

# LOPPURAPORTTI

## KOKEILUKIIHDYTTÄMÖ DATAEKOSYSTEEMIKOKEILUT 2025

# 107. Pohjatutkimusdatan datankäsittelyn kehittäminen

### Kokeilun tiimi

Staran pohjatutkimusyksikkö, KYMP Geo

### Toteutuskumppani(t)

Gofore, Geomachine Oy

Helsinki



# Sisällysluettelo

1. **Tiivistelmä**
2. **Kokeilun tavoitteet**
3. **Kokeilun keskeiset opit**
4. **Kokeilun eteneminen**
5. **Kokeilun tuotokset**
6. **Opit kokeiltavan ratkaisun tai toimintatavan mahdollisuuksista**
7. **Opit asiakkaiden tai palvelun käyttäjien tarpeista**
8. **Opit ratkaisun kehittämisestä teknisesti**
9. **Opit kokeilemisestä yleensä**
10. **Opit kokeiluprojektin arjen pyörittämisestä**
11. **Kokeilun tekninen ympäristö**
12. **Kokeilun data**
13. **Jatkopäätökset ja -ideat**
14. **Liitteet:**
  1. **Dataekosysteemin systeemikuva**
  2. **Opit dataekosysteemistä ja sen toimijoista**

# 1. Tiivistelmä 1/2

## Kokeilun ydintavoitteet:

- **Selvittää** voiko pohjatutkimuksen prosessia tehostaa koneoppimisen ja tekoälyn menetelmin poistamalla siitä nykyisiä manuaalisia toimenpiteitä.
- **Kokeilla**, onko purheijari- ja porausdatan korjaaminen ja tulkinta mahdollista automatisoidusti ja voidaanko koneen tallentamaa tulostiedostoja laajempaa datamassaa käyttää tulosten laadun parantamiseen.

**Kokeilun nimi:** Pohjatutkimuksen prosessien tehostaminen koneoppimisen ja tekoälyn menetelmin.

**Uskomme, että** hyödyntämällä koneoppimista ja tekoäly pohjatutkimuksen prosesseissa, voimme:

1. Tehostaa nykyistä toimintaa poistamalla manuaalisia toimenpiteitä.
2. Parantaa käsiteltävän datan tarkkuutta poistamalla virheitä, jotka johtuvat manuaalisesta työstä tai matkan varrella hukkuvasta datasta.
3. Ymmärtää koneoppimisen ja tekoälyn mahdollisuuksia pohjatutkimuksen toimialalla.

**Kokeillaksemme tätä aiomme** hyödyntää koneoppimisen ja tekoälyn menetelmiä datan korjaamiseen, täydentämiseen ja rikastamiseen siten, että mallia koulutetaan ensin laajalla määrällä jo käsiteltyä dataa. Koulutettua mallia testataan pohjatutkimuksessa kerättyyn raakadataan, jonka perusteella sen tulisi tunnistaa maalajit sekä korjata mahdolliset datan vääristymät.

**Olemme oikeassa, jos** koulutettu malli pystyy raakadatatista automaattisesti tunnistamaan maalajit, ja korjaamaan datan virheitä siten, että lopputuloksen laatu vastaa manuaalisesti käsitellyn datan tasoa tai ylittää sen.

# 1. Tiivistelmä 1/2

**Kokeilun opit:** maalajien tunnistaminen saatiin toimimaan purheijari- ja porausdatalla suhteellisen hyvin, ja koneen mittaaman datan käyttöä datan laaduntarkkailuun saatiin hieman edistettyä. Jotta tulokset saataisiin vietyä tuotantoon, tarvittaisiin jatkokehitystä järjestelmien välisen dataliikenteen sujuvoittamiseksi sekä automaattitulkinnan laajentamiseksi muihin menetelmiin.

**Suositukset jatkotoimenpiteiksi:** tarvittaisiin projekti, jossa dataliikenne suunnittelijalta järjestelmiin saataisiin tehokkaammaksi, automatisoitu tulkinta saataisiin upotettua osaksi muita järjestelmiä, ja tulkintamenetelmiä saataisiin laajennettua purheijarin ja porauksen lisäksi muihin menetelmiin.

**Kokeilun nimi:** Pohjatutkimuksen prosessien tehostaminen koneoppimisen ja tekoälyn menetelmin.

**Uskomme, että** hyödyntämällä koneoppimista ja tekoäly pohjatutkimuksen prosesseissa, voimme:

1. Tehostaa nykyistä toimintaa poistamalla manuaalisia toimenpiteitä.
2. Parantaa käsiteltävän datan tarkkuutta poistamalla inhimillisiä virheitä, jotka johtuvat manuaalisesta työstä.
3. Ymmärtää koneoppimisen ja tekoälyn mahdollisuuksia pohjatutkimuksen toimialalla.

**Kokeillaksemme tätä aiomme** hyödyntää koneoppimisen ja tekoälyn menetelmiä datan korjaamiseen, täydentämiseen ja rikastamiseen siten, että mallia koulutetaan ensin laajalla määrällä jo käsiteltyä dataa. Koulutettua mallia testataan pohjatutkimuksessa kerättyyn raakadataan, jonka perusteella sen tulisi tunnistaa maalajit, korjata mahdolliset datan vääristymät, sekä tunnistaa ja lisätä puuttuvat koordinaattitiedot.

**Olemme oikeassa, jos** koulutettu malli pystyy raakadatasta automaattisesti tunnistamaan maalajit, korjaamaan havaittuja virheitä ja täydentämään puuttuvat koordinaattitiedot siten, että lopputuloksen laatu vastaa manuaalisesti käsitellyn datan tasoa tai ylittää sen.

# 2. Kokeilun tavoitteet 1/2

## Ongelman kuvaus

- Pohjatutkimuksen nykyinen prosessi sisältää useita manuaalisia ja aikaa vieviä vaiheita, jotka hidastavat työnkulkua ja lisäävät virhemahdollisuuksia. Datat siirtäminen, korjaaminen, täydentäminen ja rikastaminen tehdään pitkälti käsityönä, ja merkityksellistä dataa häviää siirroissa.

## Rajaus

- Kokeilu rajattiin niin, että
  - Gofore vastasi koneoppimis- ja tekoälyratkaisun toteutuksesta pohjatutkimuksen keräämän datan käsittelyyn.
  - Geomachine laajensi kairakoneen dataloggereiden toimintoja datan virheiden poistamiseksi ja laadun parantamiseksi käyttäen koneen keräämiä lisätietoja. Heidän kanssaan pyritään järjestämään myöhemmin myös datan siirron automatisointia.

# 2. Kokeilun tavoitteet 2/2

## Oletukset

Oletamme, että kokeilun tuloksena pystymme osoittamaan, että pohjatutkimuksen prosessia voidaan tehostaa merkittävästi hyödyntämällä koneoppimisen ja tekoälyn menetelmiä siten, että datan korjaaminen, täydentäminen ja rikastaminen voidaan toteuttaa automatisoidusti, ja että pohjatutkimussuunnitelma voidaan siirtää kairajalle ilman manuaalisia toimenpiteitä.

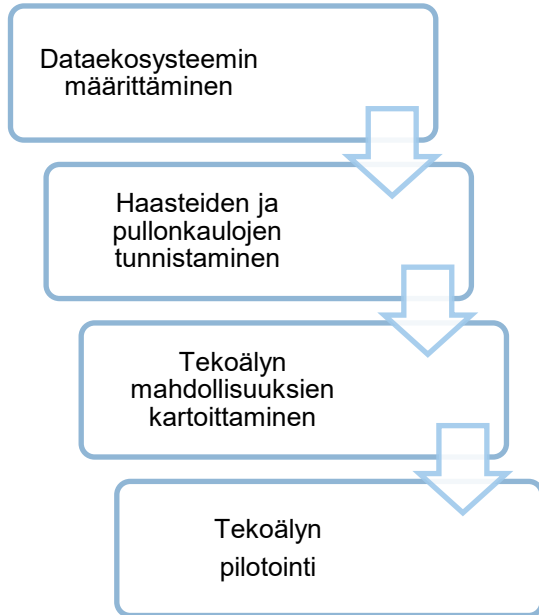
## Tavoitteet

Kokeilun tavoitteena on selvittää, voidaanko pohjatutkimuksen eri vaiheita tehostaa ja automatisoida koneoppimisen ja tekoälyn menetelmillä siten, että datan laatu paranee, prosessi nopeutuu ja tiedonsiirto kentälle tapahtuu sujuvasti ilman manuaalista vaiheita.

# 3. Kokeilun keskeiset opit

- Kokeilu jäi alun perin toivottua suppeammaksi lähinnä aikataulu- ja kustannusrajoitusten vuoksi, mutta perusteellisimmin testattu osio, datan tekoälytulkinta, saatiin toimimaan suhteellisen hyvin.

# 4. Kokeilun eteneminen



## Dataekosysteemin tunnistaminen

- Tunnistetaan pohjatutkimuksen dataekosysteemin prosessi ja sidosryhmät.
- Tunnistetaan, miten data jaetaan ja hyödynnetään eri sidosryhmien välillä.

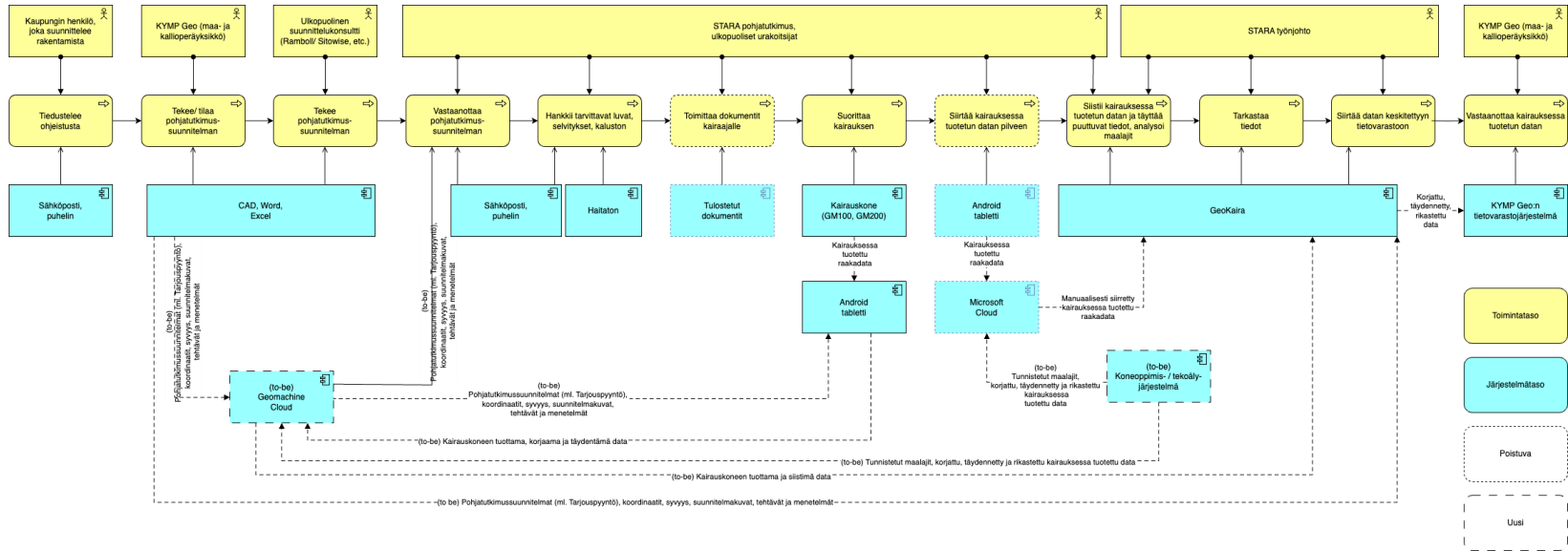
## Tekoälyn mahdollisuuksien kartoittaminen toimialalla

- Tunnistetaan nykyisen dataekosysteemin haasteet ja pullonkaulat.
- Kartoitetaan ja priorisoidaan käyttötapaukset tekoälyn hyödyntämiselle.

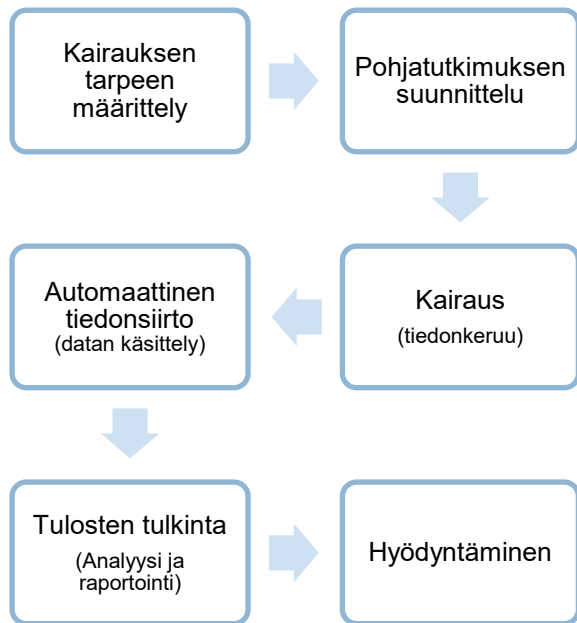
## Koneoppimisen/ tekoälyn pilotointi valitussa käyttötapauksessa

- Pilotoidaan kokeilu valitulle käyttötapaukselle.

# 5. Kokeilun tuotokset - Dataekosysteemin kuvaus



# 5. Kokeilun tuotokset – Tekoälyn mahdollisuudet pohjatutkimuksessa



## Suunnitelma:

Tekoäly lukee suunnitteluraportit ja muuttaa tiedot rakenteiseksi dataksi, jotka voidaan toimittaa kairajalle ilman manuaalisia välivaiheita.

Kairauksen yhteydessä tekoäly tunnistaa puuttuvat, väärinformatoidut tai epäloogiset tiedot, sekä korjaa ne automaattisesti.

Tekoäly tunnistaa automaattisesti mittaustulosten perusteella maalajit, korjaa kairauksesta johtuvat tyypilliset poikkeamat ja yhdistää eri lähteistä saatavan datan.

## Toteutuma:

Keskimmäistä tavoitetta saatiin hieman edistettyä, alin tavoite toteutui automaattisen tulkinnan osalta.

# Teknisestä toteutuksesta



## 1. Ympäristön valmistelu

- Luotiin virtuaaliympäristö (.venv), johon asennettiin tarvittavat Python-kirjastot (esim. pandas, numpy, scikit-learn).
- Konfiguroitiin Jupyter Notebook - sovellus, jolla PoC esitellään interaktiivisesti.
- Koodattiin Python-sovellus, joka aukaisee html-sivulle koontinäytön (dashboard) -> havaintoaineiston lähempi tarkastelu.
- Koodattiin 'postprocessing'-sovellus, joka korjaa ML-algoritmin (heuristisesti) projisoimat loogisesti virheelliset maaperätyypit todennäköisemmiksi maaperätyypeiksi -> validointi -> korjaus

## 3. Raakadatan luku, parsiminen ja yhdistäminen

- Useista kairausdatan tiedostoista (esim. \*.tek) luetaan mittausarvot ja metatiedot (TX, FO, KJ, TY, XY).
- Jokainen tiedosto sisältää TT-lohkoja (HP ja PO), joista poimitaan syvyys, vastus, iskujen määrä, koodi ja maalaji.
- Metatiedot liitetään jokaiseen datariviin, jolloin tiedot pysyvät yksilöitävinä ja yhdisteltävinä.
- Rivien pituudet normalisoidaan ja tekstimuotoiset numeeriset arvot muunnetaan oikeiksi luvuiksi (esim. Pandasin to\_numeric), jotta datan käsittely ja analyysi onnistuu luotettavasti.
- Maalajikenttä puhdistetaan: poistetaan numerot ja ylimääräiset merkit, täydennetään puuttuvat arvot.
- Kaikkien tiedostojen lohkot yhdistetään yhdeksi siivotuksi datasetiksi, joka tallennetaan jatkokäsittelyä varten.
- Datasetti esikäsitellään ML-prosessia varten (esim. maalajienkoodaus LabelEncoderilla).

# Teknisestä toteutuksesta: Analyysiprotokolla ja menetelmät

Tavoite: Automaattinen maaperätyyppien luokittelu porausdatan perusteella XGBoost-pohjaisella ML-pipeline -ratkaisulla

Data ja haasteet:

Piirteet (*feature-muuttujat*): *Depth, Value, Resistance, Blows*

- Kohde (*vastemuuttuja*): *SoilTypeEncoded* (maaperätyypin koodi)
- Haasteet:
  - Epätasapainoinen luokkajakauma → stratifioitu otanta
  - Harvinaiset luokat → heikko ennustettavuus
  - RAM-rajoitteet suurilla aineistoilla
- Pipeline-vaiheet:
  - Esikäsittely: puuttuvat arvot → 0, label mapping
  - Train-Test Split: stratifioitu jako
- Mallinnus:
  - Algoritmi: XGBoost (multi:softmax)
  - Parametrit: max\_depth=8, eta=0.05, subsample=0.8, tree\_method='hist'
  - Early stopping → *tehokkuus*<sup>1)</sup>
- Arviointi:
  - Mittarit: Accuracy, sekaannusmatriisi (*confusion matrix*)
  - Visualisoinnit: *piirteiden tärkeys*, luokkajakauma (selitysvoimien vertailu)
- Tulosteet: *projisointi ("ennustaminen")* → `soil_type_predictions_xgb.csv` → loogisesti mahdollottomat projisoidut kerrostumat korjataan automaattisesti validointialgoritmin perusteella → `postprocess_predictions.py` → `soil_type_predictions_xgb_corrected.csv`

<sup>1)</sup> [Koulutusprosessi lopetetaan ennen kuin kaikki ennalta määritellyt iteraatiot (boosting-kierrokset) on suoritettu; tämä silloin, jos mallin suorituskyky ei enää parane tietyn määrän jälkeen peräkkäisiä kierroksia.]

**Huomio suorituskyvystä:** suuret datasetit kannattaa esisuodattaa tai käyttää *server-side filtering*-mallia; turhat NaN-rivit poistetaan jo latauksessa.

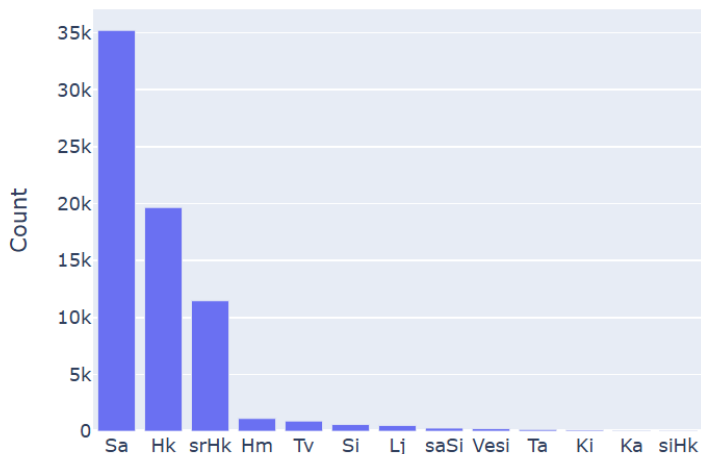
• **Rajoitteet ja vastuut:** heuristinen suositus ei korvaa geoteknistä ammattilaisarviota; CSV:n kenttien laatu (mm. SoilType-puhdistus ja ffill) vaikuttaa tulokinnan luotettavuuteen.

Maaperätyyppien jakauma aineistossa:

### Soil Type Distribution



Soil Type Distribution



### Summary Statistics

- Filtered Records: 70268
- Selected Soil Types: All
- Selected Locations: All
- Average Depth: 7.14 m
- Average Resistance: 4.27
- Average Blow Count: 25.66

## Lisättiin muutamia loogisia rajoituksia datan parantamiseksi

Kalliota ei yli 3m pohjasta (johtuen kairaustavasta)  
Kalliopinnan alla ei muita maalajeja  
Liejua vain pinnassa (tai veden alla)  
Jne.

### Mallinnuksen tunnuslukuista:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Observations}}$$

#### Mitä tarkoittaa?

- Oikein projisoitujen tapausten määrä jaettuna havaintojen lkm:llä.
- Käsittää sekä true positives että true negatives jakolaskun osoittajassa.

#### Esimerkki:

- 100 havainnosta oikein projisoituja on 90 kpl, silloin accuracy = 0.90

#### Milloin johtaa harhaan?

- Epätasapainoisissa aineistoissa, accuracy näyttää korkealta jos yleiset luokat projosoidaan oikein mutta pienet ryhmät väärin.
- Esimerkki: jos 95 % datasta on luokkaa A, ja nämä projosoidaan kaikki oikein (mutta muut väärin), antaa 95 % Accuracyn, mutta lähes mitättömän käyttökelpoisuuden.

**Macro F1-score** yhdistää **precision-** and **recall-** suuret yhdeksi mittariksi, joka kuvaa mallin käyttökelpoisuutta harmonisesti eritoten luokittelutehtävissä

**Precision** = Kuinka monta projisoiduista positiivisista () on mennyt oikein.

**Recall** = Kuinka monta todellisuudessa positiivista on tunnistettu oikein.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

### Miksi hyödyllinen?

Balansoi sekä precisionin että recallin, joten se ottaa huomioon kummankin tyyppiset virheet, false positives and false negatives.

Korkea F1-score tarkoittaa, että malli toimii hyvin molemmista näkökulmista.

Tärkeä mittari etenkin epätasapainoisissa dataseteissa, joissa accuracy yksin voi olla harhaanjohtava.

### Tulkinta:

**F1 = 1** → Täydellinen onnistuminen

**F1 = 0** → Malli sakkaa täydellisesti

## Tulokset ja tarkastelu

- Testitarkkuus (Accuracy): 79,31 %
- Macro F1-score: 0,4453

→ Hyvä tarkkuus, mutta F1 kertoo epätasapainosta luokkien välillä.

## Piirteiden tärkeys

Piirteiden tärkeys (Feature Importance)

- Normalisoitu gain:
- Resistance: 37,44 %
- Value: 35,61 %
- Blows: 13,79 %
- Depth: 13,15 %

Osa arvoista muita tärkeämpiä ennustavia tekijöitä.

LocationID	Depth	Value	Resistance	Blows	SoilTypeEn	SoilType	PredictedE	PredictedSoilType
20280	Vall	0,2	11			14 srHk	14	srHk
20280	Vall	0,4	8			14 srHk	14	srHk
20280	Vall	0,6	4			14 srHk	14	srHk
20280	Vall	0,8	7			14 srHk	14	srHk
20280	Vall	1	12			14 srHk	14	srHk
20280	Vall	1,2	10			14 srHk	14	srHk
20280	Vall	1,4	12			0 Hk	14	srHk
20280	Vall	1,6	4			0 Hk	14	srHk
20280	Vall	1,8	7			6 Sa	14	srHk
20280	Vall	1,84		0,926	0	6 Sa		6 Sa
20280	Vall	1,88		0,939	0	6 Sa		6 Sa
20280	Vall	1,92		0,914	4	6 Sa		6 Sa
20280	Vall	1,96		0,914	4	6 Sa		6 Sa
20280	Vall	2		0,907	4	6 Sa		6 Sa
20280	Vall	2,04		0,907	0	6 Sa		6 Sa
20280	Vall	2,08		0,926	0	6 Sa		6 Sa
20280	Vall	2,12		0,914	0	6 Sa		6 Sa

# 6. Opit kokeiltavan ratkaisun tai toimintatavan mahdollisuuksista

Menetelmät ja mitä opittiin niiden käyttömahdollisuuksista

- Koneoppiminen:  
Vaikuttaa lupaavalta menetelmältä, epävarmaa kuinka pitkälle sen varmuus on vietävissä. Ihmiselle ilmeiset seikat saattavat olla koneelle vaikeita ymmärtää (sorapinnan alla ei esiinny luonnossa irtovettä)
- Logiikkapohjainen toiminta:  
usein löytyy erikoistapaus, jossa ei toimi. Esim. ajatus että mitataan vain, kun laitteessa on paino päällä on hieno, mutta entäpä jos maa onkin niin pehmeää että tangot painuvat siihen itsestään? -> vaara tulosten hukkaamisesta entistä pahemmin.

# 7. Opit asiakkaiden tai palvelun käyttäjien tarpeista

- Ei päästy niin pitkälle, että ratkaisua olisi saatu käytettyä asiakkaiden tarpeisiin, tähän saakka tehdyille toimenpiteille me olemme loppuasiakas.

# 9. Opit kokeilemisesta yleensä

Tarvitaan paljon enemmän aikaa. Ja tietysti jotta tästä saataisiin käytännön hyötyä, homma pitäisi viedä paljon pidemmälle. Joten tarvittaisiin myös rahaa.

# 12. Kokeilun data

- Mitä kokeilun datasta ja sen käsittelystä opittiin?
  - Dataa on paljon. Jotta voisi olla varma siitä, mitä koneelle oikeastaan opetetaan, pitäisi ehtiä käymään melko hyvin läpi sekä lähtödata että se, miten ohjelma sitä lukee. Melko viime tingassa esim. huomattiin, että kohteet, joissa ei ollut maalajeja lainkaan merkittyinä (joita ei pitänyt opetusdatan joukossa olla), ohjelma luki kokonaan kallioksi, joka tietysti haittasi tuloksia.

# 13. Jatkopäätökset ja -ideat

- Millaisia päätöksiä jatkosta on tehty ja millä perustein?  
Ei ole tehty päätöksiä, jatkorahoituslähdettä ei tunneta.

# Skaalauspohja 1/2

## 1. Eniten hyötyvä asiakas

Urakoitsija (esim. Stara) itse sekä datojen vastaanottaja (Geo)

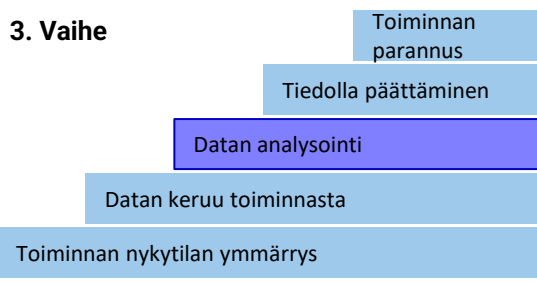
## 2. Kokeilun tyyppi

Nykyisen toiminnan tehostaminen

Toiminnan parantaminen

Strateginen parannus tai täysin uusi lähestymistapa

## 3. Vaihe



## 4. Toteutus

Helppo

Keskivaikea

Vaikea

Tavoiteltu tilanne, jossa data juoksee järjestelmien välillä automaattisesti, työn sisältö on kairaaajilla jatkuvasti hallussaan, työnjohto näkee välittömästi työn vaiheen ja kaiken tavallisen kairadatan tulkinta on mahdollista, käyttää maksimimäärää koneen mittaustuloksista hyväkseen ja toimii suurelta osin automatisoidusti.

## 5. Nykytilanne lukuina

Pohjatutkimusyksikön keskeistä toimintaa. Kymppä ostaa tätä vuosittain meiltä yli miljoonalla, urakoitsijoilta useilla miljoonilla. Lisäksi suomen kairausalalla käytetään nykyistä tulkintaohjelmaa yleisesti muutenkin kuin Helsingin kanssa asioitaessa, joten vaikutus on maan tasolla paljon suurempi.

## Oletukset ja epävarmuudet

Ei tiedetä, saadaanko tähän koskaan rahaa.

Alle 6kk

1-2 vuotta

ei tiedetä

Aika ensimmäiseen hyötyyn

# Skaalaus pohja 2/2

## 6. Nykytilanteen Flow

1. Geon projektipäällikkö lähettää datat sähköpostilla Staran työnjohtolle, joka printtaa tekstit kairaajalle
2. Kairaaja kairaa käyttäen käytettävissä olevia tietoja, työnjohto ja asiakas pysyttelevät tilanteen tasalla kysymällä.
3. Data otetaan järjestelmästä infra-formaatissa, iso osa laitteen mittaamista tiedoista menetetään
4. Data tulkitaan eri ohjelmassa käsin. Kaikelle datalle ei ole tulkintatoimintoja, tulkitsijan osaamisen taso vaihtelee.
5. Data lähetetään kaupungin järjestelmiin.

## 7. Tavoitetilanteen Flow

1. Järjestelmä lukee datapaketin sisällön järjestelmään automaattisesti. Data kulkeutuu kairaajan koneelle valmiiksi, josta hän näkee reiän tiedot työtä tehdessään.
2. Työnjohto ja asiakas näkevät työn etenemisen reaaliajassa
3. Data saadaan käyttöön mahdollisimman täytenä, data ei huku.
4. Data tulkitaan samassa järjestelmässä. Kaikille datatyypeille löytyy tulkintatoiminnot. Tulkinta on suurelta osin automaattista, käyttäjä tarvitaan vain tarkastamaan ja korjaamaan koneen mahdolliset väärinkäsitykset.
5. Data lähetetään kaupungin järjestelmiin

## 8. Hyödyt tavoitetilassa

Datansiirto nopeutuu, työn seuranta tarkentuu, dataa tulkittaessa on käytettävissä maksimimäärä informaatiota, tulkinta nopeutuu, tulkinnan henkilöriippuvuus vähenee. Kaupunki saa parempaa data halvemmalla.

# Liite 1 - Dataekosysteemin systeemikuva

