Lunds Universitet
Informatik

VT 2009

Datakvalitetsarbete vid kunskapsutvinning ur databaser

Kandidatuppsats, 15 högskolepoäng, INFK01 i informatik

Framlagd: 06-2009

Författare: Alexander Lindén
Filip Lagerlöv

Handledare: Anders Svensson

Examinator: Agneta Olerup, Erik Wallin

Abstrakt

Titel	Datakvalitetsarbete vid kunskapsutvinning ur databaser
Författare	Alexander Lindén
	Filip Lagerlöv
Utgivare	Institutionen för Informatik
Handledare	Anders Svensson
Examinator	Agneta Olerup, Erik Wallin
Publiceringsår	2009
Uppsatstyp	Kandidatuppsats
Språk	Svenska
Nyckelord	datakvalitet, datadimension, KDD, DM, mönster

Statistik visar att data ökar exponentiellt i omfattning varje år. Det finns ett stort värde i data, speciellt för organisationer som med rätt tekniker kan använda sig av data för att kartlägga kunder eller sina egna affärsangelägenheter. För att kunna dra nytta av data i form av kunskap behöver data först sammanställas och analyseras. Ett viktigt mått vid analysering av datamängder är kvaliteten. Framförallt eftersom både djupet av data i form av kvantiteten men också bredden i form av olika typer av data ständigt växer. Vi har därför använt oss av en process som analyserar data och vars slutgiltiga mål är att skapa kunskap av data. Med hjälp av denna process har vi undersökt hur datakvalitetsarbetet bör ske parallellt med processen för att förhoppningsvis förbättra resultatet.

Målet med uppsatsen har varit att först och främst undersöka hur datakvalitet kan påverka en dataanalys och den produkt som bildas. För att genomföra denna uppgift behövde vi även undersöka vilka faktorer som är speciellt intressanta för att bedöma datakvalitet.

Undersökningen består även av åtkommande relationer mellan data, information och kunskap. Vi har därför behandlar hur dessa förhållanden ter sig.

Med hjälp av tre olika företag som är verksamma inom områdena Artificial- och Business Intelligence har vi kunnat undersöka hur tekniken för kunskapsutvinning ur data fungerar. Vi har även fått ta del av hur datakvalitet påverkar tekniken och hur arbetet med att förbättra datakvalitet går till. De underlag som vi fått förmedlat från företagen har tillsammans med ett teoretiskt ramverk och våra egna uppfattningar bildat resultatet av undersökningen.

Förutom att fastställa att datakvalitet är ett viktigt och problemartat område inom analyser av stora datamängder har vi även föreslagit ett antal faktorer som är viktiga vid bedömning av datakvalitet. Våra iakttagelser har även resulterat i en modell beskrivande hur datakvalitet bör förebyggas löpande under analyseringen för att skapa ett informativt och säkert resultat.

Innehåll

Figurförteckning

Tabellförteckning

1 Introduktion

Kapitlet avser att ge en bakgrund till vårt valda uppsatsämne för att få en inblick i vad vi vill undersöka och vilka tidigare undersökningar som genomförts på området. Därefter följer vår problemformulering och våra frågeställningar för att mer ingående förstå vilka frågor vi tänker besvara under arbetet samt hur de hänger ihop med uppsatsens syfte. Avslutningsvis behandlas de avgränsningar i uppsatsen som vi inte kommer att undersöka.

1.1 Bakgrund

Det råder inte någon brist på data i världen, utan snarare finns det för mycket och för olik data tillgänglig för att de ska kunna behandlas på användbara sätt (Chen et al, 1996; Cooley et al, 1997; Kosala & Blockeel, 2000). Frawley et al (1992) uppskattar att informationen i världen fördubblas var 20:e månad. Mer aktuell statistik, mellan år 2003 och 2005, visar på en ökning av världens största kommersiella databas med 300 %. Den största mäter över 100 terabyte (TB) i storlek (WinterCorp TopTen Program, 2005). Med så pass stora datamängder behövs det kraftfulla verktyg för att kunna hantera dessa data.

Till stor del sker statistiska analyser av datamängder via olika typer av Business Intelligence (BI) system. BI kan beskrivas som strävan efter att använda data för att skapa insikt om nutida, framtida och historiska mönster (Michalewicz et al, 2006; Negash & Gray, 2004).

Kvantiteten av de datamängder som samlas in till förmån för analys via BI-system ökar drastiskt och organisationer förlorar samtidigt stora summor pengar på att använda data med dålig kvalitet (Vassiliadis, 2000). Detta vill vi försöka motverka och visa på hur datakvaliteten kan påverka organisationers kunskapsutvinning via BI. Samtidigt som datamängderna ökar ställs det större krav på tekniken, men frågan om kvalitet kontra kvantitet tenderar att genomsyra måttet på den slutprodukt som bildas.

Tekniken som behandlar statistisk analys av stora datamängder kallas för Data Mining (DM). DM går fram för allt ut på att skapa användbara mönster av information för att undersöka prediktiva företeelser (Fayyad et al, 1996). Det kan även förenklat beskrivas med den process som skapar kunskap av information, information som tidigare förädlats ur data. Data måste alltså först bilda information innan analyser av informationen kan bilda korrelationsmönster via DM. Genom de mönster som bildas och de upptäckter de formar, är tanken att kunskap ska ses som den slutliga produkten.

Det övergripande syftet med DM är att upptäcka information som kan klassas som beslutsgrundande inom olika områden. En viktig del är dock om datamängderna som påträffas verkligen kan kvalitetssäkras. Det är framförallt den stora komplexiteten i datamängder, i form av dess storlek och härkomst som gör det komplicerat att kvalitetssäkra dessa. (Frawley et al, 1992)

Det finns inte heller ett enskilt verktyg för att ta tillvara på datamängder, utan flera olika tekniker behöver användas beroende på typ och härkomst (Chen et al, 1996). Tekniken finns tillgänglig och har funnits sedan en lång tid tillbaka men det finns ändå vissa frågetecken rörande kvaliteten i datamängderna som samlas in till förmån för DM.

Kvaliteten av data har t.ex. betydelse när olika datamängder tillsammans bildar information som senare används av organisationer inom t.ex. Business Intelligence (BI) för att skapa mönster som kan relateras till kunskap inom kundbeteenden eller att kartlägga olika aktiviteter i organisationen. Den kunskap som genereras måste även vara relevant för den kontext de används inom. De måste även kunna säkerställas så att de upptäckter som påträffats är korrekta med verkligheten. (Frawley et al, 1992)

Tidigare forskning som sträcker sig från början av 1980-talet har konstaterat att det råder problem gällande datakvaliteten i databaser (Frawley et al, 1992; Strong et al, 1997). Det finns dock hjälpmedel för att säkerställa att värdefull och med hög kvalitet information kan skapas ur de datamängder som återfinns i databaser, alltså data mining (Frawley et al, 1992). Men det råder ändå delade meningar om all information verkligen kan säkerställas. Tidigare forskning har visat att datakvalitetsproblemen är av en flerdimensionell härkomst, där flera olika infallsvinklar kan användas(Strong et al, 1997; Wand & Wang, 1996).

Andra forskare tar sig istället an frågan ifall det går att sätta en kvalitetsmärkning på data. van Well & Royakkers (2004) anser att den information som påträffas kräver att källan verifierar att en specifik upptäckt verkligen är sann. Pazzani (2000) hävdar att kunskap om det mänskliga beteendet borde integreras i bearbetningsprocessen av datamängder eftersom kunskap är nära relaterat till mänskligt beslutsfattande och på så sätt bör kvaliteten på informationen förbättras. Det är även viktigt att lära sig mer om vad användare vill, vad de gör och deras faktiska kunskap inom olika områden (Kosala & Blockeel, 2000). Detta är ett antal av de olika infallsvinklar som finns rörande dagens datakvalitetsproblem och kan således även användas gällande DM.

Den tidigare forskningen tenderar att söka lösningar i symbiosen mellan systemet och dess omgivning, medan den senare forskningen inriktat sig allt mer vid mänskligt beslutsfattande och hur de kan påverkar data. Tidigare har forskare även fokuserat på var problemen uppstår i insamlingsprocessen av data, varför de uppstår samt delgett lösningsförslag på hur tekniken för DM bör förändras. Utifrån det givna området och förslag i tidigare undersökningar vill vi istället undersöka hur kvaliteten av de datamängder som utvinns ur databaser påverkar ett skapande av korrelationsmönster i data.

Vi anser att detta är ännu ett perspektiv till problemet rörande datakvalitet inom Informations System (IS) men vill samtidigt betona att vi är neutrala i frågan gällande hur mönster påverkas av datakvalitet. Med det menar vi att våra påträffanden kan vara av både negativ och positiv art relaterat till datakvalitet.

Området anses intressant och relevant eftersom stora datamängder finns tillgängliga via dagens IS och vi är intresserade att undersöka i vilken utsträckning dessa datamängder kan användas. Med användas menar vi i den mån data kan vara grunden för beslutsfattande information. För att detta ska kunna ske, måste data och information kunna bilda informativa och säkra mönster via DM. Detta är skälet till att vi har intresserat oss för detta problem. I detta perspektiv ingår även hur precisa och tillförlitliga data är, vilket leder oss in på hur detta bestäms och utifrån vilka preferenser data kan styrkas.

1.2 Problemområde

All statistik visar att mängden data ökar exponentiellt i omfattning varje år, och de största databaserna kommer att nå petabyte inom ett par år om utvecklingen fortsätter i samma takt (WinterCorp TopTen Program, 2005).

Men hur kan data vara värdefull, då data i sig själv är oanvändbar och simpel. Det är först när olika data sammanställs och bildar information som den kan användas (Daft, 2006; Zeleney, 1987). Sedan är det upp till människan att själv skapa kunskap om den information han eller hon erhåller. Data kan därför lite grovt ses som en råvara till information och precis som de flesta andra råvaror behöver de vara av en god kvalitet för att slutprodukten ska bli bra.

Kvalitetsproblem i data är dock en kraftigt ökande trend inom olika IS (Strong et al, 1997). Problemen ökar drastiskt i framför allt databaser, men även Internet ses som en bidragande faktor. Organisationer förlorar samtidigt stora summor pengar på att använda data där kvaliteten är bristande (Vassiliadis, 2000). Men är det bara problem som uppstår eller finns det även fördelar med detta som kan lyftas fram?

Till vår hjälp för att undersöka detta kommer vi använda oss av en process vars syfte är att skapa kunskap av data. Denna process kallas Knowledge Discovery in Databases (KDD) och kommer att beskrivas utförligt i kapitel 2. Överskådligt kan processen kallas för en slags dataanalys. Processen kommer även att fungera som ett analysverktyg där resultaten från våra undersökningar ska evalueras utifrån processen. KDD kommer även att utvärderas för att beskriva hur datakvalitetsarbetet bör bedrivas genom processen.

Våra resultat kommer även mynna ut i hur bedömning av datakvalitet bör genomföras samt vilka åtgärder som kan förbättra resultatet av dataanalyser.

1.3 Frågeställningar

De frågeställningar vi har arbetat med är följande:

- Hur påverkar datakvaliteten i databaser skapandet av datamönster?
- Vilka faktorer är relevanta vid bedömning av datakvalitet?
- Vilka förebyggande åtgärder krävs för att förbättra resultatet av datamönster?

1.4 Syfte

Syftet med uppsatsen är att öka förståelsen för datakvalitetsarbetet vid kunskapsutvinning ur databaser. Vi vill således beskriva hur kvaliteten av de data som utvinns ur databaser påverkar mönsterskapande och statistiska analyser. Med hjälp av dessa påträffanden vill vi i sin tur belysa de faktorer som ses som speciellt intressanta, relevanta och användbara vid bedömning av datakvalitet.

Till sist vill vi visa på vilka förebyggande åtgärder som krävs för att förbättra resultatet av de datamönster som utvinns via DM. Detta med hjälp av en reviderad modell av KDD innehållande de aspekter vi anser är speciellt talande för att uppnå ett mer kvalitetsinriktat resultat av processen.

1.5 Avgränsningar

Vi avser inte att ställa olika tekniska eller funktionella aspekter av datainsamling mot varandra för att beskriva olika fördelar, nackdelar osv. eftersom det är tekniken i sin helhet och dess kvalitetsrelaterade frågor vi anser är relevanta för vår uppsats.
Vi kommer inte att undersöka kunskapen som kan komma att genereras via DM. Detta eftersom det är mönstren som bildas som är det väsentliga i vår undersökning, inte kunskapen som i sin tur skapas. Vi kommer inte heller att se till etiska aspekter kring datainsamling eftersom lagstiftare är ansvariga för området.
Vi kommer inte att gå igenom datakvalitet mer ingående utan endast redogöra de definitioner som finns. Detta eftersom ett allt för omfattande inslag av datakvalitet skulle resultera i en helt egen uppsats och detta är inte vår mening.
Vi avser inte heller att behandla datakvalitetsrelaterade åtgärder överlag, utan vi är intresserade av dessa som kan användas parallellt med KDD-processen då den ska fungera som ett evalueringsverktyg.

1.6 Litteraturgranskning

För att uppnå vårt syfte med undersökningen har vi tagit del av teoretiska tidsskrifter beskrivande vårt problemområde. Läsningen av dessa tidsskrifter har också skapat vidare kunskap om området och hjälpt oss att specificera våra frågeställningar.

En viktig del i detta arbete var att genomföra en kritisk granskning av den litteratur vi påträffat. För att undersöka litteratur och teorier av relevans för vår studie har vi använt oss av Google Scholar som är en artikeldatabas på Internet, ELIN som är Lunds Universitets artikel- och journaldatabas samt LOVISA som är en databas över tidsskrifter som finns tillgängliga via Lunds Universitet olika bibliotek.

Den litteratur vi framförallt letat efter är den om KDD, DM och datakvalitet. Vi har även sökt efter litteratur för att fastställa skillnader mellan data, information, kunskap och vishet.

Detta för att relationerna mellan dessa element behandlas till stor del i vår undersökning och vi ansåg att vi behövde fastställa hur vi ställde oss till dessa.

Med hjälp av vårt teoretiska underlag har vi skapat oss en bred förståelse inom vårt problemområde samtidigt som vi har kunnat utforma vår undersökningsmall på ett mer kvalitetsinriktat sätt. Vårt teoretiska underlag används även för att analysera de svar som vår undersökning resulterat i. På så sätt har vi kunnat ställa teorin mot undersökningen och kommit fram till slutsatser som besvarar våra frågeställningar.

2 Teoretisk referensram

Kapitlet innehåller en kritisk granskning av den litteratur vi använt oss av för att undersöka vårt problemområde. De områden som behandlas är omvandling av data till vishet för att beskriva relationerna mellan de olika elementen. Datakvalitet behandlas även och vilka olika dimensioner det finns av datakvalitet. Slutligen beskrivs även KDD-processen utifrån dess definition, hur den används, vilka problem som kan uppstå i den och hur framtiden ser ut.

2.1 Från data till vishet

Genom vår uppsats kommer vi att behandla relationerna mellan data, information och kunskap. Detta beroende av att mönsterskapande via DM behandlar alla dessa element. Vi anser därför att vi behöver klargöra hur dessa relationer mellan elementen data, information och vishet ter sig.

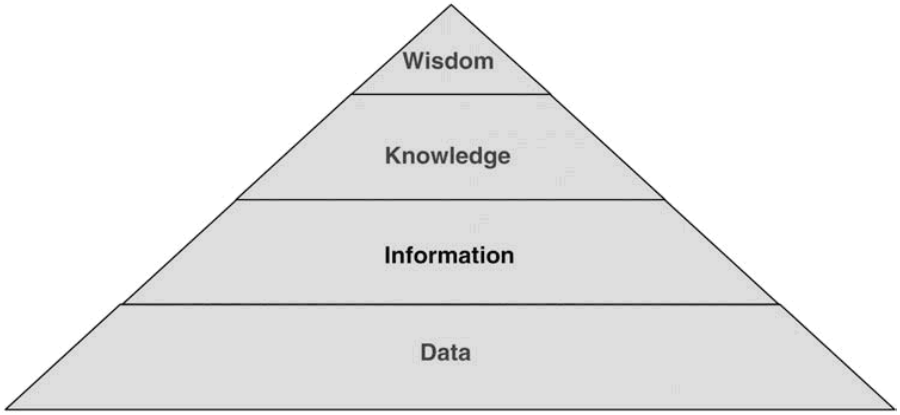
Både distinktionerna och samspelet mellan data, information, kunskap och vishet (DIKV) skiljer sig åt genom historien. Viktigt är dock att beskriva några av de mest erkända fastställanden som gjorts mellan dessa förhållanden. Vi kommer här även beskriva vishet som ett element eftersom de har en stor betydelse i forskningen. Men det är också för att visa på varför vishet börjar glida isär från dagens forskning.

"While data and information are piecemeal, partial and atomized by their very nature, knowledge and wisdom are 'holistic', related to and expressed through systemic network patterns, integrative by definition" (Zeleny 1987, p. 59)

Zelenys definition av relationerna mellan elementen i DIKV är en av de äldsta. Data och information ses som enkla uppsättningar som liknar atomer, och som i sig själva inte är användbara, atomism. Kunskap och vishet beskriver i sin tur helheten av delarna relaterat till data och information, holism. Liknelsen kan göras där data och information liknas med atomer, medan kunskap och vishet liknas vid molekyler, där helheten är större än summan av delarna. Människan kan ha tillgång till data eller information, men kan inte ha tillgång till kunskap. Delarna av data och information bildar en helhet och det är då kunskapen uppnås. vishet i sin tur, är när kunskap kan sammankopplas, beskrivas eller förstås.

Trots många olika definitioner och skiljaktigheter inom forskningen kring DIKV är de flesta forskare överens om hur arvshierarkin i förhållandet mellan DIKV ser ut *(figur 2.1)*. Det debatteras dock fortfarande om vem som egentligen kläckte de ursprungliga idéerna bakom dessa förhållanden. Två av de första forskarna som började använda sig av grafiska modeller över förhållandena var Zeleny (1987) och Ackoff (1989).

Zeleny (1987) förespråkar en fyrfaldig hierarki med ett femte steg inom parantes. Hierarkin innehåller elementen data, information, kunskap, vishet och upplysning som det femte steget. Med upplysning menas de värde som går att sätta på visheten, antingen sant eller falskt. Ackoff (1989) är av samma uppfattning som Zeleny, men presenterar även förståelse som ett inslag i modellen, i det avseende att förståelse av kunskapen måste uppnås innan vishet kan uppnås.



Figur 2.1 Hierarkin – Data, Information, Knowledge, Wisdom (Rowley, 2007)

Den ursprungliga modellen är enkelt utformad och lämnar stora möjligheter till fri tolkning. Det generella antagandet är att data används för att skapa information, som i sin tur används för att skapa kunskap och som i sin tur används för att skapa vishet (Rowley, 2007).

Elementens ordningsföljd i modellen råder det alltså inga tvivel om, men vad varje enskilt element betyder för modellen och hur varje element bestämmer nästa råder det delade meningar om. Här är det information och kunskap de två element som diskussionen främst handlat om (Rowley, 2007). Vi kommer att beskriva varje enskilt element och vad de har för betydelse utifrån en modell Rowley skapat, där hon jämfört Zelenys och Ackoffs definitioner av de olika elementen *(figur 2.2)*.

Tabell 2.1: Jämförelse av Zeleneys och Ackoffs definitioner av DIKV (Rowley, 2007)

	Zeleny	**Ackoff**
Data	Know nothing	Symbols
Information	Know what	Data that are processed to be useful; provides answers to who, what, where, and when questions
Knowledge	Know how	Application of data and information; answers how questions
Understanding		Appreciation of why
Wisdom	Know why	Evaluated understanding
Enlightenment	Attaining the sense of truth, the sense of right and wrong, and having it socially accepted, respected and sanctioned	

Zelenys (1987) diskussion utgår ifrån hur och varför människors administrativa arbete bör inrikta sig på att skapa kunskapstillgångar i form av integration. Här skall en sammanställning

av data och information bilda en helhet i form av kunskap och kunskapen i sin tur ska leda till att vishet skapas. Vishet i det hänseende att människan förstår varför kunskapen är viktig. Han menar alltså att kunskap inte är användbar i sig själv, utan människor måste även veta hur de ska använda kunskapen. Ackoff (1989) hävdar att informationen måste bli användbar innan den kan bilda kunskap men menar även att ett eget element, förståelse, behövs innan kunskap kan skapa vishet. Bellinger et al (2004) har utvärderat Ackoffs definition av DIKV och menar att förståelse borde vara en förutsättning för att ta nästa steg i hierarkin, istället för ett eget element. De menar att förståelse om relationer i data måste uppnås innan information kan bildas etc.

Förståelse är mer underförstått i Zelenys teori eftersom han beskriver vishet som förståelsen om varför kunskapen är viktig. Kunskap kan i sin tur erhållas på två olika sätt, antingen förmedlas den via en annan person eller så skapas den via nya erfarenheter (Ackoff, 1989).

Att skapa kunskap ur information handlar i sin tur om relationer och tolkningar av information, endast vetskapen om information i sig är inte relevant eftersom den är obestämd utifrån ett subjektivt perspektiv (Zeleny, 1987). Information bildas i sin tur ur data, vilket Zeleny beskriver som smådelar eller atomer. Data i sig själv är oanvändbar och enkel, det är först när olika data sammanställs som den bildar information (Daft, 2006; Zeleney, 1987). Eller som Ackoff (1989) beskriver förhållandet mellan data och information, information är data som har tilldelats en mening och kan därför bli användbar. Ett exempel på detta är hur en databas skapar information utifrån de data som finns lagrad i den. Fayyad (1996) anser att data som fångas från vår omgivning är de grundläggande bevis som används för att skapa teorier och modeller över det universum vi befinner oss i.

På senare år och speciellt relaterat till publikationer som behandlar Informations System (IS) och Knowledge Managegment (KM) så har vishet i allt större utsträckning exkluderats från hierarkin. Inte på ett sätt som tar avstånd till att vishet borde vara ett element i DIKV, men det definieras inte längre i lika stor utsträckning. Rowley (2007) har genomfört en studie på detta där hon sammanställt innehållet i 15 stycken böcker inom områdena IS och KM. Endast 3 av böckerna beskriver hur vishet är en viktig del av DIKV för IS och KM, där merparten av böckerna alltså tar avstånd till vishet. Och detta trots att det är det högsta elementet i hierarkin. Vidare menar hon att om IS och KM ska stå som en lämplig grund för människors och organisationers handlingar behöver forskare engagera sig i debatten huruvida vishet är en viktig beståndsdel inom dessa områden.

Trenden rörande vishetens utelämnande i litteraturen om IS och KM visar just på att kunskap är den högsta graden av beslutsunderlag som organisationer riktar in sig på (Michalewicz et al, 2007). Detta har bland annat att göra med organisationers BI, där just kunskap om nutida, framtida och historiska mönster i organisationers affärsangelägenheter är det centrala (Michalewicz et al, 2006; Negash & Gray, 2004). Inom BI är olika IS och KM viktiga inslag, men även DM och annan form av informationsinsamling (Negash & Gray, 2004). Detta gör att kunskap är det som prioriteras av systemen och beslutsfattandet lämnas åt människan. Michalewicz et al (2006) argumenterar bland annat för att system i högre grad borde vara kapabla att ta egna beslut utifrån den kunskap som skapas. Dock är detta något som fortfarande inte ses i speciellt stor utsträckning inom dagens IS, utan är något som lämnas till det mänskliga arbetet. För att dra en parallell till tidigare forskning inom området, så var Ackoff (1989) inne på ett liknande spår när han argumenterade för expertsystemens

kapacitet. Han menar att expertsystem har kunskapen hos en expert programmerad inuti dem och är sällan lärande system, alltså inga intelligenta system. Drivkraften kring att göra allt bättre och intelligentare system är fortfarande stark i branschen. Speciellt Michalewicz et al (2006) beskriver systemens möjligheter att ta beslut och anpassa sig efter situationer som ledande förutsättningar för dagens IS.

2.2 Data

I detta avsnitt kommer vi mer tydligt redogöra för vilka olika kategorier eller typer av data som finns. Vi kommer även att definiera hur olika forskare ser på datakvalitet eftersom detta är ett återkommande inslag genom vår uppsats.

2.2.1 Kategorier av data

Batini & Scannapieca (2006) anser att det förekommer tre olika typer av data, nämligen: strukturerade, semistrukturerade och ostrukturerade data. Strukturerade data är data som har en fixerad struktur och där relationstabeller anses vara den mest poplära form av strukturerad data. Semistrukturerade data anses vara data med en flexibel struktur, och ett exempel på ett verktyg för förvaring av semistrukturerade data är eXtensible Markup Language (XML). Ett tydligt karaktärsdrag av semistrukturerade data är att strukturen oftast inte är fullständig. Det kan därför förekomma olika typer av modifiering av attributen i data för att reparera strukturen. Ostrukturerade data är data som saknar en struktur eller ett domänområde. Exempel av ostrukturerade data kan vara data som är representerad i skriftligt språk. Enligt Batini & Scannapieca (2006) behöver olika tekniker anpassas för att förbättra datakvaliteten, beroende på vilka data som används. Kategoriseringen eller typ av data har därför en viktig betydelse i kvalitetsarbetet. Men vad är då hög datakvalitet?

2.2.2 Datakvalitet

Data som har en hög kvalitet, är data som är redo för användning av en datakonsument (Herzog et al, 2007; Strong et al, 1997; Wang Y, 2000). Herzog et al (2007) anser att det går att definiera datakvalitet efter hur pass bra data efterföljer den standarden som organisationen har för att data ska kunna användas.
Det finns ett antal dimensioner som används för att bedöma datakvalitet. Skulle det uppstå något problem i någon av kvalitetsdimensionerna har troligen ett datakvalitetsproblem påträffats (Strong et al, 1997). Strong et al (1997) presenterar även en tabell med fyra olika datakvalitetskategorier som kategoriserar de olika datakvalitetsdimensionerna. Denna tabell presenteras nedan (*tabell 2.1*).

Tabell 2.1 Kategorier och dimensioner för Datakvalitet (Strong et al, 1997)

Datakvalitet kategori	**Datakvalitet Dimensioner**
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise, Representation, Consistent of representation

Herzog et al (2007) anser istället att det förekommer sju erkända attribut för att bedöma datakvalitet. Dessa är: relevance, accuracy, timeliness, accessibility, clarity of results, comparability, coherence, completeness. Batini & Scannapieca (2006) har samma syn som Strong et al (1997) och menar att olika dimensioner förekommer beroende på vilket domänområde som är aktuellt.

Batini & Scannapieca (2006) hävdar att en första utgångspunkt i en datakvalitetsrelaterad aktivitet är att urskilja dimensioner i data för att kunna mäta kvaliteten. De utvalda dimensionerna används sedan i modeller, tekniker, verktyg eller ramverk. Modellerna används i databaser för att beskriva datascheman. Dimensionerna måste integreras in i modellen för att vara användbara och kunna bedöma datakvaliteten. Först då kan problemen bli uppmärksammade. Med tekniker menar Batini & Scannapieca (2006) algoritmer eller kunskapsbaserade procedurer som ska användas till att lösa ett specificerat datakvalitetsproblem.

Herzog et al (2007) har ett annat tillvägagångssätt och presenterar tre olika alternativ för att uppnå hög datakvalitet.

- *Förhindra*: Förhindra att dålig data kommer in i systemet genom att använda ett system som kontrollerar de data som tillkommer i databasen.

- *Upptäcka*: Aktivt leta efter data med dåligkvalitet. Med hjälp av Metoder och tekniker låta en dataanalys aktivt arbeta med att hitta fel i data.

- *Reparera*: Reparera data då data används.

Författarna anser även att tillvägagångssättet *förhindra* är att föredrar då det är mer kostnadseffektivt. Batini & Scannapieca (2006) och deras förslag för att uppnå bättre datakvalitet överensstämmer med Herzogs et al (2007) andra alternativ, *upptäckt*, där data aktivt genomsöks för att finna bristande data. Batini & Scannapieca (2006) är dock något djupare och mer förklarande i sina förslag.

2.3 Kunskapsutvinning i databaser

Genom åren har benämningarna för att skapa användbara mönster av data varierat. DM, kunskapsutvinning, informationsupptäckt, informationsinsamling, data arkeologi och bearbetning av datamönster är ett antal av dessa benämningar. DM är den vanligaste benämningen och har sitt ursprung från statistiker, data analytiker och arbete inom Management Information Systems (MIS). Det används även i stor utsträckning som ett verktyg för databaser men är kanske minst lika viktigt för att generera information från webbsidor på Internet. (Fayyad et al, 1996)

År 1989 myntades termen Knowledge Discovery in Databases (KDD) på en workshop för att lyfta fram kunskap som den slutgiltiga produkten av datadrivna upptäckter (Piatetsky-Shapiro, 1991). På så sätt bildades även ett nytt samlingsnamn och en ny strategi för att framställa kunskap från data.

> *"Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."*(Frawley et al 1992, p. 58)

KDD är en framställning av information från data som:

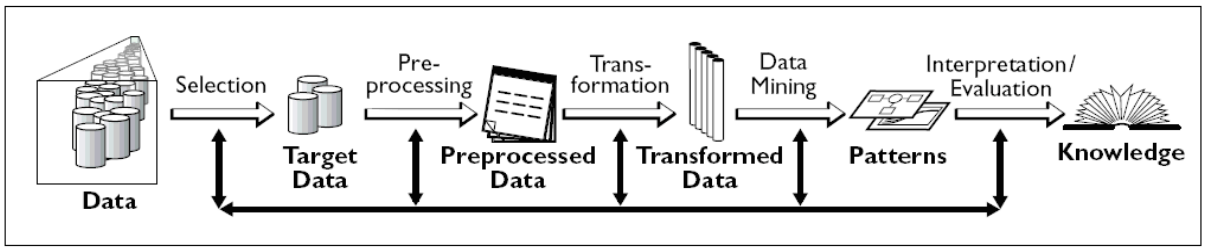
- inte är alldaglig
- är gömd
- är inte tidigare känd
- kan vara användbar eller relevant inom en given kontext

För att belysa ut definitionen ytterligare kan alldaglig eller trivial data bestå av data som är för vanlig eller i generella antaganden, irrelevant. Dessa data är inte relevanta för KDD utan data måste uppfylla ett visst värde eller syfte för att vara relevant. Detta kan även beskrivas som att informationsinsamling är endast en del av KDD om det finns ett syfte bakom informationsinsamlingen. Allmänna data eller fakta grundar sig inte i KDD utan det måste finnas en strävan efter att länka samman data för att på så sätt bilda information. KDD-processen strävar inte efter att samla in data på ytan, utan verkar snarare för att söka i stora data mängder efter olika typer av data och bilda mönster ur dessa. Processen strävar också efter att skapa ny information ifrån de upptäckter som görs, vilket ska kunna ses som användbar kunskap inom olika sammanhang. Information som tidigare redan är känd har troligen redan används i processen och har antingen behållits eller raderats. Slutligen måste information som framställts från de data som påträffats vara relevant inom den kontext de används. (Frawley et al, 1992)

KDD består av den totala processen för att upptäcka användbar kunskapa i stora datamängder, och en av de viktigaste delprocesserna är DM (Fayyad et al, 1996; Piatetsky-Shapiro, 1991; Frawley et al, 1992; Goebel & Gruenwald, 1999; Maimon & Rokach, 2005). DM är benämningen för att skapa användbara mönster av information. Detta är i sin tur ett steg mot slutprodukten av hela processen, nämligen att utvinna kunskap av den information

som skapats av påträffad data (Fayyad et al, 1996). Maimon & Rokach (2005) hävdar att det är viktigt att utveckla en modell beskrivande hur och vilken information som ska samlas in för att effektivisera arbetet. Chen et al (1996) anser att algoritmerna som används i kunskapsutvinningen bör vara effektiva och skalbara gentemot stora databaser. Med detta menas att löptiden för DM- algoritmerna måste kunna förutses och fungera gentemot stora databaser. Goebel & Gruenwald (1999) tydliggör även vikten av att skapa ett mål med processen där det behöver vara klargjort vad för kunskap som eftersöks. Detta kan även beskrivas som hur eller på vilket sätt data eftersöks i databaser. Vilken komponent eller metod som är den viktigaste under delprocessen DM har alltså debatterats genom åren.

Den faktiska processen i KDD består av nio olika delprocesser (Fayyad et al, 1996; Goebel & Gruenwald, 1999). Hela processen är interaktiv, iterativ och involverar ett flertal olika steg innan den slutgiltiga kunskapen kan utvinnas (Fayyad et al, 1996). Processen och dess delprocesser återfinns nedan (*figur 2.2)*. För att få djupare förståelse av figuren består delprocessen data mining av ytterligare tre delprocesser (delprocess 5,6 och 7 nedan).



Figur 2.2 KDD-processen (Fayyad et al, 1996b)

Nedan beskrivs varje enskild delprocess i KDD och dess betydelse för den totala processen (Fayyad et al, 1996).

- Den första delprocessen syftar till att sätta upp mål med processen och skapa förståelse över den domän som är relevant.

- I delprocessen *Selection* sker ett urval av data som är relevanta för den domän som bestämts i föregående steg.

- Den tredje delprocessen *Preprocessing,* består av att ta bort data som inte är relevant för den valda domänen.

- Nästa delprocess, som också är en typ av preprocessing, *Transformation* innebär att reducera data och välja hur de ska bli presenterade utifrån de mål som vill uppnås med processen.

- Den femte delprocessen består av att passa ihop målen som angavs i första delprocessen med olika metoder inom data mining.

- Den sjätte delprocessen präglas av tekniska aspekter vilket innebär: val av DM-algoritm och metod som ska användas vid sökningen efter datamönster.

- Nästa delprocess består av själva DM. Där sökning och insamling av data genomförs efter olika mönster som är av intresse för den utvalda domänen.

- Delprocessen som följer benämns *Interpretation/evaluation* där förekommer tolkning av de insamlade mönstren, för att säkerställa att mönstren stämmer överens med den utvalda domänen.

- Den sista delprocessen behandlar den slutgiltiga kunskapsupptäckten. Kunskapen används antingen direkt, implementeras i ett annat system eller dokumenteras för senare användning. I detta steg kontrolleras även om lösningsförslaget kommer i konflikt med tidigare antaganden eller andra kunskaper.

När hela KDD-processen har genomförts analyseras resultatet utifrån de mål som fastställts vid processens början. Detta för att undersöka att resultatet överensstämmer med det uppsatta målet. Om detta inte är fallet itereras processen om igen, alternativt läggs ned.

Under processens gång kan det förekomma betydelsefulla iterationer mellan olika delprocesser. Dock är inte iteration mellan delprocesserna alltid nödvändigt beroende på om resultatet är tillfredsställande eller inte. Forskare anser ofta att delprocessen där DM förekommer är den viktigaste men så är inte fallet utan alla delprocesser fyller sitt eget syfte för att kunna utvinna korrekt kunskap. (Fayyad et al, 1996)

2.3.1 Delprocessen Data Mining (DM)

DM är en process som hjälper bland annat organisationen att identifiera korrelationer eller mönster mellan olika fält i relationsdatabaser. Informationen bearbetas därefter till kunskap om framtida trender eller historiska mönster. (Fayyad et al, 1996; Frawley et al, 1992; Lou, 2008)

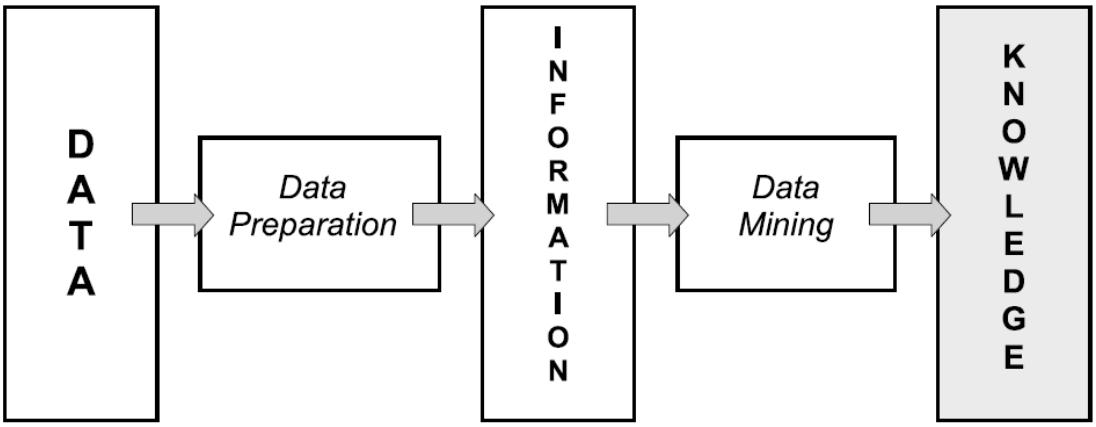
De mönster eller modeller som skapas av data är själva slutprodukten eller syftet med DM. Det är dessa som måste vara relevanta och unika för en given KDD-process för att kunskap ska kunna utvinnas. Definitionen av mönster relaterat till DM skiljer sig något åt i olika publikationer kring området, beroende på undersökningsmetoderna. Frawley et al (1992) har dock fastställt en definition som generellt sett kan användas som ledmotiv:

Given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset FS of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in FS. (Frawley et al, 1992, p. 58)

Tanken är att mönstren ska presenteras som ”påståenden” som beskriver relationer i delmängder av data. Dessa ”påståenden” ska i sin tur vara enkla att förstå (språkligt) och vara användbara (säkra).

De mönster som skapas ska alltså sedan kunna fungera som en avbildning av relationer av data som ska vara kunskapsbärande. Visar det sig att mönstren inte uppfyller något användbart eller unikt kan de alltså inte klassas som kunskap. En förenklad bild av

förfarandet att transformera data till kunskap genom DM presenteras nedan (*figur 2.3*). Modellen är ursprungligen tagen från ett sammanhang som beskriver BI, men visar ändå på vikten av DM i relationen mellan information och kunskap i organisationer.



Figur 2.3 Förenklad modell av KDD-processen (Michalewicz et al, 2006)

Det finns ett stort antal olika metoder och modeller av DM, men de flesta utgår ifrån två generella mål, nämligen förutsägande- och beskrivande DM. Förutsägande DM går ut på att leta efter mönster som riktar sig mot specifika egenskaper i information för att kunna förutsäga framtida beteenden. DM som en beskrivande operation, går istället ut på att låta systemet själv söka igenom information efter återkommande kännetecken t.ex. under vilka tider en butik säljer mest varor. (Fayyad et al, 1996)

DM omfattar testning av modeller (förutsägande eller beskrivande), skapande av olika typer av mönster (t.ex. användare) samt de data som observeras eller den information som eftersöks. Modellens eller metodens syfte är att beskriva vilken sorts kunskap KDD strävar efter att upptäcka. Huruvida kunskapen som genereras via den valda modellen är till nytta eller inte, bestäms utifrån det mänskliga omdömet och är en del av den interaktiva processen KDD. (Fayyad et al, 1996)

Som tidigare nämnts, finns det idag ett flertal olika modeller eller metoder som används i stor utsträckning för att bedriva DM. DM består inte av en enskild teknik, utan varje användbar metod som kan påträffas för att hitta mer information kring ett ämne bör användas (Goebel & Gruenwald, 1999). De flesta metoder inom området är baserade på beprövade och testade tekniker från maskinkunskap, mönster igenkännande och statistik i from av: klassifikation, kluster, regression osv. (Fayyad et al, 1996). Olika metoder används för att framhäva olika syften med DM, där varje metod erbjuder sina egna för- och nackdelar. En grundläggande distinktion mellan olika metoder är hur information eftersöks i en databas, explicit eller implicit (Goebel & Gruenwald, 1999).

I databaser förekommer det både explicita och implicita metoder. Implicit är kunskap som inte går att uttrycka på papper eller är underförstådd. Den explicita kunskapen kan relativt enkelt nås genom att skicka en enkel förfrågan till databasen, gällande t.ex. ett

personnummer. Den implicita kunskapen är till skillnad från den explicita något svårare att ta fram. En mer komplicerad fråga behöver istället ställas till databasen, vilket skiljer sig från mer enkla och vardagliga operationer. Denna förfrågan kräver att flera databaser samarbetar, vilket blir mer komplicerat. Denna typ av frågor är svårare för en databas att besvara eftersom databasen inte är designad för att besvara frågan, vilket även gör att DM blir svårare att bedriva utifrån implicita frågor. (Birrer, 2005)

2.3.2 Data Cleaning och Preprocessing

Data cleaning omfattar aktiviteter som gemensamt strävar efter att förbättra kvaliteten i data (Maletic & Marcus, 2000; Rahm & Do, 2000). Maletic & Marcus (2000) anser att processen bör försöka sträva efter att upptäcka och reparera data som inte är sammanhängande eller är felaktiga så att de blir fullständiga. Rahm & Do (2000) hävdar istället att de data som påträffas som felaktig eller icke sammanhängande istället borde raderas. Detta eftersom det råder brist på användbara verktyg för att utföra automatisk reparation av felaktig data. För att reparation ska bli effektivt behövs manuella metoder som reparerar felen, vilket är kostsamt både penga- och resursmässigt.

Kvalitetsproblem i data uppstår framförallt i databaser och kan bero på att data lagras på fel sätt eller att de inte bildar något sammanhang och kan på så sätt inte genererar information. Dessa problem är framförallt viktiga att beakta och försöka rätta till när flera olika datakällor används t.ex. när flera olika databaser behöver samverka för att svara på en förfrågan. När flera olika datakällor används, ökar även risken för redundant data och därför blir data cleaning extra nödvändigt i dessa sammanhang. (Rahm & Do, 2000)

Data cleaning eller preprocessing, som processen som raderar felaktiga data kallas, är den tredje delprocessen i KDD. Där raderas all data som inte är användbar för vidare analys under KDD-processens gång (Fayyad et al, 1996). Viktigt att påpeka här, är att Preprocessing strävar efter att radera felaktig data istället för att reparera den (se ovan). Det finns dock inga standardiserade definitioner eller tillvägagångssätt på hur data cleaning utförs i KDD, utan det kan variera från olika domäner eller olika system som utför KDD (Maletic & Marcus, 2000).

2.3.3 KDD – Relevanta mönster

De mönster som påträffas via DM måste vara relevanta för den kontext de används inom (Fayyad et al, 1996; Frawley et al, 1992; Siberschartz & Tuzillin, 1996). Via DM genereras en mängd olika mönster och det är därför viktigt att kunna sortera dessa och inrikta sig på de mönster som kan skapa kunskap för den domän som är relevant. Frawley et al (1992) beskriver tre kriterier som bör uppfyllas för att mönster ska kunna klassas som relevanta, nämligen att de är:

- Ovanliga
- Användbara
- Icke-triviala

Ovanliga och användbara mönster beskriver sig själva, men dessa två kriterier är alltså inte tillräckliga. De mönster som påträffas måste även vara icke-triviala att beräkna.
För att system ska behandla information icke-trivialt, måste systemet kunna göra annat än att endast beräkna statistik. Ett system som används för att genomföra relevanta upptäckter i mönster måste vara utformat på ett sådant sätt att det kan bestämma vilken information som behöver beräknas och huruvida resultatet är relevant för sammanhanget eller området. Alldaglig information, som andra redan känner till, är inte av intresse, utan nya upptäckter måste göras och då behöver systemen kunna genomföra dessa beräkningar för att ny kunskap ska kunna upptäckas. (Frawley et al, 1992)

Verkligheten är dock inte lika enkel vilket stora delar av litteraturen påvisar. Forskare inom KDD har länge arbetat med metoder som används för att ”mäta” hur pass relevant ett mönster är (Siberschartz & Tuzillin, 1996). De två generella mått som bl.a. Siberschartz & Tuzillin beskriver är hur pass relevanta mönster är, utifrån objektiva och subjektiva mått. De objektiva måtten fastställs utifrån relationer i data med t.ex. hjälp av associationsregler. Hur pass relevant ett mönster är, mäts alltså i data och av andra data. Ett objektivt fastställande kan t.ex. vara att kunder som köper kaffe, även köper kaffefilter (kaffe - > filter). Subjektiva mått utgår istället ifrån människans egna preferenser. Detta leder till att analysen av mönster kan komma att skilja sig från person till person beroende på hur pass relevanta denne person tycker att mönstren är.

McGarry (2005) hävdar att valet att använda objektiva eller subjektiva mått kan skilja sig åt. KDD strävar efter att finna ovanliga och överraskande mönster som tidigare inte är kända. Detta behöver ofta fastställas genom subjektiva mått men det kan vara svårt eftersom människan sällan reagerar på om något är överraskande. Upptäckten behöver sedan även fastställas utifrån användbarhet vilket även de utgår ifrån subjektivitet (Siberschartz & Tuzillin, 1996). En upptäckt för en viss analytiker behöver inte vara användbar för en annan, beroende på användningsområdet.

Guyon et al (1996) beskriver en användbar metod som kan appliceras i KDD. Metoden går ut på att framställa informativa mönster av data genom användningen av specifika algoritmer för genomsökning av data. De mönster som inte klassas som informativa benämns istället för skräpmönster och bör raderas eftersom de antingen genererar meningslös data som inte kan bilda ett sammanhang eller bara innehåller helt felaktig data.

2.3.4 KDD – Säkra mönster

Frawley et al (1992) hävdar att säkerheten i data är ett viktigt mått för att fastställa att de data som samlats in är pålitliga och korrekta. Författarna beskriver att det finns olika sätt att fastställa säkerheten i data, beroende på användningsområdet. Vissa områden kräver att data säkerställs mer noggrant än andra områden. De faktorer som används för att säkra data är bland annat integriteten i data, storleken på de data som samlats in för att fastställa en speciell upptäckt och de stöd upptäckten har från redan generellt känd kunskap inom det specifika området där upptäckten gjorts. Om data inte kan säkerställas blir mönstren oberättigade och kan senare inte bilda kunskap.

Frawley et al (1992) hävdar även att processen för att finna mönstren måste vara snabb och effektiv. Det handlar om att passa rätt sorts algoritmer till KDD för att relationer mellan data ska kunna påträffas under förutsatta och accepterade förhållanden.

van Well & Royakkers (2004) tar upp en annan aspekt av problemet kring säkerställande av information. Personers privata information kan spridas mycket enkelt på t.ex. Internet om de inte är uppmärksamma. Problem kan uppstå när personer inte får chansen att samtycka med den informationen som de kopplas ihop med. När personen inte kan godkänna den information som samlats in så dras alltså olika slutsatser. Dessa slutsatser kan stämma överens eller bestå av olika antaganden utifrån personens agerande på t.ex. en viss hemsida. Detta kan i sin tur göra anspråk på problem kring personers privata information då den kan vara felaktigt formulerad. I många fall är detta något som personen aldrig får reda på och kan därför inte heller påpeka att den insamlade informationen är falsk.

I sin tur kan detta leda till att information används på fel sätt, vilket den inte var ämnad till från början. van Well & Royakkers (2004) hävdar dock att detta inte är ett problem i nuläget eftersom tekniken inte finns tillgänglig i så pass stor utsträckning. Det kan däremot vara något som kan komma att ses i större utsträckning i framtiden.

2.3.5 Kognitionspsykologins inverkan på DM

DM har som uppgift att utvinna användbara mönster ur data för att skapa kunskap i KDD-processen. Eftersom kunskap är nära relaterat till mänskligt beslutsfattande argumenterar flera forskare bland annat Pazzani (2000) att kognitiv psykologi är ett relevant område för DM. Pazzani (2000) hävdar att KDD-processen borde involvera inslag av den mänskliga kognitiva processen för att bli mer användbar. Det är trots allt människors uppfattningar om *ovanlighet, nytta och förståelse* som bestämmer ifall användbar kunskap har erhållits via DM processen. Han presenterar även en modell med artificiell intelligens, databaser, statistik och kognitiv psykologi, vilket är de fyra element som borde ingå i KDD-processens grundkoncept. Databaser, statistik och AI är något som tillhört DM sedan begynnelsen, medan kognitiv psykologi är ett element som egentligen också alltid funnits i bakvattnet till tekniken, men har först på senare år börjat analyseras utifrån deras betydelse för DM (Pazzani, 2000).

2.4 Sammanställt teoretiskt ramverk

För att knyta samman vår teoretiska studie har vi sammanställt de olika delarna av teorin som vi kommer att utveckla i vår undersökning. Vi har valt att exkludera vishet i förhållandet mellan DIKV genom vår undersökning. Dels eftersom KDD-processen endast behandlar data, information och kunskap, men även för att kunskap är den högsta graden av beslutsunderlag som organisationer behandlar (Michalewicz et al, 2007). Forskare har även beskrivit hur vishet allt mer försvinner från forskningen kring IS och KM, vilket är ytterligare en anledning till vårt beslut (Rowley, 2007).

Tabell 2.3 Teoretiskt ramverk över KDD-processen

Delprocess	Beskrivning
1) Purpose and Objectives	Bestämma processens syfte och mål samt skapa förståelse över den domän som är relevant (Fayyad et al, 1996). Klargöra vad för kunskap som eftersöks i processen, även relaterat till hur eller på vilket sätt data eftersöks i en databas (Goebel & Gruenwald, 1999).
2) Selection	Urval sker av de data som är relevanta för den domän som bestämts i föregående steg (Fayyad et al, 1996). Utveckla modeller beskrivande hur och vilken information som ska samlas in för att effektivisera arbetet (Maimon & Rokach, 2005).
3) Preprocessing	Radera data som inte är relevant för den valda domänen (Fayyad et al, 1996; Rahm & Do, 2000). Upptäcka och reparera data som inte är sammanhängande eller är felaktig så att de blir fullständiga (Maletic & Marcus, 2000).
4) Data Mining	Klargöra om det är förutsägande eller beskrivande DM som ska utföras (Fayyad et al 1996; Frawley et al 1992; Goebel & Gruenwald, 1999). Passa ihop syfte och mål som angavs i första delprocessen med olika metoder inom DM, val av DM-algoritm och metod som ska användas vid sökningen efter datamönster, sökning och insamling av data som är relevanta för den utvalda domänen samt mönsterbildning av resultatet (Fayyad et al, 1996). Presentera mönstren i form av ”påståenden” som beskriver relationer i delmängder av data (Frawley et al, 1992). Algoritmerna som används vid kunskapsutvinningen bör vara effektiva och skalbara gentemot stora databaser (Chen et al, 1996). Beskriva hur information eftersöks i en databas, explicit eller implicit (Goebel & Gruenwald, 1999).
5) Interpretation and Evaluation	Tolkning av de insamlade mönstren, för att säkerställa att mönstren stämmer överens med den utvalda domänen (Fayyad et al, 1996). Anpassa bedömningen av ett säkert mönster till den specifika domänen (Frawley et al, 1992). Säkerhetsåtgärder i form av verifiering att en viss upptäckt är sanningsenlig (van Well & Royakkers, 2004). Säkerställa att relevanta upptäckter gjorts, annars itereras processen (Fayyad et al, 1996; Frawley et al, 1992; Siberschartz & Tuzillin, 1996). Mönstren som skapas bör vara användbara, ovanliga och icke-triviala (Frawley et al, 1992). Objektiva och subjektiva mått som används för att fastställa mönstrets relevans (Siberschartz & Tuzillin, 1996). Framställning av informativa mönster för att styrka relevansen, icke informativa mönster benämns skräpmönster och raderas (Guyon et al, 1996).

Eftersom vi kommer att använda oss av KDD-processen som ett analysverktyg genom uppsatsen har vi valt att samla de olika delar vi kommer att fokusera vår undersökning vid i

ett teoretiskt ramverk (se tabell 2.3). I den vänstra kolumnen återfinns de olika delprocesserna vi har valt att inkludera i ramverket. I den högra kolumnen finns beskrivningar av respektive delprocess och de åtgärder som bör beaktas. Ramverket består av fem olika delprocesser, till skillnad från den verkliga KDD-processen som består av nio. De delprocesser vi har valt är inkluderade eftersom de speciellt återspeglar datakvalitetsarbetet ur olika synvinklar. Kvalitetsaspekten kommer således att vägas in och utvärderas under respektive delprocess genom vår undersökning. Vi har även valt att inkludera aspekter som säkra och relevanta mått av resultatet eftersom det är en viktig aspekt för att fastställa att processen genererat ett kvalitativt resultat. Det kognitiva perspektivet är också en intressant aspekt som kommer att undersökas utifrån dess datakvalitetsförbättrande funktion, om någon sådan är möjlig. Vi kommer mer detaljerat gå in på hur ramverket används i vår undersökning i avsnitt 3.2.

3 Metod

I detta kapitel argumenterar vi för den metod vi har använt oss av för att undersöka vårt problemområde. Här beskriver vi vilken forskningsstrategi vi använt, hur undersökningen är upplagd och genomförd, de data vi samlat in samt de kriterier vi har för att den ska vara trovärdig.

3.1 Forskningsstrategi

För att undersöka vårt problemområde och besvara vår frågeställning har vi först behandlat en teoretisk studie som presenterades i föregående kapitel. Den teoretiska studien mynnade ut i ett eget konstruerat teoretiskt ramverk (avsnitt 2.4). Tanken är att ramverket ska fungera som en sammanställning av vårt teoretiska underlag, men samtidigt en introduktion till vår undersökning som vi presenterar i detta kapitel.

Genom att studera teoretisk litteratur beskrivande forskningsmetodik bestämde vi oss för att genomföra kvalitativa expertintervjuer för att kunna få kvalitativa svar på de frågor vi ställt oss och på så sätt nå vårt mål med undersökningen.

Anledningen till att vi valt att arbeta efter en kvalitativ metod är dels att tyngdpunkten läggs vid ord istället för kvantifiering i form av data och siffror (Bryman, 2001). Vi behövde alltså kunna diskutera de frågor vi har för att kunna komma fram till svar på våra frågor. På så sätt har de kvalitativa intervjuerna gett oss möjligheten att ha öppna samtal med våra informanter och på så sätt skapat en flexibel miljö för ventilering av frågor och svar. Detta hade inte varit möjligt om vi istället genomfört en kvantitativ metod.

En annan anledning till att vi valt en kvalitativ forskningsstrategi är att skillnaden mellan planering, genomförande och analys är förhållandevis liten (Jacobsen, 2002). På så sätt har vi kunnat planera, genomföra och analysera våra intervjuer och sedan gå tillbaka för att undersöka att vi fått ihop de material vi önskade. På detta sätt har vi kunnat avgränsa och specificera vår undersökning ytterligare. Vilket har varit viktigt eftersom området är både stort och komplext. Genom att ytterligare specificera och avgränsa området blir det både lättare för oss att undersöka området men även för läsaren att följa uppsatsens upplägg.

3.2 Undersökningens uppläggning

Vi har med hjälp av kvalitativa expertintervjuer undersökt hur datakvaliteten i databaser påverkar skapandet av datamönster. För att undersöka detta har vi konstruerat ett teoretiskt ramverk bestående av fem delprocesser från KDD-processen (se tabell 2.3). Delprocesserna är, som tidigare nämnts, speciellt utvalda för att kunna relateras till datakvalitetsaspekter, men även mått på resultatet som bildas i processen (se avsnitt 2.4).

Med hjälp av ramverket har vi även utformat två stycken intervjuguider som har används vid expertintervjuer med personer som är verksamma inom området för DM och dataanalys (avsnitt 3.3.2). Ramverket har även fungerat som ett analysverktyg då vi diskuterat och

analyserat vår insamlade data från expertintervjuer (se avsnitt 3.6 för mer detaljerad analysmetod).

Resultatet från analysen och det totala intrycket av studien mynnar ut i våra slutsatser där en reviderad modell av KDD-processen presenteras, vars syfte är att beskriva vilka förebyggande åtgärder som krävs för att förbättra resultatet av de datamönster som utvinns.

3.3 Undersökningens perspektiv på DIKV

Vi kommer att använda oss av Zenelys definition av relationerna mellan DIKV i vår undersökning. Där ses data och information som enkla uppsättningar som liknar atomer, och som i sig själva inte är användbara, atomism. Kunskap och vishet beskriver i sin tur helheten av delarna relaterat till data och information, holism.

Människan kan ha tillgång till data eller information, men kan inte ha tillgång till kunskap. Det är upp till varje enskild individ att skapa kunskap av den information som han eller hon erhåller.

Vi har valt arbeta efter dessa tolkningsperspektiv eftersom vi tycker de stämmer bra överrens med hur data, information och kunskap behandlas i KDD. Där data i sig själv inte är användbar, data behöver först sammanställas med andra data innan information kan bildas. Informationen måste sedan i sin tur också sammanställas. De relationer och likheter som upptäcks bildar därefter mönster i informationen och det är först då den utgör ett sammanhang. Det är sedan upp till varje enskild människa att skapa kunskap utifrån dessa sammanhang. Bara för att KDD har funnit ett ovanligt, icke trivialt och användbart mönster kan det inte klassas som kunskap. Vi anser att det bestäms utifrån en subjektiv parameter och är alltså upp till var och en.

3.4 Expertintervjuer

Vår undersökning innefattar tre olika expertintervjuer som grundas på två olika intervjuguider. Vi kommer nedan beskriva hur vi gått tillväga för att samla in den information vi erhållit.

3.4.1 Intervjuernas strategi

Vi har valt att använda oss av semistrukturerade intervjuer vilket Bryman (2001) hävdar är en intervjuform, där man använder en intervjuguide som innehåller en lista med generella teman och övergripande frågor. Guiden ska vara till hands under intervjun och se till att alla ämnen har berörts. Vi har på så sätt även kunnat hålla oss mer flexibelt till intervjuguiden för att intervjuerna skulle resultera i ett gott kunskapsunderlag. Frågorna behöver inte heller följa varandra utan kan ställas när som helst så länge som intervjun pågår. Denna semistruktur har även gjort att en mer öppen dialog mellan oss och informanten har kunnat genomföras.

Informanten har stor frihet att utforma de svaren och på så sätt blir intervjuformen därför flexibel.

För att våra informanter ska vara förberedda inför intervjun, har vi sänt mail till dem i god tid innan för att kunna beskriva vad vårt mål med intervju är och de ämnena som kommer att beröras under intervjun.

3.4.2 Intervjuguidernas utformning

Intervjuguiderna är utformade efter det teoretiska material som vi använder och är strukturerade efter frågeställningarna (se bilaga 3). Vi har, med hjälp av intervjuguiderna, stegvis både undersökt vårt problemområde och specificerat våra frågeställningar. Den första intervjuguiden är utformad efter vårt problemområde och med hjälp av denna lyckades vi komma fram till våra frågeställningar. Den andra består av våra frågeställningar där vi har skapat teman som hör ihop med våra frågeställningar. Detta förlopp beskriver Kvale(1997) som vanligt inom vetenskaplig forskning där två intervjuguider skapas som komplimenterar varandra.

Intervjuguide 1 (bilaga 2) behandlar KDD och DM mer ingående för att vi skulle få en bättre inblick i tekniken och kunna ställa oss mer kritiskt till litteraturen.
De övergripande inslag som ses som ett tema i denna intervjuguide var:

- KDD som en process
- Förutsägande- kontra beskrivande DM
- Relevanta och säkra mönster
- Kognitiv och subjektiv inverkan

Intervjuguide 2 (bilaga 2) utgår ifrån den första men behandlar istället mer övergripande datakvalitetsarbetet och hur det påverkar KDD. De mest övergripande rubrikerna i denna intervjuguide var:

- Datamängdernas utbredning, positivt/negativt
- Klassificering/kvalificering av data
- Bedömning av datakvalitet
- Analys av resultatet
- Framtida perspektiv och förändringar

Intervjuguiderna överlappar dock varandra vilket gör att de kan användas parallellt vid analysen av vårt material. Två olika intervjuguider lämpade sig även bäst eftersom kunskapen skiljde sig åt hos informanterna. Vi ansåg att vi bara ville ha informationsrika svar på våra frågor och därför kunde vi inte ställa samtliga frågor till alla informanter.

3.4.3 Frågornas utformning

Ett syfte med våra intervjuer, var att utforska mer om själva tekniken bakom DM och om det i vardagsspråk finns något som kallas KDD, och om det verkligen är detta som företag använder sig av. Vi hade därför en rad olika frågetecken då teorin inom detta område kan vara både komplex och svårförståelig. Ett viktigt led vid utformningen av frågor till våra intervjuer var därför att ställa öppna frågor. Genom att använda denna typ av frågor var det enklare för oss att ta in ny information om området samtidigt som vi gav informanten mer fritt spelrum för att visa sin kunskapsnivå (Bryman, 2001).

Vi har även varit intresserade av att ta reda på hur datakvalitetsarbetet behandlas i KDD-processen. Detta för att besvara vår övergripande frågeställning: *hur påverkar datakvaliteten i databaser skapandet av datamönster?*
KDD är, som beskrivits tidigare, en hel process vilket gjort att vi behövt undersöka processen från olika synvinklar. Från syfte och mål med processen, till val och specificering av data och slutligen till måtten av resultatet. På så sätt har vi utformat frågor som ska täcka alla de olika delar vi är intresserade av att undersöka. Vi vill därför påstå att frågorna avseende KDD är mer specificerade utefter vårt teoretiska ramverk. Frågorna angående datakvalitet är inspirerade av teorin men består däremot till större del av våra egna funderingar kring hur datakvalitet undersöks, om det finns några typer av kategoriseringar eller kvalificeringar av data, vilka verktyg som används för att bedöma och förbättra data osv. (se bilaga 3 för samtliga intervjufrågor). Ett led i denna utveckling beror på att de sistnämnda frågor har varit svåra att få bra svar på i vår teoretiska granskning, samtidigt som vi anser att vi behöver vara kritiska och värdera de teoretiska svaren mot de vi erhållit via expertintervjuerna.

Med hjälp av frågorna har vi även kunnat komma till nya insikter beträffande vårt problemområde och även specificerat våra frågeställningar ytterligare. Detta har också varit ett bakomliggande syfte då vår undersökning handlar om att vi skapar oss en egen bild av den process vi är intresserade av. Resultatet ska inte direkt komma från våra informanter, utan de frågorna vi ställer till dem ska hjälpa oss att finna nya tankebanor. På så sätt tillåts vi även att agera mer subjektivt genom vår analys för att sedan fastställa våra slutsatser (Backman, 1998). Något vi anser är viktigt, eftersom det ska vara vi som utforskat ämnet som kommer fram till våra egna slutsatser.

3.4.4 Urval

Vi har valt att genomföra tre olika expertintervjuer på tre olika företag som är verksamma inom området DM. Vi har via Internet och egna kontakter funnit ett antal företag som vi kontaktat via telefon och berättat om vår undersökning. De tre olika företagen som vi valde för intervju, är verksamma inom skilda områden, vilket gör att vi fått olika perspektiv på våra svar. Det har även varit ett bakomliggande syfte eftersom vi konstruerat två stycket intervjuguider till skilda experter. Den ena riktar sig mer till experter inom KDD-processen och den andra till experter inom datakvalitet och statistisk dataanalys.

Bland de företag vi valt finns två internationellt ledande företag inom databaser, och utveckling av DM-verktyg samt ett företag som arbetar med framtida lösningar inom Artificiell Intelligence (AI). På så sätt har vi kunnat skapa oss kunskap om många delar av det spektrum som DM ingår i, samt de aspekter vi intresserat oss för, datakvalitetsarbetet.

3.4.5 Beskrivning av informanter

Informant A

Informant A arbetar på ett relativt litet företag med ca 10 anställda. Företaget är verksamma inom området AI. Med hjälp av AI skapar de mobila lösningar, framtida Internetlösningar med integrerad mänsklig kunskap och framtida DM-lösningar med bland annat inriktning på kundinsikt. En stor del i företagets vision är alltså att arbeta mot framtida lösningar och här har de blivit väldigt framgångsrika.
Ett av deras stora projekt för tillfället är skapandet av sökmotorer för mobiltelefoner där AI, machine learning och DM är integrerat. Att genomföra en intervju på detta företag var mycket bra, eftersom de anställda hade bred kunskap om tekniken DM och kunde på så sätt svara mycket utförligt på våra frågor.
Vi har därför använt oss av *intervjuguide 1* för att få reda mer kring just tekniken bakom DM och hur förfarandet går till mer förenklat.

Informant B

Informant B arbetar på ett av de företag som är ledande i världen på utveckling av affärssystem. De är även en av giganterna inom databaslösningar och Database Management Systems (DBMS). På senare år har företaget arbetat med att integrera DM-verktyg i deras DBMS. Vi ansåg det därför högst relevant att ta reda på hur deras kunskap skulle kunna hjälpa oss med vår undersökning. Den person vi intervjuat på företaget har lång erfarenhet av just DM och analyser av data. Under denna intervju använde vi oss av i*ntervjuguide 2*, då vi ansåg att denna person skulle kunna hjälpa oss med de frågor vi hade angående datakvalitet och hur den påverkar analysen av data.

Informant C

Informant C arbetar även han på ett företag som är en av de ledande leverantörerna av mjukvaror i världen. Företagets nisch är från begynnelsen statistik och de har tagit sina kunskaper inom detta område och implementerat detta i sina beslutstödssystem. Statistik är ett oerhört viktigt inslag i DM, då alla datasammanställningar just är av denna art. Vi ansåg det därför viktigt att undersöka hur en statistiker ser på datakvalitetsarbetet med DM. Vår informant på detta företag har ca 10 års erfarenhet inom olika områden av beslutsstöd och är utbildad statistiker. Vi har valt att använda oss av i*ntervjuguide 2*, då datakvalitetsperspektivet lämpar sig bäst för denna informant, men även för att det är vårt huvudsakliga undersökningsområde.

3.4.8 Genomförande av intervjuer

Intervjun med informant A genomfördes på dennes kontor på IDEON i Lund och vi samtalade ungefär 1 timma om vårt område. Som tidigare nämnts ville informant A inte att vi skulle spela in samtalet, men vi ansåg ändå att vi fick ett så pass bra underlag från intervjun och därför utgjorde det aldrig några problem vid vår senare sammanställning.

Informant B och C var lokaliserade i Stockholm och därför fick vi komma överrens om en tid med båda dessa för en telefonintervju. Samtalen med dessa informanter varade också ungefär 1 timme och kunde spelas in då båda informanterna samtyckt till detta.
Vi använde oss av en dator och Skype[1] när vi genomförde dessa två telefonintervjuer. Vi ansåg att detta var lämpligast eftersom vi båda kunde prata med informanten samtidigt och vi kunde även spela in samtalet direkt på datorn.

3.4.9 Intervjuernas resultat

De resultat som intervjuerna mynnade ut i har gett oss en bra bild över hur DM fungerar i praktiken och hur olik den process vi kallar KDD kan se ut inom olika affärsområden. Resultatet av intervjuerna har även hjälpt oss att specificera våra frågeställningar ytterligare men framför allt att se på litteraturen mer kritiskt.

En sammanställning av intervjuguider och transkriberingar från våra intervjuer återfinns i bilagorna. En mer detaljerad sammanställning som är direkt kopplad till vår undersökning finns i kapitel 4.

3.5 Sammanställning av insamlade data

När intervjuerna genomfördes, har vi först försökt att använda oss av en dator för ljudupptagning och i efterhand transkribera intervjun. Om en informant inte gillat en ljudupptagning, så har vi istället antecknat under intervjuns gång och sen gjort en fullständig redogörelse om vad som sades under intervjun . Anteckningar medför dock brister, t.ex. vid en redogörelse av den sociala kontexten vid intervjutillfället samt svårigheter i att bibehålla intervjupersonens ordalag och uttryckssätt (Bryman, 2001). Vi anser dock att det är innehållet i sig, som är det väsentliga i intervjun och de resultat vi kommer fram till. Det har därför inte varit några problem att sammanställa vare sig inspelade eller icke inspelade intervjuer.

---

1 Mer information om denna tjänst finns på www.skype.com

3.6 Analysmetod

Då materialet från våra genomföra intervjuer har sammanställts behövde vi någon form av angreppssätt för att lyfta fram de relevanta delarna. Vi har därför reducerat komplexiteten av materialet samt förenklat och strukturerat upp innehållet för att få en bättre överblick (Jacobsen 2002).

Materialet har kategoriserats efter vårt problemområde, frågeställningar och teoretiska underlag. Vi har på så sätt fastställt ett antal rubriker eller teman som överensstämmer med vårt sammanställda teoretiska ramverk och syftet med uppsatsen. Systematisering och kategorisering av materialet är viktigt för att kunna redogöra för våra upptäckter (Jacobsen 2002).

Rubrikerna har fungerat som ett verktyg där vi kunnat kategorisera de olika teman som vi anser överensstämmer med vår undersökning (se kapitel 4). Därefter har vi kunnat undersöka vårt insamlade material och placera informanternas respektive resonemang under rätt rubriker. Vi har på så sätt även kunnat urskilja vad som ansetts som extra intressant i det insamlade materialet eller om vi funnit några oväntade resultat.

Vårt insamlade undersökningsmaterial har därefter analyserats kopplat till vårt teoretiska underlag. Vi har där valt att använda oss av vårt ramverk som analysverktyg och kommer att undersöka hur kvalitetsförbättringar kan ske parallellt med processen (se kapitel 5).

3.7 Kriterier för studiens trovärdighet

Vi har tagit fram intervjuguider för våra expertintervjuer där vi har valt att ställa relativt öppna frågor. Detta eftersom vi behövde klara ut vissa definitionsfrågor rörande DM ville samtidigtf skapa oss en för snäv bild av vårt problemområde.

För att vår undersökning ska bli trovärdiga och uppfylla sitt syfte, har vi använt oss av ett antal kriterier för att kunna styrka detta. Det första kriteriet är reliabilitet vilket syftar till tillförlitlighet och knyter an till frågan huruvida resultaten från en undersökning blir likartad om undersökningen genomförs på nytt (Bryman, 2006). För att höja reliabiliteten vid våra intervjuer har vi ställt öppna frågor för att inte styra informanten åt någon riktning. Vi har även till stor del tillfrågat informanterna om liknande saker och har på så sätt kunnat jämföra svaren.

Informanterna har även fått ta del av de sammanställningar vi gjort av våra insamlade data. På så sätt har informanterna kunnat granska och verifiera resultatet för att undersöka att det stämmer överrens med deras uppfattning.

Det andra kriteriet vi ämnar förhålla oss till är validitet, vilket Bryman (2006) hävdar är frågan om huruvida de slutsatser som genererats från en undersökning hänger ihop eller inte dvs. att klara slutsatser kan dras utifrån undersökningens resultat.
En fråga som vi därför ställt oss: *Svarar verkligen våra genomförda intervjuer på våra forskningsfrågor?*

Detta anser vi vara viktigt för att resultatet av våra intervjuer verkligen ska spegla den syfte vi haft viljan att utföra. Vi har även kunnat reda ut eventuella otydligheter och frågat om då vi anser att vi inte fått svar på de vi verkligen frågade om. På så sätt har vi även kunnat höja validiteten i uppsatsen.

3.8 Etik

Enligt Bryman (2001) förekommer det tre etiska principer som gäller för forskning. Informationskravet vilket innebär att forskaren ska informera de berörda personerna om undersökningens aktuella syfte. Samtyckeskravet, som kräver att deltagarna själva har rätt att bestämma om de vill medverka i undersökningen. Konfidentialitetskravet vilket innebär privata uppgifter om personerna som deltar i undersökningen.

För att säkerställa att de etiska principerna efterföljdes genomförandet av våra expertintervjuer, informerade vi våra deltagare om vårt syfte till vår undersökning och har även haft mail kontakt med vår intervjuperson.

Identifierbara kännetecken av informanterna har även uteslutits i vår undersökning samt i våra transkriberingar för att möta konfidentialitetskravet. Vi ansåg de inte nödvändigt att använda oss av informanternas eller företagens namn då undersökningen inte specifikt avser evaluering av dessa. Vi anser inte heller att anonymiteten har reflektera vårt resultat på något sätt.

Vi har även låtit informanterna granska de transkriberade resultat som vi sammanställt efter intervjuerna. Informant A ville dock inte att vi spelade in samtalet med honom, vi har därför inte kunnat genomföra en transkribering ordagrant av detta samtal utan endast återberättat intervjun i sin helhet. Informant A har dock fått samma möjlighet som de andra att granska de resultat som vi sammanställt efter intervjun.

3.9 Metodkritik

Ämnet vi har berört i denna undersökning har varit KDD, DM och datakvalitet vilket var helt nya ämnen för oss om man ser till teoriinnehållet. Vi har därför stegvis format vårt teoretiska underlag eftersom vi själva först har behövt sätta oss in i ämnet innan vi kunnat beskriva de utförligt. Detta har även medfört att teorin ändrats till stor del genom undersökningens gång. Vi har även använt oss av metaspråk och förklaringar av teorin till stor del för att underlätta för läsarna.
Våra genomförda expertintervjuer har även hjälpt oss att utforma teorin mer utförligt eftersom många frågetecken angående litteraturen har fallit på plats.

Skälet att endast använda oss av expertintervjuer som undersökningsunderlag kan ifrågasättas. Dock föll det ganska naturligt eftersom DM är ett ganska litet område, sett till marknaden. Genomförande av enkäter eller observationer hade därför varit svårt eftersom ett stort antal respondenter med god kunskap inom området skulle vara svårt att finna. Det skulle förmodligen resultera i ett stort bortfall vid en enkätundersökning, vilket sänker validiteten på undersökningen (Bryman, 2001).

4 Resultat av Expertintervjuer

I detta kapitel presenterar vi de undersökningar som vi har genomfört, samt de resultat som dessa har genererat. Avsnittet är uppdelat i fyra delar som är speciellt talande för vår undersökning. Dessa är ökade datamängder i världen, KDD och DM, datakvalitetens inverkan på mönsterskapande och validering av mönster.

4.1 Disposition för intervjuer

De intervjufrågor vi har valt att presentera har kopplingar till teoriavsnittet i vår uppsats rörande KDD-processen(DM), relevant- och säker information, kognitivt beteende samt datakvalitet. Syftet och utformning av frågorna återfinns i kapitel 3.

Precis som vi beskrev i metodkapitlet, så har vi använt oss av två olika intervjuguider. Intervju A används framförallt för att beskriva KDD-processen och DM mer ingående. Intervju B och C används istället mer för att ta reda på mer kring vårt egentliga problem, datakvalitetens inverkan vid skapande av mönster via data DM.

4.2 Ökande datamängd i världen

Kvantiteterna av data växer ständigt inom dagens IS. Detta gör att det inte råder någon brist på denna råvara, men frågan gällande kvantiteten kontra kvaliteten är dock en viktig fråga. Både informant B och C hävdar att ökningen av data är lavinartad. Informant B beskriver organisationen Winter Corporation som vartannat år genomför en undersökning med syfte att ta reda på vilka de största kommersiella databaserna är och hur stor ökningen av data är. Mellan år 2005 och 2007 var ökningen så stor som 300 %, där den största databasen mättes till 300 TB. Med denna ökning kan det komma att finnas databaser på petabyte nivå 2010.

Informant C tar även ostrukturerad data i beaktande och menar att alla mail, texter, webbloggar osv. också bidrar till dataökningen. Fingeravtryck och andra biometriska data sparas i stor utsträckning i olika databaser, något som knappt syntes för 10 år sedan. Nästa steg i utvecklingen inom dataanalys är att analysera dessa ostrukturerade data i större grad.

4.2.1 Konsekvenser av stora datamängder

Informant B och C anser båda att det finns både positiva och negativa aspekter rörande användning av stora datakvantiteter vid utförande av DM. Informant B hävdar att både djupet och bredden i data är viktiga mått för att kunna få fram ett bra resultat av en analys. Djupet i den meningen att data finns tillgänglig från flera år tillbaka och med bredden menas variationen av data. Är det ett bra djup och bredd i data kan sedan algoritmerna köras bättre. Informant C hävdar att stora datakvantiteter är fördelaktigt om de kan användas på bra sätt. På så sätt kan bättre mönster påträffas och bättre beslut genomföras. Det ställer dock stora krav på både kompetensen hos analytiker och även verktygen som används. Finns det dock brister hos analytiker eller i verktygen kan stora datakvantiteter istället ses som en nackdel. Det beror även mycket på situationen och vilket problemområde som DM kretsar kring.

Visar det sig att det är något specifikt eller undangömt en analytiker vill finna kan det vara svårt att hitta den lilla nålen i allt för stora datamängder.
Informant B beskriver istället hur stora datakvantiteter kan vara problematiska då data behöver flyttas eller göras om till andra format. Han förespråkar därför att själva motorn som utför miningen bör implementeras och köras i databasen där datamängden är lagrad. Sammanfattningsvis blir det statistiska underlaget bättre ju mer data som finns tillgänglig, men det kan samtidigt bli svårhanterligt med allt för stora datavolymer.

4.3 KDD-processen och DM

I detta avsnitt behandlas hur våra informanter se på de olika inslagen i KDD, beskriver skillnader och likheter samt hur framtiden inom tekniken kommer att se ut.

4.3.1 KDD-processen, en översikt

KDD-processen ser olika ut beroende på kontexten eller situationen där den används. Syftet med processen är att uppnå kunskap från data som utvunnits och behandlats på olika sätt genom processens gång. Alla tre informanterna beskriver vikten av att först och främst skapa ett syfte med processen, dvs. att definiera vilken kunskap som eftersöks. Informanterna B och C betonar även detta, och menar att DM inte bör bedrivas, om det inte finns något problem att lösa.

Informant A har en något annorlunda filosofi vad det gäller detta. Han menar istället att de bör fastställas huruvida det är input eller output som är relevant för sammanhanget. Visar det sig att det är output som är det relevanta bör en föreställning skapas rörande slutmålet med processen. När KDD-processen sedan har fullföljts behövs en uppföljning av arbetet. Detta för att kontrollera att den information som genererats kan användas som kunskap för att uppfylla processens grundläggande syfte. Visar det säg att detta inte är fallet, itereras processen till dess att relevant information har påträffats som kunskap kan utvinnas ifrån, berättar informant A.

Han hävdar vidare, om det är input som istället är det relevanta för processen så behöver mönster skapas för att utrycka syftet med processen. Då är det istället återkommande kännetecken som är det relevanta, mönstren i detta fall används för att beskriva vilka kännetecken som är relevanta för KDD-processen. De olika kännetecken som sedan genererats bildar processens output.

Både Informant B och C påpekar att det måste finnas ett problem som vill lösas för att DM ska få utföras. Processen är tidskrävande och dyr och därför är det inte hållbart att köra mining för att hitta spännande relationer i data, ett problem måste lösas.

4.3.2 Förutsägande kontra beskrivande DM

Förutsägande DM kan generellt beskrivas som sökandet efter framtida kunskap, medan beskrivande DM söker igenom datamängder efter åtkommande kännetecken som senare ska bilda kunskap. Informant A hävdar att skillnaden mellan förutsägande- och beskrivande DM egentligen kan beskrivas på ett enklare sätt i en naturlig process för att få mer klarhet.

I en given kontext kan beskrivande DM fungera som en input. Olika kännetecken som systemet ska söka efter bland data programmeras in, därefter genomförs själva transformationen och data eftersöks efter återkommande kännetecken utifrån de uppgifter som programmerats in. Resultatet av detta blir sedan process output.

Förutsägande DM kan istället beskrivas bakvänt. Inputen i detta fall består av syfte och föreställningar beskrivande processens mål, vilka framtida utfall som eftersöks. Transformationen består sedan av sökande efter data som kopplas till det syfte som fastställts. De data som genereras består av processens output.

Informant A beskriver även vad som är viktigt att utgå ifrån när det handlar om att genomföra DM, samt i vilken situation förutsägande- eller beskrivande DM bör användas. Han hävdar att valet mellan förutsägande- eller beskrivande DM först och främst beror på vad som vill eftersökas. Det handlar om att sätta saker i kontext till andra, att fastställa ett syfte med DM, men även att förstå och få insikt i en naturlig process. Därför är det svårt att beskriva vilken typ av DM som generellt sett mest används. Först när kontexten är underförstådd kan olika algoritmer användas för att utvinna data.

4.3.3 DM och nyskapande

Både informant B och C anser att datakvalitet är en av de viktigaste bitarna i ett framtidsperspektiv för att uppnå bättre och säkrare resultat via DM.

> *”Datakvalitet, datakvalitet och datakvalitet. Det är det viktigaste.” (Informant B, 2009-05-06)*

Algoritmerna, som består av matematiska formler, har varit ganska oförändrade de senaste åren. De är även stabila och analytiker vet vad de får ut av algoritmerna. Informant B förklarar bland annat att algoritmerna blir i stort sett värdelösa om data med dålig kvalitet matas in för beräkning. Han tror även att förändringen kommer att ligga i att kunna köra renare data och större datamängder. Informant C betonar även vikten av att analytiker är medvetna om hur pass vital datakvaliteten är och därför arbetar man hela tiden mycket med att säkra kvaliteten ytterligare.

Informant A hävdar att det sker mer utvecklingen av nya metoder, snarare än att tekniken för DM förändras. Metoderna utgår från olika typer av modeller beskrivande hur DM ska genomföras, medan teknikerna mer är olika verktyg, programvaror och algoritmer. A hävdar att dagens utveckling använder sig allt mer av konceptuella modeller, vilket informant C också betonar. Modellerna används för att skapa en bra bild över den relevanta domänen. Vidare beskriver A att valet av algoritmer inte längre är en lika viktig del i utvecklandet av

nya metoder. Det finns ett stort antal färdiga algoritmer och välja mellan som lätt kan passas till den framställda modellen. Det som är viktigt i dagens utveckling är att fokusera på den konceptuella modellen och se till att de variabler som är viktiga för processen används som grund för analysen.

Informanterna B och C anser istället att det snarare är i tekniken som den största förändringen sker. Informant C betonar bland annat vikten av att bygga upp varje DM projekt kring en analysfråga, vilket har varit grundläggande inom metodiken en lång tid tillbaka och har inte förändrats. B och C anser även att val av algoritm inte är lika viktigt nu som förr, men det finns mycket mer teknik bakom än just algoritmerna i form av olika verktyg som ständigt förbättras. Tekniken är de område som förändras mest, enligt B och C.

Samtliga informanter tror att DM kommer att förbättras i framtiden. Informant C anser att fler organisationer kommer att börja använda sig av det och då kommer även systemkraven att öka från leverantörer. Informant B tror även på mer halvfabrikat inom området. För tillfället är det dyrt för organisationer att starta en egen DM-verksamhet. Finns det enklare verktyg, som vem som helst med någorlunda datorkompetens kan använda så kommer efterfrågan att öka. Kanske till och med färdiga DM-lösningar för olika branscher. Konkurrensen mellan olika organisationer är också en viktig faktor menar informant B. DM kommer att behövas för att vara konkurrenskraftig.

Informant A beskriver även en annan viktig del i utvecklingen kring nya metoder. Han hävdar att intelligenta och smarta tekniker bör användas mer. AI används bland annat i utvecklingen av DM, vilket marknaden förmodligen kommer att se mer av i framtiden. Många av dagens system är gamla och hindrar dock utvecklingen av mer intelligenta DM-metoder. Det är därför svårt att använda DM med AI på system som inte behärskar tekniken. Han hävdar dock att framtiden kommer att bestå av allt mer intelligenta system vilket kommer leda till att DM blir ännu mer användbart och effektivt. Men för att detta ska kunna ske behöver som sagt domänen stöda dessa metoder. Informant A beskrivet t.ex. Internet som ett område där stora förändringar behöver ske innan intelligentare tekniker kan börja användas.

4.3.4 Kognitivt beteende

Forskare har på senare tid hävdat att kognitiv psykologi borde inkluderas i KDD-processen eftersom kunskap är nära relaterat till mänskligt beslutsfattande.
Informant A hävdar att kunskap om människors beteende för att förbättra DM- metoder men även resultatet är viktigt.

> *”Egentigen är all DM till för att hitta beteende och kunna sälja mer eller andra saker, är det inte så?” (Informant B, 2009-05-06)*

Informant B hävdar att det är något som används speciellt inom detaljhandeln för att följa kunders mönster och lyckas behålla dem. Det används även inom tv, relaterat till tittarsiffror. Informant C anser att denna psykologiska aspekt kommer att bli mer användbar i framtiden och speciellt på nätet. Där används det till stor del redan, men inte fullt ut.

Informant A beskriver även tanken bakom förloppen, hur de psykologiska aspekterna kan fungera. Han hävdar att det främst handlar om hur människor uppfattar systemet. Får människor en dålig uppfattning av systemet eller hemsidan känner de sig mindre engagerade och kommer troligen inte att lägga ner tid på att använda sig av olika funktioner. Visar det sig istället att människor får en bra uppfattning av systemet är sannolikheten större att de data dessa lämnar efter sig kan användas i större utsträckning. A hävdar att det i grunden ligger en hel del psykologiska aspekter i vilken utsträckning användbar information kan genereras från olika människor till DM- processen. Vilket bekräftar påståendet att människa-data-interaktion (MDI) även har en betydande roll i DM. Det rör sig speciellt om kognitivt psykologi eller kognitivt beteende, berättar han. Genom att förstå hur människor tänker och uppfattar system eller hemsidor kan olika DM-metoder på så vis dra nytta av kunskapen om mänskligt beteende och bli mer användbara.

Både informant B och C anser även att detta kan hjälpa till med en förbättring av datakvaliteten, då saknade värden och konstigheter i data blir mindre.

4.4 Datakvalitetens inverkan på mönsterskapande

Datakvalitet är ett stort och komplext område. Vi vill här ta reda på hur kvaliteten går att förbättras, hur kategorisering av data sker och hur informanterna ser på kvaliteten kontra kvantiteten.

4.4.1 Förebygga datakvalitet

Både Informant B och C hävdar att datakvaliteten är en vital del vid mönsterskapande i DM. Är datakvaliteten bristande går det kanske inte ens att utföra DM eller så kan resultatet bli felaktigt. För att komma till bukt med dessa problem anser informant B att dataprofilering bör genomföras innan själva DM görs. Vid profilering plockas extremvärden bort, eftersom dessa värden sticker ut för mycket från majoriteten. Genom erfarenhet vet man att dessa värden kan ge konstiga resultat och därför tas de bort från analysen och datakvaliteten kan på så sätt förbättras.

Informant C påpekar även vikten av att kontrollera saknade värden i datatabeller innan DM påbörjas. Visar det sig att en variabel eller tabell har ett väldigt högt bortfall av data går den förmodligen inte att använda.

Konstigheter i data är också något som bör undersökas, som t.ex. annorlunda tecken eller ett tal som anses vara konstigt. C hävdar vidare att det finns en uppsjö av olika verktyg som används för att optimera datakvaliteten. På mer avancerad nivå använder sig analytiker av olika hjälpmedel för att systematiskt granska datakvaliteten men precision. Det kan t.ex. handla om hur saker med samma namn men olika innebörd bedöms, dvs. hur svårare analyser utförs. Informant B beskriver även fuzzy logic som fundamentalt vid tillämpning av datatvätt. Vid fuzzy logic undersöks data från dess grad av sanning. Det som är speciellt med

denna gradering är att data kan bedömas delvis sann eller delvis falskt, vilket används när det är svårt att fastställa om vissa data är helt sann eller helt falsk beroende på kontexten.

> *”Ja fuzzy logic används till stor del. Det finns klassiska saker som att man plockar bort alla vokalerna ur ett namn och så gör man jämförelse på konsonantsidan för att få bättre träff, hit rate”*
> *(Informant B, 2009-05-06)*

Han beskriver vidare hur fuzzy logic används för att rätta till problem i variabler för att motverka felaktiga resultat av en analys.

Informant A hävdar att data kan bedömas som osäker då det saknas datavärden i variabler. När detta händer kan slutsatsen av en analys få en helt annan utgång än vad en analys med fullständig hade fått. En lösning till att ersätta saknade datavärden finns inte. När bristfällig data används för analysering kan detta leda till att fel slutsatser dras beroende på syftet med analysen och vem som utför den.

4.4.2 Kategorisering av data

Data finns lagrad i olika former, och i olika system. Informant C beskriver framför allt skillnaden mellan att använda sig av strukturerad och ostrukturerad data vid en dataanalys. Där den bästa och säkraste datan är strukturerad data i form av transaktionsdata, så som siffror och kvittorader. Den är både säker och håller en hög kvalitet. Ostrukturerad data, vilket vi beskrivit kort ovan består mer av mail, loggar osv. alltså data som är svårare att använda då den inte är bearbetad.

Informant C hävdar även att spårbarheten i data är ett viktigt inslag när data kvalitetssäkras och kategoriseras. Detta för att avgöra den ursprungliga källan där data har påträffats och hur de har uppkommit. Informant B förtydligar även detta och menar på att data kan kategoriseras utifrån dess ursprung. Den lägsta nivån består endast av enkla tupler eller rader i en databas. Dessa data används t.ex. inom detaljhandeln i så kallade basketanalyser där organisationen kan undersöka vilka varor som säljs tillsammans.

Nästa nivå består av aggregerad data där olika rader eller dataset kombineras från olika system, sammanställd data. Utan aggregering av dessa dataset kommer det förmodligen inte kunna gå att utläsa något från dessa data och därför är aggregeringen nödvändig. Kategorisering av data är därför en viktig del eftersom ursprunget uppdagas. Både hur data lagras inom systemet men även från vilken bransch data härstammar från.

Informant B beskriver även ett verktyg som kallas för ”Attribute importance” som kan användas för att rangordna attribut i data. Framförallt vill man klassificera de attribut som kan komma att påverka en analys mest, viss data kommer inte påverka analysen och dessa är man inte intresserade av. Sorteringen av data är viktigare ju större datamängderna blir och det är inte ofta all data som är relevant för analysen.

4.4.3 Kvalitet kontra kvantitet

Informanterna B och C hävdar båda att ett drömscenario är när det finns mycket data tillgänglig och kvaliteten samtidigt är så pass bra att en analytiker känner att resultatet av DM känns säkert. Informant B menar att kvaliteten nästan är ett större problem än kvantiteten och beskriver ett datawarehouse som exempel. Antag att data samlas in från en rad olika länder med en rad olika system. Hur man än bär sig åt kommer den insamlade datan vara av olika typ och format. Då är det enormt tidskrävande att sammanställa all dessa data i en och samma struktur. Med hjälp av bra ”tvättning” av data kan dock detta genomföras och sammanställd data kan lagras i ett datawarehouse. Sedan kan DM ske på dessa data för att söka efter kännetecken.

Informant B hävdar vidare att kvaliteten behöver höjas en hel del om data kommer från olika källor och därför bör datawarehouse användas i dessa fall. Handlar det istället om enklare DM, i form av basketanalys, kan istället kvantiteten vara att föredra. Informant C anser även att kvaliteten är den viktigaste länken till ett bra resultat av DM. Han anser, att hur fina modeller och bra verktyg man än har, så spelar det ingen roll om datakvaliteten är dålig.

4.5 Validering av mönster

De mönster som återfinns i data måste först analyseras innan de går att använda. Detta görs för att en specifik upptäckt ska vara relevant och säkerställd gentemot domänområdet.

4.5.1 Analys av resultatet

Det går att genomföra både manuell och automatisk analys av den produkt DM resulterar i. Alla informanter anser dock att i slutändan måste någon form av manuell analysering ske för att kunna säkerställa resultatet.

> *”Det går ju att sätta upp regelmotorer så att man kan automatisera en bit till. Men någonstans på slutet, måste man fråga sig ”Är det här rimligt?” (Informant B, 2009-05-06)*

Informant C anser att automatisering kan komma att bli mer populärt inom vissa områden eftersom efterfrågan på marknaden ökar, speciellt när det gäller ostrukturerad data från nätet. Han anser dock att som analytiker vill man hela tiden ha koll på processen och följa sin modell.

> *”Jag vill själv sitta i förarsätet. Jag anser att det är lite farligt att bara trycka på en knapp och sen kommer det ut något ur processen som man ska använda, men man har ingen aning varför resultatet blev som de blev” (Informant C, 2009-05-07)*

Informant C anser dock att vissa automatiserade funktioner kan förekomma för att hjälpa till i processen, men i det stora hela behövs manuell analysering för att kunna säkerställa ett bra resultat.

Informanterna hävdar även att i slutändan är det ofta det mänskliga omdömet som bestämmer resultatet av analysen. Subjektivitet i analyseringen av data kommer därför alltid att finnas kvar för att kunna säkerställa resultatet av en analys. Informant A menar att generellt sett så skapas konceptuella modeller som kan tolkas på olika sätt. Det kan både vara positivt eller negativt för resultatet och det är därför kvalitetssäkring alltid måste göras noggrant. Informant C är inne på ett liknande spår, men menar även att erfarenhet är minst lika viktigt. En mer erfaren person kan med sannolikhet konstruera en bättre analysmodell jämfört med en mindre erfaren person.

Informant B anser även han, att den subjektiva parametern kan både hjälpa och stjälpa resultatet. I slutändan måste ändå någon med verksamhetsförståelse analysera resultatet, men den personen kan ju också ha fel. Analysering av resultatet kan därför vara en knepig process och därför valideras de mönster som DM resulterar i, i två övergripande mått, nämligen relevans och säkerhet.

4.5.2 Relevanta mönster

Vilken information som är relevant och för vem, är ett ämne som till stor del berörs i KDD-processen. Informanterna hävdar att de mönster som skapas via DM bör överensstämma med vad som räknas som relevant i den domän där KDD-processen används. Informanterna B och C anser även att DM inte ens bör köras om det inte finns ett klarlagt problem eller syfte med processen. För att kunna säkerställa om resultatet som genereras via DM är relevant, beskriver informant A en typ av verktyg som kan användas för att uppnå detta. Ett sätt är att använda sig av en sorts transformationsbox som är kunskapsdriven. Data kommer på så sätt från källan, där den utvinns och passerar transformationsboxen på vägen mot målsystemet. Transformationsboxen består av olika filter som är programmerade för att upptäcka viss sorts data som är relevant för målsystemet. De data som inte är relevanta sorteras bort medan relevant data tas upp av målsystemet.

Informanterna B och C beskriver att testning av mönstren förekommer i stor utsträckning. Resultatet testas mot tidigare kända företeelser inom det specifika området, på så sätt går det att avgöra om mönstren är relevanta eller inte.

Samtliga informanter hävdar även att ostrukturerad data är ett relevant område för DM, men det kommer ta tid innan tekniken kan använda dessa data till fullo. Skälet är att det nästan finns för mycket data att tillgå och att de data som genereras ofta är irrelevant. Det mesta som genereras är chatloggar eller liknande där alldagliga konversationer äger rum mellan användare. DM bygger inte på att utforska denna typ av data, då den inte är relevant eller triviala på något sätt.

4.5.3 Säkra mönster

Ett annat ämne som berörs till stor del i KDD-processen är säker data, i den bemärkelsen att de mönster som genereras via DM måste vara sanna eller korrekta. Informanterna hävdar att det kan vara problematiskt att fastställa säkra mönster. Informant A menar att det beror på vilket område som berörs. Handlar det mer om informationskänsliga områden som t.ex. sjukhusjournaler, används olika verifieringsverktyg för att fastställa att informationen som utvunnits är korrekt. Mindre informationskänsliga områden, som t.ex. filmbranschen, kräver inte denna typ av verifiering. Där kan det istället handla om någon sorts verktyg som genererar feedback från användarna.

Informanterna B och C förespråkar testning av resultatet mot data som man redan vet är sanningsenlig. På så sätt kan informationen kontrolleras och sannolikheten om den är korrekt eller inte kan fastställas. Informant C relaterar även till den modell som DM utgår ifrån. Använder analytiker en redan beprövad modell där resultatet varit lämpligt vid ett tidigare projekt, går det att förlita sig till den modellen i ganska stor grad.

Experter inom området kan även analysera resultatet för att undersöka om det är sanningsenligt, om det t.ex. rör sig om avancerad kunskap. Informant C anser vidare att när modellen är färdigställd så krävs ett system som ska övervaka modellens giltighet. Detta eftersom domänområdet kan ha ändrats. När omgivningen förändras kommer modellen också att behöva göra det. Informant C hävdar vidare att testning av mönstrens resultat är en viktig del där man förlitar sig mycket på beprövade metoder och modeller. Visar det sig att en redan välanpassad modell får felaktiga utslag kan den behövas korrigeras. Olika branscher har även olika trösklar för säkerheten, då modellen behöver korrigeras, vilket kallas för modellens livscykel.

5 Analys & Diskussion

Avsnittet består av en analys av intervjuerna kopplat till vårt teoretiska underlag.
Vi har valt att tillämpa KDD-processen som ett analysverktyg där vi kommer att undersöka hur kvalitetsförbättring kan ske parallellt med processen.

5.1 Ökade kvantiteter av datamängder

1992 bedömde Frawley et al att informationen i världen fördubblades var 20:e månad. 2005 mätte WinterCorp upp en tredubbling av storleken på de största databaserna under en tvåårsperiod. Det visar på vilken lavinartad utveckling datamängderna i världen har. Informanterna B och C anser att det både finns positiva och negativa aspekter av denna trend. Där det finns större djup och bredd i data, kan analyser genomföras med ett säkrare underlag i data anser B. Informant B anser även att det finns negativa aspekter. När data finns lagrad på så många olika platser och när de sammanförs och sammanställs kan problem uppstå. Detta överrensstämmer med flera av forskarnas syn. De menar att de stora datamängderna är svåra att behandla eftersom det finns för mycket och för olik data tillgänglig (Chen et al, 1996; Cooley et al, 1997; Kosala & Blockeel, 2000).

En annan negativ aspekt är, som informant C påpekar, de ostrukturerade data som lagras i allt större utsträckning, t.ex. data från privatpersoner. Det är intressant att informant C påpekar detta, eftersom van Well & Royakkers (2004) har varit inne på sammas spår. De menar att det är svårt att säkerställa data som framtagits om och från en privatperson (se avsnitt 2.3.4). Denne bör helst samtycka att informationen är korrekt för att den ska kunna säkerställas. van Well & Royakkers (2004) anser dock att detta inte är ett problem i nuläget, men det kan bli i framtiden.

Med hjälp av informant C:s uttalanden finns det anledning att antaga att detta kanske är ett problem, då ostrukturerad data allt mer börjar analyseras och kommer troligen att öka i framtiden. Detta perspektiv är intressant eftersom dagens analyser behandlar strukturerad data till störst del, men i framtiden kommer troligen ostrukturerad data att analyseras i allt högre grad. Kontentan av detta är att analysen av data kommer förmodligen inte bara att handskas med större datamängder och mer olik data, utan strukturen för analys av mönstren kommer troligen även att förändras då nya parametrar kommer in i bilden, så som t.ex. integriteten.

5.2 Löpande kvalitetsarbete i KDD-processen

Resultatet från våra expertintervjuer kommer att analyseras utifrån vårt teoretiska underlag. KDD kommer här att användas som den övergripande modellen varvid vi kommer att visa hur varje enskilt steg påverkas av de resultat intervjuerna har påvisat.

5.2.1 Mål, syfte och problematisering

Samtliga informanter beskriver åtskilliga gånger hur viktigt det är att ha ett affärsproblem eller en uppgift att lösa med DM. Det handlar om att ett klarlagt syfte finns framtaget som stödjer strävan med det framtida arbetet så att man vet vilket mål som skall uppnås. Inom KDD betonas detta syfte med vilken typ av information som är relevant för processen och målet är att kunna skapa kunskap utifrån denna information. Informant A och C hävdar även att en stor del i arbete går ut på att skapa konceptuella modeller över domänområdet för att lättare just beskriva syfte, mål och hur processen bör fortlöpa.

Stora delar av vad informanterna berättar stämmer överens med vad (Fayyad et al, 1996) beskriver om första steget i KDD-processen (se avsnitt 2.3.1). De hävdar att det handlar om att skapa förståelse över domänområdet som arbetet kretsar runt, samt identifiera och bestämma målet med KDD. Frawley et al (1992) understryker detta, ty de mönster som påträffas via DM måste vara relevanta för den kontext de används inom (se avsnitt 2.3). Samtliga informanter poängterar även att arbetet med att passa algoritmer till KDD-processen, borde inte längre vara speciellt omfattande. Det finns ett stort antal färdiga algoritmer som lätt går att passa till processen. Det som är viktigt är syftet med processen. Denna syn på algoritmernas inverkan skiljer sig från Fayyad et al (1996) beskrivning. De menar att algoritmerna är en viktig del i arbetet.

Algoritmerna har dock alltid varit en viktig del, men allt eftersom DM-tekniken har förändrats så har algoritmerna ofta bestått. Detta eftersom de består av matematiska formler och förändras inte i lika stor utsträckning.

5.2.2 Selection

Det andra steget i KDD-processen, selection, består av att välja data som är relevant för det valda domänområdet. Fayyad et al (1996) beskriver hur val och fokusering av utvald data sker och därefter skapas datatabeller som kan kopplas till det relevanta området. Vilket förbereder processen för det tredje steget, preprocessing, där irrelevant och överflödig data rensas bort (se avsnitt 2.3).

Innan urval av data sker behöver någon typ av kategorisering av data ske. Dels för att fastställa vilket ursprung data har men även vilken form den är lagrad i.
Informant C beskriver framförallt skillnaden att använda sig av strukturerad eller ostrukturerad data. Skillnaden på typ av analys skiljer sig markant åt beroende på om data är strukturerad eller ostrukturerad. Därför behöver strukturen fastställas tidigt i processen. Detta överrensstämmer med Batini & Scannapieca (2006) och deras syn på kategorisering av data. De anser att data först måste specificeras som antingen strukturerad, semistrukturerad eller ostrukturerad innan kvaliteten kan höjas (se avsnitt 2.2.1).

Informant B påpekar även betydelsen av ursprunget, varifrån data har påträffats och hur de lagras. Dessa steg anser vi faller bort något i stora delar av teorin kring KDD, där fokuset under andra steget ligger på att välja data som ska användas i processen. Men vi anser att det borde särskiljas vilken typ av data processen handskas med, eftersom analysen skiljer sig

beroende på vilken data som används. Vi anser att någon form at kategorisering av data kunde ske under denna delprocess, alternativt under den första delprocessen. Anledningen till detta är att fastställa vilken typ av data som används.

5.2.3 Preprocessing

Maletic & Marcus (2000) anser att processen bör försöka sträva efter att upptäcka och reparera data som inte är sammanhängande eller är felaktig så att den blir fullständig. Rahm & Do (2000) hävdar istället att de data som påträffas som felaktig eller icke sammanhängande istället borde raderas (se avsnitt 2.3.2). Åsikter om preprosessing och arbetet med att förbättra datakvalitet skiljer sig alltså ganska mycket beroende på vem som tillfrågas. Vi är därför av den uppfattningen att det beror mycket på erfarenheter och vilken typ av verktyg som används.

Vissa skiljaktigheter återfanns även hos våra informanter. Informant C anser att modellen ska fungera som ett underlag vid datatvätt men menar även att olika typ av verktyg även behöver användas för att optimera datakvaliteten. Informant B hävdar också att olika verktyg och hjälpmedel bör användas för att tvätta data. Han förespråkar även användningen av datawarehouse, där är datakvaliteten redan är säkerställd, vilket förbättrar datakvaliteten avsevärt. Vi tror att skillnaderna har mycket att göra med vilken bransch och vilken yrkesgrupp som tillfrågas. Dock är informanterna och flertalet av forskarna inom området eniga om att datakvalitet är ett viktigt inslag för att uppnå mer kvalitetsinriktade mönster av DM.

Informant A har dock en annan syn på det andra (selection) och det tredje steget (preprocessing) i KDD-processen. Han hävdar att med hjälp av den konceptuella modell som skapats i början av KDD-processen kan steg två och tre flyta samman på ett smidigare sätt för att effektivisera processens slutprodukt. Han beskriver hur en modell för relevant data framställs. I denna modell finns även komponenter som säkerställer att endast relevant information utvinns. Dessa komponenter kan bestå av transformationsboxar som är kunskapsdrivna och ser till att endast relevant data kan passera igenom boxen på vägen från källan till målet.

Strong et al (1997) anser även att datakvalitetsarbetet är en viktig aktivitet och hänvisar till deras fyra olika dimensioner av data. Detta är dock mer en bedömning av datakvaliteten och alltså varken en kategorisering eller en typ av datatvätt (se avsnitt 2.2.2). Bedömning av datadimensioner skulle därför kanske placeras mellan selection och preprocessing i KDD. Detta för att kunna undersöka kvaliteten på de data som kategoriserats i selection från olika dimensioner. Data undersöks därmed från olika synvinklar och på så sätt går det att urskilja vilken data som kan vara bristande för den aktuella KDD-processen. Skulle det visa sig att kvaliteten är bristande tas data bort eller repareras under nästa steg, preprocessing. Vi presenterar ett förslag på detta i avsnitt 6.1.2.

5.2.4 Transformation och val av modell

Delprocesserna transformation och val av modell är två delprocesser som består av relativt standardiserat arbete. Under transformationen reduceras de data som redan undersökts i preprocessing. Val av modell är dock ett viktigt inslag, men det beror mycket på syfte med processen och vilken typ av data som används. Det mesta av arbetet är därför redan utfört och en modell som kommer att användas är ofta klar sen tidigare.

5.2.5 Val av algoritm

Frawley et al (1992) beskriver vikten av att passa algoritmer till KDD-processen (se avsnitt 2.3.4). Samtliga informanter hävdar också detta, men tycker inte att det lika viktigt i dagens läge, ty det finns så pass många färdiga algoritmer att använda och testa data mot. Algoritmerna har även varit i stort sett oförändrade de senaste åren eftersom analytikerna vet att de fungerar. Informant A förklarar att det viktiga är att modellen över området presenteras ingående, när syftet är klarlagt är det enkelt att passa färdiga algoritmer till processen. Val av algoritm är ingen svår process.

5.2.6 DM

Frawley et al (1992) anser att de mönster som bildas via DM ska beskriva relationer i delmängder av data och senare fungera som kunskapsbärare. Flertalet forskare förespråkar både förutsägande och beskrivande DM (Fayyed et al 1996; Frawley et al 1992; Goebel & Gruenwald, 1999).

Informanterna B och C anser att DM inte bör bedrivas om det inte finns något problem att lösa. Resonemanget som då kan föras är att B och C anser att förutsägande DM är den typen som bör bedrivas. Det är nämligen då som data eftersöks för att finna framtida kunskap och det finns ett klarlagt syfte med processen. Under beskrivande DM är syftet inte klarlagt och därför kan resultatet bli varierande. Informant A anser istället att både förutsägande – och beskrivande DM kan användas beroende på hur information eftersöks. Han anser vidare att det är viktigt att beskriva processens input respektive output, där beskrivande DM kan fungera som input, och förutsägande DM som output i en transformation.

Detta kan kopplas till Goebel & Gruenwald (1999) och deras beskrivning av beskrivande jämfört med förutsägande DM. De tar bland annat upp kluster som en väletablerad metod inom beskrivande DM för att dela upp klasser i mindre partitioner med hjälp av återkommande kännetecken. De behandlar även klassifikation vilket är en typ av förutsägande DM, där klasser skapas och sedan söker igenom datamängder för att förutsäga var data hör hemma inom de olika klasserna. Både Goebel & Gruenwald (1999) och Fayeed et al (1996) behandlar beskrivande och förutsägande DM på ett liknande sätt, vilket kan vara krångligt att förstå (se avsnitt 2.3.1). Genom att istället tänka sig en transformation med input och output tar informant A resonemanget till en ny nivå och beskriver förekomsterna på ett nytt och smart sätt som vi inte tänk på tidigare.

Det är intressant att stora delar av litteraturen och informant A förespråkar både förutsägande – och beskrivande DM. Informanterna B och C anser däremot att endast förutsägande DM bör användas eftersom det alltid måste finnas ett klarlagt syfte med DM. De hävdar framförallt att kostnaderna är stora för att utföra DM och det är inte hållbart att söka efter kännetecken utan att ha ett specifikt mål uppsatt. Vad som dock är intressant är att beskrivande DM förmodligen kommer att upptäcka mer triviala mönster. Eftersom det inte är klarlagt vad som egentligen eftersöks, utan processen söker själv igenom data efter återkommande kännetecken. Samtidigt kan det vara ett syfte eller mål med KDD, att söka efter okända relationer i data.

Förutsägande kontra beskrivande DM kan därför diskuteras i stor utsträckning och vi anser att det är upp till företaget, analytikern eller processens syfte att bestämma vad som bör användas.

5.2.7 Interpretation/evaluation

I den sista delprocessen i KDD analyseras de mönster som påträffats i data. Analys av resultatet är en omfattande process där mönstren måste vara relevanta (användbara) och uppnå ett validerat resultat (säkra). Analysen skiljer sig även beroende på om objektiva eller subjektiva mått av resultatet används.

Siberschartz & Tuzillin (1996) anser att objektiva mått används när analysen av ett resultat sker av redan tidigare känd data (se avsnitt 2.3.3). De subjektiva måtten är istället när det mänskliga omdömet används. Detta kan kopplas till de resonemang informant B och C förespråkat kring automatiskt och manuell analysering av resultatet. Där ett objektivt mått liknar mer en automatisk analysering, och där manuellt mått liknar mer en subjektiv.

Samtliga informanter anser att automatisk analysering kan ske till viss del, men i slutändan bör det mänskliga omdömet bestämma om resultatet är relevant och korrekt.

5.2.8 Sammanfattning

KDD-processen har en stomme med ett antal inslag som på ett eller annat sätt behöver beaktas för att utföra DM. Även om processen ser annorlunda ut och de olika stegen har olika namn, så finns tankesättet från niostegsmodellen alltid i bakgrunden. Det handlar om att först skapa sig en överblick över området, kategorisera och välja data, undersöka datakvaliteten, sammanställa data, analysera och skapa mönster av data och till sist analysera och säkerställa resultatet.

Modeller är ett inslag som blivit allt vanligare i processen och kan skilja sig åt i vissa av stegen. Informanterna hävdar dock att tankesättet i grunden alltid är av liknande slag och att det snarare är tekniken som förändrats de senaste åren. En anledning till detta är att olika syften med DM tagits fram. Fayyad et al (1996) hävdar att från begynnelsen förknippades DM med ett verktyg som användes i databaser för att söka efter relevant information (se avsnitt 2.3). Nuförtiden återfinns DM inom en rad olika områden.

Informant A beskriver att deras företag för tillfället använder sig av DM i mobiltelefoner för att samla information. Informant C talar även om video och voice mining, där bild och ljud alltså ska kunna analyseras. Tekniken och metoderna kommer därför säkerligen att behöva förändras i och med att nya appliceringsområden utforskas. Informant B och C hävdar även att ostrukturerad data och speciellt Internet säkerligen kommer att bli ett allt vanligare område att utföra mining på i fortsättningen.

Ett inslag som vi inte tycker har behandlats tillräckligt i litteraturen vi använt är datakvaliteten. Både informant B och C anser att bra datakvalitet är den mest betydande faktorn för att uppnå ett bra resultat av DM. Som analytiker kan man ha tillgång till de bästa verktygen som finns tillgängliga, men är datakvaliteten dålig så blir resultatet opålitligt (Informant C, 2009-05-07) (se bilaga 4). Det finns förvisso inslag av datakvalitetsarbetet i KDD, i form av delprocessen preprocessing.

Vi anser dock att arbetet med att förebygga problem med datakvaliteten borde vara löpande genom hela processen och introduceras tidigt för att eliminera problem vid ett senare skede.Kontentan av resonemanget är därför att DM, precis som de flesta andra tekniker, går framåt i utvecklingen eftersom större krav ställs på tekniken och metoder i och med att användningsområdet blir större.

5.3 Relevanta mönster

Frawley et al (1992) hänvisar till vissa kriterier som används för att fastställa hur ett relevant mönster har upptäckts (se avsnitt 2.3.3.). Dessa kriterier behandlades inte nämnvärt hos informanterna. Informanterna B och C anser istället att vad som är ett relevant mönster beror på domänområde. Det måste finnas ett bakomliggande syfte till att använda sig av DM, t.ex. ett verksamhetsproblem. B och C anser även att det är väldigt dyrt att leta efter mönster i stora datamängder och därför är det också viktigt med ett klarlagt syfte.

Domänområdet fastställs redan i början av processen för att undersöka de attribut som berör området. I detta skede anser vi att Frawley et al (1992) och deras ena kriterium för relevanta mönster är lämpligt att använda (se avsnitt 2.3.3). Detta kriterium bestämmer användbarheten i mönstret och vi anser att användbarheten av mönstret kan säkerställas då domänområde har blivit specificerat. På så sätt går det att urskilja vilka mönster som kommer att vara av stor betydelse för domänområdet.

Informanterna B och C hävdar att det ibland förekommer verksamhetsexperter som granskar de mönster som utvunnits för att bedöma hur relevanta mönstren är. Detta anser vi kan överensstämma med Frawleys et al (1992) och deras kriterier för icke triviala, ovanliga och användbara mönster. Eftersom en expert har verksamhetskompetens inom det område som mönstret berör kan denne på så sätt bestämma om mönstret är icke trivialt, ovanligt och användbart.

Informanterna B och C poängterar också att det förekommer testning av de utvunna mönstren. De anser att dessa bör testas mot data där kända företeelser redan förekommer.

På så sätt menar informanterna B och C att mönstrets relevans kan säkerställas. Detta kan ses som de objektiva mått på relevanta mönster som Siberschartz & Tuzilin (1996) beskriver (se avsnitt 2.3.3). Vilket just innebär att data testas mot redan tidigare känd data inom det specifika området. Relationer bland data mäts även med olika typer av associationsregler. Informant B hävdar att det går att implementera en typ av regelmotorer i processen för att beakta detta. Dock används det främst inom automatisk analysering av resultatet.

Samtliga informanterna hävdar att i slutändan är det ofta det mänskliga omdömet som bestämmer om resultatet av analysen är relevant eller inte. Det kan leda till att analysen blir annorlunda beroende på vem som utför den. Detta anser vi överrensstämmer med Siberschartz & Tuzilin (1996) och deras syn på subjektiva mått av resultatet (se avsnitt 2.3.3). Subjektivitet i analyseringen av data kommer därför alltid att vara en viktig parameter som kan problematisera resultatet. Det är något man helt enkelt får finna sig i, alla människor är olika och uppfattar saker på olika sätt.

5.4 Säkra mönster

Som tidigare nämnts behöver de mönster som framställs ur data vara relevanta för den kontexten de ska användas inom. Det är dock lika viktigt att fastställa att de mönster som framställts är korrekta och sanningsenliga.

Informant A beskriver även att ett av de främsta skälen till att data bedöms som osäker är att det saknas datavärden i variabler. När detta händer kan slutsatsen av en analys få ett annat resultat än vad en analys med fullständiga data hade fått. Frawley et al (1992) hävdar att data måste kunna säkerställas. Annars blir mönstren oberättigade och kunskap kan då inte genereras (se avsnitt 2.3.4).

Informant A anser även att det kan vara problematiskt att bestämma om informationen som utvunnits verkligen är korrekt. Hur pass känsligt det är med missvisande information beror främst på området som behandlas.

Detta uttalande stämmer väl överens med vad Frawley et al (1992) anser gällande säker data (se avsnitt 2.3.4). Dock ser förfarandet för att fastställa säker data lite annorlunda ut. Frawley et al (1992) menar att det finns olika faktorer i data som bör undersökas för att säkerställa dess äkthet. Informant A hävdar att olika typer av verifieringar bör genomföras för att säkerställa att data är korrekt. Informanterna B och C anser istället att testning mot tidigare känd data inom det aktuella området används i stor utsträckning. Alternativt kan även experter inom området analysera resultatet om det handlar om riktigt områdesspecifik information. Det sistnämnda anser vi överenstämmer med Frawley et al (1992) förslag om att kontrollera hur bra upptäckten stämmer överens med tidigare kända kunskaper (se avsnitt 2.3.4). Experter kan verifiera den nya upptäckten mot sin kompetens och avgöra om mönstret är säkert.

Det kan dock visa sig att det är ny kunskap som påträffas där det inte finns några experter. På så sätt blir upptäckten svår att säkerställa, men samtidigt behöver kunskapen inte

säkerställas mot redan känd kunskap inom området eftersom det inte existerar. Kunskapen bör dock säkerställas i det hänseende att den är relevant och användbar.

Ibland är det ändå svårt att fastställa hur pass säker data är. van Well & Royakkers (2004) hävdar att privatpersoner ofta inte får chansen att samtycka med den information som lagras om dem på nätet (se avsnitt 2.3.4). När personen inte kan godkänna den information som samlats in kan felaktiga slutsatser dras och informationen blir osäker. Detta är ett problem som det inte finns några bra lösningar på eftersom stora mängder data ständigt samlas in på Internet och det är sällan användarna får samtycka till denna information. Dessa problem uppstår i huvudsak på Internet men kan även återfinnas i andra medium. Viktigt att påpeka är dock att van Well & Royakkers (2004) understryker att detta inte är ett problem i nuet eftersom DM för tillfället inte används till sin fulla potential. När tekniken börjar utnyttjas än mer i framtiden och med hjälp av Internets ständiga förändring kommer troligen denna typ av problem att bli vanligare.
Enligt informant C blir DM av ostrukturerad data, främst Internet allt vanligare. Detta kan leda till allt större problem vid säkerställandet av datamönster.

Säkra mönster är ett problematiskt område och på sikt finns det inte några garantier att all information som sammanställs från data verkligen kan märkas med en säkerhetsstämpel. Problem vid användningen av denna information uppstår om validiteten inte kan styrkas .

Säkra mönster är också ett komplext område och därför tror vi att både kontrollera data och utveckla förfaranden för hur den ska kontrolleras, är viktigt. Vi har därför valt se på teorin och informanternas uttalanden som ett sammansatt koncept. Först bestäms vilken typ av verifiering som ska utföras och därefter kontrolleras data efter dessa beskrivningar. Eftersom varje situation och information är unik behövs en väletablerad process för att detta förfarande ska fungera.

5.5 Sammanställning av relevanta och säkra mönster

Vi anser att verifieringen av relevanta och säkra mönster överlappar varandra till stor del. Det går att urskilja ett klart samband mellan teknikerna som används för att fastställa att ett relevant och säkert mönster har upptäckts. Vi anser dock att båda av dessa faktorer är viktiga för att verifiera ett mönster. Vi vill dock poängtera att ett säkert mönster är ännu viktigare än ett relevant. Detta eftersom ett icke säkert mönster kan som Frawley et al (1992) beskriver det leda till utebliven eller ”fel” kunskap. Utebliven kunskap kan i sin tur leda till att affärsmöjlighet försvåras eller som informanterna hävdar, att felaktigt beslut kan fattas.

5.6 Kognitivt beteende

Pazzani (2000) hävdar att kognitiv psykologi borde vara ett inslag i KDD-processen eftersom kunskap är nära relaterat till mänskligt beslutsfattande (se avsnitt 2.3.5). Precis som Pazzani (2000) hävdar även informant A att psykologi och speciellt kognitiv psykologi har stor inverkan i KDD-processen. Han beskriver att kunskap om människors beteende och hur de tänker kan dels användas för att framställa mer intelligenta system, men också för att förstå hur människorna integrerar med systemen.

Både Pazzani (2000) och informanterna lämnar dock stora delar av sambandet mellan DM och kognitiva beteenden åt framtiden. Pazzani (2000) hävdar att det finns mycket information kring mänskligt lärande och psykologi som kan knytas till KDD-processen och att det borde vara ett intressant område att utforska. Informant C tror också att detta är något som kommer att beröras mer i framtiden och speciellt när ostrukturerad data börjar analyseras i allt högre grad.

Informant A behandlar istället ett antal problem som det kognitiva införandet kan medföra. Bland annat att stora delar av de system som finns idag inte är anpassade för införandet av intelligentare system. Han nämner Internet som ett område där stora förändringar behöver ske innan DM-tekniken kan utnyttjas till full kapacitet.

Informanterna B och C anser båda att datakvaliteten kan komma att förbättras genom kognitivt beteende, speciellt ostrukturerad data. Den kognitiva inverkan är därför ett relevant inslag som kan komma att förbättra hela processen av statistiska dataanalyser på sikt.

6 Slutsats

Kapitlets avsikt är att presentera de slutsatser vi fastställt relaterat till vårt problemområde och syfte med uppsatsen. Vi ger även förslag till vidare forskning inom området.

När vår undersökning började utgick vi från att datamängderna i världen ständigt växer. På så sätt bör det finnas mer data tillgänglig för organisationer att använda sig av, både i det interna arbetet och i arbetet med kunder. Vi antog även att data finns lagrad i många olika medium och att kvaliteten på data är mycket varierande. Därför ansåg vi att en undersökning relaterat till hur datakvaliteten påverkade resultatet av en dataanalys var relevant.

Den grundläggande slutsatsen av detta resonemang är att datakvaliteten är oerhört viktig för en relevant, säker och informativ dataanalys. Är kvaliteten bristande kan resultatet av en analys bli en helt annan än om data med bra kvalitet användes. Det är därför viktigt med bra datakvalitet. Men hur vet man då vad bra datakvalitet är?

6.1 Faktorer för att bedöma datakvalitet

Vi har kunnat fastställa ett antal faktorer som är betydelsefulla när det handlar om att bedöma datakvalitet. Dessa är:

- Vilken bransch handlar det om?
- Förändringar i domänområdet
- Strukturerad kontra ostrukturerad data
- Aggregerad eller icke aggregerad data
- Dimensioner i data
- Bredden och djupet på data

Branschspecifika bedömningar riktar sig till hur pass känslig data är inom olika områden. Datakvaliteten vid kategoriseringar av sjukdomar via sjukhusjournaler är känsligare, jämfört med t ex speltiden på DVD filmer. Branschen eller domänområdet där analysen sker är viktig att ta hänsyn till. Det kan även ske förändringar i domänområdet som leder till att resultatet av en analys skiljer sig från de mål som fastställs i början av analysen.

Huruvida data är strukturerad eller ostrukturerad avspeglas även i kvaliteten. Strukturerad data, som transaktionsdata, siffror eller kvittorader är lättare att kvalitetsgranska och på så sätt lättare att använda sig av. Ostrukturerad data kan dock komma att bli mer relevant och lättare att kvalitetssäkra på sikt.

Det kognitiva perspektivet och en strävan efter att göra mer intelligenta system bedömer vi kommer att vara viktiga inslag i framtida kvalitetsåtgärder. Det är även en fördel att använda aggregerad data som i kort beskrivs som sammanställd data. Det kan t.ex. vara data som är sammanställd från olika system, med andra ord redan bearbetad data. Dessa data är oftast mer kvalitetssäkra än icke aggregerad data. När man använder aggregerad data bör man försäkra sig om kvalitetssäkerheten i den ursprungliga datan.

Dimensionerna i data är ett komplext område och beskrivs i litteraturen som ett av de grundläggande sätten att bedöma datakvalitet. Vi har dock valt att presentera det som ett inslag i våra kriterier eftersom vi anser att det inte endast data som behöver undersökas, utan även domänen, subjektivitet m.m.

Mängden av data som finns tillgänglig i form djup och bredd är också en bidragande faktor till kvalitetshöjning. Med djup menas att data finns lagrad flera år tillbaka och bredd refererar till hur stort överensstämmande det finns i data från andra källor. Viktigt är dock att poängtera att stora datamängder i form av djup och bredd även kan sänka datakvaliteten. Här kommer den subjektiva parametern och styrkan att arbeta med en bra modell in i bilden. Vi har därför reviderat KDD-processen och modifierat modellen för att visa på vilka förebyggande åtgärder som krävs för att förbättra resultatet av en dataanalys. Eftersom vi har arbetat parallellt med KDD-processen genom stor del av vår undersökning, ter det sig naturligt att vi presenterar våra förebyggande åtgärder som en modifiering av denna process.

6.2 Reviderad modell av KDD-processen

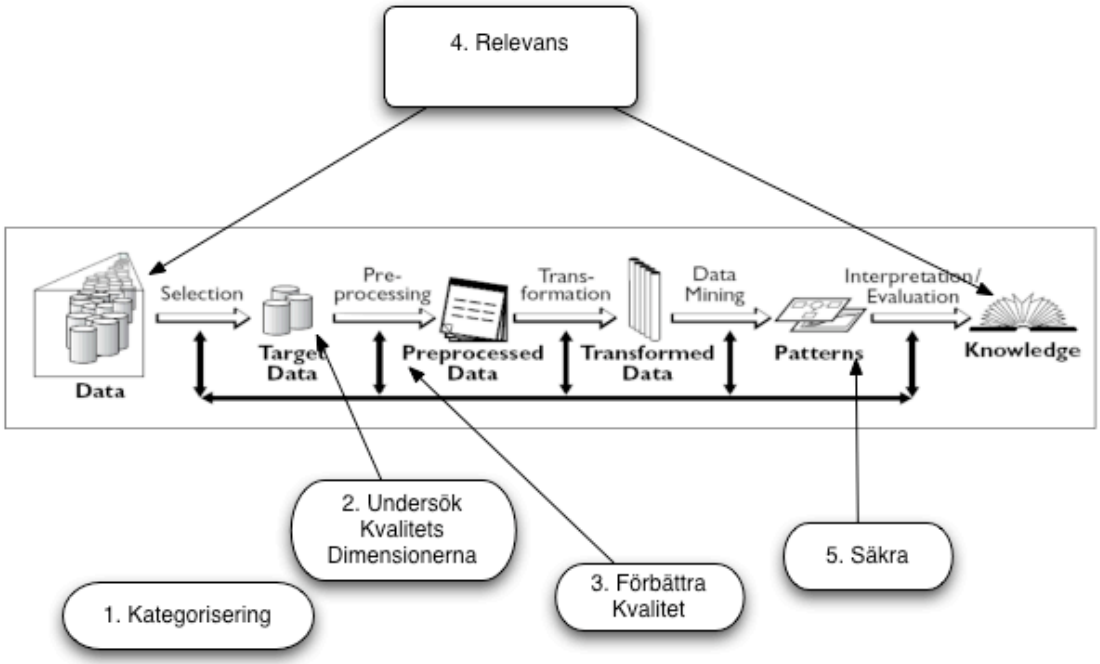
Modellen (återfinns i figur 6.2) utgår från nödvändigheter av att kategorisera data och förbättra datakvaliteten. Kategoriseringen av data (1) fastställer vilken typ av data som processen använder sig av. Detta är viktigt för att kunna spåra data till dess ursprung och kontrollera hur data sammanställts. Då data först kategoriserats bör kvalitetsarbetet bli enklare eftersom typ av data beskrivits. Det blir på så sätt även enklare att applicera rätt verktyg för att reparera, förbättra eller radera data under delprocessen *preprocessing*.

Vi har även föreslagit en ny delprocess, undersökning av datakvalitetsdimension (2). Syftet med delprocessen är att bedöma kvaliteten på de data som kategoriserats i delprocessen *selection*. Skälet till att bedömning av kvaliteten sker här och inte innan processen börjar är att dimensionerna relaterar till härkomst och typer av data. Vi tror därför att det är viktigt att först välja och kategorisera data innan kvaliteten bedöms. På så sätt kan det bli enklare att reparera eller tvätta data under nästa delprocess, *preprocessing*.

Kvaliteten förbättras (3) genom arbete med olika verktyg för att förbättra de data som kommer att användas i processen. Skälet till att vi poängterar denna aktivitet är att vi bedömer att kvalitetstänkandet inte får tillräckligt stor plats i den ursprungliga modellen (figur 2.2). I den verkliga modellen präglas denna aktivitet till stor del av att antingen reparera eller radera data. Det är förvisso kvalitetsförbättrande åtgärder, men vi anser att en bakomliggande strävan efter att förbättra kvaliteten saknas. Kvaliteten förbättras endast för att felaktigheter uppenbaras, något som görs för att det måste göras. Viktigt att poängtera är att vårt resonemang bygger på att data kategoriseras i ett tidigare skede av processen, vilket är beskrivet ovan (se 1).

Syftet styr datarelevansen (4), som är tvåriktad och relaterar både till att relevant data används, men även att intressanta eller relevanta mönster bildas i analysen. Detta är viktigt eftersom relevanta data för den specifika processen bör användas, likaså måste även resultatet av processen vara relevant i förhållande till processens ursprungliga syfte.

De mönster som påträffas behöver även vara säkra (5) i det avseende att resultatet är riktigt och kan användas. Skälet till att vi vill belysa denna aktivitet är dels för att det är ett viktigt inslag vid dagens DM, men även för att vi tror att det kommer bli allt viktigare i framtiden då DM kommer att användas inom nya områden. Det kommer då troligen även att bli svårare att fastställa att ett säkert mönster påträffats eftersom osäker och ostrukturerad data förmodligen kommer att användas i en större utsträckning.



Figur 6.2 Reviderad modell av KDD-processen

De inslag som vi framförallt tittat på i modellen, är kategoriseringen av data, arbetet med att förebygga datakvaliteten tidigt i processen samt att undersöka kvalitetsdimensionen. Detta är tre omfattande faktorer som vi inte anser får tillräckligt stor plats i kvalitetsarbetet. Genom att beakta dessa faktorer och arbeta med högre kvalitetsinriktning av data i KDD tror vi att mer relevanta och säkra mönster ska kunna framställas.

Bilaga 1

Förkortningar och Ordförklaring

AI – Artificial Intelligence

BI - Business Intelligence

DBMS – Database Management Systems

DM - Data Mining

DIKV – Data, Information, Kunskap och vishet

IS – Information System

KDD – Knowledge Discovery in Database

KM – Knowledge Management

MDI – Människa Data Interaktion

MIS – Management Information System

TB – Terabyte

Attribut – Kan också kallas för variabel i visa sammanhang. Ett Attribut beskriver en egenskap hos en instans. Attribut förekommer i objekt orienterade programspråk och i relationsdatabaser.

Data warehouse – En eller flera databaser sammanlänkade, där databaserna delar en gemensam struktur. För att migrera data till ett Data warehouse måste data oftast tvättas och transformeras för att pass den nya strukturen

Datakvalitetsdimension – Vissa författare kallar även dimensionerna för egenskaper. Datakvalitetens olika dimensioner bedömer datakvalitet från olika perspektiv. Vid bedömning av datakvalitet är det viktigt att välja de mest relevanta dimensionerna för det domänområde där data ska tillämpas.

Datakvalitetsproblem – Då det förekommer olika fel i en eller flera datakvalitetsdimensioner pratar man oftast om ett datakvalitetsproblem.

Mönster – En serie av återupprepande element i data.

Tupel – Kan också kallas för post. En tupel representeras i en relationsdatabas i form av tabeller, kolumner och rader. En tupel består av flera attribut med definierade värden.

Bilaga 2

Intervjuguide 1

- Vad skiljer KDD-processen åt inom olika domänområden?
- Hur utvecklas nya DM metoder?
- Vill man främst förutsäga framtiden(predictive) med DM eller vill man hellre komma på nya upptäckter om personer/grupper(descriptive)?
- Hur skiljer sig beskrivande(descriptive) och förutsägande(predictive) DM åt i ett givet sammanhang?
- Hur garanteras att den information som utvunnits via DM är intressant?
- Hur garanteras att den information som genererats via DM är sann eller korrekt?
- Blir kunskapen mindre subjektiv när information analyseras med automatisk analysering gentemot manuell analysering? Hur vet man att kunskapen som man utvinner är äkta?
- Hur kan kunskap om människors beteenden vara till nytta för olika tekniker, modeller, metoder av DM?
- Hur ser framtiden inom data minig ut relaterat till dagens syn, relaterat till förbättringar kring relevanta och säkra mönster?

Intervjuguide 2

- Vad vi förstått växer ständigt kvantiteterna av data inom olika IS. Hur uppfattar du detta påstående?

- Är det positivt eller negativt med stora datakvantiteter vid utförande av DM? Isåfall varför?

- Vi har uppfattat att organisationer förlorar både resurser och pengar på att använda sig av data med bristande kvalitet. Hur ser du på detta?

- Blir metoder och tekniker inom BI allt effektivare på att möta organisationers krav relaterat till kvalitativa och säkra resultat? Hur har detta förändrats de senaste 5 åren?

- Finns det någon kvalificering för att data ska lämpa sig speciellt för DM? T.ex nivåer, eller rangordning av data.

- Förekommer de någon typ av klassificering av de data som används vid data minig? Isåfall beskriv mer.

- Finns det olika faktorer i data som bör undersökas för att fastställa dess äkthet? Isåfall vilka?

- Finns det någon typ av verktyg som används för att bedöma kvaliteten på de data som används vid DM?

- Hur skiljer sig datakvaliteten inom olika områden?

- Hur upplever du att datakvalitet påverkar skapande av mönster och i sin tur kunskapsutvinning?

- Hur garanteras att den information som utvunnits via DM är intressant?

- Hur garanteras att den information som genererats via DM är sann eller korrekt?

- Vad anser du används mest, manuell eller automatisk analysering av resultatet som framställs via DM?

- Blir kunskapen mindre subjektiv när information analyseras med automatisk analysering gentemot manuell analysering? Hur vet man att kunskapen som man utvinner är äkta?

- Hur kan kunskap om människors beteenden, det kognitiva perspektivet, vara till nytta för olika tekniker, modeller, metoder av DM?

- Hur ser framtiden inom data minig ut relaterat till dagens syn, relaterat till förbättringar kring relevanta och säkra mönster?

Bilaga 3

Sammanställning – Intervjuföretag A

Vad skiljer KDD-processen inom olika domänområden?

Alla situationer är olika, det är viktigt att förstå syftet med den information som vill nås. Börja med att reda ut om det är input eller output som är intressant. Om det t.ex. är outputen som är intressant kan det vara bra att skapa sig en föreställning över vad man vill få fram och sen iterera processen till den förutsedda informationen genererats. Är det istället input som är intressant behöver mönstren kunna uttrycka syftet med processen.

Hur utvecklar man nya DM metoder?

Ett sätt för att utvecklar nya DM metoder är att koncentrera sig på att skapa konceptuella modeller över det området som man är intresserad av att samla in data från.
Valet av algoritmer är inte det viktiga, det finns mängder med färdiga algoritmer, vissa är snabbare än andra men i grund och botten är det viktigt att fokusera sig på den modell man skapar och se till att de variabler som ligger i grunden för analysen kommer med.

Vill man främst förutsäga framtiden(predictive) med DM eller vill man hellre komma på nya upptäckter om personer/grupper(descriptive)?

Det beror främst på vad man vill eftersöka eller uppnå med DM. Det gäller att sätta saker i kontext till andra, vilka syften man vill uppnå med DM, vad man vill få fram osv. Inom förutsägande DM handlar om att förstå och få insikt i en naturlig process. Detta kan handla om att förstå sammanhanget och situationen inom det område där DM bedrivs. När man förstår kontexten kan olika algoritmer sedan användas för att utvinna data. Inom beskrivande DM handlar det istället om att upptäcka nya mönster och ny information där systemet ständigt söker efter dessa upptäckter.

Hur skiljer sig beskrivande(despriptive) och förutsägande(predictive) DM åt i ett givet sammanhang?

I en given kontext kan beskrivande DM fungera som en input, där en transformation genomförs och söker igenom data efter återkommande kännetecken. Dessa kännetecken blir sedan processens output. Förutsägande DM är tvärt om. Där finns en input i form av syfte och beskrivning av vad man vill uppnå med DM. Det som sedan genereras utifrån detta mål, outputen, är förutsägande DM.

Hur kan man avgöra att den information som utvunnits via DM är intressant?

Det beror mycket på det område där DM bedrivs, vad som är intressant eller inte för det specifika området. Det finns vissa verktyg som kan appliceras i DM processen som säkerställer att den information som genereras är intresserat. Data genomgår en transformation på vägen från källan till målsystemet. Detta sker i ett sorts filter eller transformationsbox som är kunskapsdriven. Syftet med detta är att sortera igenom de data som genereras från källan och endast ta tillvara på de data som är intressant för målsystemet.

Hur ser framtiden inom data minig ut relaterat till dagens syn?

Det finns en tydlig tendens som visar på att framtiden kommer bestå av mer intelligenta och smartare tekniker(AI). Den typen av tänk används redan idag men det är för tillfället många gamla system, som hindrar utvecklingen. För att kunna dra nytta av nya metoder av DM behöver appliceringsområdet kunna stöda dessa metoder. Ett exempel är webben som för tillfället är mitt uppe i webb 2.0 hysterin. Inom AI pratar vi redan om webb 4.0 då intelligens krav kommer att finnas på systemen. Innan dess måste vi dock passera webb 3.0, den semantiska webben, där webben förstår och uppfyller användarnas önskemål bättre.

Kan man säga att kunskapen blir mindre subjektiv när man analyserar informationen med hjälp DM gentemot manuell analysering? Hur vet man att kunskapen som man utvinner är äkta?

Subjektivitet i analyseringen av data kommer alltid att finnas kvar. Konceptuella modeller skapas och kan tolkas på olika sätt. När det saknas datavärden i variabler kan slutsatsen av en analys bli en helt annan än vad en analys med fullständiga data skulle ha gett. En direkt lösning till att ersätta det saknade värdet finns inte. När bristfällig data används för analysering kan detta leda till fel slutsatser beroende på syfte som analysen har och vem som gör analysen.

Hur garanteras att den information som genererats via DM är sann eller korrekt?

Det beror mycket på systemområdet. Handlar det t.ex. om system som behandlar känslig information som sjukhus, används olika typer av verifiering för att kontrollera att den information som samlats in är riktig. I mindre känsliga områden som t.ex. filmindustrin kan det räcka med någon typ av feedback från användarna.

Hur kan kunskap om människors beteenden vara till nytta för olika tekniker, modeller, metoder av DM?

Det handlar mycket om vilken uppfattning människor får av systemet. Människor vill ofta uttrycka sig, om deras inställning till systemet är bra. Det ligger en hel del psykologiska aspekter bakom, i vilken utsträckning användbar information kan genereras från olika människor. Speciellt rör det sig om kognitiv psykologi eller kognitiva beteenden som är ett

oerhört intressant område inom DM. Genom att förstå hur människan tänker och uppfattar system eller hemsidor kan olika DM metoder på så vis dra nytta av detta och bli mer användbara. När dessa aspekter är avklarade behöver man sedan titta på informationsvärdet från det som har genererats. I vilken utsträckning är den information som genererats värdefull för syftet med DM processen som bedrivits.

Transkribering – Intervjuföretag B

A: Vad vi förstått växer ständigt kvantiteterna av data inom olika IS. Hur uppfattar du detta påstående?

I: Ja, visst är det så, Har ni sett ”Winter corp”?

A: Vad för något?

I: Winter Corperation. Vart annat år gör dom en undersökning för att ta reda på hur stora dom största databaserna är. Bland annat Data warehousen och där gjordes 2005 , 2007 undersökningar. Mellan dom åren växte de största databaserna med 300 %,
I: Så Winter corperation, om du söker det på webben, så finns dom mättningarna där. Så vad man gör då är att man tittar på vilka behov företaget har och så skickar man ut script till företaget så att alla mäter på samma sätt.
I: Så kategoriserar man dem lite på olika sätt om det vore Data warehouse eller något annat. Den är ganska intressan för den visar hur mycket det har växt under dom här åren och där finns snygga grafer och sånt. Det är en exponentiell tillväxt och där den största databasen 2007 är på 300 terabyte av den kommersiella biten då. Och där man tror att dom kommer att nå petabyte storlekar först 2010 kanske.

A: Oj Oj.

I: Ja det växer, det växer nått enormt.

A: Jo det har vi också förstått, men vi har inte hittat några bra siffror.

I: Du kan hänvisa till denna för det är en oberoende analysfirma.

A: Okej. Är det positivt eller negativt med stora datakvantiteter vid utförande av DM? I så fall varför?

I: Jo det är både och, naturligtvis. Det positiva är, ju djupare dataset du har. Alltså jag tänker inte på bredden utan på djupet, jag har alltså inte 1 års data utan 15 års data. Bredden är också intressant, ju mera data jag har både på bredden och på djupet desto bättre kan mina algoritmer köra. Baksidan är ju då share volumen, hur mycket data det är, att kunna bearbeta mycket data. Traditionella DM verktyg har varit separata DMmotorer vid sidan om. Då ska man göra uttag ur data från massa olika system som ska göras om till ett annat format och som sen ska ges till DMmotorn att tugga på. Det är inte helt trivialt när det handlar om terabyte tabbeler. Det ska göras om till ett speciellt format och sen ska det flyttas igenom i ett nätverk och sen till DMmotorn som ska sitta och tugga igenom det, det blir som en begränsande faktor, att man inte kan ta all data. Istället för att köra motorn någon annanstans har vi stoppat in DMmotorn i databasen. Vi kör algoritmerna där data ligger, man behöver alltså inte flytta den.
I: Det blir bättre ju mer data man har, för det statistiska underlaget, men det blir mer svår hanterligt för volymerna blir förstora. Det är därför det blir mer och mer intressant att köra

DM där data redan ligger, att slippa flytta på det. Annars kanske man aldrig kommer att kunna använda sig av dessa stora volymer och man skulle inte kunna köra dessa stora volymer så ofta som man skulle vilja.

A: Vi har uppfattat att organisationer förlorar både resurser och pengar på att använda sig av data med bristande kvalitet. Hur ser du på detta?

I: Så klart skulle man vilja ha mycket data och bra datakvalitet för att kunna säkra DMs resultatet. Här har vi första problemen med kvalitén och kvantiteten. Kvaliteten är nästan ett ännu större problem, man har väldigt dålig kvalité på data. Detta problem förekommer även i Data warehouse sammanhangen. När man plockar in data från, ta ett exempel ett kundsystem i 15 länder och så ska kunden konsolideras i ett Data warehouse och där ser man att det har varit olika systemen som har olika affärsregler som har registrerat kunderna. Och det är detta som är jätte jobbigt i alla Data warehouse. För att på något sätt ensa data, vilket är 80 % av projektet, få in data i en ny struktur. I min värld är DM något som man kör i ett Data warehouse. Av just denna här orsaken att man har lagt ner ett jätte jobb på att höja kvalitén på data. Sen kan det vara så att man kör visa basketanalyser som man kör mycket i detaljhandeln dom går att köra direkt i tabellerna.
Men så fort du vill börja använda andra attribut då måste du hämta kund information från fyra andra ställen, och då blir det ganska naturligt att använda sig av ett Data warehouse där allt redan är sammanställt och någorlunda tvättat. Så datakvalitet är ett jätteproblem. Så stora delar av ett datawarehouse är tvättmaskinen för att tvätta data. Så använd detta istället för att bygga separat silo för DM där gör man ju om dom tvättstegen.

A: Blir metoder och tekniker inom BI allt effektivare på att möta organisationers krav relaterat till kvalitativa och säkra resultat? Hur har detta förändrats de senaste 5 åren?

Jag gjorde mitt första Data warehouse åt Ericsson någon gång på 90-talet. Redan då visste vi egentligen hur vi skulle göra, att man var tvungen att ensa data.
Så fort det var mer än ett ämnesområde, säg att vi har tillverkning och försäljning, dom ligger oftast i två olika system. Då hamnar man i det problemet att det inte går att ställa en fråga till två olika system för då sitter du och klipper och klistrar i Excel i sluttampen. Ska det där vara trovärdigt måste man föra ihop informationen till en ny struktur, det är där warehouset dyker upp, det är det som är Data warehouse. Den metoden har funnits länge, hur man skulle göra. Det som egentligen skiller sig, vi har ju dom guruna på Data warhouse sidan med Bill Immon som har kommit ut med en bok. Honom har ni kanske hört talas om? Hans böcker: ”Building the Data warehouse” den kom ut i början av 90-talet. Han är egentligen pappa till Data warehouse, dom tankarna om hur det görs finns fortfarande kvar. Sen finns det en ny guru som heter Ralph Kinball och han skrev en bok som heter ”Data warehouse toolkit” och den kom ut 2000 och då har liksom tekniken förändrats men inte metoden. Databaser har förändras från 92 till 2000 ganska mycket men metoden och metodiken är den samma. Så jag skulle vilja säga så här, jag tror inte att metodiken har förändrats någonting under 5 år men tekniken. Det som kanske ändras mest i metodiken, normalt sätt i Data warehouse projekt. Då början man med verksamhetens behov av information. Sen försöker man modulera upp det i en struktur. Sen nästa steg , ”-jaha var ligger denna informationen?” Och så får man gå in och rota i alla källsystem. Sen gör man alla tvättstegen för att kunna få in informationen i den nya strukturen. Sen när allt det där är inladdat kan man göra de första rapporterna och visa intressenterna. ”- ja vi har den och den informationen så här ser det ut”

och det första dom kommer at säga är ”Det ser jättebra ut, men jag skull också vilja ha…” och så kommer dom med fler önskemål.
Den här cykeln är ganska lång, och detta kan hålla på i från 3 mån till flera år innan det händer något. Det är vad som har hänt, Immon körde på detta viset och Timo försöker korta ner den här cykeln och han säger att ”Gör inte hela informationsmängden på en gång utan försök prioritera och ta små steg i taget. Men man måste ändå göra de viktiga saker som ta reda på vilken struktur vi har. Det är lite som att inte göra långa projekt, utan göra korta iterationer. För en sak är säkert, när kunder ser resultatet kommer han be om mer, det där är lite förändringen. Kravet på förändrings hastighet har ökat. Allting går fortare, jag vill ha det nu.

Det är väl inte så konstigt, det är som allt annat i samhället. Data warehouse och beslutstödet har blivit allt viktigare i systemet speciellt i lågkonjekturen. Det som just nu är drag på är Data warehouse, beslutstödsystem och DM. När det går dåligt vill man ta reda på var det är mest lönsamt, vad är det som är lönsamt? Vad ska vi skära ner på? Har man inte beslutstöd så kan det bli fel.

A: Finns det någon kvalificering för att data ska lämpa sig speciellt för DM? T.ex nivåer, eller rangordning av data.

I: Om den är aggregerade eller inte aggregerad? Det beror lite på vilken DM som ska göras. Om vi tar det exemplet om basketanalysen som detaljhandeln använder. Vilka produkter som säljs tillsammans. Då är ju den på allra lägsta nivå, dvs. typelrader, annars kan du liksom inte hitta det. Telco kör ju rätt så mycket, och dom kör mer på en aggregerad nivå för att kunna se användarmönster. Då behöver dom inte den där CDR:en, Call detail record, det vet ni vad det är va? För varje telefonsamtal blir det en log som säger: nummer A ringde till nummer B och samtalet var så här långt och massa annan information. Det motsvarar alltså en kvittorad i Telco sammanhang. Dom kanske inte kör DM på den nivån utan lite högre T ex den här kunden tillhör den här kategorin eller man kör DM för att kategorisera kunder och det gör man utifrån ” Ja han ringer 150 samtal i veckan till 30 telefonnummer och här är en annan kund som ringer 150 samtal i veckan men som bara ringer till två telefonnummer”. Det här exemplet är taget från Mobile Italia alltså den största telefon operatören i Italien. Dom har ett speciellt attribut på sina kunder som heter ”frequent mama caller” och då kör dom DM under en månad igenom dom samtals loggarna och så kategoriseras kunden i en klass stege mellan ett till tio. Om kunden är av klassen 10 betyder det att kunden bara ringer till ett ytterst få telefonsamtal. Klassen 1 betyder att kunder ringer många samtal till många olika. Detta använder man sig då av om kunden säger: ”Jag vill inte vara kvar hos er, jag vill byta operatör” då kan kundsupporten titta på frequent mama caller attributet och se ”jaha, du är frequent mama caller 8,” och då kan kundsupporten säga ”Du kan få två telefonnummer som du kan ringa gratis till”
Det beror alltså helt och hålet på vad DMen är ute efter? Om man ska använda sig av detaljerad data eller aggregerad data.

I: Ni har säkert hört om Wallmart Exemplet?

A: Vad sa du?

I: Blöj och öl exemplet

A: Jo, det har vi läst om.

I: Där används en mer detaljerad information från kvitto raderna men samtidigt också massa annan information. T ex Var låg butiken i förhållande till bostadsområdena? Demografin eller hur många människor som bodde i området. Det räcker alltså inte med information om själva händelsen, någon som ringt ett samtal eller någon har köpt någonting, man behöver alltså veta massa annan information.

A: Det är ju sant att det är utökat, man kanske stirrar sig för mycket blint på det här med blöjorna och ölen, Det är mycket annat som spelar in också.

I: Facit för blöjorna och ölen var ju att dom stod i samma sektion. Alltså måste du ha information om hur butiken ser ut. Detta var kanske inte det mest naturliga attribut, så antagligen måste man ha ganska många attribut. DM är ju något som du inte vet. Du försöker hitta något sammanhang. Jag skulle gissa på att man måste ta med fler attribut än vad man själv tror ”Ä, det där kan inte spela någon roll” är ni med? Det är alltså fel tänk, så det kanske är bättre att slänga med ännu fler attribut för att det kan vara något som man inte trodde initialt att i kombination med något annat.

A: Sen har vi ju nästa fråga om klassificering.

I: Det finns ju något inom DM som heter ”Attribute importent” som rangordnar de attribut som påverkar mest.

A: Vad hete det sa du?

I: Attribute …… ja jag vet inte… men det innebär att man rangordnar vilka attribut som påverkar mest. Och det är en sorts fler steg körning, att man först försöker hitta de attributen som verkar vara de attribut som är påverkar och sen gör man en annan körning ut efter det.

I: Men ju fler attribut som du slänger i desto tyngre kommer det bli liksom mer och mer komplext. Så det kan vara så att man försöker få bort dom som inte behövs.

I: Attribute…

I: Attribute, någonting, nä ni får söka på det .

I: Och då blir det ju det med klassificering av data. Attribute importence att man försöker klassificera och får hum om vilka attribut som går att använda.

A: Hur upplever du att datakvalitet påverkar skapande av mönster och i sin tur kunskapsutvinning?

Data kvalitet är jätte viktigt. Dålig datakvalitet kan till och med göra så att det inte ens går att köra mining, så illa kan det vara. Man kör ju något som heter data profiling. Profilingen hjälper ju dig med att ta reda på fördelningen. Hur är semantiken i data? T ex Av dom här kunderna så är det 1000 som är klassificerade så här, och så är det 15 som är klassificerade så

här och så finns det bara en som är klassificerade så här. Då kan det vara så här att man inte tar med dom udda mönstrarna i DMen. För att man vet att det kan ge väldigt konstiga resultat. Man får helt enkelt skala av dom yttervärdena, dom lägsta och dom högsta.
Sen går man in och kör DM på 90 eller 80 % av data, där dom flesta ligger. För dom där ytterligheterna är kanske såna man ändå inte vill ha med. För dom kommer att förtycka resultatet. Men det betyder då att du måste ha gått igenom data och veta var ytterligheterna är och var den stora mängden data ligger. Ligger dom i den mittersta delen, var är dom 80 % som vi ska använda? Så att du sätter dina begränsningar. Det är där data profiling kommer in som ett försteg, före DMen för att förbättra data kvalitet. I detta fallet tar man bort helt enkelt dom dåliga raderna, och där med få ett bättre resultat.

A: Hur garanteras att den information som utvunnits via DM är intressant?

Det som är intressant ur DM perspektivet är att det inte körs av en händelse utan man har ett affärsproblem som man på något sätt vill få reda på. Hur man ska kunna gå runt det här. Det finns alltså ett problem och det problemet försöker man lösa med mining. Vi pratar alltså inte om en körning utan man testar olika algoritmer , man kör dom i testdataset och försöker jämföra det med sanningar som man redan vet för att det kan faktiskt ge helt tokiga resultat.
Det är alltså inte helt lätt att tolka det och inte helt lätt att veta om det är intressant om man inte redan från början har jobbat efter en tes som man testar. Man måste alltid ha en tes och i affärssammanhang något att bygga på.
I affärssammanhang vet man oftast varför man gör en körning. Detta pågrund av att det är ganska dyrt och någon gång så ska det här betalas tillbaka.
Och det där med att garantera att informationen är korrekt efter DM, det är ju de testen där man testar av mot redan sann data.
Man har ett litet dataset och man vet att det blev så här… vad ska man ta för exempel … Vilken kund kategori är mest trovärdiga att köpa en Volvo bil igen? Då kan man titta tillbaka till alla som har köpt en Volvo och titta på deras attribut och så kör man en körning på en större datamängd och så ser man att man får samma attribut igen, det blir då en sorts testfas.

A: Men då kan det bli lite problematiskt att hitta nya mönster?

I: Absolut, vilket innebär att det måste analyseras och säkerställas, och det blir lite som blöjor och ölen det fanns inget att testa mot. Man viste att andra butiker sålde mer öl än de andra men varför, det hade man ingen aning om och det kunde man inte göra med vanliga BI-verktyg. Stolpen var högre, dom sålde mer men varför sålde dom mer? Att analysera det manuellt är ju omöjligt, det var ju miljarders rader. Sen när det kom fram att belägenheten var den samma, då räckte inte det heller utan man var tvungen att förstå köpmönstret. ”jaha det är papporna som handlar på fredag” om det nu är ölen eller blöjorna som åker med, jag vet inte. Det kan man tolka som man själv vill. I vissa av miningarna finns det alltså ingen absolut sanning, du kan alltså inte veta resultatet i förväg.

A: Man tar alltså då i största beakt att det är mycket analyser innan man kan använda sig av det.

I: Detta är inte såna grejor som man kan göra på ett par dar utan det är mycket jobb och det är därför det görs så lite DM idag. I Sverige är det ju inte så speciellt utbrett.

A: Vad anser du används mest, manuell eller automatisk analysering av resultatet som framställs via DM?

I: Det går ju att sätta upp regelmotorer så att du kan säkert automatisera en bit till. Men någonstans i sluttampen måste man fråga sig ”Är det här rimligt?”.

A: Jo men då kommer vi in på det här om kunskapen blir mindre subjektiv när information analyseras med automatisk analysering gentemot manuell analysering? Hur vet man att kunskapen som man utvinner är äkta?
I: Ja jag tror att man alltid har den där subjektiva parametern som kan förrycka resultatet eller ger analysen totalt fel eller inte rätt. Jag tror heller inte att den automatiska analyseringen fungerar så särskilt bra. Någonstans måste det finnas en klok gumma eller gubbe som har verksamhetsförståelse och processförståelse och se om det är någonting eller om det är totalt fel av någon anledning, sen kan dom ha fel också.

A: Hur kan kunskap om människors beteenden, det kognitiva perspektivet, vara till nytta för olika tekniker, modeller, metoder av DM?

I: Ja det som jag har stöt på är ju lite det där som Telco gör. Alltså tolka användarbeteenden och vad kommer dom göra här näst, retail gör samma sak. Antagligen så körs det DM på tittarsiffror också. Egentigen är all DM till för att hitta beteende och kunna sälja, är det inte så? Det beror ju på att det är affärerna som driver det annars skulle ingen göra det, dom vill sälja mer, eller behålla kunden, det är billigare att behålla kunden än att skapa en ny kund, så även där blir det också beteende.

A: Men hur mycket används det som inom företaget som business intelligens, för att förbättra affärsverksamheten?

I: Det är ju mera BI, så visst gör man saker som att titta på leveransprecision, flera steg i en tillverknings process. Har ju t ex ett stålverk, det inte bara ett stålverk utan valsverk, plåtverk klippverk, det är liksom flera steg i processen. Det vi gjorde där var en dellösning, att det som kom ut från stålverket skulle vara klart precis när det skulle valsas och sen skulle. E ni med? så att något steg inte blev väntande, för att det förra steget inte var klart. Så att processen hela tiden rullar. Det ska inte köras förtidigt och då ska det inte heller köras försent. Det ska inte ligga ett lager som väntar för det binder kapital och i sluttampen gäller det att leverera exakt den vara som kunden vill ha vid den tidpunkten som är bestämd. Det är ju en sak att ha leverans precision gentemot kunden, men sen har du leverans precisionen internt, alla interna steg. Där behöver man inte köra mining utan det är ju BI. Där går man in och kollar och ser vad man har för tider, vad har kommit försent, så det ser jag lite mer som BI. Mining är ju något ting som du inte vet vad du söker efter.

A: Hur ser framtiden inom data minig ut relaterat till dagens syn, relaterat till förbättringar kring relevanta och säkra mönster?

I: Jag tror att DM har varit förknippat med något som har varit dyrt, komplext, att man behöver massa data. Det har varit lite hokus pokus över dom som har kunnat göra det. Jag tror att det kommer bli mer vanligt att man gör det i det perspektivet att för olika branscher vet man vilka DM grejer som är intressanta. Istället för att ha fått en verktygslåda och virke

och så säger vi att vi bygger ett hus så är det mer monteringsfärdiga saker. Så för dom olika branscherna har vi färdiga DMlösningar för branschområdet. Det behöver inte vara någon som sätter upp algoritmerna, det är redan färdigt, utan ”Det här är en intressant siffra att titta på” Man har alltså ett halvfabrikat, så här bör man alltså göra, det här är best practies. Det blir alltså en kortare startsträcka för dom som skulle vilja börja med mining. Då behöver dom inte göra det här jobbet. Dom behöver alltså inte köpa en statistiker som vill sätta upp det här . Om detta blir allemansgods, då har man inte heller någon competative edge mot dom andra, men om din konkurrent gör det måste du också göra det.
Så jag tror att om business intelligence tidigare har varit något som ytterst få i verksamheten fick ta del av rapporterna men nu är det något som är mera spritt till massorna. En naturlig del av din verksamhetsprocess, att man hela tiden följer upp och mäter. Det har blivit spritt sig till en större del av verksamheten. Det är inte alltså regioncheferna som tittar på det utan nu är det mellannivåcheferna eller gruppchefer som har sin grupp, för att kunna se hur det ”går för min grupp” och på samma sätt tror jag att DMen och resultat kommer att användas mera.

A: Om vi ser på hela processen, var kommer förändringen att liggas ..behövas för att uppnå bättre resultat.

I: Jag tror datakvaliteten, jag tror att… Algoritmer har varit ganska oförändrade dom sista 10-15 åren, det är inte så konstigt det är ju matematiska formler. Och dom tror jag är ganska stabila, utan det är mera data som matas in. Har du dålig data då får man börja sortera. Själva algoritmen, den matematiska formeln förändras inte så särskilt mycket, det är just data som du kör igenom formeln. Det handlar då om hur ren data är och hur stor data mängd det rör sig om. Att kunna köra igenom mera data och köra igenom renare data, kommer vara där förändringarna kommer att göras.

A: Men sen när det ska rätts till?

I: Ja det är ju det som man har kämpat med hela tiden och fortsätter med att kämpa med och där man försöker automatisera de här rättelserna eller använder sig av program som rättar till felaktiga koder och man använder sig av fuzzy logic och för att rätt felstavningar.

F: så Fuzzy logic används fortfarande?

I: Ja jamän, där har du tillämpningen i data tvätt. Det finns såna klassiska saker som att man plockar bort alla vokalerna ur ett namn och så gör man jämförelse på konsonant sidan för att få bättre träff, hit rate, allt är för att ”
Om det står K. Petterson eller Kalle Peterson, ett med 2 T:en och så råkar dom ha samma adress ”- Är inte detta samma kund”. Om man inte reparerar detta får man felaktiga resultat i sluttampen.

I: Så datakvalitet, datakvalitet och datakvalitet. Det är det viktigaste.

F: ICA har väl precis börjat med att skrädda sy sina reklamblad åt sina kunder och hela deras data baseras på att kunden måste dra sitt ica kort.

I: Dom här medlemskorten är till för att få mera information från sina kunder. Om vi tittar på det här med Retail. Om du inte vet vem kunden är vet du bara att någon har kommit in

och köpt en kundvagn och i den kundvagnen låg dom här sakerna och du vet ingenting mer. Om du nu kan koppla det till en Person, Kalle Petterson, han är 43 år, han bor i radhus, han har fru och barn, dom är i den och den åldern. Han tjänar så här så mycket, till slut kan man börja klassificera kunderna, han köper mycket blöjor och öl… ha ha. Detta handlar om att beredda data, för att få information runt omkring.

A: Är det något som du vill tillägga som vi inte hunnit prata om men som anses intressant?

I: Nja det är väl det att DM tillhör business intelligence och att det handlar om är att DMen skapar ett nytt attribut, om det är på produkt eller kund som säger någonting som t ex det där med Frequent mama caller. Så Efter man har kört DM, då är det bara ett extra attribut på en kund så ju flera attribut ger ju oss mera kunskap om kunden i det här fallet. Man måste ju hitta det där som triggar en kund mer och mer. Då fungerar det inte att gå med bombmatta längre, nu måste man verkligen träffa rätt för att det kostar så pass mycket att göra dessa kampanjer. Istället för att skicka ut till 10 miljoner människor skickar man ut till 1 miljon människor och har en ännu högre hit rate för att man har träffat dom rätta.

Hade en kollega i tyskland som berättade om Quelle . Quelle är en jättestor tysk postorderfirma. Dom har funnits i hur många år som helst, deras katalog är jättebred, ungefär tusen sidor och den kosta ungefär 4 eller 5 euro att trycka. Då vill man ju inte skicka ut den till alla kunder. Man ville bara skicka ut till en viss utvald kundtyp. Då körde man DM och hitta vilka kunder som är mest troligen att köpa igen och för ett visst belopp. Så istället för att skicka ut till alla kunderna skickar man ut det till dom och istället betalar man för projektet.

A: Vi tror nog att vi har fått det mesta, tack så mycket

I: Så är det något mer så hör ni av er, dyker det upp några frågor, så får ni ringa igen så bollar vi det också.

F: Tack så mycket.

Transkribering – Intervjuföretag C

A: Vad vi förstått växer ständigt kvantiteterna av data inom olika IS. Hur uppfattar du detta påstående?

I: Jo det finns ju såna hypoteser om den där lavinen, att data fördubblas på, jag vet inte på hur många månader eller år. Mer och mer lagras även också all ostrukturerad data alla våra mail. All text som finns det är också det som bygger upp datavolymerna. Men sen är det också att allt vi gör på nätet eller går och handlar, alla våra fingeravtryck sparas någonstans . Så det är ju mycket mer som sparas idag än vad det gjordes för 10 år sen. Sen med alla dom här automatiserade kanalerna som finns, så alla har ju tillgång till mer data. Jag håller verkligen med påståendet och jag tycker att man kan se det också. Nästa utmaning är att analysera icke strukturerad data som t.ex. texter och även det som vi kan spela in. Definitivt, det är lavinartat.

A: Är det positivt eller negativt med stora datakvantiteter vid utförandet av data minin? I så fall varför?

I: Det positiva är att om man kan använda det på rätt sätt borde man kunna fatta bättre beslut, hitta bättre mönstren att göra bra saker med det.
Men det kräver också att man kan analysera den mängden data och ta till vara på den. Det krävs verktyg, kompetens och systemstöd. Men det ger möjligheter om man gör på rätt sätt.

A: Något Negativt?

I: Negativt kan ju vara integritet, såna saker som att allt lagras, hela den biten. Man kanske känner sig att storebror som kan spara allt vi gör, det är väl något som man kan tycka. Negativt då, det är väl att informationen kan missbrukas det kan vara en negativ aspekt.

A: Men om man ser det vid användning av mining att det är stora data kvantiteter?

I: Det ställer ju större krav på analytikerna och mjukvaran att kunna hantera mängden. Att kunna sola och hitta fram det relevanta, det kan vara större mängd att leta efter den där nålen. Det som är negativt är att det kommer utkristalliseras vilka saker som kommer att klara av det och vilka som inte kommer att klara av det i så fall. Dom som kommer att klara av det kommer få fördelar, men det krävs mer utav användare och av mjukvaran.

A: Vi har uppfattat att organisationer förlorar både resurser och pengar på att använda sig av data med bristande kvalitet. Hur ser du på detta?

I: Det är en bra fråga och även om man är ett geni och har dom mest sofistikerade verktygen som går att uppnå.
Data som man stoppar in blir ändå inte bättre än vad data är. Det förutsätter ju överhuvudtaget att kunna bygga en modell. Modellen ska man använda sig av att fatta beslut, för att på något sätt skapa fördelar i sin organisation.

Det är ju det fundamentala. Har man inte bra data spelar det ingen roll. Data kvalitén är enormt viktig.

Legiskerar man den biten så kommer man stötta på problem. Det som kommer att hända då är att modellerna kommer att göra fel saker.
Dom kanske attraherar in fel kunder och vi missar affärsmöjligheter. Så det ligger alltså i data kvalitén. Har du dåligt data så kommer det komma ut och det kan vara farligt alltså, att inte ta hänsyn till data kvaliteten.

A: Men du som jobbar inom olika områden, hur skiljer sig data kvalitén beroende på vilken område?

I: Man kan väl tänka sig att inom områden där data har lagrats sen tidigt som t ex Telecom och finanse. Dom har oftast väldigt god datakvalité. Utan där handlar det mer om att kunna tillrättalägga data så att det blir analyserbart och aggregera upp det med tiden. Där har man transformerat data så att det blir analyserbart, där är det bättre datakvalité. Men sen kanske mindre utvecklade branscher, där är kvalitén sämre. Man har inte system som kan fånga upp data på ett bra sätt.

Vilka branscher det skulle kunna va? Jag har väl haft turen att jobba med ett bra warehouse och en struktur. Jag har ju mest jobbat inom bank, finanse där man har tillgångar . Mindre outvecklad bolag där man inte har samma systemstöd för databaser. Man har kanske Excelsidor som skickas hit och dit. Samma saker som finns lagrade på olika ställen men heter olika saker. Har man inte gjort jobbet och fått en ordentlig databas så kan det bli problem, även om man har hyfsad kvalité men man förstår inte vad som är samma saker.
Tittar man inom retail har man inte kommit lika långt på att få upp allting på kund nivå för att kunna göra bra analyser men ändå så fångar man upp siffror på ett automatiserat sätt men man har inte kommit så långt att använda det i DM.

A: ja det tycker vi är konstigt när man håller på med kundkort och grejor

I: Ja det tycker vi också, tittar man utomlands så har alla dom största kedjorna t ex Wallmart. Dom har en avdelning med analytiker. Det finns ett företag i England som heter Tesco dom har 100 till 200 data minners som sitter och analyserar kundbeteende. Ica har kanske en halv som kanske gör någonting.

Tur att det har gått så bra för dom. Dom har inte haft någon konkurrens.
En handlare i Sverige tycker nog att man har koll på sin bransch och behöver inte använda sig av prediktiva analyser. Men det kommer att ändra sig så småningom när det blir större konkurrens osv.

A: Blir metoder och tekniker inom BI allt effektivare på att möta organisationers krav relaterat till kvalitativa och säkra resultat? Hur har detta förändrats de senaste 5 åren?

I: Senaste 5 åren så har företagen fått upp ögonen över att dom behöver modeller för att hjälpa med att analysera sin data. Det går inte att ha kommunikationsplaner, att man ska ha en strategi för varje kund.

Man behöver hitta trender och mönster i data, det går inte att göras manuellt. Man behöver system stöd och modeller för att kunna göra detta. Det som man behöver göra idag för att kunna agera är att ta reda på vad som kommer att hända och inte titta på vad som har hänt och då behöver man modeller.

Jag har hållit på med detta i 10 år. I början var man ute och missionera och folk förstod knappt vad man snackade om. Men nu har även organisationer som inte har kommit så långt blivit intresserade. Tittar man på rena algoritmer så kommer det hela tiden nya, men tittar man på marknaden så kör folk med 99 % med regression och beslutstöd

A: så det är kanske metoderna mer än tekniken som förändra?

I: Jag tror att tekniken förbättras hela tiden. Men metoderna allting börjar med en analysfråga eller en affärsfråga: ”Varför tappar vi kunder ” i den ändan. Sen använder vi mining för att kunna besvara denna fråga. Men fler är nyfikna på att pröva på detta än vad det var för 5 år sen .
Metoderna för att göra mining börjar med att hitta ett affärsproblem som man ska utforska och sen anpassa en modell över, som man sen ska använda.

Men tittar vi hos oss så ser metodik lika dan ut, man kanske gör det på ett bättre eller snabbare sätt, har bättre funktionalitet för att komma snabbare till mål och få bättre resultat. Men det stora är att det är fler som vill pröva på detta. Men det här är nu högre upp i organisationen på lednings nivå där man vill få reda på vilka prediktiva modeller som används.

De kreditgivande bankerna använder prediktiva modeller när man ska ansöka om lån och om man har sköts sina lån, det har ju dom högt upp i banken koll på. Det är mera utbrett nu. Det är inte en statistiker i vitt rock som sitter i ett hörn och grubbla. Utan de prediktiva modellerna ser man som en tillgång. Det är den stora skillnaden de senaste åren .

A: Vi kan hoppa tillbaka då till datakvalitet då, Finns det någon kvalificering för att data ska lämpa sig speciellt för DM? T.ex nivåer, eller rangordning av data.

I: Man kan säga så här att den bästa data är ju transaktions data, så som siffror som är kvalitet säkra, det har ju en hög kvalité hög säkerhet.
Sen vet jag inte om ni menar kvalitet som ”Missing, extremvärden, skeva fördelningar, konstigheter i data, det är mer som man använder metoder för att hitta.

A: Jo men typ lite så rangordning nivå? Om det är något sånt man använder sig av.

I: Det man brukar göra för att titta på kvalitén i data är att se hur många procent missing man har, det är ju den första delen. Har man ungefär 50 % missing i en variable då bör man överväga om man överhuvudtaget ska ha med den. Men sen om man har mycket konstigheter som T ex konstiga tecken, kanske en nolla som kan betyda något speciellt. Ibland betyder den missing och då måste man också tänka till hur man ska hantera det. Man har ju det som första steg i mining det är att kvalitetssäkra data
Även om man fått data från ett etl-system eller om vi har en analystabell så gör vi ändå det som analytiker undersöker antalet missing. Om det är för mycket missing då går variabeln

bort. Sen tittar vi på fördelningen, är den extremt skev? Det kommer att påverka mina parametrar, vi kanske måste transformera variabeln eller kanske gruppera den på något sätt för att kunna hantera ”missing”.

Sen har vi konstiga värden, ja då får vi gå in och titta på den biten också . Man får helt enkelt transformera variable så att den blir så optimal som möjligt. Vilken modell vill vi använda? Så det är lite upp till vilken teknik man ska använda och sätta upp en agenda för datakvaliteten och eventuella modifieringar av data.

Finns det olika faktorer i data som bör undersökas för att fastställa dess äkthet? I så fall vilka?

I: Då har du ju spårbarhet till datas ursprungliga källa. Den har ju varit i massa instanser på vägen där man gjort vissa funktioner och där kan det ju ha blivit fel.
Så det räcker inte med att bara titta i ett analysverktyg utan jag måste också veta hur dessa siffror har kommit till. Så man pratar alltså om spårbarhet och transparanse. Det är oftast ett krav på vissa modeller att man kan gå tillbaks och härleda, så här har det kommit till.

A: Finns det något verktyg för att bedöma kvalitén?

I: Jag relaterar bara det som vi har på vårt företag. Vi har flera delar som granskar kvalitén. Man kan titta på extrem värden, missing, konstiga saker, vi har en hel verktygslåda som vi kan applicera för att få data optimalt. Det finns något som heter Data FLUX där man kan systematiskt granska kvalité och mer avancerad kvalitetskontroll. Man ser T ex att Mathias Lanner stavas på ett sätt men annorlunda på ett annat ställe och med logik kan man förstå att det är samma observation. Alla mjukvare leverantörer har nog stöd för datakvalitet på något sätt, det behöver man ha.

Men sen när vi pratar om att bygga modeller och kollar kvalitén på resultatet så har man någon slags metod där man validerar modellen. Man bygger modellen på en typ av data och validerar den på andra data och även ett testdataset där man testar modellen på data som är osedd för modellen och som använder sig av kvalitetsmått som t.ex. informations strukturering, ROI index osv. Det finns alltså en uppsjö med verktyg för att se hur bra ens modell är. Man kan också använda sig av grafer för att se hur pass bra modellen är.
Men det som är viktigast är att när du bygger modellen är att du måste validera den vid samma skede. Att man inte bygger modellen på ett dataset och tror att det ska fungera till nästa data. Där har man också sin metodik, att ha med validerings biten så att det vi får fram ska kunna generaliseras och ska kunna användas på ny information för att kunna fatta bra beslut,

A: Hur upplever du att datakvalitet påverkar skapandet av mönster? själva kunskapsutvinningen.

I: Är det dålig kvalitet så är det alltså fel mönster som vi hittar. Så det har en enorm påverkan. Har vi låg kvalité på data så kommer vi inte att hitta några mönster, eller de vi hittar är felaktiga. Så det finns en enorm påverkan, allt ligger i data. Data sätter ribban för hur bra det blir.

A: Hur garanteras att den information som genererats via DM är sann eller korrekt?

I: Jag kan ju säga att man inte ska DM för skoj skull utan det är framför allt att det måste finnas en uppdragsgivare, en beställning, ett business case eller vad som helst som vi behöver använda oss av analytiska metoder så att vi kan besvara. Då har vi redan där en koppling mot verkligheten. Så finns det också det vi kommer framtill som folk är väldigt intresserade av för det kommer för hoppnings vis ge nya intäkter eller minskade kostnader. Så det finns inte så många som minnar för saken skull.

I: Ta in en stor fil också hittar man utan allt utgår i någon form av affärsproblem. Så fungerar det oftast inom närringslivet där vi kör då. Det finns ingen som testar lite för skoj skull och kanske testar ”är det här något ting att ha; Ja eller Nej”

A: det måste alltså finnas ett syfte.

I: ja, det är så vi börjar. Vi börjar med en hypotes som vi besvara då genom att bygga en modell då så småning om.

A: Hur säker ställer man det då att det är korrekt då?

I: Det man får göra då, dels att titta på olika anpassnings mönster och se hur pass bra modellen hittar det där sambandet. Oftast när vi bygger en modell med miningen, utgår vi från ett dataset där vi känner till vilka kunder som var bra vilka var dåliga. Sen har vi uppsjö av förklarande variabler och då kan vi se hur bra modellen klassificerar rätt. Då får vi ett mått på om modellen fungerar eller inte och vi kan också validera modellen på ny data. Sen kan vi också ta in experter som kanske har verksamhetskompetens och som kan gå in och titta om det är relevanta faktorer som påverkar beteendet. Då kanske dom säger: ”ja det är relevanta” eller ”Den här variablen borde vi inte ha med för att den kolliderar med vårt mål variable”. Så man kan använda affärskompetens för att granska en modell. Men sen har du alla anpassningsmått som man får av olika analyser. Sen är det också att testa modellen du ska ju använda modellen för att spåra upp nya kunder och fatta bra beslut och då får man ha system för att övervaka moddelen, är det de kunderna som den plockar in eller som den predikerar rätt? Så även om modellen är klar måste man ha ett system som övervakar eller monitorer modellen. Det kan ha gått en tid så att det har hänt saker . Så det är som en livscykel för en modell, som ska skapas som sen ska används men sen underhållas, sköttas och sen går modellen i pension och man behöver bygga en ny. Den är inte för evigt och olika branscher går olika snabbt.

I: Pratar då om modell livscykeln .

A: Vad anser du används mest, manuell eller automatisk analysering av resultatet som framställs via DM?

I: Det är intressant, jag som är statistiker och jag vill inte bara få trycka på en knapp och så kommer det ut någonting. Jag vill ha kontroll över den processen. Jag vill kanske inte sitta och bygga regressions metoder och bygga modeller. Jag vill själv sitta i förarsättet och bestämma själv hur jag vill bygga dessa och vilka variabler jag ska använda och det jag ska transformera. Jag tycker att det är lite farligt att trycka på en knapp så kommer det ut

någonting som man sen ska använda. Men jag vet att det finns att många av våra konkurrenter har mycket att det automatiskt bygger en modell som tittar på datakvalitet, gör transformationer med ett antal olika tekniker och sen kommer det ut något resultat. Men jag tror att det kan vara lite farligt om man inte har den kompetens som krävs, det kan bli att man fattar felaktiga saker. En modell som är automatiskt kan aldrig känna av relevansen, att visa variabler inte ska användas, det kan vara en variable som är kolliderar med en målvariable, så att det kan bli felaktigheter.

Så jag vill som analytiker, ska man sitta i förarsättet och bestämma hur analysen ska gå till väga men sen ska jag ha bra verktyg som gör så att vi kan göra det på ett smart automatiserat sätt. Du ska tala om vad som ska ske och vad som ska göras, inte ta in en fil och trycka på en knapp och så tar det två minuter och så kommer det in en lista med kunderna som ska bearbetas då tror jag att det kan vara lite farligt. Om man inte förstår ”varför blir de här utvalda? ”, att man inte har nån dokumentation eller log för att se vad fan som har hänt. Jag vet inte hur mycket ni har tittat på olika verktyg på nätet. Såg en del verktyg på nätet där allt bara gick på automatik. Men det kommer att efterfrågas på marknaden att man vill ha mer automatiserat då och jag tror också att många av våra konkurrenter har börjat göra det, att man har deras generella mining verktyg men man bygger in vissa automatiserade funktioner. Du ska ha olika typer av användare men jag tycker personligen att de är lite vanskligt. Men jag ser att trenden är att ”Pang” så kommer det ut en rapport eller en analys,

A: det leder in då också på nästa fråga : om kunskap blir mindre subjektiv om information analyseras automatiskt

I: Man kan ju inte lägga in sin egen affärs kompetens riktigt då man kanske då genom åren skaffat sig en jädra koll. När jag började bygga min första modell på Nordea så kanske den tog 3 veckor, den sista kunde jag bygga på nästan en timme men man lär sig under tiden. Man lär sig hur optimalt hantera olika variabler för olika syften och analyser, och det tappar man ju när man automatiserar. Det är liksom någon som har programmerat några sköna regler, så här ska det vara liksom. Det tappar den mänskliga insikten och logiken.

A: Hur kan kunskap om människors beteenden, det kognitiva perspektivet, vara till nytta för olika tekniker, modeller, metoder av DM?

I: Ni har kanske hört talas om webmining, att man analyserar det som händer på hemsidor. Då kan man fånga hur personer har rört på sin mus, hur dom har pekat och den informationen vill man också använda för att nästa gång jag kommer tillbaks så är det här det som ska highlightas. Jag tror att det är något som kanske kommer att komma mer i framtiden. Men det kräver också ett system för att ta in den informationen på något smart sätt. Jag tror att det kommer massor framförallt inom hemsidor, den biten där man har såna krav.

A: Det är bara en hypotes som vi har kollat runt om också spolat tillbaka till just att data skulle bli mer användbar om människor interagerade med sidor och liknande.

I: Definitivt, det man vill göra är att fånga in livsrörelser från hemsidor. Det kan ju används för att förbättra och öka mervärdet på hemsidan. Så att man tycker att den är bättre, så att man vill komma tillbaka. Det är något jag tror kommer att komma.

A*: jo det är kanske inte så stort nu men det är något som vi har intresserat oss från början faktiskt,*

A: Det kanske används inom företaget och sen användes mot kunder. Två olika delar, var tror du att det används mest ?

I: Tittar man på företag rent globalt så är det dom företagen som har många kunder, många kundrelationer, dom behöver använda sig av sofistikerade analyser för att kunna analysera detta. Då har banker och Telecom övertiden kommit ganska långt för dom har mycket kunder och dom har mycket interaktioner. Vi ringer ofta, vi gör många transaktioner på våra konton tillskillnad från försäkringsbolag som också har mycket kunder men där är det ganska sällan man har kontakt med sitt bolag. Det är kanske en gång per år eller när man betalar sin räkning eller när man råkat ut för en olycka. Det är liksom inte alls på samma sätt. Där har man liksom inte kommit lika långt, det är ju mer utmanande där. Men sen börjar det komma inom handeln och tillverkningsindustrin där det handlar om att hitta nya optimala tillverknings processer och ha system för att ha koll på när man ska börja underhålla olika delar i sina maskiner så att inte hela tillverkningsprocessen stannar. Så där tror jag att det kommer att komma mer inom tillverkning, det har jag sett på konferens. Att det är flera i tillverkningsindustrin som börjar tillämpa denna typ av analys.
Men fortfarande är det på strukturerad data. Men text mining har ju funnits ganska länge men den börjar få mer gehör det är ju näst steg att börja analysera den. Sen kommer då nästa steg det är ju att börja med voice mining, av inspelade samtal. Att ha algoritmer för att analysera detta. Även video mining är väl också något som man pratar om i framtiden som kanske kommer om några år, så hela tiden sker det en utveckling.

A: det var ju häftigt, det har vi inte hört talts om Voice och Video

A: Hur ser framtiden inom data minig ut relaterat till dagens syn förbättringar kring relevanta och säkra mönster?

I: Jag tror att det kommer bli bättre och säkrare än vad det är idag. Flera organisationer kommer att börja med detta och det kommer bli högre krav på dom som levererar system stöd och verktyg för detta.
Om man tittar på den tiden som jag har jobbat och på vad som har hänt så kommer det hela tiden nya aktörer och nya tillämpningar på det, detta är bara början.

A: Vilken del av processen kommer att vara viktigast från data kvaliten till algoritmer och liknande?

I: Jag tror generellt att algoritmerna, dom är inte så viktiga det är ju data kvalitén
Att man har tillgång till säker och god data som beskriver det man försöker åstadkomma. Algoritmerna är ju det enkla och som hjälper en att ta fram associationerna mellan det man vill åstadkomma till dom förklarande variablerna. Men det handlar om att producera mer, snabbare och effektivare. Fortfarande kan man komma ganska långt med en gammal vanlig regressions analys.

A: så data kvalitet ses som en stort problem?

I: Ja för att den aspekten måste man ta in för att data växer ju hela tiden och om inte kvalitén är god blir det inget bra resultat. Men det jobbas mycket på att kvalitetssäkra som affärsanalys att datakvalitet är viktigt.
Det kommer komma fram ännu mer i framtiden, skulle jag tro
Flera organisationer börjar mer och mer tänka till hur man ska försöka säkerställa det. Vad behöver man för systemstöd för att göra det? Jag tror att det är ett område som börjar ta fart.

A: Några speciella tankar på hur den ska förbättras?, kanske svår fråga.

I: Jag vet inte, Eniro där behöver dom bra data kvalitet, när dom har uppgifter som namn och telefonnummer. Dom har en lösing för att kunna säkerställa att det blir rätt i alla olika adress uppgifter , men jag tror att man behöver systemstöd för att göra det alltså .

A: Okej men då tror jag att vi har fått med alla frågor

I: det var ju jättebra.

A: Tack så mycket

Referenser

Ackoff, R. (1989): From Data to Wisdom, *Journal of Applied System Analysis*, Vol 16, pp 3-9

Backman, J. (1998): *Rapporter och Uppsatser*, Lund: Studentlitteratur

Batini, C. & Scannapieca, M. (2006): *Data Quality*: *Data-Centric Systems and Applications*, Berlin: Springer Heidelberg

Bellinger, G., Castro, D., & Mills, A. (2004), *Data, Information, Knowledge, and Wisdom*, www.systems-thinking.org/dikw/dikw.htm, < 2009-02-16 >

Birrer, A.J. F. (2005): Data mining to combat terrorism and the roots of privacy concerns, *Ethics and Information Technology*, Vol 7, No 4, pp 221-220

Bryman, A. (2001), *Samhällsvetenskapliga metoder*, 1:3, Malmö: Liber

Chen, MS., Han, J., & Yu, PS. (1996): *Data* Mining: An Overview from a Database Perspective*, IEEE Transactions on Knowledge and Data Engineering*, Vol 8, pp 866-883

Cooley, R., Mobasher, B., Srivastava, J. (1997): Web Mining: Information and Pattern Discovery on the World Wide Web, *Tools with Artificial Intelligence, Proceedings, Ninth IEEE International Conference*, pp 558-567

Daft, R.L. (2006): *Understanding the theory and design of organizations*, Mason: Thomson South-West

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases*, AI Magazine*, Vol 17, No 3, pp 37-57

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996b): The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, Vol 39, No 11, pp 27-34

Frawley, J. W., Piatetsky-Shapiro, G., and Matheus, J. C. (1992): Knowledge discovery in databases: An overview, *AI Magazine*, Vol 13, No 3, pp 57-70

Goebel, M. & Gruenwald, L. (1999): Survey of DM and Knowledge discovery Software Tools, *ACM SIGKDD Explorations Newsletter*, Vol 1, No 1

Guyon, I., Matic, N. & Vapnik, V. (1996): Discovering informative patterns and data cleaning, *Advances in knowledge discovery and data mining*, Menlo Park: American Association for Artificial Intelligence, pp 181-203

Herzog, N., Scheuren, J. & Winkler, E. (2007): *Data Quality and Record Linkage Techniques*, New York: Springer E-books

Jacobsen, D.I. (2002): *Vad, hur och varför? Om metodval i företagsekonomi och andra samhällsvetenskapliga ämnen*, Lund: Studentlitteratur

Kosala, R. & Blockeel, H. (2000): Web mining research: a survey, *ACM SIGKDD Explorations Newsletter*, Volume 2, No 1, pp 1-15, portal.acm.org

Kvale, S. (1997): *Den kvalitativa forskningsintervjun*, Lund: Studentlitteratur

Maimon, O., & Rokach, L. (2005): *Data mining and Knowledge Discovery Handbook*, New York: Springer E-Books Science & Business

Maletic, J.I., & Marcus, A. (2000): Data cleansing: Beyond integrity analysis, *In Proceedings of the Conference on Information Quality (IQ2000)*, Department of Mathematical Sciences, University of Memphis

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006): *Adaptive Business Intelligence*, Berlin: Springer-Verlag

McGarry, K. (2005): A survey of interestingness measures for knowledge discovery, *The Knowledge Engineering Review*, Vol 20, No 1, pp 39–61

Negash, S., & Gray, P. (2005), *Handbook on Decision Support Systems 2*, Part 7, pp 175-193, Berlin: Springer E-book

Pazzani, M. (2000): Knowledge discovery from data?, *IEEE Intelligent Systems*, Vol 15, No 2, pp 10-13

Piatetsky-Shapiro, G. (1991): Knowledge Discovery in Real Databases:
A Report on the IJCAI-89 Workshop, *AI Magazine*, Vol 11, No 5, pp 68-70

Rahm, E. & Do, H.H. (2000): Data Cleaning: Problems and Current Approaches, *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol 23, No 4, pp 1-11

Rowley, J. (2007): The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information Science*, Vol 33, No 2, pp 163-180

Silberschatz, A. & Tuzhilin, A. (1996): What makes patterns interesting in Knowledge Discovery Systems, *IEEE Transactions on Knowledge and Data engineering*, Vol 8, No 6, pp 970-974

Strong, D., Lee, Y., & Wang, R. (1997): Data quality in context, *Communications of the ACM*, Vol 40, No 5, pp 103-110

Van Well, L., & Royakkers, L. (2004): Ethical issues in web data mining*, Ethics and Information Technology*, Vol 6, No 2, pp 129-140, Netherlands

Vassiliadis, P. (2000): *Data Warehouse Modeling and Quality Issues*, Ph.D. Thesis, Department of Electrical and Computer Engineering, National Technical University of Athens, Greece

Wand, Y., & Wang, R.Y. (1996): Anchoring data quality dimensions in ontological foundations, *Communications of the ACM*, Vol 39, No 11, pp 86–95.

WinterCorp (2005): *TopTen Program,* www.wintercorp.com/PressReleases/ttp2005_pressrelease_091405.htm <2009-05-08>

Zeleny, M. (1987), Management Support Systems: Towards Integrated Knowledge Management, *Human Systems Management*, Vol 7, No 1, pp 59-70