MENSELIJKE DATAWETENSCHAP

De meningen over menselijke datawetenschap zijn verdeeld in grofweg twee kampen: het ‘alles is fantastisch’-kamp en het ‘alles is waardeloos’-kamp. Als we verder willen komen met *data science* over mensen moeten we af van beide kampen, en menselijke datawetenschap benaderen als een echte wetenschap. Daarvoor moeten we (nog) meer samenwerken over de grenzen van de disciplines heen.

Daniël Oberski

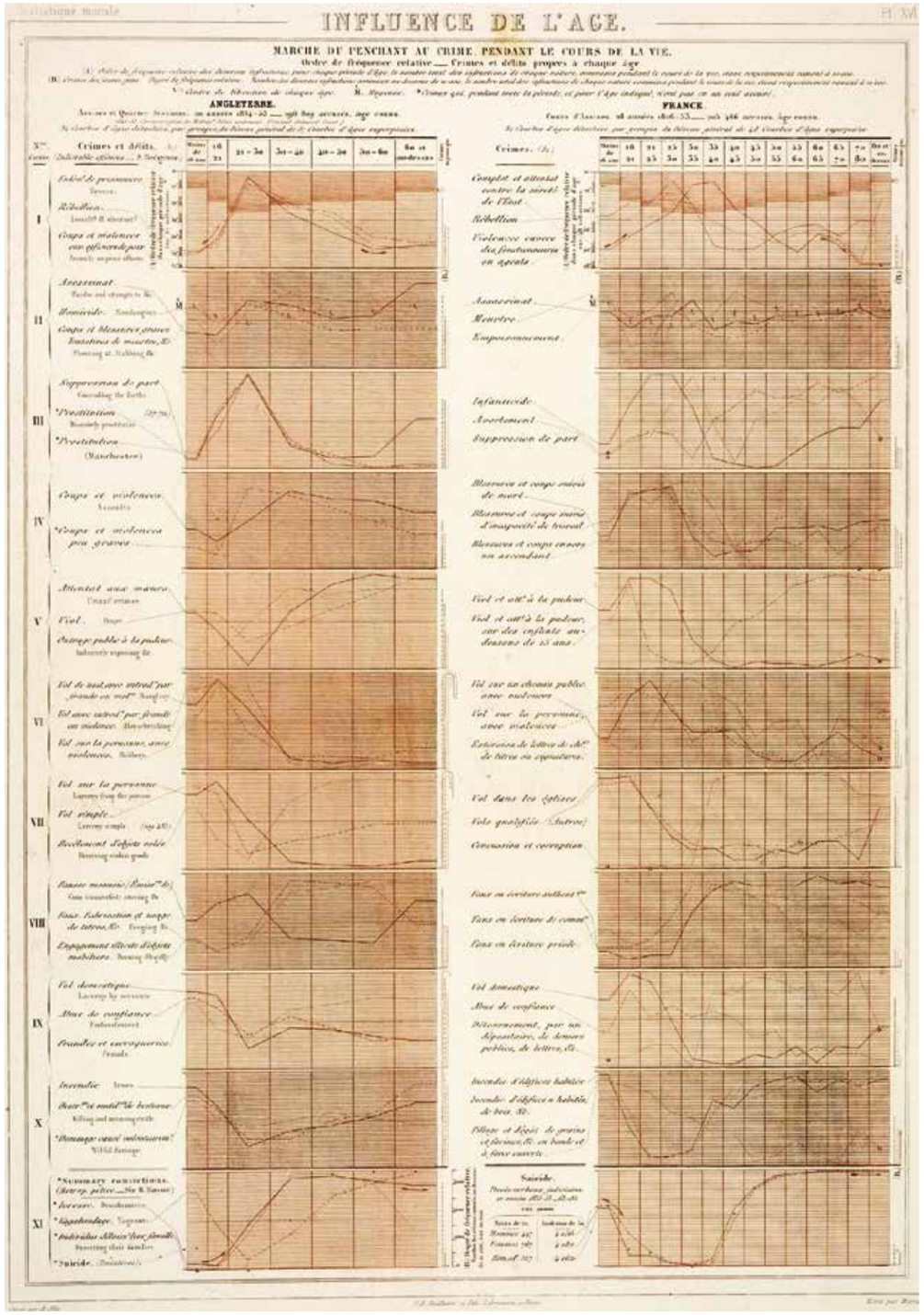
Op 2 juli 1832 presenteerde de 29-jarige datawetenschapper avant la lettre André-Michel Guerry een kort manuscript voor de prestigieuze Académie Française. Zijn presentatie, gepubliceerd als *Essai sur la statistique morale de la France* (Guerry, 1833), veranderde het leven van de spreker, die een prijs en een felbegeerde plek in de academie verdiende. Tot dat moment was Guerry een bescheiden advocaat uit een provinciestad geweest, met de ogenschijnlijk sombere taak departementale misdaadgegevens te archiveren. Door zijn tijd als ‘datamanager’ was Guerry zich echter iets gaan realiseren: data over mensen lagen voor het oprapen, en daarmee nieuwe kennis over mens en maatschappij. Een mooi voorbeeld toont figuur 1. Dit inzicht stond niet alleen aan de wieg van de sociale wetenschap, maar ook aan die van de moderne visualisatietechniek (Friendly, 2007). Je zou dus kunnen zeggen dat het praatje van Guerry niet alleen zijn eigen leven heeft veranderd.

Als sinds de tijd van Guerry en zijn bekendere tijdgenoten Quetelet, Nightingale, en Playfair zijn er twee kampen: zij die datawetenschappers als de apostelen van de menselijke vooruitgang zien, en zij die vinden dat data-analyse over mensen bij voorbaat waardeloos, bedreigend of onwetenschappelijk is. Tegenwoordig staat de krant bol van de vele gevaren waaraan kwaadaardige data scientists de mensheid onderwerpen – of juist weer van de revoluties in gezondheid en geluk, die, zo lijkt het, steeds op het punt staan *bijna* te gebeuren. Zelfs Lacroix, de persoon die Guerry in 1832 introduceerde bij de academie, schrijft in het voorwoord van *Essai sur la statistique morale de la France*: ‘Parmi les différen[t]s objets qui sont du ressort de la statistique, un des plus importants et des plus difficiles à traiter, consiste dans l’énumération et le classement des actions humaines.’ Hij was er duidelijk ook niet helemaal uit of menselijke datawetenschap nu dom was of briljant.

Geen van beiden weten we nu. Menselijke datawetenschap moet precies dat worden: een *wetenschap*. En zoals elke wetenschap zullen we die we moeten benaderen met optimisme waar gerechtvaardigd, en met zorg waar geboden.

Datawetenschap op mensen: dom of slim?

Wat zou je kunnen bereiken als data science met menselijke gegevens goed werkt? Daar zijn talloze voorbeelden van te bedenken, van automatische segmentering



Figuur 1. Guerry’s visualisatie van relatieve misdaadcijfers (verticale as) voor verschillende leeftijdsgroepen (horizontale as) voor Engeland (linker kolom) en Frankrijk (rechter kolom). Vergelijkbare misdaden zijn samengevoegd. Collectie: Universitat de Barcelona, Biblioteca Patrimonial Digital. Creative Commons Public Domain Mark 1.0

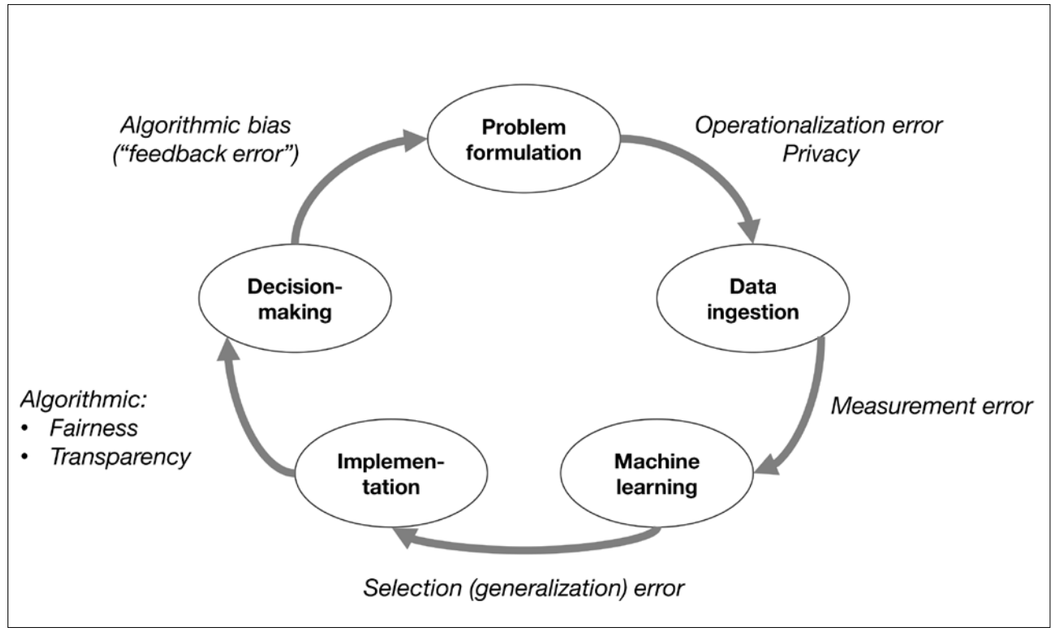
van radiologische beelden, verwerking van -omics data in bio-informatica, vroegtijdige waarschuwing voor sepsis op neonatale intensive care-afdelingen, tot voorspelling van de huizenmarkt, of modellering van de verspreiding van meningen over infectieziekten én vaccinaties via sociale netwerken. De lezer zal zelf nog wel meer en betere voorbeelden kunnen bedenken.

Een opvallend kenmerk van veel datawetenschappers is een optimistische houding ten opzichte van data. Bij het maken van de prachtige grafiek weergegeven in figuur 1 bijvoorbeeld, realiseerde Guerry (1864) zich vast wel dat Engeland en Frankrijk verschillende definities van misdrijven hanteerden, dat er behoorlijk wat ontbrekende gegevens waren, en dat de Engelse gegevens minder nauwkeurig waren. Let bijvoorbeeld op de verschillen in *binning* tussen de twee kolommen (zie ook Friendly, 2007). Maar gelukkig lijkt geen van deze problemen hem op een mentaal dwaalspoor van onredelijke angsten te hebben gebracht. De oorsprong van deze optimistische houding zou kunnen liggen in het feit dat data science deels geworteld is in de ingenieurswetenschappen (Breiman, 2001). Immers, als MacGyver moest ontsnappen uit een haaienkooi verspilde hij ook geen tijd met filosofische traktaten over de adhesieve eigenschappen van ducttape, en wat ducttape eigenlijk betekent, metafysisch gezien. Het was plakken of zakken. Die *can do attitude* speelt volgens mij een belangrijke rol in de huidige populariteit van menselijke data science.

Figuur 2 toont een benadering van (menselijke) datawetenschap in een geïdealiseerde cyclus van onderzoek en 'productie' (waarbij een 'product' ook een inzicht kan zijn). Denk bijvoorbeeld aan een systeem dat uit gegevens van een app en elektronisch patiëntendossier leert om te voorspellen wie moet worden doorverwezen en wie naar huis kan. Of een systeem dat voorspelt waar politie het beste ingezet kan worden. Of een systeem dat pesten op Instagram detecteert, of dat wetenschappers helpt te begrijpen hoe groepscultuur menselijk handelen beïnvloedt. Er zijn talloze systemen in omloop die deze cyclus doorlopen, met als doel de menselijke conditie te verbeteren.

Daar kunnen natuurlijk ook de ons welbekende problemen bij ontstaan. Sterker nog: bij elke keuze in figuur 2 kan een probleem ontstaan, waaronder:

- De probleemstelling kan slecht geformuleerd zijn;
- De gegevens kunnen niet beschikbaar zijn vanwege privacyproblemen, of ze kunnen verkeerd gekozen zijn;
- De gegevens zullen meetfouten bevatten, die de analyse kunnen verstoren;
- De gegevens kunnen onrepresentatief zijn, of het model te idiosyncratisch voor de verkregen data. Beide verstoren generalisaties naar de toepassingssituatie;
- Het adviesalgoritme kan discrimatoir zijn, of te mysterieus en ondoorzichtig om vertrouwd te worden;
- Het geven van het advies zelf kan de gegevens verstoren, denk aan algoritmische politie-inzet in een straat met hoog inbraakrisico.



Figuur 2. Bij elke stap van deze gestileerde menselijke data science-cyclus treden fouten op

Om met een twintigste-eeuwse wijsgeer te spreken: elk nadeel heb zijn voordeel. Daarom moeten we de volgende vragen beantwoorden:

- Hoe groot zijn de voordelen? Bestaan ze wel echt in de praktijk?
- Hoe groot zijn de nadelen? Zijn ze groot genoeg om de positieve effecten weg te nemen?
- Wegen de voordelen (bijv. betere gezondheidszorg) op tegen de nadelen (bijv. gebrek aan transparantie)?

Naar mijn mening is er geen manier om deze vragen in het algemeen te beantwoorden. Soms zullen ze positief uitpakken voor de menselijke datawetenschap, soms negatief. Het zal bijvoorbeeld best waar zijn dat de UK biobank selectiefouten bevat. Maar dat betekent niet automatisch dat elke analyse ervan nutteloos is. Evenzeer is het ongetwijfeld waar dat de kosten die je verzekeraar voor je maakt een *proxy* vormen voor je gezondheid, en voorspeld kunnen worden uit je patiëntendossier. Maar helaas blijkt de meetfout in die proxy het voorspellingmodel te verleiden tot onacceptabele discriminatie (Obermeyer et al., 2019; Boeschoten et al., 2020).

Met andere woorden, zowel het huidige blindstaren op de voordelen als het huidige blindstaren op de nadelen in abstracto is zinloos. Voor elk project zullen we bovenstaande vragen opnieuw moeten beantwoorden. Gelukkig hebben we daarvoor een geweldig instrumentarium, het beste zelfs dat we kennen: de wetenschap.

Een wetenschappelijker benadering

Goede menselijke datawetenschap:

- Erkent dat fouten – in gegevens, modelbeslissingen en implementatie – onvermijdelijk zijn;
- Onderzoekt de mate waarin fouten de output daadwerkelijk beïnvloeden;
- Verwijdert deze effecten waar nodig door de fouten te voorkomen of, wanneer dit niet mogelijk of kosteneffectief is, door de effecten te corrigeren.

In overeenstemming met de technische wortels van data science is goede menselijke datawetenschap ook goede engineering. Een goede ingenieur ploegt niet zomaar voort, ongeacht het terrein. In plaats daarvan ontwerpt ze een oplossing waarbij gebruik wordt gemaakt van de beschikbare materialen en waarbij de risico's worden beheerst.

Hier zijn enkele mogelijke dingen die je zou kunnen doen om risico's te beheersen:

- Werk samen met domeinexperts om een nuttig probleem te formuleren en het op gepaste wijze te operationaliseren in een data science-project;
- Ga er niet vanuit dat het effect van meetfouten verwaarloosbaar is; gebruik bijvoorbeeld modellering om deze effecten te corrigeren;
- Gebruik causale modellen en onderzoeksontwerpen om de effecten van selectiefouten en ontbrekende waarden te corrigeren;
- Onderzoek bedreigingen voor de eerlijkheid van beslui-ten en gebruik bestaande technieken om modelbevooroordeling zoveel mogelijk te voorkomen;
- Gebruik geschikte technieken om modelresultaten uit te leggen aan betrokkenen;
- Werk samen met cognitieve wetenschappers om te zien of dat ook echt werkt;
- Onderzoek met sociale wetenschappers en (andere) domeinexperts of en hoe je product werkt in de dagelijkse praktijk.

Hoe werken we aan het bereiken van deze verheven doelen?

Ten eerste zouden menselijke datawetenschappers routinematig een veel bredere kijk op 'modelevaluatie' moeten hebben. Een succesvol data science-project moet niet alleen goed voorspellen in out-of-sample data, maar ook: generaliseren naar andere contexten waarin het zou kunnen worden toegepast, generaliseren naar het bedoelde concept, onzekerheid juist weerspiegelen, beter zijn dan wat er al was, voordelen hebben die opwegen tegen de nadelen, en daadwerkelijk de gestelde doelen in de sociale realiteit bereiken. Om dit alles te verifiëren, zou de **uitgebreide wetenschappelijke studie van data science-projecten een routine moeten worden**.

Ten tweede kunnen **latente variabelenmodellen en causale modellen** een handig raamwerk bieden om over

een aantal van deze problemen na te denken. Latente variabele modellen zijn bijvoorbeeld handig om meetfouten te schatten en te corrigeren (Oberski et al. 2017), en causale modellering is een goede manier om na te denken over selectiefouten (Mohan and Pearl, 2014). Beide zijn belangrijk bij het onderzoeken van de eerlijkheid van modeloutput (zie Boeschoten et al., 2020).

Ten derde moeten we meer samenwerken. Veel meer. Ik pretendeer absoluut niet de eerste persoon te zijn die erop wijst dat deze kwesties belangrijk zijn. Veel van de bovenstaande punten zijn immers al hele onderzoeksgebieden, die zich uitstrekken over statistiek, informatica, ethische en juridische wetenschap, domeinwetenschappen zoals geneeskunde en biologie, en sociale wetenschappen. Maar naar mijn mening werken we niet genoeg samen en brengen we in de praktijk onvoldoende diversiteit aan perspectieven binnen (en dan heb ik het ook over mezelf!). Bij het doorlopen van de data science-cyclus, ook als statistica of -us, is het namelijk maar al te gemakkelijk om je te concentreren op wat je toevallig interesseert, ten nadele van het project als geheel. Niemand, inclusief ikzelf dus, is immuun voor dit probleem. Sommige onderzoeken naar de ethiek van gegevenswetenschap bijvoorbeeld richten zich uitsluitend op ethische, juridische, of maatschappelijke aspecten, en gaan volledig voorbij aan bestaande wiskundige tools om problemen af te wegen in een technische context. Anderzijds richt ander onderzoek zich volledig op de wiskundige tooling zonder na te gaan of het praktisch nut zal hebben in een sociale context: heb ik recht gedaan aan de intelligentie en autonomie van mensen; zullen mensen echt reageren zoals ik dat veronderstel? Natuurlijk kan niemand alles weten. Dat is dan ook precies waarom we harder zouden moeten werken om ***alle*** **velden die aan deze problemen werken te verenigen**.

Wat nu?

Van de statistische heldengalerij Guerry, Nightingale, Quetelet, en Playfair staat alleen Florence Nightingale bekend als iemand die niet alleen een uitvinding deed (het roosdiagram), maar ook daadwerkelijk bijdroeg aan het oplossen van een concreet probleem (sterfte aan het front). De anderen genieten (enige) bekendheid om hun knappe visualisaties van sterfte, criminaliteit, en opleiding, maar niet omdat we denken dat ze die dingen hebben verbeterd. Dat komt, denk ik, omdat hun tijd er nog niet rijp voor was. Knap als Guerry was kon hij in zijn eentje niet op tegen de obstakels in figuur 2, zoals de schrijver van zijn eigen voorwoord al – zeer beleefd – aanvoelde. Onze tijd is anders. Bij zijn presentatie tegenover de opgetrokken wenkbrauwen van de academie, stond Guerry alleen. In 2020 zijn we wereldwijd letterlijk met duizenden. Het wetenschappelijke werk dat de basisvragen van de datawetenschap kan beantwoorden, kan nu beginnen.

Literatuur

Boeschoten, L., Kesteren, E.-J. van, Bagheri, A., & Oberski, D. L. (2020). Fair Inference on Error-Prone Outcomes. *arXiv:2003.07621 [stat.ML]*. http://arxiv.org/abs/2003.07621.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the Author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726.

Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis. *Statistical Science, 22*(3): 368–99. https://doi.org/10.1214/07-STS241.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France précédé d'un Rrapport à L'Académie des Sciences, Par Mm. Lacroix*. Crochard.

Guerry, A.-M. (1864). *Statistique morale de L'Angleterre comparée avec la statistique morale de la France*. Baillière.

Mohan, K., & Pearl. J. (2014). Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data. In *Advances in Neural Information Processing Systems*, 1520–8. http://papers.nips.cc/paper/5575-graphical-models-for-recovering-probabilistic-and-causal-que-ries-from-missing-data

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm Uused to manage the health of populations. *Science, 366*(6464), 447–53. https://doi.org/10.1126/science.aax2342.

Oberski, D. L. (2020). Human Data Science. *Patterns, 1*(4), 100069. https://doi.org/10.1016/j.patter.2020.100069.

Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association, 112*(520), 1477–89. https://doi.org/10.1080/01621459.2017.1302338.

Daniel Oberski heeft een gedeelde aanstelling als universitair hoofddocent data science aan de Universiteit Utrecht, afdeling Methoden & Statistiek, en aan het Universitair Medisch Centrum Utrecht (UMCU), afdeling Biostatistiek. Zijn werk richt zich op latente variabele modellen en data science-toepassingen in de sociale, gedrags- en biomedische wetenschappen.
E-mail: daniel.oberski@gmail.com

Henk Tijms

column

Vrouwe Justitia (Hildesheim, Duitsland). Foto: Falco via Pixabay

Hoe betrouwbaar zijn loterijen?

In een eerdere column heb ik aandacht besteed aan het loterijschandaal 'Triple Six Fix' in 1980 in Pennsylvania. In de loterij werden tijdens een tv-show in drie containers die elk 10 ballen genummerd van 0 tot en met 9 bevatten de ballen met een luchtblazer door elkaar gehusseld totdat uit elk van de containers een bal in een vacuümbuis was getrokken en in een display was uitgeworpen. De winnende combinatie van drie nummers werd bepaald door de volgorde waarin de ballen de display binnenkwamen. De groep fraudeurs stond onder leiding van Nick Perry die de presentator van de tv-show was. Perry had witte latex verf in elk van de ballen geïnjecteerd behalve in de ballen met de nummers 4 en 6. Dit betekende dat alleen ballen met één van deze twee nummers de vacuümbuis konden bereiken. Vervolgens had de groep heel veel loten gekocht met de acht combinaties 666, 664, 646, 644, 466, 464, 446 en 444. In de bewuste trekking was de winnende combinatie 666, vandaar de naam 'Triple Six Fix'. Het syndicaat van fraudeurs kon niet lang genieten van de ongeveer 1,8 miljoen dollar die ze gewonnen hadden. Ze vielen door de mand door geen voorzichtigheid te betrachten in onderlinge telefoongesprekken. Nick Perry werd veroordeeld tot een gevangenisstraf van zeven jaar. Op het loterijschandaal van Triple Six Fix is een film gebaseerd, zoals ook binnenkort ook een film zal uitkomen over het politieke schandaal en de hypocrisie rond de Winfall-loterij in de Amerikaanse staat Massachusetts. In deze loterij werd jarenlang door de leiding van de loterij doelbewust toegelaten dat de gewone loterijspeler legaal een oor werd aangenaaid door een kleine groep van investeerders. Het script van de film is gebaseerd op de interessante longread '*Jerry and Marge go Large*' van Jason Fagone in de *Huffington Post* van 1 maart 2018.

Het grootste loterijschandaal in de Verenigde Staten kwam in 2017 aan het licht nadat Eddie Tipton, de voormalige beveiligingsdirecteur van een loterij die actief was in meerdere staten, bekende dat hij de uitkomsten van meerdere trekkingen van de loterij had gemanipuleerd. De loterij gebruikte de computer om met een toevalsgenerator de winnende getallen te produceren. Dit ging als volgt: de computer leest een registratie af uit een Geigerteller die straling in de omgevingslucht meet. Deze registratie leidt tot een echt toevalsgetal. Dit getal is het startgetal (seed) voor de Mersenne Twister toevalsgenerator die