Tieto tietojenkäsittelytieteessä

Jesse Hauninen

14.4.2008

Joensuun yliopisto
Tietojenkäsittelytieteen ja tilastotieteen laitos
Pro gradu -tutkielma

Tiivistelmä

Tiedosta kuulee puhuttavan jatkuvasti. Yhteiskunnan kehittyessä myös tiedon määrän sanotaan kasvavan. Sanonnan mukaan tieto on valtaa. Filosofit eivät kuitenkaan tähän tyydy. Tiedon ontologiasta ja epistemologiasta on käyty keskustelua antiikin ajoista lähtien aina tähän päivään saakka. Miten tieto on olemassa, miten tietoa voi saada vai onko ylipäätänsä mielekästä puhua tiedosta?

Tieteen tehtävänä on tuottaa uutta tietoa. Tieto muuttuu kuitenkin uusien tutkimustulosten myötä. Hyvän kuvan tästä saa, jos seuraa ravitsemussuosituksia. Tutkimukset vaikuttavat antavan jatkuvasti ristiriitaista tietoa. Tieto ei kuitenkaan voi olla näin ristiriitaista. Ehdoton vaatimus totuudesta tarkoittaa, että väite on joko totta tai sitten se ei ole. Tiedollisen väitteen totuusarvoa ei voida jatkuvasti muuttaa, koska tällöin tiedon käsitteestä häviää merkitys. Siksi kuuleekin välillä puhuttavan informaatiosta ja datasta. Tiedon, informaation ja datan käsitteet kuitenkin sekoittuvat hyvin usein keskenään. Varsinkin suomenkielisessä tietojenkäsittelytieteessä puhutaan usein tiedosta tarkoittaen dataa tai informaatiota. Suomen ulkopuolella esimerkiksi englannin ja ruotsin kielessä ero vielä korostuu, koska siellä data tarkoittaa dataa ja tieto tietoa. On varsin kiusallista, että kotimaiset tiedemiehet eivät aina itsekään tiedä mitä termiä tarkoittavat.

Tutkielmassa käydään läpi yleisesti miten tiede tuottaa tietoa ja mitä käytäntöjä tieteessä vallitsee. Samalla pohditaan itse tiedon käsitteen mielekkyyttä ja verrataan sitä informaation ja datan määritelmiin. Tietojenkäsittelytiedettä käytetään esimerkkitieteenä ja tutkitaan sen eri prosesseja sekä prosessien tuloksia tietoteoreettisesta näkökulmasta. Tarkoituksena on osoittaa että termien käyttöä tulisi miettiä enemmän ja pohtia milloin todella olisi mielekästä puhua tiedosta, milloin informaatiosta ja datasta.

ACM -luokat (ACM Computing Classification System, 1998 version): A.0, E.0, H.0, H.1, H.2.8, K.4.0.

Avainsanat: Tieto, Data, Informaatio, käsiteanalyysi.

Sisällysluettelo

1. Johdanto

Vaikka sana tieto esiintyykin tietojenkäsittelytieteen nimessä, on sen merkityksen pohdinta jäänyt taka-alalle tietojenkäsittelytieteilijöille. Moni pitääkin termiä ”tieto” niin itsestään selvänä ettei sen pohdintaa koeta mielekkääksi. Tarkempi tarkastelu osoittaa kuitenkin ettei näin ole. Filosofit ovat jo antiikin ajoilta pohtineet mitä tieto on, ja voiko meillä ylipäätänsä olla tietoa. Tähän päivään mennessä ei ole kyetty esittämään aukottomia perusteita edes tiedon määritelmälle ja alkuperälle. Tutkielmassa esitellään filosofien antamia erilaisia määritelmiä tiedolle sekä niiden puutteita. Lisäksi pohditaan hieman miten tietoa on mahdollista saavuttaa esittelemällä rationaalisesti tapahtuvan päättelyn ja empiirisen kannan hyödyt ja haitat.

Tieteen tärkein tehtävä on tuottaa tietoa ympäröivästä todellisuudesta. Pääsääntöisesti tietoa tuotetaan tutkimalla tarpeeksi yksittäisiä ilmiöitä ja niistä laaditaan yleispäteviä sääntöjä, joita myös tiedoksi nimitetään. Historia on kuitenkin usein osoittanut, että tieteen on vaikea, ja käytännössä mahdotonta, täyttää tiedolle esitetty tosi vaatimus. Tämän vuoksi onkin mielekkäämpää puhua tieteen tuottavan erilaisia todellisuutta koskevia hypoteeseja. Tutkielmassa pohditaan onko tieteen edes mahdollista tuottaa tietoa, millaisia ovat tieteen käyttämät metodit tiedon tuottamiseen ja tieteen tuottaman tiedon merkitystä.

Tutkielmani pääpaino on kuitenkin tietojenkäsittelytieteen tuottamassa tiedossa, jota pohdinnat tiedosta yleisesti ja tieteellisestä tiedosta täydentävät. Miten tieto ilmenee tietojenkäsittelytieteessä? Miten tiedon käsitettä on käytetty tietojenkäsittelytieteessä? Entä mikä on tiedon suhde dataan ja informaatioon? Kyseiset käsitteet esiintyvät usein rinnakkain tiedon kanssa, mutta ainakaan merkityksen perusteella ei voida sanoa käsitteiden olevan synonyymeja keskenään. Pyritään siis analysoimaan tietojenkäsittelytieteessä olevia tieto-sanoja verrattuna tiedon määritelmään, jotta lukija tietäisi milloin tietojenkäsittelytieteessä on kyse tiedosta, milloin datasta tai informaatiosta.

Tietojenkäsittelytieteilijälle tutkielma selkeyttää datan, tiedon ja informaation eron. Tästä on hyötyä varsinkin englannin- ja ruotsinkielisten tietojenkäsittelytieteilijöiden kanssa keskusteltaessa. Tutkielmasta voi olla myös apua omien tuloksien tulkintaan. Kun tiedetään mitä tiedolta vaaditaan on helppo suhtautua kriittisesti omiin koetuloksiin ja näin on mahdollista saada kokonaan uusia näkökulmia tulosten analysointiin.

2.Tieteellinentieto

Luvussa kaksi esitellään tieteellistä tietoa, sekä tarkastellaan miten tieteellinen tutkimus tapahtuu. Luku käsittelee yleisesti tieteen tapaa tuottaa tietoa, ei niinkään tietojenkäsittelytieteen näkökulmasta. Toki yhtäläisyyksiä tietojenkäsittelytieteen kanssa on, mutta luku on tarkoitettu antamaan taustatietoa lukua kolme ajatellen, jossa käsitellään tarkemmin tietojenkäsittelytieteen ja tiedon suhdetta.

Kappaleessa 2.1 käydään läpi tieteenfilosofiaa muun muassa esittelemällä tieteessä käytetyt päättelysäännöt, joilla tieteellisiä päätelmiä tehdään. Kappaleessa pohditaan myös tieteen teorioita: mihin teorioissa esitetyt termit viittaavat ja mitä itse teoriat ovat. Kappaleen lopuksi esitellään mitä erilaisia tapoja tieteellä on selittää ilmiöitä.

Totuusteorioita esitellään kappaleessa 2.2. Siinä käydään läpi kolme totuusteoriaa, totuuden korrespondenssiteoria, totuuden koherenssiteoria ja pragmaattinen totuusteoria. Korrespondenssiteoriassa verrataan teorian ja todellisuuden vastaavuutta, koherenssiteoriassa verrataan teorian yhteensopivuutta muihin teorioihin ja pragmaattisessa teoriassa yhteiskunnassa vallitseviin käytäntöihin.

Kappale 2.3 käsittelee lyhyesti empirismiä. Empirismin mukaan tietoa todellisuudesta saavutetaan aistihavainnoilla. Kappaleessa käydään läpi empirismin historiaa sekä eri suuntauksia.

Popperin näkemystä tieteellisestä tiedosta esitellään kappaleessa 2.4. Niin sanotun falsifikationismin mukaan tieteellinen tieto on voitava osoittaa vääräksi, ja näin ollen teoriat kehittyvät kohti totuutta. Meidän ei kuitenkaan ole Popperin mukaan mahdollista osoittaa koskaan teorian olevan lopullisesti totta.

Kuhn esitteli tieteen kehittyvän paradigmojen avulla, jotka vaihtuvat tieteellisten vallankumousten avulla. Mitä tieteelliset vallankumoukset ovat, miten ne syntyvät, kehittyvät ja mikä niiden merkitys on? Entä mitä paradigmalla tarkoitetaan? Näihin kysymyksiin vastataan kappaleessa 2.5 missä käydään läpi tarkemmin Kuhnin näkemystä tieteen kehittymisestä.

2.1 Tieteenfilosofia

Tieteenfilosofia on tietoteorian ala, jossa keskitytään tieteellisen tiedon luonteeseen, yleisiin metodologisiin ongelmiin, teorioiden ja todellisuuden suhteeseen, tiedon kasvun ongelmiin sekä tieteen ja yhteiskunnan väliseen suhteeseen (Hallamaa et al. 2002, s. 119). Deduktio ja induktio ovat keskeisessä asemassa tieteellisestä päättelystä puhuttaessa. *Deduktiivinen päättely* säilyttää

totuuden. Premisseistä, jotka ovat tosia, seuraa myös tosi johtopäätös. Jos on hyvät johtopäätökset pitää jotakin teoriaa totena, on myös samat perusteet pitää totena teoriasta deduktiivisesti tehtyjä päätelmiä (Räikkä 1991, s. 21-22, 24). *Induktio* päättelymenetelmänä ei välttämättä säilytä totuutta. Induktiossa tehdään yksittäisistä havainnoista johtopäätös ja näihin havaintoihin sisältyy aina virhemahdollisuus (Saarinen 1996, s. 261-262). *Induktivistinen tieteenkäsitys* perustuu ajatukseen, että induktiivinen yleistäminen on paras tapa tehdä tieteellistä päättelyä. Yleistyksen pohjana tulee kuitenkin olla tarpeeksi paljon luotettavia havaintoja, eikä päätelmiä saa tehdä kevein perustein (Määttänen 1995, s. 127). *Induktivismissa* korostetaan teorioiden yhteyttä empiirisiin havaintoihin ja pyritään näin sulkemaan pois turha spekulointi (Määttänen 1995, s. 127). *Hypoteettis-deduktivistisen tieteenkäsityksen* mukaan teoriat ovat hypoteeseja eli oletuksia, joille ei välttämättä tarvitse hankkia empiiristä tukea etukäteen, induktiivisen päättelyn keinoin (Hallamaa et al 2002, s. 128). Olennaista hypoteettis-deduktiivisessa päättelyssä on teorioiden testaaminen. Teorioita ei voida suoraan verrata todellisuuteen, koska niissä puhutaan asioista joita ei voida aistein havaita. Toki teoriat käsittelevät aisteinkin havaittavia asioita. Teorian testausta varten johdetaan eli dedusoidaan lauseita L teoriasta, jonka totuus on selvitettävänä (Hallamaa et al 2002, s. 128-129). Deduktiohan on totuuden säilyttävää päättelyä, joten mikäli teoria on tosi, on siitä deduktiivisesti johdetut seurauslauseetkin tosia.

Tieteessä *teoreettiset termit* viittaavat usein asioihin, joita ei voida havaita. Esimerkiksi elektroneja, ei kukaan ole omin silmin nähnyt. Toki kuka tahansa voi katsella elektronien jälkiä hiukkasilmaisimista, mutta mistä loppujen lopuksi tiedämme että kyseiset jäljet ovat juuri elektronien jättämiä? *Teorian* voi ajatella olevan joukko lauseita, joiden yhteys todellisuuteen välittyy tieteellisten havaintojen ja mittausten välillä (Hallamaa et al 2002, s. 127). Esimerkiksi elektronien tapauksessa teorioilla voidaan selittää elektronien massa, varaus ja spin. Näitä ominaisuuksia kukaan ei kuitenkaan voi havaita omilla aisteillaan, kuten näkö-tai kuuloaistilla. Avuksi tarvitaan tieteellisiä havaintovälineitä. Hiukkasilmaisimien tuottamien kuvien havaitsemiseen liittyy havaintojen teoriapitoisuuden ongelma. Kuvassa näkyvän jäljen osaa tulkita elektronien jäljeksi vain henkilö jolla on tietoa kyseisestä teoriasta ja kuvan tuottaneesta laitteesta. Alaan erikoistumaton ihminen ei ymmärrä kuvaa. Kuvien kanssa asetelma on analoginen, kaksitulkintainen (Määttänen 1995, s. 133-135). Yksi ratkaisu olisi tulkita teoriat osaksi tutkimusta, johon kuuluu oleellisena osana myös kokeellinen toiminta, tutkimusvälineiden avulla suoritettava todellisuuden manipulointi (Määttänen 1995, s. 135). Teoria ei siis olisi vain todellisuutta kuvaava lausejoukko. Näin ollen riittää, että elektronista tiedetään joitain ominaisuuksia, joiden perusteella voidaan rakentaa kyseisten ominaisuuksien tutkintaan soveltuva laite. Näin ollen käsitykset

elektronista voivat muuttua, mutta aina voidaan pitää kiinni siitä oletuksesta että tarkastelun kohteena on yksi ja sama olio. Jos elektroneja voi suihkuttaa niin oletettavasti niitä on olemassakin.

Yksi syy sille, ettei teoria välttämättä aina vastaa täysin todellisuutta on, että teoriat ovat usein idealisoituja (Yrjönsuuri 1996, s. 131). Toisin sanoen, teoria kuvaa idealisoitua todellisuutta, mistä on jätetty pois ”merkityksettömiä” tekijöitä. Teorian tarkoitus ei välttämättä ole kuvata todellisuutta aina kaikkine ominaisuuksineen tai piirteineen. Esimerkkinä Määttäsellä (2005, s. 137) on kaasujen tilayhtälö. Se kuvaa kaasujen lämpötilan, tilavuuden ja paineen välistä suhdetta. Lähestyttäessä asteikon ääripäätä ei kaasujen tilayhtälö enää pidä täysin paikkaansa. Tämä ei kuitenkaan tee kaasujen tilayhtälöstä hyödytöntä teoriaa, koska se kaikesta huolimatta kuvaa todellisuutta riittävän tarkasti.

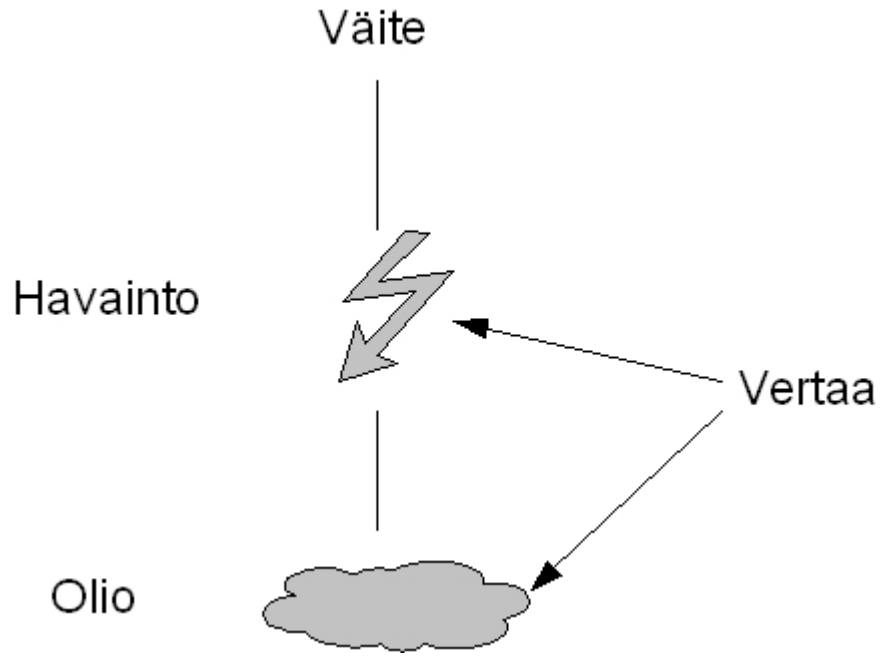
Tiede pyrkii selittämään ilmiöitä, ja antamaan näin tietoa maailmassa vallitsevista säännöistä ja laeista. Mitä tarkoittaa tieteen tapa selittää ilmiöitä? *Selittämisen klassisen mallin* muodosti viime vuosisadalla niin sanottu peittävän lain malli eli Hempel-Oppenheim- malli (Paakkola & Turunen 1995, s. 114-115). *Peittävässä mallissa* peitetään ilmiö alkuehtojen ja yleisten lakien alle. Esimerkiksi naapurin ammuttua kiväärillään luodin halutaan selitys sille, miksi luoti tipahtaa lopulta maahan. Vastausta tähän saadaan yleisellä painovoimalailla, jonka mukaan kaikki esineet joiden lähtönopeus on riittävän pieni tipahtavat lopulta maahan. Näin ollen tästä ehdosta seuraa loogisesti, että myös luoti tipahtaa maahan. Peittävää mallia on mahdollista täydentää kausaalisella selittämisellä. *Kausaalisuus* on syyn ja seurauksen yhteyttä, ja syy ennen seurausta antaa selityksen ilmiölle (Hallamaa et al 2002, s. 102). Esimerkiksi savu voi olla tulen seurausta. Kausaalisessa selityksessä ilmiö siis selitetään sitä edeltävällä syyllä, jonka seuraus ilmiö on. Funktionaalisessa selitysmallissa vedotaan ilmiön tarkoituksiin ja seurauksiin (Määttänen 1995, s. 149). *Funktionaalisessa selittämisessä* ei vedota ilmiön syihin, sillä ilmiö itse on syy. Esimerkkinä funktionaalisesta selityksestä voisi olla vastaus kysymykseen miksi ihmisellä on korvat. Vastauksessa vedotaan korvien toiminnan seurauksiin, korvat auttavat ihmistä kuulemaan. Toisin sanoen, vastataankin kysymykseen, mikä on korvien funktio.

2.2 Totuusteoriat

Klassisen tiedon määritelmän mukaan tieto on tosi, perusteltu ja uskomus. Tässä vaiheessa ei puututa sen tarkemmin muihin kuin totuusvaatimukseen. Tieteellisen tiedon on myös oltava totta, kuten klassinen tiedon määritelmä vaatii. Induktion muodostama havaintojen määrän ja totuuden ongelma kuitenkin tekee mahdottomaksi osoittaa tieteellisen tiedon ja totuuden yhtäläisyyden. Erilaisilla totuusteorioilla on pyritty osoittamaan perusteita tieteellisen tiedon totuudesta.

Seuraavaksi esitellään kolme totuusteoriaa: totuuden korrespondenssiteoria, totuuden koherenssiteoria ja pragmaattinen totuusteoria.

Tieteessä totuudella tarkoitetaan samaa kuin arkikielessä. Väite on tosi, jos se vastaa asioiden tilaa (Määttänen 1995, 142). Esimerkiksi väite ”jotkut karhut nukkuvat talviunta” pitää paikkansa jos maailmassamme on karhuja, jotka nukkuvat talven yli. Tällaista käsitystä nimitetään totuuden *korrespondenssiteoriaksi*, jolla on kannattajansa empiirisen tieteen piirissä, ja se on myös tieteellisen realismin kanta (Hallamaa 2002, s. 67). Kuvassa 1 on kuvattuna väitteen syntyminen oliosta havainnoinnin kautta.



Kuva1: Väitteen syntyminen oliosta havainnon kautta (mukailtu Määttänen 1995, s. 142).

Hallamaa et al:n (2002, s.69) korrespondenssiteoria eroaa muista totuuskäsityksistä, erityisesti koherenssiteoriasta, ei-episteemisellä totuuskäsityksellä. Tällä tarkoitetaan, että totuuden ja tiedon käsite pitäisi erottaa toisistaan. Tällä Hallamaa tarkoittaa, että totuus on ihmisen tietämyksestä riippumatonta. Se mikä on totta tai epätotta, on totta tai epätotta siitä riippumatta voidaanko totuutta tai epätotuutta inhimillisin kyvyin tietää. Esimerkiksi Mount Everestin huipulla on lunta, riippumatta ihmisten tietämyksestä. Todellisuutta kuvaavien väitelauseiden totuusarvot, eli ovatko lauseet totta vai epätotta, ovat määräytyneet sen mukaan millainen maailma todellisuudessa on, vaikka totuusarvoja ei välttämättä pystytä koskaan ratkaisemaan.

Korrespondenssikäsitys ei kuitenkaan ole täysin ongelmaton. Empiristisen tietoteorian kannalta korrespondenssiteoriaan liittyy vertaamisen ongelma. Jos haluttaisiin tietää onko havainto

tosi, olisi havaintoa korrespondenssiteorian mukaan kyettävä vertaamaan todellisuuteen. Empiristi jonka mukaan havaitseminen on ainoa tapa saada tietoa todellisuudesta, joutuu verratakseen tekemään uuden havainnon (Määttänen 1995, 142). Pystymme vain havaitsemaan havaintoja. Kyseisiä havaintoja voi verrata keskenään, mutta mistä tiedämme, että vertaamisen vuoksi tehty havainto on tosi? Emme voi myöskään tietää että kyseinen havainto vastaa riippumatonta todellisuutta. Vertaamisen ongelma johtaa empirismissä siihen, että perimmäistä selvyyttä havaintojen ja todellisuuden vastaavuudesta eli korrespondenssista ei voida saada (Määttänen 1995, s. 142). Vertaamisen ongelma liittyy totuuden kriteerien etsintään. Korrespondenssiteorian kannattajat voivat sanoa, että totuus on mahdollista määritellä korrespondenssin avulla, mutta totuuden kriteerit pitää määritellä jotain muuta kautta (Määttänen 1995, s. 142).

Totuuden *koherenssiteorian* mukaan väite on tosi, mikäli se ei ole aiemman tiedon kanssa ristiriidassa. Totuus määritellään väitteen yhteensopivuudeksi, koherenssiksi, muun totena pidetyn tiedon kanssa (Yrjönsuuri 1996, s. 49). Vertaamisen ongelmaa ei koherenssiteoriassa muodostu, koska väitteitä voi verrata toisiin väitteisiin ja vastaavasti havaintoja toisiin havaintoihin. Määttäsen (1995, s. 143) mukaan koherenssiteoria sopii hyvin instrumentalismiin, jonka kannattaja voi myös sanoa tieteellisen teorian olevan tosi tai epätosi. Instrumentalistin ja tieteellisen realismin edustajan tarkoittava totuus tarkoittaa kuitenkin eri asiaa. Tieteellisen realismin edustaja siis kannattaa totuuden korrespondenssiteoriaa ja instrumentalisti koherenssiteoriaa. Teorian totuudesta tai epätotuudesta puhuminen ei välttämättä tarkoita, että puhuja väittäisi teorian kuvaavan riippumatonta todellisuutta oikein tai väärin. Vaikka totuuden määritteleminen yhteensopivuudeksi tuntuisi meistä oudolta, ei koherenssi ole arkielämässä kuitenkaan outo asia. Esimerkiksi, jos joku väittää kesällä olevan pakkasta, harva viitsii vilkaista lämpömittaria ja tarkistaa väitettä, koska kyseinen väite on jo valmiiksi ristiriidassa sen kanssa mitä tiedämme kesästä. Koherenssi toimittaa tässä kuitenkin totuuden kriteerin virkaa, mikä sopii myös korrespondenssiteorian kannattajalle (Määttänen 1995, s. 144). Korrespondenssista totuuden ehdosta ei tarvitse luopua. Koherenssia käytetään tieteenteossa yhtenä totuuden kriteerinä. Mikäli tulos ei sovi yhteen vallitsevan teorian kanssa, on syytä selvittää onko työssä tehty virheitä. Jos uudelleentarkistuskaan ei ratkaise ristiriitaa aiemman vallitsevan teorian kanssa, joudutaan miettimään toisen muokkaamista tai hylkäämistä. Totuuden määritteleminen koherenssiksi ei kuitenkaan ole ongelmatonta. Määttänen (1995, s. 144) esittää, että kahden kilpailevan teorian olosuhteissa on epäselvää, kumpaan käsitykseen väitettä tulisi verrata. Väitteen totuus voi vaihdella riippuen siitä kumpaan teorian käsitykseen sitä verrataan. Tällainen epämääräisyys ei ole eduksi totuusteorialle. Yrjönsuuri (1996, s. 49) huomauttaa ettemme pidä ristiriidatonta kuvitteellista tarinaakaan totena. Totuudelta vaaditaan enemmän kuin pelkkää

väitteiden keskinäistä yhteensopivuutta, sillä totuus liittää ymmärryksemme maailmaan.

Viimeisenä totuusteoriana esitellään *pragmaattinen totuusteoria*. Siinä totuus perustuu tiedon ja toiminnan väliseen suhteeseen. Totuus määritellään käyttökelpoisuudeksi, eli väite on tosi jos se mahdollistaa menestyksellisen toiminnan (Saarinen 1996, s. 280-281). Pragmaattisuus eli toimivuus totuuden määritelmänä on kuitenkin ongelmallista. Vääränkin teorian avulla on mahdollista toimia menestyksellisesti. Esimerkiksi joku voi mennä naimisiin rikkaan kanssa sillä perusteella, että raha takaa rakkauden säilymisen suhteessa. Rakkaus voi toki rahankin avulla säilyä, mutta teorian totuus kuitenkin on kyseenalainen. Näin ollen käyttökelpoisesta totuuden kriteeristä ei kannata tehdä totuuden määritelmää. Määttänen (1995, s. 144) esittää realismin kiistäjälle koherenssin ja toimivuuden tarjoavan mahdollisuuksia totuuden kehittämiselle, mutta realismin kannattajalle ne sopisivat lähinnä totuuden kriteereiksi. Pragmatismin vaikutus näkyy erityisen voimakkaana analyyttisessä filosofiassa (Saarinen 1996, s. 280-281). Pragmatistisesti sävyttyneessä ajattelussa korostetaan todellisuuden kuvailuun ja selittämiseen käytettyjen käsitteiden ja teorioiden sopimuksenvaraisuutta. Samoin pragmaatikot korostavat ettei ole olemassa mitään ehdottoman varmaa perustaa tiedolle, vaan että kaikki tieto on olennaisesti epävarmaa (Saarinen 1996, s. 280-281). Saarisen (1996, s. 280-281) mukaan tällä pragmaatikot tarkoittavat, että käsitteet, teoriat ja tiedoksi kutsutut käsitykset ovat pragmaatikoille erilaisiin tarkoituksiin valittuja välineitä, joita voidaan tarpeen mukaan vaihdella.

2.3 Empirismi

Empirismi on tietoteoreettinen näkemys, jossa korostetaan tiedon ja kokemuksen välillä vallitsevaa suhdetta. Tietoa syntyy aistihavainnoin ja kokemusperäisellä tutkimuksella (Runes 1983, s. 104-105). Empiristit pitävät riittämättöminä rationalistien järjen käsitteitä, intuitioita, sisäisiä näkemyksiä tai uskontoja. Varsinainen tieto saadaan aistihavainnoista ja kokemuksista induktiivisen päättelyn avulla. Induktion perusteella yksittäisistä tapauksista pystytään johtamaan yleisiä päättelysääntöjä. Esimerkiksi biologin tehtävänä voi olla selvittää minkä värisiä korpit ovat. Hän menee lintutorniin ja havaitsee tuhat mustaa korppia. Tästä hän tekee johtopäätöksen, että kaikki korpit ovat mustia. Edellä mainitun esimerkin perusteella myös tiedemiehet toimivat. Heillä on joukko yksittäistapauksia, joista pitäisi tehdä jonkinlainen yleinen johtopäätös. Induktiivisen päättelyn ongelmana on, että miten paljon pitäisi havaintoja olla luotettavan tiedon saavuttamiseksi. Biologihan voikin huomata seuraavan korpin olevan valkoinen, jolloin hänen aikaisemmin tekemänsä johtopäätös kumoutuu. Mikäli tiedolta edellytetään ehdotonta totuutta pitäisi induktiivisen päättelyn olla pysyvää, eikä tieteelle tyypillistä myöhemmin itseänsä korjaavaa.

Klassisen empirismin juuret ulottuvat antiikin Kreikan sofisteihin asti. Nykyään tunnetuimpia klassisia empiristejä ovat Aristoteles ja Tuomas Akvinolainen. John Locke muotoili perinteisen empirismin opinkappaleet 1600-luvulla. Hän kiisti valmiit, sisäiset, ideat ja esitti että ihmismieli on tyhjä taulu, tabula rasa, mihin kokemus painaa jälkensä. Muita kuuluisia perinteisen empirismin edustajia ovat George Berkeley ja David Hume (Runes 1983, s 104-105 & Flew 1979, s. 104-105). Empirismi on totuttu asettamaan vastakkain rationalismin ja sen eri muotojen kanssa, erityisesti mannermaisen rationalismin kanssa, jossa esitetään että suurin osa tiedostamme perustuu järkeen, aisteista riippumatta. Nykyisin ero rationalismin ja empirismin välillä on kuitenkin kaventunut, sillä on huomattu monien rationalistien kannattaneen myös empiristisiä teorioita.

Antiikin Kreikassa 400-luvulla eläneet sofistit olivat ensimmäisiä empiristejä (Flew 1979, s. 330). He hylkäsivät esisokraattiset rationaaliset maailmanselitykset, ja keskittyivät konkreettisiin asioihin kuten yhteiskuntaan ja ihmisiin. Sata vuotta heidän jälkeensä Aristoteles esitti kritiikkinsä Platonin ideaoppia vastaan ja syntyi induktiiviseksi päättelyksi kutsuttu menetelmä, jota käyttämällä nykyaikainen tiede pyrkii tuottamaan tietoa. Samalla määriteltiin empirismin perusajatus Aristoteleen toimesta, jonka mukaan inhimillinen tieto todellisuudesta perustuu kokemukseen (Flew 1979, s. 25-27). Aristoteleen jälkeen stoalaiset ja epikurolaiset muotoilivat selvemmin empiristisiä selityksiä ideoiden ja käsitteiden muodostamiselle. Stoalaiset esittivät kuinka kokemus painaa jälkensä mieleemme ja epikurolaiset opettivat miten oliot tuottavat aistimuksia (Flew 1979, s. 108 & 339). Keskiajalla Tuomas Akvinolainen esitti, että jumalan olemassaolo voidaan todistaa aistihavainnoista (Saarinen 1985, s.100).

Brittiläiset empiristit, Locke, Berkeley ja Hume kävivät keskustelua sisäsyntyisistä ideoista muiden empiristien kanssa 1600-1700- luvuilla. Locke esitti, että kaikki tieto on kokemukseen perustuvaa. Ihmisen mieli on tyhjä taulu johon kokemus jättää jälkensä. Berkeley päätteli, että Locken näkemykset voisivat johtaa ateismiin jolloin hän esitti teorian, että olemassaolo perustuu havaitsemiseen ja havaituksi tulemiseen (Saarinen 1985, s. 206-207). Kun ihminen ei ole havaitsemassa, suorittaa Jumala havaitsemisen eikä kappaleiden olemassaolo muutu. Kaikki järjestys minkä ihmiset luonnossa näkevät, on Jumalan kieltä tai käsialaa Berkeleyn mielestä (Saarinen 1985, s. 206-207). Hume ei kyseistä näkemystä hyväksynyt, vaan esitti että meillä on tietoa vasta kun voimme osoittaa sen perustuvaan suoraan ja välittömään kokemukseen. Käytännössä tämä osoitus on mahdotonta, jolloin suurin osa antiikin filosofeista nykyajan filosofeihin sai tylyn tuomion, ja metafysiikka kyseenalaisen maineen.

Fenomenalistien mukaan kaikki fyysinen, kuten oliot ja ominaisuudet, voidaan supistaa mielessä oleviksi olioiksi, ominaisuuksiksi ja tapahtumiksi (Flew 1979, s. 266). Lopulta on

olemassa vain mielen tapahtumat, oliot ja ominaisuudet. Esimerkiksi erilaiset aistikokemukset kuten kuuleminen ja näkeminen jostain tietystä objektista kuuluvat vain tietyntyyppisten kokemusten joukkoon. Tässä mielessä vallitseva kokemusten joukko on muuttumaton ja johdonmukainen, toisin kuin esimerkiksi joukko johon aistiharhat kuuluvat. Olemme tekemisissä vain todellisuutta koskevien ilmentymien kanssa, emme todellisuuden kanssa sellaisenaan. Esimerkiksi Mill katsoi, että mikään tieto ei tule suoraan kokemuksesta, vaan kokemuksesta tehdyistä induktiivisista johtopäätöksistä (Flew 1979, s. 266).

Loogiset empiristit esittivät 1900 -luvulla, että filosofian pitäisi tukea ja selventää tiedettä, esimerkiksi tieteessä käytettyjä käsitteitä ja oivalluksia (Runes 1983, s. 302). Heidän tarkoituksensa oli luoda uusi täydellinen kieli, josta puuttuisi luonnollisen kielen epämääräisyydet ja muodottomuudet, jotka aiheuttavat metafyysisiä näennäisongelmia, sekä sekaannuksia käsitteiden käytössä. Lauseet jaettiin mielettömiin ja mielekkäisiin, niin sanotun todentamisperiaatteen mukaisesti. Lauseet jotka eivät olleet puhtaasti loogisia tai joita ei pystynyt suoraan aistein todentamaan, olivat mieltä vailla. Filosofian perinteiset ongelmat muuttuivatkin näin ollen näennäisongelmiksi.

Edellä esitettyjen erilaisten empiristisen näkemysten perusteella empiristiset teoriat voisi jakaa kahteen eri kategoriaan:

1) Jyrkkään muotoon, jossa kaikki inhimillinen tieto perustuu kokemukseen. Tähän kategoriaan kuuluu esimerkiksi looginen positivismi. Määritykset olisi voitava kääntää havaintokielelle tai määritellä havaintokäsitteiden kautta. Jyrkässä muodossa ei hyväksytä teoreettisia käsitteitä, eli käsitteitä joita ei ole voitu todentaa havainnon kautta (geenit, virukset, atomit).
2) Naiivin empirismin kannattajat uskovat että ideat ja teoriat pitää pystyä testaamaan todellisuudessa, eivätkä ennalta omaksutut ajatukset saa niihin vaikuttaa.

2.4 Popperin falsifikationismi

Induktion, ja tieteen, ongelmana on pidetty tiedon epäluotettavuutta suhteessa uusiin havaintoihin. Useinhan käy niin, että uusi teoria jollei suoraan kumoa, niin ainakin korjaa vanhaa olemassa olevaa teoriaa. Tämä vie tieteeltä uskottavuutta, sillä onhan tiedettä pidetty yhtenä tapana tuottaa tietoa. Miten monta havaintoa tiedemiehen tulisi tehdä, jotta havainnoista voitaisiin tehdä yleinen johtopäätös? Esimerkiksi riittääkö korppien värejä selvitettäessä tuhannen, miljoonan vai miljardin korpin havainnointi? Mitä enemmän havaintoja on, sitä luotettavampi väittämästä tulee, mutta kaikkea on kuitenkin mahdoton havainnoida. Sir Karl Popper myönsi tämän ongelman. Hän

tyytyi toteamaan ettei mitään faktaa tai fiktiota voida todistaa lopullisesti oikeaksi. Aina on mahdollista, että löytyy uusi havainto, joka kumoaa aiemman teorian (Paakkola & Turunen 1995, s. 95). Ei siis ole olemassa vastausta induktio-ongelmaan. Tulisikin luottaa edes teoriassa kumottavissa oleviin malleihin. Popperin mukaan tieteelliset teoriat voivat olla peräisin metafysiikasta tai henkilön omista psykologisista lähteistä (Geneshan & Kleindorfear, 1993). Olennaista Popperille ei ole mistä teoriat ovat peräisin, vaan miten ne toimivat käytännössä. Tieteen tehtävä on näin ollen testata empiirisesti teorioiden toimivuus käytännössä (Paakkola & Turunen 1995, s. 96). Jotta teoriaa voisi pitää totena, olisi se voitava osoittaa vääräksi. Aiemman käyttämäni esimerkin, että havaintojen perusteella on tehty päätelmä korppien olevan mustia, voisi falsifoida löytämällä sinisen korpin. Mikäli teoriaa ei voida osoittaa vääräksi, ei se täytä *falsifioinnin* vaatimuksia. Esimerkiksi väittämää violetteja korppeja on olemassa ei voida osoittaa vääräksi, koska väittämästä ei selviä milloin ja missä violetteja korppeja esiintyy.

Popperin teoria kumoaa verifikationismin, jonka mukaan tieteelliset teoriat on voitava osoittaa todeksi havaintojen avulla. Verifikationismi on osoittautunut kuitenkin liian tiukaksi, koska emme pysty esimerkiksi havainnoimaan kaikkia korppeja nyt, ennen ja tulevaisuudessa ja osoittamaan että ne ovat mustia. Falsifikationismin voisi ajatella johtavan skeptismiin, eli emme voisi tietää mitään varmaksi. Popperin mielestä näin ei ole, vaan tiede oppii aina virheistään ja uudet teoriat ovat aina lähempänä totuutta kuin aiemmat teoriat olivat (Geneshan & Kleindorfear, 1993). Popperin tarkoituksena ei siis ollut puolustaa skeptismiä, vaan kiinnittää huomiota teorioiden kritisoimiseen (Geneshan & Kleindorfear, 1993). On myös hyvä erotella toisistaan tiedon tavoittelu ja tiedollisen varmuuden tavoittelu. Ihminen on erehtyväinen, mutta tämä ei kuitenkaan estä tiedon tavoittelua, vaikka tiedolta ehdotonta totuutta vaadittaisiinkin. Kotkavirta (1999, s. 134) toteaa ettei Popperin falsifikationismi onnistu kuitenkaan ratkaisemaan induktion ongelmaa, eli että se osoittaisi ne perusteet joiden nojalla päättelemme tähänastisista kokemuksista tulevia tapahtumia. Popperin näkemys, jonka mukaan toistaiseksi falsioitumattomat väitteet ovat empiirisesti tosia muotoilee ongelman uudella tavalla. Miksi tähänastisten kokemusten pitäisi päteä myös tulevaisuudessa? Perustuuko luottamuksemme asioiden kulkuun tulevaisuudessa enemmänkin tottumukseen kuin vain järkeen?

Popperin falsifikationismi soveltuu hyvin tietojenkäsittelytieteeseen. Laatimamme ohjelmat eivät koskaan ole täydellisiä, tai ainakin niiden osoittaminen täydellisiksi on mahdotonta. Voimme osoittaa ohjelmamme toimivan tietyillä syötteillä tietyssä tilanteessa tietyllä alustalla, mutta kaikkien syötteiden testaaminen kaikissa tilanteissa jokaisella mahdollisella erilaisella tietokoneella on mahdotonta. Ohjelmassa olevat virheet voivat myös olla sopusoinnussa keskenään siten, että

toinen kumoaa toisen vaikutuksen ja ohjelma näyttää toimivan. Kuitenkin tilanne jossa ohjelma voidaan osoittaa täysin toimivaksi on yhtä mahdoton kuin induktion ongelma. Ohjelmia korjataan päivityksillä aina virheiden löydyttyä, joten niiden lähestyminen virheettömyyttä on verrattavissa tieteen lähestymistä kohti totuutta. Ei voida kuitenkaan osoittaa, että tieteen saama tulos olisi se lopullinen totuus, samoin kuten ei voida osoittaa ohjelman päivityksen tekevän ohjelmasta täysin virheettömän. Aina on muistettava etteivät korjaukset välttämättä paranna ohjelman toimivuutta. Ohjelmien korjaus voikin tuoda mukanaan uusia ongelmia, samoin kuten teoriaa vastaan voidaan esittää aiheetonta kritiikkiä joka vie vain kauemmas totuudesta.

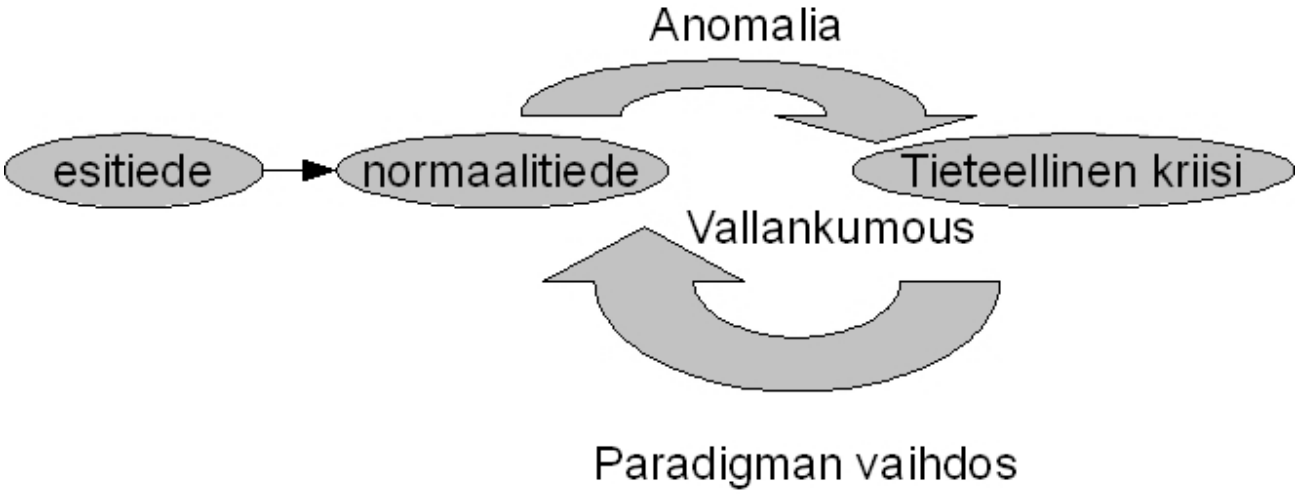
2.5 Paradigmat

Käsitteen paradigma esitteli Thomas Kuhn teoksessaan tieteellisten vallankumousten rakenne (Kuhn 1994, s. 56). *Paradigmalla* hän kuvasi tieteessä olevia tiettyjä vakiintuneita toiminnan tapoja (Kuhn 1994, s.36-42, 191-197):

- Tarkkailtava kohde
- Tutkimusaiheeseen liittyvät kysymykset
- Miten tutkimusaiheisiin liittyvät kysymykset esitetään.
- Miten saadut tutkimustulokset tulkitaan.

Kuhnin mielestä tiede ei kehity yksittäisillä teorioilla vaan suurilla muutoksilla, vallankumouksilla. Pikkuhiljaa tulee vastaan ongelmia, anomalioita joita ei pystytä enää vanhalla paradigmalla selittämään. Tämä aiheuttaa vallankumouksen ja uuden paradigman synnyn (Kuhn 1994, s. 94-98, 102).

Kuvassa 2 esitellään Kuhnin näkemys tieteen kehittymisestä. Kehittymässä olevasta tutkimusalueesta ei vielä puhuta tieteenä vaan *esitietieteenä* (pre-science). Siinä kilpailevat joukko erilaisia teorioita ja selityksiä. Esimerkiksi tietojenkäsittelytieteessä 1900-luvulla oli lukuisia erilaisia lähestymistapoja automaattiseen laskentaan. Monet näistä teorioista selittivät ja ratkaisivat ongelman yhtä hyvin, jolloin niistä olisi voinut tulla tieteellinen paradigma. Kuhnin mukaan esitieteessä tiedon hankkiminen on satunnaista ja rajoittuu jo saatavilla olevaan tietoon (Kuhn 1994, s. 28).



Kuva 2: Tieteen kehitys Kuhnin mukaan (käännetty Tedre 2007, s. 122).

Kuhn esitti, että se teoria mikä esitiedevaiheessa vaikuttaa muita paremmalta, kerää myös muita enemmän kannattajia. Kun selvästi muita enemmän kannatusta saanut teoria vakiintuu ja on saavuttanut yhteisymmärryksen tiedemiesten keskuudessa, on siirrytty esivaiheesta *normaalitieteeseen* (Kuhn s. 31-32). Tiedemiehet jotka hyväksyvät uudet teoriat, uskomukset ja tutkimustavat harjoittavat uuden paradigman mukaista normaalitiedettä (Kuhn 1994, s. 29-35). Normaalitieteen muodostanut uusi paradigma koostuu siis yleisistä teoreettisista laeista ja uskomuksista, sekä tekniikoista näiden soveltamiseksi. Uusi paradigma ei kuitenkaan pysty ikinä selittämään kaikkia vastakkaisuuksia, toisaalta sitä ei vielä alussa siltä vaaditakaan (Kuhn 1994, s. 30).

Oli paradigma miten hyvä tahansa, törmää se ennemmin tai myöhemmin ongelmiin joihin ei ratkaisua löydy nykyisellä, normaalitieteen, paradigmalla. Näille ongelmille Kuhn antaa nimeksi *anomaliat*. Kun anomalioita ilmenee tarpeeksi monta, rapistuu tiedemiesten usko nykyiseen paradigmaan ja tiede ajautuu kriisiin (Kuhn 1994, s.80).

Tieteen kriisissä eri teoriat kilpailevat siitä, mikä niistä osaa parhaiten selittää anomaliat. Useat uudet teoriat pyrkivät ainakin osittain selittämään anomaliat. Kun kilpaileva paradigma kerää taakseen muita enemmän tutkijoita, voidaan puhua tieteellisestä vallankumouksesta. Vasta kun viimeisetkin vanhan ajatusmallin kannattajat ovat siirtyneet kannattamaan uutta paradigmaa tai uuden sukupolven tutkijat ovat syrjäyttäneet vanhan paradigman kannattajat, voidaan puhua paradigman vaihdoksesta (Kuhn 1994, s. 165-169).

2.5.1 Paradigmojen ominaisuuksista

Kuhn esittää, että tutkijoilla on usein niin tarkat ennusteet kokeiden tuloksista, ettei tutkimustulosten tutkiminen ole enää mielekästä. Sen sijaan itse tutkimusmetodit ovat hänestä mielenkiintoisimmat itse tutkimuksissa (Kuhn 1994, s. 64-65, 74-77). Kuhn vertaakin tieteessä tapahtuvaa tutkimusta palapelinkokoamiseen (puzzle-solving) (Kuhn 1994, s. 49-55). Palapelissä on tietyt säännöt miten palaset sopivat yhteen sekä millaisen lopputuloksen pitäisi olla, kuten tieteessäkin. Paradigma määrittelee miten eri palojen tulisi yhteen sopia. Mikäli tulos ei olekaan ennusteen kaltainen, pyritään selityksillä kiertämään ongelma. Esimerkiksi tutkittava ongelma ei kuulukaan kyseiseen tieteenalaan tai ongelman ratkaisemista pidetään vain resurssien tuhlaamisena. On selvää, ettei tieteenala kehity tärkeiden kysymysten jäädessä vaille vastausta.

Kuhnin mukaan palapelin kokoamiseen verrattavissa oleva tiede ei tuota uutta tietoa. Koska kaikki paradigmaan sopimattomat kysymykset sysätään pois, ei myöskään synny uusia mullistavia totuuksia tai teorioita. Toisaalta kun resursseja käytetään vain yhteen hyvin tunnettuun alaan, johtaa se nopeampaan ja syvällisempään kehitykseen.

Kuhn esittää tieteessä olevan kolme erillistä painopistettä (Kuhn 1994, s. 38-41). Kyseiset painopisteet eivät välttämättä riipu toisistaan. Ensimmäinen painopiste on paradigmasta saadut totuudet, jotka kertovat asioiden luonteista. Näiden totuuksien avulla tutkija pystyy tarkentamaan ja laajentamaan teorioitaan koskemaan uusia ilmiöitä, jotka liittyvät samaan alaan.

Toisen painopisteen muodostamat totuudet ovat suoraan verrattavissa ja rinnastettavissa paradigmassa oleviin teorioihin. Paradigmassa kehitetyt aiemmat teoriat saattavatkin ratkaista ennen mahdottomina pidettyjä ongelmia. Esimerkiksi heuristiikalla on pyritty ratkaisemaan kauppamatkustajan ongelmaa.

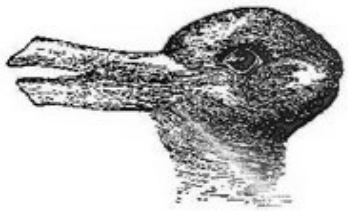
Kolmannen painopisteen muodostavat tutkimusmetodit. Käytännössä tämä tarkoittaa sitä empiiristä työtä joilla osoitetaan paradigman ja käytännön vastaavuus, ratkaistaan empiirisesti teorioissa olevat epäselvyydet sekä pyritään ratkaisemaan uudet ja yllättävät ongelmat. Esimerkiksi algoritmien aikavaativuuksien parantaminen, graafisten käyrien analysointi sekä käytettävyystestien järjestäminen kuuluvat kolmanteen tyyppiin. Tutkimusmetodien avulla paradigma lisää merkitystään myös uusilla tutkimusalueilla.

2.5.2 Tieteelliset vallankumoukset

Normaalitieteessä on pyritty rajaamaan ja supistamaan tieteellistä tutkimusta paradigman asettamiin rajoihin (Kuhn 1994, s. 48-50). Silti tutkijat törmäävät ennemmin tai myöhemmin

paradigmaan sopimattomiin ilmiöihin, anomalioihin. Kuhn kuitenkin toteaa, että mikäli tiedemies ei saa anomaliaa sovitettua vallitsevaan paradigmaan, pidetään tiedeyhteisössä anomalian aiheuttanutta tutkimusta virheellisenä. Tiedeyhteisössä ei hyväksytä ajatusta, että paradigma voisi olla puutteellinen (Kuhn 1994, s. 74-77). Mikäli tutkimustulos ei ole sovitettavissa paradigmaan, on vika tutkimustuloksessa, ei paradigmassa.

Kuhnin teorian mukaan tiedemiesten kohdatessa anomalioita jatketaan niiden tutkimista niin kauan, että anomalia ja sen aiheuttanut tutkimustulos on sovitettu paradigman kanssa yhteen (Kuhn 1994, s. 64-65). Toisin sanoen, niin kauan kuin tutkimustulosta ei ole sovitettu paradigmaan, ei siitä voida puhua tutkimustuloksena eikä tieteellisenä tosiasiana. Kuhnin oma esimerkki tutkimustuloksen ja paradigman aiheuttamasta epäyhteensopivuudesta on jänis-ankka kuva, joka on kuvattuna kuvassa 3. Siinä katsojan on mahdollista hahmottaa samasta kuvasta joko ankka, jänis tai molemmat. Tällä Kuhn pyrkii osoittamaan, että jos tiedemiehet vaihtavat tutkimusnäkökulmaa, he voivat nähdä tutkimustulokset hyvinkin erilaisina verrattuna aiempaan näkökulmaan.



Kuva 3: jänis-ankka kuva (Kuhn 1996, s. 114).

Mitä enemmän anomalioita löytyy, sitä enemmän tiedemiehet alkavat epäilemään käytettävissä olevan paradigman täydellisyyttä (Kuhn 1994, s. 78-80). Joitain anomalioita on mahdollista pakottaa paradigman luomaan muottiin, mutta mitä enemmän siihen pakotetaan anomalioita, sitä enemmän paradigma rakoilee tutkijoiden silmissä ja ajaudutaan kohti tieteellistä kriisiä. Osa tiedemiehistä pitää silti viimeiseen asti kiinni vanhasta paradigmasta vaikka se olisi selvästi puutteellinen ja anomalian paremmin selittävä teoria olisi olemassa (Kuhn 1994, s. 91-98). Kuhn kuitenkin toteaa, että vaikka kaikki tiedemiehet eivät hyväksyisi vanhasta paradigmasta luopumista, ei se silti ole este tieteelliselle vallankumoukselle. Enemmän kannatusta saava paradigma voittaa. Viimeistään siinä vaiheessa vaihtuu paradigma, kun häviävän paradigman kannattajat ovat kuolleet. Tällöin uusi sukupolvi korvaa vanhan (Kuhn 1994, s. 167-169).

3. Tieto tietojenkäsittelytieteessä

Tässä luvussa tutustutaan tietojenkäsittelytieteen tuottamaan tietoon. Tarkoituksena on käydä läpi miten tietojenkäsittelytieteessä tuotetaan tietoa, sekä tutkia eroavaisuuksia tavallisen ja tieteellisen tiedon välillä. Aikaisempien osien tapaan ei paneuduta syvällisesti mihinkään yksittäiseen osa-alueeseen, vaan pyritään esittämään jokaisesta asiasta pääkohdat. Näin lukijalle hahmottuu mahdollisimman kokonaisvaltainen kuva tietojenkäsittelytieteen tuottamasta tiedosta.

Kappaleessa 3.1 tutkitaan datan, informaation ja tiedon suhdetta. Lisäksi esitellään erilaisia määritelmiä kyseisille termeille. Samalla tehdään ero informaation ja tiedon välille.

Kappaleessa 3.2 käydään läpi Platonin luolavertausta tietojenkäsittelytieteen kannalta. Luolavertauksessa joukko luolaan sidottuja ihmisiä luulee auringosta heijastuvia varjoja todellisiksi. Voisiko tietojenkäsittelytieteessä olla varjoja joita tietojenkäsittelytieteilijät erehtyvät pitämään totena?

Kappaleessa 3.3 selvitetään miten tietojenkäsittelytiede tuottaa tietoa. Tiedon tuottamista tutkitaan käymällä läpi tiedonlouhinta ja asiantuntijajärjestelmät. Mitä kyseiset käsitteet tarkoittavat, mihin niitä käytetään ja mitä tekemistä niillä on tiedonlouhinnan kanssa? Ovatko niiden antamat tulokset tietoa?

Kappaleessa 3.4 esitellään tiedon pakkaamista. Tiedon pakkaamisesta esitellään kaksi erilaista tapaa pakata tietoa: häviöllinen ja häviötön tapa. Lopuksi pohditaan termin tiedonpakkaus mielekkyyttä ja tiedonpakkauksen vaikutusta pakattavaan tietoon.

Kappaleeseen 3.5 on niputettu joukko termejä, joissa esiintyy sana tieto, mutta jotka eivät ole kuitenkaan sen mielekkäämpiä syvällisempään analyysiin. Jokainen termi määritellään lyhyesti ja osoitetaan miksi kyseisen termin yhteydessä on puhuttu harhaanjohtavasti tiedosta.

3.1 Data, tieto ja informaatio

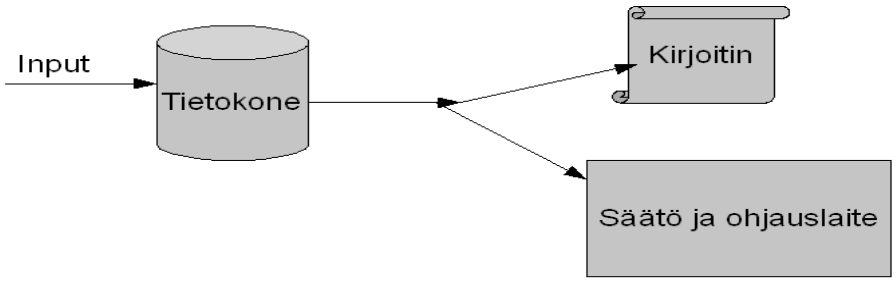
Tässä kappaleessa määritellään termit data, tieto ja informaatio. Samalla esitellään erilaisia informaation-ja tiedon lajeja. Kappaleen päätarkoituksena on kuitenkin tehdä ero näiden käsitteiden välille eli tehdä käsiteanalyysi määrittelemällä käsitteet yksikäsitteisesti. Näin lukija tietää milloin mitäkin käsitettä tulee käyttää ja miten käsitteet toisistaan eroavat.

3.1.1 Data

Termin data juuret ovat latinassa, jossa 'dare' tarkoittaa suomeksi antaa. Sen monikko on 'data' eli suomeksi annetut. Tämän perusteella dataa voitaisiin ajatella kanavaan syötettävänä

koodattuna viestinä (Ikeda & Konishi, 2007). Tilastotiedettä soveltavat yhteiskuntatieteilijät ovat jaotelleet datan vielä erikseen pehmeään ja kovaan dataan. Esimerkiksi tutkijan kokoama verbaalinen haastatteluaineisto on pehmeää dataa ja kvantitatiiviseen muotoon koodattu aineisto kovaa dataa (Huberman & Miles 2002, s. 81-83).

Tietokoneeseen esimerkiksi näppäimistöltä syötettyjä koodattuja viestejä nimitetään *dataksi*. Tietokoneen voidaan ajatella olevan erikoislaatuisella muistilla varustettu viestintäkanava, jonka toimintaa säädellään muistissa olevalla ohjelmalla (Konishi & Ikeda, 2007). Ohjelmasta kone saa käskyt joiden perusteella se käsittelee sisään syötettyä dataa (input). Tietokone voi joko tuottaa uuden kielellisen tuloksen (output), jossa on muutettu vastaanotettu viesti haluttuun muotoon, tai tietokone voi liittää saadun tuloksen johonkin kausaaliseen säätö- ja ohjausjärjestelmään. Prosessi saattaa usein alkaa alusta sillä tietokoneen antama tulos voi olla seuraavan vaiheen syötedataa. Jos tulosta rupeaa ihminen analysoimaan, tulee siitä informaatiota. Kuvassa 4 on kuvattuna miten tietokone käsittelee dataa.



Kuva 4: Datan käsittely tietokoneessa (Niiniluoto 1996, s. 12).

Yllämainitun perusteella tietokone on laite, joka suorittaa automaattista datan käsittelyä. Mekaaniset, nopeat ja virheettömät toiminnat mahdollistavat esimerkiksi tekstinkäsittelyn ja matemaattisen laskemisen (Konishi & Ikeda, 2007). Niiniluoto (1996, s. 12) toteaa tietokoneen käsittelevän dataa pelkkänä merkkijonona. Hän jatkaa, että esimerkiksi ATK-sanakirjassa data on määritelty ”asian säännönmukaiseksi esitykseksi viestitettävissä tai käyttökelpoisissa muodoissa”. Näin ollen tietokoneen lukema, käsittelemä ja tulostama data on viesti, jonkin asian esitys, jolla on sisältö. Tietokoneella itsellään ei ole tietoisuutta ja ”minuutta” eli se ei itse ymmärrä millaista tietoa sen piirisiruihin on pakattu. Tietokoneen kieli on verrattavissa eläimien käyttämään

kommunikointiin, eli indeksien tasolla tapahtuvaan signaalien siirtoon (Niiniluoto 1996, s. 12).

Data on usein käännetty suomen kielessä tiedoksi. Esimerkiksi data processing käännetään tietojenkäsittelyksi, data mining tiedonlouhinnaksi ja database tietokannaksi. Tämä on kuitenkin harhaanjohtavaa, sillä data ja tieto eivät ole sama asia. Esimerkiksi filosofien tieto-oppi ei ole dataoppi. Kun ajattelemme henkilöllä olevan tietoa jostakin emme ajattele automaattisesti että hänellä myös on dataa kyseisestä asiasta. Kielitoimisto on suositellut tässä yhteydessä tieto-sanan korvaamista sanalla ”anne” (Niiniluoto 1996, s. 13). Kuitenkin esimerkiksi anteidenkäsittelytiede ja annekanta kuulostavat hieman oudoilta. Pääsyy lienee siinä, että olemme jo tottuneita käyttämään tieto-sanaa harhaanjohtavasti tietojenkäsittelytieteessä.

Myös informaatio- sanaa on käytetty datan vastineena. Usein kuulee tietokoneen varastoivan ja käsittelevän informaatiota. Tässä yhteydessä informaatiolla tarkoitetaan informaation kantajia, eli merkkejä ja merkkijonoja joita kutsutaan myös dataksi (Loh et al. 2000). Niiniluoto (1996, s. 13) esittää, että yksi mahdollisuus datan kääntämiseksi olisi *merkki* minkä myös ATK-sanakirja hyväksyy merkityksessä ”pienin tiedon esityksen alkio, joka voi kantaa informaatiota”. Niiniluoto (1996, s. 13) esittääkin tietokoneen suorittaman datakäsittelyn olevan syntaksisella tasolla tapahtuvaa automaattista merkkien käsittelyä, eli tietokone on merkeillä operoiva laskulaite.

3.1.2 Tieto

Tässä kappaleessa käydään läpi tiedon eri lajit. Filosofisempaa pohdintaa tiedon alkuperästä ja ylipäätänsä sen olemassaolosta voi lukea kappaleesta yksi. Lammenranta (1993, s. 72) jakaa tiedon lajit kolmeen eri osaan:

1) *Tuttuustieto*, tunnetaan jokin asia tai ihminen. Esimerkiksi ”tunnen juoksija Hartosen”.
2) *Taitotieto*, osaaminen joka ilmaistaan usein olla verbin yhteydessä. Esimerkiksi ”osaan juosta”.
3) *Propositionaalinen tieto*, jonkin lauseen tai proposition tietäminen todeksi. Esimerkiksi ”tiedän, että nopeimmat juoksijat ovat kenialaisia”.

Lammenrannan määritelmiin voisi vielä lisätä omaksi kohdakseen metatiedon ja tieteellisen tiedon. Metatiedolla tarkoitetaan tietoa jostain toisesta tiedosta; esimerkiksi tiedostoissa olevat tiedostopäätteet kertovat tiedoston sisällöstä kuten sisältääkö tiedosto tekstiä vai onko se ohjelman käynnistystiedosto. Tieteellinen tieto on oma tapauksensa jo sen takia, että tieteelliseen tietoon kohdistetaan tavallisesta tiedosta poikkeavia vaatimuksia, kuten toistettavuus. Klassisen tiedon vaatimukset ovat seuraavat (Lammenranta 1993, s. 79): tieto on tosi, perusteltu uskomus. Erityisesti

tieteen tuottamalla tiedolla on vaikeuksia täyttää tiedolta vaadittu totuus. Miten osoitamme kokeellisesti jonkin vallitsevan asiantilan siten ettei seuraava tieteellinen koe joskus kumoa sitä? Mikäli tieteen tuloksista puhutaan tietona, täytyy sen myös täyttää tiedon vaatimukset. Tuntuu oudolta sanoa väitteen A olevan enemmän tietoa kuin väitteen B. Silti molemmat väitteet olisivat tietoa. Yksi mahdollisuus on ajatella A:n olevan paremmin perusteltua eli varmempaa tietoa kuin B:n. Tiedon vaatimuksena on perustelu, mutta eikö tällöin molempiin A:n ja B:n vaatimuksiin jo sisälly perustelu? Molemmat väitteethän ovat tietoa. Jos tietoa mitataan perustelujen määrällä joudutaan pohtimaan myös riittävien perustelujen rajaa. Tiedon paremmuus ei ole suoraan verrannollinen perustelujen lukumäärään. Klassisen tiedon vaatimus on selkeä, ja mikäli tieteellinen tieto ei sitä voi täyttää tulisi puhua tieteellisestä informaatiosta.

Taito ja osaaminen, taitotieto

Lammenranta (1993, s. 73) nimittää ihmisten ja eläinten käyttäytymiseen liittyviä kykyjä ja valmiuksia *osaamisiksi* tai *taidoiksi*. Niiniluodon (1996, s.20) mukaan elottomaan luontoon kyseisiä käsitteitä ei voida kuitenkaan hyväksyä. Tämän perusteella esimerkiksi auringolla ei ole kykyä polttaa ihoa, vaikka rannalla olija olisi varmasti eri mieltä Niiniluodon kanssa auringonpistoksen saatuaan. Esimerkkinä kyvyistä ja valmiuksista voisi olla ihmisen kyky puhua, kissan hämäränäkö tai lintujen muuttovaellus. Jotkin kyvyt voivat olla myös yksilöllisiä, kuten sirkuseläimillä olevat taidot, joita voi olla mahdotonta jälkipolville siirtää. Niiniluoto (1996, s. 20) esittää, että tällaiset kyvyt eivät ole taitoa vaan ”ehdollisia refleksejä eli omaksuttuja käyttäytymistottumuksia”. Ihminen on oppinut parantamaan omaa taitoaan erilaisten artefaktien avulla (Niiniluoto 1996, s. 19-20). Esimerkiksi töitä tehdessämme meillä on käytössämme artefakteja, kuten tietokoneet, tehostamassa työntekoamme. Apunamme on artefakteja myös vapaa-ajalla, kuten kahvinkeitossa kahvinkeitin. Niiniluoto (1996, s. 20) esittää, että myös artefakteilla on taitoja. Ne voivat olla esimerkiksi taitavia pesemään pyykkejä tai tietokoneen tavoin tehokkaita laskutoimituksissa.

Niiniluodolle (1996, s.20) taito on tiedon esiaste. Vastaavasti Lammenrannan (1993, s. 73) mukaan taitotieto on propositionaalisen tiedon erikoistapaus. Taidon omaavalta ei vaadita ymmärtämistä tai kielellistä kuvausta sen toiminnasta tai säännöistä. Niiniluoto esittääkin tällaisten taitojen perustuvan yritykseen ja erehdykseen, eikä julkilausuttuihin ohjeisiin. Esimerkkinä Niiniluodolla (1996, s. 20) on lapsen ensimmäisen kielen oppiminen. Pieni lapsi oppii kasvuympäristönsä kielen, sanaston ja kieliopin ilman, että hänelle kukaan opettaisi taivutusmuotoja ja lauseiden rakenteita koskevia sääntöjä. Tällöin lapsella on Niiniluodon mukaan *piilevää tietoa* (tacit knowledge), jonka mukaan hän toimii, mutta hän ei kykene ilmaisemaan sen sääntöjä tai sisältöjä.

Piilevä tieto on ilmaisun suhteen vastakohta propositionaaliselle tiedolle, joka on ilmaistu väitelauseiden muodossa (Niiniluoto 1996, s. 20). Piilevä tieto ilmenee ei-kielellisessä muodossa. Esimerkiksi urheilussa, käsitöissä ja taiteissa näkyy hyvin piilevän tiedon merkitys. Kuhn on esittänyt piilevän tiedon merkitystä tavassa, jolla tieteenharjoitus opitaan normaalitieteellisissä tutkimustraditioissa (Kuhn 1994, s. 36-46).

Taidon ja propositionaalisen tiedon eroa on pohdittu antiikin ajoista lähtien. Esimerkiksi Platon käytti aidon tiedon mallina tekijän tietoa (Kotkavirta 1999, s. 62). Lääkärillä on taito parantaa potilas eli lääkärillä on tietoa terveydestä. Veneenrakentajalla on taito tehdä vene joten hänellä on myös oltava tietoa veneistä. Toisaalta, vaikka katsoisimme opetusvideon sydänleikkauksesta, emme kuitenkaan osaisi kyseistä leikkausta suorittaa. Herääkin kysymys, että eikö tällöin tulisi puhua vain pelkästä kokemuksen tuomasta taidosta eikä tiedosta. Kotkavirran (1999, s. 60) mukaan on perusteltua kuitenkin puhua myös tiedosta, koska esimerkiksi käsityöläisen mestaruus ja hänen kykynsä opettaa muita perustuu viime kädessä siihen, että hän tietää oman tekemisensä idean. Lääkärillä on selkeä kuva siitä, miten sydänleikkaus tehdään onnistuneesti, ja jos hän on tarpeeksi taitava niin leikkaus todennäköisesti onnistuu.

Pragmatismissa korostetaan teorian ja käytännön välillä vallitsevaa yhteyttä. Kotkavirran (1999, s.59) mukaan ”kokemus on subjektin ja ympäristön jatkuvaa toiminnallista vuorovaikutusta, jonka kuluessa kumpikin muodostuu ja muuntuu jatkuvasti”. Toisin sanoen, tietoon liittyy aina toiminnallisia intressejä ja arvoja, ja tiedon oikeutus ja pätevyys ratkeavat toiminnan yhteydessä. Pragmatismissa tieto on välineellistä, sillä sen avulla järjestämme kokemuksiamme toiminnan kannalta tarkoituksenmukaisesti. Näin ollen totuutta ei ole mahdollista määritellä pelkästään tiedollisten kriteerien avulla, koska se liittyy aina myös toiminnallisiin päämääriin ja arvoihin (Kotkavirta 1999, s. 59). Arvojen Kotkavirta (1999, s. 59) esittää syntyvän ja muuttuvan kulloistenkin olosuhteiden mukaan. Muuttuvat arvot aiheuttavat sen, etteivät tiedolliset uskomukset ja totuudet voi olla yleispäteviä, koska ne saavat vahvistuksensa toiminnallisiin yhteyksiin liittyvistä kokemuksista. Peircen (Kotkavirta 1999, s. 59) mukaan todet uskomukset kuvaavatkin todellisia asiantiloja niin hyvin kuin toistaiseksi on mahdollista.

Aristoteelle taito on oikeajärkiperäistä tekemisvalmiutta johon liittyy käsitys siitä miten valmistettava tulos syntyy (Aristoteles 1989, VI. 4-5) . Eli meillä on aina enemmän tai vähemmän taitoon tai sen harjoittamiseen liittyvää käsittämistä tai tietoa. Aristoteles erottelee varsinaiset taidot mekaanisesta osaamisesta. Nykyään tällaista tietoa kutsutaan teknologiaksi. Sana on peräisin ”tekhnen logoksesta” eli ”oppia tekniikasta” (Niiniluoto 1996, s. 22). Teknologia on määritelty suppeasti suomen kielessä 1880-luvulla opiksi ”raaka-aineiden mekaanisista ja kemiallisista

jalostuskeinoista". Etymologisessa mielessä teknologiaa voidaan pitää kuitenkin minkä tahansa taidon alan tehostamisena (Niiniluoto 1996, s. 22). Näin ollen teknologiaan voidaan liittää yrityksen ja erehdyksen kautta saatuja osaamisen taitoja, kuten Hippokrateen lääkintäoppi ja Pythagoraan geometria. Aluksi näihin liittyvät säännöt perustuivat uskontoihin ja myytteihin, mutta nykyään tieteen tulokset ovat korvanneet ne. On syntynyt "tieteellistettyä teknologiaa", mitä ns. soveltavat tieteet (applied science, esim. Kasvatustiede) toteuttavat (Niiniluoto 1996, s. 22).

Propositionaalinen tieto

Propositionaalisen tiedon käsitteen mukaan tieto edellyttää kieltä, jonka avulla muotoillaan maailmaa koskevat väitelauseet (Hallamaa et al 2002, s. 106). Lammenranta (1993, s. 73) määrittelee propositionaalisen tiedon todellisuutta koskevaksi informaatioksi. Kieli voi sisältää esimerkiksi indeksejä, ikoneja tai symboleja kunhan merkkijonot voidaan tulkita maailmassa vallitsevia asiantiloja koskeviksi väitteiksi. Tästä seuraa, että tietoa voidaan ilmaista esimerkiksi kirjoituksilla, kuvilla, piirroksilla tai vaikkapa äänten avulla. Yleisin ilmaisuväline on luonnollinen kieli, vaikka tiede tarjoaakin useita vaihtoehtoisia tapoja. Luonnollinen kieli ei kuitenkaan automaattisesti ilmaise aina maailmassa vallitsevia totuuksia. Esimerkiksi huudahdukset, kysymykset, pyynnöt ja kiitokset saattavat olla jotain muuta kuin tietoa (Hallamaa et al, s. 107).

Kun sanotaan X:n tietävän että Y, niin Y edustaa mielivaltaista väitelausetta, mikä kuvaa tiedon kohteena olevan tosiseikan. X:n ilmaisemaa väitettä nimitetään propositioksi, joka on väitelauseen tiedollinen sisältö tässä tapauksessa (Räikkä 1991, s. 11). Lauseet ja muut kielelliset merkit kuten sanat voidaan yksilöidä kahdella eri tavalla. Niillä voidaan tarkoittaa joko lauseen tai sanan esiintymää eli instanssia tai tyyppiä (Räikkä 1991, s. 12). Instanssi tarkoittaa lauseen tai sanan käyttöä tietyssä tilanteessa ja tyyppi samaan sana- tai lausetyyppiin kuuluvia lauseen tai sanan esiintymiä. Propositiot ja asiantilat vastaavat yleensä lauseiden esiintymiä pikemminkin kuin lausetyyppejä. Lauseella pitää olla myös totuusarvo. Mikäli lause vastaa maailmassa vallitsevaa asiantilaa on se tosi ja jos ei, niin lause on epätosi. *Tautologiset* eli loogisesti todet lauseet sallivat kaikki asiantilat. Esimerkiksi lause "kirjoitus naurattaa tai ei naurata" on tautologinen. Mikäli lause kieltää kaikki asiantilat on se ristiriita, eli *loogisesti epätosi*. Esimerkiksi "kirjoittaminen naurattaa ja ei naurata" on ristiriita. Muut väitelauseet ovat Niiniluodon (Niiniluoto 1996, s. 23) mukaan faktuaalisia eli tosiasiaväittämiä. Faktuaaliset väittämät voivat koskea (Popper 1992):

1) Aineellista maailmaa, esimerkiksi lause "maa kiertää aurinkoa".
2) Mentaalista maailmaa, esimerkiksi lause "veteen pistetty keppi näyttää taipuneelta".
3) Mielen tuotteiden maailmaa, esimerkiksi lauseet "Sibelius sävelsi 7 sinfoniaa" ja "89 on

alkuluku”.

Niiniluoto (1996, s. 23-24), luokittelee indikatiivilauseiden ilmaiseman tiedon seuraavanlaisiin tyyppeihin, joista tärkeimmät ovat hänen mielestä singulaarinen ja yleinen tieto.

1) *Singulaarinen tieto* kuvaa yksittäisiä asioita, tosiseikkoja ja tapahtumia kuten historiaa koskeva tietomme tai ihan arkielämän havaintotietoa omasta ympäristöstämme. Se voi olla joko kuvailevaa tai tulkitsevaa tietoa. Esimerkiksi ensimmäinen voisi olla ”millainen x on? ” ja jälkimmäinen ”mitä x tarkoittaa”. Humanististen tieteiden tarjoamaa tulkitsevaa tietoa nimitetään myös ymmärrykseksi.

2) *Yleinen tieto* käsittää luonnontieteiden ja systemaattisten yhteiskuntatieteiden lait ja teoriat. Esimerkiksi lause ”kalat elävät vedessä”.

3) *Tilastollinen tieto* kuvaa jonkin populaation luonnetta. Tiedon käsite voidaan yleistää sellaisiin lauseisiin jotka eivät ole indikatiivisessa muodossa, mutta joilla silti on totuusarvo. Esimerkiksi, ”60% Suomen väestöstä on naisia”.

4) *Modaalinen tieto* kuvaa maailmassa vallitsevia mahdollisuuksia ja välttämättömyyksiä. Esimerkiksi ”kasvihuoneilmiö saattaa aiheuttaa maailmanlopun”.

5) *Konditionaalinen tieto* ilmaistaan tosiasioiden vastaisella, eli ”kontrafaktuaalisilla” ehtolauseilla. Esimerkiksi ”jos Suomessa olisi despootti, ei Suomi olisi enää demokratia”.

6) *Selitykset* kertovat miksi tai mistä syystä tietyt asiantilat vallitsevat ja tapahtumat sattuvat. Esimerkiksi muotoa ” A, koska B” olevat selitykset voivat kohdistua sekä luontoa että ihmistä koskeviin tosiasioihin.

7) *Välineellinen tieto*, kuten soveltavien tieteiden tarjoama taitotieto kertoo mitä pitää tehdä ”jos A haluaa X:n”. Väite on tosi, mikäli A:n tekeminen on X:n saavuttamiseen välttämätön tai riittävä ehto.

8) *Arvioiva tieto*, kertoo X:n (asiantila, esine tai teko) olevan hyvä tai arvokas suhteessa johonkin arvojärjestelmään S. Kyseiset kriteerit voivat olla esimerkiksi taloudellisia tai terveydellisiä. Niiniluodon (1996, s. 24) esimerkissä taidearvostelussa, joka perustuu eksplisiittisesti muotoiltuun arviointijärjestelmään, voi olla kyse arvioivasta tiedosta.

Ovatko kaikki kahdeksan tiedon lajia kuitenkaan tietoa sen klassisen määritelmän mukaan? Singulaariseen ja yleiseen tietoon on helppo liittää klassisen tiedon määritelmä. Tilastollista tietoa on kuitenkin vaikea pitää tietona, koska tilastot saattavat valehdella. Esimerkiksi tilastolliseen tietoon liitetään virhemarginaali. Tällä perusteella mikä tahansa saadaan näyttämään tiedolta, kun

virhemarginaalia on tarpeeksi. Modaalista- ja konditionaalista tietoa on myös vaikea pitää klassisen määritelmän mukaisena tietona. Voimme toki ajatella esimerkiksi, että ”tuli sammui, koska happi loppui”. Humen giljotiinin mukaisesti emme kuitenkaan voi havaita syy-seuraussuhdetta. Näemme siis vain, että tuli sammui ja happi loppui muttemme sitä että tulen sammuminen olisi hapen loppumisen syynä. Mahdollisuuksia ja selityksiä ei voi pitää klassisena tietona, koska mikään ei takaa niiden teoriassa kuvaavan tilanteen toimivan täysin käytännössä. Tämä muistuttaa tieteessä olevia teorioita ja niihin liittyviä käytännön ongelmia. Välineellinen tieto eli soveltavien tieteiden tarjoama taitotieto kuuluu enemmän taitotietoon, kuin propositionaaliseen tietoon. Arvioiva tieto kertoo X:n olevan hyvä suhteessa johonkin arvojärjestelmään. Niiniluodon esimerkki arvioivasta tiedosta oli taidearvostelu. Eikö taidearvostelu kuitenkin perustu henkilön omiin mielipiteisiin? Tieto ei voi olla pelkästään kulttuuriin sidonnainen, vaan asiantiloja kuvaavat väitteet ovat globaaleja.

Hume (1739, kirja 3 luku 1 kappale 1) esittää ehdottomien arvoväittämien ja normien olevan tavallisen tiedon piirin ulkopuolella, koska niillä ei ole totuusarvoa. Esimerkiksi raamatun esittämät kymmenen käskyä kuuluvat edellä mainittuihin. *Giljotiinissaan* Hume erotteli arvot tosiasiaväittämistä; eli siitä miten asia ovat ei voi loogisesti johtaa sitä miten asioiden pitäisi olla (Hume 1739, kirja 3 luku 1 kappale 1). Tällaiset arvoväittämät eivät ole kognitiivisesti eli tiedollisesti mielekkäitä koska kyseisillä lauseilla ei ole totuusarvoa eli ne eivät väitä tästä maailmasta mitään. Näiden lauseiden merkitys korostuu emotionaalisella tasolla eli ne herättävät tunteita ja ilmaisevat niitä.

Gettierin vastaesimerkit perinteisen tiedon määritelmään

Edmund Gettier osoitti vuonna 1963 artikkelillaan ”Is justified true belief knowledge” ettei tosi ja oikeutettu uskomus ole aina tietoa. Hallamaa et al. (2002, s. 109) esittävät Gettierin esiin nostaman ongelman seuraavalla esimerkillä. Henkilö uskoo linja-auton numero 67, johon hän nousee Helsingin rautatientorilta, kulkevan Pakilantien kautta. Perusteluinaan henkilöllä on Helsingin kaupungin liikennelaitoksen kartta, aikataulukirja ja reittikartta. Arkisissa tilanteissa perustelut olisivat riittäviä. Mikäli linja-auto kulkee Pakilantien kautta on uskomus tosi, ja voidaan sanoa, että henkilö tiesi bussin kulkevan Pakilantietä. Oletetaan kuitenkin, että linja-autojen reitit ovat juuri muuttuneet siten ettei bussi numero 67 enää kuljekaan entistä reittiä eikä esimerkkihenkilömme tiedä tätä. Oletetaan ettei hänellä ole myöskään hyviä perusteita uskoa reitin muutokseen, esimerkiksi ettei muutoksesta ole tiedotettu riittävästi. Seuraavaksi oletetaan, että linja-auton kuljettaja ajaa tottumuksesta Pakilantien kautta, eikä noudatakaan uutta reittiä. Henkilön uskomus, että linja-auto ajaa Pakilantietä, on tosi ja hyvin perusteltu, mutta sen pitäminen tietona

olisi kuitenkin outoa. Esimerkissä totuus on satunnaista, sillä se on seurausta vain kuljettajan vahingosta, vaikka matkustaja luuli sen perustuvan linja-auton normaaliin reittiin. Perustelu uskomuksen totuudelle on väärä, vaikka henkilö onkin oikeutettu pitämään sitä hyvänä perusteluna.

Hallamaa et al.:n ylläolevan esimerkin mukaan tieto ei ole sama asia kuin hyvin perusteltu tosi uskomus. Miksi sitten yhä puhutaan perinteisestä tiedon määritelmästä, vaikka Gettier on osoittanut sen puutteet? Siksi, ettei Gettier ole osoittanut muuta kuin etteivät ehdot ole riittävät. Äskeinen esimerkki ei siis viittaa siihen etteivätkö ehdot pitäisi paikkansa, vaan että ne eivät ole yhdessäkään riittäviä tiedon määrittelemiseksi. Siksi filosofiassa on keskusteltu neljännestä ehdosta tiedon määrittelemiseksi, mutta ehdosta ei ole vielä päästy sopuun (Steup 2006). Koska tiedolta voidaan edellä mainitun perusteelta vaatia yhä että se täyttäisi kolme alkuperäistä ehtoa, pysytään myös tutkielmassa perinteisessä tiedon määritelmässä ja sen perusteella analysoidaan mikä on tietoa ja mikä ei.

Metatieto

Metatieto on tietoa jostain toisesta tiedosta, toisin sanoen kuvailevaa tietoa jostain tietovarannosta tai sisältöyksiköstä (Salminen 2005). Esimerkiksi CD-levyjen tiedot kuten esittäjä, säveltäjä ja julkaisuvuosi ovat metatietoa. Vastaavasti tekstidokumenteissa olevat tiedot kuten tekijä, julkaisija, versionumero ja julkaisupäivämäärä ovat metatietoa. Metatietojen kehittämisellä pyritään tehostamaan jonkin tietovarannon käyttöä (Duval et al. 2002). Metatiedoilla esimerkiksi voidaan helpottaa tiedostojärjestelmien välisiä tiedonsiirtoja sekä eri paikoissa olevien sisältöjen yhdistämistä. Parantamalla metatietoa voidaan kehittää tietojen arkistointia, versiohallintaa, prosessien toimintaa ja asiakäsittelyä. Metatietoa voi olla tallennettu myös tiedosta jota ei enää ole. Internetin kasvun myötä myös metatiedon merkitys on kasvanut, sillä laadukkaat metatiedot auttavat hakukoneita löytämään entistä laadukkaampaa ja monipuolisempaa informaatiota. Entistä laadukkaammalla informaatiolla voisi tarkoittaa informaatiota, mikä sulkee toista informaatiota enemmän maailmassa vallitsevia asiantiloja tai sisältää vähemmän kohinaa.

Tiedostojärjestelmissä olevat tiedostonimet ovat tyypillistä metatietoa (Duval et al. 2002). Tekstitiedoston tiedostonimi voi esimerkiksi kertoa mitä aihetta tiedosto käsittelee, esimerkiksi ”metatieto.txt” voisi käsitellä nimensä mukaisesti metatietoa. Vastaavasti tiedostonimeen voidaan kirjoittaa tekijä sekä päivämääriä, ”2007- metatieto-Hauninen Jesse.doc” kertoo muokkausvuoden ja tekijän. Edellisessä esimerkissä käytettiin muotoa ”vuosi, asiayhteys ja kirjoittaja” tyyppistä metatietoa. Ohjelmatiedostoissa oleva metatieto voisi kertoa, että ohjelma käynnistyy kyseisellä tiedostolla, kuten voisi kuvitella tiedoston ”launch.exe”:n tekevän Windows ja DOS-ympäristössä. Tiedostojen lopussa olevat päätteet ovat nekin tyypillistä metatietoa. Ne kertovat minkä tyyppinen

tiedosto on kyseessä, esimerkiksi onko kyseessä tekstidokumentti vai suoritettava ohjelma sekä millä ohjelmalla tiedostoa tulisi käsitellä. Esimerkiksi doc päätteinen tiedosto on tarkoitettu avattavaksi Microsoft Word-ohjelmalla. Perinteisten tiedostojärjestelmien metatietojärjestelmä perustuu hakemistorakenteeseen, jossa samanlaiset tiedostot ovat samoissa hakemistoissa, sekä kuvaaviin tiedostonimiin.

Metatieto on tallennettu useimmiten joko sisälle tiedostoon tai erillisesti keskitettynä tietokantaan (Duval et al 2002). Esimerkiksi musiikki- ja kuvatiedostoissa on usein tallennettuna itse tiedostoon metatietoa joka voi liittyä esimerkiksi tekijään, kohteeseen, tiedostokokoon tai albumiin. Tosin vastaavat tiedostot saattavat joskus tallentaa metatietonsa erillisiin mediakirjastoihin tai albumitietokantoihin. Mikäli metatiedot ovat sisällytetty itse tiedostoon siirtyvät ne tiedoston mukana, esimerkiksi musiikkikappaleen tiedot siirtyvät kyseisen kappaleen mukana kannettavaan soittimeen. Mikäli metatietoa ei ole sisällytetty tiedostoon, ei se useinkaan siirry tiedostoa siirrettäessä. Monien metatietoa sisälleen tallentavien ohjelmien ongelmana ovat erilaiset epäyhteensopivuudet muiden ohjelmien kanssa sekä rajalliset tallennusmahdollisuudet.

Salminen (2005) jakaa metatiedon seuraavasti:

1) Semanttinen metatieto (Sisällön merkitystä kuvaava tieto kuten asiasanat, aihe ja tiivistelmä).
2) Rakennemetatieto (Sisältöyksikön fyysistä tai loogista rakennetta tai sisällön kieltä kuvaavaa tietoa).
3) Kontekstuaalinen metatieto (Kuvaa sisältöyksiköiden jossain tietyssä tilanteessa. Esimerkiksi sisältöyksikön luomisaika, tuottaja, käyttäjä ja suhteet muihin sisältöyksiköihin).

Metatietoa tuotetaan kahdella eri tavalla, joko automaattisesti tai manuaalisesti (Duval et al 2002). Automaattinen metatieto tuotetaan joko dokumenttien ominaisuustiedoista tai rakenteisista dokumenteista erottelemalla. Manuaalisesti metatietoa tuotetaan kirjoittamalla dokumentteihin erilaisia kuvaustietoja. Perinteinen metatiedon kehittämismenetelmä on olemassa olevien sisältöjen analysointi. Nykyään analysointi tapahtuu pääsääntöisesti automaattisesti erilaisten tietokoneohjelmien avulla, mutta ihmisten tekemää luokittelua ja kuvailua ei tulisi kuitenkaan väheksyä.

Metatietojen käytössä esiintyy muun muassa seuraavia ongelmia (Duval et al 2002):

1) Luonnollisen kielen runsaus ja monimutkaisuus

2) Koneellisen tulkinnan vaikeudet
3) Ongelmat sanastojen käytössä ja kehittämisessä
4) Ohjelmat tallentavat metatiedot sellaisessa muodossa ettei niitä voi avata kuin kyseisellä ohjelmalla.

Kohta 4) on tyypillinen tietokoneenkäyttäjän perusongelma. Ohjelmat tallentavat metatietoa tiedostoihin, mutta vain itseään varten jolloin toiset ohjelmat eivät kykene avaamaan kyseistä tiedostoa. Esimerkiksi Word-asiakirjat sisältävät niin paljon metatietoa (kuten tekijä, muokkaaja, kirjasimet, rakenne ja muotoilu) ettei niiden avaaminen kaikilla tekstinkäsittelyohjelmilla onnistu (Salminen 2005). Ongelma on pyritty ratkaisemaan kehittämällä openDocument-tiedostomuoto. Metatietojen yhteensopivuutta on myös pyritty parantamaan valmiiksi luoduilla sanastoilla. Pääsääntöisesti näillä parannetaan järjestelmien välistä yhteensopivuutta eli puhutaan niin sanotusta tiedostojärjestelmien välisestä semanttisesta yhteensopivuudesta. Valmiita sanastoja ovat esimerkiksi Dublin Core, IPTC, RDF ja OWL (Salminen 2005).

Onko metatieto tietoa sen varsinaisessa merkityksessä eli tosi, perusteltu ja uskomus? Kuvitellaan, että henkilö on avaamassa lue_minut.txt nimistä tiedostoa. Henkilö on avannut aikaisemmin kyseisen tiedoston joten hänellä on hyvä syy olettaa sen sisältävän ohjelman kannalta tärkeää informaatiota. Henkilö on myös tietojenkäsittelytieteen opiskelija, joten hänellä on perusteet pitää txt päätteisiä tiedostoja tekstitiedostoina. Tämä ei kuitenkaan takaa, että lue_minut.txt pitäisi uudella avaushetkellä sisällään tekstiä. Onhan se voinut vaikka joutua viruksen saastuttamaksi, eikä näin ollen sisällä enää yhtään mitään. Toisaalta virustutka on voinut korjata viruksen vahingot käyttäjän tietämättä, eikä käyttäjä huomaa mitään eroa. Metatieto ei aina olekaan tietoa, vaikka tiedon vaatimukset täyttyisivät. Täten metatietoon tulee suhtautua yhtä epäilevästi kuin tietoonkin.

3.1.3 Informaatio

Tässä kappaleessa käsitellään informaatio-termiä. esitellään hieman kyseisen termin historiaa, käydään läpi informaatioteoriaa sekä informaation lajeja. Lisäksi tutustutaan erilaisiin informaation kantajiin ja pohditaan mitä informaatio todella on.

Informaatio-termin historia

Latinan kielessä 'forma'-sanalla tarkoitetaan muotoa ja 'informare':lla formuloimista, muodostamista ja muotoilua. Tästä juontaa myös ranskan kielen vanha verbi enformer sekä keskiajan englannin information, jotka suomeksi tarkoittavat sitä mikä on muodostettu, kerrottu tai ilmaistu (Niiniluoto 1996, s. 2). Niiniluoto tototeaa Augustinuksen käyttäneen keskiajalla informatio-sanaa opettamisen

synonyymina. Ciseron sanotaan käyttäneen termiä ilmaisemaan miellettä, sanan merkitystä ihmisen mielessä, vuosisadan ajan ennen Jeesuksen syntymää. Keskiajan skolastikot käyttivät informaatio-sanaa tiedotusprosessin tuloksena.

Tietosanakirjojen mukaan suomen kielessä informaatio-sanalla tarkoitettiin alun perin yksityisopettajien antamaa opetusta, mutta nykyään sillä tarkoitetaan myös muiden informaattoreiden tarjoamaa ”tiedonantoa” kuten tavaratalojen neuvonta-informaatiota tai liikeyrityksien tiedottajan järjestämää informaatiotilaisuutta, infoa (Niiniluoto 1996, s. 3). Tietojenkäsittelytiedettä ajatellen määrittelen *informaation* dataksi, jolle on annettu merkitys. Tämä tulkinta voitaneen rinnastettaa skolastikkojen tulkintaan.

Informaation olemus

Mitä informaatio loppujen lopuksi on? Ensimmäisenä mieleen tulee, että informaatio olisi ainetta tai energiaa. Kyberneetikko Wienerin mukaan informaatio ei ole kumpaakaan, vaikka sitä voidaan materiaalisten prosessien avulla siirtää ja tallentaa (Niiniluoto 1996, s. 16). Mitä sitten informaatio voi olla, jollei se ole aineen tai energian muotoja? Yksi selitys voisi olla fysikaalinen suure, joka liittyy materiaalisten järjestelmien makro- tai mikrotilojen muutokseen. Niiniluoto (1996, s. 16) kritisoi ettei tällainen ei-kielellinen käsite kerro mitä merkkien, merkkijonojen tai datan kantama syntaktinen, semanttinen ja pragmaattinen informaatio on. Niiniluoto (1996, s. 17) ei hyväksy myöskään näkemystä, että informaatio edellyttäisi aina kieltä puhuvien ja ymmärtävien, tajunnalla varustettujen olentojen olemassaoloa, sillä hänen mielestään se on seurausta vain hyvin suppearajoitteisesta informaation määritelmästä. Mikäli informaatiota olisi vain ihmisten päässä, niin atk-alalla työskentelevien puheet informaation ”siirrosta” viestintävälineiden avulla tai informaation ”käsittely” ja ”tallentaminen” tietokoneissa eivät olisi enää mielekkäitä. Tässä sotketaan keskenään datan ja informaation käsite.

Informaation olemassaolon sijoittaminen vain yhteen kategoriaan ei ollut ongelmatonta. Olemassaoloa esiintyy usealla eri tasolla, joten informaation sijoittaminen eri todellisuuden tasoihin ei ole mahdoton ajatus. Esimerkiksi Popper erottelee kolme todellisuuden piiriä (Popper 1992, s. 7-11) :

1) Maailma 1 johon sisältyvät ajassa ja avaruudessa esiintyvät fysikaaliset objektit, tapahtumat ja prosessit, aine ja energia, epäorgaaninen ja orgaaninen luonto. Esimerkiksi planeetat, ihmisen ruumis, ja vesi.

2) Maailma 2 johon sisältyvät yksilöllisen tajunnan tilat, mentaaliset tapahtumat ja psyyke. Esimerkiksi elämykset ja ajatukset.

3) Maailma 3, johon sisältyvät ihmisen sosiaalisen toiminnan kautta syntyneet kulttuuriesineet, artefaktit ja abstraktiot, kulttuuri ja yhteiskunta. Esimerkiksi tieteelliset teoriat, luvut ja käsitteet.

Materialismin mukaan todellisuus on aineellinen. Materialismista on eri käsityksiä, radikaalit väittävät vain maailman 1 olevan olemassa, emergenttimaterialistit hyväksyvät myös maailman 2 ja 3. Yhteistä eri materialismin näkemyksien kannattajalla on, että psyykkiset ilmiöt eivät voi esiintyä ilman materiaalisia kannattajia (Hallamaa et al 2002, s. 21-22). Maailman 2 ajatuksien olemassaolo on riippuvainen ihmisten aivojen toiminnasta. Maailman 3:n oliot säilyvät vain niin kauan kuin ne ovat tallennettuja maailmaan 1 ja 2.

Idealistien mukaan todellisuus on henkistä eli se muodostuu yksinomaan mentaalisista tiloista tai tapahtumista. Näin ollen vain maailma 2 on olemassa. ”Todellisuus on viime kädessä henkinen.” (Saarinen 1996, s. 260-261). Objektiivisten idealistien mukaan todellisuuden perusaine on jonkinlainen korkeampi prinsiippi, esimerkiksi Jumala.

Dualistit erottavat todellisuudesta aineen ja hengen. Heidän mukaan molemmat ovat itsenäisiä, toisiinsa palautumattomia perusosia, jotka kehittyvät joko samansuuntaisesti toisistaan riippumatta (parallelismi) tai ovat keskenään jatkuvassa kausaalisessa vuorovaikutuksessa (interaktionismi) (Kotkavirta 1999, s.197-198).

Käyttämällä hyväksi edellä tehtyä maailmojen erottelua voidaan määritellä missä todellisuuden piirissä eri informaation lajit sijaitsevat. Gregorin (2006) mukaan ensimmäiseen maailmaan kuuluu kaikki fysikaalinen joten hänen mielestään on loogista sijoittaa fysikaalinen informaatio (negeopia, materian järjestys) maailmaan yksi. Samoin kielelliset merkit, merkkijonot tai datan esiintymät fysikaalisina objekteina kuuluvat ensimmäisen maailmaan. Esimerkiksi kirjainmerkit ovat maailman yksi olioita. Näiden lukijassa herättämät ajatukset kuuluvat maailmaan kaksi. Niiniluoto (1996, s. 18) toteaakin pragmaattisen informaation subjektiivisessa mielessä kuuluvan maailmaan kaksi. Semanttisen informaation voisi sijoittaa kolmanteen maailmaan. Semanttinen informaatiohan edellyttää symbolifunktiota, ihmisen luomaa merkityksellistä kieltä, mutta muutoin se on objektiivista, yksilöllisestä ihmismielestä riippumatonta (Gregor 2006). Samaan kolmosmaailmaan voidaan sijoittaa pragmaattinen informaatio, joka perustuu jonkun kulttuurin piirissä vallitseviin intersubjektiivisiin merkityksiin ja merkittävyyden arvioihin.

Klausin ja Buhrin filosofisessa sanakirjassa todetaan ettei informaatio ole ainetta, eikä aineellisten kappaleiden ominaisuus, vaan aineellisten kappaleiden ominaisuuksien ominaisuus (Niiniluoto 1996, s. 18-19). Fysikaaliset objektit ovat merkkejä, fysikaalisten objektien

ominaisuudet ovat merkkityyppejä ja fysikaalisten objektien ominaisuuksien ominaisuudet informaatiota. Niiniluoto (1996, s. 19) toteaa ettei tämä puhetapa ole aina kovinkaan luontevaa kaikissa suhteissa.

Bakan lähestyy informaation käsitettä dualismin ja idealismin pohjalta (Niiniluoto 1996, s. 19). Hän väittää informaation olevan oma erillinen realiteetti, itsenäinen substanssi, jota voidaan varastoida vapaana ihmismieleen ja sidottuun materiaan. (Niiniluoto 1996, s. 19). Bakanin informaatio muistuttaa Popperin kolmatta maailmaa, koska se vaikuttaa todellisuuteen, mutta se ei ole kuitenkaan ihmisen tekemää, vaan on itsenäisesti olemassa oleva kuten esimerkiksi Platonin ideamaailma. Luonnonlait ovat Bakanin (Niiniluoto 1996, s. 19) mukaan sitä, että objektiivinen informaatio informoi fysikaalista liikettä. Bohm (Niiniluoto 1996, s. 19) esittää, että ”maailma muodostaa ykseyden, yhtenäisen energiavirran, kokonaisuuden, jonka aineellisen ja mentaalisen puolen yhteisen perustan muodostaa aktiivinen, organisoiva ja liikkuva informaatio, merkitys (meaning) tai piilevä järjestys (implicate order)”.

Edellä esitetyn perusteella voidaan todeta ettei pelkästään radikaali materialismi pysty selittämään informaation olemassaoloa. Niiniluoto (1996, s. 19) toteaa emergentin materialismin antavan tyydyttävän perustan informaation eri lajien ontologisen aseman selvittämiselle.

Informaatioteoria

Informaation käsite on 1900 luvulla tieteissä yhtä keskeinen kuin energian ja evoluution käsite 1800- luvulla (Niiniluoto 1996, s. 4). Niiniluoto (1996, s. 4) esittää lähtökohdaksi aiheen tutkimiselle yrityksen mitata välitetyn informaation määrää. Ensimmäiset ehdotukset tästä tehtiin 1920-luvulla ja varsinainen läpimurto 1940-luvun lopulla. Bell System Technical Journal:in julkaisemissa H. Nyquistin artikkelissa (1924) ” Certain Factors Affecting Telegraph Speed” ja R. V. Hartleyn artikkelissa ”Transmission of Information” (1928), tutkittiin tiedonsiirron tehokkuutteen ja nopeuteen liittyviä kysymyksiä. Esimerkkinä Niiniluodolla (1996, s. 4) on, että jos lennättimen kautta lähetetään viesti joka on poimittu n:n mahdollisen vaihtoehdon joukosta, niin Hartleyn mukaan viestiin liittyvän informaation määrä on n:n kymmenkantainen logaritmi. Claude Shannon loi tästä artikkelissaan ” The Mathematical Theory of Communication” vuonna 1948 varsinaisen systemaattisen teorian toisin sanoen tilastollisen kommunikaatioteorian (Shannon & Weaver 1963, s. 1). Shannon tutki esimerkiksi erilaisia tiedonvälityskanavia ja niiden kykyä välittää informaatiota sekä tiedonvälityksen luotettavuuden ja tehokkuuden riippuvuutta käytetystä koodaustekniikasta. Shannonin perustulokset todistettiin täsmällisesti 1950- luvulla jolloin syntyi matemaattinen *informaatioteoria* (Shannon & Weaver 1963, s. VIII-IX).

Jääkaapin termostaatit, automaattiset ilmapuolustusjärjestelmät tai ihmisaivot ovat kommunikaatiosysteemejä, jotka ovat erikoistapauksia järjestelmästä joka käsittelee, vastaanottaa ja välittää informaatiota. Näihin systeemeihin kuuluvat itsesäätelyyn kykenevät automaatit ovat kiinnostavimpia kommunikaatiosysteemejä, sillä niiden toimintaa ohjaa takaisinkytkennän (feedback) avulla jokin sisäänrakennettu tavoite (Wiener 1961, s. 6-7). Esimerkiksi jääkaappi pyrkii säilyttämään lämpötilan tietyn asteisena. Norbert Wiener esitti teoksessaan ”Cybernetics, or the Control and Communication in the Animal and the Machine” (1948) ajatuksen uudesta yleistieteestä, jossa tutkitaan koneissa ja elävissä olennoissa esiintyvien säätö- ja ohjausjärjestelmien yhteisiä piirteitä. Tämän uuden tutkimusalueensa hän nimesi *kybernetiikaksi* (Wiener 1961, s. VII, 1-4).

Kybernetiikan keskeisen osan muodostaa informaatioteoria. Teemmehän eron ”kybergeneettisten” systeemien esimerkiksi tietokoneiden ja ”energeettisten” systeemien kuten höyrykoneiden välillä juuri viestien välittämisen perusteella (Wiener 1961, s. 6-13). Insinööritieteisiin kuuluvat systeemi- ja säätötekniikka tutkivat kyberneettisten systeemien teknistä toteuttamista. Näiden systeemien abstraktia matemaattista tutkimusta kutsutaan automaattien teoriaksi, mikä myös liittyy läheisesti tietojenkäsittelytieteeseen. Itsesäätelyjärjestelmät voivat kyetä älykkäältä näyttävään toimintaan, esimerkiksi tietokonepeleissä tietokone osaa reagoida pelaajan tekemisiin ja robottien toiminta saattaa välillä olla hyvinkin inhimillistä. Tällöin puhutaan keinoälyn (artificial intelligence) tutkimuksesta (Niiniluoto 1996, s. 5).

Informaation lajit

Fysikaalinen informaatio on ei-kielellistä (Eckert 2006, s. 102-105). Lähtökohtana *fysikaalisen informaation* käsitteelle on ollut Rudolf Clausiuksen vuonna 1865 käyttöön ottama termodynamiikan suure entropia. Fysikaalinen informaatio ilmaisee lähtökohtaisesti aineellisten systeemien järjestäytyneisyyttä, organisaatiotasoa tai monimutkaisuutta (Mäkipää & Ruohonen 2004, s. 2). Kuvauksen kohteena voi olla joko elollinen tai eloton luonto. Esimerkiksi elottomassa luonnossa oleva tähtisumu ja elollisessa erilaiset solut ja organismit. Fysikaalinen informaatio kelpaa siihen, kun arvioidaan ihmisten suunnittelemia ja luomia artefakteja, esimerkiksi erilaisten koneiden kompleksisuutta.

Kielellinen informaatio voidaan jakaa kolmeen eri osaan: syntaktiseen, semanttiseen ja pragmaattiseen. Alla olevissa kappaleissa esitellään pääpiirteet kyseisistä informaatiolajeista.

Carnap ja Bar-Hillel olivat käsitteen ”semanttinen informaatio” ensimmäisiä kehittäjiä. *Semanttisen informaation* mukaan lause on sitä informatiivisempi, mitä enemmän kielessä

erotettavia asiantiloja saadaan suljetuksi pois (Niiniluoto 1996, s. 14). Mikäli lause ei sulje mitään pois, on sen informaatioarvo nolla. Kuitenkin jos jotain asiatiloja saadaan suljettua pois, saadaan jokin maailmassa vallitseva asiantila myös lausuttua. Toisin sanoen, asiantilojen mahdollisuuksien avaruus kapenee ja epävarmuutemme vähenee. Semanttinen informaatio lähenee näin ollen tiedon käsitettä. Heinonen (2007, s. 5) erottelee nämä kaksi termiä totuusarvon mukaan. Tieto ei voi koskaan olla epätotta, mutta semanttinen informaatio ei ota kantaa totuuteen, vaan väitelauseen informaatiosisältö on riippumaton totuudesta. Näin ollen hyvinkin informatiivinen lause voi olla epätotta. Esimerkiksi ristiriitaisella eli loogisesti epätodella väitteellä on maksimaalinen informaatiosisältö. Vastaavasti tautologialla eli loogisesti todella väitelauseella ei ole informatiivista sisältöä.

Shannon ja Weaver (1963, s. 8-12) tekevät eron kanavassa välitetyn merkeistä muodostetun viestin (message) ja kanavan välittämän informaation välillä. *Syntaktinen informaatio* on jälkimmäistä. Sen ollessa kyseessä ei ole väliä mitä siirtokanavassa kulkee, kunhan vain informaatio välittyy nopeasti. Näin ollen täysin järjetönkin viesti voi olla informaatiota. Pääpaino onkin käytettyjen merkkien ja siirtokanavan suhteessa, eli miten viestit tulisi lähettää jotta lähetys olisi mahdollisimman luotettavaa ja tehokasta. Pyritään siis ratkaisemaan miten pisteestä X lähetetty viesti saataisiin mahdollisimman tarkasti toistettua pisteessä Y. Itselleni tulee syntaktisesta viestistä mieleen lähinnä data, joka on määrittelemätöntä ”informaatiota”.

Carnap ja Bar-Hillel toteavat heidän teoriansa vain koskevan ”lauseen kantamaa informaatiota, sekä itsessään että suhteessa toiseen lauseeseen tai lausejoukkoon”, ei ”informaatiota, jonka lähettäjä aikoi välittää lähettäessään tietyn viestin”, eikä myöskään "informaatiota, jonka vastaanottaja sai tästä viestistä” (Niiniluoto 1996, s. 14). Vastaavasti *pragmaattisessa informaatioteoriassa* painotetaan viestin ja vastaanottajan välistä suhdetta. Viestin voidaan sanoa sisältävän pragmaattista informaatiota sillon, kun vähennetään vastaanottajan epätietoisuutta viestin kuvaaman kohteen osalta (Niiniluoto 1996, s. 15). Ero semanttiseen informaatioon ei näin ollen aina ole selvä. Hyvin yleisessä mielessä pragmaattisella informaatiolla voidaan tarkoittaa henkilö- ja kulttuurisidonnaista merkityksellisyyttä tai merkittävyyttä. Esimerkki pragmaattisesta informaatiosta on keskisormen näyttäminen. Voidaan myös puhua yksilöllisyyden merkityksellisyydestä, jonka mukaan informaatio on datan tuottama mielle tai merkitys vastaajalle (Niiniluoto 1996. s. 15-16). Amerikkalaisen pragmatismin vaikutus näkyy määritelmissä, joissa viesti liitetään hyödyllisyyteen. Informaatioarvo (value of information) on suoraan verrannollinen informaation säilymisaikaan systeemissä ja kääntäen verrannollinen jo käytössä olevan yhtäpitävän informaation määrään (Niiniluoto 1996, s. 16).

Informaation kantajat

Kielellisen informaatiokäsitteen perusolettamuksen mukaan informaatiolla pitää olla jonkinlainen kantaja, joka välittää tai tallentaa viestejä tietyissä olosuhteissa (Niiniluoto 1996, s. 8). *Kantajat* voivat olla esimerkiksi aineellisia olioita, tapahtumia tai prosesseja. Laajassa mielessä informaation kantajia voidaan kutsua merkeiksi (sign), ja merkkijärjestelmiä kieliksi. Meille tunnetuimpia merkkijärjestelmiä ovat ihmisten käyttämät kielet, niin puhutut kuin kirjoitetutkin. Puhutussa kielessä äänteistä muodostuu sanoja ja lauseita, jotka kantavat informaatiota. Kirjoitetussa kielessä taas kirjaimista muodostuu informaatiota kantavia sanoja ja lauseita. Viestejä on mahdollista kuitenkin välittää usealla muullakin tavalla kuten kuvilla, musiikilla tai tietokoneen eri kielillä kuten kone- ja ohjelmointikielet. Lisäksi eläimet välittävät viestejä omalla tavallaan, esimerkiksi soidinmenoissaan.

Peirce ehdotti nimeä semiotiikka merkkijärjestelmien yleiselle teorialle. Hänen mukaansa *merkki* esittää aina jotakin jossakin suhteessa jollekin (Houser et al 1998, s. 4-10). Merkin viittaamaa kohdetta kutsutaan referenssiksi ja tulkitsijan synnyttämiä ideoita ja mielteitä interpretanteiksi. Peirce jakaa merkit kolmeen eri luokkaan sen mukaan millainen peruste liittää merkin kohteeseensa (Houser et al 1998, s. 8-9):

1) *Ikoni* on jonkinlaisessa samankaltaisessa suhteessa kohteensa kanssa, esimerkiksi kuvat ja diagrammit.
2) *Indeksi* liittyy kohteeseensa syy- vaikutus- suhteen, kausaalisuuden, perusteella. Esimerkiksi väsymys on unenpuutteen merkki.
3) *Symbolin* viittaama kohde on kieliyhteisön keskenään sopima, esimerkiksi morsen koodi aakkosille.

Charles Morris jatkoi Peircen oppien pohjalta semiotiikkaa ja jakoi kielen tutkimisen syntaktiselle ja semanttiselle tasolle (Kasher 1998, s. 15-16) :

1) Syntaktisella tasolla tutkitaan kielen merkkien keskinäistä suhdetta. Esimerkiksi luonnollisessa kielessä on kielioppi ja logiikan formaalikielessä päättelysääntöjä syntaksia vastaamassa.
2) Semanttisessa tasossa tutkitaan syntaksin esittämien merkkien viittauksia kielen ulkopuolelle. Esimerkiksi kohteita joihin sanat ja nimet viittaavat. Lisäksi semantiikassa tutkitaan miksi kahdella eri ilmauksella voi olla sama merkitys, eli ne viittaavat samaan kohteeseen, sekä toisinpäin, miksi kahdella ilmauksella voi olla sama referenssi, mutta eri

merkitys.

Sanoista muodostetaan väitelauseita, jotka mahdollisesti kuvaavat maailmassa vallitsevia asiantiloja. Maailmaa koskevan väitelauseen totuus määräytyy sen perusteella vallitseeko väitelauseen ilmoittama asiantila maailmassa. Tästä johtuen totuus on myös kielen esittävään tehtävään liittyvä semanttinen käsite (Kasher 1998, s. 16-17).

Pragmatiikassa tutkitaan kieltä tasolla, jossa otetaan huomioon kielen käyttäjät ja kielen todellinen käyttö erilaisissa kommunikatiivisissa tehtävissä. Näin ollen pragmatiikan tutkimuspiirissä olevat huudahdukset, kysymykset ja kehotukset, eivät kuvaa maailmassa vallitsevia asiantiloja (Kasher 1998, s. 19-20). ”Merkitys pragmaattisessa mielessä on merkittävyyttä tai tärkeyttä jonkin ihmisen tai ihmisryhmän näkökulmasta.” (Niiniluoto 1996, s. 10).

Niiniluodon (1996, s. 10-11) oma esimerkki viestinnästä on haaksirikkoinen henkilö A, joka on pelastunut yksinäiselle saarelle. Hänellä on kolme eri vaihtoehtoa hälyttää apua:

1) Ampumalla punaisen hätäraketin.
2) Morsettamalla lennättimellä apua merkkisarjalla ...---... ...---...
3) Lähettämällä radiopuhelimella tiedonannon ”Haaksirikkoutunut saarelle. Tulkaa apuun. Herra A”.

Viestit 1-3 kuuluvat kaikki eri merkkijärjestelmiin, eli syntaktisella tasolla ne ovat eri lauseita. Semantiikan tasolla ei kuitenkaan eroja ole, koska kaikki sisältävät saman sanoman. Vastaanottajan kannalta pragmaattinen merkitys liittyy pelastustoimiin joihin hätäsanoma johtaa. Henkilölle A pragmaattisesti merkittävin on se viesti, jonka ansiosta hän pelastuu.

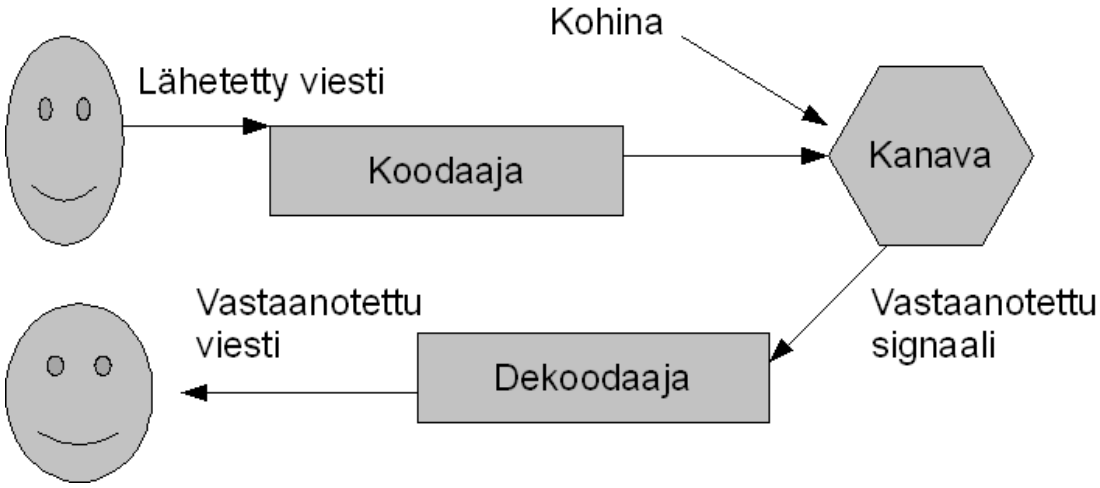
Seuraava esimerkki syntaktisesta ja semanttisesta informaatiosta on van der Lubbelta (1997, s. 1). Tarkastellaan seuraavia lauseita:

1) John tuotiin taksilla rautatieasemalle (John was brought to the railway station by taxi).
2) Taksi toi Johnin rautatieasemalle (The taxi brought John to the railway station).
3) Liikenneruuhkaa on moottoritiellä A3, Nürnbergin ja Münchenin välillä Saksassa.
4) Liikenneruuhkaa on moottoritiellä A3 Saksassa.

Lauseet yksi ja kaksi ovat syntaktisesti erilaisia. Kuitenkin pragmaattiselta ja semanttiselta kannalta lauseet ovat samoja. Lauseilla on sama merkitys ja molemmat ovat yhtä informatiivisia. Lauseet kolme ja neljä eroavat niin syntaktiselta, kuin semanttiselta kannalta. Lauseen kolme antama informaatio on tarkempaa kuin lauseen neljä informaatio. Pragmaattiselta kannalta lauseiden kolme

ja neljä merkitys on riippuvainen ympäristöstä. Saksalaisille tieto oman maan liikenneruuhkista on huomattavasti tärkeämpää kuin meille suomalaisille.

Alla olevassa kuvassa 5 on kuvattuna edellä esitetyn viestintäsysteemin rakenne Shannonin kommunikaatioteorian mukaisesti. Inhimillisissä viesteissä lähettäjänä on yleensä ihminen. Lähetetty ja vastaanotettu viesti kuuluvat tavallisesti luonnolliseen kieleen. Lähetetty viesti etenee koodattuna signaalina kanavassa (channel). Vastaanottajan on lopuksi purettava (decode) kanavasta saatu koodattu viesti sen tulkintaa varten. Kanavassa esiintyviä, merkkien oikeata tunnistamista vaikeuttavia, häiriöitä kutsutaan kohinaksi (Weaver 1949).



Kuva 5: Viestintäsysteemi (Niiniluoto 1996, s. 11).

Informaation paikka kommunikaatiojärjestelmässä on avoin. Mikäli informaatiolla tarkoitetaan syntaktisia olioita, kuten yllä olevassa kuvassa, vastaa informaatio merkkijonoja ja signaaleja joita syötetään kanavaan ja tulkitaan. Näitä merkkijonoja kutsutaan myös raakainformaatioksi ja dataksi (Niiniluoto 1996, s. 11). Informaatioteoriassa merkkijonoja pidetään informaation kantajina, ja ne kykenevät tallentamaan ja välittämään informaatiota (van der Lubbe 1997, s. 1-4). Kuvassa 5 lähettäjällä on informaatiota, mutta siitä syntyy dataa hänen lähettäessään viestinsä koodaajaan. Näin ollen kanavassa ei välitetä informaatiota, vaan dataa. Kun vastaanottaja on vastaanottanut puretun (decode) viestin ja tulkinnut sen, syntyy kanavassa välitetystä datasta informaatiota.

3.1.4 Datan, informaation ja tiedon suhde

Datalla itsellään ei ole merkitystä, vaan sille annetaan aina merkitys. Dataa ovat esimerkiksi kirjaimet tekstissä tai äänteet puheessa. Tietotekniikassa datalla on tarkoitettu erilaisia lukuja,

kirjaimia ja kuvia sellaisenaan ilman erityistä merkitystä. Kun datalle annetaan merkitys tulee siitä joko informaatiota tai tietoa. Esimerkiksi kun näyteikkunassa on tietokone, jonka hinta on 950 euroa ovat numerot ”9”, ”5” ja ”0” dataa. Kun annamme hintalapuissa oleville numeroille merkityksen, tulee siitä informaatiota, eli tulkitsemme hinnaksi 950 euroa. Toinen esimerkki on kirjasta jota olemme lukemassa. Siinä olevat kirjaimet ovat dataa, mutta kun lukiessamme annamme kirjaimille merkityksen, niistä muodostuu sanoja ja edelleen informaatiota.

Informaatio ja tieto muodostuvat datasta. Vaikka puhekielessä informaatio ja tieto samaistetaan hyvinkin usein, ovat nämä kaksi termiä merkityssisällöltään kaksi eri käsitettä. Tieto on aina informaatiota, mutta informaatio ei välttämättä ole aina tietoa. Mikäli propositionaalilauseen muodossa oleva informaatio vastaa todellisuutta eli on totta, on se silloin tietoa. Tietoa voisi tämän perusteella pitää informaation erikoistapauksena. Informaatiolta ei kuitenkaan aina vaadita, että sen olisi vastattava todellisuutta, vaan informaatio voi myös olla fiktiota. Esimerkiksi ”Suomessa käytetty raha on euro”, on sekä informaatiota ja tietoa, koska on olemassa maa nimeltä Suomi, jonka rahayksikkö on euro. Vastaavasti informaatio ”ihmissusi voidaan tappaa vain hopealuodeilla” ei ole tietoa, koska emme ole voineet osoittaa ihmissusien olevan ylipäätänsä olemassa, ja vaikka niitä olisikin, emme pysty osoittamaan etteikö niitä voisi tappaa jollain muullakin tavalla kuin hopealuodeilla. Kuitenkin lause ”ihmissusi voidaan tappaa vain hopealuodeilla” on informaatiota, vaikka se on ainakin tämänhetkisten tietojen mukaan fiktiota.

Voidaanko väite osoittaa todeksi jolloin se olisi tietoa? Tästä aiheesta filosofit kiistelevät vielä nykyäänkin. Mikä takaisi sen ettei tule uutta informaatiota joka korvaisi tai korjaisi aiemman informaatiomme. Mikäli on niin ettei voida osoittaa väitteen täyttävän tiedon määritelmiä, ei meillä tällöin ole tietoa, jolloin kaikki tähän asti pitämämme tieto on todellisuudessa informaatiota. Tiedon erottaminen datastakaan ei aina ole helppoa, esimerkiksi tutkiessamme Jumalaa. Tällöin päätelmät tapahtuvat mielessämme ja todistelu on pääsääntöisesti rationaalista. On kuitenkin epäselvää voimmeko sanoa mielikuviemme olevan dataa, varsinkin jos datalla tarkoitetaan jotain konkreettista.

3.2 Platonin luolavertaus

Tässä luvussa esitellään Platonin luolavertaus ja pohditaan onko sillä mitään annettavaa tietojenkäsittelytieteilijälle. Platonhan oli antiikin filosofeja, joka pohti muun muassa tiedon käsitettä päätyen klassiseen tiedon määritelmään. Filosofina Platon oli rationalisti ja dualisti. Hän uskoi tiedon alkuperäisen lähteen olevan järki. Todellisuuden Platon jakoi kahteen osaan: fyysisesti aistein havaittavaan maailmaamme, sekä järjellä saavutettavaan ideamaailmaan, jonka heijastus oma

maailmamme on. Tästä voisi ajatella, että tietojenkäsittelytieteilijän tekemät ohjelmat ovat heijastus ohjelmoijan omasta ohjelman ideasta. Ohjelmoija ajattelee mielessään miten ohjelma toimisi täydellisesti ja virheettömästi. Kirjoitettu ohjelma ei kuitenkaan koskaan toimi täydellisesti, vaan sisältää virheitä jolloin se olisi heijastus ohjelmoijan mielessä olevasta täydellisestä ohjelmasta.

3.2.1 Luolavertaus

Seuraavaksi esitellään Platonin luolavertaus Huardin (1996, s. 6-8) mukaan. *Platonin luolavertaus* on kuvitteellinen tilanne, jossa joukko ihmisiä on kahlehdittu syntymästään asti luolan perälle. Heidän päänsä on lisäksi kahlehdittu siten, että he näkevät vain luolan seinän. Kuvassa 6 on kuvattuna luolavertauksen asetelma vangeista luolassa ja siitä miten he näkevät vain luolan seinälle heijastuvat varjot todellisuuden sijasta. Nämä vangit eivät ole koskaan päässeet ulos luolasta ja ulkomaailma hahmottuu heille heidän takanaan palavan tulen luomien varjokuvien avulla. Samoin ainoa puhe mitä nämä vangit kuulevat, on se mitä luolan perukoille kantautuu. Samoja muotoja nähdessään vangit alkavat nimeämään niitä ja pikkuhiljaa varjokuvista ja luolan vääristämistä äänistä muodostuu vankien todellisuus. On muistettava ettei yksikään vanki ole käynyt luolan ulkopuolella, vaan kaikki ovat viettäneet koko elämänsä luolassa, joten heillä ei ole tietoa varjojen ja äänien todellisista aiheuttajista. Kuvitellaan nyt että yksi vangeista päästetään vapaaksi ja hän näkee auringonvalon ja sen, mistä muodot todella aiheutuvat. Vapautettu vanki haluaa vapauttaa toverinsa ja näyttää heille varjojen heijastaman todellisen maailman, mutta Platon tähdentää etteivät luolaan kahlehditut vangit halua vapautua eivätkä usko vapautuneen vangin kertomusta luolan ulkopuolisesta todellisuudesta.

Kuva 6: Luolavertaus (Wikipedia.org, allegory of truth).

Filosofiassa Platonin luolavertausta on tulkittu siten, että ihmiset ovat luolan sisälle kahlehdittuja vankeja, jotka erehtyvät pitämään luolan perälle heijastuvia varjoja todellisina. Todellisuus löytyy kuitenkin luolan ulkopuolella. Siellä oleva aurinko edustaa hyvän ideaa, ei luolassa palava tuli. Platon myös tähdensi, että todellisuus avautuu järkeilemällä, ei aisteilla. Nykyään voimme kuvitella erilaisten instituuttien, laitosten ja auktoriteettien olevan meille eräänlaisia muodonkantajia. Platonia mukaillen nämä muodot heijastavat meille todellisuutta (Huard 1996, s. 2). Kyseiset instituutiot ja auktoriteetit määrittelevät sen miten me maailman näemme. Platon toteaakin, että meidän pitäisi pystyä hahmottamaan omat rajamme järkemme avulla. Kun tiedämme rajoituksemme voimme siirtyä niiden toiselle puolelle, eli vapautua kahleista kuten luolaan sidotut vangit. Todellisuus on meille tällöin riippumaton instituuteista ja auktoriteeteista.

Mitä luolavertaus merkitsee tietojenkäsittelytieteelle? Seuraavaksi esitellään kaksi erilaista tapaa tulkita luolavertaus tietojenkäsittelytiedettä ajatellen. Ensimmäisessä voimme ajatella tietojenkäsittelytieteilijöitä vangitsijoina, jotka määrittelevät mitä vangit, muut ihmiset, näkevät. Tietojenkäsittelytieteilijän moraali on siis koetuksella. ”Kahlehditut” ihmiset uskovat helposti heille koneiden avulla luodun varjotodellisuuden. Vaikka muutama ihminen näkisi koneiden taakse kätkeytyvän todellisuuden, ei valtaosa ihmisistä uskoisi häntä. Ihmisten hakua varjojen luomaan haamutodellisuuteen kuvaavat hyvin erilaiset pelit. Ihmiset hakeutuvat virtuaalimaailmoihin, joista tulee heille todellisuutta. Esimerkiksi World of Warcraft-peliin on moni ihminen paennut todellisuutta. Toisaalta ei välttämättä tarvitse paeta virtuaalimaailmaan. Yhden ihmisen on mahdotonta tietää kaikesta kaikkea joten hänen on pakko uskoa mitä muut hänelle vieraan alan ammattilaiset sanovat. Tiedonhakuun ja tuottamiseen kun usein käytetään erilaisia ohjelmia, joten tietojenkäsittelytieteellä on viime kädessä moraalinen vastuu ohjelmien tuottaman tiedon oikeellisuudesta. On tietenkin selvää, että tietojenkäsittelytieteilijöille sattuu joskus virheitä, jolloin tulokset väärentyvät ja varjot eivät enää vastaakaan todellisuutta. Platonin voisikin kuvitella viestittävän, että juuri nämä virheet ”luolaihmisten” tulisi huomata. Tämä näkemys johtaa toiseen tapaan tulkita luolavertausta tietojenkäsittelytieteen alalla.

Koska tietojenkäsittelytieteilijä joutuu aina jollain tapaa mallintamaan todellisuutta ohjelmissaan sekä järjestelmissään, voidaankin sanoa, että tietojenkäsittelytieteilijän tekemän työn tuloksena syntynyt ohjelma tai järjestelmä on aina varjo todellisuudesta. Joskus varjot kuvaavatkin hyvin todellisuutta, mutta täyttä vastaavuutta on aina mahdoton saada. Koska jo syntyneitä järjestelmiä käytetään pohjana uusien järjestelmien luontiin, syntyy vanhoista varjoista uusia varjokuvia jolloin vieraannutaan yhä kauemmaksi todellisuudesta. Tietojenkäsittelytieteilijä ei siis välttämättä sido muita varjotodellisuuteen, vaan myös itsensä. Esimerkiksi tietokoneella voidaan

mallintaa aidon näköistä ruohoa, mutta ruohon hajua on tällä hetkellä mahdotonta saada koneelle mukaan. Ihmiset jotka koskaan eivät ole ruohoa nähneet, esimerkiksi aavikolla asuvat, saavat kyllä kuvan ruohosta. Emme voi kuitenkaan sanoa heidän tietäneen tai kokeneen mitä ruoho todellisuudessa merkitsee. On kuitenkin vaara, että tietojenkäsittelytieteilijä erehtyy pitämään omia luomuksiansa alkuperäisinä kuvina. Toisen tulkinnan voisi tiivistää, että tietojenkäsittelytieteilijä on sekä kahlitsija että kahlehdittu.

3.3 Tiedon tuottaminen

Tässä kappaleessa esitellään miten tietoa tuotetaan tietojenkäsittelytieteessä. Kappaleessa ei käsitellä tietojenkäsittelytieteessä vallitsevia paradigmoja, jotka määrittelevät metodit joilla tieteellinen tutkimus suoritetaan tietojenkäsittelytieteessä. Kappaleessa pääpaino on tietoa tuottavien metodien, tiedonlouhinnan ja asiantuntijajärjestelmien, esitteleminen.

Tiedonlouhinnassa pyritään löytämään uutta tietoa datajoukosta. Tutkielmassa käydään lävitse miten tiedonlouhinta määritellään, mihin sitä käytetään, millaisilla eri menetelmillä tiedonlouhintaa suoritetaan ja mikä on tiedonlouhinnan suhde tietoon.

Asiantuntijajärjestelmät ovat asiantuntijoiden opettamia järjestelmiä, jotka pyrkivät tukemaan asiantuntijoiden päätöksentekoa. Esimerkiksi lääkärit käyttävät taudin selvittämiseksi apunaan asiantuntijajärjestelmiä, jotka yrittävät oireiden avulla ennustaa mitä tautia potilas sairastaa. Seuraavaksi esitellään muutama erilainen asiantuntijajärjestelmä ja pohditaan voisivatko näiden antamat vastaukset olla tietoa.

3.3.1 Tiedonlouhinta

Jatkuva teknologian kehitys on tuottanut luonnollisesti suuren määrän uutta informaatiota. Kyseinen informaatio tallennetaan datana tietokantoihin, mikä onkin johtanut tietokantojen koon huimaan kasvamiseen. Informaation tulkitseminen tietokannoissa olevasta datasta ei kuitenkaan ole enää niin helppoa. Halutun datan seulomista raakadatan seasta kutsutaan *tiedonlouhinnaksi*. Hand et al. (2001, s. 1-4) esittävät tiedonlouhinnan (data mining) tarkoittavan suurten tietojoukkojen analysoimista, minkä tarkoituksena on löytää odottamattomia suhteita sekä tiivistää dataa uusilla ymmärrettävillä tavoilla. Tiedonlouhinnan tuloksena saadaan erilaisia malleja ja hahmoja, jotka voivat olla tilastollisia tai loogisia (Fayyad et al 1996). Nurminen (2005, s. 9) esittää omassa tutkielmassaan tiedonlouhinnan lähitieteiksi tilastotiedettä, tietokantoja, koneoppimista, hahmontunnistusta, tekoälyä ja visualisointia. Tiedonlouhinnan kohteeksi olevaa dataa ei yleensä

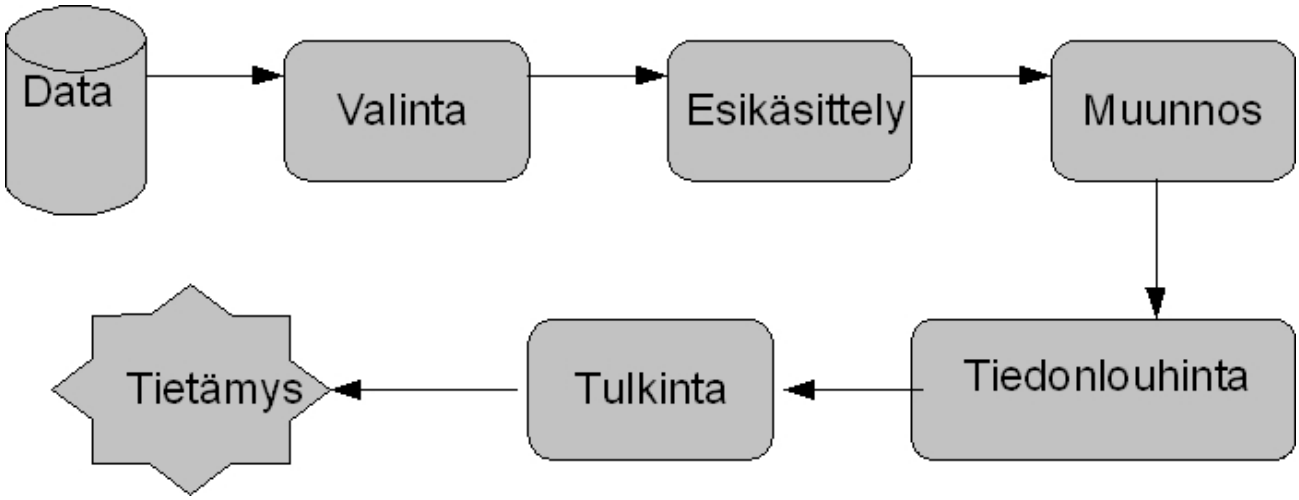
ole kerätty data-analyysia silmällä pitäen, joten tiedonlouhintaa on nimitetty myös sekundääriseksi data-analyysiksi (Nurminen 2005, s. 9).

Tietämyksen muodostaminen tietokannoista (Knowledge Discovery in Databases, KDD), jonka osana tiedonlouhintakin voidaan nähdä, määritellään Fayyad et al:n (1996) toimesta seuraavasti: ”Tietämyksen muodostaminen tietokannoista on epätriviaali prosessi, jossa pyritään muodostamaan päteviä, uusia, potentiaalisesti käyttökelpoisia ja lopulta ymmärrettäviä malleja datasta”. Tietämyksen muodostaminen tietokannoista on iteratiivinen sekä interaktiivinen, ja se voidaan jakaa seuraavaan viiteen osaan (Nurminen 2005, s. 9):

1) *Valinta*: valitaan kohde prosessille joko suuremmasta joukosta tai kerätään prosessin kohteena oleva tietojoukko.
2) *Esikäsittely*: datan puhdistus, poistetaan datasta turha ”kohina” sekä täytetään tyhjät havainnot.
3) *Muunnos*: datasta valitaan käsiteltävät ominaisuudet sekä suoritetaan datan mahdollinen muunnos datasta toiseen muotoon jatkokäsittelyä varten. Käytetään esimerkiksi tekstidatassa, missä tiedonlouhinta-algoritmit eivät käsittele suoraa tekstiä vaan siitä johdettua numeraalista esitystä. Tämä vaihe vastaa tilastollisen hahmontunnistuksen piirteiden valintaa ja erottelua.
4) *Tiedonlouhinta*: missä esitellään saadut tulokset, mikäli mahdollista, visualisoinnin avulla.
5) *Tulkintavaihe*: arvioidaan saadut tulokset ja mietitään mahdollista jälkikäsittelyä.

Kyseinen prosessi on myös kuvattuna kuvassa 7. Prosessista saadulla *tietämyksellä* tarkoitetaan tässä yhteydessä mallia tai hahmoa, joka on käyttäjän kannalta kiinnostava sekä riittävän varma (Frawley et al 1992). Ryan (2007) on määritellyt tietämyksen tarkoittavan, että henkilöllä on jonkinlaista käytännön kokemusta tukemaan tietoa. Tieto ei siis ole pelkästään faktojen muistamista. Esimerkiksi katsoessani keihäänheittoa tiedän, miten keihästä kuuluisi heittää. Minulla on tietämystä keihäänheitosta vasta, jos osaan itse heittää keihään pitkälle. Tietämys on mahdollista tulkita myös perinteiseksi nimitykseksi käytettävissä olevien tai hyväksyttyjen tietojen kokonaisuudelle. Tietämys ei siis ole mikään erikoinen lisäehtoja toteuttava tiedonlaji, vaan tietojen kokonaisuuden määrää ja laajuutta ilmaisema termi (Mäkipää & Ruohonen 2004, s. 2). Vastaavasti Kodratoff:n (1999) mukaan tietämyksen täytyy liittyä todelliseen maailman vaikuttaakseen sovelluksen käyttäjän toimintaan. Kodratoff:n kanssa voidaan olla myös eri mieltä. Esimerkiksi voi olla tietämystä, joka ei välttämättä liity todelliseen maailmaan, mutta vaikuttaa silti käyttäjän toimintaan. Platon kuvitteli toisen maailman, ideamaailman, joka ei liittynyt meidän maailmaamme,

mutta oli järjellä saavutettavissa. Tämä vaikutti kuitenkin varmasti jollakin tapaa Platonin toimintaan tässä maailmassa. Vaikka Platonin päättely ei tullutkaan edellä mainitun prosessin mukainen, voidaan silti tietämyksen muodostamisella tietokannoista saada tietämys, joka ei välttämättä liity tähän maailmaan, mutta vaikuttaa silti toimintaan. Käyttäjä voi saada jonkinlaisen tietämyksen tuonpuoleisesta elämästä, esimerkiksi näyn tai henkilökohtaisen kokemuksen ja alkaa näin elämään hurskasta elämää tässä maailmassa.



Kuva 7: Tietämyksen muodostamisprosessi (käännetty Fayyad el al 1996).

Nurminen esittää (2005, s. 9), että termit tiedonlouhinta ja tietämyksen muodostaminen ovat usein synonyymeja kirjallisuudessa. Termit eivät kuitenkaan ole täysin sama asia. Tietämyksen muodostamisprosessissa on painopiste tietokannoissa olevan tiedon analysoinnissa, kun taas tiedonlouhinta on laajempi käsite eikä näin ollen riippuvainen syötetiedon formaatista.

Kohteet

Perinteinen tiedonlouhinnassa käytetty tietojoukko on matriisi johon on koottu lueteltujen tai numeeristen muuttujien arvoja (Hand et al. 2001). Tietojoukoksi soveltuvat myös yksittäinen relaatiotietokannan taulu tai tietovarasto. Nykyään louhinta ei ole pelkästään numeerista, vaan on kehitetty uusia monimuotoisen syötetiedon louhintaan keskittyviä tutkimusalueita kuten tekstitiedostonlouhinta (text mining), web-louhinta (web mining) ja relaatiotiedon louhinta (multi-relational datamining) (Nurminen 2005, s. 10) .

Tekstitiedostonlouhinnassa (dokumenttien louhinta, tekstianalyysi) sovelletaan tiedonlouhinnan menetelmiä pääsääntöisesti rakenteettomiin ja puolirakenteisiin tekstidatoihin ja dokumenttijoukkoihin. Yleisimpiä sovelluksia ovat lyhennelmien generointi, dokumenttien

luokittelu, tekstitiedonhaun tehostaminen ja klusterointi (Dörre et al 1999).

Web-louhinnassa louhitaan tietoa nimen mukaisesti web-ympäristössä (Kosala & Blockeel 2000). Web-louhinnalle tunnusomaista on HTML- dokumenttien, linkkirakenteiden, web-sivustojen ja aihehakemistojen analyysi. Rakenteisen tekstin lisäksi WWW-sivustoissa esiintyy monimuotoista tietoa kuvien, äänien, animaatioiden, ohjelmakoodien (Java ja Active-X) ja valmistajakohtaisten sisältölaajennuksien (Flash) muodossa. Dokumentit ja sivustot ovat jatkuvasti muutoksen alla ja sisältävät kohinaa; web-dokumentit eivät välttämättä noudata kovinkaan tarkasti HTML-kielen skeemaa ja relevantti tieto on mainosten tai harhaanjohtavan metatiedon seassa (Chakrabarti 2003, s. 11-12). Etzioni (1996) jakaa web-louhintaprosessin seuraaviin vaiheisiin: resurssien löytäminen ja tiedon eristäminen sekä yleistäminen. Näistä kaksi ensimmäistä vastaavat web-hakukoneiden indeksointia tai KDD-prosessin kahta ensimmäistä vaihetta. Jälkimmäinen vuorostaan vastaa tiedonlouhinnan mallien löytämistä. Tähän voi vielä lisätä analyysivaiheen joka vastaa KDD-prosessin tulkintavaihetta (Kosala & Blockeel 2000). Nurminen (2005, s. 11) jakaa web-louhinnan vielä sisällön, rakenteen ja käytön analysointiin. Rakenneanalyysi voi koskea dokumenttien osien hausta aina useita sivustoja kattavien WWW-yhteisöjen kartoittamiseen. Web-palvelujen käyttäjälokien analysointi on yksi esimerkki käytönanalysoinnista.

Relaatiotiedonlouhinnassa etsitään malleja ympäristöstä, jossa useat tietojoukot ovat liittyneet toisiinsa erilaisilla suhteilla. Esimerkiksi relaatiotietokannan taulut voisivat olla yksi edellä mainitun kaltainen tietojoukko. Perinteiseen tiedonlouhintaan liitetään yleensä oletus, että tutkittavat näytteet ovat riippumattomia ja samasta jakaumasta (Independent and Identically Distributed, IID). Relaatiodatalla tämä oletus voidaan kuitenkin rikkoa, sillä kyseinen data voidaan ”pakottaa” yhteen tauluun korvaamalla viitetiedot toisessa taulussa olevilla ilmentymillä. Nurmisen (2005, s. 11) mukaan tämä prosessi on käänteinen operaatio tietokannan muodostamisen aikana tehtävälle normalisoinnille ja näin mahdollistetaan relaatiotiedon analysointi tavallisilla tiedonlouhintamenetelmillä. Hän jatkaa ettei IID-oletus ole enää voimassa, koska datassa on jo lähtökohtaisesti riippuvuuksia. Laitteistamisen yhteydessä kuitenkin eksplisiittinen riippuvuustieto häviää. Induktiivinen logiikkaohjelmointi (Inductive Logic Programming, ILP) on perusmenetelmä relaatiotiedonlouhintaan. Siinä etsitään ja yleistetään päättelysääntöjä annettujen esimerkkien ja taustatiedon avulla. Vastakohtana on deduktiivinen päättelysääntö, joka hyödyntää vain annettuja aksioomia ja päättelysääntöjä. Induktiivisessa päättelyssä pyritään löytämään uusia päättelysääntöjä eikä tyydytä vain valmiiksi annettuihin päätelmiin (Džeroski 2003).

Menetelmät

Tiedonlouhinnan yhteydessä puhutaan usein malleista ja hahmoista. Ero näiden kahden

välillä ei kuitenkaan aina ole selkeä, mutta kyseiset käsitteet ovat laajassa käytössä tiedonlouhintaa koskevassa kirjallisuudessa (Nurminen 2005, s. 12). Mallilla tarkoitetaan globaalia tiivistelmää tietojoukosta kuten regressiomallia $Y = aX + b$, jossa Y ja X ovat satunnaismuuttujia sekä a ja b mallin parametrejä. Hahmo (pattern) on väite tai sääntö, jolla kuvataan rajoitettua osaa datasta (Hand et al 2001, s. 9-11).

Tiedonlouhinnan menetelmät voidaan jakaa kahdella eri tavalla. Ensimmäisessä menetelmät jaetaan kuvaileviin ja ennustaviin. *Kuvailevissa menetelmissä* etsitään datan yleisiä ominaisuuksia. *Ennustavien menetelmien* avulla datasta pyritään tekemään hypoteeseja. Toinen tapa jakaa tiedonlouhinnan menetelmät on kategorioida ne seuraavanlaisesti: klusterointi, luokittelu ja assosiaatiosääntöjen etsiminen (Han & Kamber 2000, s. 21-28). Hand et al. (2001, s. 11-15), nimeää vuorostaan kategoriat kuvaileva mallinnus, ennustava mallinnus ja hahmojen ja sääntöjen etsiminen.

- *Klusteroinnissa* tietoalkioita ryhmitellään äärelliseen määrään klustereita niiden keskinäisen samanlaisuuden perusteella mikä määritellään etäisyysfunktiolla. Klusterointi on kuvaileva menetelmä kuten muuttujien välinen riippuvuusanalyysikin.
- *Luokittelu ja regressio* ovat ennustavia menetelmiä. Luokittelussa on ennalta määrättyjä luokkia joihin tietoalkio pyritään sijoittamaan. Regressiomalli on yleistys luokittelumallista ja siinä on korvattu luokat numeerisella arvolla.
- *Assosiaatioiden* ja *peräkkäisten toimintojen* etsintä. Assosiaatiomalli on kuvaileva malli toistuvasti yhdessä esiintyville tietueille (frequent itemsets). Peräkkäiset toimintosäännöt ovat ”ennustavia assosiaatioita, joiden kohdalla on tiedossa tietueiden suhteellinen järjestys.” (Nurminen 2005, s. 12). Assosiointisääntöihin liittyy kaksi tärkeää validointimittaa, tuki ja luottamus. Luottamuksella mitataan assosiaation todennäköisyyttä saatavilla olevan datan perusteella ja tuella mitataan kuinka suuressa osassa tietokantaa X ja Y esiintyvät yhdessä. Perusmenetelmä yhdessä esiintyvien tietueiden etsimiseen on Apriori-algoritmi. Siinä edetään rekursiivisesti yksittäisistä tietuista joiden esiintymistodennäköisyys, tuki, on kynnysarvon yläpuolella. Kokoelmaa laajennetaan asteittain suurempiin yksiköihin yhdistämällä kokoelman joukot ristitulolla itseensä ja samalla poistetaan ne joukot joiden osajoukot on jo valmiiksi poistettu. Vastaavasti tulosjoukoista poistetaan ne joiden tuki on kynnysarvon alapuolella (Han & Kamber 2000, s. 230-235).

  Lisäksi on muutama vähemmän käytetty menetelmä, esimerkiksi Hanin ja Kamberin (2000)

esittämä kuvausten ja koosteiden menetelmä, niin sanottu eksploratiivinen data-analyysi. Nurminen (2005, s. 13) kuitenkin kritisoi kuvausten ja koosteiden liittymistä lähes aina tietämyksen muodostamisprosessiin joten hänen mielestä sitä ei ole järkevää erotella omaksi ryhmäkseen. Samalla perusteella Nurminen (2005, s. 13) kritisoi Hanin ja Kamberin ehdottamaa poikkeamien (outliers) analyysia.

Yksittäiset tiedonlouhintamenetelmät voidaan vielä jakaa kolmeen osaan (Fayyad 1996): mallin esittämiseen, mallin arviointiin ja hakuun. *Mallin esittämisellä* tarkoitetaan kieltä jolla löydetyt hahmot kuvataan kuten esimerkiksi luonnollisella kielellä, loogisilla säännöillä tai kaavioilla (Frawley et al 1992). *Mallin arvioinnissa* arvioidaan nimensä mukaisesti mallia eli miten hyvin löydetyt hahmot täyttävät tietämyksen muodostamisprosessin tavoitteet. Ennustavia malleja arvioidaan ennustustarkkuudella ja kuvaavia malleja käyttökelpoisuudella ja ymmärrettävyydellä. *Haku* voidaan tehdä joko mallien tai parametrien etsintänä. Lisäksi on *tiedonhallintastrategia*, jolla pyritään ratkaisemaan vastaan tulevat ongelmat, kun tietojoukko ei mahdu kerralla muistiin (Hand et al. 2001, s. 15-18). Tiedonhallinnalla on kuitenkin omat ongelmansa varsinkin jos tiedonlouhintatehtävää ei voida tulkita tilastollisessa ympäristössä ja arviointikriteerit ovat epäselvät, kuten esimerkiksi klusteroinnissa on tilanne usein (Nurminen 2005, s. 13).

Suhde tietoon

Edellisissä kappaleissa esiteltiin tiedonlouhinnan perusajatus. Tutkielman kannalta ei ole kuitenkaan mielekästä keskittyä tiedonlouhinnan toimintaan syvällisemmin, vaan arvioidaan sen tuottamia tuloksia, malleja tieto-opillisesti. Voivatko mallit olla tietoa?

Mallit jaetaan joko kuvaileviin tai ennustaviin. Kuvailevat menetelmät keskittyvät etsimään datasta yleisiä ominaisuuksia ja ennustavissa menetelmissä pyritään tekemään datasta päätelmiä. Näyttää kuitenkin siltä, että ennustavissa menetelmissä on jo itsessään ristiriita. Sana ennuste viittaa johonkin joka voi olla totta tai epätotta. Tästä ennusteesta pitäisi tehdä päätelmä, jonka tulisi olla totta. Koska tiedonlouhinnasta saatu ennuste voi olla epätotta ei siitä seuraa automaattisesti, että päätelmä olisi totta, vaan päätelmästä voi, ja todennäköisesti tuleekin, epätotta. Tieto ei kuitenkaan voi olla ikinä epätotta, joten tämän vuoksi ennustavat menetelmät eivät tuota tietoa, vaan informaatiota.

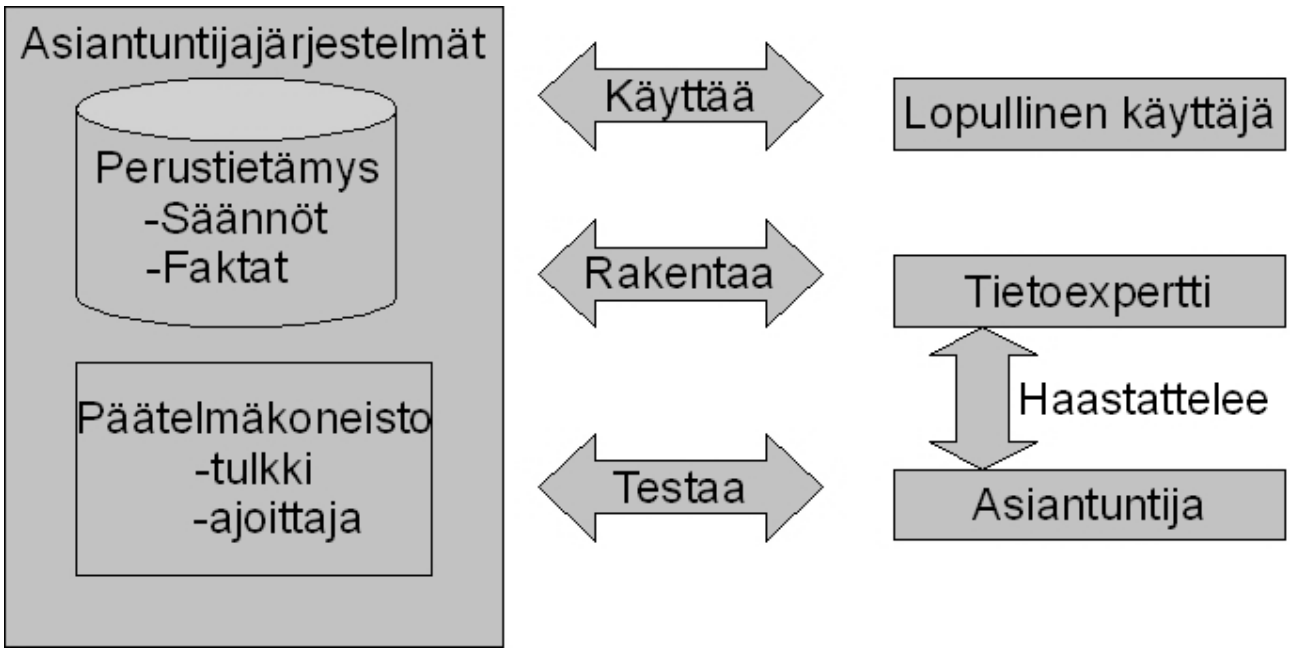
Kuvailevissa malleissa etsitään datasta yleisiä ominaisuuksia. Esimerkkinä voidaan mainita klusterointi, jossa data luokitellaan keskinäisten ominaisuuksien perusteella. Tämän voidaan ajatella tuottavan samanlaista tietoa kuin esimerkiksi lajien luokittelu. Vaikka filosofiassa yleiskäsitteistä voidaan kiistellä, niin tieteelle riittää että yleiskäsitteet ovat jollakin tapaa olemassa kuten sanoissa

tai mielessä. Konkreettista on yksilöllisten olioiden olemassaolo. Tämä antaa pohjan tiedonlouhinnan luokittelulle, jos luokiteltaville asioille on vastine todellisuudessa. Ainoa ongelma luokittelussa ovat rajatapaukset. Jos rajatapaukset voidaan luokitella aina oikein, voidaan olettaa että ainakin teoriassa kuvailevilla malleilla tuotetaan tietoa. Mikäli luokittelu kuitenkin menee edes joltain kohdalta väärin, on järkevää puhua kuvailevan tiedonlouhinnan tuotoksista informaationa.

3.3.2 Asiantuntijajärjestelmät

Asiantuntijajärjestelmät (expert systems, ES), ovat tietämysjärjestelmiä joihin on varastoitu jonkin kapean ongelma-alueen inhimillinen tietämys (Varpa 2005, s. 12). Tämä tietämyksen avulla asiantuntijajärjestelmien pitäisi pystyä ratkomaan ongelmia asiantuntijoiden tavoin. Ei kuitenkaan riitä, että järjestelmä matkisi asiantuntijoiden toimintaa, vaan sen tulisi myös pystyä perustelemaan ratkaisunsa, jolloin niitä vasta pidetään asiantuntevina (Varpa 2005, s. 12).

Kuvassa 8 esitellään yleisesti sääntöpohjaisen asiantuntijajärjestelmän rakenne, toiminta sekä eri osien välinen suhde (Laurikkala 2001, s. 5-6). Perustietämykseen (knowledge base) on tallennettu asiantuntijajärjestelmän erikoisalaan liittyvä tietämys ”if-then” muotoisiin sääntöihin. Päätelmäkoneistossa (Inference engine) on komentotulkki, joka määrittelee miten sääntöjä käytetään. Esimerkiksi mitä sääntöjä käytetään ongelman ratkaisemiseen ja missä järjestyksessä sääntöjä sovelletaan. Kysymystä ratkoessa muodostuu säännöistä päätelmäketjuja, joista saatu uusi tieto tallennetaan tietokantaan. Asiantuntijajärjestelmät käyttävät heuristista tapaa ratkoa ongelmia. Tämä mahdollistaa ratkaisujen pelkistämisen, millä saavutetaan nopeutta ongelman ratkomiseen. Onkin tärkeää, että asiantuntijajärjestelmät kykenevät käsittelemään puutteellista dataa, sillä järjestelmän tulisi toimia luontevasti, vaikkei kaikkiin päättelyn kannalta kriittisiin kysymyksiin olisikaan vastattu (Varpa 2005 s. 13). Algoritmeja käyttämällä on mahdollista saavuttaa täysin oikeat ratkaisut, mikäli kaikki säännöt on annettu, mutta haittapuolena on, että oikean ratkaisun selvittäminen vie aina oman aikansa. Se voi pahimmillaan tarkoittaa ettei ratkaisua pystytä selvittämään inhimillisessä ajassa. Asiantuntijajärjestelmät eivät kuitenkaan osaa käsitellä kunnolla tapahtumaa joka ei ole deduktiivisesti ratkaistavissa. Toisin sanoen asiantuntijajärjestelmät pyrkivät löytämään yksikäsitteisen ratkaisun tapaukselle, jolle ei ole yksikäsitteistä ratkaisua olemassa (Varpa 2005, s.13). Säännöt asiantuntijajärjestelmään laatii tietoexpertti haastattelemalla kyseisen alan asiantuntijaa. Asiantuntija huolehtii järjestelmän testauksesta, kun asiantuntijajärjestelmään on laadittu säännöt. Viimeiseksi testattu järjestelmä luovutetaan lopullisen käyttäjän käyttöön.



Kuva 8: Asiantuntijajärjestelmien rakenne (Käännetty Laurikkala 2001, s. 6).

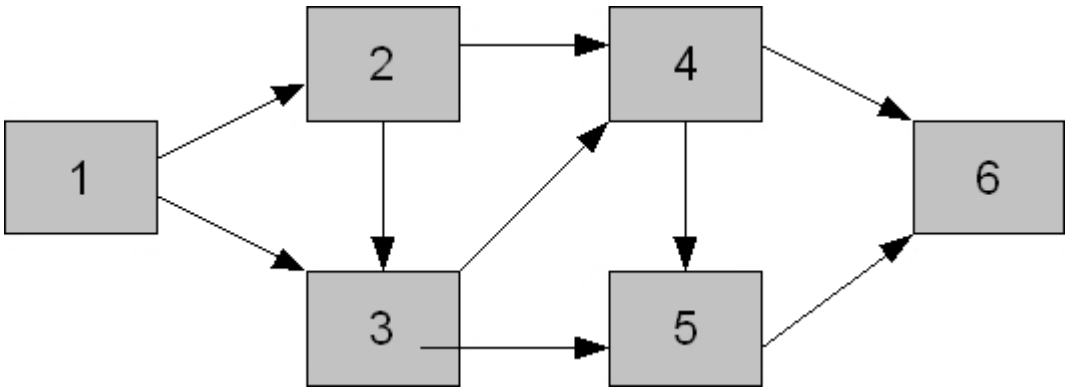
Aikaisemmin asiantuntijajärjestelmiä käytettiin vain vastaamaan kysymyksiin, samalla tapaan kuin antiikissa oraakkeleilta kyseltiin vastauksia. Nykyään asiantuntijajärjestelmiä käytetään tukemaan päätöksen tekoa eikä antamaan suoraa vastausta kysymyksiin. Siksi niitä onkin kutsuttu toiselta nimeltä päätöstukijärjestelmiksi (Decision support systems, DSS). *Päätöstukijärjestelmillä* ei ole mitenkään maailmanlaajuisesti hyväksyttyä yksikäsitteistä määritelmää, vaan teoriassa jokaista ohjelmaa joka auttaa päätöksenteossa voisi sanoa päätöstukijärjestelmäksi (Varpa 2005, s. 10) . Näin ollen asiantuntijajärjestelmiä voidaan ajatella päätöstukijärjestelmien erikoistapauksina. Toisaalta on myös esitetty näkemyksiä siitä, että asiantuntijajärjestelmät ja päätöstukijärjestelmät ovat toisiaan täydentäviä.

Suurin osa asiantuntijajärjestelmistä (30%) on jonkinlaisia sovelluksia, joita on kehitetty tukemaan diagnooseja ja ratkomaan diagnooseihin liittyviä ongelmia (Laurikkala 2001, s. 7). Suurimpia alueita missä asiantuntijajärjestelmiä käytetään on konetekniikka (engineering), kaupankäynti ja lääketiede (Laurikkala 2001, s. 7). Asiantuntijajärjestelmät ovat helpottaneet käyttäjäystävällisten järjestelmien suunnittelua. Tämän seurauksena asiantuntijajärjestelmät ovat kehittyneet prototyypeistä kaupallisiksi ohjelmiksi. Niiden tulevaisuus näyttää hyvältä, sillä uusia järjestelmiä kehitetään kokoajan. Lääketeollisuus säilyy kuitenkin jatkossa yhä alueena missä asiantuntijajärjestelmiä käytetään eniten (Laurikkala 2001, s. 7). Asiantuntijajärjestelmien apudiagnooseja suoritettaessa on usein korvaamaton. Seuraavassa kappaleessa esitellään erilaisia asiantuntijajärjestelmiä sekä tarkastellaan niiden tuottamaa tietoa verrattuna perinteiseen tiedon

käsitteeseen.

Bayesin verkot

Bayesin verkko on todennäköisyyttä kuvaava graafinen malli, joka kuvaa joukon muuttujia ja niiden välillä vallitsevia todennäköisiä suhteita. Bayesin verkot ovat suunnattuja syklittömiä verkkoja (directed acyclic graph, DAG) (Goldenberg & Moore, 2005). Syklittömyydellä tarkoitetaan sitä, ettei verkossa ole silmukoita. Kuvassa 9 on esitetty suunnattu verkko G, missä on pari (V,E), missä V on äärellinen joukko ja E on joukko pareja (v, w), ja v ∈ V ja w ∈ V. V:n alkiot ovat nimeltään solmuja (node) ja E:n alkiot ovat särmiä (edge). Suunnatussa verkossa särmiä nimitetään myös kaariksi (arc). Verkko on suunnattu, mikäli E:n parit ovat järjestettyjä. Suuntaamattomassa verkossa parit ei ole vastaavasti järjestettyjä, eli käytännössä tämä tarkoittaa yhtäläisyyttä (v,w) ja (w,v) välillä.

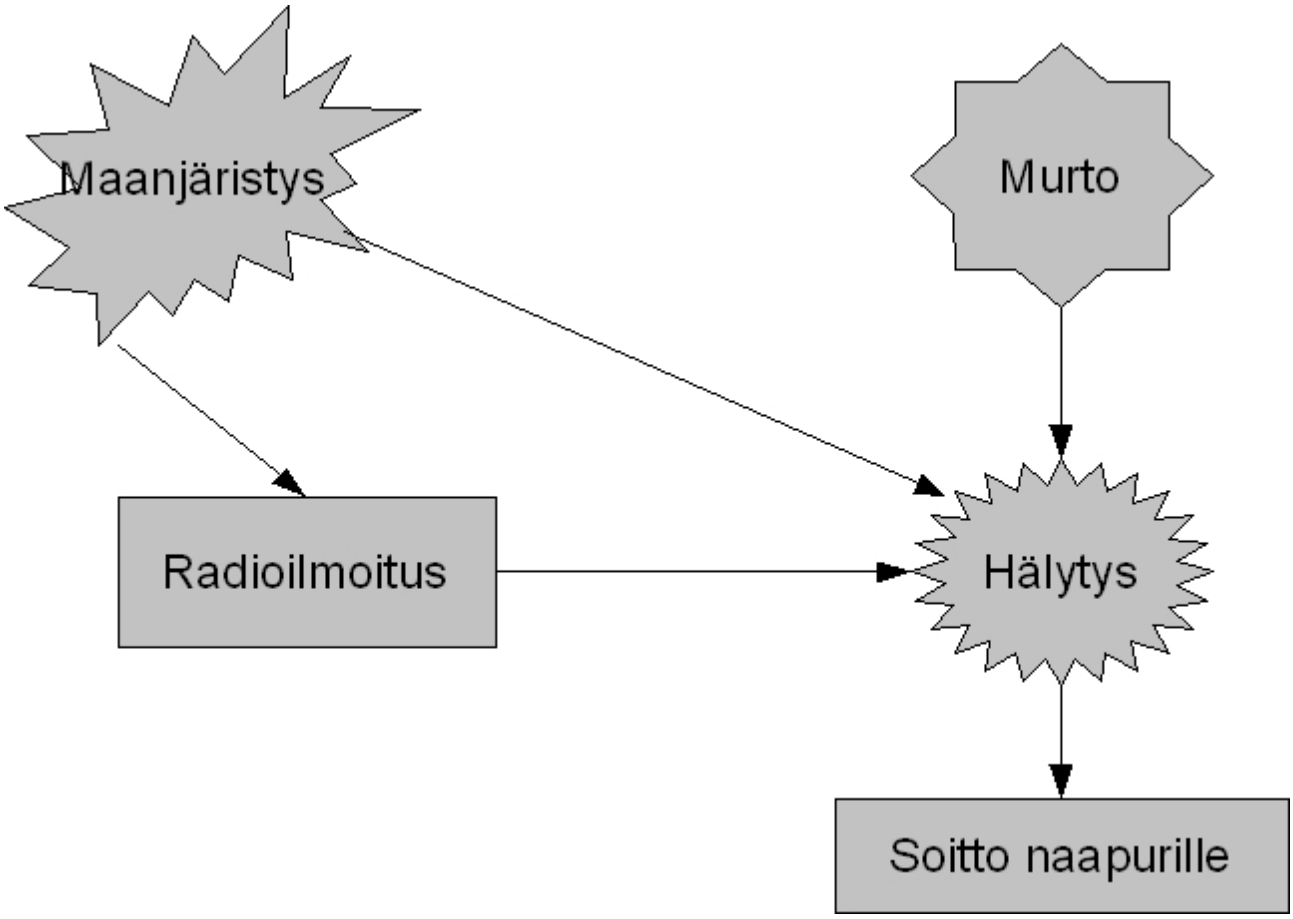


Kuva 9: Suunnattu syklitön verkko.

Bayesin verkot ovat siis suunnattuja syklittömiä verkkoja. Verkossa olevat solmut ovat Bayesin mallissa muuttujia ja solmuja yhdistävät kaaret kuvaavat muuttujien välillä olevia todennäköisyyksiä (Goldenberg & Moore, 2005). Bayesin verkkojen yksi erikoisuus on, että solmuissa olevat muuttujat voivat olla melkein mitä tahansa kuten esimerkiksi mitattuja tai piileviä parametreja, sekä erilaisia hypoteeseja. Bayesin verkkoa käydään sellaisten algoritmien avulla lävitse, joiden avulla saadaan verkosta todennäköisyystuloksia ja päätelmiä. Sellaisia Bayesin verkkoja, joissa on kuvattu muuttujien järjestys, nimitetään dynaamisiksi Bayesin verkoiksi (Harva et al, 2007). Dynaamisia Bayesin verkkoja käytetään esimerkiksi puhesignaaleissa (speech signals). Sellaiset Bayesin verkot, jotka ratkaisevat todennäköisyyksiin liittyviä ongelmia, mutta myös päätöksentekoon liittyviä ongelmia nimitetään vaikutuskaavioiksi (influence diagrams) (Howard & Matheson, 2005).

Kuvassa 10 esitellään typistetty Bayesin verkko, josta puuttuvat kaarien ja solmujen väliset todennäköisyydet. Verkossa on kuitenkin kuvattuna Bayesin verkkojen idea. Sitä luetaan ylhäältä

alaspäin ja siinä on kuvattuna yksinkertaistettuna murron ja maanjäristyksen aiheuttamat seuraukset. Maanjäristyksen sattuessa seurauksena on kuvan mukaan radioilmoitus sekä hälytys. Varkauden sattuessa tehdään myös hälytys ja soitetaan naapurille. Verkkoa voi lukea myös toisinpäin, eli alhaalta ylös. Esimerkiksi hälytyksen sattuessa voidaan päätellä, että syynä on joko varkaus tai maanjäristys. Kun tiedossa on vielä maanjäristyksen ja varkauden todennäköisyydet, pystytään laskemaan millä todennäköisyydellä seuraa mistäkin mitä.



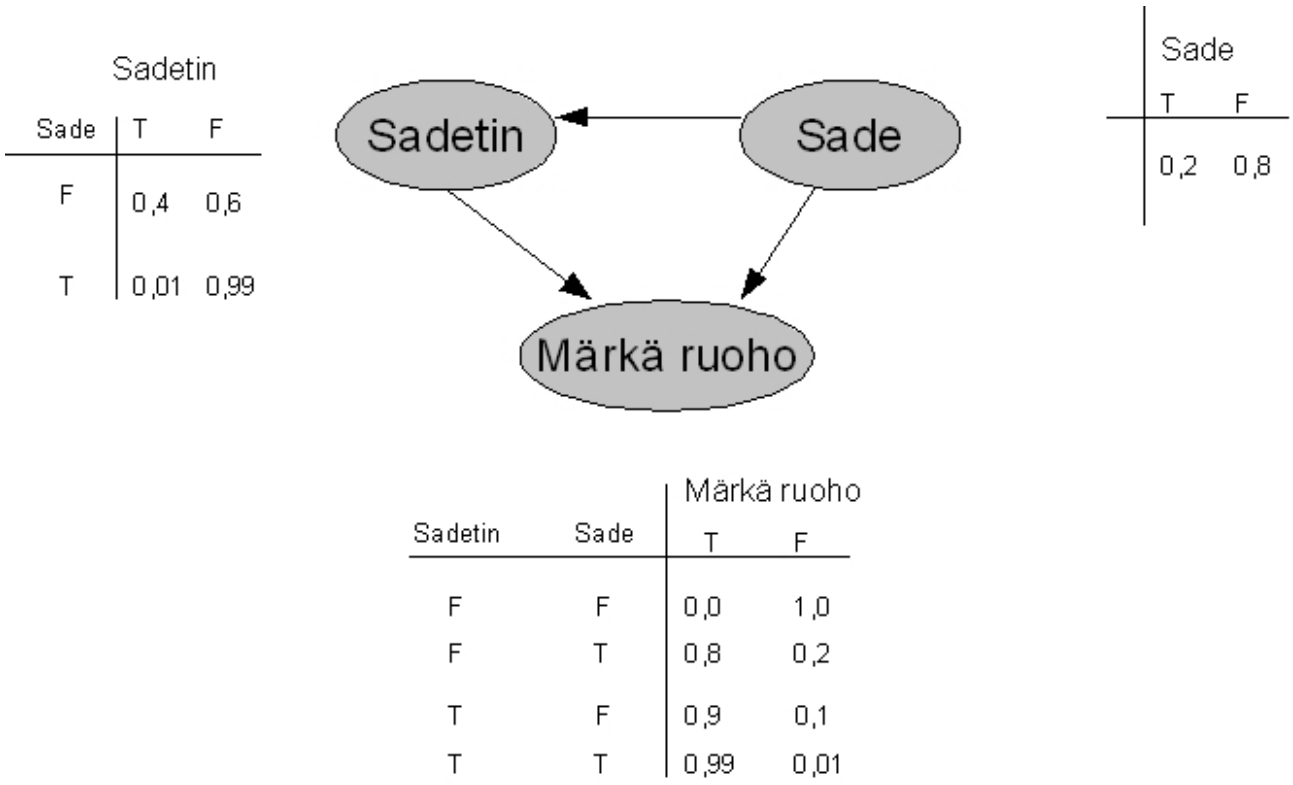
Kuva 10: Yksinkertainen Bayes-verkko (käännetty Niedermayer 1998).

Kuvassa 11 esitellään toinen Bayesin verkko. Tällä kertaa mukana on myös eri kaarien ja solmujen todennäköisyydet. Kuvassa 11 on kuvattu kaksi syytä miksi ruoho voi olla märkää. Kuvan mukaisesti syynä voi olla joko sade tai sadetin. Lisäksi voidaan olettaa, ettei sateella käytetä sadetinta, eli sateen sattuessa on sadetin pois päältä. Kaaviossa on siis kolme muuttujaa, sade, sadetin ja märkä ruoho. Näitä kuvataan siten, että G = märkä ruoho, S = sadetin ja R = sade. Jokaisella muuttujalla on kaksi arvoa, joko se on tosi eli T tai epätosi F. Eri solmujen välillä olevat todennäköisyydet on kerrottu kuvassa 11. Siinä kuvatun verkon avulla on mahdollista esimerkiksi

selvittää millä todennäköisyydellä sade on kastellut ruohon tai millä todennäköisyydellä ruoho ylipäätänsä on märkää. Se, millä todennäköisyydellä sade olisi kastellut ruohon saadaan selville käyttämällä ehdollisen todennäköisyyden kaavaa P(A, B). Kun tähän sijoitetaan kuvan 12 muuttujat niin saadaan :

P (R = T | G = T)

Toisin sanoen, vastaamme kysymykseen: ” Millä todennäköisyydellä sade on kastellut ruohon?”.



Kuva 11: Painotettu Bayesin verkko (käännetty Wikipedia.org, SimpleBayesNet).

Bayesin verkko on täydellinen malli muuttujista ja niiden välillä vallitsevista suhteista ja todennäköisyyksistä. Näitä tietoja käytetään myös hyväksi ennustavia kysymyksiä tehtäessä. Esimerkiksi verkkoa hyväksi käyttämällä voidaan saada päivitettyä tietoa jostain kyseisen verkon osajoukosta, kun muita muuttujia tutkitaan. Tällaista tutkintametodia kutsutaan nimellä todennäköisyyspäätelmät (probabilistic inference) (Kozlov & Singh, 1994) . Se muistuttaa induktiivista päättelyä, eli tietystä havaintojoukosta tehdään yleisiä sääntöjä. Nyt havainnot vastaavat vain Bayesin verkon muuttujia joihin uutta tutkittavaa muuttujaa verrataan.

Yksi tärkeimpiä Bayesin verkkoja koskevia yksittäisiä metodeja on muuttujien eliminointi

(variable elimination) (Stracuzzi & Utgoff 2004). Yleensä Bayesin verkot toteutetaan erilaisina puina, joten yksinkertaisuus on yksi niiden etu, joka seuraa lokaalisuudesta. Puussa olevat alikomponentit ovat vain rajoitetun määrän verran tekemisissä muiden komponenttien kanssa, riippumatta komponenttien kokonaismäärästä. Tästä seuraa kompleksisuuden lineaarinen kasvaminen exponentiaalisen kasvamisen sijasta. Tämän takia tulisi heikkoja riippuvuuksia, kuten muuttujia joita ei tutkita tai joihin ei kohdisteta kyselyjä, poistaa (Stracuzzi & Utgoff 2004). Tämä tietysti heikentää hieman verkon tarkkuutta, mutta pienentää kompleksisuutta. Vähäisen kompleksisuuden etuina ovat mahdollisuus tehdä samanaikaisia kyselyjä tehokkaasti, sekä tilan ja muistin säästäminen. Myös Bayesin verkkoon tehtävät lisäykset tulisi tehdä oikein, ettei kompaktisuus kärsisi. Muihin muuttujiin vaikuttavat muuttujat olisi saatava mahdollisimman alas puussa, eli juuriin. Lehtiin jäisivät näin ollen vain ne muuttujat, jotka eivät vaikuta enää mihinkään muuhun muuttujaan.

Yksinkertaisin tapa rakentaa Bayesin verkko on antaa asiantuntijan tehdä tarvittavat verkkoa koskevat määritykset, minkä jälkeen verkko on valmis käytettäväksi. Aina tämä ei kuitenkaan ole mahdollista, vaan verkon määrittely voi hyvinkin olla liian vaikeaa ihmisille. Näissä tapauksissa verkon rakentaminen, rakenne ja sille syötettävät parametrit opetellaan tutkittavasta datasta. Bayesin verkon rakenteen oppiminen on hyvin tärkeää koneoppimisessa (machine learning). Kun oletetaan, että data on generoitu Bayesin verkosta ja kaikki muuttujat ovat näkyvissä jokaisessa iteraatiossa, voidaan optimointiin perustuvan etsimismetodin avulla löytää verkon rakenne. Käytännössä tämä tarkoittaa, että on oltava funktio, joka mittaa mallin sopivuutta dataan ja mallin ennustuskykyä. Tällaista funktiota nimitetään pistemääräfunktioksi (Scoring Function). Tämän lisäksi on oltava jonkinlainen etsimisstrategia siitä, miten informaatiota/rakennetta lähdetään selvittämään, ja mikä auttaa pistefunktion käytössä (Search Strategy) (de Campos 2006). Yksi yleinen tapa laatia pistefunktio on harjoitusdatan rakenteen posteriori todennäköisyyden ratkaiseminen (de Campos 2006). Esimerkiksi Markov Chain Monte Carlo (MCMC) on eräs etsimisalgoritmi.

Bayesin verkoista on erityisesti hyötyä lääketieteessä diagnoinneissa, kun on analysoitava potilaan oireiden perusteella häntä vaivaava tauti. Lääkäri antaa verkolle tiedot parametreina oireista, jolloin hän saa vastaukseksi erilaisia oireisiin täsmääviä tauteja. Toinen yleinen käyttötapa Bayesilaisille verkoille on roskapostisuodatus. Käytännössä se tarkoittaa sähköpostien lajittelua eri kategorioihin Bayesin tilastomenetelmien avulla (Sundström 2008, s. 26-28). Esimerkkeinä Sundströmillä (2008, s. 28) on POPfile, SpamProbe, Bogofilter, DSPAM ja dbacl, jotka perustuvat Bayes menetelmään.

Bayesilaisessa roskapostisuodatuksessa käytetään hyväksi Bayesin teoreemaa. Bayesilaisen

roskapostisuodatuksen mukaan todennäköisyys sille, että tietyssä viestissä jossa on tietyt sanat on yhtä suuri kuin todennäköisyys löytää kyseiset sanat mistä tahansa roskapostista kerrottuna todennäköisyydellä sille, että mikä tahansa viesti on roskapostia, jaettuna todennäköisyydellä sille, että kyseiset sanat esiintyvät missä tahansa sähköpostiviestissä (O'Brien & Vogel 2003). Formaalissa muodossa edellinen on:

P(roska |sanat) = (P (sanat | roska) P(roska)) / P (sanat)

Tietyillä sanoilla ja otsikoilla on tietty todennäköisyys esiintyä niin roskapostiviestissä kuin luvallisessakin viestissä. Esimerkiksi sukupuoliyhteyteen viittaavat sanat esiintyvät usein roskaposteissa, mutta harvoin asiallisissa sähköpostiviesteissä. Suodatin ei kuitenkaan osaa laskea tyhjästä todennäköisyyksiä, vaan suodattimien opetus jää käyttäjän vastuulle. Käytännössä tämä tarkoittaa, että käyttäjällä pitää olla joitakin varmoja roskapostiviestejä jotka hän antaa suodattimen tutkittavaksi. Kun suodatin tietää viestin olevan varmasti roskapostia, se säätää annettujen viestien perusteella tietokantaan jokaisen sanan kohdalla niiden esiintymistodennäköisyyksiä luvalliselle ja luvattomalle postille. Näin ollen sanat kuten viagra, porn saavat hyvin suuren roskatodennäköisyyden toisin kuin esimerkiksi luvallisissa postissa esiintyvät sanat kuten sukulaisten ja ystävien nimet. Kyseisillä sanoilla on siis vastaavasti korkea todennäköisyys, sille että posti olisi luvallista. Suodatin oppii myös tuntemaan nopeasti kielen tuottaman eron, mikäli käyttäjän saamat asialliset postit ovat suomeksi ja roskapostit usein englanniksi. Käyttäjän pitää olla kuitenkin tarkkana, koska suodatin saattaa helposti luokitella asiallisiakin sähköposteja roskaposteiksi. Hyvä esimerkki erilaisista asiallisista posteista, jotka saatetaan luokitella roskaposteiksi, ovat käyttäjän tilaamat englanninkieliset mainoslehtiset ja uutiskirjeet.

Kun Bayesin verkolle on syötetty tarpeeksi oppimateriaalia eli roskapostia ja tavallista postia voi varsinainen sähköpostin suodatus alkaa. Oppimateriaalista saatujen tietojen perusteella lasketaan todennäköisyyksiä, sille että sähköposti jossa on tietyt sanat olisi roskapostia. Tätä vaikutusta kutsutaan posteriori- todennäköisyydeksi ja se lasketaan Bayesin teoreemaa soveltaen, minkä jälkeen lasketaan todennäköisyys sille että sähköposti kaikkine sanoineen olisi roskapostia (Yao et al. 2004, s. 249). Jos saatu yhteenlaskettu todennäköisyys ylittää ennalta sovitun rajan, vaikkapa 90%, luokitellaan viesti roskapostiksi ja siirretään jatkokäsittelyä varten eli se joko poistetaan tai siirretään roskapostikansioon.

Mitä enemmän käyttäjä kouluttaa roskapostisuodatintaan, sitä paremmin se oppii suodattamaan käyttäjälle lähetettyjä viestejä. Jokaisella käyttäjällä on oma erikoissanastonsa ja mieltymyksensä, minkä seurauksena kahta täysin identtistä suodatinta ei synny vaan käyttäjän posteissa on henkilökohtaista riippuvuutta. Se määrittelee kullekin käyttäjälle ominaiset

todennäköisyysfunktiot. Esimerkiksi potenssiapua hakenut voi saada sähköpostiinsa erilaisia potenssimainoksia, jolloin suodatin oppii karsimaan tarkemmin erilaiset potenssimainokset ennen kuin ne alkavat haitata käyttäjää. Samoin suodatin oppii nopeasti luvallisissa posteissa usein esiintyvät sanat. Esimerkiksi jos käyttäjällä on harrastuksena bridgeilloissa käynti ja hän keskustelee niistä ystäviensä kanssa sähköpostin välityksellä, oppii suodatin nopeasti ettei bridgeviestejä tulisi suodattaa. Suodattimet tekevät virheitä, mutta kehittyvät kunhan käyttäjä jaksaa niitä opettaa. Mikäli bayesilaisessa suodattimessa käytetään hyväksi Markov-ketjuja on mahdollista kehittää suodatinta siten että se huomaa myös roskaposteissa esiintyvien sanojen muodostamia yhdistelmiä. Useat roskaposteissa esiintyneet sanat ovat yksinään harmittomia, mutta roskaposteissa usein esiintyvät samat sanat yhdessä, ja juuri Markov-ketjujen avulla nämä sanojen yhdistelmät on mahdollista havaita bayesin verkoilla.

Bayesin verkkoja käytetään lääketieteen ja sähköpostisuodattimien lisäksi esimerkiksi konetekniikassa ja luonnontieteissä luokittelemaan dataa. Esimerkiksi AutoClass-ohjelmalla luokitellaan tähtiä spektriominaisuuksien mukaan, jotka olisivat muuten niin hienovaraisia, ettei niitä huomattaisi. Viime aikoina on myös esitetty aivojen käyttävän hyväkseen bayesin verkkoja luokitellessaan aistiärsykkeitä ja tehdessään päätelmiä käyttäytymisreaktioista (Knill & Pouget, 2004).

Mikä on Bayesin verkkojen ja tiedon suhde? Tiedoltahan vaaditaan, että se on tosi, perusteltu ja uskomus. Päätösten perustelu eli miksi tuotettu informaatio olisi tietoa onnistuu Bayesin verkoilta. Deduktiivisen päättelyn voidaan ajatella tapahtuvan syllogismi muodossa eli kun tiedetään, että jos A niin siitä seuraa B. Kun verkko tietää että X:llä on A niin se pystyy päättelemään minkä seuraus A on tai mitä A:sta seuraa. Päätöksille on siis jonkinlainen perustelu kuten tiedon käsite vaatii. Perustelu ontuu hieman, koska jo Hume opetti, että vaikka A ja B havaittaisiin yhdessä, ei se tarkoita että A:sta seuraisi B. Bayesin verkothan eivät väitäkään A:sta seuraavan B:tä, vaan esittävät todennäköisyyden, että A:sta seuraisi B. Tämä ei kuitenkaan riitä, vaan tiedon vaatimushan on myös totuus mikä ei ole sama asia kuin todennäköisyys. Bayesin verkkojen tuottama ”tieto” on näin ollen rinnastettavissa kappaleessa 3.1.2 esiteltyyn tilastolliseen ja kausaaliseen tietoon. Mikäli muuttujien välillä vallitseva todennäköisyys olisi 100% voitaisiin puhua tiedon tuottamisesta. Tämä tietysti on ristiriidassa Bayesin verkkojen ajatusta vastaan. Kuitenkin todennäköisyydet aiheuttavat sen ettei voida puhua tiedon tuottamisesta, vaan ennemminkin ennusteista.

Neuroverkot

Alun perin termiä *neuroverkko* on käytetty viittaamaan biologisiin neuroneihin. Nykyään

termiä käytetään usein viittaamaan niin sanottuihin keinotekoisiin neuroverkkoihin jotka muodostuvat keinotekoisista neuroneista ja solmuista eli yhtymäkohdista (Fu, 1999). Edellisen perusteella neuroverkoilla on siis kaksi erilaista merkitystä:

1) *Perinteinen biologinen neuroverkko*, mikä koostuu keskushermostoon yhteydessä olevista oikeista neuroneista. Neurotieteessä niitä käytetään usein suorittamaan erilaisia tehtäviä laboratorioanalyyseissa.

2) *Keinotekoiset neuroverkot*, mitkä koostuvat keinotekoisista neuroneista. Niillä voi kuitenkin olla joitain oikeiden neuroneiden ominaisuuksia. Keinotekoisia neuroverkkoja käytetään pääsääntöisesti ymmärtämään biologista neuroverkostoa sekä ratkomaan tekoälyyn liittyviä ongelmia ilman että tarvitsisi yrittää simuloida oikeaa hermoverkostoa.

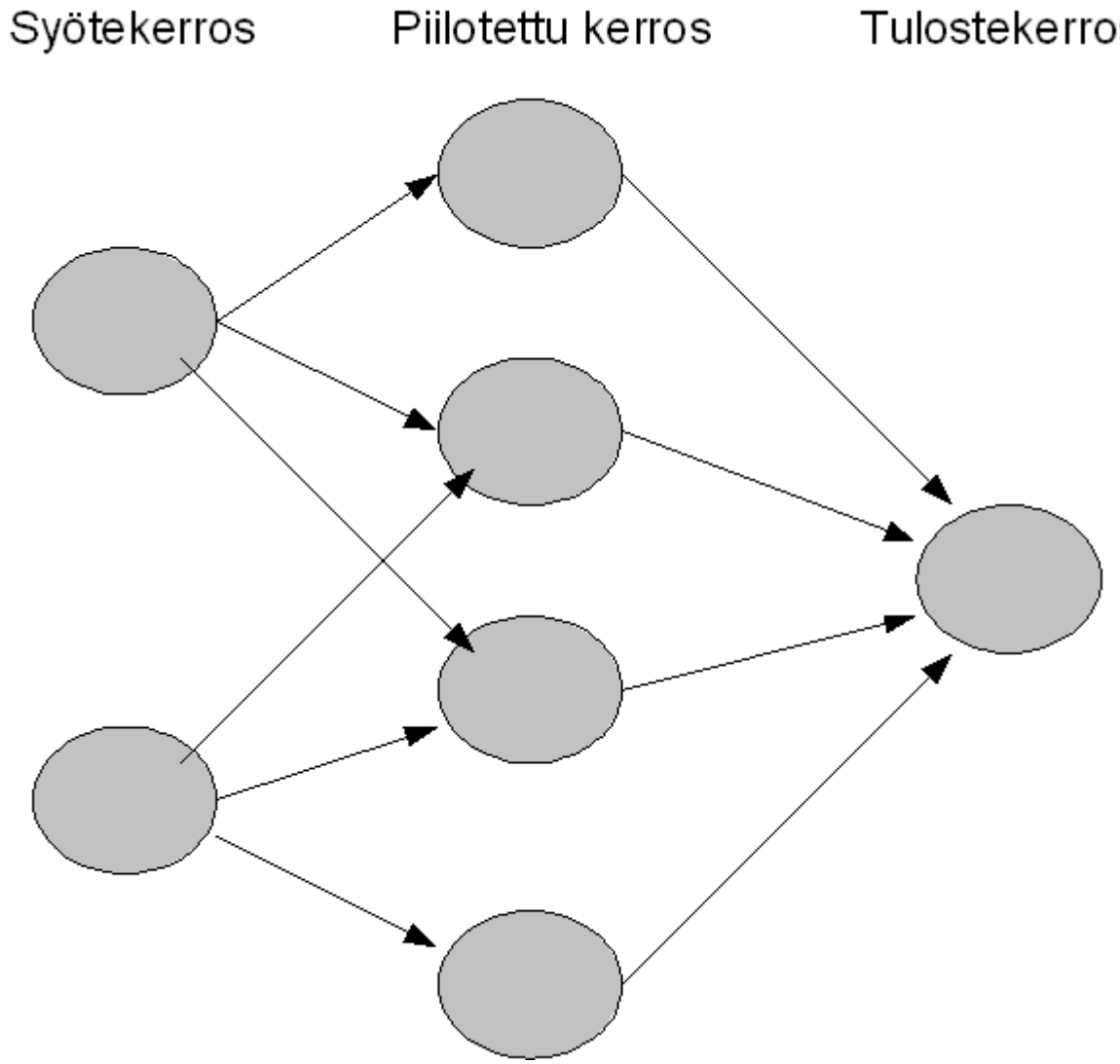
Tutkimuksen kannalta keinotekoiset neuroverkot ovat kiinnostavampia mutta niiden ymmärtämiseksi olisi hyvä esitellä hieman biologista neuroverkkoa.

Biologinen neuroverkko siis koostuu ryhmästä tai ryhmistä joko kemiallisesti tai funktionaalisen toiminnan avulla toisiinsa liittyneistä neuroneista. Neuroni voi olla yhteydessä useisiin muihin neuroneihin jolloin verkossa olevat neuronien väliset yhteydet saattavat muodostua hyvinkin laajoiksi. Neuroverkossa vallitsevia yhteyksiä nimitetään synapseiksi (Fu, 1999). Synapsit muodostuvat aksonien ja dendriittien välille. Sähköisen viestinnän lisäksi myös erilaiset välittäjäaineet huolehtivat viestinnästä. Kaiken kaikkiaan neuroverkot ovat hyvin monimutkaisia järjestelmiä eikä niitä ole pystytty selittämään läheskään täydellisesti.

Keinoäly- ja kognitiivinen, tiedollinen, mallinnus pyrkivät molemmat jäljittelemään ainakin joitain ihmiselle tyypillisiä älykkyyttä vaativia toimintoja. Vaikka molemmat jäljittelevätkin neuroverkkoja samanlaisilla tekniikoilla on molemmilla eri päämäärä. Keinoälyssä pyritään ratkomaan yksittäisiä tehtäviä neuroverkkojen avulla, kun taas kognitiivisessa mallissa pyritään mallintamaan neuroverkkoja matemaattisiksi malleiksi. Keinoälyssä neuroverkkoja on käytetty menestyksekkäästi esimerkiksi puheen tunnistuksessa sekä kuva-analyysissa. Nykyisin kuitenkin neuroverkkoja hyödyntävää keinoälyä on käytössä estimointiteoriassa (statistical estimation), optimoinnissa (optimization) ja kontrolliteoriassa (control theory) (Miikkulainen 2007).

Keinotekoista neuroverkkoa (An artificial neural network, ANN) myös nimitetään simuloiduksi neuroverkoksi (simulated neural network, SNN) tai neuroverkoksi (neural network, NN). Keinotekoisessa neuroverkossa on toisiinsa yhdistettyjä keinotekoisia neuroneja jotka käyttävät apunaan matemaattisia tai tietokoneella tehtyjä laskennallisia malleja informaation käsittelyyn. Useimmissa tapauksissa keinotekoinen neuroverkko on oppivainen järjestelmä, joka

muuttaa rakennettaan sisäisen tai ulkoisen informaation perusteella (Fu, 1999). Toisin sanoen, keinotekoinen neuroverkko on epälineaarinen datan mallinnusväline ja päätöksenteon apuväline. Sitä käytetään mallintamaan syötteiden ja tulosteiden suhteita sekä etsimään datassa olevia kaavioita. Kuvassa 12 on kuvattuna keinotekoinen neuroverkko ja sen rakennetta miten dataa käsitellään eri osissa. Syötekerroksessa verkolle annetaan syöte, jonka muokkaus tapahtuu piilotetussa kerroksessa ja siitä saatu tulos välittyy tulostekerrokselle ja lopulta käyttäjälle.



Kuva 12: Keinotekoinen neuroverkko (Wikipedia.org, Neural network example).

Neuroverkkoja käyttävän mallin etuna on, että pystytään esittelemään erilaisia funktioita havainnoimalla verkossa olevaa dataa. Tämä on hyödyllistä, jos dataa on epäkäytännöllistä tutkia käsin.

Niemenlehdon (2004, s. 23-24) tutkielman perusteella keinotekoisten neuroverkkojen käyttö voidaan jaotella seuraavasti:

1. Funktioiden arviointi ja regressioanalyysi. Tähän ryhmään kuuluu myös mallinnus (modelling).
2. Luokittelu, kuten kaavioiden ja jatkuvuuksien tunnistaminen.
3. Datan käsittely, kuten suodattaminen ja ryhmittely.

Neuroverkot kykenevät mukautumaan uuteen tietoon, eli toisin sanoen, neuroverkkoja on mahdollista opettaa. Verkko siis muuttaa painokertoimia saamiensa uusien syötteiden perusteella (Fu, 1999). Opetusparadigmoja on kolme (Niemenlehto 2004, s. 22):

1. *Ohjatussa oppimisessa* (Supervised learning) verkolle annetaan syötteitä ja opetetaan se vastaamaan annettuihin syötteisiin halutulla tavalla.
2. *Ohjaamattomassa oppimisessa* (Unsupervised learning) verkko joutuu ohjauksen puuttuessa mukautumaan jonkin tietyn kriteerin mukaisesti annettuihin syötteisiin. Opetusjoukossa ei kuitenkaan ole haluttuja vasteita, toisin kuin ohjatussa oppimisessa.
3. *Vahvistus eli palauteoppimisessa* (Reinforcement learning) ideana on, että verkko tutkii ympäristöään, syötteitä, havaitsee jonkinlaisen tilan ja toimii sen mukaisesti. Toiminnasta saadaan aina palaute, joko positiivinen tai negatiivinen. Ideana on löytää ratkaisu josta saadaan eniten positiivista palautetta. Vahvistusoppimisen ja ohjatun oppimisen erona on, ettei vahvistusoppimisessa käytetä syöte- tulos- pareja, eikä heikkoja ratkaisuja välttämättä aina korjata. Oppimisprosessi voi olla elinikäinen työ, jossa tasapainotellaan tunnettujen reittien käyttämisen ja tutkimattomien reittien kartoituksen välillä.

Mikä sitten on neuroverkkojen suhde tiedon tuottamiseen? Tässä tapauksessa tiedon vaatimusten täyttämisen määritteleminen ei ole yksinkertaista. Keinotekoiset neuroverkothan pyrkivät jäljittelemään mahdollisimman tarkasti biologista neuroverkkoa. Miten tarkasti tämä jäljittely onnistuu jääkin epäselväksi. Mikäli jäljittely olisi täydellistä niin keinotekoisten neuroverkkojen tiedon tuottamisprosessi olisi samanlainen kuin aidoillakin biologisilla neuroverkoilla. Tällöin neuroverkko tuottaisi samalla tavalla tietoa, kuin ihminenkin. Biologisten neuroverkkojen toimintaa ei ole voitu täydellisesti selvittää, joten niiden jäljittely tuskin onnistuisi täydellisesti tämänhetkisten tietojen perusteella. Toisin sanoen, neuroverkot täyttävät oppivien ja älykkäiden menetelmien kriteerejä, vaikka eivät ihmisen tasolle ylläkään. Neuroverkkojen toimintaa tehostetaan opettamalla niitä. Ne siis oppivat toimimaan tietyllä tavalla tietyllä syötejoukolla. Tällöin tiedon tuottaminen ei ole enää riippuvainen verkosta, vaan opettajasta. Neuroverkot vain toimivat opettajan ajattelun

jatkeena. Epäselväksi kuitenkin jää mikä on neuroverkon opettajan suhde opettamaansa asiaan. Opettajalla on oltava tietoa asiasta X ennen kuin hän voi sitä eteenpäin opettaa. Ongelma palautuu näin ollen juuritasolle, eli voiko meillä olla tietoa ylipäätänsä, mistä filosofit ovatkin vuosisatoja kiistelleet.

Päätöspuut

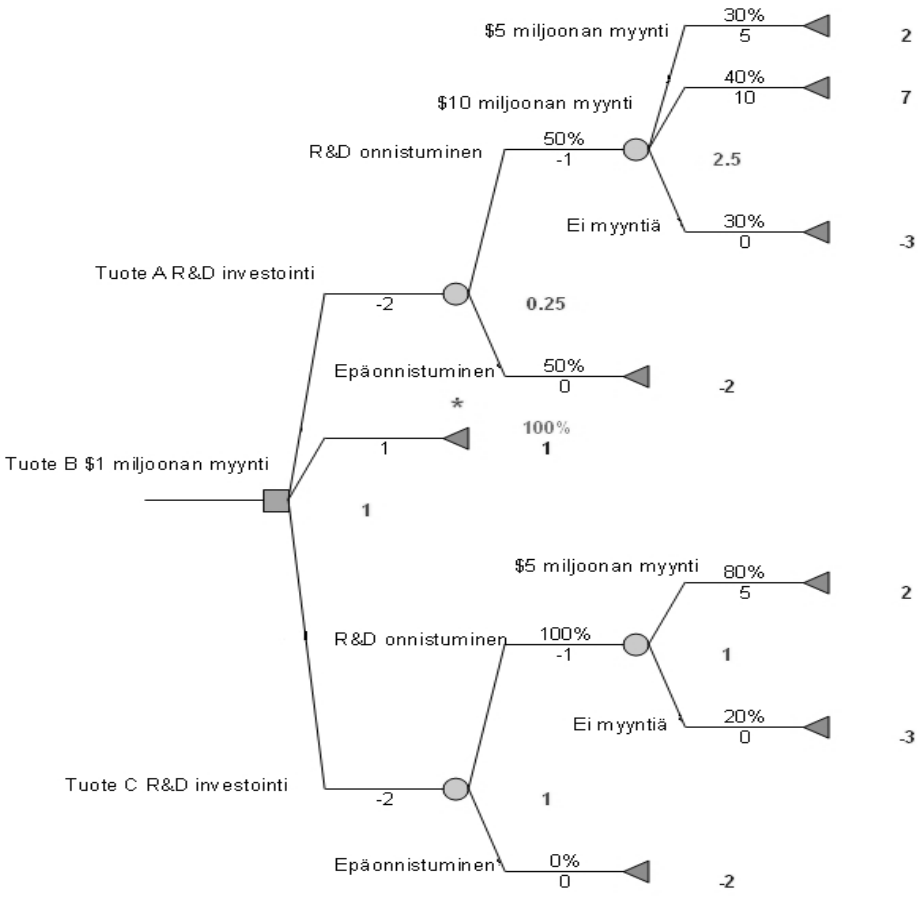
Päätöspuut (decision trees) helpottavat päätöksenteossa. Niissä on kuvattuna tai mallinnettuna eri päätökset ja niiden seuraukset. Päätöspuita käytetäänkin päämäärään vaadittavan strategian etsimiseen sekä todennäköisyyksien laskemiseen. Niiden avulla muodostetaan yleisiä luokittelusääntöjä sovellusalueen esimerkkijoukosta. Tällöin ne ovat induktiivista päättelyä hyödyntäviä menetelmiä (Quinlan 1986). Päätöspuut ovat yksi tehokkaimmista oppimismenetelmistä (Varpa 2005, s. 49). Puun eri osat Varpa (2005, s. 49) luokittelee seuraavanlaisesti: puiden lehtisolmut ovat miellettävissä luokiksi, sisäsolmut ominaisuuksiksi tai niiden arvoja tutkiviksi testisolmuiksi ja oksat ominaisuuksien arvoiksi tai arvoväleiksi. Varpan (2005, s. 49-50) mukaan puut ovat myös tulkittavissa eräänlaiseksi sääntöjoukoksi jolloin niiden avulla on mahdollista johtaa uusia sääntöjä tietämyskantaan. *Tietämyskannan* on määritellyt Laukkanen (2007, s. 11) tietokannaksi minne esimerkiksi yrityksissä tallennetaan kaikki hyödylliset ohjeet joiden avulla jonkin tietyn asian tekeminen on mahdollista. Laukkasen mukaan (2007, s.11) tietämyskannoilla pyritään vähentämään hiljaisen tiedon määrää. Laukkasen määritelmän mukaan tietämyskanta on tietokanta tiedonhallintaan. Päätöspuun puurakenteiset luokittelusäännöt ovat myös purettavissa erillisiksi sääntöjoukoiksi, jolloin on mahdollista yksinkertaistaa sääntöjen ehtolauseita alkuperäisestä rakenteesta (Varpa 2005, s. 51). Quinlanin (1986) mukaan päätöspuita on mahdollista käyttää luokitteluun lähes millä tahansa sovellusalueella, kunhan opetusjoukon tapaukset ovat esitettävissä muuttujapareina ja tapausten attribuutit ovat riittäviä luokkien erotteluun. Päätöspuiden kyky käsitellä epätäydellistä dataa tekee niistä yhden hyödyllisimmistä koneoppimismenetelmistä (Varpa 2005, s. 51). Varpan (2005, s. 51) mukaan päätöspuiden prosessointi on nopeaa, koska sen ei tarvitse käydä läpi kaikkia kohteesta annettuja muuttujia ja niiden yhdistelmiä. Päätöspuiden muodostama tietämys häviää ilmaisuvoimassa semanttisille verkoille (Quinlan 1986). Fayyad et al:n (1996) mukaan rajallisempi ilmaisuvoima helpottaa oppimaan ja hyödyntämään puun oppimismenetelmää.

Quinlanin (1986) mukaan päätöspuu rakentuu seuraavasti. Lähtökohtana ovat opetusjoukon attribuuttien arvot. Kyseisestä joukosta valitaan puun juurisolmuksi parhaaksi todettu attribuutti. Todentaminen tapahtuu esimerkiksi erilaisia informaation määrään tai tilastolliseen merkittävyyteen pohjautuvien heuristiikkojen avulla (Varpa 2005, s. 50). Kun attribuutti on valittu, muodostetaan

puuhun solmu johon lisätään oksiksi attribuuttiin liittyvät arvot tai sen arvovälit. Puun muodostaminen tapahtuu rekursiivisesti opetusjoukosta. Aluksi kaikki opetusjoukon tapaukset sijaitsevat samassa juurisolmussa. Muuttujan valinnan jälkeen opetusjoukkoa aletaan jakamaan alijoukkoihin muuttujan arvojen perusteella. Mikäli alijoukon tapaukset kuuluvat samaan luokkaan tai lopetusehto täyttyy luodaan kyseisen oksan päähän lehtisolmu. Muulloin jatketaan puun muodostamista etsimällä alijoukon paras attribuutti. Puun muodostaminen tapahtuu rekursiivisesti opetusjoukosta. Näin jokaisen opetusjoukon tapaus sijoitetaan johonkin lehtisolmuun ja se pyritään määrittelemään vain yhteen tulosjoukkoon kuuluvaksi kerrallaan (Varpa 2005, s. 50). Puuttuvan tiedon tapauksessa opetusjoukon tapaus luokitellaan useampaan tulosluokkaan ja liitetään tulosjoukkojen esiintymistodennäköisyys. Quinlanin (1986) mukaan puun koon rajoittaminen, luokittelusäännön yksinkertaistaminen, on mahdollista karsinnan (tree pruning) avulla. Esikarsinta suoritetaan puun muodostamisen yhteydessä. Kun valmiista puusta poistetaan haaroja on kyseessä jälkikarsinta.

Päätöspuut pystyvät perustelemaan päätöksensä. Polku lehtisolmusta juurisolmuun muodostaa luokittelusäännön ja perustelee annetun lopputuloksen. Päätöspuumenetelmän tavoitteena on muodostaa mahdollisimman matala puu. Tämä onnistuu jakamalla muuttujat omiin luokkiinsa mahdollisimman suppealla muuttujamäärällä (Varpa 2005, s. 50). Quinlan (1983) toteaakin, että mitä yksinkertaisempi opetusjoukosta muodostettu luokittelusääntö on, sitä varmemmin tulevat uudet tapaukset luokiteltua oikein.

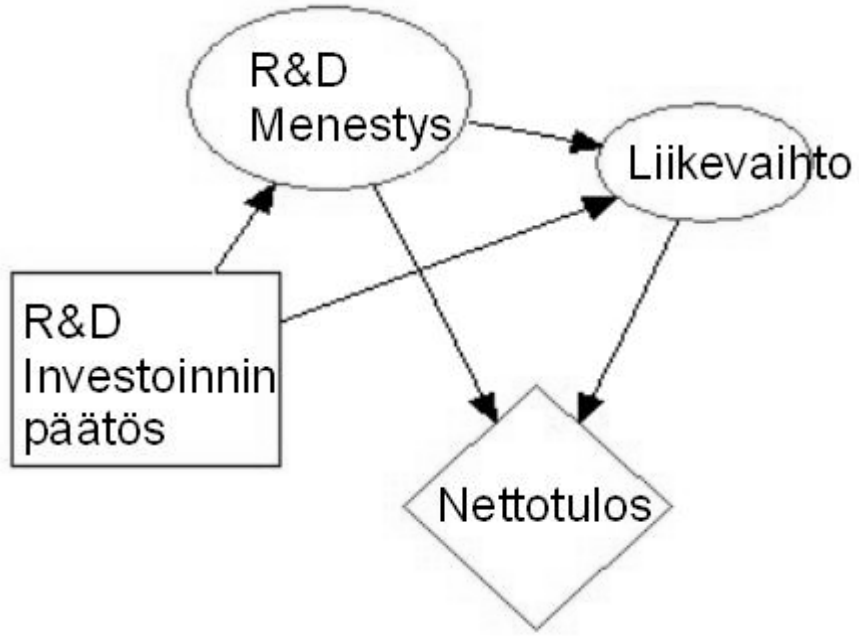
Alla olevassa kuvassa 13 on kuvattuna päätöspuu. Siinä johtajan pitäisi päättää sijoittaako hän rahansa tuotteeseen A vai tuotteeseen C. Tuotte A vaatisi 2 miljoonan dollarin tutkimus- ja kehityssijoituksen (Research and Development, R&D), ja todennäköisyys tutkimisen onnistumiselle on vain 50%. Mikäli tutkimus onnistuu niin tuotteella A on 30% todennäköisyys tuottaa 5 miljoonan dollarin voitto, 40% todennäköisyys 10 miljoonan dollarin voittoon ja 30% todennäköisyys ettei tuote myy. Tuote C vaatii myös 2 miljoonaa dollaria kehitykseen, mutta tuotteen kehittäminen onnistuu 100% varmuudella. Se tuottaisi 80% todennäköisyydellä 5 miljoonan dollarin voitot ja 20 % todennäköisyys olisi ettei tuote myisi ollenkaan. Molempiin tuotteisiin olisi lisäksi laitettava miljoona dollaria tuotantokustannuksia. Kumpaan tuotteeseen johtajan tulisi sijoittaa, kun oletetaan että yrityksen tavoitteena on aina maksimoida odotusarvo? Kuvassa 13 näkyy tarvittavat vaihtoehdot, todennäköisyydet, maksut ja odotusarvot. Tässä esimerkissä molemmat tuotteet A ja C oletettavasti tuottaisivat voittoa, mutta C:llä on korkeampi odotussarvo miljoonalle dollarille.



Kuva 13: Päätöspuu (käännetty Wikipedia.org, Dtree2).

Päätöspuut voidaan myös tiivistää *vaikutuskaavioihin*, joissa keskitytään kuvaamaan välttämättömät elementit, päätökset, epävarmuudet ja päämäärät sekä näiden vaikutukset toisiinsa. Kuvassa neljätoista on kuvattu vaikutuskaavio. Siinä neliö, R&D investoinnin päätös, on päätöksentekijän kontrolloitavissa. R&D:n investoinnin päätös vaikuttaa R&D:n menestykseen sekä liikevaihtoon. Jollei kehitykseen investoida ei myöskään ole odotettavissa onnistumisia tuotekehityksen suhteen. Sen sijaan ympyrät, R&D menestys ja liikevaihto, eivät ole

päätöksentekijän hallinnassa. Onnistunut tuotekehitys vaikuttaa liikevaihtoon ja nettotulokseen. Salmiakin muotoinen neliö, nettotulos, on päämäärä jonka päättäjä pyrkii maksimoimaan liikevaihdolla ja tuotekehityksen menestymisellä.



Kuva 14: Vaikutuskaavio (käännetty wikipedia.org, Factory2 Influence Diagram).

Vaikutuskaaviot kuvaavat päätöstilannetta jollain tietyllä hetkellä, eivät koko prosessia. Vaikutuskaavioiden osien suhdetta kuvataan solmuilla ja kaarilla (Howard & Matheson 2005). Vaikutuskaavioissa ei esiinny syklejä. Esimerkkejä tilanteista joissa voidaan käyttää vaikutuskaavioita käytetään esimerkiksi seuraavanlaisissa tilanteissa: riskitilanteissa kun pohditaan rahan investointia, epätäydellisen informaation tilanteissa, peräkkäisissä päätöksissä ja laskutoimituksissa joissa mietitään lasketaanko uusi tuote markkinoille kuten kuvassa 14.

3.4 Tiedon pakkaus

Tietojenkäsittelytieteessä tiedonpakkaus on yksi yleisimmistä tavoista käsitellä tietoa. Miten tiedolle käy tiedon pakkauksessa? Kappaleessa esitellään häviöllinen ja häviötön tapa pakata tietoa. Lisäksi käydään läpi mitä tiedon pakkauksella tarkoitetaan tietojenkäsittelytieteessä ja pohditaan samalla mitä tekemistä tiedon pakkauksella ja tiedolla on keskenään. Onko siis käsite tiedon pakkaus ylipäätänsä mielekäs ja mitä loppujen lopuksi pakataan?

3.4.1 Häviötön ja häviöllinen pakkaus

Tietojenkäsittelytieteessä *tiedon pakkauksella* tarkoitetaan menetelmää jossa tietoaines korvataan lyhyemmällä kuvauksella (Harris 2001). Tiedon tiivistäminen onkin mahdollista, koska useimmiten tallennettavan tiedon kuvaamiseen käytetään enemmän tilaa kuin todellisuudessa tarvittaisiin. Tieto on siis tilastollisesti redundanttia. Esimerkiksi ”Jesse kirjoittaa tutkielmaa” voitaisiin korvata muotoon ”J kirjoittaa T”. Tässä on ”Jesse” korvattu J:llä ja tutkielma T:llä, sillä oletuksella että kyseisillä merkeillä ei ole muuta merkitystä tekstissä (Harris 2001). Tiedon tiivistäminen on tärkeää, koska näin pyritään vähentämään tallennus- ja tiedonsiirtokapasiteetin käyttöä. Pakkauksen haittapuolena on kuitenkin laskentatehon vaatiminen, minkä takia voidaan joutua investoimaan kalliisiin laitteisiin. Siksi tiedon pakkaaminen vaatii monien asioiden huomioimista, jotta maksimaalinen hyöty saavutettaisiin.

Tiedon pakkaamiseen on kaksi erilaista menetelmää, häviöllinen (lossy compression) ja häviötön (lossless compression) (Harris 2001). *Häviöttömässä* menetelmässä todellista tietoainesta ei häviä pakkaamisen yhteydessä, vaan se kuvataan toisella tavalla. Kun pakkaus puretaan, saadaan sama tietoaines kuin ennen pakkausta. *Häviöllisessä* menetelmässä tietoa hävitetään tarkoituksella. Siinä pyritään kuitenkin hävittämään ihmisen kokemuksen kannalta merkityksetöntä tietoa. Esimerkiksi televisiolähetyksestä on mahdollista poistaa värejä ilman, että ihmisen silmä sitä huomaisi. Samoin musiikista on mahdollista poistaa korkeita ja matalia ääniä, joita ihminen ei muutenkaan kuulisi. Teoriassa häviölliset menetelmät heikentävät aina pakattavan tiedon laatua. Kuitenkin kun häviöllisin menetelmin pakataan siten, että ihminen ei huomaa eroa, päästään usein häviöttömiin menetelmiin verrattuna parempiin pakkaussuhteisiin (Harris 2001). Häviöllisiä menetelmiä ei kuitenkaan voida soveltaa symboliseen tietoon kuten tekstiin, taulukoihin tai ohjelmakoodiin, koska ne eivät kestä yhdenkään bitin muuttamista yksittäistapauksia lukuun ottamatta (Harris 2001).

Eräs yksinkertaisimmista pakkausmenetelmistä on juovakoodaus (RLE) (Hertell 2005, s. 9). Sen keskeisin idea on, että tiedostoissa esiintyy usein samaa arvoa peräkkäin. Esimerkiksi kuvissa on usein sama väriarvo peräkkäisissä pisteissä (Hertell 2005, s. 9) . Näin ollen voidaan kuvata yksi samanvärinen juova tiedolla sen väristä ja pituudesta, sen sijaan että kuvattaisiin jokaisen pisteen väri erikseen. Juovakoodaus onkin esimerkki häviöttömästä pakkauksesta mikä on tärkeää esimerkiksi tietokoneohjelmien ja mittaustulosten pakkaamisessa, koska pienikin muutos sisällössä voi aiheuttaa virheellistä tulkintaa. Häviöllistä pakkausta käytetään usein ihmisen havaintoon perustuvaan tietoon. Esimerkiksi äänenpakkauksessa käytetään psykoakustiikkaa hyväksi hävittämään ääniä joita ei ihmiskorva kuule. Näin saadaan esimerkiksi arkistoitua CD-levy

kiintolevylle pienempään tilaan Ogg Vorbis muodossa (Xiph 2003). Puheenpakkausta käytetään esimerkiksi internet-puheluissa. Ihmiset ovat usein valmiita hyväksymään häviöllisen pakkausmenetelmän, mikäli muutokset eivät häiritse ymmärtämistä tai niitä ei koeta epämiellyttäväksi. Pienempi levytila tai vähäisempi kaistankäyttö korvaa usein pakkauksen aiheuttaman häviön. Häviöllistä pakkausmenetelmää sovelletaan kuitenkin jossain määrin myös kuviin (Harris 2001). Esimerkiksi edullisissa digitaalikameroissa käytetään häviöllistä menetelmää, koska näin muistiin mahtuu enemmän kuvia. DVD-video-levyissä käytetään MPEG-2 menetelmää videon pakkaamiseen. Häviöllisiä menetelmiä käytettäessä on tärkeintä ottaa huomioon suorituskyvyn tarve sekä tiivistyneen tiedon ja häviön suhde.

On selvää että jos häviötön pakkaus on tehty oikein, ei sitä purettaessa tieto muutu mihinkään, vaan säilyy samanlaisena kuin ennen pakkaustakin. Kuitenkin tilanne on toinen häviöllisessä pakkauksessa. Materiaalista hävitetään niin sanottu ihmiselle ”turha” tieto. Tällöin väite: ”tässä Mozartin kappaleessa on paljon korkeita ääniä joten en suosittele soittamaan sitä eläinten kuullen”, muuttuu epätodeksi, koska häviöllisessä pakkauksessa korkeat äänet on poistettu. Onko kyse enää tiedosta, kun siitä poistetaan osia, vaikka ihminen ei niitä huomaisi? Ei ole, koska tieto kuvaa maailmassa vallitsevaa asiantilaa, ja kun siitä poistetaan osia ei se välttämättä enää kuvaa kyseistä asiantilaa. Vaikkei ihminen kuulisi jotain ääntä, ei se tarkoita etteikö jokin muu voisi kuulla. Esimerkiksi eläimen kutsuhuutoa pakattaessa voi siitä hävitä ääniä, joita vain eläimet voivat kuulla. Kutsuhuuto menettää merkityksensä, vaikkei ihminen eroa huomaisikaan. Onko eläimen kutsuhuuto tietoa? Mikään ei estä sitä etteikö eläimillä voisi olla tietoa, kunhan tiedon vaatimukset täyttyvät. Eläimen kutsuhuuto on tosi, jos muut eläimet vastaavat kutsuhuutoon. Esimerkiksi sudet kutsuvat ulvomalla toisiaan. Sudella on varmasti perustelut huudolleen kuten uskomuskin. Esimerkiksi lauma on kehittänyt yhteiset säännöt ja määritelmät kutsuhuudolle. Näin kutsuhuuto ainakin teoriassa täyttää tiedon vaatimuksen, eli eläimellä on tietoa lajinsa kutsuhuudosta. Toinen esimerkki elävästä elämästä on koirapilli, jonka äänen vain koirat kuulevat. Tieto voi siis muuttua informaatioksi käytettäessä häviöllistä pakkausta. Häviöllistä pakkausta käytettäessä kannattaa miettiä onko se sen arvoista, sillä jotain olennaista saattaa hävitä.

3.5 Tieto eri termeissä tietojenkäsittelytieteessä

Tähän kappaleeseen on koottuna muutamia tietojenkäsittelytieteen käsitteitä, joissa esiintyy sana tieto harhaanjohtavasti. Näillä esimerkeillä pyritään havainnollistamaan miten tieto-sanan väärinkäyttö on jäänyt luontevasti kielenkäyttöömme. Termejä on tietenkin paljon enemmän kuin tutkielmassa on esitetty, mutta niissä käytetään pääsääntöisesti vain tieto-sanaa väärin datan tai

informaation sijasta. Esimerkiksi tietokone, tiedonhaku ja tietorakenne ovat termejä, joissa tieto-sana on korvannut data- tai informaatio-termin.

3.5.1 Tietoverkot

Tietoverkko on määritelty WSOY:n suuressa tietosanakirjassa tietokoneiden muodostamaksi kokonaisuudeksi, johon tavalliset käyttäjät voivat tietokoneensa liittää modeemin avulla (Halinen et al 2001, s. 824). Toisin kuin WSOY:n suuressa tietosanakirjassa sanotaan, tietoverkot eivät enää rajoitu vain pelkkiin tietokoneisiin ja modeemeihin, vaan verkkoon on mahdollista liittyä muillakin laitteilla kuten matkapuhelimilla ja pelikonsoleilla. Nämä eri laitteet voidaan kuvitella erilaisiksi pisteiksi, ja niitä yhdistävät viivat tietoverkoiksi.

Tietoverkolle synonyymeja ovat tiedonsiirtoverkko, dataverkko, datansiirtoverkko ja tietoliikenneverkko. Tietoverkkoja on myös erilaisia. Ajoneuvojen ja koneiden sisäistä väylää nimitetään CAN-väyläksi (Controller Area Network) (Guesmi & Rezig, 2006). Esimerkiksi CAN-väylää käytetään autossa moottoriohjausyksikön, ABS-jarruyksikköjen ja vaihteistonohjausyksikön väliseen kommunikointiin. CAN-väylää käytetään myös busseissa, hisseissä, maatalouskoneissa ja roboteissa. Sitä voidaankin käytännössä käyttää missä tahansa koneessa, kun sanomat ja tiedonsiirtoetäisyydet ovat lyhyitä ja tarvitaan reaaliaikaista prosessorien välistä kommunikointia. Likiverkkoa (PAN, Personal Area Network), käytetään yhdistämään henkilökohtaisia elektronisia laitteita. Tietokoneeseen voidaan liittää esimerkiksi tulostimia, kameroita ja matkapuhelimia. Kommunikointi tapahtuu USB-kaapeleilla ja porttitoistimilla, mutta voi olla myös langatonta (WPAN, Wireless Personal Area Network) (Manasis et al. 2004). Maantieteellisesti pienellä alueella toimivia verkkoja kutsutaan lähiverkoiksi (LAN, Local Area Network). Esimerkiksi yrityksessä samassa rakennuksessa olevien koneiden muodostamaa verkkoa nimitetään lähiverkoksi. Tunnetuimpana verkkona voidaan pitää internettiä. Se on yleisnimitys yhteenliitetyille alueellisille ja paikallisille tietoliikenneverkoille, jotka on liitetty toisiinsa internet-protokollan, IP:n, avulla.

Tietoliikenneverkot välittävät meille dataa eri laitteiden välityksellä. Välitetystä datasta muodostamme informaatiota. Mikään ei kuitenkaan takaa, että tämä informaatio olisi tietoa. Siksi nimitys ”tietoverkko” on hämäävä. Teoriassahan on mahdollista, että välitetty data ei milloinkaan täyttäisi tiedon vaatimuksia, eli tietoverkoilla ei olisi mitään tekemistä itse tiedon kanssa. Tämän takia nimitys dataverkko olisi kuvaavampi kuin tietoverkko.

3.5.2 Tietokanta

Tietojenkäsittelytieteessä tietokokoelmat, joilla on jonkinlaista yhteyttä toisiinsa nimitetään *tietokannaksi* (database). Esimerkiksi yrityksien keräämät tiedot asiakkaistaan muodostavat tietokannan. Jotta tietokanta toimisi, olisi sen osien välillä oltava looginen yhteys. Tietokokoelmien ei kuitenkaan tarvitse olla sähköisessä muodossa muodostaakseen tietokannan. Tietokantaa voidaan pitää myös yllä kynän ja paperin avulla. Esimerkiksi kalenteri on ei-sähköinen tietokanta. Tietokantojen koko vaihtelee hyvin suuresti yhteen tiedostoon tallennetuista taulukosta usealla eri kovalevyllä sijaitseviin tietueisiin.

Tietokantoja voidaan toteuttaa usean eri mallin mukaan, vaikka nykyiset tietokannat on pääsääntöisesti toteutettu yhden mallin määräämällä tavalla. Tuki useammalle kuin yhdelle mallille on kuitenkin yleistymässä. Lisäksi loogisesti suunnitellut mallit on mahdollista muuttaa myös fyysiseen muotoon. Alla on esiteltynä erilaisia malleja, joiden avulla tietokantoja voidaan laatia.

- ”Flat File” -malli on kaksiulotteinen taulukko, jossa oletetaan sarakkeissa olevan samanlaisia arvoja ja rivissä olevien arvojen liittyvän toisiinsa. Tietokannan tieto siis esitellään yhdellä rivillä (Schahczenski 2000) . ”Flat File” tietokanta sijaitsee usein tiedostossa puhtaassa tekstimuodossa. Esimerkki ”Flat File” tyyppisestä tietokannasta on puhelinluettelo, jossa nimeä seuraa numero.
- Hierarkkisessa mallissa data on jaoteltu kuten puu- tietorakenteessa (Schahczenski 2000). Tämä mahdollistaa tiedon toistamisen käyttämällä lapsi-vanhempi-suhdetta. Jokaisella lapsella on vain yksi vanhempi, mutta yhdellä vanhemmalla voi olla useampiakin lapsia. Esimerkiksi ihmisistä voi olla sarakkeissa tiedot kuten nimi, osoite jne. Erillisessä taulussa on rikosrekisteri, jossa on ylhäällä tiedot henkilön mahdollisista rikkeistä. Esimerkiksi IMS (Information Management System) on hierarkkinen tietokantamalli.
- Relaatiotietokantamallissa on useita eri taulukoita (Schahczenski 2000). Tarkoituksena on tallentaa tiedot siten, että yksi tieto tallennetaan aina yhteenpaikkaan. Lisäksi on aina tallennettava viite siitä, miten eri taulukot liittyvät toisiinsa. Relaatiotietokannan etuna on nopeus. Taulukoiden välille luotu yhteys nopeuttaa ja tarkentaa tietojen päivittämistä, koska päivitys on tehtävä vain yhteen paikkaan. MySQL ja Oracle ovat

relaatiotietokantoja.

Tietokanta koostuu tauluista (table), jotka ovat useimmiten kaksiulotteisia. Taulukossa yhdellä rivillä olevia tietoja, jotka liittyvät henkilöön, asiaan tai tapahtumaan nimitetään tietueeksi. Esimerkiksi yksi tietue asiakastaulussa sisältää vain yhtä asiakasta koskevan informaation kuten nimen, osoitteen ja asiakasnumeron. Informaatio on tietokannassa tallennettu sarakkeisiin eli kenttiin. Asiakastaulukossa tämä tarkoittaisi että ensimmäisessä sarakkeessa, kentässä, olisi asiakasnumero, toisessa nimi ja niin edelleen. Taulukoissa ei saa olla yhtään täysin samanlaista riviä, vaan ainakin yhden rivin on oltava jokaisessa taulukossa erilainen. Perus- eli pääavain yksilöi jokaisen taulukon rivin. Siksi pääavain ei voi olla tyhjä. Esimerkiksi asiakastaulukossa asiakasnumero voisi olla pääavain, joka yksilöi aina jokaisen asiakkaan. Viittaus eri taulukoiden välillä tapahtuu viiteavainten avulla. Halutuista sarakkeista tai niiden yhdistelmistä voidaan luoda hakemistotauluja, indeksejä, joiden avulla nopeutetaan tietokannasta hakua. Kun puhutaan tietokannan eheydestä, tarkoitetaan taulukoiden oikeellisuutta, yhdenmukaisuutta, ristiriidattomuutta ja tuoreustasoa.

Loogiseen tietokannan *transaktio*, eli tietokantatapahtuma (transaction) on tietokantaan kohdistuva loogisten tietokantaoperaatioiden sarja, jonka vaikutusten halutaan muodostavan yhden jakamattoman kokonaisuuden (Jensen et al. 2000). Tietokantasovelluksen laatija merkitsee ohjelmaan transaktioiden rajat. Transaktio alkaa aloituskirjoituksella (begin) ja päättyy sitoutumispyyntöön (commit) tai keskeytys- ja peruutuspyyntöön (abort/rollback) (Jensen et al. 2000). Transaktioilta vaadittavat ominaisuudet voidaan tiivistää ACID- sääntöön, vaikka kaikki tietokannat eivät sitä noudata saavuttaakseen maksimaalisen tehokkuuden (Jensen et al. 2000):

1) Atomisuus (Atomicity): jokainen transaktionissa suoritettava tehtävä pitää suorittaa loppuun asti tai sitä ei suoriteta ollenkaan. Eli mikäli jotain osatehtävää ei saada loppuun asti suoritettua transaktionissa, perutaan silloin koko transitionin suoritus.

2) Oikeellisuus (Consistency): transaktin suorittamisen jälkeen täytyy tietokannan olla oikeassa tilassa.

3) Eristys (Isolation): transaktioiden samanaikainen suoritus ei saa vaikuttaa toisiinsa eivätkä samaan aikaan suoritettavat transaktiot saa näkyä edes toisilleen.

4) Kestävyys (Durability): transaktioiden aiheuttamat ja tekemät muutokset pitää säilyä tietokannassa mahdollisista häiriöistä kuten tietokannan kaatumisen ja uudelleen käynnistyksen jälkeen.

Miten tietokanta liittyy tietoon? Tietokantaa voisi ajatella datakokoelmana, kuten englanninkielinen

nimi database viittaakin. Ihmiset tekevät siitä tulkintoja, jolloin syntyy informaatiota tai tietoa. Näin ollen nimitys tietokanta on harhaanjohtava ja on tapahtunut tyypillinen tiedon ja datan sekoittaminen. Henkilöllä on mielessään tietoa tai informaatiota, kun hän tekee datasta tulkintoja. Nämä tulkinnat tallennettaessa tietokantaan tallennettava informaatio tai tieto muuttuu dataksi. Tietoa tai informaatiota haettaessa tietokannasta, hakija saa ensin dataa. Vasta kun hän tekee datasta tulkintansa, syntyy tallennetusta datasta informaatiota tai tietoa. Tietokannassa itsessään ei siis ole informaatiota tai tietoa, vaan pelkkää dataa, oli tietokanta sitten sähköisessä tai sähköttömässä muodossa.

3.5.3 Tiedosto

WSOY:n suuressa tietosanakirjassa (Halinen et al, s. 822) *tiedosto* (file) määritellään ”joukoksi tietokonemerkkejä, jotka voidaan tulkita joko tietokoneohjelmina, kirjoitusmerkkeinä, numeroina, kuvina tai musiikkina”. Toisin sanoen tiedosto pitää sisällään dataa, jota tiedoston tekijä on halunnut talletettavan. Tiedosto onkin tietokonejärjestelmissä massamuisteille tallennettu datayksikkö. Viittaaminen kyseiseen datayksikköön onnistuu nimen avulla. Jokaisen tiedoston on oltava nimetty jotta viittaaminen olisi mahdollista. Tiedoston nimi onkin metatietoa, joka kertoo tiedostossa olevasta datasta. Esimerkiksi tiedostoon voidaan tallentaa nimen yhteyteen tekijä, päivämäärä ja muokkaaja. Myös tiedoston nimen lopussa oleva pääte on metadataa, joka kertoo millä ohjelmalla tiedostoa pitää käsitellä tai mitä tiedosto mahdollisesti pitää sisällään. Esimerkiksi txt tiedostot ovat tavallisia tekstitiedostoja ja doc päätteiset tiedostot ovat Microsoft Word:llä käsiteltäviä asiakirjoja.

Tiedostossa ei kuitenkaan ole tietoa tai informaatiota. Ihminen tulkitsee tiedoston sisällä olevan datan, jolloin siitä tulee joko tietoa tai informaatiota. Henkilön tallettaessa tietoa tai informaatiota ne muuttuvat vastaavasti dataksi. Aina ei kuitenkaan tallenneta tietoa tai informaatiota tiedostoon. Jotkut ohjelmat tekevät suorituksen yhteydessä omia varmuuskopioita käyttäjän tietämättä. Tällöin tallennetaan dataa, koska käyttäjä ei ole sitä tulkinnut informaatioksi tiedostoon.

Miten tiedostot ovat olemassa? Näemme niiden sisällön näytöltä, mutta itse tiedostoa emme voi nähdä. Tiedostoja hallitaan esimerkiksi Windows:in resurssihallinnan avulla, mutta siitäkin näemme vain missä tiedostot sijaitsevat ja minkä nimisiä ne ovat. Tiedostoihin on mahdotonta päästä konkreettisesti käsiksi. Tiedostoja ei voi esimerkiksi käsin kosketella. Tiedostojen voidaan ajatella olevan olemassa sähköimpulsseina tai magneettisina tiloina, jotka tietokone tulkitsee. Tietokoneen tulkinta ilmenee meille datana, jonka me määrittelemme tiedoksi tai informaatioksi.

4. Yhteenveto

Tieto-termi on huomattavasti monimutkaisempi kuin ensinäkemältä vaikuttaa. Pidämme itsestään selvyytenä mitä kyseisellä termillä tarkoitetaan, jolloin monet sen määritelmistä sekoittuvat keskenään. Tämä on yksi syy sille, että tieto-termi on sekoitettu niin informaatioon kuin dataankin.

Tieto on jaettu kolmeen eri lajiin. Ensimmäiseen lajiin kuuluvat erilaiset tuntemiset, niin sanottu tuttuustieto. Tunnen esimerkiksi naapurini Jaanan. Tuttuustietoa ei tutkielmassa sen tarkemmin käydä läpi.

Taitotiedolla tarkoitetaan, että osataan tehdä jokin asia. Esimerkiksi osaan juosta ja puhua suomen kieltä. Vaikka osaa tehdä jonkin asian se ei kuitenkaan välttämättä tarkoita, että osaa selittää sanallisesti taitoaan. Usein taitotieto perustuu yritykseen ja erehdykseen. Esimerkiksi pieni lapsi oppii kielen ja kieliopin ilman, että kukaan opettaa hänelle sanojen taivutusta ja muita kielioppisääntöjä. Tällöin lapsella on piilevää tietoa, jonka mukaan hän toimii. Piilevä tieto onkin siinä mielessä vastakohta propositionaaliselle tiedolle, että piilevää tietoa ei ilmaista väitelauseilla, kuten propositionaalista tietoa. Piilevä tieto ilmenee myös urheilussa, käsitöissä ja taiteissa.

Tutkielman kannalta mielenkiintoisin tiedon laji on kielellä ilmaistu propositionaalinen tieto. Kieli voi olla niin kirjoitettua kieltä kuin indeksejä tai symboleitakin. Tärkeintä on, että kieli on tulkittavissa maailmassa vallitsevia asiantiloja koskeviksi väitteiksi. Luonnollinen kieli ei automaattisesti kuvaa maailmassa vallitsevia asiantiloja, vaan erilaiset huudahdukset, pyynnöt ja kiitokset pitää sulkea tiedon ulkopuolelle. Propositionaalisen tiedon väittämät voivat koskea:

1) Aineellista maailmaa.
2) Mentaalista maailmaa.
3) Mielen tuotteiden maailmaa.

Perinteisen tiedon määritelmän mukaan tieto on tosi, perusteltu uskomus. Tosi-kriteerillä tarkoitetaan, että väitteen on vastattava maailmassa vallitsevaa asiantilaa. Väitteen esittäjällä pitää olla myös perustelut väitteelleen. Samoin on väittäjän uskottava omaan väitteeseensä, minkä vuoksi tiedolta vaaditaan sen olevan uskomus. Tämä niin sanottu tiedon klassinen määritelmä syntyi jo antiikin ajoilla. Sen puutteet huomasi vasta Gettier vuonna 1963 jolloin hän osoitti, että klassisen tiedon määritelmä on riittämätön. Aikaisemman kolmen ehdon rinnalle on löydettävä neljäs ehto. Neljännestä ehdosta ei kuitenkaan vielä päästy yksimieliseen ratkaisuun filosofien keskuudessa.

Tiedon alkuperä on kautta historian askarruttanut filosofeja. Vastakkain ovat olleet

rationalistinen ja empiristinen näkemys. Rationalistien mukaan järki on vähintäänkin etusijalla tietoa hankittaessa. Empiristisen näkemysten mukaan kokemuksella saatu tieto on oikeaa ja aitoa tietoa. Kokemus tarkoittaa empiristeille aistihavaintoja. Nykyään erot näiden kahden ajatusmallin välillä ovat kaventuneet, eivätkä ne ole enää toisiansa poissulkevia.

Tiede luottaa empiristiseen näkemykseen tiedon tuottamisessa. Tieteessä tarkastellaan pääsääntöisesti yksittäisiä havaintoja joiden perusteella tehdään päätelmät eli maailmaa koskevat tosiasiaväittämät. Päättely on induktiivista, millä tarkoitetaan, että yksittäisistä tapauksista tehdään yleiset säännöt. Induktion ongelmana on ettei milloinkaan voida tietää, koska on tutkittu tarpeeksi yksittäisiä tapauksia yleisen säännön laatimiseksi. Esimerkiksi riittääkö tuhannen mustan korpin havainnointi yleistämiseen, jonka mukaan kaikki korpit ovat mustia? Entä jos tuhannes ensimmäinen korppi onkin valkoinen? Tieteelliselle tiedolle onkin tyypillistä, että se korjaa itseään uusimpien tutkimustuloksien mukaiseksi. Tämä ei ole kuitenkaan hyväksyttävää, mikäli tiedolta vaaditaan ehdotonta totuutta. Tieteellistä tietoa tutkii tarkemmin tieteenfilosofia. Se tutkii muun muassa tieteellisen tiedon luonnetta, yleisiä metodologisia ongelmia sekä teorioiden ja todellisuuden suhdetta.

Tieteellisen tiedon ongelmana on pidetty luotettavuutta. Miten tieteellinen tieto voi olla tietoa sen klassisessa määritelmässä, jollei tieteellinen tieto ole totta? Tieteellisen tiedon totuutta on pyritty määrittelemään erilaisilla totuusteorioilla. Korrespondenssiteorian mukaan väite on totta, mikäli se vastaa maailmaa. Todellisuutta kuvaavien väitelauseiden totuusarvot määräytyvät sen mukaan millainen maailma todellisuudessa on, vaikkei totuusarvoja pystyttäisikään ratkaisemaan. Totuus ei ole riippuvainen ihmisen tietämyksestä. Totuuden koherenssiteorian mukaan väitteen totuus on riippuvainen sen yhteensopivuudesta aikaisempien väitteiden kanssa. Mikäli uusi väite on ristiriidassa vanhojen väittämien kanssa, on todennäköistä ettei uusi väite ole totta. Koherenssiteorian ongelma ilmenee, kun vanhat väittämät ovatkin epätotta. Näin uusi väittämä voi turhaan tulla tuomituksi epätodeksi, koska se ei ole yhteensopiva aikaisempien epätosien väittämien kanssa. Pragmaattisen totuusteorian mukaan totuus perustuu toiminnan ja tiedon väliseen suhteeseen. Totuus määritellään käyttökelpoisuudeksi, eli väite on tosi jos se mahdollistaa menestyksellisen toiminnan. Pragmaattista totuusteoriaa on kritisoitu siitä, että myös epätoden teorian mukaan on mahdollista toimia menestyksellisesti.

Popper luopui ajatuksesta, että tieteellinen väite olisi todistettavissa todeksi. Sen sijaan hän esittikin tieteelliselle tiedolle vaatimuksen, että se on voitava edes teoriassa osoittaa epätodeksi. Mikäli väitteen ei voida edes teoriassa osoittaa olevan epätotta, ei se myöskään ole tieteellistä tietoa. Popperin teoriaa kutsutaan falsifikationismiksi. Teoriallaan Popper ei halunnut tukea skeptisismiä,

vaan pyrki altistamaan tieteelliset teoriat kritiikille. Kun teorioita kritisoidaan, voidaan niitä myös korjata kritiikin mukaisesti. Näin tieteellinen tieto saavuttaisi kokoajan totuuden raja-arvoa.

Tiede ja sen menetelmät ovat kehittyneet aikojen kuluessa. Kuhn esitti tieteen kehittyvän vallankumousten avulla. Tiedettä hallitsee paradigma, joka määrittelee muun muassa mitä kohteita tutkitaan, tutkimusaiheeseen liittyviä kysymyksiä, miten kysymykset esitetään ja miten saatuja tutkimustuloksia tulkitaan. Kuhnin mukaan uudessa tieteenalassa vallitsee aluksi kaaos, kunnes yksi paradigma nousee muiden yläpuolelle ja tieteenala kehittyy tällöin normaalitieteeksi. Normaalitieteessä hallitseva paradigma joutuu taistelemaan erilaisia anomalioita vastaan. Kuhn toteaa kuitenkin, että ennemmin tai myöhemmin anomalioita on niin paljon ettei vallitseva paradigma pysty niitä enää selittämään tai tukahduttamaan. Tällöin alkaa vallankumous, mikä johtaa tieteen kriisiin, jossa vastakkain ovat uusi ja vanha paradigma. Lopulta vanha paradigma on syrjäytetty uuden paradigman tieltä ja kierros alkaa taas alusta.

Informaatio eroaa tiedosta siinä, ettei informaation tarvitse olla totta. Tiedon tapaan informaatiokin on tulkinta datasta. Informaatio jaetaan kahteen eri lajiin, fysikaaliseen ja kielelliseen informaatioon. Fysikaalista informaatiota ei ilmaista kielellisesti, vaan fysikaalinen informaatio kuvaa esimerkiksi aineen järjestäytyneisyyttä ja organisaatiotasoa tai monimuotoisuutta. Kielellinen informaatio vaatii kielen välittäjäkseen. Kielellinen informaatio jaetaan vielä kolmeen eri osaan:

1) Semanttinen informaatio. Lauseen informatiivisuus määritellään semanttisessa informaatiossa sen mukaan miten hyvin asiantiloja saadaan maailmasta suljettua. Mitä enemmän lause sulkee maailmassa vallitsevia asiantiloja, sitä informatiivisempi se on. Semanttinen informaatio ei kuitenkaan ota tiedon lailla kantaa lauseen totuuteen kuten tieto tekee. Siten epätosi väite voi olla semanttisessa mielessä kuitenkin hyvinkin informatiivinen.
2) Syntaktinen informaatio. Syntaktisessa informaatiossa tehdään ero merkeistä muodostetun viestin ja kanavassa välitetyn informaation välillä. Välitetyn viestin merkityksellä ei ole väliä syntaktisessa informaatiossa, vaan tärkeintä on välittää viesti mahdollisimman muuttumattomana. Esimerkiksi kohinan vaikutuksen minimoiminen on syntaktisen informaation kannalta tärkeää.
3) Pragmaattinen informaatio. Pragmaattinen informaatio tarkoittaa henkilö-ja kulttuurisidonnaista merkityksellisyyttä tai merkittävyyttä. Keskisormen näyttäminen on esimerkki pragmaattisesta informaatiosta.

Tiedon ja informaation suhde on mielenkiintoinen. Mikäli ei voida olettaa yhdenkään väitteen täyttävän tiedon vaatimuksia, tulee kaikesta pitämästämme tiedosta informaatiota. Toisaalta,

informaatiohan voi olla totta tai epätotta, joten pitäisikö silti väitteen joka täyttää tiedon vaatimukset ajatella olevan informaation erikoistapaus.

Kun informaatio tai tieto tallennetaan esimerkiksi tietokoneella tai johonkin kirjaan syntyy siitä dataa. Data on tietoa tai informaatiota mille emme ole vielä antaneet merkitystä. Annamme kirjassa esiintyneelle datalle eli kirjaimille merkityksen vasta kirjaa lukiessamme. Dataa yksinään ei kuitenkaan voi pitää informaationa tai tietona, vaan se vaatii aina tulkitsijan, joka synnyttää datasta informaation tai tiedon.

Suomenkielisessä tietojenkäsittelytieteessä käsitteet data, tieto ja informaatio sekoittuvat keskenään. Esimerkiksi database käännetään tietokannaksi. Tarkempi tarkastelu kuitenkin osoittaa, että tiedosta puhuminen tietojenkäsittelytieteessä on harvoin mielekästä. Tutkielmassa tarkasteltiin tietojenkäsittelytieteen tuottamaa tietoa. Tarkastelun kohteena olivat tiedonlouhinta ja asiantuntijajärjestelmät. Tutkielmassa käytiin tietojenkäsittelytieteen käsittelemää tietoa läpi tarkastelemalla tarkemmin tiedon pakkausta. Tietojenkäsittelytieteen ja tiedon suhdetta tarkasteltiin myös Platonin luolavertausta apuna käyttäen.

Teknologian kehityksen myötä ovat tietokantojen koot kasvaneet siinä määrin, ettei mielekästä dataa ole aina helppo löytää. Kun haluttua dataa ruvetaan seulomaan tietokannoista kutsutaan prosessia tiedonlouhinnaksi. Tiedonlouhinnan tuloksena saadaan hahmoja ja malleja. Mallit ovat globaaleja tiivistelmiä tietokannoista. Hahmoilla tarkoitetaan väitettä tai sääntöä, jolla kuvataan rajoitettua joukkoa datasta. Tiedonlouhinnan kohteet jaetaan seuraavaan kolmeen osaan:

1) Tekstitiedonlouhinta. Käsittelee muun muassa dokumenttien luokittelua ja klusterointia.
2) Web-louhinta. Käsittelee HTML-dokumenttien, linkkirakenteiden, web-sivustojen ja aihehakemistojen analyysia.
3) Relaatiotiedonlouhinta. Käsittelee mallien etsimistä tietojoukoista, jotka liittyvät toisiinsa erilaisilla suhteilla.

Tiedonlouhinta on osa tietämyksen muodostamista tietokannoista. Näitä kahta on myös pidetty synonyymeina, mutta tiedonlouhinta on käsitteenä laajempi kuin tietämyksen muodostaminen tietokannoista. Tämän takia tietämyksen muodostamista tietokannoista käytetään usein tietokannoissa olevan tiedon analysoimissa, kun tiedonlouhinta ei ole riippuvainen syötetiedon formaatista. Tiedonlouhinnan menetelmät jaetaan kuvaileviin ja ennustaviin. Tarkemman jaottelun mukaan tiedonlouhinta jaotellaan seuraavasti:

1) Klusteroinnissa jaotellaan alkioita niiden keskinäisen samanlaisuuden perusteella.

2) Luokittelussa tietoalkio sijoitetaan ennalta määritettyihin luokkiin. Regressiomalli on yleistys luokittelumallista ja siinä on korvattu luokat numeroilla.

3) Assosiaatioiden ja peräkkäisten toimintojen etsintä on malli toistuvasti yhdessä esiintyville tietueille.

Ennustavat menetelmät tuottavat ennusteita, jotka voivat olla epätotta eivätkä näin ollen ole tietoa vaan informaatiota. Kuvailevissa menetelmissä dataa luokiteltiin ominaisuuksien mukaan. Mikäli luokittelu onnistuu joka kerta oikein, on mahdollista tuottaa tietoa kuvailevien menetelmien avulla.

Asiantuntijajärjestelmät ovat järjestelmiä joihin on varastoituna jonkin kapean ongelma-alueen inhimillinen tietämys. Asiantuntijajärjestelmät eivät pyri pelkästään matkimaan asiantuntijoita, vaan niiden on myös kyettävä perustelemaan päätöksensä. Asiantuntijajärjestelmät ovat erikoistapauksia päätöstukijärjestelmistä, joilla on tarkoitus helpottaa päätöksentekoa. Tutkielmassa käsiteltiin seuraavia asiantuntijajärjestelmiä:

1) Bayesin verkot ovat täydellisiä malleja muuttujista ja niiden välillä vallitsevista suhteista ja todennäköisyyksistä. Verkkoa hyväksi käyttämällä saadaan päivitettyä tietoa jostain verkon osajoukosta, kun verkon muita muuttujia tutkitaan.

2) Neuroverkkoja käytetään erityisesti tekoälyn yhteydessä jäljittelemään inhimillistä toimintaa. Neuroverkkojen käyttö jaotellaan seuraavasti: funktioiden arviointi ja regressioanalyysi sekä datan luokittelu ja käsittely.

3) Päätöspuissa on kuvattuna tai mallinnettuna eri päätökset ja niiden seuraukset. Päätöspuita käytetään päämäärään vaadittavan strategian etsimiseen sekä todennäköisyyksien laskemiseen. Niiden avulla muodostetaan yleisiä luokittelusääntöjä sovellusalueen esimerkkijoukoista. Kun päätöspuusta kuvataan vain välttämättömät elementit, päätökset, epävarmuudet sekä päämäärät ja niiden vaikutukset, puhutaan vaikutuskaaviosta. Vaikutuskaaviot ovat tiivistettyjä päätöspuita.

Asiantuntijajärjestelmät auttavat pääsääntöisesti päätöksenteossa erilaisten mallien, tilastoiden ja ennusteiden avulla. Tietona ei näitä ennusteita, malleja, tai tilastoja voida pitää . Parhaimmillaan ne voivat olla totta, mutta tällöinkin niitä tulisi ajatella vain informaationa.

Tiedonkäsittelystä tietojenkäsittelytieteessä on esimerkkinä tiedonpakkaus. Se jaetaan kahteen eri osaan: häviölliseen ja häviöttömään pakkaukseen. Häviöllisessä tiedonpakkauksessa hävitetään tahallaan ihmiselle merkityksetöntä dataa. Esimerkiksi musiikista hävitetään ihmiskorvalle kuulumattomat äänet. Tieto ei kuitenkaan ole ihmisestä riippuvaista, joten

häviöllisessä pakkauksessa on mahdollista hävittää sellaista dataa ettei sitä enää voida tiedoksi tulkita. Häviöttömässä pakkauksessa tietoa pakataan siten, ettei mitään häviä. Näin ollen pakattu data säilyttää merkityksensä, jolloin se on tulkittavissa samoin kuin ennen pakkausta.

Platonin luolavertauksessa joukko vankeja on kahlehdittuna luolaan. Näiden vankien todellisuus muodostuu luolan seinille heijastuneista varjokuvista sekä luolan vääristämistä äänistä. Ihmisiä voidaan ajatella tällaisina vankeina. Todellisuus muodostuu tieteen antamista tuloksista, jotka ovat vastaavasti hyvin usein riippuvaisia tietokoneista. Tekemillään ohjelmillaan tietojenkäsittelytieteilijät määräävät sen mitä ihmiset näkevät. Toisaalta myös tietojenkäsittelytieteilijät ovat algoritmiensa ja tietorakenteidensa vankeja.

Pääsääntöisesti suomenkielisessä tietojenkäsittelytieteessä termit data, tieto ja informaatio sekoittuvat keskenään. Tämän vuoksi olisi hyvä suorittaa perusteellinen käsiteanalyysi ja pohtia tarkemmin käsitteiden merkitystä. Samalla olisi syytä miettiä tiedon määritelmää ja verrata sitä tietojenkäsittelytieteen eri vaiheisiin kuten tiedon tuottamiseen ja käsittelemiseen. Pohdinnalla voisi hyvinkin saavuttaa lisää luotettavuutta saatuihin tieteen tuloksiin, jolloin tieteellisen tiedon ja tiedon ero kaventuisi.

5. Lähteet

Aristoteles (1989): *Nikomakhoksen etiikka*, Gaudeamus, Helsinki.

L. M. de Campos (2006): A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests, *The Journal of Machine Learning Research*, 2149 -2187.

S. Chakrabarti (2003): *Mining the Web - Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.

E. Duval, W. Hodgins, S. Sutton, S. Weibel (2002): Metadata Principles and Practicalities, *D-lib Magazanine april 2002 volume 8 number 4*, http://www.dlib.org/dlib/april02/weibel/04weibel.html (27.3.2008).

J. Dörre and P. Gerstl and R. Seiffert: Text mining: finding nuggets in mountains of textual data. In U. Fayyad, S. Chaudhuri, and D. Madigan (eds.) (1999): *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and datamining*, 398–401, http://doi.acm.org/10.1145/312129 (27.3.2008).

S. Džeroski (2003): Multi-relational data mining: An introduction. *ACM SIGKDD Newsletter, 5(1),* 1–16*,* http://www.acm.org/sigs/sigkdd/explorations/issue5-1/Dzeroski.pdf (27.3.2008).

M. Eckert (2006): *Theories of Mind: An Introductory Reader*, Rowman & Littlefield.

S. Gregor (2006): *The Nature of Theory in Information Systems*, research essay, Australian National University, Canberra.

O. Etzioni (1996): The world-wide web: quagmire or gold mine?, *Commun. ACM 39(11)*, 65–68, http://doi.acm.org/10.1145/240455.240473 (27.3.2008).

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (1996): From data mining to knowledge discovery in databases. *AI Magazine 17*, 37–54, http://citeseer.ist.psu.edu/fayyad96from.html (27.3.2008).

A. Flew (1979): *A Dictionary of Philosophy*, MacMillan, London.

W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus (1992): Knowledge discovery in databases: An overview. *AI Magazine 13(3)*, 57–70, http://citeseer.ist.psu.edu/frawley92knowledge.html (27.3.2008).

L. Fu (1999): Knowledge discovery based on neural networks**,** *Communications of the ACM archive Volume 42 Issue 11 (November 1999)*, 47-50.

R. Geneshan, G. B. Kleindorfear (1993): The philosophy of science and validation in simulation, *Winter Simulation Conference Proceedings of the 25th conference on Winter simulation*,

50-57.

A. Goldenberg, G. W. Moore (2005): Bayes net graphs to understand co-authorship networks?, *Conference on Knowledge Discovery in Data archive Proceedings of the 3rd international workshop on Link discovery*, 1-8.

T. Guesmi, H. Rezig (2006): Design and implementation of a real-time notification service within the context of embedded ORB and the CAN bus, *Symposium on Applied Computing archive Proceedings of the 2006 ACM symposium on Applied computing*, 773-777.

A. Halinen, L. Honkala, I. Hetemäki, J. Rossi, S. Salin (2001): *Suuri tietosanakirja*, Werner Söderström Osakeyhtiö, Jyväskylä.

J. Hallamaa, S. Pihlström, U. Pulliainen, E. Salmenkivi, J. Sihvola (2002): *Tiedon Odysseia*, Edita, Helsinki.

J. Han and M. Kamber (2000): *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

D. Hand, H. Mannila, and P. Smyth (2001): *Principles of Data Mining*, MIT Press.

T. Harris (2001): *How file compression works*, WWW-sivusto, http://www.howstuffworks.com/file-compression.htm (27.3.2008).

M. Harva, J. Karhunen, T. Raiko, H. Valpola (2007): Building Blocks for Variational Bayesian Learning of Latent Variable Models, *The Journal of Machine Learning Research*, 155-201.

J. Heinonen (2007): *Semanttinen tiedonhaku*, Pro gradu -tutkielma, Tietojenkäsittelytieteen laitos Helsingin yliopisto.

J. Hertell (2005): *Äänen streamaus internet -radio Vastavirta ry:lle*, viestinnän koulutusohjelman tutkintotyö, Tampereen ammattikorkeakoulu.

N. Houser, J. W. Kloesel, S. P. Peirce (1998): *The Essential Peirce: Selected Philosophical Writings*, Indiana University Press.

R. L. Huard (1996): *Plato's Political Philosophy: The Cave*, Algora Publishing.

A. M. Huberman, M. B. Miles (2002): *Qualitative research*, Sage.

D. Hume (1739): *A Treatise of Human Nature*, eBooks@Adelaide, http://etext.library.adelaide.edu.au/h/hume/david/h92t/ (27.3.2008).

R. A. Howard, J. E. Matheson (2005): Influence diagrams, *Decision Analysis vol. 2, No3*, 127-143.

N. F. Ikeda, K. Konishi (2007): Data model and architecture of a paper-digital document management system, *Document Engineering Proceedings of the 2007 ACM symposium on Document engineering table of contents SESSION: Paper documents: capture and physical-digital-coexitence*, 29-31.

S. Jensen, R. T. Snodgrass, K. Torp (2000): Effective timestamping in databases, *The VLDB J ournal — The International Journal on Very Large Data Bases Volume 8 , Issue 3-4*

(February 2000), 267-288.

A. Kasher (1998): P*ragmatics: Critical Concepts*, Routledge.

D. C. Knill, A. Pouget (2004): The Bayesian brain: the role ofuncertainty in neural coding and computation, *Trends in Neurosciences Vol.27 No.12 December*, 712-719.

Y. Kodratoff (1999): Knowledge discovery in texts: A definition and applications. In Z. W. Ras and A. Skowron (eds.): *Foundations of Intelligent Systems, 11th International Symposium, ISMIS '99 Proceedings vol. 1609 of Lecture Notes in Computer Science*, 16–29, http://citeseer.ist.psu.edu/kodratoff99knowledge.html (27.3.2008).

R. Kosala, H. Blockeel (2000): Web mining research: a survey, *SIGKDD Explor Newsl 2(1)*, 1–15, http://doi.acm.org/10.1145/360402.360406 (27.3.2008).

J. Kotkavirta (1999): *Tietoteoria*, WSOY, Porvoo.

A. V. Kozlov, J. P. Singh (1994): A parallel Lauritzen-Spiegelhalter algorithm for probabilistic inference, *Conference on High Performance Networking and Computing Proceedings of the 1994 ACM/IEEE conference on Supercomputing*, 320-329.

T. S. Kuhn (1994): *Tieteellisten vallankumousten rakenne*, Art House, Helsinki.

T.S. Kuhn (1996): The Structure of Scientific Revolutions, The university of Chicago Press, Chicago.

M. Lammenrant(1993): *Tietoteoria*, Gaudeamus, Tampere.

H. Laukkanen (2007): *Mom-palvelujen tuottaminen*, opinnäytetyö, Tampereen ammattikorkeakoulu, Tampere.

J. Laurikkala(2001): *Knowledge Discovery for Female Urinary Incontinence Expert System*, Väitöskirja, Tampereen yliopisto, Tampere.

S. Loh, L. K. Wives, J. P. M. Oliveira (2000): Concept-based knowledge discovery in texts extracted from the Web, *ACM SIGKDD Explorations Newsletter Volume 2 Issue 1 (June 200*0), 29-39.

J. C. A. Van der Lubbe (1997): *Information theory*, Cambridge University Press.

C. Manasis, P. Trakadas, S. Voliotis, Th. Zahariadis (2004): Efficient routing in PAN and sensor networks, *ACM SIGMOBILE Mobile Computing and Communications Review archive Volume 8 , Issue 1 (January 2004)*, 10-17.

R. Miikkulainen (2007): Evolving neural networks, *Genetic And Evolutionary Computation Conference archive Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation*, 3415 – 3434.

M. Mäkipää, M. Ruohonen (2004): *Organizational Learning and Knowledge Management in Contexts*, verkkojulkaisu, Tampereen yliopisto,

http://newwww.cs.uta.fi/is/julkaisut/2004/2004_Makipaa_Ruohonen.pdf (27.3.2008).

P. Määttänen (1995): *Filosofia, johdatus peruskysymyksiin*, Gummerus kirjapaino Oy, Jyväskylä.

D. Niedermayer (1998): *An Introduction to Bayesian Networks and their Contemporary Applications*, WWW-sivusto, http://www.niedermayer.ca/papers/bayesian/ (27.3.2008) .

P. Niemenlehto (2004): *Tahdonalaisen lihasaktiviteetin havaitseminen EMG-signaalista neuroverkon avulla*, Pro Gradu -tutkielma, Tampereen Yliopisto, Tampere.

I. Niiniluoto (1980), *Johdatus Tieteenfilosofiaan*, Otava, Helsinki.

I. Niiniluoto (1996): *Informaatio, tieto ja yhteiskunta, filosofinen käsiteanalyysi*, Valtion painatuskeskus ja Valtionhallinnon kehittämiskeskus, Helsinki.

M. Nurminen (2005): *Tiedonlouhinta rakenteisista dokumenteista*, Pro gradu tutkielma, Jyväskylän yliopisto.

C. O'Brien, C. Vogel (2003): Spam filters: bayes vs. chi-squared; letters vs. Words**,** *ACM International Conference Proceeding Series; Vol. 49 archiveProceedings of the 1st international symposium on Information and communication technologies*, 291- 296.

E. Paakkola, K. E. Turunen (1995): *Miten ihminen tietää, johdatus tiedon ja tieteen käsitteisiin*, Gummerus Kirjapaino Oy, Saarijärvi.

K. Popper (1992): *In Search of a Better World: Lectures and Essays from Thirty Years*, Routledge.

J. R. Quinlan (1986): Induction of decision trees, *Machine learning 1*, 81-106 .

D. Runes (1983): *Dictionary of Philosophy*, Philosophical Library, New York.

S. Ryan (2007): *Wisdom*, Stanford Encyclopedia of Philosophy, WWW-sivusto, http://plato.stanford.edu/entries/wisdom/ (27.3.2007).

J. Räikkä (1991): *Filosofian ongelmia*, Turun yliopiston offsetpaino, Turku.

E. Saarinen (1985): *Länsimäisen filosofian historia, huipulta huipulle Sokrateesta Marxiin*, WSOY, Juva.

E. Saarinen (1996): *Filosofia*, WSOY, Porvoo.

A. Salminen (2005): Metatiedot organisaatioiden sisällönhallinnassa, Metatiedot suomalaisen lainsäädäntöprosessin tiedonhallinnassa. RASKE2 -projektin II väliraportti, *Eduskunnan kanslian julkaisu 7/2005*, 4-13.

C. Schahczenski (2000): Object-oriented databases in our curricula, *Journal of Computing Sciences in Colleges Volume 16 Issue 1 (October 2000)*, 170-176.

C. E. Shannon, W. Weaver (1963): *The Mathematical Theory of Communication*, University of Illinois Press.

M. Steup (2006): *The Analysis of Knowledge*, Stanford Encyclopedia of Philosophy, WWW-sivusto, http://plato.stanford.edu/entries/knowledge-analysis/ (27.3.2008).

D. J. Stracuzzi, P. E. Utgoff (2004): *Randomized Variable Elimination*, *The Journal of Machine Learning Research*, 1331 – 1362.

H. Sundström (2008): *Roskapostin estäminen*, Pro gradu -tutkielma, Tampereen yliopisto, Tampere.

M. Tedre (2007): *Lecture Notes in The Philosophy of Computer Science*, Joensuun yliopisto, Joensuu.

K. Varpa (2005): *Tietämysjärjestelmien tietämyksen esittäminen ja hankinta sekä huimaustautien päätöstukijärjestelmän ja sen tietämyksen uudistaminen*, Pro gradu -tutkielma, Tampereen yliopisto, Tampere.

W. Weaver (1949): *Recent Contributions to The Mathematical Theory of Communication*. WWW-sivusto, http://grace.evergreen.edu/~arunc/texts/cybernetics/weaver.pdf (27.3.2008).

N. Wiener (1961): *Cybernetics or Control and Communication in the Animal and the Machine*, MIT PRESS.

Wikipedia.org (2006): *daughters of albion*, WWW-sivusto, http://en.wikipedia.org/wiki/Image:Blake_Daughters_of_Albion_2.jpg (27.3.2008) .

Wikipedia.org (2007): *Dtree2,* WWW-sivusto, http://en.wikipedia.org/wiki/Image:Dtree2.png (27.3.2008).

Wikipedia.org (2006): *Factory2 Influence diagram*, WWW-sivusto, http://en.wikipedia.org/wiki/Image:Factory2_InfluenceDiagram.png (27.3.2008).

Wikipedia.org (2007): *Neural network example*, WWW-sivusto, http://en.wikipedia.org/wiki/Image:Neural_network_example.png (27.3.2008).

Wikipedia.org (2006): *SimpleBayesNet*, http://en.wikipedia.org/wiki/Image:SimpleBayesNet.svg (27.3.2008).

Xiph (2003): vorbis.com, WWW-sivusto, http://www.vorbis.com/faq/#what (27.3.2008).

T. Yao, L. Zhang, J. Zhu (2004): An evaluation of statistical spam filtering techniques, *ACM Transactions on Asian Language Information Processing (TALIP) Volume 3 Issue 4 (December 2004)*, 243-269.

M. Yrjönsuuri (1996): *Tiedon rajat johdatus tietoteoriaan*, Gummerus Kirjapaino, Jyväskylä.