De combinatie van big data en collective intelligence

Annemieke Roobeek, **Jacques de Swart**, **Myrthe van der Plas** en **Michel Bourgonje**[1].

In dit artikel geven we een inkijkje in een recente casus van verantwoord gebruik van big data met collective intelligence om een potentieel gevaarlijke situatie snel onder controle te krijgen en de verdachte in te rekenen. Deze aanpak biedt ook perspectief voor andere toepassingsmogelijkheden, zoals het opsporen van fraude en het bestrijden van terrorisme.

NS worstelde met een urgent veiligheidsprobleem: een reeks treinbranden, met als gevolg maandenlang uitval van materieel, overlast door vertragingen, hoge kosten. De garantie op de veiligheid van reizigers en eigen personeel werd op de proef gesteld. Omdat de incidenten elkaar snel opvolgden, was haast geboden, maar er was geen zicht op een oplossing. Een nachtmerrie voor elke vervoerder.

Elk incident was tot op dat moment door lokale bevoegde instanties bekeken. Daardoor was er geen zicht op een mogelijk patroon. Je hebt data nodig en een intelligente aanpak om er snel iets zinnigs mee te doen. De vraag die door de NS gesteld werd, luidde: is er een andere manier of methode die meer licht op de incidenten kan laten schijnen? Wat kan de combinatie van big data en collective intelligence in zo'n geval meer opleveren dan een conventionele opsporingsaanpak? Wat komt erbij kijken? Hoe ziet zo'n gecombineerde aanpak eruit en wat levert het op?

Kenmerken van big data aan de hand van de 5 V's

Big data speelden een belangrijke rol in de aanpak. Maar wat verstaan we onder big data in deze context? Er zijn immers vele definities van big data. Wij beperken ons hier tot een illustratie aan de hand van de vijf V's van Anil Jain (https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/) en refereren meteen aan de big-data-analyse in onze casus.

> ***» Urgent NS-veiligheidsprobleem: data nodig en een intelligente aanpak***

- **Variety**. Deze V heeft niet alleen betrekking op het koppelen van data uit verschillende bronnen, zoals OV-chipdata, tweets, incidentmeldingen en dienstregelingen, maar ook op de aard van de data, bijvoorbeeld of deze gestructureerd of ongestructureerd zijn, en of ze uit beeldmateriaal, tekst of cijfers bestaan.
- **Volume**. Deze V is geen doel op zich, maar al snel worden databestanden groot in omvang. Dit geldt bijvoorbeeld voor de OV-chipdata, waarin miljoenen reisbewegingen per dag worden vastgelegd. Omdat je volgens de inductieve aanpak je in eerste instantie niet wilt beperken tot die reisbewegingen die je bij voorbaat verdacht acht, neem je al deze reisbewegingen mee.
- **Veracity**. Hoe waarheidsgetrouw zijn de data? Uiteraard is het streven om zoveel mogelijk schone, waarheidsgetrouwe data te gebruiken. Bij bijvoorbeeld Twitterberichten is dit streven niet haalbaar. Rondom incidenten blijken ook veel onzintweets gepost te worden. De kunst van het gebruik van big data is deels ook data die niet 100% schoon zijn te benutten.
- **Velocity**. De snelheid waarmee nieuwe data worden toegevoegd, maakt het minder efficiënt om de data eerst te bevriezen, dan analyses de definiëren die reproduceerbare resultaten opleveren. Het is juist de kunst om de analyses ook te draaien op data die zojuist zijn ontstaan. Vlak na een incident geldt dit bijvoorbeeld voor het betrekken van camerabeelden.
- **Value**. Dit is veruit de belangrijkste V. Hoe zorg je ervoor dat je niet eindeloos blijft schrapen en schaven om inzichten te verwerven uit de oneindige mogelijkheden in de zoektocht naar de *golden nuggets* die in de data verstopt zijn, maar dat je uiteindelijk waarde creëert in de vorm van *actionable insights*: patronen die getoetst zijn en waarop actie ondernomen kan worden. Het antwoord op de vraag: welke verdachte(n) kunnen we gaan volgen in de datastromen? Uiteraard is hier de afweging tussen *false positives* (mensen die ten onrechte op de lijst van verdachten staan) en *false negatives* (mensen die onterecht niet op de lijst van verdachten staan) cruciaal. Hier bewijst zich de meerwaarde van de collective intelligence door met elkaar enerzijds de hypothesen ruim te houden, maar ook uitsluitingen aan te geven.

Aanpak

We behandelen de vijf belangrijkste ingrediënten van deze gecombineerde kwantitatieve en kwalitatieve aanpak:

1. Het creëren van een interdisciplinaire setting om collective intelligence met deskundigen te optimaliseren.
2. Het faciliteren van een exploratieve dialoog over mogelijke hypotheses om tunnelvisie te voorkomen.
3. Het uitvoeren van diverse big-data-analyses om het grotere plaatje te visualiseren.
4. Een geavanceerd *data experience lab* dat een creatieve manier van werken bevordert.
5. Tot slot het belang van snelheid in een interactief en iteratief proces.

1. Creëren van een interdisciplinaire setting voor collective intelligence

Onze big-data-aanpak met collective intelligence startte vanuit de overtuiging dat data alleen niet veel zeggen, zolang er niet vanuit interdisciplinaire kennis, kunde en ervaring een aanzet gegeven kan worden voor een brede set hypothesen om meerdere zoekrichtingen te geven aan de datadeskundigen. Bij het verschijnen van patronen in de data was de gezamenlijke, interdisciplinaire interpretatie essentieel, omdat hierbij feiten met formele en informele kennis geconfronteerd werden. De combinatie van big data en collective intelligence was vereist in een proces dat interactief en iteratief verliep. Het wederzijds testen en toetsen van diverse hypothesen en sets van verschillende data was noodzakelijk om niet in een eenzijdige tunnelvisie terecht te komen. Men moest in heel korte tijd als een netwerkend team op elkaar ingespeeld raken en men moet elkaar kunnen uitdagen op de uitgangspunten van de hypotheses en de bevindingen uit de data-analyses.

2. De exploratieve dialoog stimuleert een brede zoekrichting

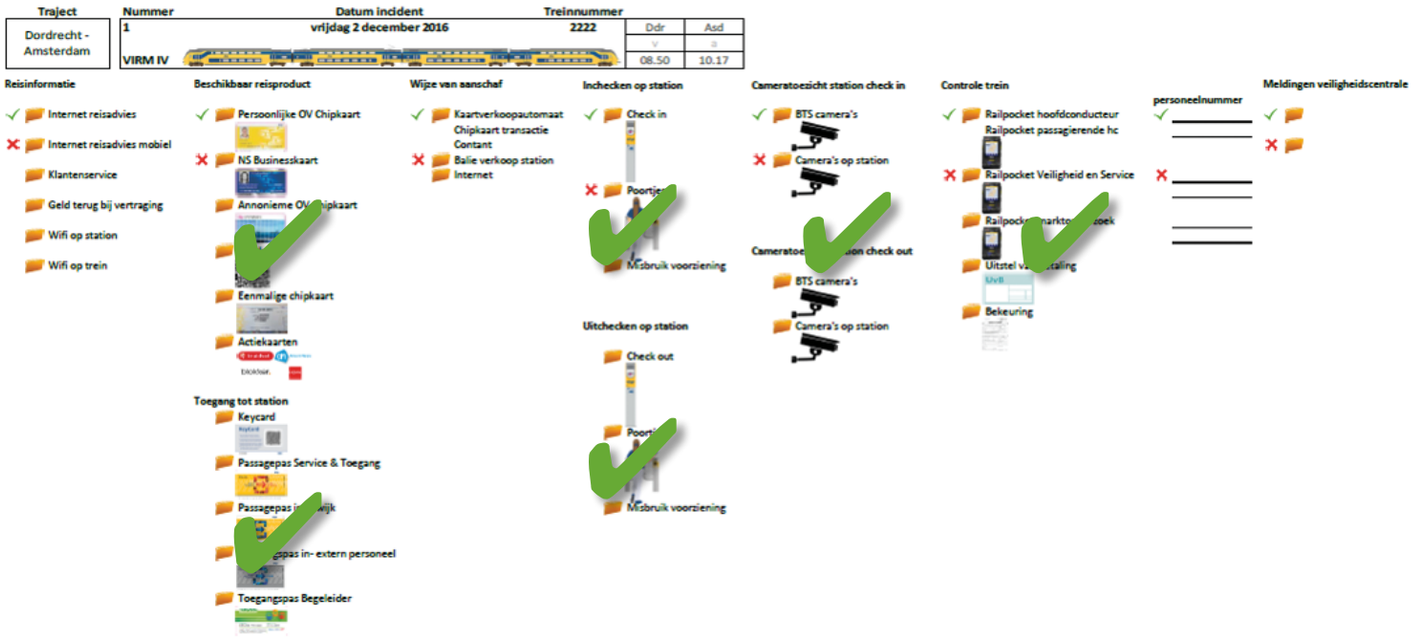
Een exploratieve dialoog tussen de datadeskundigen en de materiedeskundigen bevordert dat ze worden uitgedaagd hun kennis, kunde en ervaring optimaal in te zetten. Dit in plaats van de traditionele aanpak om vanuit een hypothese het bewijs deductief te gaan vinden. Out-of-the-box denken wordt gestimuleerd, meerdere hypothesen worden geformuleerd, kennis uit diverse kennisgebieden wordt aangeboord en de data zijn een basis om vooral creatief na te denken en meerdere wegen uit te proberen. Alleen wanneer iedereen op een gelijkwaardige en dus absoluut niet-hiërarchische manier actief meedoet, is de kwaliteit van de collective intelligence te garanderen en kan de interpretatie van de data-analyse betekenisvol plaatsvinden. Hiervoor is een ervaren procesbegeleider nodig om de datadeskundigen te voeden met de hypothesen van de materiedeskundigen en om de interpretatie van de data-analyse met de materiedeskundigen via doorvragen scherp te houden.

3. Big-data-analyse om patroonherkenning in het grotere plaatje te zien

Een grote variëteit aan data werd zowel gebruikt om de tijdens de sessies gedefinieerde hypothesen te toetsen als ook om nieuwe hypothesen te voeden. Hierbij kan gedacht worden aan data van vervoerbewijzen (OV- kaarten), data van de tijdstippen van de incidenten, data over de trajecten waarop de incidenten plaatsvonden, data van de controles voordat een incident had plaatsgevonden, data over boetes, stationslocaties, agressiemeldingen, mogelijke routes, data over reizigers die hetzelfde traject als verdachte(n) aflegden, camerabeelden, data van controlepoortjes, data over afwijkend reisgedrag, sociale-media-data, en informatie die door conducteurs genoteerd werd.

Er zijn diverse analyses uitgevoerd met verschillende ondersteunende software.

1. Er is een geïntegreerde analyse uitgevoerd, waarbij verschillende databronnnen aan elkaar gekoppeld zijn om incidenten te plotten op tijd en locatie en daaromheen zoveel mogelijk informatie te voorzien.
2. Er heeft tekst mining plaatsgevonden om te zien welke informatie er naar boven komt bij bepaalde zoektermen. Hiermee kon een completere incidentenlijst ontwikkeld worden. Belangrijk is ook dat hiermee aanwijzingen



kwamen voor informanten, waarbij er *één* uiteindelijk als kroongetuige aangewezen kon worden.

3. Er is een '*strangeness*'-analyse uitgevoerd om op basis van de OV-kaart-data het (afwijkende) reisgedrag van de vervoersbewijsnummers in kaart te brengen. Hiermee kon bijvoorbeeld een patroon van ongebruikelijke overstappen ('loops') zichtbaar gemaakt worden, maar ook kon hiermee een medereiziger geïdentificeerd worden die op tenminste twee van de trajecten met incidenten meegereisd is. Dit kon later ook bevestigd worden met data over in- en uitchecken bij toegangspoortjes.

Tegelijkertijd lieten de verschillende analyses ook zien dat het belangrijk is om de hypothesen in het begin breed te houden, omdat de data ook veel ruis opleveren. De materiedeskundigen, bijvoorbeeld met kennis van treinnetwerk en de specifieke trajecten en stations, konden deze ruis vaak direct verklaren. Dat voorkwam dat op verklaarbare afwijkingen in de data verdere analyse gepleegd ging worden. Tegelijkertijd konden mogelijke hits ook meteen gecheckt worden en bij positief resultaat konden andere hypothesen afgevoerd worden. Dit hield snelheid in het zoekproces zonder dat er tunnelvisie ontstond.

4. Meerwaarde van samenwerken in een data experience center

Door in een en dezelfde ruimte tijdelijk als interdisciplinair team te werken, in dit geval een geavanceerd data experience center, kon er voortdurend onderling geschakeld worden tussen datadeskundigen en materiedeskundigen die aan de hypothesen werkten. Een enorm touchscreen ('the wall') besloeg een groot deel van de creatief ingerichte ruimte. Krachtige computerapparatuur en de beschikbaarheid van big-data-analyse programma's ondersteunden het proces. Zo kon direct met elkaar en met externe collega's overlegd worden om bijvoorbeeld actuele ontwikkelingen mee te nemen. Op deze manier konden diverse incidenten meteen getoetst worden met de verschillende dataprogrammatuur. Resultaten werden tijdens het proces direct op grote schermen geprojecteerd. Gezamenlijk werden resultaten uit de data-analyse besproken om tot een overwogen interpretatie te komen.

De inrichting van het data experience center speelde een belangrijke rol in het informeel samenwerken met elkaar en het voeren van een open dialoog. Zo'n laboratorium doorbreekt het traditionele patroon waarin analisten vooraf gedefinieerde analyses draaien en die rapporteren aan het management dat beslissingen neemt. Deze stappen liepen nu dwars door elkaar heen en iedereen was gelijkwaardig in deze informele, creatieve setting vol high tech. Zo onstond een stimulerende werksfeer van samenwerking aan een complexe opgave.

5. Snelheid

Het proces werd gekenmerkt door een enorme snelheid. De totale exercitie om van vraagstelling tot oplossing te komen heeft in totaal 21 uur gezamenlijke tijd gekost. Daarnaast zijn er werkzaamheden verricht om aanvullende informatie naar boven te halen en zijn er data-analyses gedraaid. Op vrijdagmiddag 15.00 werd de hulpvraag gesteld, op maandagochtend om 09.30 begon het interdisciplinaire team in het Data Experience Lab. Twee sessies volgden hierop. Een week later op de donderdag was de zaak opgelost.

Niet alleen tijdens de drie sessies was sprake van snelheid in de samenwerking, maar ook tussen de sessies door werd er direct opvolging gegeven als er een belangrijke ontwikkeling was. Na een nieuw incident werd er onmiddellijk contact opgenomen met de procesbegeleiding. Het plannen van een nieuwe sessie ging razendsnel, ondanks dat veel partijen deelnamen.

Het proces in het Data Experience Lab

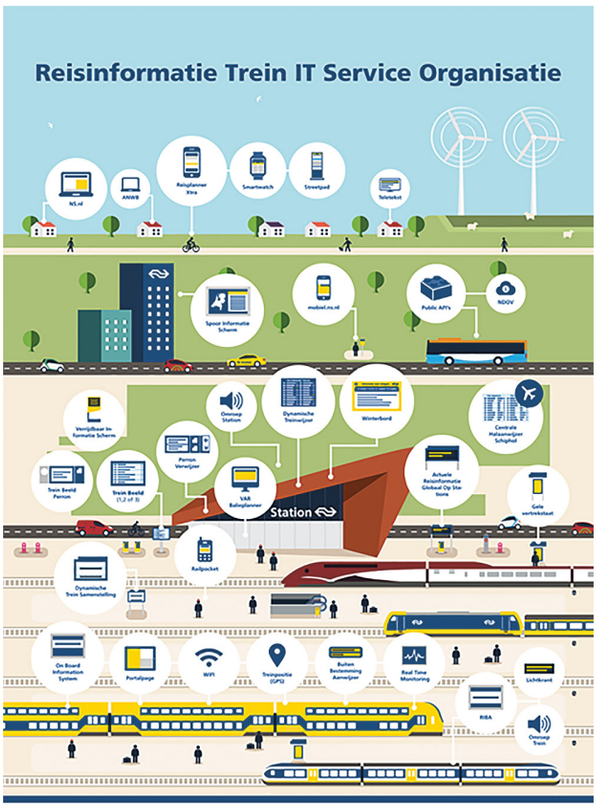
De eerste sessie op maandag duurt 8 uur. Er zijn ruim 20 mensen aanwezig met diverse expertise: security, data-analyst, cybersecurity, forensische specialisten, IT, expertise van het OV-net, landelijke politie, expert in benutten van collective intelligence, strateeg. Er wordt vooral verkend: welke data er voorhanden zijn, hoe de incidentenlijst volledig gemaakt kan worden, welke technieken er kunnen worden gebruikt , welke tijdsintervallen interessant zijn, wat er uit tekstanalyse kan worden gehaald en hoe de verschillende hypotheses met reisgedrag op basis van

OV-kaartinformatie onderbouwd kunnen worden. Parallel werkt men aan het breed houden van de hypotheses om tunnelvisie te voorkomen. Hypotheses worden voorgelegd aan data-experts en pas weggestreept wanneer de data een hypothese niet kunnen onderbouwen. Na deze sessie zijn vragen over informatie en data uitgezet.

De tweede sessie is 8 dagen later en duurt 9 uur. Er zijn ongeveer 12 mensen aanwezig, een kleiner aantal omdat nu beter is in te schatten welke expertise precies benodigd is voor het kwalitatieve en kwantitatieve deel van de opgave. Bovendien is er tussen de eerste en de tweede sessie veel data-analyse verricht, zodat er in het begin al meer gepresenteerd kan worden. De uitwisseling tussen de verschillende experts verloopt uitstekend. Hierin komt ook de meerwaarde van de eerste dag actief samenwerken in het Data Experience Lab terug. De hypotheses worden in deze tweede sessie opnieuw eerst breed gehouden om tunnelvisie te voorkomen. In de loop van de sessie worden de persona's voor verdachten teruggebracht naar drie, waarbij al de hypothese wordt ingebouwd dat het feitelijk om twee mensen zou kunnen gaan, omdat de tweede en de derde persona dezelfde kunnen zijn. Voor twee informanten is nader onderzoek geadviseerd, omdat deze personen de branden actief gezien hebben.

Er worden die dag meerdere hits gevonden voor een van de verdachten in de relevante tijdsvakken op de locaties rondom de incidenten. Er zijn tenminste twee hits voor een andere verdachte. Het reisgedrag van deze verdachten is verschillend en worden vooralsnog niet met elkaar in verband gebracht. Daarom worden de hypothesen voor de beide verdachten open gehouden. Na deze sessie zijn aanvullende vragen uitgezet.

De derde (slot)sessie is twee dagen later op donderdagmiddag, en duurt 4 uur. De uitkomsten van de tweede sessie worden nog verder doorgeëxerceerd om te checken of er niets over het hoofd wordt gezien, er geen tunnelvisie is opgetreden en er mogelijke aanvullende relevante informatie uit de data-analyses is over verdachte personen en reizigers op hetzelfde traject. Alle verbanden uit de data-analyses en de hypotheses worden naar boven gehaald zijn en er wordt een match gevonden tussen het reisgedrag van een van de verdachten en de reeks incidenten. Verder wordt in deze sessie bevestigd dat een eerdere verdachte als kroongetuige beschouwd kan worden, omdat die met de verdachte samen heeft gereisd op de trajecten waar brand is geweest. Een andere verdachte kan worden weggestreept als verwarde reiziger.



RTSO: Reisinformatie van deur tot deur en IT op de trein

In 3 sessies binnen tien dagen is de missie volbracht: de dader van de treinbranden is bekend en in beeld bij de politie. De kroongetuige is benaderd.
Tijdens het proces zijn de privacyregels in acht genomen. Er is voor het big-data-onderzoek gewerkt met geanonimiseerde bestanden. De resultaten van het data-onderzoek zijn aan de NS overgedragen. Alle beschikbare data zijn vervolgens gevorderd door het OM bij NS.
Uiteindelijk is de verdachte op heterdaad betrapt, aangehouden en voorgeleid. «

> » *In drie sessie binnen tien dagen missie volbracht: de dader is bekend*

1. Prof. dr. Annemieke Roobeek is hoogleraar Strategie en Transformatiemanagement verbonden aan Nyenrode Business Universiteit en directeur van MeetingMoreMinds. Zij was degene die bovenstaand project leidde. Prof. dr. Jacques de Swart is hoogleraar Applied Mathematics, eveneens verbonden aan Nyenrode en hij leidt de Data Analytics Practice van PwC, waar hij partner is. Myrthe van der Plas is als manager verbonden aan de Data Analytics Practice van PwC te Amsterdam. De sessie in het Data Experience Lab bij PwC is mede door Dr. Andre Mikkers, forensisch deskundige en partner van PwC, georganiseerd en gefaciliteerd. Michel Bourgonje is Programmamanager Sociale Veiligheid bij NS Security.