Inhoud

Inleiding

Je hoort regelmatig de term 'big data', maar weet je ook echt wat het is en waarom het zo belangrijk is? Kan het verschil maken in je organisatie, kan het resultaten verbeteren en concurrentievoordeel opleveren, en kan het zo zijn dat het níet gebruiken van big data je een significant concurrentienadeel oplevert?

Het doel van dit boek is de term 'big data' te ontrafelen en praktische manieren te vinden om deze data in te zetten met behulp van data science en machine learning.

De term 'big data' verwijst naar een nieuwe categorie data: grote, zich snel vermeerderende hoeveelheden data, die vaak niet in een traditionele structuur passen. De term 'big' is een understatement dat geen recht doet aan de complexiteit van de situatie. De hoeveelheden data waar we mee te maken hebben zijn niet alleen groter dan we gewend zijn; ze zijn ook fundamenteel anders, zoals een motorfiets niet simpelweg een grotere fiets is en een oceaan meer is dan een dieper zwembad. Big data brengt nieuwe uitdagingen en mogelijkheden met zich mee, vervaagt traditionele concurrentiegrenzen en vereist een paradigmaverschuiving met betrekking tot hoe we tastbare waarde ontlenen aan data. Deze oceaan aan data – in combinatie met technologieën die zijn ontwikkeld om ze te kunnen gebruiken – biedt op grote schaal inzichten en heeft een stortvloed aan machine learning mogelijk gemaakt, waardoor computers auto's kunnen besturen, hartinfarcten beter kun-

nen voorspellen dan artsen en buitengewoon complexe spellen als Go beter beheersen dan welke mens dan ook.

Waarom is big data een gamechanger? Zoals we later nog zullen zien, maakt big data het mogelijk om meer diepgaande inzichten uit onze data te halen, inzichten in wat onze klanten motiveert en wat onze productielijnen afremt. Ze stellen bedrijven wereldwijd in staat om miljoenen klanten uitermate gepersonaliseerde ervaringen te bieden en ze leveren het computervermogen dat nodig is om miljarden datapunten wetenschappelijk te analyseren op gebieden als kankeronderzoek, astronomie en deeltjesfysica. Big data biedt de data en het rekenvermogen die de recente heropleving in kunstmatige intelligentie mogelijk heeft gemaakt – vooral als het gaat om vooruitgang op het gebied van **deep learning**, een methodiek die sinds kort wereldwijd de media haalt.

Het big data-ecosysteem stelt ons in staat immense waarde te halen uit big data

Onderzoekers en technici hebben de afgelopen twee decennia niet alleen aan de data zelf gewerkt, maar ze hebben een heel ecosysteem ontwikkeld van hardware- en softwareoplossingen voor het verzamelen, opslaan, verwerken en analyseren van deze stortvloed aan data. Ik noem deze hardware- en softwaretools samen het **big data-ecosysteem**. Dit ecosysteem stelt ons in staat immense waarde te halen uit big data voor toepassingen in het bedrijfsleven, de wetenschap en de gezondheidszorg. Maar om deze data te gebruiken, moet je eerst de onderdelen van het big data-ecosysteem samenvoegen die voor jouw toepassingen het beste werken, en je moet de juiste analysemethoden op de data toepassen – een werkwijze die **data science** wordt genoemd.

Al met al wordt de geschiedenis van big data veel meer dan een simpel verhaaltje over data en technologie. Het gaat over wat er al wordt gedaan in het bedrijfsleven, de wetenschap en onze samenleving, en welk verschil het kan maken voor jouw bedrijf. Je beslissingen moeten verder gaan dan het aanschaffen van een technologie. In dit boek geef ik een overzicht van tools, toepassingen en processen en leg ik uit hoe je waarde kunt ontlenen aan moderne data in al hun vormen.

De meeste organisaties zien big data als een integraal onderdeel van hun digitale transformatie. Veel van de succesvolste organisaties zijn al een heel eind op weg in het toepassen van big data en data science-technieken, waaronder machine learning. Uit onderzoek is een sterke correlatie gebleken tussen het gebruik van big data en omzetgroei (een 50 procent hogere omzetgroei[1]) en het is niet ongebruikelijk dat organisaties die data science-technieken toepassen een verbetering van 10 tot 20 procent in key performance indicators (KPI's) laten zien.

Voor organisaties die het pad van big data en data science nog niet zijn ingeslagen, is de voornaamste barrière dat ze gewoonweg niet weten of de voordelen opwegen tegen de kosten en de inspanningen. Ik hoop in dit boek de voordelen duidelijk te maken, waarbij ik gaandeweg praktijkvoorbeelden geef om de bijbehorende waarde en risico's te onderstrepen.

In de tweede helft van dit boek beschrijf ik praktische stappen voor het ontwikkelen van een datastrategie en om dataprojecten binnen je organisatie gerealiseerd te krijgen. Ik zal bespreken hoe je de juiste mensen samenbrengt en een plan kunt opstellen om data te verzamelen en te gebruiken. Ik bespreek specifieke gebieden waarin big data- en data science-tools binnen je organisatie kunnen worden gebruikt om de resultaten te verbeteren, en ik geef advies over het zoeken en aannemen van de juiste mensen om deze plannen uit te voeren.

Daarnaast bespreek ik andere aspecten die je zult moeten aanpakken, zoals data governance en privacybescherming, om je organisatie te beschermen tegen juridische, concurrentie- en reputatierisico's.

Tot slot geef ik aanvullend praktisch advies voor het succesvol uitvoeren van big data-initiatieven binnen je organisatie.

Samenvatting van de hoofdstukken

Deel 1: Big data ontrafeld

Hoofdstuk 1: De geschiedenis van big data
Hoe big data is uitgegroeid tot een fenomeen, waarom big data de afgelopen jaren zo'n belangrijk thema is geworden, waar die data vandaan komen, wie ze gebruiken en waarom, en wat er is veranderd waardoor vandaag mogelijk is wat in het verleden niet mogelijk was.

Hoofdstuk 2: Kunstmatige intelligentie, machine learning en big data
Een korte geschiedenis van kunstmatige intelligentie of artificial intelligence (AI), hoe het zich verhoudt tot machine learning, een inleiding in neurale netwerken en deep learning, hoe AI tegenwoordig wordt gebruikt en hoe het zich verhoudt tot big data, en enkele woorden van waarschuwing bij het werken met AI.

Hoofdstuk 3: Wat is het nut van big data?
Hoe ons dataparadigma verandert, hoe big data nieuwe kansen blootlegt en bestaande analytische technieken verbetert, en wat het betekent om datagestuurd te werken, mét succesverhalen en praktijkvoorbeelden.

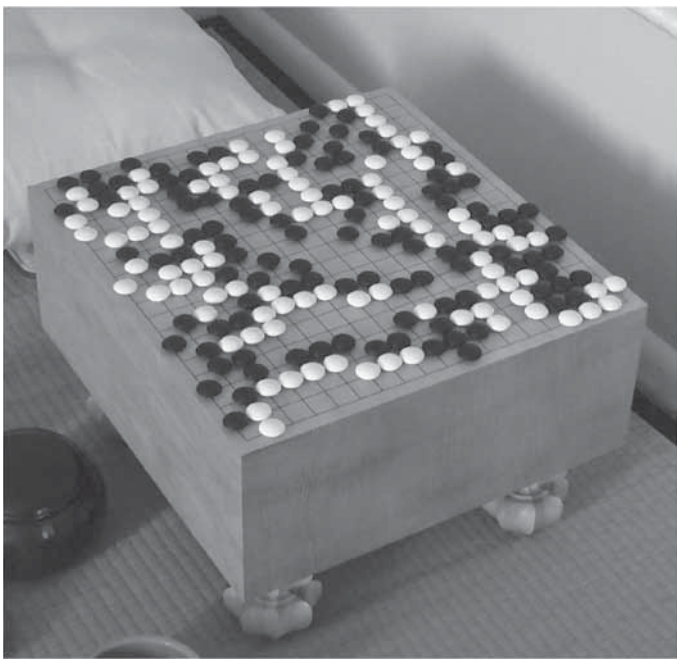
Hoofdstuk 4: Use cases voor (big) data-analyse
Een overzicht van twintig veelvoorkomende bedrijfstoepassingen van (big) data, analytics en data science, met extra aandacht voor manieren waarop big data bestaande analysemethoden verbeteren.

Hoofdstuk 5: Inzicht in het ecosysteem van big data
Een overzicht van belangrijke concepten die verband houden met big data, zoals open source code, distributed computing en cloud computing.

2

Kunstmatige intelligentie, machine learning en big data

Op 11 mei 1997 schreef een IBM-computer genaamd Deep Blue geschiedenis door Gerry Kasparov, de toenmalig wereldkampioen schaken, te verslaan tijdens een partij die werd gespeeld in New York City. Deep Blue won op basis van pure rekenkracht. Het programma beoordeelde maximaal 200 miljoen zetten per seconde terwijl het een lijst met ingeprogrammeerde regels volgde. Maar Deep Blue kon maar één kunstje en werd al snel afgedankt. Computers konden mensen bij lange na niet verslaan in de meest elementaire taken of bij gecompliceerdere spellen zoals het Chinese bordspel Go, dat meer mogelijke spelsituaties heeft dan er atomen in het universum zijn (zie figuur 2.1).



Figuur 2.1 *Een Go-bord.*

We maken een sprong voorwaarts van negentien jaar, naar een wedstrijd in Seoul, Zuid-Korea, toen een programma genaamd AlphaGo de toenmalige wereldkampioen Go, Lee Sedol, versloeg. Kunstmatige intelligentie was niet simpelweg beter geworden in die negentien jaar sinds Deep Blue, het was wezenlijk veranderd. Terwijl Deep Blue beter was geworden door aanvullende expliciete instructies en snellere

processoren, leerde AlphaGo zelf. Het programma bestudeerde eerst zetten van bedreven spelers en trainde die vervolgens tegen zichzelf. Zelfs de ontwikkelaars van AlphaGo konden de logica achter bepaalde zetten van het programma niet verklaren. Het had zichzelf geleerd die te zetten.

Wat zijn kunstmatige intelligentie en machine learning?

Kunstmatige intelligentie of artificial intelligence (AI) is een brede term voor een machine die intelligent kan reageren op zijn omgeving. We maken gebruik van AI als we communiceren met Siri van Apple, Echo van Amazon, zelfrijdende auto's, online chatbots en game tegenstanders. AI helpt ook op minder zichtbare manieren. Het filtert de spam uit je inbox, corrigeert je spelfouten, en besluit welke posts bovenaan je socialemediafeeds verschijnen. AI heeft een heel arsenaal aan toepassingen, inclusief beeldherkenning, taalverwerking, medische diagnostiek, robotbewegingen, fraudedetectie en nog veel meer.

Machine learning (ML) houdt in dat een machine haar prestaties blijft verbeteren ook als je hem niet langer programmeert. ML zorgt ervoor dat de meeste AI zo goed werkt, vooral als er ruimschoots trainingsdata voorhanden zijn. Deep Blue was gebaseerd op regels. Het was AI zonder machine learning. AlphaGo maakte gebruik van machine learning en verwierf zijn vaardigheid door eerst te trainen op een grote dataverzameling bekende zetten en vervolgens tegen zichzelf te spelen om te ervaren wat werkte en wat niet. Aangezien technieken van machine learning beter worden naarmate ze meer data gebruiken, versterkt big data machine learning. Het meeste AI-nieuws vandaag de dag en vrijwel alle AI die ik in dit boek bespreek, verwijzen naar toepassingen van machine learning.

De herkomst van AI

Onderzoekers ontwikkelen sinds de jaren vijftig van de vorige eeuw AI-methoden. Veel technieken die vandaag de dag worden gebruikt

zijn al decennia oud en grijpen terug op de zelfverbeterende algoritmen die zijn ontwikkeld in de onderzoekslaboratoria van Marvin Minsky van MIT en John McCarthy van Stanford.

AI en ML hebben verscheidene valse starts gemaakt. Onderzoekers verwachtten er veel van, maar computers hadden beperkingen en de aanvankelijke resultaten waren teleurstellend. Begin jaren zeventig begon wat wel 'de eerste AI-winter' wordt genoemd, die tot het eind van dat decennium duurde.

Het enthousiasme voor AI herleefde in de jaren tachtig, meer specifiek na successen met expertsystemen. De Britse, Amerikaanse en Japanse overheden investeerden honderden miljoenen dollars in onderzoekslaboratoria van universiteiten en overheden, terwijl ondernemingen vergelijkbare bedragen uitgaven aan interne AI-afdelingen. Er ontstond een bedrijfstak van hardware- en softwarebedrijven die AI ondersteunden.

De AI-bubbel barstte al snel opnieuw. De ondersteunende hardwaremarkt stortte in, expertsystemen werden te duur om te onderhouden en grote investeringen leverden te weinig op. In 1987 bezuinigde de Amerikaanse regering drastisch op de financiering van AI, en de tweede AI-winter was begonnen.

Vanwaar deze recente opleving van AI?

AI kreeg halverwege de jaren negentig een nieuwe impuls, onder andere als gevolg van de toenemende capaciteit van supercomputers. De schaakoverwinning van Deep Blue in 1997 was in feite een returnpartij. Vijftien maanden eerder had het de eerste partij verloren, waarna IBM het programma een grote hardware-upgrade had gegeven.[12] Met tweemaal de verwerkingscapaciteit, won het de returnpartij door bruut rekenvermogen in te zetten. Hoewel het gebruik had gemaakt van gespecialiseerde hardware en hoewel de toepassing van het programma erg beperkt was, had Deep Blue de toenemende kracht van AI aangetoond.

Big data gaven AI een nog grotere impuls dankzij twee belangrijke ontwikkelingen:

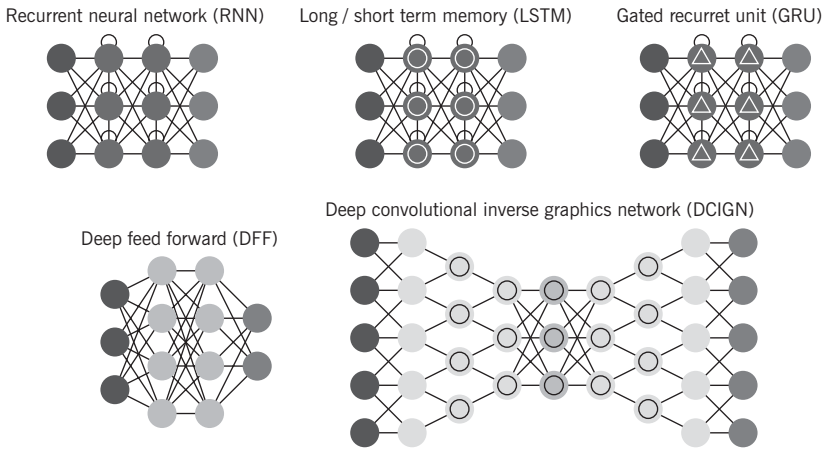
1. We begonnen grote hoeveelheden data te vergaren die gebruikt konden worden voor machine learning.
2. We ontwikkelden software die gewone computers in staat zouden stellen om samen te werken met het vermogen van een supercomputer.

Krachtige methoden voor machine learning konden nu op betaalbare hardware worden gedraaid en konden zich tegoed doen aan enorme hoeveelheden **oefendata**.

Om de schaal aan te geven: toepassingen van machine learning kan tegenwoordig op netwerken van honderdduizenden computers draaien. Een specifieke techniek voor machine learning waarover veel gezegd en geschreven is en die tegenwoordig in toenemende mate wordt gebruikt, is **artificial neural networks** (kunstmatige neurale netwerken), een techniek die kortgeleden is uitgebreid tot grotere (diepere) netwerken en die deep learning wordt genoemd. Deze techniek heeft bijgedragen aan de overwinning van AlphaGo in 2016.

Kunstmatige neurale netwerken en deep learning

Kunstmatige neurale netwerken of artificial neural networks (ANN) zijn er al sinds eind jaren vijftig van de vorige eeuw. Het zijn verzamelingen zeer eenvoudige bouwstenen die samen grotere netwerken vormen. Elke bouwsteen voert slechts een paar basisberekeningen uit, maar het netwerk als geheel kan worden 'getraind' om te ondersteunen bij ingewikkelde taken: foto's labelen, documenten interpreteren, autorijden, een spel spelen enzovoort. Figuur 2.2 geeft voorbeelden van ANN-architecturen. Kunstmatige neurale netwerken worden zo genoemd vanwege hun overeenkomsten met de onderling verbonden neuronen binnen de hersenen van dieren. Ze functioneren als patroonherkenningsinstrumenten vergelijkbaar met de oudste lagen van de visuele cortex van ons brein, maar niet vergelijkbaar met die delen van onze hersenen waar cognitief redeneren plaatsvindt.



Figuur 2.2 *Voorbeelden van architecturen voor kunstmatige neurale netwerken.*[13]

De uitdaging van het bouwen van een ANN is het kiezen van het juiste netwerkmodel (architectuur) voor de basisbouwstenen, en vervolgens het netwerk trainen op de gewenste taak.

Het getrainde model wordt toegepast op een computer, een smartphone of zelfs een chip die in productieapparatuur wordt ingebouwd. Er is een groeiend aantal tools beschikbaar om dit proces van bouwen, trainen en toepassen van ANN's te faciliteren. Dit zijn onder andere Caffe, ontwikkeld door Berkeley Vision Lab, TensorFlow, ontwikkeld binnen Google en in november 2015 vrijgegeven aan Apache, en Theano.

Het 'trainen' van een ANN betekent dat het miljoenen gelabelde voorbeelden ingevoerd krijgt. Als je bijvoorbeeld een ANN wilt trainen om dieren te herkennen, moet je het miljoenen plaatjes laten zien en de plaatjes labelen met de namen van de dieren die erop staan. Als alles goed gaat, zal de getrainde ANN me vervolgens kunnen vertellen welke dieren er op nieuwe, ongelabelde foto's staan. Gedurende de training verandert het netwerk zelf niet, maar de kracht van de verschillende verbindingen tussen de 'neuronen' wordt aangepast om het model nauwkeuriger te maken.

Grotere, meer complexe ANN's kunnen modellen voortbrengen die beter presteren, maar het kan veel langer duren om ze te trainen. De gelaagde netwerken zijn nu meestal veel dieper, vandaar de nieuwe benaming van ANN's als 'deep learning'. Het gebruik van deze netwerken vereist big datatechnologieën.

ANN's kunnen worden toegepast op een breed scala aan problemen. Vóór het tijdperk van big data zeiden onderzoekers wel dat neurale netwerken 'de op een na beste manier waren om elk mogelijk probleem op te lossen'. Dit is veranderd; ANN's bieden nu sommige van de beste oplossingen. Behalve dat beeldherkenning, vertalingen en spamfilters zijn verbeterd, heeft Google in 2015, met de implementatie van RankBrain, ANN's opgenomen in de belangrijkste zoekfunctionaliteiten. RankBrain, een neuraal netwerk voor zoekfuncties, heeft bewezen de grootste verbetering in de kwaliteit van rangschikking te zijn die Google in jaren heeft doorgevoerd. Het is volgens Google de op twee na belangrijkste factor bij het bepalen van de zoekrangschikking.[14]

Praktijkvoorbeeld: De grootste uitdaging ter wereld op het gebied van beeldherkenning

De grootste uitdaging ter wereld op het gebied van beeldherkenning is de jaarlijkse ImageNet Large Scale Visual Recognition Challenge (ILSVRC), waar onderzoeksteams wereldwijd met elkaar strijden om programma's voor machine learning te bouwen om meer dan 14 miljoen afbeeldingen te labelen. In 2012 werd de uitdaging voor het eerst gewonnen door een ANN, en op zeer indrukwekkende wijze. Terwijl de beste classificeringsfoutscore voor eerdere ML-algoritmen 26 procent was geweest, had de ANN een classificeringsfoutscore van slechts 15 procent.

ANN's hebben alle daaropvolgende wedstrijden gewonnen. In 2014 won het GoogleNet-programma met een foutscore van slechts 6,7 procent door gebruik te maken van een ANN met 22 lagen en miljoenen kunstmatige neuronen. Dit netwerk was drie keer dieper dan dat van de winnaar van 2012 en, vergeleken met het aantal neuronen in het dierenbrein, lag hun netwerk iets voor op dat van de honingbij, maar nog altijd achter op de kikker.

Het winnende ML-algoritme van 2016 (CUImage) had de classificeringsfoutscore teruggebracht tot minder dan 3 procent door gebruik te maken van een ensemble van AI-methoden, waaronder een ANN met 269 lagen (10 keer dieper dan de winnaar van 2014).

Hoe AI helpt big data te analyseren

De meeste big data is **ongestructureerde data**, waaronder afbeeldingen, tekstdocumenten en weblogs. We slaan deze in ruwe vorm op en halen er gedetailleerde informatie uit wanneer we die nodig hebben.

Veel traditionele analysemethoden steunen op data die gestructureerd zijn in velden zoals *leeftijd, geslacht, adres* enzovoort. Om een model beter af te stemmen, maken we vaak extra datavelden, zoals *gemiddelde uitgave per bezoek* of *tijd verstreken sinds laatste aankoop*, een proces dat **feature engineering** wordt genoemd.

Bepaalde AI-methoden vereisen geen feature selectie en zijn met name bruikbaar voor data zonder duidelijk gedefinieerde kenmerken. Een AI-methode kan bijvoorbeeld leren een kat op een foto te herkennen door foto's van katten te bestuderen, zonder concepten als kattengezichten, -oren of -snorharen te leren.

Een woord van waarschuwing

Ondanks het prille enthousiasme is de AI die we vandaag de dag hebben nog altijd 'narrow AI'. Elk programma kan slechts worden gebruikt voor de specifieke toepassing waarvoor het is ontworpen en getraind. Deep learning heeft marginale verbeteringen aangebracht in narrow AI, maar wat we nodig hebben voor volledige AI is een wezenlijk andere gereedschapsset.

Gary Marcus, onderzoekspsycholoog aan New York University en medeoprichter van Geometric Intelligence (later overgenomen door Uber), beschrijft drie fundamentele problemen met deep learning.[15]

1. Er zullen altijd bizarre resultaten uit voortkomen, met name als er onvoldoende trainingsdata zijn. Zelfs nu AI een steeds hogere mate

van nauwkeurigheid in beeldherkenning bereikt, zien we bijvoorbeeld nog altijd foto's die worden getagd met labels die geen enkele gelijkenis met de foto vertonen, zoals te zien in figuur 2.3.

Figuur 2.3 *Voorbeeld van mislukte AI.*[16]

Figuur 2.4 *Hond of struisvogel?*

Of kijk naar figuur 2.4 hierboven. Door de afbeelding van de hond op de linker foto zo subtiel te veranderen dat het voor het menselijk oog niet waarneembaar is, misleidden onderzoekers het beste AI-programma van 2012, dat dacht dat de foto rechts een struisvogel voorstelde.[17]

2. Het is erg moeilijk om deep learning-processen te creëren. Fouten opsporen is lastig, evenals stapsgewijs reviseren en verifiëren.

3. Er is geen echte vooruitgang geboekt op het gebied van taalbegrip of causaal redeneren. Een programma kan een man en een auto op een foto herkennen, maar het zal zich niet afvragen hoe het kan dat de man de auto boven zijn hoofd houdt.

> **Let op**
> Kunstmatige intelligentie blijft nog altijd beperkt tot specifieke taken met duidelijke doelen. Voor elke toepassing is een speciaal ontworpen en getraind AI-programma nodig.

Denk er verder om dat AI in hoge mate afhankelijk is van grote hoeveelheden diverse, gelabelde data en dat AI dat wordt getraind met onvoldoende data meer fouten zal maken. We hebben al gezien dat zelfsturende auto's cruciale fouten maakten als ze in ongebruikelijke (niet getrainde) situaties navigeerden. Onze foutentolerantie bij dit soort toepassingen is buitengewoon laag.

AI vereist vaak een waardensysteem. Een zelfsturende auto moet weten dat mensen overrijden erger is dan van de weg afrijden.

Commerciële systemen moeten omzet en risicoreductie afwegen tegen klanttevredenheid

Toepassingen van AI in de geneeskunde hebben hun eigen beloften en valkuilen. Een team van Imperial College London heeft kortgeleden AI ontwikkeld dat met 80 procent nauwkeurigheid longhypertensie kan diagnosticeren, aanzienlijk hoger dan de 60 procent nauwkeurigheid van cardiologen. De toepassing van dit soort technologie brengt ook ingewikkelde vragen met zich mee, die we later zullen bespreken.[18]

Alle toepassingen hebben de afgelopen jaren het nieuws gehaald en dat zal ongetwijfeld zo blijven. In hoofdstuk 8 ga ik verder in op AI, als ik de keuze van **analytische modellen** bespreek die bij je zakelijke uitdagingen passen. Maar AI is slechts een van de vele analytische instrumenten in onze gereedschapskist, en het bereik ervan is nog beperkt. Laten we even een stap terug doen en naar het grotere plaatje

kijken van manieren waarop big data meerwaarde kunnen bieden door middel van een bredere verzameling tools in een breder scala van toepassingen.

Belangrijke punten

- AI wordt al zestig jaar actief onderzocht, maar heeft al twee winterperiodes van stagnatie doorgemaakt.
- AI gaat vaak samen met machine learning, waarbij een programma leert van voorbeelden die het krijgt in plaats van expliciete instructies te volgen.
- Big data is een natuurlijke katalysator van machine learning.
- Deep learning, een moderne verbetering van een oudere methode die bekendstaat als neurale netwerken, wordt in veel van de huidige AI-technologie gebruikt.
- AI-programma's hebben een beperkt bereik en zullen altijd niet-intuïtieve fouten maken.

Stel jezelf de volgende vragen

- Waar heb je in je organisatie grote hoeveelheden gelabelde data die zouden kunnen worden gebruikt om een machine learningprogramma te trainen, zoals patroonherkenning in plaatjes of tekst, of het voorspellen van een volgende klanthandeling op basis van eerdere handelingen?
- Als je al een AI-project in je organisatie bent begonnen, hoe verhoudt zich dan de geraamde return on investment (ROI) tot de kosten van het project?
- Als je de verwachte kans op succes vermenigvuldigt met de geraamde ROI, zou daar een getal uit moeten komen dat hoger ligt dan de geraamde kosten.