



18.6.2014

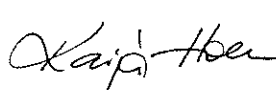
Liikenne- ja viestintäministeriö

Viite: Liikenne- ja viestintäministeriön lausuntopyyntö 3.6.2014, LVM138/05/2014

Tampereen yliopiston lausunto kansallisen big data –strategian luonnoksesta

Ohessa Tampereen yliopiston lausunto.

Rehtori


Kaija Holli

Liite

Tampereen yliopiston lausunto

Lausunto kansallisen big data -strategian luonnoksesta

Tampereen yliopisto, 18.6.2014

Kommentteja big data -strategian luonnoksesta

Yleiskommenttina pidämme strategiaa hyvänä ja tarpeellisena. Alla on muutamia kommentteja yksityiskohdista.

1. Strategialuonnoksessa on hyvin tunnistettu lääketieteellisen tutkimuksen teemat, ja toimenpide-ehdotukset niiden suhteen ovat kannatettavia. Kuten luonnoksessa mainitaan, Tampereen yliopistoon on perustettu Masters Programme in Computational Big Data Analytics -koulutusohjelma, jolla varmistetaan tarvittava analyysiosaaminen paikallisesti.
2. Luonnoksessa mainitaan ”Tutkimuksessa ei välttämättä tarvitse samalla tavalla muodostaa teoriaa, kun voidaan käydä valtavia tietomassoja läpi ilman ennakko-oletuksia ja ns. ”löytää” malleja, yhteyksiä ja selityksiä.”

Tähän tarkennamme että suuria tietoaaineistoja käsitellessä tarvitaan edelleen hyviä lähtöoletuksia ja johdonmukaista todennäköisyyksiin ja malleihin perustuvaa päättelyä.

Päättely säilyy haastavana kahdesta syystä: 1) Vaikka aineistot sisältävät suuria määriä näytteitä, samalla mitattujen piirteiden määrä on kasvanut, esimerkiksi bioinformatiikassa mittausaineiston kudosnäytteiden määrä saattaa edelleen olla huomattavasti pienempi kuin geenien määrä, joiden aktiivisuus mitataan kullekin näytteelle. 2) Suurilla aineistoilla halutaan usein päätellä aiempaa monimutkaisempia ("yksityiskohtaisempia") malleja kuin pienillä aineistoilla, ja näillä monimutkaisilla malleilla on enemmän parametreja, jotka tulee päätellä aineistosta. Nämä kaksi syytä yhdessä aiheuttavat sen, että parhaimpien selityksien (mallien) löytäminen aineistoille on edelleen haastava päättelytehtävä, jonka ratkaisemiseksi tulee tuntea mm. tilastotieteen menetelmiä. Johdonmukaista todennäköisyyspohjaista ennakko-oletuksien ja uuden aineiston yhdistämistä tarvitaan, jottei monimutkaisissa malleissa tehdä harhaan meneviä päätelmiä, joissa yksittäisten näytteiden ominaisuuksia erehdytään luulemaan osaksi yleistä trendiä. Yksinkertaisemmat lähestymistavat puolestaan saattavat toimia kelvollisesti, mutteivät välttämättä havaitse niitä hienovaraisia yllättäviä löydöksiä, jotka "big data" mahdollistaa.

Näin ollen olemassa oleva asiantuntemus kuten teoriat ja ennakko-oletukset kannattaa ottaa mukaan päättelyyn: olemassa olevaan aiempaan tietoon perustuva parempi lähtöoletus ja johdonmukainen tilastotieteen keinoin tehty päättely parantaa edelleen tuloksia varsinkin kun halutaan päätellä aineistoista hienovaraisia tai monimutkaisia ominaisuuksia. Aiempia teorioita ja oletuksia ei käytetä jyrkkinä rajoitteina löydöksille, vaan kuvauksena alkuperäisistä uskomuksista, joita päivitetään aineiston avulla - tällöin mahdollistetaan se, että aineisto voi joko vahvistaa teorioita tai tuottaa niistä poikkeavia uusia löydöksiä.

Tampereen yliopiston tuleva Computational Big Data Analytics -maisteriohjelma painottaa opetuksessaan tällaisia hyvin käyttäytyviä päättelymenetelmiä, joiden avulla voidaan yhdistää aiemmat oletukset aineistojen tuomaan uuteen tietoon.

3. Big datan mahdollisuuksista: sen lisäksi että big data mahdollistaa parempien päätösten tekeminen ennalta tiedossa olleista kysymyksistä, suurten tietoaaineistojen eksplorointi

mahdollistaa yllättävien löydösten tekemisen, ja näin ollen tarttumisen aiemmin havaitsemattomiin tilaisuuksiin.

4. Edellytykset -osion koulutusohjelmalistassa on pieni kirjoitusvirhe: Tampereen yliopiston tarjoa maisteriohjelma on nimeltään Masters Programme in Computational Big Data Analytics (sivulla tulisi siis lukea "Analytics", ei "Analysis").
5. Sivun 12 kuva ja sivun 13 esimerkki big data -tiimistä on hyvä, mutta on huomioitava että usein yksittäistenkin tehtävien paras suorittaminen vaatii monialaista osaamista: esimerkiksi parhaimmat tiedon visualisointimenetelmät saattavat toimia tilastollisilla periaatteilla ja optimoida aineiston visualisointeja algoritmisilla keinoilla; mm. Tampereen yliopistolla prof. Jaakko Peltonen on kehittänyt tällaisia edistyneitä tiedon visualisointimenetelmiä. Tällaisten menetelmien kehittäjä saattaa olla yhtä aikaa tilastotieteen, algoritmien ja visualisoinnin osaaja; koulutustarjonnan monipuolistuessa esim. Tampereen yliopiston Computational Big Data Analytics -ohjelman kautta tällaisia usean osa-alueen taitajia tulee saataville aiempaa enemmän. Toisaalta menetelmät kehitetään helppokäyttöisiksi jotta niitä voi käyttää ilmankin laajaa osaamispohjaa, ts. menetelmät tehdään palvelemaan useanlaisia big data -tutkimus- ja yrityskäyttäjää.
6. Kommentti ajatuksesta "yhteiskunnan luoma luotettu dataoperaattori" (sivu 20): eräs hankaluus tässä on että analysointitapa ja analysointimenetelmät kehittyvät ja ovat saattaa olla tarve kehittää uusia menetelmiä uuden aineiston tai analyysin tarpeisiin; yhteinen dataoperaattori ei aina helposti pystyisi hyödyntämään kasvavaa kirjoa analysointimenetelmiä, koska niiden asiantuntemus on menetelmien kehittäjillä.
7. Sivun 20, Suomen laaja start-up-kenttä: tässä voisi mainita AaltoES:n lisäksi myös Tampereella toimivat Demola Tampere (www.demola.fi) ja Protomo Tampere (www.protomo.fi).

Big data -toimenpiteet Tampereen yliopistossa

Tampereen yliopisto panostaa suurten tietoaaineistojen analysoinnin tutkimukseen ja opetukseen erityisesti kahdella käynnissä olevalla uudella aloitteella.

Tampereen yliopiston Informaatitieteiden yksikköön perustetaan uusi kansainvälinen maisteriohjelma **International Master's Programme in Computational Big Data Analytics**. Tämä ohjelma onkin lyhyesti mainittu nykyisen strategialuonnoksen edellytyksiä -osiossa. Ohjelma toteutetaan yhteisvoimin erityisesti tilastotieteen ja tietojenkäsittelytieteen asiantuntijoiden vetämänä ja yhteistyössä tilastotieteen sovellusalojen kuten bioinformatiikan, biometrian ja biostatistiikan, yhteiskuntatieteiden ja datajournalismin kanssa. Ohjelma kouluttaa opiskelijoita asiantuntijoiksi, joilla on sekä tilastotieteellinen valmius tehdä haastavia analyyskejä laajoista ja monimuotoisista tietoaaineistoista, tietojenkäsittelytieteen taidot toteuttaa analyyskejä olemassaolevia ohjelmistoja käyttäen ja/tai muokaten ohjelmistoja aineistojen ja analyysien erityistarpeita varten, sekä ymmärrys suuren datan analysoinnin käytännöstä sovellusaloilla. Ohjelman sisällössä yhdistyy sekä metodologinen opetus joka toteutetaan Informaatitieteiden yksikön kautta sekä tutustuttaminen opetuksen kautta suuria tietoaaineistoja tuottaviin sovellusaloihin ja niiden erityispiirteisiin ja menetelmiin, mikä toteutetaan yhteistyössä muiden yksiköiden opetuksen kanssa. Aineistoja saadaan yritysten kanssa yhteistyössä toteutettavista tutkimushankkeista (esim. Informaatitieteiden yksikössä toimivalla pelitutkimusryhmällä on tällaisia pelaaja-metriikkaan ja analyysiin liittyviä tutkimushankkeita), sekä Informaatitieteiden yksikön ja muiden yksikköjen

opetushankkeissa toteuttavista kokeilupalveluista ja -sovelluksista. Suuret tietoaaineistot ovat ohjelman erityinen fokusalue, joka otetaan huomioon opetuksen sisällössä käsittelemällä suurten aineistojen tilastollisia ominaisuuksia, tehokkaita käsittelymenetelmiä, ja sovellusalojen tiedonkeräystapoja jotka tuottavat suuria aineistoja. Ohjelman ensimmäinen hakudeadline on tammikuussa 2015 ja jonka ensimmäiset opiskelijat aloittavat syksyllä 2015. Ohjelmaa mainostetaan jo aktiivisesti kansainvälisillä foorumeilla.

Tampereen yliopistoon perustetaan myös uusi **Tilastotieteen keskus**, joka koordinoi tilastotieteen opetusta ja tutkimusta. Keskuksen toiminnan aloittamisen tavoiteaikataulu on syyslukukauden 2014 alussa. Tällä hetkellä Tampereen yliopiston tilastotieteen opetusta annetaan Informaatitieteiden yksikössä matematiikan ja tilastotieteen kandidaattiohjelman, tilastotieteen maisteriohjelman, ja informaation ja järjestelmien tohtoriohjelman yhteydessä, sekä Terveystieteiden yksikössä erityisesti biometrian ja biostatistiikan alalla, Johtamiskorkeakoulussa, ja Yhteiskunta- ja kulttuuritieteiden yksikössä; tilastotieteen tutkimusta ja tilastotiedettä soveltavaa tutkimusta tehdään näissä yksiköissä. Perusteilla oleva Tilastotieteen keskus tulee järjestämään tätä opetusta ja osallistumaan tilastotieteellisten opetustavoitteiden laadintaan kaikissa yliopiston tutkinto-ohjelmissa. Keskus koordinoi opetuksen ja ajankohtaisen korkeatasoisen tutkimuksen vuorovaikutusta. Keskus myös koordinoi tilastotieteen asiantuntijoiden käyttöä tutkimusprojekteissa ja korkeatasoisen tilastotieteellisen näkökulman ja tavoitteiden huomioonottamista tutkimusrahoituksen hakemuksissa. Keskus erityisesti toimii yhteistyöelimenä suurten tietoaaineistojen hyödyntämiseksi: keskuksen kautta kootaan ja välitetään tietoa suurten tietoaaineistojen saatavuudesta, niiden käytön haasteista ja järjestetään metodologisten ja sovellusalojen voimavarojen yhdistämistä suurten aineistojen tehokkaaksi hyödyntämiseksi tutkimuksessa ja opetuksessa. Yksikön johtoon palkataan uusi soveltavan tilastotieteen professori (tenure track), ja yksikön johtoryhmään kuuluu edustaja kustakin yllämainituista tieteenalaysiköistä; tämän lisäksi tutkimuksen vahvistamiseksi voidaan koota keskukselle erityinen scientific advisory board.

Näiden aloitteiden lisäksi huomattavaa on, että Tampereen yliopistolla tuotetaan ja käsitellään tietoaaineistoja, jotka edustavat hyvin strategialuonnoksessa käsiteltyjä big data -aineistojen mahdollisuuksia ja aineistojen käsittelyn näkökohtia. Eräs tärkeä aineisto on lääketieteen yksikön tutkimukseen liittyvä biopankki, lääketieteellinen kokoelma, joka sisältää yksittäisten ihmisten kudosperäisiä näytteitä sekä tietoja kunkin näytteenantajan terveydestä ja elintavoista. Lääketieteen näkökulmasta biopankki tulee tulevaisuudessa sisältämään merkittävän määrän tietoa, joka on pyrittävä hyödyntämään tehokkaasti. Tiedon käsittelyssä on kiinnitettävä erityistä huomiota tietojen hallintaan ja salassapitoon, koska kysymyksessä ovat henkilöihin liittyvät asiat. Jatkossa saattaa olla tarpeen käsitellä biopankin sisältämiä tietoja useampien keskustusten välillä ja myös kansainvälisesti, jolloin erilaiset tiedonsiirto- ja harmonisointiasiat tulevat ajankohtaisiksi.

Yhteystiedot:

International Master's Programme in Computational Big Data Analytics -ohjelman vastuuhenkilöitä ovat prof. Jaakko Peltonen (jaakko.peltonen@uta.fi), prof. Jyrki Nummenmaa (jyrki.nummenmaa@uta.fi) ja prof. Martti Juhola (martti.juhola@uta.fi).

Tilastotieteen keskuksen valmistelun vastuuhenkilöt ovat Informaatitieteiden yksikön johtaja Mika Grundström (mika.grundstrom@uta.fi), Yhteiskunta- ja kulttuuritieteiden yksikön johtaja prof. Risto Kunelius (risto.kunelius@uta.fi), ja prof. Pekka Rissanen (pekka.rissanen@uta.fi).

Tampereen yliopiston rooli kansallisessa big data -kehittämistyössä

Tampereen yliopisto haluaa olla mukana aktiivisena ja korkean profiilin toimijana big data -koulutuksessa ja -tutkimuksessa, mihin myös uudet aloitteemme (maisteriohjelma ja tilastotieteen keskus) tähtäävät.

- Uuden maisteriohjelman tarjoama koulutus voi tuoda yliopiston merkittäväksi big data -asiantuntijoiden kansalliseksi kouluttajaksi tutkimusta ja yritysmaailmaa varten.
- Tampereen yliopistolla tehdään aktiivista tutkimusta useihin big data -sovelluksiin kuten esimerkiksi avoimiin yhteiskunnallisiin aineistoihin (muun muassa liikenneaineistoihin) liittyen ja suuriin tieteellisten artikkeleiden kokoelmiin liittyen, ja luodaan uusia big datan käsittely- ja analysointimenetelmiä kuten tehokkaita suurten korkeaulotteisten tietoaaineistojen visuaalisen eksploroinnin menetelmiä.

Lisäksi yliopisto toimii aktiivisesti kommentoijana ja asiantuntijana yhteiskunnallisissa big data -päättöissä; yliopiston asiantuntijoista erityisesti prof. Jyrki Nummenmaa on jo nyt mukana strategialuonnosta valmistelevassa big data -työryhmässä ja prof. Jaakko Peltonen on ollut luonnoksen kommentoijana sille perustetuilla Opatnet-wikisivuilla.

Yliopisto osallistuu mahdolliseen datalähettilästoimintaan, big data -klusterin muodostamisessa ja toiminnassa. Yliopisto osallistuu big data -koulutusohjelmien kansalliseen koordinointiin ja big dataan erikoistuvien yliopistojen verkoston luomiseen. Yliopisto osallistuu big data -hautomojen toimintaan mm. datalähettiläiden kautta. Yliopisto osallistuu big data infrastruktuurin verkostotoimintaan. Yliopisto osallistuu mm. asiantuntijoin mahdollisiin big data standardointiprosesseihin ja näihin liittyviin open data standardointiprosesseihin kuten älyliikenteen standardointiin ja mahdollisiin kansallisiin dataportaaleihin. Yliopisto osallistuu myös data-analytiikan toimijoiden verkoston luomiseen.