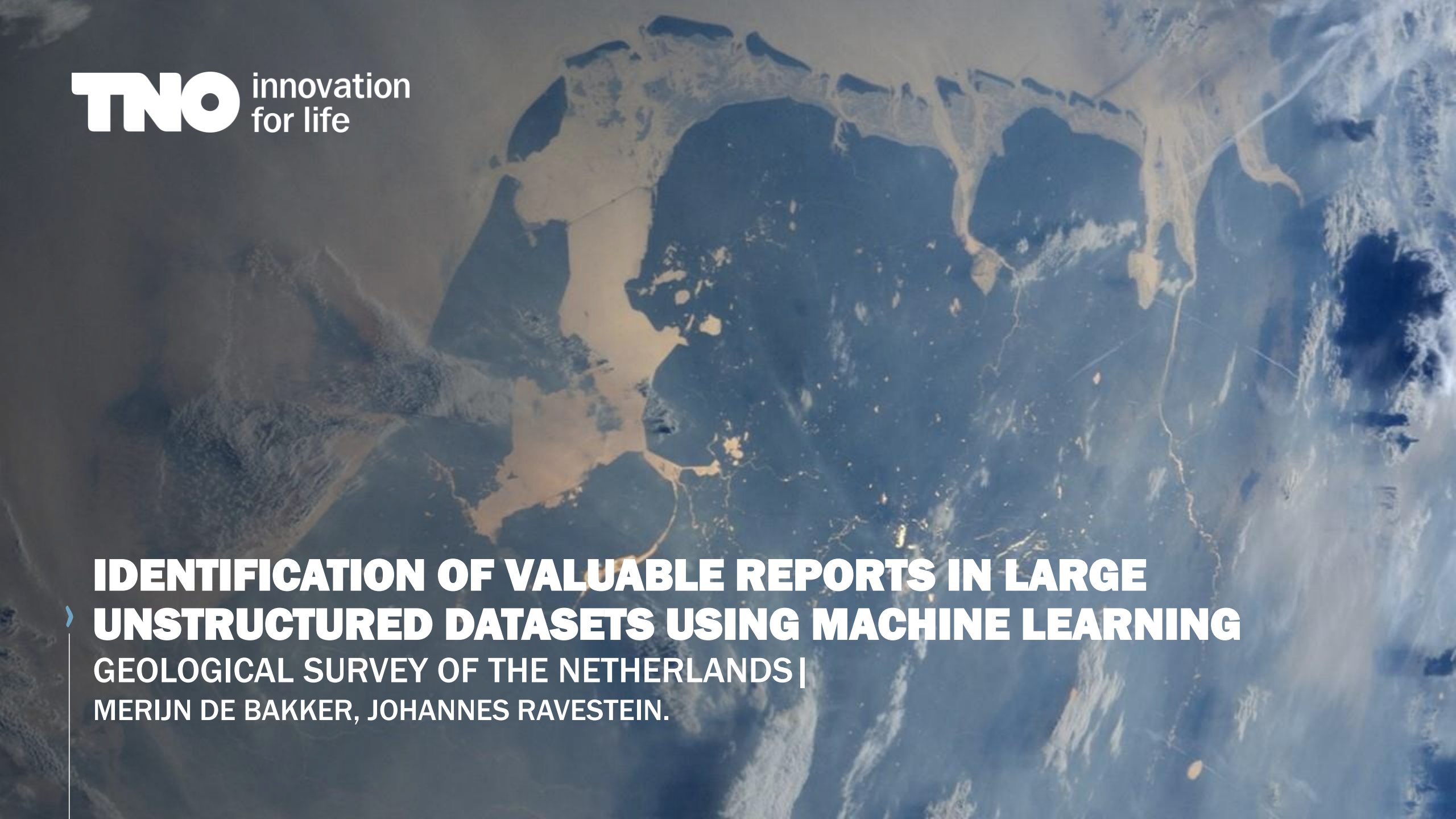


TNO innovation for life

IDENTIFICATION OF VALUABLE REPORTS IN LARGE UNSTRUCTURED DATASETS USING MACHINE LEARNING

GEOLOGICAL SURVEY OF THE NETHERLANDS |
MERIJN DE BAKKER, JOHANNES RAVESTEIN.

GEOLOGICAL SURVEY OF THE NETHERLANDS

The Geological Survey of the Netherlands (GSN) independently develops and manages data and knowledge of the subsurface and subsurface technologies, for societal questions of today and tomorrow.

GSN is part of TNO, an independent research organization.

ECONOMIC

e.g. resources: geo-energy, raw materials, groundwater

SOCIETAL

e.g. climate, energy transition

RISK MANAGEMENT

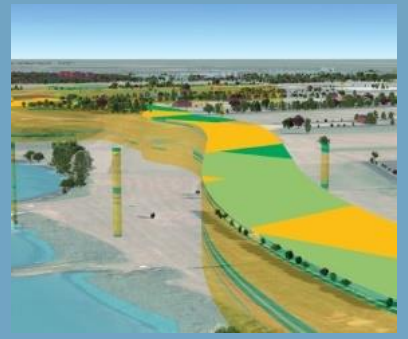
e.g. earthquakes, subsidence, safe abandonment

BUILT ENVIRONMENT

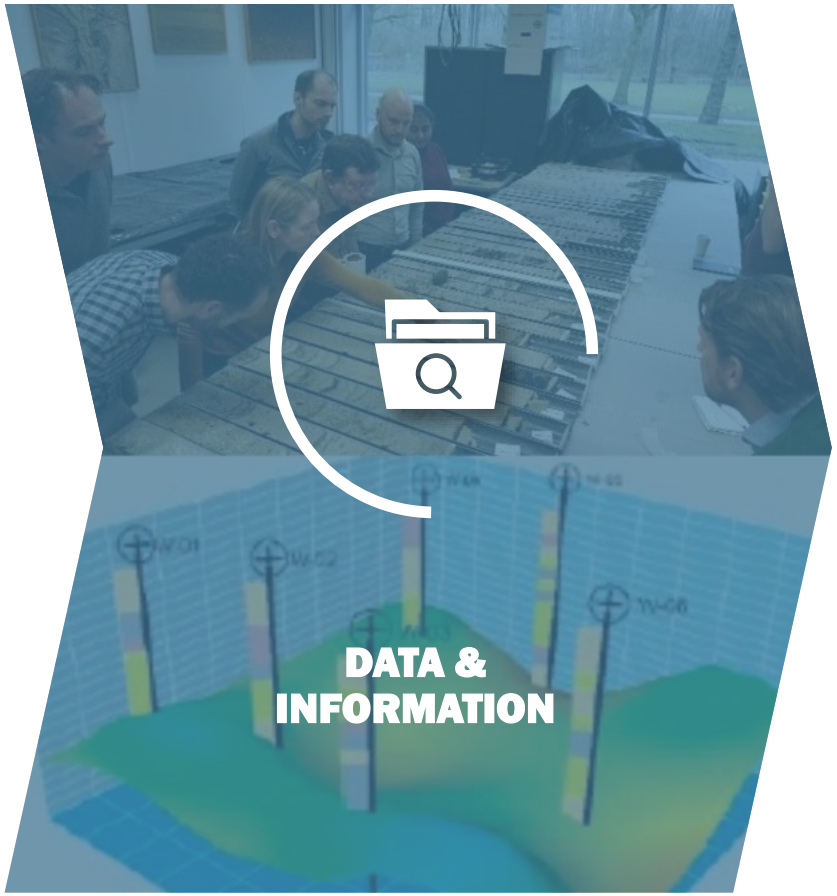
e.g. subsurface spatial planning, ground conditions

FUTURE CHALLENGES

e.g. digital twins, emerging societal questions



GEOLOGICAL SURVEY OF THE NETHERLANDS



Key Register
of the Subsurface

National Data
Repository (Mining Law)

Core Store

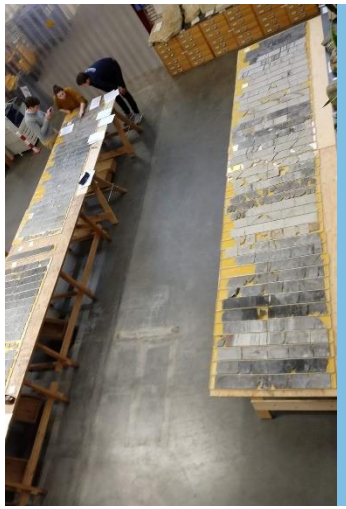
Nation wide
3D models of the
subsurface

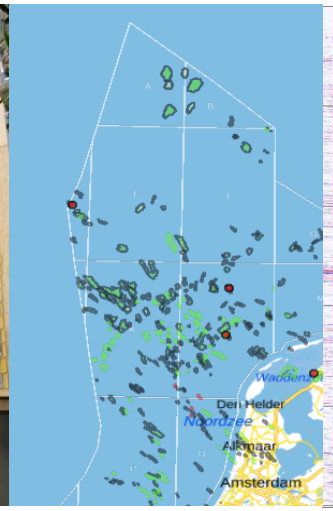
GEOLOGICAL SURVEY OF THE NETHERLANDS
NATIONAL DATA REPOSITORY - MINING LAW

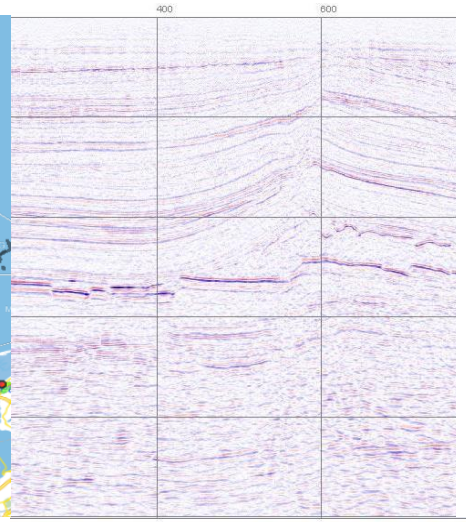
A copy of data obtained during the reconnaissance, exploration, production or storage of mineral resources or geothermal energy has to be provided to TNO-GSN on behalf of the Ministry of Economic Affairs and Climate. After a certain period the data is released to the public domain through the Dutch Oil and Gas portal (www.NLOG.nl).

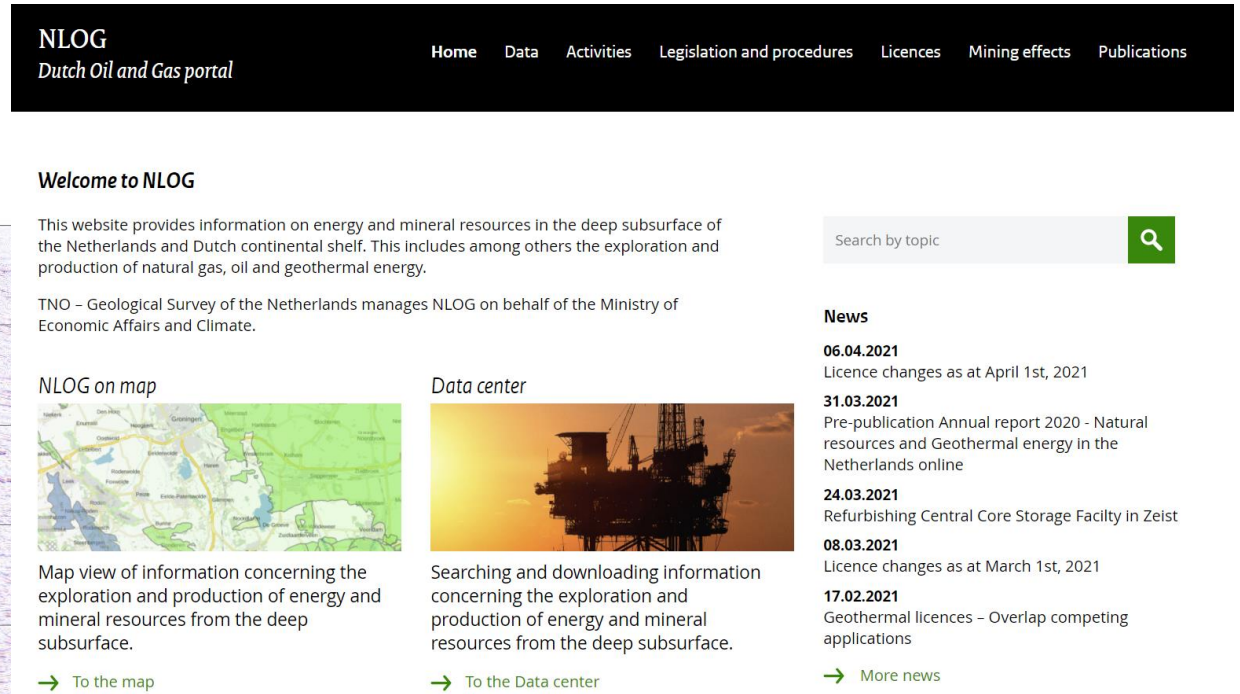
Main data types:

- Well/borehole data
- Seismic data
- Production data





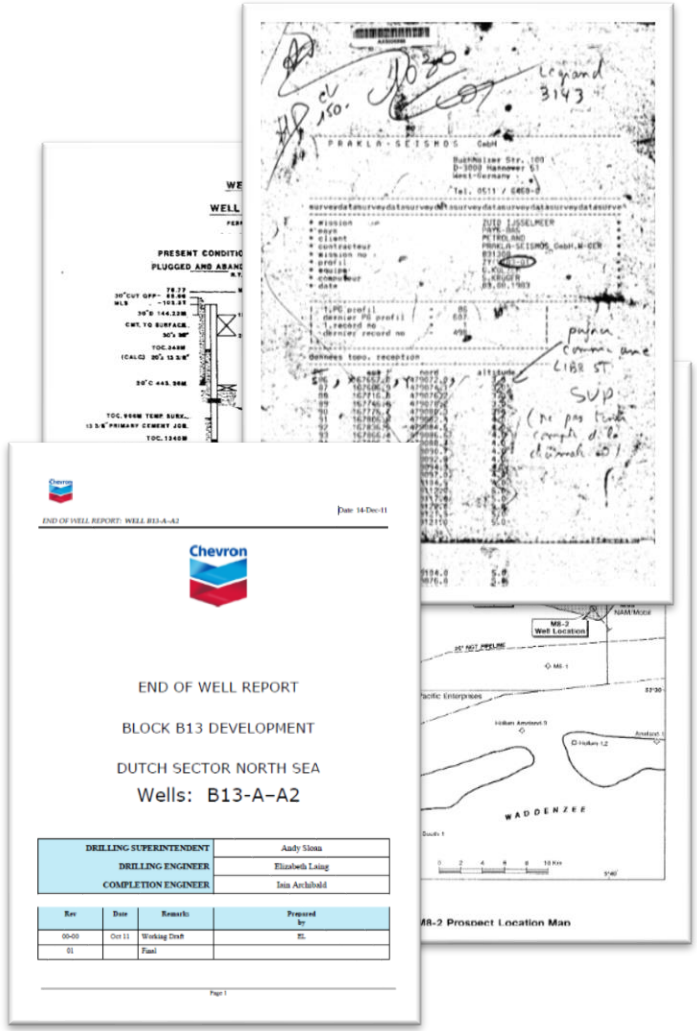




PROBLEM SETTING

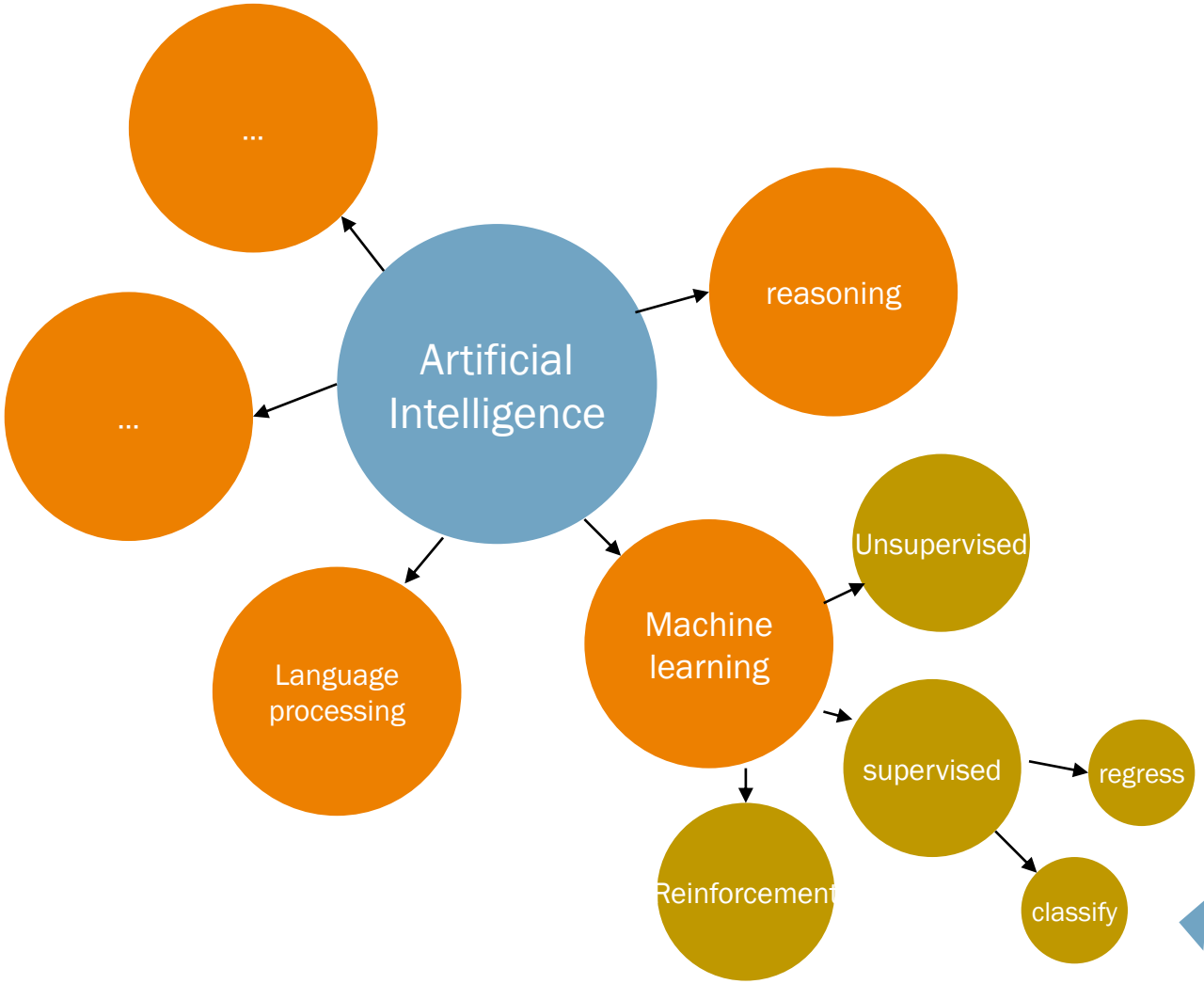
IDENTIFICATION OF VALUABLE REPORTS IN LARGE UNSTRUCTURED DATASETS USING MACHINE LEARNING

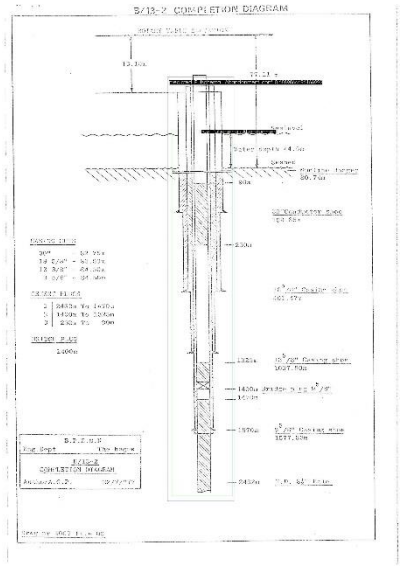
- GSN occasionally receives large diverse unstructured data sets of over 500 000 files:



- **Problem:** *No capability to go through all files to identify the valuable reports.*
- Can this problem be solved?: Machine learning
  - To recognize characteristic figures and tables in order to identify valuable reports.
  - And aid the processing and QC of the reports.
- Pilot project.

AI AND MACHINE LEARNING



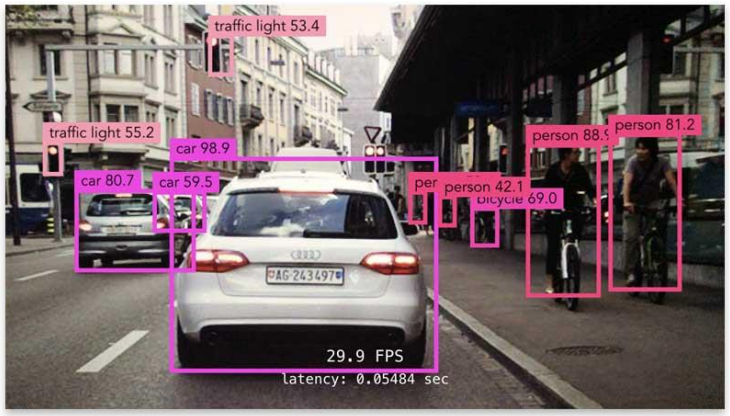


In this project:
Classification of information objects in documents

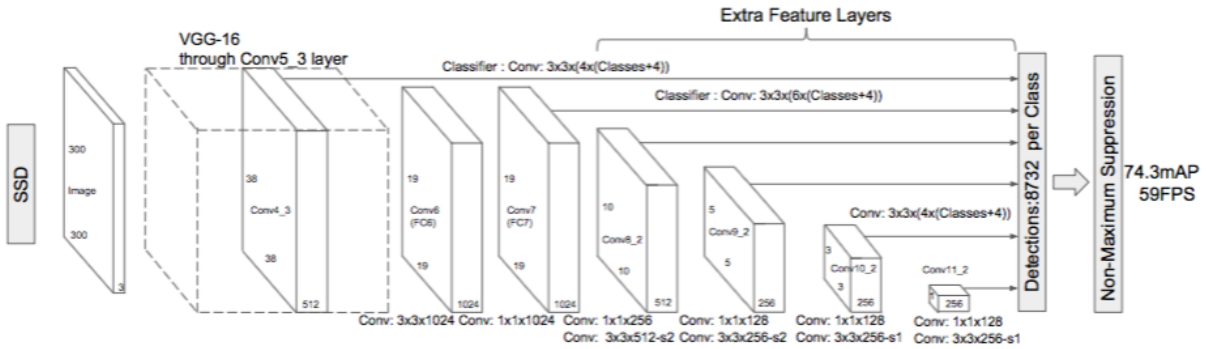
OBJECT DETECTION

- Object detection is a supervised machine learning technique to *classify* and *localize* objects in images
- We use this technique to find interesting elements in documents

- Our implementation is a Single Shot Detector MultiBox, as presented in Liu et al., 2016
  - SSD is based on Convolutional Neural Networks
  - Requires a labelled training set to learn the model
  - Outputs a series of bounding boxes with class labels and confidence value



From: https://machinethink.net/blog/object-detection/



From: Liu et al., 2016. DOI: 10.1007/978-3-319-46448-0_2

WORKFLOW PILOT

Image conversion

↓

Annotation

↓

Supervised learning

↓

Validation

- Image conversion to uniform filetype as raster graphic (PNG)
- Annotation of the important tables and figures for the data set
- Training the model using part of the annotated images
- Validating the model using the remaining annotated images

Open source and free tools:

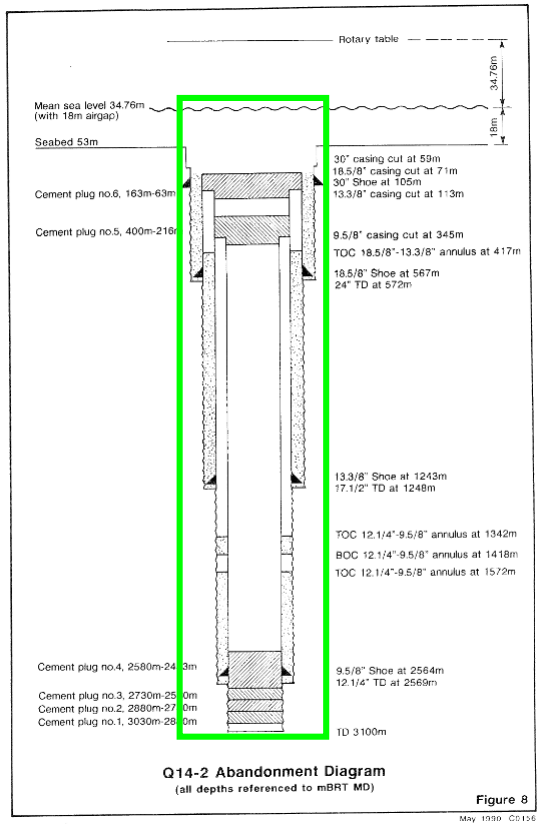
Python

Image Magick

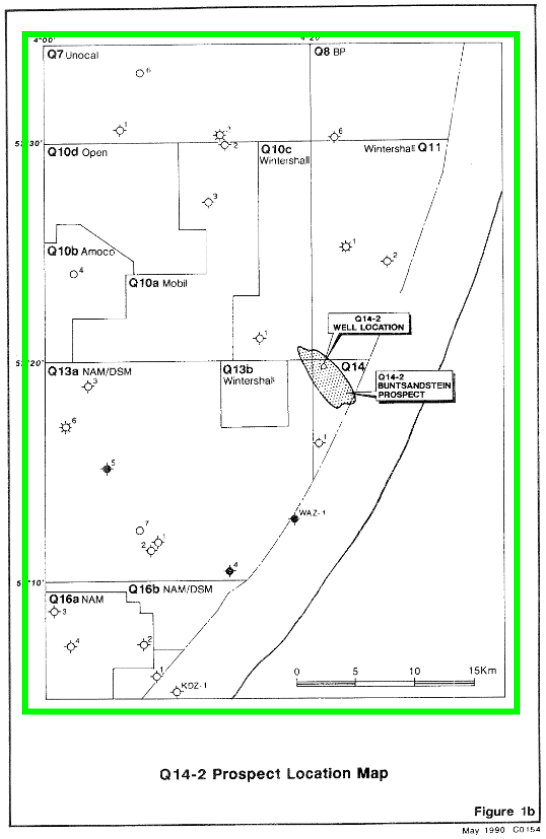
Labelimg

Tensorflow research API

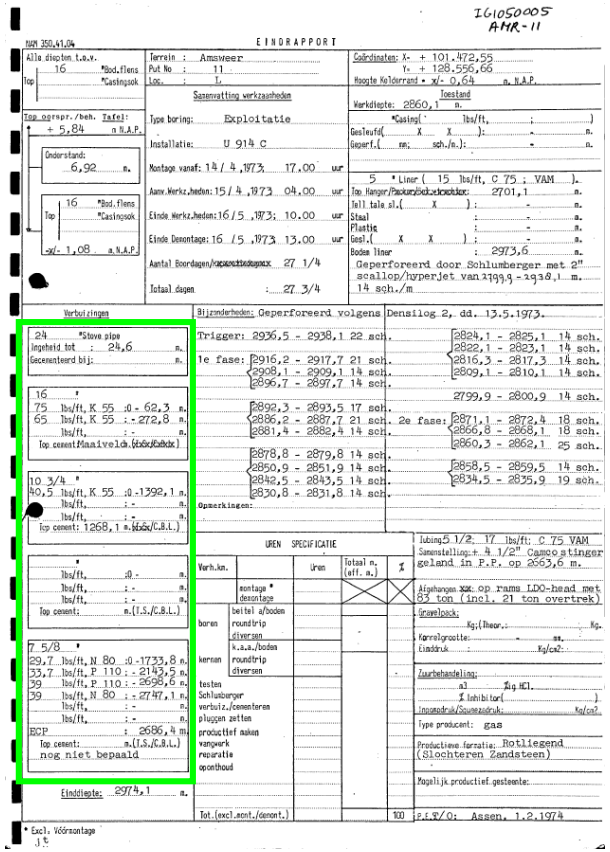
ANNOTATION & CLASSES



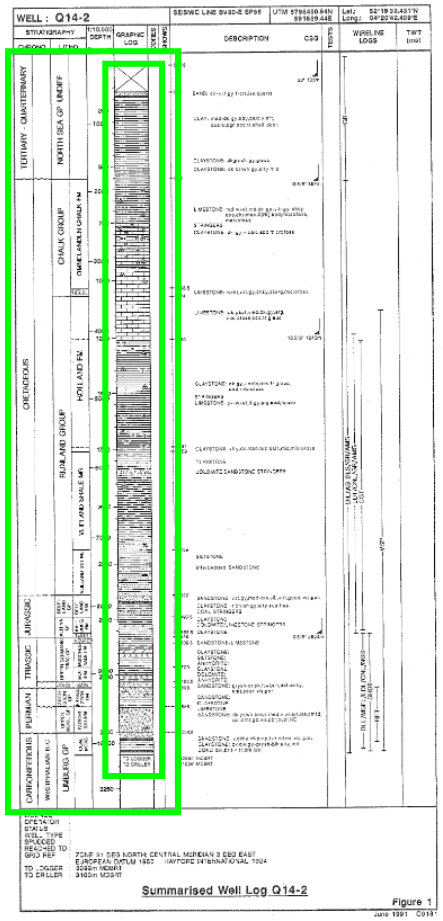
Well schematic



Location Map



Casing table

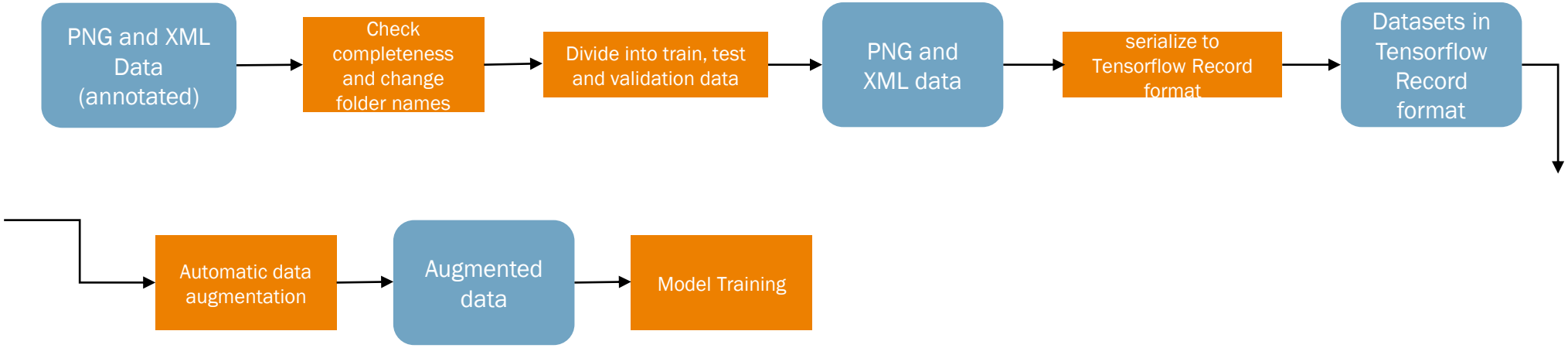


Stratigraphic column

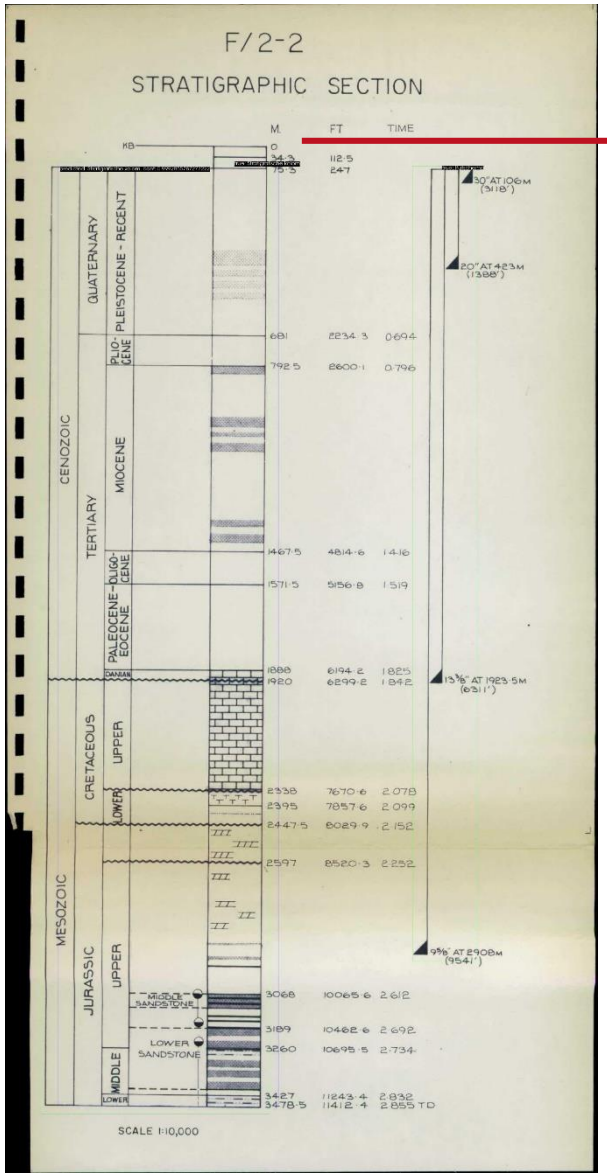
TRAINING THE NETWORK

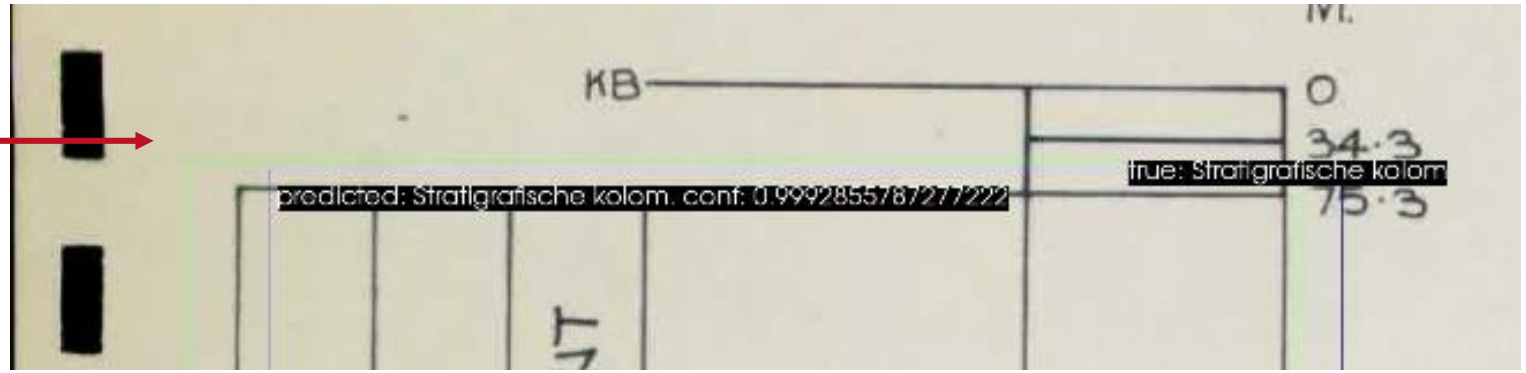
- Preprocessing the data
  - Creation of training, testing and validation dataset.
  - Validation set is used for publication/comparing final models
- Training of the network
  - Training TNO HPC with GPU
  - ~ 12 hours training
  - In this pilot we did not optimize for the model hyperparameters

Set	*n*
Train	658
Test	183
Val	74



RESULTS

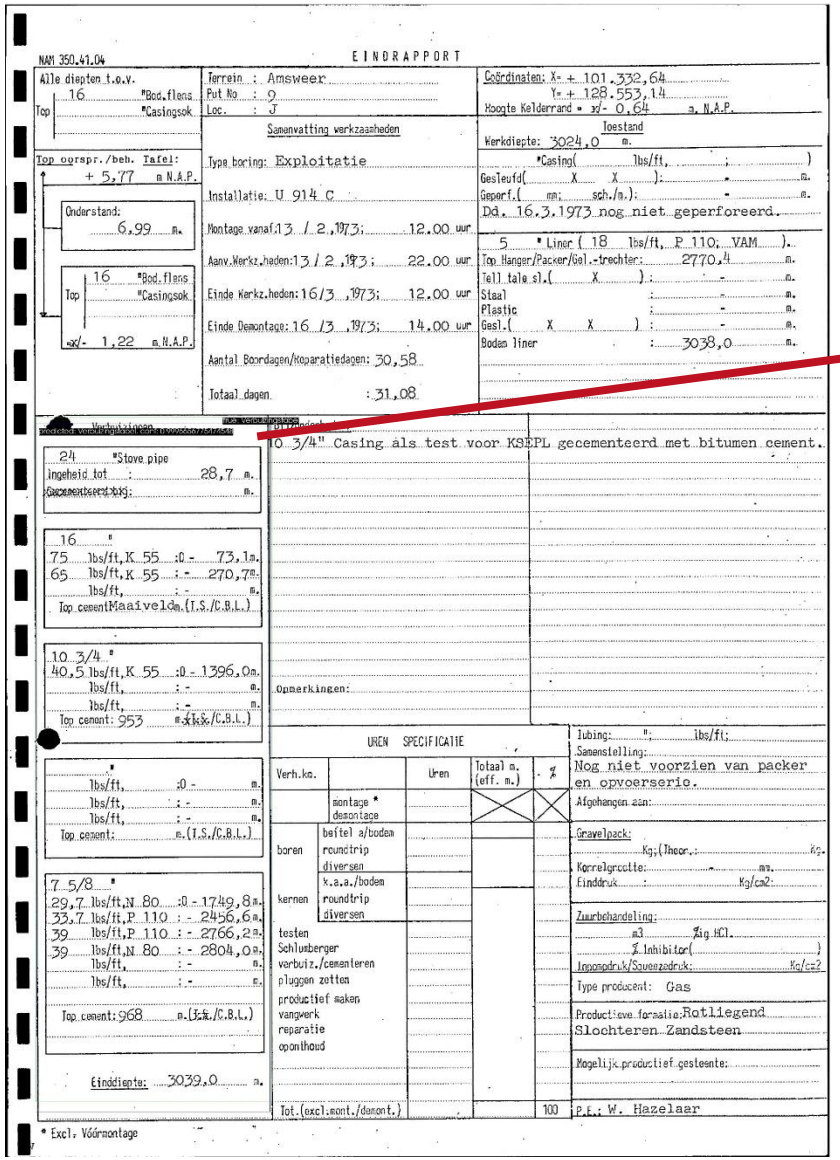




Blue line: predicted bounding box and class label
Green line: true bounding box and class label

In this example, the stratigraphical column is predicted with a 99% confidence

RESULTS



NAM 350.41.04

EINDRAPPORT

Alle diepten t.o.v.
Top 16 "Bod.flens
"Casingsok

Terrein : Amsweer
Put No : 9
Loc. : J

Coördinaten: X= + 101.332,64
Y= + 128.553,14
Hoogte Kelderrand = x/- 0,64 m. N.A.P.

Samenvatting verkzaamheden

Type boring: Exploitatie

Installatie: U 914 C

Montage vanaf 13 / 2 ,1973; 12.00 uur

Aanv.Werkz.haden: 13 / 2 ,1973; 22.00 uur

Einde Werkz.heden: 16/3 ,1973; 12.00 uur

Einde Demontage: 16 /3 ,1973; 14.00 uur

Aantal Boordagen/Reparatiedagen: 30,58

Totaal dagen : 31,08

Toestand

Werkdiepte: 3024,0 m.

"Casing(lbs/ft, ;)

Gesleufd(X X): - m.

Geperf.(mm; sch./m.): - m.

Dd. 16.3.1973 nog niet geperforeerd.

5 " Liner (18 lbs/ft, P 110, VAM).

Top Hanger/Packer/Gel.-trechter: 2770,4 m.

Tell tale s.(X): - m.

Staal : - m.

Plastic : - m.

Gesl.(X X) : - m.

Bodem liner : 3038,0 m.

Top oorspr./beh. Tafel: + 5,77 m N.A.P.

Onderstand: 6,99 m.

Top 16 "Bod.flens
"Casingsok

x/- 1,22 m. N.A.P.

predicted: verbuizingstabel. conf: 0.9996656775474548

Verbuizingen

24 "Stove pipe
Ingeheid tot : 28,7 m.
Gecementeerd tot: m.

16 "
75 lbs/ft,K 55 :0 - 73,1m.
65 lbs/ft,K 55 : - 270,7m.
lbs/ft, : - m.
Top cement Maaiveld m.(T.S./C.B.L.)

10 3/4 "
40,5 lbs/ft,K 55 :0 - 1396,0m.
lbs/ft, : - m.
lbs/ft, : - m.
Top cement: 953 m.(T.S./C.B.L.)

" lbs/ft, :0 - m.
lbs/ft, : - m.
lbs/ft, : - m.
Top cement: m.(T.S./C.B.L.)

7 5/8 "
29,7 lbs/ft,N 80 :0 - 1749,8m.
33,7 lbs/ft,P 110 : - 2456,6m.
39 lbs/ft,P 110 : - 2766,2m.
39 lbs/ft,N 80 : - 2804,0m.
lbs/ft, : - m.
lbs/ft, : - m.
Top cement: 968 m.(T.S./C.B.L.)

Einddiepte: 3039,0 m.

predicted: verbuizingstabel. conf: 0.9996656775474548

Verbuizingen

10 3/4" Casing als test voor KSEPL gecementeerd met bitumen cement.

Opmerkingen:

UREN SPECIFICATIE

Verh.ka.		Uren	Totaal m. (eff. m.)	%
	montage * / demontage			
boren	beitel a/bodem			
	roundtrip			
	diversen			
kernen	k.a.a./bodem			
	roundtrip			
	diversen			
testen				
Schlumberger				
verbuiz./cementeren				
pluggen zetten				
productief maken				
vangwerk				
reparatie				
openhouden				
Tot.(excl.mont./demont.)				100

* Excl. Vóórmontage

Tubing: "; lbs/ft;
Samenstelling:
Nog niet voorzien van packer en opvoerserie.
Afgehangen aan:

Gravelpack:
Kg; (Theor.: Kg.
Korrelgrootte: - mm.
Indruk : Kg/cm2.

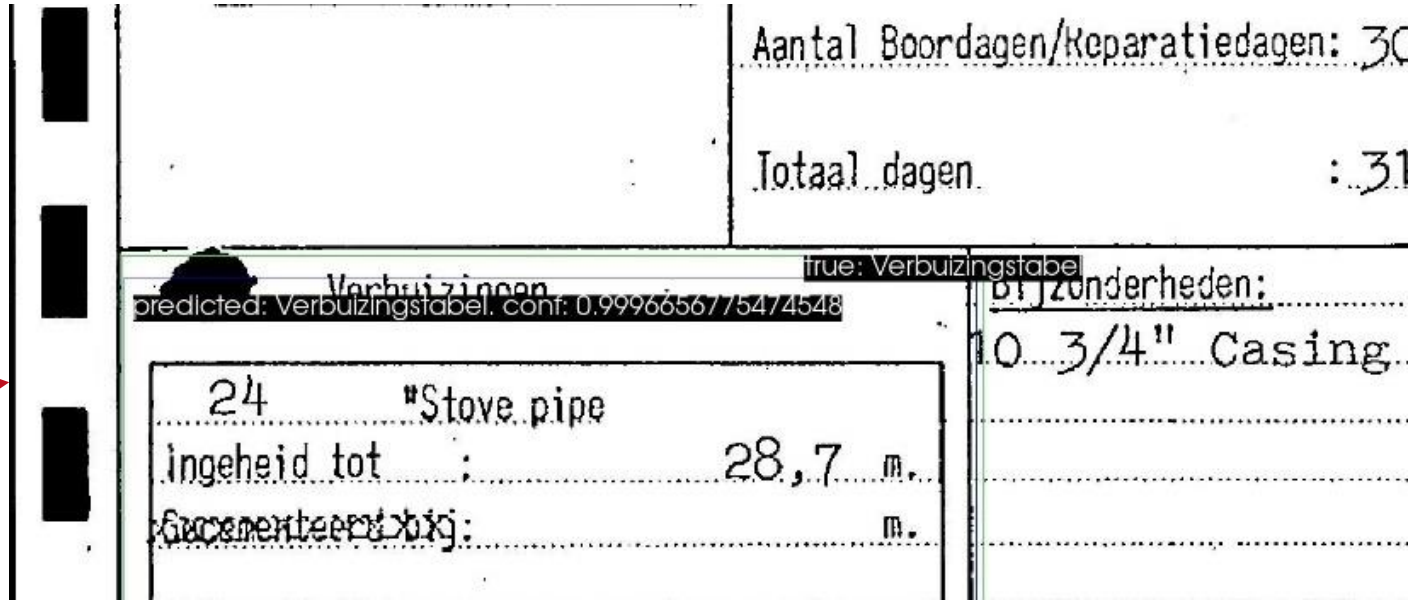
Zuurbehandeling:
m3 %ig HCl.
% Inhibitor( )
Inpompdruk/Squeezedruk: Kg/cm2

Type producent: Gas

Productieve formatie: Rotliegend
Slochteren Zandsteen

Mogelijk productief gesteente:

P.F.: W. Hazelaar



Aantal Boordagen/Reparatiedagen: 30

Totaal dagen : 31

true: Verbuizingstabel

predicted: Verbuizingstabel. conf: 0.9996656775474548

Verbuizingen

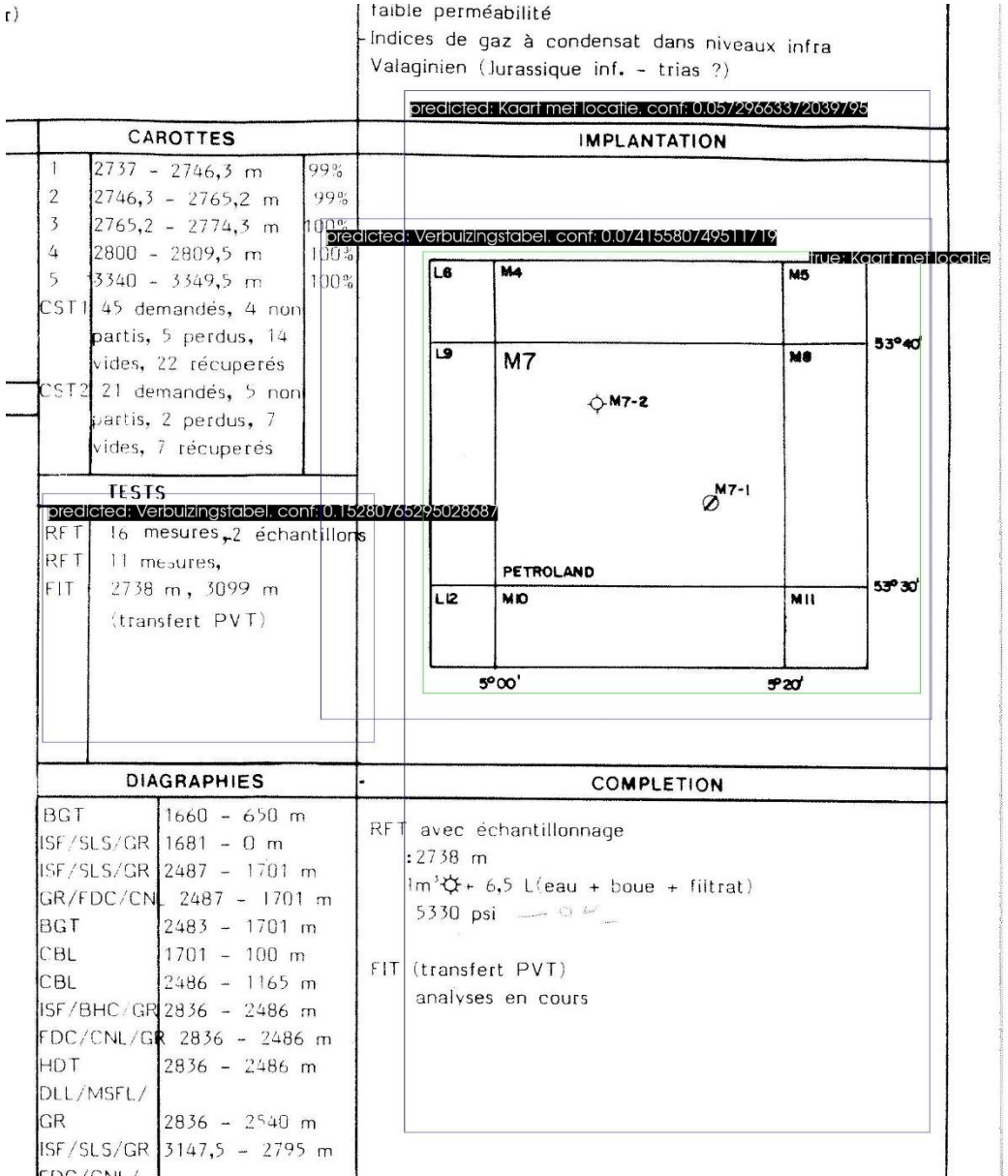
24 "Stove pipe
Ingeheid tot : 28,7 m.
Gecementeerd tot: m.

Bijzonderheden:

10 3/4" Casing

TNO innovation for life

RESULTS



r)

faible perméabilité
- Indices de gaz à condensat dans niveaux infra Valaginien (Jurassique inf. - trias ?)

predicted: Kaart met locatie. conf: 0.05729663372039795

CAROTTES

1	2737 - 2746,3 m	99%
2	2746,3 - 2765,2 m	99%
3	2765,2 - 2774,3 m	100%
4	2800 - 2809,5 m	100%
5	3340 - 3349,5 m	100%
CST1	45 demandés, 4 non partis, 5 perdus, 14 vides, 22 récuperés	
CST2	21 demandés, 5 non partis, 2 perdus, 7 vides, 7 récuperés	

predicted: Verbuizingstabel. conf: 0.07415580749511719

TESTS

predicted: Verbuizingstabel. conf: 0.15280765295028687

RFT	16 mesures, 2 échantillons
RFT	11 mesures,
FIT	2738 m, 3099 m (transfert PVT)

IMPLANTATION

true: Kaart met locatie

L6 M4 M5
L9 M7 M8
M7-2
M7-1
PETROLAND
L12 M10 M11
53°40'
53°30'
5°00'
5°20'

DIAGRAPHIES

BGT	1660 - 650 m
ISF/SLS/GR	1681 - 0 m
ISF/SLS/GR	2487 - 1701 m
GR/FDC/CNL	2487 - 1701 m
BGT	2483 - 1701 m
CBL	1701 - 100 m
CBL	2486 - 1165 m
ISF/BHC/GR	2836 - 2486 m
FDC/CNL/GR	2836 - 2486 m
HDT	2836 - 2486 m
DLL/MSFL/GR	2836 - 2540 m
ISF/SLS/GR	3147,5 - 2795 m

COMPLETION

RFT avec échantillonnage
: 2738 m
1m³ + 6,5 L (eau + boue + filtrat)
5330 psi

FIT (transfert PVT)
analyses en cours

RESULTS
CONFUSION MATRIX

Predicted class

True class

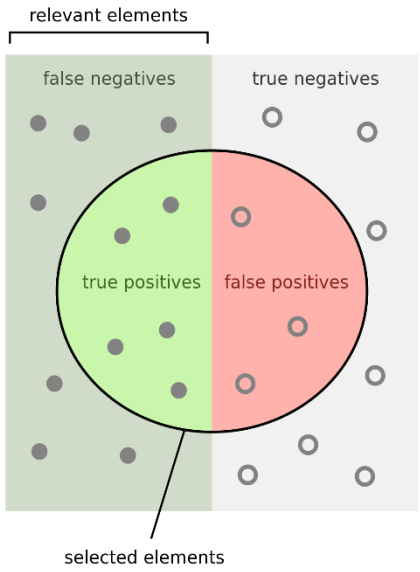
	Map with location	Well scheme	Casing table	Strat. Column
Map with location	30	0	1	0
Well scheme	2	26	1	7
Casing table	0	0	75	3
Strat. Column	2	1	0	23

METRICS

	F1	support
Map with location	0.92	31
Well scheme	0.79	37
Strat. column	0.77	26
Casing table	0.96	79
Overall weighted	0.88	183

Note: intersection over union metric is not taken into account here.

F-score: harmonic mean of precision and recall.
= measure for classification accuracy



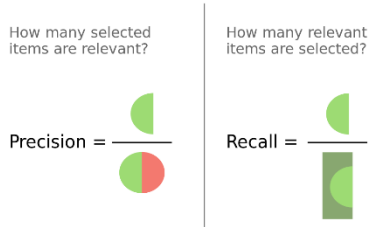
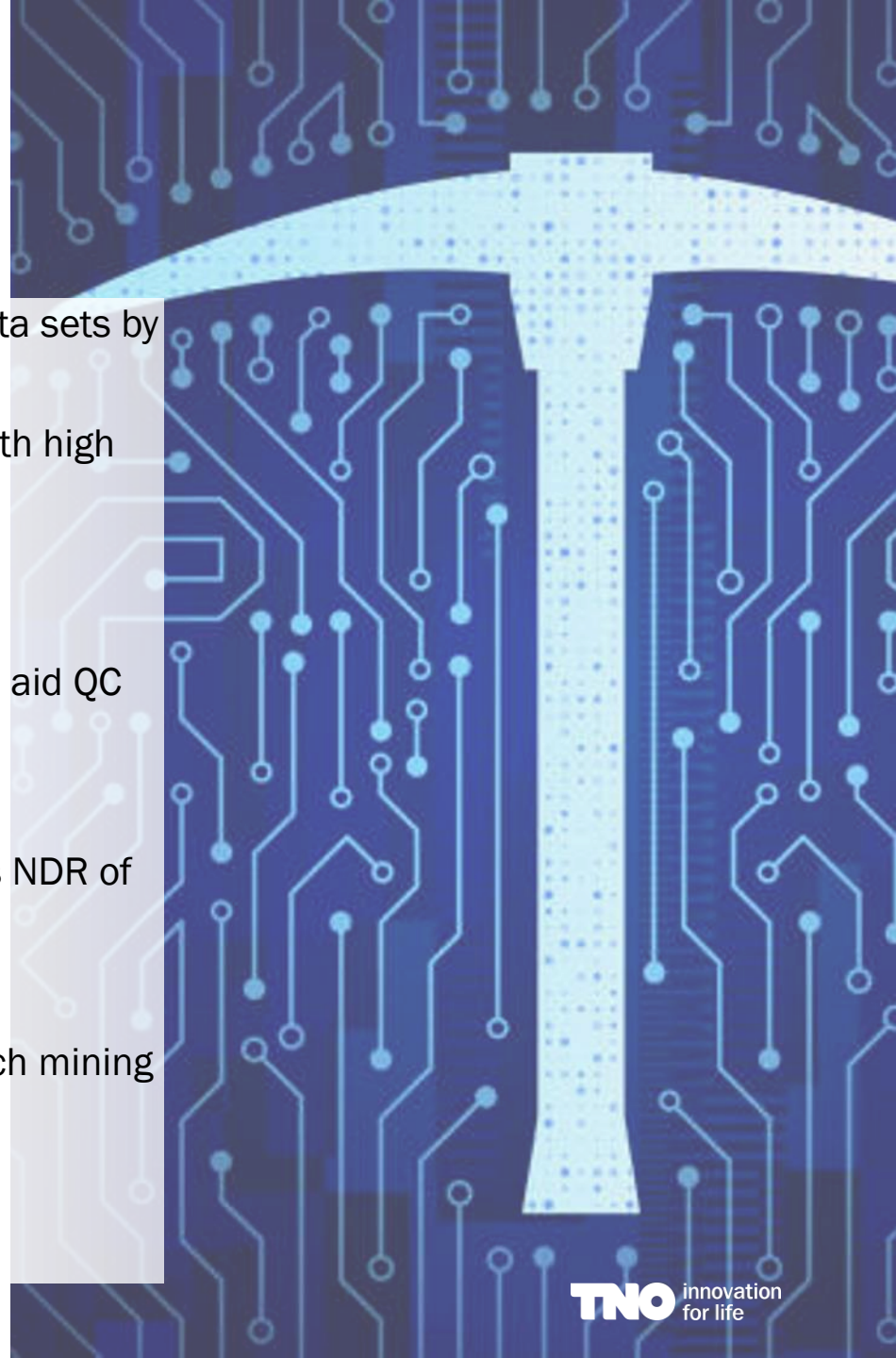


Image source: https://en.wikipedia.org/wiki/Precision_and_recall

CONCLUSIONS

- The pilot has shown that image recognition can help us process these large data sets by identifying specific valuable reports fast and reliably
  - The SSD algorithm performs well in recognizing distinct tables and figures with high confidence.
  - The combined presence of these tables and figures can be used to identify documents of interest.
  - Location or page of specific information can be extracted from documents to aid QC and processing of the document.

- The applied workflow represents a purpose specific tool, tuned to our needs as NDR of the Netherlands
  - Low cost, flexible and under our control.
  - Tuned to unique documents: in Dutch and made to specifications of the Dutch mining law.
  - Not a ‘one click’ solution for all unstructured data.
  - All new document types to identify require partial retraining of the model.

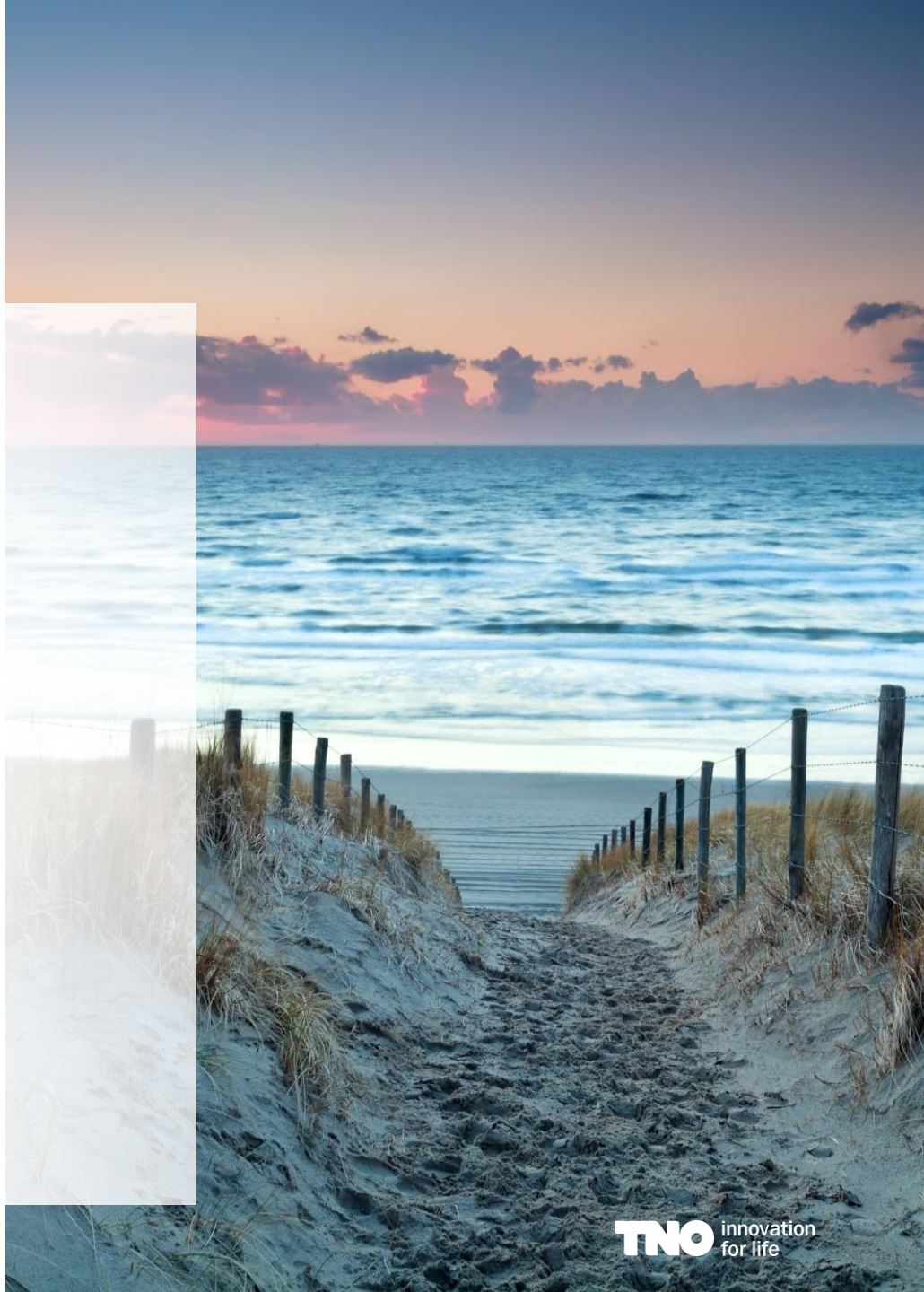


OUTLOOK: FUTURE STEPS

- Apply the trained model to a huge dataset of 500 000+ files (in progress)
- Apply in a study into the status of old wells

- Investigate how to integrate this approach in future workflows
  - E.g. How to retrain the model when new data or class labels arrive?
  - Automatization of preprocessing, training, classification pipeline

- Extend classification of documents with data extraction:
  - Automated extraction of the data found by combining with an OCR pass.



THANKS FOR LISTENING!
QUESTIONS, REMARKS?

MERIJN.DEBAKKER@TNO.NL
JOHANNES.RAVESTEIN@TNO.NL

TNO innovation for life