KNOWLEDGE DISCOVERY IN DATA

VAN AD HOC DATA MINING NAAR REAL-TIME PREDICTIEVE ANALYSE

Johan BLOMME

2005

www.johanblomme.net

INHOUD

Inleiding

Tegenwoordig functioneren organisaties in een zeer dynamische omgeving waarbij marktposities snel kunnen veranderen. Klanten worden kritischer, producten lijken steeds meer op elkaar en technologische ontwikkelingen volgen elkaar snel op. Het tijdig herkennen van en het snel kunnen aanpassen aan deze veranderingen is van cruciaal belang.

Als bedrijven zich moeilijk kunnen diversifiëren op basis van producten alleen, wordt het probleemoplossend vermogen een doorslaggevend verkoopsargument. Meer en meer wordt de klant het middelpunt van elke bedrijfsactiviteit. Het besef groeit dat naast verkoop ook service, informatie en advies belangrijk zijn. Klantenservice wordt een zelfstandig marketinginstrument, ja zelfs als merk naar voren geschoven. Hierdoor kunnen ondernemingen een concurrentievoordeel opbouwen of behouden. De vaststelling dat meer en meer de nadruk komt te liggen op bedrijfsactiviteiten die een concurrentievoordeel opleveren, heeft tot gevolg dat informatietechnologie (die voorheen vooral werd aangewend voor de automatisering van (deel)processen in de administratie en de productie), gebruikt wordt om relevante bedrijfsinformatie te genereren. Het succes van relationele databases heeft geleid tot een overweldigende hoeveelheid operationele en historische data waarover bedrijven kunnen beschikken. Echter het ontsluiten van informatie op basis van operationele databases levert allerlei problemen op, o.m. op het vlak van prestaties en veiligheid. Het is op dit punt dat de nieuwste ontwikkelingen op het gebied van informatietechnologie kunnen ingezet worden om de juiste informatie over markten en consumenten ter beschikking te stellen en om nieuwe vormen van klantgericht denken en handelen te ondersteunen.

Tegen de achtergrond van het toenemend belang van kennis als economische productiefactor, schetsen we in hetgeen volgt de betekenis van informatietechnologie voor gegevensanalyse. In het bijzonder blijven wij stilstaan bij de ontwikkelingen die geleid hebben tot het gebruik van toepassingen die gegroepeerd worden onder de noemer van bedrijfsintelligentie of *business intelligence*. Met deze term wordt verwezen naar de processen en technieken om strategisch bruikbare informatie en kennis aan gegevens te onttrekken. Door het inrichten van gegevenspakhuizen (*datawarehousing*) worden data uit diverse bronnen samengebracht in een centrale gegevensencyclopedie die voor de gebruiker snel en gemakkelijk toegankelijk is. M.b.t. de betekenis van databasetechnologie voor beslissingsondersteuning wordt de aandacht gericht op gebruikersgestuurde OLAP-technieken en het zoekproces dat aangeduid wordt als *data mining*.

1. Verandering als uitdaging

De afgelopen decennia ontwikkelden Westerse maatschappijen zich gestadig naar op dienstverlening gerichte samenlevingen. De wereld is in economisch opzicht één grote markt geworden. “The market place” wordt steeds meer “the market space”. Kenmerkend voor de overgang van arbeidsintensieve naar kennisintensieve economieën is het steeds toenemend belang van kennis als economische productiefactor. De explosieve ontwikkeling van de informatie- en communicatietechnologie fungeert in dit opzicht als een katalyserende factor. Veranderingen die reeds op gang waren gebracht, voltrekken zich door de stuwende kracht van nieuwe technologieën in een hoger tempo.

De veranderingen die lijken samen te komen in een beweging die leidt tot een in hoge mate door technologie gedomineerde toekomst staan niet op zichzelf en zijn verstrengeld met andere maatschappelijke transformatieprocessen. Niet alleen op economisch vlak, ook op cultureel vlak vallen grenzen weg. De toegenomen welvaart heeft zich o.m. vertaald in een culturele diversificatie met een pluriformiteit aan levensstijlen, culturele uitingen, vrijetijds- en consumptiepatronen. Deze verscheidenheid tekent zich niet alleen af op maatschappelijk niveau. Culturele verscheidenheid leidt eveneens tot een grotere verscheidenheid aan individuele identiteiten. Door het proces van individualisering zien individuen zichzelf minder als onderdeel van vaststaande collectiviteiten en meer als zelfstandige individuen met zelfgekozen sociale verbindingen en netwerken. De hier genoemde processen, en in het bijzonder de vaststelling dat economieën steeds meer gedragen worden door kennis, blijven niet zonder gevolgen voor de bedrijfsvoering.

Hoe is, onder invloed van de verschillende transformaties die onder de noemer van de kennissamenleving schuilgaan, het vak marketing het laatste decennium veranderd ? Aan het begin van de jaren negentig werd door de *Gartner Group* vooropgesteld dat het overwinnen van veranderingen de hoofdtaak van de bedrijfsvoering zal uitmaken. Kortere productlevenscyclussen, een toenemende concurrentiedynamiek en de vaststelling dat veel markten veranderd zijn van een verkoopmarkt in een koopmarkt, zijn slechts enkele factoren die aangeven dat de marktsituatie voor veel bedrijven een stuk moeilijker geworden is. De afgelopen jaren is veel geschreven over de gewijzigde omgeving waaraan de marketing zich moet aanpassen. Individualisering, demassificatie, marktversplintering, fragmentatie, graascultuur, e.a. zijn vaak gebruikte termen om de hedendaagse consument te omschrijven. In een samenleving met vrijwel onbeperkte keuzemogelijkheden en onvoorspelbare varianten wordt doelgroepselectie veel minder eenvoudig. De ongrijpbare consument. Producten gaan steeds meer op elkaar lijken (en worden gemakkelijker gekopieerd) terwijl de concurrentie heviger wordt.

Bepalend voor de veranderingen in het vak marketing is de afnemende rol van de traditionele marketing mix-instrumenten (product, prijs, promotie, plaats). Deze blijken onvoldoende houvast te bieden in snel veranderende markten. Bedrijven moeten zich in toenemende mate concentreren op de wijze waarop zij hun dienstverlening organiseren. Een goed functionerende organisatie is immers veel moeilijker te kopiëren. Kwaliteit zal in de hele (klanten)benadering kwantiteit als hoofdprioriteit verdringen. In deze context heeft Schultz (1990 : 12) het over de noodzaak van een nieuw type marketing :

“In the past, traditional marketeers, in an effort to organize their marketing activities, tried to aggregate customer wants and needs. In other words, they tried to find common products and concepts and areas which would appeal to large numbers of people who could be served economically and efficiently with a product or service. Because customers and their social norms were rather homogeneous, that management approach resulted in the development of the mass market for consumer products. … Today, there’s a new look at the marketplace. A new look to consumers. … Today, and increasingly tomorrow, the market for many consumer products is and will be driven by time, not money. It is also driven by a desire by people to be different and unique, not more like others. This dramatic upheaval in the marketplace has resulted in an equally dramatic change in marketing. … we’re moving from what was historically mass marketing to a new form of specialized marketing”.

Wat uit het bovenstaande kan afgeleid worden, is dat het meer dan ooit belangrijk wordt om de markt en de ondernemingsmissie duidelijk te definiëren. Schultz (1990 : 19) heeft het in dit opzicht over het belang van strategisch management :

“Two basic principles guide the development of any strategic marketing plan. 1. The organization’s planning view and orientation must be external, that is, toward customers and markets, not internal or

toward what has been done or what facilities or technologies are now available. 2. Strategic marketing plans must be based on identifying and making use of some sustainable competitive advantage".

De nieuwe eisen overstijgen de combinatie van product-prijs-promotie-plaats. Om in te spelen op de marktontwikkelingen moeten bedrijven zich met de hele organisatie richten op de markt. Bedrijfsprocessen dienen de klant als vertrekpunt te nemen. Het managen van klantrelaties wordt daarmee een essentieel onderdeel van marketing. Massamarketing is daarmee niet verdwenen, maar wél van sterk afnemend belang geworden. Het ultieme doel van de marketeer wordt het op individuele basis benaderen en bedienen van de klant. Dit brengt een fundamentele verandering teweeg in de marktbewerking. Het gaat niet langer (alleen) om marktaandeel (*market share*) maar (in toenemende mate) om aandeel in de bestedingen van klanten (*share of wallet*). Dit betekent dat niet zozeer beoogd wordt om zoveel mogelijk producten aan eenieder te verkopen of af te zetten binnen een bepaald marktsegment, maar om een relatie met een klant zodanig te ontwikkelen dat deze zoveel mogelijk producten van een bepaalde leverancier afneemt. Hiermee verschuift het accent van de marketingactiviteiten meer van het winnen van nieuwe klanten naar het behouden van bestaande klanten (*retention marketing*). Aangezien het vervullen van behoeften van individuele klanten centraal staat i.p.v. het verkopen van producten is het, met het oog op het kennen van de *lifetime value* van een klant, belangrijk diens behoeften nauwkeurig in kaart te brengen. Een klant kan immers beschouwd worden als een generator van een aantal transacties gedurende een bepaalde periode, en het komt erop aan de behoeften van de klant te kennen en inzicht te verwerven omtrent welke klanten (potentieel) winstgevend zijn of niet.

2. Informatietechnologie

De veranderingen aan de vraagzijde van de markt resulteren de afgelopen jaren in een steeds nadrukkelijker wordende noodzaak tot klantgerichtheid. Hierdoor zijn ook de eisen t.a.v. bedrijfsinformatiesystemen veranderd. Werden deze laatste vooral in verband gebracht met het automatiseren van bedrijfsprocessen, waarbij automatisering rond de bestaande processen werd gedrapeerd, dan wordt informatie meer en meer als een productiefactor beschouwd. In die zin krijgt informatie een strategische betekenis aangezien het een middel wordt tot ondersteuning van de besluitvorming. Het snel anticiperen op de wisselende behoeften van consumenten waardoor een bedrijf een concurrentievoordeel kan verwerven, is afhankelijk van informatie. De revoluties die zich voltrekken in de informatietechnologie, en daarmee samenhangend in de media, grijpen diep in in het marketingproces. Het is daarbij niet dat de informatietechnologie voorschrijft wat wel en niet kan. Veeleer moet informatietechnologie gezien worden als een *enabling technology* waardoor bedrijfsprocessen (zoals de klantenadministratie) aansluiten op hetgeen vanuit klantenperspectief wenselijk is (en niet omgekeerd) en besluitvormingsprocessen door informatie ondersteund worden.

De voorbije decennia werd automatisering vooral aangewend voor de ondersteuning van eerst de administratieve en daarna de overige bedrijfsprocessen. Het resultaat ervan was de snelle verwerking van gegevens en het gebruik van deze laatste voor het aansturen van de verschillende onderdelen van het productie- en distributieproces. Door de technologie kregen bedrijfsprocessen en de daaruit resulterende producten en diensten een informatiecomponent waardoor 'waarde' werd toegevoegd aan de fysieke keten. Geheel in overeenstemming met de kenmerken van een aanbodgerichte marktstrategie was de toegevoegde waarde van de informatiecomponent intern gericht en werd deze

voornamelijk aangewend voor het kwantitatief en kwalitatief optimaliseren van bestaande producten en diensten om uiteindelijk het marktaandeel te vergroten.

Door de toepassing van direct marketing technieken bleek het mogelijk om de informatiecomponent ook afzonderlijk te exploiteren. Door te sturen vanuit speciaal daartoe gebouwde marketing databases, nemen de mogelijkheden tot effectieve marktbewerking aanzienlijk toe. Aan de hand van de in een database opgeslagen informatie over prospecten en klanten, is het mogelijk om in grote markten een individuele benadering toe te passen. Aanvankelijk ging het om gepersonaliseerde massacommunicatie, waarbij eenzelfde boodschap wordt verstuurd aan individueel geadresseerde personen. Maar ook massale persoonlijke communicatie vindt thans op grote schaal plaats. Hierbij wordt de boodschap individueel afgestemd op de kenmerken die van de ontvanger bekend zijn. Met de mogelijkheden die de informatietechnologie vandaag biedt, kunnen prospecten en klanten niet alleen individueel benaderd worden, maar wordt het ook mogelijk om op massale schaal maatwerk te leveren. Daardoor kan massamarketing in consumentenmarkten vervangen worden door één-op-één marketing[1].

De individuele relatie op basis van data in marketing databases kan niet alleen tot stand gebracht worden omdat technologische mogelijkheden grootschalig datagebruik mogelijk maken, maar ook omdat deze een nieuwe dimensie scheppen voor de communicatie tussen aanbieder en afnemer. Werd in de tijd van de overheersende massamarketing een medium vooral als advertentiemedium gebruikt, dan worden media in toenemende mate gebruikt om één-op-één relaties te leggen met prospects en klanten. In de jaren negentig zijn het vooral de communicatiemogelijkheden die tot een verandering leiden[2]. De massamedia zullen in de (direct) marketing een plaatsje moeten opschuiven voor de intussen steeds drukker bereden informatiesnelweg. Niet in het minst omdat het een bij uitstek interactief medium is, waardoor de gewenste direct response vanzelfsprekend wordt. Multimedia, interactieve media, elektronische snelwegen en interactieve marketing zijn termen die steeds vaker gebruikt worden om aan te geven dat de informatietechnologie nu gebruikt wordt als een middel om op een nieuwe manier te communiceren[3]. De consument krijgt meer vormen en methoden van communicatie tot zijn beschikking en ontwikkelt zo, op basis van zijn behoeften en wensen, een eigen wijze van interactie met de omgeving. Afstand, plaats en tijd vormen geen belemmering meer waardoor de informatiecomponent een eigen leven gaat leiden parallel aan en in interactie met de fysieke waardeketen[4]. Hierdoor ook beweegt de informatiecomponent van *back office*- naar *front office*-processen. Aan de voorkant van de organisatie komt een laagdrempelige klanteninterface te staan. De oriëntatie verschuift van een product- naar een klantoriëntatie. Bedrijven moeten zich meer en meer richten op het reactieve van de consument. Zij moeten zo bereikbaar mogelijk zijn, zowel in medium als in tijd. Telematica-instrumenten zorgen voor een betere interactie met de consument en zijn daarom belangrijke instrumenten in de (elektronische) distributieketen. Door callcenters en internet ontstaan virtuele marktplaatsen waardoor gegevens over prospecten en klanten ingewonnen worden. De uitgebreide communicatie-, databeheer- en analysemogelijkheden laten toe een dieper inzicht in de markt en de eigen organisatie te verkrijgen. Dit verklaart het toegenomen belang van de klantendatabase waarop informatietechnologie met succes kan toegepast worden voor zowel het aanboren van informatie en kennis over klanten als het aansturen van marketingacties.

3. Gegevensanalyse

De besproken ontwikkelingen blijven inderdaad niet zonder gevolgen voor de databasetechnologie. De sterk gestegen aandacht voor datawarehousing (gegevenspakhuizen) en data mining (gegevensdelving) houdt verband met het toegenomen belang van informatie ter ondersteuning van beslissingen (*decision support*).

Zowel het feit dat veel gegevens van prospecten en klanten geautomatiseerd verzameld worden als de vaststelling dat operationele systemen heel wat (verborgen) informatie in zich houden, verklaren de belangstelling voor nieuwe technologieën die toelaten gegevens te ontsluiten en hieraan informatie te onttrekken die een gerichte en directe benadering van bestaande en nieuwe klanten toelaten. De ontwikkelingen op het vlak van databasetechnologie voor beslissingsondersteuning markeren in dit opzicht een omslag, een technologische verschuiving ten opzichte van het relationele model dat in de jaren tachtig opgang maakte.

Het relationele model is ontwikkeld in een periode dat databanken een andere functie in de informatievoorziening van bedrijven hadden dan tegenwoordig het geval is. Relationele database management systemen (RDBMS) richten zich bij uitstek op de ondersteuning van administratieve processen *(On Line Transaction Processing*, OLTP). Sterk samenhangend met de administratie van transactionele gegevens is ook de betekenis van normalisatieprocessen die de gegevensintegriteit van relationele databases moeten verzekeren enerzijds en de structuur van de zoektaal SQL anderzijds.

Het belang van informatie ter ondersteuning van beslissingen voert evenwel tot de conclusie dat een operationele database niet meer volstaat. Vooral het feit dat beslissingsondersteuning de toegang tot een brede verzameling gegevens vereist, is hiervan de oorzaak. Om die reden wordt de ontwikkeling van gegevenspakhuizen (datawarehousing) door velen beschouwd als de spil van de hedendaagse IT-architectuur. Terwijl transactionele gegevens (OLTP) zodanig georganiseerd zijn dat ze snel opgeslagen en opgeroepen kunnen worden, zijn de gegevens in een gegevenspakhuis georganiseerd op een wijze die uiteenlopende analyses mogelijk maken.

Het multidimensioneel modelleren van gegevens dat kenmerkend is voor OLAP *(On Line Analytical Processing*) maakt het mogelijk door aggregatie en samenvatting snel toegang te krijgen tot gegevens en deze vanuit verschillende gezichtspunten te benaderen. Naast de meer traditioneel op hypothesentoetsing geënte verificatie-technieken bieden met name explorerende technieken voor kennisontdekking (data mining) de mogelijkheid tot extractie van verborgen patronen in gegevens.

3.1. Datawarehousing

De afgelopen jaren sterk gestegen belangstelling voor datawarehousing wordt onderbouwd door de groeiende oriëntatie op klanten en, daarmee samenhangend, het strategisch belang van informatie voor besluitvorming. Als een continu proces voorziet datawarehousing in het ter beschikking stellen van gegevens om informatie voor beslissingsondersteuning te leveren. Datawarehousing is bedoeld om productie-georiënteerde gegevensbestanden voor analysedoeleinden beschikbaar te maken. De gegevens waarmee een datawarehouse wordt gevuld, zijn grotendeels afkomstig van de in de organisatie geëxploiteerde operationele/transactionele systemen, de bronsystemen. Daarnaast

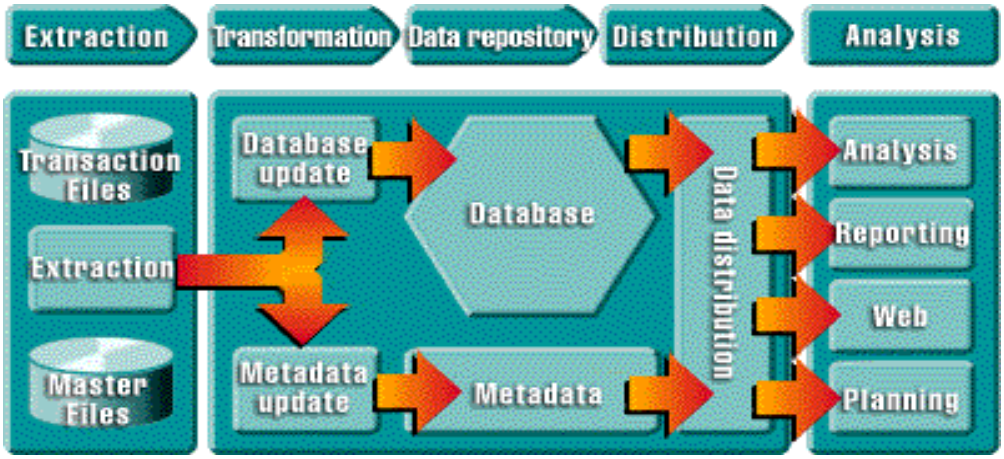
worden ook externe gegevens (bv. marktonderzoekgegevens en gegevens uit externe databanken) in een datawarehouse opgenomen. Eén van de redenen voor het opzetten van een datawarehouse is dat de operationele processen niet of zo weinig mogelijk verstoord worden door activiteiten die gericht zijn op het genereren van managementinformatie. Daarom worden queries voor dit laatste doel bij voorkeur niet rechtstreeks op de gegevens in productie-omgevingen uitgevoerd. Deze gegevens worden daarom gekopieerd naar een omgeving waar zij zonder storende werking kunnen worden geanalyseerd. Dit heeft (zoals o.m. het geval is met multidimensionele databases) het bijkomende voordeel dat de gegevens op een wijze worden opgeslagen die meer geschikt is voor dit analysewerk.

In zijn veel geciteerde studie, *Building the Data Warehouse* (1996) , bakent W.H. Inmon een gegevenspakhuis af van een relationele database door een gegevenspakhuis te definiëren als een onderwerp gerichte, gëintegreerde, tijdsgebonden en statische verzameling van gegevens ter ondersteuning van besluitvorming. Terwijl de gegevens in een operationele database transactie-georiënteerd zijn, zijn de data in een gegevenspakhuis gericht op de behoeften van de eindgebruiker en als zodanig gemodelleerd. Dit laatste houdt in dat gegevens over eenzelfde onderwerp (bv. klantgegevens) die in een OLTP-omgeving in verschillende productiesystemen opgeslagen en verwerkt worden, in een gegevenspakhuis samengebracht worden per onderwerp. De data zijn anderzijds geïntegreerd, hetgeen betekent dat gegevens die in OLTP-omgevingen vaak in een verschillend formaat worden beheerd, in een gegevenspakhuis zodanig geconsolideerd worden dat ze op eenduidige wijze te benaderen zijn. Om die reden maakt het aanleggen van gegevensdefinities (metadata, gegevens over gegevens) een cruciaal onderdeel uit van datawarehousing.

Een ander verschilpunt met gegevens opgeslagen in operationele databases is dat data in een gegevenspakhuis tijdsgebonden zijn en derhalve een historische dimensie hebben. Weerspiegelen de data in OLTP-systemen een momentopname, dan hebben analyses van data in gegevenspakhuizen vaak tot doel verschuivingen en trends op te sporen. Alleen historische data laten toe veranderingen in kaart te brengen.

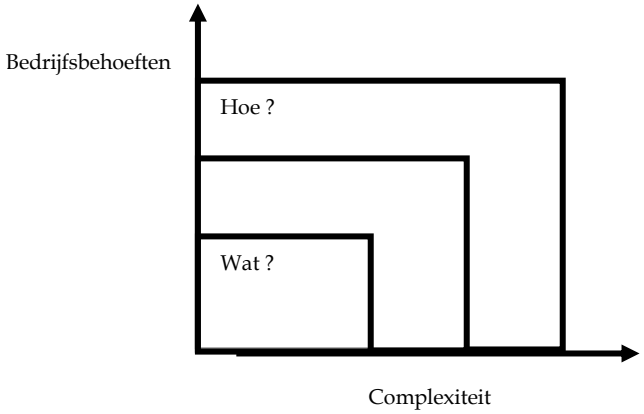
Tenslotte zijn de data in een gegevenspakhuis duurzaam (statisch). Er worden, in tegenstelling tot databases voor operationeel gebruik, geen gegevens veranderd, noch verwijderd ; er worden enkel gegevens toegevoegd. In een operationele database fungeren normalisatie- en integriteitsregels als condities voor het wijzigen, toevoegen en verwijderen van gegevens op recordniveau. Tegenover de eisen voor update-optimalisatie in RDBMS, staat query-optimalisatie centraal in een datawarehouse-omgeving.

Ten aanzien van de analyses die op de gegevens in een datawarehouse plaatsvinden, kan een tweedeling gemaakt worden tussen op verificatie gerichte analyses en analyses gericht op het ontdekken van nieuwe kennis. Het onderscheid tussen beide benaderingen kan verduidelijkt worden door de behoeften op het vlak van beslissingsondersteuning weer te geven. In de eerste plaats is het belangrijk te weten wat er in de markt gebeurt. Dergelijke wat-vragen (bv. Wat is de evolutie van het marktaandeel van de eigen producten t.o.v. concurrerende producten gedurende de afgelopen vijf jaren ?) worden beantwoord aan de hand van vraaggestuurde data-analyses, geïnduceerd door gebruikers die aan de hand van op voorhand geformuleerde hypothesen de gegevens in een datawarehouse benaderen. Zowel traditionele queries als OLAP-technieken ondersteunen deze op verificatie gebaseerde gegevensanalyse.



Figuur 1 : Datawarehousing

Behalve het antwoord op wat-vragen heeft de bedrijfsvoering ook behoefte aan analyses die een antwoord geven op de vraag waarom ontwikkelingen zich voordoen in de markt en hoe hierop kan gereageerd worden. Het antwoord op waarom- en hoe-vragen vereist inzicht in de markt en kennis van het gedrag van klanten om te kunnen voorspellen hoe deze zich in de toekomst zullen gedragen en hoe de verworven inzichten kunnen vertaald worden in een strategisch voordeel.



Figuur 2 : Vraagstellingen bij beslissingsondersteuning

In tegenstelling tot de gebruikersgerichte aanpak die centraal staat in traditionele querying en OLAP, hebben de kennisontdekkende algoritmen die onder de noemer 'data mining' worden ondergebracht een gegevensgestuurd karakter. Gegevens worden zonder vooraf geformuleerde hypothesen doorzocht op ongekende/onverwachte verbanden en patronen. Kennisontdekkende technieken werken autonoom (zonder begeleiding) en de relaties en patronen die erdoor aan de oppervlakte gebracht worden, leiden tot nieuwe inzichten en het nemen van de daarbij passende beslissingen en acties.

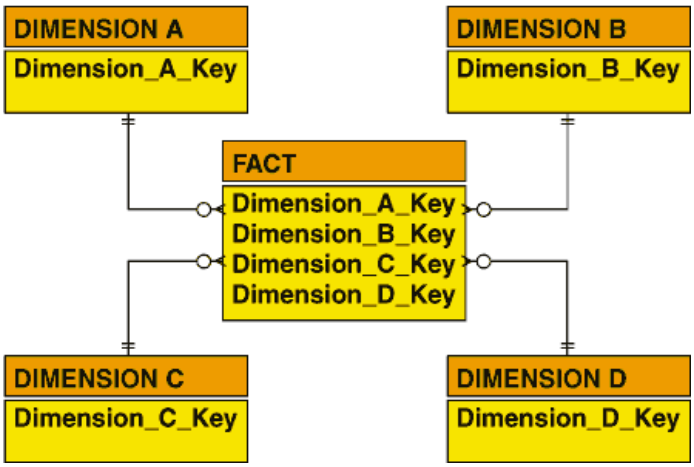
3.2. Multidimensionele gegevensanalyse

Uit de omschrijving van een gegevenspakhuis kan afgeleid worden dat de organisatie van een datawarehouse verschilt van een operationele database. Aangezien de inhoud van een operationele database door de gebruiker wordt gewijzigd, is het gegevensmodel van een operationele database gericht op het ondersteunen van transacties die vele malen dezelfde bewerkingen op gegevens uitvoeren. Om die reden ligt de nadruk bij de modellering van operationele databases op normalisatie om redundantie en verstoring van de database-integriteit te voorkomen. De onderliggende genormaliseerde structuur wordt voor de gebruiker verborgen gehouden. De door normalisatie ingebouwde beveiligingsmechanismen schermen de fysieke gegevensstructuur van een operationele database af tegen directe benadering door eindgebruikers. In het geval van het gebruik van een gegevenspakhuis wijzigt de gebruiker geen gegevens en wordt de toegang tot de gegevens gedefinieerd vanuit het gezichtspunt van de gebruiker. De ontwerpfilosofie van een datawarehouse bestaat er dan ook in een open toegang tot gegevens te bieden ter ondersteuning van een breed scala van queries. Bij de analyse van gegevens die in een datawarehouse opgeslagen liggen, neemt de multidimensionele modellering van gegevens een belangrijke plaats in.

De techniek die bekend staat als het multidimensionele stermodel sluit aan bij het uitgangspunt om de gegevens in een datawarehouse te modelleren vanuit het perspectief van de eindgebruiker. In het stermodel worden feiten zoals verkopen, facturen, betalingen, e.d. gekwalificeerd langs meerdere dimensies. Het centrum van de ster wordt de feitentabel (*fact table*) genoemd. In de (gedenormaliseerde) feitentabel worden naast de eigenschappen van het centrale object ook de verwijzende sleutels (*foreign keys*) naar de dimensietabellen bijgehouden. Dimensietabellen bevatten attributen die de dimensiewaarden beschrijven (en worden in SQL-opdrachten vaak als zoekcriteria gebruikt).

In figuur 3 wordt een voorbeeld gegeven van een eenvoudig stermodel[5]. In de feitentabel worden bijvoorbeeld gegevens opgeslagen over de verkoop van producten in een supermarktketen. In de centrale feitentabel worden naast gegevens over aantal verkochte artikelen en omzet ook de verwijzende sleutels bijgehouden naar dimensietabellen (mogelijke dimensies zijn tijd, (winkel)locatie, product en klant).

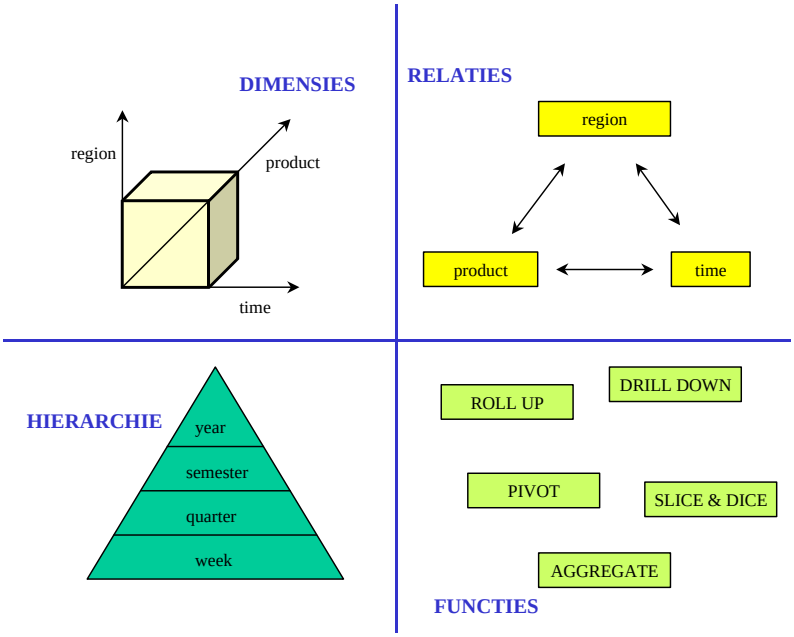
Een benadering die toelaat een potentiële feitentabel en kandidaten voor dimensies in het datawarehouse te identificeren, is het zgn. Starnet-model. Een dergelijk model geeft aan welke de verschillende dimensies (en de onderdelen ervan) zijn die bij ieder aandachtsgebeid horen. Op die manier vormt het Starnet-model een uitgangspunt voor het modelleren van het datawarehouse. Het Starnet-model is ook vanuit het standpunt van gegevensanalyse een belangrijk hulpmiddel. Het Starnet-model wordt gebruikt om de behoeften te formuleren m.b.t. de samenvoeging van gegevens ten einde het aantal te analyseren dimensies te verminderen en de mate waarin gegevens op een lager dan wel hoger aggregatieniveau geanalyseerd worden.



Figuur 3 : Weergave van een sterschema

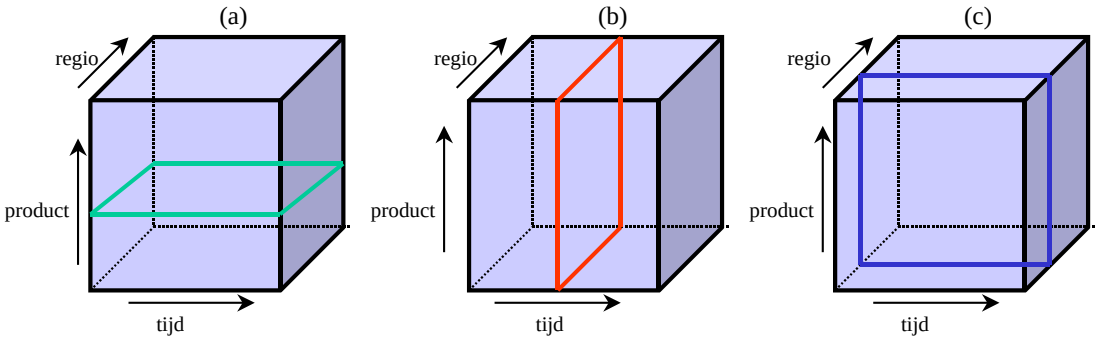
De methode die bekend staat als On Line Analytical Processing (OLAP) is gebaseerd op de multidimensionele analyse van gegevens, onafhankelijk van de wijze waarop de gegevens fysiek zijn opgeslagen. Wat dit laatste betreft, kan een onderscheid worden gemaakt tussen multidimensionele en relationele opslag. In het geval van multidimensionele opslag, die bekend staat onder de naam MOLAP, worden frequent benaderde gegevens uit het datawarehouse door voorberekening, samenvatting en aggregatie naar meerdere dimensies opgeslagen in de multidimensionele opslagcapaciteit van de OLAP-server. Deze laatste is hierbij dikwijls ingericht als een *data mart*, d.w.z. een datawarehouse met een beperkt functioneel of organisatorisch aandachtsgebied, zoals een business unit. In het geval van relationele OLAP (ROLAP) presenteert de OLAP-server de gegevens die relationeel opgeslagen liggen in een datawarehouse onder de vorm van een multidimensioneel model.

Zoals kan afgeleid worden uit het hierboven beschreven stermodel, worden bij meerdimensionele analyses meerdere gegevensdimensies onderscheiden. Binnen deze laatste kunnen vaak ook nog hiërarchieën aangebracht worden. Illustratief hiervoor is de tijdsdimensie die toelaat gegevens per dag, per maand, per kwartaal, enz. te analyseren. Het is gebruikelijk om in het geval van OLAP gegevens te aggregeren. Detailgegevens verliezen immers hun waarde in de tijd en de naar meerdere dimensies geconsolideerde gegevens bevorderen zowel de efficiënte opslag van data als snelle responstijden.



Figuur 4 : Multidimensionele gegevensanalyse

OLAP-toepassingen voorzien in een aantal navigatietechnieken. Behalve de mogelijkheid om gegevens vanuit meerdere dimensies te analyseren (*pivoting*) en hierbij meerdere gegevensdoorsnijdingen te gebruiken (*slice en dice)* wordt met *drill down* verwezen naar analyses waarbij gegevens (stapsgewijs) vanuit een hoger aggregatieniveau op een lager, meer detaillistisch niveau worden bestudeerd. Het tegenovergestelde wordt *roll up* genoemd. Een OLAP-tool voorziet in de mogelijkheid om snel gegevens te analyseren vanuit meerdere invalshoeken. Stel bijvoorbeeld dat verkoopcijfers worden geanalyseerd naar het soort product, de regio en de tijdsdimensie. Zoals hieronder afgebeeld (a) kunnen verkoopcijfers gerapporteerd worden per product, over alle regio's en alle tijdsdimensies. De afzet kan anderzijds ook beschouwd worden in een bepaalde periode, over alle producten en regio's (b). Tenslotte (c) kan de afzet in een bepaalde regio, voor alle producten en periodes, gerapporteerd worden.



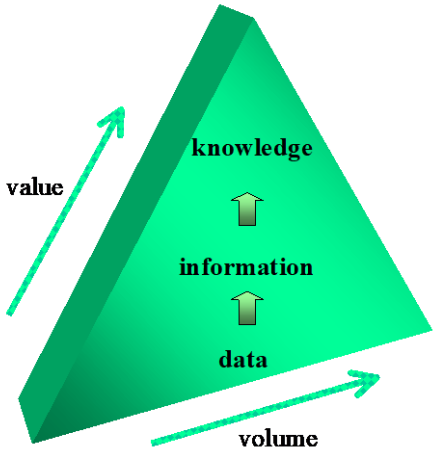
Figuur 5 : OLAP

3.3. Kennisontdekking

3.3.1. Data mining

Vaak in één adem genoemd met OLAP maar niettemin wezenlijk verschillend ervan is het zoekproces dat aangeduid wordt als data mining. In het algemeen kan data mining omschreven worden als het destilleren van onbekende informatie uit grote gegevensbestanden :

"Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (Gartner Group).



Figuur 6 : Data volume / knowledge value

Dat de aandacht voor data mining en in het bijzonder de mogelijkheden ervan om informatie aan te wenden voor het verwerven van een concurrentieel voordeel steeds meer onderkend worden, is o.m. toe te schrijven aan de convergentie van een drietal technologische ontwikkelingen. In de eerste plaats maakt datawarehousing het mogelijk om op massale basis data op te slaan en deze data toegankelijk te maken voor eindgebruikers. In combinatie met een eveneens sterk toegenomen verwerkingscapaciteit (parallelle databasetechnologie) en het gebruik van kennisontdekkende algoritmen, biedt data mining een geschikte oplossing voor de analyse van grote hoeveelheden gegevens waarvoor querymethoden minder geschikt zijn.

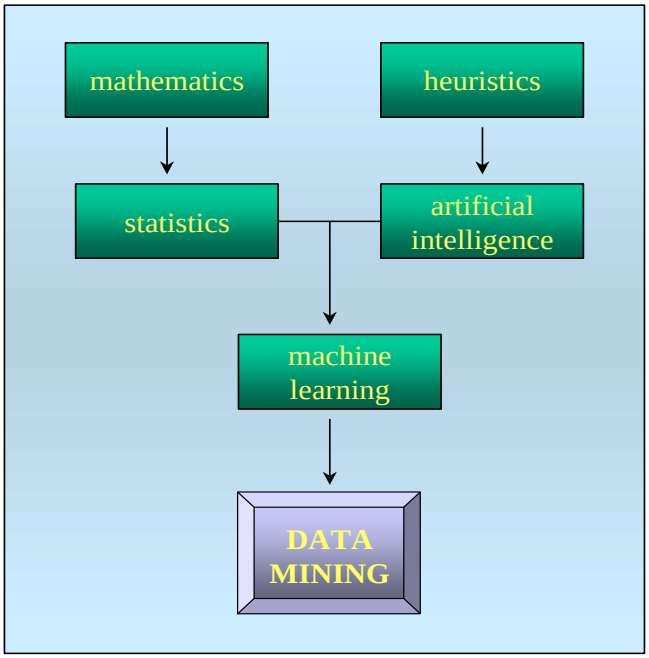
Ook data mining markeert een omslag t.o.v. het relationele model. Het in dit laatste centraal staande begrip 'sleutel' leidde tot het formuleren van verschillende normaalvormen die als leidraad dienen voor het ontwerp van relationele databases. Deze normaalvormen fungeren als *constraints*, o.m. om update- en verwijderanomalieën te vermijden (sleutels identificeren records uniek)[6]. Vanuit het perspectief van data mining gaat de aandacht niet naar het extraheren van unieke records maar naar het aantal keer dat objecten of events voorkomen. Het gaat er niet langer om functionele afhankelijkheden te gebruiken voor het ontwerpen van databases maar om ongekende

afhankelijkheden vanuit statistisch oogpunt te traceren in de gegevens[7]. Dit impliceert eveneens dat in het geval van data mining geen genormaliseerde maar gedenormaliseerde tabellen het startpunt van analyse uitmaken.

3.3.2. Data mining, statistiek en OLAP

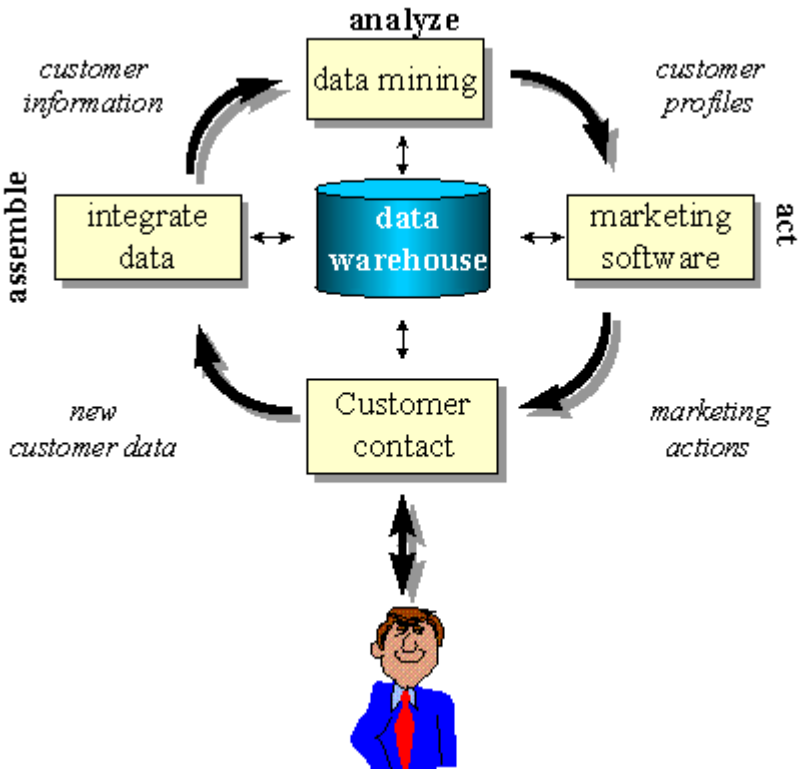
Zoals vermeld kan data mining omschreven worden als het op geautomatiseerde wijze zoeken naar patronen en verbanden in grote gegevensverzamelingen. In dit opzicht wordt bij data mining gebruik gemaakt van de inzichten uit de statistiek, de kunstmatige intelligentie en machine-leren. De meeste data mining-technieken ontlenen aan de statistiek het begrippenkader en de methoden aan de hand waarvan variabelen en relaties tussen variabelen geanalyseerd kunnen worden (bv. standaardafwijking, variantie,betrouwbaarheidsintervallen, e.a.). Terwijl steekproefonderzoek de basis vormt voor statistische analyse (*inferences from data*, inferentiële statistiek) worden data mining-technieken aangewend om patronen op te sporen in meestal volledige gegevensverzamelingen die veelvouden van gigabytes groot kunnen zijn (VLDB, *very large databases*). Een hiermee verbonden verschilpunt is dat de analist-gebruiker van statistische technieken verondersteld wordt een idee te hebben van de vorm van het model dat zal gebouwd worden, en dus een inzicht heeft in zowel de variabelen die hiertoe zullen gebruikt worden als de combinaties tussen deze laatste. Aan de basis van de analyse van VLDB ligt daarentegen juist de doelstelling om (langs exploratieve weg) ongekende verbanden en patronen aan de oppervlakte te brengen. Ook omdat traditionele statistische technieken vaak niet goed kunnen omgaan met zeer grote gegevensbestanden, de verwerking van duizenden velden hiermee niet mogelijk is en deze technieken gevoelig zijn voor afwijkende gevallen (*outliers*) in zeer grote databestanden, wordt in het geval van data mining gebruik gemaakt van zgn. adaptieve technieken die – zonder tussenkomst van de gebruiker – toelaten te achterhalen welke variabelen invloedrijk zijn en welke de belangrijke combinaties zijn.

Adaptieve technieken vinden hun oorsprong in het domein van de kunstmatige intelligentie (*artificial intelligence*), een onderzoeksdomein waarin onderzocht wordt hoe de menselijke denkwijze en het leerproces door machines kan gereproduceerd worden. Door de enorme rekenkracht die toepassingen van AI opeisen, bleven commerciële successen ervan uit. Aan het eind van de jaren tachtig, toen de prijs-prestatie verhouding van computers gunstiger werd, werden de technieken die ten grondslag liggen van AI geïmplementeerd binnen het gebied van het machine-leren. Dit laatste kan beschouwd worden als een voortzetting van de inzichten uit het domein van de kunstmatige intelligentie waarbij grondbeginselen uit de statistiek ingebouwd worden in geavanceerde algoritmen voor patroonherkenning.



Figuur 7 : Data mining

Aan de basis van data mining ligt de toepassing van een geheel van kennisontdekkende algoritmen. Om die reden wordt data mining niet zelden vereenzelvigd met *Knowlegde Discovery in Databases* (KDD), waarvan het evenwel een onderdeel vormt (zie figuur 8). KDD is een iteratief proces dat begint met het bepalen van de te beantwoorden vraagstelling. Vervolgens vindt gegevensselectie plaats uit een datawarehouse maar ook databases met transactiegegevens kunnen als bronbestanden dienen (*assemble*). Aangezien de kwaliteit van de gegevens van essentieel belang is, worden de gegevens veelal in meerdere stappen voorbewerkt (opschonen, transformatie, datatype-conversies, aggregatie, denormalisatie). Nadat de gegevens geconsolideerd zijn, kunnen ze door een (geschikte) techniek onderzocht worden op trends en patronen (*analyze*). Deze dienen eerst geïnterpreteerd en op de bruikbaarheid ervan gevalideerd te worden. Op basis van de nieuw aangeboorde informatie worden o.m. productverbeteringen en klantprofielen ontwikkeld die in een volgende fase tot marketingacties leiden (*act*). Deze laatste resulteren op hun beurt in klantcontacten die nieuwe klantgegevens genereren die terug in de centrale gegevensbank geïntegreerd worden waar ze o.m. gebruikt worden voor het bijsturen van modellen en profielen waarop marketingacties gebaseerd zijn[8].

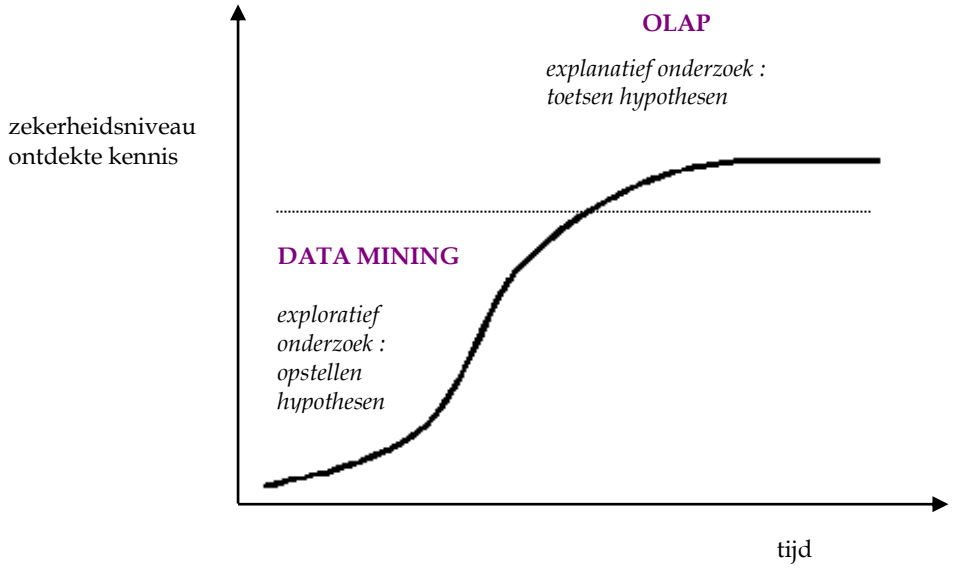


Figuur 8 : Data mining als een onderdeel van KDD[9]

In vergelijking met OLAP en traditionele statistische technieken waarbij het beantwoorden van tevoren gestelde vragen een zeker inzicht in de gegevensstructuur veronderstelt – en de analyse derhalve gebruikersgestuurd is – wordt het zoekproces naar onderliggende patronen en verbanden in het geval van data mining uitgevoerd zonder veel bemoeienis van de gebruiker. De analyse is gegevensgestuurd. Een belangrijke randvoorwaarde hierbij is dat de gebruiker voldoende domeinkennis heeft van het bedrijfsproces dat geanalyseerd wordt. De gebruiker moet in staat zijn om voor de te beantwoorden vraagstelling de relevante gegevens te verzamelen. Deze gegevens dienen veelal eerst bewerkt te worden om ze geschikt te maken voor analyse. Pas daarna vindt analyse plaats, waarbij voor de interpretatie van de resultaten eveneens deskundigheid op het toepassingsgebied een vereiste is.

Bij data mining wordt een inductieve werkwijze gevolgd die ook bekend staat als exploratieve data-analyse (EDA)[10]. Het verschil met de hypothetisch-deductieve methode die gevolgd wordt voor de toetsing van verbanden kan verduidelijkt worden aan de hand van een empirische cyclus. In een eerste fase van deze cyclus, waartoe ook data mining kan gerekend worden, worden waarnemingen en verbanden tussen waarnemingen onder een gemeenschappelijke noemer geplaatst. Dit proces wordt inductie genoemd. Het resultaat ervan is een theorie, d.w.z. een geheel van uitspraken waarvan het geldigheidskarakter evenwel nog niet vaststaat. Het afleiden van hypothesen uit de abstracte gedeelten van de theorie wordt deductie genoemd. Ondanks de verschillen tussen data mining en

OLAP zijn beide ook complementair aan elkaar. De hypothesen die gegenereerd worden uit het data mining- proces kunnen met behulp van OLAP-tools geverifieerd worden.



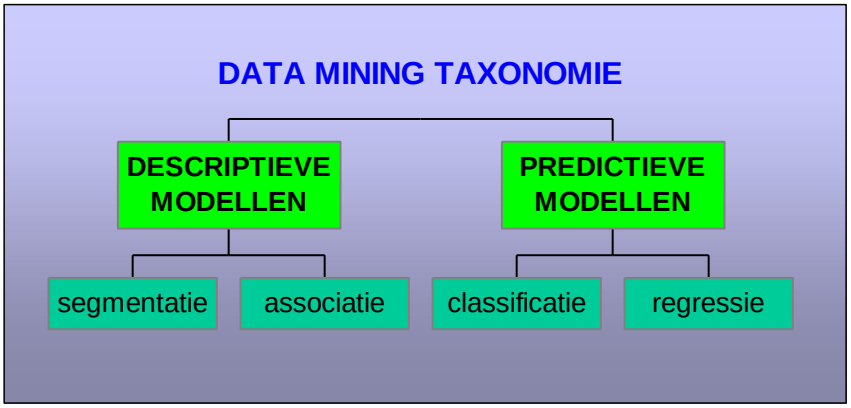
Figuur 9 : Empirische onderzoekscyclus

Een belangrijk verschil tussen data mining en exploratieve data-analyse is dat de doelstelling van data mining er niet in bestaat de interrelaties tussen variabelen inzichtelijk te maken of te verklaren. Veeleer is het de bedoeling op basis van het samenspel van variabelen tot een predictie te komen. Vandaar ook dat heel wat toepassingen van data mining gebaseerd zijn op scoringsmodellen. Scoringsmodellen zijn voorspellende modellen. Het doel ervan is een afhankelijke variabele te voorspellen aan de hand van een aantal onafhankelijke variabelen. Op grond van deze laatste kenmerken kunnen klantscores bepaald worden die een voorspellende waarde hebben. Zo kunnen klantprofielen opgesteld worden die de waarschijnlijkheid weergeven dat een prospect of klant reageert op een mailing of een product aankoopt. Banken en verzekeringsmaatschappijen maken o.m. gebruik van credit-scoringsmodellen om de kredietwaardigheid van hun klanten te beoordelen.

3.3.3. Patroonherkenning

Hoewel data mining wordt toegepast voor het beantwoorden van een breed scala aan vraagstukken, hebben vrijwel alle toepassingen gemeenschappelijk dat gezocht wordt naar patroonherkenning. De methoden die hierbij gebruikt worden, kunnen in een viertal categorieën onderverdeeld worden, nl. segmentatie (clustering), associatie, classificatie en regressie. Het onderscheid tussen deze methoden houdt in de eerste plaats verband met het al of niet aanwezig zijn van een opdeling tussen afhankelijke (te verklaren) en onafhankelijke (verklarende) variabelen (zie figuur 11, blz. 20).

In het geval van de toepassing van segmentatie en associatieve technieken wordt geen onderscheid gemaakt tussen afhankelijke en onafhankelijke variabelen. Het doel van de toepassing van interdependentietechnieken zoals deze ook genoemd worden, is het opsporen van relaties tussen variabelen om een mogelijk aanwezige, maar nog niet bekende structuur in de gegevens te ontdekken. Segmentatie-analyses worden aangewend om relaties te identificeren tussen gegevens ten einde groepen te kunnen onderscheiden. Een vaak gebruikte techniek hierbij is clusteranalyse, die tot doel heeft groepen (segmenten) te construeren op basis van kenmerken. Elementen die deel uitmaken van een groep zijn zo homogeen mogelijk ten aanzien van de beschrijvende kenmerken terwijl tussen de groepen een maximale heterogeniteit naar deze kenmerken bestaat. Clusteranalyse is o.m. een geschikte techniek voor het afbakenen van doelgroepen en het opstellen van klantprofielen.



Figuur 10 : Data mining taxonomie

Associatie- of affiniteitsanalyses worden gebruikt om te bepalen welke kenmerken of gebeurtenissen in samenhang voorkomen. Op grond hiervan worden regels opgesteld die deze samenhangen beschrijven. Dergelijke regelinductietechnieken worden veel toegepast in het geval van *market basket*-analyses, waarbij onderzocht wordt welke producten door klanten in samenhang worden afgenomen. Regelinductietechnieken worden ook gebruikt voor *cross-selling* analyses. Aan de hand van gegevens over historisch klantgedrag wordt bepaald welke combinaties van klantkenmerken en producten leiden tot interesses voor andere producten. Deze profielen kunnen o.m. gebruikt worden om tijdens klantcontacten gerichte aanbiedingen te doen.

Een belangrijke reden voor het succes van data mining ligt in het feit dat de patronen die erdoor aan de oppervlakte gebracht worden, voorspellingen toelaten in consumentengedrag. Veel toepassingen van data mining hebben dan ook betrekking op het ontwerpen van predictieve modellen. Dependentietechnieken hebben tot doel de invloed van één of meerdere onafhankelijke variabelen (predictoren) na te gaan op een afhankelijke variabele (criteriumvariabele). Met predictieve modellen wordt beoogd ofwel het behoren tot een klasse (classificatie) ofwel een waarde te voorspellen (regressie) (zie hierover Brand & Gerritsen, 1998b). In het geval van classificatie is de klasse een categoriale variabele die uit twee of meerdere elkaar uitsluitende categorieën bestaat. Scoringsmodellen die voorzien in de predictie van de respons op een mailing, voorspellen het behoren van prospecten of klanten tot de klasse “ja” of “nee”. Op analoge wijze kunnen klanten in kaart

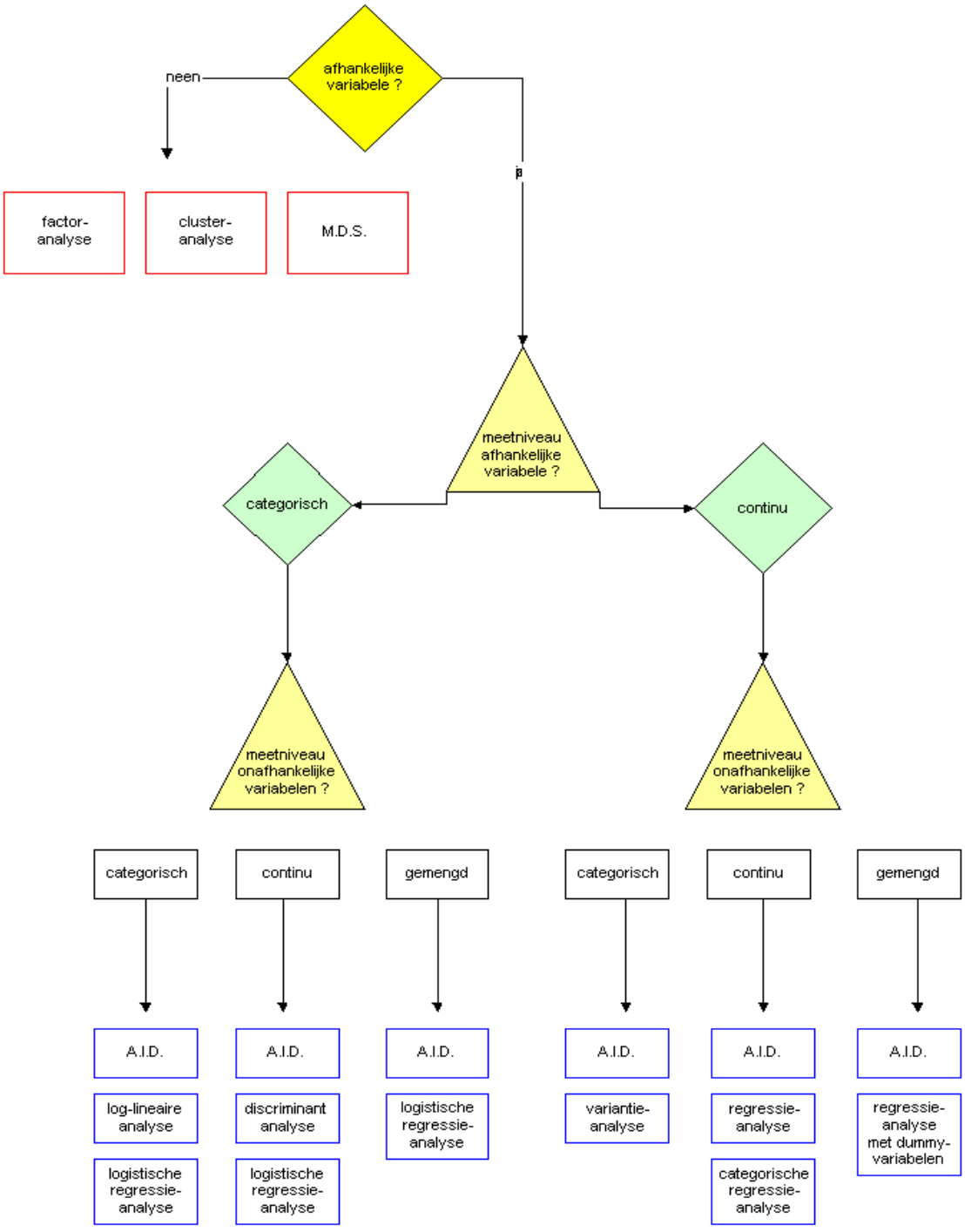
gebracht worden die de grootste kans hebben om binnen een bepaalde termijn te vertrekken (*attrition, churning*). Door de klanten die binnen dit profiel passen extra aandacht te geven kan het verloop teruggedrongen worden. Regressie wordt toegepast in het geval de te verklaren waarde (de afhankelijke variabele) een veelheid aan (numerieke) waarden kan aannemen (continue variabele). Een voorbeeld hiervan is het voorspellen van de beursnotering van aandelen.

3.3.4. Data mining-technieken

Sinds geruime tijd worden statistische technieken gebruikt om patronen in gegevensverzamelingen op te sporen. Veel gebruikte technieken zijn clusteranalyse, factoranalyse en regressie-analyse. Zoals vermeld, wordt clusteranalyse toegepast om segmenten of profielen te construeren aan de hand van kenmerken. Een voorbeeld hiervan is het opdelen van een klantenbestand in groepen en het beschrijven van de profielen van deze klantgroepen in termen van socio-demografische kenmerken, lifestyle-gegevens en aankooppatronen.

Ook factoranalyse is een exploratief-beschrijvende techniek om de dimensionaliteit in de data te analyseren. Het doel van factoranalyse is een hoeveelheid variabelen samen te vatten in een kleiner aantal onderliggende dimensies, die alle een lineaire combinatie zijn van de oorspronkelijke variabelen. Factoranalyse biedt o.m. de mogelijkheid om onderling sterk samenhangende variabelen te reduceren, hetgeen de interpretatie van de onderzoeksbevindingen ten goede komt.

Regressie-analyse is een predictieve techniek die gebruik maakt van het optimaliseringsprincipe dat bekend staat als de methode van de kleinste kwadraten. De bedoeling hiervan is de waarde van een continue variabele te voorspellen aan de hand van een lineaire combinatie van onafhankelijke variabelen. Het verschil tussen de verwachte en geobserveerde waarden van de afhankelijke variabele geldt als criterium voor de beoordeling van het regressiemodel. Een bezwaar dat aan de toepassing van regressie-analyse kleeft, is dat de gegevens die ermee gemodelleerd worden vaak niet voldoen aan de lineariteitsassumptie van de techniek. Het oplossen van de problemen die hiermee gepaard gaan, vereist statistische expertise. Bovendien blijken heel wat gegevens in de marketingpraktijk eerder van categoriale dan van parametrische aard te zijn. Voor het induceren van modellen uit grote gegevensverzamelingen worden om die reden technieken toegepast die voorzien in de niet-lineaire analyse van variabelen[11]. In hetgeen volgt geven we van deze laatste een overzicht.



Figuur 11 : Symmetrische en asymmetrische analysetechnieken[12]

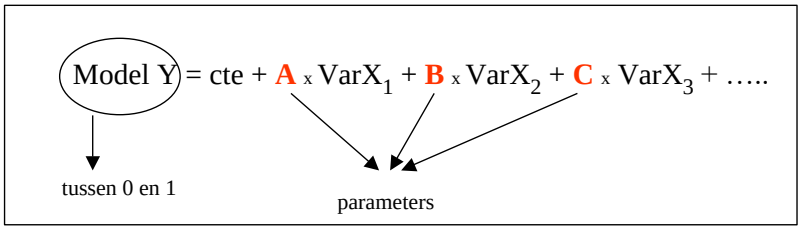
3.3.4.1. Logistische regressie-analyse

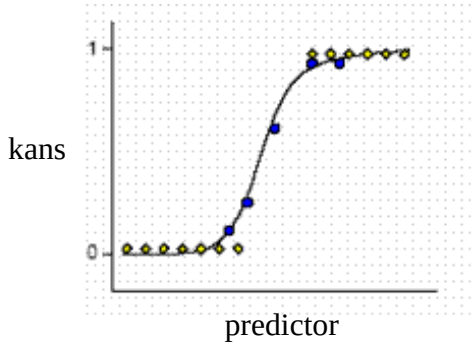
Een tegenwoordig veel gebruikte classificatiemethode is logistische regressie. Deze techniek wordt toegepast voor het voorspellen van categoriale variabelen. Behalve de toepassing ervan bij responsanalyse, wordt de techniek van logistische regressie door bank- en verzekeringsinstellingen vaak aangewend voor het opstellen van modellen voor kredietacceptatie (*credit scoring*). De doelstelling hierbij is het optimaliseren van het percentage geaccepteerde aanvragen zodat het maximaal toegestane infectiepercentage niet wordt overschreden. Aan de hand van klantgegevens (geslacht, leeftijd, beroep, e.d.), het product (doorlopend krediet, persoonlijke lening) en gegevens van in het verleden verstrekte kredieten waarbij voor ieder krediet ook de afloop is vastgelegd, stelt het algoritme van logistische regressie de gebruiker in staat om profielen te ontdekken met een sterk verlaagde of verhoogde kans op wanbetaling. De kansen op wanbetaling worden afgeleid uit het logistische model dat de vorm aanneemt van een regressievergelijking. Door de waarden van de variabelen die in het model opgenomen zijn, te wegen overeenkomstig de parameters van de regressievergelijking worden logitscores bekomen. De logaritmische transformatie van deze laatste laat toe logitscores te vertalen in waarschijnlijkheden. De resultaten van een logistische regressie-analyse kunnen teruggeschreven worden naar de prospecten- of klantendatabase om deze te verrijken, waardoor voor iedere prospect of klant op basis van zijn/haar karakteristieken een kans op wanbetaling wordt berekend.

Nemen we het voorbeeld van de opbouw van een model ter voorspelling van de respons op een direct marketing-actie. Stel dat een bedrijf op basis van steekproefonderzoek over gegevens beschikt van geslacht, leeftijd en de reactie van respondenten op een eerder toegestuurde mail. Op basis hiervan kan een logistisch model opgesteld worden. Veronderstel dat onderstaande schattingen voor de twee predictoren (geslacht en leeftijd) van respons bekomen worden :

$$R = -10{,}83 + 2.30(\text{geslacht}) + 0.28(\text{leeftijd})$$

Voor een mannelijke klant van 45 jaar oud krijgen we op die manier een logitscore van R = -10.83 + (2.30 x 0) + (0.28 x 45) = 1,77. Om betekenisvolle conclusies uit de logitscore van 1,77 af te leiden, is het noodzakelijk deze waarde om te zetten in een responswaarschijnlijkheid. Met een logit van 1,77 correspondeert een responswaarschijnlijkheid van $P = e^{1,77}/1 + e^{1,77} = 0.85$, hetgeen betekent dat voor een vijfenveertigjarige mannelijke klant een kans van 85 % responswaarschijnlijkheid bestaat.





Figuur 12 : Responsmodellering door logistische regressie-analyse

Een nadeel van logistische regressie is dat het opbouwen van een model sterk gebruikersgestuurd is. De kwaliteit van het uiteindelijke model hangt sterk af van de inhoudsdeskundigheid van de onderzoeker. Het is aan de onderzoeker om vast te stellen welke variabelen in aanmerking komen om in het model te worden opgenomen. Zoals dit het geval is bij lineaire regressie-analyse dienen interacties tussen variabelen door de gebruiker opgespoord en in de regressie-vergelijking opgenomen te worden. Een zwak punt is ook de beperkte schaalbaarheid van de techniek. Bij een toenemend aantal predictoren gaat de kwaliteit van de modellen achteruit en wordt het eveneens aan de gebruiker overgelaten om de problemen die daarmee gepaard gaan (bv. multicollineariteit) te ondervangen.

3.3.4.2. Beslissingsbomen

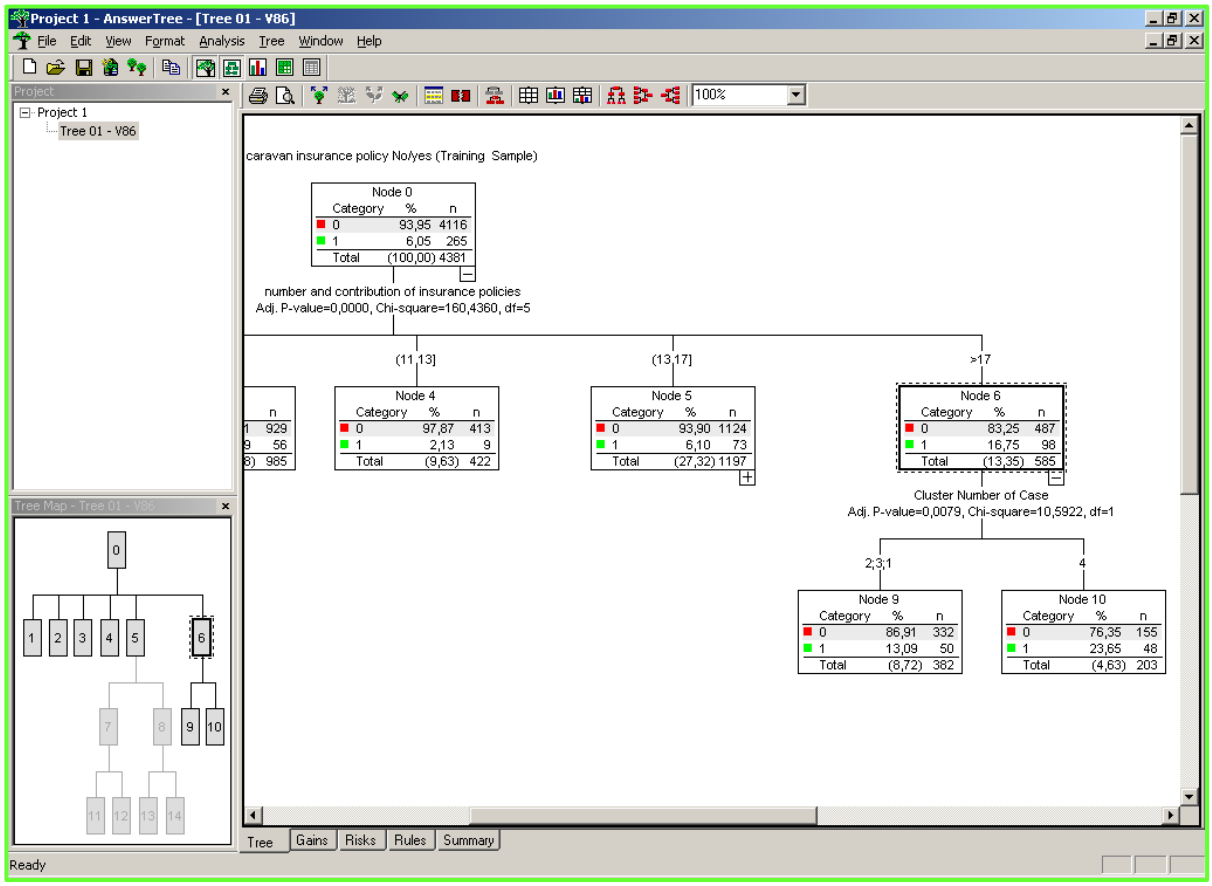
Veel minder gebruikersafhankelijk is het regelinductiemechanisme van beslissingsbomen. Een beslissingsboom is te vergelijken met een scoringsmodel waarbij de inputvariabelen (predictoren) gebruikt worden om waarnemingen te splitsen in verschillende groepen die discrimineren naar de te verklaren variabele (doelvariabele). Het gegevensbestand wordt opgedeeld in elkaar uitsluitende segmenten op basis van een doelvariabele. Deze segmentatie komt tot stand door na te gaan welke variabele en welke samenvoeging van waarden binnen deze variabele het hoogst mogelijk onderscheidend vermogen heeft t.a.v. de doelvariabele. Dit proces wordt iteratief toegepast totdat er geen onderverdeling meer te maken is die voldoende onderscheidend vermogen bezit.

In tegenstelling tot regressie-analyse, worden bij de toepassing van boomstructuren geen vooronderstellingen gemaakt over de functionele vorm van de relatie tussen de onafhankelijke variabelen en de afhankelijke variabele. Een voordeel van beslissingsbomen is dat de meest belangrijke variabelen en de combinaties (en interacties) ertussen voor de voorspelling van de doelvariabele gedetecteerd worden[13]. Op basis van beslissingsbomen worden regels opgesteld die aangeven aan welke voorwaarden een waarneming dient te voldoen om in een bepaald segment terecht te komen. Het segment geeft de voorspellende waarde aan voor de waarnemingen die

daarbinnen vallen (te bedenken valt evenwel dat met boomstructuren enkel groepsvoorspellingen en geen individuele prognoses kunnen gemaakt worden).

Responsvoorspelling is met voorsprong de meest populaire toepassing van data mining voor direct marketing. Als illustratie nemen we het voorbeeld van een case waarbij het de bedoeling was op kostenefficiënte wijze de markt voor een verzekeringspolis te vergroten (Van Der Putten & Den Uyl, 2000). De doelstelling was een voorspellingsmodel te bouwen dat nieuwe klanten kan identificeren onder de bestaande klanten van een verzekeraar. Om kansrijke prospects te kunnen selecteren werd een model gebouwd om de variabele 'bezit van een caravanpolis' te voorspellen. Hierbij werd gebruik gemaakt van interne, bedrijfseigen variabelen over het productportfolio van de individuele gebruiker enerzijds en externe socio-demografische gegevens die op postcode-niveau verzameld werden anderzijds. Een random steekproef werd getrokken uit het klantenbestand en aan de hand van de variabelen werd een voorspellingsmodel opgesteld. De interne variabelen kregen doorgaans een zwaarder gewicht (hetgeen voor de hand ligt aangezien toekomstig gedrag zich beter laat voorspellen vanuit historisch gedrag dan vanuit socio-demografische kenmerken). Het voorspellingsmodel werd ook getest op een tweede steekproef, een testverzameling. Het model moet immers niet slechts goede resultaten geven op de groep klanten waarmee het is getraind maar moet dit ook doen voor nieuwe klanten.

De dataset zoals beschreven door Van Der Putten en Den Uyl is door ons geanalyseerd met behulp van SPSS-CHAID. De beschikbare gegevens van 5822 klanten zijn opgedeeld in een training- en testset van resp. 70 % en 25 % van de klanten. In figuur 13 is een gedeelte van de beslissingsboom voor de trainingset weergegeven. Hieruit kan o.m. opgemaakt worden dat terwijl de totale respons 6,1 % bedraagt, dit stijgt naar 23,6 % door gebruikmaking van de beschikbare informatie in het klantenbestand.



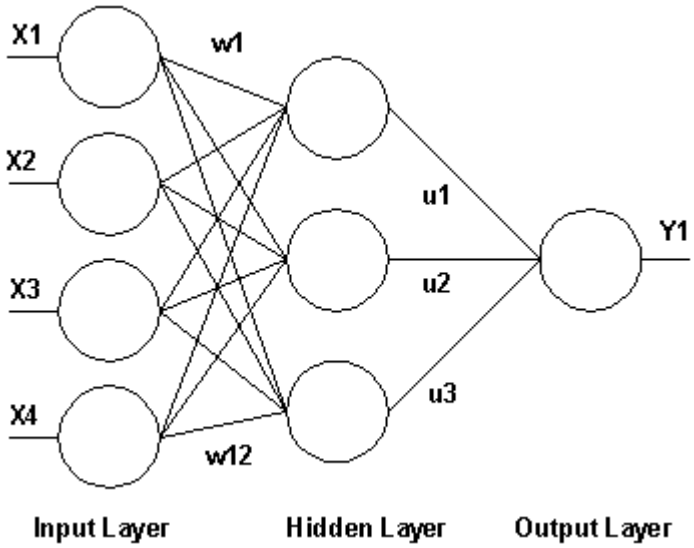
Figuur 13 : Beslissingsboom trainingset (eigen bewerking dataset
Van Der Putten & Den Uyl,2000 ; analyse uitgevoerd met behulp van SPSS Anwer Tree®)

De visualisering van de resultaten en de hoge interpreteerbaarheid van beslissingsbomen hebben ertoe geleid dat deze techniek tegenwoordig veel gebruikt wordt. Zowel voor classificatie als voor regressie is de techniek van beslissingsbomen een aangewezen analysemiddel[14]. Maar ook als hulpmiddel voor het exploreren van gegevens bewijst deze techniek zijn nut, o.m. vanwege de mogelijkheid ervan interacties tussen de verklarende variabelen op te sporen om deze naderhand als input te gebruiken voor logistische regressie-analyse.

3.3.4.3. Neurale netwerken

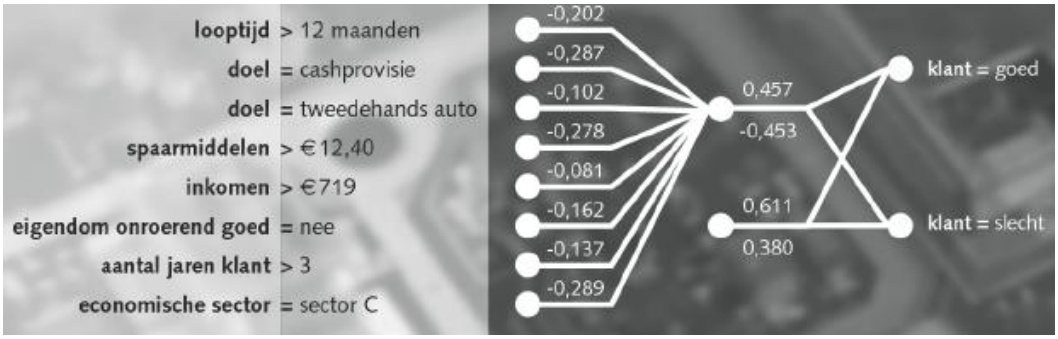
Een techniek met een breed toepassingsgebied is neurale netwerken. Omdat neurale netwerken niet-lineaire relaties modelleren en hierbij een veelheid aan variabelen kunnen gebruiken, zijn de voorspellingen via deze methode in de regel accurater dan bij traditionele regressie-analyse. Een neuraal netwerk is een input-output model . De inputvariabelen zijn aan de outputvariabele

gekoppeld door één of meer verborgen lagen (*hidden layers*). In elke laag bevinden zich neuronen. De neuronen staan in contact met elkaar via verbindingen. Deze hebben een bepaalde sterkte, ook gewicht genoemd. Het neuraal netwerk berekent voor een gegeven input een voorspelling van de afhankelijke variabele. Voor het trainen van het netwerk wordt de dataset gesplitst in een trainingset en een testset. Tijdens de eerste iteratie worden de waarnemingen uit de trainingset aan het netwerk aangeboden. Per waarneming wordt de netwerk-output vergeleken met de gewenste waarde. Op basis van het verschil worden de gewichten aangepast (*backpropagation*). Bij een volgende iteratie zal het netwerk een kleinere fout maken bij dezelfde input. De training stopt zodra de gemiddelde fout niet meer afneemt. Nadat het model voor de trainingset is ontwikkeld, wordt het verder gevalideerd door de toepassing van de modelparameters op de waarnemingen van de testset[15].



Figuur 14 : Neuraal netwerk

Een gekend toepassingsdomein van neutrale netwerken is het opstellen van modellen die de kredietwaardigheid van toekomstige klanten voorspellen. Gebaseerd op de kenmerken en het terugbetalingsgedrag van klanten uit het verleden wordt betracht modellen te construeren die de kans op succesvolle terugbetaling van potentiële klanten zo nauwkeurig mogelijk berekenen (*credit scoring*). In onderstaande figuur worden de resultaten weergegeven van een modelschatting voor het voorspellen van de kredietwaardigheid (Baesens e.a., 2003).



Figuur 15 : Neuraal netwerk voor het voorspellen van kredietwaardigheid (Baesens e.a., 2003)

Een voor de hand liggend criterium bij de beoordeling van een scoringmodel is de nauwkeurigheid van het ontwikkelde model. De accuraatheid van het model uitgedrukt in termen van het aantal correct geclassificeerde klanten in een onafhankelijke testset is uiteraard belangrijk. Toch kleven een aantal tekortkomingen aan dit criterium. Slechts een klein aantal klanten zal wanbetaler zijn, hetgeen ervoor zorgt dat een weinig informatieve regel zoals "elke klant is een goede klant" reeds een goede prestatie oplevert. Er moeten derhalve ook andere factoren, zoals de kosten verbonden aan misclassificatie, in beschouwing genomen te worden. Dergelijke kosten zijn doorgaans moeilijk te kwantificeren. Een ander aandachtspunt is het gegeven dat de geëxtraheerde modellen bij toepassing van neurale netwerken wel accuraat maar moeilijk interpreteerbaar zijn. Het black box-karakter van neutrale netwerken is een factor die een terughoudendheid tegenover het gebruik ervan in de hand werkt. Om die reden worden de modellen bekomen door toepassing van neutrale netwerken tegenwoordig vaak aangevuld met regelextractiemethoden. In onderstaande figuur worden de "als-dan"-regels die geëxtraheerd werden uit het netwerk zoals voorgesteld in figuur 15 gevisualiseerd onder de vorm van een beslissingstabel die eenvoudig te interpreteren is.

als	dan
als looptijd > 12 maanden **en** doel = cashprovisie **en** spaarmiddelen ≤ € 12,40 **en** aantal jaren klant ≤ 3	**dan** klant = slecht
als looptijd > 12 maanden **en** doel = cashprovisie **en** eigendom onroerend goed = nee **en** spaarmiddelen ≤ € 12,40	**dan** klant = slecht
als doel = cashprovisie **en** inkomen > € 719 **en** eigendom onroerend goed = nee **en** spaarmiddelen ≤ € 12,40 **en** aantal jaren klant ≤ 3	**dan** klant = slecht
als doel = tweedehands auto **en** inkomen > € 719 **en** eigendom onroerend goed = nee **en** spaarmiddelen ≤ € 12,40 **en** aantal jaren klant ≤ 3	**dan** klant = slecht
als spaarmiddelen ≤ € 12,40 **en** economische sector = sector C	**dan** klant = slecht
default klasse: klant = goed	

Figuur 16 : "Als-dan" regels geëxtraheerd uit het neuraal netwerk in figuur 15 (Baesens, 2003)

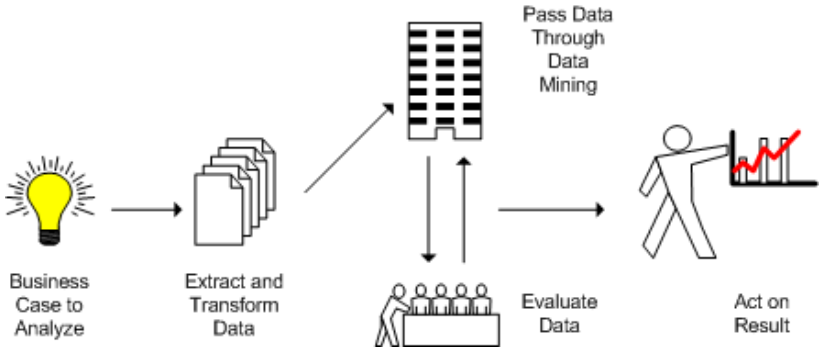
De hier besproken technieken blijken doorgaans onderling weinig te verschillen in piekprestatie. Verschillen tekenen zich wel af m.b.t. snelheid, gebruiksgemak en inzichtelijkheid.

Gesteld kan worden dat technieken die voorzien in de niet-lineaire analyse van variabelen, en met name technieken van regelinductie en neurale netwerken, doorgaans beter in staat zijn tot patroonherkenning als het gaat om interne bestanden met omvangrijke gegevens over klanten en de historiek van klantbestedingen. Naarmate de transitie van *list-based* marketing naar *customer-based* marketing zich verder aftekent, zullen niet alleen klantgegevens (verzameling gegevens over klanten) maar vooral klantmodellen (weergaven van kenmerken en behoeften van klanten) centraal komen te staan. Voor het opbouwen en onderhouden van klantmodellen moeten analysetaken met een hoge frequentie uitgevoerd worden. Het ligt daarom in de lijn van de verwachtingen dat een deel van de analysetaken zal overgelaten worden aan zelflerende, adaptieve technieken[16].

3.3.5. Data mining : Een iteratief proces

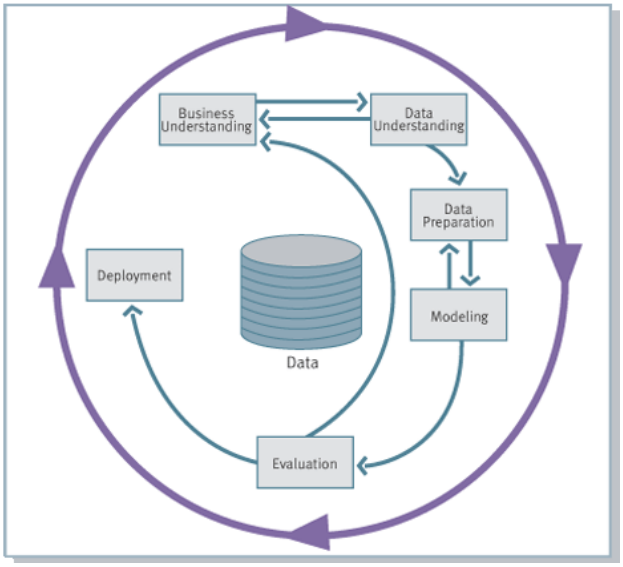
Data mining heeft veel meer dan enkel met technische kwesties te maken. De opvatting als zou data mining een automatisch proces zijn, is allerminst juist (er kan alleen gesteld worden dat data mining-algoritmen zelf automatisch verlopen voor zover er geen assumpties gelden m.b.t. de behandelde data zoals normaliteit, lineariteit, e.a.). In het geval van data mining komt het er in de eerste plaats op aan een beleidsprobleem te vertalen naar een onderzoekbaar probleem. Hiermee wordt afgebakend *wat* men *waarom* wil analyseren. In veel publicaties wordt de illusie gewekt dat deze taak automatisch kan worden uitgevoerd door data mining-systemen. Het definiëren van een 'onderzoekbaar' probleem blijft een creatief proces waarbij vooral inzicht en ervaring een belangrijke rol spelen. De grote kracht ligt daarbij in de beperking en de juiste afbakening van het probleem. Dit veronderstelt de nodige achtergrondkennis van de organisatie en de processen die daarbinnen plaatsvinden.

Wanneer het eenmaal duidelijk is wat men waarom wil onderzoeken, kan begonnen worden met het inventariseren welke gegevens noodzakelijk zijn om de vragen te beantwoorden. Eenmaal de gegevens samengevoegd zijn tot een werkbestand, dienen deze nog te worden gecontroleerd, getransformeerd en bewerkt (bv. dubbele records, verouderde gegevens, afwijkende codes, ontbrekende waarden). Vervolgens is het tijd voor het uitvoeren van aanvullende berekeningen en het toevoegen van nieuwe variabelen. Is het analysebestand klaar, dan kunnen de geschikte technieken toegepast worden. En vooraleer de bevindingen kunnen gerapporteerd worden, dienen de analyseresultaten teruggekoppeld te worden naar de oorspronkelijke vraagstelling. Aan het vertalen van de door data mining bekomen inzichten in de operationaliteit gaat derhalve een proces vooraf waarbij het niet uitgesloten is dat op eerdere stappen wordt teruggekomen. Data mining is een iteratief proces.



Figuur 17 : Data mining als iteratief proces

Een belangrijke ontwikkeling die de acceptatie van data mining kan vergroten, is het afspreken van standaarden in de vorm van een procesmodel waardoor de kans van een succesvolle toepassing van data mining in een bedrijfscontext toeneemt. Vaak wordt immers geconstateerd dat een draagvlak voor data mining in de organisatie ontbreekt. Een industrie- en tool-onafhankelijke standaard die de kwaliteit van de uitvoering van data mining verbetert, is CRISP-DM (*CRoss Industry Standard Process for Data Mining ;* voor een bespreking : zie Shearer, 2000).



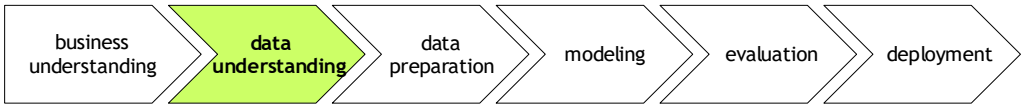
Figuur 18 : CRISP-DM procesmodel voor data mining

CRISP-DM voorziet in een methodiek die het hele data mining-proces beschrijft vanaf de fase van ''business understanding'' tot en met het toepassen van de resultaten van data mining. Het data mining-proces in de context van een lerende organisatie laat zich beschrijven aan de hand van onderstaand procesmodel, dat bestaat uit een zestal stappen.

3.3.5.1. Opstartfase

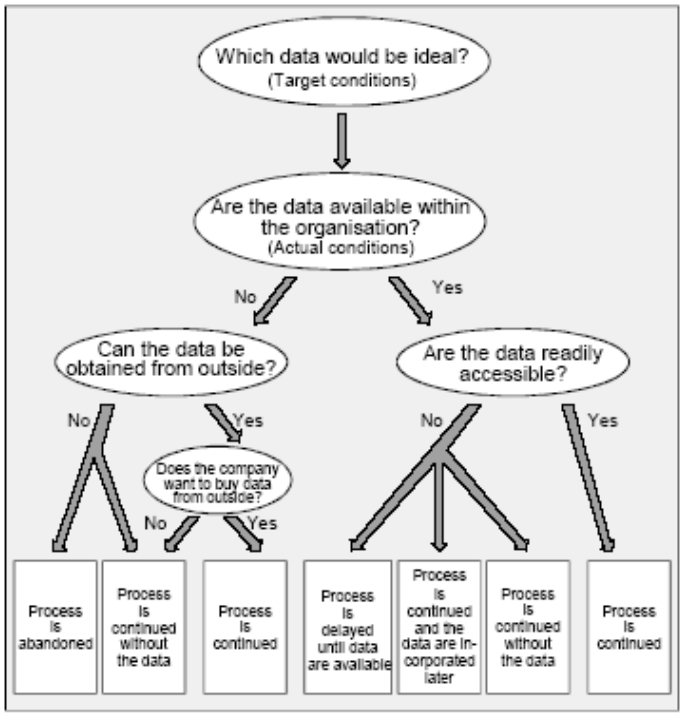
Zoals hierboven reeds werd aangegeven, is het van cruciaal belang om een data mining-exercitie steeds te beginnen met het vastleggen van de bedrijfsdoelstellingen die eraan ten grondslag liggen. Specifieke data mining-doelstellingen dienen altijd een rechtstreekse afgeleide te zijn van bedrijfsdoelstellingen. Het is derhalve raadzaam een projectplan op te stellen met een omschrijving van het probleem, een kosten/baten-analyse en de succescriteria. Een data mining-exercitie begint derhalve met de herkenning van een organisatorisch probleem, zoals een te duur acquisitieproces, het vasthouden of verhogen van de klantwaarde van bestaande klanten, het beoordelen van de kredietwaardigheid van (potentiële) klanten, het voorkomen van frauduleuze transacties, e.d. Het gaat om problemen die betrekking hebben op de uitkomsten van bepaalde processen of het verhogen van de effectiviteit of efficiëntie van deze processen, met als doel het verbeteren van de resultaten voor de organisatie. Een voorbeeld hiervan is : voorspel het klantprofiel van de top 25 % respondenten die zullen reageren op een direct mailing ten einde de acquisitiekosten in vergelijking met voorgaande acties met minstens 15 % te verminderen.

3.3.5.2. Gegevensoriëntatie



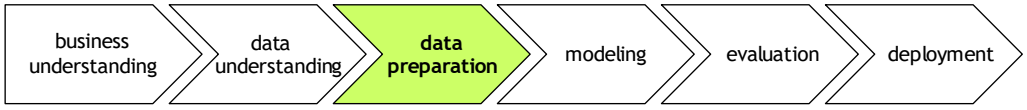
De fase van de gegevensoriëntatie behelst het selecteren van de gepaste data om de gestelde bedrijfsdoelstelling(en) te analyseren, het beschrijven van de data en het uitvoeren van de nodige bewerkingen om ''gevoel te krijgen'' met de gegevens. In de eerste plaats dient bepaald te worden welke data noodzakelijk zijn voor het data mining-proces. Deze vraagstelling heeft betrekking op de relevantie en de informatiewaarde die gegevens dienen te bezitten om opgenomen te worden in het data mining-proces. Eveneens moet nagegaan te worden of deze data al of niet in de organisatie beschikbaar zijn. Indien niet, dient nagegaan te worden of deze data extern beschikbaar zijn en of de organisatie desgevallend bereid is de data aan te schaffen.

Bij het toegankelijk maken van relevante gegevens kunnen data-extractie en koppelingsprobleem optreden, bijvoorbeeld om gegevens die in administratieve systemen opgeslagen zijn, te gebruiken voor marketingdoeleinden. Ook wanneer gegevens geïntegreerd zijn in een data warehouse, kunnen dergelijke problemen opduiken.

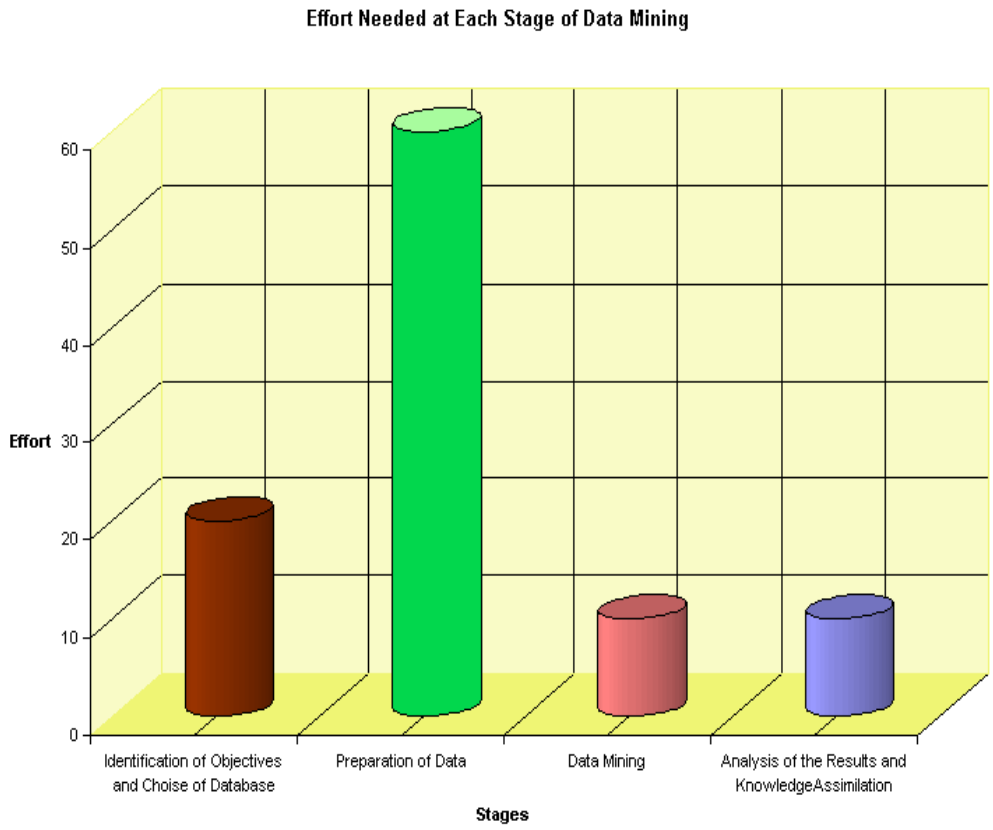


Figuur 19 : Scenario's voor gegevensverzameling in het DM-proces[17]
(Bron : Barth, 1998)

3.3.5.3. Gegevenspreparatie



De fase van de voorbewerking van gegevens behelst de behandeling van uiteenlopende vormen van datavervuiling zoals verdubbeling, veroudering en ontbrekende gegevens (*missing values).* Eveneens wordt bepaald welke variabelen (*features*) zullen meegenomen worden in het eigenlijke mining-proces. Variabelen worden tijdens de fase van de voorbewerking van data bestudeerd en gevisualiseerd naar de verdeling ervan, hetgeen wenselijk is om extreme waarden (*outliers*) op te sporen waarbij de gegevensdistributie van variabelen passend wordt aangepast. Tijdens de fase van de gegevenspreparatie worden eveneens vaak variabelentransformaties uitgevoerd zoals de bepaling of wijziging van het meetniveau van variabelen (*discretization*), het samenvoegen van variabelen die eenzelfde gegeven indiceren (variabelenreductie door bv. factoranalyse) en het coderen van gegevens ten einde er naderhand de gewenste behandelingen op toe te kunnen passen (bv. het coderen van tijdreeksgegevens naar seizoen wanneer de vraag onderzocht zal worden of seizoensinvloeden al of niet aanwezig zijn). De fase van de gegevenspreparatie dient gerekend te worden tot de meest tijdsintensieve van het hele data mining-proces.



Figuur 20 : Aandeel van dataverzameling, –bewerking en modellering in het DM-proces[18]

Ter illustratie van de fase van de gegevensvoorbereiding nemen we het voorbeeld van de eerder vermelde case waarbij het de bedoeling was op kostenefficiënte wijze de markt voor een verzekeringspolis te vergroten (Van der Putten & Den Uyl, 2000). M.b.t. de socio-demografische variabelen zullen we aan de hand van een clusteranalyse nagaan in hoeverre een trainingset (N=5822) kan ingedeeld worden in een aantal segmenten die intern zo homogeen mogelijk zijn en onderling zoveel als mogelijk verschillen vertonen naar de socio-demografische kenmerken. Anderzijds zullen we m.b.t. de bedrijfsinterne gegevens over het productgebruik onderzoeken hoe deze variabelenset kan gereduceerd worden. In beide gevallen ligt aan de variabelenreductie de doelstelling ten grondslag de interpreteerbaarheid van de onderzoeksresultaten te bevorderen door zo weinig als mogelijk van de informatie vervat in de oorspronkelijk beschikbare variabelen te verliezen.

3.3.5.3.1. Variabelenreductie door clusteranalyse

Zoals vermeld is het de bedoeling om aan de hand van clusteranalyse de overeenkomsten en verschillen in waarden van de invoervariabelen (in het hierna uit te werken voorbeeld socio-demografische kenmerken verzameld op postcode-niveau) te bepalen. Er is gebruik gemaakt van *k-means* clusteranalyse, een iteratieve optimaliseringsprocedure die rechtstreeks in de data naar clusters zoekt. Voor iedere cluster wordt gestart met een willekeurig gemiddelde, waarna iedere case wordt toegekend aan het meest nabije gemiddelde. Iteratief worden vervolgens nieuwe gemiddelden berekend en worden de cases aan het meest nabije gemiddelde toegekend, totdat ze niet meer worden opgeschoven naar andere clusters. Er is dan een stabiele situatie bereikt. Omdat het optimale aantal clusters echter niet vooraf bepaald kan worden, zijn meerdere classificaties, met een oplopend aantal clusters berekend. De keuze van het optimaal aantal clusters kan gebeuren met behulp van drie hulpmiddelen, nl. de berekening van het percentage verklaarde variantie (R^2), de pseudo F-statistic[19] en de interpreteerbaarheid van de oplossing. In onderstaande tabel zijn de berekende waarden voor R^2 en de pseudo F-statistic weergegeven voor de oplossing met resp. 2, 3, 4 en 5 clusters.

cluster metric	2 clusters	3 clusters	4 clusters	5 clusters
R^2-overall	0,127	0,193	0,229	0,261
pseudo F-statistic	846,67	689,29	586,92	501,92

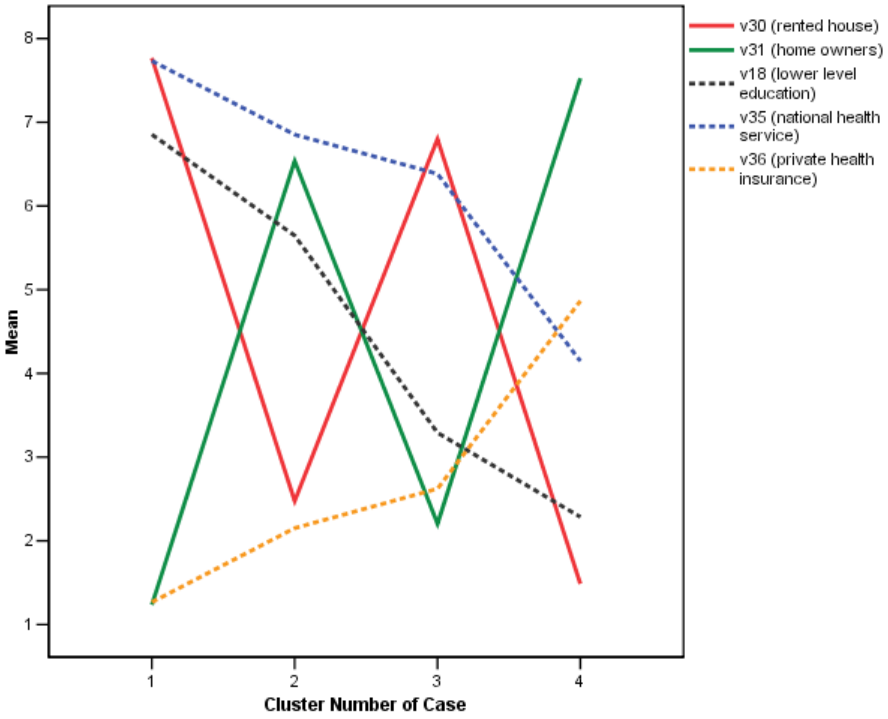
Tabel 1 : Resultaten k-means clusteranalyse (eigen bewerking op socio-demografische variabelen dataset Van Der Putten & Den Uyl,2000)

De waarde van de pseudo F-statistic (die een indicator is voor de homogeniteit binnen clusters) heeft de hoogste waarde in het geval van een 2 cluster-oplossing, hetgeen betekent dat een oplossing met 2 clusters de meest optimale is. Uit tabel 1 lezen we echter ook onmiddellijk af dat de R^2-waarde voor de 2 cluster-oplossing ver beneden de R^2-waarde ligt voor de oplossing met 3 of meer clusters. Rekening houdend met deze hogere R^2-waarden gaat de voorkeur in dit opzicht uit naar een 3- of 4-clusteroplossing. Een kruistabulatie van beide oplossingen toont aan dat de eerste cluster uit de 3 cluster-oplossing in het geval van een 4-clusteroplossing wordt afgesplitst in ongeveer twee gelijke groepen. Van de twee andere clusters is het aantal eenheden die naar een andere cluster opschuiven veel kleiner. Eenzelfde kruistabel voor de verschillen tussen resp. de 4-clusteroplossing en de 5-clusteroplossing wijst uit dat de verschuivingen tussen beide clusteroplossingen veeleer te maken hebben met de toepassing van de clusterprocedure dan met werkelijke verschillen tussen de clusteroplossingen. Om die reden gaat de voorkeur uiteindelijk naar de 4-clusteroplossing.

variable	cluster mean square	df	error mean square	df	F	sig.
6 Roman catholic	149,141	3	0,930	5818	160,351	0,000
7 Protestant ...	380,926	3	2,749	5818	138,558	0,000
8 Other religion	61,617	3	1,004	5818	61,367	0,000
9 No religion	202,453	3	2,449	5818	82,654	0,000
10 Married	1595,028	3	2,826	5818	564,504	0,000
11 Living together	142,958	3	0,860	5818	166,274	0,000
12 Other relation	1127,242	3	2,388	5818	472,087	0,000
13 Singles	1183,255	3	2,631	5818	449,689	0,000
14 Household without children	180,364	3	2,533	5818	71,215	0,000
15 Household with children	974,827	3	3,521	5818	276,894	0,000
16 High level education	1734,046	3	1,741	5818	995,871	0,000
17 Medium level education	1781,770	3	2,184	5818	815,773	0,000
18 Lower level education	5936,948	3	2,223	5818	2670,970	0,000
19 High status	2204,747	3	2,099	5818	1050,495	0,000
20 Entrepreneur	66,772	3	0,567	5818	117,844	0,000
21 Farmer	139,674	3	1,046	5818	133,576	0,000
22 Middle management	1014,156	3	2,863	5818	354,202	0,000
23 Skilled labourers	1372,634	3	2,290	5818	599,518	0,000
24 Unskilled labourers	1409,837	3	2,140	5818	658,733	0,000
25 Social class A	2364,284	3	1,751	5818	1350,455	0,000
26 Social class B1	548,841	3	1,489	5818	368,703	0,000
27 Social class B2	256,193	3	2,208	5818	116,034	0,000
28 Social class C	2706,189	3	2,353	5818	1150,134	0,000
29 Social class D	833,185	3	1,270	5818	656,301	0,000
30 Rented house	13223,702	3	2,730	5818	4843,810	0,000
31 Home owners	13264,474	3	2,712	5818	4890,476	0,000
32 1 car	361,316	3	2,226	5818	162,307	0,000
33 2 cars	225,617	3	1,332	5818	169,409	0,000
34 No car	1230,531	3	1,926	5818	638,941	0,000
35 National Health Service	3176,402	3	2,279	5818	1393,593	0,000
36 Private health insurance	3199,792	3	2,280	5818	1403,430	0,000
37 Income < 30.000	2575,478	3	3,026	5818	851,107	0,000
38 Income 30-45.000	481,623	3	3,298	5818	146,040	0,000
39 Income 45-75.000	1754,888	3	2,813	5818	623,805	0,000
40 Income 75-122.000	296,154	3	1,200	5818	246,762	0,000
41 Income >123.000	43,729	3	0,282	5818	155,165	0,000
42 Average income	1184,258	3	1,127	5818	1050,997	0,000
43 Purchasing power class	2780,118	3	2,597	5818	1070,434	0,000

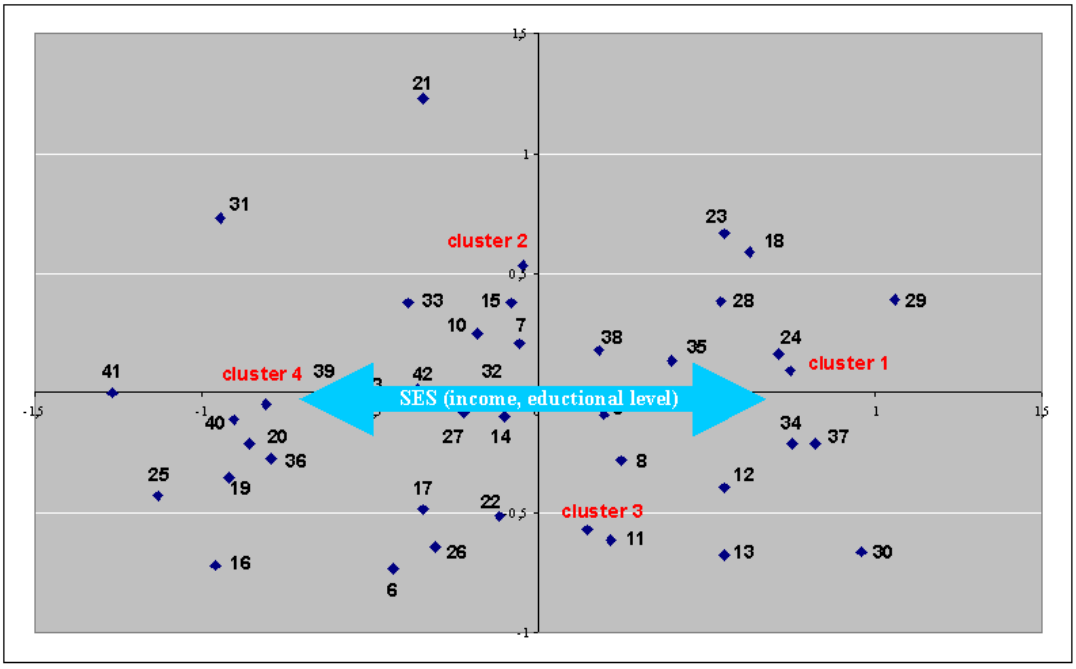
Tabel 2 : F-ratio's voor de 4-cluster-oplossing (eigen bewerking op socio-demografische variabelen dataset Van Der Putten & Den Uyl, 2000)

In tabel 2 is voor de 4 cluster-oplossing de F-ratio weergegeven (ratio van de variantie binnen de clusters en de variantie tussen de clusters) voor elk van de inputvariabelen. De F-waarden zijn een aanduiding voor het bepalen van de onderling grootste contrasten tussen de clusters m.b.t. deze variabelen. Uit de tabel valt af te lezen dat de voornaamste socio-demografische verschillen tussen de 4 clusters in eerste instantie samenhangen met het eigenaarschap van een woning dan wel het huren van een woning (v31 : *home owners* vs v30 : *rented house*), het niveau van genoten onderwijs (v18 : *lower level education*) en publieke sociale zekerheidsvoorzieningen dan wel bijkomende private verzekeringen (v35 : *national health service* vs v36 : *private health insurance*). In figuur 21 worden de verschillen van de vier clusters m.b.t. de gemiddelde score ervan op elk van deze variabelen gevisualiseerd.



Figuur 21 : Grafische voorstelling van clusterverschillen m.b.t. socio-demografische variabelen (eigen bewerking dataset Van Der Putten & Den Uyl,2000)

Ten behoeve van de interpreteerbaarheid van de 4 cluster-oplossing zijn de cluster centroids van de 38 socio-demografische variabelen als input gebruikt voor een correspondentie-analyse[20] waarvan de resultaten grafisch voorgesteld worden in figuur 22.



Figuur 22 : Plot van de relaties tussen clusters en socio-demografische kenmerken (toepassing van correspondentie-analyse met SPSS Categories®)

Uit de plot blijkt dat de vier clusters verdeeld worden over de kwadranten van het assenstelsel. Aangezien de eerste (horizontale) dimensie de meest verklarende is[21], zien we de sterkste verschillen m.b.t. de socio-demografische variabelen optreden tussen clusters 1 en cluster 4, waarbij clusters 2 en 3 een tussenpositie innemen. De horizontale dimensie kan omschreven worden als een dimensie waarbij de verschillen tussen de clusters vooral kunnen afgemeten worden aan sociale klasse- en inkomensverschillen. De (geringere) verschillen tussen de clusters die door de tweede (verticale) dimensie aangegeven worden, hebben in eerste instantie te maken met socio-demografische variabelen die de gezinssamenstelling indiceren. In dit opzicht treden de sterkste verschillen op tussen cluster 2 en cluster 3.

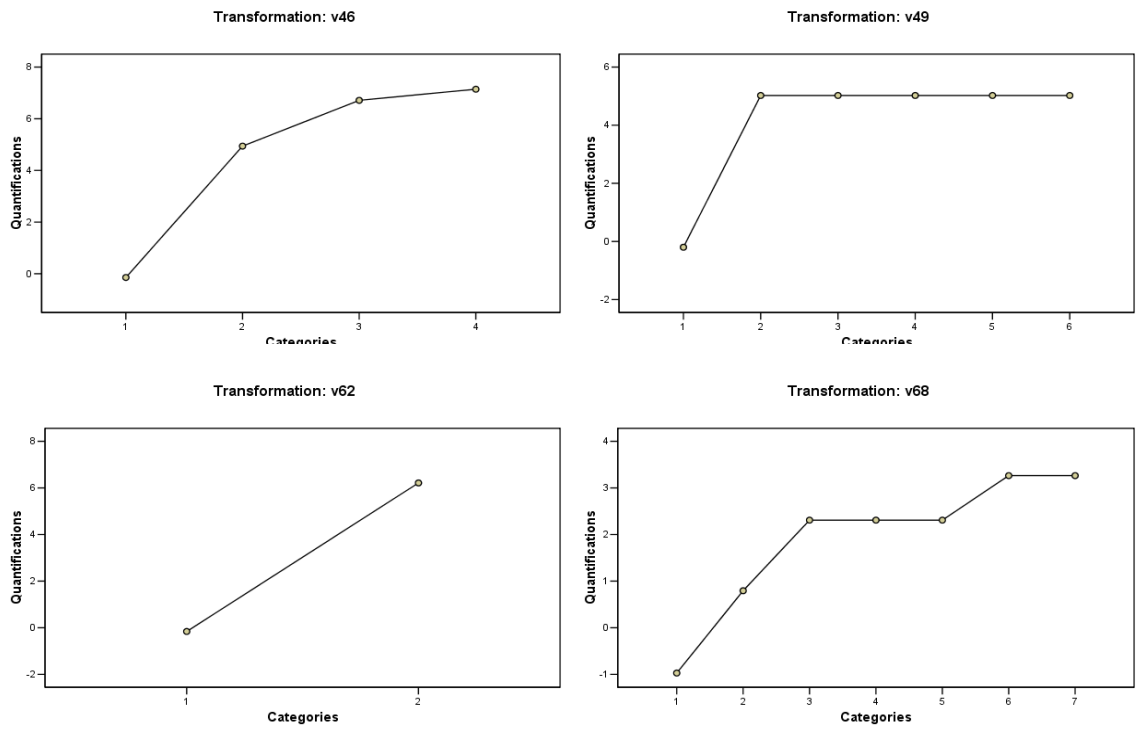
3.3.5.3.2. Variabelenreductie door principale componentenanalyse

Voor het opbouwen van een voorspellingsmodel wordt meestal gebruik gemaakt van interne klantgegevens die het gedrag van klanten indiceren. In het geval van voorliggende case zijn dit variabelen die het productportfolio van bestaande klanten aangeven. Deze bedrijfseigen informatie maakt het mogelijk het feitelijk gedrag van klanten te meten. Vooraleer deze data te betrekken in het opbouwen van een predicitef model is het eveneens noodzakelijk de dimensionaliteit van de gegevens te onderzoeken. Het beschikken over meerdere variabelen die klantgedrag aangeven houdt immers het gevaar in dat de onderling sterke samenhang tussen verschillende variabelen (ook multicollineariteit genoemd) de interpretatie van de analyseresultaten bemoeilijkt. Daarom wordt getracht om de informatie die vervat is in meerdere variabelen te combineren tot één of meerdere schalen. Een schaal kan bv. geconstrueerd worden door iedere categorie van een variabele een waarde te geven en de som van alle waarden geeft dan de schaalwaarde aan. Een eis bij deze

methode is dat alle items minstens op intervalniveau gemeten zijn. Dit is niet steeds het geval. Er kan niet worden aangenomen dat bijvoorbeeld de afstanden tussen vijf waarden op een variabele zoals koopfrequentie allemaal even groot zijn. Bovendien zijn niet alle variabelen waarvan de waarden opgeteld worden noodzakelijk op dezelfde wijze gemeten waardoor de waarden niet zomaar mogen gesommeerd worden tot een schaal. Een datareductiemethode voor (ordinaal) gemeten waarnemingen ligt meer voor de hand.

In het voorbeeld van de bedrijfsinterne gegevens die het productgebruik indiceren van potentiële klanten voor een verzekeringspolis, hebben we de relaties tussen de verschillende variabelen onderzocht met behulp van categoriale principale componentenanalyse. Deze analysetechniek kan ordinale en zelfs nominaal gemeten gegevens verwerken. Kenmerkend voor categoriale principale componentenanalyse is dat er gezocht wordt naar een zo klein mogelijk aantal nieuwe variabelen ter vervanging van de oorspronkelijke variabelen. Hierdoor treedt datareductie op zonder dat veel verklarende variantie verloren gaat. De procedure verloopt als volgt. In een eerste stap worden de categorieën van de variabelen optimaal gekwantificeerd.

Tranformatiegrafieken (zie figuur 23 voor enkele voorbeelden van de optimale kwantificatie van de bedrijfsinterne productgegevens) geven aan of de waarden van variabelen een ordinaal dan wel een intervalniveau representeren[22]. De tweede stap is die van het samenvatten van de verschillende variabelen in één of meerdere dimensies. Hiervoor wordt door de techniek van categoriale principale componentenanalyse voor elke variabele een gewicht bepaald. In tabel 3 zijn de principale componentenladingen vermeld voor de productkenmerken. Deze zijn te lezen als correlaties tussen de oorspronkelijke variabelen en de dimensies. De componentenladingen zijn eenvoudig te interpreteren. Hoe groter de lading in absolute zin, hoe sterker de relatie tussen de oorspronkelijke variabele en de dimensie. Als twee of meer variabelen hoog laden op dezelfde dimensie, dan hangen deze variabelen ook onderling sterk samen. Variabelen die nagenoeg onafhankelijk zijn van elkaar laden nooit hoog op dezelfde dimensie. In tabel 3 is ook Cronbach's alfa vermeld, een maat die de betrouwbaarheid van de dimensies weergeeft. Bij een hoge alfa-waarde kunnen de metingen op de verschillende variabelen gezien worden als herhalingen van elkaar. Er wordt in de regel uitgegaan van een alfa-waarde vam minimaal 0,60.



Figuur 23 : Tranformatiegrafieken door toepassing van categoriale principale componentenanalyse op bedrijfsinterne productgegevens
(eigen bewerking op dataset Van Der Putten & Den Uyl,2000)

	dim 1	dim 2
44 Contribution private third party insurance	0,732	-0,416
46 Contribution third party insurane (agriculture)	0,286	0,840
47 Contribution car policies	0,384	0,060
52 Contribution tractor policies	0,280	0,825
55 Contribution life insurances	0,312	-0,127
59 Contribution fire policies	0,809	-0,112
65 Number of private third party insurance	0,730	-0,416
67 Number of third party insurane (agriculture)	0,285	0,838
68 Number of car policies	0,378	0,074
73 Number of tractor policies	0,281	0,828
76 Number of life insurances	0,312	-0,129
80 Number of fire policies	0,796	-0,134
Cronbach's alfa	0,736	0,852

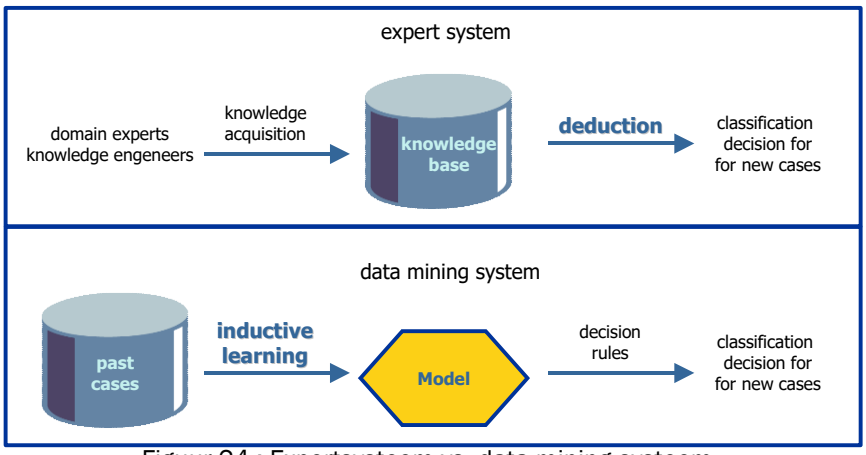
Tabel 3 : Componentenladingen van bedrijfsinterne productgegevens per dimensie
(eigen bewerking op dataset Van Der Putten & Den Uyl,2000)

3.3.5.4. Modelontwikkeling

Tijdens de fase van de modelontwikkeling worden de data geanalyseerd en wordt voor het ontwikkelen van een model een data mining-algoritme op de gegevens losgelaten. Wat gegevensanalyse betreft, dient een onderscheid gemaakt te worden tussen predictieve en descriptieve data mining[23]. In het geval van predictieve data mining is het doel een voorspellend model op te stellen (bv. het opstellen van een respons scoring-model in het geval van mailing, het voorspellen van kredietwaardigheid voor leningaflossing, het voorspellen van het omzetpotentieel van klanten, e.a.). Descriptieve of beschrijvende data mining is geschikt om klantprofielen in kaart te brengen, doelgroepsegmenten te ontdekken, e.a. Omdat het ontwikkelen van voorspellende modellen met voorsprong de meest populaire toepassing is van data mining voor zakelijke toepassingen, heeft onze bespreking van gegevensmodellering in hoofdzaak betrekking op predictieve data mining.

3.3.5.4.1. Deductie en inductie

M.b.t. de fase van de gegevensmodelling dient een onderscheid gemaakt te worden tussen de deductieve en de inductieve werkwijze, waarbij data mining-systemen die gebaseerd zijn op deze laatste werkwijze zich laten onderscheiden van kennissystemen. Kennissystemen, ook expertsystemen genoemd, zijn doorgaans computerprogramma's die zijn opgebouwd uit ''if-then-else''-instructies die menselijke deskundigen hebben bedacht. Het systeem leert niet direct van de gegevens, maar van de kennis die afkomstig is van menselijke deskundigen. In het geval van kennissystemen wordt voor de modellering van gegevens de deductieve werkwijze gehanteerd, d.w.z. gegevens worden gemodelleerd op basis van de kennis die vervat is in het systeem. Daartegenover staat dat in data mining-systemen op inductieve wijze tewerk gaan, d.w.z. de gegevensmodellering is gebaseerd op patronen (en de herkenning ervan) die in de gegevens zelf besloten ligt.



Figuur 24 : Expertsysteem vs. data mining-systeem

3.3.5.4.2. Patronen en modellen

Kenmerkend voor data mining-algoritmen is dat ze patronen kunnen ontdekken en modellen kunnen bouwen. Modellen verwerken gegevens om uiteindelijk efficiënte acties te kunnen ondernemen. Een model kan gedefinieerd worden als een beschrijving van de (historische database) waarbij de patronen (op basis van dewelke het model gebouwd is) toegepast kunnen worden op nieuwe gegevens ten einde een voorspelling te maken van verwachte waarden. Patronen zijn gebeurtenissen of combinaties van gebeurtenissen in een database die vaker plaatsvinden dan verwacht. Patronen zijn slechts interessant als ze met succes in nieuwe situaties kunnen toegepast worden. Het grootste verschil tussen een model en een patroon is dat patronen vaak minder complex zijn en dat er doorgaans veel van zijn. Een model van klantengedrag kan bijvoorbeeld zeer complex zijn en kan honderden patronen bevatten die uit de database zijn gehaald.

3.3.5.4.3. Steekproeven

Modellen maken per definitie een deel uit van een groter geheel waarin gebeurtenissen plaatsvinden. Daarom gaan modelleren en steekproeven nemen hand in hand. Als elke mogelijke situatie in de database verzameld zou kunnen worden, is het niet nodig een voorspellend model te bouwen. Het zou dan voldoende zijn om in de database de specifieke situatie op te zoeken en het antwoord te vinden. Omgekeerd worden steekproeven gebruikt om de omvang van de database te beperken waarin de voorspellende patronen zich bevinden. Neem het voorbeeld van een *creditcard*-maatschappij met miljoenen klantenrecords. Het beste statistische model dat kan gebouwd worden, zou op alle records gebaseerd moeten zijn maar dit zou heel lang kunnen duren en veel kosten. Door het nemen van steekproeven kan een betrouwbaar antwoord verkregen worden dat, met gebruik van veel minder records, toch vaak heel dichtbij het optimale antwoord ligt (*law of diminishing returns*).

Een goed doordacht en correct uitgevoerd experimenteel ontwerp is belangrijk voor het slagen van data mining. Een experimenteel ontwerp refereert naar de manier waarop de gegevens die voor analyse nodig zijn, verzameld en getransformeerd worden. Soms heeft de analist weinig controle over

de manier waarop de gegevens verzameld zijn en/of getransformeerd. Soms is controle hierover mogelijk zodat een beter voorspellend model kan worden gebouwd. Het is eenvoudiger een goed voorspellend model te bouwen wanneer een willekeurige steekproef is uitgevoerd dan met een steekproef die op een bepaalde manier beperkt is en niet de populatie weergeeft die uiteindelijk gemodelleerd zal worden.

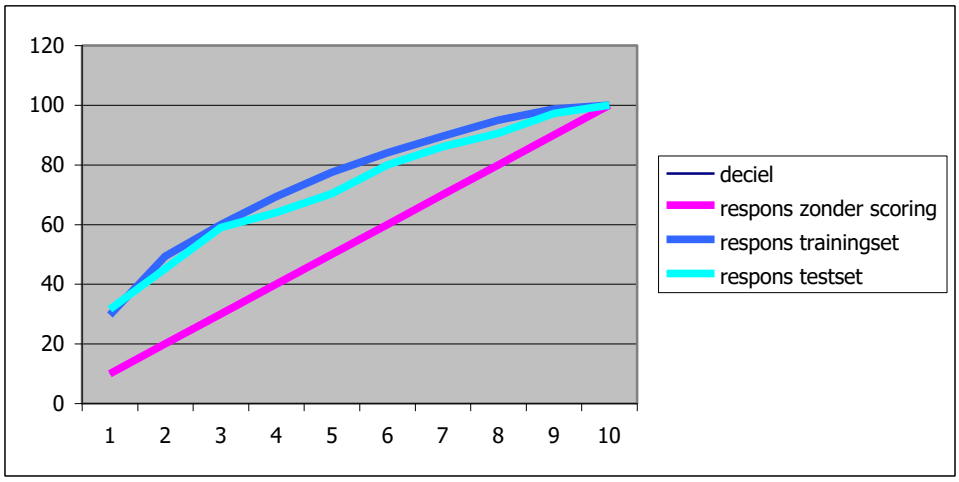
Behalve toevalssteekproeven (*random sampling*) wordt ook vaak gebruik gemaakt van gelaagde steekproeven (*stratified sampling*). Dit is met name het geval wanneer een waarde van de te voorspellen variabele waarvoor gemodelleerd wordt, in zeer lage concentraties voorkomt. Een respons van één procent is niet ongebruikelijk bij een direct mail-campagne waarin 100.000 (potentiële) klanten een aanbod toegestuurd krijgen. Als er een model wordt gecreëerd met behulp van een steekproef van 1000 eenheden, zijn er slechts 10 records die informatie bevatten over klanten die gereageerd hebben. In een dergelijk geval is het beter om meer records te gebruiken van klanten die gereageerd hebben ten einde een beter model te kunnen bouwen. In dit geval zou het zinvol zijn om 500 records te gebruiken van klanten die gereageerd hebben en 500 records waarop geen antwoord is gekomen. Als het model eenmaal gebouwd is, is er nog een stap waarin het model wordt gecorrigeerd naar de oorspronkelijke concentraties van respons en niet-respons records.

Een andere vorm van gelaagde steekproeven is de *cluster* steekproef, waarin de oorspronkelijke database wordt gegroepeerd en er vervolgens gelijke aantallen records uit elke cluster worden gehaald, zodat alle belangrijke subgroepen binnen de database voorkomen. Als er veel van deze subgroepen zijn, en sommige van die groepen bevatten slechts enkele records, moet deze methode zeker gebruikt worden om te kunnen garanderen dat alle groepen in de steekproef voldoende vertegenwoordigd zijn.

3.3.5.4.4. Modelvalidatie

Om de nauwkeurigheid van een model te beoordelen, wordt gebruik gemaakt van een trainingsset en een testset. Als het voorspellende model wordt gebouwd, wordt hierbij de voorspelling op de trainingset gebaseerd en het model wordt vervolgens op de testset toegepast. Een vereiste hierbij is dat beide datasets onafhankelijk van elkaar zijn (d.w.z. geen enkel record bevindt zich in beide databases). In dit geval zal de performantie van het voorspellende model op de testverzameling een indicatie geven van de nauwkeurigheid van het model.

Een specificatie van de wijze waarop aan de hand van klantgegevens een marketingactie kan opgesteld worden, bestaat erin de respons van klanten in te delen in decielen op basis van hun responskans, vertrekkend van de best scorende groep tot de slechtst scorende groep. In figuur 25 is een dergelijke ‘’lift’’-grafiek weergegeven voor de gegevens uit de eerder beschreven training- en testset (zie blz. 25) die we gebruikten voor het opstellen van een beslissingsboom[24]. Uit de grafiek valt o.m. af te lezen dat 40 % van de steekproef met scoring in zowel het geval van de trainingset als de testset ruim twee derde (resp. 69,4 % en 64 %) van de respons genereert. Aan de hand van de scoringsanalyse is het mogelijk de optimale mailingdiepte te berekenen, d.w.z. het aantal mailings dat aan de best responderende groepen toegestuurd dient te worden zonder daarbij onder het *break even*-niveau te zakken.

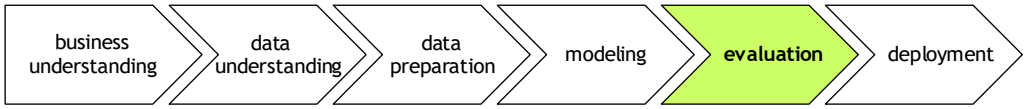


Figuur 25 : Respons-analyse en scoring
(eigen bewerking op dataset Van Der Putten & Den Uyl,2000)

Het opstellen van een scoringsmodel kan tot een aanzienlijke kostenbesparing leiden. Door de respondenten in een mailing te betrekken waarvoor overeenkomstig het scoringmodel een grotere respons kan verwacht worden, wordt een belangrijke kostenbesparing gerealiseerd. Dit veronderstelt evenwel ook dat nadat een model opgesteld is de database in staat moet zijn om de resultaten van het model te implementeren op prospecten en klanten ten einde hieruit de geschikte respondenten te extraheren voor een mailingactie.

Als een voorspellend model gebouwd wordt, is het belangrijk dat de patronen in de gegevens die de volgende keer dat het model wordt toegepast ook nog van kracht zijn, worden gevonden. Een probleem dat zich kan voordoen bij het opbouwen van een model is het effect dat *overfitting* wordt genoemd. Nemen we het voorbeeld van een model dat op basis van een neuraal netwerk is gebouwd. Een neuraal netwerk ‘’leert’’ door herhaaldelijk voorbeelden van de voorspellende variabelen en de voorspellingen (de onafhankelijke variabele en de afhankelijke variabele) voorgeschoteld te krijgen. Als het netwerk meer en meer voorbeelden te zien krijgt, wordt subtiel een verbinding gelegd tussen de verschillende gegevens die binnenkomen (*input*) en het gevraagde antwoord (*output*). Hierdoor maakt het netwerk steeds minder en minder fouten bij de voorspelling van de afhankelijke variabele. In sommige gevallen zal het netwerk helemaal geen fouten meer maken, nadat de traininggegevens verschillende keren door het netwerk zijn gedraaid. De performantie van het model op de validatiegegevens laat echter een ander gedrag zien. Dit effect wordt overfitting genoemd. Naarmate het netwerk steeds meer onthoudt, loopt de kans op fouten in de trainingset steeds meer terug. Dit betekent dat het netwerk patronen begint te ontdekken die niet voorspellend zijn voor het probleem dat aan het netwerk aangeboden wordt, en dat de gevonden patronen enkel voorspellend zijn vanwege enkele typerende kenmerken in de trainingset.

3.3.5.5. Evaluatie



Een belangrijke overweging bij de evaluatie van een data mining-exercitie is dat het resultaat ervan niet enkel beoordeeld wordt op de voorspellende nauwkeurigheid ervan maar dat de bijdrage die data mining oplevert voor het hele bedrijfsproces beoordeeld wordt. Tijdens de evaluatiefase wordt derhalve een terugkoppeling gemaakt naar de bedrijfsdoelstellingen die aan de basis liggen van de data mining-exercitie. De evaluatie van een data mining-exercitie kan plaatsvinden vanuit het gezichtspunt van de sterke en zwakke punten van het toegepaste algoritme, d.w.z. onafhankelijk van de toepassing waarvoor het algoritme gebruikt wordt (zie tabel 4). Nauwkeurigheid is hierbij een belangrijke maatstaf, maar ook andere criteria spelen een rol. Wat hebben we immers aan nauwkeurigheid als het algoritme niet kan omgaan met gegevens die tot op zekere hoogte onzuiver zijn of ontbrekende waarden hebben ? Anderzijds is het mogelijk dat een data mining-algoritme snel een voorspellend model kan opbouwen, maar dit voordeel verzwakt fel als blijkt dat de voorbewerking van gegevens een hele tijd in beslag neemt.

beoordelingselement	omschrijving
nauwkeurigheid	hoe nauwkeurig is het toegepaste algoritme voor het verkrijgen van een juiste voorspelling ?
duidelijkheid	hoe duidelijk is het algoritme in het verklaren van het voorspellend model dat ermee gebouwd wordt ?
vervuilde gegevens	kan het algoritme omgaan met ontbrekende en/of vervuilde gegevens ?
dimensionaliteit	hoe goed kan het algoritme werken met veel voorspellende variabelen ?
onbewerkte gegevens	kan het algoritme werken met de voorspellende variabelen zoals ze in de oorspronkelijke database of steekproef voorkomen, of moeten er voorafgaandelijk bewerkingen uitgevoerd worden ?
RDBMS	is de data mining-techniek geschikt om direct in RDBMS te worden opgenomen ?
schaalbaarheid	werkt het algoritme of grote aantallen records ?
snelheid	is het algoritme (in termen van de tijd die nodig is om een model te bouwen) snel of langzaam ?
validatie	ondersteunt het algoritme mogelijkheden om het voorspellende model te valideren ?

Tabel 4 : Eisen voor beoordelingen van data mining-technieken op algoritmescorelijst
(Bron : Berson & Smith, 1997 : 349)

Om een data mining-techniek en de op basis daarvan ontwikkelde voorspellende modellen te beoordelen op de waarde ervan voor een bedrijf worden andere criteria gehanteerd (zie tabel 5). Hierbij wordt in eerste instantie niet naar snelheid gekeken. Een eerste element voor het bedrijfsmatig succes heeft te maken met het gebruiksgemak en de automatische werkwijze van een data mining-techniek. De techniek moet anderzijds ook begrijpelijke antwoorden verstrekken waarmee een bedrijf iets kan doen. Ten slotte moet de techniek ook een antwoord kunnen geven dat omgezet kan worden in een ROI-analyse.

beoordelingselement	omschrijving
automatische werkwijze	is de techniek relatief automatisch en gemakkelijk te gebruiken of is er veel ervaring met data mining noodzakelijk ?
duidelijkheid	zijn de resultaten die voortkomen uit het gebruik van de data mining-techniek duidelijk en begrijpelijk of zijn ze complex en niet-intuïief ?
ROI	kan de techniek gebruikt worden voor winst- en verliesresultaten en verbeterd rendement op geïnvesteerd kapitaal ?

Tabel 5 : Eisen voor beoordeling van data mining-technieken op bedrijfsscorelijst
(Bron : Berson & Smith, 1997 : 347)

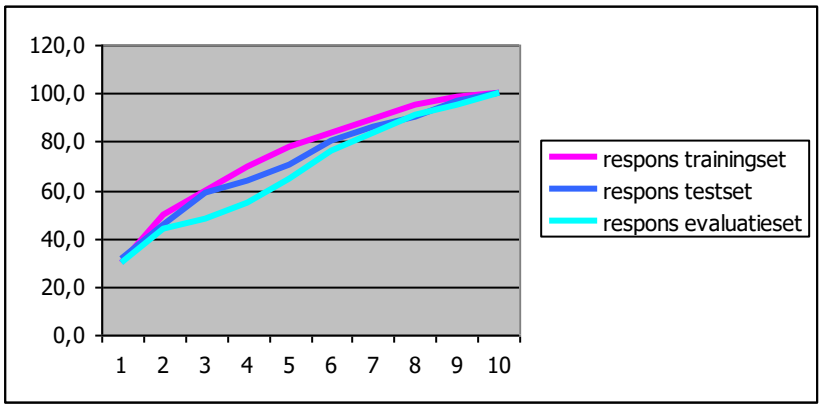
Ongeacht een data mining-techniek beoordeeld wordt overeenkomstig de algoritmescorelijst of de bedrijfsscorelijst, gelden drie aspecten als voornaamste evaluatie-criteria, nl. nauwkeurigheid, verklaring en integratie.

3.3.5.5.1. Nauwkeurigheid

Bij data mining gaat het in de eerste plaats om voorspellende nauwkeurigheid, d.w.z. een data mining-techniek moet in de eerste plaats een model produceren dat zo nauwkeurig mogelijk is. Voor ongerichte data mining (zoals clustering) zijn directe beoordelingen moeilijker dan in het geval een techniek gebruikt wordt voor gerichte data mining.

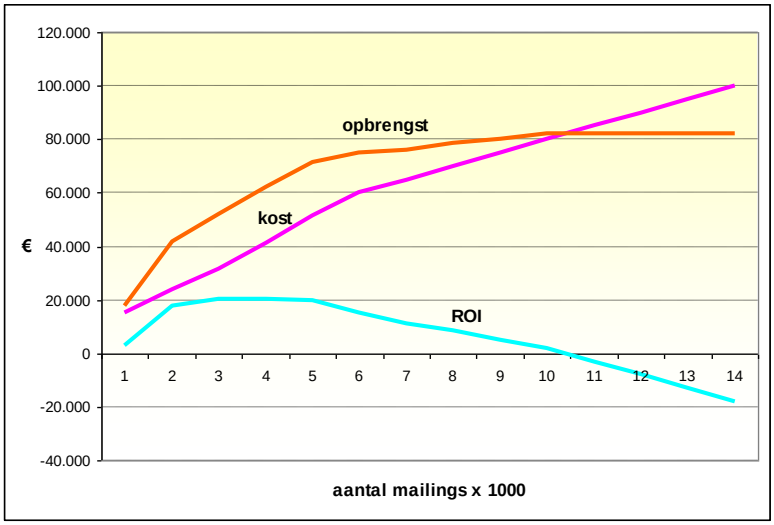
Als een techniek wordt gebruikt voor voorspelling, kan de nauwkeurigheid op verschillende manieren vastgesteld worden. Zo kan nauwkeurigheid vastgesteld worden als het totaal aantal correcte voorspellingen. Voor een binaire voorspelling kan de correctheid of incorrectheid direct berekend worden[25] en voor een voorspelling met meerdere waarden kan de nauwkeurigheid berekend worden als de gemiddelde gekwadrateerde fout (*mean squared error*). De verhoging (*lift*, zie figuur 25 blz. 40) meet in hoeverre een voorspellingsmodel de responsdichtheid voor een bepaalde deelverzameling van een database verhoogt t.o.v. hetgeen zonder model (willekeurige selectie) zou bereikt worden. Om te vermijden dat er interferentie ontstaat tussen het opstellen van een classificatieregel en het valideren ervan, wordt naast een training- en testset gebruik gemaakt van een evaluatieset waarbij het ontwikkelde model toegepast wordt ter voorspelling van de waarde van de afhankelijke variabele. M.b.t. het hiervoor uitgewerkte voorbeeld van de voorspelling van de klantenrespons op een verzekeringspolis, is het ontwikkelde model toegepast op een evaluatieset van 4.000 klanten waarbij de waarde van de (binaire) afhankelijke variabele voorspeld werd aan de hand van de

classificatieregels. In figuur 26 worden de resultaten van deze analyse grafisch voorgesteld. Hieruit kan opgemaakt worden dat de ontwikkelde classificatieregels bij toepassing ervan op een evaluatieset goed standhouden, al moet eraan toegevoegd worden dat de responsdichtheid lager ligt in de evaluatieset.



Figuur 26 : Respons-analyse en scoring : vergelijking tussen training-, test- en evaluatieset (eigen bewerking op dataset Van Der Putten & Den Uyl,2000)

Een andere mogelijkheid om de nauwkeurigheid vast te stellen bestaat erin de maximale winst of het rendement op het geïnvesteerde kapitaal uit het voorspellende model te berekenen.



Figuur 27 : ROI-analyse

Bij een gekende mailingdiepte met bijhorende responsverwachting kunnen kosten, opbrengsten en *return on investment* grafisch voorgesteld worden. Uit het (fictieve) voorbeeld in figuur 27 kan afgeleid worden dat de grootste ROI bereikt wordt bij een mailinggrootte van 5.000.

3.3.5.5.2. Verklaring

M.b.t. het door een data mining-tool ontwikkelde model moet de eindgebruiker op een duidelijke manier kunnen verklaren hoe het model werkt. Het gaat erom dat inzichtelijkheid dient opgebouwd te worden. Eveneens moet de winst of de ROI-berekening op een duidelijke wijze uitgelegd worden.

3.3.5.5.3. Integratie

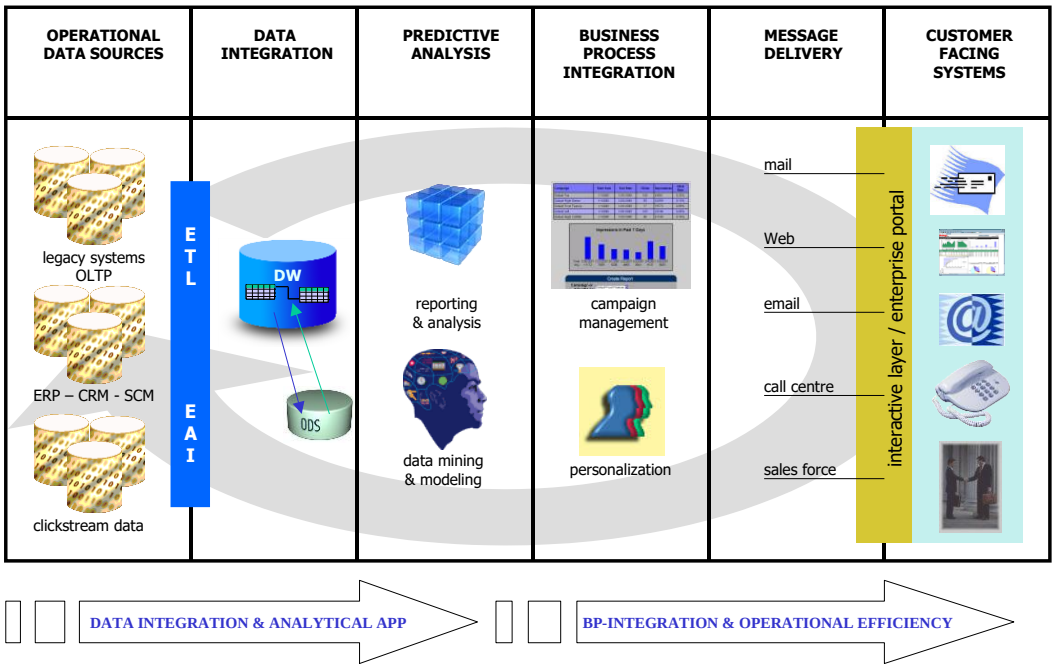
Data mining-tools en de resultaten en modellen die erdoor bekomen worden, moeten kunnen geïntegreerd worden in de bestaande gegevens- en informatiestromen en de bedrijfsprocessen (zie hierover verder onder 3.3.5.6.).

3.3.5.6. Actie

De resultaten van een data mining-exercitie dienen uiteindelijk ook daadwerkelijk gebruikt te worden in de bedrijfspraktijk. Waar het om essentie om draait, is dat er twee processen zijn die een gesloten kringloop moeten vormen waardoor de onderneming klantgeoriënteerd en dynamisch, d.w.z. met een korte *time-to-market* kan reageren.

Aan de ene kant staat het integreren van gegevensstromen waardoor de grondstoffen voor handen zijn om kennis te extraheren aan data. Aan de andere kant volgt het implementeren van deze kennis in de bedrijfsprocessen. Wat betreft dit laatste, kunnen data mining-tools ingedeeld worden in generieke systemen die een brede range van toepassingsgebieden ondersteunen en applicatie-specifieke systemen, die ontwikkeld zijn voor een specifiek toepassingsgebied. Deze laatste kennen de afgelopen jaren een sterke opgang. Het ontwikkelen van gerichte oplossingen wordt ook verticalisatie genoemd (zie ook blz. 51).

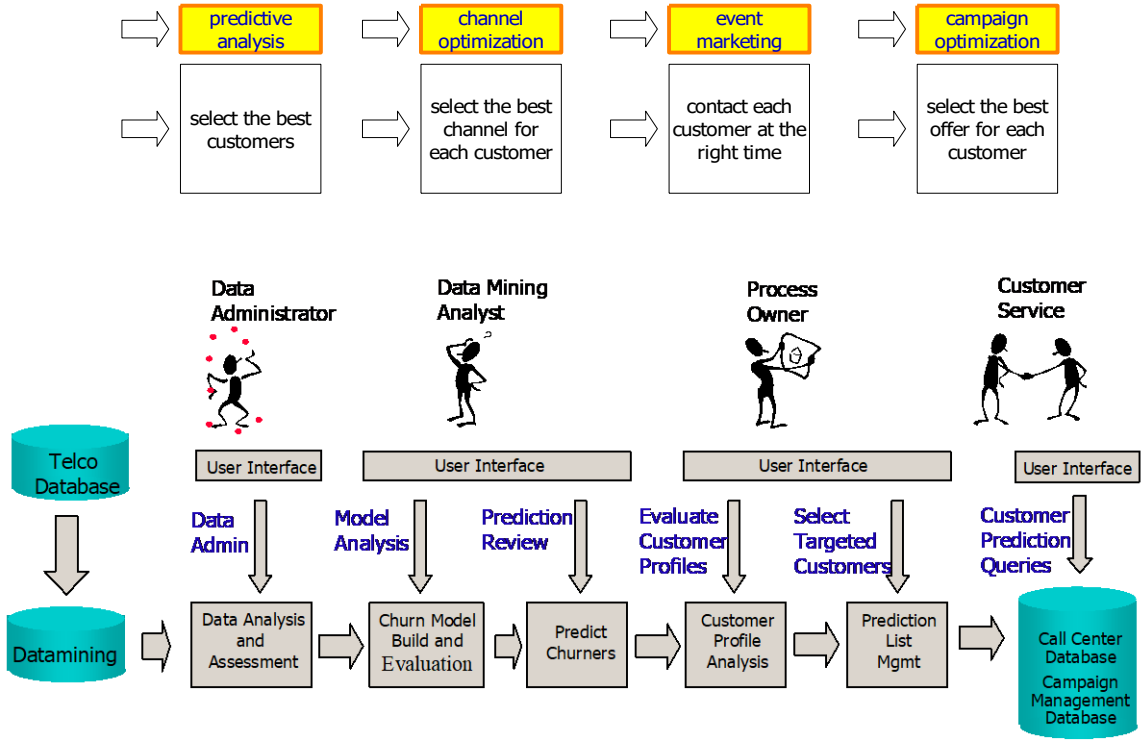
In figuur 28 worden de twee genoemde kringloopprocessen, nl. gegevensintegratie en daarop gebaseerde (predictieve) analyse die resulteert in het implementeren van de resultaten in bedrijfsprocessen grafisch voorgesteld.



Figuur 28 : Integratie van databaronnen, predictieve analyse en implementatie in operationele toepassingen

Verschillende ontwikkelingen zoals *multi-channel management* en *e-business* leiden tot een toenemende behoefte aan recente en geëintegreerde data voor zowel analytische als operationele toepassingen. Een voorbeeld is een toepassing waarbij in het *customer contact center* of op de website aan *cross-selling* wordt gedaan. Het klantprofiel is dan essentieel. Dit profiel is enerzijds gebaseerd op eerdere aankopen maar is anderzijds afhankelijk van ander klantgedrag dat (zeer) recent heeft plaatsgevonden, zoals klikgedrag op de website van het bedrijf. Om het klantprofiel te bepalen, is derhalve meer nodig dan (historische) informatie uit het datawarehouse. Zoals dit het geval is voor data-integratie in een traditioneel datawarehouse, dienen recente gegevens uit diverse operationele systemen opgenomen te worden. Is het inrichten van een ETL-proces in een DW-omgeving al complex, dan is het oppikken van *real-time* transacties dit in verhevigde mate omdat voor dit laatste een batchproces geen optie is aangezien de transactionele data zo snel mogelijk na het plaatsvinden ervan opgepikt moeten worden. Om te voorzien in zowel recente als geïntegreerde data, is een aantal benaderingen mogelijk. Eén ervan bestaat erin gebruik te maken van *message brokers* die op recordniveau zorgen voor real-time communicatie tussen applicaties. Het gebruik van message brokers in projecten voor applicatie-integratie (EAI : *Enterprise Application Integration*) komt hierop neer dat informatiecomponenten (zoals bijvoorbeeld naam, leeftijd, aankoopbedrag, e.a.) binnen één toepassing gewijzigd worden en via een *message queue* doorgegeven worden aan alle andere relevante toepassingen (zie hierover Kamst, 2002). Met het oog op de analyse ervan met andere data, komen de aldus behandelde transacties terecht in een afzonderlijke database die doorgaans aangeduid wordt onder de naam *operational data store* (ODS). Aangezien het gewenst is dat de data zo snel mogelijk in het ODS worden opgenomen, worden op gegevens in het ODS nauwelijks transformaties en aggregaties uitgevoerd. En vermits het ODS steeds *up-to-date* wordt gehouden, kan ook de laadfrequentie van het datawarehouse opgevoerd worden (zie o.m. Den Hamer, 2001).

Het komt er echter niet alleen op aan om data (historische en recente) te integreren en nieuwe kennis te onttrekken aan gegevens. Het gaat er ook om deze kennis ook aan te wenden. Het doel van predictieve analyse bestaat erin om data om te zetten in inzicht zodat beslissingsprocessen op operationeel, tactisch en strategisch niveau ondersteund worden en dat tijdig kan geanticipeerd worden op gedrag en gebeurtenissen. Het implementeren van de resultaten van predictieve analyse in bedrijfsprocessen komt de efficiëntie en de effectiviteit van deze laatste ten goede. Grijpen we terug naar het voorbeeld over cross-selling dan kan een inkomend gesprek in het customer contact center aan de hand van een DM-applicatie geanalyseerd worden in combinatie met gegevens over de klanthistorie en andere klantinformatie uit verschillende kanalen. Het systeem evalueert op deze wijze de slagingskans van cross-sell en retentie-aanbiedingen en bepaalt eveneens op basis van *business rules* welke aanbieding de hoogste kans heeft door de klant geaccepteerd te worden én de meeste omzet genereert voor het bedrijf. De specifieke aanbieding wordt, samen met bijhorende verkoopadviezen, zichtbaar gemaakt op het scherm van de call center-medewerker. Door de combinatie van real-time technologie en predictieve analyse worden de slaagkansen van aanbiedingen verhoogd. Real-time gegevensintegratie zorgt ervoor dat aanbevelingen gebaseerd zijn op de meest actuele klantinformatie (zoals gegevens verkregen via andere kanalen). Aan de hand van predictieve analyse worden de voorkeuren en behoeften van klanten voorspeld en kunnen de geschikte doelgroepen voor aanbiedingen afgebakend en geselecteerd worden en aan de hand van ''best practice''-scenario's wordt de complexiteit van data en modellen onttrokken aan de beslissers en worden de resultaten van analyse omschreven in begrijpelijke termen die eigen zijn aan het bedrijfsproces.



Figuur 29 : Predictieve analyse en target marketing

3.4. Trends in business intelligence

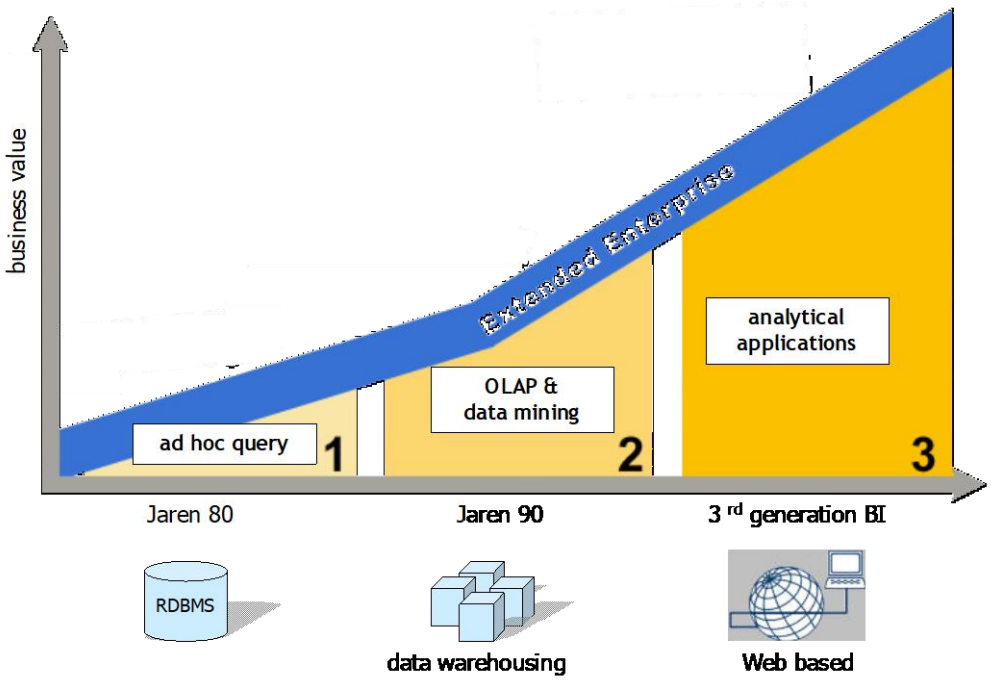
3.4.1. Van ad hoc-rapportering naar predictieve analyse

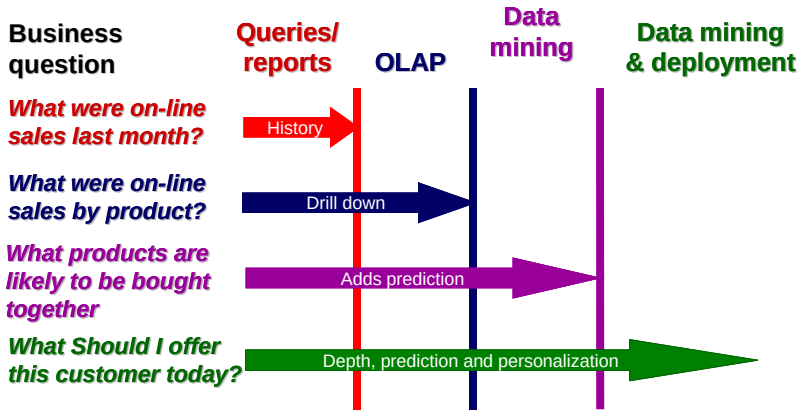
Data mining maakt deel uit van een proces dat bekend staat onder de naam *Knowledge Discovery in Databases* (KDD). Het doel van KDD is het afleiden van nieuwe informatie en kennis door middel van analyse van in databases opgeslagen gegevens. Het proces van KDD is gebaseerd op de veronderstelling dat databases veel verborgen informatie bevatten. Met KDD probeert men deze verborgen informatie aan de oppervlakte te brengen. Behalve het opsporen van deze verborgen informatie brengt KDD de resultaten van verfijnde analyses ook binnen het bereik van eindgebruikers. Zo stelt Fayyad : ''The idea is to put effective analysis-tools in the hands of end-users that are typically not statisticians, or machine learning researchers'' (Fayyad, 1996 : 11). Hiermee wordt een eerste trend in de toepassing van bedrijfsintelligentie aangegeven, nl. de opkomst van voorspellende analyses in plaats van rapportering en de directe inzet van de resultaten van voorspellende analyses door de integratie ervan in operationele systemen.

Het toegenomen belang van analyse in het algemeen en predictieve (voorspellende) analyse in het bijzonder dient geplaatst te worden tegen de achtergrond van de verschuiving die zich de afgelopen jaren heeft voltrokken van het efficiënt communiceren met klanten naar het kennen van de individuele klantbehoeften en voorkeuren en het toepassen van deze kennis in de klantcommunicatie. Op het moment dat klantkennis het uitgangspunt is voor commerciële klantactiviteiten, verandert ook de focus van het marketing- en verkoopproces. Was er voorheen een productfocus, dan moet deze nu veranderen naar een klantfocus. De doelstellingen zullen dan ook niet meer op productniveau liggen maar op klantniveau.

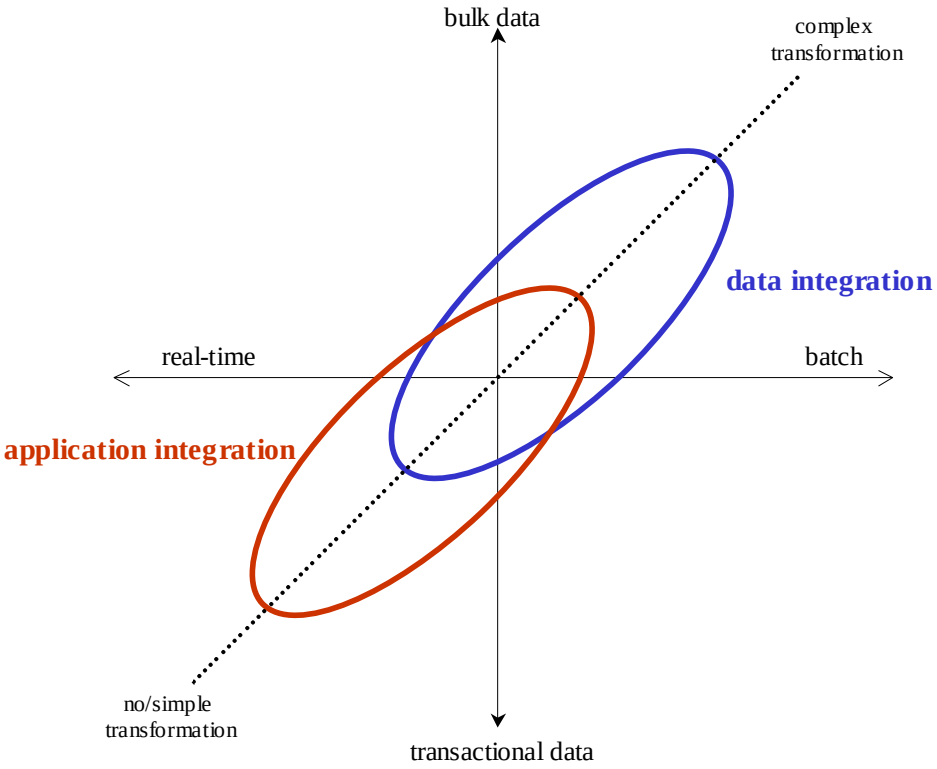
Klanten zijn veeleisend en willen gericht benaderd worden met aanbiedingen die aansluiten bij hun individuele behoeften. Klanten verwachten ook een onmiddellijke reactie op hun behoeften. Internet heeft de serviceverwachting van klanten verhoogd : ze verwachten een directe respons op hun serviceverzoek. Klanten verwachten ook consistente communicatie over de verschillende kanalen. Een naadloze integratie van alle kanalen, met de beschikbaarheid over real-time klantinformatie, biedt hier een oplossing.

Personalisering van klantcontacten en *real-time* analyses zijn derhalve belangrijke elementen in de klantcommunicatie. Voor het uitvoeren van analyses dienen dan ook recente operationele data beschikbaar te zijn (cfr. supra). Historische informatie (aanwezig in het datawarehouse) wordt gecombineerd met real-time gegevens. Voor real-time analyse is real-time integratie vereist. Een overgrote meerderheid (95 %) van de informatie die voor een dergelijke integratie in aanmerking komt, wordt onttrokken en geladen in een typisch ETL-batchproces. Het restant van up-to-date bedrijfsinformatie wordt via EAI-*message queues* onttrokken aan ERP-, SCM en CRM-applicaties en samen met de historische informatie uit het datawarehouse geïntegreerd in een ODS (cfr. supra). De ODS functioneert dus als een centrale opslagplaats voor bedrijfsinformatie van waaruit toepassingen geïntegreerde informatie kunnen onttrekken[26].





Figuur 30 : Evolutie vraagstellingen voor bedrijfsintelligentie
(Bron : Purcell, 2002)



Figuur 31 : De convergentie tussen data integration en application integration

De basisvraag luidt hoe de informatie die vele organisaties en bedrijven verzamelen (via klantenkaarten, klik- en koopgedrag op websites, demografische informatie, respons op mailings, marktonderzoek, e.d.) kan omgezet worden in direct toe te passen activiteiten. Inmiddels is gekend dat data mining voor veel organisaties en bedrijven een middel is om maximaal rendement te halen uit de opgeslagen informatie. Er zijn evenwel nog veel organisaties die de vruchten van diepgaande analyse nog niet plukken. Zo duurt het krijgen van resultaten en vooral de integratie van resultaten in operationele processen vaak zo lang dat het te laat is om nog van de resultaten gebruik te maken. Om die reden worden tools voor predictieve analyse tegenwoordig steeds vaker ingebed in analytische applicaties waardoor het analyseproces veel dichter aansluit bij het operationele proces dan voorheen mogelijk was. Zoals blijkt uit onderstaande tabel is een analytische tool in meerdere opzichten anders dan een analytische applicatie.

Criterium	ANALYTISCHE TOOL	ANALYTISCHE APPLICATIE
oriëntatie	generiek gebruik van analytische technologieën	specifieke bedrijfsprocessen
doelgroep	''power users''	''business users''
analytische technologie	smal en diep	breed en meestal minder diep (integratie van alle relevante technologie : ETL, data opslag, reporting, data mining)
analytisch proces	wordt aan de gebruiker overgelaten	ingebouwd
inzetbaarheid	brede inzetbaarheid	gerichte inzetbaarheid
integratie in operationele systemen	doe het zelf	volledig en voorgeprogrammeerd

Tabel 6 : Analytische tool vs analytische applicatie

Een belangrijke ontwikkeling binnen data mining is verticalisatie, d.w.z. het ontwikkelen van maatwerk-producten met een focus op een specifieke doelstelling. Analytische applicaties zorgen voor een terugkoppeling van de analyseresultaten naar de (operationele) processen waardoor ze bijdragen tot het uitvoeren van acties binnen de directe context van bedrijfsprocessen. Het analyseren van bijvoorbeeld klantgegevens is geen doel op zich. Waar het om gaat is dat de conclusies van analyses (in combinatie met kennis over de profitabiliteit van de klant) wordt teruggebracht naar direct en indirecte verkoopkanalen. Op deze wijze kunnen bedrijfsprocessen aangestuurd worden vanuit de resultaten van analyses.

Operationele processen zoals marketing en verkoop (operationele CRM) worden ondersteund en gestuurd vanuit opgebouwde klantenkennis (analytische CRM). In dit opzicht wordt gesproken over *cross-channel* of *multi-touchpoint* personalisatie die ervoor zorgt dat klantgegevens op de juiste plaats, op het juiste moment en in de juiste vorm en kwaliteit beschikbaar en toegankelijk zijn. De met behulp van data mining uitgevoerde klantsegmentaties en opgestelde scoringsmodellen kunnen geïntegreerd worden in marketing- en verkoopacties. Binnen het kader van *campaign management* kunnen klantprofielen en scores geëxporteerd worden naar operationele applicaties waarbinnen de te benaderen populaties gedefinieerd worden. De door personalisering gerealiseerde verhoging van de conversie kan hierbij aanzienlijke financiële voordelen opleveren. Naast de toepassing van campaign management kunnen de resultaten van analyses ook geïntegreerd worden in andere communicatiekanalen, zoals call center en Internet. Wanneer een klant zich bij het call center of op de website aanmeldt, worden de actueel beschikbare klantgegevens on-line vergeleken met de door data mining vastgelegde profielen. Op basis van het resultaat van deze vergelijking wordt de voor de klant meest effectieve boodschap gegenereerd en aangeboden via het call center of de website.

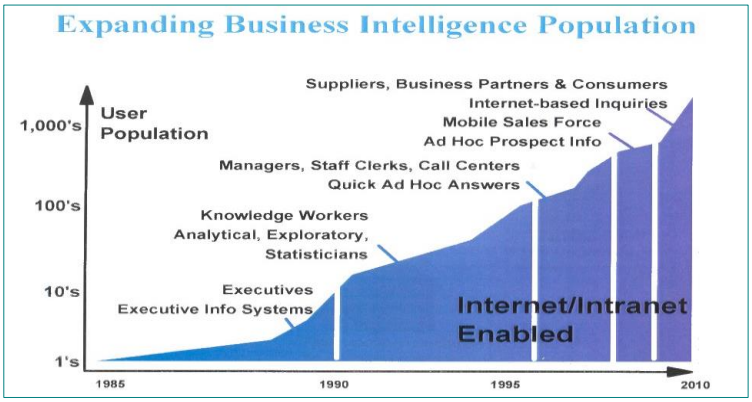
Waar data mining voorheen als *ad hoc*-proces bedoeld werd, wordt data mining zoals hierboven beschreven steeds vaker toegepast in verticale applicaties en *on-line* gebruikt. In het geval van on-line data mining is een belangrijk deel van het data mining proces reeds voltooid vooraleer de mining vraag is gesteld. De database wordt geanalyseerd en regels worden opgeslagen in een patroonbestand (zie hierover o.m. Parsaye, 1999). Een onmiskenbaar voordeel van on-line data mining is dat het meer geïntegreerd is in het bedrijfsproces en dat de doorlooptijd (heel) kort is.

De evolutie van data mining van de ad hoc- naar de on-line variant loopt parallel met het toegenomen belang van *enterprise information portals* (EIP) voor de verspreiding van informatie. Via deze laatste hebben gebruikers toegang tot informatie die voor hen beschikbaar is op eigen verzoek, automatisch op vaste tijdstippen of *event-driven* (waarbij de gebruiker de informatie pas ontvangt als zich een bepaalde gebeurtenis voordoet). Op basis van deze informatie kunnen de gebruikers beslissingen nemen en de acties die daaruit voortvloeien direct in transactiesystemen uitvoeren. De ontwikkelingen op IT-gebied maken het ook mogelijk om beslissingssystemen op te zetten die in staat zijn om op basis van regels, zelfstandig beslissingen te nemen en acties uit te voeren in de operationele systemen (zie hierover o.m. Schipper, 2001 ; Allen Bonde Group, 2002, Bates, 2003, Blomme, 2003 ; Rudin & Cressy, 2003)[27].

3.4.2. Business intelligence en corporate performance management

Een tweede belangrijke trend die zich aftekent m.b.t. business intelligence is dat BI-systemen zeer frequent actuele informatie over de prestaties van de organisatie (of organisatieonderdelen) bieden, waardoor business intelligence niet alleen een technisch begrip is maar ook een concept is dat zich richt op het ondersteunen van kritische bedrijfsprocessen op basis van tijdige, juiste en volledige informatie. Business intelligence wordt meer en meer aangewend als middel om een resultaatgerichte bedrijfsvoering te ondersteunen, zowel door de performatie van bedrijfsprocessen te meten als door de indicatoren van de bedrijfsvoering te delen met partners uit de waardeketen.

Omdat bedrijfsintelligentie niet alleen vanuit een technische hoek moet benaderd worden maar ook vanuit een organisatorische, procesmatige invalshoek wordt BI ook steeds vaker in één adem genoemd met managementstrategieën zoals *corporate performance management* (deze strategie richt zich op procesverbetering voor efficiënter en slagvaardiger opereren van de organisatie, hetgeen de gestegen belangstelling voor *dashboards* en *scorecards* verklaart) en *knowledge management* (kennismanagement : het integreren van kennis- en ervaringsbronnen zowel binnen als buiten de onderneming). Om die reden kan BI als volgt gedefinieerd worden : ''Business intelligence is a process to monitor key performance indicators about business environment, customers, investors, partners, suppliers, regulators, competitors, industry trends, and their impact on business strategy to help define and improve a profitable business model'' (Hashmi, 2003).



Figuur 32 : De uitbreiding van business intelligence tot het ecosysteem van de onderneming (Bron : Altius Consulting, 2001)

Sterk samenhangend met het belang van BI ter ondersteuning van bedrijfsprocessen is de uitbreiding van bedrijfsintelligentie tot het ecosysteem van de organisatie. In dit opzicht dient het onderwerp van bedrijfsintelligentie gekoppeld te worden aan het door de Gartner Group gelanceerde begrip *zero latency enterprise* en de organisatie van wat omschreven wordt als *the extended enterprise.* De ZLE kan gedefinieerd worden als een onderneming die in staat is gegevens en informatie in zeer korte tijd (zero latency) te laten stromen naar de plek waar ze nodig zijn om bedrijfsprocesen te stroomlijnen, kostenreducties te realiseren en een hoge response te verwezenlijken met partijen buiten de onderneming (klanten, leveranciers).

De extended enterprise is een organisatie die zich bewust is van het op de juiste manier communiceren met zijn externe relaties. In de extended enterprise is een technische infrastructuur ontworpen die er zorg voor draagt dat de eigen informatiesystemen kunnen communiceren met die van de relaties. Relaties met partijen, klanten en leveranciers zullen meer en meer gemanaged worden door een E-business infrastructuur.

Extended enterprise-toepassingen waarop zero latency zijn vruchten afwerpt zijn *supply chain management* (SCM) en *customer relationship management* (CRM). De kritische succesfactoren van SCM zijn de zichtbaarheid waar processen en gegevens zich bevinden in de keten en de snelheid waarmee producten en gegevens/informatie zich bewegen door de keten. Zero latency op het vlak van CRM kan worden bereikt door de medewerkers van de organisatie die in contact staan met de klant op het juiste moment van de juiste gegevens te voorzien.

4. Besluiten

In marketing is een ontwikkeling gaande van een massabenadering naar een gesegmenteerde aanpak en verder naar een individuele benadering. De lange tijd dominerende *push*-benadering maakt meer en meer plaats voor een *pull*-benadering waarbij het initiatief niet langer bij de producent maar steeds meer bij de klant komt te liggen. De toenemende invloed van de klant en de intensiever wordende concurrentie eisen dat het aanbod nauwkeurig wordt afgestemd op de wensen en de verwachtingen van de klant. Bedrijfsprocessen dienen zodanig ingericht te worden dat deze zo effectief en zo efficiënt mogelijk gericht zijn op het opbouwen en onderhouden van duurzame relaties met klanten.

De ontwikkeling van een klantgerichte marketingstrategie (*customer relationship management,* CRM) is afhankelijk van de invulling van zowel een interactieve als informatieve dimensie. De omslag van een aanbod- naar een vraaggeoriënteerde benadering impliceert in de eerste plaats dat bedrijven in staat moeten zijn om interactieve media aan te wenden in de communicatie met prospecten en klanten. Het principe van klantgerichtheid houdt in dat de onderneming elke klant individueel benadert en aan de individuele wensen van de consument tegemoet komt met een op maat gesneden aanbod. Dit is alleen mogelijk als de onderneming voldoende inzicht heeft in de voorkeuren van prospecten en klanten. Waar het in toenemende mate over gaat, is kennis te verwerven in de zeer dynamische omgeving van de organisatie en deze om te buigen in een competitief voordeel.

De vaststelling dat klantcommunicatie in de toekomst grotendeels elektronisch zal geschieden en het gegeven dat het behoud van een competitief voordeel wel eens zou kunnen afhangen van het kleine verschil in markt- en klantinformatie, bewerkstelligen een toenemende vervlechting tussen informatie-technologie en bedrijfsprocessen. Wat dit laatste betreft, kunnen datawarehousing en data mining aangeduid worden als belangrijke ontwikkelingen ter ondersteuning van de beschikbaarheid van gegevens en de extractie van kennis aan deze laatste.

Mede door de invloed van het Internet en E-commerce die een stimulerend effect hebben op het genereren en opslaan van “clickstream data” zal het belang van data mining in de komende jaren nog toenemen. Tegenwoordig zijn de meeste producten voor data mining nog toegesneden op expertgebruikers, zoals statistici. De resultaten van data mining krijgen pas waarde als er zinvolle actie op ondernomen kan worden. Een belangrijke voorwaarde voor de acceptatie van data mining is het integreren van data mining in zgn. verticale applicaties, d.w.z. toepassingen gericht op een specifieke bedrijfsfunctie. Rekening houdend met het strategisch belang van data mining maar ook met de mogelijkheden om data mining direct toe te passen in het operationele domein, kan in de nabije toekomst verwacht worden dat de technologie van data mining in bedrijfsprocessen zal ingezet worden voor specifieke doelstellingen zoals lening- en polisacceptatie in de bank- en verzekeringswereld, basketanalyse en voorraadbeheer in retail marketing, loyaliteitsanalyses op basis van betaalgedrag in de telecommunicatie, e.a.. Data mining is dan niet langer een proces dat ad hoc wordt uitgevoerd, maar is één van de bouwstenen om *real-time marketing* uit te voeren aan de hand van de combinatie van historische data en gegevens van de *on-line* gebruiker. Om hieraan inhoud te geven, wordt niet alleen gewerkt met klantbestanden (verzamelingen gegevens over klanten) maar ook met klantmodellen. Deze laatste zijn een weergave van de kenmerken, attitudes en mogelijke behoeften van een klant en laten derhalve toe om marketingacties op één-op-één basis aan te sturen. Data mining is hierbij eveneens een hefboom in de ontwikkeling van een lerende organisatie : een hulpmiddel om alle kennis die in een bedrijf aanwezig is gecontroleerd en efficiënt te gebruiken en om leerprocessen te versterken. Nauw samenhangend met de evolutie van database management naar kennismanagement, is het onderwerp van het beheer en de distributie van kennis binnen organisaties via intra/internet, *groupware* en *intelligent agents* [28].

De hier besproken ontwikkelingen blijven tenslotte ook niet zonder gevolgen voor het marktonderzoek. Met het in de plaats treden van klantconcepten voor productconcepten, zullen ook marktonderzoekers meer te maken krijgen met de analyse van klantgegevens en -modellen. Een hiermee verbonden ontwikkeling is dat het opstellen van klantprofielen en clusters naar aankoopgedrag samen met analyses van de klantlevenscyclus zullen primeren boven het verzamelen van steekproefgegevens. Het type informatie verschuift van opgegeven informatie naar gedragsinformatie[29]. Anderzijds vervangt informatie op individueel prospect- of klantniveau de meer generieke modellen en veronderstellingen. Klantsegmentaties zullen meer *bottum-up* gebeuren door samenvoeging van vrijwel identieke cases en minder *top-down* vanuit de opdeling van markten in homogene subgroepen. In hun geheel genomen hebben deze ontwikkelingen tot gevolg dat de marktonderzoeksfunctie in toenemende mate integreert met business intelligence en schuift het marktonderzoek op in de richting van informatiemanagement.

Noten

[1] Zie hierover Peppers & Rogers, 1993.

[2] Het is evenwel niet alleen de versnelde technologische vooruitgang die het ontwikkelingspad van de nieuwe media bepaalt. Ook de demassificatie van de consumentenmarkt drukt zijn stempel op de ontwikkeling van de nieuwe media. De door Faith Popcorn (1995) beschreven *cocooning* als een belangrijke trend in het hedendaagse consumentengedrag, vormt een gunstige voedingsbodem voor de groei van nieuwe interactieve media. Van 'nieuwe' media in de echte zin van het woord kan niet gesproken worden. De meeste nieuwe interactieve mediavormen zijn ontstaan als uitbouw of integratie van de bestaande communicatievormen : televisie, telefoon en computer. Vandaar dat men de nieuwe media ook wel multimedia noemt. Met nieuwe media wordt dus verwezen naar elk medium dat een directe, interactieve communicatie toelaat tussen de prospect of klant en de onderneming. Interactieve media worden in de eerste plaats gekenmerkt door informatieverstrekking op maat. De gebruiker gaat alleen in op datgene dat hem of haar interesseert. Een ander kenmerk van interactieve media is de vrijheid van timing. De gebruiker wordt niet op ongewenste momenten met een boodschap geconfronteerd. Het gebruik van het medium vindt plaats op het moment dat de gebruiker het zelf wil of op het moment dat daarvoor het meest geschikt is. Een interactief medium kan tenslotte de door gebruikers opgegeven input opslaan. Het medium wordt dan een krachtig instrument om meer te weten te komen over het gedrag van prospecten of klanten. Kenmerkend voor de hedendaagse kenniseconomie is de paradigmatische omslag waarbij de productoriëntatie (oude economie) geruild wordt voor informatie als centraal sturend element (nieuwe economie). Door de toegenomen connectiviteit in de netwerkmaatschappij dient informatie niet langer de lineaire stroom van de fysieke waardeketen te volgen (oude economie) maar kan informatie zich onafhankelijk van de goederenstroom voortbewegen (nieuwe economie) (Evans & Wurster, 2000). Een gevolg hiervan is dat de ''make and sell''-aanpak uit de oude economie de plaats ruimt voor een ''sense and respond''-benadering waarin het creeëren van klantwaarde vooropstaat (zie hierover o.m. Marcos, 2003).

[3] Lester Wunderman die in de jaren zestig het begrip 'direct marketing' introduceerde als uitbreiding van het direct mail-concept, heeft het in dit opzicht over het gebruik van het wereldwijde internet (De Standaard, 3 mei 1996). De voortschrijdende ontwikkeling van direct marketing , o.m. via internet, is in de visie van Wunderman een veruitwendiging van een nieuw sociaal-economisch model, nl. dat van de informatiemaatschappij waar we opnieuw tot een producent-consument relatie komen zoals we die hadden vóór de industriële revolutie. De markten waren toen klein, zowel in omvang als in geografische uitgestrektheid. Dit maakte dat producenten en verkopers nauwkeurig de koopgewoonten van elke klant kenden. Aan dit op landbouw geënte economisch model kwam na 1840 een einde met de uitvinding van de machines. Die stelden de producent in staat op grote schaal te mechaniseren. Het gevolg was seriewerk, daling van de prijzen, en uiteindelijk de ontwikkeling van grootschalige distributiesystemen zoals supermarkten en winkelketens en de aanwending van de massamedia om het aanbod bij een groot publiek kenbaar te maken.

[4] Voor een heldere uiteenzetting over de betekenis van telematica voor bedrijfsprocessen verwijzen wij naar een bijdrage van De Wit (1996).

[5] Een uitbreiding van het sterschema is het zgn. sneeuwvlokschema. In een sneeuwvlokschema worden de dimensietabellen die deel uitmaken van een stermodel genormaliseerd. Doordat queries in het geval van een sneeuwvlokschema kunnen uitgevoerd worden op kleinere, genormaliseerde tabellen in plaats van grote ongenormaliseerde tabellen, wordt een betere query-performance bekomen dan dit het geval is met een sterschema (zie hierover Gill & Rao, 1996, hfst. 5).

[6] Zie hierover Blomme (1997).

[7] Zie hierover Adriaans, Knobbe & Van der Hulst, 1996.

[8] Belangrijke producenten van software voor data mining zoals SAS en SPSS ontwikkelen een eigen methodologie die als raamwerk fungeert voor toepassing van data mining-technieken. In het geval van SAS wordt het proces van kennisontdekking opgedeeld in een vijftal stappen die onderdeel uitmaken van de zgn. *SEMMA*-methodologie (*Sample, Explore, Modify, Model, Assess*). Op analoge wijze wordt het KDD-proces door SPSS omschreven met behulp van hetgeen de 5A-benadering wordt genoemd (*Assess, Access, Analyze, Act, Automate*). Verder werd in

1998 door een Special Interest Group van een honderdtal leveranciers en consultants een standaard, genaamd CRISP-DM (*Cross Industry Standard Process for Data Mining* ; voor een bespreking : zie Shearer, 2000), ontwikkeld. Het afspreken van standaarden en procesmodellen is een factor die de acceptatie van data mining kan vergroten. Vaak gaat de aandacht bij data mining uit naar verkeerde factoren en wordt veelal aandacht geschonken aan de tools die kunnen/zullen worden ingezet. Het succes van data mining wordt slechts gedeeltelijk bepaald door de ingezette tools. Belangrijker in eerste instantie is de vraag welke de specifieke behoefte is aan kennis in de organisatie. Het proces van data mining moet worden voorafgegaan door de zgn. business vraag en een kosten-baten analyse. CRISP-DM is een aanzet om het data mining proces te laten vertrekken door het stellen van de business vraag. Tijdens de opstartfase (*business understanding*) wordt een projectplan opgesteld waarin het probleem nader omschreven wordt, samen met een kosten-baten analyse en de specificatie van de doelen en succescriteria. De fase van de gegevensoriëntatie (*data understanding*) behelst het selecteren van de data om het probleem te analyseren en op te lossen. Tijdens deze fase worden de gegevens geselecteerd, beschreven en verkend. In de daaropvolgende fase (*data preparation*) worden de gegevens voorbereid voor de werkelijke analyse- en modelleringsfase. Gegevens dienen opgeschoond te worden (bv. behandeling *van missing value*s) en vaak worden transformaties uitgevoerd (bv. variabelenreductie d.m.v. factoranalyse). Tijdens de modelleringsfase (*modelling*) wordt het algoritme op de gegevens toegepast. Hierbij wordt in veel gevallen gebruik gemaakt van een trainingset waarop modellen uitgeprobeerd worden en met een testset kan vervolgens nagegaan worden in hoeverre een geselecteerd model voldoet. Tijdens de evaluatiefase (*evaluation*) wordt nagegaan of het vereiste doel behaald is of niet. Data mining is een iteratief proces, hetgeen impliceert dat stappen kunnen worden herhaald. De zesde en laatste fase (*deployment*) betekent dat acties worden ondernomen om het probleem dat aan de basis lag van de data mining-exercitie op te lossen. Rekening houdend met het toenemend gebruik van verticale applicaties wordt de kennis opgedaan in het data mining proces geïntegreerd in bedrijfsprocessen (bv. door de toepassing van workflow-systemen). Data mining in de online variant maakt deel uit van primaire processen en vergt derhalve een andere inbedding dan de ad hoc-variant.

[9] Bron : Holsheimer, M. & Molenaar, C., 1998.

[10] Zie hierover Tukey (1977).

[11] Voor een uitgebreide beschrijving van data mining-technieken verwijzen wij naar Berry & Linoff (1997) en Brand & Gerritsen (1998a,b).

[12] A.I.D. staat voor *Automatic Interaction Detection* of contrastgroepenanalyse, tegenwoordig gekend als beslissingsbomen (*decision trees*). Een asymmetrische probleemstelling, een groot aantal variabelen en geen afgeronde hypothesen over de verwachte relaties tussen de variabelen kenmerken de situatie waarbij de behoefte aan exploratieve data-analyse aanwezig is en waarin contrastgroepenanalyse kan voorzien.

[13] In het geval van onderling sterk samenhangende onafhankelijke variabelen wordt het vertakkingsproces wel instabiel.

[14] Beslissingsbomen die gebruikt worden voor de predictie van categoriale variabelen worden “classification trees” genoemd. In het geval de predictie betrekking heeft op een continue afhankelijke variabele wordt gesproken van “regression trees” (zie Breiman e.a., 1984 ; voor een heldere beschrijving van classificatie- en regressietechnieken, zie Brand & Gerritsen, 1998b).

[15] Het “black box”-karakter van neurale netwerken waardoor een verklaring voor de gevonden verbanden achterwege blijft, wordt vaak aangevoerd als een zwak punt van deze techniek.

[16] Zie hierover Den Uyl & Langendoen (1997).

[17] Zie ook Jackson, 2002.

[18] Zie hierover ook Baragoin e.a., 2001, pp. 44.

[19] Zie hierover Calinski & Harabasz, 1974.

[20] Een correspondentie-analyse is toegepast op de gemiddelde waarden (*final cluster centers*) van de socio-demografische variabelen binnen elk van de vier clusters. Uit de analyse blijkt dat een tweedimensionele voorstelling 95,7 % van de variantie verklaart (resp. 76,4 % voor de eerste dimensie en 19,3 % voor de tweede dimensie).

[21] Het relatieve belang van beide dimensies blijkt eveneens uit de correlatieratio (*singular value*) die 0,309 bedraagt voor de horizontale dimensie en daalt tot 0,156 voor de verticale dimensie.

[22] Uit de tranformatiegrafiek voor v46 kan opgemaakt worden dat de waarden van deze variabele een ordinale schaal uitmaken. Het ordinale meetniveau is eveneens van toepassing op v49 en v68, maar in beide gevallen kunnen meerdere categorieën gehercodeerd worden (voor v49 is dit het geval voor de categorieën 2 tot en met 6 en voor v68 voor de categorieën 3 tot 5 en de categorieën 6 en 7). De lineaire tranformatiegrafiek voor v62 geeft aan dat deze als een numerieke variabele kan beschouwd worden.

[23]Data mining kan op twee manieren ontdekkend zijn, nl. gericht en ongericht. Gericht wil zeggen dat een bepaalde variabele wordt voorspeld, terwijl bij ongerichte data mining er vooraf geen te voorspellen variabele wordt meegegeven. In plaats hiervan wordt een algoritme op de data losgelaten dat een bepaalde structuur in de data tracht te vinden. Bij verzekeringsmaatschappijen kan gerichte data mining zorden gebruikt om de kans te voorspellen dat een klant, gegeven een bepaalde historie (zoals andere verzekeringspolissen die de klant bij deze verzekeraar heeft) een bepaalde nieuwe verzekering zal nemen. De klanten met de hoogste waarschijnlijkheid tot het nemen van deze nieuwe verzekering kunnen dan worden gemaild of gebeld. Dit kan het aantal respondenten op een dergelijke actie (ook 'lift' genoemd), aanzienlijk doen toenemen. Ongerichte data mining kan worden uitgevoerd om klanten in te delen in segmenten. Een tool kan dan de clusters ontdekken van klanten die heel interessant kunnen zijn. Een bank kan zo op het spoor komen van een cluster klanten die zowel een zakelijke als een persoonlijke rekening hebben, waarbij deze klanten ook vaker een hypotheeklening bij de bank afsluiten. Op grond hiervan kan de hypothese opgesteld worden dat die klanten eerder een hypotheek nemen om het vrijgekomen kapitaal te gebruiken voor het opstarten van een eigen zaak. De verschillende soorten data mining worden vaak in combinatie met elkaar gebruikt. Zo wordt in veel gevallen een ongerichte techniek op de data losgelaten, om daarna een gevonden cluster verder te onderzoeken met een gerichte data mining-algoritme.

[24] Een uitbreiding van deze validatiemethode bestaat erin de database niet alleen op te splitsen in twee delen, maar deze opdeling ook verscheidene keren willekeurig te herhalen waarbij de nauwkeurigheid van het voorspellende model na elke opdeling vastgesteld wordt. Elke keer dat de beschikbare records willekeurig in twee groepen worden verdeeld, wordt het voorspellende model op het ene deel gebouwd en wordt de nauwkeurigheidstest op het andere deel uitgevoerd. Deze methode wordt kruisvalidatie (*cross validation*) genoemd.

[25] In het geval van de toepassing van logistische regressie-analyse kan de performantie van de classificatieregels direct afgelezen worden uit de *confusion matrix*.

[26] EAI-producten worden ontwikkeld vanwege de behoefte aan transactie-integratie tussen diverse bedrijfsapplicaties. Zo moet een nieuw order in het ERP-systeem doorgegeven worden aan het financiële systeem en moet het ook ingevoerd worden in het verkoop informatie systeem. EAI is gericht op het propageren van berichten (transacties) met het doel om acties teweeg te brengen en niet zozeer op het verzamelen en transformeren van grote hoeveelheden data voor analyse. EAI-producten zijn gebaseerd op gebeurtenissen (*events*) waarbij hoofdzakelijk kleine hoeveelheden data in real time worden doorgegeven (transactie per transactie) van de ene applicatie naar een of meerdere andere applicaties (zie hierover Oosterhof, 2002). Een onderscheid dient gemaakt te worden tussen real time data warehousing (RTDW) en active data warehousing (ADW). RTDW verwijst naar het geheel van operaties, uitgevoerd met behulp van ETL- en EAI-software, om het datawarehouse te voorzien van historische en vooral de meest recente data. Hoewel ADW verwant is met RTDW verwijst ADW niet naar een technologie om het datawarehouse up-to-date te houden naar is het een concept dat verwijst naar de rol van het datawarehouse in de organisatie. Wat het ADW echt 'actief' maakt, is dat het verwijst naar de essentiële rol van het datawarehouse, in real time, met andere operationele systemen zoals CRM, ERP en SCM. Het gegeven dat recente data geladen worden in een datawarehouse (RTDW) impliceert niet onmiddellijk dat met die recente gegevens gewerkt wordt. Daartegenover staat dat in het geval van ADW de meest recent beschikbare data aangewend worden in operationele toepassingen waarbij deze gegevens mede de basis vormen voor event-gedreven beslissingen (zie hierover Raden, 2003). ADW wordt ook aangeduid als business activity

minitoring (BAM) : ''The need for real-time visibility and immediate response capabilities has given rise to business activity monitoring (BAM), a process by which key operational business events are monitored for changes or trends indicating opportunities or problems, and enabling business managers to take corrective action. … For example, a BAM system in a customer service environment might quickly alert a manager to a top customer's call. … By contrast, a non-BAM (or batch) system requires extra, time-comsuming steps to perform the correlation, and the manager might not be alerted to the issue until hours later, when the customer has been lost. … BAM systems won't replace data warehouses. By definition, a data warehouse is a place to accumulate or aggregate data for subsequent analysis. Typically, users look to data warehouses for historical analysis and planning functions within a business. As a result, the data warehouse generally contains fixed data models and permanent data storage, available for user-specified analytical ''roll-ups''. However, for operational functions (e.g. supply chain, customer interaction, logistics) within a company, the need for up-to-date information and business rules governing event-specific action becomes very important, requiring a BAM solution that is designed to serve operational needs'' (Nesamoney, 2004).

[27] De evolutie waarvan de de distributie van kennis via een portaalstructuur het voorlopige eindpunt is, kan meer in het algemeen ingepast worden in de ontwikkeling van functionele naar netwerkende organisaties (zie hierover o.m. Van Der Zee, 2000 ; Molenaar & Molenaar, 2003).

dimension	DP era	IT era	network era
primary role of IT	data (transaction) processing	information provision	knowledge distribution
technology focus	Mainframe	distributed IT (ERP)	omnipresent and ubiquitous (Web connected)
restructuring tools	TQM (improving functions)	BPR (restructuring processes)	restructuring industry chains
business objectives	cost reduction (efficiency)	service improvement (effectiveness)	customer satisfaction (flexibility)
organization focus	Functions	business processes (intra-enterprise)	extended enterprise

Tot in de jaren tachtig van vorige eeuw had automatisering een hoofdzakelijk interne oriëntatie. Kenmerkend voor de automatisering van de hiërarchisch-functionele organisatie was de nadruk op kostenefficiëntie en het inrichten van de verschillende bedrijfsfuncties (inkoop, logistiek, verkoop) rond product-marktcombinaties. De verkoopsorganisatie was transactie-georiënteerd en werd IT-matig ook als zodanig ondersteund. De technologische ontwikkelingen maakten het vanaf de jaren tachtig mogelijk om informatie voor meerdere functies beschikbaar te stellen. In het tijdperk van de gedistribueerde IT (dat langzamerhand ten einde loopt) staan niet langer bedrijfsfuncties centraal maar vormt automatisering de basis voor de structurering en reorganisatie van bedrijfsprocessen (BPR, *Business Process Reengeneering*). De ontwikkelingen op IT-vlak (met name de gedistribueerde verwerkingskracht) maakten het immers mogelijk om sneller in te spelen op maatschappelijke veranderingen. Door het toenemend belang van het vermogen tot innovatie (nieuwe producten en diensten, nieuwe distributiekanalen) werd automatisering de basis van het structureren van processen over bedrijfsfuncties heen. En dat dit alles gebeurt met inachtneming van de verwachtingen van de klant duidt op het strategisch belang van informatie. In het netwerktijdperk neemt de verstrengeling van IT met het zakendoen en vooral de omgeving waarin dit gebeurt nog verder toe. De grenzen van de individuele organisatie vervagen. In de netwerkorganisatie grijpt een toenemende vervlechting plaats tussen IT en producten, diensten, distributiekanalen, werkprocessen en organisatiestructuren waardoor de samenwerking tussen de eigen organisatie en leveranciers, partners en klanten de kenmerken aanneemt van een virtuele organisatie waarin de toegankelijkheid en beschikbaarheid van (klanten)kennis centraal staat.

[28] De toenemende vraag naar een individuele klantbenadering en faciliterende trends zoals een betere gegevensverzameling, snellere computers en goedkopere software spelen een belangrijke rol bij de stijgende populariteit van data mining. Als we deze trend consequent voortzetten, ligt het voor de hand te voorspellen wat er de komende jaren zal veranderen aan de toepassing van data mining : het wordt een instrument van de klant. Een instrument om zelf de behoefte te kunnen bepalen, het juiste productaanbod bij de vraag te zoeken en te

onderhandelen over prijs en productspecificaties. Het valt te verwachten dat op elektronische marktplaatsen van de toekomst intelligent agents de klant zullen helpen bij het zoek-, selecteer- en onderhandelingswerk. Namens de klant zullen deze assistenten interessante product- of behoefte-ideeën genereren en gaan onderhandelen met de best matchende aanbieders met de best matchende producten om tot shortlists te komen waaruit de klant een keuze kan maken. Als de koop gesloten is en het product gebruikt wordt, geeft de agent melding van service momenten en slaat de ervaringen met het product op. Naarmate meer en meer koopprocesen doorlopen zijn, leert de agent zijn ''baasje'' steeds beter kennen. Ook bij dit leergedrag zal data mining, als intelligente methode om kennis uit data te destilleren, een belangrijke rol spelen. Er is dus sprake van synergie tussen elektronische marktplaatsen, data mining en intelligent agents (zie hierover Van Der Putten & Den Uyl).

[29] Ondanks de verschillen tussen marktonderzoek (waarin vooral attitudinale kenmerken het voorwerp van analyse uitmaken) en data mining (waarin klantgedrag central staat), komen de inzichten van beide disciplines elkaar ten geode en verdient het zonder meer aanbeveling om zowel attitudes als gedrag als basis voor predictieve analyse te gebruiken (zie hierover Elliott, Scionti & Page, 2003 ; SPSS Magazine, 2004).

Geraadpleegde Literatuur

Adriaans, P., Knobbe, A. & Van Der Hulst, M.P., *Data mining and fuzzy databases*, Syllogic (Nl.), 1996.

Alan Bonde Group, *Real-world business intelligence and enterprise reporting : Key trends and best practices*, Research Study, april 2002.

Altius Consulting, , *The changing face of business intelligence*, UK Compass 2001 World Tour, 2001.

Atre, S. & Malhotra, D., Real-time analysis and data integration for BI, *DM Review*, February 2004.

Baesens, B., Mues, Chr. & Vanthienen, J., Knowledge discovery in data, *Informatie*, januari-februari 2003, pp. 30-35.

Baragoin, C., Andersen, Ch.M., Bayerl, S., Bent, G., Lee, J. & Schommer, Ch., *Mining your own business in banking*,IBM Redbooks, 2001.

Barth, T., *Guidelines for the data mining process*, Critikal Consortium, University of Stuttgart, 1998.

Bates, J., Business in real time. Realizing the vision, *DM Review*, may 2003.

Berry, M.J. & Linoff, G. , *Data mining techniques for marketing, sales and customer support*, New York, John Wiley & Sons, 1997.

Berson, A. & Smith., S.J., *Datawarehousing, datamining en OLAP*, Schoonhoven, Academic Service, 1997.

Blomme, J., *Het relationele model en normalisatie,* Damme*,* 1997.

Blomme, J., *E-business integration. Enabling the real-time enterprise*, Damme, 2003.

Brand, E. & Gerritsen, R., Data mining and knowledge discovery, *DBMS*, vol. 11, nr. 9, 1998a, pp. 52-55.

Brand, E. & Gerritsen, R., Classification and regression, *DBMS*, vol. 11, nr. 9, 1998b, pp. 56-61.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., *Classification and regression trees*, Belmont, CA, Wadsworth, 1994.

Calinski, T. & Harabasz, J., A dendrite method for cluster analysis, *Communication in Statistics*, vol. 3, 1974, pp. 1-27.

Den Hamer, P., Enterprise information portals en realtime datawarehousing. Kraakverse data uit éé loket, *Database Magazine*, mei 2001, nr. 3, pp. 16-18.

Den Uyl, M.J. & Langendoen, E., De inzet van adaptieve analysetechnieken in direct marketing, in A.E. Bronner e.a. (red.), *Recente ontwikkelingen in marktonderzoek*, Jaarboek 1997 van de Nederlandse Vereniging voor Marktonderzoek en Informatiemanagement, Haarlem, Uitgeverij de Vrieseborch, 1997, pp. 107-121.

De Wit, R.M., *Het creëren van toekomst voor uw business met telematica*, Beam'IT Position Paper, 1996.

Elliott, K.E., Scionti, R. & Page, M., *The confluence of data mining and market research for smarter CRM*, SPSS Inc & The Kantar Group, april 2003.

Evans, Ph. & Wurster, Th.S., *De nieuwe economie*, Amsterdam, Business Contact, 2000.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., *From data mining to knowledge discovery in databases*, American Association for Artificial Intelligence, 1996.

Fransen, E., *Analytic applications : Derde generatie business intelligence*, CIBIT Business Intelligence Whitepaper, juni 2003.

Gill, H.S. & Rao, P.C., *De client/server gids voor data warehousing*, Schoonhoven, Academic Service, 1996.

Hashmi, W., *Business intelligence in collaborative business model and business performance*, Information Frameworks, october 2003.

Holsheimer, M. & Molenaar, C., *Marketing and data mining produce a client profile*, 1998.

Hostmann, B. & Buytendijk, F., *Management update : Effective BI approaches for today's business world*, Gartner Note (IGG-04142004-01), 14 april 2004 .

Inmon, W.H., *Building the data warehouse*, New York, John Wiley & Sons, 1996.

Jackson, J., Data mining : A conceptual overview, *Communications of the Association of Information Systems*, vol. 8, 2002, pp. 267-296.

Kamst, F., Trends in ETL-tools, *Database Magazine*, juni 2002, nr. 4, pp. 23-27.

Klemenhagen, B., *Business intelligence – The missing link*, Cherry Tree & Co., july 2000.

Kohavi, R., Rothleder, N.J. & Simoudis, E., Emerging trends in business analytics, *Communications of the ACM*, vol. 45, no. 8, 2002, pp. 45-48.

Malhotra, Y., Knowledge management for E-business performance : Advancing information strategy to ''Internet Time''*, Information Strategy*, vol. 16, nr. 4, summer 2000, pp. 5-16.

Marcos, A., *A new maketing paradigm in the knowledge economy*, 2003.

Michielsen, T., Internet voert ons terug naar de preïndustriële tijd, *De Standaard*, 3 mei 1996.

Mohanty, S., Multitier architecture for high performance data mining, *DM Direct Special Report*, july 2004.

Molenaar, C., *Interactieve marketing. Het einde van de massamarketing*, Brussel, Management Bibliotheek, 1992.

Molenaar, C. & Molenaar, S.*, De impact van de ik-cultuur op maatschappij, marketing en organisatie*, Amsterdam, Pearson Education, 2003.

Moncla, B., *The convergence of E-business and business intelligence : The I-business hurricane*, ThinkFast Consulting, 2000.

Nesamoney, D., BAM : Event-driven business intelligence for the real-time enterprise, *DM Review*, march 2004.

Oosterhof, B., EAI en ETL : De integratie-uitdaging, *Beyond*, jrg. 7, nr. 4, september 2002, pp. 33-37.

Ortiz, S., Is business intelligence a smart move ?, *Industry Trends*, july 2002.

Parsaye, K., From data management to pattern management, *DM review*, january 1999.

Peppers, D. & Rogers, M., *The one-to-one future. Building business relationships one customer at a time*, London, Piatkus, 1993.

Popcorn, F., *Trends van overmorgen*, Amsterdam, Contact, 1995.

Postma, P., *Het nieuwe marketing tijdperk*, Amsterdam, Contact, 1996.

Pursell, D., *Turning information into knowledge*, SPSS, june 2002.

Raden, N., *Exploring the business imperative of real-time analytics*, Hired Brains, Inc, october 2003 (E2231 1103).

Rudin, K. & Cressy, D., Will the real analytic application please stand up ?, *DM Review*, march 2003.

Shearer, C., The CRISP-DM model : The new blueprint for data mining, *Journal of Data Warehousing*, fall 2000, pp. 13-22.

Schipper, H., De informatiecirkel gaat sluiten, *Database Magazine*, november 2001, pp. 40-47.

Schultz, D.E., (1990), *Strategic newspaper marketing*, I.N.M.A., Reston, Virginia.

Data mining en marktonderzoek : blijven het twee gescheiden stromen of komen ze samen ?, *SPSS Magazine*, 2, 2004, pp. 10-11.

Thearling, K., (1998), *Increasing customer value by integrating data mining and campaign management software*, Boston, Exchange Applications.

Tukey, J., *Exploratory data analysis*, Cambridge, MA, Addison Wesley, 1977.

Van Der Putten, P. & Van Someren, M. (eds.), *COIL challenge 2000 : The insurance company case*, Sentient Machine Research, Amsterdam, 2000.

Van Der Putten, P. & Den Uyl, M., Mining E-markets. *IT Monitor*, march 2001.

Van Der Zee, H.T.M., *Business Transformatie en IT*, Rede, Katholieke Universiteit Brabant, 2000.

Van Lent, R. & Zwietering, D., Technical integration of data mining, *DM Review*, september 2002.

Van Nieuwenhuysen, W. & Smulders, T., Dataminingtechnieken verplichten eens te meer tot nadenken, *Beyond*, Jrg. 4, nr. 4, december 1999.