**Jorn Lelong** / j.lelong@student.rug.nl

# Data Science

## Een frisse wind door fysica, genetica en geschiedenis

Weinig onderzoekers die met zoveel verschillende wetenschapsgebieden in aanraking komen als de data scientists van het CIT, het Centrum voor Informatie Technologie van de RUG. Naast onderwijs in Data Science aan de IT Academy Noord-Nederland en Campus Fryslân en eigen projecten, zoals met het CBS, voert het team een dertigtal projecten uit voor de meest uiteenlopende afdelingen van de universiteit – denk aan genetica, fysica en geschiedenis.

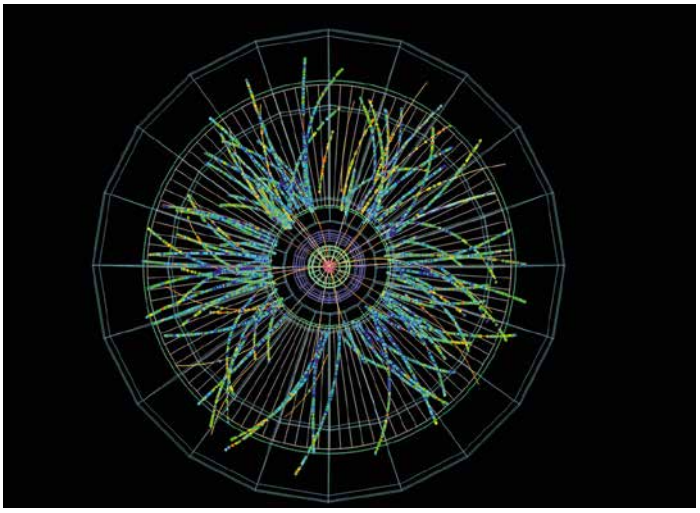
In november 2016 blies de RUG het team Data Science leven in, op vraag van de IT-strategiecommissie van de universiteit. Jonas Bulthuis, die als consultant voor de CIT-afdeling Research and Innovation Support al langer had opgemerkt dat data science een serieuze opmars maakte, nam hierin het voortouw. Er kwam budget voor het aantrekken van twee data scientists en de IT-strategiecommissie koos enkele beloftevolle projecten uit.

### Verschillende achtergronden

Sindsdien is er een jaarlijkse call for proposals, waarvoor onderzoekers van alle RUG-faculteiten een voorstel kunnen inleveren. Steeds meer faculteiten ontdekken hierdoor de mogelijkheden die data science kan bieden, zegt Dimitrios Soudis, een van de data scientists van het CIT. “De eerste ronde hadden we 14 aanmeldingen, bij de tweede al 40. Steeds meer mensen komen naar ons met de vraag: hoe kunnen jullie ons helpen?”

Toen Dimitrios anderhalf jaar geleden begon, bestond het team Data Science slechts uit 3 medewerkers. Dat aantal is ondertussen gegroeid naar 10 mensen - met elk hun eigen expertise. “Als behoorlijk klein team is het noodzakelijk dat we elkaar goed aanvullen”, legt Jonas Bulthuis uit.

Daarom vind je binnen het team Data Science zowel mensen met een achtergrond in econometrie, fysica of wiskunde. Bij de verdeling van projecten is er dan ook aandacht voor de achtergrond en interesse van elk teamlid. “We willen mensen projecten geven waar ze voeling mee hebben, waar ze energie uithalen. Dan ben je gelijk veel gemotiveerder.”



*Visualisatie van banen die worden afgelegd door subatomaire deeltjes*

### Een baanbrekende speurtocht naar gluonen

Dat zie je bijvoorbeeld bij het dataproject voor het Kernfysisch Versneller Instituut (KVI). Dit project kwam in handen van Leslie Zwerver en Cristian Marocico, vanwege hun affiniteit met het onderwerp. Cristian heeft bijvoorbeeld zelf een achtergrond in fysica, wat volgens kernfysicus Johan Messchendorp, opdrachtgever voor dit project, erg van pas komt. “Als fysicus heb je een bepaald verwachtingspatroon, wat helpt om te zoeken naar relevante informatie. Of als ik praat over energiebehoud of impulsen, weet Cristian gelijk waar ik het over heb.”

Al enkele jaren legt Johan Messchendorp zich toe op deeltjesfysica, een onderzoeksveld binnen de fysica dat zich richt op subatomaire deeltjes,

vervolg van pagina 7 Data Science

zoals protonen en neutronen. Die subatomaire deeltjes komen vaak alleen vrij door botsingen in deeltjesversnellers. Maar aangezien die deeltjesversnellers steeds meer botsingen per seconde aankunnen, wordt de hoeveelheid data die verzameld wordt ook steeds groter. Klassieke onderzoeksmethoden stuiten hierdoor op hun limieten. “Dan moet je wel iets slims bedenken om met die complexe data aan de slag te gaan”, vertelt Johan Messchendorp.

Daarom riep hij de hulp in van het team Data Science. Met machine learning, een onderzoeksmethode binnen artificiële intelligentie die toelaat om grote hoeveelheden data te analyseren en daarin patronen bloot te leggen, kunnen data scientists Leslie en Cristian interessante deeltjes filteren van de enorme hoeveelheid achtergrondinformatie. Dat moet allemaal online, want geen computer kan de data aan die hun algoritme erdoor jaagt.

Concreet willen ze machine learning gebruiken om gluonen te reconstrueren. Gluonen zijn kerndeeltjes die erg veel informatie bevatten, maar erg moeilijk te zien zijn. Ze komen namelijk alleen in combinatie met andere deeltjes voor. Daarom wil Messchendorp protonen met antiprotonen laten botsen om glueballs te creëren, exotische toestanden die gluonen bevatten. Toch blijft het zoeken naar een speld in een hooiberg: zo’n exotische toestand komt slechts een in de 100 miljoen botsingen voor.

## Nieuwe wind

Het onderzoek naar gluonen zelf moet voorlopig nog even wachten. In afwachting dat een gloednieuwe, hypersnelle deeltjesversneller in Darmstadt afgewerkt wordt, baseren Cristian en Leslie zich op experimenten bij een deeltjesversneller in China, waar je lagere snelheden hebt. “Die data is nog behapbaar. Daar kunnen we mee spelen om de techniek te optimaliseren en deep learning te vergelijken met de traditionele methoden”, zegt Messchendorp. De eerste tekenen zijn alvast positief. In enkele maanden hebben Leslie en Cristian een algortime ontwikkeld dat zich staande houdt tegenover de klassieke methoden.

Volgens Messchendorp zijn zulke vergelijkende tests noodzakelijk om een nieuwe wind door de fysica te laten waaien. “De fysica-community

*DNA*

houdt zich toch graag vast aan bekende formules. Als je dan met een nieuwe methode komt, krijg je al snel de vraag: werkt dat wel? Dat moet je overtuigend kunnen aantonen op basis van bestaande data, alvorens het op complexere data toe te passen.”

## Een nieuwe screeningtool voor erfelijke ziektes

Ook data scientist Dimitrios Soudis merkt dat de academische wereld zich doorgaans traag openstelt voor nieuwe technieken. “Academici willen voornamelijk in wetenschappelijke tijdschriften publiceren. Als die nog geen interesse hebben in bepaalde nieuwe technieken, dan wacht de academicus ook met zich daarin te verdiepen. Daarom duurt het altijd enige tijd voordat innovaties van het ene gebied overslaan op het andere.”

Het Genomics Coordination Centre, de afdeling Bio-informatica van het UMCG, vormt daar in zekere zin de uitzondering op. Al meer dan tien jaar doet de afdeling uitgebreid dataonderzoek. Zo was het UMCG in 2014 betrokken bij Genome of the Netherlands, een gigantisch nationaal dataproject.

Toch is er een groot verschil tussen data verzamelen en meer geavanceerde technieken zoals machine learning. “Door onze jarenlange dataverzameling kunnen we goed goedaardige van kwaadaardige genvariaties in het DNA van mensen onderscheiden. Maar wat we nog niet konden, is voorspellen of een onbekende genvariant tot ziektes kan leiden”, zegt Joeri van de Velde, genspecialist van het UMCG.

Daarom klopte hij samen met promovendus Li Shuang aan bij het CIT. Het doel was om CADD, de meest gebruikte annotatiemethode om genetische mutaties op te delen, aan te passen aan hun onderzoeksvraag. “Maar in plaats van die bestaande methode aan te passen, gebruikten we de achterliggende informatie en goten we het in een nieuw machine learning-model. Eigenlijk startten we dus opnieuw”, aldus Dimitrios Soudis.

Dimitrios en zijn team gaven zichzelf een week de tijd. Al snel volgden veelbelovende resultaten, waarna ze die aan Joeri en Shuang voorlegden om te vergelijken met bestaande methodes. “Twee weken lang was ik vrij paranoïde”, zegt Dimitrios. “Ik dacht echt dat er ergens een foutje in geslopen was. Maar toen we ons algoritme toe op nieuwe data toepasten, bleek het ook te werken. “Ik dacht echt: hoe kan het dat niemand dit eerder geprobeerd heeft?”

## Toegevoegde waarde

Ook Joeri van de Velde was verbaasd door de snelle resultaten. “Voor veel van onze onderzoeksvragen werkt het nu al beter dan bestaande

*Cosimo III de'Medici*

*Landkaart uit de collectie van de familie de' Medici*

methodes. Dat is veel sneller dan verwacht." De toegevoegde waarde van machine learning-experts zit volgens hem vooral in de toepassing van het algoritme. "Ik dacht altijd dat het bouwen van het algoritme het moeilijkste was. Maar daarna moet je precies weten hoe je dat algoritme toepast op je onderzoeksvraag. Dat is eigenlijk een hele kunst op zich."

Het algoritme is nu al in staat om op efficiënte wijze pathogenische (ziekteverwekkende) genvarianten te onderscheiden van de duizenden onschadelijke. Bovendien kun je op basis van dit model die genetische variaties aan specifieke eigenschappen linken en voorspellen of nog onbekende genvarianten binnen het DNA mogelijk ziekteverwekkend zijn.

Hun onderzoeksresultaten gaan nu naar labspecialisten. Daarna kunnen onderzoekers het verder ontwikkelen tot een screeningstool voor de dokterspraktijk. De nauwe band met de praktijk bij dit onderzoeksproject maakt het voor Van de Velde extra interessant. "Dankzij dit project kunnen we erfelijke ziektes op termijn veel sneller vaststellen en voorspellen. Dat is best een grote stap."

Ook voor Dimitrios Soudis smaakt het project om die reden naar meer. "Als statisticus kun je in iedereens achtertuin spelen, zei Tukey. Dat is ook zo: ik heb al van verschillende wetenschappen kunnen proeven. Maar dankzij dit project heb ik toch een voorkeur voor medisch onderzoek gekregen. Hier kan ik echt een verschil maken."

## Een virtuele reis naar het zeventiende-eeuwse Nederland

Niet alleen de bètawetenschappen en de medische wereld doen steeds meer beroep op data science. Ook bij de faculteit der Letteren van de RUG groeit het besef dat data science een interessante aanvulling kan zijn op de klassieke onderzoeksmethodes. Sabrina Corbellini, docent middeleeuwse geschiedenis, doet historisch onderzoek naar de reizen van Cosimo III de' Medici. Als deel van de roemrijke Florentijnse bankiersfamilie werd hij groothertog van Toscane aan het eind van de 17de eeuw.

Voor dit project kijken ze vooral naar zijn band met Nederland. Tussen 1667 en 1669 maakte hij twee reizen naar de Republiek. Hij keek op naar Nederland, zoveel is duidelijk. Zo noteerde hij nauwkeurig hoe Nederlanders hun landschap bewerkten met waterbeheer en windmolens. Ook zou hij later de regeringsvorm van de VOC proberen toe te passen als hertog van Toscane.

Volgens Corbellini bieden de verschillende kaarten en dagboeken, die eeuwenlang bewaard werden in de Biblioteca Medicea Laurenziana in Florence, een schat aan informatie over die reis. "Cosimo ging langs bij kunsthandelaars, schrijvers en krantenverzamelaars. Hij bezocht verschillende steden en sprak met stadhouders. Alles wat gaande was in Nederland, wilde hij weten. En dat is allemaal quasi per uur genoteerd in zijn dagboeken."

Aan de hand van die dagboeken wil Corbellini dus achterhalen met wie Cosimo in contact kwam, en – als het even kan – om welke reden. Maar Oud-Italiaanse dagboeken doorlezen en analyseren is een erg tijdrovende taak. Daarom vroeg ze aan data scientists Venustiano Soancatl Aguilar en Nicoletta Giudice een script te maken om snel locaties, personen en plaatsnamen te labelen in de dagboeken.

Voor Venustiano Soancatl Aguilar is het de eerste keer dat hij aan historisch onderzoek deelneemt. "Eerder deed ik projecten met realtime data over lichaamsbeweging, of dataonderzoek over sterrenkunde. Daar werk je met veel grotere datasets. Toch betekent dat niet dat dit project makkelijker is. Cijfers zijn universeel, betekenen altijd hetzelfde. Maar werken met taal is gecompliceerder, want een woord kan verschillende betekenissen hebben."

## Primeur

De dagboeken zijn bovendien in Oud-Italiaans geschreven, wat het er niet makkelijker op maakt. Gelukkig kan Nicoletta Giudice, zijn Italiaanse collega, daar een handje bij helpen. "Machine learning alleen zal je niet vertellen of je fouten maakt. Daarom controleert Nicoletta en geeft ze feedback, zodat we het model kunnen optimaliseren", legt Venustiano uit.

Het model optimaliseren is de moeilijkste taak. Maar op lange termijn levert het sowieso tijdwinst op. "Als het model eenmaal werkt, kun je het binnen een paar seconden op andere documenten toepassen. Dat gaat veel sneller dan manueel 100 boeken analyseren."

Voor Sabrina Corbellini is de samenwerking met data scientists een primeur. In de toekomst ziet ze echter mogelijkheden om het vaker te doen. "Bij mijn recente onderzoeksprojecten ben ik telkens bezig met ruimte, als plaats waar kennis overgedragen wordt, bijvoorbeeld gemeentehuizen, apotheken of universiteiten. En wat betreft geolokalisatie hebben data scientists heel wat expertise die bij klassieke historici ontbreekt. We moeten elkaar vaker ontmoeten om te weten wat we voor elkaar kunnen betekenen."