science analyse , anderzijds (Jeffery et al., 2021; Wittenburg & Strawn, 2021).

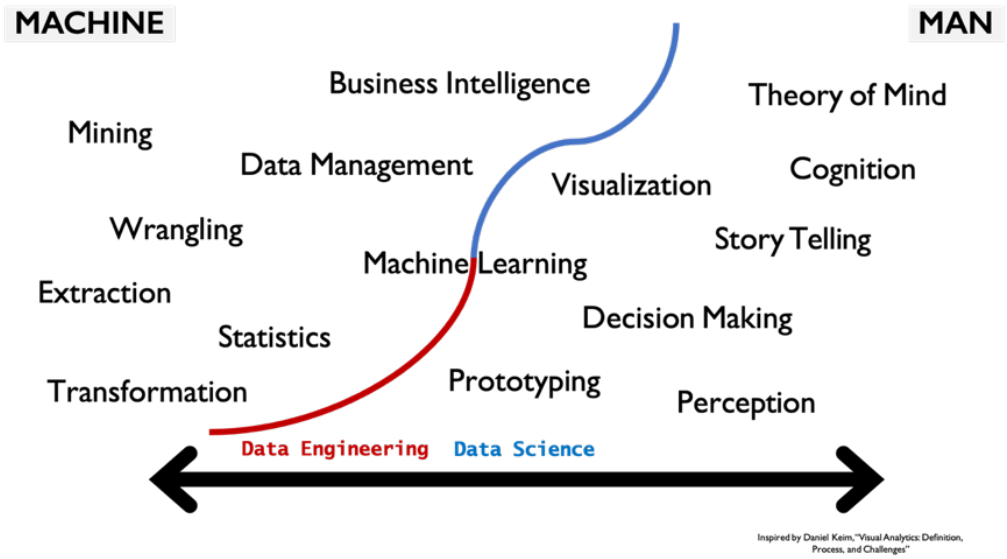
Tabel 1.1 Overzicht onderwijs + onderzoek stakeholders; direct betrokken bij de Data Fabric pilot.

Data Science [DS] Toepassingsdomein	Kenniscentrum	Bachelor Onderwijs Hogeschool Rotterdam [HR]
Computer Vision [CV]	Duurzame Havenstad	Instituut voor Engineering en Applied Science [EAS] Cursus: Mechanical Engineering — Docentonderzoekers: N = 2 — Studenten: N = 33
Natural Language Processing [NLP]	Creating 010	Instituut voor Communicatie, Media en Informatietechnologie [CMI] Minor: Digital Humans — Docentonderzoekers: N = 5 — Studenten: N = 10
Statistical Machine Learning [ML]	Business Innovation	Instituut Business School Rotterdam [BSR] Minor: Data-Driven Solutions — Docentonderzoekers: N = 3 — Studenten: N = 35
Natural Language Processing [NLP]	Data Supported Healthcare	Instituut voor Gezondheidszorg [IVG] Minor: Data Engineering — Docentonderzoekers: N = 3 — Studenten: N = 5

Probleemaanpak

De totstandbrenging van een data fabric

Een data fabric vereenvoudigt de toegang tot Cloud computing om zo selfservice analyse mogelijk te maken van heterogene researchdata. Het is ***agnostisch*** ten aanzien van ***data-omgevingen, -processen, -nut en -locatie***, terwijl tegelijkertijd end-to-end mogelijkheden voor datamanagement worden geïntegreerd (Priebe et al., 2021). Een data fabric (zie **Figuur 1.1**) verbindt ***Data Science knowhow & Computational Thinking skills*** waarover ***Data Scientists*** beschikken met ***infrastructurele kennis*** waarover ***Data Engineers*** beschikken (Keim et al., 2008).



FIGUUR 1.1 Een data fabric integreert twee wetenschappelijke disciplines: Data Science + Data Engineering. Het creëert synthese op het snijvlak tussen deze twee disciplines (rood-blauwe lijn).

Data Fabric Touchpoint Roadmap

De hieronder beschreven Data Fabric ***touchpoints*** zijn grafisch weergegeven in **Figuur 1.3**:

1. De pilot stuurt aan op hands-on disseminatie van ***Data Science (DS)*** methodieken door studenten + lectoren/docentonderzoekers hands-on kennis te laten maken met Cloud Computing gebaseerd op computational thinking paradigma's. Dit in combinatie met het uitbouwen van de professionele ontwikkeling van lectoren/docentonderzoekers.
2. De pilot is ***—selfservice-based—*** vraag-gestuurd. Dat wil zeggen, het HR-brede programma voor AI & Ethiek zal zorgdragen voor de Data Science kennisoverdracht binnen de desbetreffende lectoraten/Minoren om de benodigde Jupyter-notebooks en/of GitHub Repositories te ontwikkelen die kunnen worden benut met de SURF "JupyterHub for Education" Service. Zie https://servicedesk.surfsara.nl/wiki/display/WIKI/JupyterHub+for+education.
3. Elk deelnemende lectoraat/Minor levert de benodigde research data + context beschrijving van de gewenste Data Science Analyse/Visualisatie. Dit zal worden vastgelegd in een ***Definition-of-Done*** in combinatie met een gebruikersovereenkomst. De onderzoeksdata *—inclusief de uitkomster ervan—* worden ondergebracht op Research Drive en beheerd onder supervisie van een datasteward.

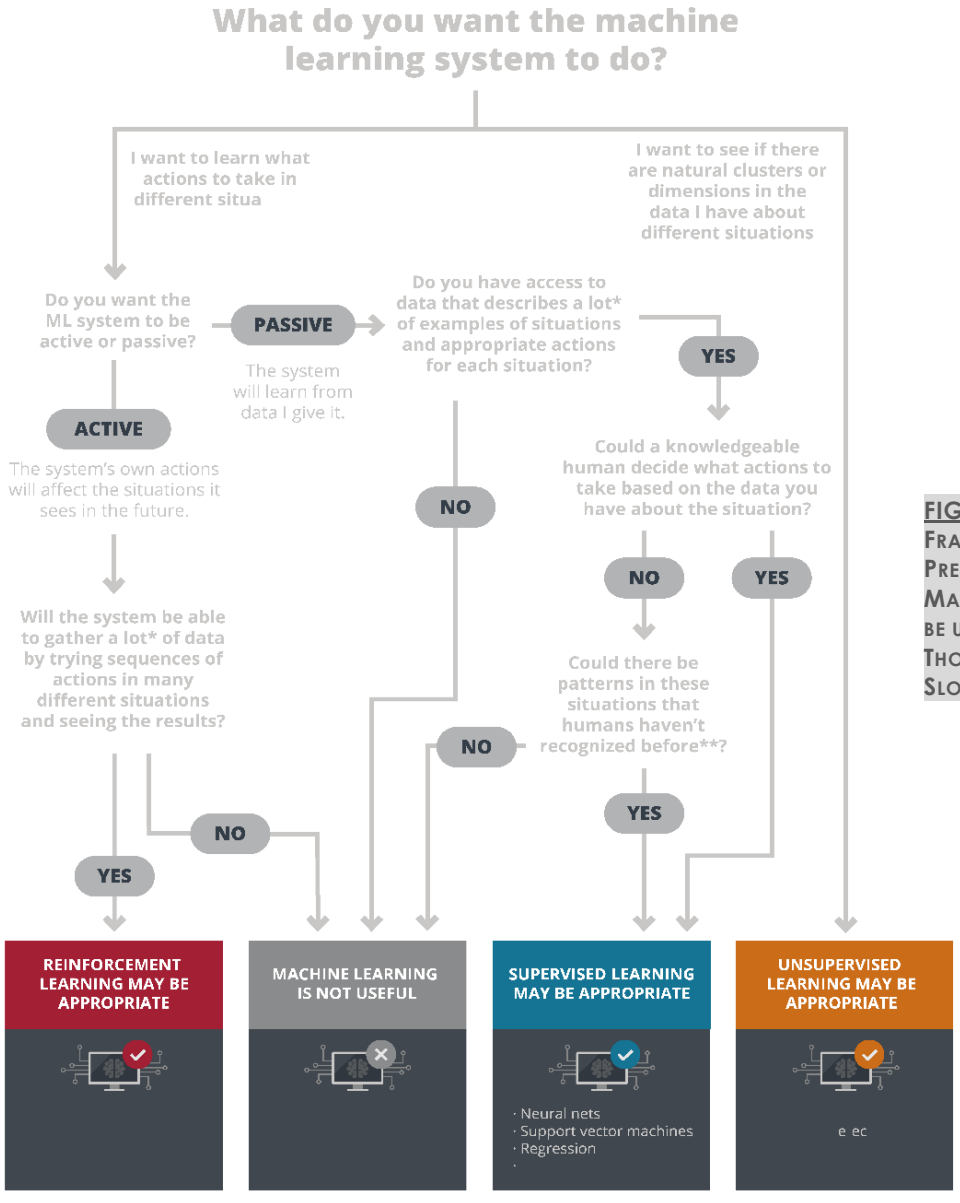
4. Het ***HR-brede programma voor AI & Ethiek*** organiseert vervolgens een brainstormsessie —*met de duur van minimaal een dagdeel*— om te komen tot een plan-van-aanpak voor de benodigde Jupyter-notebooks waarmee de Data Science analyse en/of visualisatie tot stand kan worden gebracht.

5. Voorafgaand aan deze op Agile gebaseerde projectmethode zal een intakegesprek plaatsvinden met de lector/docentonderzoeker door gebruikmaking van een binaire beslisboom (zie **Figuur 1.2**). Op basis hiervan wordt bepaald of er een kennis te kort is bij het betrokken kenniscentra; en zo ja hoe dit te kort ongedaan te maken.

6. **Proof-of-Concept**. Voorafgaand aan de pilot zal een nulmeting gedaan worden bij alle direct betrokken stakeholders met betrekking tot kennis over Data Science & Computational computing. Na afloop van de pilot wordt deze op een Likertschaal-gebaseerde enquête herhaald worden.

7. Het HR-brede programma voor AI & Ethiek vervult de liaison functie naar SURF + andere hogescholen. Het draagt zorg voor een nauwgezet en verantwoord verloop van de pilot en het gebruik van de SURF services. Het zal ook zorgdragen voor de benodigde personele inzet en de financiële afdekking ervan + analyse en verslaglegging van de verkregen uitkomsten.

8. De deelnemende lectoraten hebben zich gecommitteerd aan de hierboven beschreven inspanningsverplichting. Een ***touchpoint roadmap*** hiervan is weergegeven in **Figuur 1.3**.

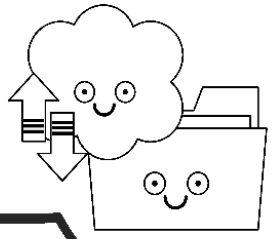


FIGUUR 1.2
FRAMEWORK FOR PREDICTING WHERE MACHINE LEARNING WILL BE USEFUL. SOURCE: THOMAS MALONE MIT SLOAN 2019.

Kenniscentrum stuurt aan op samenwerking
in een virtuele onderzoeksomgeving (VRE)
waarin Lectoren/docentonderzoekers + studenten
transdisciplinair aan onderzoeksprojecten
werken met als basis research data

lecturer/docentonderzoeker

1 Reseach data, context beschrijving
is digitaal beschikbaar via
Research Drive (SURFsara)service

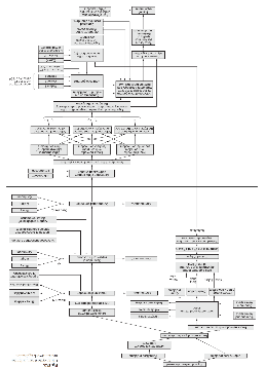


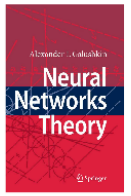
Data Steward

2 Definition of Done
+ Gebruikersovereenkomst

Data Wetenschapper

3 Intakegesprek met de lector en/of
docentonderzoeker door gebruikmaking
van een binaire beslisboom





Data Wetenschapper

4 Brainstormsessie om te komen tot
een plan-van-aanpak voor de
benodigde profesionalisering +
Jupyter-notebooks waarmee
de Data Science analyse en/of visualisatie
tot stand kan worden gebracht.

Data Wetenschapper + JupyterHub SURFsara Teach-the-Teacher project

5 Data Science professionaliseringstraject
+ workshops.

lecturer/docentonderzoeker

6 Data Aanalysis wordt vastgelegd
in stap-voor-stap code-based uitwerkingen
van ruwe researchdata om tot custom-made
betekenisvolle visualisaties te komen en/of
NLP, CV-based machine learning analyses.

Studenten

7 Studenten doorlopen de
Jupyter-notebooks.
Dit op basis van
Computational Thinking
leerdoelen +workshops.

Data Wetenschapper

8 Proof-of-Concept.
Voorafgaand aan de pilot zal een nulmeting
gedaan worden bij alle betrokken stakeholders
(docentonderzoeker/lector/student)
met betrekking tot kennis over Data Science
& Computational computing.

Na afloop van de pilot zal deze op een
Likertschaal-gebaseerde enquête herhaald worden.
Daarnaast zal steekproefsgewijs een aantal
computational thinking skills worden getest.

FIGUUR 1.3 ROADMAP FOR IMPLEMENTING DATA SCIENCE TOOLS & COMPUTATIONAL THINKING SKILLS.

2. Resultaat & Impact

Lessons Learned

Het tot stand brengen van een ***Data Fabric*** binnen het bachelor onderwijs is problematisch door een gebrek aan primaire software skills. Hiermee wordt bedoeld basale softwarevaardigheden die nodig zijn om taken uit te voeren die het schrijven van broncode mogelijk maken; nodig voor het opschonen en analyseren van research-datasets. Het goed beheersen van Data Science tools om bestaande broncode te kunnen aanpassen is een vereiste waaraan vaak niet voldaan wordt. Het bijbrengen van softwarevaardigheden maakt geen integraal onderdeel uit van de onderbouw-fase van ***—niet technische—*** bacheloropleidingen. Een onderliggend probleem is dat open-source, Integrated Development Environment **[IDE]** software ***—zoals Anaconda | Visual Studio Code | Rstudio—*** in een hoog tempo wordt doorontwikkeld waarbij achterwaartse compatibiliteit — *backward compatibility*— vaak niet mogelijk is omdat software-updates vaak gepaard gaan met aanpassingen aan de onderliggende ***"cloud computing"*** infrastructuur. Deze complicerende factor wordt ook benoemd in het "Need Analysis Rapport, 2021" een deliverable geproduceerd door "the European Software skills Alliance" **[ESSA]**. Met andere woorden, software skills moeten voortdurend worden bijgehouden om toepasbaar te blijven.

> ***Bacheloropleidingen —niet direct gerelateerd aan het technische domein— besteden weinig of geen aandacht aan het onderwijzen en/of stimuleren van softwarevaardigheden.***

Computational thinking ***—het praktisch en creatief inzetten van Data Science tools om research data met computertechnologie te analyseren & visualiseren—*** is pas effectief wanneer dit binnen de context van praktijkgericht onderzoek wordt toegepast. Dit is consistent met wat in de literatuur wordt omschreven als **Project Based Learning** [PBL] (Berikan & Özdemir, 2020; Shin et al., 2021).

> ***Het ontwikkelen van een Data Fabric heeft de grootste kans van slagen binnen de context van een toepassingsdomein waar docentonderzoekers —domein experts — samenwerken met Minor studenten onder begeleiding van een ervaren Data Scientist.***

GitHub vormt anno 2022 de grootste open source community voor hosting van broncode, met meer dan 80 miljoen gebruikers. Dit cloud-platform is eigendom van Microsoft dat ***Git —in de vorm van GitHub Repositories—*** als onderliggende softwaretechnologie gebruikt. Git is een gedistribueerd versiebeheersysteem, waardoor meerdere softwareontwikkelaars gezamenlijk aan broncode kunnen werken (Borges et al., 2016).

> ***GitHub Repositories in combinatie met Jupyter Hub vormen een adequaat vehicle voor zowel kennisdeling over Data Science als ook het benutten & disseminatie van research data.***

GitHub Repositories vormen zo een open-source opslagplaats voor broncode + links naar research datasets. De populariteit van GitHub in combinatie met een Easy-to-Use cloud computing component ***—Jupyter Hub—*** vormen een krachtig instrument voor het hoger onderwijs om een data fabric tot stand te brengen (Cardoso et al., 2019; National Academies of Sciences & Medicine, 2018; Tu et al., 2022; Weiss, 2017).

Results

In het kader van deze pilot zijn ***GitHub Repositories*** ontwikkeld (zie **Tabel; 2.1**). Conceptversies van de Repositories kunnen worden uitgeprobeerd via URL: https://github.com/robvdw?tab=repositories. Medio Januari 2023 zullen deze online komen via de URL: https://github.com/HR-DATA-FABRIC, na afronding van de deelnemende Minoren.

Tabel 2.1 Voorbeelden van de ontwikkelde open-source GitHub Repositories.

Data Science [DS] Toepassingsdomein	Developer	Repository
Neural Networks [NN] Machine Learning [ML] Deep Learning [DL]	Programma AI & Ethiek	Matlab Regression Learner IDE: `Matlab` Software Libraries: `Machine learning`, `Deep Learning` Data-sets: `yacht_hydrodynamics.csv`, `blood_donation.csv`, `bodyfat.csv`, `breast_cancer.csv`, `combined_cycle_power_plant.csv`, `concrete_properties.csv`, `creditcard-fraud.csv`, `fault_detection.csv`, `iris_flowers.csv`, `power-plant-gas-emissions.csv`, `telecommunications_churn.csv`, `tree_wilt.csv`
Natural Language Processing [NLP]	Programma AI & Ethiek	Decision Support Systems [DSS] in Allied Healthcare IDE: `Anaconda Navigator`, `Visual Studio Code` Software Libraries: `Python 3.9.15`, `NLTK`, `spaCy`, `Cupy`, `Pytorch`, `Tensorflow`, `ipywidgets`, `jupyterlab` Data-sets: `healthcare free-texts.docx`
Natural Language Processing [NLP]	Programma AI & Ethiek	Digital Humans IDE: `Jupyter Hub`, `Anaconda Navigator`, `Visual Studio Code` Software Libraries: `Python 3.9.15`, `NLTK`, `spaCy`, `numpy`, `pandas`, `Chatterbot`, `Scikit-Learn`, `Rasa` Data-sets / Models: `nl_core_news_sm`, `nl_core_news_md`, `nl_core_news_lg`

Elke *GitHub Repository* bevat een reeks van Jupyter Notebooks. Dit kunnen Data Science workshops zijn en/of gedetailleerde stap-voor-stap broncode uitwerkingen van research-data analyses in combinatie met (geanonimiseerde) online beschikbare research datasets. Met Binder ***—Binder-ready Repository—*** is het mogelijk de broncode uit te proberen in de vorm van een Jupyter Notebook via een standaard webbrowser.

Follow-Up

Op termijn kunnen de GitHub Repositories **—Tabel 2.1—** worden gekoppeld aan EduBadges als gebleken is dat ze effectief zijn in het bevorderen van Data Science Skills. Zo kunnen Nederlandse Hogescholen EduBadges aanbieden binnen opleidingen die niet direct gerelateerd zijn aan het technische domain, zonder zelf te moeten investeren in kennis en knowhow over Computational Thinking educatie.

Het installeren van additionele software libraries is vaak problematisch. Dit kan ondervangen worden door gebruik te maken van Docker images. Docker is een software-engine die gebruik maakt van containertechnologie. Een container, die wordt uitgevoerd op een online-hostbesturingssysteem, is een ***geïsoleerde software-omgeving*** waarin code en alle bijbehorende afhankelijkheden ***—software libraries—*** zijn verpakt. Daardoor kunnen specifieke Dat a Science tools snel en betrouwbaar worden gebruikt. Zo voorkom je dat de eindgebruiker zelf software libraries moet gaan installeren om broncode werkend te krijgen. Nadeel is wel dat Docker containers expliciet moeten worden gemount met de harddrive van het hostbesturingssysteem om bestanden te kunnen lezen en schrijven.

Impact op praktijkgericht onderzoek

De totstandbrenging van een Data Fabric binnen het ZorgTech domein heeft een versnelling van de ontwikkeling van data gedreven toepassingen ***—in het bijzonder op het gebied van Elektronische Patiënten Dossiers (EPD's)—*** teweeggebracht. Het is daarmee de primaire aanjager van vrije-tekst ontsluiting door gebruikmaking van Natural Language Processing (NLP) technologie binnen de hogeschool Rotterdam.

State-of-the-Art (SotA) AI die human-level performance benaderd is tot nu toe alleen mogelijk met zeer grote datasets in combinatie met High Performance Computing (Han et al., 2021). Het vermogen om met ***relatief kleine*** research datasets ***Deep learning AI-modellen*** *te benutten* kan een grote impact hebben op praktijkgericht onderzoek binnen het hoger onderwijs.

Het gebruik van kleine datasets voor het trainen AI-modellen heeft als groot voordeel dat in korte tijd ***—dagen i.p.v. maanden—*** human-level performance kan worden bereikt. Dit is nog niet mogelijk met open-source software-frameworks zoals Pytorch en/of Tensorflow + Keras. Er zijn wel propriétaire **"multi-paradigm programming languages and numeric computing environments"** beschikbaar die gebruik kunnen maken ***van pre-trained deep-neural-networks***, zoals ***Matlab, Mathematica***.

Het HR-brede programma voor AI & Ethiek heeft als eerste stap een GitHub Repository ontwikkeld ***—MATLAB REGRESSION LEARNER—***. Ons streven is om het trainen van *"pre-trained deep-neural networks"* ***—met custom research datasets—*** te gaan implementeren zodat Kenniscentra hierover kunnen beschikken . Als follow-up zal worden getest hoe de ***Matlab IDE*** als ***JupyterHub kernel*** kan worden gebruikt. Groot voordeel van het werken met Matlab is de mogelijkheid om broncode om te zetten in ***standalone applicaties*** voor zowel ***Linux, Windows*** als ***MacOS*** platforms

3. Diensten

Jupyter Hub

Jupyter Notebook is een browser-based Application Programming Interface (API) waarmee je de werking van broncode interactief kunt uitproberen. Een **'hub'** van notebooks wordt gedeeld door meerdere gebruikers ***—Multi user—***; wat het geschikt maakt voor zowel data-analyse als onderwijs. Notebooks worden veelal toegepast voor opschoning en transformatie van datasets, numerieke simulaties, statistisch modelleren, data visualisatie en machine learning.

Vier verschillende HR-instituten hebben met enige regelmaat gebruik gemaakt van de ***JupyterHub voor educatie* service** *—zoals weergeven in* **Tabel 1.1**— ten behoeve van hun onderwijstaak. Door gebruik te maken van Jupyter notebooks via de SURF Jupyter Hub is het mogelijk gebleken om bachelor studenten Data Science (DS)/ Computational Thinking (CT) skills te laten ontwikkelen zonder zelf iets te moeten downloaden en/of te installeren. Installatie van de benodigde software libraries wordt verzorgd door IT-professionals van SURFsara of door docentonderzoekers. Na te hebben ingelogd op de JupyterHub-server, kunnen studenten dan meteen aan de slag met data zonder gefrustreerd te raken bij het uitzoeken wat te installeren en een te trage computer. Met deze nieuwverworven DS-skills kunnen studenten duurzaam worden ingezet om ruwe researchdata te analyseren.

Research Drive

De Hogeschool Rotterdam (HR) benut Research Drive om onderzoeksgegevens op een veilige en verantwoorde manier op te slaan en/of te kunnen delen met projectleden en externe partijen. Per project is er een gegevens-verantwoordelijke (docentonderzoeker/lector) die verantwoordelijk is voor het beheer van de data en het verder delen met projectleden en externe partijen.

Meer in het bijzonder, de **Jupyter Notebook plug-in** is uitgeprobeerd om ***Natural Language Processing —NLP–*** research van klinische vrije-teksten verantwoord en effectief te kunnen analyseren. Dit praktijkgerichte onderzoeksproject ***—Decision Support Systems [DSS] in Allied Healthcare—*** richt zich op het verbeteren van de geallieerde gezondheidszorg door het toepassen van ***state-of-the-art (SOTA) AI-technologieën***. Het is een transdisciplinair samenwerkingsverband tussen het Instituut voor Gezondheidszorg ***—IVG—***, de ***Minor Data Enginering*** en ***het Prometheus Data-Lab*** van de Hogeschool Rotterdam ***—RUAS—.*** Ondersteuning wordt gegeven door het RUAS Programma voor AI & Ethiek, het ***Digital Competence Centre*** (DCC) voor Praktijkgericht Onderzoek ***—DCC SURF-pilot project—*** en het ***RUAS Data Supported Healthcare team —Zorgtech010 data-science unit—.***

4. Betrokkenen en Inzet

Rollen & Functies (Taken)

Het HR-brede programma voor AI & ethiek is hoofdverantwoordelijk voor de uitvoer en monitoring van de pilot. Dit programma levert de benodigde Data Science knowhow en de daarbij benodigde bemensing.

De aan de pilot verbonden kenniscentra & lectoraten (**zie Tabel 1.1**) dragen zorg voor een **context en datarijke omgeving *—opvraagbaar als GitHub Repositories inclusief datasets—*** en vormen een directe koppeling met (docent)- onderzoekers + bachelor studenten via hun Minoren onderwijs. Voor een overzicht van de rolverdeling & taken zie **Tabel 4.1**.

Expertise vanuit SURF geleverd

De pilot heeft hoofdzakelijk gebruik gemaakt van de Lisa support van de SURFSARA-servicedesk. Zie **Appendix III**: *SURF Servicedesk Email verkeer aangemaakte tickets*.

Betrokken lectoraat(en)/kenniscentra & Minoren (zie Tabel 1.1)

Het lectoraat ***Asset Management*** *—kenniscentrum* ***"Duurzame havenstad"***— heeft deelgenomen via de Minor **Mechatronica**. Het kenniscentrum ***"Creating 010"*** *—Maatschappelijke Digitalisering en Maatschappelijke Transformaties*— heeft deelgenomen via de Minoren **Digital Humans + Data engineering**. Het ***Zorg Tech domein*** *— kenniscentrum* ***"Data Supported Healthcare"***—heeft als opdrachtgever voor de minor **Data engineering** deelgenomen**.** Kenniscentrum "***Business Innovation***" deelgenomen via de minor **Data-Driven Solutions**.

Tabel 4.1 Overzicht Rollen &taken inclusief de uitvoerende HR organisatorische eenheden.

Organisatorische Eenheid	Rol	taak
Dienst Onderwijs & Ontwikkeling [OeO]	Projectcoördinator & Strategisch Adviseur	Overziet de werkzaamheden van de projectmanager en de projectteams. Is de HR-liaison voor "het geven van ruchtbaarheid naar andere hogescholen".
Bedrijfsbureau Ondersteuning Kenniscentra (OKC) Hoofddocent Instituut voor Communicatie, Media en Informatie-technologie (CMI).	Projectmanager	Is de Tech lead Data Scientist van het HR-brede programma AI & Ethiek en daarmee de hoofdverantwoordelijke voor de inhoudelijke voortgang van het project en de eindrapportage naar SURF. Ontwikkelt een planning, stelt de projectteams samen en beheert de werklast gedurende de hele levenscyclus van de pilot. Een projectteam bestaat uit een of meerdere docentonderzoeker(s) + studenten van de desbetreffende minor en/of onderwijseenheid. Draagt zorg voor de professionele ontwikkeling —Data Science professionaliseringstraject— van de betrokken onderzoekers, inclusief de implementatie van de benodigde Jupyter-notebooks, nulmeting + validatie van de Proof-of-Concept analyse ervan direct na afloop van de Pilot.
Kenniscentrum [KC]	Docent-Onderzoeker	Teamleider aan wie het projectteam rapporteert. Ontvangt inhoudelijke begeleiding van de projectmanager m.b.t. data science knowhow — kennisvermeerdering+ vaardigheidsverbetering— in combinatie met de implementatie ervan in de vorm van Jupyter Notebooks/GitHub Repositories. Levert de benodigde datasets en/of bepaald in samenspraak met de projectleider welke opensource dataset gehanteerd zal worden + beschrijving van —Doel van het onderzoek— de gewenste data science Analyse/Visualisatie.
Minor	Bachelor Student	Ontvanger van de aangeboden data science basis technieken & tools m.b.v. de SURF "JupyterHub for Education" Service. Doorloopt een gedetailleerde stap-voor-stap code-based uitwerking van ruwe researchdata. Dit op basis van Computational Thinking leerdoelen. Deelt bevindingen & uitkomsten via een custom-made GitHub Repository.

5. Aandachtspunten

Een belangrijk punt van aandacht is dat een streaming/en of online koppeling van externe-datasets met de SURF JupyterHub niet mogelijk is. De standaardoplossing is dat lokaal datasets geüpload moet worden naar de harde-schijf van het hostbesturingssysteem. Dus het verplaatsen van lokaal beschikbare datasets is nodig om ze te kunnen benutten binnen Jupyter Notebooks van de Hub. Dit is ongewenst omdat data zo verspreid wordt over meerdere plaatsen tegelijkertijd. Vaak is dit niet toegestaan i.v.m. AVG-compliance en/of non-disclosure gebruikers-overeenkomsten.

Een mogelijk oplossing is Jupyter Notebook plug-in voor de SURF Research Drive Dashboard. Echter deze ***Jupyter Data Science Notebook plug-in*** is nog in een pilot-fase. Dat wil zeggen, binnen de pilot is intensief met de plug-in geëxperimenteerd maar de werking ervan is nog niet stabiel genoeg om betrouwbaar en reproduceerbaar Data Science Tools te kunnen benutten.

6. Geraadpleegde Literatuur

1. Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N. M., & Fangohr, H. (2021). Using Jupyter for Reproducible Scientific Workflows. *Computing in Science & Engineering*, *23*(2), 36-46. https://doi.org/10.1109/MCSE.2021.3052101
2. Berikan, B., & Özdemir, S. (2020). Investigating "Problem-Solving With Datasets" as an Implementation of Computational Thinking: A Literature Review. *Journal of Educational Computing Research*, *58*(2), 502-534. https://doi.org/10.1177/0735633119845694
3. Borges, H., Hora, A., & Valente, M. T. (2016). Understanding the Factors That Impact the Popularity of GitHub Repositories. *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 334-344. https://doi.org/10.1109/ICSME.2016.31
4. Cardoso, A., Leitão, J., & Teixeira, C. (2019). Using the Jupyter Notebook as a Tool to Support the Teaching and Learning Processes in Engineering Courses. In M. E. Auer & T. Tsiatsos, *The Challenges of the Digital Transformation in Education* ICL 2018. Advances in Intelligent Systems and Computing, vol 917. Springer, Cham https://doi.org/10.1007/978-3-030-11935-5_22.
5. Christensen, C. M. (1997). *The innovator's dilemma : when new technologies cause great firms to fail*. Harvard Business.
6. Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., & Zhang, L. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250. https://doi.org/https://doi.org/10.48550/arXiv.2106.07139
7. Hollaway, M. J., Dean, G., Blair, G. S., Brown, M., Henrys, P. A., & Watkins, J. (2020). Tackling the Challenges of 21st-Century Open Science and Beyond: A Data Science Lab Approach. *Patterns*, *1*(7), 100103. https://doi.org/https://doi.org/10.1016/j.patter.2020.100103
8. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 154-175). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7
9. Kim, B., & Henke, G. (2021). Easy-to-Use Cloud Computing for Teaching Data Science. *Journal of Statistics and Data Science Education*, *29*(sup1), S103-S111. https://doi.org/10.1080/10691898.2020.1860726
10. Looman, A. G. E., de Kloet, M., Yang, S., & Klienbannink, L. R. (2022). *Towards data literacy in higher education*. Kenniscntrum Business Innovation, Hogeschool Rotterdam. Retrieved from https://www.hbo-kennisbank.nl/details/sharekit_hr:oai:surfsharekit.nl:b5797e26-fa07-458f-a048-af3683e3cfb3?q=data+driven
11. National Academies of Sciences, E., & Medicine. (2018). *How People Learn II: Learners, Contexts, and Cultures*. The National Academies Press. https://doi.org/doi:10.17226/24783

12. Priebe, T., Neumaier, S., & Markus, S. (2021, 15-18 Dec. 2021). *Finding Your Way Through the Jungle of Big Data Architectures* 2021 IEEE International Conference on Big Data (Big Data), https://doi.org/10.1109/BigData52589.2021.9671862
13. Shin, N., Bowers, J., Krajcik, J., & Damelin, D. (2021). Promoting computational thinking through project-based learning. *Disciplinary and Interdisciplinary Science Education Research*, *3*(1), 7. https://doi.org/10.1186/s43031-021-00033-y
14. Tu, Y.-C., Terragni, V., Tempero, E., Shakil, A., Meads, A., Giacaman, N., Fowler, A., & Blincoe, K. (2022). GitHub in the Classroom: Lessons Learnt. *Australasian Computing Education Conference*, 163-172. https://doi.org/10.1145/3511861.3511879
15. Weiss, C. J. (2017). Scientific Computing for Chemists: An Undergraduate Course in Simulations, Data Processing, and Visualization. *Journal of Chemical Education*, *94*(5), 592-597. https://doi.org/10.1021/acs.jchemed.7b00078
16. Wing, J. M. (2006). Computational thinking. *Commun. ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215
17. Wittenburg, P., & Strawn, G. (2021). Revolutions Take Time. *Information*, *12*(11), 472. https://www.mdpi.com/2078-2489/12/11/472

APPENDIX

SURF Servicedesk: gecondenseerd overzicht aangemaakte tickets.