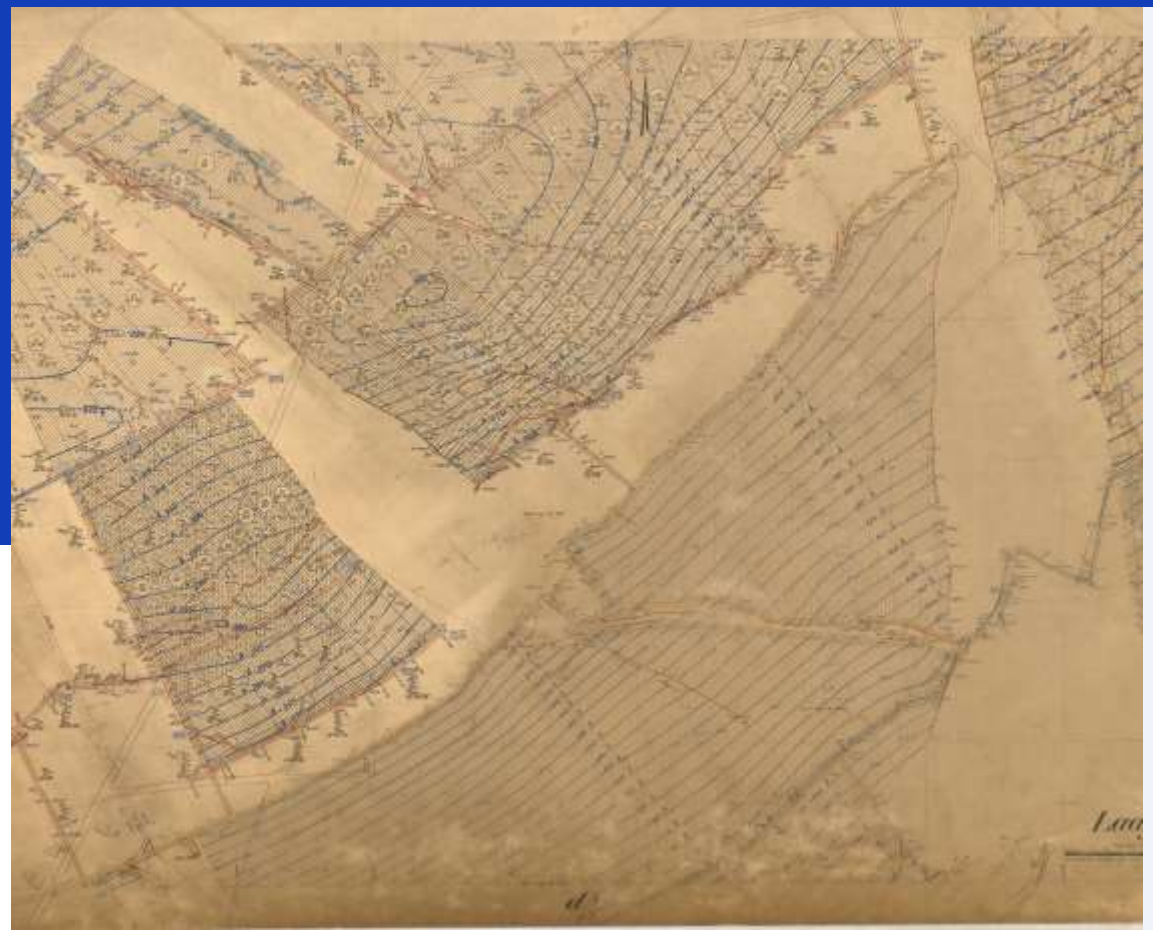
Vectorizing Mining maps

TNO GDN | GDIM

Data Science Team & AGE
Cees van Middelkoop, David Demmers,
Samantha Kim, Joana Esteves Martins, Erik van Linden

Agenda



1. Team
2. Background & Scope
3. Features
4. Point data – Regex & Clustering
5. Point data – Vision Language model
6. Shape data – Semantic segmentaiton
7. Results
8. Take aways

Introduction of the Team

Joana Esteves Martins
- Remote Sensing Specialist
- EO Scientist

Joop Hasselman
- Manager International Projects

Erik van Linden
- Geologist
- Expert Mining Maps

Samantha Kim
- Scientist Innovator
- Data Assimilation

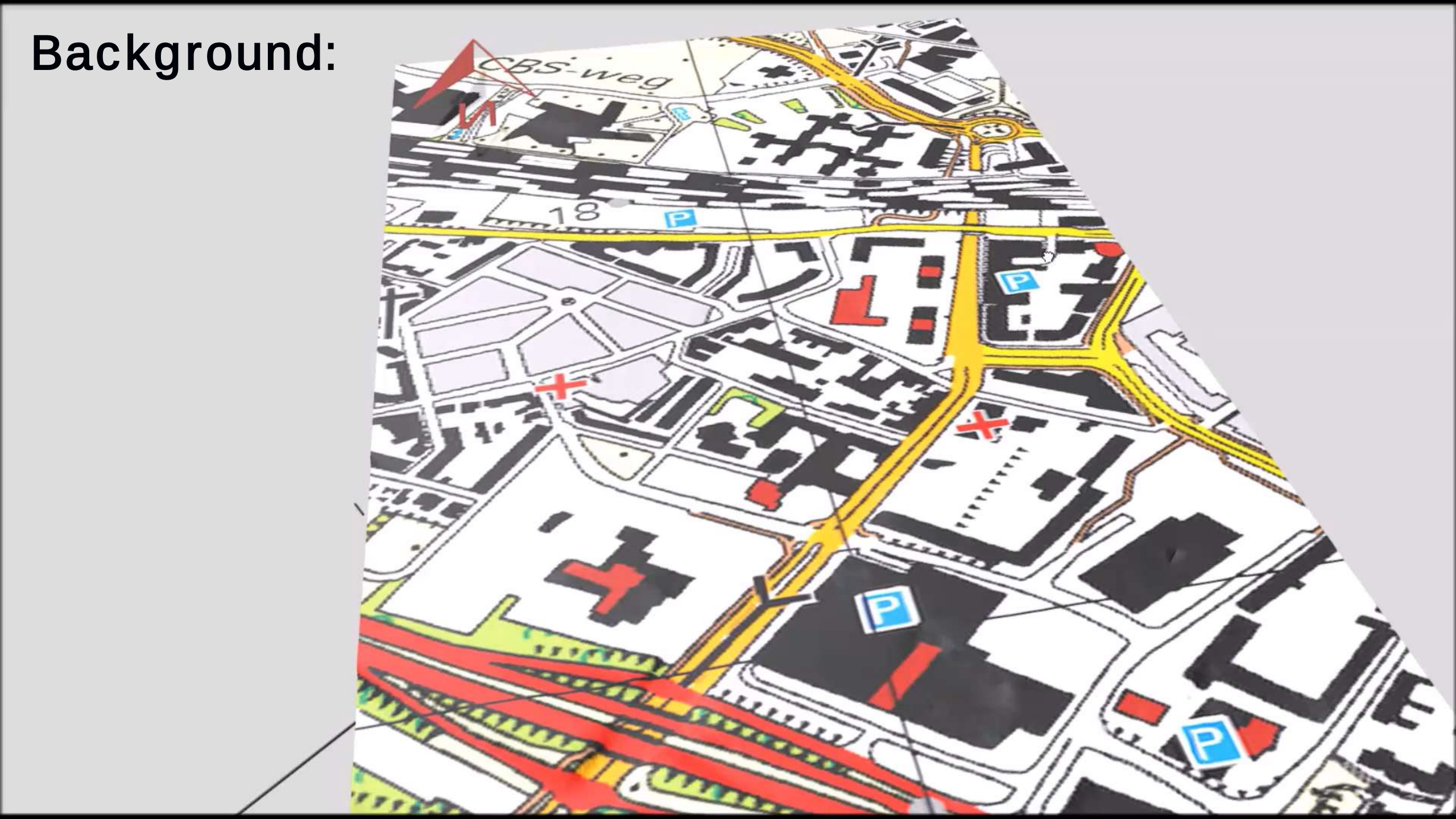
Cees van Middelkoop
- Data Scientist
- Python Developer

David Demmers
- Data Scientist
- Python Developer

Wilfred Visser
- Product Owner

TNO Geologische Dienst Nederland

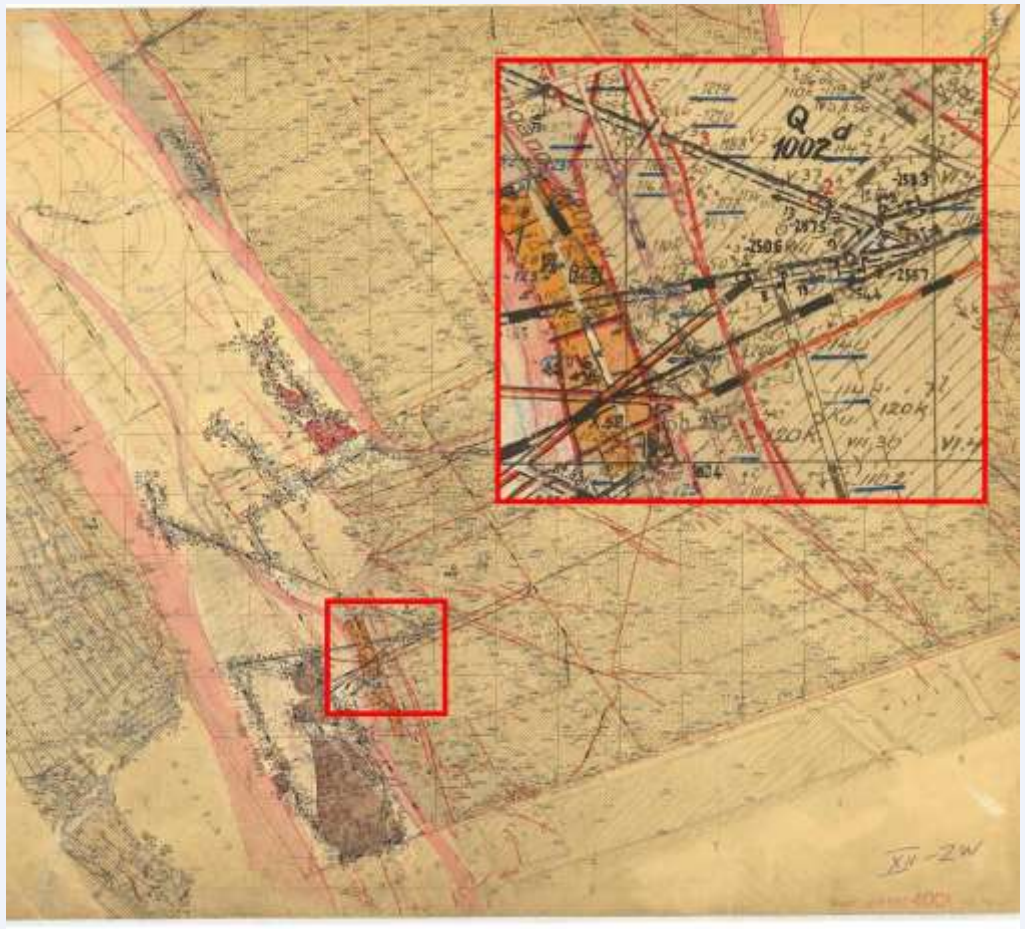
Background:



Project scope

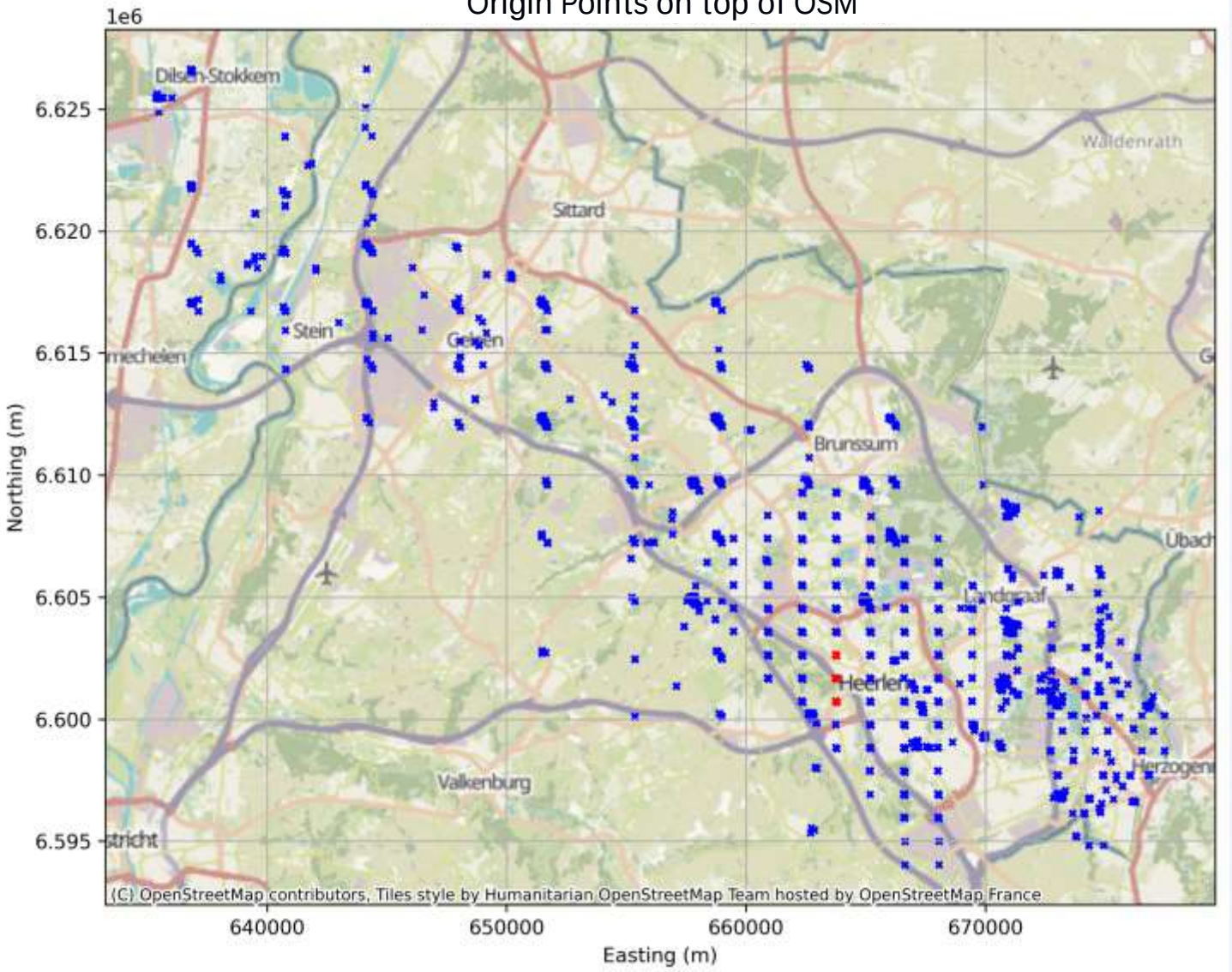
Digitalisation/ Vectorization of scanned mining maps.

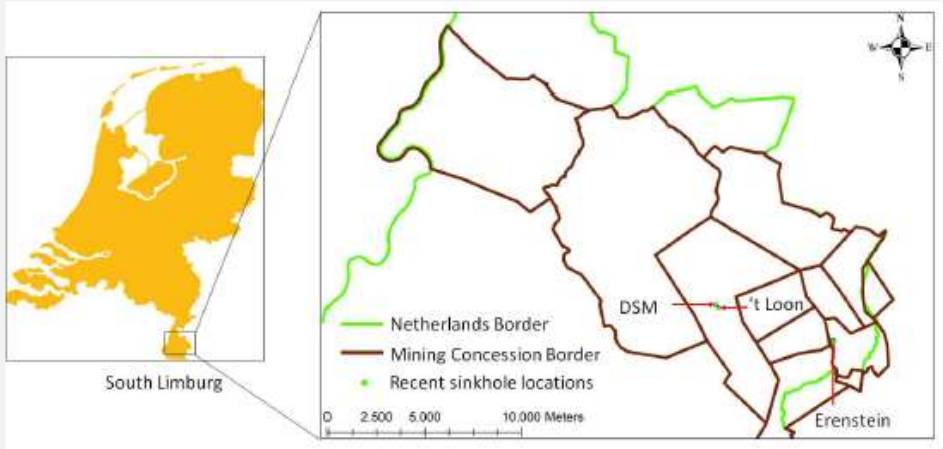
- A subset of n ≈ 3.500 scanned maps to be processed, sliced into 345.323 [1024px *1024px] imgs.
- Significant variation in maps due to differing mining operations over a long timespan.
- With the purpose of; internal research, relative probability map of latent mining effects, and external engagement & usage.
- Several features each with variation in representations: mining panels, galleries, depth values and temporal data



Maps in the real world

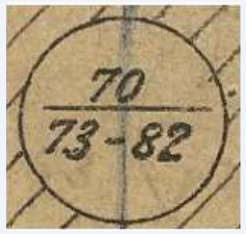
Origin Points on top of OSM

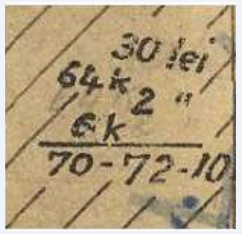


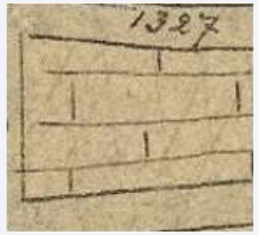


TNO Geologische Dienst Nederland

Feature types

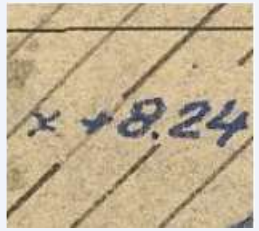










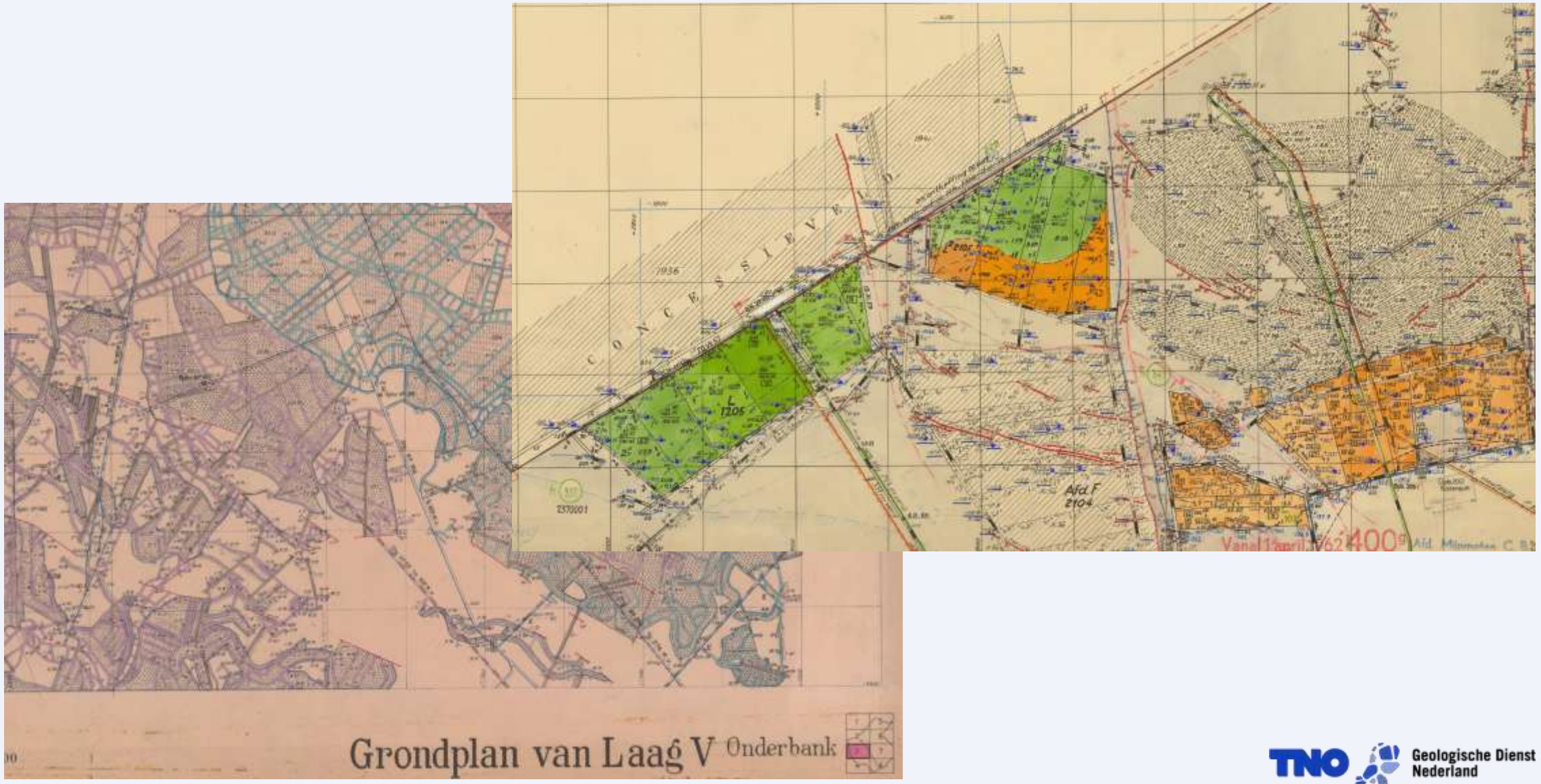


Feature of interest

Galleries & Panels:

Dates & Depths:

Challenge I: Variety in feature expression

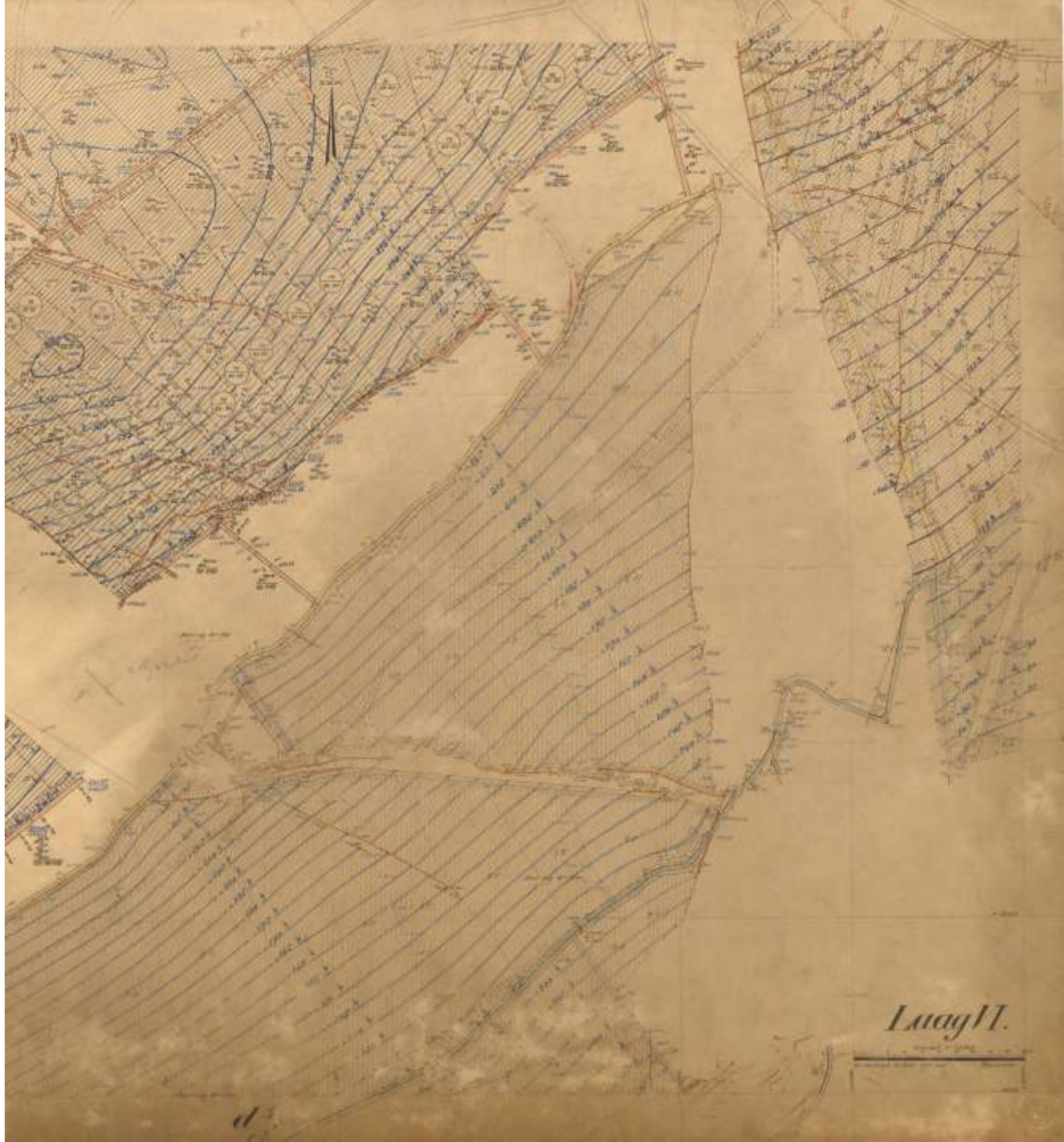


Challenge II: Real-world data is messy, and sometimes missing

Vectorizing Mining Maps

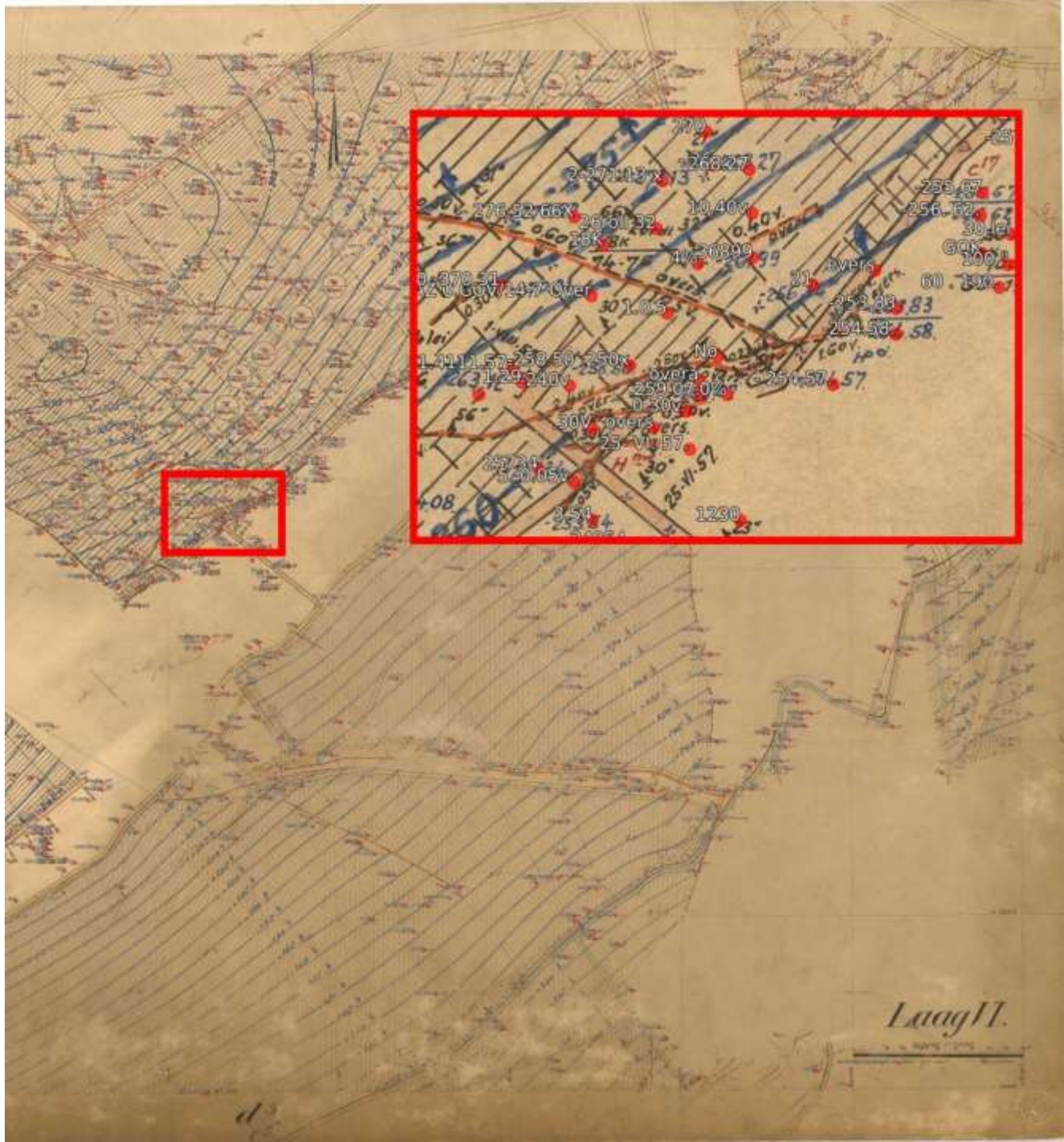
Point data extraction:
OCR, Regex & Clustering

- Original maps are only images
- Extracting text data



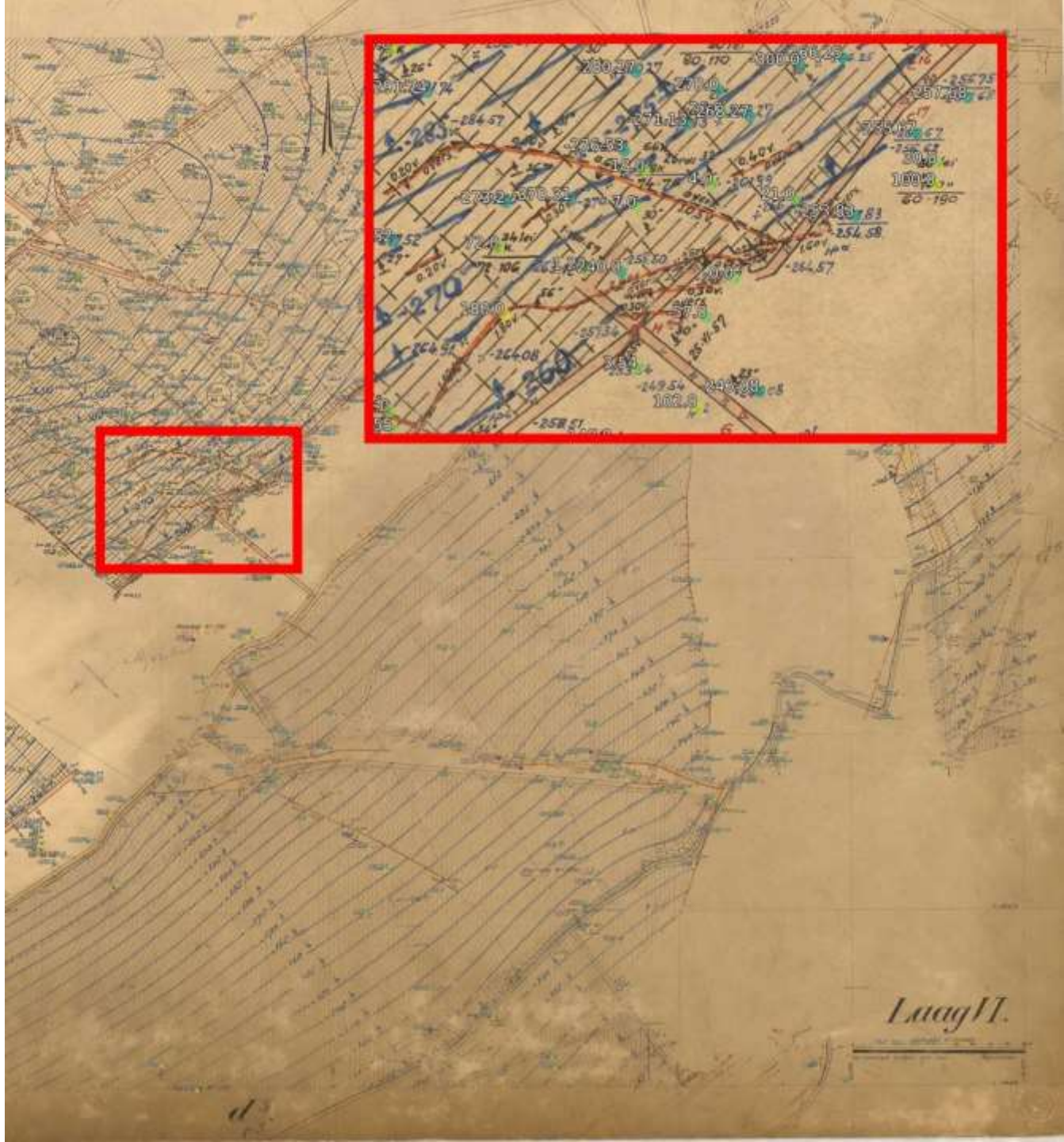
Point data extraction:
OCR, Regex & Clustering

- Original maps are only images
- Extracting text data
- Regex rules for transformations & grouping:
  - Several date standards.
    (YYYY/ YY/ YYYY-MM/ YY-MM/ Roman YYYY/ Roman YY-MM/ ...)
  - Several depth standards
    - Negating possible depths



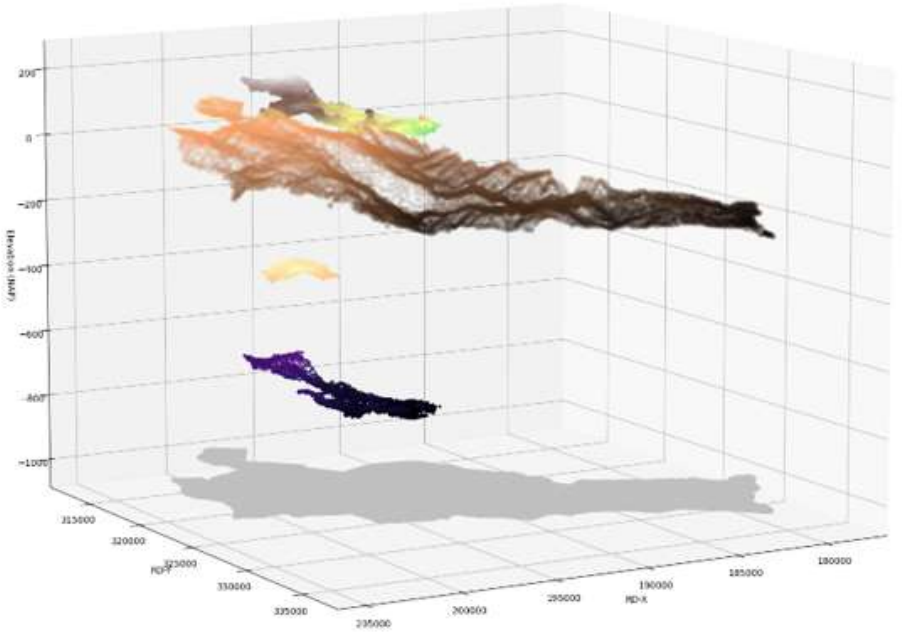
Point data extraction:
OCR, Regex & Clustering

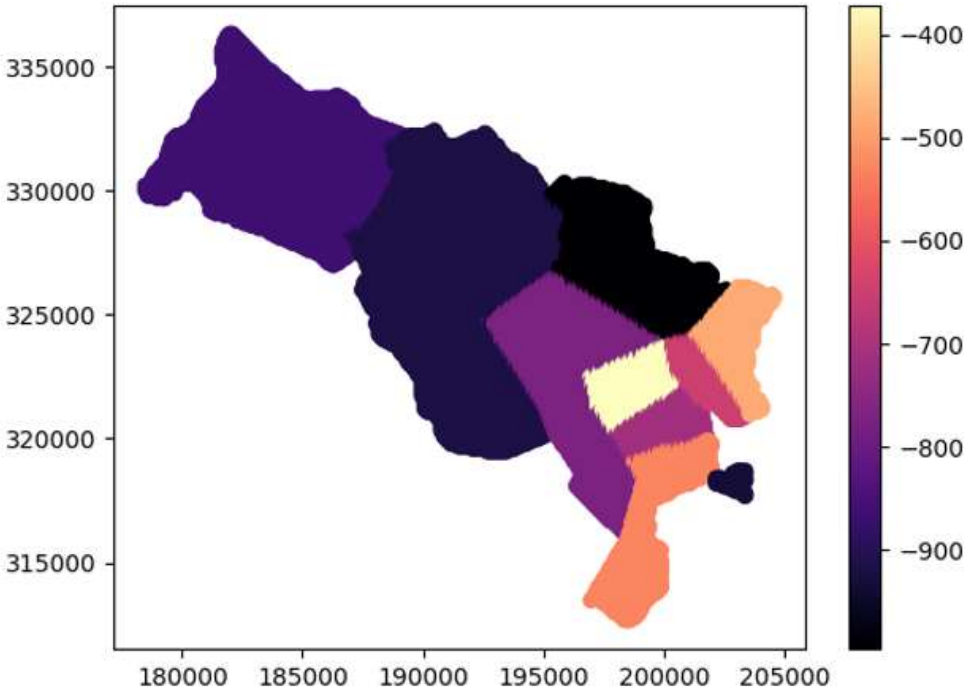
- Original maps are only images
- Extracting text data
- Regex rules for transformations & grouping :
  - Several date standards.
    (YYYY/ YY/ YYYY-MM/ YY-MM/ Roman YYYY/ Roman YY-MM/ ...)
  - Several depth standards
    - Negating possible depths
  - Removing: Outliers, Angles, Yields



Point data extraction:
OCR, Regex & Clustering

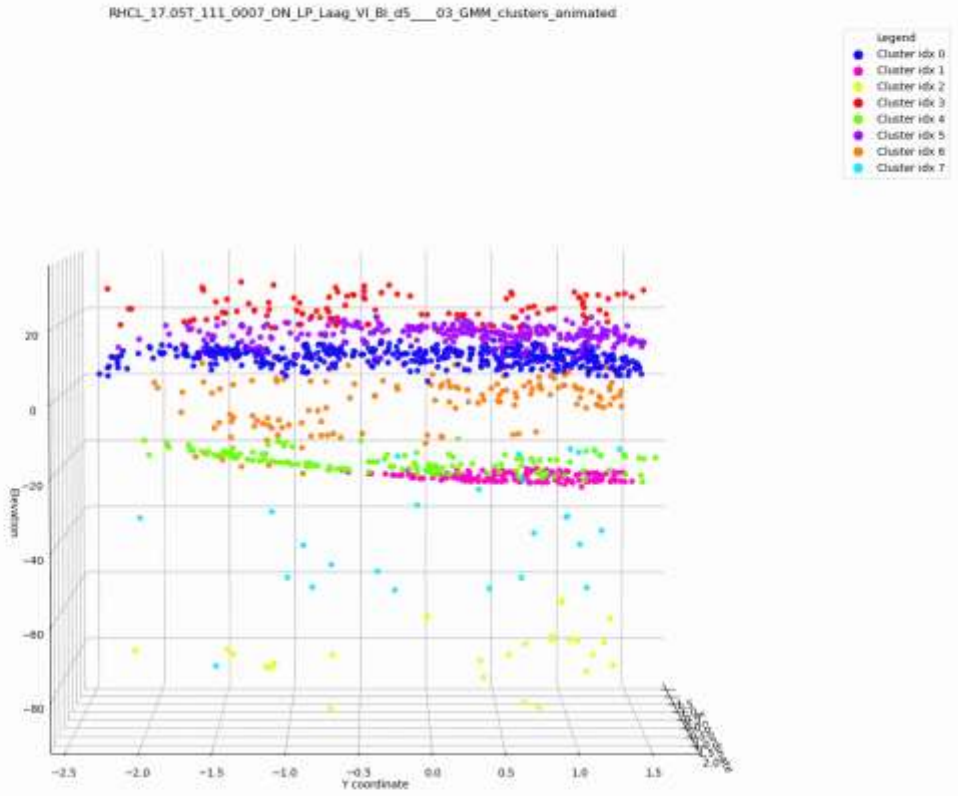
- Original maps are only images
- Extracting text data
- Regex rules for transformations & grouping:
  - Several date standards.
    (YYYY/ YY/ YYYY-MM/ YY-MM/ Roman YYYY/ Roman YY-MM/ ...)
  - Several depth standards
    - Negating possible depths
  - Removing: Outliers, Angles, Yields
  - Constraining depths based on carboniferous layer & deepest shaft per concession.





Point data extraction:
OCR, Regex & Clustering

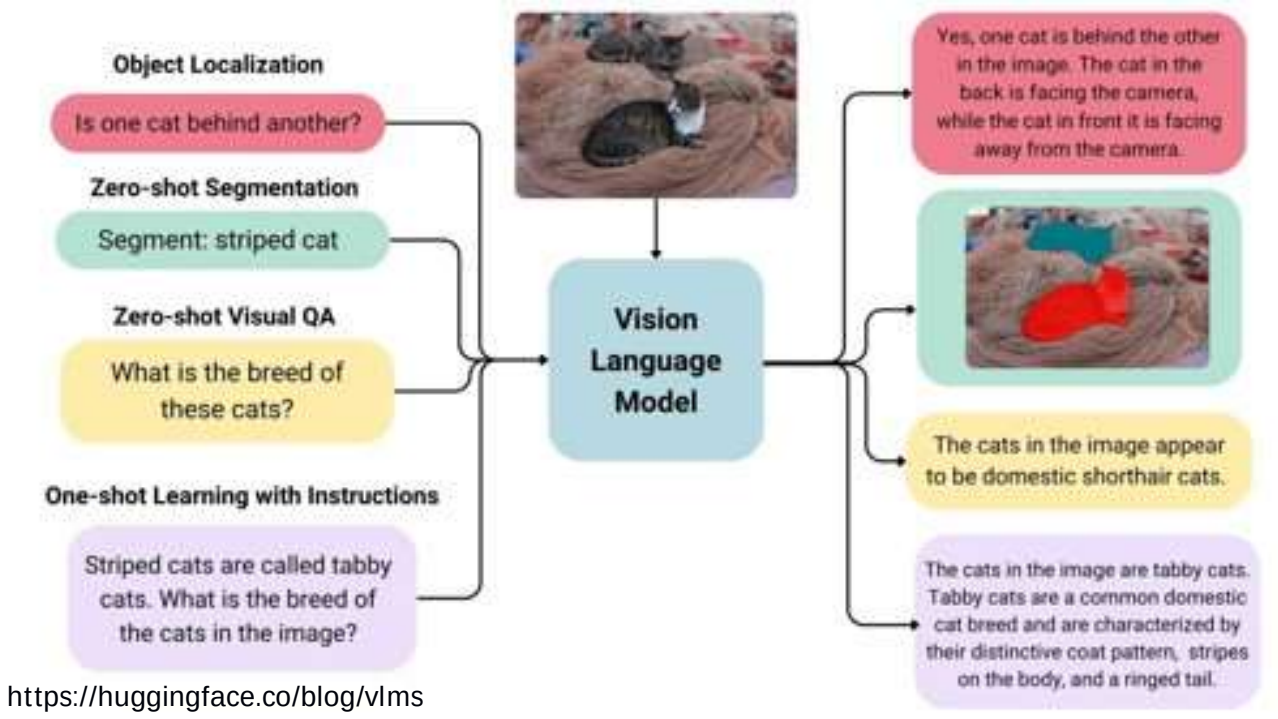
- Original maps are only images
- Extracting text data
- Regex rules for transformations ~~& grouping~~:
  - Several date standards.
    (YYYY/ YY/ YYYY-MM/ YY-MM/ Roman YYYY/ Roman YY-MM/ ...)
  - Several depth standards
    - Negating possible depths
  - ~~Removing: Outliers, Angles, Yields~~
  - Constraining depths based on carboniferous layer & deepest shaft per concession.
- Group point data with clustering.



Issues with rule based and unsupervised methods.

- Output quantity limited by OCR.
- Non exhaustive rule set for text transformation by Regex.
- Complex geometry of point values → dept values often follow complex shapes.
- Feature engineering to select clusters of interest → hard on large varied nonlinear sets.

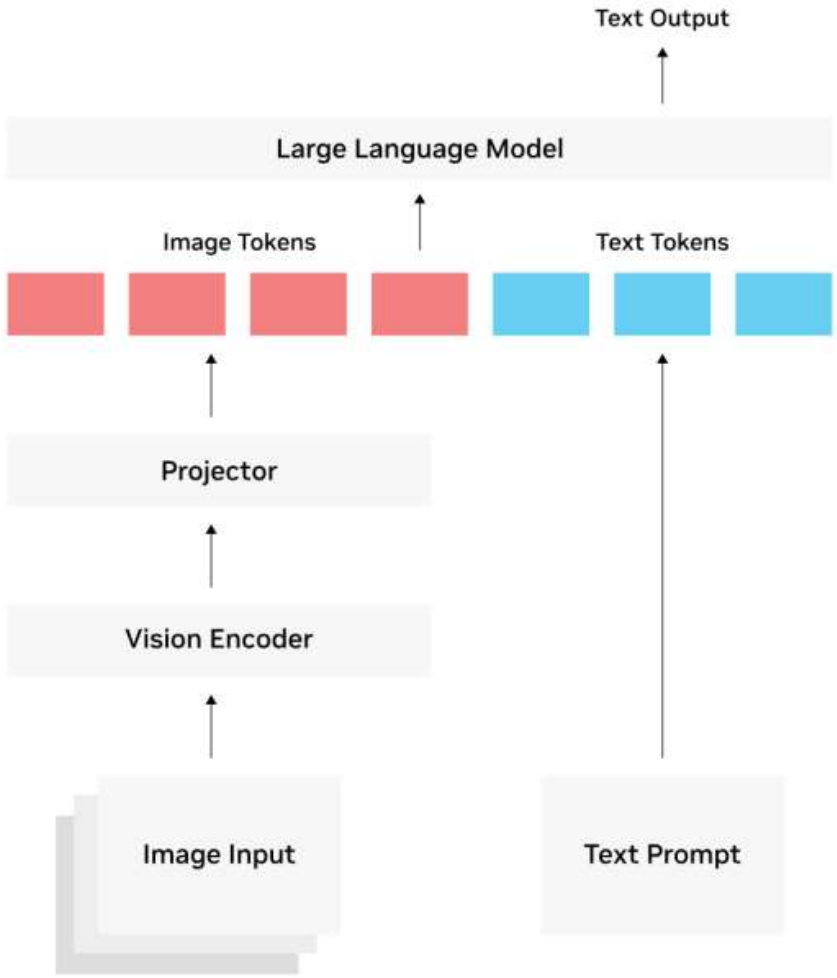
Sunk cost fallacy → Pivot to Vision language model for point data extraction & classification.



https://huggingface.co/blog/vlms

Point data extraction:
Vision Language Model

- Combining all available data sources.
  - Image, OCR, Labelled data
- Constraining the output with OCR.
- More context from the image
- Learning by example

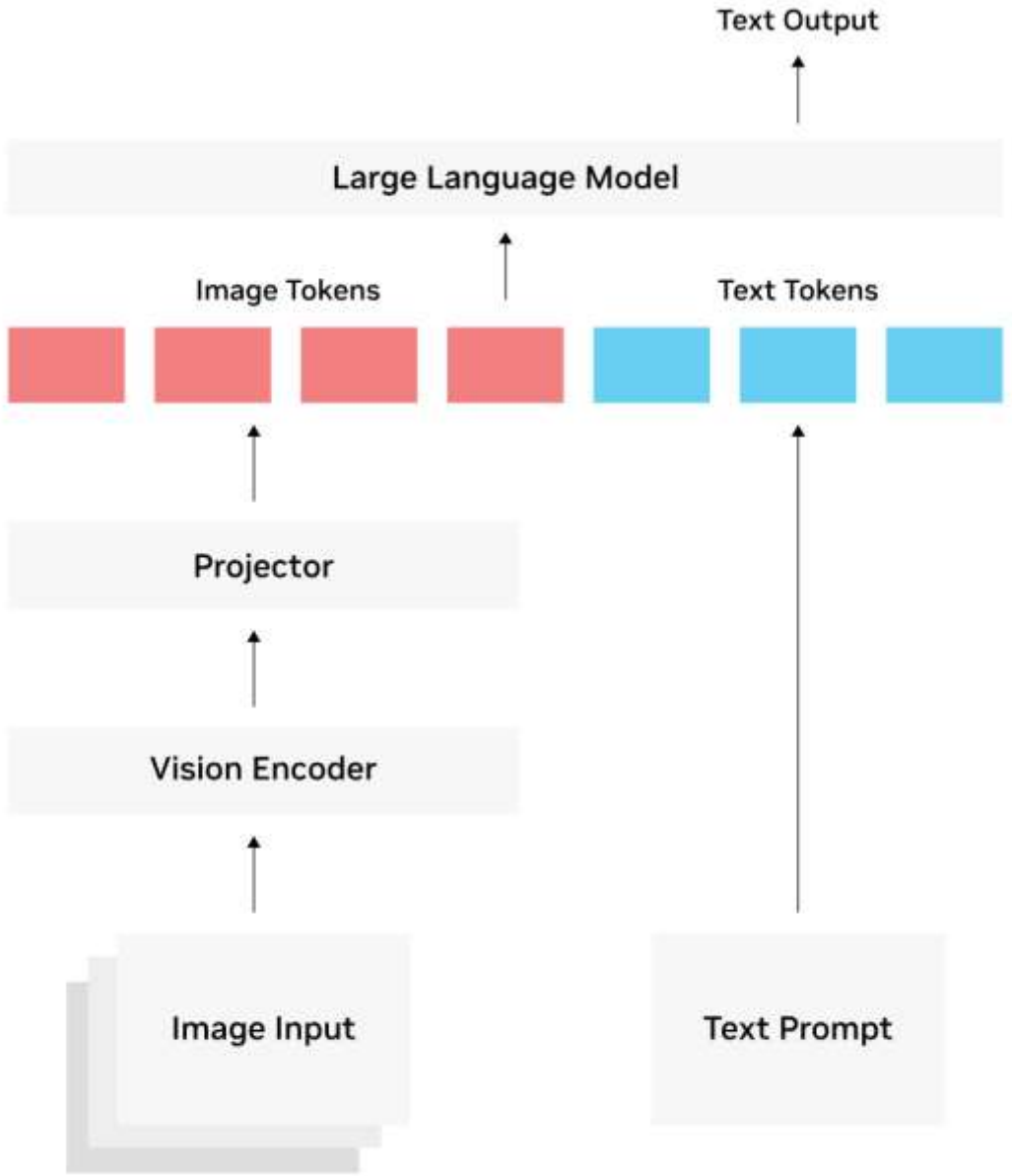


Point data extraction:
Vision Language Model

Qwen 2.5 72B

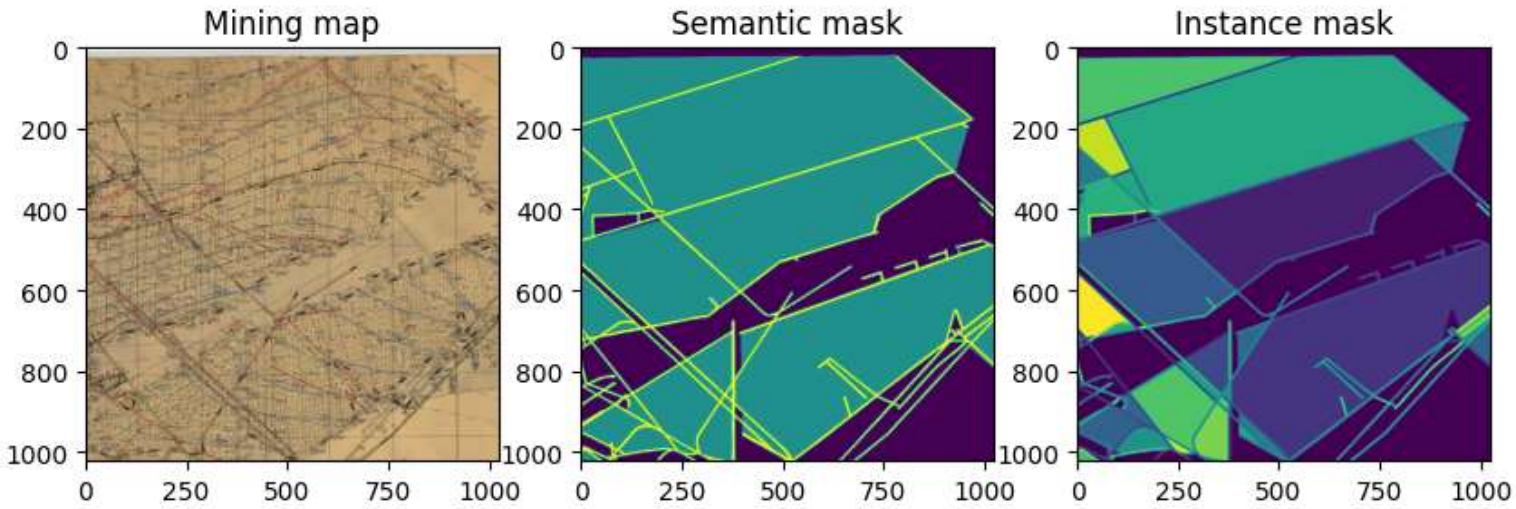
Practically:

- Due to time-crunch & budget constraints no finetuning or distilling was possible.
- Requirement of the largest available model to approach best performance.
- Computational (cost) challenges of large cloud compute clusters & their availability.
- Measures of scale.
- Required JSON format + Hard task = hallucinations → Low temperature with invalid JSON.



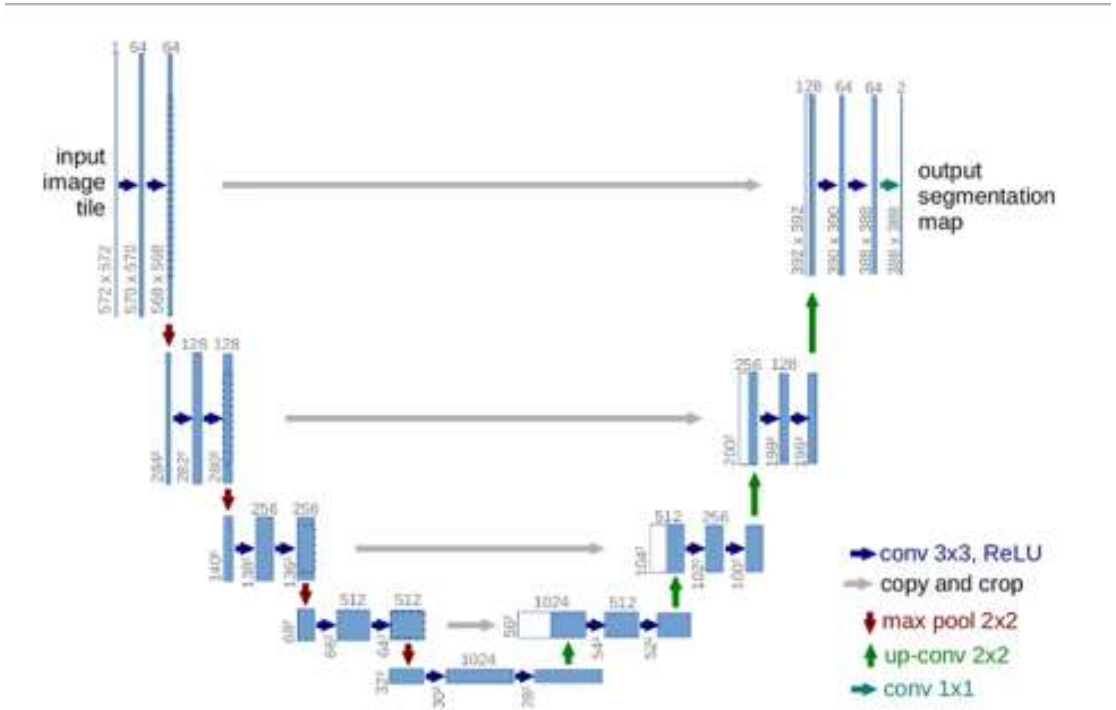
Shape data extraction:
Semantic Segmentation

- Acquisition of training data
- Multiple scale variation - spatially informed Train, Test & Validation split selection.
  - A base selection of all patch from its origin maps.
    (To avoid spatial correlation & data leakage)
- Labelling service: 4096px x 4096px imgs.
  - Down sampling (linear interpolation)
  - 2048px → 512px
  - 1024px → 512px
  - 512px → 512px



Shape data extraction:
Semantic Segmentation

- Acquisition training data
- Choosing a model architecture – U-Net:
  - Proven structure for pixel-wise class prediction.
  - No multiscale objects.
  - Generalisation through abstraction.
  - Keep context of higher order features from skip connections.

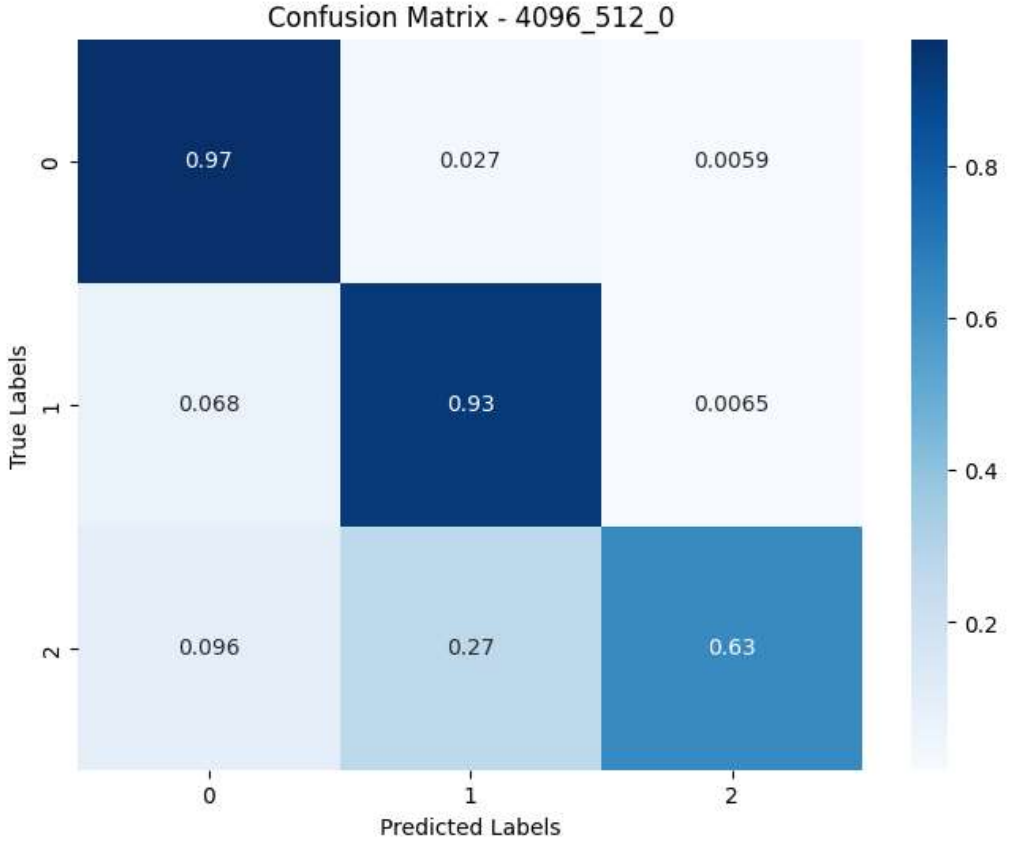


Shape data extraction:
Semantic Segmentation

- Acquisition training data
- Choosing a model architecture – U-Net:
  - Proven structure for pixel-wise class prediction.
  - No multiscale objects.
  - Generalisation through abstraction.
  - Keep context of higher order features from skip connections.
- Experimenting with scale – Trade-off between detail & context.

Gain detail → lose context

- Slower processing – More patches
- Original details are useful for informing minority classes. Will trip-up majority class prediction.

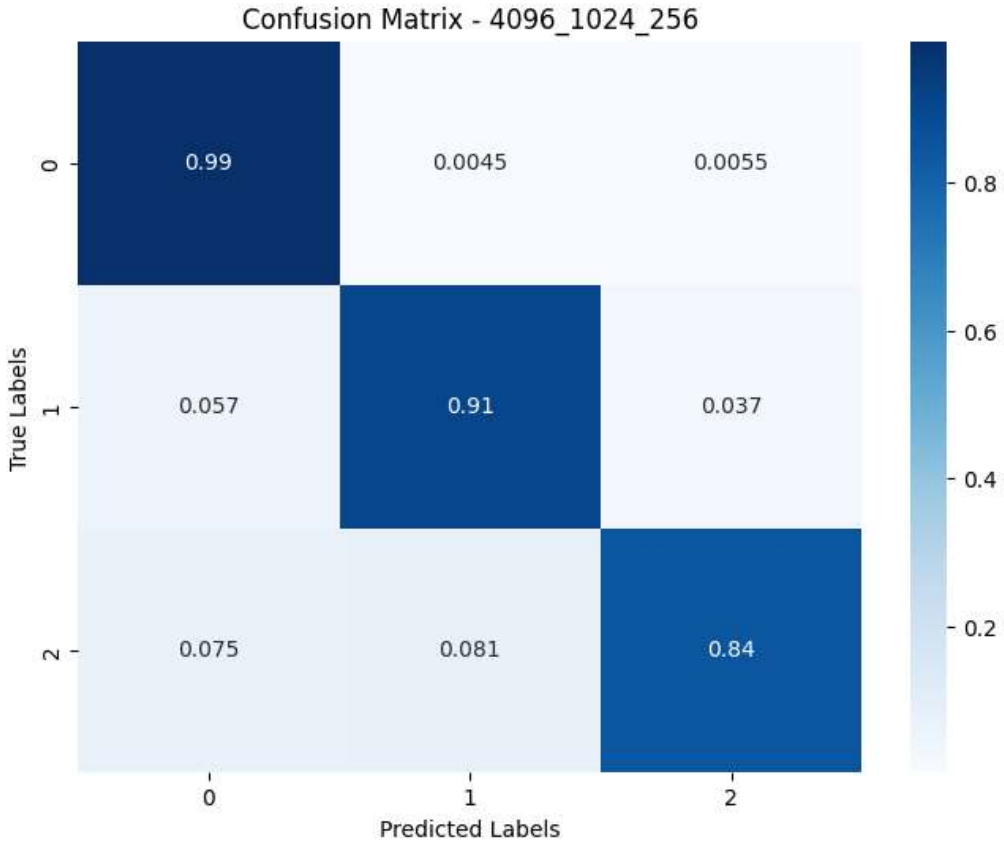


Shape data extraction:
Semantic Segmentation

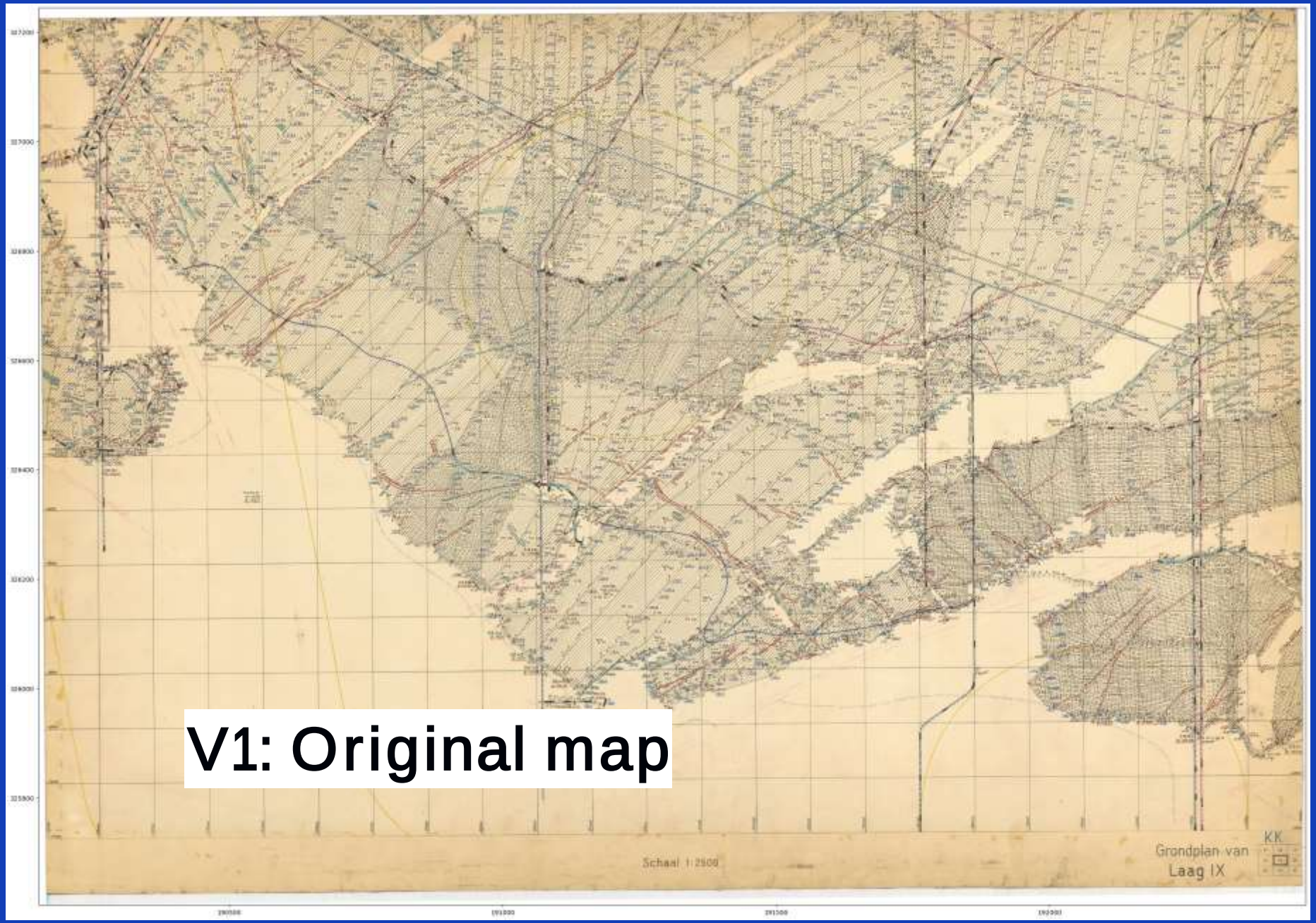
- Acquisition training data
- Choosing a model architecture – U-Net:
  - Proven structure for pixel-wise class prediction.
  - No multiscale objects.
  - Generalisation through abstraction.
  - Keep context of higher order features from skip connections.
- Experimenting with scale – Trade-off between detail & context.

Lose detail → gain context

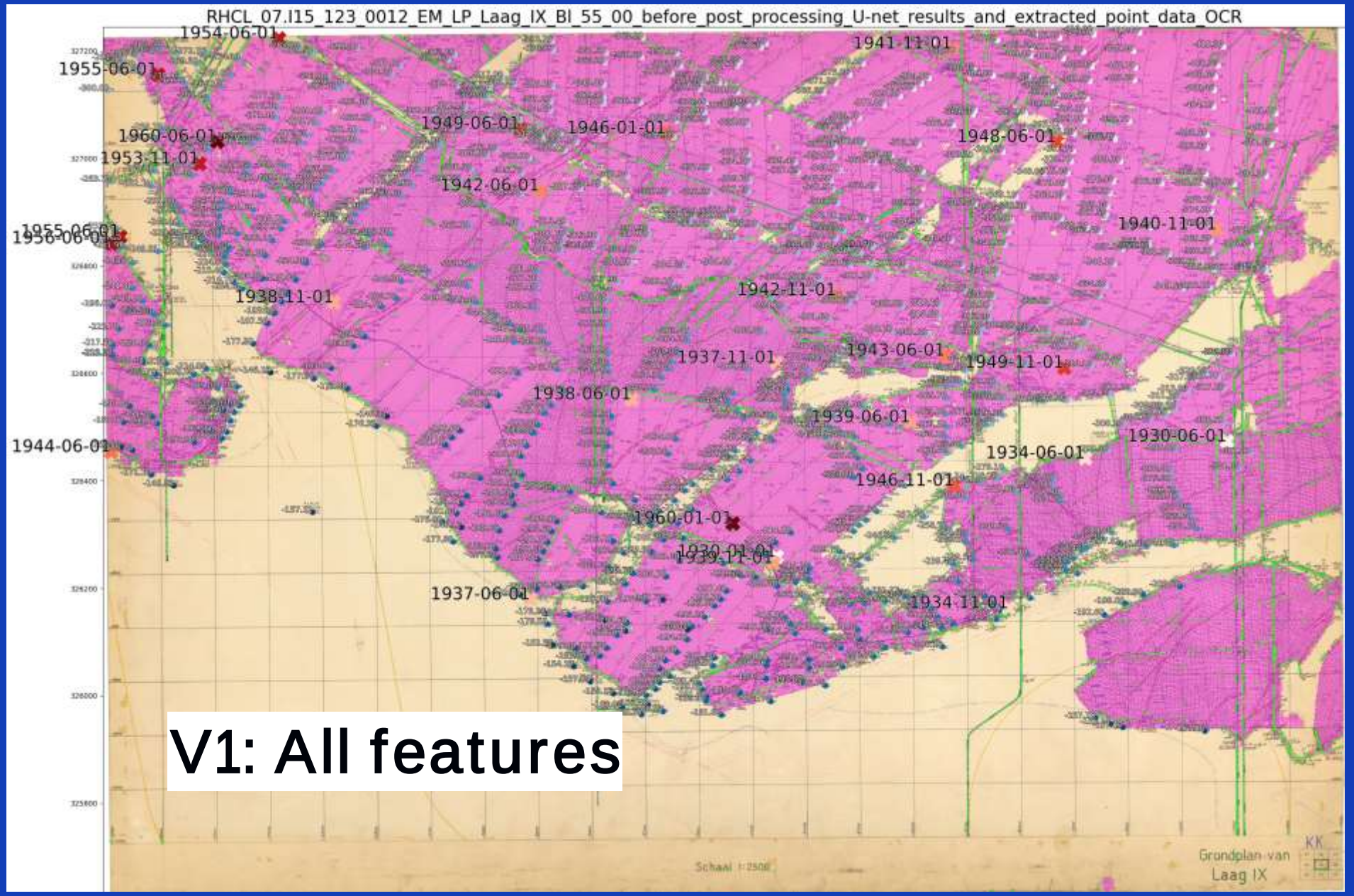
- Faster processing – Fewer patches
- Averaged out information benefit generalization on majority classes



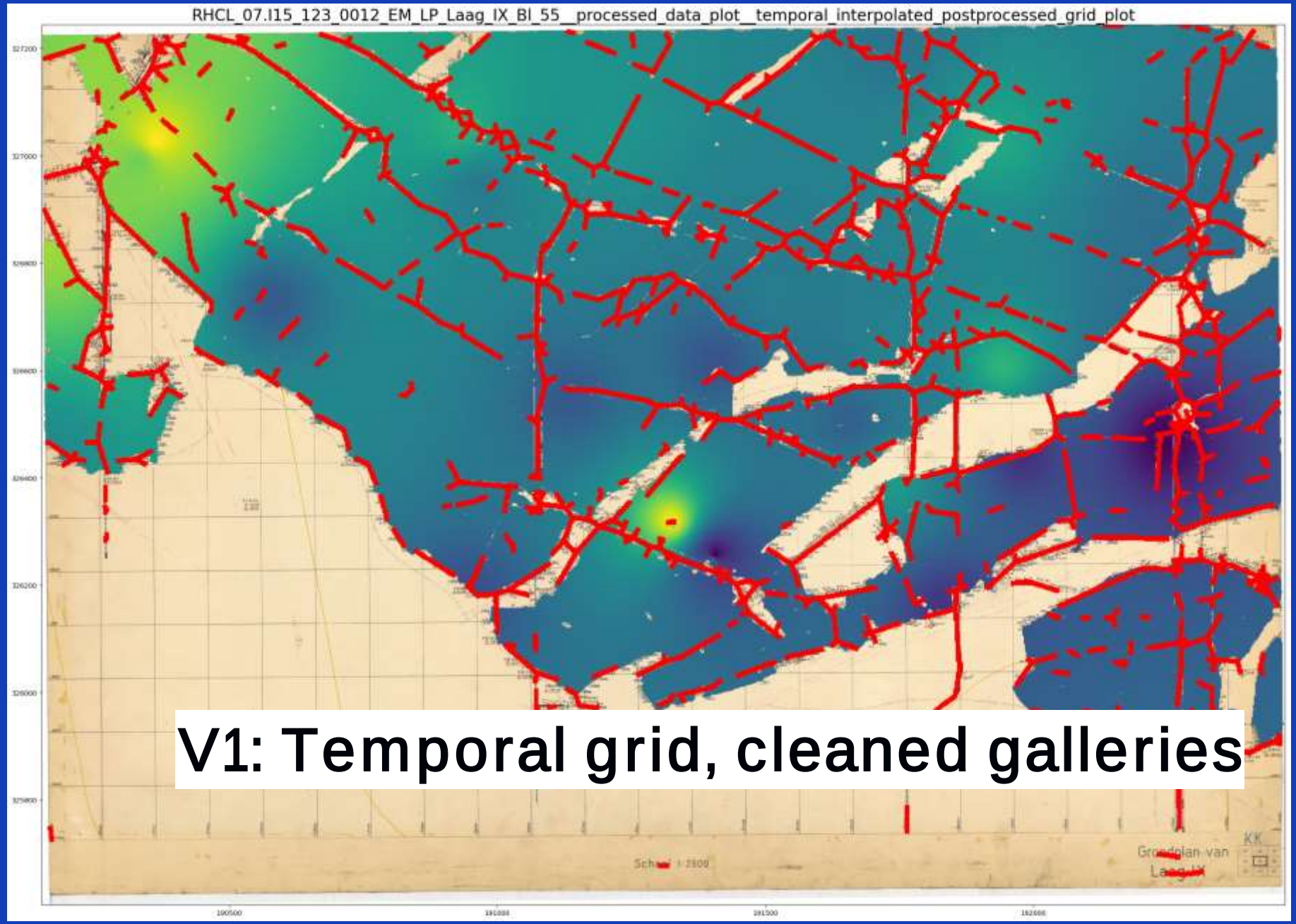
Results

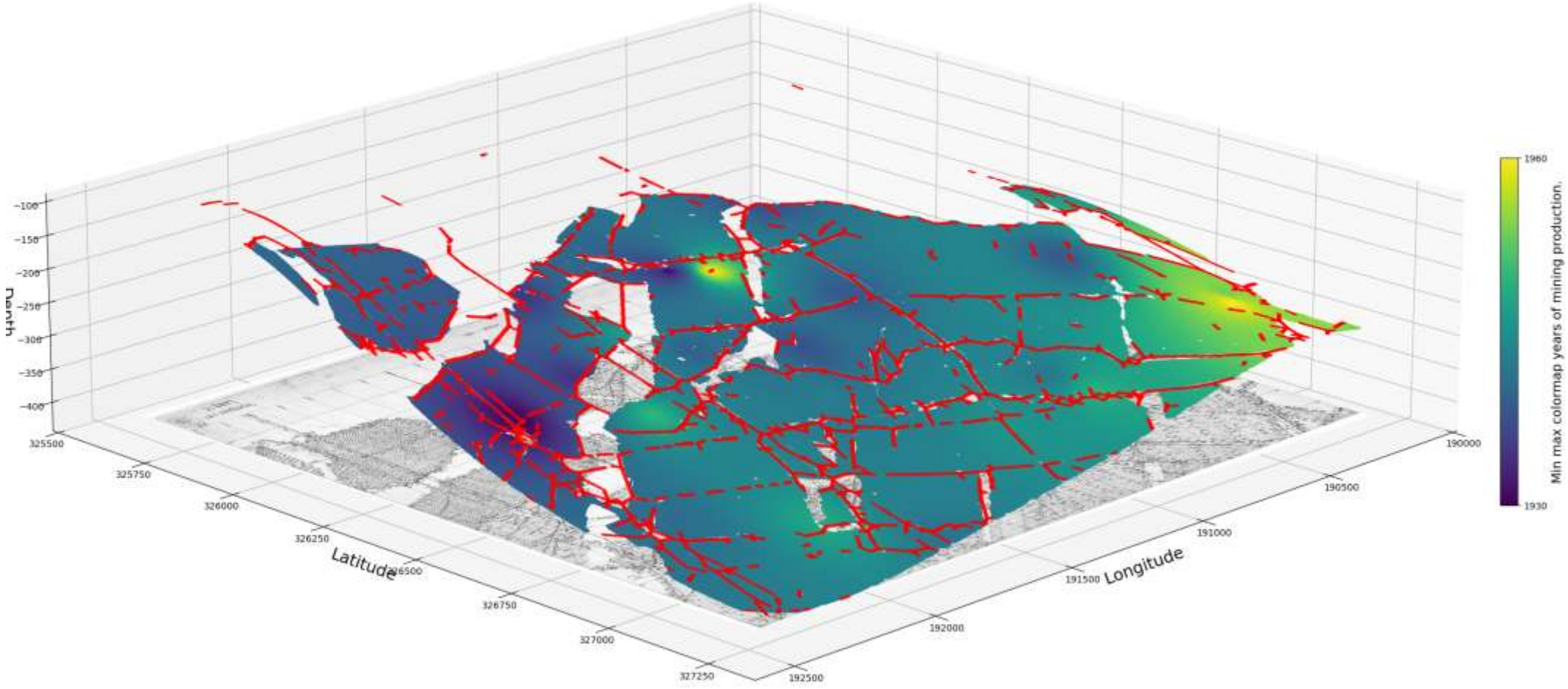


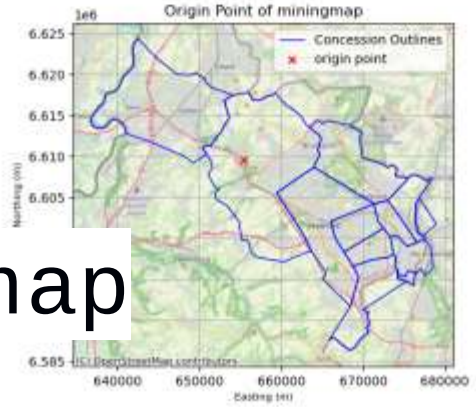
V1: Original map



V1: All features

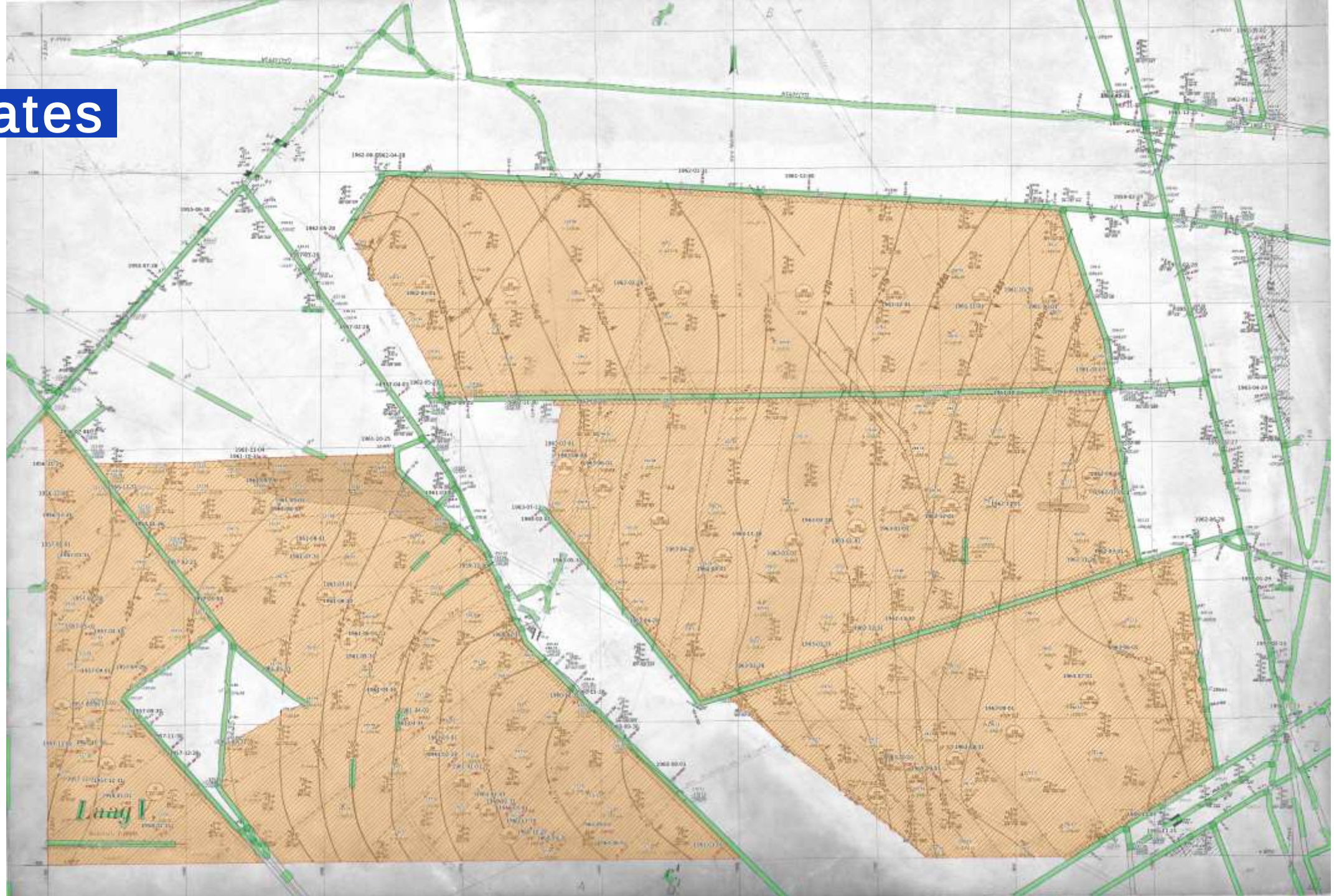

RHCL_07.I15_123_0012_EM_LP_Laag_IX_Bl_55__processed_data_plot__temporal_interpolated_postprocessed_grid_plot
V1: Temporal grid, cleaned galleries



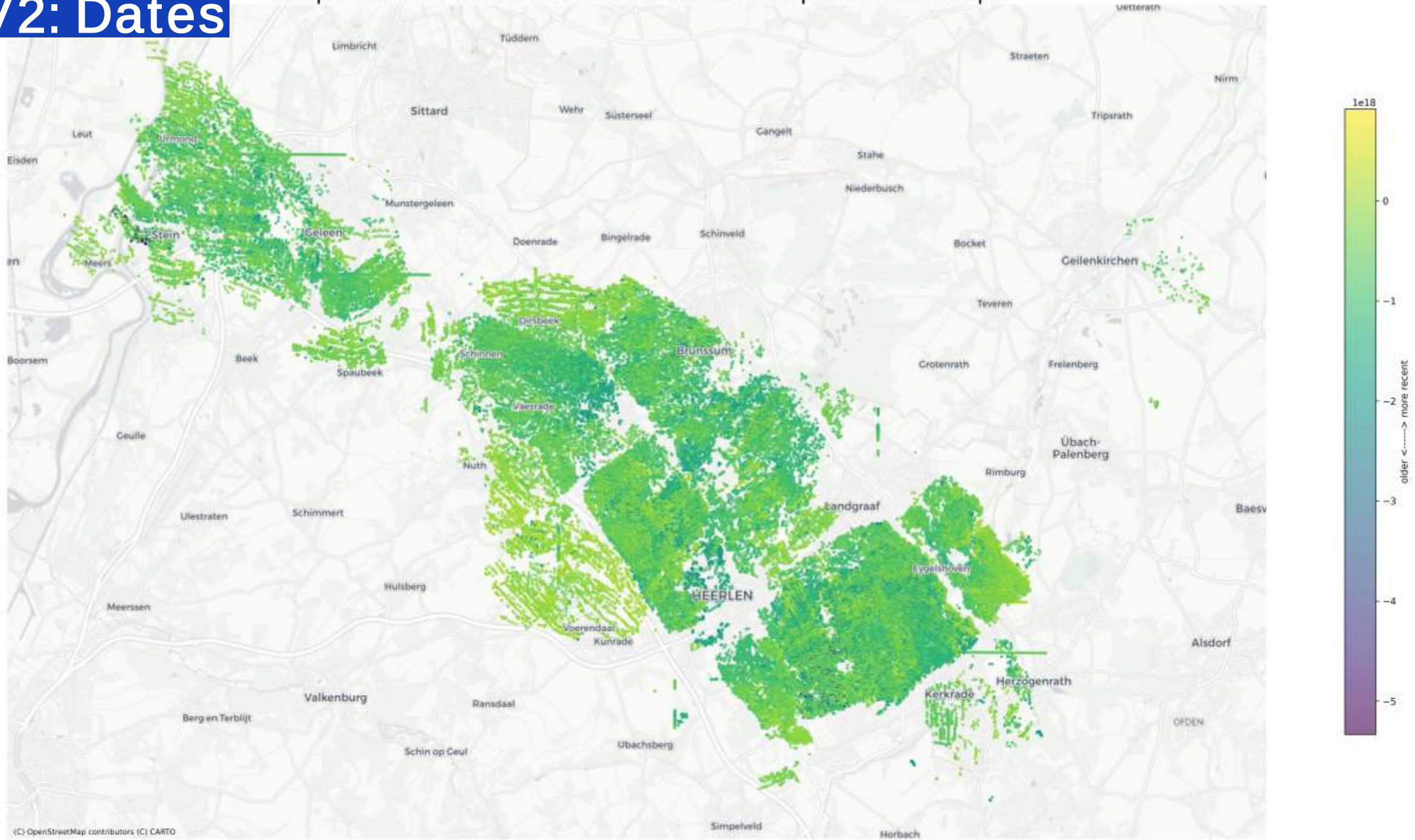


Temporal grid & galleries
Projected in x,y,z with corresponding map

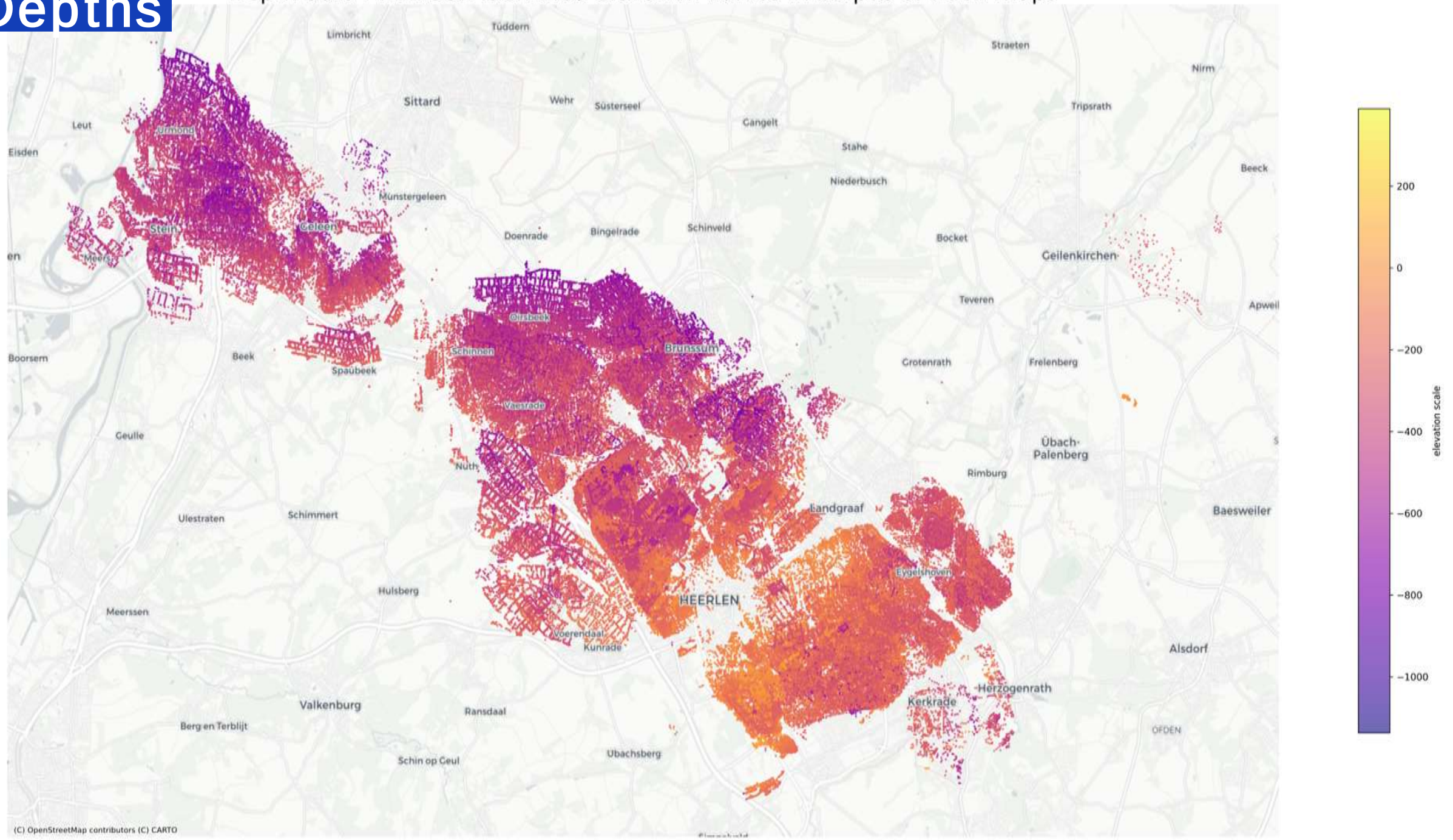
V2: Dates



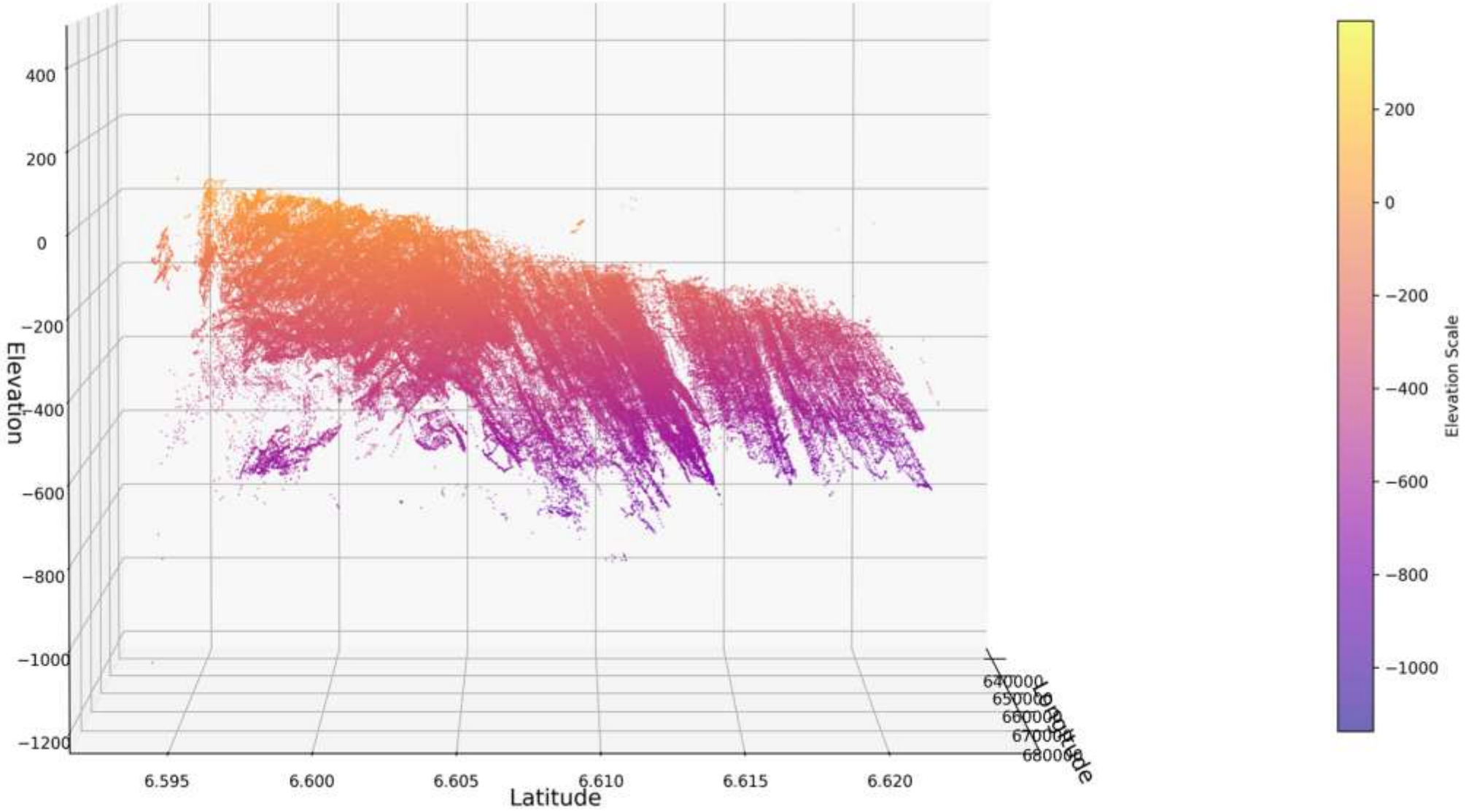
V2: Dates


Temporal data : 137.881 Identified dates in corpus of 2861 maps
Limbricht
Tüddern
Straeten
Nirm
Sittard
Wehr
Süsterseel
Tripsrath
Gangelt
Leut
Eisden
Stahe
Niederbusch
Munstergeleen
Geleen
Stein
Doenrade
Bingelrade
Schinveld
Bocket
Geilenkirchen
Meers
Teveren
Oirsbeek
Schinnen
Brunssum
Boorsem
Beek
Spaubeek
Crotenrath
Frelenberg
Vaesrade
Geulle
Übach-
Palenberg
Nuth
Rimburg
Landgraaf
Ulestraten
Schimmert
Hulsberg
Eygelshoven
HEERLEN
Meerssen
Voerendaal
Kunrade
Alsdorf
Herzogenrath
Valkenburg
Ransdaal
Kerkrade
Berg en Terblijt
Ubachsberg
Schin op Geul
Simpelveld
Horbach
(C) OpenStreetMap contributors (C) CARTO
1e18
0
−1
−2
−3
−4
−5
older <------> more recent

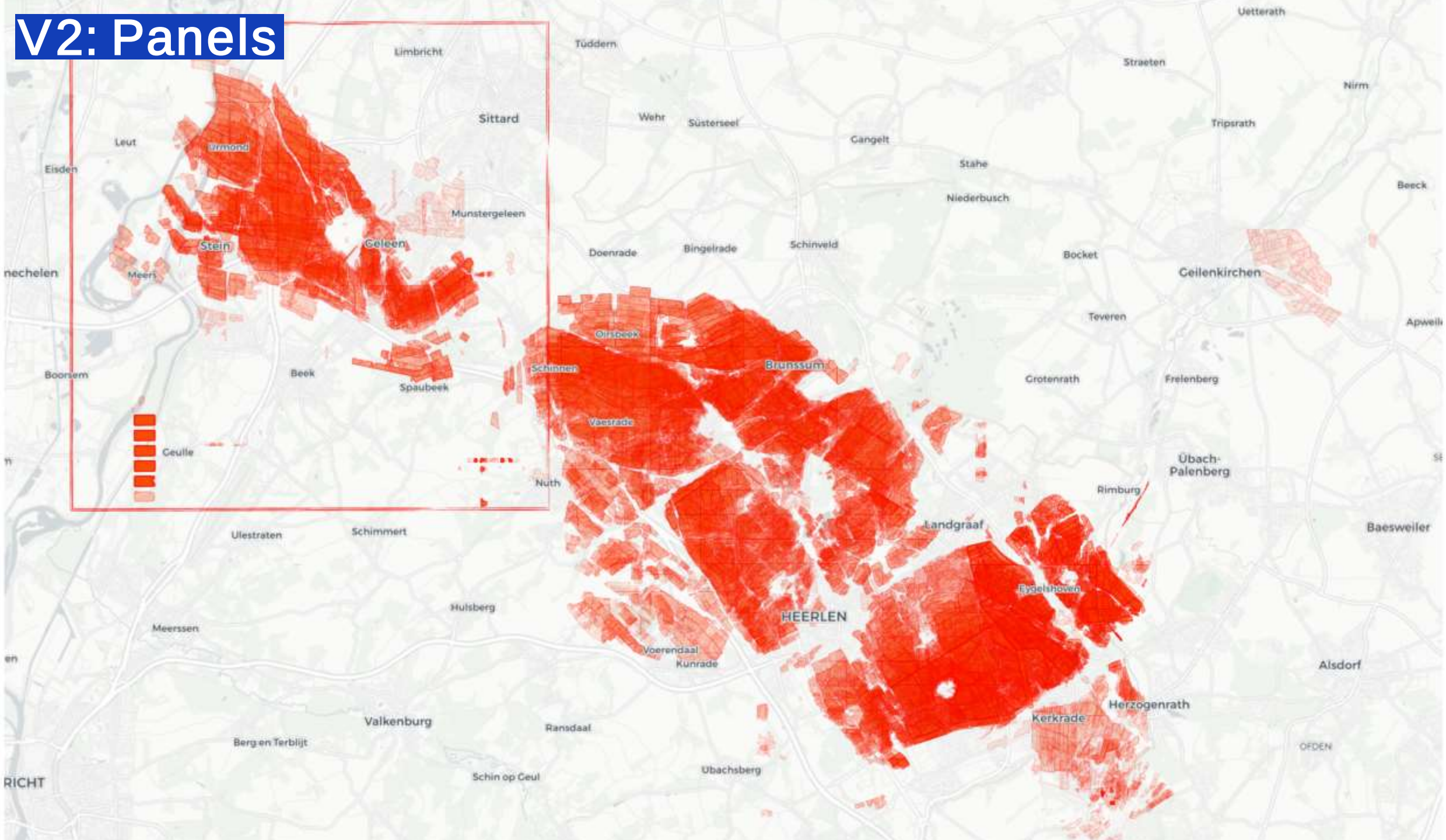
Depths


Depth data : 211.382 Identified elevation values in corpus of 2861 maps
Limbricht
Tüddern
Straeten
Nirm
Sittard
Wehr
Süsterseel
Tripsrath
Leut
Gangelt
Eisden
Stahe
Beeck
Niederbusch
Munstergeleen
Stein
Geleen
Doenrade
Bingelrade
Schinveld
Bocket
Geilenkirchen
Meers
Teveren
Apweil
Oirsbeek
Schinnen
Brunssum
Boorsem
Beek
Spaubeek
Crotenrath
Frelenberg
Vaesrade
Geulle
Übach-
Palenberg
Nuth
Rimburg
Landgraaf
Baesweiler
Ulestraten
Schimmert
Eygelshoven
Hulsberg
HEERLEN
Meerssen
Voerendaal
Kunrade
Alsdorf
Herzogenrath
Valkenburg
Kerkrade
Ransdaal
Berg en Terblijt
OFDEN
Schin op Geul
Ubachsberg
(C) OpenStreetMap contributors (C) CARTO
200
0
−200
−400
−600
−800
−1000
elevation scale

V2: Depths


400
200
0
−200
−400
−600
−800
−1000
−1200
Elevation
6.595
6.600
6.605
6.610
6.615
6.620
Latitude
640000
650000
660000
670000
680000
Longitude
200
0
−200
−400
−600
−800
−1000
Elevation Scale

V2: Panels


Limbricht
Tüddern
Uetterath
Straeten
Nirm
Sittard
Wehr
Süsterseel
Tripsrath
Leut
Ursmond
Gangelt
Eisden
Stahe
Beeck
Niederbusch
Munstergeleen
Stein
Geleen
Doenrade
Bingelrade
Schinveld
Bocket
Meers
Geilenkirchen
Teveren
Apweil
Oirsbeek
Schinnen
Brunssum
Boorsem
Beek
Spaubeek
Grotenrath
Frelenberg
Vaesrade
Geulle
Übach-
Palenberg
Nuth
Rimburg
Landgraaf
Ulestraten
Schimmert
Baesweiler
Eygelshoven
Hulsberg
HEERLEN
Meerssen
Voerendaal
Kunrade
Alsdorf
Herzogenrath
Valkenburg
Ransdaal
Kerkrade
Berg en Terblijt
OFDEN
Schin op Geul
Ubachsberg
RICHT

Take aways of applying & developing AI

The role of project management.

Developing custom AI model or even applying pretrained solutions is still a research field.

Using domain knowledge to inform the datascience process.

Fundamentals & SOTA in AI.

Working with data as you are supposed to enabled us run experiments for scale.

Using the SOTA can be beneficial if the internal knowledge is there.