Kennis maken met big data

In de wereldwijd explosief groeiende berg van digitale gegevens zitten nieuwe inzichten en kennis verstopt. Om die te vinden zijn slimme software en rekenkracht nodig, en heel veel data-experts. ‘Over tien jaar is 80 procent van het onderzoek gebaseerd op het analyseren van databestanden.’

TEKST RIK NIJLAND **ILLUSTRATIES** KAY COENEN

Begin dit jaar werd Alphabet het meest waardevolle bedrijf ter wereld. Voor het eerst voerde niet een maak- of oliebedrijf de ranglijst aan, maar een nieuwkomer die handelt in informatie. Alphabet is het moederbedijf van Google. De opmars van het bedrijf – tien jaar geleden kwam het bedrijf de top-100 binnen – is een duidelijk signaal hoeveel waarde er wordt toegekend aan informatie, informatietechnologie en databestanden.

Op tal van terreinen groeit de hoeveelheid gegevens die digitaal worden opgeslagen en verwerkt met duizelingwekkende snelheid. Bijvoorbeeld op het gebied van DNA (zie kader), maar ook dichter bij ons alledaagse leven. Moderne auto's geven dag in dag uit informatie door aan de fabrikant over onder meer toerental en de lengte van ritten. Interessante informatie om het onderhoudspakket op af te stemmen, maar ook handig voor studies naar bijvoorbeeld verschillen in rijgedrag. Leuk materiaal dus voor onderzoekers maar ook voor verzekeraars.

BROCCOLI WOKKEN

In de databestanden van Google, grote winkelketens, sms, twitter of autofabrikanten ligt een schat aan kennis opgeslagen over onze voorkeuren. Voor snel optrekken bijvoorbeeld; waar we op internet naar zoeken; of wat we graag eten. Bij een analyse van Nederlandse twitterberichten bleek het woord broccoli vaak voor te komen in combinatie met wokken. Niet wereldschokkend, maar voor de groentesector een eyeopener over consumentengedrag.

Die sterk uitdijende digitale gegevensberg om ons heen wordt aangeduid als big data. Big staat voor groot, maar de naamgeving is ook een verwijzing naar Big Brother, de alwetende overheid die volgens schrijver George Orwell al ons doen en laten zou gaan bepalen.

Wageningen UR zet inmiddels stappen om grote databestanden beter te exploreren. Verwacht wordt dat de informatiestroom ook binnen het typisch Wageningse domein sterk zal groeien. Dankzij moderne ICT worden nu al sensoren ingezet die in de kas of op de trekker de gewasgroei in de gaten houden. Alle melkrobots samen weten bijna alles over honderdduizenden koeien. In die brij aan gegevens zitten nieuwe inzichten en kennis verstopt. Zeker als je databestanden kunt koppelen, bijvoorbeeld melkgift of voeropname met genetische informatie.

Het is broodnodig dat Wageningen die nieuwe kennis verder aanboort, vindt Karin Andeweg. Samen met Sander Janssen is zij aanjager en kwartiermaker voor big data, door de organisatie benoemd tot speerpunt. 'Big data lijkt een hype, maar over een paar jaar is het gemeengoed geworden. De verwachting is dat over tien jaar 80 procent van het onderzoek gebaseerd is op het analyseren en combineren van databestanden om zo nieuwe kennis te genereren.'

De komende tien jaar zijn een half miljoen data-experts nodig

NIET MEER HET VELD IN

Die toekomst tekent zich nu al af. Wageningen UR doet mee aan een proef in Amsterdam om op drukke dagen aan de hand van het gebruik van honderdduizenden mobiele telefoons te bepalen waar de mensenmassa gevaarlijk aangroeit. Ook wordt zo gevolgd of de genomen maatregelen effect sorteren. Geen waarnemer hoeft de straat op.

Ook om de gewasgroei, zeg van tarwe te achterhalen gaat de onderzoeker in de nabije toekomst niet meer het veld in om metingen te doen. Nu al kan hij via sensoren in het veld, maar ook van een drone of de satelliet gegevens binnenkrijgen, niet van een paar onderzoeksplotjes, maar van de hele akker of meer tarweakkers tegelijk. Voor dergelijke gegevensbergen draait de computer in principe zijn hand niet om.

Koppel daaraan de gegevens over de hoeveelheid kunstmest, bodem, bespuitingen, neerslag en in een handomdraai is bekend hoe een bepaald tarweras het doet onder wisselende omstandigheden. Nu nog moeten onderzoekers daarvoor jarenlang deelaspecten onderzoeken. Uiteindelijk zal ook de boer daarvan profiteren als al die kennis kan worden vertaald naar handzame adviezen.

BAKENS VERZETTEN

Om databestanden ook daadwerkelijk zo te kunnen benutten, dient de wetenschap de bakens te verzetten, denkt de EU. De Europese Commissie kondigde half april een miljardeninvestering aan in datamanagement.

FOTO ANP

GROTE GETALLEN

Astronomen en fysici grossieren al heel lang in enorme hoeveelheden data. Tijdens experimenten in de deeltjesversneller van CERN worden de resultaten van 600 miljoen botsingen per seconde geregistreerd. Deze eeuw is de biologie met een inhaalrace begonnen. In 2003 werd voor het eerst het DNA van een mens ontrafeld; in 2011 volgde de duizendste mens; volgend jaar komt waarschijnlijk 1 miljoen in zicht.

Elk jaar verdrievoudigt de capaciteit om via sequencing DNA te ontleden. Een vergelijkbare data-explosie vindt plaats rond eiwitten en stofwisselingsproducten in het lichaam. De biologie krijgt daardoor het karakter van een datawetenschap, zegt Dick de Ridder, hoogleraar bioinformatica aan Wageningen University. Miljarden gegevens over genomen, genen, eiwitten en andere moleculen worden in grote bestanden bij elkaar gebracht en systematisch onderzocht. 'De verwachting is dat we dit jaar een miljoen miljard DNA-basen kunnen aflezen', aldus De Ridder. 'Het is de kunst van de bioinformaticus om op basis van de terabytes aan data nieuwe biologische hypotheses op te stellen.' Waar nu biologen data-analyse vaak uitbesteden aan bioinformatici, verwacht ik dat er steeds meer onderzoekers vanachter hun computer voorspellingen opstellen en juist de experimentele validatie daarvan uitbesteden aan onderzoekers in het laboratorium.'

'Als je mee wilt doen, moet je investeren'

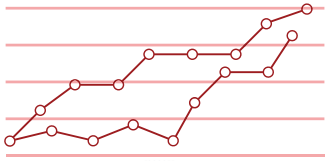
Resultaten van onderzoek worden lang niet altijd goed opgeslagen, wetenschappelijke tijdschriften bieden daar niet de ruimte voor of de bestanden zijn door verschillende formats slecht toegankelijk of niet te koppelen. Om vervolgens de relevante informatie eruit te lichten, is slimme software nodig en rekenkracht. Ook in Wageningen is dat het geval, denkt Andeweg. 'We moeten investeren in hardware, in personeel met ICT-skills en in data-wetenschappers. En we zullen mensen moeten gaan opleiden die met deze nieuwe manier van onderzoek overweg kunnen.' Tijdens de dies natalis van Wageningen University begin maart noemde rector magnificus Arthur Mol data-wetenschap '*a new frontier*'. Om dat nieuwe terrein te versterken, onderzoekt Wageningen UR de mogelijkheden om een data sciences center op te zetten.

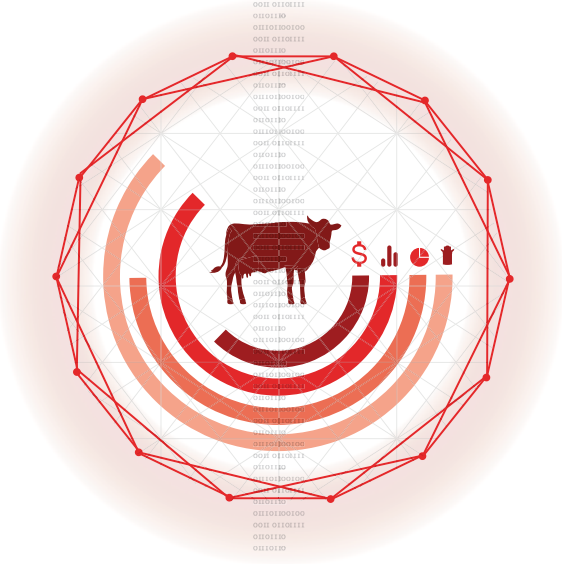
Gastspreekster Laxmi Parida van IBM Research legde tijdens die bijeenkomst uit welke mogelijkheden haar bedrijf al biedt. Centraal daarin staat de zelflerende supercomputer Watson, vernoemd naar een voormalige CEO van het bedrijf. Die kan razendsnel informatie verwerken, literatuur over de hele wereld doorzoeken (40 miljoen documenten in 15 seconden) maar ook verbanden leggen, en informatie op waarde wegen dankzij geavanceerde kunstmatige intelligentie.

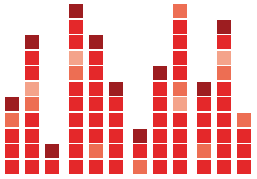
'Om computerprogramma's te ontwikkelen die redeneren en leren complexe vraagstukken op te lossen, heb je gecombineerde kennis nodig van computerdeskundigen en vakspecialisten', meent Parida. 'Die twee partijen moeten voortdurend samenwerken om programma's als Watson te 'leren' hoe die vraagstukken moet oplossen.' Datawetenschap wordt een nieuwe wetenschappelijke discipline, verwacht zij, waarbij onderzoekers en bedrijven uit meerdere disciplines betrokken zijn. 'Alles draait om context, context, context.'

PEUTER MET POTENTIE

Watson is een peuter die je alles eerst moet leren, maar wel met een geweldige potentie', stelt Richard Visser, >







hoogleraar plantenveredeling bij Wageningen UR. Vorig jaar klopte hij aan bij IBM Research om de supercomputer in te zetten voor de veredeling van aardappel. Visser verwacht dat Watson Potato, zoals het systeem inmiddels informeel heet, een belangrijk hulpmiddel wordt om efficiënter en met een beter resultaat te veredelen. Maar eerst moet Watson leren zich te focussen. 'Bij het woord 'knol' moet hij niet uitkomen bij de dahlia' aldus Visser.

Na dat leerproces zal Watson Potato een waardevol hulpmiddel zijn om literatuur te scannen, bijvoorbeeld over de positie van een bepaald gen, verwacht Visser. 'Watson kan veel sneller zoeken dan wij en in veel meer literatuur. Wij komen er ook wel achter dat dat gen bijvoorbeeld bovenin chromosoom 3 ligt, maar dan heb je het over een gebied met wel duizend genen. Dat is onhandelbaar veel', aldus Visser. 'Watson kan verder graven; die duizend genen vergelijken met DNA bij andere planten, maar ook bij paddenstoelen, mossels of vogels, om te zien wat er bekend is over wat die genen daar doen. Dan beperk je de lijst waarschijnlijk tot vijf of tien kandidaatgenen, en naarmate de computer meer kennis vergaart, wordt het misschien wel meteen een schot in de roos.'

De computer leert van zijn fouten. Als hij van experts te horen krijgt dat de antwoorden in de goede richting wijzen, dan borduurt hij daarop voort. Bij negatieve feedback laat hij het gevolgde spoor rusten. 'Hij leert dus, echt net als een mens en bouwt expertise op', aldus Visser. 'Door kennis over uiterlijke kenmerken, DNA-samenstelling, opbrengst en groeiomstandigheden te koppelen kunnen we veel beter ons uitgangsmateriaal kiezen.'

En op termijn zit er misschien meer in het vat, bijvoorbeeld door Watson zijn tanden in complexe problemen te laten zetten. 'Welke genen zijn er verantwoordelijk voor dat de ene aardappel hardkokend is en de andere afkookt? Daar hebben we nog eigenlijk geen idee van.'

AANSPRAAK MAKEN

Ondanks de mooie vooruitzichten met big data, zijn er ook beren op de weg. Wie mag aanspraak maken op de gegevens over een tarweakker? De eigenaar van de drone, degene die de gegevens kan interpreteren of de boer? Zijn de data die de melkrobot dag in dag uit verzamelt van de fabrikant of van de veehouder? Dat is nog nagenoeg onontgonnen gebied. Van wie zijn de resultaten als een bedrijf en een universiteit hun datasets combineren?

Volgens Ben Schaap is innovatie ermee gediend als databestanden toegankelijk zijn voor iedereen. Schaap is door Wageningen UR gedetacheerd bij Global Open Data for Agriculture and Nutrition (GODAN), een mondiale lobby-organisatie voor open data in de landbouw- en voedingssector. Sponsors zijn onder meer de VS, Groot-Brittannië, Nederland en de FAO, en onder de 250 partners zijn bedrijven als IBM en Syngenta maar ook lokale Afrikaanse ngo's.

'Openheid zorgt voor een gelijkwaardig playing field voor iedereen', zegt Schaap. 'Als gegevens niet toegankelijk zijn, hebben partijen met meer macht en geld de grootste controle. Een multinational kan informatie kopen, een eenmansbedrijfje kan daar niet tegenop. Als de informatie van weerstations of satellieten openbaar is, kan iedereen ermee aan de slag. Niet alleen een multinational maar ook een slimme whizzkid. Open data zorgt ervoor dat iedereen applicaties kan ontwikkelen en de boer niet afhankelijk is van één partij die bijvoorbeeld ook zaaizaad , meststoffen of gewasbeschermingsmiddelen levert.'

Daarom is het erg belangrijk dat publiek gefinancierde gegevens openbaar zijn zonder voorwaarden vooraf, vindt Schaap. 'Daar horen ook de datasets bij die universiteiten en instituten verzamelen. De Nederlandse overheid en NWO stellen inmiddels open science als voorwaarde voor subsidiëring, maar ook onderzoekers die geld willen van de Gates Foundation of van het Europese onderzoeksprogramma Horizon 2020 zijn verplicht hun data te publiceren.'

COMMERCIEEL BELANG

Dat bedrijven waarschijnlijk niet happig zijn, om eigen gegevens door anderen te laten gebruiken, begrijpt Schaap. Maar hij verwacht wel dat die bereid zijn uitzonderingen te maken als ze zelf geen competitief belang hebben. 'Syngenta heeft een dataset vrijgegeven over een pesticide tegen muggen, geen speerpunt voor het bedrijf en malariaonderzoekers waren er erg blij mee. Je moet dat zien als een bijdrage om de wereld te verbeteren, net zoiets als aan CO_2-reductie doen.'

Toch ziet hij openheid als meer dan liefdadigheid. 'Er zijn ook bedrijven die bereid zijn een open-data-landschap te creëren', aldus Schaap. 'Zij vinden dat data beter uitwisselbaar moeten zijn zodat meer partijen met elkaars gegevens uit de voeten kunnen voor de ontwikkeling van nuttige toepassingen, bijvoorbeeld in de precisielandbouw. Innoveren kun je tegenwoordig niet meer alleen, is de gedachte. Open science zorgt voor samenwerking met slimme uitvinders. Syngenta zegt tegen startups: maak maar gebruik van onze onderzoeksgegevens. Komt daar een interessante toepassing uit, dan willen we jullie misschien wel overnemen.'

> 'Als gegevens niet toegankelijk zijn, hebben partijen met meer geld de controle'

Plantenveredelaar Visser is ook voorstander van open access, maar onder voorwaarden. 'Je wilt niet dat de eerste de beste whizzkid in Rusland ermee aan de haal gaat. Dat is op dit moment de frustratie van Amerikaanse onderzoekers die worden gefinancierd door de National Science Foundation. Sequensen ze vandaag een genoom dan moet dat morgen op het web. Ze krijgen geen tijd er eerst zelf goed naar te kijken. Anderen zitten te wachten, zeggen dankjewel en publiceren een leuk resultaat.'

Ook bij open science horen spelregels, vindt Visser. 'Wellicht moet je als gebruiker eerst vertellen, wat je van plan bent ermee te gaat doen. Komt dat dicht op ons terrein dan is het logisch afspraken te maken over samenwerking óf wij moeten eerst de tijd krijgen een paar artikelen te publiceren.'

HARDWARE NODIG

Om samenwerking in goede banen te leiden, zijn in maart door wetenschappelijke instellingen waaronder ook Wageningen UR in Nature de FAIR Guiding Principles gepubliceerd. Data moeten *findable*, *accessable*, *interoperable* en *reusable* zijn. Een goed begin, vindt Visser. 'Maar daarnaast heb je ook hardware nodig. Het maakt niet uit waar je data huisvest, of dat bij IBM is, in de cloud, of bij SARA, maar we zullen in Wageningen ook eigen computerinfrastructuur moeten aanschaffen en mensen hebben die daarmee kunnen omgaan. Anders ben je voor alles afhankelijk van anderen. Als je mee wilt doen, moet je investeren.' ■

www.wageningenur.nl/bigdata