

Tietoalustakehitys Varhalla

Arho Virkki

Analytiikkajohtaja, Varha

Lääketieteellisen matematiikan dosentti, Turun yliopisto

Varsinais-Suomen hyvinvointialue • Egentliga Finlands välfärdsområde

```
arho@pollux:~$ df -h -x tmpfs
Filesystem                Size      Used Avail Use% Mounted on
devtmpfs                  498G         0   498G   0% /dev
/dev/mapper/vg_sys-lv_root 207G       36G   172G  18% /
/dev/sdb2                 1014M     319M   696M  32% /boot
/dev/sdb1                 200M       6.9M   193M   4% /boot/efi
/dev/sda1                 42T       29T    14T  68% /data

[arho@pollux ~]$ free -h
               total        used        free      shared  buff/cache   available
Mem:           995Gi       960Gi       4.5Gi         0.0Ki        30Gi        296i
Swap:           16Gi        13Gi       2.3Gi
```

Puhujan esittely

Arho Virkki

Analytiikkajohtaja, Varha

Lääketieteellisen matematiikan dosentti, Turun yliopisto

Muodollinen koulutus: Sähkötekniikan DI, Väitös Turun yliopiston matematiikan ja tilastotieteen laitokselle 2007 unitutkimuksesta matemaattisten ja tilastollisten mallien avulla (tekoäly mukaan lukien)

Kaksi viimeisintä työnantajaa: VTT: (Erikois)tutkija ja projektipäällikkö 2009 – 2016, Auria tietopalvelun johtaja VSSHP / Varha syyskuusta 2016.

Matemaattiset mielenkiinnon kohteet: Viimeisimpänä tilastollinen tietosuoja ja analyttisen lukuteorian alkeet, mutta yleisesti ottaen matemaattinen analyysi, epäparametrinen tilastotiede ja suurteholaskenta.

Ohjelmointikielet: R, Python, Java, JavaScript, SQL, Unix shell, C/C++, Matlab/Octave, Modula-3, Fortran 77/95, 80296-assembly. Käyttöjärjestelmät ja työkalut: Fedora, Ubuntu ja RHE Linux, Mac OS X ja Windows 3.1 – 11; Git, SVN, Vim, Eclipse. Julkaisuvälineet: Open Office, Inkscape, Gimp, HTML5/CSS3, LaTeX

Harrastukset: Ruoanlaitto, puukäsityöt (pienää hirsimökkiä viimeistellessä), hupiliikunta koiran kanssa tai ilman sekä indie-, proge- ja elektropopmusiikki.



Varhan tietotaltaan ja alustan kehitys 2014 →

Sote-tiedon toisiokäyttö Varsinais-Suomen Sairaanhoidopiirissä ja Varhalla

Tutkimus, kehitys, innovaatiotoiminta ja opetus: Kliinisen tiedon varasto (KTV 2014) → Kliinisen tiedon palvelu (KTP 2016) → Auria Tietopalvelu (ATP 2018) → Tutkimuksen ja opetuksen tietopalvelu Auria (TOTA 2023)

Tietojohtaminen: Tietopalvelutiimi (2013) ja ositettu toteutus (2M-IT, talous, tekniikka, Auria-palvelut) → Varhan järjestämisen toiminnon Tietojohtamisen palvelualue (=TIJO ja TOTA)

ATP Amiraali



Luovutus-/tutkimustilasto

Tietopalvelumyynti

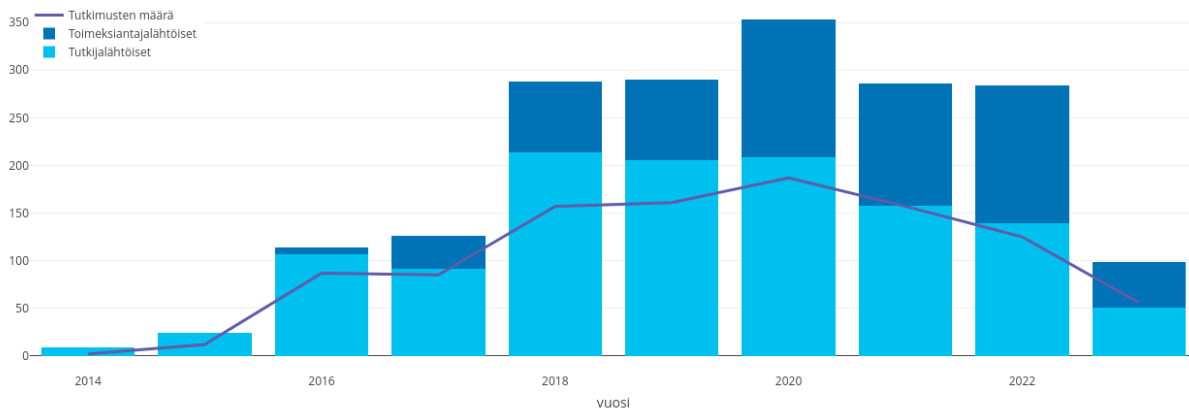
Sisäiset asiakkuudet

Toimeksiantaja-asiakkuudet

Valitse aikaväli:



Toimeksiantaja vs. tutkijalähtöisten luovutusten määrät



1873 kpl

Luovutuksia yhteensä aikavälillä 2014 - 2023

67 %

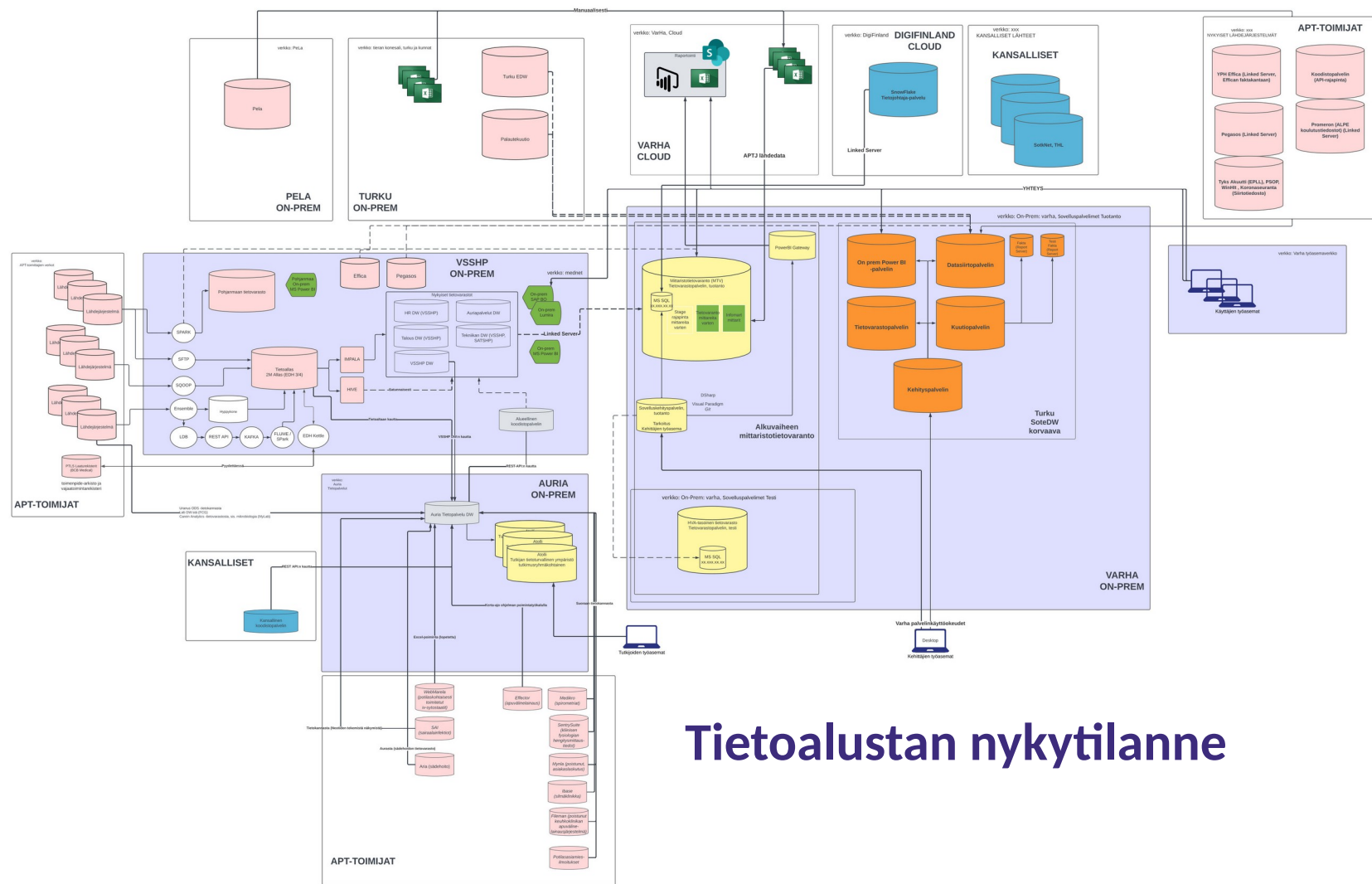
Tutkijalähtöisten luovutusten osuus

16 %

Toimeksiantajalähtöisten osuus

17 % Tietojohtaminen ja tietoturvalliset ympäristöt

Tilasto 31.5.2023



Tietoalustan nykytilanne

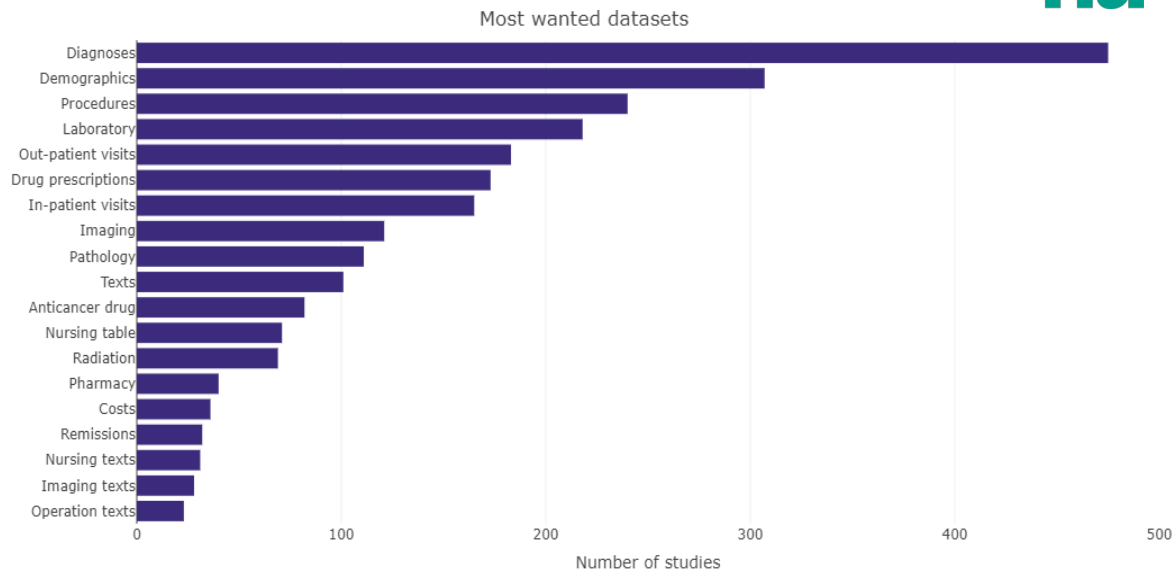
Tietoalustan nykytilanne

Tarjolla olevat aineistot

- Erikoissairaanhoidon tärkeimmät aineistot (<https://aineistokatalogi.fi>)
- Varhan aineistot aggregaattitasolla (n.k. mittaristotietovaranto)

Mihin soveltuu tällä hetkellä

- Tietojohdamisen raportointiin
- Varhan omaan tutkimukseen ja toiminnan kehittämiseen
- Akateemiseen tutkimukseen
- Lääkeyritysten tutkimuksiin
- Lääkeyrityksiä palvelevien yritysten tutkimuksiin
- Viranomaistoimintaan, esim. Findata, Fimea, ja viimeisimpänä Euroopan lääkevirasto EMA (Darwin-yhteistyö)

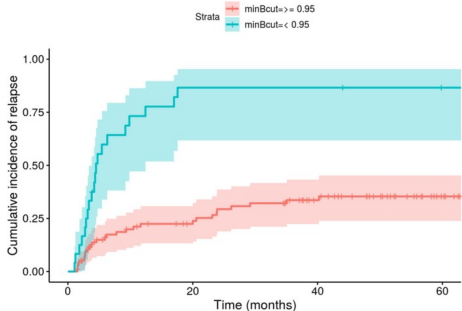


Nykytoimintaa

AURIA PALVELUT

Relapse	no (N=73)	yes (N=49)	Overall (N=122)
Gender			
female	28.0 (38.4%)	25.0 (51.0%)	53.0 (43.4%)
male	45.0 (61.3%)	24.0 (49.0%)	69.0 (56.6%)
Age at transplantation (years)			
Median (Q1, Q3)	53.0 (42.0, 60.0)	57.0 (46.0, 62.0)	55.0 (43.0, 61.0)
Min, Max	19.0, 69.0	23.0, 68.0	19.0, 69.0
Diagnosis class			
g1	30.0 (41.1%)	29.0 (59.2%)	59.0 (48.4%)
g2	7.00 (9.6%)	5.00 (10.2%)	12.0 (9.8%)
g3	5.00 (6.8%)	3.00 (6.1%)	8.00 (6.6%)
g4	31.0 (42.5%)	12.0 (24.5%)	43.0 (35.2%)
Donor			
mud	59.0 (80.8%)	36.0 (73.5%)	95.0 (77.9%)
sibling	14.0 (19.2%)	13.0 (26.5%)	27.0 (22.1%)
Graft source			
blood	64.0 (87.7%)	46.0 (93.9%)	110 (90.2%)
bone marrow	9.00 (12.3%)	3.00 (6.1%)	12.0 (9.8%)
Graft cell count			
Median (Q1, Q3)	6.00 (4.29, 9.20)	6.48 (4.70, 8.76)	6.34 (4.70, 9.09)
Min, Max	0.750, 15.6	1.00, 13.5	0.750, 15.6
Missing	2.00 (2.7%)	0 (0%)	2.00 (1.6%)
Condition regimen			
MAC	26.0 (35.5%)	14.0 (28.6%)	40.0 (32.8%)
RIC	34.0 (46.5%)	22.0 (44.9%)	56.0 (45.9%)
sequential	13.0 (17.8%)	13.0 (26.5%)	26.0 (21.3%)
DRI stage			
high	21.0 (28.9%)	18.0 (36.7%)	39.0 (32.0%)
low	52.0 (71.2%)	31.0 (63.3%)	83.0 (68.0%)
EBMT risk score			
Median (Q1, Q3)	4.00 (3.00, 5.00)	3.00 (3.00, 5.00)	4.00 (3.00, 5.00)
Min, Max	1.00, 6.00	1.00, 6.00	1.00, 6.00

Minimum B all



```
Median survival with 95% CI
median lower upper
[1,] "minBcut<= 0.95" "NA" "NA"
[2,] "minBcut< 0.95" "4.50" "3.80" "12.42"

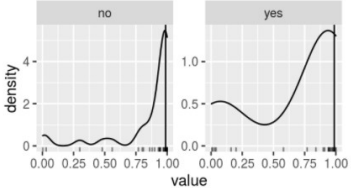
Log-rank test

Call:
survdifff(formula = Surv(relapsetime, relapse_stat) ~ minBcut,
data = analysisdata)

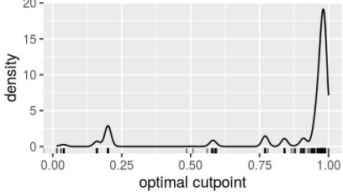
N Observed Expected (O-E)^2/E (O-E)^2/V
minBcut<= 0.95 99 29 42.4821 4.2787 32.913
minBcut< 0.95 24 20 6.5179 27.8871 32.913

Chisq= 32.9 on 1 degrees of freedom, p= 9.64e-09
```

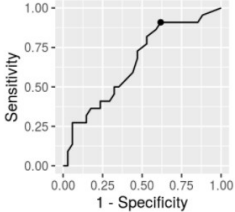
Independent variable
optimal cutpoint and distribution by class



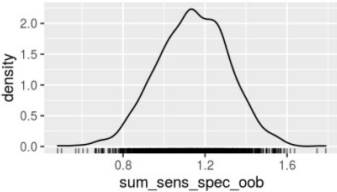
Bootstrap
distribution of optimal cutpoints



ROC curve



Bootstrap
out-of-bag estimates



Nykytoimintaa

Aurian Atolli

Findatan vaatimusten mukainen tietoturvallinen käyttöympäristö

Q https://atolli.auria.fi

Atolli

Tietoturvallinen käyttöympäristö Aurian Atolli

Aurian Atolli on Varsinais-Suomen hyvinvointialueeseen kuuluvan Auria Tietopalvelun tarjoama tietoturallinen käyttöympäristö tietoluvalla saatavien aineistojen käsittelyyn. Palvelu täyttää Findatan määräyksen mukaiset vaatimukset.

Käyttämällä palvelua sitoudut palvelun käyttöehtoihin.

Palvelun käytön aloittamista varten tulee olla yhteydessä osoitteeseen atolli@varha.fi.

Tunnistautuminen palveluun

Tunnistautuminen palveluun tapahtuu Suomi.fi-tunnistautumisen avulla.

Mitä tietoja käyttäjistä tallennetaan tunnistautumisen yhteydessä?

Suomi.fi-tunnistautumisen yhteydessä palveluun välittyy nimesi ja henkilötunnuksesi. Henkilötunnuksesi perusteella muodostetaan kryptattu tunnist. Henkilötunnustasi ei tallenneta palveluun, ainoastaan siitä muodostettu kryptattu tunnist.

[Kirjaudu palveluun](#)

Olet tunnistautumassa palveluun

Auria Tietopalvelu - Aurian Atolli

Valitse tunnistustapa



Varmennekortti



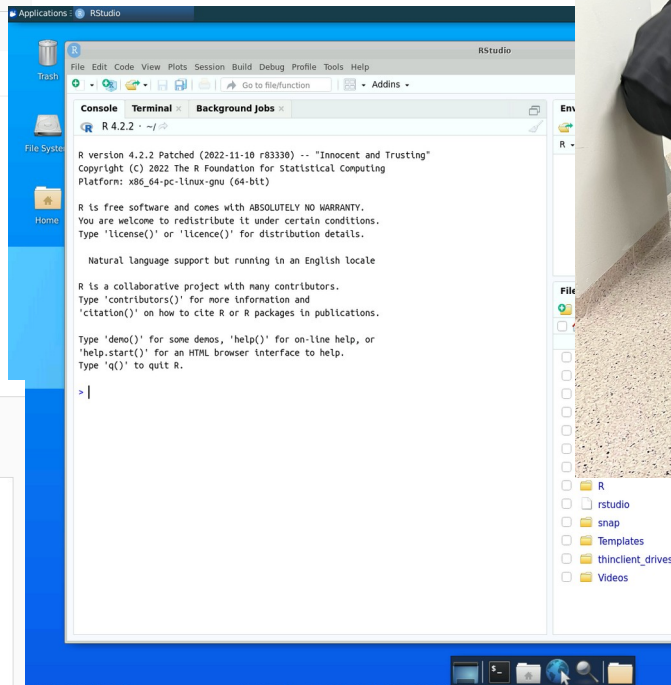
Mobiilivarmenne



Osuuspankki

Nordea

Nordea



AURIA
TIETOPALVELU

Nykytoimintaa

Tietojen siltaus kansainvälisiin koodistoihin

OMOP Common Data Model



The OMOP Common Data Model allows for the systematic analysis of disparate observational databases.



The goal of EHDEN is to make the large-scale analysis of health data in Europe a reality. The project aims to do this by building a federated data network of allowing access to the data of 100 million EU citizens standardised to a common data model.



DARWIN EU will deliver real-world evidence from across Europe on diseases, populations and the uses and performance of medicines.

This will enable EMA and national competent authorities in the the European medicines regulatory network to use these data whenever needed throughout the lifecycle of a medicinal product.

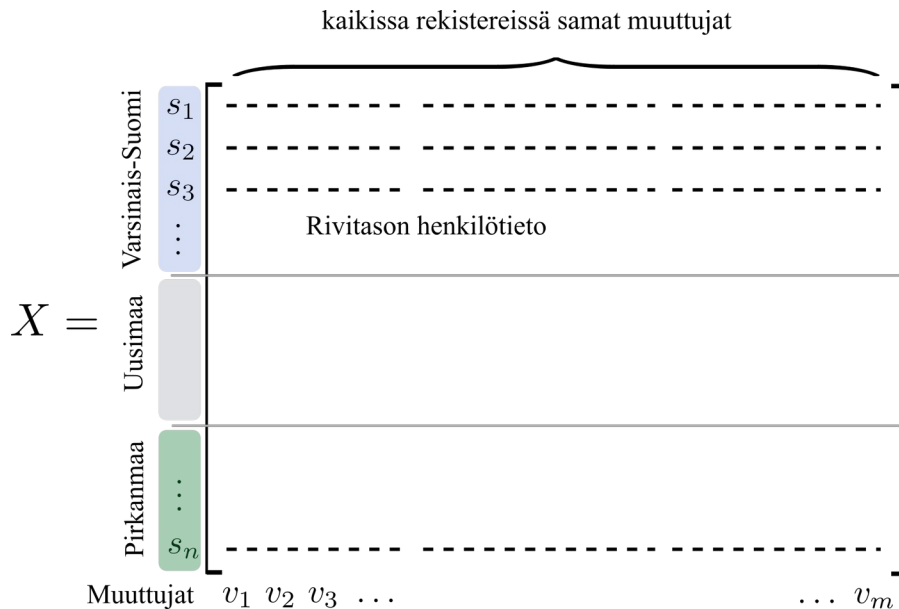
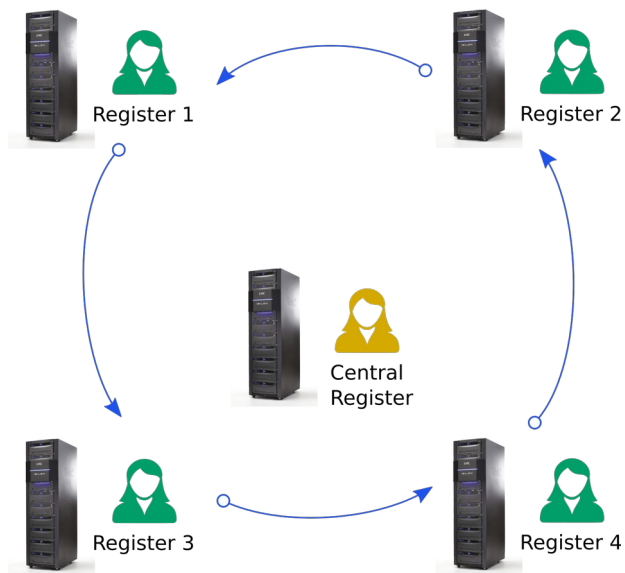


Nykytoimintaa

Hajautettu laskenta

OMOP CDM -dataan perustuvat parvioppimisen pilottihanke

Yhdistetty laskenta (federated computing)



Tekstinlouhinta ja tietojen rakenteistus

Semanttiset verkot ovat tehokkaita, mutta säännölliset lausekkeet ovat helpommin validoitavissa

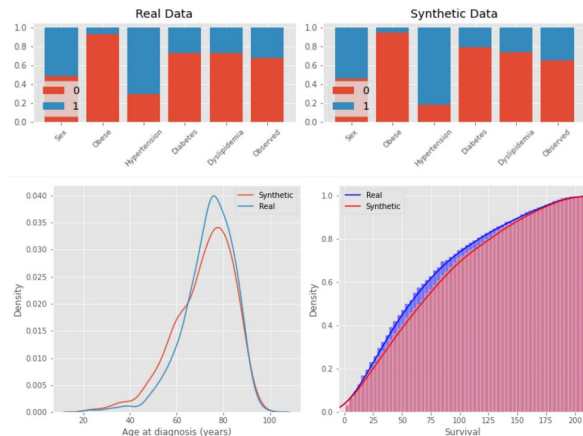
```
# A tibble: 6,765 × 1
  arvo
<chr>
1 Positiivinen (2+)
2 ei tehdä
3 Negatiivinen (1+)
4 tilattu
5 posit (1+)
6 neg.
7 negatiivinen
8 negat.
9 potiivinen
10 (1+ - 2+)
# ... with 6,755 more rows
```

```
# A tibble: 6,765 × 7
  arvo      roska raja_arvo pos  neg  min_arvo max_arvo
<chr>    <lgl> <lgl>    <lgl> <lgl> <chr>    <chr>
1 Positiivinen (2+) FALSE FALSE TRUE  FALSE NA      2+
2 ei tehdä      TRUE FALSE FALSE FALSE NA      NA
3 Negatiivinen (1+) FALSE FALSE FALSE TRUE  NA      1+
4 tilattu       TRUE FALSE FALSE FALSE NA      NA
5 posit (1+)    FALSE FALSE TRUE  FALSE NA      1+
6 neg.          FALSE FALSE FALSE TRUE  NA      NA
7 negatiivinen  FALSE FALSE FALSE TRUE  NA      NA
8 negat.        FALSE FALSE FALSE TRUE  NA      NA
9 potiivinen    FALSE FALSE TRUE  FALSE NA      NA
10 (1+ - 2+)    FALSE TRUE  FALSE FALSE (1+)  (2+)
```

NLP-prosessin
tulokset



Oikean datan pohjalta mallinnettu keinotekoinen anonyymi, mutta alkuperäistä muistuttava aineisto



Binary variable
proportions
in real (left) and
synthetic (right) data.

Age at diagnosis
feature (left)

Cumulative survival
distribution (right)



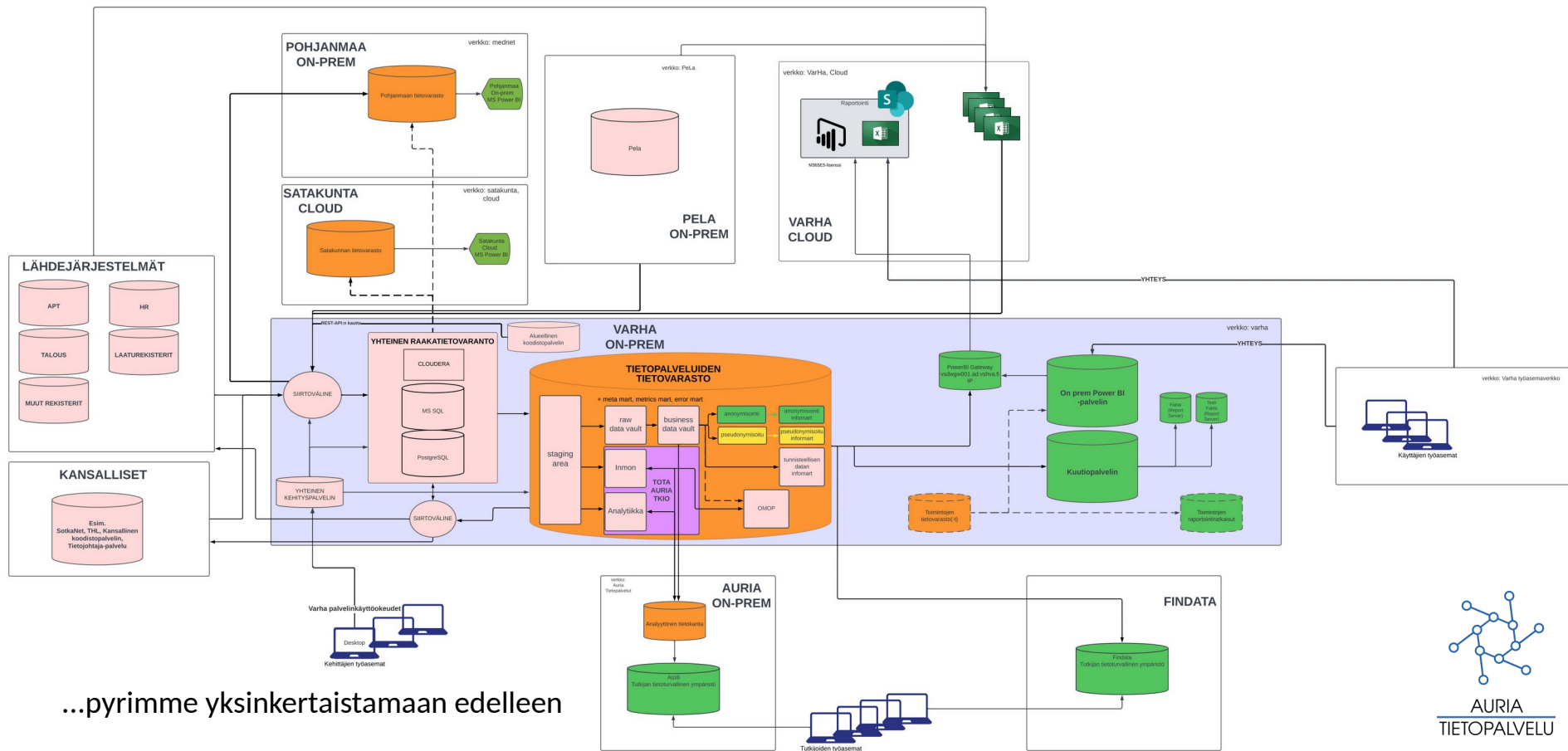
Rajoitukset ja suorat ongelmat

- 1) Perusterveydenhuollon data puuttuu rivitasoisena!
- 2) Talouden, henkilöstöhallinnon ja tekniikan data on integroimatta
- 3) Tekninen arkkitehtuuri on hajanainen
- 4) Hallintamalli on osin implementoimatta
- 5) Helppokäyttöistä tietohakemistoa (datakatalogia) ei ole
- 6) Suurteholaskenta puuttuu Aurian Atollista (Findata-akkreditoitu tietoturvallinen käyttöympäristö)

Mitä tietoaltaalla pitäisi pystyä tekemään tulevaisuudessa

- Johtamaan tiedolla, kaikilla tasoilla (eli tekemään operatiivisen toiminnan suunnittelua, taktista johtamista yksikötasolla sekä ylimmän johdon strateginen suunnittelua)
- Tukemaan tutkimusta, kehitystä, innovaatiotoimintaa ja opetusta koko rekisterinpitäjän tietopohjan laajuudessa
- Yhdistämään aineistoja federoitusti yli rekisterinpitäjien (OMOP-mallinnettu data ja hajautettu laskenta)
- Palauttamaan tietoa ensiökäytön tarpeisiin

Tietoalustan seuraava askel



Käyttörajoitusten alaisen tiedon suojaaminen

Käsitteiden määritelmät

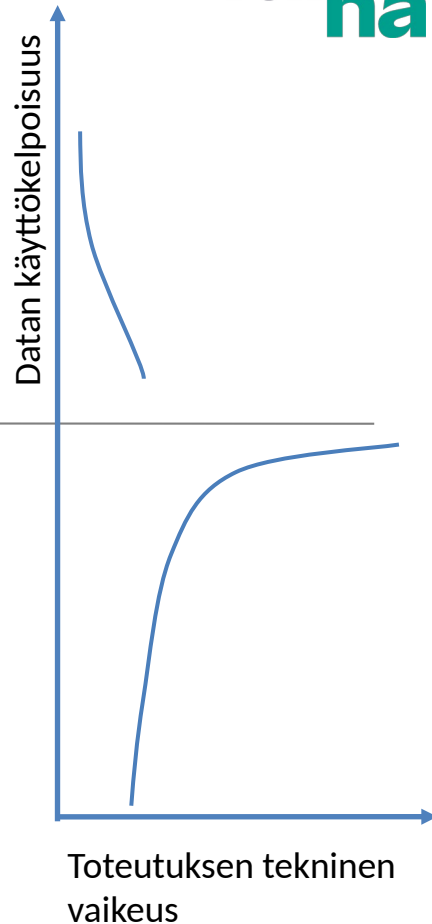
Henkilötietoa

1. **Alkuperäinen data:** Yksilöityä ja ajantasaista henkilötietoa suoraan lähdejärjestelmistä
2. **De-identifioitu data** eli peitenimillä suojattu data: suorat tunnisteet, kuten henkilötunnukset on muutettu koodeiksi (P001, P002,..., C001, C002,...) tai poistettu.
3. **Ei-anonyymi synteettinen data**, (esim. uusio-otanta, huono GAN-malli,...)

Ei henkilötietoa

Ei jännösriskiä Jännösriski

4. **Anonyymi synteettinen data:** (i) *Oikeasta datasta* tilastollisesti mallinnettu (ii) *anonyymi* satunnainen data, jolla on (iii) *todellisen kaltaisia tilastollisia ominaisuuksia*.
5. **Anonyymi** eli peruuttamattomasti tunnistamattomaksi tehty data: Datapisteestä on mahdotonta päätellä, kenestä se alun perin on. Myös purkuavain on hävitetty. Jos purkuavain on tallessa, puhutaan (GDPR:n mukaan) **pseudonymistä** datasta.
6. **Simuloitu data:** Matemaattisesta (kuten mekanistisesta differentiaaliyhtälö)mallista generoitu data, jolla on todellisen kaltaisia tilastollisia ominaisuuksia. (Esim. pandemiaennusteet)
7. **Fabrikoitu** eli tekaistu data: Esim. oikean muotoiset sähköpostiosoitteet ja henkilötunnukset ohjelmistotestauksen käyttöön. Ei tilastollista ennustusvoimaa.





varha

Turkimuksen ja opetuksen
tietopalvelu Auria
<https://auria.fi>



AURIA CLINICAL
INFORMATICS