Intro:

Tervetuloa kuuntelemaan CSC:n For a Future -podcastia.

Teemu Roos: Mun nimi on Teemu Roos, ja mä olen tietojenkäsittelytieteen proffa Helsingin yliopistolla. Tänään puhutaan LIFEPLAN-hankkeesta. Puhutaan luonnon monimuotoisuudesta ja koneoppimisesta, ja miten niiden avulla voidaan seurata ilmaston ja maankäytön muutosta.

Mulla on tänään juttukaverina kolme LIFEPLAN-hankkeessa olevaa henkilöä: professori Otso Ovaskainen Jyväskylän yliopistolta, Bess Hardwick, joka on LIFEPLAN-hankkeen projektisuunnittelija ja lisäksi CSC:ltä Hanna Koivula, joka on ensinnäkin tutkimusaineistojen hallinnan asiantuntija, mutta taustaltaan myös biologi.

Otso, voitko kertoa ihan ensimmäiseksi, että mikä tää LIFEPLAN-hanke on ja miksi se on niin tärkeä?

Otso Ovaskainen: Joo, eli LIFEPLAN on tämmöinen globaali, maailmanlaajuinen biodiversiteetin kartoittamishanke. Ja tietenkin biodiversiteettiä on kartoitettu iät ja ajat, mutta se mikä tässä LIFEPLAN-hankkeessa on uutta on se, että se käyttää näitä uusia tietojen keräämisen menetelmiä. Elikkä me kartoitetaan luonnon monimuotoisuutta DNA:n, äänen ja kuvan avulla. Ja nyt tämä auttaa meitä ottamaan tämmöisen täysin uudenlaisen systemaattisen otoksen koko maailman luonnon monimuotoisuudesta. Meillä on parisataa tutkimusryhmää ympäri maailmaa, jotka kerää näitä aineistoja ja niiden avulla sitten on tarkoitus ymmärtää paremmin tätä luonnon monimuotoisuutta ja sitä miten vaikka ilmastonmuutos siihen vaikuttaa.

Teemu Roos: Aika laaja hanke - parisataa tutkimusryhmää ympäri maailmaa.

Otso Ovaskainen: Joo, tää on tämmöinen Euroopan tiedeneuvoston eli ERC:n rahoittama tämmöinen synergyhanke. Eli tässä on aika iso Euroopan laajuinen rahoitus. Ja tässä on kolme vetäjää tällä hankkeella, eli meillä on biologinen kärkivetäjä Tomas Roslin Ruotsista. Sitten meidän mallinnuksen ja tilastopuolen vetäjä on David Dunson Amerikasta. Ja mä oon sitten tämmöinen tilastoekologi, joka varmasti ei ymmärrä hirveän paljon tilastoista eikä ekologiasta, mutta jollain tavalla osaa niitä yhdistää, niin sitten se minä olen se kolmas henkilö hankkeessa.

Teemu Roos: Mielenkiintoista. Mainitsit koneoppimisen mallinnuksen. Miten niitä hyödynnetään tässä hankkeessa?

Otso Ovaskainen: No tämä hanke tulee tuottamaan ihan valtavan määrän dataa, eli tämä tulee tuottamaan noin 1000 vuotta äänimateriaalia 5 vuoden aikana, kun niitä ääninauhureita on iso joukko ympäri maailman. Se tulee tuottamaan ehkä noin 100 miljoonaa riistakamerakuvaa, ja sitten noin 10 biljoonaa dna-sekvenssiä, eli näitä ei tietenkään voi käydä manuaalisesti läpi. Eli se ensimmäinen vaihe mihin koneoppimista tarvitaan, on tunnistaa mitä lajeja näissä näytteissä on.

Ja nyt kun puhutaan osittain sienistä ja hyönteisistä, kun puhutaan siitä DNA-aineistosta. Itse asiassa suuri osa lajeista on edelleen tieteelle tuntemattomia, eli me ei voida pelkästään tunnistaa, että mitä tieteelle tunnettuja lajeja siellä näytteissä on, vaan myös sitä, että mitä sellaisia lajeja on, jotka on tällä hetkellä tieteelle täysin tuntemattomia. Niin tämän materiaalin läpikäyntiin tarvitaan sitten niitä koneoppimismenetelmiä.

Se on se ensimmäinen vaihe, mutta sitten tietenkin mallintamista tarvitaan myös siihen seuraavaan vaiheeseen, että sitten kun ne lajit on tunnistettu, niin mitä me opitaan tästä aineistosta vaikkapa juuri ilmastonmuutoksen näkökulmasta?

Teemu Roos: Super mielenkiintoista! Bess Hardwick, voitko sä kertoa enemmän siitä, että miten sitä dataa kerätään? Mistä se data on peräisin? Kuletteko te jonkun mikrofonin kanssa metsässä keräilemässä jotain, vai onko siinä jotain systematiikkaa, että mistä sitä kerätään ylipäänsä, että miltä se data näyttää tai kuulostaa sitten ihan konkreettisesti?

Bess Hardwick: No sitä äänidataa kerätään tämmöisillä autonomisilla nauhureilla, jotka on sellaisia pieniä, luottokortin kokoisia laitteita, joihin menee paristo ja muistikortti. Ja ne voi ohjelmoida silleen, että ne nauhoittaa tietyllä aikataululla, ja ne pystyy myös nauhoittaa lepakkotaajuuksia ja lintutaajuuksia. Ja ne pystyy myös silleen triggeröitymään, että kun ne kuulee lepakkotaajuuksia, niin ne menee päälle ja lähtee nauhoittaa.

Teemu Roos: Siis mikä? Tämä ei liity varmaan lepakkomieheen. Siis mikä on tämä lepakkotaajuus?

Bess Hardwick: Lepakkojen äänet kuuluu vähän eri taajuuksilla, että että ne ei kuulu samalla tasolla kuin linnut.

Teemu Roos: Siis se on se tutkajuttu, mitä ne lepakot käyttää, vai?

Bess Hardwick: Joo.

Teemu Roos: Okei, vau. Kiehtovaa.

Bess Hardwick: Ja näitä nauhureita ollaan lähetetty maailmanlaajuisesti satakuntaan paikkaan, missä jokaisessa on viisi tämmöistä nauhuria semmoisessa neliöasetelmassa ja yksi keskellä, ja siellä ne sitten nauhoittaa. Ja näissä 100 paikassa on paikallisia tutkijoita, jotka oman työnsä ohessa haluaa tehdä tätä näytteenottoa. He käyvät kerran viikossa siellä nauhureilla vaihtamassa paristot ja muistikortit, sekä lähettävät sitä dataa meille.

Ja sitten siellä on myös riistakameroita, jotka laukeavat, kun eläin kulkee siitä ohi – niin valossa kuin pimeällä, eli niissä on tällainen infrapunapuoli myös. Ja niissä on myös muistikortit, joilta lähetetään näitä riistakamerakuvia meille valtavia määriä.

Teemu Roos: Niitä kuvia me ei voida esitellä tässä podcastissa, mutta meillä on ääninäyte itseasiassa, joka me ollaan saatu. Me voidaan kuunnella tähän väliin vähän sitä ääninäytettä.

- Ääninäytettä soitetaan -

Hanna Koivula: Ääninäytteestä tunnistin ilman koneoppimista peipon.

Teemu Roos: No niin, hyvä!

Eli Hanna Koivula… Sen lisäksi, että olet tutkimusaineistojen hallinnan asiantuntijana, olet myös biologi, eli tavallaan tutkimuskohdekin on tuttu. Onko siitä tässä työssä paljon hyötyä?

Hanna Koivula: No ainakin se tekee tästä äärimmäisen kiinnostavan projektin itselleni olla mukana. Mä oon aikaisemmassa työssäni tehnyt myös niin sanotun kansalaistieteen parissa aika paljon, ja mä näen paljon sellaisia synergioita eli se mitä tästä opitaan, niin sitä voidaan mahdollisesti sitten hyödyntää vielä

laajemmasti siinä, että voidaan harrastajille ja kelle tahansa tarjota välineitä, joilla voi oppia lisää luonnosta ja samalla kerätä havaintoja tieteen käyttöön. Mutta se ei ole vielä tässä hankkeessa se ydin ymmärtääkseni.

Teemu Roos: Jos puhutaan siitä teknisestä puolesta, niin mitä haasteita tällainen aineisto tuottaa? Onko se aineiston määrä, vai onko siinä jotain muita sellaisia haasteita, kuin että täytyy hankkia isompi kovalevy?

Hanna Koivula: No niinku tosta Otson kuvauksesta kuulitte, niin se tulee olemaan aivan järkyttävä se datamäärä oikeasti.

Teemu Roos: Ettei ehkä yksi kovalevy riitäkään?

Hanna Koivula: Että lajistoa ja luontoa on kartoitettu koko sen ajan, kun on ollut tiedettä, tai kun on syntynyt biologiaa ja ekologiaa. Biodiversiteetti käsitteenä on syntynyt vasta 80-luvulla, ja pikkuhiljaa on alettu ymmärtää sitä monimuotoisuuden merkitystä ja sitä, että oikeasti edelleen on lajeja vaikka kuinka paljon. Me ei tiedetä kuinka paljon lajeja on, eikä me tiedetä mitä me ei tiedetä, ja sen kartoittaminen on äärimmäisen haastavaa ja sitä tietoa tarvitaan. Vasta ihan viime aikoina ollaan alettu ymmärtää sitä, että mitä se merkitsee, että monimuotoisuus rapautuu ja niitä seurauksia. Me ollaan kaikki osa sitä oikeasti.

Sen asian tai kokonaisuuden ymmärtäminen paremmin on tässä mun mielestä se todella kiehtova asia. Ja CSC:n palveluna me voidaan tarjota sille datamäärälle paikka, jossa se voidaan organisoida, sekä kapasiteetti, millä sitä koneoppimista pyöritetään. Eli se haaste tulee kyllä siitä datamäärästä. Datatyyppejä ei ole hirveän montaa erilaista, eli se työn kulku saadaan kyllä järjestettyä. Mutta se, että sitä on niin paljon, niin se tekee tästä haastavaa.

Ja sitten toisaalta se lajin tunnistaminen koneoppimisellakin ei ole mikään yksinkertainen asia, vaan sitä on yritetty pitkään ja vasta nyt ehkä alkaa olla eväitä ja osaamista siihen, että että voidaan opettaa koneelle, että mitkä ne tunnuspiirteet on äänestä tai valokuvasta, tai DNA:sta määritellä, että mitä lajeja. Sitten on vielä semmoinen pieni ongelma, että sen lajin käsite ei ole mitenkään yksiselitteinen. Ehkä ei mennä siihen vielä tässä kohtaa.

Teemu Roos: Ehkä palataan vielä siihen. Mutta kuten Otso sanoi tuossa alussa, että siellä on paljon sellaisia lajeja, jotka ei ole vielä tieteelle tunnettuja, niin mun on vaikea kuvitella, että miten sitten sellaista pystyy koneoppimisella käsittelemään? Että kun pitäisi kuitenkin olla sitä opetusdataa, joka pitäisi olla “labeloitya” niin, että siinä olisi aina nimetty se laji. Miten ihmeessä sellaisessa tilanteessa voidaan tehdä jotain? Jos se on aiemmin tieteelle tuntematon laji, tai muuten laji, josta ei ole mitään opetusaineistoa.

Hanna Koivula: Toki ensin täytyy lähteä liikkeelle siitä, mitä me tunnetaan, eli poimia sieltä ne lajit.

Ja sitä mä en oikeasti tiedä miten on ajateltu hoitaa sitä, että sieltä muusta kohinasta sitten alkaa löytyä jotain.

Teemu Roos: Mitä tehdään silloin, Otso?

Otso Ovaskainen: Ensinnäkin jos puhutaan linnuista ja nisäkkäistä, niin sieltä me ei varmaan hirveästi tieteelle uusia lajeja löydetä. Mutta sitten kun puhutaan sienistä tai hyönteisistä, joita tunnistetaan

DNA-menetelmillä, niin se on ihan tyypillistä, että ehkä vaan 10% siitä lajistosta, joka on siellä meidän näytteessä, niin tunnetaan tieteelle. Että se ei ole millään tavalla eksoottista, että siellä on tieteelle tuntemattomia lajeja vaan se on tyypillistä. Se olisi hyvin eksoottista, jos me kaikki pystyttäisiin tunnistamaan sieltä näytteistä. Näin ei juuri koskaan käy.

Ja nyt olet ihan oikeassa siitä, että eihän me voida niitä tieteelle tuntemattomia lajeja tunnistaa, eli me ei voi saada niille nimiä kuin niitä nimiä ei ole vielä olemassa, eikä tällaista treenausaineistoa ole olemassa.

Teemu Roos: Anteeksi kun keskeytän, mutta oletteko te löytänyt sellaisia uusia lajeja, ja saatteko te sitten keksiä niille nimet?

Otso Ovaskainen: No me ei olla tehty sitä, koska siis se mitä me tehdään on se, että me näillä tilastomenetelmillä tai koneoppimismenetelmillä voidaan antaa todennäköisyys sille, että tämä laji on tieteelle tuntematon, eli sitä ei löydy niistä referenssitietokannoista. Ja sitten me voidaan usein sijoittaa se uusi tieteelle tuntematon laji johonkin kohtaan sitä tunnettua fylogeniaa, eli me voidaan ehkä sanoa, että se kuuluu tähän lahkoon tai se kuuluu tähän sukuun.

Mutta siis se itse laji on tuntematon. Mutta sitten taas se itse lajien nimeäminen, mikä on sitten taas niinku taksonomien leipätyötä, niin se taas vaatii sitten sen kyseisen lajin paljon syvällisempää tutkimusta ennen kuin sille voidaan antaa nimi. Eli ne nimet, joita me annetaan on enemmän tämmöisiä koodeja esimerkiksi, eli niin sanottuja binejä. Barcode index number. Eli ne on tällaisia työnimiä niille lajeille ennen kuin joku taksonomi sitten varsinaisesti kuvaa ne tieteelle.

Teemu Roos: Ah, olisi ollut niin houkuttelevaa, että olisi voinut itse keksiä nimen jollekin koppakuoriaiselle tai edes sienelle.

For brilliant minds – For a Better Future!

Teemu Roos: Kenen tehtävä on rakentaa sitä siltaa juuri tällaisten sovellettavien hankkeiden ja teknologian välille. Ja miten se teidän yhteispeli vaikka tässä hankkeessa on toiminut?

Hanna Koivula: CSC:n rooli on se, että me tarjotaan tekninen mahdollisuus tehdä näitä asioita. Se, että se data kulkee, sitä voidaan paketoida ja muokata sellaiseen muotoon, että sitä voidaan sitten analysoida koneoppimisen ja mallintamisen avulla. Ja se kapasiteetti, jolla se tehdään. Ja sitten se, mikä jäi mainitsematta, on se, että tämä aineisto on tarkoitus viedä raakadatana pitkäaikaissäilytykseen, joka sitten mahdollistaa ehkä tulevalle koneoppimiselle jotakin, mitä vielä ei osata tehdä. Ja se on myös yksi CSC:n palveluista, mikä on mun mielestä aika tärkeä. Opetusministeriö arvottaa tutkimusaineistoja ja maksaa hankkeelle tämän pitkäaikaissäilytyksen, ja se on oikeasti todella arvokas asia.

Teemu Roos: Siitä ei varmaan tule ajatelleeksi, mutta jos ajattelee sitä toisaalta niin päin, että jos katsottaisiin vaikka tästä vaikka 20-30 vuotta taaksepäin ja mitä aineistoja silloin on käsitelty ja tallennettu jonnekin, niin itseasiassa niistä varmaan aika pieni osa olisi nykyään hyödynnettävissä, koska ne on jossain sellaisissa formaateissa, jotain korppuja ja lerppuja tai jotain magneettinauhoja, joiden kuunteluun ei sitten muutaman kymmenen vuoden päästä välttämättä olekaan laitteistoa.

Teemu Roos: Esimerkiksi LIFEPLAN-aineistojen pitkäaikaissäilytys… mitä pitkäaikaissäilytys tässä käytännössä tarkoittaa?

Hanna Koivula: Siis sen palvelun tarkoitus on oikeasti tyyliin 100 vuoden mittakaavassa, eli siinä huolehditaan siitä, että ne on sellaisissa formaateissa, että ne ei korruptoidu, kun ne viedään eteenpäin ja että ne on luettavissa vielä 100 vuodenkin päästä.

Sitä on kulttuuripuolella ehkä tehty enemmän, mutta luonnontieteen puolella ei niinkään.

Meidän se perinteinen pitkäaikaissäilytyshän on museokokoelmat, eli luonnontieteellinen keskusmuseo, jossa on miljoonia ja taas miljoonia hyönteisiä. Tai siis kasveja, hyönteisiä, eläimiä ja kaikkia tieteellisiä näytteitä.

Teemu Roos: Siis niitä itse eläimiä?

Hanna Koivula: Itse eläimiä ja niistä voidaan toki sitten ottaa DNA:ta ja valokuvia, ja tunnetuille lajeille ne museonäytteet on se niin sanottu prototyyppi, joka on säilytetty, että tämä on nyt tämän lajin prototyyppi, että tästä katsomme mitkä muut ovat tätä samaa lajia.

Mutta se, että saataisiin tätä dataa säilymään digitaalisessa muodossa 100 vuotta eteenpäin tai 200 vuotta, niin se sellainen palvelu, että toivottavasti sit myöhemmillä menetelmillä osataan hyödyntää sitä sitten. Toivottavasti se tulee olemaan kultakaivos.

Teemu Roos: Kiehtovaa. Tulee mieleen tällainen vertaus niinku arkki, että sinne arkkiin otetaan sitten ainakin 2 kappaletta jokaista lajia ja nyt ne sitten oikeasti siirtyy ajassa todella todella kauas.

Bess Hardwick: Mä voisin ehkä sanoa vielä tuosta yhteistyöstä CSC:n kanssa sen verran, että heti kun tämä projekti sai rahoituksen, niin oltiin heti alkuun yhteydessä CSC:hen. Ne keskeiset haasteet oli juuri tämä valtavan datamäärän säilyttäminen ja sen siirtäminen joka puolelta maailmaa, ja sitten tää metadatan mukana pitäminen, että miten jokaisen tiedoston pysyy mukana se tieto, että mistä ja milloin se on kerätty. Ollaan monta etäkokousta pidetty CSC:n kanssa, missä näitä on ihan alusta asti mietitty ja keksitty yhdessä ratkaisut siihen. Ja myös luonnontieteellisen keskusmuseo Luomuksen ICT-tiimi on ollut mukana siinä kehittämässä tätä. Että alusta piti miettiä koko homma kyllä.

Teemu Roos: Alustahan se pitää miettiä, eikö niin? Olen kuullut liian monta tarinaa siitä, että on ollut joitain isojakin aineistoa, joita on kerätty ja sitten kun niitä on yritetty lopulta hyödyntää, niin on huomattu, että meidän olisi kannattanut tämäkin muuttuja lisätä sinne tai tämäkin asia kirjata muistiin, ja sitten se on jo myöhäistä, kun sitä dataa on kerätty jo valtavat määrät. Eli se on luonnollisesti hyvä miettiä etukäteen, niin tulee ehkä vähemmän virheitä.

Bess Hardwick: Joo. Meillä on automatisoitu sillä tavalla, että meillä on QR-koodit joka ikisessä nauhurissa, samplerissa, muistikortissa ja pullossa ja purkissa, ja tiimit sitten vaan mobiilisovelluksella skannaa ne QR-koodit keräten näin koko ajan metadatan automaattisesti, niin siten ei tarvitse minkään maastomuistiinpanovihkojen kanssa pelata.

Teemu Roos: No jos palataan vielä tämän hankkeen merkitykseen ja tärkeyteen, niin mitä tästä voidaan tulevaisuudessa oppia? Miten tätä voidaan hyödyntää? Onko se vaan niin kuin tämmöistä ihan perus tutkimuksellista, että me ymmärretään miten elävä luonto muuttuu, vai tuleeko mieleen jotain sellaisia konkreettisempia sovelluksia, miten tätä voidaan hyödyntää nyt, lähitulevaisuudessa tai vielä myöhemmin?

Otso Ovaskainen: No tässä hankkeessa on tosiaan ikään kuin kaksi kärkeä, että tässä pyritään oppimaan jotain luonnosta, mutta tässä sitten myöskin yhtä lailla pyritään kehittämään parempia koneoppimisen ja tilastotieteen menetelmiä. Tämä biodiversiteettiaineisto tuottaa sellaisia haasteita koneoppimisen ja tilastotieteen menetelmille, joka vaatii uutta menetelmäkehitystä. Näille uusille tilastotieteen ja koneoppimisen menetelmille, joita kehitetään, on varmasti käyttöä sitten myös tämän luonnon monimuotoisuuden ulkopuolella. Meillä on tässä tosiaan mukana yksi tilastotieteilijä, joka ei ole pääosin tämmöinen luonnontieteilijä, vaan hän on yhtä lailla tekemisissä teknologian, tähtitieteen ja lääketieteen koneoppimisessa ja tilastomenetelmien kanssa. Sitä kautta sitten näitä menetelmiä, joita me kehitetään yhdessä, voidaan hyödyntää myös sillä puolella, koska nehän ovat ikään kuin yleismenetelmiä.

Sitten kun ajatellaan luonnon monimuotoisuutta, niin ehkä se suurin motivaatio tässä on se perustutkimus, että me tunnetaan maapallon lajiston huonosti ja olisi tosi hienoa, että me tunnettaisiin se paremmin. Ja sitä tässä on lähdetty tekemään ja erityisesti oppimaan niistä prosesseista, että mitkä on ne prosessit siellä luonnon monimuotoisuuden dynamiikan taustalla. Miten lajit vastaa ympäristöolosuhteisiin? Miten ne vuorovaikuttaa keskenään? Miten evoluutio on kaiken tämän takana? Se on se, mistä me ollaan eniten kiinnostuneita.

Mutta ilman muuta tällä on myös kaikenlaista soveltavaa arvoa erityisesti juuri liittyen lajien uhanalaistumiseen - ymmärretään paremmin, miten voidaan suojella lajeja ilmaston- ja maankäytön muutokselta. Mutta myöskin ehkä sitten löydetään sellaisia lajeja, jotka on ihmisen kannalta muutenkin erityisen kiinnostavia tai hyödyllisiä. Esimerkiksi lajeja, jotka vaikuttaa ihmisen terveyteen tai ruoantuotantoon. Ne on kaikki mukana tässä meidän kartoituksessa. Esimerkiksi sienten näytteitähän me otetaan suoraan ilmasta, joista kaikki sen alueen sienet tunnistetaan DNA:n avulla, ja se on erittäin tehokas ja uusi menetelmä käytännössä reaaliajassa monitoroida alueiden sienipopulaatioita. Siellä on mukana sitten ne kaikki patogeenit ja sienet, jotka vaikuttavat ihmisen talouteen ja terveyteen joko positiivisesti tai negatiivisesti, että sieltä voi löytyä kaikennäköisiä sovelluksia.

Teemu Roos: Onko home sieni?

Otso Ovaskainen: Home on sieni ja jäkälä on sieni näiden tutumpien tattien ja kantarellien lisäksi. Sieniä on ehkä 10 miljoonaa lajia maailmassa ja niistä tällä hetkellä tosiaan vain pieni osa tunnetaan. Mutta ne on meille elintärkeitä. Esimerkiksi hiilen kierrossa sienillä on tosi keskeinen rooli ja sitä kautta ilmastonmuutoksesta ja sen hillinnässä.

For brilliant minds – For a Better Future!

Teemu Roos: Mä oon itsekin koneoppimistutkija ja kiehtoo tämä, että miten muilla tieteenaloilla voidaan hyödyntää näitä menetelmiä. Minkä tyyppisistä menetelmistä me nyt puhutaan? Puhutaanko me jostain tämmöisistä taksonomia- tai fylogenetiikka-algoritmeista, jotka luo tällaisia eliöiden sukupuita vai puhutaanko me jostain hahmontunnistusalgoritmeista? Voitko sä avata jotain mitä siellä on tutkimuspuolella sillä osastolla menossa?

Otso Ovaskainen: No jos me puhutaan vaikka siitä, että miten linnut tunnistetaan noista ääninäytteistä, mitä mä äskenkin kuultiin, niin silloin me puhutaan aika perinteistä neuroverkoista. Silloin ehkä se suurin haaste on se, että mikä on se treenausdata eli se aineisto, jolla me opetetaan se kone tunnistamaan ne linnut. Ja tässä itse asiassa voisin mainita, että meillä on aika suuri kansalaistiedehanke tässä mukana, eli

sen lisäksi että ne hyödynnetään noita olemassa olevia maailmanlaajuisia referenssitietokantoja, jossa voi olla siis esimerkiksi nauhoitus, jossa peippo laulaa ja tämän nauhoituksen avulla voidaan opettaa tietokoneita tunnistamaan peippoa, niin me myös sitten itse hyödynnetään maailman lintuharrastajaverkostoja sillä tavalla, että luonnontieteellisen keskusmuseon kanssa me ollaan jo luotu tämmöisiä webbisivuja, joihin nämä harrastajat voi kirjautua sisään ja tuottaa meille treenausdataa esimerkiksi kuuntelemalla 10 sekunnin äänipätkiä ja kertomalla, että mitkä linnut siellä laulaa.

Tällaisen ikään kuin kansalaistieteen avulla tuotetun aineiston avulla me ollaan jo opetettu malli tunnistamaan 100 suomalaista lintua ja todettu, että näin opettaen se toimii paremmin kuin vaikka esimerkiksi BirdNET-sovellus, joka on kenen tahansa kännykkään ladattavissa ja joka ei ole siis meidän tekemä vaan aikaisemmin olemassa oleva vastaava sovellus.

Mutta se, miksi me nyt kehitetään uusia sovelluksia itse on se, että tämä BirDNET esimerkiksi ei toimi riittävän hyvin tämmöiseen ikään kuin passiivisesti kerättyyn aineistoon, jossa mikrofonia ei suoraan suunnata siihen linnun suuntaan, vaan se vaan äänittää kaiken siitä maisemasta, niin ollaan todettu, että se ei riittävän hyvin tunnista niitä lintuja ja siksi me tarvitaan tällaista parempaa itsetehtyä opetusdataa, jolla ihan vastaavanlaisia neuroverkkomalleja opetetaan.

Teemu Roos: Muistan, että olen itsekin juuri tätä BirdNET:iä taisi käyttänyt. Tunnistin jotain lintuja sieltä, jotka suurin piirtein kyllä itsekin arvasin, että mikä mikä lintu siellä on nyt sitten laulussa. Taisi olla juuri mustarastas. Halusin vaan kokeilla että toimiiko, ja ihan kivastihan se toimi. Mutta sitten siellä tuli myös pitkä lista sellaisia lintuja, jotka oli tavallaan “ehkä tällainen ja tällainen lintu”, mutta ne oli nyt selkeästi sitten jotain pingviinejä tai jotain muita, joita ei siinä ympäristössä todennäköisesti ollut paikalla. Ei se nyt ehkä ollut pingviini, mutta joku muu eksoottinen lintu kuitenkin siellä oli.

Mutta okei, eli me voidaan odottaa todennäköisesti, että joku tällainen kuluttajillekin, jokaiselle luonnon seuraajalle tuleva sovellus, jonka avulla voidaan tunnistaa vaikka sitten ääni tai kuvanäytteistä, että mitä lajeja siinä on, niin tämmösiltä sovelluksia nyt ainakin voidaan varmaan sitten odottaa, että ne kehittyy entistä paremmiksi jatkossa tällaisten hankkeiden kautta.

Hanna Koivula: Mä voisin lisätä tuohon vielä sen kansalaistieteen roolista, että sitten kun saadaan tällaisia apuvälineitä harrastajille, niin sitten me voidaan myös valjastaa harrastajat keräämään sitä dataa, koska vaikka sitä tulee ihan valtavasti, niin silti tiedetään ihan liian vähän siitä eliöiden levinneisyydestä nykyään, menneisyydessä ja eritoten tulevaisuudessa. Että sitten kun me saadaan sitä dataa niin sitten voidaan katsoa, että miten esimerkiksi mallinnusten nämä ennustukset käy toteen, ja siinä voidaan hyödyntää sitten laajemmin kansalaisia. Toivottavasti.

Teemu Roos: Onko se niin, että kysellään että “onko näkynyt tällaista lintua?” vai kerätään ihan niin kuin ylipäänsä sitä dataa vaikka sitten lintubongareiden toimesta?

Hanna Koivula: No harrastajille on erilaisia sovelluksia olemassa jo nyt, muun muassa tämmöinen, onkohan se California Universityn kehittämä iNaturalist-sovellus, joka on nyt myös kustomoitu suomeen sopivaksi. Luonnontieteellinen keskusmuseo on sen luonut, tai siis suomentanut. Eli sinne lähinnä voi ottaa sellaisista lajeista valokuvia kännykällä, jolloin siihen kännykän kuvan exif-tiedostoon tallentuu se paikka ja aika, ja sitten siellä on muistaakseni Googlen kuvakirjasto taustalla, eli se osaa ehdottaa silloin lajeja mitä siinä kuvassa voi olla. Ja sitten siinä on myös luotu semmoinen vertaisarviosysteemi, että

muut kokeneemmat harrastajat validoi sen, että jos se ehdottaa jotain niin sitä ei suoraan hyväksytä siksi lajiksi, vaan jonkun täytyy sitten sanoa että se myös on se.

Teemu Roos: Nää on hirveän hyödyllisiä. Me ollaan juuri itse hankittu tällainen niinku kesäpaikka, ja siellä on valtavan hieno puutarha. Mutta itse asiassa siinä kohdassa me ei välttämättä tarvita niin paljon näitä iNaturalist- ja muita sovelluksia, koska siellä on ollut niin ystävällinen puutarhuri, että on kirjoittanut semmoisille lapuille, että mitä kasveja ne on, että me tiedetään sitten ylläpitää niitä tai hoitaa niitä kasveja oikein, mutta varmaan useammassa paikassa ei ole ja se on tietenkin sitten puutarhurillekin tärkeä tieto, että mitä kaikkea siellä kasvaa.

Mut nää on varmasti jokaiselle meistä – ainakin niille, jotka on kiinnostunut luonnosta ja puutarhan hoidosta – hyödyllisiä tuloksia mitä sitten voidaan odottaa tästäkin hankkeesta.

Kuinka pitkään tämä hanke muuten on tarkoitus jatkua? Ei varmaan ihan lyhyen ajan hanke, kun tässä puhutaan näin hitaista prosesseista kuin luonnon monimuotoisuuden muutoksesta. Tai toivottavasti hitaista, ettei tässä nyt kaikki romahda muutamassa vuodessa.

Otso Ovaskainen: Joo, tämä on kuusivuotinen hanke ihan sitä kautta, että nää ERC:n rahoittamat synergyhankkeet on 6 vuotisia. Että silloin 6 vuoden päästä loppuu rahoitus, mut et ilman muuta sitten toivotaan että on joku muu rahoitus, jolla sitten voidaan jatkaa näytteenottoa. Ainakin ehkä vähän pienemmässä mittakaavassa kuin tässä globaalissa.

Teemu Roos: Eli terveisiä tutkimusrahoittajille vaan, että varautukaa 6 vuoden päästä. Kannattaa budjetoida TKI-määrärahoja riittävissä määrin.

Hanna Koivula: Ja myös meille CSC:lle siihen datan säilyttämiseen.

Teemu Roos: Hyvä lisäys.

Bess Hardwick: Mulla tuli mieleen muuten yksi semmoinen ihan vinkeä tulevaisuusvisio vielä liittyen tohon DNA-viivakoodaukseen sekä sekvensointiin. Paul Hebert, joka on DNA-viivakoodauksen semmoinen isä, jonka labrassa meidänkin näytteet sekvensoidaan, niin hänellä on ollut pitkään semmoinen visio, että joku päivä meillä on semmoinen taskuun mahtuva bioskanneri, jonka kanssa voit mennä metsään ja sitten vaikka jonkun lehden nypätä puusta ja syöttää sen sinne ja se DNA-viivakoodaa sen ja tunnistaa sulle sen lajin. Ehkä jos tämä joskus tulee, niin mitä enemmän näitä kirjastoja DNA-viivakoodeista kerätään, niin sitä paremmaksi kehittyy sekin.

Teemu Roos: Joo, ja tämähän on varmaan ollutkin jo pitkään tämmöinen visio ja trendi, että toi DNA-sekvensointi miniaturisoituu, ja ne alkaa olla jo ilmeisesti sellaisia taskuun mahtuvia ne pienimmät, jotka ei ehkä ole maailman tarkimpia, mutta tämmöiseen tarkoitukseen varmaan riittäviä.

Sitten ei tarvitsisi myöskään ottaa turhia riskejä niiden sienien kanssa, että voisi sillä skannerilla tsekata, että mikä laji tämä nyt sitten varmuudella on.

Otso Ovaskainen: Sitä varten on tosi keskeistä, että CSC on tässä mukana. Tai miksi tämä hanke ei olisi mahdollista, jos CSC ei olisi tässä mukana, niin siinä on nämä kaksi puolta, että dataa on paljon ja sitten analyysit on mutkikkaita. Että ihan tuosta datan määrästä, että kaikillehan on tuttu esimerkiksi megan käsite, että sulla on joku kuva tietokoneella ja se on ehkä pari megaa. Sitten kun mennään kertaluokkaa isompaan, niin tulee giga. Se on vielä tyypillisessä päivittäisessä käytössä monella. Sit mennään

kertaluokkaa isompaan, puhutaan terasta. Sitten se on jo semmonen aika iso kovalevy, että sinulla on siellä jotain teroja. Sitten mennään siitäkin kertaluokkaa isompaan ja sitten tulee peta, niin tässä hankkeessa on tarkoitus siis kerätä muutama peta aineistoa.

Eli ihan tämä datan tallentaminen on jo valtava haaste, se ei onnistuisi meidän omille koneille, kun taas CSC:llä on tähän resursseja ja mahdollisuuksia. On tää Allas niminen iso kovalevy, jos sitä nyt näin yksinkertaistaa, jonne sitten tämmöinen muutaman petan aineisto saadaan tallennettua. Tämä on se yksi asia, mikä ei olisi mahdollista ilman CSC:n yhteistyötä. Ja sitten toinen on se datan käsittely. Osa näistä menetelmistä on laskennallisesti aika haastavia, että vaikka niitä kuinka optimoisi, niin se, että niistä algoritmeista työntää nämä muutama peta aineistoa läpi, niin se kestää pitkän aikaa jos sen tekee läppärillä. Ja sitä varten sitten tarvitaan näitä supertietokoneita. Ja nythän Suomessa on tämä uusi uusi kone LUMI, jota me toivotaan pääsemään hyödyntämään tässä hankkeessa ja sitä kautta tekemään näitä analyyseja enemmän reaaliajassa kuin että odotetaan 5 vuotta, että saadaan tuloksia.

Teemu Roos: Tää LUMI onkin tuttu, siitä ollaan kaikki kovasti innoissaan tekoälypiireissäkin. Tää Allas oli itseasiassa mulle uusi asia. Uima-altaista olen kyllä kuullut ja käynyt uimassa, mutta tämä on ilmeisesti vähän eri tyyppinen allas. Mikä on Allas ja kuinka iso allas se on ja missä missä se on?

Hanna Koivula: No itse asiassa Allas on Data Lake - siis se paikka jossa organisoidaan sitä dataa ja missä ne metatiedot liitetään siihen dataan ja se valmistellaan sekä sinne pitkäaikaissäilytykseen että sitten LUMI:a varten käsiteltäväksi. Ja sinne sitä ollaan nyt luotu putki sisään, ja nyt yritetään luoda sitten putkea sinne pitkäaikaissäilytyksen suuntaan, ja kun aika on niin sitten sinne LUMI:in.

Altaan teknisistä tiedoista eli se on tämmöinen CEPH objektitallennustekniikkaan perustuva paikka, “datajärvi”, ja se tarjoaa 12 petatavua tallennuskapasiteettia. Eli tilaa löytyy muillekin kuin LIFEPLAN-hankkeelle. Voin vielä tietysti revitellä, että sitten kun se pitkäaikaissäilytetään, niin sittenhän siitä 2 petasta tehdään vielä varmuuskopioita.

Teemu Roos: Aa, totta kai. Totta kai, että jos joku vahingossa painaisi jotain delete-nappia, niin ei sitten mene koko hanke sinne roskikseen. Jep, ihan hyvä idis.

Bess Hardwick: Ja siinä Altaassa on sitten vielä se kiva, että paitsi että sinne mahtuu tosi paljon, niin sitten myös kun sitä dataa käsitellään jossain Rahdissa, Puhdissa tai Mahdissa, niin sieltä Altaasta pystyy hakemaan sitä dataa hyvin laskentaan ja se liikkuu sieltä joustavasti, niin se ei ole pelkästään semmoisessa syväjäässä vaan että se on aktiivisessa käytössä myös sieltä käsin.

Teemu Roos: Allas ei ole syväjäässä. Se on ihan hyvä.

Otso Ovaskainen: Sitten ehkä yksi toinen asia, joka haluaisin tästä CSC yhteistyöstä nostaa esiin niin sehän ei niinku toimi näin, että CSC sanoo, että “Hei, tässä on teille Altaat ja supertietokoneet, käyttäkää näitä”, vaan me ollaan itse oltu tosi iloisesti yllättyneitä siitä, että miten paljon syvempää tämä yhteistyö on kuin pelkästään noiden fasiliteettien tarjoaminen, koska ne laskentaalgoritmit ja analyysialgoritmithän alun pitäen tulee meiltä. Me kehitetään näitä menetelmiä millä biodiversiteettiä tutkitaan.

Mutta sitten se ei ole todellakaan ihan suoraviivaista, että miten näitä menetelmiä sitten ajetaan läpi noista supertietokoneista, koska niitähän pitää sitten muokata niin, että nää supertietokoneet pystyy mahdollisimman tehokkaasti tekemään sitä laskentaa, koska niissä kuitenkin se laskenta tehdään vähän

eri tavalla, kun sitten vaikka tavallisessa läppärissä. Niin se, mistä me ollaan oltu tosi iloisesti yllättyneitä on se, että CSC:ltä on löytynyt kokonainen tiimi ihmisiä, jotka tavallaan auttaa meitä siirtämään näitä algoritmeja niille CSC:n koneille, joka on tosi iso osa tätä koko hanketta.

Teemu Roos: Ja mä voin kuvitella miten arvokasta toi on. Mulla on itselläkin niinku monta kertaa tullut vastaan tämmöinen… tavallaan se on varmasti meille akateemiselle tutkijoille tuttu tilanne, että meillä on se oma prototyyppi siellä labrassa kehitetty, ja sitten jossain vaiheessa ehkä tietyissä tilanteissa meille tulee sellainen ajatus, että tätä voisi ihan viedä tuotantoon jossain. Joko kaupallistaa tai sit jonkun startupin kautta tai sitten yhteistyössä joidenkin yritysten kautta. Niin se onkin yllättävän vaikeata, jos ei sitä ole tullut hirveän montaa kertaa aikaisemmin tehneeksi.

Että se prototyypistä tuotantoon vieminen, niin se on ihan oma taiteenlajinsa. Sen mäkin oon huomannut muutaman kerran, ja sen mä voin kuvitella, että se on tosi arvokas lisäarvo, mitä nyt tässä tilanteessa sitten esimerkiksi CSC pystyy tarjoamaan, eli ikään kuin akateemisille tutkijoille niiden prototyyppien tuotantoon viemistä ja ylläpitoa ja kaikkea sellaista, mitä me ei ehkä siinä labrassa protoillessa tulla ajatelleeksikaan, että tällaisia toimia täytyy tehdä. Et se on varmasti tosi hyvä yhteistyön lähtökohta, että nää tämmöiset tarpeet otetaan huomioon.

Otso Ovaskainen: Niin, että tässä ehkä vielä ikään kuin kaksi tasoa. Toinen on se, että miten näitä CSC:n fasiliteetteja hyödynnetään meidän omaan käyttöön, että me saadaan tämä meidän LIFEPLAN-data analysoitua. Se on se ensimmäinen taso.

Mut sitten se seuraava taso on se, että no miten sitten kaikki maailman muut biodiversiteettitutkijat, niin miten heillä olisi sitten pääsy näihin menetelmiin. Se on sitten se seuraava taso, joka vielä vie tämän, niin kuin sanoit, prototyypistä sinne tuotantoon. Se on valtava haaste.

For brilliant minds – For a Better Future!

Teemu Roos: Okei, yksi kysymys on vielä. Nyt me ollaan puhuttu tästä pitkäaikaisesta säilytyksestä.

Ja se tavallaan johtaakin tähän tähän kysymykseen, joka me kysytään kaikilta tämän podcastsarjan vierailta liittyen siihen, että tänä vuonna on ilmeisesti CSC:n 50-vuotisbileet. Ja kun CSC selvästi haluaa ajatella tulevaisuutta, niin seuraavaksi on tietenkin mielessä ne seuraavat 50 vuotta.

Eli milloin sitten onkaan… 2071 on satavuotisjuhlat. Ja ollaan tekemässä tällaista aikakapselia, minne tallennetaan muun muassa näitä podcasteja niiden viidenkymmenen vuoden päässä täällä planeetalla tallaavien ihmisten ihmeteltäväksi. Niin meillä olisi nyt tässä tilaisuus lähettää tämän aikakapselin kautta terveisiä sinne viidenkymmenen vuoden päähän. Mitä terveisiä te haluaisitte lähettää sinne viidenkymmenen vuoden päähän?

Bess Hardwick: No mä lähettäisin semmoisia, että kyllä teillä nykynuorilla on helppoa, kun ei tarvitse kun bioskannerilla vähän pyyhkäistä jotain sientä tai puhelimella nauhoittaa vähän jotain lintua, niin heti aivoimplantti kertoo teille saman tien, että mikä peippo se mustarastas nyt oli. Että silloin oli kunnollista kun minä olin nuori. Piti muovihanskoilla kerätä näytteet ja kiikuttaa ne jäissä labraan ja odottaa päiviä ja sitten sieltä tulee vastaus että “tuntematon laji”. Että me vaivalla ollaan kerätty teille kirjastot, että olkaa nyt kiitollisia, että sitten kun oltaisiin vielä saatu säästettyä ne lajit sukupuolta niin se olisi ollut kiva.

Teemu Roos: Okei, no mun mielestä tää on hyvä. Hyvät terveiset, että ennen oli kunnollista ja sitten vielä, että voi kun oltaisiin säästetty sukupuutolta. Vähän pessimististä, mutta…

No niin, ehkä me päätetään näihin tunnelmiin. Terkkuja vaan ja jaksamista sinne tulevaisuuteen.

Tämä podcast tallennetaan CSC:n 50-vuotisjuhlan kunniaksi aikakapseliin, joka avataan seuraavan kerran 100-vuotisjuhlan yhteydessä vuonna 2071. Kiitos kun kuuntelit For a Better Future -podcastia!