”Data Mining”
En revolution eller ännu ett analysverktyg?

Abstrakt

I vår tid råder det ingen brist på information, utan snarare ett överflöd. Företag fyller sina databaser med data som de tror sig kunna dra nytta av. Problemet är att förmågan att analysera och dra nytta av data är lägre än förmågan att samla in och lagra den. Med data mining kan företag analysera data och få fram mönster och samband som kan användas som underlag för beslut. I försäkrings- och kreditkortsbranschen finns det stora potentialer för användningen av data mining eftersom stor möda läggs på marknadsundersökningar, där stora mängder information om kunders behov bearbetas. I denna uppsats har vi studerat två bolag, ett i försäkrings- och ett i kreditkortsbranschen, för att undersöka deras möjligheter att använda data mining. Vi har även undersökt hur bolagen kan använda en data mining produkt, SAS Institutes Enterprise Miner. Studien visar att det är viktigt vid användningen av data mining att vara införstådd med vad som egentligen undersöks samt att besitta den kompetens som krävs för att tolka resultatet. Det är också viktigt att inte underskatta behovet av tids- och kompetensresurser för att implementeringen av data mining ska lyckas.

Catrin Andersson och Annika Elfström

Institutionen för Informatik **Göteborgs Universitet**
Magisteruppsats 20 p, vt 1998

Abstract

In our age there is no lack of information rather an information overload. Companies are filling their databases with data they think some valuable assets are coded within. The problem is that the ability to analyse and understand massive datasets lags far behind their ability to gather and store data. With data mining companies can easily analyse the data and find new patterns and relationships to use as a decision support. Specially suitable for data mining are insurance- and creditcard companies, because they work with a large amount of data and put a large amount of time on campaignmarketing. In this examination paper we have studied two companies, one insurance company and one creditcard company to investigate how they can use data mining to improve their ability to analyse their data and if they can use SAS Institutes Enterprise Miner, a product for data mining. The results show the importance to understand what to analyse and how to interpret the results of data mining. The results also show the need for time and competence for a successfully implementation of data mining.

Förord

Det finns ett antal personer vi skulle vilja tacka:

Ulf Johanisson och Bosse Falk, vår handledare på Volvo IT, för stöd under arbetets gång.

Etnogruppen med Magnus Bergqvist och Nina Lundberg, för uppmuntran, stöd och glada tillrop. Speciellt tack till Anette Stahl och Ilona Wearn för värdefulla förslag under skrivandets gång.

Henrik Fagrell, vår handledare på Institutionen för Informatik, du har varit oerhört hjälpsam.

Sist men inte minst vill vi tacka:
Birgit Andersson, för hökögon och tålamod.
Sune Andersson, för "coaching" och en snabb bil.

Innehållsförteckning

1. Introduktion

Under vår utbildning vid Institutionen för Informatik har vi läst kurser som hanterar beslutstödssystem och artificiell intelligens. Vi fann detta intressant och beslutade oss för att skriva en uppsats om beslutstödssystem. Data warehouse är ett koncept utvecklat för att stödja ledningens informationsbehov som ett beslutstödssystem. För att effektivt kunna använda data warehouse behövs det en teknik som kan finna mönster, trender och samband i data, en sådan teknik är data mining. Vi fann detta intressant och ville undersöka hur tekniken fungerar och skriva vår magisteruppsats om data mining. Vi tog kontakt med Volvo Data AB (Volvo IT fr o m januari 1998) som i januari startade ett projekt för att betatesta en data mining produkt, SAS Enterprise Miner. Projektet gick ut på att undersöka möjligheterna att använda Enterprise Miner på Volvias kampanjdata. Vi har varit med i projektet för att undersöka hur Volvia och Volvokort kan använda sig av data mining.

1.1 Tidigare arbeten

Under de senaste åren har data mining varit ett ofta diskuterat och debatterat område inom forskningen kring beslutstödssystem. År 1996 ägnades ett specialnummer av Communications of the ACM åt data mining och Knowledge Discovery in Databases. Det finns ett flertal web-sidor och forum som tillägnats området, där de mest framstående auktoriteterna inom området debatterar. Det har gjorts ett flertal examensuppsatser inom området data warehouse och data mining, dels på Chalmers tekniska högskola och dels på andra universitet och högskolor runt om i Sverige. De flesta examensuppsatser som gjorts handlar om data warehouse, till exempel ”Trettio års problem med datoriserat beslutstöd - Kan Data Warehouse vara lösningen?” som skrevs av David Ericson och Robert Ericsson vid Umeås Universitet 1995. Uppsatsen undersöker möjligheterna att använda data warehouse för att förbättra användningen av beslutstödssystem. På Chalmers tekniska högskola i Göteborg skrev Magnus Björnsson 1997 en uppsats med titeln “En jämförelse av data mining algoritmer för klassifikation”. Magnus Björnssons uppsats fokuseras på algoritmer som testas på stora och medelstora datamängder och sedan jämförs utifrån testresultatet. På Institutionen för Informatik skrevs under hösten 1997 en uppsats av Jonas Landgren med titeln “Data Warehouse and Data Mining”. Jonas Landgren förklarar i sin uppsats begreppen data warehouse och data mining samt hur de kan användas i en organisation för att bland annat förkorta led- och leveranstider.

1.2 Syfte

Syftet med vår magisteruppsats är att undersöka hur data mining kan användas inom kreditkorts- och försäkringsbranschen. För att kunna undersöka de två olika branscherna kommer vi att studera två olika bolag inom Volvokoncernen, Volvia som är ett försäkringsbolag och Volvokort som är ett kreditkortsföretag. I undersökningen kommer vi även att studera om det finns ett verktyg för data mining som kan uppfylla kraven för användning och producera resultat, i vår uppsats har vi testat SAS Enterprise Miner. Vi kommer att undersöka det begreppet data mining både praktiskt och teoretiskt.

I den här rapporten undersöker och diskuterar vi kring följande frågeställningar.

1) Hur kan bolagen använda sig av data mining och kan de använda sig av SAS Enterprise Miner ?
2) Vilka är de kritiska faktorerna för att lyckas införa data mining i en verksamhet?

Vi kommer även att presentera ett lösningsförslag samt kritiska faktorer för att lyckas införa data mining i verksamheten. I vårt teoriavsnitt presenteras mer av tidigare studier och forskning inom data mining.

1.3 Presentation av bolagen

Volvia och Volvokort är en stödjande verksamhet till Volvo, för att kunna erbjuda en helhetslösning för Volvos kunder. De köper en Volvobil och försäkrar den hos Volvia och får därigenom en förmånlig bilförsäkring. En Volvoägare kan också skaffa sig ett Volvokort som är ett förmånskort inom Volvohandeln. Här nedan kommer vi att presentera bolagen lite mer ingående.

Volvia

Volvia är ett försäkringsbolag beläget i östra delarna av Göteborg. Grunden till Volvia lades då Volvo införde ett femårigt garantisystem för alla sina nya bilar år 1954. Volvia startade sin verksamhet år 1959 för att ta hand om kunder med Volvogaranti. De har idag 170 anställda samt ett antal personer som arbetar kvällstid med tele-marketing (TM), vilket går ut på att sälja försäkringar via telefon. Volvia försäkrar Volvo- och Renaultbilar och har 35 % av alla Volvobilar försäkrade, vilket gör dem till ett förhållandevis stort försäkringsbolag då det gäller bilförsäkringar. Volvia har en fördel jämfört med andra bilförsäkringsbolag eftersom de är ett Volvobolag med ett nära samarbete med Volvohandeln och kan tillsammans med dem erbjuda en helhetslösning för sina kunder.

Volvokort

Volvokort startade sin verksamhet i Göteborg 1985 och har idag ca 75 anställda. Verksamheten startades då idén att sälja bensin via ett kort skulle testas. Efter visat intresse för kortet kopplades det även andra förmåner till det. De som kan få ett Volvokort är ägare till Volvo- eller Renaultbilar samt Volvoanställda. Med kortet får kortägarna rabatt på bensin och förmånliga priser på tillbehör och reservdelar i Volvohandeln. De kan även utnyttja kortet i verkstäder och butiker, som tillhör Volvos verksamhet eller som har avtal med Volvokort, med fyra räntefria månader. Andra verksamheter utanför Volvohandeln där ett Volvokort kan användas är till exempel biluthyrning, parkerings- och telefonautomater.

1.4 Rapportens disposition

Uppsatsen är i huvudsak uppdelad i fem huvuddelar: introduktions-, teori-, metod-, resultat- och diskussionsdel. Förutom dessa delar finns det en del med referenser och en del med appendix och bilagor. En kortfattad genomgång av de olika delarna följer här nedan.

Kapitel 2 är teoridelen, där olika begrepp presenteras och där vi beskriver vad data mining är, bakgrund till data mining, olika tekniker, metoder och teorier. I teoridelen presenteras även olika områden data mining traditionellt sätt använts på, samt kritiska faktorer för att lyckas med att införa data mining i en verksamhet.

I **Kapitel 3** förklarar vi den metod vi använde oss av vid litteraturstudier och vid studien av de två Volvobolagen.

Kapitel 4 innehåller de resultat vi kom fram till under vår studie och analys av Volvia och Volvokort.

I **Kapitel 5** diskuterar vi kring våra resultat och varför vi har fått de resultaten. Här presenterar vi också vår slutsats.

Kapitel 6 visar vår referenslista.

Bilaga 1 Rich Picture för Volvia, det vill säga en överblick av Volvias verksamhet såsom vi ser den.

Appendix A innehåller kortfattade förklaringar till de begrepp vi använder oss av i uppsatsen.

2. Teori

I detta avsnitt redovisar vi det teoretiska ramverk vi använder oss av under resultat- och diskussiondelen. Vi har även inkluderat en presentation av data mining och dess ursprung för att belysa området och tidigare forskning. Efter inledningen kommer vi att presentera skillnaden mellan data, information och kunskap för att definiera hur vi kommer att hänvisa till de olika begreppen i uppsatsen. I teoriavsnittet kommer vi att förklara begreppen data warehouse och data mining samt hur de är relaterade till varandra.

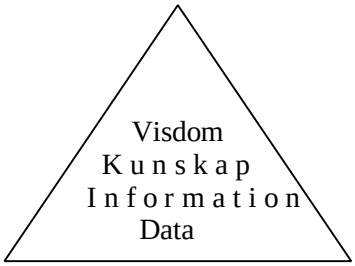
2.1 Data - Information - Kunskap

Data, information och kunskap används flitigt inom vårt område men definitionerna skiljer sig dock något åt beroende på författare. Vi kommer här nedan att presentera några författares definitioner av begreppen för att påvisa skillnaderna mellan data, information och kunskap samt att göra läsaren medveten om svårigheten att särskilja begreppen.

Ordet *Data* kommer ursprungligen från latinets do, dare som översätts till "att ge", vilket enligt Schoderbeck med flera (1990) ska ses som den stora mängd ostrukturerad data datorerna ger oss. *Information,* menar de, är data som har en form, struktur och organisation med ursprung från latinets informo, informare vilket översätts till "att ge form", det vill säga data som fått en betydelse för oss och blir information, som senare kan bli kunskap. Data skall, enligt Andersen (1991), ses som en samling tecken och symboler utan något värde för oss men som fungerar som bärare av information. Data som tolkas ger oss ett värde i form av information. Det finns flera författare som motsäger Andersens uppfattning att data endast är symboler utan betydelse för oss[1].

Samband mellan data, information och kunskap

För att beskriva sambandet mellan data, information, kunskap och visdom har Knight & Silk (1990) illustrerat sambandet med följande hierarki (se figur 2.1).



Figur 2.1 Beskriver sambandet mellan data, information, kunskap och visdom enligt Knight & Silk (1990).

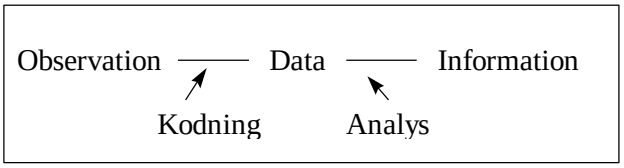
[1] Bland annat Langefors (1966) som vi presenterar senare.

Kunskap är något som bygger på relevant information, teoretisk såväl som praktisk. De tre översta lagren skiljer sig åt på så sätt att de bygger på mänsklig inblandning. Visdom beskrivs som kunskap som bygger på erfarenhet tillsammans med en persons omdöme. De lägre delarna, data och information finns i större utsträckning än kunskap och visdom, enligt Knight och Silk (1990).

Tolkningsprocessen

För att få fram information ur en datamängd måste en tolkningsprocess genomföras. Det råder delade meningar bland författare om hur tolkningsprocessen ser ut. Här nedan följer två exempel på hur två författare beskriver processen, dels David Shepard Associations och dels Börje Langefors, som ses som grundaren av den akademiska inriktningen System theory i Skandinavien (Dahlbom 1993).

David Shepard (1995) beskriver dataanalys-processen med följande steg (se figur 2.2):



Figur 2.2 Beskriver tolkningsprocessen enligt David Shepard Association (1995)

Processen utgår från observation av en händelse eller ett resultat som kodas till data i form av variabler. Dataanalysen genomförs och slutligen får man fram informationen. I bilden beskriver Shepard (1995) data som ett resultat av en observation. Denna uppfattning utgår från att data har ett värde eftersom en verklighet har observerats. Här kan man inte se på data som något som saknar värde (Andersen 1991), vilket illustrerar författares olika uppfattningar om data.

Infologisk Ekvation

En annan tolkningsprocess beskrev Börje Langefors (1966), då han skapade den infologiska ekvationen som säger att:

$$\mathbf{I = i\ (\ D, S, t)}$$

Där **I** är den information man får ut av tolkningsprocessen då **i,** operatorn på data **D** med **S** som förkunskap under tiden **t**. Tolkningsprocessen styrs av individens förkunskaper och världsbild. Langefors menar vidare att för att två personer ska få ut samma information I, från samma data D, måste de ha samma förkunskaper S. Detta kan leda till problem eftersom förkunskaper avgörs utifrån vilken erfarenhet personerna har och det är inte troligt att deras erfarenheter är identiska med varandra. Enligt Magolus och Pessi (1998) kan information I, beskrivas som "skillnaden mellan den verklighetsbild som individen hade innan denne motog meddelandet (S) och den verklighetsbild som individen skapande efter tolkningen av

meddelandet (S'). I den meningen är information ett tillskott till människans kunskaper" (Magulas & Pessi 1998).

I = S' + S samt S = S + I

2.2 Information och konkurrens

I företag finns det flera informationsmiljöer som består av människor, objekt (som kan vara artefakter) och händelser. Informationssystem är artefakter som befinner sig i en informationsmiljö. Det finns olika sorters informationmiljöer och deras förhållande till information skiljer sig något åt. De är främst tre olika som presenteras av Magoulas och Pessi (1998), dels ser man på information som en resurs eller som ett kunskapstillskott eller så ser man information som en maktfaktor[2].

Information som konkurrensmedel

Under de senaste årtiondena har konkurrensen mellan företag hårdnat, vilket har lett till att företag har utvecklats från att vara produktfokuserade till att bli mer kundfokuserade (Shepard 1995). Det har satsats mer tid och pengar än någonsin på att få fram information om företagets kunder som kan stödja verksamheten och utveckla en framgångsrik kundrelation. Konkurrensen om kundernas lojalitet har hårdnat både på produktmarknaden och på den tjänsteproducerade marknaden, vilket leder till att allt mer investeras i att behålla kunderna och att locka till sig nya lönsamma kunder. En nyckeluppgift för att öka konkurrenskraften är effektiv marknadsföring.

Enligt Shepard (1995) har marknadsföringen sedan 1980-talet blivit alltmer inriktad på kunder där marknadsföringen ska ske med så kallad 1-till-1 metod, vilken går ut på att den blir mer personligt riktad efter kundens förutsättningar. Företag går i allt större utsträckning ifrån att skicka ut massupplagor av till exempel kampanjer, på grund av den dåliga lönsamheten, till att ha riktade kampanjer. Många mindre företag har en fördel då det gäller att få en god kundrelation, vilken bygger på att de noterar kundens behov, kommer ihåg produkter och tjänster kunden tidigare köpt. På så sätt lär de sig hur de skall tillfredsställa kunderna bättre i framtiden, vilket för ett mindre företag inte är något problem eftersom kundgruppen är så liten och att företagets anställda mer eller mindre har personlig kontakt med kunderna. Det är företagets kunskap om kunderna som får dem att vara trogna företaget. Då företag måste bli mer kundfokuserade gäller det för större företag att efterlikna den kundrelation mindre företag kan bygga upp. Berry och Linoff (1997) ställer sig frågan hur ett stort företag ska kunna efterlikna ett litet företag. Ett första steg är att samla på sig så mycket information som möjligt om sina kunder, vilket redan görs i stora företag. OLTP, Online Transaction Processing system, används för att samla in all sorts information som kan vara användbar i ett senare läge. Kunders transaktioner sparas för att senare analyseras. Kundfokuserade företag måste se varje interaktion med en kund eller en blivande kund som en möjlighet att lära sig något nytt om kunden, exempelvis varje samtal till kundsupporten, varje transaktion, varje katalogbeställning och varje besök på företagets web-sida (Berry & Linoff 1997).

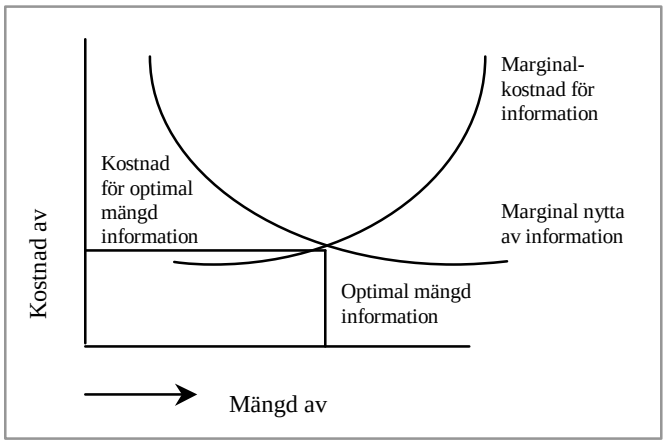
[2] För vidare referens se Magoulas och Pessi (1998).

Företags Informationsbehov

Behovet av information har förändrats radikalt från att bara vara en del av företaget till att nu vara en Critical Success Factor, CSF, det vill säga en nyckel till framgång. Vad beror förändringen på? En anledning att företagets förutsättningar har förändrats är på grund av den digitala revolutionen, som har bidragit till att företag…: (Dilly 1997)

1. …lätt kan samla på sig och förhållandevis billigt lagra data i stora databaser med exempelvis OLTP (Online Transaction Processing System) för att snabbt, enkelt och säkert föra in data i databasen. En svaghet hos OLTP system är att de inte kan användas för att få fram en användbar analys av data som lagrats.
2. …i allt större utsträckning kan använda sig av insamlingsverktyg exempelvis så kallad Point-of-sales, som går ut på att företag sparar all data då en kund till exempel gör ett inköp i en affär.
3. …kan använda DBMS (Database Management System) för att ta fram en mindre del av det som sparats i företagets databas.

Det gäller att få fram rätt information vid rätt tidpunkt. Företagens viktigaste tillgångar är i många fall inte längre land, råmaterial och arbetare, utan information, vilken bör ses som en resurs och måste hanteras därefter. Beslut som gäller att skaffa fram mer information ska behandlas på samma sätt som ett beslut att skaffa fler maskiner i en maskinpark (Schoderbeck m.fl. 1990). Innan man skaffar fram information måste nyttan av informationen tänkas igenom. Information är en färskvara och den måste behandlas som en. Den optimala mängden information enligt Schoderbeck m.fl. (1990) fås genom en marginalanalys så kallad "trade-off" analys (se figur 2.3). Figuren illustrerar problemet med den kompromiss företag tvingas göra mellan kostnaden för informationen och nyttan av införskaffad information.



Figur 2.3 beskriver sambandet mellan optimal kostnad för information och optimal informationsmängd (Schoderbeck 1990)

Ett problem som följer med stora mängder data som samlas är *information overload* (Denning 1982), det vill säga informationsmängden är så stor att det inte går att ta fram något meningsfullt ur den och informationen går förlorad. En negativ effekt av information overload

är att företag lagrar hundratals gigabyte utan att lära sig något om sina kunder. Gregory Piatetsky-Shapiro uttrycker problemet med information overload så här:

"Computers promised us a fountain of visdom, but delivered us a flood of data"
Gregory Piatetsky-Shapiro (1996)

Det krävs analyser och verktyg för att hantera den stora mängd data som lagras i företagets databas. Under de senaste åren har det utvecklats en mängd verktyg och tekniker, bland annat verktyg för data mining, som inte bara organiserar data utan som även skall användas till att ta fram nya samband och mönster i datamängden.

2.3 Beslutstödssystem (Decision Support System, DSS)

Fuori och Gioia (1994) förklarar att syftet med beslutstödssystem är att låta användaren interagera med en dator. Med ett beslutstödssystem kan användaren ställa frågor och genast få svar. Informationen presenteras i form av diagram eller grafer.

År 1978 skrev Keen och Scott-Morton (Turban 1995) en klassisk definition på beslutstödssystem:

"*Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semi-structured problems.*"

En annan definitionen ger Little från 1970 (Turban 1995) som förklarar ett beslutstödssystem som:

"*model-based set of procedures for processing data and judgement to assist a manager in his decision making*"

Det finns ett antal kriterier som ett beslutstödssystem bör uppfylla för att kunna klassificeras som ett och för att det kan vara ett stöd när ledningen ska fatta beslut. Enligt Little (Turban 1995) finns det sex stycken kriterier:

1) Det ska vara enkelt att förstå.
2) Systemet ska vara robust.
3) Det ska vara lätt att kontrollera.
4) Ett beslutstödssystem ska vara användarvänligt.
5) Det ska innehålla viktig och relevant information.
6) Systemet ska även vara lätt att interagera med.

Uppfyller systemet alla sex kriterierna ses det som ett välfungerade beslutstödssystem.

2.4 Data warehouse

Data warehouse är utvecklad för att ta fram viktig information ur stora datamängder som finns lagrade i databaser. Alla anställda i ett företag bör ha tillgång till att använda företagets data warehouse, men det är speciellt utvecklat för att tillgodose ledningens informationsbehov och anses därför vara ett beslutsstödssystem (Hadden 1997a). De olika användarna hanterar olika analysverktyg för att ta fram den information som de söker. De flesta företag har enorma mängder data och information av olika kvalitet och därför är det svårt att ta fram en specifik uppgift om man inte i förväg vet precis vad man letar efter. Med hjälp av data warehouse underlättas aggretionen av data, det vill säga arbetet att summera och presentera data på olika sätt och synvinklar som de ursprungligen inte var strukturerade för.

Ett data warehouse kan förklaras som

”*the logical link between what the managers see in their decision support EIS* [executive information systems] *applications and the company’s operational activities*”

John McIntyre, SAS Institute Inc. (Dilly 1997)

EIS, som nämns i citatet, är främst utvecklat för att stödja den högsta ledningens informationsbehov (Turban 1995).

För att konstruera en modell för en organisations data warehouse måste man identifiera och definiera organisationens strategier och mål, organisationens struktur, de geografiska platser som den finns på, enheter och relationer. När modellen är färdig skall företaget bestämma vilket som är det bästa verktyget för att underlätta tillgängligheten till informationen. Företag bör välja huruvida de ska använda sig av ett data warehouse eller ett data mart, som är ett mindre och mer välavgränsat data warehouse för ett specifikt område eller uppgift (Hadden 1997a).

Earl Hadden har utvecklat en omfattande metod för hur större företag ska bygga ett data warehouse. (Earl Hadden är av analysföretaget Gartner Group rankad som en av de främsta auktoriteterna inom området.) Han har utvecklat metoden tillsammans med Sean Kelly. Det finns många olika metoder för hur ett företag ska utveckla ett data warehouse, där Earl Haddens och Sean Kellys metod är en av dem. Hadden - Kelly metoden består av fyra olika huvudfaser, förberedelse, planering, konstruktion och underhåll, som förklaras i nästa avsnitt (se figur 2.4). Anledningen till att vi presenterar denna metod är att den är företagsoberoende.

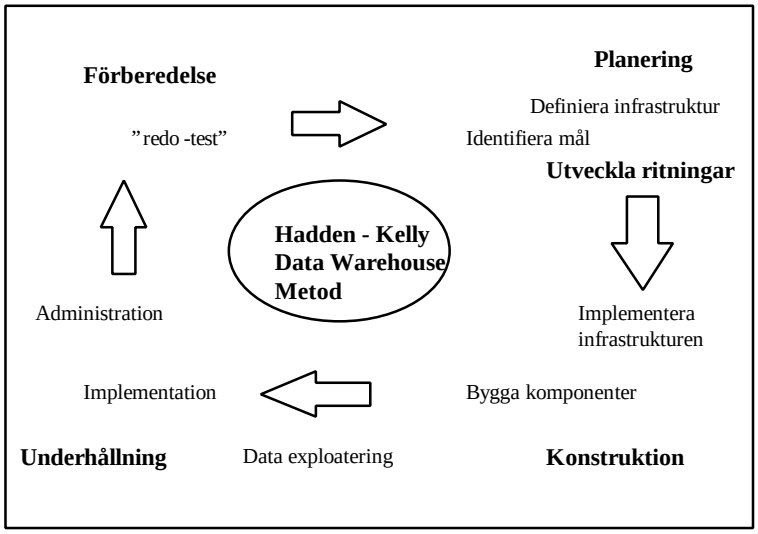
2.4.1 Faser i data warehouse

Den första fasen enligt Hadden (1997b) består av ett ”redo - test” som skall undersöka huruvida en organisation är redo att införa ett data warehouse. Syftet med det här testet är att undersöka om organisationen förstår och är medveten om de kritiska faktorer som krävs för att införandet av data warehouse ska lyckas. I den första fasen ska man först och främst genomföra en bedömning av organisationen genom att göra intervjuer samt att granska den befintliga dokumentationen. I den här fasen skall man även förbereda en projektplan som skall användas i planeringsfasen.

Nästa steg är planeringsfasen som består av tre delmoment, definiera infrastruktur, identifiera mål och utveckla ritningar. Syftet med planeringsfasen är att identifiera den information som ledningen är i störst behov av och identifiera möjligheten att uppfylla deras informationsbehov. I planeringsfasen ingår även att definiera den infrastruktur som behövs för att leverera informationen (Hadden 1997b).

Den tredje fasen, konstruktion, har till syfte att utveckla ett data warehouse som överensstämmer med organisationens definierade mål. Först bör man definiera de tekniska komponenter som behövs vid konstruktion av ett data warehouse. Efter det ska man förbättra den befintliga datamodellen tills den, och organisationens mål, överensstämmer. I denna konstruktionsfas utvecklas applikationer och program. I fasen förbereds även möjligheten att införa data mining vid ett senare tillfälle (Hadden 1997b).

Den fjärde och sista fasen enligt Hadden (1997b) är underhåll av data warehouse. Detta är en fas som är konstant och vars syfte är att tillgodose organisationen med ett alltid lika aktuellt data warehouse, som bör överensstämma med organisationens kravspecifikation.



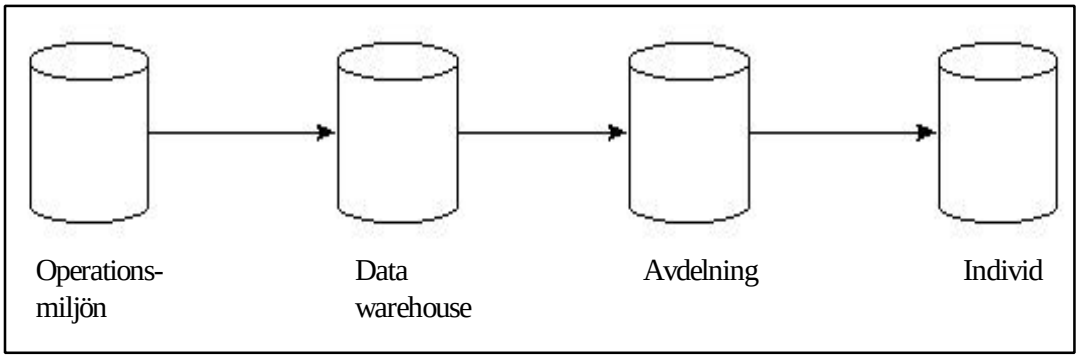
Figur 2.4. Bilden visar Hadden - Kelly - metoden för att konstruera ett data warehouse.(Hadden 1997b)

2.4.2 Data i data warehouse

Inmon (1996a) beskriver att det finns två former av data, dels den data som används vid transaktioner för att hantera de dagliga funktionerna i operationsmiljön och dels den data som härleds från transaktionsdata. Den härledda datan ingår i ett DSS och är summerad för att tillgodose ledningens behov[3]. Härledd data uppdateras inte utan innehåller ofta historisk data.[4]

[3] Här påvisar vi ytterligare än gång att data warehouse är en form av beslutstödssystem, uppgiften är hämtad från Inmon (1996a).
[4] I Inmon (1996a) förklarar han de två formerna av data och det finns vissa skillnader mellan begreppen. I tabell 2.2 förklarar vi de viktigaste kännetecknen för data i ett data warehouse.

Enligt Inmon (1996a) finns det en naturlig förekomst av dessa två former av data i de fyra nivåerna i följande arkitektur: operationsmilön, data warehouse, avdelningen och den individuella nivån (se figur 2.5, Inmon 1996a):



Figur 2.5 visar den arktitektur som finns när data transformeras från den operationella nivån till individnivån (Inmon 1996a)

I operationsmiljön finns det data innehåller som används vid de transaktioner som ska utföras i operationsmiljön. Ett data warehouse innehåller data från operationsmiljön som inte ska uppdateras, detta innebär att det finns data som är härledd. Avdelningsnivån innehåller nästan enbart härledd data. På den individuella nivån är datan härledd och används för analyser (Inmon 1996a). När data går från att vara i den operationella miljön till miljön för data warehouse blir data integrerad[5].

För att tydliggöra hur data passeras i de fyra olika miljöerna, ska vi förklara det med ett exempel som är hämtat från Inmon (1996a). I operationsmiljön finns det en post som innehåller uppgifter om en kund, Sven Svensson. Posten innehåller enbart aktuella uppgifter om kunden, det vill säga de förhållanden som finns just nu. I ett data warehouse finns det flera olika poster om Sven Svensson som visar kundens historia, till exempel vilken adress kunden har under olika tidsperioder. I avdelningsnivån kan det finnas en fil som summerar uppgifterna om Sven Svensson under en månad. Den sista nivån, den individuella, är ofta temporär och används för att analysera kunden och vad han har gjort.

Inmon (1996a), som ibland kallas för skaparen av begreppet data warehouse, ställer upp ett antal kriterier för data i ett data warehouse. Den data som har samlats in kan man inte veta om den är sann eller inte, men det finns olika sätt som man kan behandla data för att få så bra data som möjligt[6]. Det finns fyra generella kännetecken som beskriver ett data warehouse (se tabell 2.1):

[5] Se förklaring och exempel till integrerad data i tabell 2.1 och i Inmon (1996a).
[6] Enligt Andersen (1991) är data endast symboler utan värde med den synen kan man inte fälla något omdöme om data är sann eller falsk.

Kännetecken	**Förklaring**	**Exempel**
Objektorienterat (subject-oriented)	All data i ett data warehouse blir kategoriserad objektvis istället för i operationsmiljön där data är kategoriserad applikationsvis.	Ett försäkringsbolag kan organisera sin data i termer som kunder, premier och fordran.
Integrerad (integrated)	När hänvisningar till samma data existerar i olika applikationer har det tidigare varit svårt att få kodningen att vara konstant. När data är flyttad från operationsmiljö för att lagras i ett data warehouse enas man om en gemensam begreppsapparat för kodningen av applikationer. För att få en gemensam begreppsapparat integreras de olika definitionerna till en gemensam förklaring och benämning.	I olika applikationer betecknas kön ibland som m och f, ibland som 0 och 1. För att underlätta programmeringen enas man om att könen alltid ska betecknas som m och f när de ingår i applikationer.
Tidsvarianter (time-variant)	I ett data warehouse lagras data som kan vara upptill fem år eller äldre. Historisk data används för att göra jämförelser, visa trender och göra förutsägelser. Äldre data ska inte uppdateras.	I ett data warehouse har varje datapost en variabel som visar vilken tidpunkt en datapost härstammar från.
Icke ombytlig (non-volatile)	Den data som lagras i ett data warehouse skall inte uppdateras eller ändras. Uppdateringen sker i operationsmiljön innan data överförs till ett data warehouse. I	Data som lagrats i ett data warehouse kan man endast manipulera och inte ändra innehållet på. Detta skiljer sig från operationsmiljön där manipulering sker med en datapost åt gången.

Tabell 2.1 visar de kännetecken som ett data warehouse har enligt Inmon (1996a).

Data i ett data warehouse har olika detaljnivåer

Ett data warehouse har enligt Inmon (1996a, 1996b) olika detaljnivåer för att skilja olika sorters data (se tabell 2.2).

Detaljnivå	Förklaring till begrepp.	Exempel
Integrerad data (integrated data)	Med integrerad data kan man med ett analysverktyg få en enkel och snabb bild över den data som finns lagrad. Data warehouse har behandlat och integrerat data så att verktyget för data mining kan koncentrera sig på sin uppgift att hitta mönster.	Med integrerad data får man en övergripande bild av data. Ett exempel är en kund som har flera konton. På den här nivån får man en sammanfattning av en kunds totala konton, det vill säga man får resultatet av en summering[7].
Sammanfattad data (summarized data)	Sammanfattat data används för att ge en överblick av den data som är befintlig. Med sammanfattad data kan ett verktyg för data mining använda sig av tidigare resultat istället för att göra allt från grunden.	Den här nivån ger en summering av en helhet, vilket innebär att det har skett en uträkning av en mängd och i den här nivån redovisas delmängden. Ett exempel är den årsinkomst en person har, årsinkomsten är en sammanfattning av den månatliga lönen.
Detaljerad data (detailed data)	Data som finns i den här nivån är hämtad direkt från operationsmiljön och det är utifrån den här nivån som datan levereras till de andra nivåerna. Detaljerad data är viktig när analysverktyget behöver utforska data i dess minsta beståndsdel. Det kan finnas mönster i data som bara hittas när man undersöker data in i minsta detalj.	I den detaljerade data finns det uppgifter som alltid är aktuella. Ett exempel på den här nivån är vad en anställd har för timlön som leder till dess månadslön (sammanfattningen blir ju årsinkomst som gavs som exempel i föregående nivå).
Historisk data (historic data)	Används för att hitta kluster från historiska händelser som kan innehålla viktig information. Det är också en definitionsfråga vad som är historisk data, men det är upp till varje företag vad de anser och vilken definition de har. Historisk data behövs för att öka förståelsen för hur företaget har fungerat, vad som hänt och vilka cykler företaget har gått igenom.	Ju äldre den detaljerade data blir, desto mer intressant är det att använda den av historisk tidsaspekt och den förflyttas från detaljnivån till den historiska nivån. Ett företag kan ha bestämt att efter varje årslut ska data överföras till den här historiken. Ett exempel är att det i den här nivån finns uppgifter om vad den anställde, från föregående nivå, tjänade de tidigare åren han har varit anställd.
Meta data	Meta data beskriver sammanhanget av informationen.	Meta data ger en summering av de övriga nivåerna.

Tabellen 2.2 visar vilka nivåer av data det finns i ett data warehouse (Inmon 1996a, 1996b).

[7] Inmon (1996a) förklarar begreppen med ett annat exempel som hanterar de olika datanivåerna i ett produktsäljande företag.

2.4.3 Förhållandet mellan data warehouse och data mining

Berry och Linoff (1997) förklarar att data warehouses syfte är att hämta data från hela organisationen för att stödja beslutsfattandet. Målet med data warehouse kan formuleras som: ”*… data is available but not information - and not the right information at the right time. This is the goal of the data warehouse.*”

Syftet med data mining är att hitta funktionella mönster i data genom att identifiera, förvärva och behandla den. Finns det ett data warehouse i grunden kan det underlätta för analysen med data mining. Data warehouse bidrar med nödvändig och förberedd data (integrerad, detaljerad och sammanfattad, historisk och metadata) som data mining kan konvertera till användbar information[8]. Enligt Inmon (1996b) så kan förhållandet betecknas som ett symbios-förhållande. Observera dock att data warehouse inte är nödvändigt för att kunna använda sig av data mining, däremot ökar chanserna att nå ett bättre resultat om det finns ett data warehouse i grunden. [9]

Finns det inget data warehouse i botten kan det ta upp mot 80 % av data mining teknikens tid för att samla in rätt data. Data mining är en iterativ process som behöver använda sig av samma data flera gånger - data warehouse underlättar den processen (Berry & Linoff 1997).

2.5 Data mining

År 1997 listade Gartner Group data mining och artificiell intelligens i toppen av framtida nyckelteknologier som ”*clearly have a major impact across a wide range of industries within the next 3 to 5 years*”. Ur Gartner Group Advanced Technologies and Applications Research Note, 1995 (Pilot software 1997).

Data mining har en stark potential att hjälpa företag att fokusera på den viktigaste informationen i ett data warehouse. Data mining verktyg kan svara på affärsfrågor som traditionellt sett är för tidskrävande att lösa. Begreppet data mining ingår i en process som går ut på att omvandla data till kunskap. Processen kallas Knowledge Discovery in Databases (KDD) (Fayyad m.fl. 1996). Data mining och KDD-processen beskrivs mer i avsnitt 2.4.3. Intresset för data mining och KDD har resulterat i en djungel av definitioner och begrepp som är snarlika varandra. Sökandet efter mönster i data beskrivs i litteraturen med flera benämningar, däribland data mining, men också *knowledge extraction, information discovery, information harvesting, data archology* och *pattern processing*.

[8] Se även avsnitt 2.1 där hänvisning sker till Langefors definition för skillnaderna mellan data och information.
[9] I Berry och Linoff (1997) förklarar de vilken relation data warehouse och data mining har, i boken förklaras även exempel för hur det kan fungera.

2.5.1 Definitioner av data mining

Data mining är ett högaktuellt område men hittills är det mer ett samtalsämne än något som faktiskt är i bruk hos företag. Att så är fallet märkte vi tydligt vid den litteraturstudie som vi genomförde, se metoddel kapitel 3. Det finns många olika definitioner inom ämnet, vilket gör att det tar tid att reda ut vad olika författare menar och vad de lägger in i olika begrepp de använder. Data mining är ingen nyhet i sig, teknikerna som används har sitt ursprung i andra ämnen som statistik och artificiell intelligens (se avsnitt 2.5.2). Det som är nytt är att sammanslagningen och tillämpningen av tekniken används på ett nytt sätt och på olika områden.

Här nämner vi några definitioner:

”Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency net works, analysing changes, and detecting anomalies.”

William J Frawley, Gregrory Piatetsky-Shapiro and Christopher J Matheus (Dilly 1997)

Data mining kan beskrivas som:

”*Data mining refers to using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is hidden information in the data that is useful”*

Clementine User Guide, a data mining toolkit (Dilly 1997)

Nedanstående citat överensstämmer med vår syn på vad data mining innebär och det är denna förklaring vi använder oss av i uppsatsen:

”Data mining is the search for relationships and global patterns that exist in large databases but are ‘hidden’ among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database.”

Marcel Holshemier & Arno Siebes 1994, (Dilly 1997)

Anledningen till att vi anser att den överensstämmer med vårt sätt att se på data mining är att den poängterar att relationerna och sambanden som söks mellan data är dolda och tidigare okända. Om databasen är en "sann spegel" av verksamheten kommer man att finna värdefull information med hjälp av data mining.

Teknikerna som ingår i data mining är resultatet av en lång process av undersökning och utveckling av teorier inom artificiell intelligens, statistik och maskinlärande. Tabellen nedan är hämtad från Pilot Software (1997) och speglar deras förklaring till utvecklingen av data mining (se tabell 2.3).

Utvecklingssteg	**Företagsfrågor**	**Användbara tekniker**	**Företag med produkter**	**Kännetecken**
Data samling (Data collection) sextiotalet	"Vad var vår totala vinst för de senaste fem åren?"	Datorer, magnetband och disketter.	IBM och CDC	Levererar historisk data som är statisk.
Tillgång till data (Data access) åttiotalet	"Vad var enhets-försäljningen i Västra Götaland i april?"	Relationsdatabas (RDBMS), *Structured Query Language* (SQL) och ODBC	Oracle, Sybase, Informix, IBM och Microsoft	Levererar historisk, dynamisk data på datapost-nivå.
Data warehouse & Beslutstödssystem (Decision Support System, DSS)	"Vad var enhets-försäljningen i Västra Götaland i april? Titta närmare på Göteborg."	*On-line analytic processing* (OLAP), multidimensionella databaser och data warehouse	Pilot, Comshare, Arbor, Cognos och Microstrategy	Levererar historisk, dynamisk data.
Data mining (Utvecklas idag)	"Vad är sannolikt att Göteborgs-enheten säljer för i nästa månad? Varför?"	Avancerade algoritmer, datorer med multiprocessor och stora databaser	Pilot, Lockheed, IBM, SGI och ett antal företag som är i utvecklingsstadiet.	Levererar information om framtida händelser..

Tabell 2.3 visar data mining utvecklingssteg (Pilot software 1997)

2.5.2 Bakgrund till data mining

Data mining är ingen nyhet i sig och har sitt ursprung i andra områden. I detta avsnitt förklarar vi vilka de är. Vid användandet av data mining är det flera olika tekniker som används för att lösa olika problem, dessa tekniker härstammar från de områden som beskrivs i det här kapitlet. Data mining teknikerna beskrivs i avsnittet 2.4.5.

Statistik

Statistik används för att behandla och analysera observerad data, som är den data som statistikerna har samlat in från undersökningar med ett antal utvalda personer. För att välja ut antalet testpersoner sker det ett stickprov bland en grupp människor, till exempel studenter, politiker eller tjänstemän. Det finns olika tekniker för att göra detta stickprov, en av dem är

klustertekniken, då den utvalda gruppen har liknande egenskaper till exempel är alla studenter som har problem med CSN. En annan teknik är gruppering (sampling), som innebär att ett urval görs för att minska testgruppen (Dilly 1997).

Flera av de olika statistikmetoderna är viktiga vid data mining, framförallt kluster och gruppering. Datamängden, som data mining ska använda, är ofta flera gigabyte stor och med dessa tekniker kan data mining istället välja ut ett antal tusen poster att arbeta med. En annan viktig del som data mining har hämtat från statistik är verifikationsmetoden, vilket innebär att försöka bevisa eller motbevisa en uppställd hypotes (Dilly 1997).

Artificiell intelligens och andra metoder för lärande

Artificiell intelligens, AI, är en term, som likt data mining, har många definitioner. Enligt Turban (1995) handlar AI om två stycken basidéer. Den första innebär att man studerar den mänskliga tankeprocessen och den andra handlar om att presentera dessa processer med hjälp av maskiner, till exempel robotar och datorer. En definition av artificiell intelligens hämtad ur Turban (1995) är:

”*Artificiell intelligens is behavior by a machine that if performed by a human being, would be called intelligent.*”

Det finns ett test som kan användas som exempel för att ge en förklaring till citatet ovan, det så kallade Turing-testet. Designern av testet var Alan Turing och dess uppgift var att bestämma om en dator visades ha ett intelligent beteende. Enligt det här testet kan en dator klassas som smart endast om en mänsklig intervjuare, som kommunicerade med både en osedd dator och en osedd människa, inte kunde bestämma vem som var vem.

Winston och Prendergast (1984) definierar tre mål med AI (Turban 1995):
1. Göra maskiner smartare (det primära målet).
2. Förstå vad intelligens är.
3. Göra maskiner mer användbara.

Följande definition av AI fokuserar på mönstermatchande tekniker (Turban 1995), vilket innebär att det finns tekniker som hjälper till att hitta mönster i data och det är även syftet med data mining.

”*Artificiell intelligence works with pattern-matching methods which attempt to describe objects, events, or processes in terms of qualitative features and logical and computational relationships.*”

Det finns flera tekniker i data mining som är hämtade från AI och de är neurala nätverk, beslutsträd och minnesbaserat resonemang.

Maskinlärande är en annan metod för lärande som har sitt ursprung i AI. Maskinlärande refererar till en uppsättning av metoder som försöker lära maskiner hur de ska lösa problem och stödja problemlösning genom att använda sig av historiska fall. Problemet med den här tekniken är att det finns många modeller för lärande. Ibland kan det vara svårt att hitta rätt modell för ett visst problem, som behöver lösas. Tidigare exempel av maskinlärande är de

schackspelande programmen. Dessa program förbättrar sina prestationer med hjälp av erfarenheter (Turban 1995).

Det finns många metoder och algoritmer i maskinlärande. Flera av dem har varit under utveckling men används nu i data mining, några exempel som Turban (1995) nämner är lärande vid induktion (inductive learning), case-based reasoning och genetiska algoritmer (beskrivs i avsnittet 2.4.5.7).

Enligt Turban (1995) förklaras lärande vid induktion (inductive learning) genom följande definition (han sätter likhetstecken mellan lärande vid induktion och ”*knowledge acqusition*”):

”*Knowledge acqusition is the accumulation, transfer, and transformation of problem-solving expertise from some knowledge source to a computer program for constructing or expanding the knowledge base.*”

Lärande vid induktion innebär att man erhåller kunskap från mänskliga experter, böcker, dokument eller datafiler. Processen involverar utdrag av kunskap från expertkällor och överföring av kunskapen till en kunskapsbas. Processen innehåller flera steg: identifiering, konceptualisering, formalisering, implementering och testning. De metoder som används för att få fram kunskapen är manuella, semi-automatiserade eller automatiserade. I den manuella metoden används intervjuer för att kunna erhålla kunskap, i semiautomatiserade bygger experterna själva kunskapsbasen och i den automatiserade metoden är expertrollen helt eller delvis minimerad. Lärande vid induktion använder sig av historiska resultat för att lösa aktuella problem (Turban 1995).

Lärande vid induktion och case-based reasoning är två olika sätt att lösa nya problem med hjälp av historiska problem och lösningar. Case-based reasoning skiljer sig från lärande vid induktion på det sättet att den här metoden använder sig av historiska resultat och anpassar resultatet till den aktuella situtaionen (Turban 1995).

2.5.3 Data mining och Knowledge Discovery in Databases (KDD)

Data mining och Knowledge Discovery in Databases (KDD) har i vetenskapligt sammanhang varit ett hett ämne under de senaste åren. Det har under åren skrivits ett flertal böcker och en stor mängd artiklar, konferenser och seminarier har tillägnats ämnena.

Usama Fayyad & Evangelos Simoudis, två av de ledande experterna inom området, gör skillnad mellan begreppen KDD och data mining. Deras definition och förklaring är att KDD är en process, som utgår från data med kunskap som resultat genom stegen selektion, förberedning, transformation, data mining och evaluering. Enligt deras definition är data mining en del av KDD-processen. Vi kommer att använda oss av deras definition och sätter därmed *inte* likhetstecken mellan KDD och data mining.

Knowledge Discovery in Databases Process

Här nedan beskrivs de sju steg som ingår i KDD-processen enligt Fayyad (1996). Under varje steg i KDD-processen omvandlas data till att komma ett steg närmare kunskap. Processen är iterativ och varje steg är lika viktig som det tidigare steget. Varje stegs resultat fungerar som input för nästa steg (se figur 2.6):

1. Undersök och förstå problemdomänen. (Learning the application domain)
Lär känna det område applikationen ska appliceras på, det vill säga hur ser verksamheten ut, vad är relevant kunskap och vad är målet med applikationen. Här krävs ett samarbete mellan analytiker och användare för att komma underfund med vad det till exempel finns för flaskhalsar, vilka mål användaren har och vilka kriterier som är viktiga för att lyckas. Vad vill man att den slutliga produkten ska utföra för funktion, till exempel klassificering, summering eller visualisering. Hur mycket av processen behöver användaren förstå? Kan man använda sig av en black-box variant (Schoderbeck m.fl. 1990) där användaren endast vet vad som matas in och vad resultatet blir utan att behöva bekymra sig för vad som sker däremellan?

2. Skapa måldata (Creating a target dataset)
Välj datamängd eller delmängd som ska bearbetas. Det krävs att man funderar på om datan förändras över tiden eller om den är homogen. Efter urvalet av en delmängd har man fått fram måldata som används i steg 3, där data tvättas och förbereds.

3. Datan tvättas och förbereds (Data Cleaning and Preprocessing)
Data filtreras för att ta bort avvikelser och störande värden. Här måste man fatta ett beslut om hur värden som saknas ska behandlas och presenteras.

4. Reducering och transformation av data (Data reduction and transformation)
Här försöker man hitta sätt att representera data, beroende på vilka mål man har. Se till att ta bort variabler som inte är relevanta för målet och på så sätt reducera antalet variabler till de som är intressanta. Efter steg 3 och 4 är datamängden i ordning för att analysera med data mining verktyg.

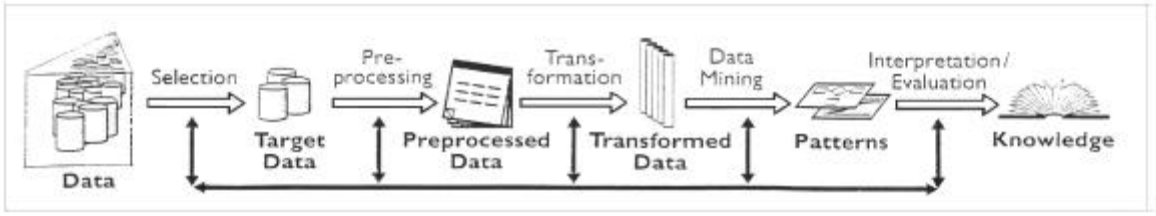
5. Val av data mining uppgift (Choosing the function of data mining)
I detta steg ska man bestämma vilken funktion data mining ska ha till exempel klassifikation, summering, regression eller klustering och vad som är syftet med modellen.

6. Val av data mining algoritm (Choosing the data mining algorithm(s))
Här väljer vi metod för att hitta mönster i data. Valet är oftast avgörande för resultatet. Data mining metoden måste vara kompatibel, det vill säga passa med de övergripande kriterierna för KDD- processen.

7. Data mining
Sökning efter mönster som kan vara intressanta visas till exempel som beslutsträd, regression eller kluster. Beskrivs mer i senare avsnitt. Resultatet av detta steg är mönster som måste utvärderas och tolkas vilket görs i det sista steget.

8. Tolkning (Interpretation)
Resultatet från steg 7 tolkas, evalueras och visas samt redundanta mönster avlägsnas, tolkningen sker av en analytiker. Svårbegripliga mönster översätts på ett sätt som användarna kan förstå och tolka utifrån deras kunskap om problemdomänen. Här bestäms vad som är kunskap av det processen tagit fram, vilket är en komplex uppgift som kräver att man förlitar sig på tekniken och på användarens kunskap och möjlighet att avgöra vad som är intressant nog att arbeta vidare med. Ett hjälpmedel för att avgöra om resultatet är kunskap är visualiserings verktyg för att få resultatet presenterat på ett mer överskådligt sätt. Resultatet kan även jämföras med tidigare kunskap som kan motsäga den nya kunskapen.

9. Använd kunskapen (Using discovered knowledge)
Arbeta in kunskapen i verksamheten eller dokumentera och rapportera till intresserade parter beroende på vilket målet var. Jämför med tidigare resultat eller uppfattningar.

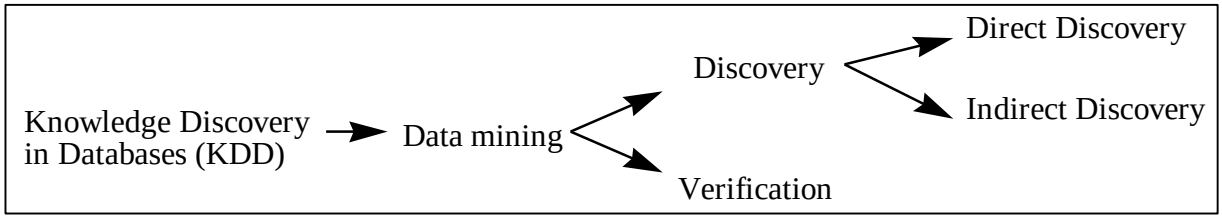


Figur 2.6. Beskriver Knowledge Discovery in Databases (KDD) processen enligt Fayyad (1996)

Fayyad menar att alla steg i processen är lika viktiga trots att mest arbete och uppmärksamhet läggs på steget för data mining. För att förklara de olika steg som ingår i data mining kommer vi att presentera de steg av KDD-processen som direkt berör data mining.

Verification and Discovery - Hypothesis and Knowledge Discovery

Steg 7 av Knowlegde Discovery in Databases (KDD) är data mining. I litteraturen skiljer man på två olika sorters data mining, Brachman m fl (1996) kallar dem verification och discovery och Berry och Linoff (1997) kallar dem hypothesis testing och discovery (se figur 2.7). Till innehållet är de näst intill identiska. Vi har valt att presentera de olika sorterna enligt Brachman med verifikation (verification) och discovery. Vi har valt att kalla discovery för sitt ursprungsnamn på engelska. En översättning av betydelsen skulle på svenska bli upptäckande metod. Anledningen till att vi valt att inte översätta discovery är att det inte finns någon tillräckligt bra och allmänt accepterad översättning av begreppet i svensk facklitteratur.



Figur 2.7. visar sambandet mellan de olika data mining sorterna och hur de är relaterade till varandra.

Verifikationstest (Verification testing)

Verifikationstest är en top-down metod som försöker bevisa eller motbevisa förutfattade idéer, det vill säga man testar en hypotes. Verifikation består av följande steg (Berry & Linoff 1997):

1. Ta fram en bra hypotes.
2. Bestäm vilken data som hypotesen ska testas med.
3. Ta fram data.
4. Förbered data för analys.
5. Bygg en dator-baserad modell.
6. Utvärdera modellen, bekräfta eller refusera hypotesen.

När verifikationsmodellen används tar användaren fram en hypotes som de sen testar mot den data de har lagrat. Tyngdpunkten ligger på användarna som skall avgöra om man ska anta resultatet eller om man ska förkasta hypotesen och börja om från början med en ny hypotes.

Exempel på verifikationstest: Då en marknadsföringskampanj skall startas är det viktigt att undersöka vilka kunder som troligtvis kommer att acceptera det kampanjen erbjuder. Användaren formulerar en hypotes för att identifiera potentiella kunder och vad som speciellt för kunderna. Historisk data från tidigare kampanjer och demografisk information kan användas som underlag för att ta fram ett urval för kampanjen. Testet fortgår tills man fått ett resultat och ett urval man är nöjd med.

Problemet med den här modellen är att den inte hittar och skapar någon ny information utan bara bearbetar existerande. Användaren lär sig dock känna vad som finns lagrat och lär sig studera den med hjälp av olika verktyg för visualisering.

Discovery

Discovery är en bottom-up metod där man börjar med data och försöker få den att säga oss något vi inte redan visste. En discovery-modell söker igenom data för att hitta trender, generaliseringar och mönster i data som tidigare var okända utan användarens inblandning.

Exempel på discovery modellen är då man ur en stor databas skall söka efter potentiella kunder för en kampanj. Modellen söker igenom datamängden utan någon hypotes om vad som ska finnas utan söker fritt efter grupper som liknar varandra.

Discovery är den metod som ägnas mest uppmärksamhet och studie, vilket gör att den har mest utrymme i litteraturen. En av anledningarna, enligt Berry & Linoff (1997), är att discovery-modellen tros ge enkla och billiga lösningar för företag att få fram information om hur man skall förbättra lönsamheten. Discovery kan antingen vara *indirekt* eller *direkt* (se figur 2.7). Indirekt discovery används för att upptäcka relationer och samband mellan data. Direkt discovery förklarar de relationer och samband som hittats med en indirekt discovery modell. De flesta data mining tekniker är direkta, det vill säga de söker svar på konkreta frågor, till exempel vad kommer att hända om man gör en viss förändring istället för att fråga vad som kommer hända i stort som den indirekt modellen gör. Anledningen till att direkt är mer användbart är att det är lättare att använda sig av de klassiska teknikerna, som beslutsträd och neurala nätverk, då man vill få relationer förklarade än att ta fram nya, vilket gör att direkt discovery är mer målinriktad.

Direkt Discovery

Processen går ut på att hitta meningsfulla mönster som förklarar händelser som inträffat på ett sätt som gör det möjligt att förutse framtida händelser.

1. Identifiera källor av fördefinerad data.
2. Förbered data för analys.
3. Bygg och lär upp en datorbaserad-modell.
4. Utvärdera modellen.

Indirekt Discovery

Processen för indirekt discovery ser ut på följande sätt:

1. Identifiera datakällan.
2. Förbered för dataanalys.
3. Bygg och lär upp en datorbaserad-modell.
4. Utvärdera modellen.
5. Identifiera potentiella mål för direkt discovery.
6. Generera nya hypoteser att testa.

Steg 1 till 4 är samma för direkt- och indirekt discovery. Steg 5 och 6 i den indirekta modellen beskriver möjligheten att efter indirekt discovery utföra direkt discovery för att ta reda på vad som ligger bakom de samband man fann.

2.5.4 Metod för att införa data mining i en verksamhet

Det finns ingen specificerad metod eller affärsprocess för att införa data mining i en verksamhet, det finns ungefär lika många förslag som det finns författare i ämnet. Men vad som är lika för dem är i stora drag de fyra följande stegen (Berry & Linoff 1997, Simoudis 1996):

1) Identifiera affärsproblem data mining ska underlätta eller lösa. Under detta första steg är det viktigt att företagets domänexperter är med och analyserar fram problem och möjligheter som ska lösas respektive utvecklas. För att man ska lyckas med data mining måste man ha en realistisk bild på vad i verksamheten data mining ska underlätta.

   "Domain experts who identify the business opportunity should have some idea of how to act on and measure the results from a data mining stage."
   (Berry & Linoff 1997)

2) Använd data mining för att transformera data till information[10]. Välj vilken modell som ska presentera resultatet, till exempel beskrivande eller förutsägande modeller, beroende på vad målet med data mining lösningen är.

3) Använd informationen för att förbättra verksamheten inom de affärsproblem som identifierades i steg 1). I denna fas ska informationen verkligen arbetas in i verksamheten. Här är det oerhört viktigt att personer med gedigen branschkunskap är med och tolkar resultatet. Eftersom det kan vara svårt att tolka resultatet av data mining är en analytiker viktig i denna fas, eftersom de kan tolka och översätta resultatet så att de branschkunniga kan förstå och använda informationen för att fatta beslut.

4) Mät resultatet och utvärdera tekniken och modellen. Berry och Linoff (1997) menar att mäta resultatet är en viktig del av metoden för att införa data mining eftersom man kan dra fördelar av sina erfarenheter till nästa gång.

   "*It is a good idea to think of every data mining effort as a small business case. By comparing our expectations to the actual resultat, we can often recoqnize promising opportunities to exploit on the next round of the virtous cycle [författarnas namn på metoden för att införa data mining i en verksamhet]*"
   (Berry & Linoff 1997)

   Om man ska utvärdera en marknadsföringskampanj kan man ställa sig följande frågor för att ta reda på om data mining kan förbättra verksamheten:

   - Fick rätt kunder utskicken ?
   - Är de kunderna mer lojala än genomsnittet ?
   - Hur ser demografen ut för de kunder som fick kampanjen ?
   - Köper kunden andra produkter ?
   - Hur jämför man olika kunder som man hittat med olika tekniker ?

   Med svar på dessa frågor kan ledningen besluta sig för om data mining tillför något nytt eller om det finns andra sätt att få fram samma resultat.

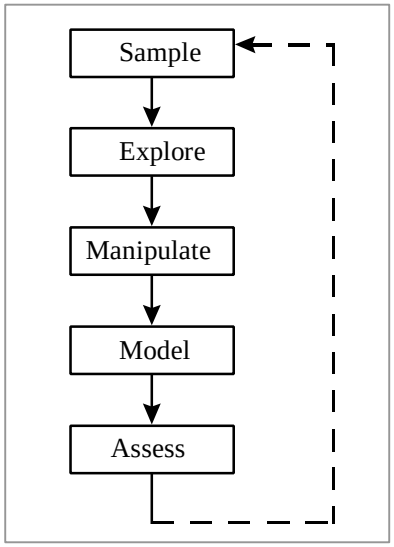
[10] Se även delkapitlet 2.1 där vi förklarar skillnaden mellan data, information och kunskap.

De ingående stegen i data mining metoden är direkt beroende av varandra, vilket leder till att resultatet av ett steg är direkt beroende av steget innan.

SEMMA, SAS Metod för Data mining

Produkten Enterprise Miner som vi har testat är uppbyggd kring SAS Institutes metod för data mining, SEMMA (Sample, Explore, Manipulation, Modeling & Assess). SAS Insititutes metod definierar de steg som de anser ska ingå i processen för data mining (se figur 2.8). SEMMA innebär (SAS 1996):

- **Urval** (Sample) -
  Användaren gör ett urval av den totala datamängden för att enbart använda en delmängd för data mining. Detta gör man då det bara är intressant med generella mönster.
- **Undersökning** (Explore) -
  Undersöker vilken typ av data som finns för att användaren ska få en förståelse för vilka mönster som kan bli resultatet från datatypen. För att inte behöva undersöka all data kan man använda statistiska metoder som kluster för att summera den befintliga datan.
- **Manipulering** (Manipulation) -
  Efter undersökningen av data finns det möjlighet att manipulera den för att kunna inkludera ny information som upptäcktes i förra fasen. Vid manipulationen kan man upptäcka vilken grupp av data som är intressant för fortsatt analys.
- **Modellering** (Modeling) -
  I den här fasen väljer man tekniker som sedan används för att hitta mönster och samband i data. Teknikerna används på olika former av datatyper och därför bör man veta vilken typ av data som ska bearbetas och välja teknik därefter.
- **Jämförelse** (Assess) -
  Här kan man jämföra de tekniker som har använts och dess resultat. Hela SEMMA-metoden kan upprepas till man får ett tillfredsställande resultat.



Figur2.8 visar processen för data mining som SAS har byggt upp, SEMMA (1996)

2.5.5 Data mining tekniker

Data mining teknikerna är anpassade för att lösa olika problem, somliga är mest lämpade inom vissa områden, vilka kommer att förklaras här nedan. Teknikerna kan vara av två olika sorter, Berry och Linoff (1997) definierar dem som, direkt discovery och indirekt discovery (förklarades i 2.4.3).

Enligt Berry och Linoff (1997) kan teknikerna för data mining delas in i sex olika funktioner. I det här avsnittet förklarar vi vilka de sex funktionerna är och vilka tekniker som kan lösa dem. Den första och vanligaste är *klassificiering* (classification) av poster. Vid klassificering tilldelas varje post en klasskod beroende på vilken typ posten är av. De tekniker som kan lösa den här funktionen är minnesbaserat resonemang (2.5.5.3), länkanalyser (2.5.5.4) och beslutsträd (2.5.5.5). Vid en given input används *bedömning* (estimation) för att få fram värdet på en okänd variabel, till exempel inkomst. Bedömning används för att utföra en klassificeringsuppgift och tillämpas av neurala nätverk (2.5.5.6). Den tredje funktionen är *förutsägelse* (prediction) som delar in posterna efter hur deras framtida värden kommer att se ut. Framtiden får utvisa om indelningen är korrekt. Tekniker, som sekvensbaserad analys (2.5.5.2), minnesbaserat resonemang (2.5.5.3), beslutsträd (2.5.5.5) och neurala nätverk (2.5.5.6) är användbara vid förutsägelser. Till skillnad från klassificering använder inte *kluster* (2.5.5.1) några fördefinierade klasser, utan posterna är grupperade tillsammans efter de gemensamma egenskaper de har. *Sekvensbaserad analys* (2.5.5.2) visar vilka saker som hör ihop, till exempel vilka varor som säljs tillsammans. Med den här formen av analys får man ett resultat i form av associationsregler. Den sista funktionen är *beskrivning* (description) av ett beteende, som även förklarar varför ett visst beteende uppstår. De olika teknikerna är antingen rent beskrivande, som sekvensbaserad analys (2.5.5.2), eller inte alls beksrivande, som neurala nätverk (2.5.5.6).

2.5.5.1 Kluster (Cluster detection)

Pilot softwares (1997) definition av klustertekniken:

”*The process of dividing a dataset into mutually exclusive groups such that the members of each group are as ‘close’ as possible to one another, and different group are as ‘far’ as possible from one another, where distance is measured with respect to all available variables.*”

En stor databas kan innehålla många variabler och poster som gör det svårt för teknikerna att finna meningsfulla mönster. Ett problem som kan uppstå med data mining är att man hittar för många mönster som, i slutändan, inte bidrar med relevant information. För att lösa detta problem kan man tillämpa klustertekniken, en av de få teknikerna som använder sig av indirekt discovery-modellen (Berry & Linoff 1997).

Klustertekniken går ut på att bygga en modell som hittar dataposter som liknar varandra. Målet är att hitta likheter i data som tidigare var okända. Det finns flera olika metoder för att hitta kluster, till exempel statistiska och neurala nätverk. Fördelen med kluster är att man delar in data i olika grupper och får därigenom en god uppfattning om hur datamängden ser ut (Berry & Linoff 1997).

Det finns flera olika metoder vid tillämpandet av klusterteknikken, den vanligaste metoden är ”K-means” och skapades år 1968 av J.B. MacQueen (Berry & Linoff 1997). K-means består av flera steg som ska utföras. I det första steget väljs K datapunkter som utgångspunkter, vilka efterhand kommer att utvidgas till kluster. MacQueens algoritm låter den första K posten utgöra startpunkten. En efter en delas alla poster in i det kluster vars centrum ligger närmast posten. När alla poster är indelade i grupper ska centrumpunkten räknas ut i de nya klustren. När de är funna börjar man om från början med att dela in varje post i olika kluster tills de slutar att förändras. Resultatet av tekniken är att de poster som ingår i samma kluster har likartade egenskaper, vilket kan underlätta processen för data mining (Berry och Linoff, 1997).

2.5.5.2 Sekvensbaserad analys (Sequence based analysis)

Sekvensbaserad analys kallas även för *market basket analysis*. Tekniken härstammar från statistik och sannolikhetslära och är anpassad för både indirekt och direkt discovery.

Berry och Linoff (1997) förklarar att tekniken går ut på att analysera vad kunder handlar, för att kunna förklara vilka kunderna är och varför de handlar som de gör. Resultatet från analysen beskriver vilka produkter som köps tillsammans och vilka som är mest lämpliga för reklam. Resultaten från sekvensbaserad analys presenteras i form av associationsregler som består av if-then satser.

Tekniken är en form av kluster som används för att hitta grupper av saker som brukar förekomma tillsammans i en transaktion. När det gäller detaljhandeln kan transaktionsdata vara den enda informationen som finns tillgänglig för att lära känna kunderna. Annan information såsom demografisk och historisk data finns ej, eftersom transaktionerna är anonyma. Som en klusterteknik, är analysen användbar vid problem där man vill veta vilka saker som påträffas tillsammans eller på ett speciellt sätt. Resultatet är lättare att agera på eftersom det blir klart och specifikt. Resultatet kan även användas till ett flertal uppgifter, till exempel för att ändra butikslayout eller reducera utbudet av varor, som man vet går att sälja tillsammans. Då transaktionerna är anonyma kan man lägga till historik med en tidskomponent så man vet under vilken tidsperiod transaktionen genomfördes (Berry och Linoff, 1997). Resultatet blir då mer användbart eftersom man kan skilja på exempelvis veckodagar och månader.

2.5.5.3 Minnesbaserat resonemang - (Memory-based reasoning)

Berry och Linoff (1997) kallar den här tekniken för minnesbaserat resonemang (MBR), men av andra experter benämner den här tekniken som närmaste grannen (”nearest neighbour”). I vår uppsats kommer vi att använda termen minnesbaserat resonemang vid förklaring och benämning av den här tekniken. Minnesbaserat resonemang är en direkt discovery metod. Pilot softwares (1997) definition av MBR är:

”*A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1).*”

Det väsentliga med minnesbaserat resonemang är hur man kan bearbeta nya händelser på liknande sätt som tidigare erfarenheter har bearbetats. Det första steget är att identifiera de händelser från tidigare upplevelser som motsvarar den situation som nu uppstått. Tekniken använder sig av grannposter i register för att kunna klassificera och förutse framtida händelser. Minnesbaserat resonemang fungerar bra vid försäkringsbedrägerier och förutsägelsen om huruvida en kund svarar på en kampanj (Berry & Linoff 1997).

Enligt Berry och Linoff (1997) finns det endast två stycken funktioner som den här tekniken hanterar. Den första funktionen kallas för *distansfunktionen* och innebär att det tilldelas en distans mellan två poster i samma register. Den andra heter *kombinationsfunktionen* och den kombinerar resultatet från näraliggande poster för att komma fram till ett gemensamt resultat. För att kunna klassificera okända kategorier behöver man en databas som innehåller historiska poster som redan är indelade i kategorier**.** Man letar bland de historiska posterna efter en eller flera händelser som påminner om den aktuella situationen. När MBR hittar en historisk post som matchar den nya situationen, klassificeras den nya posten efter vilken kategori den historiska posten tillhör. För att hitta den kategorin ska två steg utföras**.** Det första steget heter *lärande*, vilket innebär att leta efter historiska värden i databasen, och det andra steget är *förutsägelse*, som beskriver hur tekniken ska kunna använda sig av nya händelser. För att lösa ett problem med den här tekniken är det tre saker som skall lösas. Det första, är att välja rätt mängd historiska poster som utgör testdatan, vilken bör vara en del av alla tillgängliga poster och innehålla lika många poster från varje kategori (Berry & Linoff 1997). Det andra, är att bestämma det bästa sättet att presentera de historiska posterna på. Det sista problemet är att bestämma distans- och kombineringsfunktionen samt antalet grannar en post skall ha. Dessa nyckeldelar bestämmer hur ett minnesbaserat resonemang kan producera ett bra resultat.

2.5.5.4 Länkanalys (Link anlysis)

Länkanalyser försöker etablera relationer mellan poster i en databas, vilket är det primära målet med den här formen av analys. Genom att studera relationer mellan olika poster kan länkanalysen utveckla modeller som baseras på de funna relationerna. Vissa algoritmer kan spåra samband mellan poster över tiden. (Simoudis 1996)

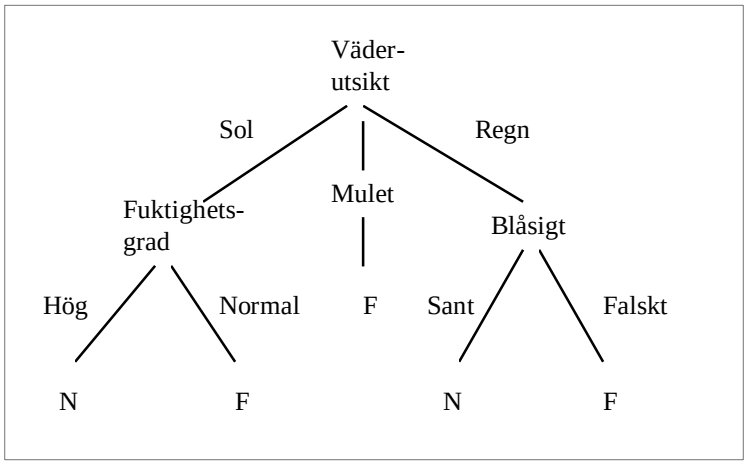
Tekniken baseras på matematiska grafer, vilket leder till att resultaten av applikationen visas i form av grafer. Noderna i en graf representerar den enhet som är av intresse och länkarna mellan noderna representerar deras förhållande eller transaktioner. Både länkarna och noderna kan ha specifika attribut som är unika för en domän eller som är relevanta för insamlingsmetoden. Datan som används kan vara enkel men i stora mängder, annan data kan vara rik och varierad. Exempel på områden där länkanalys är tillämpningsbar är inom telekommunikation och försäkringsbedrägerier. I länkanalyser inom telekommunikation representeras abonnenterna som noder och linjen mellan den som en relation (Berry och Linoff, 1997).

2.5.5.5 Beslutsträd och regelformulering (Decision Tree and Rule induction)

Beslutsträd är en teknik som använder direkt discovery-modellen och är ett alternativ att använda vid klassificering och förutsägelse. Tekniken använder sig av regler för att komma fram till ett resultat. Reglerna kan uttryckas i vanlig språkform eller som databasspråk i form av SQL-satser.

Enligt Berry och Linoff (1997) finns det olika algoritmer för att bygga beslutsträd. Tre av de mest använda är CART, C4.5 och CHAID som används vid klassificering och förutsägelser. Algoritmen CART (Classification And Regression Tree) producerar binära träd till skillnad från C4.5 som bildar träd med två eller flera förgreningar för varje nod. CHAID (CHi-squared Automatic Interaction Detection) försöker, till skillnad från de två andra algoritmerna, stoppa förgreningarna innan trädet växer och blir för stort.

Trädet ritas med roten som topp och löven i botten (se figur 2.9). Varje post börjar vid roten där det sker ett test för att bestämma vilken subnod posten ska gå vidare till. Målet är alltid att välja ett test som leder till bäst klassificering av posterna. Processen upprepas tills varje post har kommit till en lövnod, alla poster som slutar på samma nod är därmed klassificerade på samma sätt. Det finns en unik väg från roten till varje lövnod där vägen är uttryckt som en klassificeringsregel för posterna (Berry & Linoff 1997).



Figur2.9 illustrerar en trädstruktur från roten till lövnoderna. (Dilly 1997)

En av de största fördelarna med beslutsträd är att modellen är enkel att förklara eftersom det finns explicita regler. Det gör att man kan evaluera resultatet och identifiera nyckelattribut i posterna.

2.5.5.6 Neurala nätverk

Pilot software (1997) definierar neurala nätverk enligt följande:
”*Non-linear predictive models that learn through training and resemble biological neural networks in structure.* ”

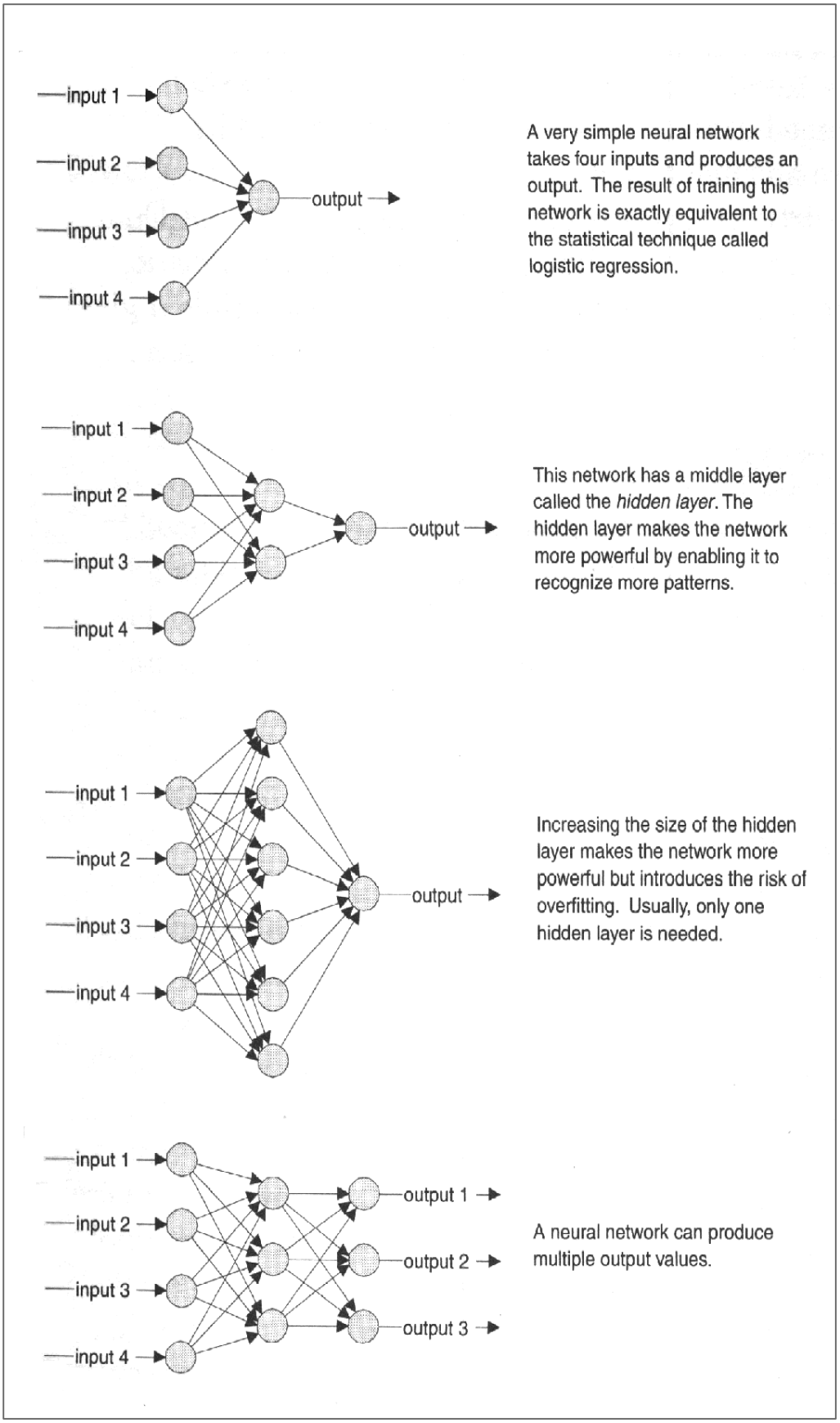
Brooks (1997) har en utförligare definition som lyder:
”*Neural networks are a series of software synapses used to create a prediction model by clustering information into natural groups and then predicting into which groups new records will fall.*”

Enligt dessa två definitioner är neurala nätverk bra att använda vid förutsägelser och klassifikationer genom att efterlikna biologiska neuroner i strukturen. För att kunna göra förutsägelser måste man dela in posterna i olika grupper och sedan förutsäga till vilken grupp en okänd post ska placeras in i. Innan man kan placera in okända poster måste det neurala nätverket ha använt sig av en mängd historisk data som testdata, för att kunna lära sig känna igen situationer och därefter placera in de nya posterna. Neurala nätverk är en direkt discovery metod.

Enligt Berry och Linoff (1997) är det första steget i neurala nätverk att producera en modell för testdata. Resultatet av ett neuralt nätverk blir inte bättre än den data som användes vid testet. Att välja rätt data är ett av de viktigaste valen som görs i den här tekniken. Testdatan består av observationer och klassificeringar av kända poster. För att välja rätt data finns det några saker man bör tänka på, det viktigaste är att testdatan ska täcka alla värden som kännetecknar nätverket. Att träna ett neuralt nätverk är en tidsödande process, där tiden beror på hur många kännetecken ett specifikt nätverk har. Den sista punkten är ju fler kännetecken nätverket har desto fler testomgångar behövs det för att täcka alla samband mellan posterna.

Neurala nätverk består av basenheter av biologiska neuroner. Varje enhet har ett flertal input-värden som kombineras till ett output-värde (se figur 2.10). Vissa enheters output fungerar som input till nästa enhet. Den enklaste varianten av ett neuralt nätverk är ett så kallad feed-forward nätverk, som innebär att det endast finns en väg genom nätverket från input till output. Aktiveringsfunktion kallas den funktion som gör att enheternas input kombineras till ett output värde. En aktiveringsfunktion består av två delar, *kombineringsfunktion* och *överföringsfunktion*. Den första funktionen slår ihop inputen till ett värde och den andra funktionen för över värdet från kombineringsfunktionen till enhetens output. (Berry & Linoff 1997).

För att kunna använda ett neuralt nätverk måste man för det första välja rätt testdata. För det andra måste man representera data på ett sätt som leder till att man hittar maximalt antal samband i datan. För det tredje måste man kunna tolka resultaten från neurala nätverket, vilket innebär att förstå specifika detaljer inne i nätverket som kan resultera i bättre resultat (Berry & Linoff 1997). Ett problem med neurala nätverk är att det kan vara svårt att tolka resultatet och ge en förklaring till varför resultatet blir som det blir. Eftersom det är svårt att förstå resultatet kan det även vara svårt att motivera ett beslut, till exempel kan man inte motivera ett avslag på ett banklån, genom att säga “mitt neurala nätverk sa mig så” (Berry & Linoff 1997).



Figur 2.10 beskriver ett neuralt nätverk som är uppbyggda på olika sätt (Berry & Linoff 1997)

2.5.5.7 Genetiska algoritmer

Precis som minnesbaserat resonemang och neurala nätverk försöker genetiska algoritmer efterlikna biologiska processer. Genetiska algoritmer är en artificiell variant av evolution, där man använt sig av strängar och datorprogram som förändras över tiden istället för naturliga organismer. Genetiska algoritmer är en direkt discovery metod som söker efter optimala förutsägelser. Likt statistik behöver algoritmerna ha en från början känd modell. Genetiska algoritmer använder sig av selektion och mutation för att utveckla generationer av lösningar. Allteftersom nya generationer utvecklas överlever endast den som är mest förutsägbar. Genetiska algoritmer har också använts för att förbättra MBR och neurala nätverk. Experter inom data mining beräknar att genetiska algoritmer kommer användas i allt större utsträckning inom de närmaste åren (Berry & Linoff 1997).

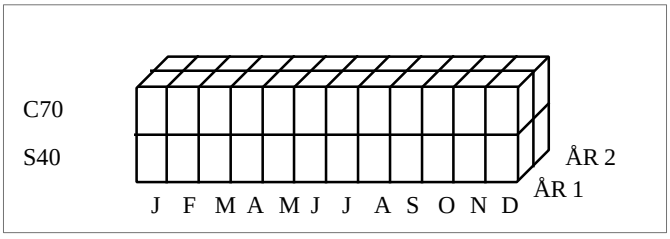
2.5.5.8 OLAP

OLAP, On-Line Analytic Processing, är egentligen inte någon data mining teknik, men vi har valt att presentera den i detta stycke eftersom Volvokort använder sig av OLAP för att analysera data. Vi kommer att hänvisa till begreppet i resultatdelen för att jämföra OLAP med data mining tekniker som Volvokort kan använda sig av.

Pilot softwares (1997) definition av OLAP:

”*Refers to array-oriented database application that allow users to view, navigate through, and analyze multidimensional databases.*”

Enligt Berry och Linoff (1997) är OLAP ingen teknik som kan användas istället för data mining utan bör fungera som ett komplement till data mining. OLAP är ett analysverktyg med ett grafiskt gränssnitt som använder sig av en kub för att visualisera data**.** Kuben implementeras som en relationsdatabas eller en multidimensionell databas som optimerar OLAP-funktioner. OLAP tekniken bidrar med funktioner som är svåra eller omöjliga att uttrycka i SQL-satser. I en kub definieras vilka rapporterar som användaren är intresserad av (Berry och Linoff, 1997). Dessa rapporter tilldelas var sin dimension i kuben (se figur 2.11), vilket motsvarar en axel i kuben. Kuben består av många små delkuber. Varje delkub innehåller en unik nyckel som definierar den specifika delkuben och en sammanfattad information för den data som kategoriserats in i den.



Figur 2.11 visar ett exempel för hur man kan analysera med OLAP. Bilden visar hur man kan jämföra försäljningen av bilar under samma månad men olika år etc. (Berry & Linoff 1997).

En subkub kan förklaras som en del av en multidimensionell databas där alla dimensioner har fixerade värden. I en välformulerad kub faller varje post i exakt en delkub, detta är kardinalregeln för multidimensionella databaser. För att kuben ska kunna användas i analytiska syften är det tre saker, förutom kardinalregeln, som bör fungera (Berry & Linoff 1997). Kuben bör kunna hantera kontinuerliga värden, hierarkiska dimensioner och dimensionerna skall kunna sträcka sig över multipla fält.

Tabell för översikt av teknikerna

Tekniker	Fördelar	Nackdelar
Kluster	• En av få tekniker som är indirekt discovery. • Lätt att använda.	• Känslig för vilken startpunkt som väljs i algoritmen K-means. • De funna klustren garanterar inte att det leder till någon nyttig information.
Sekvensbaserad analys	• Analysen ger klara och tydliga resultat. • Analysen är en indirekt discovery metod.	• Kan inte användas på alla typer av problem. • Är svårt att bestämma vad som ska användas till analysen.
Minnesbaserat resonemang	• Kan användas på de flesta typerna av data. • Resultatet är lättförståeligt.	• Det krävs stor kapacitet i form av datorminne vid den här proceduren. • Resultaten är helt beroende av vilken distans- och kombineringsfunktion och antal poster som väljs.
Länkanalys	• Är ett bra användningsverktyg vid visualisering av data. • Länkanalysen är bra då man vill beskriva relationer mellan poster t.ex. inom telekommunikation där man studerar telefonsamtal mellan olika abonnenter.	• Det är inte många datatyper som kan användas. • Det finns få verktyg för data mining som stödjer länkanalys.
Beslutsträd	• Genererar användbara regler. • Utför klassificering på ett lättförståeligt sätt.	• Det är dyrt att utföra klassificering med beslutsträd eftersom det är tidskrävande. • Vissa algoritmer kan bara hantera en viss form av regler.
Neurala nätverk	• Kan appliceras på de flesta problem. • Producerar bra resultat även vid komplicerade domäner.	• Det är svårt att förklara resultaten från nätverket. • Det måste finnas en expert på området som kan kunna tolka resultaten.
Genetiska algoritmer	• Ger förståeliga resultat. • Resultatet är lätt att använda.	• Ingår i få data mining produkter. • Tar tid och blir kostsam.

Tabell 2.4 beskriver för- och nackdelar med de olika data mining teknikerna (Berry & Linoff 1997)

2.5.6 Data mining områden

Traditionellt sett har majoriteten av företag som har använt sig av data mining varit företag som vill förbättra sin kundrelation, marknadsföring och för att upptäcka bedrägerier. Utveckling av data mining tekniker pågår hela tiden och fler användningsområden upptäcks. De områden som hittills varit mest uppmärksammade som lämpliga för data mining har varit:

Kundrelation : Företag som vill ta fram kundprofiler har varit de som mest sagt sig behöva data mining, enligt Edelstein och Millenson (1997). En applikation företagen är intresserade av är avvikelseanalys (Churn or attrition analysis). Genom att studera kunder som lämnat företaget kan man se på sina andra kunder och analysera vilka som liknar dessa och som eventuellt funderar på att byta företag. På så sätt kan företaget sätta in rätt kampanj vid rätt tillfälle för att förmå kunden att stanna kvar i företaget. Försäkringsbolag i USA använder data mining för att ta reda på var i livscykeln deras kunder befinner sig i, med hjälp av den informationen kan de räkna ut vilka behov kunderna har. Har de barn som ska börja skolan eller ska de snart pensionera sig? Ett annat exempel är företag som använder sig av data mining för att ta reda på vilka produkter en kund föredrar för att kunna skicka erbjudanden om liknade produkter, det vill säga en form av rekommendationssystem.

Marknadsföring: Intresset för riktade kampanjer har under de senaste åren ökat (Shepard 1995). Det finns ett stort intresse för riktad marknadsföring som inkluderar att förbättra svarsfrekvensen på brevkampanjer, val av produkt för att passa kunden eller val av område där kampanjen kan ge ett bättre resultat. En teknik som kan stödja marknadsföringen är sekvensbaserad analys som till exempel används för att ändra om butikslayouten i en affär för att placera varor som säljer bra tillsammans. Ett klassiskt exempel på det här är en stormarknad i USA. De undersökte vilka varor som sålde tillsammans och med en sekvensbaserad analys kom de fram till att det på fredagskvällarna var unga pappor som tog bilen till affären för att handla. De köpte stora paket med blöjor och passade samtidigt på att handla öl till fredagskvällen. Affären ändrade om sin butikslayout och placerade ölen tillsammans med blöjorna (Berry & Linoff 1997).

Upptäcka försäkringsbedrägerier: Tydliga bedrägerier går att spåra utan data mining, men fördelen med data mining är att de hittar mer diskreta bedrägerier. Den här sortens bedrägeri går ut på att göra mycket små förändringar varje månad, till exempel små uttag varje månad för att systematisk tömma ett konto. När man utnyttjar tekniker som dessa, för att upptäcka diskreta bedrägerier, måste en kunnig person kunna avgöra om hypotesen är riktigt eller ska förkastas.

Kreditundersökning: Exempel på data mining applikationer som utvecklats för att förutse vem som kommer att bli en pålitlig kund och vem som kommer att ligga efter med betalningar. Varje månad kan kundernas transaktioner undersökas och dokumenteras. Med hjälp av analysverktyg kan man sedan se om kunderna följer sin vanliga rutiner. Stora kreditkortsbolag har börjat använda den här tekniken för att se om en kund plötsligt ändrar sitt kreditbeteende, till exempel används kortet med kortare intervaller än vanligt och då med stora belopp eller att kortet plötsligt börjar används i ett helt annat land under en längre tid. Plötsliga förändringar i användandet kan bero på att kortet har blivit stulet och företaget kan då hjälpa kunden att spärra det så snabbt som möjligt (Berry & Linoff 1997).

Övriga områden: Det finns ingen direkt begränsning av områden data mining appliceras på. Andra potentiella verksamheter som kan använda sig av data mining är till exempel sjukvården, för att underlätta diagnostisering av patienter. Inom industrin kan data mining användas för process- och kvalitetskontroll samt kontroll av led- och leveranstider. Data mining kan även vara ett stöd vid rekrytering av ny personal.

2.5.7 Fallgropar med data mining.

Piatetsky-Shapiro (Koo 1998) skiljer på två områden som är kritiska för att lyckas med att införa data mining i verksamheten, dels tekniska och dels organisatoriska. Inom det tekniska området handlar det främst om saknad data, felaktig data eller andra problem som kan uppstå vid insamling och lagring av data. Inom det organisatoriska området kan problem uppstå om inte problemet är väldefinierat eller om ledningen är svag.

Enligt Piatetsky-Shapiro (Koo 1998) är gynnsamma förhållande när

"there is sufficient data, relatively error-free, and there are knowledge-based decision with high payoff, and enviroment is changing"

Tekniska fallgropar

Fallgropar som kan nämnas då det gäller den tekniska biten är:

- Datamängdens form och innehåll: Felaktiga dataformat, fem-positionsfält där man hade behövt nio-positionsfält. Otydliga datafält, där leveransdatum egentligen betyder planerat leveransdatum. Saknad data eller störande värde (Dilly 1997). Här är det viktigt att man beslutar om hur man ska hantera data som till exempel saknas eller är felaktig. En annan viktig aspekt är att besluta hur man ska uppdatera databasen och avlägsna till exempel redundanta fält. Det bör, för stora företag, finnas någon ansvarig för att kontrollera databasens fält och förändringar, så att allt som finns lagrat är aktuellt.

  En regel för att förbereda data för analys med data mining är "Garbage In Garbage Out" (Shepard 1995), det vill säga "Skräp in Skräp ut", lägger man inte ner tid på att förbereda datan kan man inte heller räkna med att resultatet blir tillförlitligt. Det gäller även om man inte får tag i all den information man behöver för att genomföra en analys. Form och innehåll av datan är kritiskt för att kunna använda data mining på ett lyckat sätt. Samma sak gäller att datan är aktuell. Kommer den en månad för sent kanske det inte går att grunda några beslut på den, eftersom den är inaktuell.

- Brist på funktionalitet, man tillåts till exempel inte att summera två fält eller dra ut en rapport per kund, vilket leder till att man inte kan få ut de rapporter man önskar. För att undvika detta är det viktigt att man lägger ner tid på att undersöka vad det är man vill ha och hur man vill att det skall se ut. (Berry, Linoff 1997)

Organisatoriska fallgropar

När det gäller de organisatoriska aspekter av data mining är det främst ledning och kompetens som är kritiska. När det gäller ledningen måste de vara insatta och engagerade så att de avsätter tid och pengar för att genomföra ett data mining projekt. Ledningen måste vara med då problemdomänen definieras eftersom det oftast är i ledningen den övergripande synen på företaget och affärsmöjligheterna finns

Kompetens

"Data mining don´t understands business, data mining products reguire statistical skills"
(Thearling 1998 a)

För att företaget, i slutändan, ska ha någon verklig nytta av det som kommer fram under analysen måste de kunna tolka och förstå resultatet. Svårigheter att tolka resultatet gäller för hela KDD-processen och inte bara för data mining. För att kvalificeras som kunskap måste de funna mönstren och sambanden förstås. Det här är en av anledningarna till att neurala nätverk inte har använts i någon större utsträckning eftersom det har varit för svårt att tyda resultatet (Askara-Gelman 1998).

I en artikel av Init Askara-Gelman (1998) diskuteras begreppet "Comprehensibility" som enligt Norstedts engelsk-svenska lexikon betyder *Begriplighet* och *förståelighet*. Hon ställer frågan hur funna mönster, samband och metoder ska definieras och mätas för att kunna bidra till ökad begriplighet. När det gäller KDD-processen och data mining är de generella kriterierna validitet, nyhetens behag, potential användbarhet och ultimat förståelse. Nyhets behag innebär att KDD-processen skall tillföra något nytt (Uthurusamy 1995).

Författaren beskriver två skäl varför förståligheten är en kritisk del av KDD-processen. För det första, resultatet kommer inte bara att användas av människor utan kommer också användas som input till ett annat program (Frawley m.fl. 1991). För det andra, den process, där data tolkas till kunskap, är iterativ och interaktiv med användaren, vilket leder till att mönster och samband tolkas samt utvärderas utifrån vad som ses som kunskap genom en människa-data interaktion och blir därigenom subjektiv (Uthurusamy 1995).

Då en ny teknik kommer ut på marknaden skapas det ofta en övertro till vad tekniken kan lösa, så också med data mining (Small 1997). Det finns en övertro på att data mining är så sofistikerade, att den ersätter domänkunskap och erfarenhet, när det gäller att analysera resultatet och konstruera modeller för att lösa olika domänproblem. Enligt Robert Small (1997) kan inga analystekniker ersätta erfarenhet och kunskap, tvärtom, data mining gör att utbildning och erfarenhet är viktigare än någonsin. Han menar vidare att en person, som inte är expert på analysverktyg, inte är till någon större nytta om han inte har någon kunskap om marknaden och det specifika affärsproblemet.

Det kan vara svårt att i verkligheten hitta en person som är både expert på analysverktyg och affärsområdet, vilket gör att man kan behöva sätta ihop ett team av kompetens, för att täcka in den analytiska förmågan och förmågan att identifiera problemen samt att se möjligheterna med de resultat som presenteras (enligt Piatetsky-Shapiro, Koo 1998).

3. Metod

De metoder vi använde oss av var intervjuer och möten med en etnografisk ansats. Vi började att studera litteratur för att öka kunskapen om data warehouse och data mining.

3.1 Våra källor

Vid litteraturstudien använde vi oss av böcker, tidningsartiklar och web-sidor inom vårt ämne. Litteraturen sökte vi främst på Göteborgs universitets söktjänst, Libris, via internet och Chalmers Bibliotek. Då vi använt oss av internetkällor har vi främst använt oss av sökmotorer med sökord som data mining, data warehouse, knowledge discovery, men även kända författare inom området, Fayyad, Piatsky-Shapiro, Inmon och Hadden.

3.2 Etnografi

Anledningen till att vi valde en etnografiskt ansats vid intervjuerna och mötet beror främst på att vi ville ha en kvalitativ metod. Vi ansåg att det var viktigt att intervjupersonerna själva skulle ha möjlighet att berätta om sin situation och sitt arbete. Etnografisk metod poängterar studier av de förutsättningar under vilka användarna arbetar och deras egna miljö, vilket skapar förståelse för arbetet. Vi ansåg att det var viktigare att få fram en djupare förståelse och ett sammanhang mellan olika faktorer och skulle av denna anledning inte vara behjälpta av en kvantitativ metod, där man får fram en stor datamängd med statistiska resultat.

Etnografisk metod jämfört med en traditionell metod

För att få en uppfattning om vad som skiljer den etnografiska metoden från den traditionella tänkte vi kort beskriva dem.

I den traditionella metoden sker de flesta observationerna inte på användarnas arbetsplats utan på en plats som ska vara tänkt att efterlikna arbetsplatsen. Kvaliteten på informationen blir då inte den bästa, eftersom det kan vara svårt för användarna att sätta sig in i hur den nya tekniken till exempel kommer att förändra deras sätt att arbeta. I den traditionella metoden är utrymmet för användarmedverkan litet eller mycket litet och istället för att förlita sig på att användarna vet vad de har för behov, utförs det tester vid några fåtal tillfällen, som sedan får ligga till grund för designen. En annan skillnad är att i den etnografiska metoden har utvecklarna kontinuerlig kontakt med användarna där en relation utvecklas som baseras på att designern finns på plats hos användaren. Designern lär känna miljön och förutsättningarna, vilket resulterar i god kännedom om vad kunden efterfrågar och behöver.

I sin artikel “Etnographic Field Methods and Their Relation to Design” beskriver Jeanette Blomberg sex skäl till varför etnografisk metod är bättre än de traditionella metoderna för att få en uppfattning om hur användarnas behov och beteende ser ut. För det första måste designern förstå hur situationen ser ut, så att tekniken, som ska implementeras, passar in. För det andra är det viktigt att vara medveten om att nya tekniker och applikationer förändrar arbetssättet för dem som ska använda dem. Det är därför viktigt att designerns och användarnas uppfattning om hur det ser ut stämmer överens. För det tredje, då applikationer

skapas utan att slutanvändaren är känd, är det svårt att passa in den hos användarna när den är klar. Vet man däremot vilka användarna är kan man förbättra designen vilket leder till att tekniken passar bättre in i den omgivningen den ska användas. För det fjärde, eftersom användarnas erfarenheter av tekniken influeras av dess innehåll, är det viktigt att få en bredare uppfattning om teknikens innehåll än vad man får genom traditionella tester. För det femte, är det svårt för användarna att ge meningsfulla svar på frågor om hur de kommer använda en teknik som är helt ny för dem. De behöver veta mer om hur tekniken kan användas för att de ska kunna diskutera vidare om den. Den sjätte och sista anledningen är, att då man använder sig av andra metoder, får man inte någon helhetssyn på arbetet utan löser en specifik uppgift istället för att se om man skulle behöva ändra andra delar av sättet att arbeta, för att införandet av en teknik ska lyckas.

Flertalet av Jeanette Blombergs sex skäl stämmer in på de förutsättningar vi hade då vi gjorde intervjuerna på Volvobolagen. Det femte skälet var speciellt uppenbart i och med att vissa av intervjupersonerna inte hade funderat i termer kring data mining tidigare och då man förklarar vad tekniken kan ge, skapar man ett behov i sig.

3.3 Intervjuer och möten

Vi använde oss av intervjuer och möten som en metod för att ta reda på hur de anställda på Volvobolagen såg på data mining och hur långt de kommit i tankar kring data mining metoder. Syftet med intervjuerna och mötet var även att få en bild av hur samling, lagring och analys av datan gick till på de olika Volvobolagen idag. På Volvia genomförde vi flera informella intervjuer, vilket är en del av den etnografiska metoden. Intervjufrågorna var indelade i två delar, en mer specificerad och en mer ospecificerad. Den första delen av intervjun frågade vi om arbetsuppgifter, utbildning och vad de visste om data mining och källan för kunskapen. Den andra delen av intervjun var mer informellt formad och tillrättalagd efter vem vi intervjuade och vilken kunskap de hade om ämnet och vilka arbetsuppgifter de hade. Det viktigaste var att intervjupersonerna fick berätta utifrån sin uppfattning och erfarenhet. På Volvia genomförde vi sammanlagt fyra intervjuer på mellan en halvtimme till en timme, varav alla spelades in på bandspelare och senare transskriberades. Vi har under intervjuerna valt att anta en sk. ”Interview guide approach”(Patton 1990) då det passat vårt syfte att inte styra intervjuerna för hårt utan strävat efter att vara flexibla.

Till skillnad från Volvia fanns det ingen möjlighet att genomföra intervjuer med enskilda personer på Volvokort. Vår analys av Volvokort har vi baserat på ett möte som varade i drygt två och en halv timme med personer som var ansvariga för data warehouse, data mining och val av tekniker som passar bolagets behov. Mötet var informellt och inleddes med att de ansvariga berättade hur verksamheten såg ut, vilka planer de hade med data mining och andra tekniker. De frågor vi ställde var inriktade på att ta reda på hur de hade tänkt använda sig av data mining i framtiden. Mötet spelades in på bandspelare och transskriberades senare.

Vi har även varit med på möten med SAS Institute som marknadsför produkten Enterprise Miner som vi varit med och beta-testat. Under dessa möten har det handlat om hur Volvo ska kunna använda sig av Enterprise Miner på bästa sätt. Vårt resultat har inte baserats på dessa möten eftersom SAS representerar ett företag som konkurrerar med andra om att hitta den bästa lösningen på data mining problem.

3.4 Kritik mot metoden

Som kritik till eget arbete kan nämnas att vad människor säger och vad de gör skiljer sig ofta åt. Vi har inte haft möjlighet att kontrollera om det de säger överensstämmer med vad de gör utan de har fritt fått berätta sin uppfattning. För att få ett bättre underlag skulle intervjuerna kunna kompletteras med studier av deras vardagliga arbete och på så sätt ge oss en djupare förståelse för deras arbete. Den metod som vanligtvis används i sådana situationer är deltagarobservationer, då forskaren strävar efter att delta i så många situationer som möjligt över en längre tid, från några månader till något år. Vi har inte haft möjlighet att genomföra en så lång studie utan har baserat vårt material på informella intervjuer och möten.

Vi upplevde vid flera tillfällen att intervjupersonerna inte alltid var så insatta i ämnet och att vi därför är medvetna om att vi påverkat intervjupersonerna något. Vi lade därför stor vikt på att intervjuerna skulle vara öppna och informella, mer i form av en diskussion. Vid flera tillfällen lade vi fram ett ämne och intervjupersonerna fick berätta utifrån sina egna referensramar om hur de uppfattade situationen och tekniken. De vi intervjuade utgjorde inte någon homogen grupp utan hade olika bakgrund och skilda arbetsuppgifter.

Ett problem vi stött på är att det är svårt att gå in i en organisation och vara helt objektiv. Under en etnografisk studie ska man inte påverka personerna man intervjuar, vilket kan vara svårt att undvika. Det kan även vara svårt att få fram den information man söker eftersom det ofta finns så kallade “gatekeepers” (Morgan 1986) i företag, det vill säga personer som kontrollerar informationsflödet. I vårt fall har det handlat om att bolagen vill skydda information som de ser som en konkurrensfördel mot andra företag inom samma bransch och det måste vi respektera.

4. Resultat

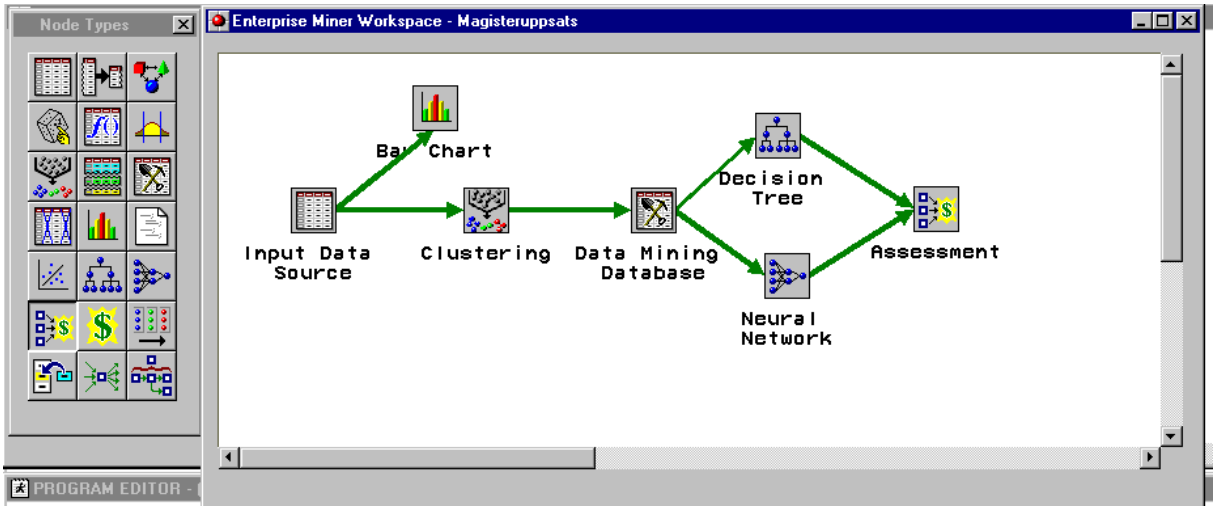
Utgångspunkten för den här uppsatsen var att studera hur data mining fungerar och hur två olika Volvobolag kan dra nytta av data mining för sina behov. Nedan kommer vi att beskriva bolagen lite närmare var för sig och redovisa resultaten från intervjuerna och mötena vi genomfört. Vi kommer även att presentera den produkt som vi varit med och beta-testat på Volvo IT, SAS Enterprise Miner. Vidare kommer vi att presentera ett lösningsförslag. Lösningsförslaget är uppdelat i två delar, en del som är inriktad på att presentera de data mining tekniker vi anser att de olika bolagen kan använda sig av för att lösa de problem vi identifierade under vår studie av bolagen (Tekniska Lösningsförslag). Den andra delen är mer inriktad på de organisatoriska faktorer som vi anser vara viktiga att tänka på när man ska införa ett data mining verktyg i verksamheten (Organisatoriska Lösningsförslag).

Vid närmare undersökning insåg vi att bolagen hade kommit olika långt med tankar, utveckling och beslut om data mining. På Volvia har man kommit långt i sina tankar kring data mining och har tillsammans med Volvo IT under våren testat en data mining produkt, Enterprise Miner, från SAS Institute, för att i första hand stödja kampanjerna och förbättra kundrelationen. På Volvokort har man inte kommit lika långt, vilket påverkar deras uppfattning om data mining. Vi kommer först att presentera bolagen var för sig, men under avsnittet med lösningsförslag, presentera dem tillsammans.

Presentation av data mining produkt, SAS Enterprise Miner

Enterprise Miner är en integrerad mjukvaruprodukt som erbjuder företag en helhetslösning för data mining (http://www.sas.com). Produktens syfte är att hjälpa företags beslutsfattare att hitta trender i stora mängder data. Det finns vissa nyckeldelar som ingår i den här produkten. Det första är att programmet använder sig av ett antal tekniker för data mining som, till exempel beslutsträd och kluster men också regression och neurala nätverk. Det finns även funktioner som gör urval (sample) från stora datamängder. Visualiseringen sker i form av grafer, trädstrukturer och tabeller.

Produkten har ett grafiskt gränssnitt som är användarvänligt med ”klicka och dra- symboler” som motsvarar de tekniker man använder för data mining. Det hela visas i form av projekt där man tillämpar teknikerna för att på olika sätt manipulera data för att fram ett resultat (se figur 4.1). Programmet är uppbyggt kring SAS Institutes data mining metod SEMMA som nämndes tidigare.



Figur 4.1. Exempel på hur ett projekt kan se ut i Enterprise Miner.

Hur Enterprise Miner tillämpar SEMMA

I figuren ovan visas hur Enterprise Miner tillämpar de olika stegen i SEMMA och i det här avsnittet förklarar vi på vilket sätt den gör det.

- **Urval** (Sample)-
  Om man gör ett urval av data som ska användas för data mining reduceras processtiden och resultaten produceras snabbare. Produkten kräver inte att man ska göra urval, men det underlättar för efterföljande procedurer. De tekniker man kan använda är till exempel att reducera de avvikande värden som inte har någon påverkan på resultatet.
- **Undersökning** (Explore) -
  Diagram och kluster är exempel på tekniker som Enterprise Miner använder för att undersöka data för att få perspektiv på den befintliga datan. Använder man dessa tekniker underlättas sökning av mönster i data.
- **Manipulering** (Manipulation) -
  För att manipulera data kan man i produkten välja vilka variabler som är intressanta att undersöka och vilka kombinationer av variabler som kan ge intressanta resultat.
- **Modellering** (Modeling) -
  Det finns många algoritmer i produkten, framför allt statistiska modeller men även tekniker som neurala nätverk och beslutsträd. Alla modeller som konstrueras lagras automatiskt i en så kallad Model Manager. Modellerna är sedan alltid tillgängliga.
- **Jämförelse** (Assess) -
  Det går att visuellt jämföra modellernas resultat som finns lagrade i Model Manager. Det går att jämföra resultaten från olika tekniker som användes, till exempel resultaten från neruala nätverk jämförs med resultaten från modellen för beslutsträd. Resultaten från den här fasen ska vara till hjälp för att se vilken teknik som är mest lönsam för företaget.

4.1 Volvia, Volvos försäkringsbolag

Bakgrund

Volvia genomför varje år ett antal kampanjer. Innan kampanjerna genomförs görs först ett urval. Volvia har tidigare satsat på ett brett urval men vill i framtiden satsa på mer lönsamma grupper. En målgrupp de är intresserade av och vill nå är Volvokorts kunder. Urvalen har baserats på tidigare erfarenheter från kampanjer samt den information de får från Volvohandeln som de har ett samarbete med. Efter urvalet görs ett brevutskick till kunderna som därefter blir kontaktade av en person ur telemarketing-gruppen som kontrollerar så att kampanjen nått fram och hör efter om kunden är intresserad. Visar kunden intresse skickas det därefter en offert till kunden (Se bilaga med rich piture).

Kampanjresultet bedöms efter hur lyckad den är. En lyckad kampanj bedöms efter vilket förhållande Volvia och kunden har. Kunder som sedan tidigare har en relation med Volvia, det vill säga redan har en del av sina försäkringar där, är de mer benägna att svara på en kampanj.

För ett par år sedan beslutade Volvia att inskaffa mer traditionell försäkringskunskap. I och med detta beslut anställdes en försäkringsanalytiker och arbetet att förvärva och hantera data om kunder och försäkringar sattes igång. På Volvia finns det sedan tidigare ett system för den operativa datan som heter FINESS. Nu startades ett samarbete med Volvo Data, numera Volvo IT, för att bygga ett stödjande system till FINESS, som fick namnet FINAL. Stödsystemet FINAL innehåller tariffanalys, riskstatistik, reservanalys, affärsresultat och kundsegmentering. FINESS matar FINAL med data för analyser med olika intervall för olika ändamål. Mellan de båda systemen finns ett gränssnitt, en urvalsgenerator, som underlättar samarbete mellan de båda. Volvia har lagrat allt som hänt med kunderna och deras försäkringar under de senaste tre åren. Det är flera miljoner poster och det krävs ett verktyg för att kunna dra nytta av all data de lagrat. Det är, som en del i steget för att kunna utnyttja den lagrade datan som Volvia även har börjat undersöka möjligheterna med data mining.

4.1.1 Intervjuer

De intervjuer vi genomförde på Volvia, var som nämnts tidigare, till största del informella bortsett från den första delen av intervjun då vi ställde frågor för att veta vad intervjupersonen hade för arbetsuppgifter, bakgrund, utbildning och kunskap om området data mining.

Arbetsuppgifter: Anledningen till att vi ställde frågor om intervjupersonernas arbetsuppgifter var för att ta reda på hur mycket de hade att göra med frågor som rör data mining och spar- och lagringsfunktioner i företaget. De visade sig att uppfattningarna skilde sig något då det gällde nyttan av data mining, vilket vi tror beror på vilka arbetsuppgifter de hade.

Bakgrund och utbildning: För att få reda på vad intervjupersonerna hade arbetat med tidigare, frågade vi om deras bakgrund. Vi märkte att intervjupersonerna hade väldigt olika bakgrund och utbildning. Det var dels personer med en gedigen statistisk utbildning och dels personer med rent ekonomisk bakgrund med inriktning på marknadsföring. Vi anser att svaren beror på vilken bakgrund man har och vilka referensramar man har till ämnet. Framförallt är personer med statistisk utbildning mer insatta i ämnet eftersom data mining bland annat utvecklas från statistik. Personer med ekonomisk utbildning inom marknadsföring har ett annat

perspektiv på data mining. Det är även skillnad mellan vad de med ekonomisk och statistik bakgrund tror att data mining kan lösa för problem.

Kunskap om Data mining och källa för kunskapen: Anledningen till att vi ville veta vad intervjupersonerna kände till om data mining var för att få en klarare bild över hur insatta de var i ämnet samt varifrån de lärt sig det. I och med att data mining är ett förhållandevis nytt begrepp för många företag var det inte så många av intervjupersonerna som var insatta i ämnet.

4.1.2 Förutsättningar och problem

Vi har främst tittat på hur Volvia kan förbättra sin marknadsföring med hjälp av data mining. Inom marknadsföring har vi skilt på två olika delar, dels den del som rör kampanjer (Direct Marketing) och dels den del som berör kundrelation (Relation Marketing). Kampanjer och kundrelation ligger nära varandra och ingen av dem utesluter den andra. För att genomföra lyckade kampanjer måste man känna sina kunder, vilket innebär att man behöver ha en god kundrelation. I resultatavsnittet kommer vi först att presentera resultatet för kampanjer och sedan för kundrelationen.

Kampanjer

En svårighet Volvia har är att genomföra lyckade kampanjer. De har under intervjuerna uttryckt viljan att öka svarsfrekvensen genom att satsa på ett data mining verktyg. En av intervjupersonerna beskriver det så här:

"Man kastar ju pengarna i sjön om man vänder sig till personer som aldrig svarar. Jag vill skapa ett verktyg för att öka träffsäkerheten. Idag kan vi bara läsa av den sk hit-raten, det vill säga hur många procent som svarar på olika aktiviteter. I framtiden ska vi kunna klassificera kunderna, få kunskap om vad som kännetecknar de som svarar."

För att kunna förbättra urvalet och svarsfrekvensen kan man använda sig av olika tekniker för data mining, lösningsförslagen presenteras här nedan. Ett annat problem Volvia har med kampanjerna är svårigheten att snabbt kunna följa upp en kampanj på ett effektiv sätt. Idag tar det upp till en månad innan de som är ansvariga för kampanjerna kan se hur utfallet blivit. Det finns ett behov att interaktivt kunna se hur utfallet blir under tiden kampanjen pågår. En intervjuperson uttryckte det så här:

"Jag skulle vilja ha en kurva ungefär hur snabbt kunderna svarar och se hur den fortsätter på en viss linje, det kan jag inte göra idag. Jag vill kunna avbryta en kampanj på någon vecka eller bara på någon dag bara. Idag får jag se kanske två, tre veckor och det är alldeles för lång tid för mig"

Med en funktion där man kan följa kampanjernas utfall interaktivt kan man även välja att avbryta en kampanj om man ser att utfallet är dåligt eller mycket dåligt.

"Vi ska kunna analysera interaktivt under kampanjens gång, så att vi varje dag kan avläsa resultatet och successivt zooma in rätt målgrupp. Om vi till exempel gör ett DM-utskick [Vilket avser på de kampanjutskick (direct marketing) de gör till sina kunder] *och får in*

spontana svar kan vi göra analyser varje dag och se vart det pekar. Sedan kan TM-avdelningen börja ringa mot de målgrupper som vi fått fram i analyserna. Ju mer detaljerat vi segmenterar kunderna, desto mer ökar träffsäkerheten. På sikt kommer vi att bygga upp en kompetensbas, med kunskap om vilka erbjudande och budskap som attraherar vilka kunder. Då kommer vi att kunna maximera utfallet per kampanj"

Att följa upp kampanjerna interaktivt är ett effektivt sätt att ta reda på om Volvia har valt att satsa på rätt kundgrupp.

De problem Volvia har med kampanjerna kan lösas med att förbättra kunskapen om sina kunder och göra en kundsegmentering, det vill säga att välja ut kunder och potentiella kunder utifrån vilka de är och vilka behov de har. Här kommer vi in på det andra område vi anser att Volvia kan dra nytta av data mining, vilket är att förbättra kundrelationen. För att kunna förbättra sin kundrelation måste man lära känna sina kunder. Resultatet kan bli en kundsegmentering som kan resultera i ökad svarsfrekvens på kampanjerna.

Kundrelation

Volvia har liksom många andra stora företag svårt att etablera en personlig kundkontakt med alla sina kunder. I teoriavsnittet förklaras problemet med större företag som försöker efterlikna mindre företags förmåga att etablera en personlig kundkontakt för att få lojala kunder. Vid flera tillfällen under intervjuerna uttrycktes viljan att lära känna sina kunder bättre för att på så sätt kunna ta fram potentiella kunder i framtiden.

"Genom att analysera vår kundstock och de kunder som lämnar oss, kan vi lära oss väldigt mycket om hur vi ska vårda våra kunder. Vi kan se vad som kännetecknar lojala kunder och vad de har för önskemål."

Att etablera en god kundrelation handlar om att lära känna kunderna, deras önskemål och behov. Det finns mycket att vinna på att känna till vad det är för kunder man har. Vi har identifierat olika sorters kunder Volvia har (se bilaga 1, rich piture)**.** Kundgrupperna delas in i nöjda-, missnöjda-, dyra- och okända kunder. Nöjda kunder är kunder som köper Volvias budskap och är trogna kunder. Genom att se vad det är som är specifikt för nöjda kunder kan man hitta tvillingsjälar i nya kunder för att se vilka som har samma förutsättningar att bli trogna kunder. Genom att upptäcka missnöjda kunder kan man förekomma sina kunder innan de lämnar bolaget och samtidigt se om det finnas andra kunder som är missnöjda. Att en kund är missnöjd kan bero på missförstånd som kan redas upp om Volvia har chans att få reda på vilka de är. När det gäller dyra kunder gäller det för Volvia att kunna spåra vilka som är olönsamma för dem. När de dyra kunderna har spårats kan man sätta rätt premie på dem. Bara för att en kund är dyr betyder det inte att kunden är olönsam så länge kunden betalar rätt premie. Ett problem i försäkringsbranschen är att det finns en grupp kunder som drar på sig mycket skador och blir en dyr och olönsam kund för försäkringsbolagen. Av denna anledning byter dessa kunder ofta bolag för att på så sätt komma undan att betala en högre premie. Om Volvia lär känna sina kunder kan de undvika att göra samma misstag och kan istället sätta rätt premie på kunden från början. När det gäller att finna de kunder som tillhör gruppen okända kunder gäller det att känna till sina egna befintliga kunder så väl att Volvia kan undersöka vilka potentiella kunder som liknar dem.

När Volvia lärt känna sina kunder kan de satsa på att förbättra kampanjer och kundrelationen. Ett exempel är att de kan ta reda på var i livscykeln en kund befinner sig i, det vill säga var i sitt liv befinner sig kunden och vilka behov har han/hon. Kunden kan till exempel vara fembarnsfar och behöver en stor bil med gott om plats och använder bilen för att köra och hämta barn till och från fotbollsträningar till åka och storhandla, eller så är kunden nyligen pensionerad och använder sin bil på soliga söndagseftermiddagar.

En annan fördel med förbättrad kundvård genom data mining är att då telemarketing-gruppen på Volvia ringer upp en kund har de redan en god bild på vem kunden är och vilka önskemål och behov kunden har. Kunden känner i sin tur att försäkringsbolaget verkligen känner och förstår sig på honom. En liknade situation uppstår om kunden själv tar kontakt med Volvia och den som tar emot samtalet vet vilka till exempel försäkringar och premier kunden haft tidigare. Kunden slipper upprepa samma uppgifter flera gånger och samtalstiden kan förkortas.

Övrigt
Ett annat problem vi identifierade, som hamnar utanför de problemområden som tidigare nämnts, är att Volvias IT-avdelning i dagsläget är något överbelastad. De som arbetar på IT-avdelningen har fullt upp med den dagliga driften och har svårt att ta på sig ytterligare uppgifter. IT-avdelningen på Volvia består idag av 13 personer varav 6 st är konsulter som är inhyrda för att minska belastningen för de anställda. Volvia har satsat mycket resurser på IT under de senaste åren och har en välutvecklad IS/IT strategi. Problemet med överbelastning gör dock att det kommer dröja innan de kan ta på sig fler uppgifter i from av nya tekniker som ska införas.

4.2 Volvokort

Bakgrund

Volvokort lagrar stora mängder data som innehåller information om kunderna, vilka bilar de äger, kreditupplysningar mm. Många företag har problem att använda all data, det vill säga de har svårt att utnyttja den data som de har lagrat undan. Problemet är att det saknas effektiva metoder för analys.

Hur har vi fått fram information om Volvokort?

För det första har vi fått reda på den allmänna informationen från deras hemsida (http://www.volvo.com). Vi har dessutom deltagit på ett möte tillsammans med de ansvariga för data warehouse och data mining hos Volvokort. Till skillnad från Volvia fanns det ingen möjlighet att göra några intervjuer med enskilda personer utan vår analys baseras på det som de berättade för oss vid mötet.

Hur fungerar det på Volvokort?

I början av årsskiftet installerade de ett nytt data warehouse där de kan lagra all sin data. Det har nyligen tagits i bruk och därför har de inte kommit så långt i tankarna kring data mining. I framtiden kommer de att undersöka huruvida de anser att de kan använda sig av data mining. Eftersom de nyss har installerat sitt data warehouse är de i ett annat stadium än Volvia som har använt sitt data warehouse i över ett år.

4.2.1 Problem

Här presenteras de problem som vi anser att Volvokort har. Ett problem som vi upptäckt vid analysen av Volvokort är att de har stora mängder av data, men det är svårt att utvinna nyttig information ur den, eftersom det inte finns någon bra datastruktur.

”*Det största problemet är att vi råkar ut för den stora mängd data, men vill hitta den kritiska affärsinformationen, det är det som är det svåra hela tiden.*”

Många företag har problem med att veta hur de ska hantera insamlad data som handlar om kunderna. Volvokort använder sig av något som de kallar förädlingsprocess, de har skapat en databas där de manuellt har kategoriserat in kunder i olika grupper beroende efter vilken typ av kund de är.

Vid analyser för att komma fram till hur vinsten är i olika regioner i landet, kan man än så länge inte gå djupare än till regionsvis. De kan inte se vilken bil eller vilken kundtyp som det är de går förlust med.

Ett annat problem är att de producerar en massa data som de inte använder. Orsaken till detta kan vara att det är en tidsödande process att analysera all data som finns lagrad, hittills har de inte behövt den informationen som de kan få fram med hjälp av analys. Nu vill de förändra strukturen för att analysera vilka kunder de har, vilka de förlorar och vilka de kan förvärva, då behövs det verktyg som kan analysera och hantera datan.

Ytterligare ett problem är att det tar tid att hitta rätt analysverktyg. Ett av Volvokorts största problem är att de har stor tidsbrist och kan inte lägga ner mycket tid på att leta efter ett tillfredsställande analysverktyg och ingen tid att undersöka hur data mining kan vara till hjälp för dem.

Nu använder de OLAP-applikationer för att analysera sin data. Nackdelen med OLAP är att den tekniken bara kan rapportera om data, data mining däremot kan hitta mönster i data.

”*OLAP tar in och analyserar. Data mining, det borde vi också ge oss in på men vi vill veta lite mer om vad vi är ute efter.*”

4.3 Tekniska Lösningsförslag

Då det gäller de olika problemområden Volvia och Volvokort har måste man angripa dem med olika metoder för data mining. För att Volvia ska kunna följa upp kampanjer passar en modell som utgår från verifikations-metoden, vilket innebär att man testar en hypotes. Då det gäller att förbättra kundrelationen behöver Volvia en discovery- metod där man söker samband som inte tidigare var kända. Volvokort behöver metoder för att kunna strukturera sin datamängd och lösa de problem som påverkas av att de har ostrukturerad data. Vi ger här ett lösningsförlag för vilka tekniker bolagen kan använda, hur de ska använda teknikerna och på vilket sätt Enterprise Miner kan vara till hjälp.

4.3.1 Klustertekniken

För att veta vad som identifierar de poster som finns i olika kluster bör man i modellen för kluster specificera vilka variabler som ska påverka grupperingen. Variablerna skall kännetecknа posterna som skall ingå i de olika klustren. När tekniken är använd finns det en övergripande struktur över kundtyperna.

Volvia: Klustertekniken är något som Volvia kan använda sig av för gruppering av de olika kundtyperna. Medlemmarna i de olika klustren har fler associationer med varandra än poster i andra kluster. Tekniken underlättar för andra data mining tekniker vid analys av data.

Volvokort: För att lösa problemet med att Volvokort har stora mängder ostrukturad data, som gör det svårt att utvinna nyttig information, kan de använda sig av klustertekniken. Med kluster kan bolaget gruppera sin data, där det är mest intressant för Volvokort, liksom för Volvia, att i första hand gruppera de olika kundtyperna.

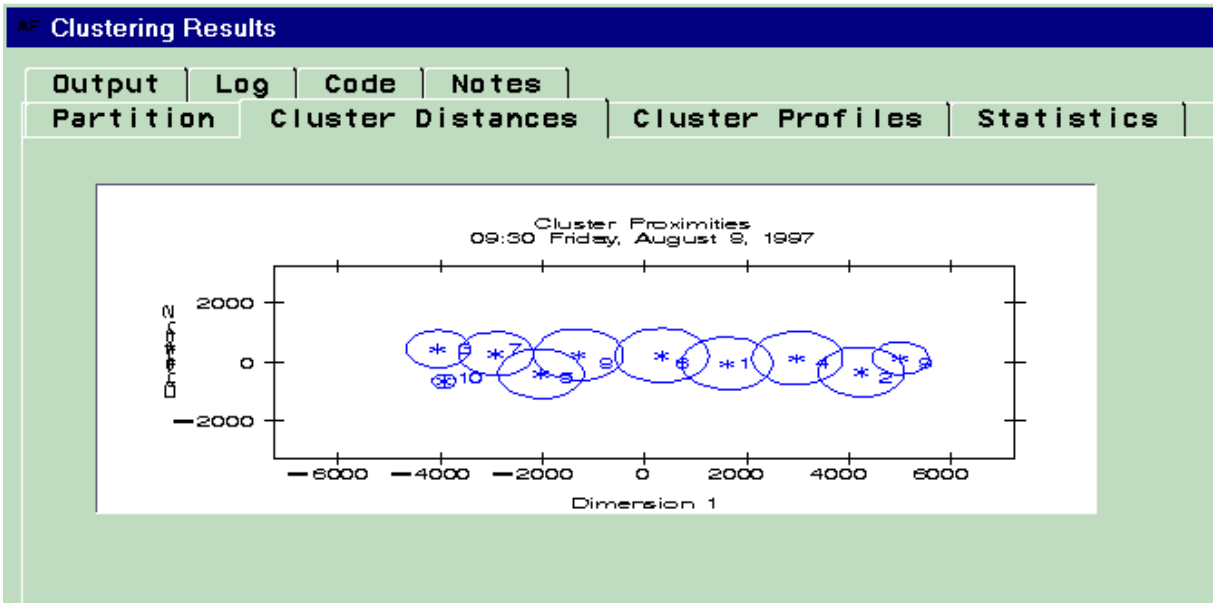
Enterprise Miner

För att underlätta för klustertekniken bör man behandla den totala datamängden först, bland annat för att få bort avvikande värden som inte kommer att påverka resultaten. I Enterprise Miner kan man behandla data på flera olika sätt, förutom att ta bort avvikande värden, kan man göra ett urval av den totala mängden poster och om det finns ett behov går det att standardisera variabler.

I produkten finns det en nod som heter ”*clustering node*”. Vid användandet av den här noden kan man i dess dialogruta specificera de val som ska gälla vid grupperingen och även bestämma totala antalet kluster som ska existera.

I en annan dialogruta ”*seeds dialog page*” specificeras kännetecknen för klustren, hur de ska uppdateras och kontrollera att rätt antal kluster skapas och att det inte blir fler än angivet.

Resultaten visas i form av tabeller och diagram. Ett diagram visar de unika kännetecknen för varje enskilda kluster, ett annat visar storleken av varje kluster och förhållandet mellan olika grupper (se figur 4.2). Det finns även en grafisk presentation över indata variablerna till de olika grupperna samt statistisk information om varje kluster. Den här tekniken är en bra utgångspunkt för fortsatt analys med data mining.



Figur 4.2. Exempel på hur ett diagram som visar kluster kan se ut i Enterprise Miner. Bilden visar förhållandet mellan olika kluster.

4.3.2 Beslutsträd

Med beslutsträd kan företag analysera trender från historisk data för att sedan kunna förutsäga framtida trender, till exempel analysera de mest lönsamma kunderna från tidigare år för att kunna förutsäga vilka kunder som är lönsamma i framtiden. Det viktiga är att formulera reglerna i trädet på ett sätt som leder till rätt kategorisering av kunderna.

Volvia: De kan använda beslutsträd för att identifiera de kunder som är positiva till Volvias försäkringskampanjer och vilka som svarar på deras kampanjer.

Volvokort: Bolaget analyserar inte vilka kunder de tjänar pengar på. För att informationen ska vara till nytta är det bättre att analysera kundgrupper istället för enskilda kunder, vilket är viktigt för större företag med dålig översikt av sina kunder. Klassificering av olika kundtyper med beslutsträd kan leda till att resultatet redovisar vilka kunder som är mest lönsamma.

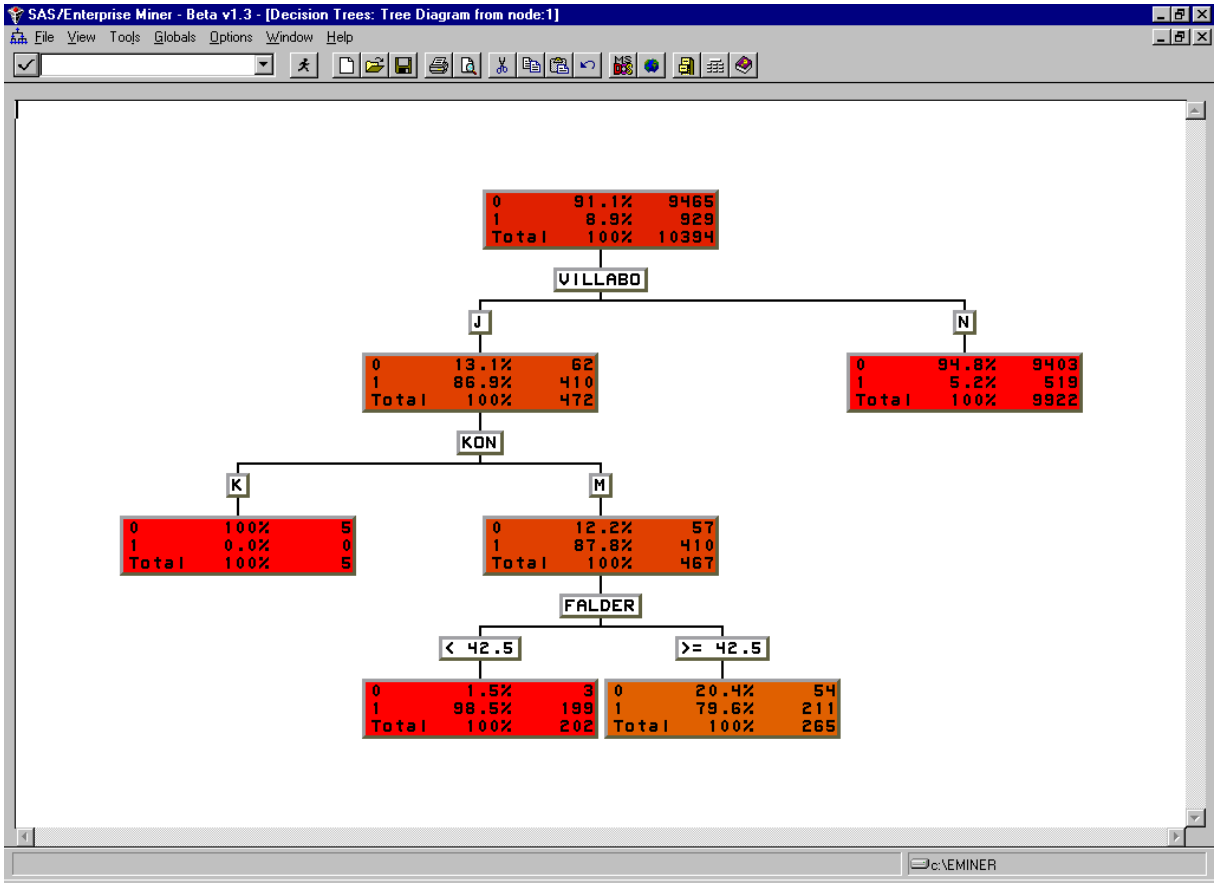
Enterprise Miner

I Enterprise Miner kan beslutsträd användas för tre olika saker, det första är klassificering av observationer, förutsäga resultat för intervall och förutsäga det bästa alternativet av flera olika beslutsalternativ.

Det första man skall göra i Enterprise Miner är att definiera den modell som ska forma reglerna som genereras från beslutsträdet och modellen kan sedan användas för klassificering av nya poster. I en dialogruta väljer användaren vilka variabler som ska användas i modellen för att utforma reglerna och klassificeringen. Det finns tre olika algoritmer som kan användas för att bygga ett beslutsträd och de är CART, CHAID eller C4.5. Efter att en algoritm har valts kan

användaren bestämma vilken struktur det producerade trädet ska ha, till exempel ska trädet producera så många löv som möjligt eller så kan användaren bestämma det maximala antalet löv som ska finnas i beslutsträdet. Modellen för beslutsträd presenterar regler och logiska påståenden med if-then satser och reglerna används till förutsägelser om nya datamängder. Resultaten visas, förutom som beslutsträd, även i form av statistiska tabeller och andra grafer.

I figur 4.3 visar vi ett exempel för hur ett beslutsträd kan se ut, där klassificeras kunderna efter om de bor i villa, vilket kön de har och hur gamla de är.



Figur 4.3 föreställer ett beslutsträd som visar hur ett resultat från SAS Enterprise Miner kan se ut.

4.3.3 Sekvensbaserad analys

Med sekvensbaserad analys kan företag identifiera vilka saker som sker samtidigt i en given händelse eller post. Analysen visar information om vad kunderna handlar för att ge en insyn till vilka kunderna är och varför de agerar som de gör.

Volvia: Företaget kan använda sig av sekvensbaserad analys för att få reda på vilka försäljningskampanjer som är lyckade och vilka som fick ett sämre resultat. Från sekvensbaserad analys presenteras resultaten i form av associationsregler som är formulerade som if-then-satser. Med resultatet från analysen kan Volvia få reda på vilka produkter och tjänster som är relaterade till varandra och vilka kunder som köper vad.

Enterprise Miner

I Enterprise Miner kallas tekniken association och används vid identifiering av saker som sker samtidigt i en given händelse. Ett exempel för hur en regel kan se ut:

if en volvobil köps hos en säljare för Volvohandeln
then sker även köp av en Volvia-försäkring
vid X procent av tillfällen

Procentsatsen beskriver sannolikheten att A och B inträffar samtidigt och definieras som sannolikhetsfaktor (confidence factor) i Enterprise Miner.

Skriver man regeln generellt ser den ut så här:
if A then B
där B en konsekvens av A.

4.3.4 Neurala nätverk

Både Volvokort och Volvia kan använda sig av neurala nätverk för att kunna placera in okända poster i de redan kända och klassificerade kategorierna. Nackdelen med neurala nätverk är att man bör ha kunskap om tekniken för att kunna tolka och förstå resultaten. För att kunna förbättra resultatet bör man förstå de specifika detaljer som finns inne i nätverket och som leder till resultatet. Ytterligare en nackdel är att man inte vet hur nätverket har kommit fram till resultatet. Det finns en modell som tekniken använder, men som användare kan man inte veta vilken det är, eftersom modellen fungerar som en ”black box”. Enterprise Miner stödjer användning av neurala nätverk, men eftersom tekniken kräver mycket av de personer som ska analysera resultatet har vi valt att inte beskriva neurala nätverk i Enterprise Miner utan återkommer till det i diskussionsdelen.

4.3.5 Minnesbaserat resonemang

Volvia vill kunna ”sikta bättre” vid kampanjurval för att ta reda på vilka kunder som ska få ett erbjudande. Det finns en teknik som heter minnesbaserat resonemang som, enligt Berry och Linoff (1997) ska användas just för kampanjurval. Nackdelen är att den här tekniken inte är representerad i Enterprise Miner. För att kunna analysera sina kampanjer bör de istället används sig av beslutsträd eller neurala nätverk.

Sammanfattning av de tekniska lösningsförslagen

För att göra en enkel bild över vilka tekniker vi har kommit fram till att de två bolagen kan använda har vi sammanställt en tabell. I tabellen visas vilka tekniker som kan tillämpas för data mining och i kolumnerna för Volvia och Volvokort förklarar vi hur de kan använda teknikerna (se tabell 4.1).

Tekniker	**Volvia**	**Volvokort**	**SAS Enterprise Miner**
Kluster	Använda kluster för gruppering av kundtyper.	Gruppera sin data för att få struktur över den.	Hanterar den här tekniken och presenterar resultatet i form av diagram och statistisk information.
Sekvensbaserad analys	För att identifiera vilka händelser som sker samtidigt kan Volvia använda analysen.		Produkten använder associationsregler för att visa resultaten.
Minnesbaserat resonemang	Volvia kan använda tekniken för att analysera sina kampanjer.		Enterprise Miner stödjer inte den här tekniken.
Länkanalys			
Beslutsträd	Med regler i beslutsträd kan Volvia bland annat identifiera vilka kundtyper som är intresserade av Volvias försäkringar.	För att få en bättre översikt av sina kunder kan Volvokort använda beslutsträd för formulering av regler.	Produkten stödjer den här tekniken och kan med hjälp av tre olika algoritmer (CART, CHAID och C4.5) presentera resultaten.
Neurala nätverk	Kan använda tekniken för att identifiera okända poster.	Volvokort kan också använda tekniken för att dela in okända poster i rätt kategori..	Enterprise Miner hanterar tekniken, men en nackdel är att man måste ha bra kunskap om hur neurala nätverk fungerar för att kunna tolka resultaten.
Genetiska algoritmer			

Tabell 4.1. Visar vilka tekniker de olika företagen kan använda och huruvida produkten stödjer dem.

Anledningen till att vi inte tar med genetiska algoritmer och länkanalyser som tekniska lösningsförslag är vi anser att de inte är tillräckligt utvecklade för att användas för Volvia och Volvokorts räkning.

4.4 Organisatoriska lösningsförslag

4.4.1 Ledning

Volvia: Ledningen är för Volvia inget problem då det gäller planeringen av data mining. Anledningen är att de är mer inne i fasen av att välja tekniker och verktyg. Då de börjar söka efter nya användningsområden för data mining behövs det mer styrning från ledningen för att förankra analyserna i verksamheten. Ett annat beslut ledningen bör fatta är huruvida de ska anställa fler personer till IT-avdelningen. Om Volvia beslutar sig för att införa data mining verktyg i verksamheten kan belastningen bli större på en redan hårt belastad avdelning.

Volvokort: Eftersom Volvokort inte har kommit lika långt med planeringen om data mining och de inte riktigt vet vad de behöver, behövs en utredning om vilka behov och problem som finns. Det är inte alls säkert att data mining är lösningen för deras problem, men det viktigaste är att de avsätter resurser för att tillsätta en grupp personer som kan reda ut vad som behövs. Det är viktigt att ledningen är med och fattar beslut om resurser som ska avsättas för ett projekt att reda ut data minings möjligheter. Ledningen bör vara med och definiera problemdomänen eftersom de besitter en unik företags- och affärsöversikt.

4.4.2 Resurs

Volvia: Eftersom det bara finns en person på Volvia som håller på med analysverktygen och undersöker vad som kommer passa in i verksamheten kan det bli ett problem med tiden. Som det är tänkt nu kommer en och samma person att sitta och tolka resultaten och översätta dem till tydliga och lätta att agera på. Risken är att den personen lätt kan bli överbelastad och de övriga som håller på med kampanjerna kommer bli tvungna att vänta på att resultatet från analysen blir klart. En annan möjlighet Volvia har undersökt är att med Enterprise Miner bygga en modell som de som sitter med kampanjurvalen får, där gränsnittet är enkelt och lättydligt så ytterligare tolkning inte behövs. Detta är dock inte klart ännu.

Volvokort: När det gäller Volvokort uttryckte de på mötet att de upplevde att de inte hade tid med att undersöka olika verktyg för data mining och dess möjligheter. Vilket innebär att om de ska kunna gå vidare måste de få tid och pengar avsatta för att kunna undersöka möjligheterna. De finns idag endast två personer som är ansvariga för data warehouse och andra analysverktyg som till exempel OLAP. För att de ska kunna utvecklas vidare måste de få mer tid och eventuellt fler personer i denna grupp.

4.3.3 Kompetens

När man ska starta ett projekt för data mining är det viktigt att det består av personer med olika kompetenser. Det är främst två områden som är viktiga, statistik- och analyskompetens och affärs- och företagskännedom.

Volvia: På Volvia finns det en person med gedigen statistisk kompetens som är ansvarig för att välja en produkt som passar. Samma person har också arbetat inom försäkringsbranschen i över tjugo år och har en god inblick i affärsproblemen och domänproblemen Volvia har. Det

kan dock, som även nämndes i resursavsnittet, bli för stor belastning på denna person om data mining användningen utvidgas och nya användningsområden söks.

Volvokort: På Volvokort finns inte samma kompetens i en motsvarande person utan den gruppen är sammansatt av marknadsförare. De skulle behöva en person som har analysförmågan att undersöka de olika statistiska resultat som data mining ger.

Sammanfattning

Som en sammanfattning visas i tabell de viktigaste delarna för att lyckas med ett data miningprojekt (se tabell 4.2).

Ledning	- Är delaktiga i beslut som rör problemdomänen. - Avsätter resurser för projektet i form av kompetens, tid och pengar
Resurs	- Projektmedlemmarna får tid och resurser att undersöka olika alternativ, för att hitta optimal lösning för företaget
Kompetens	- Projektet består av en blandning av spetskompetens, en del som kan analysvektyg och en del som kan verksamheten.

Tabell 4.2. Visar en sammanfattande bild över de organisatoriska faktorerna för att lyckas med data mining.

5. Slutsats och Diskussion

Syftet med vår uppsats var att undersöka potentialen för att bolag inom försäkrings- och kreditkortsbranschen kan använda sig av data mining för att analysera sina datamängder. För att ha möjlighet att studera detta i praktiken har vi utfört en studie av Volvia och Volvokort.

De kritiska faktorer, för att införa data mining i en verksamhet, vi identifierat under vår studie av bolagen är ledning, resurs och kompetens. Volvia har kommit långt med att införa data mining, de har undersökt möjligheterna att använda en data mining produkt, SAS Enterprise Miner. När vi startade vår studie i början av januari hade Volvokort nyligen börjat fundera kring data mining och vad det skulle innebära för dem. Vi kom fram till att de både bolagen har olika behov i nuläget och deras problem går att lösa med hjälp av olika data mining tekniker som presenterades i resultatdelen. För att införandet av data mining skall lyckas bör företagsledningen avsätta resurser såsom tid, pengar och kompetens. Resurserna ses som kritiska faktorer. Kompetensen är en viktig del om införandet och användandet av data mining ska lyckas. Det är så viktigt att ett projekt att införa en data mining teknik står och faller om det inte finns tillräcklig kompetens i företaget.

Som det ser ut i nuläget behöver Volvokort satsa mer på det organisatoriska delarna av att införa data mining, det vill säga avsätta resurser och medel för att undersöka problemdomänen och på vilka områden man har störst nytta av data mining. Vi anser vidare att de kan behöva komplettera sina analysverktyg med data mining. I dagsläget använder de sig av OLAP, men eftersom de har investerat i ett data warehouse kan man använda det med större effektivitet om de använder OLAP som ett komplement till data mining. I teoridelen presenterar vi från bland annat Inmon (1996a) att man kan använda ett data warehouse till så mycket mer om man använder sig av data mining. Berry och Linoff (1997) går så långt som att säga att data warehouse tillför ett minne till företaget men det är data mining som tillför intelligensen. Om vi ser till den produkt vi tittat på stödjer den flera tekniker som Volvokort skulle kunna använda sig av, till exempel kluster och beslutsträd (se tabell 4.1).

Under vår studie har vi undersökt hur bolagen kan använda sig av data mining och om SAS Enterprise Miner kan användas för deras behov. Volvia har, till skillnad från Volvokort, en klarar bild på vad de efterfrågar och har en god bild på vad de kan få genom att använda sig av Enterprise Miner. De tekniker de kan använda för att lösa de problem vi identifierade skulle kunna vara kluster, sekvensbaserad analys och minnesbaserat resonemang. Den sistnämnde tekniken finns inte med i Enterprise Miner från SAS. De kan komplettera med en analys som stödjer minnesbaserat resonemang men de kan också använda sig av de andra teknikerna såsom beslutsträd och neurala nätverk. Problemet med neurala nätverk är att det måste finnas någon som kan tolka resultatet eftersom resultatet av ett neuralt nätverk är svårt att tolka om man inte vet vad som finns bakom. Det fungerar som en black-box där inmatningen och resultatet är känt men processen däremellan är okänd.

I nuläget finns det inget direkt samarbete mellan de olika bolagen. Ett närmare samarbete mellan de båda anser vi skulle vara till fördel för bolagen. Vi anser att ett samarbete mellan de två företagen bör löna sig eftersom de i första hand riktar sig till samma kunder. Med ett samarbete kan de två företagen byta erfarenheter med varandra och kan även ha ett utbyte kunskapsmässigt om dels data mining, analyskunskaper etc. Parter som ansvarar för marknadsföringen har träffats vid ett tillfälle för att undersöka hur ett eventuellt samarbete

skulle se ut. Fördelen med ett samarbete är att de i många fall riktar sig till samma kundgrupper och kan med "enad front" lyckas att sälja in sitt budskap bättre. Vidare kan de titta på sina kunder för att se om man har samma svårigheter att nå en viss kundgrupp eller om de kan upptäcka dyra kunder. I många fall analyserar de samma kundgrupper vilket leder till dubbelarbete som kunde undvikas om de har ett närmare samarbete.

SAS Enterprise Miner ska erbjuda ett enkelt och användarvänligt anlaysverktyg för att ge vilken användare som helst möjlighet till att analysera den data som användaren är intresserade av. Vi har kommit fram till att det är svårt att tolka och förstå de resultat som produceras från data mining om man inte har någon form av statistisk eller analytisk bakgrund. Problemet kan lösas genom att skräddarsy en applikation som fungerar som en black-box som visar resultatet utan att blanda in komplicerade data mining algoritmer. Att använda sig av en black-box variant är inte särskilt utvecklande eller inspirerande för användaren eftersom de inte vet vad som händer utan endast får se ett resultat. Precis som med neurala nätverk kan man näst intill hoppa över att använda sig av en black-box om man inte förstår vad som händer. Användandet av black-box hämmar även kunskapsutvecklingen för de som använder den och inte förstår den, vilket inte heller är något att sträva efter i data mining sammanhang.

Då det gäller den metod vi valde för undersökningen och studien inför vår uppsats skulle vi vill säga följande. Det hade varit givande om vi hade haft möjlighet att genomför en mer omfattande studie under en längre tid så vi kunde dra nytta av alla de fördelar som finns i att använda sig av en etnografisk metod. Vi anser att vi har haft stor nytta av att använda oss av en etnografisk ansats vid intervjuerna och vid mötet, eftersom man med etnografisk metod inte styr och reglerar frågorna och svaren. Hade vi gjort det hade vi troligtvis inte kommit fram till samma resultat eftersom personerna vi intervjuade inte kunnat berättat fritt från deras uppfattning. Vi önskar också att vi hade haft möjlighet att göra fler intervjuer på Volvokort. Det hade också varit av intresse att intervjua andra försäkrings- och kreditkortsbolag för att undersöka hur andra ser på data mining och dess möjligheter.

Det finns inte enbart fördelar med en teknik som data mining. Fördelarna är förstås att vi anser att den här tekniken kan hjälpa företag att få en bättre överblick och analysering av sina datamängder. Data mining kan också hjälpa företag vid marknadsföring och vid kampanjer för att förhoppningsvis leda till förbättrande av resultat. Nackdelen med data mining är att det hanterar oftast superformaliserad data vid sina analyser, med formaliserad data menar vi att den innehåller siffror och symboler vid beskrivning av verkligheten istället för meningar som beskrivning av verkligheten. Ibland finns det ett behov att använda sig av fritext för att samla in data om omvärlden direkt istället för att formalisera all data. Det finns data mining tekniker som stödjer fritextsökningar men de är långt ifrån välutvecklade, eftersom det är enklare att genomföra analyser med formaliserad data. Nackdelen är att med formaliserad data missar man oftast intressanta aspekter av verkligheten. Vi tror att med en utveckling av tekniker i data mining som hanterar fritext kan nyttan med data mining öka avsevärt.

En intressant aspekt av data mining som vi inte tar upp i vår uppsats, som skulle vara intressant för fortsatta studier, är hur den personliga integriteten ska skyddas när alltfler företag använder sig av data mining utan kundernas vetskap. Det finns inte mycket av denna aspekt i litteraturen trots att aspekter kring datorsamhället och den personliga integriteten varit ett ofta debatterat ämne i litteraturen. I sin artikel "Data mining and Privacy: a conflict of making" diskuterar

Kurt Thearling hur utvecklingen av data mining kommer påverka kundernas möjlighet att skydda sin integritet. Ontarios informaton- och integritets kommisionär Ann Cavoukian som är författare till rapporten "Data mining: Staking a Claim on Your Privacy" menar att data mining:

"May be the most fundamental challange that privacy advocates will face in the next decade…"
Ann Cavoukian (Thearling 1998 b)

Vidare diskuterar hon att kunden måste sättas i centrum och få en möjlighet att veta vad som lagras och hur det kommer att användas. Det är ju inte säkert att all data mining kommer upplevas som negativ för kunderna eftersom företagen lär sig mer om vad kunderna efterfrågar och kan "skräddarsy" sin relation till kunden. Slutligen vill vi nämna att i nuläget använder företag data mining främst för att rikta sina kampanjer och det kan inte ses som något allvarligt hot mot den personliga integriteten.

6. Referenser

Andersen E., (1991), "Systemutveckling: Principer, Metoder och Tekniker", Studentlitteratur, sid 16-17

Askira-Gelman I., (1998), *"Knowledge Discovery: Comphensibility of the Results"*, IEEE

Berry M., Linoff G., (1997) *"Data Mining Techniques for Marketing, Sales and Customer Support"*, Wiley Computer Publishing 1997.

Blomberg J., Giacomi A., Mosher A., Swenton-Wall P., (1993), *"Etnographic field methods and their Relation to Design"*, Schuler, D. & Namioka A. (Eds.) Paticipatory Design: Perspective on system design, Lawrence Erlbaum, Hillsdale, NJ, 1993, sid 123-154

Brachman R., Khabaza T., Kloesgen W., Piatetsky-Shapiro G. och Simoudis E., (1996), "*Mining Business Databases"*, Communications of the ACM, Vol 39, No 11.

Dahlbom B., Ed., (1993), *"Essays of infology"*, Department of Information systems, Göteborgs Universitet, sid. 16-17

Denning, P.J., (1982), *"Electronic junk"*, Communications of the ACM, Vol 25, No 3

Fayyad U., Piatetsky-Shapiro G., Smyth P., (1996) *"The KDD Process for Extracting Useful Knowledge from Volumes of Data"* Communications of the ACM, Vol 39, No 11.

Frawley W., Piatetsky-Shapiro G., Matheus CJ., (1991),*"Knowledge Discovery in Databases: An overview"*, AAAI/MIT Press, Menlo Park, CA, sid 1-27

Fuori, W. M., Gioia, L. V., (1994), *"Computers and Information Systems"*, fourth edition, *Prentice Hall Internatinal Publishing*, sid 380 .

Hadden, E., (1997a), *"Modeling Techniques for Sucsessful Data Warehouses and Data Marts"*, Patricia Seybold Group Incorporated

Hadden, E., (1997b), *"Building Successful Data Warehouses and Data Marts - Using the Hadden-Kelly Data Warehouse Method"*, Hadden & Company

Inmon W. H., (1996a), *"Building the Data Warehose"*, second edition, *Wiley Computer Publishing*, sid 33-38.

Inmon W. H., (1996b), *"The Data Warehouse and Data Mining"*, Communications of the ACM, Vol 39, No 11.

Little J. D. C., (1970), *"Models and Managers: The concept of a Decision Calculus."*, Management Science, Vol 16, No 8.

Magoulas T. och Pessi K., (1998), "*Strategisk IT-management*", doktorsavhandling vid Institutionen för Informatik, Göteborgs universitet.

Morgan G., (1986), *"Images of Organization"*, Beverly Hills, SAGE Publications Ltd, sid 167-170

Patton, M.Q., (1990), *"Qualitative Evaluation and Research Methods"* New York: SAGE Publications.

SAS Institute (1996), *"Data mining with the SAS System From Data to Business Advantage"* SAS Institute White Paper for Data mining.

Schoderbeck P., Schoderbeck C., Kefalas A., (1990), *"Management Systems - Conceptual Considerations"*, fourth ed , Richard D. Irwin Inc, sid 96 och 152-156:

Shepard D., (1995), *"The New Direct Marketing - how to implement a profit-driven database marketing strategy"*, David Shepard Associaties Inc, Batra R., Deutch A., Orme G., Ratner B., Sharma D. (1995), Mcgraw-Hill Publishing 1995

Turban E. (1995) *"Decision Support Systems and Expert Systems"*, fourth edition, *Prentice Hall International Editions*, sid 10-11, 82, 443, 446, 480, 516-518, 550 och 683-684.

Uthurusamy R., (1995), *"From Data mining to Knowledge Discovery: Current Challange and Future Directions"*, AAAI/MIT Press, Menlo Park CA, sid 561-569

Länkar till Internetadressser

Url avser den internetadress där artikeln fanns på datum (Dat).

Brand E., Gerritsen R., (1998), *"Data mining and Knowledge Discovery"*, DBMS Online Data mining solutions Supplement
url: http://www.dbmsmag.com/9807m01.html
Dat: 1998-05-09

Brooks P. (1997) *"Data Mining Today"*, DBMS february.
Http://www.dbmsmag.com
Dat: 1998-02-09

Dilly R., (1997), *"Data mining Notes"*, University of Belfast.
url: http://www.pcc.qub.ac.uk, "Data mining"
Dat: 1998-02-09

Edelstein H., Millenson J., (1997), *"Lessons from the Trenches: Knowledge, Discovery and Data Mining"*, DBMS-Online
url: http://www.dbmsmag.com/9702d162.html
Dat: 1998-03-02

Edelstein H., (1997), *"Mining for Gold"*, Two Crows CMP Media Inc
url: http://www.twocrows.com/iwk9704.htm
Dat: 1998-04-27

Fayyad U., (1996), *"Data mining and Knowledge Discovery: Making Sense Out of Data"*, EEE Experts, Microsoft Research
url: http://www.computer.org/pubs/expert/1996/features/x5020/x5020.htm
Dat: 1998-01-26

Koo S., (1998), *"Interview with Knowledge Discovery Nuggets owner Piatetsky-Shapiro G"*, 11 mars 1998.
url: http://home.hkstar.com/~skoo/gps_eng.htm
dat: 1998-05-09

Pilot Software (1997), *"Data mining - White Paper"*
url: http://www.pilotsw.com/dmpaper/dmindex.htm
Dat: 1998-04-02

SAS Institute (1998), *"Getting Down to Business with Data Mining"*
url: http://www.sas.com/
dat: 1998-01-27

Simoudis E., (1996), *"Reality Check for Data mining"*, EEE Experts, IBM Almaden Research Center
url: http://www.computer.org/pubs/experts/1996/features/x5026/x5026.htm
Dat: 1998-01-26

Small R., (1997), *"Debunking Data mining Myths"*, CMP Media Inc Tech Web
url: http://www.techweb.cmp.com/iw/614/14oldat.htm
Dat: 1998-04-27

Thearling K., (1998 a), *"Increasing customer value by intergrating Data mining and Campaign Management Software - An Exchange Applications White Paper"*, Boston
url: hht://www.santafe.edu/~kurt/text/integration.shtm
Dat: 1998-05-09

Thearling K., (1998 b), *"Data mining and Privacy: A conflict in the making"*, Exchange Applications Boston Ma
url: http://www.santafe.edu/~kurt/text/dsstar/privacy.shtml
dat: 1998-05-09

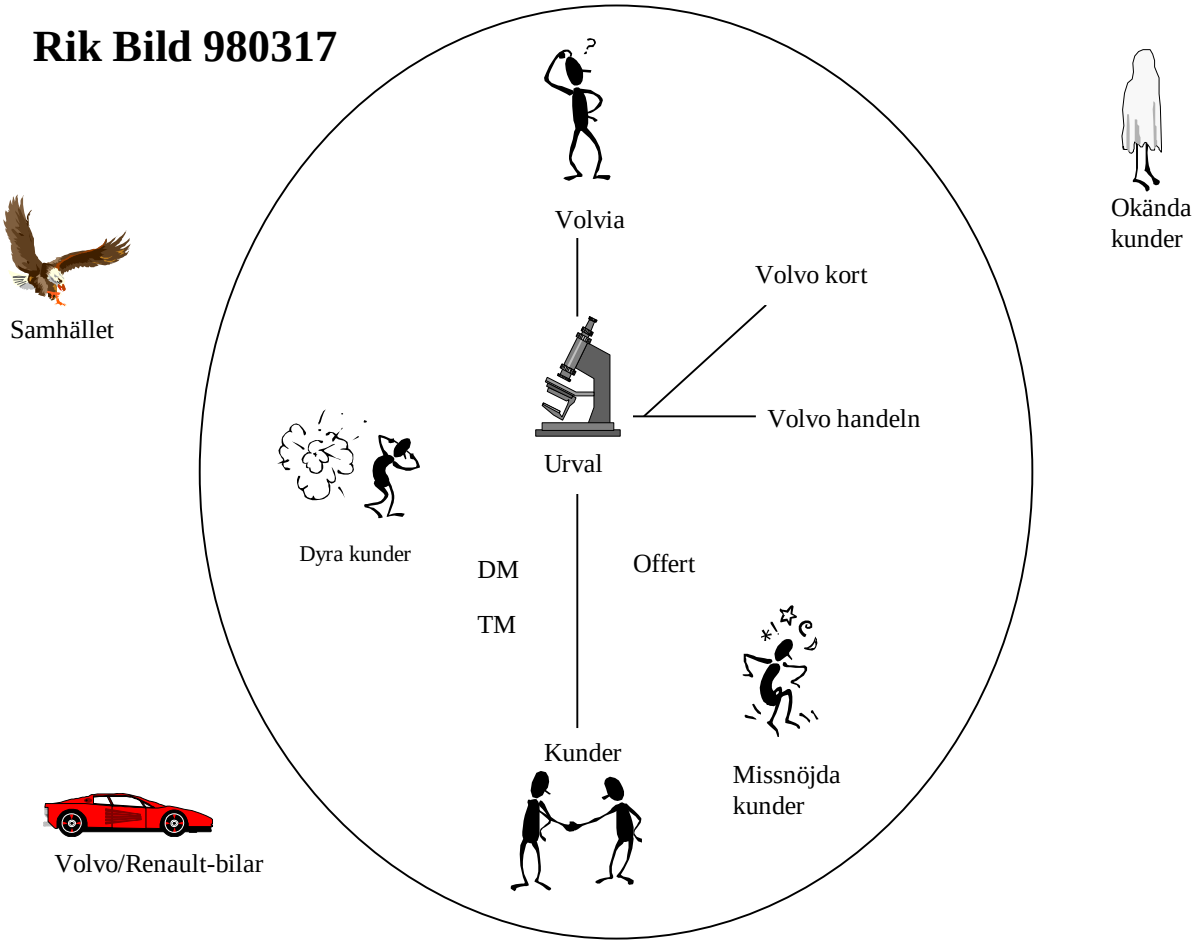
Magisteruppsatser 20p

Björnsson M., (1997), *"En jämförelse av data mining algoritmer för klassifikation"*, Chalmers tekniska Högskola

Ericson D., Ericsson R., (1995),*"Trettio års problem med datoriserat beslutstöd - Kan Data warehouse vara lösningen"*, Umeå Universitet

Landgren J., (1997) *"Data warehouse and Data mining"* Institutionen för Informatik, Göteborgs Universitet.

Bilaga 1


Rik Bild 980317
Volvia
Okända kunder
Samhället
Volvo kort
Volvo handeln
Urval
Dyra kunder
DM
TM
Offert
Kunder
Missnöjda kunder
Volvo/Renault-bilar

Bilaga 2 Ordlista

Det här avsnittet utgör en form av uppslagsverk för att på ett förenklat sätt förklara de begrepp vi använder oss av i uppsatsen.

Begrepp	**Förklaring**
OLTP	On-Line Transaction Processing. När företag samlar på sig stora mängder data och ska lagra den kan OLTP på ett enkelt och snabbt sätt överföra data till en relationsdatabas.
Point-of-sales	Innebär att företag sparar all data om till exempel en kunds inköp i en mataffär.
information overload	Det finns en risk att företag råkar ut för information overload när de lagrar en stor mängd data. Det kan vara svårt att få ut något meningsfullt från den och all information går förlorad.
beslutstödssystem	Beslutstödssystem låter en användare interagera med en dator genom att ställa frågor och snabbt få svar. Syftet med systemet är att det ska stödja användarna vid olika beslutsituationer.
data warehouse	Data warehouse är ett beslutstödssystem som lagrar data och är utvecklat för företag för att tillgodose ledningens informationsbehov..
EIS	Executive Information System, ett beslutstöd som är utvecklat för att stödja toppledningens beslut.
data mining	En analys teknik som används för att hitta mönster och samband i data som finns lagrad i ett data warehouse.
KDD-processen	Knowledge Discovery in Databases omvandlar data till kunskap genom stegen selektion, förberedning, transformation, data mining och evaluering.
artificiell intelligens	Artificiell intelligens, AI, försöker få datorer och maskiner att efterlikna biologiska neuroner i strukturen, för att få maskinerna att agera på ett intelligent sätt.
maskinlärande	Härstammar från AI och vill lära maskiner att lösa problem genom att använda sig av historiska problem och använda resultaten av dem.
lärande vid induktion	Är en del av maskinlärande. Med olika metoder försöker man erhålla kunskaper från experter, böcker och böcker och överför kunskapen till en kunskapsbas. Detta sker för att ha tillgång till experters kunskap vid problemlösning.
case-based reasoning	Härstammar som lärande vid induktion från

forts case-based reasoning	maskinkärande. Case-based reasoning använder sig av historiska problemlösningar och manipulerar dem för att anpassa lösningen till det aktuella problemet.
Verification och hypothesis	I den här metoden försöker man bevisa eller motbevisa förutfattade idé eller hypotes.
discovery och knowledgde discovery	Den här metoden går ut på att få data att säga något som vi inte visste sedan tidigare, det vill säga att man får nya mönster och samband.
indirekt knowledge discovery	Används i data mining för att upptäcka relationer och samband mellan data.
direkt knowledge discovery	Används i data mining för att förklara de relationer och samband som hittats i den indirekta metoden.
SEMMA	Är SAS Institutes metod för data mining. SEMMA står för Sample, Explore, Manipulation, Modeling och Assess.
OLAP	On-Line Analytic Processing ett analysverktyg som används för att visualisera data.