Verkenning data-gedreven onderwijsonderzoek in Nederland

Bernard Veldkamp, Kim Schildkamp, Merel Keijsers, Adrie Visscher & Ton de Jong

UNIVERSITEIT TWENTE.

Voorwoord

In onze informatiemaatschappij wordt van individuen, organisaties en processen steeds meer data vastgelegd in databestanden. Big data is de term die gebruikt wordt om de verzameling van al deze data te beschrijven. De term big data staat ook voor het koppelen van eerder gescheiden databestanden en de analyses van deze gekoppelde bestanden om antwoorden te vinden op gerichte vragen, maar ook voor het ontdekken van niet vermoede verbanden. In deze verkennende studie wordt de situatie rond big data voor het onderwijs in Nederland in kaart gebracht. Na een overzicht van de bekende (gestructureerde en ongestructureerde) onderwijs-gerelateerde databestanden, inventariseren we welke vragen er met big data beantwoord kunnen worden, of en onder welke condities data beschikbaar gesteld kunnen worden, bespreken we de technische, ethische en juridische aspecten van big data onderzoek in het onderwijs, en gaan we in op de belemmerende en bevorderende factoren voor onderzoek met big data in het onderwijs en voor het ontwikkelen van een centrale database. Bij het behandelen van deze onderwerpen gebruiken we een voor deze studie uitgevoerd literatuuronderzoek en gegevens uit 33 interviews die in de periode november 2016-februari 2017 zijn gehouden met experts en stakeholders. Big data heeft zeker potentie voor onderwijsonderzoek in Nederland, maar de visies op wat er kan, mag, en nodig zou zijn verschillen sterk. Het rapport eindigt daarom ook met een aantal big data paradoxen in het onderwijsdomein en presenteert ten slotte een aantal concrete aanbevelingen.

Enschede 10-04-2017

Bernard Veldkamp, Kim Schildkamp, Merel Keijsers, Adrie Visscher & Ton de Jong

Universiteit Twente

Inhoudsopgave

Inleiding 1

In alle maatschappelijke en bedrijfssectoren (gezondheidszorg, transport, grootwinkelbedrijven, etc.) worden tegenwoordig grote hoeveelheden data vastgelegd. Deze data kunnen op verschillende manieren zijn verkregen. Het betreft hier data die bijvoorbeeld met een doel verkregen zijn van participanten zelf. De data kunnen ook door anderen vastgelegd worden in administratieve systemen. Daarnaast kan deze data verkregen zijn uit interactie van participanten met (al dan niet mobiele) online systemen. Deze ontwikkeling, samen met het toegankelijker worden van de data in elektronische vorm en het koppelen van eerder gescheiden databestanden, wordt samengevat onder de term “big data”. Kenmerkend voor het veld van big data is a) dat het hier gaat om al beschikbare data, b) dat door het elektronisch beschikbaar zijn van de data nieuwe analyses en combinaties van verschillende soorten data en datasets gemaakt kunnen worden en c) dat er mogelijk combinaties van data gemaakt kunnen worden die niet voorzien waren bij de verzameling en die daarom ook tot nieuwe inzichten kunnen leiden.

Ook in het onderwijs wordt op vele plekken data vastgelegd. Voorbeelden van de verschillende soorten data zijn gegevens uit leerlingvolgsystemen (informatie die direct van participanten verkregen is), CBS-gegevens over zittenblijven, inkomens van ouders etc. (informatie uit administratieve systemen) en gegevens uit online leersystemen (interactie gegevens).

De potentie van big data voor het onderwijs wordt de laatste jaren steeds meer onderkend en kennis van patronen in data kan in potentie ook ingezet worden om het onderwijs te verbeteren (Bongers, Jager & Te Velde, 2015). Echter, er is ook een aantal vraagstukken rond big data dat aandacht behoeft, zoals privacy en ethische issues, verantwoordelijkheid, en de beschikbaarheid en de kwaliteit van de data, en er kunnen vragen gesteld worden rondom de behoeftes aan en waarde van big data voor het onderwijs.

In dit verkennend onderzoek wordt de situatie rond en de potentie van big data in het onderwijs in Nederland in kaart gebracht vanuit het gezichtspunt van verschillende Nederlandse experts en stakeholders. Hiervoor zijn in de periode november 2016 – februari 2017 drieëndertig interviews gehouden. Tot de geïnterviewden behoorden a) organisaties die data genereren, b) organisaties die data beheren, c) wetenschappers die onderzoek uitvoeren op (big) data, c) beleidsmakers, d) bedrijfsleven, e) juristen, f) experts m.b.t. ethische vraagstukken, en g) experts m.b.t. de technische opslag en ontsluiting van data.

Tabel 1 geeft een overzicht van de gehouden interviews en de categorie expert (bij naam) of stakeholder (als organisatie) van de respondent(en). Daarnaast is de voornaamste focus van het interview weergegeven: interviews om de bereidheid te peilen tot het beschikbaar stellen van data (hoofdcategorie 1), evenals interviews om de vragen die leven en die mogelijk met big data onderzoek kunnen worden beantwoord, alsmede interviews betreffende de factoren die het gebruik van big data in het onderwijs zouden kunnen beïnvloeden (hoofdcategorie 2) en daarnaast interviews die zich hebben gericht op de juridische, ethische en technische aspecten van big data onderzoek

(hoofdcategorie 3). Als een interview meerdere categorieën dekt dan wordt het interview in elk van deze categorieën genoemd.

Tabel 1.
Lijst van interviews

Beschikbaarheid data		**Behoeftes en bevorderende/belemmerende factoren**			**(On)mogelijkheden**		
Data genereren	Data beheren	Wetenschap	Beleid	Bedrijfsleven/ praktijk	Juridisch	Ethisch	Technisch
Oefenweb *Marthe Straatemeijer*	SchoolPoort, persoonlijke datakluis *Jeroen Schutz*	UvA, Oefenweb *Han van der Maas*	Onderwijs-inspectie *Bert Bulder*	Snappet *Martijn Allessie*	Kennisnet jurist, adviseur privacy *Job Vos*	Ethicist UT *David Douglas*	Schooluitval voorspellen *Willem-Jan Swiebel*
Snappet *Martijn Allessie*	Topicus, *Bart Broekhuis, Thomas Markus & Barthold Derlagen*	UM *Lex Borghans*	Data expeditie (PO) *Verschillende deelnemers*	KONOT basisscholen *Leonie Wenting*	Utrecht Data School *Iris Muis*	Utrecht Data School *Aline Franzke*	SchoolPoort, persoonlijke datakluis *Jeroen Schutz*
Cito *Anton Béguin & Jos Keuning*	Open state *Lex Slaghuis*	UM *Rolf van der Velden*	PO Raad *Maurits Huigsloot*	SURFnet *Jocelyn Manderveld & Christien Bok*	UL, Big data beveiliging en data re-use *Bart Custers & Helena Ursic*	RUG *Hans Beldhuis*	SURFsara *Machiel Jansen*
Dedact *Bas Vonk*	DUO *Mark de Boer*	UU *Jan van Tartwijk & Renske de Kleijn*	VO Raad *Anne Goris*	Bestuurders Stichting Carmelcollege *Tom Morskieft & Fridse Mobach*	IE/ICT-recht advocaat *Corianne Netze*		SURFsara *Axel Berg*
Stichting Beroep en Bedrijf *Ruud Baarda*	CBS *Barteld Braaksma & Ronald de Jong*	RUG *Roel Bosker*		Stichting Beroep en Bedrijf *Ruud Baarda*	Autoriteit Persoons-gegevens *3 senior onderzoekers*		UT, databases *Maurice van Keulen*
	UT *Marc-Jan Zeeman*				RUG *Esther Hoorn*		UT, data security *Andreas Peter*
	The implementation group (TIG) *Ernst-Jan Horn*						
	St. OnderwijsTTP *Hans van Vlaanderen Michiel Vlastuin Sylvia Peters* (Universiteit Utrecht)						

De resultaten van dit verkennend onderzoek worden hier gepresenteerd aan de hand van een aantal onderwerpen: (1) vragen die met big data onderzoek mogelijk beantwoord kunnen worden (2) het beschikbaar stellen van data voor big data onderzoek (3) de (on)mogelijkheden van big data: technische, ethische en juridische aspecten en (4) belemmerende en bevorderende factoren voor data gedreven onderwijsonderzoek. Deze onderwerpen komen aan de orde na een introductie over big data en een overzicht van databronnen in de Nederlandse situatie. In de besprekingen wordt literatuuranalyse gecombineerd met data uit interviews met experts en stakeholders.

Big data 2

2.1 Definitie en voorbeelden

In de eerste bekende definities wordt big data omschreven als data die aan de volgende drie V's voldoen: volume, variety en velocity (Laney, 2001). Met andere woorden, het gaat om grote hoeveelheden (volume), zeer gevarieerde (variety) data, die continue wordt aangevuld en geüpdatet (velocity). Een voorbeeld van big data in het onderwijs betreft online leersystemen. In deze systemen kunnen alle individuele acties van de leerlingen worden gelogd en worden gekoppeld aan toetsresultaten, wat allerlei mogelijkheden geeft voor verdere analyse. Dit leidt tot een groot volume aan data. De variety, oftewel de verscheidenheid aan soorten data, is enorm, omdat niet alleen het responsie gedrag van de leerling wordt bijgehouden, maar bijvoorbeeld ook de coördinaten van de mouse-clicks en de tijdstippen van de verschillende acties. Tenslotte ligt de velocity hoog, vanwege het realtime bijhouden van login- en klikgedrag van alle leerlingen kan er in korte tijd veel nieuwe data gegenereerd worden. Sommige datasets zijn zo groot, verschillend en veranderlijk dat ze tot big data gerekend kunnen worden, in andere gevallen wordt pas gesproken van big data nadat data uit verschillende sets met elkaar zijn gekoppeld. De Wetenschappelijke Raad voor het Regeringsbeleid (WRR), onderscheidde naast de definitie die vooral gericht is op het beschrijven van de data, ook nog definities die gericht zijn op de analyse en op het gebruik ervan. In dit rapport focussen we vooral op de data, maar komen ook aspecten van analyse en gebruik terug.

In de commerciële sector worden al langer grote datasets gecreëerd en gecombineerd, met als doel om producten aan te passen aan de behoeften van klanten en om de markt beter te begrijpen (Manyika, et al., 2011). In de onderwijssector is big data een nieuwer fenomeen (Enyon, 2013). Big data in de onderwijssector verschilt op een aantal aspecten van andere sectoren waarin het wordt toegepast (Romero & Ventura, 2007).

Ten eerste is het *doel* waarvoor big data onderzoek in iedere sector wordt toegepast verschillend. In de commerciële wereld gaat het bijvoorbeeld veelal om het vergroten van de winst. In het onderwijs gaat het om zowel toegepaste (verbeteren van onderwijs) als meer fundamentele onderzoeksdoelen (het beter begrijpen van bepaalde fenomenen).

Ten tweede is er een verschil in de *soorten data* die beschikbaar zijn. Onderwijsdata hebben betrekking op een specifieke school of onderwijseenheid, zijn gerelateerd aan veel andere data en hebben betrekking op verschillende niveaus van het onderwijssysteem (de kenmerken van leerlingen, docenten, en de school). Vaak gaat het ook om data die een schatting geven van een bepaald construct (Piety, 2013) (bv. leren) en niet om precies meetbare zaken, zoals het aantal verkochte producten.

Ten derde, gaat het in het onderwijs tijdens onderwijsleerprocessen om c*omplexe en sociale interacties*. Het gaat daarom om de combinatie van technische systemen en sociale systemen (Piety, 2013).

Daar komt nog bij dat onderwijsdata verzameld wordt door verschillende instanties en opgeslagen is op verschillende locaties en in verschillende systemen.

Samengevat kunnen we opmerken dat big onderwijsdata wordt verzameld en gebruikt om het onderwijs te verbeteren en onderliggende fenomenen te analyseren. De data hebben betrekking op een specifieke school of onderwijseenheid en kan gekoppeld worden met andere data op verschillende niveaus van het onderwijssysteem. De data gaan over complexe en sociale interacties en zijn verspreid opgeslagen op verschillende locaties en systemen. Deze specifieke kenmerken van onderwijsdata zorgen ervoor dat het analyseren en gebruiken van big data binnen het onderwijs zijn eigen problematiek en dynamiek kent.

Onze interviews laten zien dat de mening over wat big data is nogal kunnen verschillen. Soms hebben geïnterviewden het simpelweg over veel data. Anderen hebben het juist over de koppeling van verschillende databronnen als essentieel kenmerk van big data. Ook wordt opgemerkt dat big data (te) heterogeen is en dat er van alles onder kan vallen, anderen hebben weer een restrictievere blik op big data en realiseren zich niet hoeveel ongestructureerde data meegenomen zouden kunnen worden. Voor een aantal respondenten is big data onderzoek synoniem aan data mining, anderen vinden dat ook gericht onderzoek naar verbanden onder big data onderzoek valt. Ook wordt door een aantal respondenten onderscheid gemaakt tussen datagedreven onderwijs (op kleinere schaal) en big data in het onderwijs.

2.2 Analyse en gebruik van big data

In onderzoek met big data kan het voorkomen dat er gericht wordt gezocht naar veronderstelde verbanden, maar het kan ook gaan om het ontdekken van nieuwe verbanden tussen variabelen, die van tevoren niet vermoed werden. In dit laatste geval, tracht men door het combineren van gegevens uit verschillende databronnen nieuwe, onverwachte verbanden te vinden (Kool et al., 2015).

Bij het gericht zoeken van antwoorden op bestaande vragen worden deze vragen gesteld door onderzoekers om hun wetenschappelijk inzicht te vergroten, of door beleidsmakers om daar nieuw beleid op te baseren. Om hier te spreken van big data onderzoek zal gebruik moeten worden gemaakt van omvangrijke data, die verschillend en veranderlijk zijn (volume, variety en velocity) en zal er vaak een combinatie van verschillende databronnen plaatsvinden(zie paragraaf 3.1) .

Een voorbeeld van een wetenschappelijke vraag is bijvoorbeeld: hoe kunnen logfile gegevens gebruikt worden om inzicht te krijgen in het gebruik en effect van feedback in een online leeromgeving? Hierbij gaat het om vragen als: zijn er verschillen in studeergedrag tussen studenten die wel en die niet voor een tentamen zijn op komen dagen? Het doel hiervan is inzicht verwerven in het gebruik van feedback in een online leeromgeving.

Een voorbeeld van een beleidsvraag is: studeren studenten die veel lenen sneller af, dan studenten die weinig lenen en hangt dit samen met inkomen van ouders? De bijbehorende beleidsvraag is: moet er weer een basisbeurs komen voor kinderen met "armere" ouders.

Het 'ongericht' zoeken naar verbanden in grote sets data (educational data mining) kan interessante inzichten opleveren, maar kan ook leiden tot het vinden van toevallige verbanden zonder onderliggende oorzaak. Het onderscheid tussen deze "spurieuze" correlaties en echte verbanden zal vaak moeilijk te maken zijn. Een ander specifiek aspect van big data onderzoek en van educational data mining in het bijzonder is dat het (vrijwel) altijd correlationele data zijn waarin vaak een oorzakelijkheid wordt gezocht. Voordat er gesproken kan worden van causale relaties, moeten zorgvuldig onderzocht worden of voldaan wordt aan alle drie kenmerken van causaliteit: samenhang/correlatie tussen de variabelen, duidelijke volgorde in de tijd (oorzaak komt voor het

gevolg) en de afwezigheid van plausibele alternatieve verklaringen. Met name het laatste kenmerk is in de praktijk soms lastig aan te tonen.

> *Roberts and a colleague recently published a paper on this topic ("Social Structure and Language Structure: the New Nomothetic Approach" by Sean Roberts and James Winters, Psychology of Language and Communication 16.2 [2012], 89-112). They noted several zany positive correlations of language with behavior; for example, people who speak a subject-object-verb language (like Japanese, Turkish, or Hindi) have more children on average than do people who speak a subject-verb-object language (like English, Indonesian, or Swahili). From: Geoffrey Pullum, Spurious Correlations Everywhere: the Tragedy of Big Data. Lingua Franca.*

> *Data from PISA, for example, suggests that the "highest performing education systems are those that combine quality with equity. "What we need to keep in mind is that this statement expresses that student achievement (quality) and equity (strength of the relationship between student achievement and family background) of these outcomes in education systems happens at the same time. It doesn't mean, however, that one variable would cause the other. Correlation is a valuable part of evidence in education policy-making but it must be proven to be real and then all possible causative relationships must be carefully explored. Valerie Strauss, 'Big data' was supposed to fix education. It didn't. It's time for 'small data.' From: Washington Post; May 9, 2016*

Naast de term 'big data' en daarmee verband houdend 'educational datamining' wordt ook vaak de term 'learning analytics' gebruikt (LA) voor het analyseren van big onderwijs data. Bij LA gaat het om het meten, verzamelen, analyseren en rapporteren van data over leerlingen in hun context, met als doel het begrijpen en optimaliseren van het leren en de leeromgeving waarin het leren plaatsvindt (definitie van de first International Conference on Learning Analytics and Knowledge, 2011; in Ferguson, 2012). Bij LA worden er duizenden datapunten verzameld over het leren, zodat hier uitspraken over gedaan kunnen worden (Siemens, 2013 in Thompson & Cook, 2016). Deze data worden gebruikt om curriculummaterialen en leeromgevingen continue aan te passen aan de vastgestelde behoeften van de leerling, zodat de leerling steeds gemotiveerd blijft om te leren en krijgt wat hij/zij qua onderwijs nodig heeft (Thompson & Cook, 2016). LA heeft vaak als doel om gepersonaliseerd leren (onderwijs op maat) mogelijk te maken (Kennisnet, 2017), terwijl educational datamining meer gericht is op het verkrijgen van meer inzicht in het gedrag van leerlingen/studenten en de condities waaronder zij leren (International EDM Society, 2017).

Douglas (2015) geeft aan dat er drie modellen gebruikt kunnen worden als het gaat om het gebruik van big data, om beslissingen die in het onderwijs genomen worden te verbeteren. Ten eerste kan gebruik gemaakt worden van *beschrijvende analyses*. Hierbij gaat het om het beschrijven en analyseren van historisch verzamelde data over leerlingen, docenten, onderzoek, beleid en andere processen. Het doel is het identificeren van patronen om uitspraken te kunnen doen over bepaalde trends, zoals leerlingaantallen, slagingspercentages en de doorstroom van leerlingen. Dit is wat LaValle, Lesser, Shockley Hopkins en Kruschwitz (2010) ook wel 'aspirational use' noemen, waarbij data gebruikt worden om bepaalde patronen en acties uit te leggen of te verklaren. Ten tweede zijn *voorspellende analyses* mogelijk. Hierbij gaat het om het doen van voorspellingen op basis van trends en door het identificeren van zaken die hieraan gerelateerd zijn en het identificeren van risico's en mogelijkheden voor de toekomst. Hierbij valt bijvoorbeeld te denken aan het vroeg identificeren van leerlingen die risico lopen op uitval. LaValle et al (2010) noemen dit het 'experienced' niveau. Tot slot kan gebruik gemaakt worden van wat Douglas (2015) *'prescriptive'*, oftewel *voorschrijvende* analyses

noemen. Deze analyses kunnen organisaties helpen om hun huidige situatie te beoordelen en om geïnformeerde beslissingen te nemen m.b.t. volgende stappen op basis van voorspellingen. Hierbij wordt dus gebruik gemaakt van zowel de resultaten van de beschrijvende modellen als van de voorspellende modellen. Continu wordt bekeken op welke wijze de doelen van de organisaties behaald kunnen worden, rekening houdend met de beperkingen en risico's.

De opvattingen over het gebruik van big data lopen heel erg sterk uiteen. Opmerkelijk is dat er in de interviews vaak wordt gesproken over de behoefte van scholen om vergelijkingen tussen henzelf en andere scholen te kunnen maken (benchmarking) als een benutting van big data. Ook het vergroten van het inzicht in de onderwijsprocessen door het management wordt genoemd, vaak in relatie tot het opstellen van inspectierapporten. Dit is meer de beschrijvende analyse. Het kunnen identificeren van risicoleerlingen, het voorspellen van leerlingaantallen, en kansen op de arbeidsmarkt identificeren zijn voorbeelden van meer voorspellende analyses. De voorschrijvende analyses zien we terugkomen in de genoemde mogelijkheden om met big data ongelijkheid te bestrijden, onderwijs adaptief te maken, leermiddelen te verbeteren, rendementen te verbeteren en docenten bij te sturen. Op dit terrein zijn de suggesties talrijk. Een aantal respondenten dringt er hierbij wel op aan dat voorspellende analyses uitsluitend ter ondersteuning van de leerling gebruikt mogen worden. Hetzelfde geldt voor de analyses met betrekking op de docent; deze worden bij voorkeur gebruikt om de docent te ondersteunen en niet om hem of haar te beoordelen.

2.3 Kwaliteiten van data

De basis van big data onderzoek zijn de data zelf. De mogelijke waarde van de uitkomsten van het onderzoek worden dus bepaald door de kwaliteit van de data. In de literatuur worden acht kenmerken genoemd van de kwaliteit van big data. Traditioneel gezien wordt bij de kwaliteit van data vooral gedacht aan accuraatheid en consistentie, representativiteit en toegankelijkheid. Specifiek voor big data worden daar nog aan toegevoegd value, veracity, verifiation, volatility en granularity.

Accuraatheid en consistentie

Bij het beoordelen van de kwaliteit van de data wordt vaak in eerste instantie naar de accuraatheid gekeken. Zijn de data zonder fouten opgeslagen in de database en komen de gegevens overeen met de werkelijkheid. Een tweede kenmerk dat daaraan gekoppeld is, is consistentie. Zeker bij grote databases en na het samenvoegen van databases is het van belang dat data over hetzelfde construct op een consistente manier in de database is opgeslagen. Het Rathenau instituut verwijst in hun rapportage over de datagedreven samenleving (Kool, Timmer en van Est, 2015) naar de betrouwbaarheid en inzichtelijkheid van de databases en de analyses die er op los worden gelaten: niet alle data zijn betrouwbaar of volledig, niet alle analyses zijn doorzichtig, en niet alle gevonden verbanden weerspiegelen de werkelijkheid. Accuraatheid en consistentie van gegevens zijn daarom van groot belang. Voor big data geldt net als bij alle andere vormen van statistiek het 'garbage in garbage out'-principe; wanneer (een deel van) de data die men gebruikt onvolledig, onbetrouwbaar of biased is, zal de uitkomst van de analyse eveneens weinig waard zijn (Kool, et al., 2015). Kool en haar collega's, maar ook andere onderzoekers (zie bijvoorbeeld Boyd & Crawford, 2012), merken op dat er een wijdverbreide overtuiging bestaat dat onderzoek met big data objectief is, maar ook aan de dataverzameling en opslag bij big data liggen aspecten als accuraatheid en consistentie ten grondslag die mede bepalen wat de uitkomst zal zijn (Kool et al., 2015; Knottnerus et al., 2011).

Representativiteit

Een steekproef moet representatief zijn, wil men inferenties kunnen maken naar de populatie. Dit betekent dat er voldoende participanten uit alle lagen van de samenleving moeten hebben deelgenomen, zonder dat een bepaalde groep over- of ondergerepresenteerd is. Zo'n steekproef kan worden verkregen door te stratificeren, maar ook door het meten van achtergrondvariabelen, zodat er gewogen analyses kunnen worden uitgevoerd. Bij big data binnen de commerciële sector is het vaak onduidelijk bij welke personen en in welke context de data is verzameld. Een groot voordeel van big onderwijs data is dat de leerlingen, de docenten en de scholen unieke qualifiers zijn voor het labelen van de data, waardoor de representativiteit gecontroleerd en geborgd kan worden.

Toegankelijkheid

Onderwijsdata is opgeslagen op veel verschillende locaties. Het is bijzonder tijdrovend en omslachtig om alle locaties individueel te benaderen om de data te verzamelen. Bovendien is de vraag of alle data zonder meer beschikbaar kunnen worden gesteld voor analyse. Anonimiseren en pseudonimiseren bieden mogelijk een oplossing, maar dit kan als negatieve consequentie hebben dat de representativiteit niet langer kan worden onderzocht.

Naast deze traditionele criteria voor datakwaliteit worden bij big data nog een aantal andere kenmerken gebruikt om de kwaliteit van de data te typeren.

Value en veracity

Daniel (2015) en Kitchin (2013, in Landon-Murray, 2016) spreken van value en veracity. Value gaat over de waarde die gehecht wordt aan big data, aan wat men er mee kan. Veracity gaat over het vertrouwen dat je kunt hebben in de data. Als deze kenmerken toegepast worden op big onderwijs data, dan is er een belangrijk onderscheid tussen de verschillende soorten data. Educational surveys, cohort onderzoeken en gestandaardiseerde toetsen scoren hoog op de kenmerken value en veracity. Lesevaluaties daarentegen zijn vaak incompleet, gebiast en sterk afhankelijk van individuele interpretatie, waardoor ze laag scoren op beide aspecten.

Verification

Daniel (2016) noemt ook nog verification als belangrijk kenmerk. Dit refereert aan de verificatie en beveiliging van de data. Ook hier geldt dat de verification van gestandaardiseerde data relatief goed is, maar dat voor ongestandaardiseerde data, zoals aantekeningen of gespreksverslagen de verification veel ingewikkelder ligt.

Volatility en granularity

Douglas (2015) stelt dat nog twee andere belangrijke kenmerken een rol spelen. Zo speelt volatility een rol, waarbij het gaat om de stabiliteit, beweeglijkheid en duurzaamheid (hoe lang gaan de data mee) van de data voor analyse doeleinden. Een laatste kenmerk dat aandacht verdient is de granularity, oftewel de mate waarin de data detailgegevens over de entiteiten bevatten. Bij het verwerken van data vormt een te lage granulariteit (weinig details) van de big onderwijsdata en het daaraan gekoppelde verlies van de context een risico voor de betrouwbaarheid van de analyses.

Data uit onze interviews laten zien dat nog niet al deze kenmerken even goed ingeburgerd zijn in de discussie rond big onderwijsdata. Value en veracity werden enkele keren genoemd. Granulariteit werd genoemd door databeheerders en speelt een belangrijke rol in de discussie rond privacy en herleidbaarheid.

De scholen en de databeheerders wijzen daarnaast veel op het belang van het meenemen van de context bij het interpreteren van de data en zijn vaak van mening dat deze context noodzakelijk is bij de verification van de data, maar zeker bij de verification van de analyses.

Ook laten de data uit onze interviews zien dat bij de kwaliteit van data door de stakeholders onderscheid gemaakt wordt tussen gestructureerde en gestandaardiseerde data, zoals toetsgegevens, aan de ene kant en ongestructureerde data, zoals gespreksverslagen of aantekeningen van docenten, aan de andere kant. Met name aan de kwaliteit van de ongestructureerde data wordt sterk getwijfeld. Door de onderzoekers werd geopperd om ook standaarden voor het uitwisselen van ongestructureerde data te ontwikkelen, al wijzen zij ook op problemen m.b.t. de representativiteit. Het is lastig te controleren of de gegevens een goede afspiegeling vormen van de werkelijkheid. Bij de scholen en de beroepsorganisaties werd vooral gewezen op problemen qua toegankelijkheid. Huidige systemen zijn dusdanig ingericht dat ze niet met elkaar kunnen communiceren en de data is vaak lastig te exporteren.

Onderwijsdata in Nederland 3

Er wordt in Nederland veel data verzameld binnen een onderwijscontext. Dit zijn bijvoorbeeld examens of de eindtoets van de basisschool, maar ook data die worden verkregen in cohortonderzoeken of educational surveys, zoals PISA en TIMSS. Deze data zijn veelal onder gestandaardiseerde condities verzameld. Daarnaast zijn er bijvoorbeeld gegevens van formatieve assessments, die worden verzameld om het leerproces te ondersteunen. Deze gegevens kunnen afkomstig zijn van korte toetsen, van online oefenprogramma's waarin leerlingen aan hun reken- en taalvaardigheden werken, maar ook van klassengesprekken. De variëteit in deze gegevens is veel groter. Er zijn ook verschillende tekstbronnen, zoals social media, en verslagen van besprekingen, docentevaluaties of docentenvergaderingen. Achtergrondgegevens van leerlingen, ouders en docenten spelen eveneens een belangrijke rol. Big onderwijs data omvat bovendien financiële- en economische gegevens. Daarnaast kan van logfiles over leerlingen en interactie data (bijv. google analytics) gebruik gemaakt worden. In dit hoofdstuk geven we een overzicht van de data die in Nederland worden verzameld.

Grofweg zijn er drie groepen data: (1) input data, zoals leerlingkenmerken (bv. gender, SES, schoolgrootte) en docentkenmerken (bv. jaren werkervaring, kwalificaties), (2) proces data, zoals kwaliteit van instructie, leergedrag, informatie over het curriculum, aantal lesuren en klimaat in de klas en (3) output data, zoals leerprestaties en vervolgopleiding. Een ander belangrijk onderscheid is of de data gestructureerd dan wel ongestructureerd zijn. Gestructureerde data kunnen numeriek of textueel zijn en volgen een van te voren vastgesteld format. Ongestructureerde data zijn veel generieker, ze kunnen numeriek, textueel of audio/visueel van aard zijn, en volgen geen gespecificeerd format.

Al deze gegevens hebben hun eigen context, eigenaar(en), status en locatie. De diversiteit is groot. Dit brengt specifieke uitdagingen met zich mee op het gebied van ontsluiting, koppeling, analyse en interpretatie. In de volgende paragrafen wordt een overzicht gepresenteerd van verschillende soorten data, zowel gestructureerd als ongestructureerd, die beschikbaar zijn, op welk niveau deze data beschikbaar zijn (nationaal-, schoolbestuur-, school-, klas-, leerlingniveau; BO, VO, MBO, HO) en waar de data te vinden zijn.

3.1 Overzicht van databronnen

Onderstaand wordt een overzicht gegeven van databronnen specifiek voor de Nederlandse situatie.

3.1.1 Leerlingvolgsystemen (LVS)

Sinds het schooljaar 2014-2015 werken basisscholen verplicht met een leerlingvolgsysteem (LVS). Een LVS is een systeem waarmee de ontwikkeling van individuele leerlingen op in elk geval het gebied van Nederlandse taal en rekenen wordt bijgehouden. Na verloop van tijd hebben aanbieders van schooladministratiesystemen ook mogelijkheden in hun administratiesystemen ingebouwd om de

resultaten op LVS-toetsen te beheren, analyseren en weer te geven (ParnasSys, Dot.com, Esis). In deze systemen kunnen doorgaans ook observaties van de docent, verzuim en gedrag in de klas geregistreerd worden. Ook in het voortgezet onderwijs wordt er gewerkt met leerlingvolgsystemen. Magister en SOMtoday zijn ontwikkeld als LVS voor het voortgezet onderwijs. Hiermee kunnen onder andere toetsresultaten, observaties door de docent, oudergesprekken en absenties worden geregistreerd. Resultaten en vooruitgang kunnen opgevraagd worden voor zowel individuele leerlingen als voor klassen. Naast de LVS-toetsen van Cito, zijn er andere aanbieders op de markt, zoals Boom test uitgevers en Diataal. De leerlingvolgsysteem-toetsen leveren gestructureerde data over de leerprestaties.

3.1.2 Centrale toetsen

Scholen in het primair onderwijs zijn vanaf het schooljaar 2014-2015 verplicht om bij alle leerlingen een onafhankelijke, objectieve eindtoets af te nemen in het laatste schooljaar. Scholen kunnen kiezen voor de centrale eindtoets van het College voor Toetsen en Examens (CvTE), of voor een van de andere eindtoetsen die door de staatssecretaris van OCW goedgekeurd zijn. Naast de centrale eindtoets zijn nog vijf eindtoetsen beschikbaar waaruit scholen kunnen kiezen. Dit betreft Route8, de IEP-eindtoets, de DIA-eindtoets, de CESAN-eindtoets en de AMN-eindtoets.

Deze eindtoetsen bestaan uit twee verplichte hoofdmetingen: vaardigheden in rekenen en wiskunde (breuken, meetkunde, verbanden leggen), en taalvaardigheden (spelling, begrijpend lezen, taalverzorging) (Lenstra, Prenger, Zanten, Berkel & Verbruggen, 2015). Sommige andere eindtoetsen kennen ook aanvullende metingen van leerlingen.

De eindtoetsen hebben een tweeledig doel: enerzijds voorspellen welk niveau van voortgezet onderwijs het beste bij een leerling past en anderzijds vaststellen wat een leerling in de voorgaande jaren aan kennis heeft vergaard. Daarnaast kunnen de eindtoetsen gebruikt worden om de kwaliteit van scholen te beoordelen.

Het Centraal Examen wordt door elke scholier afgelegd aan het eind van het voortgezet onderwijs. Een aantal vakken is voor iedereen verplicht, zoals Nederlands, Engels, en (een vorm van) wiskunde. Voor een groot deel zijn de examens echter profiel specifiek. Hoewel de examens voor de meeste domeinen een betrouwbaarheid hebben van .70 of hoger, is de betrouwbaarheid van het Nederlands Centraal Examen in alle lagen van voortgezet onderwijs zwak te noemen (.43 - .69, mediaan .58, waarbij scores vanaf .70 als “acceptabel” worden aangemerkt) (Cito, 2007-2015; Tavakol & Dennick, 2011).

De Entreetoets wordt door Cito voor groep 6 en voor groep 7 van de basisschool aangeboden als aanvulling op het advies voor vervolgonderwijs van de basisschool. Leerlingen worden beoordeeld op taalverzorging, lezen, en rekenen. Daarnaast kan Wereldoriëntatie worden toegevoegd. De toets is niet verplicht, maar bedoeld als voorspelling van het schooladvies bij de Centrale Eindtoets.

Deze centrale toetsen resulteren in gestructureerde data over de leerprestaties van leerlingen.

3.1.3 PRIMA, VOCL en COOL$^{5-18}$

PRIMA (cohortonderzoek PRIMAir onderwijs en speciaal onderwijs) en VOCL (Voortgezet Onderwijs Cohort Leerlingen) zijn de voorlopers van COOL5-18 (Cohort Onderzoek OnderwijsLoopbanen), uit respectievelijk het basis- en voortgezet onderwijs. Al deze studies waren nationaal opgezet op verzoek van het Ministerie van Onderwijs en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). De studies maten rekenen/wiskunde, taalvaardigheid en welbevinden; COOL5-18 mat ook de sociale competenties en de sociaal-emotionele ontwikkeling, terwijl PRIMA tevens een non-verbale intelligentietest afnam. Al deze studies zijn cohort studies waarbij dezelfde leerlingen meerdere keren gedurende hun schoolcarrière werden getest. Bovendien werd het onderwijsnummer van participanten genoteerd, waardoor data van deze studies gekoppeld kunnen worden aan gegevens van het CBS en DUO. In aanvulling op de vorige cohorten, werd bij elke editie ook een nieuwe steekproef toegevoegd. De steekproef kwam voort uit een gestratificeerde steekproef van basisscholen waarvan alle leerlingen uit de betrokken klassen de toets maakten (voor COOL5-18 gaat het hierbij bijvoorbeeld om groep 2, 5 en 8 van het basisonderwijs; voor PRIMA om klas 2, 4, 6 en 8). Deze cohortonderzoeken resulteerden in gestructureerde data over de leerprestaties. De financiering voor COOL5-18 is in 2016 afgelopen. Er wordt verkend hoe dit in 2017 opgevolgd kan worden door het Nationaal Cohortonderzoek Onderwijs (NCO).

3.1.4 PISA, TIMSS en PIRLS

Internationaal gezien worden er meerdere onderwijskundige surveys gehouden die als doel hebben om de kwaliteit van het onderwijs in de deelnemende landen onderling te vergelijken. Deze surveys worden periodiek afgenomen, zodat ook de trends binnen een land in kaart kunnen worden gebracht. De data van deze studies worden meestal integraal beschikbaar gesteld voor het doen van aanvullend onderzoek. De uitkomsten van deze surveys hebben vaak een groot effect op het onderwijsbeleid in de verschillende landen. De bekendste surveys zijn PISA, TIMSS en PIRLS.

PISA is een internationaal opgezette studie die elke drie jaar wordt afgenomen bij een steekproef van 15-jarigen. De toets beslaat drie algemene domeinen: leesvaardigheid, wiskunde en natuurwetenschap. PISA maakt gebruik van meerkeuzevragen en open vragen. Daarnaast is er nog een set uitgebreide vragenlijsten voor de leerling, diens ouders, de school, en de docent. In Nederland worden alleen de leerling- en de schoolvragenlijst afgenomen. Daarnaast verschilt PISA van andere testen in hun steekproef. Waar andere studies een of meerdere specifieke klassen aanhouden om in te toetsen, trekt PISA een steekproef uit het bestand van alle 15-jarigen binnen een school, ongeacht de klas waarin ze zitten. Dit heeft tot gevolg dat scores van PISA niet zonder correctie voor leeftijd of klas vergeleken kunnen worden met testscores van andere toetsen.

TIMSS meet zowel de vaardigheden in rekenen/wiskunde als in natuurwetenschappen. De toets wordt sinds 1995 elke vier jaar afgenomen bij volledige klassen. Internationaal wordt TIMSS afgenomen in groep 6 van het basisonderwijs en de tweede klas van het voortgezet onderwijs. In Nederland is dit patroon een aantal malen doorbroken. In '99 werd alleen de tweede klas ondervraagd, en sinds 2007 worden alleen de basisschoolklassen meegenomen in de test (Meelissen et al., 2012). TIMSS bevat zowel meerkeuzevragen als open vragen. Naast de exacte vakken heeft TIMSS, net als PISA en PIRLS, een uitgebreide vragenlijst over achtergrondvariabelen die wordt afgenomen bij leerlingen, docenten,

ouders en de schoolleider. De toetsen van TIMSS zijn valide en zeer betrouwbaar (Sainsbury, Campbell, Kispal, Phillips & Sowerby, 1999). Het belangrijkste verschil tussen TIMSS is PISA is dat TIMSS gebaseerd is op een internationaal curriculumraamwerk en PISA een “focus on skills for future life” heeft.

PIRLS meet taalbegrip en -verwerking en hanteert hierbij ongeveer hetzelfde theoretisch kader als PISA. De test wordt elke vijf jaar afgenomen bij zesdegroepers uit het basisonderwijs. Daarnaast heeft PIRLS een uitgebreide vragenlijst voor de achtergrondkenmerken van de leerlingen, docenten, ouders en de schoolleider. Evenals TIMSS en PISA maakt PIRLS gebruik van een combinatie van meerkeuzevragen en open vragen. In 2011 viel de dataverzameling van TIMSS en PIRLS samen. Er is gebruik gemaakt van een gezamenlijke steekproef om te voorkomen dat scholen belast werden met zowel TIMSS als PIRLS (Meelissen et al., 2012).

Internationale surveys resulteren vooral in gestructureerde data. Omdat ze tegenwoordig alleen per computer worden afgenomen, zijn er ook logfile data beschikbaar, die inzicht kunnen geven in het responsie proces. Verder onderzoek is nodig t.a.v. de vraag hoe deze ongestructureerde data het beste voor analyses gebruikt kunnen worden.

3.1.5 Inspectiedata

De Onderwijsinspectie beschikt over een veelheid aan data over Nederlandse scholen in primair, voortgezet en hoger onderwijs. Belangrijk daarbij zijn de resultaten van het toezicht, de scores van de scholen op de indicatoren uit het toezichtkader van de inspectie waarin zowel de output (de prestaties van de leerlingen van de school) als de processen op klas- en schoolniveau beoordeeld worden. Deze data heeft veelal een gestructureerd karakter.

3.1.6 Dataverzameling DUO

DUO verzamelt veel gegevens over het onderwijs ten behoeve van de taken die zij uitvoert. Dit gaat om zowel leerlinggegevens als gegevens over instellingen en onderwijspersoneel. Deze gegevens worden door DUO verwerkt en een deel vervolgens weer geleverd aan o.a. het CBS en de onderwijs inspectie.

De dataverzameling van DUO omvat inschrijfgegevens van leerlingen in BRON en registratie van onderwijsinstellingen in BRIN. Daarnaast verzamelt DUO stroomgegevens, eindexamengegevens, toets- en schooladviesgegevens, jaarrekeningen van instellingen, personeelsgegevens, gegevens over studiefinanciering, gegevens over kinderopvang, gegevens van digitale examens in Facet en gegevens van inburgeringsexamens.

Duo stelt een gedeelte van deze data beschikbaar zodat anderen zelf informatieproducten kunnen ontwikkelen. Deze data is te vinden op https://www.duo.nl/open_onderwijsdata/.

3.1.7 Lesevaluaties

Om de kwaliteit van lesgeven te beoordelen kan gebruik gemaakt worden van leerlingevaluaties, van zelfevaluaties door leerkrachten en van lesobservaties door derden. Leerlingevaluaties gebeuren vaak schriftelijk, maar sinds kort zijn er ook apps beschikbaar die leerkrachten kunnen gebruiken om aan het eind van de les een beoordeling van de les te vragen aan hun leerlingen. Zelfevaluaties kunnen in

meer of mindere mate gestructureerd plaatsvinden. Ze kunnen bestaan uit de aantekeningen van een leraar of docent over bijvoorbeeld hoe de les ging en hoe de les de volgende keer (beter) voorbereid of ingericht kan worden. Een docent kan ook zichzelf beoordelen op een gestructureerde vragenlijst. Ook docenten, leerlingen of schoolleiders kunnen een dergelijke vragenlijst invullen.

Lesobservaties kunnen door een observator of met video worden uitgevoerd. Dit kan een collega docent doen, maar ook door externen, of de schoolleider. Deze observaties kunnen ongestructureerd plaatsvinden, maar er kan ook gebruik worden gemaakt van gestandaardiseerde schema's voor lesobservatie. Een belangrijk instrument hierbij in Nederland is het ICALT instrument. ICALT is in 2002 door van der Grift ontwikkeld als instrument om de kwaliteit van lesgeven in het primair onderwijs te beoordelen. Met behulp van data uit de omliggende buurlanden (Vlaanderen, Engeland, Duitsland) is het instrument gestandaardiseerd tot een internationaal meetinstrument voor leskwaliteit. De mate van standaardisatie van lesevaluaties verschilt sterk tussen scholen, zowel wat betreft de aard van de gegevens als de frequentie waarmee ze worden verzameld. Lesevaluaties kunnen resulteren in zowel gestructureerde als ongestructureerde data, afhankelijk van het instrument dat gebruikt is.

3.1.8 Domeinspecifieke leeromgevingen en MOOCs

Het aantal domeinspecifieke leeromgevingen in het onderwijs neemt snel toe (zie bijvoorbeeld SURFnet/Kennisnet, 2011). Het betreft hier online "methoden" voor het aanleren van vakspecifieke kennis en vaardigheden, bijvoorbeeld op het gebied van reken en taal. Deze programma's bevatten interactieve onderdelen (quizzen, oefeningen) waarbij op basis van de activiteiten en resultaten van leerlingen adaptief feedback en nieuwe oefeningen worden aangeboden. Docenten kunnen op basis van de verzamelde data vaak tijdens de lessen op een dashboard de voortgang van hun leerlingen volgen. Dit biedt hen de mogelijkheid om leerlingen direct feedback en hulp te geven, daar waar ze zien dat leerlingen moeilijkheden hebben. Wanneer deze programma's via een webinterface worden aangeboden en de data centraal beschikbaar zijn kunnen grote sets van interactiedata van leerlingen of studenten ontstaan. Een voorbeeld hiervan zijn de zogenaamde Moocs. Moocs, oftewel massive open online courses, zijn online leeromgevingen waarbij deelnemers vanuit hun eigen locatie, veelal in hun eigen tempo, aan een cursus kunnen deelnemen. De cursus wordt in zijn geheel online aangeboden en alle communicatie, inclusief het doen van opdrachten of het deelnemen aan discussies en chats, vindt plaats via het web. Er zijn (en ontstaan) veel van dit soort programma's, die we illustreren aan de hand van enkele voorbeelden.

Rekentuin en *Taalzee* zijn applicaties die ontwikkeld zijn door de Universiteit van Amsterdam (UvA) (van der Maas, Klinkenberg, Straatemeier, 2010; Straatemeier, van der Maas, Klinkenberg, 2009). Na twee jaar als onderzoeksproject zijn de programma's op de markt gebracht. Momenteel zijn er zo'n 1600 scholen die Rekentuin en Taalzee gebruiken. In Rekentuin en Taalzee moeten leerlingen hun tuin respectievelijk zee gezond houden door het maken van reken- respectievelijk taalopdrachten. Dit kan op elk moment van de dag, en zowel op school als in de vrije tijd. De applicaties maken gebruik van een adaptief rating model: de moeilijkheid van de opdrachten wordt afgestemd op de prestaties van een leerling.

Snappet biedt voor een groot aantal vakken uit het primair onderwijs de mogelijkheid om leerlingen opgaven te laten maken op een tablet (veelal nadat de leerkracht de stof heeft uitgelegd). Leerlingen krijgen goed/fout feedback op de door hen gemaakte opgaven en wanneer de basisopdrachten zijn

afgerond krijgen ze oefeningen aangepast aan hun niveau. Om hun vorderingen zichtbaar te maken krijgen leerlingen “sterren” toegekend voor elk niveau dat ze succesvol afronden. In het leerkracht dashboard krijgt de docent tijdens de les een gedetailleerd overzicht van de voortgang van elke leerling (o.a. welke opgaven zijn gemaakt, welke goed/fout werden gemaakt, hoeveel pogingen per opdracht zijn gedaan). Onderzoek liet een positief effect zien van het gebruik van Snappet op de ontwikkeling van wiskunde/rekenvaardigheden; zo’n effect werd niet gevonden voor spelling (Faber & Visscher, 2016).

Gynzy iPads is op dit moment beschikbaar voor rekenen en spelling in groep 4 en 5, maar wordt in rap tempo uitgebreid naar alle leerjaren en andere vakgebieden. De software is opgebouwd volgens de leerlijn, waarbij deze is opgesplitst in microdoelen. Scholen kunnen de software inzetten als verwerkingsmateriaal (adaptief of statisch) bij de methode die zij hanteren, of vanuit de leerlijn werken en de methode loslaten. Op het dashboard ziet de leerkracht per leerling per microdoel de vorderingen en het beheersingsniveau. Ook kan de leerkracht doelen en oefeningen aan- en uitzetten voor individuele en groepen leerlingen. Leerlingen kunnen bij veel opgaven op een icoontje klikken om een specifieke hint te krijgen. Daarnaast krijgen ze bij elke gemaakte opgave goed/fout-feedback, en kunnen ze in hun overzicht zien hoe ver ze zijn met het behalen van de doelen.

Voor science vakken op voortgezet onderwijs niveau zijn er veel websites die online laboratoria aanbieden. Deze laboratoria (of simulaties) bieden leerlingen de gelegenheid om zelf online experimenten uit te voeren in natuurkunde, scheikunde, biologie etc. Een bekend voorbeeld zijn de *PhET* laboratoria van de Universiteit van Colorado (https://phet.colorado.edu/). Deze website is zeer populair onder Nederlandse docenten, internationaal kent de site vele downloads per jaar. Door een overgang van technologie van Java naar HTML5/JavaScript worden deze simulatie labs nu niet meer gedownload, maar via een webinterface aangeboden waardoor de mogelijkheid ontstaat om centraal alle interactiegegevens op te slaan. Anders dan voor specifieke onderzoeksdoelen en de standaard dataverzameling door Google analytics staat de registratie van data niet aan. Gebruikers van een specifieke PhET versie (PhET-iO) hebben de mogelijkheid dit voor hun eigen gebruik aan te zetten. Data wordt dan gegenereerd in de browser en de gebruiker bepaalt welke data verzameld wordt en waar deze wordt opgeslagen. Een gerelateerd voorbeeld is de Go-Lab portal (www.golabz.eu), zie de Jong et al, (2014). Op de Go-Lab portal kunnen docenten online labs vinden, deze zelf combineren met apps die het leren ondersteunen (bijvoorbeeld een concept mapper of een hypothese kladblok) om zo een complete leeromgeving te creëren. Go-Lab groeit sterk, de portal heeft zo’n 14.000 bezoekers per maand en meer dan 5000 docenten hebben een eigen leeromgeving gecreëerd. Interactie gegevens van leerlingen en de door hen gecreëerde producten (bijvoorbeeld concept maps) worden centraal opgeslagen. Docenten kunnen via learning analytics apps de voortgang van hun leerlingen monitoren (ze kunnen bij per leerling zien hoeveel tijd deze leerling in een fase van de leeromgeving heeft doorgebracht of hoeveel leerlingen er op een bepaald moment in de tijd in een bepaalde fase zijn) en ze kunnen de prestaties van elke individuele leerling inzien. Bij elke leeromgeving kan een docent aangeven of leerling data worden opgeslagen of niet, alleen bij toestemming om data op te slaan is de learning analytics functie beschikbaar.

De gegevens die met deze domein specifieke leeromgevingen worden verzameld hebben vaak veel volume en worden met grote regelmaat aangepast en geüpdatet. Het gaat hier veelal om gestructureerde data. Een nadeel aan deze programma's is dat de steekproef niet aselect is en dat de achtergrondinformatie die wordt verzameld over leerling en school maar zeer beperkt is. Hierdoor is

het lastig om in te schatten hoe representatief de steekproef is. Daar staat tegenover dat de steekproef vaak groot is. Bijvoorbeeld Rekentuin wordt in Nederland op meer dan 1600 scholen gebruikt en telt 100.000 deelnemers (of één op de 25 scholieren) en Snappet wordt door ongeveer 2000 basisscholen gebruikt (op een totaal van iets minder dan 7000 basisscholen). De gemiddelde basisschool heeft 200 leerlingen, dus bij een volledig gebruik van Snappet in alle klassen (dit is waarschijnlijk niet het geval) gaat dit over erg grote aantallen leerlingen waarover zeer frequent en zeer gedetailleerd studievoortgangsdata worden verzameld. Bovendien gaat het hier om het gebruik van een systeem bij tal van vakken uit het primair onderwijs. Informatie over de achtergrond van de deelnemende scholen en hun leerlingen kan worden opgevraagd bij het CBS.

3.1.9 Verslagen (docent)vergaderingen

Verslagen van sectievergaderingen, rapportvergaderingen of andere docentvergaderingen bieden veel inzicht in onderwijsprocessen en in de organisatie van scholen. Daarnaast geven ze waardevolle informatie over individuele leerlingen. Bovendien bieden bijvoorbeeld verslagen van functioneringsgesprekken informatie over de ontwikkeling van de leerkrachten en docenten. De wijze waarop deze verslagen gemaakt worden, de frequentie en hoe ze worden opgeslagen in de systemen verschilt sterk over scholen en zelfs binnen scholen. Ook de inhoud van de verslagen is heel divers. Bij sommige scholen worden alle overwegingen vastgelegd, terwijl andere zich beperken tot het vastleggen van de afspraken. Deze verslagen betreffen een vorm van ongestructureerde data.

3.1.10 Registraties en officiële documenten

Jaarverslagen, verzuimregistraties, procesdata, schoolgids, schoolplan, jaarrekeningen en beleidsdocumenten geven inzicht in de operationele kant van het onderwijs. Onder procesdata vallen bijvoorbeeld gegevens als lesroosters, het aantal uren dat leerlingen les krijgen, de hoeveelheid toetsen, het aantal studiedagen, vakanties, et cetera. Deze operationele data geven informatie over waar prioriteiten gelegd worden, over het beleid en ook over het werkklimaat binnen een school. Deze ongestructureerde data die tekstueel en numeriek van aard zijn kunnen worden ontsloten met tekstmining technieken (Manning & Schütze, 1999). Tekstmining is het proces om met behulp van daarvoor ontwikkelde software waardevolle informatie te halen uit grote hoeveelheden ongestructureerd tekstmateriaal. Omdat er geen gestandaardiseerd format is waaraan deze registraties moeten voldoen, kunnen ze wat hun vormgeving betreft sterk verschillen tussen scholen.

3.1.11 Overige ongestructureerde data

Bij de voorgaande databronnen was er grotendeels sprake van data, ook de ongestructureerde, die samen te vatten zijn in een tabel of statistiek. Er ligt echter ook een schat aan informatie besloten in ongestructureerde data die zich niet gemakkelijk in een format laat samenvatten, zoals docent-oudergesprekken, of het twittergedrag van leerlingen. Hieronder vallen ook de logfiles van zoekmachines, het wifi gebruik en andere automatisch gegenereerde bestanden. Deze data zijn lastiger te ontsluiten en analyseren, maar heeft zijn kracht qua toepassing al bewezen. In het MBO wordt ongestructureerde data bijvoorbeeld gebruikt om te voorspellen of en wanneer een leerling zijn of haar diploma zal halen (“Met ongestructureerde data voorspellen”, 2016).

Alvorens ongestructureerde data kunnen worden geanalyseerd, moet het eerst geschikt worden gemaakt voor verwerking. Dit kan bijvoorbeeld door teksten door een programma te halen dat, door te zoeken naar bepaalde woordcombinaties, leestekens en syntax de structuur van een tekst in kaart brengt.

3.2 Slot

Vaak zal het zo zijn dat een van de hierboven genoemde individuele bronnen niet aan alle kenmerken van big data voldoen (volume, variety, and velocity), en dat één specifieke databron niet direct geschikt is voor het vinden van nieuwe verbanden (educational datamining) maar een combinatie van onderstaande bronnen zal hier wel al snel aan voldoen. Is dit het geval dan wordt de vraag naar hoe data gekoppeld kunnen worden direct van belang (zie paragraaf 6.1.1).

Om nieuwe inzichten te verkrijgen kunnen de hierboven genoemde bronnen onderling gekoppeld worden. Een verdere verrijking van de gegevens kan gerealiseerd worden door onderwijsdata te koppelen met andersoortige gegevens, zoals bijvoorbeeld met achtergrondgegevens over de ouders, leerlingen of docenten, of met data over economische- of geografische ontwikkelingen. Daarnaast kan onderwijsdata gekoppeld worden aan specifieke data over een doelgroep, zoals bijvoorbeeld minderjarige asielzoekers. Deze verrijking van onderwijsdata kan zowel gebruikt worden om het onderwijs te verbeteren, als om inzicht te krijgen in onderliggende fenomenen.

Tijdens de interviews komt naar voren dat de mening van de respondenten over wat big data is nogal verschilt. Data uit online leeromgevingen en ongestructureerde data worden algemeen gezien als big data. Sommige respondenten zien big data vooral als veel data. Anderen hebben het juist over het koppelen van verschillende databronnen om zo tot nieuwe inzichten te komen. Toetsdata, data uit cohort onderzoeken of internationale surveys, zo wordt opgemerkt, zijn op zichzelf nog geen big data. Maar door ze te koppelen met andere bestanden wordt het wel big data.

Welke “big data vragen” leven er bij de experts en stakeholders? 4

Bij de vraag wat big data voor het onderwijs kan betekenen is niet alleen van belang wat er is en wat er zou kunnen, maar ook wat het standpunt van de betrokken partijen is. Als vertrekpunt voor het bepalen van het draagvlak voor big data analyses is het van belang te weten welke vragen de stakeholders denken aan te kunnen pakken met behulp van big data, daarbij uitgaande van wat zij zelf als big data in gedachten hebben. Dat wordt in dit hoofdstuk verder uitgewerkt voor de verschillende doelgroepen in het onderwijs. Hierbij hebben we onderscheid gemaakt tussen leerlingen, docenten, managers en bestuurders, onderzoekers en aanbieders. We brengen telkens in kaart wat er over de doelen van specifieke gebruikers in de literatuur wordt gerapporteerd en vullen dat aan met de informatie die verzameld werd met de interviews.

4.1 Leerlingen

Voor leerlingen is het van belang dat ze inzicht krijgen in hun eigen leerproces. Daaronder valt het analyseren van het eigen leren, het vergelijken van het eigen leren met dat van anderen (in de klas), en de analyse en visualisatie van data (opsporen van bruikbare informatie die kan helpen bij het nemen van beslissingen (Liñán & Pérez, 2015; Romero & Ventura, 2010)). Daarnaast is het voor de leerlingen van belang dat ze zich kunnen verbeteren wat betreft hun leerprocessen. Big data zou gebruikt kunnen worden voor het aanbevelen van activiteiten, hulpbronnen en leertaken die bruikbaar zijn voor het optimaliseren van het leerproces, voor het doen van suggesties voor interessante leerervaringen, voor verdieping, en voor relevante cursussen en boeken etc. (Liñán & Pérez, 2015; Romero & Ventura, 2010).

Uit de interviews kwam naar voren dat big data gebruikt kan worden voor het formuleren van een beter schooladvies aan het eind van het primair onderwijs. Daarnaast kan big data de leerlingen en studenten autonomer maken. Meerdere respondenten noemen mogelijkheden voor gepersonaliseerd leren. Ook kan big data leerlingen en studenten preventief waarschuwen voor de risico’s die ze lopen tijdens het leerproces.

4.2 Docenten

In de literatuur wordt een groot aantal doelen en toepassingen genoemd waarvoor docenten big data kunnen gebruiken:

1. De analyse van de eigen instructie. Docenten kunnen feedback krijgen over de kwaliteit van hun instructie (Romero & Ventura, 2010) en kunnen de effectiviteit van hun onderwijs evalueren (Liñán & Pérez, 2015).

2. Het evalueren van de effectiviteit van het leerproces van leerlingen en het diagnosticeren van leerproblemen en het analyseren van het gedrag en het leren van leerlingen, inclusief het detecteren van ongewenst leerlinggedrag en leerproblemen (bijvoorbeeld gebrek aan motivatie, spijbelen, drop-out, tegenvallende leerprestaties) (Liñán & Pérez, 2015; Romero & Ventura, 2010).
3. Analyse en visualisatie van data (Romero & Ventura, 2010). Hierdoor krijgen docenten de mogelijkheid om leerlingen te identificeren die extra ondersteuning nodig hebben of problemen hebben (Liñán & Pérez, 2015; Romero & Ventura, 2010). Ook kunnen onderdelen van cursussen worden geïdentificeerd waarmee leerlingen moeite hebben en/of fouten die veel voorkomen (Liñán & Pérez, 2015; Romero & Ventura, 2010). Big data kan zo gebruikt worden om het onderwijs aan te passen. Dat kan door het groeperen van leerlingen (Romero & Ventura, 2010), door het aanpassen van de instructie/pedagogiek en/of lesmateriaal aan de vastgestelde behoeften van leerlingen (Liñán & Pérez, 2015; Romero & Ventura, 2010), door het het geven van feedback (er kunnen op basis van data zowel pro-actieve maatregelen als remediërende maatregelen genomen worden) (Dede, 2016; Romero & Ventura, 2010), of door gepersonaliseerd leren. Op basis van data kan bijvoorbeeld besloten worden wat de volgende taak is waaraan een leerling zou kunnen werken. (Douglas, 2015; Romero & Ventura, 2010).

Tijdens de interviews werd, als belangrijk doel voor docenten, datagedreven lesgeven genoemd. Data kan daarbij gebruikt worden om het onderbuikgevoel van de docent te bevestigen of weerleggen. Met behulp van big data kan de leerbehoefte van de individuele leerlingen beter in kaart worden gebracht. Daardoor is er een basis voor differentiatie en gepersonaliseerd leren. Gebruik van big data zou ook de werklast van docenten kunnen verlichten doordat ze met behulp van beschikbare data het niveau van hun leerlingen kunnen monitoren en er daarom minder toetsen afgenomen hoeven te worden. Ook werd big data genoemd als middel om fraude te detecteren. Tenslotte noemde één van de respondenten dat big data gebruikt kon worden als alternatief om de bevragingslast van scholen, en daarmee van docenten en leerlingen, te verminderen.

4.3 Managers en beleid

Een belangrijke toepassing van big data voor beleidsmakers is het formuleren van doelstellingen (Romero & Ventura, 2010) en de analyse van problemen en hun oorzaken (Manyika et al., 2011). Daarnaast kan informatie uit big data gebruikt worden voor een betere en effectievere inzet van human resources, materialen en hulpbronnen (Liñán & Pérez, 2015; Romero & Ventura, 2010). Tenslotte zullen managers en beleidsmakers big data in willen zetten voor het nemen van meer gefundeerde beslissingen en voor het verbeteren van de kwaliteit van hun organisatie (Manyika et al., 2011).

De geïnterviewden noemden benchmarking als een belangrijke toepassing van big data in het onderwijs. Door verschillende respondenten werd dit als een voorbeeld genoemd van hoe scholen big data nu al gebruiken voor hun bedrijfsvoering. Big data wordt zowel gebruikt om docenten als scholen te benchmarken. Daarnaast werd naar voren gebracht dat scholen zich met behulp van big data beter kunnen profileren. Ook wordt big data gebruikt om dashboards te vullen met gegevens over (ziekte)verzuim of lesuitval die gebruikt worden bij het dagelijks leidinggeven.

4.4 Onderzoekers.

Om nieuwe kennis te ontwikkelen, kunnen onderzoekers met behulp van big data theorieën testen en nieuwe hypothesen formuleren (Liñán & Pérez, 2015). Onderzoekers kunnen big data verder gebruiken voor het ontwikkelen van cognitieve modellen van personen, inclusief het modelleren van hun vaardigheden en hun declaratieve kennis (Romero & Ventura, 2010). Ook biedt big data mogelijkheden voor, bijvoorbeeld, sociale netwerkanalyses.

Tijdens de interviews werd vooral gewezen op het belang voor onderzoekers van het beschikken over data en op de mogelijkheden die een gedeelde database hen zou bieden. Ook werd meerdere keren de tijdwinst genoemd die onderzoekers boeken als ze niet meer zelf de data hoeven te verzamelen. Inhoudelijk werd in de interviews gewezen op de mogelijkheden voor onderzoek naar vroegdiagnose en het leggen van verbanden met data uit andere bronnen. Daarbij werd gedacht aan gegevens over bijvoorbeeld gezondheid, ontwikkelingen in de plaats/regio of gegevens van de ouders.

4.5 Ontwikkelaars van lesmateriaal en cursussen.

Bij het ontwikkelen van cursusmateriaal kan big data worden gebruikt voor het ontwikkelen van materialen voor cursussen en lessen (Romero & Ventura, 2010). Een tweede doel is het evalueren van (de effectiviteit van) cursussen en materialen voor het leren van leerlingen (Romero & Ventura, 2010) en het verbeteren van het leren van leerlingen (Romero & Ventura, 2010). Daarnaast kan big data gebruikt worden voor de ontwikkeling van automatische instructiemodellen (Romero & Ventura, 2010) en voor het vergelijken van en ontwikkelen van datamining tools voor het onderwijs (Romero & Ventura, 2010).

In de interviews werd gewezen op het belang van big data voor het ontwikkelen van programma's die gepersonaliseerd leren aanbieden. Daarnaast werd gewezen op de mogelijkheid om met behulp van big data instructieaanpakken en leermiddelen te evalueren en verbeteren.

4.6 Aanbieders van cursussen, trainingsinstituten, hogescholen, universiteiten.

Big data kan gebruikt worden voor het nemen van betere beslissingen in het onderwijs (Douglas, 2015; Romero & Ventura, 2010). Met behulp van big data kunnen aanbevelingen gedaan worden voor specifieke cursussen voor specifieke (groepen van) studenten (Romero & Ventura, 2010). Big data kan tevens gebruikt worden om te voorspellen wat er nodig is om het leren van leerlingen te verbeteren (Douglas, 2015). Daarnaast kan het ingezet worden om op een kosteneffectieve manier de doorstroomcijfers (zakken, uitstroom) te verbeteren en het aantal dropouts te verminderen (Douglas, 2015; Romero & Ventura, 2010). Ook wordt big data ingezet voor planning en roostering en voor selectie, zowel bij de instroom als bij de doorstroom in verschillende richtingen binnen een opleiding (Douglas, 2015; Romero & Ventura, 2010). Tenslotte is big data van belang voor kwaliteitsverbetering, onder andere door de evaluatie en verbetering van onderwijsprogramma's en van docenten (Kane, Rockoff, & Staiger, 2008; Romero & Ventura, 2010, Douglas, 2015).

Tijdens de interviews werd het belang van big data voor aanbieders gezien in de vroegsignalering van uitval of leerproblemen van leerlingen. Daarnaast biedt big data de mogelijkheid om meer gepersonaliseerd les te geven. Zwakkere leerlingen kunnen extra ondersteuning krijgen. Ook kunnen individuele leerlingen meer autonomie krijgen en meer eigenaar worden van hun eigen leerproces. Zeker als aanbieders ervoor kiezen om hun programma modulair op te zetten. Ook kunnen leerlingen verschillende opleidingen met behulp van big data beter onderling vergelijken.

Bovenstaande vragen die leven bij de verschillende experts en stakeholders laten zien dat de verwachtingen van big data divers is. Het laat ook zien dat voor de beantwoording van de diverse vragen verschillende soorten data en verschillende analyses nodig zijn die elk weer hun eigen vraagstukken hebben rond de beschikbaarheid van data en rond de technische, juridische, en ethische aspecten, onderwerpen die in de volgende hoofdstukken aan de orde komen

Beschikbaar stellen van data 5

Om big data onderzoek mogelijk te maken moeten partijen bereid zijn om data ter beschikking te stellen, dit is de vraag naar het draagvlak voor big data analyse. Daarbij is van belang wat deze partijen denken over het eigenaarschap van data.

Eigenaarschap van data is met name gerelateerd aan het beschikbaar stellen van data. Wie geen eigenaar is kan eigenlijk ook data niet beschikbaar stellen aan derden. Er zijn hier twee opties: de data zijn van degene (leerlingen of hun wettelijke vertegenwoordigers) over wie de data gaan, of de data zijn eigendom van degene die de data heeft verzameld. Een specifiek geval doet zich voor als er een partij is die data heeft toegevoegd aan bestaande data, bijvoorbeeld feedback die automatisch op basis van leerlingdata is gegenereerd. Het eigenaarschap van deze data zou dan weer kunnen liggen bij de partij die deze data heeft toegevoegd. In de praktijk is dan ook vooral het recht op gebruik van de data van belang, wie mag onder welke condities wat met de data doen.

Uit onze interviews blijkt dat de meningen over het eigenaarschap van data nogal verschilt. Sommige partijen zien de data als onvervreemdbaar eigendom van de leerlingen of hun vertegenwoordigers en die zouden dan dus in alle gevallen toestemming moeten geven, andere geïnterviewden hechten hier minder aan en leggen het eigenaarschap bij de verzamelende partij. Die laatste zou dan toestemming van de leerlingen of ouders moeten verkrijgen maar is dan verder vrij te handelen.

Een tweede belangrijk aspect bij het beschikbaar stellen van data is het zicht dat betrokkenen hebben op wat er met de data gebeurt. Dit betreft het doel waar de data voor gebruikt gaan worden. In het vorige hoofdstuk zijn verschillende vragen beschreven die bij de stakeholders leven. Het doel kan gerelateerd zijn aan het doen van theoretisch of praktijkgericht wetenschappelijk onderzoek. De data kan gebruikt worden om docenten en leerlingen te ondersteunen. De data kunnen gebruikt worden om leerlingen te categoriseren voor bijvoorbeeld het schooladvies. Leermiddelen kunnen worden verbeterd. De data kunnen gebruikt worden voor benchmarking of om leerlingen, docenten of scholen te beoordelen. Een hieraan gerelateerd mogelijk kenmerk van big data, namelijk dat er gezocht wordt naar niet vermoede verbanden, maakt deze discussie extra gecompliceerd omdat in dit geval het doel niet van tevoren kan worden vastgelegd. Om data te mogen analyseren, geldt op dit moment voor de meeste partijen het principe van doelbinding. Data mogen alleen geanalyseerd worden voor het doel waarvoor de school of de leerlingen toestemming hebben gegeven. Voor wetenschappelijk onderzoek zijn er binnen de huidige wetgeving al bepaalde uitzonderingen.

Tijdens de interviews werd opgemerkt dat het principe van doelbinding bij onderzoek met onderwijsdata misschien niet helemaal kan functioneren. Het beperkt de mogelijkheden voor big data analyses sterk. Zoeken naar verbanden zonder dit van tevoren te specificeren ligt lastig. Verschillende geïnterviewden geven aan dat er ruimte zou moeten komen voor een flexibele vorm van verantwoording, bijvoorbeeld achteraf, met daaraan gekoppeld een vorm van toezicht.

Een aantal geïnterviewden stelt dat het niet ethisch zou zijn om op grote schaal gegevens te verzamelen als het doel niet volledig helder is - er is met name angst voor het gebruiken van uitkomsten voor een ander doel dan waar ze voor bedoeld waren. Tijdens de interviews geven respondenten met verschillende achtergronden aan dat scholen huiverig staan tegenover het delen van data, omdat onduidelijk is wat mag en niet mag, en dat ze daarom op veilig spelen. Dit geldt ook voor gemeentes en andere instellingen. Daarnaast heerst er bij de onderwijsinstellingen ook enige bezorgdheid dat ze afgerekend zullen worden op basis van de data die ze aanleveren (terwijl ze zelf wel graag gegevens van anderen zien om zichzelf te benchmarken).

In de interviews bleek dat er zeker partijen zijn die (nog) niet van het mogelijk nut van big data onderzoek overtuigd zijn. Daarbij wordt als argumenten gehanteerd dat men er zelf niets mee wil en daarom data niet beschikbaar wil stellen, dat men sceptisch is over wat er mee kan waarbij meespeelt dat men vindt dat duiding zonder context praktisch onmogelijk is, en er nog weinig is gestandaardiseerd en dat met veel data "vervuiling" meekomt. Sommige onderwijsinstanties zien het als een risico dat data over hun onderwijsprestaties in beheer zijn bij derden.

Een derde overweging bij het beschikbaar stellen van data is het nut dat men inziet van big data onderzoek. Als men niet overtuigd is van het nut van dergelijk onderzoek is uiteraard de bereidheid om data ter beschikking te stellen klein.

Een tussenweg bij het beschikbaar stellen van data is dat men daar wel toe bereid is, maar alleen onder strikte en specifieke voorwaarden. Het merendeel van de scholen, data beheerders en de commerciële bedrijven gaven aan data ter beschikking te willen stellen en de voordelen van big data onderzoek te zien, maar dat ze de data beschikbaar wilden stellen onder bepaalde condities. De meningen over het beschikbaar stellen van data en de randvoorwaarden waaronder men wil meewerken lopen sterk uiteen. Ze variëren in mate van anonimiseren, granulariteit, aan wie de data beschikbaar worden gesteld, ontsluiting door een trusted third party, en de manier waarop partijen toegang kunnen krijgen.

Sommige respondenten willen hun data geanonimiseerd, gepseudonimiseerd, of juist niet geanonimiseerd beschikbaar stellen voor wetenschappelijk onderzoek. Andere respondenten geven aan dat de data ook beschikbaar zou mogen zijn voor de Onderwijsinspectie, voor leveranciers of voor scholen. De ene partij wil alle data beschikbaar stellen, terwijl andere partijen met name geaggregeerde data beschikbaar wil stellen. Sommige partijen willen data beschikbaar stellen op aanvraag, al dan niet met een strenge controle door een onafhankelijke commissie, met of zonder verdere restricties. Tenslotte zijn er partijen die een groot voorstander zijn van open data, die beschikbaar zijn voor iedereen.

Op het moment dat deze restricties wat betreft de toegang in een duidelijke bewerkersovereenkomst zijn vastgelegd, kunnen de data worden gedeeld. Een grote database waarin alle partijen op vrijwillige basis data leveren lijkt minder goed haalbaar: iedereen wil hier wel gebruik van maken, maar is huiverig om eigen data breder beschikbaar te stellen dan de nabije cirkel.

Door meerdere partijen werd erkend dat het onderling vertrouwen niet groot genoeg is om buiten de eigen groep te delen. Zo noemde een van de partijen "ik wil een honderd procent, wettelijk vastgelegde garantie dat Justitie er niet in komt." Als autoriteit voor het beheer van gedeelde data wordt geregeld het CBS genoemd, vanwege de ervaring die deze partij heeft met het koppelen en beheren van data en vanwege de specifieke rol die het CBS van de overheid heeft gekregen. Het CBS heeft bovendien de mogelijkheid om onderwijsdata te koppelen met data uit andere bronnen en ze beschikt over procedures waaronder data vrijgegeven kan worden voor onderzoek. In verschillende interviews wordt gewezen op het belang van een trusted third party die, eventueel los van het CBS, geen data beheren, maar wel verantwoordelijk is voor de koppeling en/of de encryptie van de gegevens.

Naast het delen van data, wijzen respondenten ook op het belang van het delen van algoritmes voor het analyseren van de data, bij voorkeur met een autoriteit die de algoritmes checkt. Op die manier wordt inzichtelijk gemaakt wat er met de data gebeurt. Leerlingen, docenten en scholen hebben er recht op om dit te weten en kunnen in beroep te kunnen gaan als ze het gevoel hebben oneerlijk te worden behandeld. Tenslotte wordt bij het delen van data in meerdere interviews bij wijze van voorbeeld verwezen naar de zorg, als een sector die al veel verder is, dan het onderwijs in het op een verantwoorde manier delen en analyseren van big data.

6 Mogelijkheden en onmogelijkheden van big data: technische, juridische en ethische aspecten

In dit hoofdstuk behandelen we de technische, ethische en juridische aspecten van big data zoals die door onze respondenten werden genoemd en herkend. De vragen die aan de orde zijn betreffen de mogelijkheden tot koppeling van data en de kwaliteit van data, de verantwoordelijkheid en aansprakelijkheidsaspecten van dataverzameling en - analyse, en het belang en bescherming van het individu.

6.1 Technische aspecten

Een belangrijk aspect bij de ontsluiting van data betreft de mogelijkheid om data afkomstig uit verschillende systemen aan elkaar te koppelen op basis van gemeenschappelijke indicatoren (Piety, 2013). De technologie voor het koppelen van data is in principe beschikbaar. Er zijn meerdere commerciële partijen op de markt die binnen bijvoorbeeld de zorg, of binnen de financiële wereld, reeds dergelijke koppelingen hebben gerealiseerd. De vraag is alleen in hoeverre deze technologie direct toepasbaar is op onderwijsdata.

De Amerikaanse Data Quality Campaign (zie www.dataqualitycampaign.org) beschrijft een aantal essentiële voorwaarden voor het realiseren van longitudinale datasystemen, zoals waarschijnlijk ook van belang als het gaat om het ontsluiten van onderwijsdata. Deze voorwaarden zijn:

1. Unieke leerlingnummers die per database hetzelfde zijn en gekoppeld kunnen worden.
2. Informatie per leerling over het onderwijs dat de leerling volgt, demografische kenmerken en informatie over het onderwijsprogramma.
3. Het kunnen koppelen van toetsresultaten van individuele leerlingen over de jaren heen om groei te kunnen meten.
4. Informatie over leerlingen die niet getoetst zijn en waarom ze bepaalde toetsen gemist hebben.
5. Een identificatie van docenten en de mogelijkheid om docenten aan leerlingen te koppelen.
6. Leerlinginformatie, zoals de vakken/cursussen die voltooid zijn, cijfers etc.
7. Slaag- en drop-out data per leerling.
8. Het kunnen koppelen van data uit verschillende onderwijssectoren.
9. Een audit systeem dat de kwaliteit, betrouwbaarheid en validiteit van al deze data bewaakt.

Naast een deterministische koppeling van gegevens uit verschillende databases op basis van een unieke identifier, zoals een leerlingnummer, kan data ook probabilistisch gekoppeld worden. Bij probabilistisch koppelen wordt gebruikt gemaakt van een kansmodel, waarbij aan elke koppeling van verschillende records uit twee of meer databases een waarschijnlijkheid wordt meegegeven (Willenborg & Heerschap, 2010). Bij sommige koppelingen zal het van belang zijn om gebruik te maken van data-harmonisatie. Hieronder worden technieken verstaan die het mogelijk maken om data uit verschillende heterogene bronnen op een dusdanige manier te koppelen dat consistente en ondubbelzinnige data ontstaat, die bruikbaar is voor verdere analyse.

Respondenten geven aan dat gebrek aan standaardisatie belemmerend werken kan werken voor het koppelen van de data en dat dit voor ongestructureerde data een groter probleem lijkt, gezien de grote verschillen tussen docenten onderling, tussen scholen en wat betreft de manier waarop deze gegevens zijn vastgelegd. Bij deze problemen spelen praktische zaken een rol, zoals systemen die niet ontwikkeld zijn om data vrij te geven, data formats die niet op elkaar aansluiten, of systemen die alleen maar ontsloten kunnen worden door leveranciers. De instantie die de gegevens koppelt zou daarom moeten werken met een protocol dat duidelijk gecommuniceerd wordt met gebruikers en leveranciers, zodat data gestandaardiseerd verzameld, opgeslagen, gedeeld, gekoppeld en ontsloten kunnen worden. Opslag en transport van data wordt door de experts niet als een groot probleem gezien.

Tijdens de interviews gaven verschillende respondenten aan dat de overheid een belangrijke rol zou kunnen spelen bij het formuleren van standaarden voor gegevensopslag en voor het delen van gegevens. Als deze opslag centraal zou worden gefaciliteerd dan noemen de respondenten het CBS als mogelijke locatie. Daarbij wijzen sommige respondenten op de mogelijkheid om data op meerdere plekken op te slaan en deze via slimme koppelingen samen te voegen op het moment dat een specifieke onderzoeksvraag hierom vraagt. Dat biedt voordelen bij de beveiliging. Bovendien werd aangegeven dat duidelijkheid over hoe de database er technisch uit komt te zien wel nodig is om scholen en instellingen vertrouwen te geven indien data centraal worden opgeslagen. Daarbij zou ervoor gezorgd moeten worden dat de data gekoppeld zijn aan de juiste context, zodat kan worden gegarandeerd dat data zodanig aangeboden en ontsloten wordt, dat ze op een juiste manier wordt geïnterpreteerd.

Naast issues die spelen bij het ontsluiten van de data, spelen ook de (on)mogelijkheden van het analyseren van de data een belangrijke rol. Bij het analyseren van big data wordt er vaak onderscheid gemaakt tussen verschillende fasen. Allereerst is er de pre-processing, waarin de data geschikt gemaakt wordt voor analyse. Het koppelen van data kan hierbij een rol spelen. Deze fase is vaak erg tijdrovend. Een tweede fase is de feature extractie, waarbij tekstmining, audiomining of videomining technieken gebruikt kunnen worden om ongestructureerde data om te zetten in variabelen die geschikt zijn voor analyse. Tenslotte kan gebruik gemaakt worden van datamining om relaties te vinden tussen variabelen en om onderzoeksvragen te beantwoorden. Deze drie stappen vragen om specifieke kennis op het gebied van data-handling en -analyse.

Voor het analyseren onderwijsdata wijzen respondenten bijvoorbeeld op de geneste structuur van veel onderwijsdata, waarbij leerlingen genest zijn binnen klassen, klassen genest zijn binnen scholen, etc.. Als daar geen rekening mee gehouden wordt bij de data-analysekunnen de resultaten vertekend worden.

Daarnaast geven respondenten aan dat ze zelf meestal niet beschikken over de kennis die nodig is om big onderwijsdata te analyseren, waardoor ze afhankelijk zijn van onderzoekers van universiteiten of van commerciële partijen.

6.2 Juridische aspecten

Bij juridische aspecten speelt de vraag: Hoe kan big data analyse zo worden ingericht dat er een juridische heldere en beschermde omgeving voor data analyse te realiseren valt? Het betreft hier vragen op het gebied van het eigendom van de data, privacy en randvoorwaarden voor het gebruik van de data, de grenzen die de huidige en toekomstige wetgeving legt aan de dataverzameling en -bewerking en de vraag wie toezicht houdt op het ontsluiten en bewerken van onderwijsdata. Momenteel worden deze aspecten voornamelijk geregeld door de Wet Bescherming Persoonsgegevens (WBP). Op 25 mei 2018 wordt echter de Algemene Verordening Gegevensbescherming (AVG) van kracht. Dit is een EU-brede verordening, die is opgesteld om de rechten en de privacy van het individu beter te beschermen. Deze overgang betekent dat de respondenten soms vanuit een verschillend kader hebben geantwoord, waarbij sommige respondenten bovendien meer gespecialiseerd waren in wetgeving dan anderen. De nieuwe komende wetgeving zal zich nog verder moeten uitkristalliseren.

6.2.1 Eigendomsrecht

Bij databewerking moet in eerste instantie vastgesteld worden of het om persoonsgegevens gaat. Een persoonsgegeven is een gegeven betreffende een geïdentificeerde of een identificeerbaar persoon. Dit betekent dat wanneer met redelijke inspanning achterhaald kan worden welke persoon er achter de data zit, data niet meer als anoniem mag gelden. Anonimiteit blijkt in de interviews een belangrijk aspect te zijn.

Respondenten geven aan dat anonimiteit bij het koppelen van aparte datasets moeilijker te garanderen is. Door het koppelen van grote hoeveelheden data wordt uiteindelijk zoveel informatie over een persoon verzameld, dat van anonimisering mogelijk geen sprake meer is. In de gesprekken is ook meerdere malen genoemd dat zo bezien "anonieme data" misschien wel niet meer bestaan.

Persoonsgegevens (niet anonieme gegevens) kunnen in principe op twee grondslagen bewerkt kunnen worden: na expliciete en ondubbelzinnige toestemming van de persoon van wie de data zijn, of op grond van gerechtvaardigd belang. In het eerste geval moet er per bewerkingsdoel toestemming gevraagd worden. Het doel dient dan concreet, duidelijk, en ondubbelzinnig gecommuniceerd te worden. Bovendien moet de data-leverende partij in de positie zijn om geen toestemming te geven zonder daar significante nadelen van te ondervinden (zoals afgewezen worden voor een opleiding) en om de toestemming in te trekken. Bij dataverzameling en -verwerking op basis van de tweede grondslag, moet er kunnen worden aangetoond dat inbreuk op de privacy van het individu in dit geval proportioneel en gerechtvaardigd was. Onder de huidige wetgeving kan dit belang achteraf nog worden vastgesteld, nadat de bewerking heeft plaatsgevonden. Onder de AVG moet het gerechtvaardigd belang echter van tevoren worden vastgesteld. In beide gevallen van gegevensverwerking heeft het individu een aantal rechten, die bepalen wat er wel en niet met zijn persoonsgegevens mag gebeuren. Hieronder vallen het recht op inzage van de data die over hem of haar verzameld is, het recht op aanpassing of verwijdering indien de data onjuist of onvolledig zijn, het recht op een menselijke beoordeling (en daarmee samenhangend het recht om niet onderworpen te worden aan profilering aan de hand van geautomatiseerde classificatie of voorspellingen) en het recht om te weten op welke gronden een beoordeling of beslissing tot stand is gekomen. Wanneer het specifiek over leerlingdata gaat, wordt de data bovendien als zeer gevoelig bestempeld. Data over

hoe iemand leert, worden onder gedragsgegevens gerekend. Ten aanzien van geaggregeerde data is de wetgeving minder specifiek.

6.2.2 Privacy

Een tweede belangrijk juridisch aspect is privacy. Onder de AVG wordt - sterker nog dan bij de WBP - de nadruk gelegd op de bescherming van de privacy van het individu. Dit vertaalt zich in onder andere het uitgangspunt van dataminimalisatie (alleen de data verzamelen die strikt noodzakelijk is om een bepaald doel uit te voeren). Daarnaast vereist de AVG dat er een privacy impact assessment (PIA) wordt uitgevoerd voordat er data worden verwerkt, waarbij er expliciet wordt nagedacht over hoe de privacy van het individu wordt aangetast door de verwerking, wat voor risico's dit met zich meebrengt voor het individu, en of er ook een minder ingrijpend alternatief is voor de bewerking. Verder moeten databeheerders aantoonbaar verantwoordelijk met data omgaan. Dit behelst onder andere het optuigen van een procedure waardoor mogelijke datalekken binnen 72 uur gemeld worden bij de Autoriteit Persoonsgegevens, goede beveiliging van data, en het toepassen van privacy-by-design, waarbij al tijdens de ontwikkeling aandacht besteed wordt aan privacy verhogende maatregelen en dataminimalisatie.

Uit de interviews lijkt het volgende scenario in de praktijk het meest voor te komen. De school verzamelt data over de leerling vanwege de wettelijke taak van het geven van onderwijs. Deze administratie is nodig om bij te houden welke leerlingen naar school gaan, wat hun voortgang is, en met wie er contact moet worden gezocht als er zaken besproken moeten worden. De school is daarmee verantwoordelijk voor het databeheer en de beveiliging. Bij het gebruik van leermiddelen (zoals een leerlingvolgsysteem, de tussentijdse toets van Cito, of educatieve apps) wordt ook data gegenereerd. Een groot deel van deze data ligt opgeslagen bij de leverancier van het leermiddel. Wat er met deze data mag gebeuren wordt vastgelegd in de bewerkersovereenkomst, die wordt afgesloten tussen de onderwijsinstelling en de leverancier. Hierin wordt expliciet vastgelegd of de data gedeeld mag worden (en zo ja, met wie), welke analyses mogen worden toegepast en welke data aan de onderwijsinstelling wordt aangeleverd. Wetenschappelijk onderzoek is standaard als doel van verwerking benoemd. Alle bewerkingen die niet in de bewerkersovereenkomst staan vermeld, mogen niet door de leverancier worden uitgevoerd.

Momenteel wordt data meestal opgeslagen in databases, die vervolgens in hun geheel versleuteld kunnen worden. Het is echter ook mogelijk om de data per individu te versleutelen en de sleutel aan het individu terug te geven. Op deze manier kan de databeheerder niet meer bij de individuele data zonder toestemming (en de sleutel) van het individu. Het blijft echter mogelijk om analyses over de versleutelde data uit te laten voeren. Zo kunnen er analyses worden uitgevoerd zonder dat individuele data op enig punt beschikbaar hoeft te worden gesteld. Een belangrijke noot die hierbij moet worden gemaakt is dat deze techniek nog in de kinderschoenen staat. Het is onderzocht in de academische setting, maar wordt vooralsnog nog zeer weinig in de praktijk gebruikt. Als het echter werkbaar blijkt, dan zou het een alternatief kunnen bieden voor het gerechtvaardigd belang en individuele toestemming.

6.2.3 Belang

Boven is al aangegeven dat in geval van toestemming voor het gebruik persoonsgegevens het doel moet worden gecommuniceerd. Ook in het geval van anonieme gegevens is het doel van de dataverzameling en analyse van belang, de vraag is hier welk doel gediend wordt en of dit doel in het belang is van degenen die de data geleverd hebben. Hieraan gekoppeld is het idee van proportionaliteit. Staan het verwerven van de data en de risico's die met dataopslag gepaard gaan, in verhouding met het uiteindelijke doel waar de data voor gebruikt worden?

Uit de interviews blijkt dat het hier om complexe afwegingen kan gaan. Leerlingdata door een leverancier laten gebruiken, opdat die zijn product kan verbeteren en er commercieel op vooruit heeft geen direct leerlingbelang, wellicht kan hierdoor zelfs de prijs van het product omhoog gaan. Anderzijds heeft een school (en leerlingen in het algemeen) er wel baat bij dat een leermiddel verbeterd wordt. Dit is een heel delicaat evenwicht, waarbij de belangen van de privacy van de individuele leerling worden afgewogen tegen het belang van de onderwijsinstelling, en de kosten van het commercieel gebruik van persoonlijke data tegen de baten van verbeterd onderwijs. Aangezien er op basis van gerechtvaardigd belang data wordt uitgewisseld, dient de school heel secuur te zijn in het beschermen van de privacy van haar leerlingen bij het vastleggen van de bewerkersovereenkomst.

Met het oog op de praktijk merkte een jurist tijdens een interview op: "Scholen hebben vaak geen weet van hoe groot het onderwerp van gegevensbescherming en privacy is, en hoe belangrijk het is dat je het regelt." Ook andere juristen gaven aan dat databeveiliging vaak als een lastig onderwerp wordt gezien. Enerzijds willen onderwijsinstellingen de privacy van hun leerlingen beschermen, anderzijds is er nog veel verwarring over wat wel en niet mag, en wordt privacy wetgeving - ook door andere partijen die geïnterviewd zijn - als vooralsnog niet volledig helder gezien. Onterecht, stellen verschillende geïnterviewde juristen: "Hoewel privacy in het onderwijs soms nog echt gezien wordt als een belemmering, kan het juist een enabler kan zijn om echt gepersonaliseerd leren mogelijk te maken. Als privacy gebruikt wordt om de beveiliging op te schroeven, dan kun je van tevoren eisen stellen aan de leveranciers en met de juiste waarborgen bijvoorbeeld adaptive learning toepassen, waarbij gevoeliger data zullen ontstaan."

6.3 Ethische aspecten

Volgens Franzke (2016) zijn er vier aspecten die centraal staan in de discussie over ethiek en big data. Op de eerste plaats zijn data die verzameld wordt alomvattend, het raakt alle onderdelen van een mensenleven en het is vrijwel onmogelijk om iets te doen zonder een dataspoor achter te laten. Op de tweede plaats is big data vaak expliciet; precieze persoonlijke informatie (plekken waar men geweest is, consumptiepatronen etc.) kan duidelijk worden. Op de derde plaats, wordt het door het aggregeren van data voor de betrokkenen vaak onmogelijk om in te schatten wat de consequenties zijn van het ter beschikking stellen van persoonlijke data in afzonderlijke databases. Tot slot, data worden steeds meer permanent waardoor 'the right to be forgotten' (Weber, 2011) niet meer kan worden uitgeoefend.

Deze aspecten zagen we ook terugkomen in de interviews waarin een gebrek aan privacy en een gebrek aan transparantie naar voren kwamen als ethische bezwaren. Privacy en privacyschending kwamen doorgaans het eerst in de respondenten op als het ging over ethische issues rondom big data. Wanneer elke muisklik van een leerling wordt bijgehouden, dan kan dit de persoonlijke levenssfeer bedreigen. Verschillende geïnterviewden gaven aan dat juist binnen de school, waar een leerling veilig moet kunnen oefenen, falen en leren (op zowel cognitief als sociaal vlak), privacy belangrijk is. Gerelateerd aan de ethische aspecten van privacy wordt door respondenten het probleem van 'context collapse' genoemd. Context collapse hangt samen met het feit dat het van de context afhangt welke informatie men wil delen en hoe men zich wil profileren. Met het samenvoegen van databases kan deze context echter verloren gaan en is men niet meer in staat het gedrag aan te passen aan de situatie. Informatie over uitgaansgedrag, bijvoorbeeld, wil men meestal wel delen met vrienden maar niet met de werkgever. Dit verlies aan controle kan twee verschillende effecten hebben. Of men wordt heel erg huiverig voor het delen van data of, en de respondenten constateren dit gedrag vooral onder jongeren, er ontstaat een zekere gelatenheid: het wordt inmiddels zó lastig geacht om de privacy te beschermen, dat men het maar helemaal opgeeft.

Een gebrek aan transparantie kwam weliswaar minder snel ter sprake als een ethisch bezwaar, maar werd alsnog regelmatig in de gesprekken benoemd als (potentieel) ethisch probleem. Transparantie komt er onder andere op neer dat een individu het recht heeft om te weten hoe een beoordeling tot stand is gekomen. Wanneer er data worden verzameld voor een beoordeling, weet men daarmee direct op welke gegevens de uitslag is gebaseerd. Op een toets kan een leerling zich voorbereiden, bij een kennismakingsgesprek kan een toekomstige student zich van zijn of haar beste kant laten zien. Bovendien kan men bezwaar maken als er irrelevante onderwerpen worden meegenomen bij het verzamelen van de data. Wanneer big data wordt toegepast, wordt dit proces ondoorzichtig. In de eerste plaats verliest een individu grip op welke data waarvoor wordt gebruikt; in principe kan alles wat ooit is vastgelegd de leerling benadelen. In de tweede plaats wordt het onduidelijk waar beslissingen op gebaseerd zijn. Algoritmes worden zelden openbaar gemaakt, en daarmee verliest de leerling inzicht in waar hij op beoordeeld wordt (en daarmee het vermogen zichzelf tegen die beoordeling te verdedigen). Eén van de ethici merkte op dat zelfs als de leerling precies weet welke data er zijn opgeslagen, er alsnog een gerede kans bestaat dat hij alsnog niet in staat is om de juistheid en eerlijkheid van de data en de algoritmes te beoordelen.

Een gevolg van niet-transparante beoordeling dat door ruim een derde van de geïnterviewden werd benoemd, is het risico van profilering en discriminatie. Wanneer er persoonsgebonden adviezen worden gegeven op basis van data, zal een deel van de leerlingen een stempel opgedrukt krijgen dat niet klopt, maar waar zeer lastig onderuit te komen is.

Onderzoekers uit het veld van big data (O'Neil, 2016) hebben al gewaarschuwd dat wanneer een algoritme eenmaal is vastgesteld, er zelden wordt nagegaan hoe het zit met de false negatives. Dit zijn de personen die ten onrechte werden afgewezen of als probleemgeval zijn aangemerkt. Waar een false positive snel genoeg door de mand valt, vallen de false negatives zelden op; men weet immers niet hoe zij het hadden gedaan als ze niet de speciale ondersteuning hadden gekregen, of een niveau hoger waren geplaatst. De uitdaging bij het toepassen van kennis die gebaseerd is op big data is om het aantal false negatives te minimaliseren en alle leerlingen de ondersteuning en begeleiding te bieden die ze nodig hebben.

6.4 Slot

In de literatuur en tijdens de interviews komen een aantal onmogelijkheden en risico's van het werken met big onderwijsdata naar voren. Binnen het onderwijs is een gebrek aan standaardisatie van de manier waarop data zijn opgeslagen. Systemen sluiten niet op elkaar aan en zijn niet ingericht op het exporteren van data. Dit bemoeilijkt het delen van data. Specifieke kennis op het gebied van data-analyse is bovendien te weinig aanwezig. De wetgeving op het gebied van dataverzameling en -bewerking is daarnaast sterk in ontwikkeling en de interpretatie ervan binnen de sector is niet eenduidig. Het doel van big data-analyses is bijvoorbeeld niet altijd van te voren helder, wat de mogelijkheden voor het analyseren beperkt. Vanwege de complexe aard van de analyses, is de transparantie tenslotte een issue, wat tot ethische bezwaren kan leiden.

Anderzijds zijn er ook respondenten die vinden dat er te krampachtig wordt gekeken naar de ethische bezwaren en risico's. Zeker wanneer het om wetenschappelijk onderzoek gaat, wordt het belang van de kwaliteit van het onderwijs in het algemeen vaak zwaarder gewogen dan het belang van de individuele leerling. Wanneer het om commerciële bedrijven gaat wordt men doorgaans een stuk terughoudender. Eveneens wanneer er de mogelijkheid is om op grond van de data beoordeeld of 'afgerekend' te worden (zoals bijvoorbeeld het geval is als de data worden gedeeld met de Onderwijsinspectie, of met het schoolbestuur) legt men zwaarder de nadruk op de privacybescherming van het individu. Tenslotte werd tijdens de interviews opgemerkt dat de normen en waarden voor wat acceptabel wordt geacht in verband met privacy constant aan verandering onderhevig zijn.

Belemmerende en bevorderende factoren 7

In dit hoofdstuk gaat het om het in kaart brengen van de belemmerende en bevorderende factoren voor het gebruik van big data voor verschillende groepen gebruikers. Naast de juridische en ethische aspecten die al in het vorige hoofdstuk zijn besproken, is er in dit hoofdstuk aandacht voor sociale implicaties en voor belemmeringen met betrekking tot de beschikbaarheid en de kwaliteit van de data. Ook is er aandacht voor belemmeringen t.a.v. de infrastructuur en voor risico's die te maken hebben met de capaciteit en de competenties die nodig zijn voor het analyseren en interpreteren van big onderwijs data. Tenslotte wordt ingegaan op de kansen van big onderwijsdata en op bevorderende factoren voor het gebruik van big data.

7.1 Risico's en belemmerende factoren van big data in het onderwijs

7.1.1 Sociale implicaties

Big data analyse kan sociale implicaties hebben. Met behulp van big data kunnen bijvoorbeeld goed presterende scholen, docenten en leerlingen alsmede onderpresterende scholen, docenten en leerlingen geïdentificeerd worden. Dit roept een aantal vragen op zoals de vraag of leerlingen, docenten en/of scholen hierover geïnformeerd moeten worden. (Eynon, 2013). Of sluiten we misschien op basis van big data bepaalde leerlingen uit, zoals plaatsvond in de VS, als één van de gevolgen van de No Child Left Behind Act. Goed presterende leerlingen werden hier uitgesloten van testafname en de focus lag bij sommige scholen vooral op de leerlingen die de benchmark net wel/niet zouden halen (Piety, 2013). Wat doen we met geluk en toeval in een systeem waarin we veel kunnen voorspellen? Is er nog leren dat privé is? Welke gevolgen heeft het openlijk volgen van leerprestaties op het leerproces van leerlingen? (Eynon, 2013). Is het belangrijker om een methode toe te passen die mogelijk onterecht docenten aanwijst als 'slechte' docenten, om leerlingen te beschermen, of is het belangrijker om docenten te beschermen tegen oneerlijke evaluaties, wat wel impliceert dat sommige minder goede docenten hun baan houden (Piety, 2013). Het is belangrijk om stil te staan bij dit soort vragen, aangezien zowel het verzamelen alsmede het analyseren van data nooit waardevrij is (Eynon, 2013).

Stigmatisering, profilering, labeling, self-fulfilling prophecies en circular reasoning worden door meer dan de helft van de respondenten genoemd als gevaren of risico's van big onderwijsdata. Deze fenomenen hoeven niet alleen moedwillig te gebeuren, maar ze kunnen ook optreden doordat de gebruiker niet in staat is om de uitkomsten en feedback op basis van de analyses op een juiste manier te interpreteren en toe te passen. Tijdens de interviews werd door de respondenten niet alleen gewezen op deze risico's, maar ook op de terughoudendheid die bij veel scholen heerst, omdat ze niet overzien wat de implicaties zijn van het analyseren van big data voor leerlingen en docenten. Ook wordt aangegeven dat scholen nog weinig bereid zijn om hun data te delen. Vanuit de onderzoekers wordt er op gewezen dat er weinig discussie zal zijn op het moment dat de belangen van verschillende partijen op één lijn liggen. Op het moment dat dit niet het geval is, zullen er duidelijke richtlijnen nodig zijn, die aangeven hoe er gehandeld moet worden. Bij het vergelijken is er altijd een partij die iets te verliezen heeft. Onderzoekers wijzen er bovendien op dat trends in het onderwijs lastig te interpreteren zijn, doordat het onderwijsveld continu verandert.

Big data houdt het risico in zich dat de ongelijkheid in de samenleving toeneemt. De eerste vraag is bijvoorbeeld over wie veel data beschikbaar is. Dit zijn waarschijnlijk personen met een betere positie in de samenleving. Deze personen maken bijvoorbeeld meer gebruik van sociale media als Twitter en zullen meer gebruik maken van allerlei zoekmachines op internet (Enyon, 2013). Ten tweede speelt de vraag wie toegang heeft tot deze data. Vaak zijn de eigenaren van deze databases commerciële bedrijven (Boyd & Crawford, 2012). Organisaties met veel geld kunnen toegang tot deze data kopen. Toegankelijkheid en gebruik van deze data zijn waarschijnlijk dus ongelijk verdeeld (Boyd & Crawford, 2012; Eynon, 2013).

Het verspreiden van data onder derden, waarbij het belang van de leerling niet altijd voorop staat, wordt in de interviews inderdaad genoemd als een van de risico's. Respondenten wijzen nog op het risico van function creep. Resultaten van big data analyse kunnen gebruikt worden voor andere doeleinden dan waarvoor de gegevens bij elkaar gebracht en geanalyseerd zijn. Er wordt ook aangegeven dat er ten aanzien van de voorwaarden waaronder onderzoekers of andere partijen toegang krijgen tot de data nog weinig bekend is.

Nauw gerelateerd aan de sociale implicaties van big data is de ethische kant van big data, zoals ook al omschreven in hoofdstuk 6. Dit kan een risico vormen wat betreft privacy, intellectueel eigendom, informed-consent, beveiliging en bescherming tegen misbruik (Boyd & Crawford, 2012; Douglas, 2015; Enyon, 2013; Ferguson, 2012; Liñán & Pérez, 2015; Manyika et al., 2011; Piety, 2013). Het roept vragen op met betrekking tot van wie de data zijn, welke data gecombineerd kunnen worden, en welke data gecombineerd mogen worden, geanalyseerd moeten worden en voor welke doelen big data gebruikt kunnen worden zoals ook al deels in hoofdstuk 6 is besproken (Boyd & Crawford, 2012; Enyon, 2013; Manyika et al., 2011; Piety, 2013). Een voorbeeld kan gevonden worden in publiekelijk toegankelijke data (denk aan Twitter) waarvan gebruikers zich er vaak niet van bewust zijn dat hun data op allerlei manieren gebruikt kan worden. Deze doelen, zoals het opstellen van modellen die kenmerken als inkomen, etniciteit, handicaps, geloofsovertuigingen, politieke voorkeur en seksuele geaardheid kunnen voorspellen (Liñán & Pérez, 2015; Crawford, 2013 in Landon-Murray, 2016) hoeven niet in overeenstemming te zijn met de doelen van diegenen waar de data betrekking op hebben. Een ander voorbeeld zijn data gegenereerd door docenten die hun functioneren in kaart brengen (Dede, 2016).

Een aantal respondenten benadrukt dat, ondanks het feit dat de data toegankelijk is, dit nog niet wil zeggen dat deze data gebruikt mag worden. Ook hebben wordt gemeld dat stakeholders een tegenstrijdige rollen kunnen hebben als het gaat om big data. Leerlingen genereren bijvoorbeeld data die gebruikt kunnen worden om hun leren bij te sturen, echter, leerlingen of ouders als dataleveranciers of gebruikers hebben vaak heel andere doeleinden voor datagebruik dan bijvoorbeeld managers en bestuurders.

7.1.2. Beschikbaarheid

Een mogelijk belemmerende factor m.b.t. big data betreft de beschikbaarheid van deze data. Ten eerste gaat het hierbij over het feit dat alleen data die beschikbaar is bestudeerd kan worden. We hebben niet over alles data, dus de beschikbaarheid van data is direct van invloed op de vragen die er gesteld worden en het onderzoek dat uitgevoerd wordt. Big data vertelt ons bijvoorbeeld meer over wat mensen doen, maar vertelt ons minder over de achterliggende redenen van een bepaald type gedrag (Enyon, 2013). Het is bijvoorbeeld mogelijk om het aantal relaties dat individuen hebben in een school in kaart te brengen met data (bijvoorbeeld het aantal e-mail contacten, het aantal verzonden e-mails). Echter, dit geeft nog geen informatie over de waarde die deze personen hechten aan deze relaties (Boyd & Crawford, 2012 in Enyon, 2013). In de praktijk wordt er meer data verzameld over concepten die gemakkelijker te meten zijn en is er bijvoorbeeld minder data beschikbaar over waarden, emoties, opvattingen en interacties (Kitchin, 2013 in Landon-Murray, 2016). Belangrijke doelen waarover minder data beschikbaar zijn betreffen daarnaast motivatie, zelfvertrouwen, plezier in leren, tevredenheid en persoonlijke ontwikkeling (Ferguson, 2012). Dit kan ook leiden tot wat Lavertu (2014) 'goal displacement' noemt. Hierbij focust men alleen nog maar op de doelen waarvan men data beschikbaar heeft. Over sommige uitkomsten is veel data beschikbaar, maar over andere uitkomsten en onderwijsdoelen is veel minder (goede) data beschikbaar. Het gevaar is dat de focus vooral komt te liggen op het meetbare, ten koste van andere belangrijke doelen. Ook kunnen aspecten worden vergeten. Tegenvallende leerprestaties kunnen bijvoorbeeld veroorzaakt worden door aspecten buiten de school, waarover geen data beschikbaar is (Dede, 2016).

Ten tweede waarschuwt Piety (2013) ervoor dat veel verschillen tussen scholen, docenten en leerlingen niet in data te vatten zijn. Piety (2013) geeft als voorbeeld dat lesgeven complex is en dat het verschillende dimensies omvat (bijvoorbeeld kinderen begrijpen, vakinhoudelijke kennis, didactische kennis etc.). Er is nog niet over al deze dimensies data beschikbaar. Dat leren complex is en niet altijd even goed te vatten is in data blijkt ook uit het project dat in 2009 in New York gestart is: "School of one". Centraal hierbij staat een algoritme dat op basis van data iedere dag het rooster voor zowel de docent als de leerlingen aangeeft (New York City Department of Education, 2010). Echter, evaluaties van dit programma zijn niet erg positief. Problemen zijn onder andere dat leren niet gereduceerd kan worden tot toetsdata en een algoritme. Ten tweede wordt er niet gewerkt aan de relaties tussen docenten en hun leerlingen en tussen de leerlingen onderling, omdat iedere leerling een eigen programma heeft, terwijl relaties belangrijk zijn voor leren. Tot slot vindt er geen sense making plaats op basis van de data (het algoritme bepaalt) en wordt er niet samengewerkt tussen docenten (Light, 2016).

Een ander voorbeeld betreft het gebruik van big data voor docentevaluaties. Studies (zie bijvoorbeeld Kane, Rockoff, & Staiger, 2008) laten zien dat hoewel het wel mogelijk is om het lesgeven van docenten te meten, hier nog steeds wel een grote mate van onnauwkeurigheid in zit en dat hoewel er een relatie bestaat tussen het lesgeven in de klas (gemeten met lesobservaties) en de leerprestaties van leerlingen, er nog veel variantie onverklaard blijft. De conclusie van Kane et al. (2008) is dan ook dat lesgeven te complex is om in één meetinstrument te vangen. De beschikbare big data stellen dus beperkingen aan de vragen die we kunnen stellen en beantwoorden (Enyon, 2013). Bovendien worden niet alle relevante data digitaal opgeslagen (Piety, 2013).

Respondenten wijzen op problemen die er zijn bij het koppelen van data. Vaak zijn de formats niet op elkaar afgestemd, of ze zijn niet gebruiksvriendelijk opgeslagen. Soms wordt data bijvoorbeeld wel opgeslagen, maar kan deze niet worden geëxporteerd. Van tevoren is bovendien meestal niet bekend waarvoor data in de toekomst gebruikt zouden kunnen worden. Dit heeft als gevolg dat de selectie van opgeslagen data niet altijd past bij de onderzoeksvragen. Zowel vanuit de ethische hoek als door dataleveranciers wordt gewezen op het risico van het ontbreken van een juiste context. Doordat onderwijsdata selectief opgeslagen wordt, ontbreekt vaak informatie die nodig is om de gegevens op een juiste manier te interpreteren. Scholen wijzen bovendien op de beperkte toegankelijkheid van onderwijsdata voor docenten, terwijl zij toch de primaire gebruikers zijn.

7.1.3. Kwaliteit

Aan onderwijsdata kunnen ook beperkingen zitten m.b.t de kwaliteit van de data (Boyd & Crawford, 2012; Douglas, 2015; Gibson & Webb, 2015; Piety, 2013). Piety (2013) geeft hierbij aan dat het soms lastig is om de juiste docent aan de juiste leerlingen te koppelen (één leerling heeft vaak meerdere docenten in het voortgezet onderwijs, maar ook in het primair onderwijs is dit soms lastig omdat veel docenten part-time werken, docenten wisselen tussen scholen enz.), dat sommige zaken bijvoorbeeld niet goed worden bijgehouden (missing data) en ook dat er bewust (hogere leerprestaties doorgeven, dan behaald zijn) of onbewust fouten gemaakt worden bij de invoer van data. Bij het verzamelen van data kan er al van alles misgaan: docenten kunnen leerlingen antwoorden voorzeggen als een belangrijke toets wordt gemaakt, de resultaten van laag presterende leerlingen kunnen niet ingevoerd worden, roosters kunnen slecht bijgehouden worden, enzovoort. Het feit dat veel data handmatig ingevoerd wordt maakt dat deze data gevoelig is voor fouten (Piety, 2013). In dit verband is het ook belangrijk om de term 'meetfout' te noemen. Bij iedere meting is er sprake van een zekere mate van onnauwkeurigheid en dit speelt in het onderwijs zeker een belangrijke rol, aangezien het vaak gaat om zaken die niet eenvoudig te meten zijn, zoals het leren van leerlingen (Piety, 2013). De data geeft bovendien vaak geen compleet beeld van instructie en leren, maar een gefragmenteerd beeld van met name datgene dat relatief gemakkelijk te meten en in toetsen te vatten is (Piety, 2013). Hoewel het bij big data soms om miljoenen datapunten gaat, betekent dit nog niet dat deze random of representatief zijn. Het is belangrijk om stil te staan bij de context, de kenmerken van de steekproef en de beperkingen die er kleven aan de dataset die je gebruikt (Boyd & Crawford, 2012). In verband met het combineren van verschillende datasets, zoals kan plaastvinden bij big data, is het belangrijk om vast te stellen dat iedere dataset fouten met zich meebrengt en dat door het combineren van deze datasets steeds meer fouten ontstaan (Bollier, 2010 in Boyd & Crawford, 2012).

In veel interviews kwam de gebrekkige kwaliteit van met name de ongestructureerde of zachte data aan de orde. Respondenten wijzen er op dat er veel ruis in de data zit, die er niet alleen voor zorgt dat deze data moeilijk te koppelen zijn, maar ook dat het lastig is om de verzamelde data op een juiste manier te interpreteren. Er wordt ook aangegeven dat het misschien verstandig is vast te stellen wat toelaatbaar gebruik van data is. Niet alleen vanuit privacy oogpunt, maar ook wat betreft de (on)mogelijkheden voor analyse gezien de kwaliteit van de data. Vanuit de scholen wordt ook aangegeven dat docenten op heel verschillende manieren omgaan met het archiveren van data. Onderzoekers geven aan dat ongestructureerde data op dit moment vaak ongeschikt is om te delen. Tot slot wordt opgemerkt dat de keuzes die gemaakt worden bij de analyse (bijvoorbeeld welke indicatoren worden gebruikt om bepaalde concepten te operationaliseren, of en zo ja voor welke factoren er gecorrigeerd wordt) van data per definitie het antwoord kleuren. Dit bemoeilijkt de interpretatie van de resultaten.

7.1.4. Infrastructuur

Een andere belemmerende factor heeft te maken met de infrastructuur. Data moeten opgeslagen kunnen worden, gedeeld kunnen worden, gecombineerd kunnen worden en geanalyseerd kunnen worden, zowel door één schoolorganisatie als over schoolorganisaties heen. Data zijn nu nog vaak opgeslagen in verschillende systemen, 'data silo's', die niet zomaar gekoppeld kunnen worden (Douglas, 2015; Eynon, 2013; Piety, 2013). Het betreft hier bijvoorbeeld leerlinginformatiesystemen, social media, management informatie systemen, administratieve systemen, leeromgevingen enz. Tevens is deze data in verschillende formats beschikbaar, zoals audio, video, tekst en foto's (Douglas, 2015). Bowker en Star (1999 in Williamson, 2016) hebben het bij infrastructuur niet alleen over de technische kant hiervan (datasystemen, codes, algoritmes etc.), maar ook over de sociale kant: het nieuwe soort expertise dat nodig is binnen organisaties voor het analyseren van data, het genereren van kennis, het presenteren en communiceren hierover.

Een aantal respondenten wijst op de problemen die ontstaan doordat de data moeilijk toegankelijk is en niet op elkaar is afgestemd. Ook de scholen geven aan dat hun infrastructuur nog niet zo ingericht is dat zij hun data gemakkelijk kunnen delen. Vanuit de onderzoekers wordt aangegeven dat data die met veel moeite is verzameld en gekoppeld, bovendien niet gemakkelijk met anderen wordt gedeeld. Er wordt ook gewezen op de voordelen van decentralisatie als het aankomt op het beveiligen van de gegevens.

7.1.5. Capaciteit en competenties

Een mogelijke belemmerende factor bij big data betreft het gebrek aan de vereiste capaciteit, kennis en vaardigheden (data literacy) om big data te interpreteren en hiervan gebruik te kunnen maken (Eynon, 2013; Lavertu, 2014; Liñán & Pérez, 2015; Manyika et al., 2011). Dit betreft een belemmerende factor voor wetenschap, beleid, praktijk en het bedrijfsleven. In de wetenschap wordt data science en datamining steeds belangrijker, maar niet iedereen beheerst dit even goed. De vraag is of dit opgelost kan worden door samen te werken in multidisciplinaire teams, of dat dat onvoldoende is (Eynon, 2013). Op het niveau van beleid en praktijk zien we nu al de beperkingen van datagebruik door scholen, schoolbesturen en instellingen. Er is een groot gebrek aan data literacy. Dit betreft verschillende aspecten van data literacy, zoals het opstellen van meetbare doelen, het verzamelen van data, het analyseren en interpreteren van data, het ontwikkelen en implementeren van maatregelen op basis van data en evalueren van die maatregelen (Liñán & Pérez, 2015; Ebbeler, Poortman, Schildkamp & Pieters, in press; Keuning & van Geel, 2016; Piety, 2013; Schildkamp, Karboutzki, & Vanhoof, 2014; Schildkamp & Kuiper, 2010; Schildkamp & Poortman, 2015; van der Scheer, 2016). Bij de interpretatie van data kan er bijvoorbeeld veel misgaan (Eynon, 2013; Lavertu, 2014). De personen die beslissingen moeten nemen op basis van big data hebben niet altijd de benodigde expertise om de analyses te begrijpen en te interpreteren. Dit brengt als risico met zich mee dat de data verkeerd worden geïnterpreteerd en dat er verkeerde beslissingen worden genomen (Lavertu, 2014). Bij het interpreteren van big data speelt bias en subjectiviteit altijd een rol en het is belangrijk om dit te erkennen. Ook is het belangrijk bij het interpreteren van big data om rekening te houden met context (Boyd & Crawford, 2012; Ozga, 2009). Harford (in Landon-Murray, 2016) stelt hierover dat als je niet weet wat er achter een correlatie zit, dan weet je ook niet hoe je de correlatie kunt beïnvloeden. Dit alles houdt in dat er steeds meer behoefte is aan opleiding en training op het gebied van (big) data, zowel m.b.t. de technische kant (verzamelen en analyseren) als de sociale kant (begrijpen van data, beslissingen nemen op basis van data. Hierbij valt bijvoorbeeld te denken aan

data coaches en train-de-trainer cursussen (Dede, 2016).

Een groot aantal van de respondenten wijst op de risico's die ontstaan, doordat niet iedereen beschikt over de benodigde kennis en vaardigheden om big onderwijsdata te analyseren en te interpreteren. Er wordt op gewezen dat er een kloof (power gap) is ontstaan tussen de leerlingen/studenten/docenten die de data produceren en de personen die die data analyseren. Scholen en onderzoekers benoemen dat docenten wel in staat moeten zijn om feedback op basis van big data te kunnen interpreteren om daarvan te kunnen profiteren bij hun onderwijs. Er is een kloof tussen aan de ene kant de groei van de hoeveelheden data en de gebruiksmogelijkheden en aan de andere kant de capaciteit en de expertise die daarvoor nodig is.

7.2 Kansen en bevorderende factoren m.b.t. big data

Afgezien van de belemmerende factoren wat betreft de beschikbaarheid en kwaliteit van big data, maken nieuwe technologieën het wel mogelijk dat er steeds meer data beschikbaar komen (Piety, 2013; Williamson, 2016). Het betreft hier bijvoorbeeld online tools die docenten en leerlingen gebruiken in het onderwijs, digitale presentaties, quizjes en andere interactieve werkvormen via instrumenten als Kahoot en Menti, video's, games, simulaties, online assessments, digitale leeromgevingen etc. Via al deze systemen wordt data verzameld en opgeslagen over de gebruiker. Deze data kunnen potentieel inzicht geven in voorkeuren van gebruikers, patronen van gebruik en manieren van leren (Piety, 2013). Het wordt steeds beter mogelijk om individuen m.b.t. steeds meer verschillende aspecten over de tijd heen te volgen (Williamson, 2016). Een vorm van data die waarschijnlijk in de toekomst ook steeds vaker beschikbaar komt betreft data afkomstig van Computerized Adaptive Testing (CAT) die online worden afgenomen. Deze toetsen passen zich aan het niveau van de leerlingen aan en de uitkomsten geven inzicht in de vaardigheden en problemen van individuele leerlingen. Niet alleen de resultaten van de toetsen vormen een belangrijke vorm van data, maar ook bijvoorbeeld de responstijd per vraag die bijgehouden kan worden (Thompson, 2016).

Vanuit de respondenten wordt aangegeven dat big onderwijsdata eigenlijk een instrument is naast alle andere instrumenten. Er wordt aangegeven dat dat beleidsmakers nu nog vaak op gevoel beslissingen nemen. Met behulp van big onderwijs data, kunnen ze dit in de toekomst meer gefundeerd doen. Maar ook eindgebruikers als docenten en individuele leerlingen kunnen veel baat hebben bij gevalideerde instrumenten die gebaseerd zijn op big onderwijs data.

Gerelateerd aan het beschikbaar komen van steeds meer data is de opkomst van formatief toetsen. Hierbij gaat het om het gebruiken van toetsen om informatie te verkrijgen over het verloop van het onderwijsleerproces en hieraan sturing te geven. Dit kan leiden tot betere leerresultaten van leerlingen (Bennett, 2011; Black & Wiliam, 2009). Bij formatief toetsen kent het woord toetsen een brede definitie, het gaat om het verzamelen van informatie over het leerproces van leerlingen en dit kan een docent bijvoorbeeld ook doen door het organiseren van discussies, door het geven van taken en het organiseren van activiteiten (Kippers, Schildkamp, & Poortman, 2016; Wiliam & Leahy, 2015). Dit levert veel bruikbare informatie op voor zowel docenten als leerlingen. Dit type informatie vinden we nu nog niet of minder terug in de verschillende datasystemen, maar dat is aan het veranderen doordat er steeds meer digitale tools beschikbaar komen voor formatief toetsen (Piety, 2013). Ferguson (2012) stelt bijvoorbeeld dat big data en learning analytics ervoor kunnen zorgen dat het leren van leerlingen steeds zichtbaarder wordt, door rapportages en visualisaties. Dit kan gebruikt worden om leerlingen niet alleen te beoordelen (summatief toetsen), maar juist ook om het leren van leerlingen te bevorderen (formatief toetsen).

Door scholen wordt expliciet aangegeven dat ze behoefte hebben aan tijdige gedetailleerde diagnostische informatie over hun leerlingen. Vanuit andere partijen wordt aangegeven dat zij het als een uitdaging zien om docenten te ondersteunen bij het interpreteren van deze feedback. Het analyseren van big onderwijs data kan al gebruikt worden voor signalering, maar dit kan nog veel meer voordelen opleveren voor de leraar. Bijvoorbeeld doordat het leerprogramma, met behulp van dit soort informatie, aangepast kan worden aan wat de individuele leerling nodig heeft.

Een andere mogelijke bevorderende factor is het beschikbaar komen van steeds meer geavanceerde datamining en analysetechnieken. Bij datamining gaat het om het proces van het halen van interessante, bruikbare en nieuwe informatie uit data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Gerelateerd aan het onderwijs gaat het om het ontwikkelen van methoden om onderwijsdata te exploreren met behulp van deze methoden, om zo leerlingen alsmede de context waarin zij het beste leren beter te kunnen begrijpen (Baker, 2010). Datamining is een manier om de kwaliteit van het onderwijs en de kwaliteit van het leerproces van leerlingen te verbeteren (Romero, Ventura, & De Bra, 2004). Romero en Ventura (2010) geven hierbij wel aan dat het belangrijk is dat dataminingtools gemakkelijker bruikbaar worden voor de onderwijspraktijk. De huidige tools zijn te complex. Zij geven als suggestie dat er mogelijk 'wizard tools' ontwikkeld kunnen worden door te gebruik maken van een default algoritme voor iedere taak en van parameter vrije datamining algoritmes, om zo de configuratie en uitvoering voor niet-experts gemakkelijker te maken.

In de interviews wijzen respondenten op de mogelijkheden die er nu al zijn voor benchmarking. Waar staat de school, de klas, de leraar of de leerling ten opzichte van een referentiepopulatie? Respondenten geven aan dat er nog veel winst te halen valt uit big data, mits deze gestructureerd is en met een vooropgesteld doel en met goede kwaliteit verzameld is.

Eén van de belangrijkste mogelijkheden of kansen m.b.t. big data is het uitvoeren van complexe analyses: hiermee kan inzicht verkregen worden in de kwaliteit van het onderwijs en het leren van leerlingen, die we kunnen gebruiken om het leren van leerlingen te ondersteunen. Door middel van learning analytics kunnen we bijvoorbeeld 'real time' analyses uitvoeren van leeractiviteiten. Door data te analyseren kan er voorspeld worden welke leerlingen risico lopen en kunnen direct interventies worden toegepast. Dit stelt docenten in staat om tijdig hun onderwijs aan te passen aan de behoeften van de leerlingen (Douglas, 2015).

In de interviews wordt gewezen op het kunnen identificeren van 'success stories'. Daarnaast worden de mogelijkheden genoemd die big onderwijs data bieden om meer detailinzicht te krijgen in onderwijsprocessen. De scholen noemen tenslotte de toegenomen regie van leerlingen over hun eigen leerproces. Met behulp van big data kunnen op de leerling toegesneden arrangementen en onderwijs-op-maat worden verzorgd.

Conclusie en discussie 8

In dit hoofdstuk vatten we de bevindingen uit de eerste hoofdstukken samen en formuleren we een aantal conclusies. We presenteren een aantal opvallende paradoxen rond big onderwijs data die we op basis van dit onderzoek geconstateerd hebben en eindigen met een serie aanbevelingen.

8.1 Welke data zijn beschikbaar voor datagedreven onderwijs-onderzoek?

Er is veel verschillende data beschikbaar. Een veelgebruikte indeling is het onderscheid in input data, procesdata en output data. Voorbeelden van beschikbare data die volgens de respondenten beschikbaar is betreft:

- Input data: onderwijsnummer, opleidingsniveau ouders, achtergrondkenmerken en inschrijfgegevens van leerlingen en studenten, woonplaats, etniciteit, vooropleiding, formatie, e.d.
- Procesdata: HR-data, financiële data, studieloopbaan, studievorm, studiekenmerken, stages, studiebegeleiders, mailwisselingen docenten en studenten (ook hoe vaak), verzuimfrequentie, learning analytics, bijvoorbeeld m.b.t. hoe lang een leerling is ingelogd, informatie van systemen zoals Blackboard, collegebezoek, docentkwaliteit, medewerkerstevredenheid, ziekteverzuim en dergelijke.
- Output data: (toets)data uit het leerlingvolgsysteem, methode toetsen, TIMMS data, data uit het cohortonderzoek, portfolio's, zachte data in de vorm van tekst, notities en observaties bijvoorbeeld over de sociaal-emotionele ontwikkeling van leerlingen, arbeidsmarktinformatie.

Er zit echter veel variatie in wat de geïnterviewden onder big data verstaan. Sommigen spreken bij een grote enkele dataset van big data, voor anderen moet er een combinatie van datasets plaatsvinden om van big data te spreken. Vaak wordt er van big data gesproken als het gaat om het zoeken naar onbekende verbanden, voor anderen kan het gericht zoeken naar het antwoord om een goed omschreven vraag ook big data onderzoek zijn. Ook over het gebruik van harde en/of zachte data verschillen de meningen. Gaat het ook om ongestructureerde, zachte gegevens of gaat het alleen over gestructureerde harde gegevens als cijfers en resultaten van onderwijs surveys? Het probleem met zachte onderwijsdata is dat ze lokaal verzameld worden zonder vastomlijnde architectuur. Het gevolg is dat deze data sterk verschillen tussen en zelfs binnen scholen. De value en veracity (waarde die aan deze data gehecht wordt en het vertrouwen dat erin wordt gesteld) van deze data zijn laag. Het analyseren van deze ongestructureerde zachte gegevens lijkt nog weinig meerwaarde voor onderwijsonderzoek of de onderwijspraktijk te bieden. Aan de andere kant bieden deze zachte gegevens mogelijk wel meerwaarde als het aankomt op contextualisering van harde data. Zoals sommige respondenten ook aangeven zou big onderwijsdata misschien moeten gaan over een combinatie van harde en zachte data, waarbij zachte data gebruikt kunnen worden om conclusies op basis van het koppelen van harde data in te kleuren en te verrijken.

8.2 Doelstellingen, gebruik, en de meerwaarde van big data

Het overkoepelende doel van big data kan omschreven worden als de interpretatie van verschillende operationele en administratieve data met als doel het toekomstig presteren te kunnen voorspellen en om problemen rondom programmering, onderzoek, instructie en leren te identificeren (Hrabowksi & Fritz, 2011 en Picciano, 2012 in Douglas, 2015). Echter, verschillende stakeholders hebben uiteraard eigen specifieke doelen als het gaat om het gebruik van big data in het onderwijs. Hierbij kan grofweg een onderscheid gemaakt worden in het gebruiken van big data om (1) te monitoren en meer inzicht te krijgen in bepaalde processen, inclusief het ontkrachten van mythes en assumpties, (2) te voorspellen (van leerprestaties, studiesucces, uitval etc.) en om (3) maatregelen te nemen om het onderwijs te verbeteren.

Dit onderscheid is ook relevant als het gaat om het gebruik van evaluatieresultaten. Weiss (1998) spreekt in dit geval bijvoorbeeld over het conceptueel gebruik van data (hier monitoren en voorspellen). In dit geval leidt de data (nog) niet direct tot concrete maatregelen, maar het beïnvloedt wel het denken over de onderwerpen waar data over verzameld is. Dit kan in de toekomst tot maatregelen leiden. Het daadwerkelijk nemen van maatregelen op basis van data is wat Weiss (1998) instrumenteel gebruik noemt. Dit laatste is lastiger te bereiken. Ten eerste moeten de resultaten goed te begrijpen zijn. Dit is vaak lastig als het gaat om big data en de complexe algoritmes die er gebruikt worden. Vervolgens moet men de vaardigheden, mogelijkheden en middelen hebben om deze nieuwe kennis om te zetten in concrete maatregelen. Als het gaat om de fase waarin we verkeren m.b.t. het gebruik van big data kunnen we concluderen dat we nog meer in de fase van het conceptueel gebruik zitten. Er wordt veel data verzameld en er worden nieuwe datamining tools en algoritmes ontwikkeld om deze te analyseren. Dit gaat de komende jaren mogelijk interessante nieuwe inzichten opleveren. De vraag is vervolgens hoe deze inzichten vertaald kunnen worden naar concrete maatregelen om het onderwijs te verbeteren.

De meerwaarde van big data zoals aangeven door de meerderheid van respondenten lijkt ook vooral te zitten in de nieuwe inzichten die big data kan opleveren. Door verschillende stakeholders werd bijvoorbeeld genoemd dat big data gebruikt kan worden om zaken als studiesucces, studievertraging en uitval te voorspellen en eerder problemen te detecteren/signaleren. Door big data kunnen scholen zichzelf bijvoorbeeld vergelijken met andere scholen (in Nederland, Europa, de wereld) en/of vergelijken ten opzichte van bepaalde benchmarks. Tevens kan het vragen beantwoorden zoals “wat is de toegevoegde waarde van de school?”. Ook werd genoemd dat het cohort onderzoek door slim te koppelen anders kan worden ingericht. Corhortonderzoek kan starten vanuit bestaande data en dit kan waar nodig aangevuld worden. In het koppelen van deze verschillende data zit ook de belangrijkste meerwaarde van big data. Tevens kan big data de bevragingslast m.b.t. onderwijsonderzoek verminderen voor scholen. Ook kan het voor scholen tijd besparen. Docenten moeten nu bijvoorbeeld vaak handmatig bepalen wat een individuele leerling precies nodig heeft. Het gebruik van big data kan dit proces versnellen en efficiënter maken. Dit reduceert de administratielast van docenten en geeft hen meer tijd voor andere zaken. Deze aspecten hebben allemaal te maken met het conceptueel gebruik van big data. Ook instrumenteel gebruik van big data werd door de respondenten benoemd: als belangrijkste meerwaarde werd benoemd dat big data gebruikt kan worden om het onderwijs te verbeteren. Echter, de meeste respondenten konden hier nog weinig echt concrete voorbeelden benoemen.

8.3 Technische, juridische en ethische aspecten van big data

De huidige Wet Bescherming Persoonsgegevens (WBP) speelt een belangrijke rol bij het gebruik van big data. Deze wet schrijft voor onder welke voorwaarden data die (met een redelijke inspanning) herleidbaar zijn tot een individu bewerkt mogen worden. Er is onder de niet-juridische geïnterviewden veel onduidelijkheid over wat wel en niet mag als het gaat om data verzamelen, delen en gebruiken. Ook worden bepaalde aspecten van de wet ervaren als belemmerend voor de ontwikkeling en het gebruik van big data in het onderwijs. Per mei 2018 zal de WBP vervangen worden door een Europa-brede wetgeving genaamd Algemene Verordening Gegevensbescherming (AVG). Deze neemt de ontwikkelingen omtrent dataverzameling en -verwerking in acht, en zal strengere bescherming bieden aan de privacy van het individu. Alhoewel een deel van de respondenten deze verscherping ziet als een verdere beperking van de mogelijkheden omtrent big data, dringen de geïnterviewde juristen erop aan dat deze gezien moet worden als een mogelijkheid om ethisch verantwoord met persoonsgegevens om te gaan. Deze tweedeling laat zich ook zien bij de aanbevelingen die de respondenten doen: een deel meent dat met het oog op de ontwikkelingen van big data er kritisch gekeken moet worden of de huidige wetgeving niet te beperkend is, terwijl een ander deel stelt dat met het groeien van de mogelijkheden omtrent big data analyse het tijd wordt om kritisch te kijken naar of we sommige onderzoeken en toepassingen wel moeten willen uitvoeren. In beide gevallen zal in de komende jaren de vraag ‘wat willen we als maatschappij bereiken met big data’ steeds relevanter worden.

Gerelateerd aan het bovenstaande, spelen ook privacy en eigenaarschap een belangrijke rol. Door privacyregels is het bijvoorbeeld lastig om personen in verschillende datasets aan elkaar te koppelen, hetgeen de mogelijkheden om nieuwe inzichten op te doen aan de hand van big data belemmert. De centrale vraag hier is hoe de privacy van het individu beschermd kan worden, zonder dat dit het gebruik van data voor het ontwikkelen van nieuwe inzichten en het verbeteren van het onderwijs in de weg staat. Een tweede belangrijk vraagstuk betreft eigenaarschap en verantwoordelijkheid. Hierover bestaat onder de respondenten nog veel onduidelijkheid. Van wie zijn welke data, wie is verantwoordelijk voor het beheer en de bescherming van deze data, met wie mogen data wel en niet gedeeld worden, welke data mogen wel en niet gekoppeld worden, op welke wijze en voor welke doeleinden? Hier is helderheid en transparantie gewenst. Eén van de mogelijke oplossingen voor een deel van de benoemde privacy- en eigenaarschap vraagstukken is het opzetten van een helder opt-in dan wel opt-out systeem. Hierbij kunnen individuen beslissen of hun data gebruikt mag worden voor bepaalde doelstellingen en door bepaalde partijen. Het is voor de transparantie van belang dat zowel het doel als de partijen die toegang hebben tot de data helder, volledig en concreet omschreven worden. Bovendien moet men altijd de mogelijkheid behouden om de toestemming voor datagebruik in te trekken.

Een ander aspect dat een belangrijke rol speelt bij big data is het voorkomen van het stigmatiseren en discrimineren van personen op basis van data. In het onderwijs betekent dit dat leerlingen niet al op jonge leeftijd een bepaalde stempel opgedrukt mogen krijgen, dat ze de rest van hun leven meedragen. Dit levert een zekere spanning op met het doel om leerlingen gepersonaliseerd onderwijs aan te kunnen bieden. Het wordt hier zaak om te voorkomen dat er een self-fulfilling prophecy wordt gecreëerd. Dit kan onder andere door er voor te zorgen dat de leerling te allen tijde weet op grond waarvan een beoordeling tot stand is gekomen, de mogelijkheid behoudt een advies aan te vechten, en dat de uitkomst van een big data algoritme uitsluitend tot een aanvulling en niet tot een beperking

van de onderwijsmogelijkheden mag leiden. Hieraan gerelateerd is '*the right to be forgotten*'; het recht om niet de rest van je leven achtervolgd te worden door data uit het verleden. Ook is het belangrijk om bij de interpretatie van de data rekening te houden met de persoon achter de data en de context waarin de data verzameld zijn. Een persoon is meer dan de data die over hem of haar verzameld kan worden. Eveneens moet voorkomen worden dat big data gebruikt wordt voor discriminatie en stigmatisering. Dit vereist een scherpe en kritische blik, want het is niet altijd eenvoudig te zien of bepaalde variabelen ruimte geven voor discriminatie. Tot slot speelt beveiliging van data een belangrijke rol. Beveiliging wordt steeds belangrijker, aangezien het gaat om steeds grotere databases die - zeker wanneer ze gecombineerd worden - steeds gevoeliger data bevatten. Tegelijkertijd worden databases steeds aantrekkelijker en waardevoller naarmate ze meer informatie bevatten, en wordt de beveiliging die nodig is om hackers buiten de deur te houden steeds complexer. Hier moet dan ook goed over worden nagedacht.

8.4 Centrale ontsluiting van big data

Het ontwikkelen van een centrale database met onderwijsdata is in veel van de interviews ter sprake gekomen. Respondenten zien hier voor- en nadelen. De eerste vraag is met welk doel een database ontwikkeld moet worden. Een andere vraag is of de huidige beschikbare data van voldoende kwaliteit zijn om zinvol te kunnen ontsluiten in een dergelijke database. Immers, data van lage kwaliteit leiden waarschijnlijk ook tot uitkomsten van twijfelachtige betrouwbaarheid. Ook het koppelen van data is lastig en hier worden vaak fouten bij gemaakt. Verder zijn niet alle organisaties bereid om hun data over hun eigen prestaties te delen en door derden te laten beheren. Een risico van een dergelijke database is bovendien dat men deze op een verkeerde wijze gaat gebruiken, bijvoorbeeld om personen of organisaties op af te rekenen, of om leerlingen op een negatieve wijze te profileren. Er wordt vaak gewezen op het gevaar dat er onvoldoende nagedacht wordt over de lange termijn consequenties van beslissingen die er genomen worden op basis van data. Tot slot moet een dergelijke database aan zoveel juridische en beveiligingsvoorwaarden voldoen, dat de vraag gesteld wordt of dit haalbaar is.

Aan de andere kant worden er ook voordelen genoemd van een dergelijke database. De overheid zou hier bijvoorbeeld op kunnen sturen en de kwaliteit van het onderwijs kan er mogelijk mee verbeterd worden. Het is belangrijk om na te denken over welke data er wel en niet in dit systeem terecht komen en wie er toegang krijgen tot de data om deze voor bepaalde doeleinden te gebruiken. Welke data er precies in een dergelijk systeem moeten komen is nog niet helder, maar aangegeven wordt dat het in ieder geval moet gaan om geaggregeerde data, niet terug te leiden tot personen. Aan de andere kant moet data in een dergelijk systeem juist wel persoonsgebonden zijn, want anders kunnen verschillende databronnen niet gekoppeld worden. Technisch gezien zijn er hier wel oplossingen voor. Zo wordt er gesteld dat data geanonimiseerd kunnen worden er dat er gewerkt kan worden met zogenoemde kleine 'Chinese walls', die verbindingen tussen bepaalde databases verbieden, want hoe meer gegevens, hoe makkelijker data te herleiden zijn naar individuen. In sommige gevallen (bv. het detecteren van mogelijke kindermishandeling) is het echter juist wel weer belangrijk om de data naar een individu te kunnen herleiden en dit is in die gevallen dan ook weer mogelijk.

Als een centrale database opgezet gaat worden wordt toezicht hierop een heel belangrijk onderwerp. Dit zou op nationaal niveau geregeld kunnen worden, door bijvoorbeeld instanties zoals het CBS en de overheid. Echter, de vraag wordt opgeworpen of de overheid wel de juiste instantie hiervoor is en

hiervoor de benodigde capaciteit in huis heeft. Een andere optie is dat scholen (schoolbesturen) de eigenaar worden van deze database, maar ook scholen beschikken vaak weer niet over de benodigde capaciteiten. Er zijn dus veel vragen rondom het eigenaarschap en toezicht van een dergelijke database. Men is het er wel over eens dat het essentieel is dat toezicht goed en wettelijk geregeld wordt.

Tot slot zijn door de respondenten waardevolle suggesties gedaan. Er is gewezen op 'privacy by design'. Daarnaast is over de encryptie van gegevens opgemerkt dat de key bij de leerlingen/studenten zou moeten liggen, maar dat er tegenwoordig technieken zijn om inhoudelijke analyses te doen over ge-encrypte data. Meerdere respondenten wezen op het belang van een trusted third party die de data zou moeten koppelen en op het belang van een autoriteit die de ontsluiting zou moeten regelen. Hierbij werd geregeld gewezen op de rol die scholen hierbij zouden moeten vervullen als primaire leverancier van de data, als als gebruiker van de uitkomsten van de analyses om het onderwijs te verbeteren.

8.5 Belemmerende factoren en risico versus bevorderende factoren en kansen

Wat betreft de *maatschappelijke aspecten* van big data zien we dat het belangrijk is om de sociale, juridische en ethische implicaties van big data niet uit het oog te verliezen. Vragen rondom eigenaarschap van de data, privacy en veiligheid spelen hierbij een belangrijke rol. Big data is niet per definitie waardevrij. Al bij het besluiten van welke data er op welke manieren verzameld worden, worden er bepaalde keuzes gemaakt, d.w.z. menselijke beslissingen, die niet volledig objectief of waardevrij zijn.

Waar het de mogelijkheden en kansen voor de maatschappij in het algemeen betreft zien we dat het belangrijk is om één van de belangrijkste doelen van big onderwijsdata niet uit het oog te verliezen: het verbeteren van de kwaliteit van het onderwijs en het leren van leerlingen.

Rondom de *kenmerken van big data* zien we dat er nog veel problemen opgelost moeten worden m.b.t. de beschikbaarheid van data, de kwaliteit van de data en de mogelijke bias in data (bij gestructureerde data, maar zeker ook bij ongestructureerde data), infrastructuur (bijvoorbeeld de toegankelijkheid van data en het kunnen koppelen van data) en de kosten die dit met zich meebrengt. Tegelijkertijd zien we ook dat de ontwikkelingen snel gaan: er komt steeds meer data beschikbaar, bijvoorbeeld data afkomstig van social media en allerlei online leeromgevingen. Ook de trend op het gebied van formatief toetsen zorgt er voor dat er steeds meer real time data over het leren van leerlingen beschikbaar komt en ook worden er steeds meer nieuwe en geavanceerde datamining en analysetechnieken ontwikkeld. Tot slot wordt er steeds meer geïnvesteerd in de digitale infrastructuur op scholen, zodat data op de juiste manier verzameld, opgeslagen en gebruikt kunnen worden. Dit vraagt in de toekomst echter wel om meer investeringen.

De volgende risico's die we signaleren hebben te maken met de *gebruikers van big data.* Eén van de belangrijkste risico's of belemmerende factoren betreft een capaciteitsprobleem. In alle lagen van het systeem, of het nu gaat om beleidsmedewerkers, onderzoekers, managers, docenten of leerlingen, signaleren we een groot gebrek aan capaciteit. Hierbij gaat het om de capaciteit en competenties voor het verzamelen van data, het analyseren van data, het interpreteren van data en het nemen van beslissingen op basis van deze data om het onderwijs te verbeteren. Vaak zijn big data modellen en

de gebruikte algoritmes voor de meeste betrokkenen niet inzichtelijk. Gerelateerd hieraan ligt er ook een kans m.b.t. het investeren in data science (opleidingen). Het gat tussen de groei in aanbod en de mogelijkheden van data enerzijds en anderzijds de expertise en capaciteiten van big data gebruikers wordt anders te groot. Het is in het algemeen belangrijk om te investeren in trainingen op het gebied van (big) data gebruik voor het onderwijs. Ook is het belangrijk dat scholen en docenten beseffen dat het gebruiken van data belangrijk is, bij het dagelijks werk hoort en ook niet meer weggaat. Docent-oordelen worden nu soms gevormd zonder naar de data te kijken, of zij gebruiken alleen data die bij hun opinie aansluiten. Dit is uiteraard niet voor alle docenten en scholen zo, maar hier is nog wel verbetering nodig. Tot slot kent het gebruik van big data zo veel verschillende facetten dat het raadzaam is om zowel binnen scholen, tussen scholen alsmede tussen scholen en andere partners meer samen te werken en expertise met elkaar te delen.

We signaleren ook verschillende kansen als het gaat om de gebruikers van big data. Docenten zijn bijvoorbeeld vaak heel kritisch, ze willen weten waar aanbevelingen op gebaseerd zijn. Deze kritische houding is essentieel, aangezien het kan gaan om de toekomst van leerlingen en het belangrijk is dat docenten weten waar aanbevelingen op gebaseerd zijn. Ook kunnen opleidingen meer aandacht besteden aan big data, zodat studenten hier al snel mee in aanraking komen. Hierbij gaat het om het investeren in scholing m.b.t. het verzamelen, analyseren en gebruiken van data, inclusief alle risico's die big data met zich meebrengt. Tot slot biedt teamwork veel kansen. Eén individu hoeft niet alles te kunnen, dus samenwerking tussen docenten wordt nog belangrijker. Ook samenwerkingsverbanden tussen scholen en universiteiten of andere externe partijen kunnen bevorderend werken. Universiteiten en andere externe partners kunnen scholen bijvoorbeeld ondersteunen bij het koppelen, analyseren en begrijpen van data.

Ook op het niveau van de *dataleveranciers* is er een aantal risico's te benoemen. Vanuit de literatuur wordt bijvoorbeeld gewaarschuwd voor de mogelijkheid dat commerciële bedrijven verzamelde data doorverkopen. Dit wordt ook door geïnterviewde scholen gezien als een risico. Een ander risico betreft dat het gemakkelijk is om data te manipuleren in het kader van commerciële belangen. Ook wordt een aantal bevorderende factoren en kansen gesignaleerd. Dataleveranciers kunnen er in de toekomst voor zorgen dat ze transparant(er) zijn over de gebruikte algoritmes en modellen, zodat gebruikers van big data begrijpen wat daar aan ten grondslag ligt en waar de beslissingen op gebaseerd zijn. Dataleveranciers kunnen ook klanten ondersteunen en begeleiden in het interpreteren van de data, dit gebeurt ook steeds meer. Ook zijn er ontwikkelingen gaande rondom een aantal technische mogelijkheden, zoals de ontwikkeling van benchmarks, koppelingen die mogelijk worden gemaakt tussen gestructureerde (cijfers) en ongestructureerde data (observaties, notities), digitale datakluizen (iedereen beheerder van eigen data) en dashboards voor scholen en docenten die betere signaleringsmogelijkheden bieden.

Op het niveau van de *onderzoekers* zijn eveneens risico's en kansen te benoemen. Het belangrijkste risico, te weinig samenwerking, kan omgezet worden in een kans: meer multidisciplinair onderzoek. Sommige data hebben betrekking op verschillende niveaus en zijn relevant voor verschillende organisaties. Big data vraagt bijvoorbeeld om technische kennis, onderwijskundige kennis, vakinhoudelijke kennis, organisatorische kennis en ontwerpkennis (Piety, 2013). Als we big data willen gebruiken om het onderwijs te verbeteren, dan vraagt dit om veel kennis op het gebied van leren: hoe leren leerlingen en hoe kan het leren ondersteund worden? Er moet een link gelegd worden tussen onderzoek op het gebied van leren en onderzoek op het gebied van data (Ferguson, 2012).

Big data heeft ook implicaties voor het doen van wetenschappelijk onderzoek. Vragen over wat kennis inhoudt, hoe we onderzoek moeten doen, hoe we om moeten gaan met informatie en hoe we de werkelijkheid kunnen vatten in onderzoek krijgen mogelijk een andere invulling door big data (Boyd & Crawford, 2012). De focus kan bijvoorbeeld veel meer te liggen komen op wat individuen doen, dit kunnen we steeds beter meten in tegenstelling tot het meten van waarom ze dingen doen (Boyd & Crawford, 2012). Ook de manier van onderzoek doen verandert mogelijk door de komst van big data. Onderzoek begint meestal met een theorie die onderzocht moet worden en daar wordt data bij verzameld. Bij big data is het mogelijk om te beginnen bij de data en op basis hiervan nieuwe inzichten te ontwikkelen. Een belemmering hiervoor is echter wel dat bepaalde onderdelen uit de wet (voor de verwerking van persoonsgegevens is het belangrijk om doel en noodzaak te kunnen onderbouwen) dit niet toestaan.

8.6 Big data paradoxen en suggesties voor praktijk en vervolgonderzoek

Op basis van dit onderzoek zijn we op een aantal big data paradoxen gestuit. Deze worden hieronder beschreven. De verschillende paradoxen leiden tevens tot een aantal suggesties voor de praktijk en voor vervolgonderzoek.

Privacy paradox

Vanuit juridisch oogpunt wordt er steeds meer ingezet op privacy beschermende maatregelen. Ook de toegenomen aandacht voor privacy by design speelt daarbij een belangrijke rol. Aan de andere kant wordt er zowel door databeheerders als door technici op gewezen dat het samenvoegen van data er toe leidt dat individuen herleid kunnen worden, zelfs als anonimisering of pseudonimisering wordt toegepast. Dit laatste aspect is bij veel gebruikers nog niet bekend. Op het gebied van privacy is dus verder onderzoek nodig. Hoe kan een database zo worden ingericht dat de privacy van individuen voldoende beschermd wordt, maar dat het toch mogelijk is om verschillende datasets aan elkaar te koppelen, zodat dit kan leiden tot nieuwe inzichten. Een eerste technologische ontwikkeling die hierin bijvoorbeeld meegenomen kan worden in het gebruik van de kleine Chinese walls, zoals hierboven beschreven, en het gebruik van sleutels en encrypties. Verder raden we aan dat er goed gekeken wordt naar alle aspecten van de wet in het kader van de ontwikkelingen op het terrein van big data. In de nieuwe wet, Algemene Verordening Gegevensbescherming, staat bijvoorbeeld dat het verwerken van grote hoeveelheden persoonsgegevens, zoals bij big data, zonder voorafgaand duidelijk doel en zonder toestemming van betrokkenen niet is toegestaan. Aan de andere kant mogen er volgens deze zelfde wet wel data gebruikt worden door wetenschappers voor statistische doeleinden, dus dit biedt wel wat ruimte. De vraag is welke consequenties de wetgeving precies heeft voor het gebruik van big data. Wat zijn de juridische mogelijkheden en onmogelijkheden en hoe kan hier duidelijkheid over geschapen worden?

Positioneringsparadox

Onderwijsdata kunnen samengevoegd worden in een centrale database, die bijvoorbeeld beheerd wordt door het CBS. De vraag is hoe een dergelijke database gepositioneerd moet worden. Als een dergelijke database gepositioneerd wordt als de nationale onderwijs database, dan spreekt daar een bepaald ambitieniveau uit. Als deze database benoemd wordt als een veilige centrale omgeving voor

onderwijsdata, dat wordt de database op een totaal andere manier gepositioneerd. De term ‘nationaal’ suggereert vooral dat de database alle Nederlandse onderwijsdata bevat. De term ‘veilige omgeving’ appelleert veel meer aan vertrouwen. Op het moment dat er voor gekozen wordt om onderwijsdata centraal te ontsluiten, zal daarbij in de communicatie goed gekeken moeten worden naar termen die recht doen aan het ambitieniveau met betrekking tot onderwijs en tegelijkertijd vertrouwen wekken bij de scholen, ouders en leerlingen.

Clusteringsparadox

Het samenvoegen van onderwijs data in één database, oftewel het clusteren van onderwijsdata, biedt veel voordelen op het gebied van onderzoek. Data wordt toegankelijker en er kan beter toezicht worden gehouden. Ook ten aanzien van security biedt clustering voordelen. Er kan niet alleen gebruik gemaakt worden van de nieuwste technieken, maar er kunnen ook fysieke beveiligingsmaatregelen worden genomen.

Het samenbrengen van data brengt aan de andere kant risico’s met zich mee. Bij het koppelen van gegevens door algoritmes worden nog steeds fouten gemaakt, volgens technici. Daarnaast brengt een centrale database ook meer risico met zich mee wat betreft datalekken, niet alleen vanwege de grotere massa en daarmee vanwege waarde van de data, maar ook vanwege het grote aantal betrokken partijen. Er moet dus onderzocht worden hoe een centrale database op de juiste wijze beschermd kan worden. Hiervoor komen er wel steeds meer mogelijkheden, bijvoorbeeld door te gaan werken met (biometrische) sleutels, deze mogelijkheden dienen verder verkend te worden.

Context/individu/data paradox

Het ontkoppelen van de data van zijn specifieke context biedt veel voordelen. De privacy wordt meer beschermd en data kunnen geaggregeerd worden op hogere niveaus wat voordelen biedt bij de analyse en bij het ontwikkelen van modellen. Door respondenten die data genereren wordt er aan de andere kant juist op gewezen dat de context nodig is voor een juiste interpretatie van de data. De vraag hier is dus hoe data op een zodanige wijze gekoppeld kunnen worden dat ze toch betekenisvol blijven voor de verschillende gebruikers. Mogelijk moet er nagedacht worden over verschillende soorten koppelingen en aggregatieniveaus en misschien zelfs wel over verschillende databases. Ontwikkelingen op het gebied van de hierboven genoemde datakluizen kunnen in dit verband ook meegenomen worden. Wat hierbij ook een rol speelt is dat personen nooit volledig in data gevat kunnen worden. In het onderwijs, speelt bijvoorbeeld de relatie tussen de docent en de leerling ook een belangrijke rol. Dit is moeilijk in data te vatten. Onderwijs bevat een grote menselijke factor. De vraag is dus hoe beslissingen in het onderwijs beter, efficiënter en effectiever genomen kunnen worden, waarbij (big) data één van de tools is die ingezet kunnen worden bij het nemen van deze beslissingen zonder daarbij de context uit het oog te verliezen.

Paradox rond de rol van de overheid

Meerdere respondenten wijzen op het belang van de rol van de overheid rond wet- en regelgeving. Ook wordt de overheid (en instanties zoals CBS) gezien als onpartijdig en wordt het regelen van toegang tot en ontsluiting van data gezien als een taak van de overheid. Aan de andere kant is de overheid verantwoordelijk voor het onderwijsbeleid en wordt de overheid daarom door veel scholen gezien als een belanghebbende partij. Er wordt getwijfeld of het de rol van diezelfde overheid zou

moeten zijn om een centrale onderwijs database op te zetten. Enkele respondenten herinneren zich incidenten waarbij de overheid data heeft gebruikt voor een ander doel dan was toegezegd bij het verzamelen van de data. Zij zijn dientengevolge uiterst kritisch over hoe veilig data is bij een overheidsinstantie. Er moet verder onderzocht worden of, hoe en door wie een dergelijke database opgezet zou moeten worden en hoe het toezicht geregeld zou moeten worden.

Give and take paradox

Veel respondenten willen graag data die hun eigen vragen kunnen beantwoorden, maar zijn aarzelend om hun eigen data ter beschikking te stellen voor het beantwoorden van vragen van anderen. Een belangrijke vraag is of er data zijn die in het algemeen beschikbaar moeten komen (zoals voor de belastingdienst ook data aangeleverd moeten worden). Zo ja, welke data zijn dit dan en voor welke data kunnen personen zelf bepalen of ze die willen delen of niet?

Technologische voortgang en capaciteit paradox

De ontwikkelingen op het gebied van het verzamelen en analyseren van data gaan enorm snel en leiden tot steeds meer mogelijkheden. Er komen steeds meer platformen en tools beschikbaar. Echter, de capaciteit van de gebruikers lijkt hierbij achter te blijven. Er is te weinig kennis op het gebied van het analyseren en gebruiken van big data. Slecht een beperkt aantal personen heeft hier kennis over en dit zijn vaak niet de personen die de data uiteindelijk moeten gaan gebruiken om het onderwijs te verbeteren. We raden dus aan om te investeren in de capaciteit van personen. De vraag is welke capaciteiten nodig zijn voor het gebruik van big data voor onderwijsverbetering: welke stakeholders hebben welke kennis en vaardigheden nodig als het gaat om het gebruik van big data? Kan een professionaliseringsaanbod en opleidingsaanbod ontwikkeld worden voor de verschillende stakeholders? Ook kan er over nagedacht worden over de vraag of het verstandig is om data en datagebruik al in het curriculum van middelbare scholen en universiteiten op te nemen. Aangezien het bij big data vaak om de toekomst van leerlingen gaat zouden leerlingen de big data (analyses en algoritmes) misschien moeten kunnen begrijpen, om te kunnen oordelen of de beslissingen die op basis hiervan genomen zijn eerlijk en juist zijn.

Tot slot, deze verkennende studie heeft veel inzichten opgeleverd als het gaat om de (on)mogelijkheden van big data in het onderwijs. Hierboven staan een aantal concrete suggesties voor vervolgonderzoek en de praktijk genoemd als het gaat om verschillende technische, juridische en ethische aspecten van big data en om capaciteitsontwikkeling met betrekking tot big data. Daarnaast raden we aan dat de belemmerende en bevorderende factoren voor het gebruik van big data voor het onderwijs verder onderzocht worden. De resultaten van deze studie kunnen hiervoor als basis gebruikt worden. Er lijken veel mogelijkheden en kansen te zijn voor het gebruiken van big data voor onderwijsverbetering. Big data heeft zeker potentie voor onderwijsonderzoek in Nederland, maar de visies op wat kan, mag, en nodig zou zijn verschillen sterk. Big data vraagt dus eerst nog om een investering in zowel onderzoek als in de capaciteit van mensen.

Aanbevelingen 9

Op basis van ons rapport hebben we een aantal aanbevelingen geformuleerd m.b.t. toekomstige scenario's voor big data gedreven onderwijsonderzoek. Deze aanbevelingen hebben betrekking op: visie en doelen; juridische en ethische aspecten; technische aspecten van big data; opleiding en professionalisering; en op het onderzoek.

Visie en doelen: Om optimaal te kunnen profiteren van de mogelijkheden van big data is het van belang dat er eerst een gezamenlijke visie wordt ontwikkeld en dat de doelen helder worden (wat kan er wel en wat kan/mag niet met big data). Dit leidt tot de volgende aanbevelingen.

1. Op verschillende niveaus moet er nagedacht worden over big data: Welke visie hebben stakeholders als de VO Raad, PO raad, mbo raad, universiteiten en hogescholen, schoolbesturen en scholen op big data (gemeen)? Het ministerie van OC&W neemt het voortouw bij het formuleren van een gezamenlijke visie. Een duidelijke positiebepaling en voorlichting is nodig om er voor te zorgen dat de (brede) maatschappelijke discussie gebaseerd is op een gedeelde visie.

2. Op basis van de geformuleerde visie kunnen de verschillende stakeholders nadenken over de doelen die ze willen behalen en is het van belang om te expliciteren dat resultaten niet gebruikt kunnen gaan worden voor andere doelen dan de oorspronkelijke.

Juridische en ethische aspecten: Er komen steeds meer data beschikbaar. Dit zorgt ervoor dat juridische en ethische spelregels steeds belangrijker worden. De nieuwe wetgeving die eraan komt op dit gebied voorziet in een duidelijke behoefte, maar er moet ook kritisch gekeken worden of bepaalde wet- en regelgeving het gebruik van big data niet belemmert. Op basis van dit onderzoek komen we tot de volgende aanbevelingen op juridisch en ethisch gebied:

3. Verschillende doelgroepen hebben verschillende soorten vragen die met verschillende soorten big data beantwoord kunnen worden. Bij het vaststellen van de juridische en ethische spelregels zal daarom een gedifferentieerde aanpak moeten worden gebruikt. Verschillende soorten data brengen bijvoorbeeld verschillende risico's met zich mee en het gebruik van data door scholen vraagt om andere spelregels dan het gebruik van data door onderzoekers.
4. Er moet worden vastgesteld van wie verschillende data zijn om het proces van toestemming van datagebruik eenduidig vast te kunnen leggen.
5. Om leveranciers van data vertrouwen te geven en bereid te maken om data af te staan moet duidelijk worden gemaakt hoe data beveiligd zijn opgeslagen.
6. Om afspraken tussen scholen en databeheerders/bewerkers te stroomlijnen moeten er duidelijke standaard bewerkersovereenkomsten beschikbaar worden gesteld met uitleg over de verschillende aspecten zodat scholen zich kunnen vergewissen van de strekking en correctheid van een af te sluiten overeenkomst.
7. Voor het garanderen van de anonimiteit van data wordt het inschakelen van een zogenaamde 'trusted third party' aanbevolen.

8. Het is het overwegen waard een onafhankelijke instantie (toezichthouder) op te richten die uitspraken kan doen in gevallen van twijfel over gemaakte afwegingen van belangen of beslissingen over proportionaliteit, privacy etc.
9. Het is het waard te exploreren of er een instantie kan komen die data centraal beheert. Dit zou aspecten als kwaliteit van data, beveiliging en privacybescherming beter beheersbaar maken. Bij zo'n exploratie zou een geleidelijke invoering te prefereren zijn.
10. Voor juridische aspecten met betrekking tot het omgaan met onderwijsdata wordt het aangeraden om gebruik te maken van de expertise die inmiddels ontwikkeld is binnen het Nationaal Cohort Onderzoek (NCO).
11. Aanbevolen wordt om het 'recht om vergeten te worden' te implementeren in onderwijsonderzoek. Op het moment dat een leerling of student een onderwijsinstelling verlaat, blijft de data slechts gepseudonimiseerd of geaggregeerd beschikbaar voor onderzoek.

Technische aspecten: Om optimaal gebruik te kunnen maken van big data in het onderwijs spelen verschillende technische aspecten een rol. Op basis van dit onderzoek hebben we de volgende aanbevelingen geformuleerd:

12. Bij elk big data onderzoek en bij het publiceren van de uitkomsten hiervan zou de kwaliteit van de oorspronkelijke data (accuraatheid, consistentie, representativiteit) beoordeeld en vermeld moeten worden.

13. Bij het presenteren van conclusies uit big data onderzoek moet er, naast informatie over de kwaliteit van de data, ook inzicht gegeven worden in de gebruikte data analyse technieken.

14. Voor een juiste interpretatie van data of de resultaten van big data analyses moet de context waarin de data verzameld zijn, bekend zijn.

15. Voor het gestandaardiseerd op kunnen slaan van data moet een duidelijk protocol voor het aanleveren van de data opgesteld worden.
16. Vanwege de complexheid van de materie is het in ieder geval verstandig om op kleine schaal te beginnen met projecten voor het delen van data voor onderzoek.

Opleiding en professionalisering: de ontwikkelingen op het gebied van big data gaan op dit moment sneller dan de capaciteitsontwikkeling en expertise van de mensen die big data zouden kunnen gebruiken. We bevelen daarom investering in opleiding en professionele ontwikkeling van mensen aan:

17. Er zou een initiatief moeten komen om meer experts op het gebied van big data analyse ('data science') op te leiden. Dit betreft niet alleen technisch experts maar ook experts die in staat zijn de praktijk transparant voor te lichten en te trainen in het interpreteren en duiden van big data analyses.

18. Het verdient aanbeveling tools te ontwikkelen die (beperkte) big data analyse door (getrainde) personen uit de praktijk mogelijk maken. Hierdoor komen analyses dichterbij de praktijk te staan en kunnen de data in context worden bekeken en geïnterpreteerd.
19. Om vertrouwen te winnen van schoolleiders, schoolbesturen en docenten en de meerwaarde van big data en datagedreven onderzoek moeten best practises met hen gedeeld worden.

20. Er is behoefte aan professionele ontwikkeling op het gebied van big data op het niveau van schoolbesturen en schoolleiders. Het is aan te bevelen om trainingen voor deze doelgroep te ontwikkelen.

Wetenschappelijk onderzoek: in de toekomst is er behoefte aan meer onderzoek op het gebied van big data m.b.t. bijvoorbeeld juridische en ethische aspecten, technische aspecten en opleiding en professionalisering. Op basis van dit onderzoek hebben we een aantal aanbevelingen geformuleerd voor verder onderzoek:

21. Dit onderzoek is een start, maar er is meer onderzoek nodig naar in welke mate en op welke manier er op dit moment al gebruik gemaakt wordt van big data in het onderwijs. Bijvoorbeeld, hoe maken hogescholen en universiteiten gebruik van big data? Hoe maken schoolbesturen hier gebruik van? Hoe maken schoolleiders hier gebruik van? Zijn er voorbeelden van best-practices te vinden? Waar is het misgegaan en waarom?

22. Op het gebied van technologie is er steeds meer mogelijk, maar moet er nog veel onderzocht worden. Welke mogelijkheden zijn er bijvoorbeeld om ongestructureerde data te analyseren, zodat deze bruikbaar worden voor onderwijsverbetering?
23. Tot slot wordt aanbevolen om extra onderzoek te doen naar de beste manier om een training voor schoolbesturen en schoolleiders te ontwikkelen (zie ook aanbeveling 20), die vervolgens geïmplementeerd en geëvalueerd zou moeten worden. Leidt deze training uiteindelijk tot de gewenste capaciteiten voor het gebruiken van big data voor onderwijsverbetering?

Literatuurverwijzingen 10

Baker, R (2010). *Data mining for education.* In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Boyd, D., & Crawford, K. (2012). Critical questions for big Data. Provocations for a cultural, technological, and scholarly phenomenon, information. *Communication & Society, 15*, 662–

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM, 36*, 66-77.

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46*, 904-920.

Dede, C. J. (2016). Next steps for "Big Data" in education: Utilizing data-intensive research. *Educational Technology LVI (2)*, 37-42.

de Jong, T., Sotiriou, S., & Gillet, D. (2014). Innovations in STEM education: The Go-Lab federation of online labs. *Smart Learning Environments*, *1*, 3.

Liñán, L. C., & Pérez, Á. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education, 12*, 98-112.

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (in press). The effects of a data use intervention on educators' satisfaction and data literacy. E*ducational Assessment, Evaluation and Accountability.*

Eynon, R. (2013). The rise of big data: what does it mean for education, technology, and media research? *Learning, Media and Technology, 38*, 237-240.

Fayyad, U. Piatetsky-shapiro G., Smyth. P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, 17*, 37-54.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*, 304-317.

Frederiksen, J., & Collins, A. (1998). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Franzke, A. S. (2016). *Big Data Ethicist-What will the role of the ethicist be in advising governments in the field of big data?* (Master's thesis, Utrecht University).

Gibson, D. C., & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies, 20*, 697-713.

International Educational Data Mining Society, 2017. http://www.educationaldatamining.org/

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. Ec*onomics of Education review, 27*, 615-631.

Kennisnet, 2017. https://www.kennisnet.nl/artikel/learning-analytics-wat-betekent-dat-eigenlijk/

Keuning, T. & van Geel, M. (2016). *Implementation and effects of a schoolwide data-based decision making intervention*: *a large-scale study.* Dissertation. Enschede: University of Twente.

Kippers, W. B., Schildkamp, K., & Poortman, C. L. (2016, April). *The use of formative assessment by teachers in secondary education in the Netherlands*. Artikel gepresenteerd op de AERA conferentie, 10 april, Washington D.C., USA.

Landon-Murray, M. (2016). Big Data and Intelligence: Applications, Human Capital, and Education. *Journal of Strategic Security, 9*, 92-121.

Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Variety and Velocity. *META Group Inc*, *949*, 1-4.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2010, December 21). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Piety, P. J. (2013). *Assessing the educational data movement*. New York: Teachers College Press.

Romero, C., Ventura, S., De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. User Modeling and User-Adapted Interaction. *The Journal of Personalization Research 14*, 425–464.

Romero, C., Ventura, S. (2007). Educational Data Mining: a Survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135-146

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40*, 601–625.

van der Scheer, E.A. (2016). *Data-based decision making put to the test.* Dissertation. Enschede: University of Twente.

Schildkamp, K. Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation, 42*, 15-24.

Schildkamp, K., & Poortman, C.L. (2015). Factors influencing the functioning of data teams. *Teachers College Record*, *117*, 1-42.

Schildkamp, K., & Kuiper, W (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*, 482-496.

SURFnet/Kennisnet Innovatieprogramma. (2011). Verkenning Adaptieve Leersystemen. Zoetermeer: Kennisnet.

Thompson, G., & Cook, I. (2016). The logic of data-sense: Thinking through Learning Personalisation. *Discourse: Studies in the Cultural Politics of Education*, 1-15.

Weber, R. H. (2011). The right to be forgotten. *More than a Pandora's Box*, *2*.

Weiss, C.H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*, 21–33.

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of Education Policy*, *31*, 123-141.

Willenborg, L., & Heerschap, N. (2010). *Koppelen.* Centraal Bureau voor de Statistiek: Den Haag.

Appendices

Appendix 1: Lijst met interviews

Advocaat, IE en ICT recht - *Corianne Netze-Ritsema*

Autoriteit Persoonsgegevens - *3 senior onderzoekers*

CBS - *Barteld Braaksma en Ronald de Jong*

Cito - *Anton Béguin en Jos Keuning*

Dedact - *Bas Vonk*

DUO - *Mark de Boer*

Kennisnet, Jurist en adviseur privacy - *Job Vos*

KONOT (Katholiek Onderwijs Noord-Oost Twente) - *Leonie Wenting*

Oefenweb - *Marthe Straatemeier*

Onderwijsinspectie - *Bert Bulder*

OpenState Foundation - *Lex Slaghuis*

PO-raad - *Maurits Huigsloot*

Rijksuniversiteit Groningen, Juridische aspecten datamanagement en e-learning - *Esther Hoorn*

Rijksuniversiteit Groningen, Onderwijsinnovatie en strategie - *Hans Beldhuis*

Rijksuniversiteit Groningen, Onderzoek en evaluatie van onderwijseffectiviteit - *Roel Bosker*

SchoolPoort - *Jeroen Schutz*

Snappet - *Martijn Allessie*

Stichting Beroep en Bedrijf - *Ruud Baarda*

Stichting Carmel college - *Fridse Mobach en Tom Morskieft*

St. OnderwijsTTP - *Hans van Vlaanderen, Michiel Vlastuin en Sylvia Peters (Univesiteit Utrecht)*

SURFnet - *Christien Bok en Jocelyn Manderveld*

SURFsara - *Axel Berg en Machiel Jansen*

Swiebel Advies, big data om schooluitval te voorspellen - *Willem-Jan Swiebel*

The Implementation Group - *Ernst-Jan Horn*

Topicus - *Bart Broekhuis, Barthold Derlagen en Thomas Markus*

Universiteit van Amsterdam/Oefenweb - *Han van der Maas*

Universiteit Leiden, Center for Law and Digital Technologies - *Bart Custers en Helena Ursic*

Universiteit Maastricht, Nationaal Cohort Onderzoek - *Rolf van der Velden*

Universiteit Maastricht, Onderwijseconomie - *Lex Borghans*

Universiteit Utrecht, Onderwijsadvies en training - *Renske de Kleijn en Jan van Tartwijk*

Universiteit Twente, Computerscience - *Maurice van Keulen*

Universiteit Twente, Databeheer - *Marc Zeeman*

Universiteit Twente, Data security - *Andreas Peter*

Universiteit Twente, Ethiek in big data - *David Douglas*

Utrecht Data School - *Aline Franzke en Iris Muis*

VO-raad - *Anne Goris*

Appendix 2: Personalia

Bernard Veldkamp is hoogleraar research methodology and data analytics aan de Universiteit Twente. Hij is voorzitter van de afdeling Onderzoeksmethodologie, Meetmethoden en Data-Analyse en wetenschappelijk directeur van het Research Center voor Examinering en Certificering. Hij verricht onderzoek op het gebied van onderwijskundig meten, big data en datamining. Daarbij onderzoekt hij logfiles om fraude bij digitale examens op te sporen. Hij analyseert weblogs en posts op social media om patiënten te screenen op psychische aandoeningen en hij combineert data uit verschillende bronnen (databases, logfiles en toetsen) om het leren van studenten in digitale leeromgevingen te meten. Zijn onderzoek is gebaseerd op technieken uit de Psychometrie, Operations Research, Data Science en Statistiek. Bernard is fellow van AEA-Europe. Hij heeft meer dan 100 artikelen gepubliceerd in tijdschriften en boeken en hij is editor van de boekenserie Methodology of Educational Measurement and Assessment.

Kim Schildkamp is universitair hoofddocent bij de lerarenopleiding ELAN van de Universiteit Twente. Haar onderzoek richt zich op datagebruik en formatief toetsen. Hierover heeft ze veel gepubliceerd. Ze heeft verschillende prijzen gewonnen voor haar werk, waaronder voor haar werk rondom de datateam® methode. Deze interventie heeft Kim Schildkamp ontwikkeld en uitgebreid onderzocht. Het boek hierover, getiteld "De datateam® methode: Een concrete aanpak voor onderwijsverbetering", is verschenen in het Nederlands en Zweeds en komt volgend jaar ook in het Engels uit. Kim Schildkamp is de oprichter en de voorzitter van het ICSEI (International Congress on School Effectiveness and Improvement) data use netwerk. Tevens is ze de president-elect van ICSEI en in 2019 wordt ze de president van deze internationale organisatie.

Merel Keijsers is junior-onderzoeker aan de Universiteit Twente. Zij rondde een studie psychologie en een dubbele onderzoeksmaster (in sociale- en gezondheidspsychologie, en methoden en statistiek) af aan de universiteit Utrecht. Bij een stage aan de academische werkplaats van de GGD Hollands Noorden deed ze ervaring op met het opzetten en uitvoeren van (kwalitatief) onderzoek, en voltooide ze twee studies in opdracht van de gemeenten.

Adrie Visscher is hoogleraar docentprofessionalisering aan de Universiteit Twente (ELAN, de vakgroep Docentontwikkeling). Daarnaast is hij bijzonder hoogleraar data-based decision making aan de Universiteit Groningen (vakgroep Onderwijskunde). Hij doet onderzoek naar hoe leerkrachten en scholen met feedback ondersteund kunnen worden bij het optimaliseren van hun onderwijskwaliteit en hun impact op het leren van hun leerlingen. De feedback kan bijvoorbeeld betrekking hebben op de kenmerken van hun lesgeven (bijvoorbeeld op basis van leerlingpercepties of lesobservaties) maar ook op de met leerlingen gerealiseerde leerprestaties. Omdat dergelijke feedback vaak het startpunt is voor verbeteringsacties richt hij zich in zijn onderzoek ook op de vraag hoe leerkrachten zo getraind kunnen worden in het differentiëren van hun instructie dat leerlingen onderwijs-op-maat krijgen. Hij heeft veelvuldig over deze onderwerpen gepubliceerd in wetenschappelijke tijdschriften en boeken.

Ton de Jong is hoogleraar Instructietechnologie aan de Universiteit Twente. Hij is voorzitter van de vakgroepen Instructietechnologie en Onderwijskunde en opleidingsdirecteur van de master "Educational Science and Technology". Zijn onderzoeksinteresse is het gebruik van innovatieve technologieën voor science onderwijs, met name online labs voor onderzoekend leren en het gebruik van tools voor het onderzoekend leerproces. Hij heeft verschillende EU-projecten op dit terrein

gecoördineerd en is momenteel coördinator van het H2020 project Next-Lab. Het gebruik van Learning Analytics voor het ondersteunen van 21st-eeuwse vaardigheden en het aanpassen van tools voor studenten is onderdeel van dit werk. Ton de Jong heeft meer dan 200 artikelen in tijdschriften en boeken gepubliceerd, waaronder drie artikelen in Science. Hij is fellow van de AERA en lid van de Academia Europaea. Voor meer informatie zie: http://users.edte.utwente.nl/jong/Index.htm.