Liikuntasovelluksella kerätyn big datan käytön mahdollisuudet ja haasteet

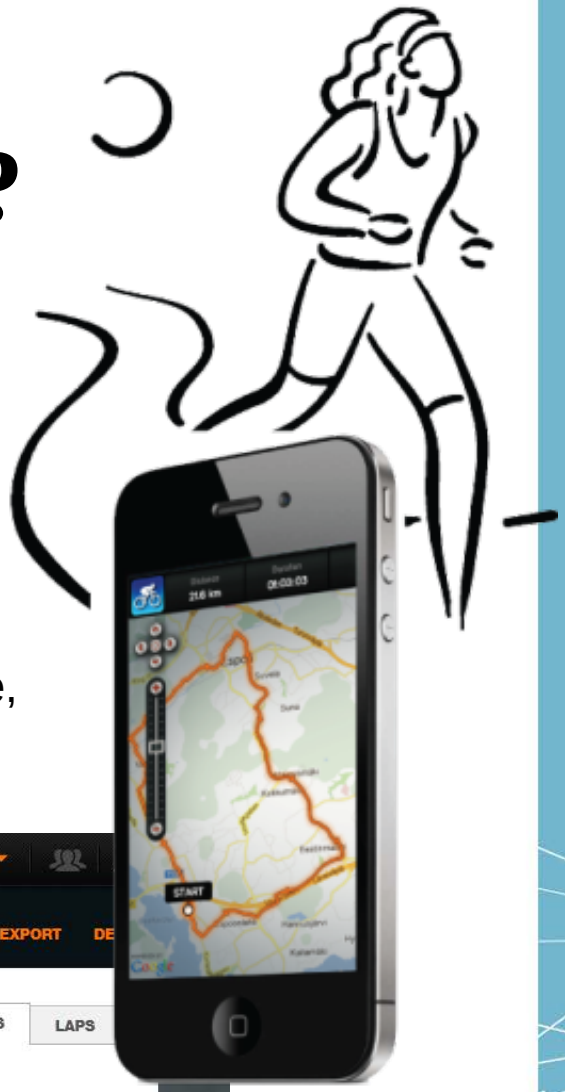
Juha Oksanen

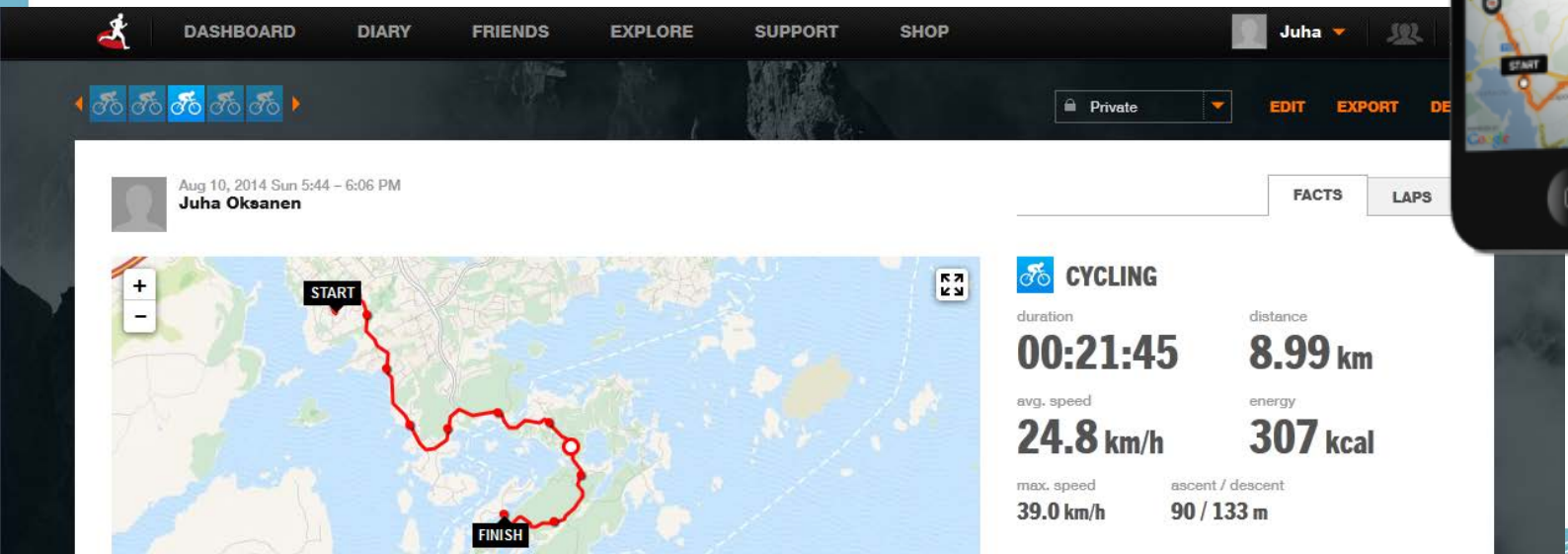
20.5.2015, Geoinformatiikan tutkimuspäivät

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Mikä on Sports Tracker?

- Ilmainen mobiili liikuntasovellus, jolla voidaan pitää harjoituspäiväkirjaa
- Online palvelu *sports-tracker.com*
- Sosiaalinen verkosto
- Aloitti 2004 osana Nokiaa, 2009 itsenäinen yritys Sports Tracking Technologies Oy, 12.5.2015 osa Amer Sports Oyj:tä
- >25 miljoonaa sovelluslatausta kaikille yleisimmille alustoille, >2 miljoona käyttäjää kuukausittain >200 maassa





NLS
FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Miltä Sports Tracker data näyttää?

```
<?xml version="1.0" encoding="utf-8"?><gpx
xsi:schemaLocation="http://www.topografix.com/GPX/1/1
http://www.topografix.com/GPX/1/1/gpx.xsd
http://www.garmin.com/xmlschemas/GpxExtensions/v3
http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd
http://www.garmin.com/xmlschemas/TrackPointExtension/v1
http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd" version="1.1"
creator="Sports Tracker" xmlns="http://www.topografix.com/GPX/1/1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3">
  <metadata>
    <name>Cycling 24/05/2013/15:46:43.18</name>
    <desc></desc>
    <author>
      <name>Juha Oksanen</name>
    </author>
    <link href="www.sports-tracker.com">
      <text>Sports Tracker</text>
    </link>
  </metadata>
  <trk>
    <trkseg>
      <trkpt lat="60.16121833333333" lon="24.5
        <ele>19</ele>
        <time>2013-05-24T15:46:43.18</time>
      </trkpt>
      <trkpt lat="60.16121833333333" lon="24.5458">
        <ele>19</ele>
        <time>2013-05-24T15:46:43.53</time>
      </trkpt>
      <trkpt lat="60.16117166666667" lon="24.545826666666667">…
```

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Miltä Sports Tracker data näyttää?

```
<?xml version="1.0" encoding="utf-8"?><gpx
xsi:schemaLocation="http://www.topografix.com/GPX/1/1
http://www.topografix.com/GPX/1/1/gpx.xsd
http://www.garmin.com/xmlschemas/GpxExtensions/v3
http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd
http://www.garmin.com/xmlschemas/TrackPointExtension/v1
http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd" version="1.1"
creator="Sports Tracker" xmlns="http://www.topografix.com/GPX/1/1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3">
  <metadata>
    <name>Cycling 24/05/2013/15:46:43.18</name>
    <desc></desc>
    <author>
      <name>Juha Oksanen</name>
    </author>
    <link href="www.sp
      <text>Sports Tra
    </link>
  </metadata>
  <trk>
    <trkseg>
      <trkpt lat="60.1
        <ele>19</ele>
        <time>2013-05-24T15:46:43.18</time>
      </trkpt>
      <trkpt lat="60.16121833333333" lon="24.5458">
        <ele>19</ele>
        <time>2013-05-24T15:46:43.53</time>
      </trkpt>
      <trkpt lat="60.16117166666667" lon="24.545826666666667">…
```

- ID/pseudo-ID
- Yksityisyyden taso:
  - yksityinen
  - jaetaan ystäville
  - julkinen (näytetään ST web-sovelluksessa)

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Miltä Sports Tracker data näyttää?

```
<?xml version="1.0" encoding="utf-8"?><gpx
xsi:schemaLocation="http://www.topografix.com/GPX/1/1
http://www.topografix.com/GPX/1/1/gpx.xsd
http://www.garmin.com/xmlschemas/GpxExtensions/v3
http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd
http://www.garmin.com/xmlschemas/TrackPointExtension/v1
http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd" version="1.1"
creator="Sports Tracker" xmlns="http://www.topografix.com/GPX/1/1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3">
  <metadata>
    <name>Cycling 24/05/2013/15:46:43.18</name>
    <desc></desc>
    <author>
      <name>Juha Oksanen</name>
    </author>
    <link href="www.sports-tracker.com">
      <text>Sports Tracker</text>
    </link>
  </metadata>
  <trk>
    <trkseg>
      <trkpt lat="60.16121833333333" lon="24.5458">
        <ele>19</ele>
        <time>2013-05-24T15:46:43.18</time>
      </trkpt>
      <trkpt lat="60.16121833333333" lon="24.5458">
        <ele>19</ele>
        <time>2013-05-24T15:46:43.53</time>
      </trkpt>
      <trkpt lat="60.16117166666667" lon="24.545826666666667">…
```

- Koordinaatit n. 1 s välein

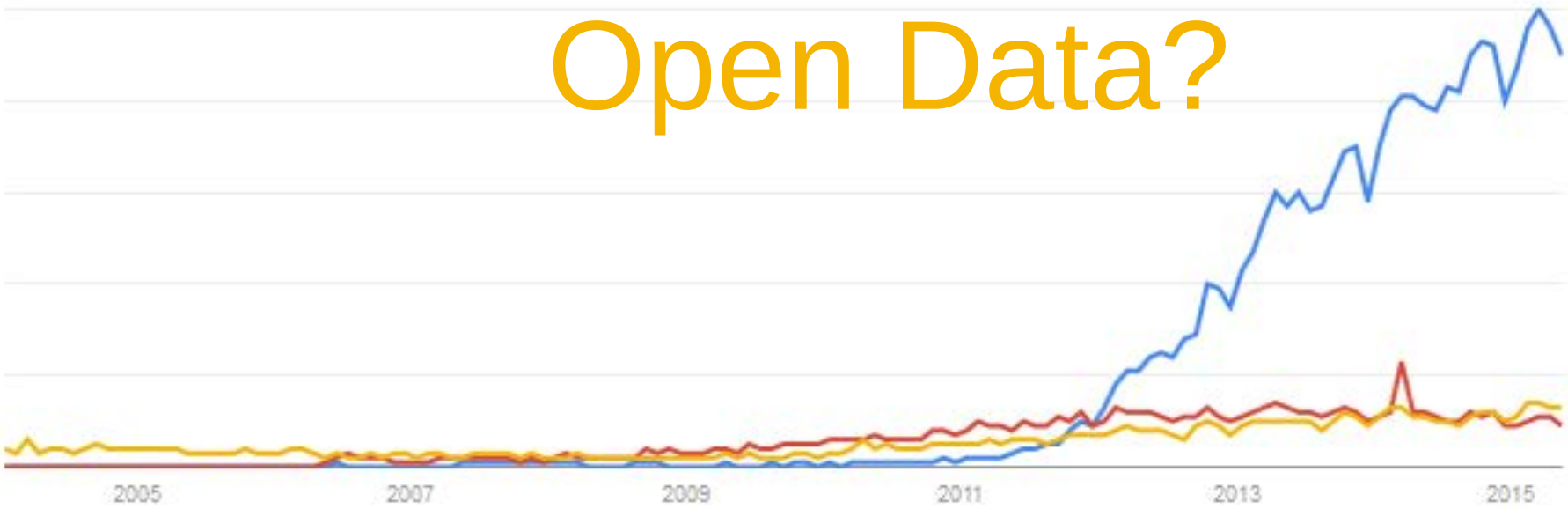
NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Mikä luonnehtii Sports Tracker dataa?

Big Data?

Crowdsourcing/VGI?

Open Data?



Google Trends

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Onko julkinen ST data avointa dataa?

- Helppo vastaus: EI
- Julkinen
  - Henkilökohtaiset suorituksen visualisoidaan ST web-sovelluksessa
  - Dataan ei enää pääsyä, paitsi omaan dataan
- Yksityisyyden suoja on todellinen haaste myös julkisessa ST datassa
  - Paljastaa käyttäytymishahmoja
    - Esim. kodin ja työpaikan sijainti helppo päätellä

Onko julkinen ST data avointa dataa?

- Helppo vastaus: EI
- Julkinen
  - Henkilökohtaiset suorituksen visualisoidaan ST web-sovelluksessa
  - Dataan ei enää pääsyä, paitsi omaan dataan
- Yksityisyyden suoja on todellinen haaste myös julkisessa ST datassa
  - Paljastaa käyttäytymishahmoja
    - Esim. kodin ja työpaikan sijainti helppo päätellä

HAASTE 1: Yksityisyys

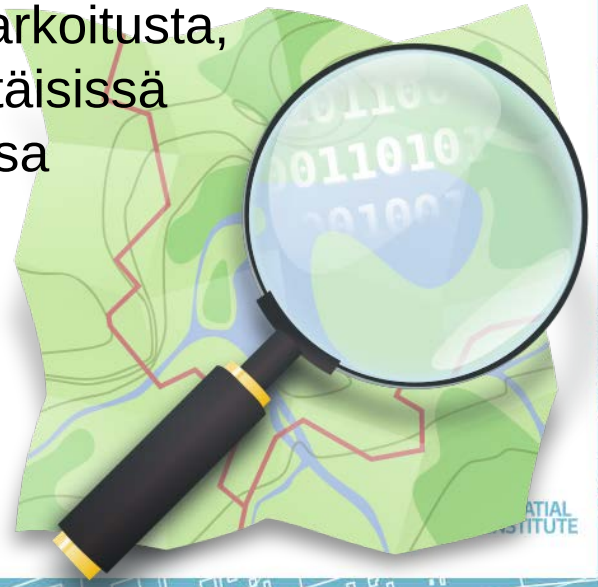
Onko julkinen ST data vapaaehtoisesti kerättyä maantieteellistä tietoa (VGI)?

Kyllä…

- ST käyttäjäyhteisö kerää datan
- Yhteisö maailmanlaajuinen, valtava määrä liikuntasuorituksia
- Ihmiset osallistuvat vapaaehtoisesti, vain ilmainen rekisteröityminen vaaditaan
- Motiivit: sosiaalinen tunnustus / lisääntynyt henkilökohtainen maine, kilpailu?

tai ehkä ei…

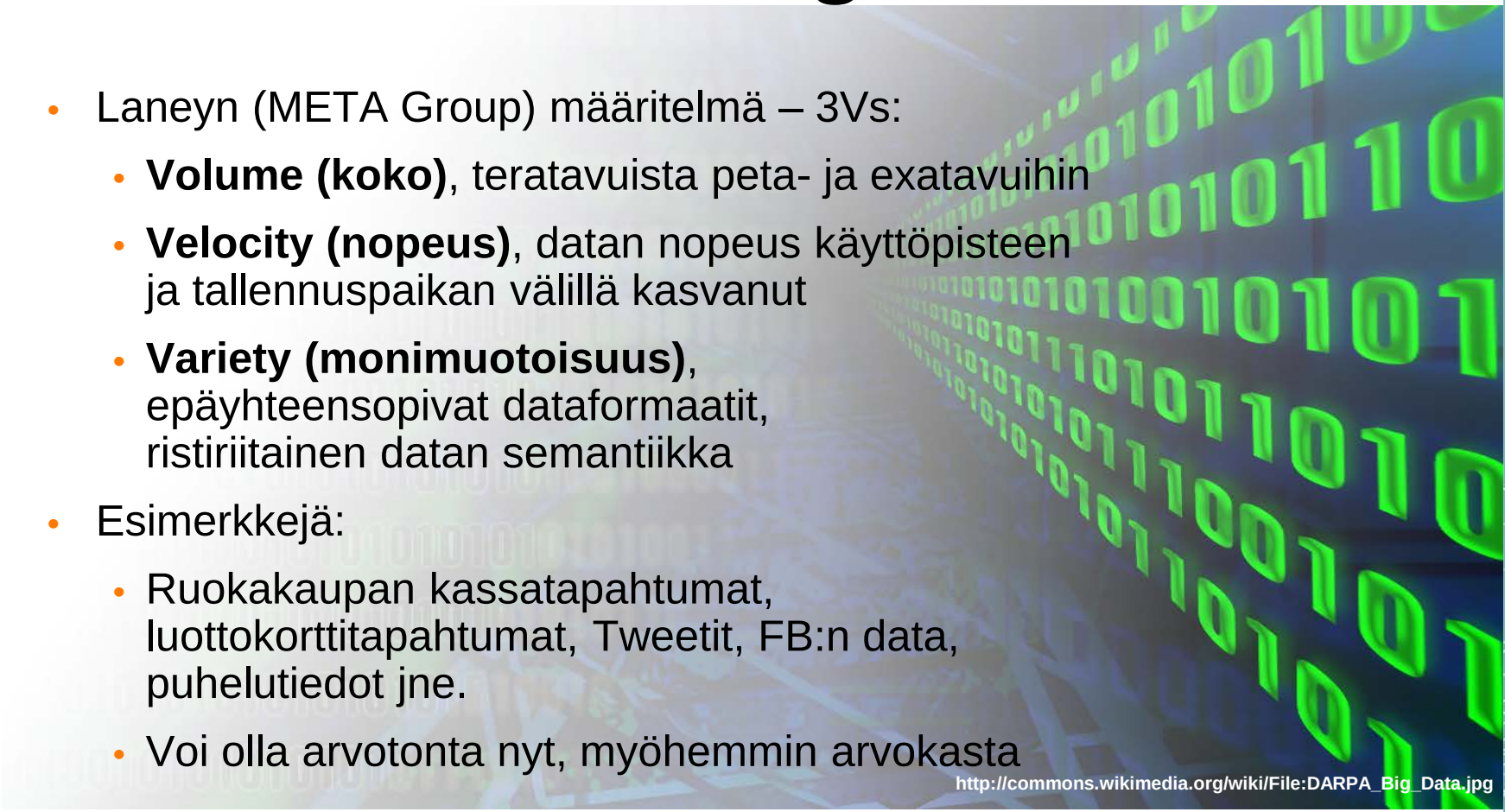
- Ei tavoitteena kerätä maantieteellisiä kohteita
- Yhteinen projekti puuttuu
- Datalla kokonaisuutena ei (vielä) ole käyttötarkoitusta, fokus yksittäisissä suorituksissa



http://commons.wikimedia.org/wiki/File:Openstreetmap_logo.svg

Onko ST data ”Big Dataa”?

- Laneyn (META Group) määritelmä – 3Vs:
  - **Volume (koko)**, teratavuista peta- ja exatavuihin
  - **Velocity (nopeus)**, datan nopeus käyttöpisteen ja tallennuspaikan välillä kasvanut
  - **Variety (monimuotoisuus)**, epäyhteensopivat dataformaatit, ristiriitainen datan semantiikka
- Esimerkkejä:
  - Ruokakaupan kassatapahtumat, luottokorttitapahtumat, Tweetit, FB:n data, puhelutiedot jne.
  - Voi olla arvotonta nyt, myöhemmin arvokasta



http://commons.wikimedia.org/wiki/File:DARPA_Big_Data.jpg

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Onko ST data ”Big Dataa”?

- Tällä hetkellä ST tietokanta sisältää noin 2 miljardia kilometriä liikuntasuorituksia
  - Data on suuri, mutta ei sentään petatavuja…
  - Data on hyvin strukturoitu…
- Toistaiseksi käytetään vain henkilökohtaisessa liikuntapäiväkirjassa tai jaetaan ystävien/kaikkien kesken
- ”Big Data” luonnehtii ST dataa paremmin, kuin VGI tai avoin data

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Onko ST data ”Big Dataa”?

- Tällä hetkellä ST tietokanta sisältää noin 2 miljardia kilometriä liikuntasuorituksia
  - Data on suuri, mutta ei sentään petatavuja…
  - Data on hyvin strukturoitu…
- Toistaiseksi käytetään vain henkilökohtaisessa liikuntapäiväkirjassa tai jaetaan ystävien/kaikkien kesken
- ”Big Data” luonnehtii ST dataa paremmin, kuin VGI tai avoin data

HAASTE 2: Iso data

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Mahdollisuuksia

Sovelluskeskeinen lähestymistapa: 

- Missä ovat parhaat/suosituimmat liikuntareitit?
  - Lämpökartat visuaalisen tiedonlouhinnan tueksi
  - Verkoston linkin suosioon perustuva reititys
  - Tehokas verkkopalvelu vuorovaikutteisten lämpökarttojen tuottamiseksi

Yhteiskuntakeskeinen lähestymistapa:

- Missä, miten, milloin, miksi liikutaan?
- Miten yhdyskuntarakenne/liikuntapaikkatarjonta vaikuttavat ihmisten liikkumisaktiivisuuteen?

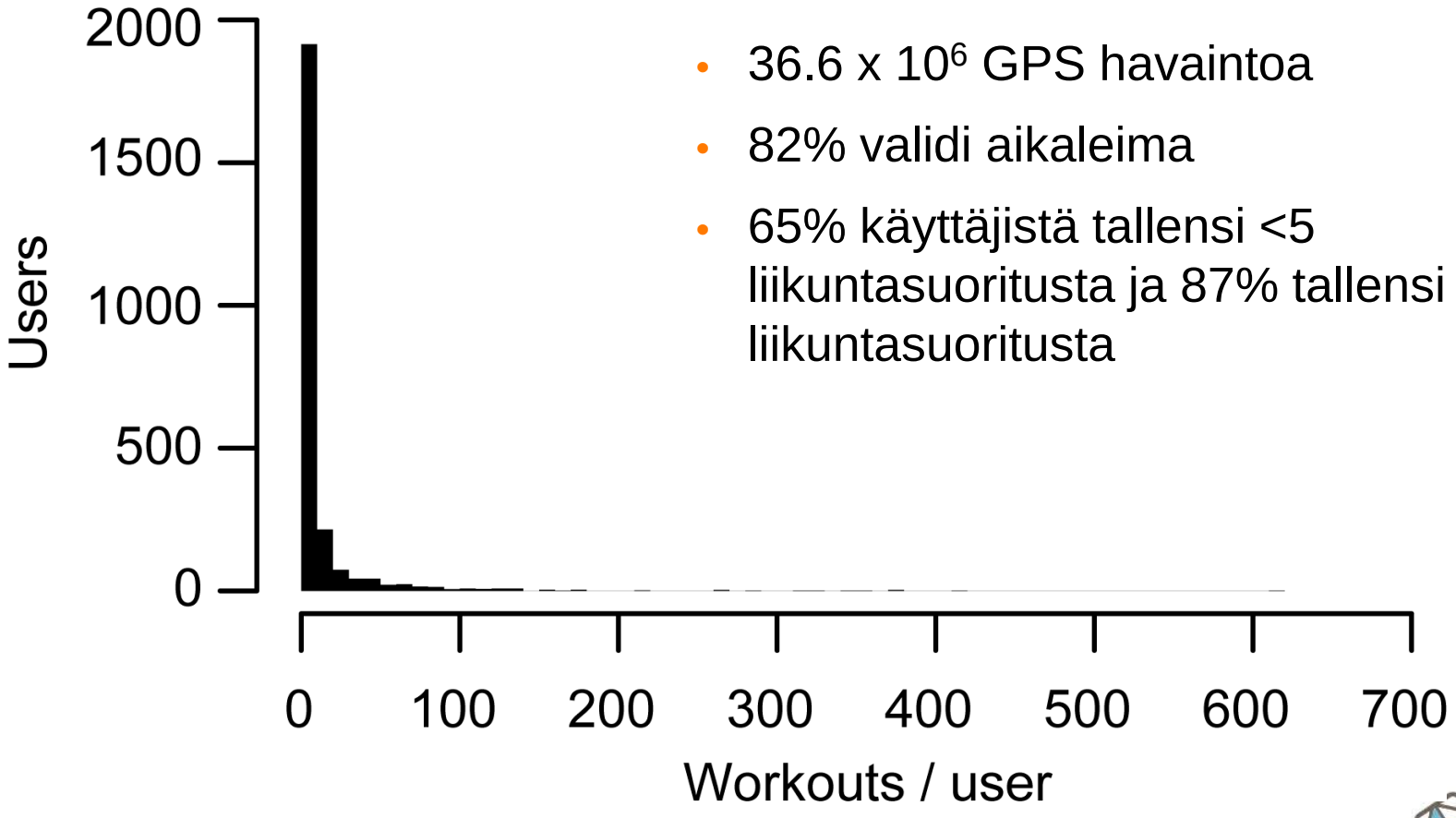
NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Sovelluskeskeinen lähestymistapa
CASE: Pyöräily Helsingissä

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Datan ominaisuuksia

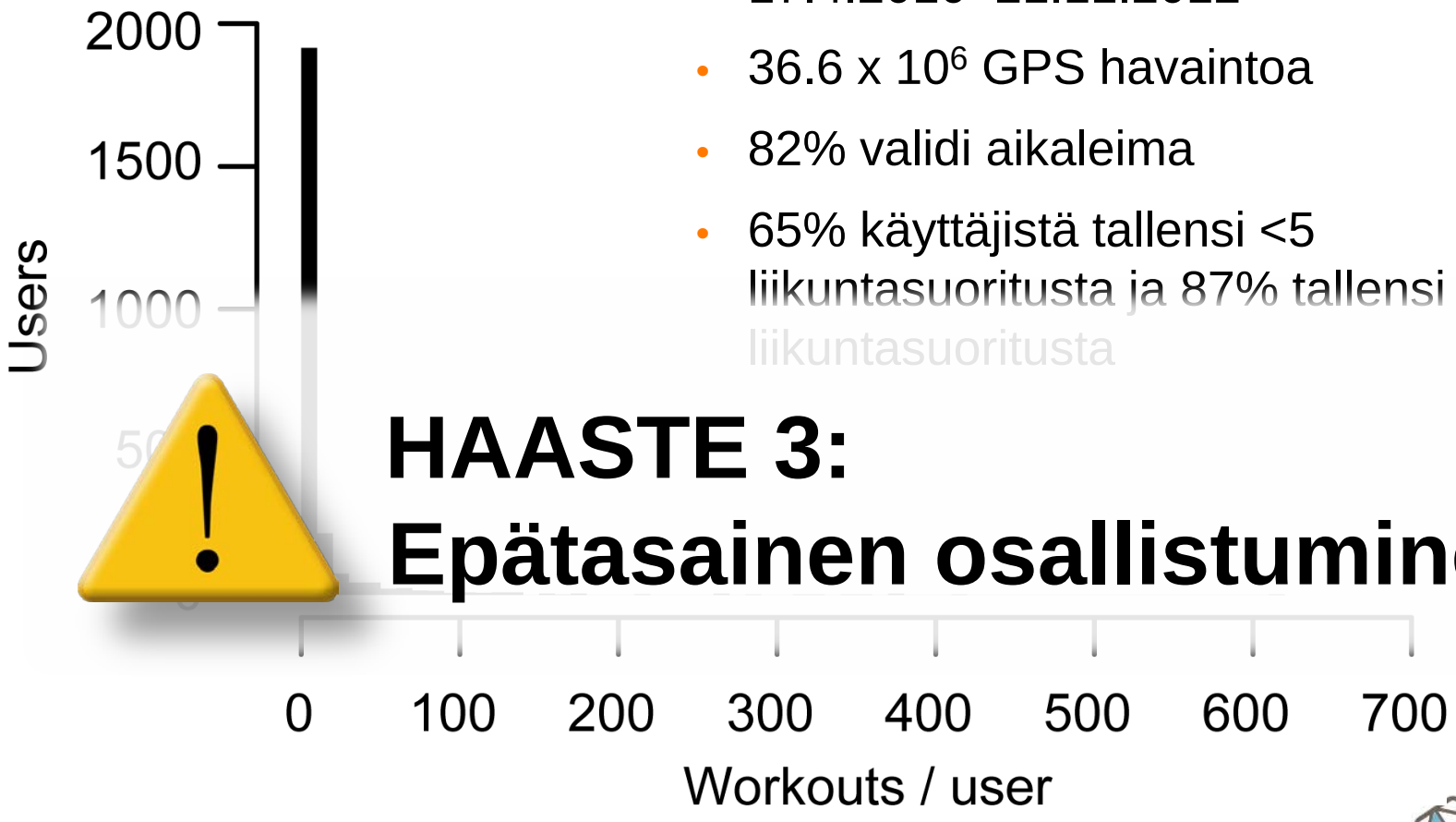
- 36 757 liikuntasuoritusta, 2424 käyttäjää
- 17.4.2010–21.11.2012
- 36.6 x 10^6 GPS havaintoa
- 82% validi aikaleima
- 65% käyttäjistä tallensi <5 liikuntasuoritusta ja 87% tallensi <20 liikuntasuoritusta



NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Datan ominaisuuksia

- 36 757 liikuntasuoritusta, 2424 käyttäjää
- 17.4.2010–21.11.2012
- 36.6 x 10^6 GPS havaintoa
- 82% validi aikaleima
- 65% käyttäjistä tallensi <5 liikuntasuoritusta ja 87% tallensi <20 liikuntasuoritusta



HAASTE 3:
Epätasainen osallistuminen

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Datan ominaisuuksia

http://www.nngroup.com/articles/participation-inequality/

The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities

by **JAKOB NIELSEN** on October 9, 2006

Topics: Social UX

Summary: In most online communities, 90% of users are lurkers who never contribute, 9% of users contribute a little, and 1% of users account for almost all the action.

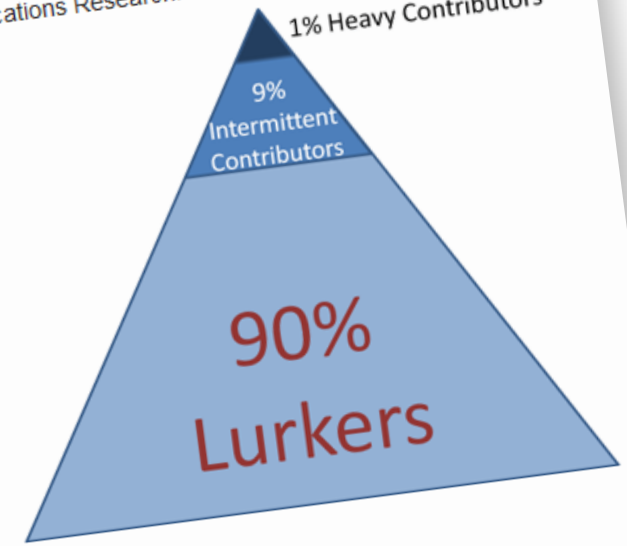
All large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one property: **most users don't participate** very much. Often, they simply **lurk** in the background.

In contrast, a tiny minority of users usually accounts for a disproportionately large amount of the content and other system activity. This phenomenon of **participation inequality** was first studied in depth by Will Hill in the early '90s, when he worked down the hall from me at Bell Communications Research.

When you plot the amount of activity for each user, the result is a Zipf curve, which shows as a straight line in a log-log diagram.

User participation often more or less follows a **90-9-1 rule**:

- **90%** of users are **lurkers** (i.e., read or observe, but don't contribute).
- **9%** of users contribute **from time to time**, but other priorities dominate their time.
- **1%** of users participate a lot and **account for most contributions**: it can seem as if they don't have lives because they often post just minutes after whatever event they're commenting on occurs.



…dia such as Usenet newsgroups, …pics. A study of

4 käyttäjää

<20

Users

2000

1500

1000

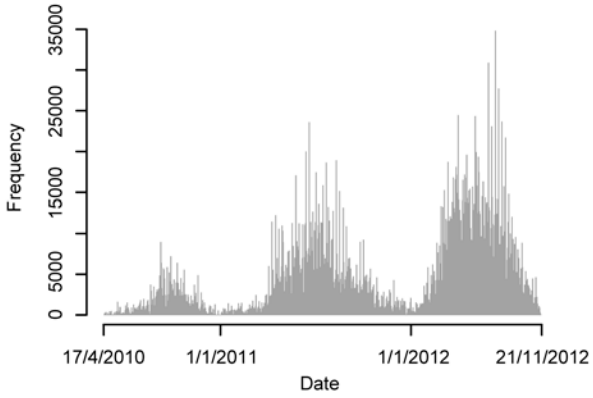
500

0

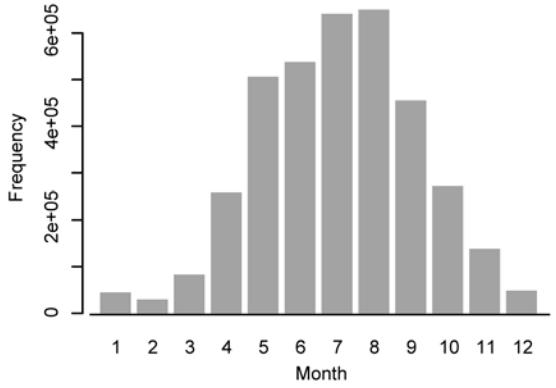
0

Datan syklisyys

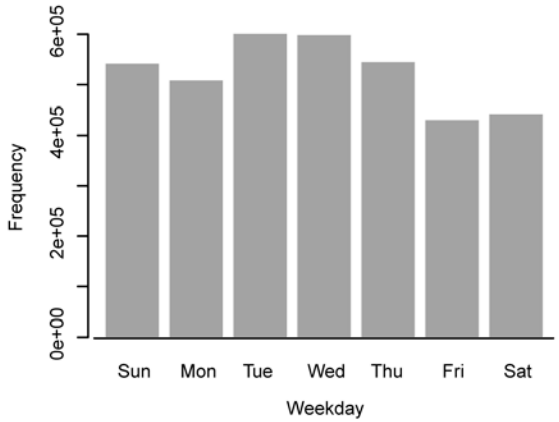
Pisteiden lkm / vuorokausi



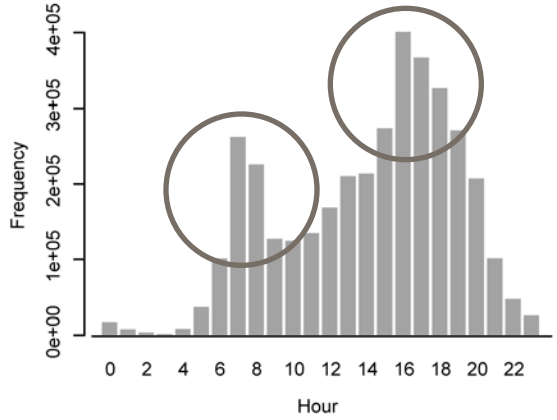
Pisteiden lkm / kuukausi



Pisteiden lkm / viikonpäivä



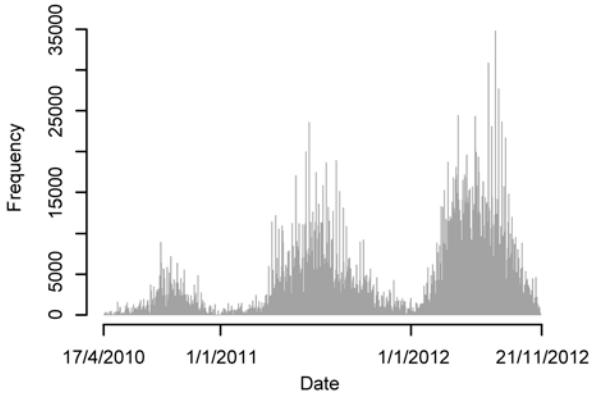
Pisteiden lkm / vuorokauden tunti



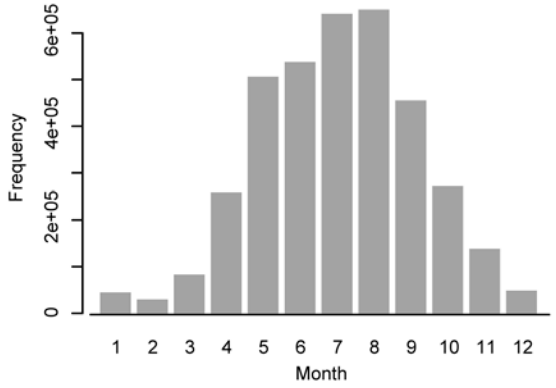
NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Datan syklisyys

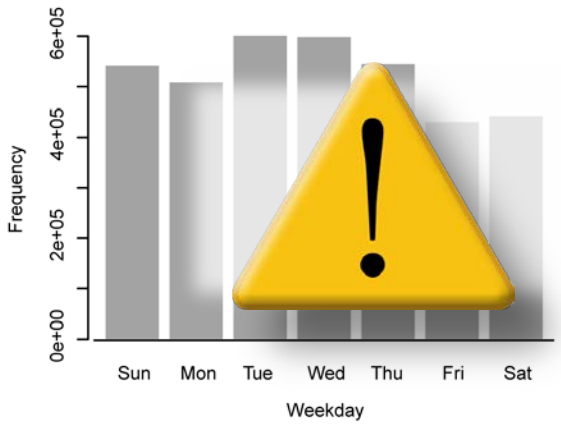
Pisteiden lkm / vuorokausi



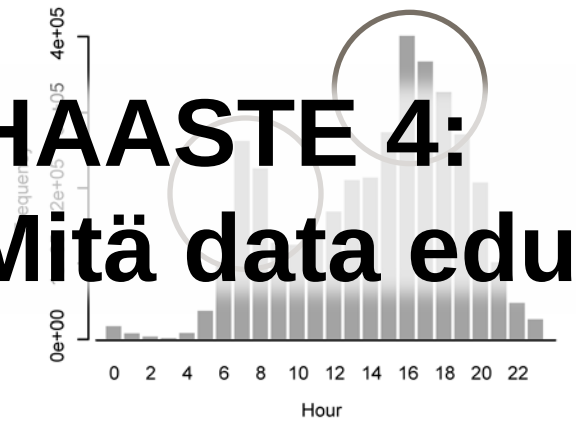
Pisteiden lkm / kuukausi



Pisteiden lkm / viikonpäivä



Pisteiden lkm / vuorokauden tunti



HAASTE 4:
Mitä data edustaa?

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Mikä on suosittu reitti?

1. Liikuntasuoritusten tallentajien lukumäärä on suuri?
2. Tallennettujen liikuntasuoritusten lukumäärä on suuri?
3. Sekä liikuntasuoritusten tallentajien lukumäärä ja tallennettujen liikuntasuoritusten määrä on suuri ja tallentajien määrä on mahdollisimman tasajakautunut?

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Lämpökarttojen laskentamenetelmät

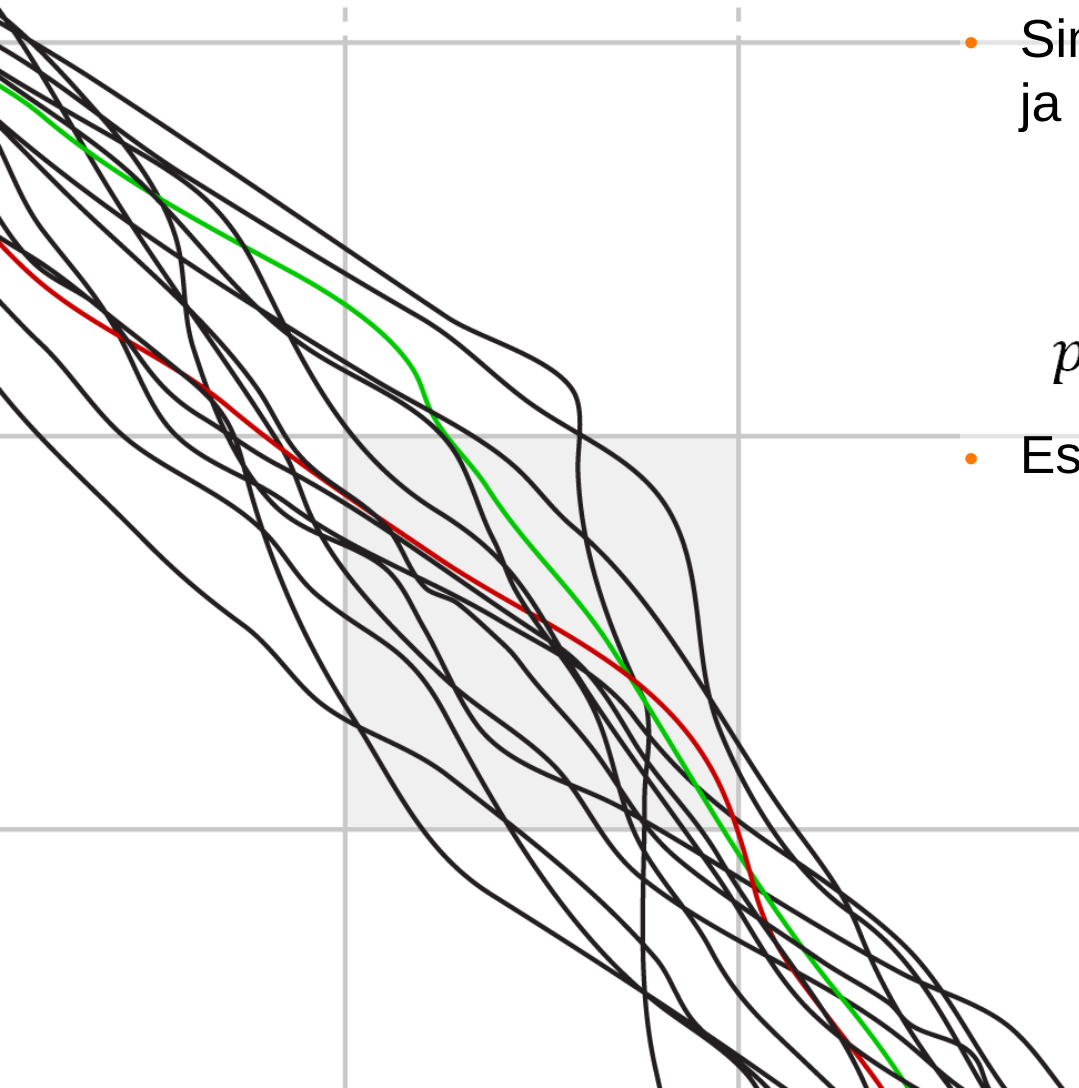
- Esikäsittely: k-anonymiteetti-pohjainen trajektorisegmenttien valinta
- **ppUCC** – privacy-preserving user count calculation
  - Yksityisyyden säilyttävä käyttäjämäärälaskenta
  - Käyttäjien ”2D histogrammi”
- **ppKDE** - privacy-preserving Kernel Density Estimation
  - Yksityisyyden säilyttävä tiheysfunktion ydinestimointimenetelmä
  - Liikuntasuoritustrajektoreiden 2D tiheys:

$$ppKDE(s) = \hat{f}(s) = \frac{1}{nh}\sum_{i=i}^{n} K\left(\frac{s-s_i}{h}\right)$$

- **ppDIV** - privacy-preserving Kernel Density Estimation modified with user diversity index
  - Uusi menetelmä, jossa yhdistyvät trajektoreiden tiheys ja käyttäjien diversiteetti

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Menetelmät: ppDIV



- Simpsonin diversiteetti-indeksi D ja ppKDE:

$$D(s) = 1 - \sum p(s)_i^2$$

$$ppDIV(s) = ppKDE(s) * D(s)$$

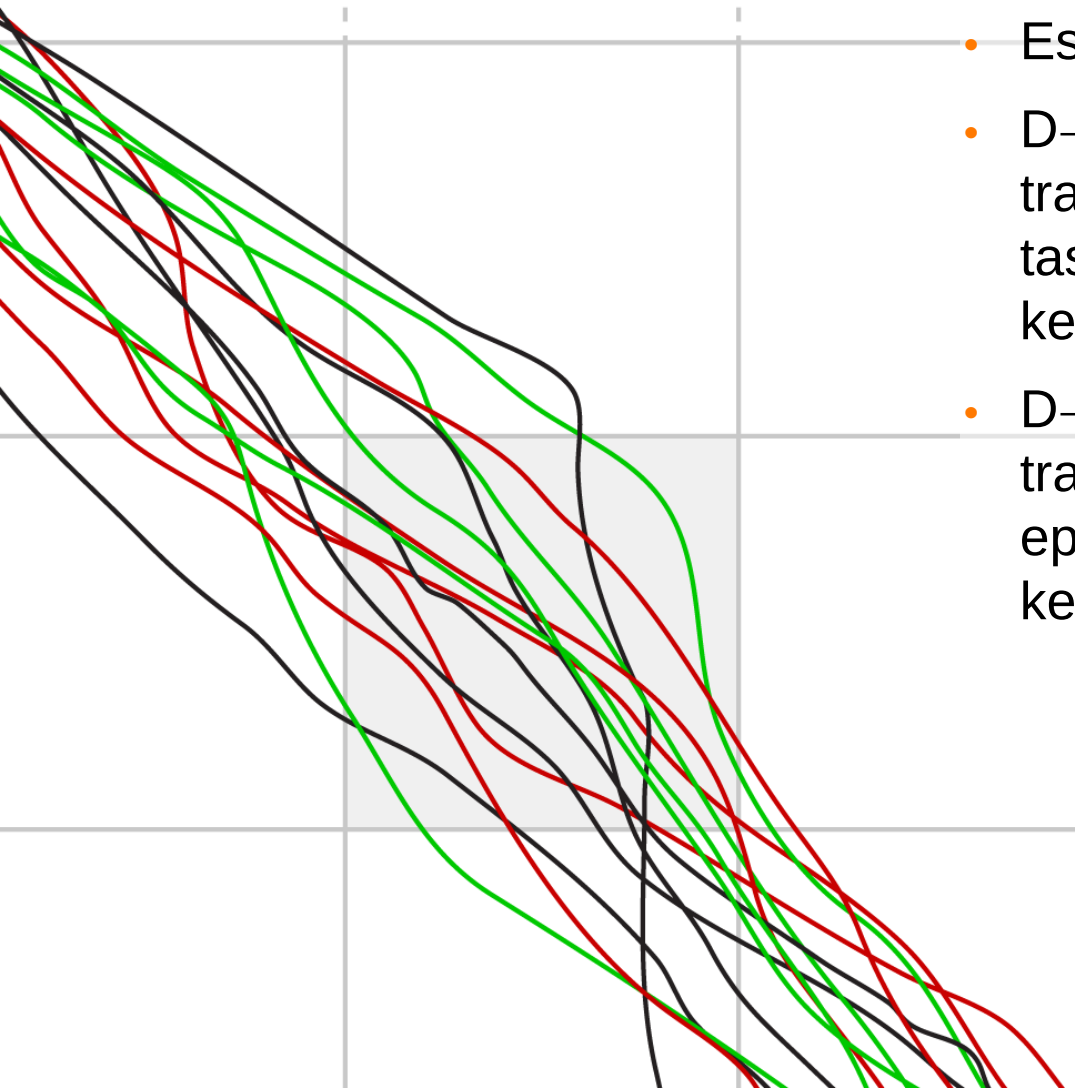
- Esimerkki: D = 0.24

13

1

1

Menetelmät: ppDIV

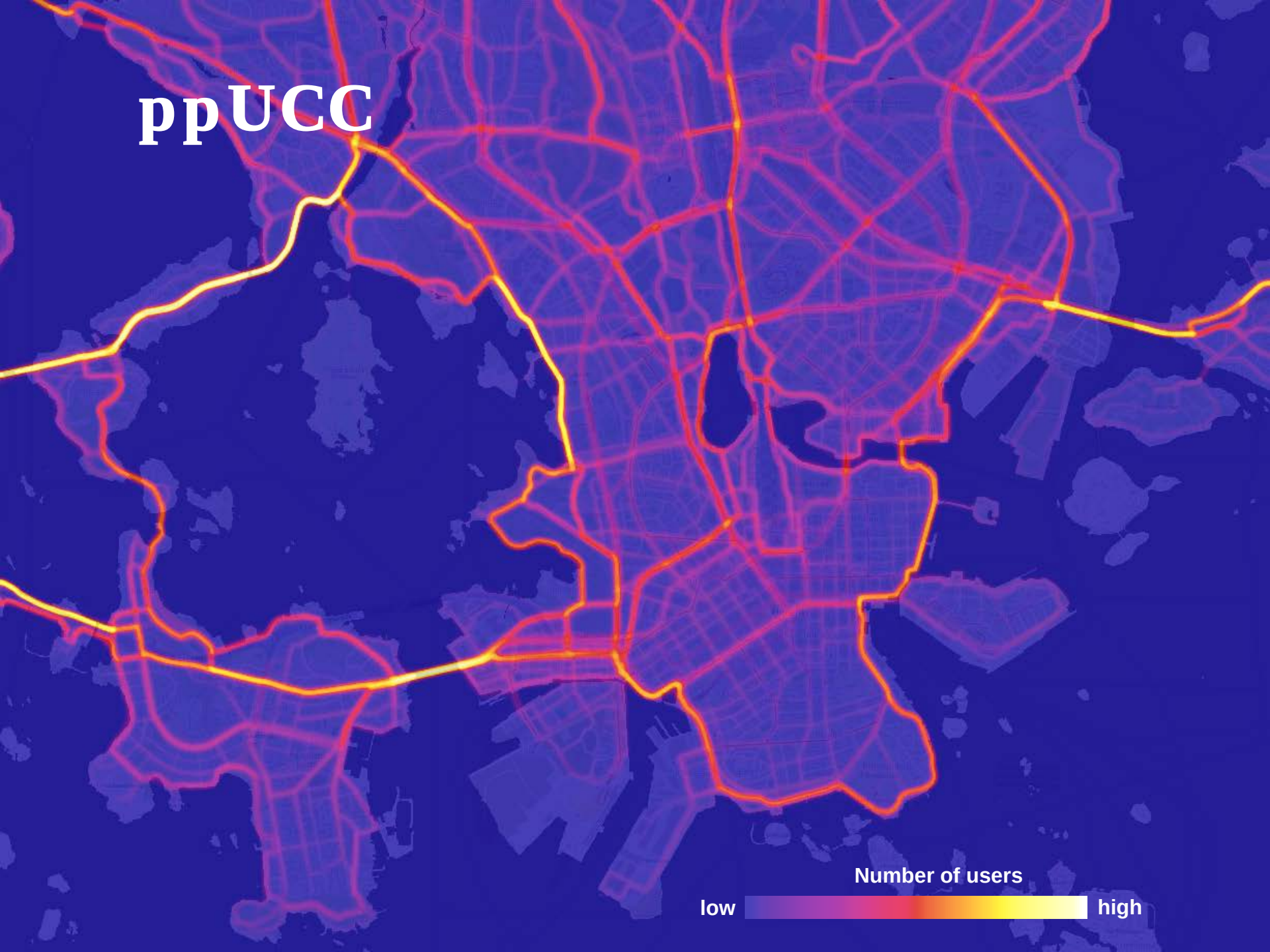


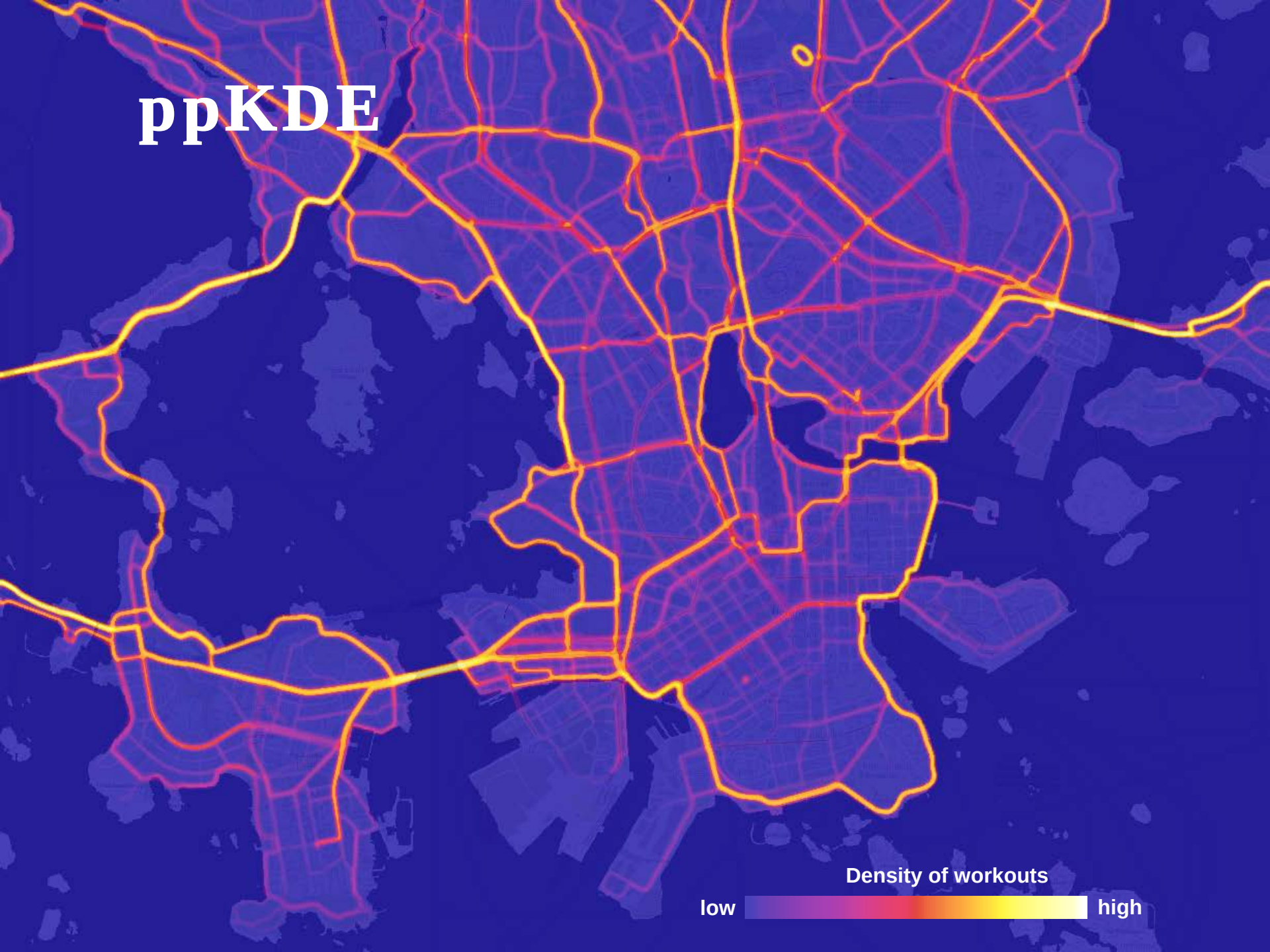
- Esimerkki: D = 0.67
- D→1, kun suuri määrä trajektoreita on jakautunut tasaisesti monen eri käyttäjän kesken
- D→0, kun pieni määrä trajektoreita on jakautunut epätasaisesti harvan käyttäjän kesken

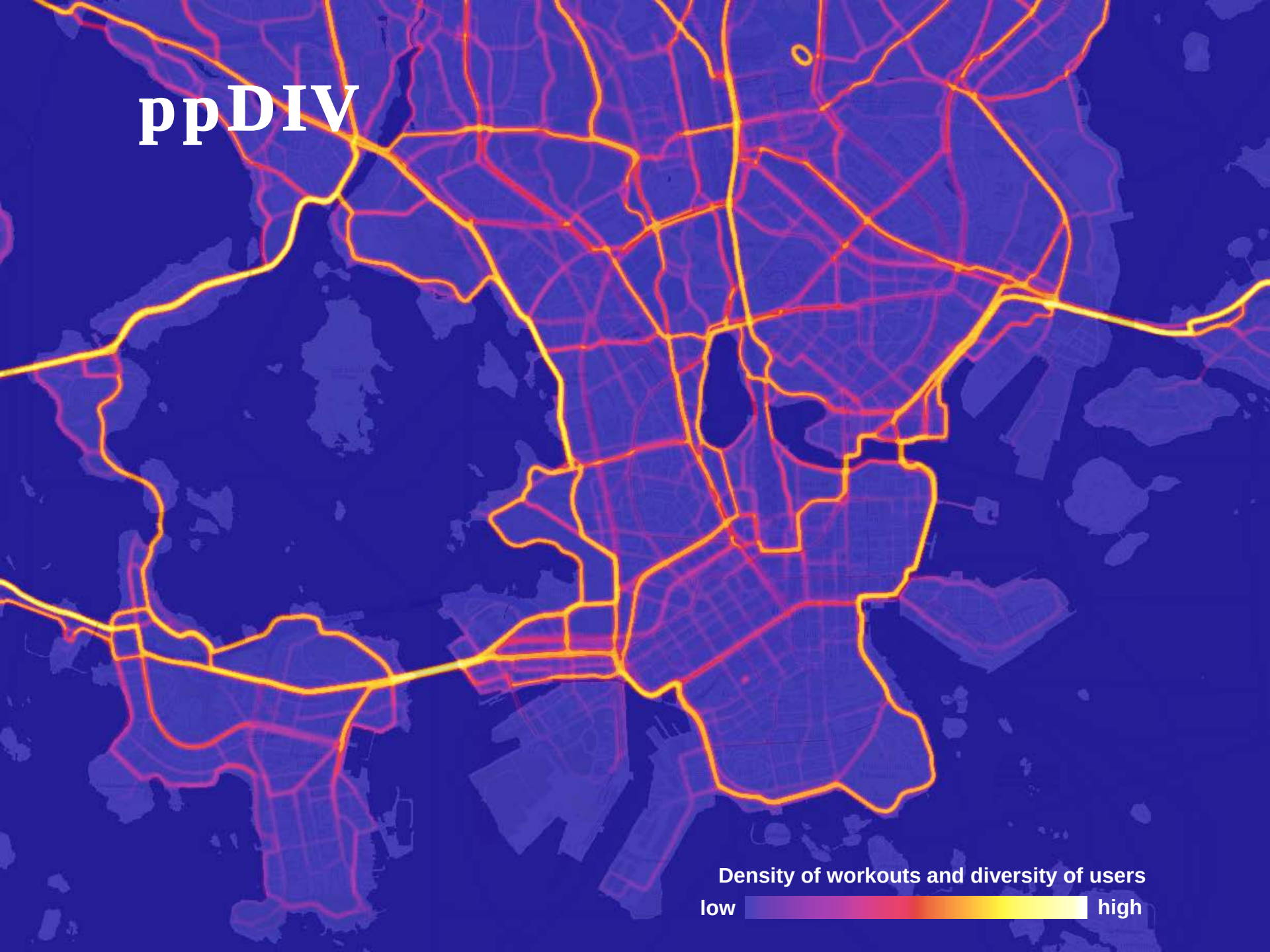
5

5

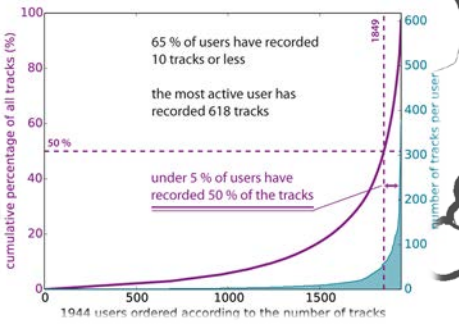
5


ppUCC
Number of users
low
high

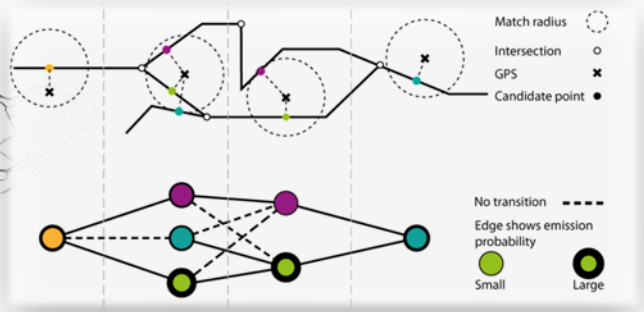

ppKDE
Density of workouts
low
high


ppDIV
Density of workouts and diversity of users
low
high

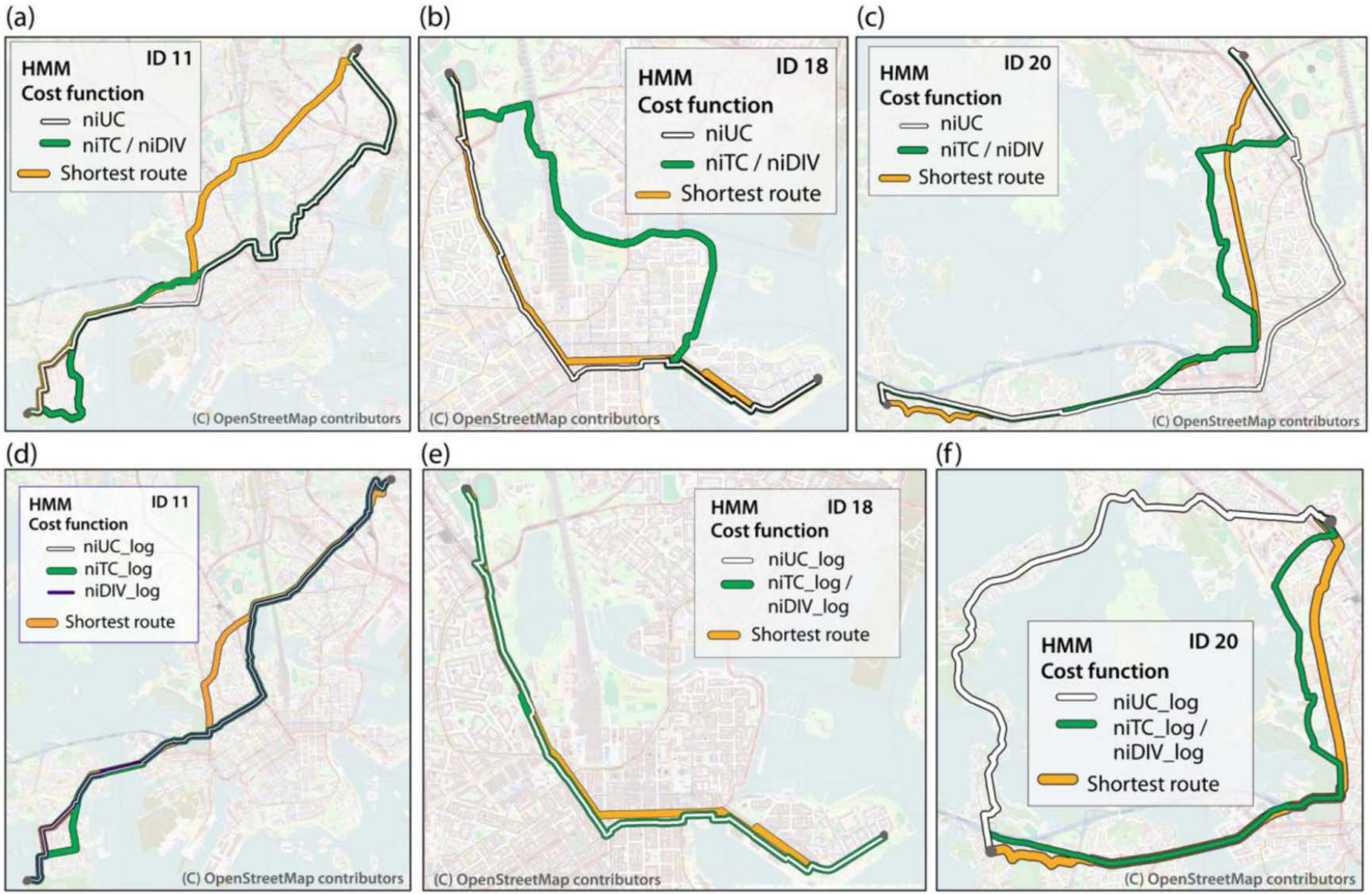
Suosioon perustuva reititys



OSM Map matching -> painotettu verkko



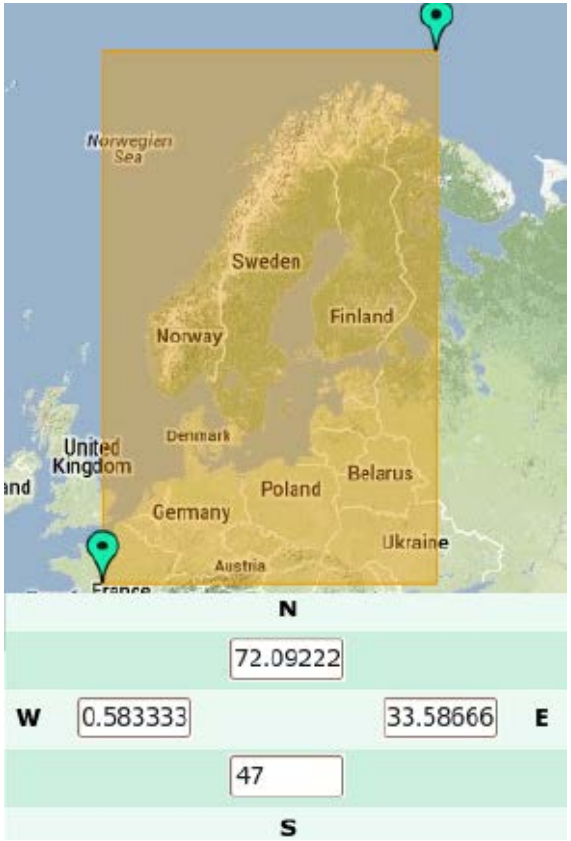
NLS
FINNISH GEOSPATIAL RESEARCH INSTITUTE
FGI



NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Verkkopalvelu lämpökarttojen tuottamiseksi

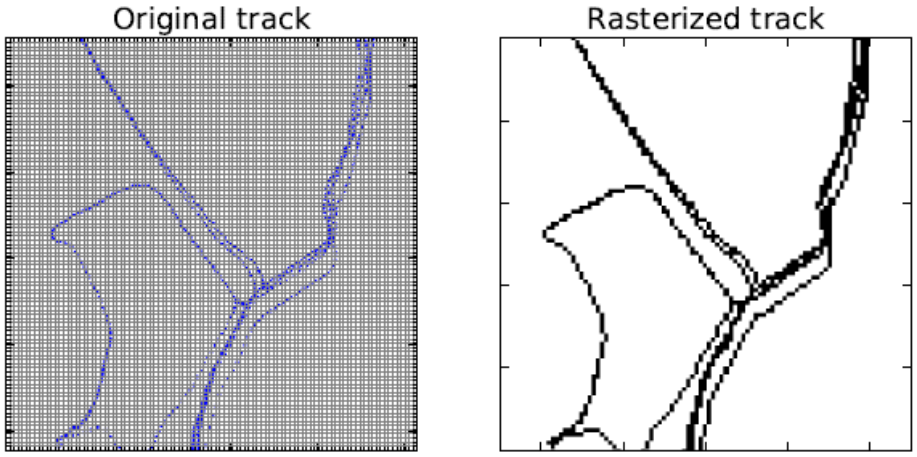
- 1017551 CSV tiedostoa, 1 liikuntasuoritus / tiedosto, 114 GB
- Karkeiden virheiden poisto:
  - Peräkkäisten pisteiden etäisyys 150 m ja aikaero < 1 h
  - Lat-Long rajat
  - Aikaleima 9.9.2001-14.5.2014
- 47801 käyttäjää, 765820 trajektoria



NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

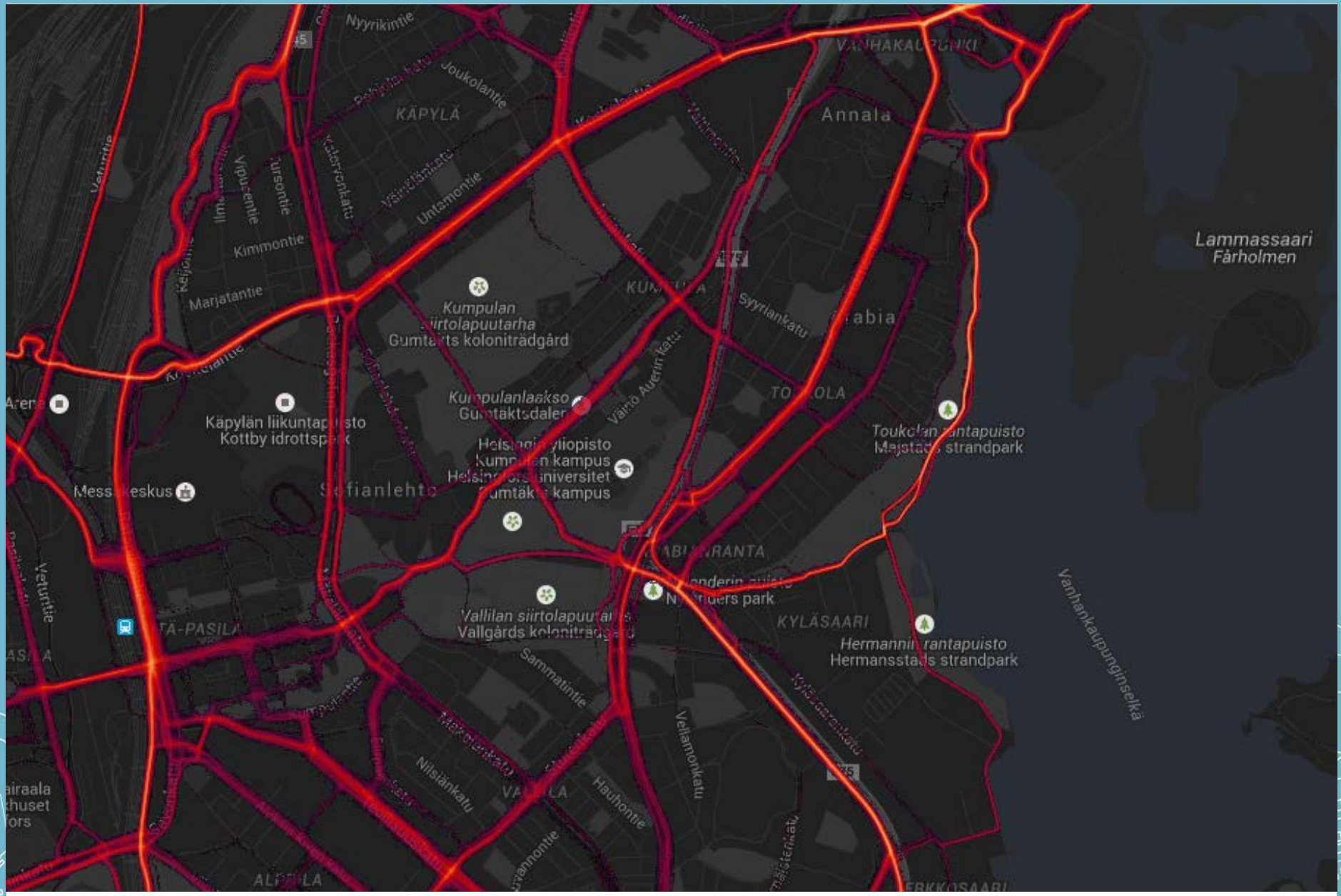
Esiprosessointi

- Trajektorit jaetaan tiiliin ja rasteroidaan valmiiksi kaikille karttapalvelun käyttämille zoomaus-tasoille



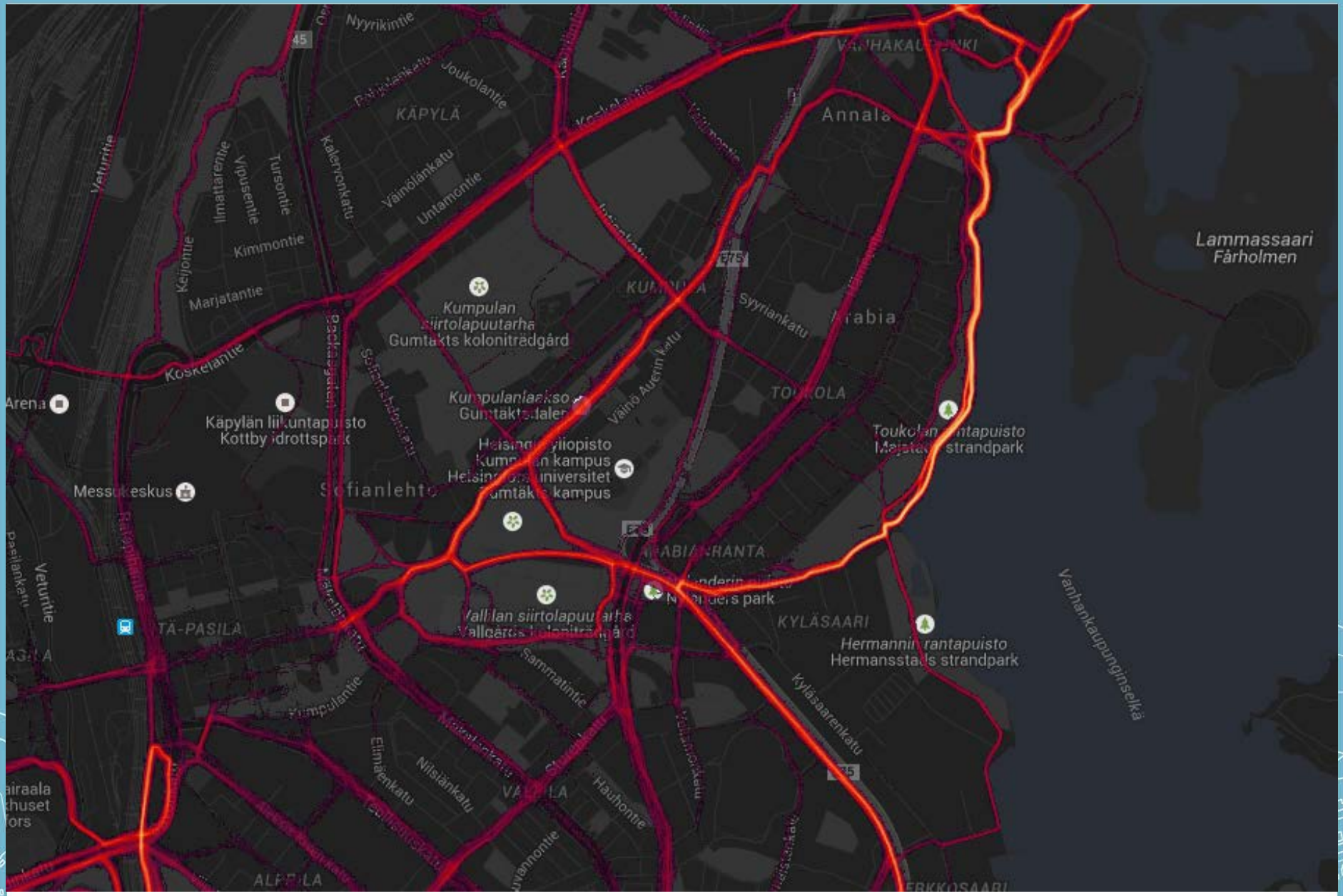
- Luodaan spatiaalinen indeksi ja tallennetaan trajektorikohtainen metatieto (laajuus, laji, keskinopeus jne.)

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI



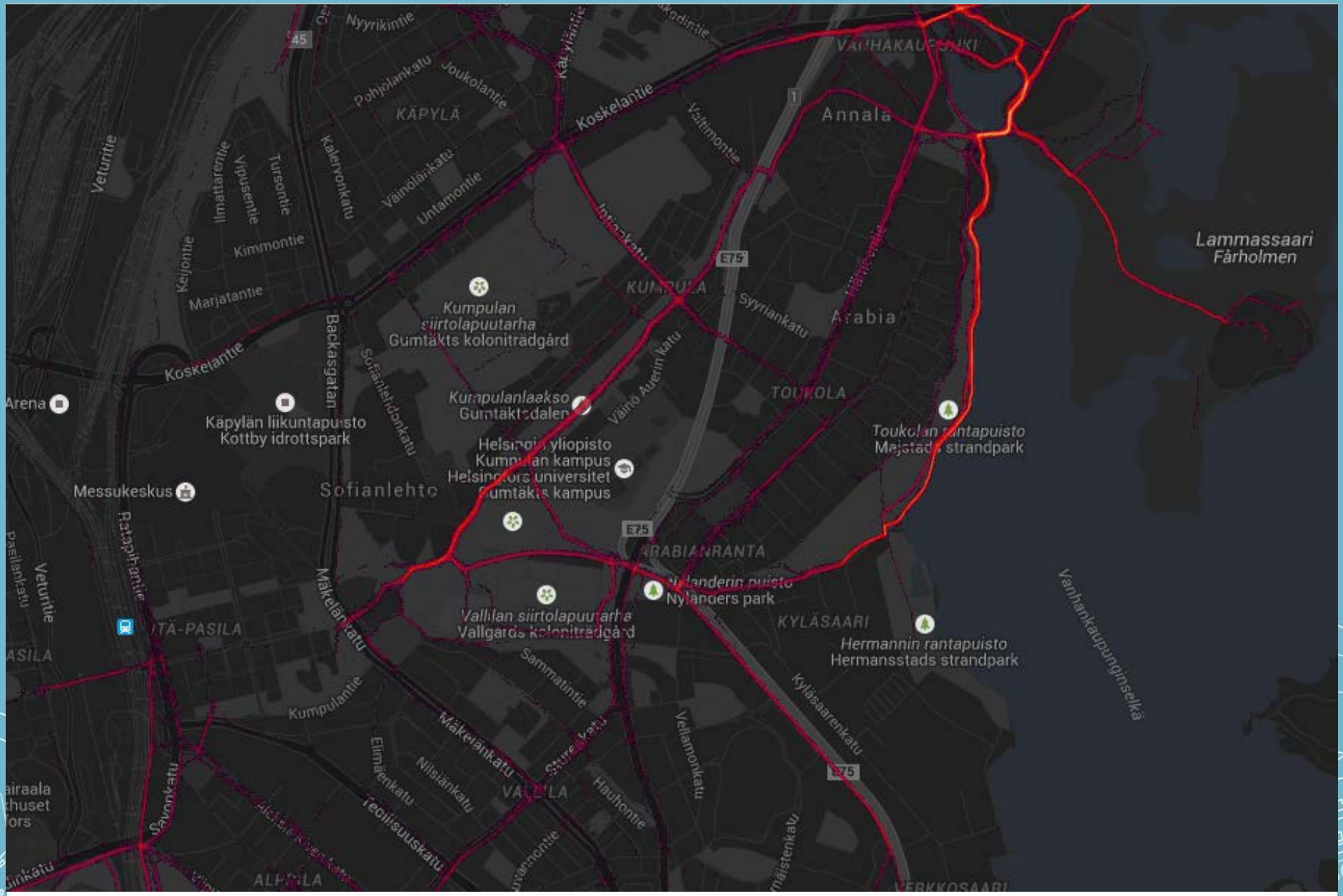
Pyöräily

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI



Juoksu

NLS
FINNISH GEOSPATIAL RESEARCH INSTITUTE
FGI

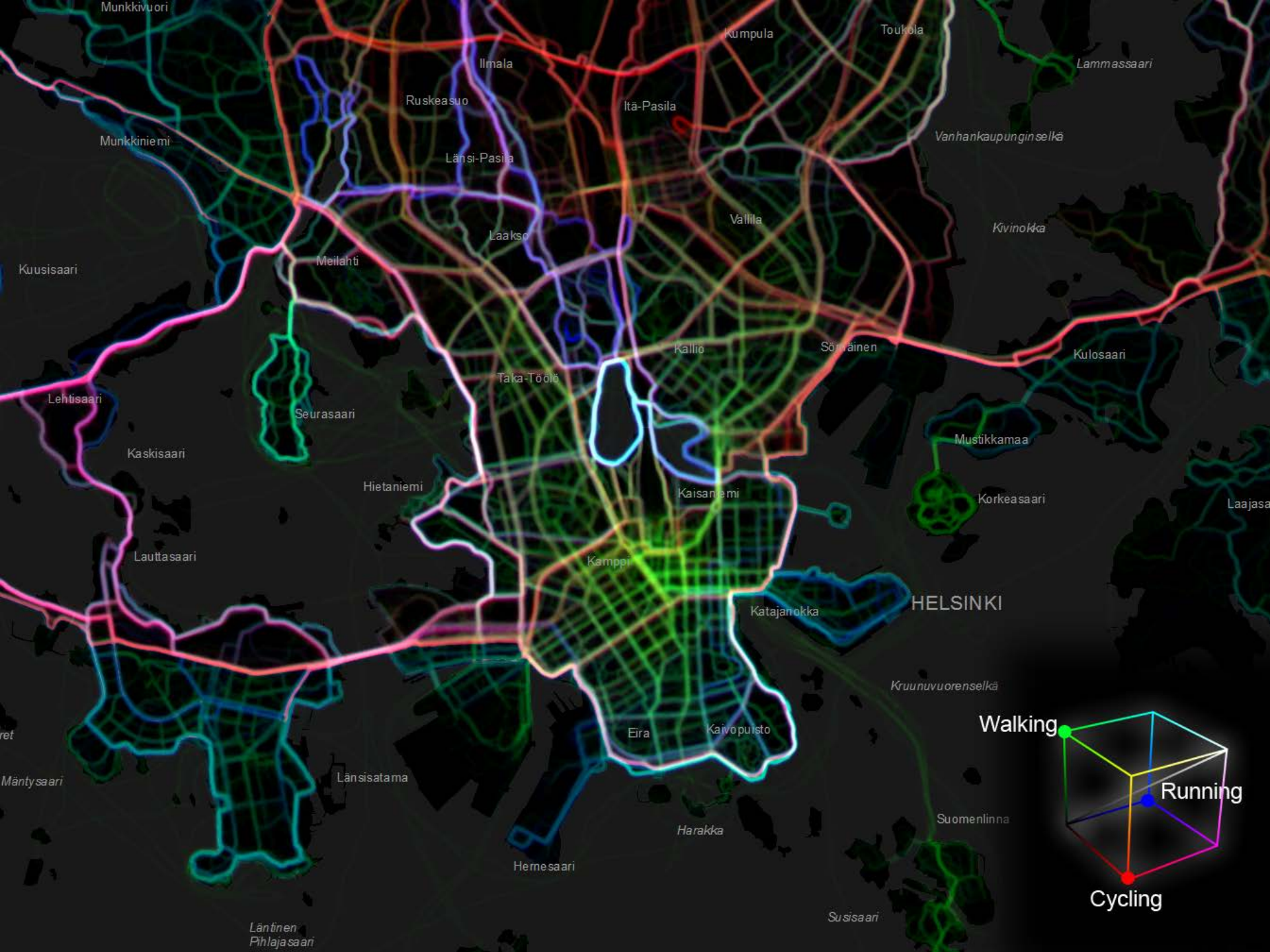


Kävely

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI

Sports Tracker Beta (Android)

- Suljettu testisovellus, jossa lämpökarttoja voi kokeilla todellisessa käyttötilanteessa
- Käyttää Supra-hankkeen lämpökarttatiiliä


Munkkivuori
Kumpula
Toukola
Ilmala
Lammassaari
Ruskeasuo
Itä-Pasila
Munkkiniemi
Vanhankaupunginselkä
Länsi-Pasila
Vallila
Kivinokka
Laakso
Meilahti
Kuusisaari
Kallio
Sörnäinen
Kulosaari
Taka-Töölö
Lehtisaari
Seurasaari
Mustikkamaa
Kaskisaari
Hietaniemi
Kaisaniemi
Korkeasaari
Laajasa
Lauttasaari
Kamppi
HELSINKI
Katajanokka
Kruunuvuorenselkä
Eira
Kaivopuisto
Mäntysaari
Länsisatama
Suomenlinna
Harakka
Hernesaari
Susisaari
Länsinen
Pihlajasaari
Walking
Running
Cycling

Yhteiskunnallinen lähestymistapa

- Liikenne-/kaupunkisuunnittelu
- Liikuntapaikkarakentaminen ja liikuntaympäristöjen tutkimus
- Yhdyskuntarakenteen tutkimus

Yhteiskunnallinen lähestymistapa

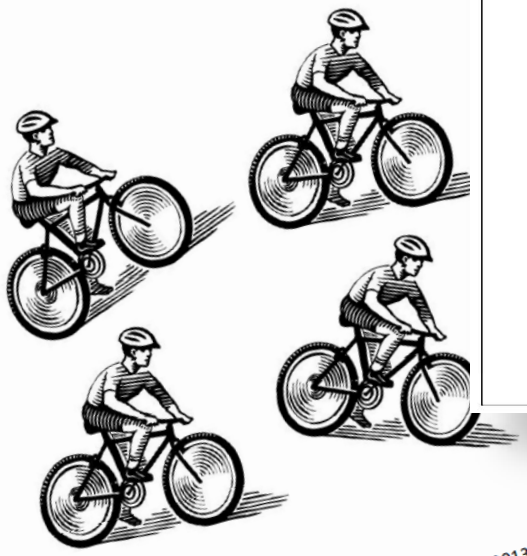
- Liikenne-/kaupunkisuunnittelu
- Liikuntapaikkarakentaminen ja liikuntaympäristöjen tutkimus
- Yhdyskuntarakenteen tutkimus

…mutta, mitä Sports Tracker datasta johdetut tulokset kertovat populaation käyttäytymisestä?

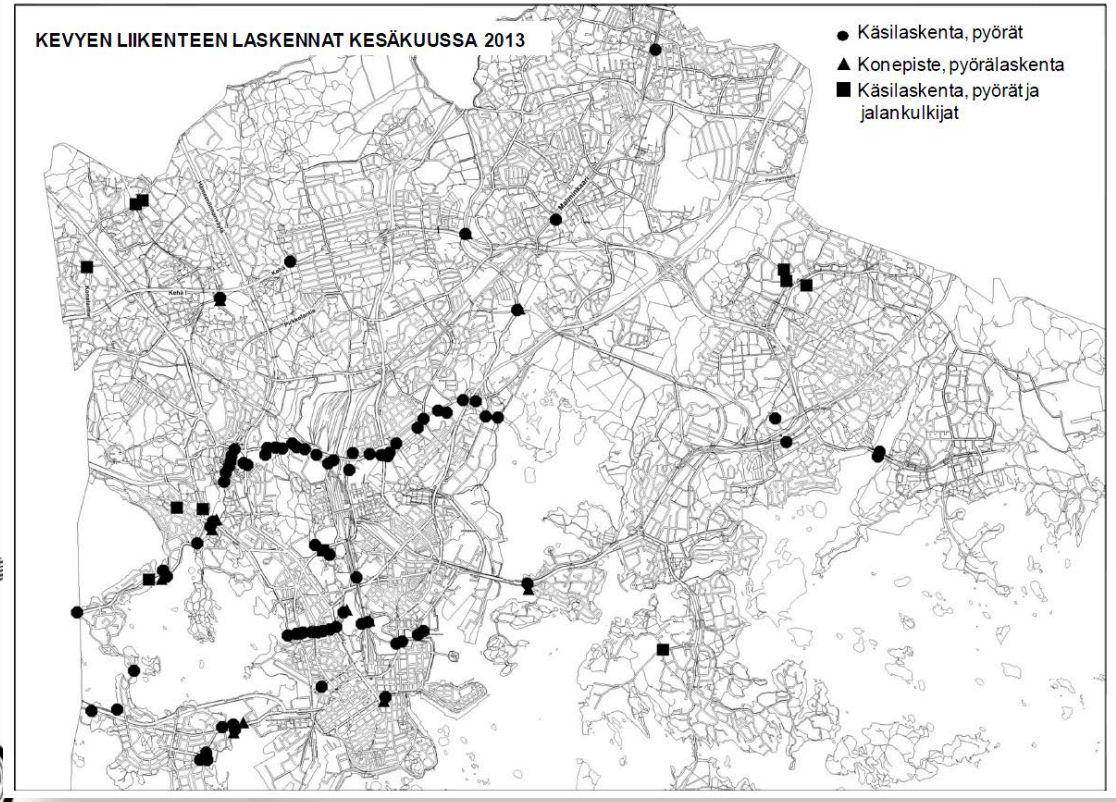
HAASTE 4:
Mitä data edustaa?

Lämpökartan kalibrointi

POLKUPYÖRÄLASKENNAT
HELSINGISSÄ
2013

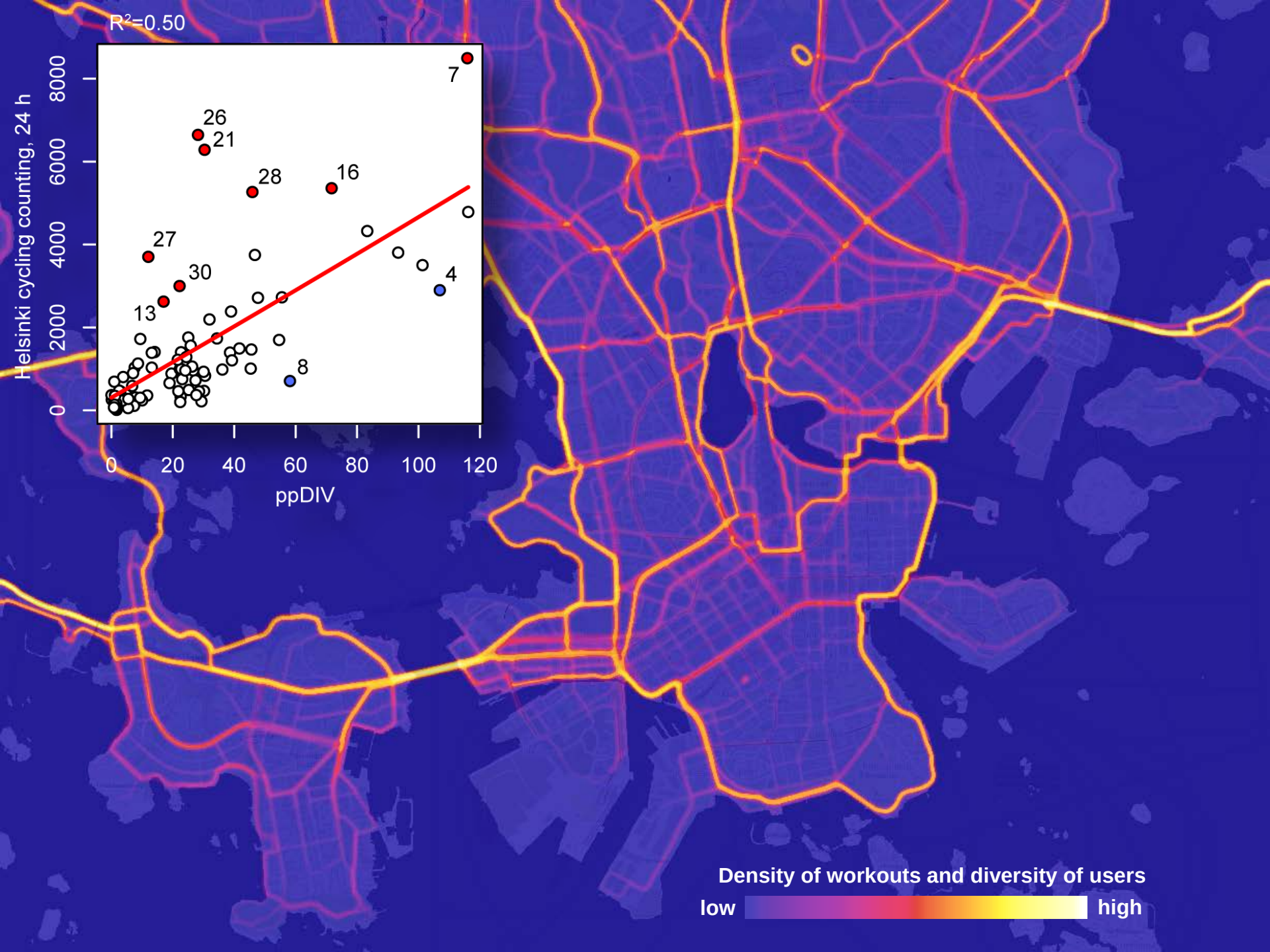


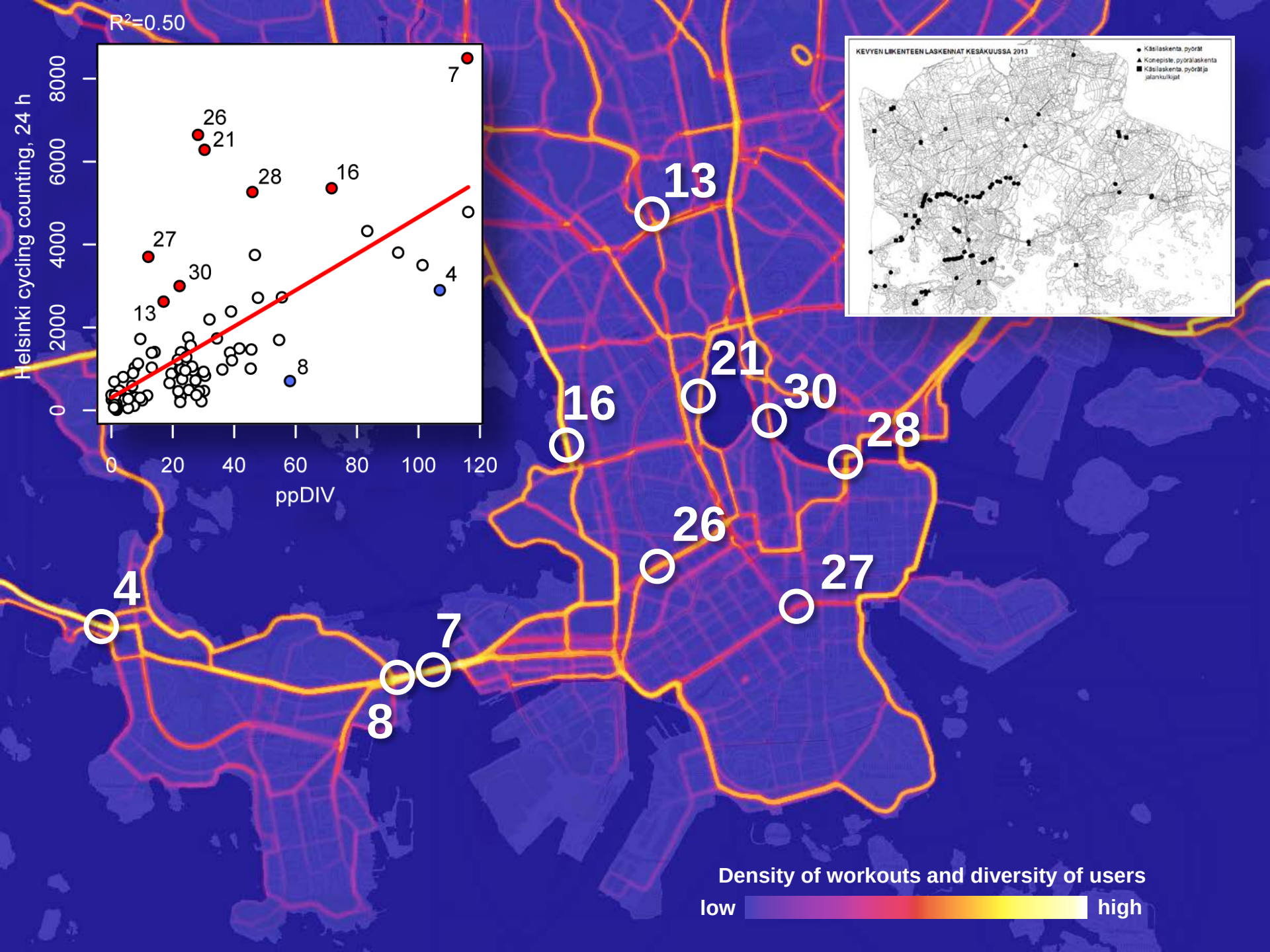
Muistio 28.10.2013 / Tuija Hellman
Helsingin kaupunkisuunnitteluvirasto
Liikennesuunnitteluosasto

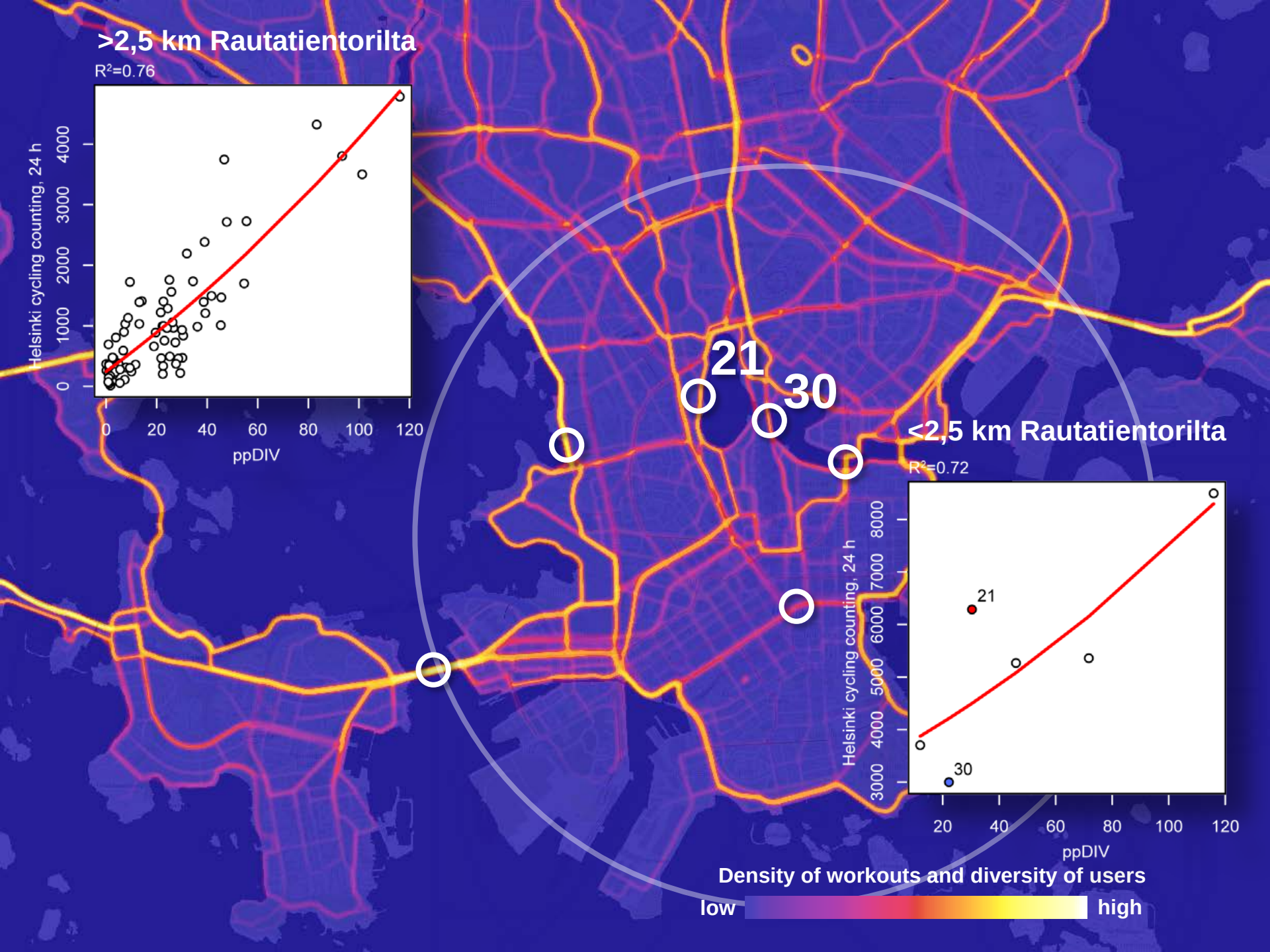


- 89 laskentapistettä
- 80 manuaalista,
  9 automaattista

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI


R²=0.50
Helsinki cycling counting, 24 h
0
2000
4000
6000
8000
0
20
40
60
80
100
120
ppDIV
7
26
21
28
16
27
30
4
13
8
Density of workouts and diversity of users
low
high


R²=0.50
Helsinki cycling counting, 24 h
0
2000
4000
6000
8000
0
20
40
60
80
100
120
ppDIV
7
26
21
28
16
27
30
4
13
8
KEVYEN LIIKENTEEN LASKENNAT KESÄKUUSSA 2013
Käsilaskenta, pyörät
Konepiste, pyöräläskenta
Käsilaskenta, pyörät ja jalankulkijat
13
21
16
30
28
26
27
4
7
8
Density of workouts and diversity of users
low
high


>2,5 km Rautatientorilta
R²=0.76
Helsinki cycling counting, 24 h
0
1000
2000
3000
4000
0
20
40
60
80
100
120
ppDIV
21
30
<2,5 km Rautatientorilta
R²=0.72
Helsinki cycling counting, 24 h
3000
4000
5000
6000
7000
8000
21
30
20
40
60
80
100
120
ppDIV
Density of workouts and diversity of users
low
high

Yhteenveto

- **HAASTE 1:** Yksityisyys
  - Kaikki ratkaisut luotava yksityisyydensuojan ehdoilla
  - Esim. k-anonymiteettiin pohjautuvat menetelmät, aggregointi, trajektoreiden päiden leikkaus jne.
- **HAASTE 2:** Iso data
  - Järkevä datan käyttö vähentämällä turhaa dataa
  - Kustomoidut ja hyvin skaalautuvat työkalut datan analysointiin, automatisointi, teholaskentaresurssien käyttö
- **HAASTE 3 ja 4**: Epätasainen osallistuminen / Mitä data edustaa?
  - Kvalitatiivinen tarkastelu mahdollista, kvantitatiivinen tarkastelu mahdollista varauksin
    - Esim. kalibrointi käyttäen in situ -havaintoja, täsmäys ajassa
    - Täytyy tuntea data ja infrastruktuurissa tapahtuneet muutokset

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Kiitokset

- SUPRA, Tekes (40261/12)
- INTUIT, Suomen Akatemia (251987)
- Cecilia Bergman, FGI
- Jani Sainio ja Jan Westerholm, ÅA
- Susanne Suvanto
- Antti, Jussi, Ramon ja muut Sports Tracking Technologies Oy:ssä

Åbo Akademi

Tekes

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Lisätietoja

juha.oksanen@nls.fi
http://supra.fgi.fi

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI