# Data science en visualisatie door de overheid

Prof. Dr Mirjam van Reisen[1]

**Een bijzondere bijeenkomst over Data Science vond plaats in de Tweede Kamer op 20 februari. De bijeenkomst was georganiseerd met de vraag wat data-analyse en visualisatie kan betekenen voor het werk van het Nederlandse Parlement. Doelstelling van de bijeenkomst was om in het kader van een pilotproject 'datavisualisatie en datadashboards' in dialoog te gaan met verschillende overheidsorganisaties over dataontwikkelingen binnen de overheid.**

Tijdens de bijeenkomst werd het gesprek gevoerd langs de volgende vragen:
1. Wat is de visie van de verschillende overheidsorganisaties op de datagedreven wereld?
2. Wat zijn de ontwikkelingen die overheidsorganisatie maken/gemaakt hebben om meer uit data te halen (met een focus op de rol van datavisualisaties en datadashboards)?
3. Hoe gaan overheidsorganisaties om met ethische vraagstukken?

Dit resulteerde in een indrukwekkende bijeenkomst met zo'n vijftig deelnemers met grote expertise in het gebruik van grote data in overheidscontext. Leiden Centre of Data Science was uitgenodigd om expertise te verlenen over de toekomst van data science en de opstart van de derde generatie internet, het internet van open data en diensten. In de afgelopen twee jaar werd 90% van de data gegenereerd – en de omvang zal nog toenemen gezien de hoeveelheid data die nu ook gegenereerd wordt via het *internet-of-things*.

**Toegankelijkheid en bescherming van privacy**
Parlement komt van het woord 'parler' – praten, en dit kenmerkt ook wel de communicatie die in een parlement belangrijk is. De Tweede Kamer heeft ten minste drie rollen ten aanzien van data. Ten eerste produceert zij data vanuit het parlementaire proces: stemmingen, verslagen en dit alles vanuit de optiek van open data om de burger zoveel mogelijk toegang en inzicht te geven in het parlementaire werk. Ten tweede, de controlerende taak van het parlement; hoe kan deze worden ondersteund door meer gebruik van data en visualisatie van data? En ten slotte, welke regels moeten eigenlijk van toepassing zijn om enerzijds de toegankelijkheid en het gebruik van data te ondersteunen en anderzijds te zorgen voor voldoende bescherming van de burger en haar rechten op privacy?

De bijeenkomst boog zich over de vraag: wat is de visie op datagestuurde informatie; welke ontwikkelingen vinden al plaats om in beleidsvorming data-analyse mee te nemen, en welke ethische vragen komt men tegen in de dagelijkse praktijk?

De rijke discussie kan worden samengevat rond een aantal thema's. Enerzijds een ontwikkeling om data-analytische modellen te construeren die de beleidsmakers en Ministers kunnen helpen in meer voorspellende zin. Hierbij doen zich ook vragen voor over de data inclusie in de modellen, en wat de data-modellen wel en niet representeren. De data-analisten bespraken in die zin ook het belang om data-modellen te zien als aanvullend op de kennis over de context en waarschuwden ervoor dat de data-modellen niet het werk kunnen overnemen van de interpretatie van de data. Daarvoor is het wel van belang dat beleidsmedewerkers beter inzicht krijgen in de processen van data-analyse en wat daarin wel en niet te lezen is, maar moet men vermijden dat data-gestuurde advisering de betekenisvolle afweging van de beleidsmedewerker overneemt, omdat de uitkomsten van de data-analyse altijd ook om een interpretatie vragen.

[1] Mirjam van Reisen is Hoogleraar Computing for Society, Leiden Centre of Data Science, Leiden Institute of Advanced Computer Science, Universiteit Leiden.

Een tweede kerngebied dat aan de orde kwam was de vraag die speelt in rijksdiensten met betrekking van welke data al dan niet gebruikt kunnen worden. Vanuit de data-analyse is de koppeling tussen allerhande data nuttig omdat het dikkere informatie kan opleveren, maar aan de andere kant is er niet altijd consensus over welke data al dan niet voor welk doel gebruikt of gekoppeld mag worden.

**Nieuwe generatie internet**
Een derde kerngebied is het belang van meer inzicht in de kennis die verstopt zit in de data, maar ook hoe de data zich vertaalt naar informatie. De informatie is altijd een sociaal proces van interpretatie. De relatie tussen data en informatie moet dus helder zijn in meer op data-analyse gestoelde beleidsprocessen. Informatie is niet alleen een kwestie van 'zenden' van data-uitkomsten maar daarin moet ook meegewogen worden wie de informatie tot zich neemt en wat de reden van de informatiedeling is.

Een vierde kerngebied vormen de snelle ontwikkelingen waarbij de inzichtelijkheid van de data-analytische processen steeds moeilijker te doorgronden is. Wie weet precies hoe een algoritme tot stand kwam? Hoe komen complexe modellen tot een bepaalde uitkomst? Wat was de input, wat was het analytisch proces en wat is de uitkomst? Begrijpen we die relaties nog? Er is een gevaar als data-analyse een zwarte doos is die niet meer inzichtelijk is en leidt tot een situatie waarin de uitkomst het startpunt wordt zonder dat de oorsprong van de data inzichtelijk is. Continu leren – op de achtergrond van de processen van data science is dus cruciaal.

Ten slotte werd de nieuwe generatie internet van data en diensten besproken. Dit belangrijke Nederlandse initiatief dat in Leiden werd ontwikkeld en ondertussen door de Europese Unie is overgenomen, gaat vanaf 2020 van start met het volgende meerjarenprogramma research waarin de wetenschap verplicht wordt om data die met publiek geld werd gegenereerd open te stellen. De principes die de basis vormen voor dit nieuwe internet zijn de volgende: vindbaarheid, openbaarheid (onder duidelijk gedefinieerde voorwaarden), interoperabiliteit en herbruikbaarheid – wat leidt tot het Engelse acroniem FAIR (Findable, Accessible, Interoperable and Reusable). De essentie is een architectuur te ontwikkelen waarmee data door machines kan worden geanalyseerd – wat noodzakelijk zal zijn gezien de omvang van de data die nu wordt gegenereerd.

**Innovatie-hackathon**
Wat betekent dit nu voor de ethische vragen omtrent overheid en data-analyse? Allereerst het belang om goed om te gaan met privacy gevoelige data. Een belangrijke conclusie was tevens dat de doelen vooraf bepaald moeten worden en dat deze duidelijk en controleerbaar moeten zijn. Er moet voorts altijd rekening worden gehouden met de bias die kan optreden door data alleen te controleren aan één bepaalde groep. Het is dus belangrijk er expliciet zorg voor te dragen dat er geen discriminatie plaatsvindt.

Ten slotte werd er uitgebreid gesproken over de voordelen van de visualisatie waarmee de overheidsdiensten ingewikkelde problemen inzichtelijk kunnen maken en die de communicatie van de overheid naar de burger kunnen ondersteunen. In een vervolg zal de Tweede Kamer een innovatie-hackaton organiseren om zo het belang van data-analyse en visualisatie ook in het werk van de Tweede Kamer te onderstrepen.