Informaatioteknologian tiedekunnan julkaisuja
No. 17/2014

Martti Lehto, Pekka Neittaanmäki

Big datan tutkimus ja opetus Jyväskylän yliopistossa

JYVÄSKYLÄN YLIOPISTO

Informaatioteknologian tiedekunnan julkaisuja
No. 17/2014

Editor: Pekka Neittaanmäki
Covers: Kati Valpe

ISBN 978-951-39-6046-9 (verkkoj.)
ISSN 2323-5004

Jyväskylän yliopistopaino, Jyväskylä 2014

Big datan tutkimus ja opetus Jyväskylän yliopistossa

Martti Lehto, Pekka Neittaanmäki

Johdanto

Tässä raportissa käsitellään Jyväskylän yliopistossa annettavaa big datan ja data-analyysin tutkimusta ja koulutusta. Jyväskylän yliopisto haluaa olla laaja-alaisesti mukana big datan tutkimuksessa ja koulutuksessa sekä alan kehittämisessä.

Liikenne- ja viestintäministeriön big datan käyttö -työryhmän raportin mukaan ”koulutuspuutteet voivat muodostua merkittäväksi esteeksi big datan laajemmalle hyödyntämiselle. Big datan murrokseen on kyettävä vastaamaan nopeilla ja nykyistä kohdennetummilla koulutustoimilla. Big data edustaa uutta hyppäystä datatutkimuksen ja -soveltamisen alalla, jollaa perinteiset koulutusohjelmat eivät vastaa enää ammatillisiin tarpeisiin. Eri toimialojen omat sovellukset edellyttävät soveltavaa osaamista myös muista aloista (esim. kaupan, teollisuuden tai oppimiseen liittyvä osaaminen).

Data-analyysin asiantuntijan tulee kyetä vaativaan tilastolliseen mallintamiseen ja osata ohjelmointia ja tiedonhallintaa. Näiden taitojen oppiminen puolestaan vaatii pohjakseen matematiikan osaamista. Alan vaativuuden vuoksi tutkijakoulutus on usein tarpeen käytännön työtehtävissä. Ohjelmistojen tekninen hallitseminen ei riitä lisäarvon tuottamiseen, vaan big datan hyödyntäminen vaatii sekä substanssialueen että analyysimenetelmien syvällistä ymmärtämistä. Vain näin voidaan varmistaa prosessien ja tulosten luotettavuus ja käyttökelpoisuus pitkällä aikavälillä.

Data-analyysin ja big datan maisteri- ja tohtorikoulutus pohjautuvat Jyväskylän yliopistossa vankkaan matematiikan, tietotekniikan ja tilastotieteen tutkimukseen. Jyväskylän yliopiston monitieteinen toimintaympäristö antaa erinomaisen lähtökohdan kehittää uusia data-analyysin ja big datan menetelmiä ja soveltaa niitä eri tieteen aloilla sekä yritysmaailman että julkisen sektorin osa-alueilla.

Tässä raportissa kuvataan Jyväskylän yliopiston Centre of Simulation and Data-analysis toimintaympäristöä.

SISÄLLYS

1 BIG DATA JYVÄSKYLÄN YLIOPISTON TUTKIMUKSESSA 3
1.1 Tieteellinen laskenta ja data-analyysi 4
1.2 Big data ja kyberturvallisuus 5
1.3 Big data ja SOTE 6
1.3.1 Sairaalasuunnittelu 6
1.3.2 Sosiaali- ja terveydenhuollon prosessit 6
1.4 Big data tilastotieteessä 6
1.5 Big data bio- ja ympäristötieteissä 7
1.6 Big data fysiikassa 7

2 BIG DATA JYVÄSKYLÄN YLIOPISTON KOULUTUKSESSA 9
2.1 Sovelletun matematiikan maisteriohjelma 9
2.2 Laskennalliset tieteet -maisteriohjelma 10
2.3 Web Intelligence and Service Engineering (WISE) -maisteriohjelma 10
2.4 Tilastotieteen koulutus 10
2.5 Ihmistieteiden metodikeskus (IHME) 11
2.6 Big dataan liittyvä opetus lukuvuonna 2014–2015 11

3 CENTRE OF SIMULATION AND DATA-ANALYSIS 12

4 BIG DATA -ALAN KEHITYKSEN SEURANTA 15
4.1 Data-analyysiin liittyvät väitöskirjat 15

LIITE 1: JYVÄSKYLÄN YLIOPISTON IT-TIEDEKUNNAN OPETUSOHJELMAAN LIITTYVÄT VERKKOKURSSIT 16

LIITE 2: YHTEENVETO BIG DATA -ALASTA JA SOVELLUTUKSISTA JYVÄSKYLÄN YLIOPISTON IT-TIEDEKUNNASSA 20

1 BIG DATA JYVÄSKYLÄN YLIOPISTON TUTKIMUKSESSA

Data-analyysin asiantuntijan tulee kyetä vaativaan tilastolliseen mallintamiseen ja osata ohjelmointia ja tiedonhallintaa. Näiden taitojen oppiminen puolestaan vaatii pohjakseen matematiikan osaamista. Alan vaativuuden vuoksi tutkijakoulutus on usein tarpeen käytännön työtehtävissä. Ohjelmistojen tekninen hallitseminen ei riitä lisäarvon tuottamiseen, vaan big datan hyödyntäminen vaatii sekä substanssialueen että analyysimenetelmien syvällistä ymmärtämistä. Vain näin voidaan varmistaa prosessien ja tulosten luotettavuus ja käyttökelpoisuus pitkällä aikavälillä.

Informaatioteknologian tiedekunnassa tehdään kansainvälisesti korkealaatuista IT-alan tutkimusta kahdella laitoksella, jotka ovat tietotekniikan ja tietojenkäsittelytieteiden laitokset. Tiedekunnan tutkimushankkeet liittyvät usein yhdessä kansallisten ja kansanvälisten tutkimuskumppaneiden ja teollisuuden kanssa tehtäviin tutkimus- ja kehityshankkeisiin.

Tietotekniikan laitoksen tutkimus perustuu pääosin analyyttis-konstruktiivisten menetelmien käyttöön teknisestä, laskennallisesta, matemaattisesta tai pedagogisesta näkökulmasta. Data-analyysin opetusta ja tutkimusta tehdään tietotekniikan laitoksella osana laskennallisten tieteiden ja ohjelmistotekniikan koulutusta. (Esimerkkejä data-analyysin teollisista sovellutuksista löytyy
http://www.mit.jyu.fi/ote/posterit/ohteposteri.pdf,
http://www.mit.jyu.fi/ote/posterit/kuvank.pdf).

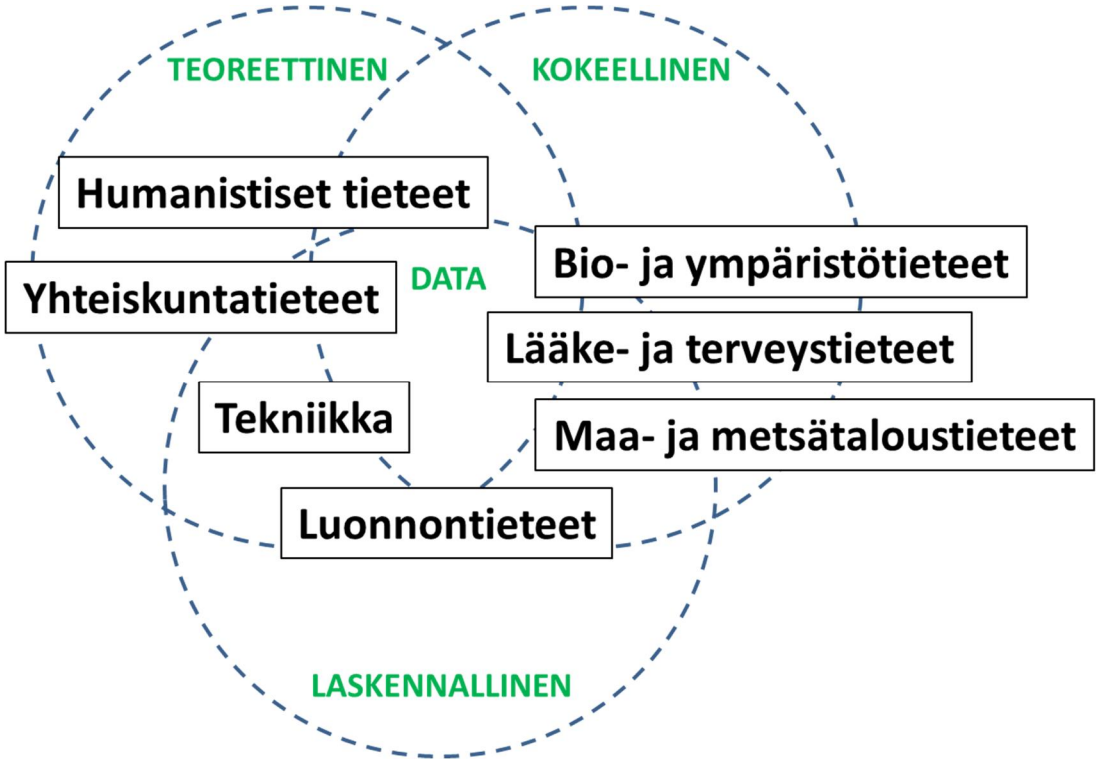
Tietojenkäsittelytieteiden laitoksen tutkimuksessa tarkastellaan tietojärjestelmiä ja tietojenkäsittelyä neljästä näkökulmasta: teknologinen, ihmiskeskeinen, liiketoiminnallinen ja informaatiokeskeinen. Nämä näkökulmat muodostavat laitoksen yleisen tehtävän: ymmärtää, kehittää, suunnitella ja hallita tietojärjestelmiä ja tietojenkäsittelyä sekä niiden vaikutuksia kokonaisvaltaisesti käyttökontekstissaan.

Data-analyysiin ja big dataan liittyvää tutkimusta tehdään IT-tiedekunnan lisäksi matematiikan ja tilastotieteen laitoksella, ihmistieteiden metodikeskuksessa (IHME) sekä sovelletaan useissa tutkimusryhmissä eri puolilla yliopistoa, mm. bio- ja ympäristötieteet, fysiikka, kemia ja sosiaalitieteet.

Perinteisesti tieteelliset tutkimusmetodit on jaettu kahteen luokkaan: teoreettiseen ja kokeelliseen tutkimukseen. Laskennallinen tiede edustaa kolmatta tieteen paradigmaa, jossa on mallipohjainen ja datapohjainen laskennallinen lähestymistapa. Neljänneksi paradigmaksi on nousemassa suurien datamassojen käsittely (Big Data Analyses).

Big dataa voidaan hyödyntää monilla tutkimusaloilla informaatioteknologian ja tietojärjestelmätieteiden ohella. Esimerkiksi biotieteiden, fysiikan, kognitiotieteen, psykologian ja taloustieteen aloilla big data -kehityksestä ja -menetelmistä on saatavissa selkeitä hyötyjä ja mahdollisuuksia tutkimuksen kehittämiseen.

Kuvassa 1 on esitetty eri tieteenalojen sijoittumista neljän paradigman alueelle, jossa data sijoittuu paradigmojen keskiöön.



KUVA 1 Eri tieteenalat tieteen neljän paradigman kentässä

Seuraavassa on kuvattu lyhyesti data-analyysiin ja big dataan liittyvää tutkimustoimintaa Jyväskylän yliopistossa.

1.1 Tieteellinen laskenta ja data-analyysi

Tieteellisen laskennan tutkimusaloja ovat matemaattinen mallintaminen, luotettava malli- ja datapohjainen simulointi, optimointi, adaptiiviset ja tehokkaat numeeriset laskentamenetelmät, epävarmuuden huomioiminen numeerisessa simuloinnissa, hajautettujen systeemien säätö, spline ja spline wavelet tekniikat signaalin ja kuvankäsittelyssä, dynaamiset systeemit ja nanoelektroniikan mallinnus.

Data-analyysin tutkimusaloja ovat analysointimenetelmien kehittäminen, erityisesti numeriikka ja massiivisen datan luokittelutekniikat, hyperspektrikameran datan analysointitekniikoiden kehittäminen ja tekniikan soveltaminen sen osa-alueilla: solubiologia, lääketiede, ympäristötiede, maa- ja metsätalous, kemialliset aseet, rikospaikkatutkimustekniikka. Lisäksi yhteistyöhankkeita on mm. fysiikan ja aivotutkimuksen alueilla.

1.2 Big data ja kyberturvallisuus

Tietotekniikan laitoksella tutkitaan tietotekniikkaa teknis-matemaattisesta näkökulmasta. Tutkimuskohteena on informaation käsittelyprosessien tehokas automatisointi. Tutkimuksen painoalat liittyvät informaatioteknologian keskeisiin alueisiin, kuten uudenlaisten tietojenkäsittelysovellusten ja ohjelmistojen suunnitteluun, tietoverkkojen tiedonsiirtojärjestelmien suunnitteluun ja hallintaan sekä tehokasta tietokonelaskentaa hyödyntävien numeeristen ja matemaattisten menetelmien ja mallien käyttöön, esimerkiksi teollisten tuotteiden suunnittelussa, teollisten prosessien ohjauksessa, luonnontieteellisessä mallintamisessa ja suurten tietoaineistojen analyysissä.

Laitoksen tutkimushankkeita, joilla on vahva yhteys big dataan, ovat mm:

Cyber Attacks Protection of Critical Infrastructures
Tutkimuksessa kehitetään innovatiivista tietojärjestelmien turvaamiseen liittyvää menetelmää. Menetelmä tutkii tietomassoista epänormaaleja käyttäytymismalleja ja tekee analyysin pohjalta päätelmiä havaintojen vakavuudesta tietojärjestelmän turvallisuudelle. Hankkeessa tutkitaan teknologioita, joiden avulla voidaan automaattisesti tunnistaa, havaita ja luokitella erilaisia haittaohjelmia.

Big Data Analytics - Data-Driven Methods for Cyber Security
Tutkimushankkeessa kehitetään automaattisia ja puoliautomaattisia laskentametodeja, analyysityökaluja ja ohjelmistoalgoritmeja, joilla voidaan analysoida suuria tietovarantoja, jotta voidaan löytää ja määritellä tuntemattomia toimintoja ja niihin liittyviä trendejä ja erityyppisten datojen välisiä suhteita mukaan lukien haittaohjelmien havaitseminen.

Organizing and analyzing massive high dimensional datasets
Tutkimushankkeen kohteena on korkea-dimensionaalisen datan analysointi. Tutkimuksessa järjestellään, klusteroidaan ja luokitellaan korkea-dimensionaalista dataa sekä tunnistetaan siitä poikkeuksia ja anomalioita. Tutkimuksessa kehitetään ydinteknologioita haittaohjelmien automaattiseen tunnistamiseen.

IT-tiedekunnan big datan sekä perustutkimuksella että soveltavalla tutkimuksella kyetään jatkuvasti tuottamaan korkeatasoisia uusia innovaatioita ja tieteellisiä läpimurtoja.

1.3 Big data ja SOTE

1.3.1 Sairaalasuunnittelu

Sairaalaympäristö on hyvä esimerkki big datan hyödyntämismahdollisuuksista. Analysoimalla dataa potilaiden, henkilökunnan, materiaalien sekä laitteiden logistiikasta, potilaille tehtävistä hoitotoimenpiteistä sekä resurssien käytöstä, on mahdollista suunnitella sairaalan hoito- sekä tukipalveluprosessit optimaalisella tavalla toteutettaviksi. Jyväskylän yliopiston IT-tiedekunnassa keskitytään sairaalatoimintojen kehittämiseen ja prosessien tehostamiseen. Hyvänä esimerkkinä tästä on Keski-Suomen sairaanhoitopiirin Uusi sairaala -projekti, jossa tiedekunta on ollut mukana jo kahden vuoden ajan: https://agoracenter.jyu.fi/projects/uuden-sairaalan-logistiikka.

1.3.2 Sosiaali- ja terveydenhuollon prosessit

Pyrittäessä potilaan hoidon kokonaisvaltaiseen optimointiin, pitää tarkastelua laajentaa organisaatiotasolta potilaan koko hoitoketjun tarkasteluun. Tässä yhteydessä big data ajattelu nousee kokonaan uudelle tasolle. Tämän tiedon perusteella luotujen laskennallisten mallien avulla voidaan lähteä etsimään optimaalista tapaa toteuttaa eri potilaiden hoito sekä luomaan ennusteita siitä missä vaiheessa hoitoon pitäisi puuttua, jotta potilaan tila ei ehtisi huonontumaan. Tällä hetkellä laskennallisen prosessianalytiikan tutkimusryhmä on kehittämässä uudenlaista työkalua tämän lähestymistavan konkretisointiin ja big datan tehokkaampaan hyödyntämiseen:
https://agoracenter.jyu.fi/projects/remaster.

Tämän lisäksi IT-tiedekunnassa on käynnistynyt hanke, jossa tähän kokonaisuuteen liitetään mukaan vielä kustannusmuuttujat ja eri rahoittaja-/maksajatahot:
https://www.jyu.fi/ajankohtaista/arkisto/2014/07/tiedote-2014-07-03-12-48-29-760690.

1.4 Big data tilastotieteessä

Tilastotiede vastaa kysymyksiin, kuinka dataa tulisi kerätä, kuinka toimia, kun data on valikoitunutta, ja kuinka epävarmuutta voi hallita. Jyväskylän yliopistossa tilastotieteen tutkimusaloja ovat tutkimusten kustannustehokas suunnittelu, biometria ja ympäristötilastotiede, rakenneyhtälömallit, puuttuvan tiedon käsittely, parametrittomat menetelmät, spatiaalinen tilastotiede ja aikasarja-analyysi,
https://www.jyu.fi/maths/tutkimus/tilastotiede.

Tutkimusta tehdään läheisessä yhteistyössä muiden tieteenalojen, tutkimuslaitosten (THL, RKTL, METLA, SYKE) ja yritysten kanssa. Yhteishankkeessa THL:n kanssa pohditaan, kuinka väestön terveydentilaa voisi seurata tehokkaasti, kun perinteisten terve-

ystarkastustutkimusten osallistumisaktiivisuus laskee. Yhteistyössä liike-elämän edustajien kanssa on selvitetty, kuinka yritykset, joilla ei ole suoraa yhteyttä asiakkaisiinsa, voisivat määrittää asiakassegmenttiensä arvon dataan perustuen ja käyttää tätä tietoa liiketoimintansa ohjaamisessa (J. Karvanen, A. Rantanen, L. Luoma (2014). Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity. Accepted for publication in *Quantitative Marketing and Economics*, http://arxiv.org/pdf/1304.5380).

1.5 Big data bio- ja ympäristötieteissä

Bio- ja ympäristötieteissä, erityisesti bio-kuvantamisen alalla datamäärän kasvu on huikea. Hyväresoluutioinen mikroskooppidata tuottaa ison määrän dataa varsinkin kun pyrkimys nykyään on tuottaa kolmiulotteista dataa ajan suhteen (neliulotteista dataa). Mikroskooppikuvantamista käytetään tänä päivänä paljolti potentiaalisten lääkkeiden testaamiseen isoissa näyteaineistossa tai kun pyritään tunnistamaan mitkä solun omat molekyylit ovat tärkeitä solun fysiologisten tai patologisten prosessien säätelijöitä. Nämä tutkimukset ovat ns. High-throughput-tutkimuksia, joissa lyhyessä ajassa tuotetaan huikeita määriä mikroskooppidataa, joista pitää tehokkaasti ja automaattisesti tulkita tutkimustulokset. Bio- ja ympäristötieteiden laitoksella on kehitetty yhteistyössä oma ohjelmistoalusta BioImageXD (www.bioimagexd.net) joka pyrkii neliulotteisen datan tehokkaaseen prosessointiin, analysointiin ja animointiin. Ohjelmistoa kehitetään lähitulevaisuudessa helpottamaan high-throughput-aineistojen analysointia. Esimerkkinä tutkimuksesta on Kankaanpää et al., BioImageXD - an open general purpose and high throughput image processing and analysis platform for biomedical images, Nat Methods. 2012 Jun 28;9(7):683-9.

1.6 Big data fysiikassa

Fysiikan laitoksella big dataan liittyvää tutkimustietoa sovelletaan mm. syklotroni laboratoriossa, nanotieteissä ja materiaalitieteissä.

Jyväskylän yliopiston fysiikan laitoksen kiihdytinlaboratorion (JYFL-ACCLAB) tieteellisenä päätehtävänä on tuottaa perustietoa aineen sub-atomaarisesta rakenteesta tutkimalla eksoottisia atomien ytimiä. Tutkimus on luonteeltaan kokeellista. Tutkimusaineistoja tuotetaan useissa mittalaitteistoissa, joiden käyttöön laboratorion kolme hiukkaskiihdytintä tuottavat ionisuihkuja. Tutkimusprojektien tuottaman datan määrä vaihtelee suuresti, muutamista kilotavuista kymmeniin teratavuihin mittausta kohden. Kokonaisuudessaan laboratorion tuottama aineistomäärä vaihtelee vuosittain 30-70TB:n välillä. Projektien elinkaari mittauksista julkaisuun saattaa olla useiden vuosien mittainen, joten aineistoille tarvitaan luotettava keskipitkän aikavälin tallennusratkaisu.

Laboratorion tutkijat ottavat osaa myös muissa laboratorioissa (esim. CERN/Isolde, GSI/FAIR) toteutettaviin kokeisiin, joissa tuotettuihin aineistoihin pätevät samat tallennuskriteerit.

Aineistot ovat pääosin mittausaineistoja ydinfysiikan kokeellisista tutkimusprojekteista. Aineistojen uudelleenkäyttö ja -analysointi tutkimusalan sisällä on normaali käytäntö. Aineistojen käyttömahdollisuudet muilla tutkimusaloilla ovat rajalliset.

2 BIG DATA JYVÄSKYLÄN YLIOPISTON KOULUTUKSESSA

Informaatioteknologian tiedekunta vastaa kehittyvän informaatioteknologian sekä digitalisoitumisen tuomiin tutkimus- ja koulutushaasteisiin. Tiedekunta yhdistää kokonaisvaltaisesti teknologian, informaation, organisaatioiden ja liiketoiminnan sekä ihmisen näkökulmat niin tutkimuksessa, koulutuksessa kuin sidosryhmäyhteistyössä. Tiedekunta kouluttaa informaatioteknologian laaja-alaisia ja kansainvälisiä osaajia sekä kauppatieteellisellä että luonnontieteellisellä koulutusalalla.

Informaatioteknologian tiedekunnalla on keskeinen rooli yliopiston painoaloihin kuuluvan ihmisläheisen teknologian kehittämisessä. Tiedekunnan keskeinen vahvuus on kyvykkyys tarkastella informaatioteknologiaa laajasti, useita näkökulmia yhdistäen ja eri ilmiöiden yhteisvaikutuksia tunnistaen. Tämä yhdistyy kansainvälisesti arvostettuun huippututkimukseen kärkialoilla ja aktiiviseen toimijuuteen ympäröivän yhteiskunnan kanssa.

IT-tiedekunta on saavuttanut johtavan aseman laskennallisissa tieteissä, kyberturvallisuudessa, tietojärjestelmätieteissä ja edustaa ainoana IT-alan tiedekuntana kognitiotieteen tutkimusta ja opetusta.

Big datan ja data-analyysin koulutusta annetaan IT-tiedekunnan seuraavissa monitieteisissä maisteriohjelmissa: sovellettu matematiikka, laskennalliset tieteet ja Web Intelligence and Service Engineering (WISE).

2.1 Sovelletun matematiikan maisteriohjelma

Sovelletun matematiikan avulla pyritään ratkaisemaan tosielämän ongelmia. Sovelletun matematiikan tavoitteena on mallintaa erilaisia ilmiöitä, kuvailla niitä ja yrittää ymmärtää niitä. Sovelletun matematiikan opiskelussa yhdistyy tieteellisen laskennan käsitteet ja menetelmät, joita käytetään kysymyksiin, jotka ilmentyvät matematiikan ja muiden tieteenalojen rajapinnoissa. Jyväskylän yliopistossa opinnoissa keskitytään sellaisiin osa-alueisiin, kuten funktionaalianalyysi, mitta- ja integraaliteoria, kompleksianalyysi, numeerinen analyysi, optimointi ja simulointi.

2.2 Laskennalliset tieteet -maisteriohjelma

Laskennallisten tieteiden maisteriohjelmassa käsitellään jatkuvan ja diskreetin simuloinnin periaatteet ja sovelluskohteet. Tavoitteena ovat jatkuvien simulointimallien tavallisimmat diskretisointimenetelmät ja niiden tehokkaan toteuttamisen perusperiaatteet moderneissa tietokonearkkitehtuureissa ja lisäksi yksi- ja monitavoitteisen epälineaarisen optimoinnin periaatteet ja ratkaisumenetelmät. Opetuksessa muodostetaan tekniikan ja luonnontieteiden ilmiöille matemaattisia simulointimalleja. Opetuksessa käsitellään laaja-alaisesti tilastotieteen, numeerisen laskennan ja ohjelmoinnin käsitteitä ja menetelmiä. Data-analyysissä opetetaan ja tutkitaan menetelmiä ja lähestymistapoja, joilla eritavoin kerätystä tiedosta (data) pyritään muodostamaan malleja ja korkeampaa tai tarkempaa informaatiota.

2.3 Web Intelligence and Service Engineering (WISE) -maisteriohjelma

Web Intelligence and Service Engineering keskittyy suunnittelemaan web-pohjaisia sovelluksia, jotka auttavat verkossa toimivaa palveluyhteiskuntaa niin julkisella kuin yksityisellä sektorilla. Maisteriohjelmalla on suora yhteys big data strategian tavoitteisiin, kuten verkossa olevan datan käsittelyssä tarvittavaan "älykkyyteen" ja tämän päälle rakennettaviin palveluihin.

On completion of the programme, the graduates will be able to use and design complex self-managed Web-based public and industrial systems, digital ecosystems 2, platforms, services and applications; will be able to connect their designs with publicly available data and Web-based capabilities as services; will be able to figure-out and approach various challenging aspects of wicked problems world-wide, which require self-managed service-based architectures for their solutions; understand and professionally utilize for that purpose knowledge on enabling technologies and tools; perform academic doctoral level studies; will be skilful in international communication due to the integrated language and communication studies. Students, who will graduate from the programme, will think beyond the routine and will be able not just to adapt to a change but to help to create and control it.

2.4 Tilastotieteen koulutus

Tilastotiede, jota Jyväskylän yliopistossa voi opiskella sekä pää- että sivuaineena, antaa hyvät valmiudet käytännön data-analyysiin. Tilastotiede vastaa kysymyksiin, kuinka dataa tulisi kerätä, kuinka toimia, kun data on valikoitunutta, ja kuinka epävarmuutta voi hallita. Tilastotieteen pääaineopintoihin kuuluu paljon myös matematiikan ja tietotekniikan opintoja ja tutkinto luo täten erinomaisen pohjan big data -tehtäviin.

Esimerkkinä big data -koulutuksesta voidaan mainita Jyväskylän yliopiston kesäkoulussa 2013 toteutetun tilastotieteen kurssin "Industrial data science", jonka luennoitsijat edustivat suomalaisen big data -osaamisen huippua yritysmaailmassa.

2.5 Ihmistieteiden metodikeskus (IHME)

Informaatioteknologian tiedekunnan panos Ihmistieteiden metodikeskuksen opetuksessa on merkittävä liittyen data-analytiikan koulutukseen ja kehittämistyöhön. Ihmistieteiden metodikeskus (IHME) tarjoaa Jyväskylän yliopiston jatko-opiskelijoille ja tutkijoille tutkimusmenetelmien ja -etiikan koulutusta yli tiedekuntarajojen ja edistää toiminnallaan tieteiden välistä tutkimusyhteistyötä ja tutkimusmenetelmien innovatiivista käyttöä. Big data -lähestymistapa sisältyy teemana IHME:en koulutukseen, jossa lähtökohtana on aineistolähtöisten analyysien hyödyntäminen monimenetelmäisessä ja -alaisessa metodikoulutuksessa. IT-tiedekunnassa tehtävään tutkimukseen perustuvan koulutuksen tavoitteina on alan kehitystä seuraten lisätä eri alojen tohtorikoulutettavien datatietoisuutta, erityisesti ymmärrystä siitä minkälaisia analyysejä ja niihin perustuvia tulkintoja voidaan big datan eri menetelmiin perustuen luotettavasti tehdä, miten big data -lähestymistapaa voidaan soveltaa eri tieteenalojen tutkimuksessa sekä minkälaisia tutkimuseettisiä valmiuksia ja menettelyitä big data - lähestymistapa edellyttää.

IT-tiedekunnan muille tiedekunnille antama opetus myös toteutuu IHME-yhteistyön kautta.

2.6 Big dataan liittyvä opetus lukuvuonna 2014–2015

Lukuvuoden 2014–2015 opetusohjelma valmistuu 15.8.2014. Ohjelmassa tulee olemaan useita big dataan, datan luokitteluun, laskennalliseen tilastotieteeseen ja tietokantoihin (mm. NoSQL) sekä sovellutuksiin liittyviä kursseja.

Lisäksi opiskelijoille tarjotaan mahdollisuus suorittaa verkkokursseja. Liitteenä 1 on lista Jyväskylän yliopiston IT-tiedekunnan opetusohjelmaan liittyvistä verkkokursseista. Päivitetty versio listasta julkaistaan 15.8 2014.

3 CENTRE OF SIMULATION AND DATA-ANALYSIS

Datan määrä ja asema yhteiskunnassa on radikaalisti muuttumassa:

- datan määrä kasvaa eksponentiaalisesti
- jalostettu ja analysoitu data on yhä keskeisempi tuottavuutta ja kilpailukykyä voimistava tekijä
- datan tuottaminen ja jalostaminen tulevat merkittäviksi liiketoiminnan alueiksi
- datan perusteella luodun tiedon esittämisen muodot ja keinot monipuolistuvat
- data-analyysi on yksi voimakkaimmin kasvavista teknologia-alueista
- suurien datamassojen käsittelystä on muodostunut uusi tieteen paradigma
- data-analyysi muuttaa merkittävästi digitaalista palvelutuotantoa

Tapahtuva muutos antaa paljon tehtäviä tutkimukselle. Toisaalta tarvitaan tutkimusta, joka liittyy datan tekniseen hallintaan, sen siirtämiseen, analysointiin ja jalostamiseen sekä turvallisuuteen erityisesti päätöksenteon tueksi. Tällainen tutkimus tukee kansantalouden kilpailukykyä ja tuotantoa. Toisaalta tarvitaan tutkimusta, joka auttaa ohjaamaan tietoyhteiskunnan kehitystä. Tällöin tutkimuskohteena on inhimillinen näkökulma datan käsittelyyn, sen luottamuksellisuuteen ja yksilön roolista data-analyysin tulosten käytön kohteena.

Big data-analyysiä voi lähestyä yleisesti käytettyjen neljän V:n määrittelyjen perusteella:

- Volume: datan määrä (sekä havaintojen että muuttujien)
- Variety: datan moninaisuus ja heterogeenisuus
- Velocity: nopeus jolla dataa syntyy
- Veracity: datan laatu

Suomessa on osaamista mm. lääketieteellisessä tutkimuksessa, mobiiliteknologioissa, peliteollisuudessa ja ympäristömonitoroinnissa, jotka kaikki ovat hyvin dataintensiivisiä ja sen monimuotoiseen analyysiin perustuvia aloja. Suomella on myös vahvaa menetelmä- ja IT-osaamista, jota muuntamalla ja hyödyntämällä koulutuksen, tutkimuksen ja asiantuntemuksen jakamisen kautta saataisiin mukaan Big Data -kehitystyöhön.

Big datan hyödyntäminen julkisella sektorilla on vasta alkutekijöissään, mutta tarjoaa suuria mahdollisuuksia niin palvelujen kuin prosessienkin parantamiseen ja tehostamiseen sekä uusiin toimintatapoihin. Suomi on ollut edelläkävijämaita avoimessa datassa ja julkinen sektori on avaamassa tietoaineistojaan. Tätä avoimuuden ja julkisten tietovarantojen saatavuuden kulttuuria tulisi hyödyntää myös Big data -kehitystyössä. Julkisten ja yksityisten data-aineistojen yhdistämisessä ja analyysissä voidaan saavuttaa merkittäviä eri osapuolia hyödyttäviä tuloksia.

Tehokas big data -tutkimus edellyttää moninaisia eri tieteenalojen tutkimusryhmiä. Tietoa louhitaan ja analysoidaan yhteistyössä muiden kanssa ja aineistolle esitetään yhä uusia kysymyksiä. Tällainen toimintatapa antaa mahdollisuuden ymmärtää laajasti digitaalista aineistoa ja tuottamaan yhä laaja-alaisempia ja tarkempia perusteita, analyysejä ja ennusteita päätöksentekijöiden käyttöön.

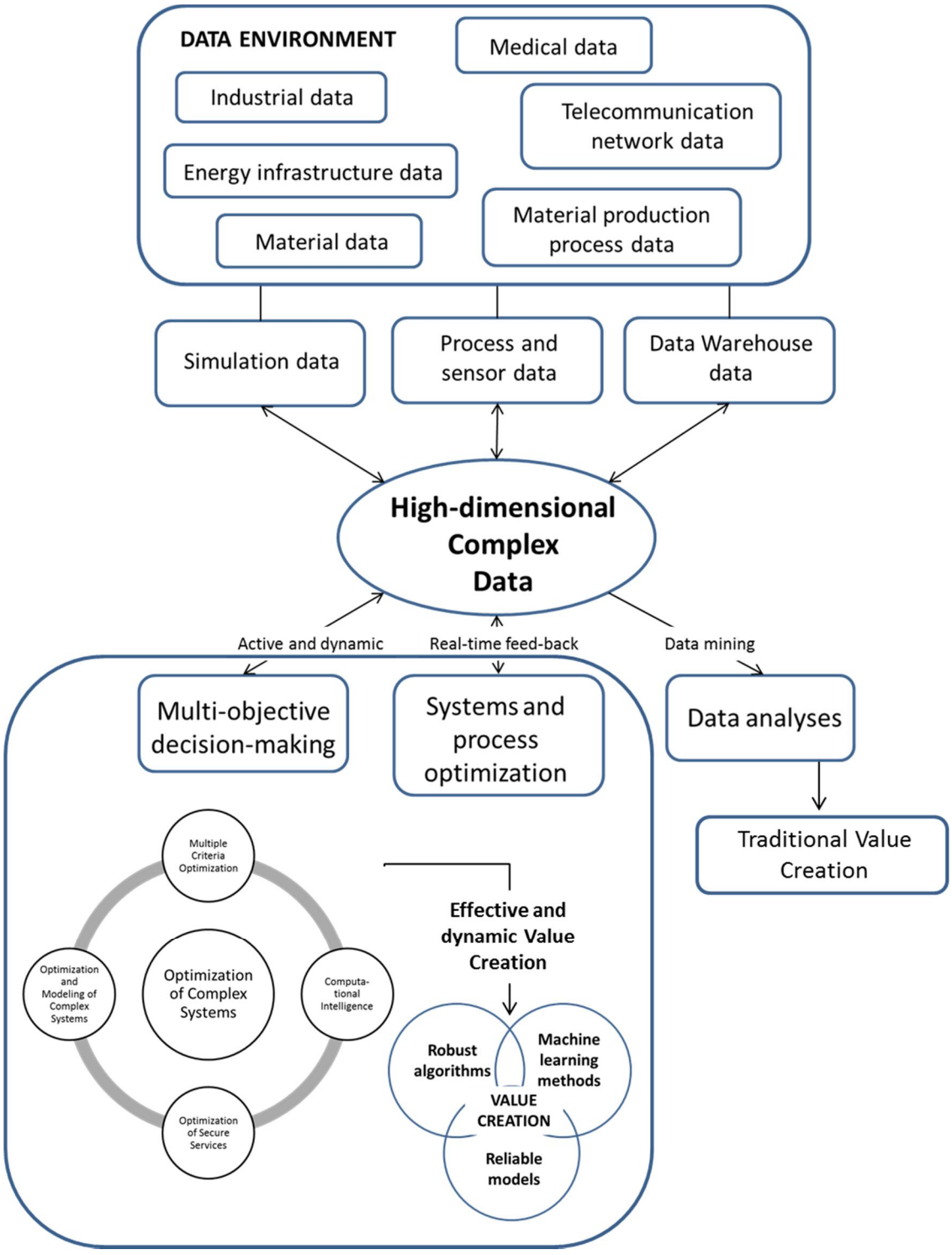
Big data -tutkimukselle on useita sovelluskohteita. SOTE-alalla Suomella on suuri potentiaali big datan suhteen. Suomesta löytyy maailmanlaajuisesti katsoen poikkeuksellisen laadukkaita ja kattavia tietokantoja. SOTE-uudistuksen myötä on mahdollisuus tutkia, kokeilla ja ottaa käyttöön big dataan perustuvia ratkaisuja. Dataan perustuvista hoitomenetelmistä ja -käytännöistä on jo saatu merkittäviä tuloksia ja hyvillä ratkaisuilla voidaan saavuttaa taloudellisia säästöjä.

Big data ajattelutapana (datatietoisuus) ja teknologiana antaa uudenlaisia näkökulmia julkishallinnolle edistää tuottavamman yhteiskunnan ja kestävyysvajeen torjumisen strategisia päätavoitteita, lisäten samalla kansalaisten tyytyväisyyttä julkisiin palveluihin. Big datan avulla on mahdollista realisoida tuottavuushyötyjä useimmilla hallinnon alueilla. Datalähtöisempää julkishallintoa voidaan tarkastella seuraavilla osa-alueilla:

- datalähtöinen päätöksenteko ja jatkuva organisaatiokehitys
- kansalaisten digitaaliset julkiset palvelut
- yritysten ja kansalaisten parempi osallistaminen julkisten palveluiden kehitykseen

Teollinen internet (IoT) antaa big data-tutkimukselle useita sovellusalueita, kuten valmistavan teollisuuden prosessit ja niiden optimointi, ennakoiva huolto, energian käytön hallinta, käyttöomaisuuden hallinta ja ennakoiva huolto. Alan tutkimukselle on laajoja mahdollisuuksia myös muualla elinkeinoelämässä, kuten kaupan ja logistiikan alueella, rakentamisessa ja kiinteistöjen hoidossa sekä kunnallisten ja muiden julkisten palvelujen tuottamisessa.

Kuvassa 2 on esitetty simuloinnin ja data-analyysin tutkimusympäristö.



KUVA 2 Centre of simulation and data-analysis

4 BIG DATA -ALAN KEHITYKSEN SEURANTA

Big data -ala ja sovellutukset kuuluvat IT-tiedekunnan strategisiin koulutus- ja tutkimusalueisiin. Alan kehityksestä tehdään puolen vuoden välein yhteenvetoja. Liitteenä 2 on 22.1.2014 tehty yhteenveto.

Jyväskylän yliopisto osallistui Liikenne- ja viestintäministeriön Big datan käyttö -työryhmän laatiman raportin laadintaan (LVM Julkaisuja 20/2014)

Kesäkuussa 2014 valmistui professori Pekka Neittaanmäen ohjauksessa Mariia Gavriushenkon tutkimusraportti: Big Data - paradigm, major topics and trends.

4.1 Data-analyysiin liittyvät väitöskirjat

Data-analyysiin ja big dataan liittyen on lukuvuonna 2013–2014 julkaistu seuraavat väitöskirjat:

- Guy Wolf: Big high-dimensional data analysis with diffusion maps
- Ilkka Pölönen: Discovering knowledge in various applications with a novel hyperspectral imager
- Tuomo Sipola: Knowledge discovery using diffusion maps
- Limor Gavish: Memcached - noSQL and big data databased particularly for caching
- Mikhail Zolotukhinin: On data mining applications in mobile networking and network security
- Antti Juvonen: Intrusion Detection Applications Using Knowledge Discovery and data Mining
- Gil Shabat: Computationally Efficient Tools for Big Data Processing
- Moshe Salhov: Manifold learning from structured kernels and out of sample extensions
- Hannu-Heikki Puupponen: Unmixing Methods in Novel Applications of Spectral Imaging

LIITE 1: JYVÄSKYLÄN YLIOPISTON IT-TIEDEKUNNAN OPETUSOHJELMAAN LIITTYVÄT VERKKOKURSSIT

Päivitetty versio listasta julkaistaan 15.8 2014.

№	Name
1	Statistics One
2	Statistics: Making Sense of Data
3	Statistics
4	Introduction to Statistics: Descriptive Statistics
5	Mathematical Statistics
6	Introduction to Probability and Statistics
7	Statistics for Applications
8	Probability and statistics
9	Probability & Statistics
10	Statistical Reasoning
11	An Introduction to Interactive Programming in Python
12	Introduction to Programming for Digital Artists
13	Creative, Serious and Playful Science of Android Apps
14	Introduction to Computer Science
15	Introduction to Computer Science and Programming
16	Introduction to Computer Science I
17	Introduction to Computer Science and Programming
18	Computer Science
19	Principles of Computing
20	Media Programming
21	Learn to Program: The Fundamentals
22	Ohjelmoinnin MOOC
23	Object-Oriented programming with Java, part I
24	Peliohjelmoinnin MOOC
25	Introduction to Systematic Program Design
26	Algoritmien MOOC
27	Algorithms, Part I
28	Algorithms, Part II
29	Algorithms: Design and Analysis, Part 1
30	Algorithms: Design and Analysis, Part 2
31	Algorithms
32	Introduction to Algorithms
33	Computer Algorithms in Systems Engineering
34	Game Theory

35 Games without Chance: Combinatorial Game Theory
36 General Game Playing
37 Advanced Algorithms
38 Introduction to Theoretical Computer Science
39 Interactive 3D Graphics
40 Foundations of Computer Graphics
41 Computational Geometry
42 Computer Graphics
43 Computer System Engineering
44 Programming Languages
45 Design of Computer Programs
46 Introduction to Data Science
47 Database Systems
48 Introduction to Computer Networks
49 Computer Architecture
50 Computer System Architecture
51 Artificial Intelligence for Robotics
52 Control of Mobile Robots
53 Cryptography I
54 Cryptography II
55 Cryptography and Cryptanalysis
56 Network and Computer Security
57 Applied Cryptography
58 Computer Security
59 Information Security and Risk Management in Context
60 Malicious Software and its Underground Economy: Two Sides to Every Story
61 Selected Topics in Cryptography
62 Advanced Topics in Cryptography
63 Designing and Executing Information Security Strategies
64 Building an Information Risk Management Toolkit
65 Network and Computer Security
66 Automata, Computability, and Complexity
67 Automata
68 Software Debugging
69 Software Testing
70 Functional Programming Principles in Scala
71 Creative, Serious and Playful Science of Android Apps
72 Computational Methods for Data Analysis
73 Computing for Data Analysis
74 Web Intelligence and Big Data
75 Data Mining
76 Statistics and Visualization for Data Analysis and Inference
77 Scientific Computing
78 Computing for Data Analysis
79 Data Analysis

80 High Performance Scientific Computing
81 Statistics: Making Sense of Data
82 Computational Methods for Data Analysis
83 Metadata: Organizing and Discovering Information
84 Machine Learning
85 Neural Networks for Machine Learning
86 Machine Learning
87 Computational Neuroscience
88 Dynamical Modeling Methods for Systems Biology
89 Introduction to Artificial Intelligence
90 Artificial Intelligence
91 Digital Signal Processing
92 Introduction to Communication, Control, and Signal Processing
93 Signal Processing: Continuous and Discrete
94 Discrete-Time Signal Processing
95 Digital Signal Processing
96 Signals and Systems
97 Machine Vision
98 Pattern Recognition for Machine Vision
99 Computer Vision
100 Computer Vision: The Fundamentals
101 Computer Vision: From 3D Reconstruction to Visual Recognition
102 Fundamentals of Digital Image and Video Processing
103 Linear and Discrete Optimization
104 Linear and Integer Programming
105 Systems Optimization
106 Optimization Methods
107 Nonlinear Programming
108 Everything is the Same: Modeling Engineered Systems
109 Introduction to Numerical Simulation
110 Introduction to Modeling and Simulation
111 Functional Hardware Verification
112 Introduction to Parallel Programming
113 Heterogeneous Parallel Programming
114 Parallel Computing
115 Theory of Parallel Systems
116 Differential Equations in Action
117 Linear Algebra
118 Coding the Matrix: Linear Algebra through Computer Science Applications
119 Calculus One
120 Calculus: Single Variable
121 Pre-Calculus
122 Visualizing Algebra
123 Web Development
124 HTML5 Game Development

125 Pattern-Oriented Software Architectures for Concurrent and Networked Software
126 CS169.1x: Software as a Service
127 Networked Life
128 Social Network Analysis
129 Videogames and Learning
130 Fundamentals of Online Education: Planning and Application
131 Gamification
132 Live!: A History of Art for Artists, Animators and Gamers
133 Online Games: Literature, New Media, and Narrative
134 How to Build a Startup
135 Startup Engineering
136 Startup
137 Leading Strategic Innovation in Organizations
138 Grow to Greatness: Smart Growth for Private Businesses, Part I
139 Grow to Greatness: Smart Growth for Private Businesses, Part II
140 Design Thinking for Business Innovation
141 An Introduction to Operations Management
142 Developing Innovative Ideas for New Companies
143 Creativity, Innovation, and Change
144 New Models of Business in Society
145 Foundations of Business Strategy
146 Design Thinking for Business Innovation
147 Content Strategy for Professionals: Engaging Audiences for Your Organization
148 International Organizations Management
149 Inspiring Leadership through Emotional Intelligence
150 Critical Perspectives on Management
151 Law and the Entrepreneur
152 Copyright
153 Markets with Frictions
154 An Introduction to Financial Accounting
155 Introduction to Finance
156 Corporate Finance
157 Organizational Analysis
158 Understanding economic policymaking

LIITE 2: YHTEENVETO BIG DATA -ALASTA JA SOVELLUTUKSISTA JYVÄSKYLÄN YLIOPISTON IT-TIEDEKUNNASSA

1. Luennot

Kesä 2013, Jyväskylän kansainvälinen kesäkoulu

- Amir Averbuch: TIEJ658 COM6: Advanced Methods for Classification of Big High Dimensional Data (JSS23), 2 op

Lukuvuosi 2013–2014

- Gil David: ITKST47 Advanced Anomaly Detection: Theory, Algorithms and Applications, 5 op
- Gil David: ITKST48 Advanced Persistence Threat, 5 op, Advanced Persistence Threat exploitation cycle
- Mauri Leppänen: ITKA204 Tietokannat ja tiedonhallinnan perusteet, 4 op
- Tommi Kärkkäinen: TIES445 Tiedonlouhinta, 5 op
- Oleksiy Mazhelis: TJTSM61 Business Analytics and Big Data Management, 5 op
- Michael Cochez: tammikuussa 2015 kurssi nimeltä "Big Data Engineering"

2. Väitöskirjat 2013

- Guy Wolf: Big high-dimensional data analysis with diffusion maps.
- Ilkka Pölönen: Discovering knowledge in various applications with a novel hyperspectral imager
- Tuomo Sipola: Knowledge discovery using diffusion maps

3. Hankkeita

- Pekka Neittaanmäki: New System for Cyber Attacks Protection of Critical Infrastructures 2012–2014, CAP-projekti (1.11.2012–31.10.2014, Tekes)
- Jari Veijalainen: Tiedonkaivuu sosiaalisesta mediasta, MineSocMed-projekti (1.9.2013–31.8.2017, SA) tarkoitus kehittää sosiaalisen median analyysialgoritmeja.
- Timo Hämäläinen: Suurien moniulotteisten datajoukkojen järjestäminen ja analysointi, HIDE-hanke (1.1.2012–31.12.2013, Tekes)
- Timo Hämäläinen: Kiinteistöautomaatiojärjestelmien datan älykäs analysointi, KIIAUDATA-hanke (1.1.2013–31.12.2014, Tekes)
- Amir Averbuch: MeBUD: Methods For Big Unstructured High Dimensional Data (1.8.2014 - 30.7.2018, haettu rahoitusta SA)

4. Maisterikoulutus

Data-analyysin monitieteellinen maisterikoulutus (DATA) on Tietotekniikan sekä Matematiikan ja tilastotieteen laitoksien yhteinen ohjelma. Opetuksessa käsitellään laaja-alaisesti tilastotieteen, numeerisen laskennan ja ohjelmoinnin käsitteitä ja menetelmiä. Data-analyysissä opetetaan ja tutkitaan menetelmiä ja lähestymistapoja, joilla eritavoin kerätystä tiedosta (data) pyritään muodostamaan malleja ja korkeampaa tai tarkempaa informaatiota.

Data-analyysin maisterikoulutus vastaa muuttuvan maailman tilanteeseen, jossa suurien data-aineistojen automaattisesta analysoinnista on tullut keskeinen työkalu useilla aloilla. Koulutuksen tavoitteena on antaa opiskelijoille data-analyysiin liittyvää erikoisosaamista sekä tilastollisista menetelmistä että niiden soveltamisesta tietokoneisiin

5. Muuta tutkimusta

- Michael Cochez: A book chapter called 'Toward Evolving Knowledge Ecosystems for Big Data Understanding' in book called ‘Big Data Computing’, 2013, http://www.crcpress.com/product/isbn/9781466578371. In this chapter we propose the use of ecosystems known from biology to tackle the big data problem.
- Michael Cochez: much of my own research is related to big data, mainly the alignment of huge ontologies.
- Timo Hämäläinen: Verkkoliikenteen analyysiin liittyvää tutkimusta (myös TIES326 tietoturvakurssilla näitä esillä). Näitä jatketaan uusien datamöykkyjen kanssa (JAMK:n laboratorio jne.).

Timo Hämäläinen: Network traffic analysis

Nowadays HTTP servers and applications are some of the most popular targets for network attacks. The easiest way to carry out such attacks is to inject malicious code into HTTP request messages. Such intrusive requests can be extremely dangerous since they can corrupt the server or collect confidential information from the server databases. One of the options to detect these attacks is to process all HTTP queries as text lines and transform them to numeric feature matrices, which then are used to find intrusions with anomaly detection algorithms. However, since there can be many different kinds of requests to different HTTP applications for each unique web resource one feature matrix is constructed and analyzed as a rule. Thus, for a huge HTTP server several thousands of matrices would need to be built, which is not efficient from the computing resources point-of-view. In addition, it is also difficult to define normal users behavior for resources that have been requested few times. In this research, we considered an algorithm for HTTP intrusions detection based on simple clustering algorithms and advanced processing of HTTP requests which allows the analysis of all queries at once and does not separate them by resource. The method proposed allows detection of HTTP intrusions in case of continuously updated web-applications. The algorithm is tested using logs acquired from a large real-life web service and, as a result, almost all attacks from these logs are detected, while the number of false alarms remains very low. [1] We have also studied online detection of anomalous HTTP requests with Growing Hierarchical Self-Organizing Maps (GHSOMs). By applying an n-

gram model to HTTP requests from network logs, feature matrices were formed. GHSOMs are then used to analyze these matrices and detect anomalous requests among new requests received by the webserver. The system proposed is self-adaptive and allows detection of online malicious attacks in the case of continuously updated web-applications. The method is tested with network logs, which include normal and intrusive requests. Almost all anomalous requests from these logs are detected while keeping the false positive rate at a very low level.[2]

In addition, real-world network logs were analyzed using dimensionality reduction technique called Diffusion Map (DM). First, some features like n-grams or character frequency was calculated to form a feature matrix. The dimensionality of this matrix was reduced for the purposes of visualization and facilitating subsequent clustering and other analysis phases. Principal Component Analysis (PCA) was used as a comparison, since it's one of the most frequently used methodologies. The main advantage of Diffusion Maps is the fact that it can handle non-linear dependencies in the data, which PCA cannot. DM is used successfully to find actual intrusions from the data, as well as visualizing the structure of the network traffic. Subsequently, a clustering algorithm, such as k-means, can be used to find anomalies and structures in the data. These will help network administrators detect intrusions and other anomalies from a network. Also, a rule extraction algorithm was implemented an d tested. This idea comes from the world of credit card fraud detection. The point is to first analyze and cluster network traffic using methods described previously. The problem is that running dimensionality reduction and clustering algorithms continuously can be computationally expensive. For this reason, conjunctive rule extraction is used to automatically generate signatures that approximate the traffic clustering result. Using these rules is very efficient, and they can be periodically updated completely automatically. This approach combines the accurate clustering of the advanced algorithms and the efficiency of using simple signature rules to classify traffic into normal and anomalous. [A,3,4]

[A] T. Sipola, A. Juvonen, J. Lehtonen: "Dimensionality Reduction Framework for Detecting Anomalies from Network Logs", Engineering Intelligent Systems, 20(1):87-97, 2012

[1] M. Zolotukhin, T. Hämäläinen: "Detection of Anomalous HTTP Requests Based on Advanced N-gram Model and Clustering Techniques". NEW2AN 2013: International Conference on Next Generation Wired/Wireless Networking, August 28, St. Petersburg, Russia, 2013

[2] M. Zolotykhin, T. Hämäläinen and A. Juvonen, "Online Anomaly Detection by Using N-gram Model and Growing Hierarchical Self-Organizing Maps", IEEE IWCMC 2012, August 27-31, Limassol, Cyprus, 2012

[3] A. Juvonen, T. Sipola: "Adaptive Framework for Network Traffic Classification Using Dimensionality Reduction and Clustering", In Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2012 4th International Congress on, pages 274-279, St. Petersburg, Russia, October 2012. IEEE

[4] A. Juvonen, T. Sipola: “Combining Conjunctive Rule Extraction with Diffusion Maps for Network Intrusion Detection”, The 18th IEEE Symposium on Computers and Communications (ISCC’13), July 7-10, 2013, Split, Croatia.

Informaatioteknologian tiedekunnan julkaisuja
No. 17/2014

ISBN 978-951-39-6046-9 (verkkoj.)
ISSN 2323-5004

JYVÄSKYLÄN YLIOPISTO