Wat is een DataLab? & Hoe werkt het?

DataLab

What-if? Scenario

Hilde, verpleegkundige en docentonderzoeker aan Hogeschool Rotterdam, analyseert gesprekken met kankerpatiënten om hun onvervulde zorgbehoeften te identificeren. Ze dacht dat computers deze gecompliceerde taak niet van haar zouden kunnen overnemen, totdat ze de *chatbot* ChatGPT ontdekte. Dit motiveerde haar om generatieve AI te gaan gebruiken voor het geautomatiseerd samenvatten van patiënt-zorgverlener dialogen. Maar ze beschikt niet over de technische kennis of infrastructuur die nodig is om deze *"data science use case"* op te lossen.

Om na te gaan of haar ongestructureerde dataset, met meer dan 1100 gesprekken, als input kan dienen voor generatieve AI-modellen, neemt ze contact op met het HR-DataLab Healthcare. Hilde krijgt op basis van een intakegesprek toegang tot een *"trusted research environment"* (TRE): *een beveiligd digitaal platform waar datawetenschappers en domeinexperts gezamenlijk kunnen werken aan data science use case oplossingen met waarborging van privacy.*

De benodigde ICT-infrastructuur wordt geleverd via een publieke clouddienst; in dit geval SURF Research Cloud. Hilde kan zo via de TRE beschikken over digitale *"healthcare workspaces"* met voldoende rekenkracht, opslagruimte en researchsoftware. Het stelt haar in staat om generatieve AI verantwoord & betrouwbaar uit te proberen onder het toeziend oog van ervaren *"data scientists"* en *"data engineers"*. Bovendien krijgt Hilde vaardigheidstrainingen in *"privacy protection"*, *"cybersecurity"* en *"datamanagement"*.

Na één semester in het Datalab actief te zijn geweest zijn Hilde en haar collega docenten ervan overtuigd dat het zich bij uitstek leent voor het ontwikkelen en testen van didactische innovaties zoals Edubadges: *digitale bewijzen van geleerde vaardigheden*. Hilde wil 2de-jaars verpleegkundestudenten laten ervaren hoe je generatieve AI betrouwbaar kunt inzetten voor deskresearch. Jupyter Notebooks, aangeraden door het DataLab-team, vormen de basis voor deze *AI & Chatbots* leerervaring.

Een Jupyter notebook is een web-based editor waarmee je computercode, opgemaakte tekst en visualisaties in één document kunt combineren. Dit interactieve format is uitermate geschikt voor het ontwikkelen van digitale geletterdheid. Studenten kunnen hierdoor experimenteren met het schrijven van code door trial-and-error, net zo lang tot het werkt. Het DataLab Rotterdam faciliteert het inregelen van een veilige, afgeschermde *"Jupyter Hub"* ICT-infrastructuur. Na te hebben ingelogd op de Jupyter Hubserver, kunnen studenten meteen aan de slag zonder gefrustreerd te raken bij het uitzoeken wat te installeren en een te trage computer. De inhoudelijke beschrijving van de Edubage *AI & chatbots* is online beschikbaar als Github repository. Deze *"Do-it-Yourself"* benadering bevordert een open cultuur van samenwerking en kennisuitwisseling.

De hier beschreven Data Science use case & Edubadge implementatie vormen een schoolvoorbeeld van hoe onderwijs en onderzoek elkaar versterken door gebruikmaking van een Datalab.

Wat is het?

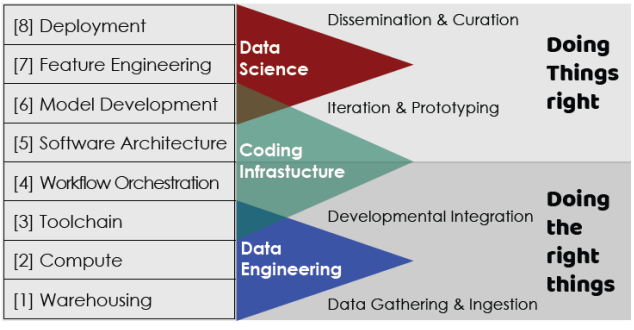
Datalabs bieden professionele samenwerkingsverbanden die eindgebruikers van data helpen oplossingen te ontwikkelen voor data science (DS) use cases met behulp van ICT-infrastructuur.

De vraag naar DS-vaardigheden is sterk toe genomen door de verregaande digitalisering van onze maatschappij: *de overgang van fysieke informatiedragers naar digitale vormen die online genereert en bewerkt kunnen worden*. Denk bijvoorbeeld aan de explosieve groei in het gebruik van generatieve AI in de vorm van populaire Chatbots zoals ChatGPT, Gemini en Copilot. De strikte scheiding tussen de tastbare, analoge wereld versus de virtuele, digitale wereld vervaagd hierdoor. Het gevolg is dat er steeds meer en sneller ongestructureerde data, digitaal beschikbaar komt in de vorm van vrije teksten, afbeeldingen, audio- en video-opnamen. Door deze transitie is kennis over DS in de vorm van het verantwoord toepassen van informatietechnologie van groot maatschappelijk belang.

Hogeschool Rotterdam investeert in een DataLab-ecosysteem gericht op het ondersteunen van een viertal transitieopgaven, zoals geformuleerd in de *"strategische agenda 2023-2028"*: (1) duurzame delta, (2) toekomstbestendige economie, (3) vitale gemeenschap en (4) slimme en sociale stad. Doel is het tot stand brengen van een *"federated data fabric"* voor de regio Rotterdam, waarbij de nadruk ligt op gestandaardiseerd en verantwoord databeheer.

Hoe werkt het?

Om aan de slag te gaan met data moet je kunnen beschikken over een digitale *workspace*, inclusief de benodigde infrastructuur (hardware + software). Datalabs gebruiken hiervoor de DS-stack. Dit conceptuele hulpmiddel vindt zijn oorsprong in software development. Het is gebaseerd op het principe van *"separation of concerns"* (SoC). Stacks zijn opgebouwd uit verticaal gescheiden lagen, elk gericht op een specifieke taak. De onderste lagen representeren hardware gerelateerde taken zoals het bewaren van datasets en rekenkracht allocatie, terwijl de hoger gelegen lagen gebruikersgerichte taken representeren, zoals feature engineering en het toepassen van datagedreven modellen, waarvoor software nodig is.



De onlinecursus: "Building a data-driven infrastructure" toont hoe je, als eindgebruiker van *"Product Shipment"* data, een DS-stack kunt samenstellen. Een Jupyter Notebook implementatie van deze *"logistics DS-usecase"* is online beschikbaar via Google Colab. Een

Het AI & ETHIEK initiatief [https://hr.nl/ai] zet zich in voor het bevorderen van verantwoorde toepassen van AI-technologie. Voor meer informatie over DataLabs neemt contact op met de Tech Lead programma AI & Ethiek