Navigating the Data Science Landscape: Personal Insights and Tips

Christian E. Westermann

University of Southern Denmark

November 17, 2023

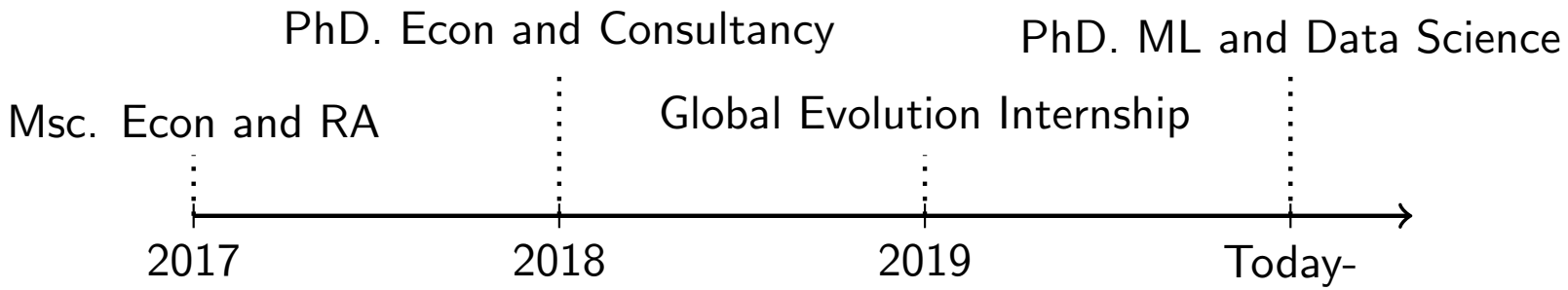
About Me

- Name: Christian E. Westermann.
- Interests: Sports (tennis) and everything Computer Science/Tech-related.
- PhD-student in Machine Learning and Data Science.

Beginning of Academic Journey

- Quickly ditched Biomedicine.
- Regression Analysis (Statistics): Sparked my interest.
- Thesis in predicting outcomes of football matches using Neural Networks.

Masters and PhD



- RA in Computer Vision.
- Consultancy in Finance.
- Internship in Finance doing NLP.
- Taught Natural Language Processing, Statistical Learning and Data Driven Decision Making
- Pivoted to a PhD in ML and Data Science (ETA summer 2024).

Interests in Data Science and Machine Learning

- Experience showed the extreme versatility within the field $\rightarrow$ Foundations!
- Introduced to Computer Science part of the field.
- Deep Learning.
- High Performance Programming: C/C++, Assembly, CUDA C.
- Bayesian Statistics.
- I strive to be a generalist. (!)

TableParser: Automatic Segmentation of Historical Documents

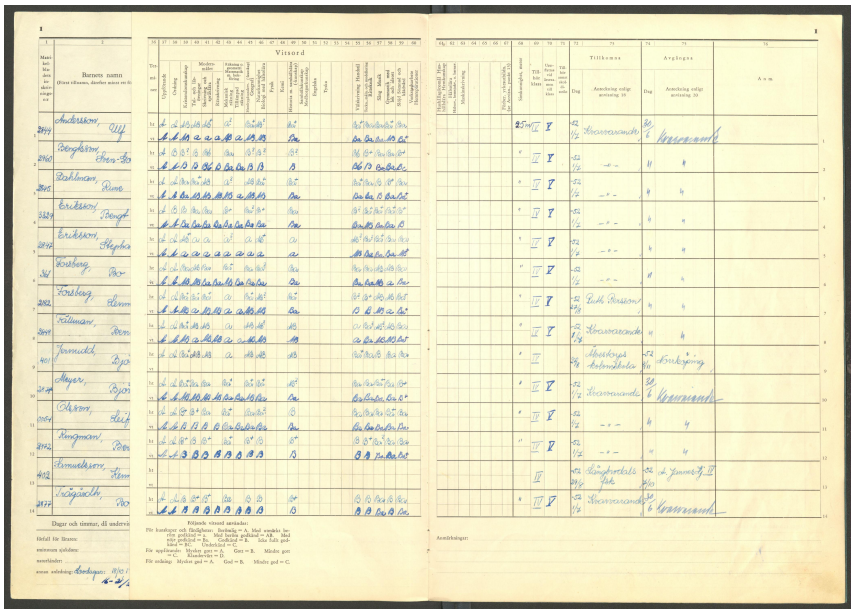


Figure: Swedish Gradesheet

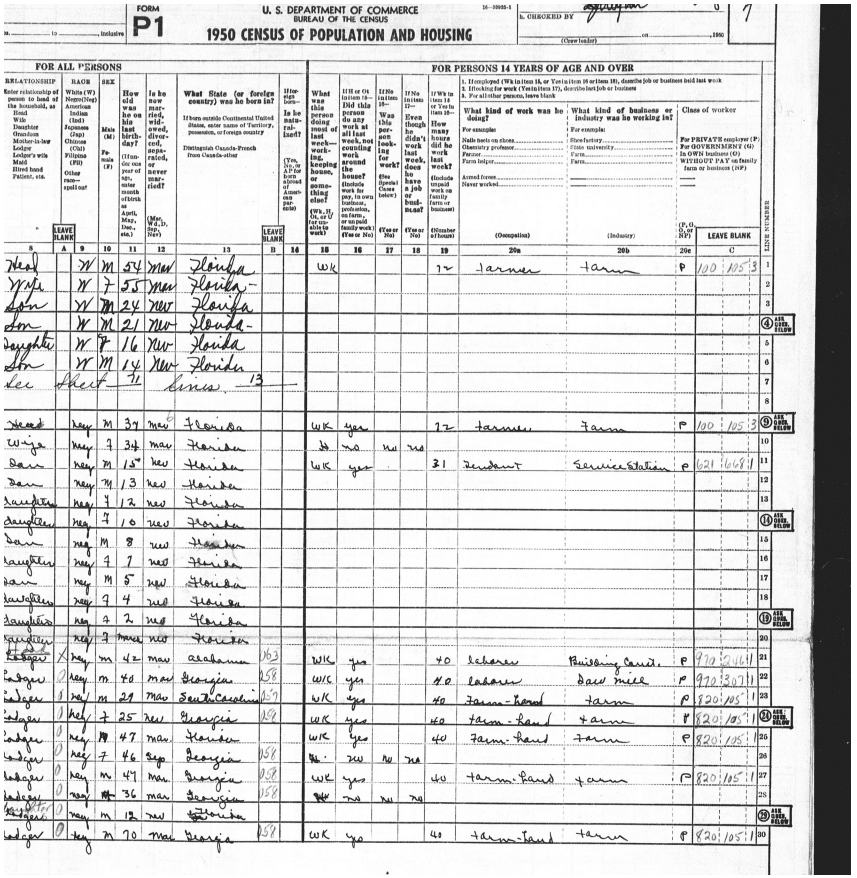


Figure: US Census

- Why not just use the Vision tools online?

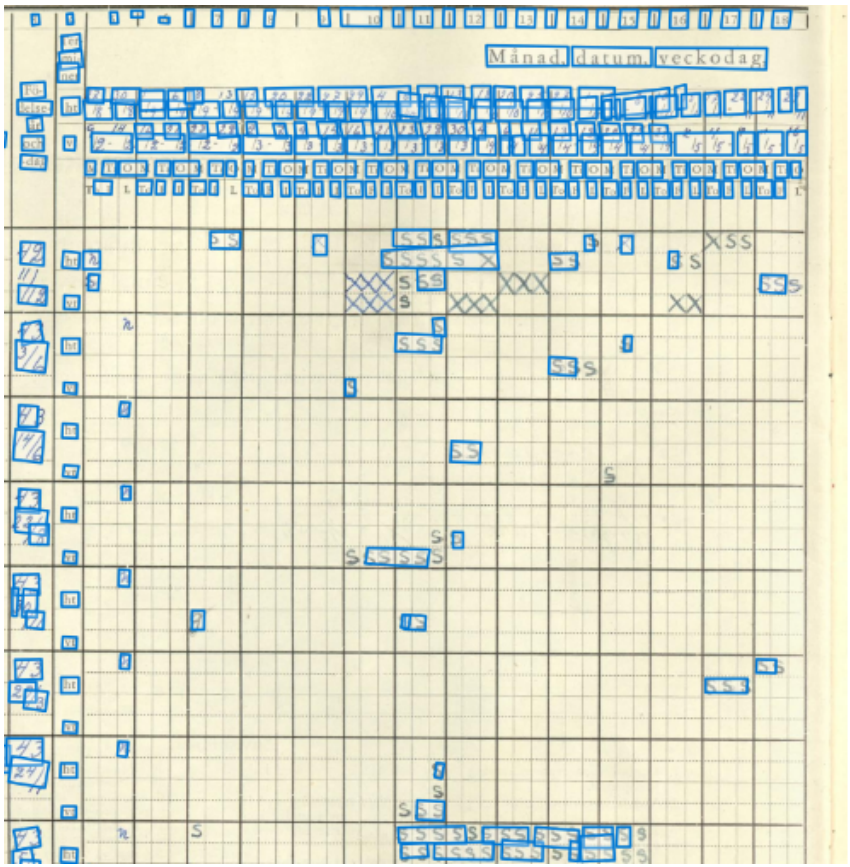


Figure: Swedish Gradesheet (Vision Studio)

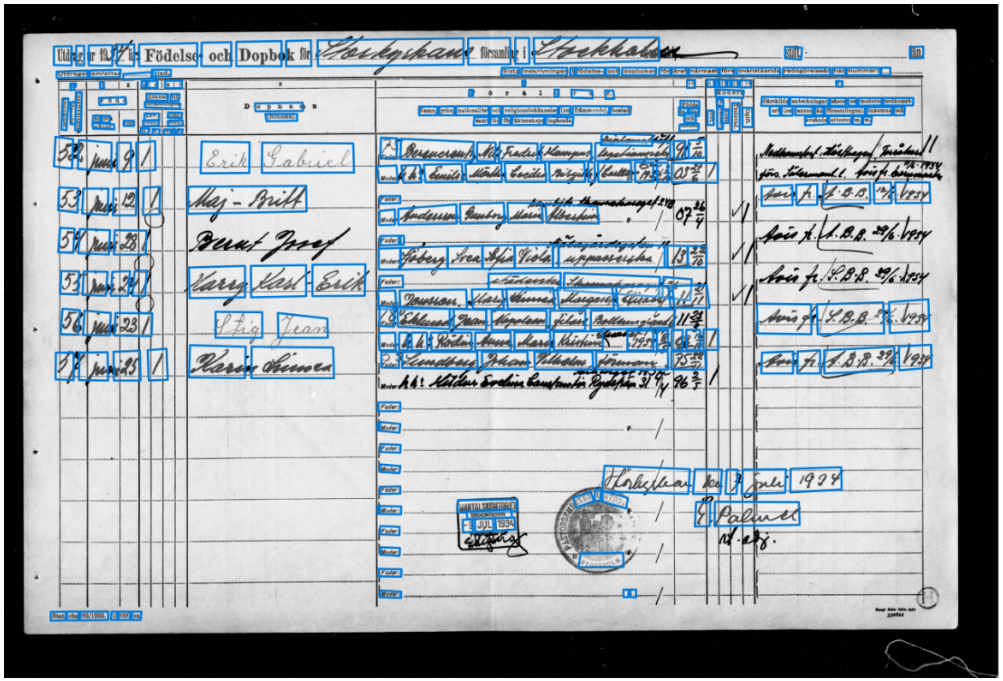
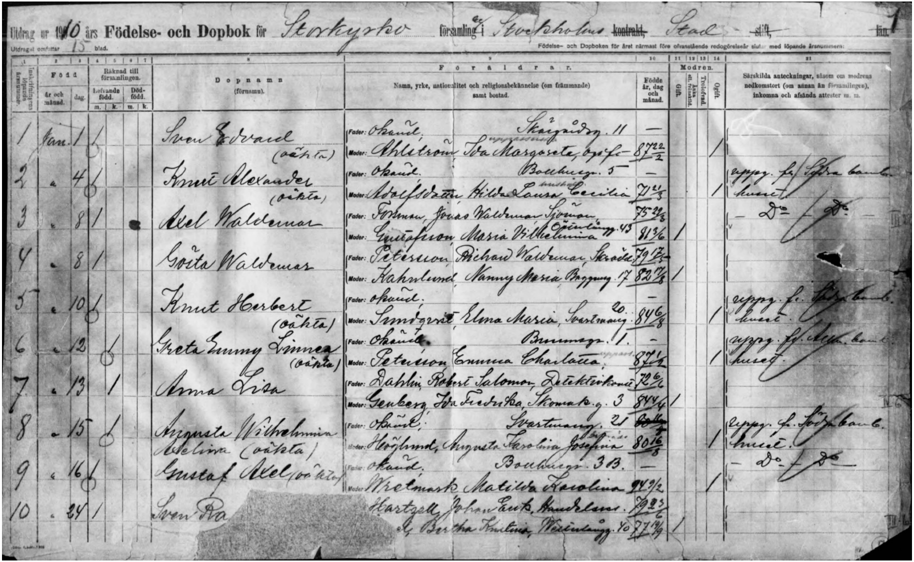


Figure: US Census (Vision Studio)

- They might even look like this



Utdrag ur 1910 års Födelse- och Dopbok för Storkyrko församling i Stockholms ~~kontrakt~~, Stad stift, län.

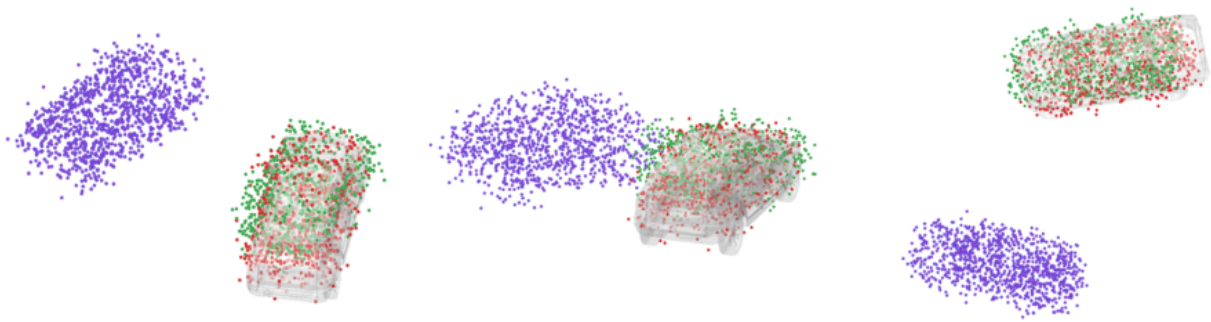
Utdraget omfattar 15 blad.

Födelse- och Dopboken för året närmast före ofvanstående redogörelseår slutar med löpande årsnummern:

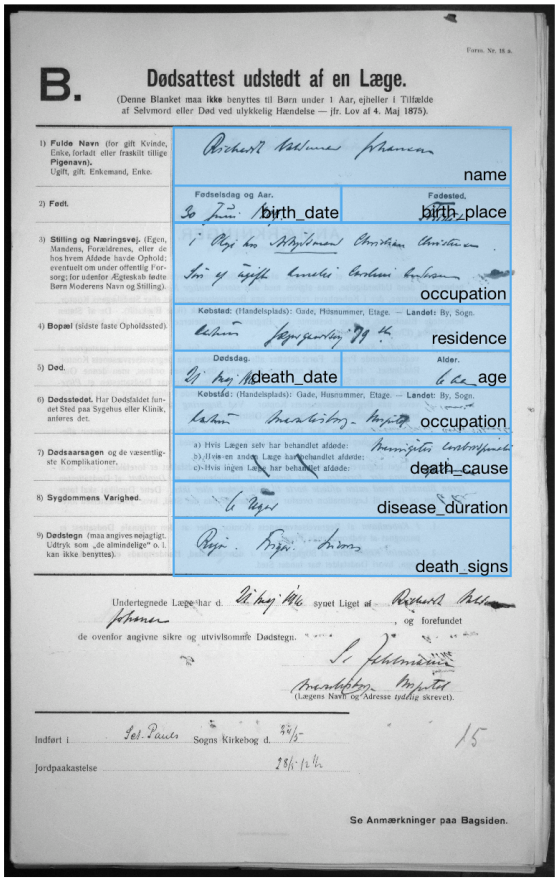
Inskrifningens löpande årsnummer	Född: År och månad	Född: dag	Räknad till församlingen: Lefvande född m.	k.	Död-född m.	k.	Dopnamn (förnamn)	Föräldrar: Namn, yrke, nationalitet och religionsbekännelse (om främmande) samt bostad	Födde år, dag och månad	Modren: Gift	Änka eller frånskild	Trolofvad	Ogift	Särskilda anteckningar, såsom om modrens nedkomstort (om annan än församlingen), inkomna och afsända attester m. m.
1	Jan.	1					Sven Edvard (oäkta)							
2	"	4					Knut Alexander (oäkta)							
3	"	8					Axel Waldemar							
4	"	8					Gösta Waldemar							
5	"	10					Knut Herbert (oäkta)							
6	"	12					Greta Emmy Linnea (oäkta)							
7	"	13					Anna Lisa							
8	"	15					Augusta Wilhelmina Adelina (oäkta)							
9	"	16					Gustaf Axel (oäkta)							
10	"	24					Sven Ra[illegible]							

TableParser: Automatic Segmentation of Historical Documents

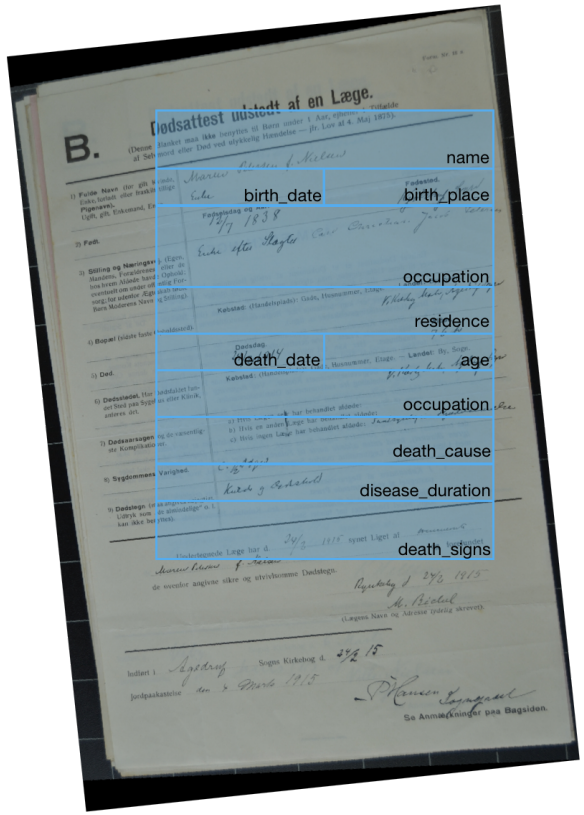
- Precise transcription of large and dense tabular documents is highly dependent on the quality of the segmentation of the tables.
- Propose a precise, fast and robust table identification and segmentation pipeline of tabular documents:
  - Semantic segmentation: Unet and SegFormer deep learning architectures for identifying table lines (on pixel level)
  - Point registration aligning target/template and source: FilterReg (likelihood based) and PCRNet (deep learning based)



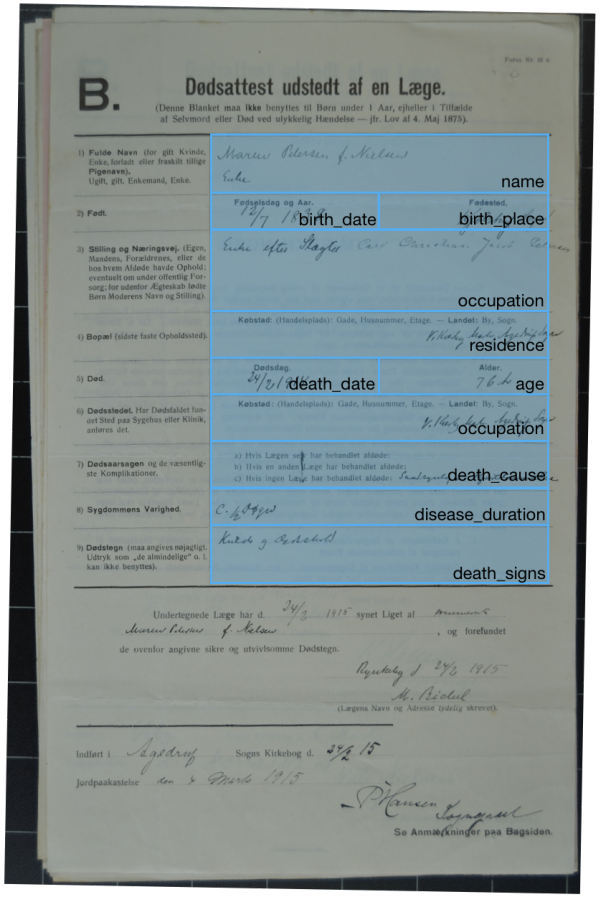
TableParser: Overview



Target

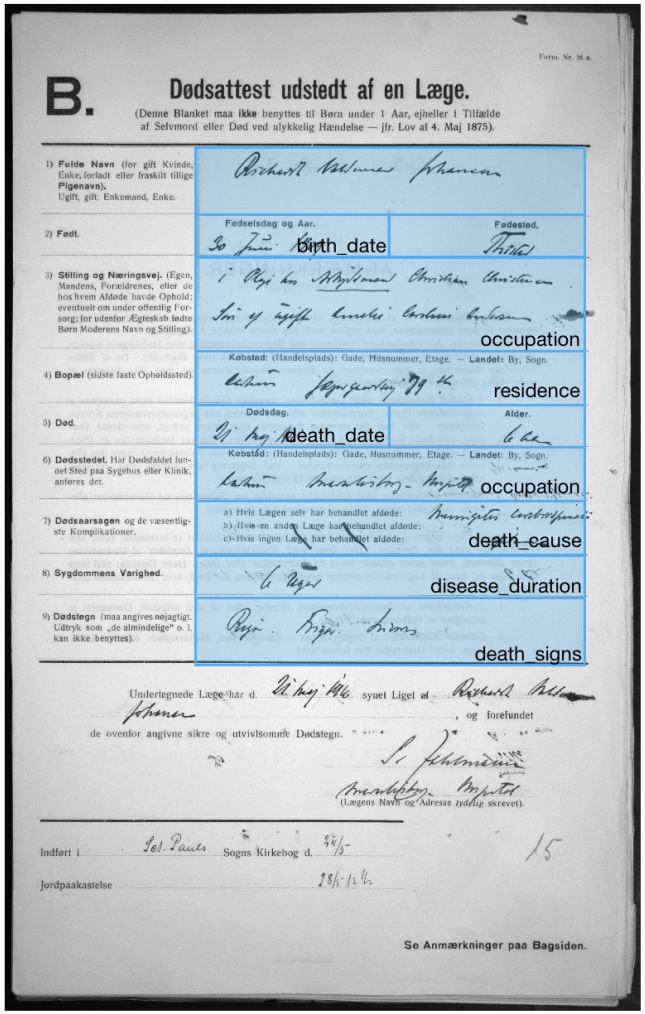


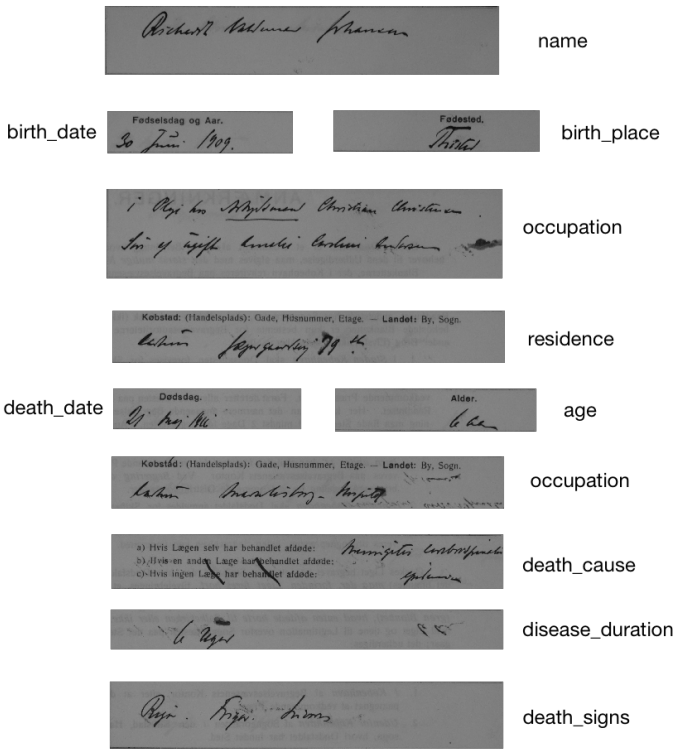
Source



Aligned

TableParser: Overview





TableParser: Overview

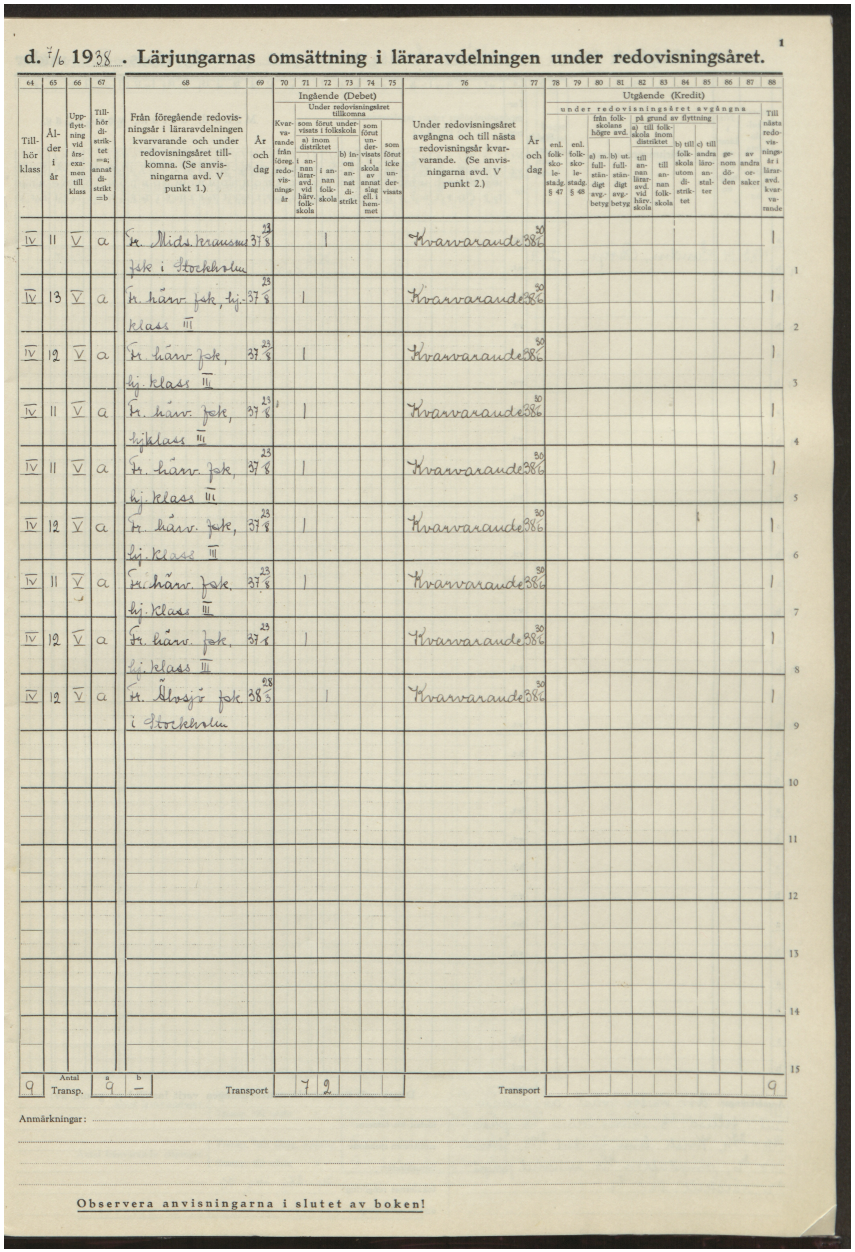
- Estimate reference document structure Y.
- Specify table cells of interest, Z.
- Estimate target document structure, X.
- Probabilistic point-to-point-set registration (href)
  - Leads to a motion parameter vector $\theta = (\theta_{rot}, \theta_{trans})$, which aligns $X(\theta)$ and Y.
  - Extract cells from image, i.e., Z from $X(\theta)$.
  - Maximize the following loglikelihood function.

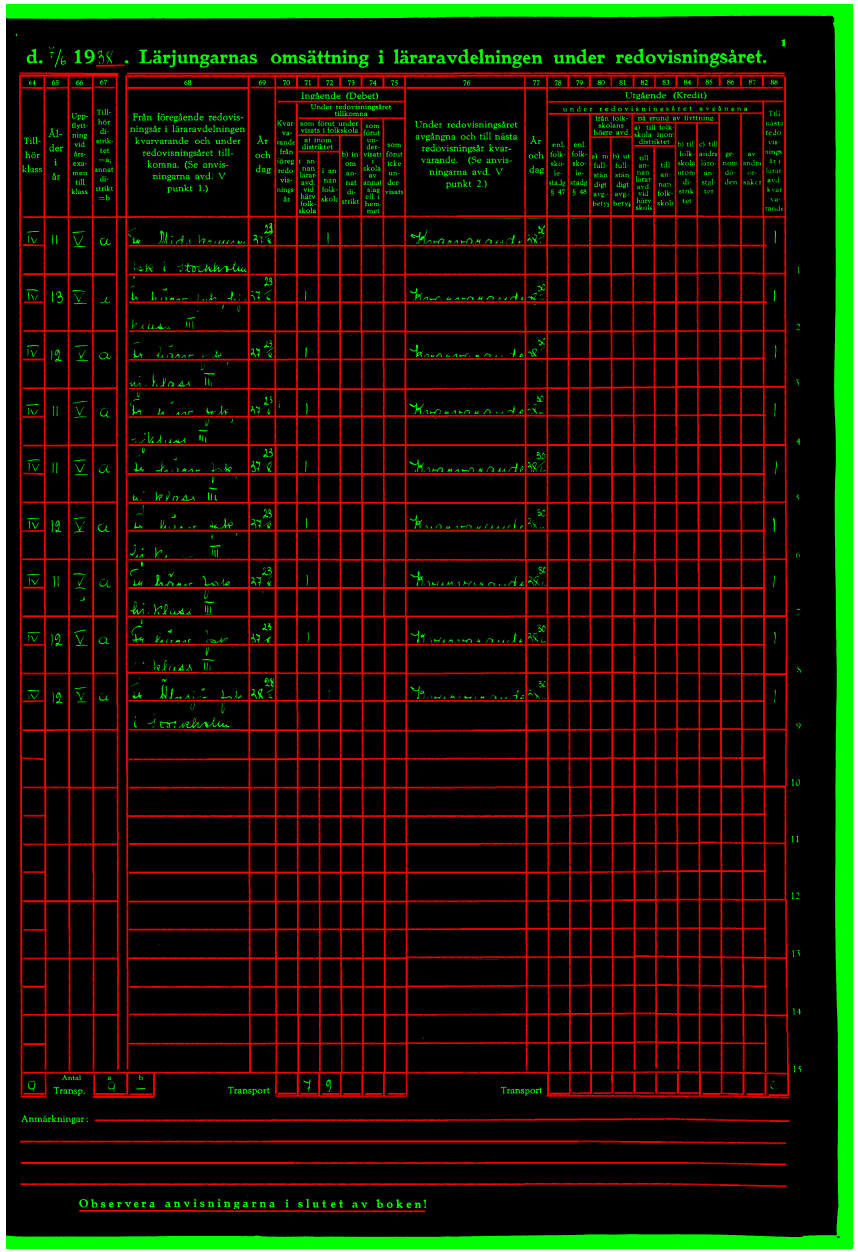
$$L = \sum_{i=1}^{M} \log(\sum_{j=1}^{N+1} P(y_j) p(x_i(\theta)|y_j))$$

What Are These Point Clouds and How Do We Get Them?

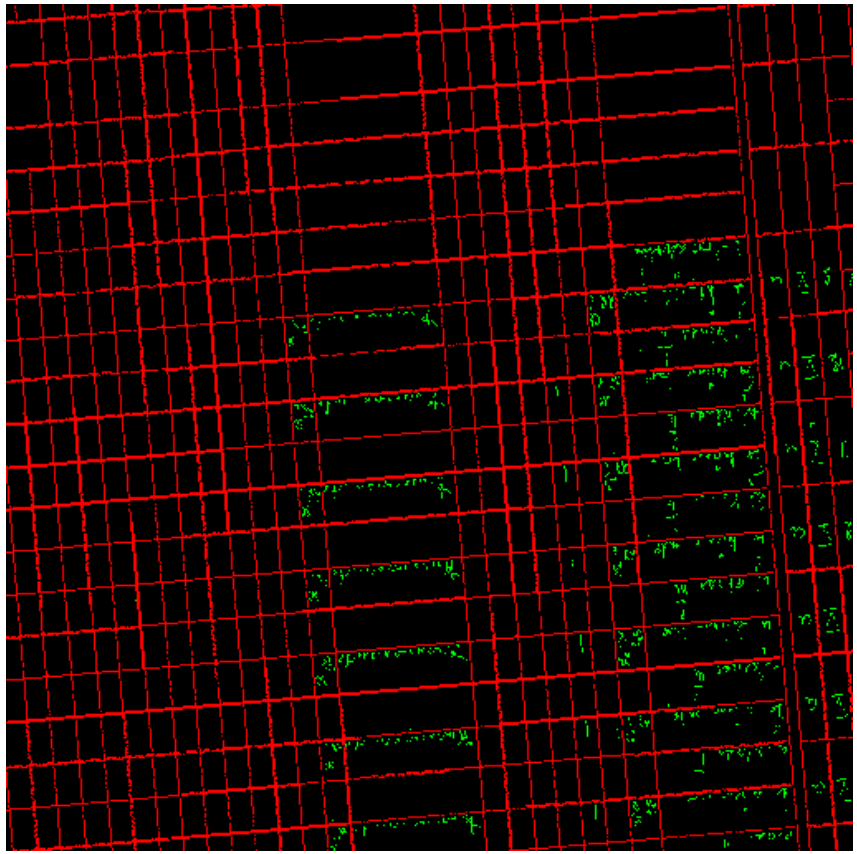
- We need point clouds that are approximately the same across documents and contain spatial information $\rightarrow$ table outlines.
- Semantic Segmentation models to go from image to cloud.
  - UNet (Ronnenberger et al. 2015)
  - SegFormer (Xie et al. 2021)
  - Both with advantages and disadvantages.

Training Semantic Segmentation Models





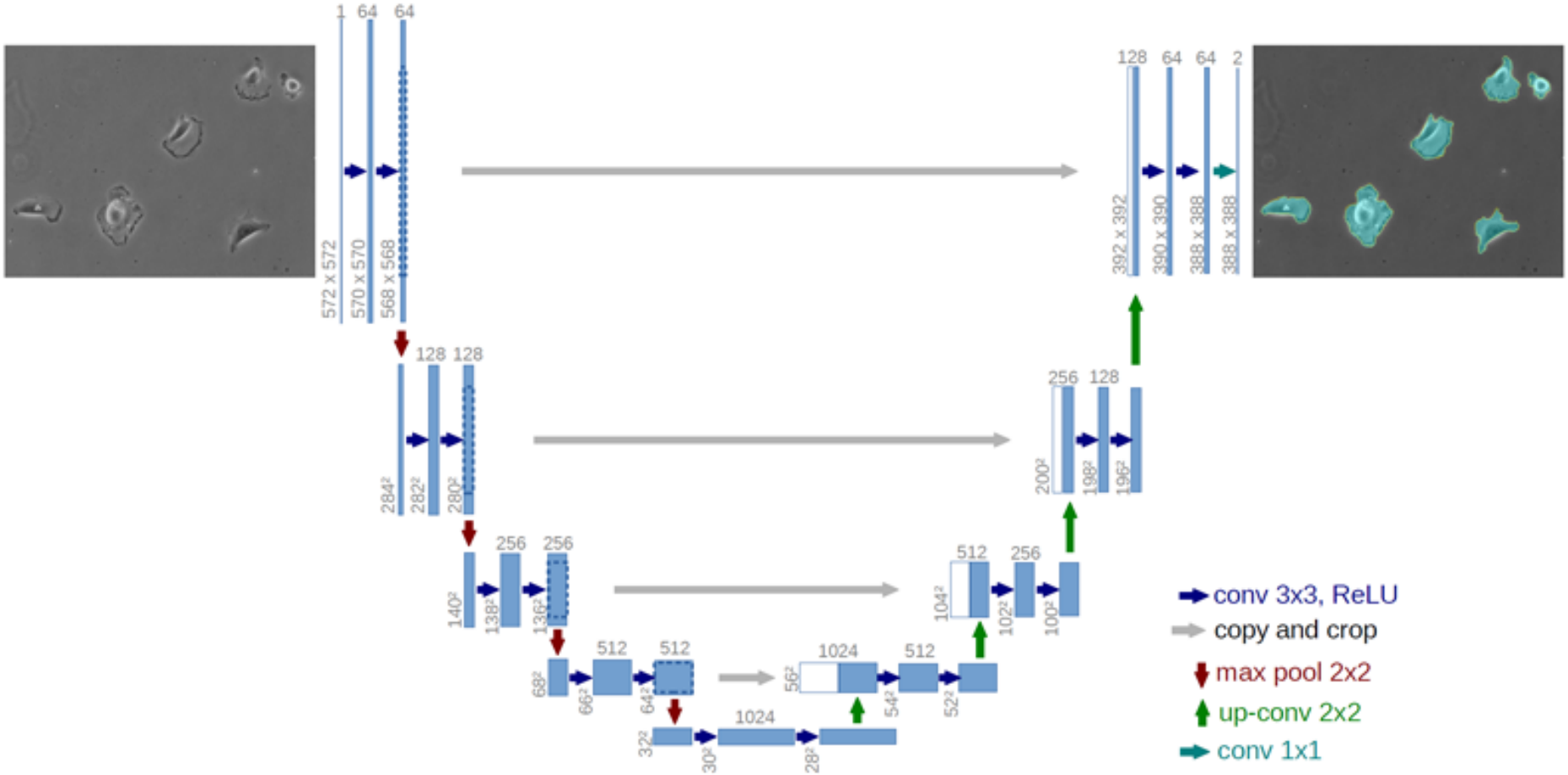
Training Semantic Segmentation Models: Augmentations



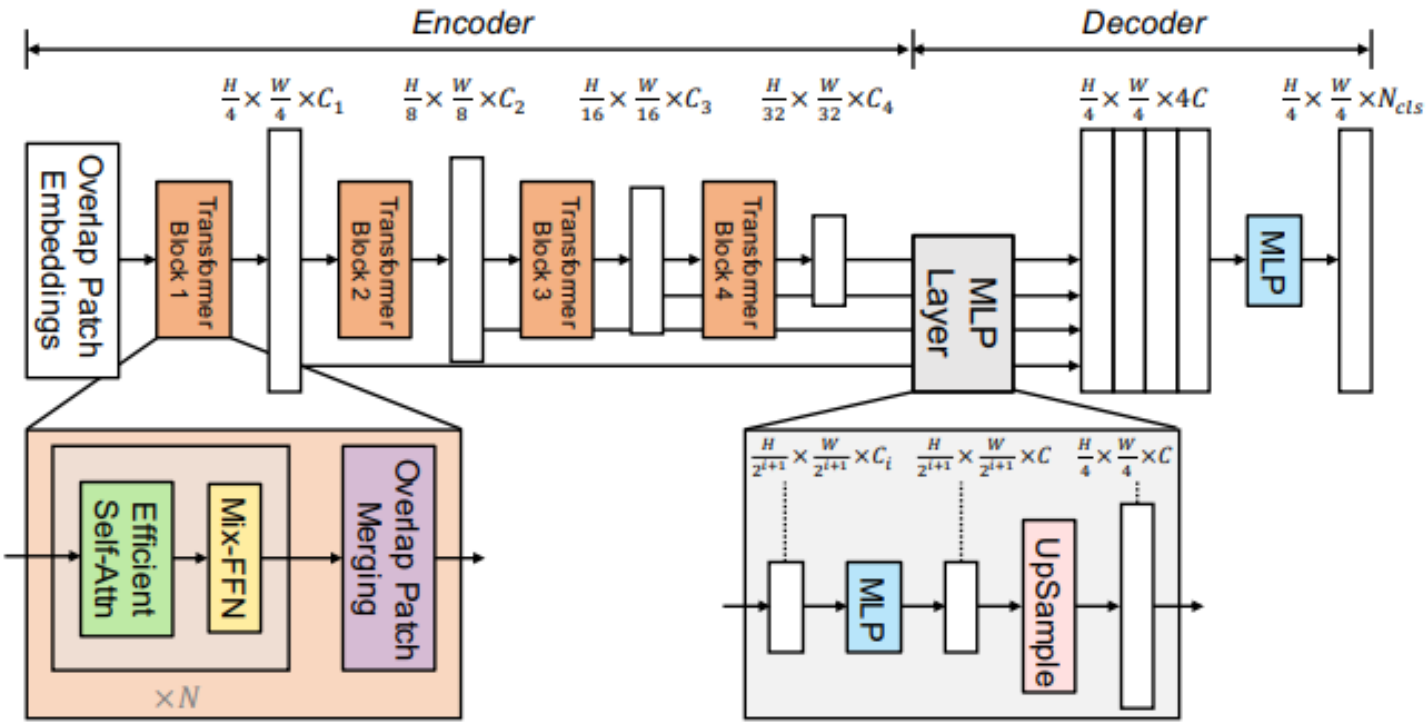
Training Semantic Segmentation Models: Computational Issues

- Documents easily reach approximately 4000×6000 pixels in size.
- That's 24 million variables.
- Vision models usually work in the ballpark of 224×224 to 512×512.
- Solution: Crop out 512×512 images randomly, which has several benefits
  - Less variability but more examples $\rightarrow$ better generalizability.
  - Combining random augmentations, different full scale documents and random crops (with random scaling) $\rightarrow$ an enormous data set.

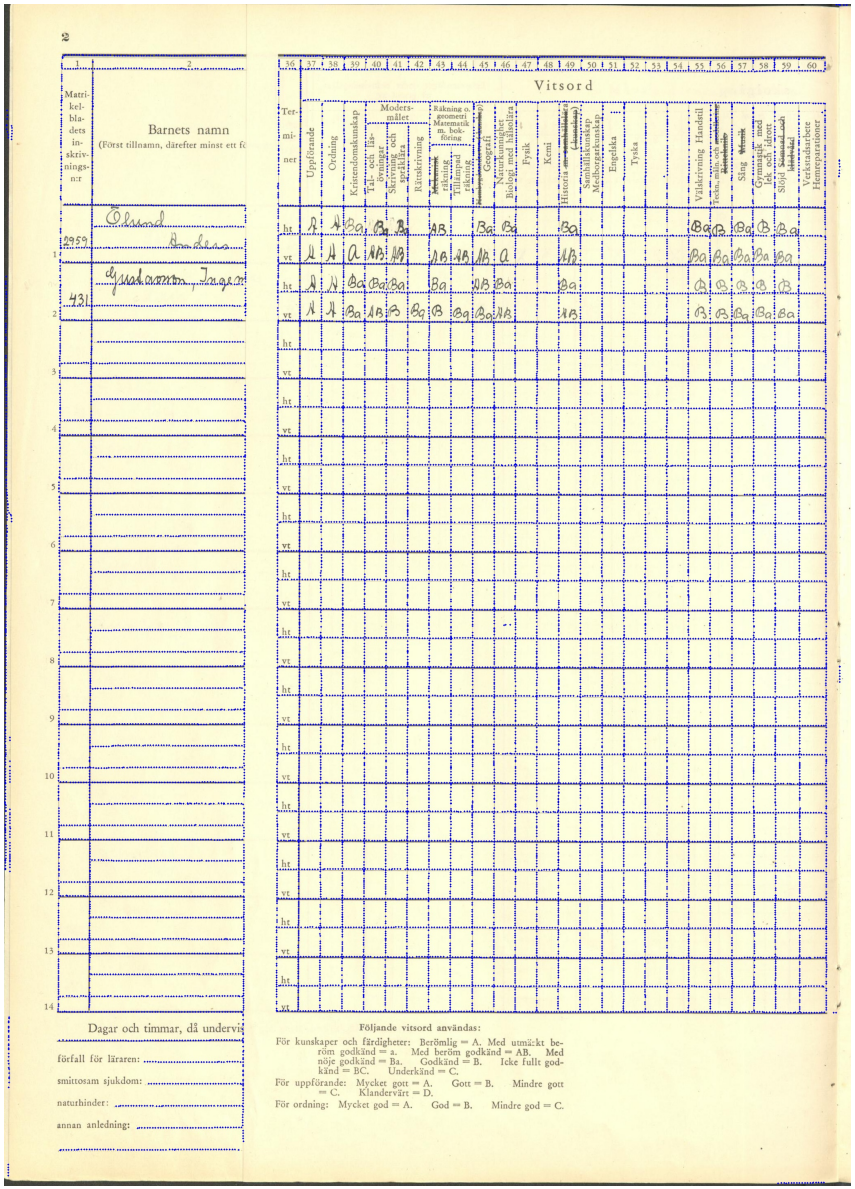
Training Semantic Segmentation Models: UNet

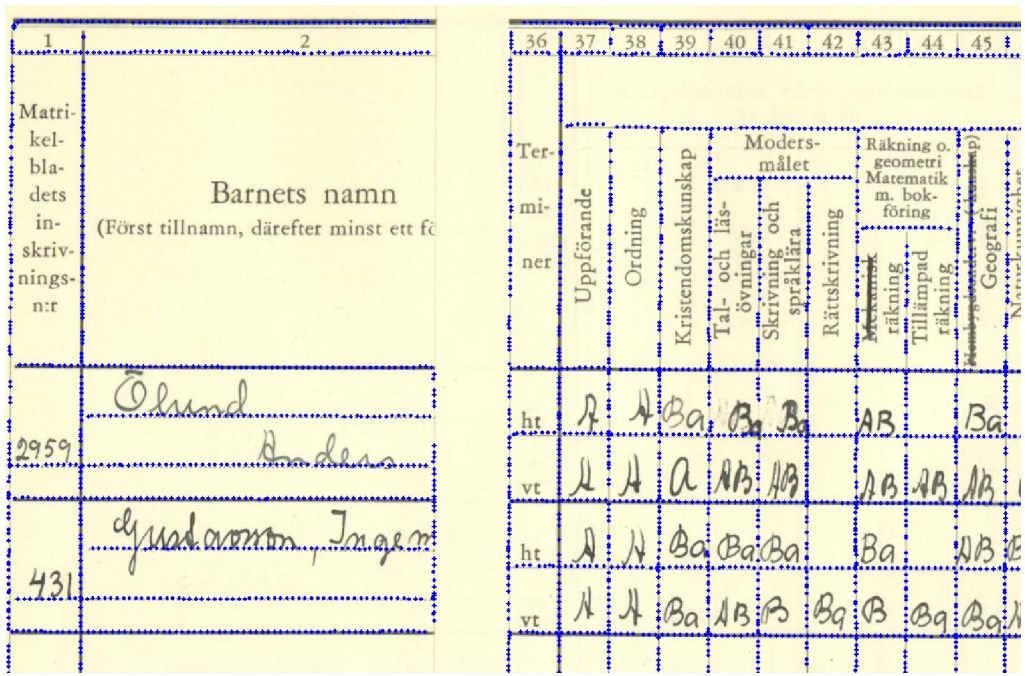


Training Semantic Segmentation Models: SegFormer

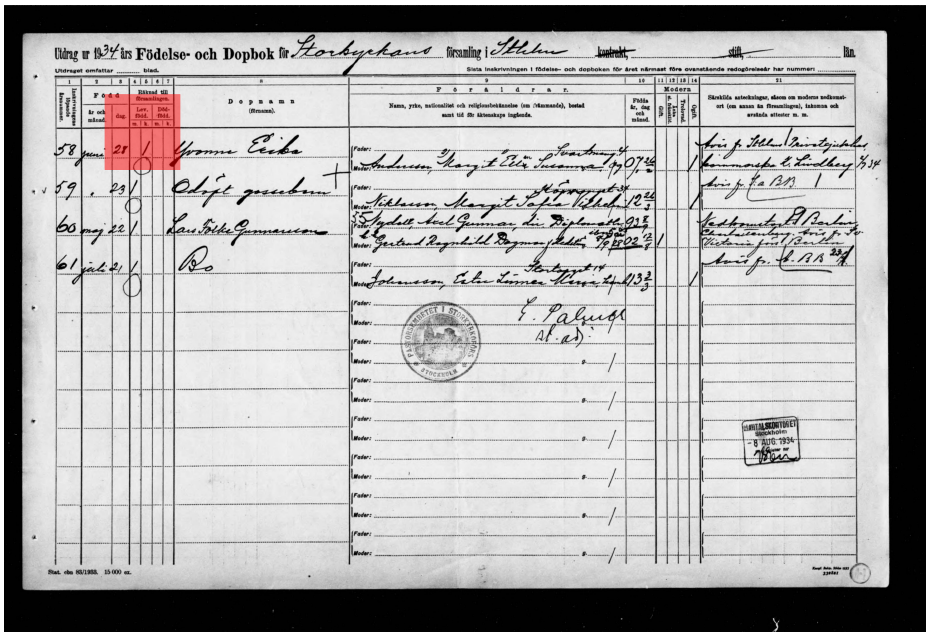


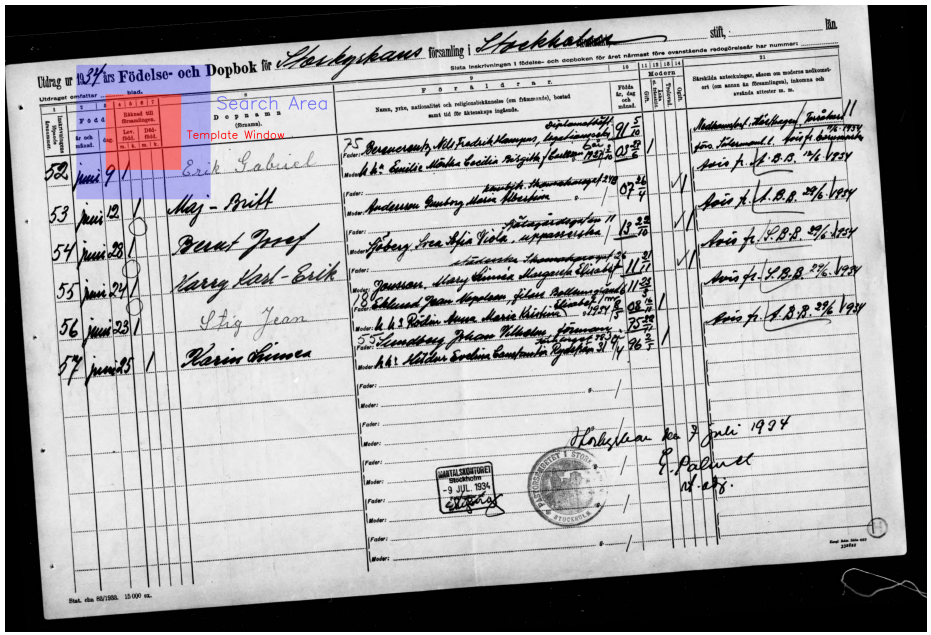
Training Semantic Segmentation Models: OOS Prediction





Applying the Models



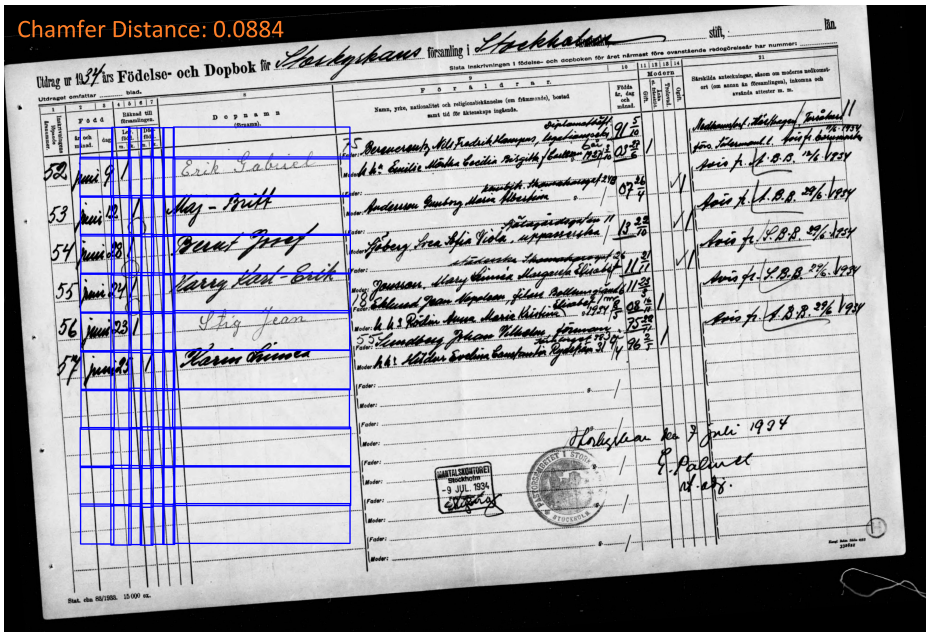


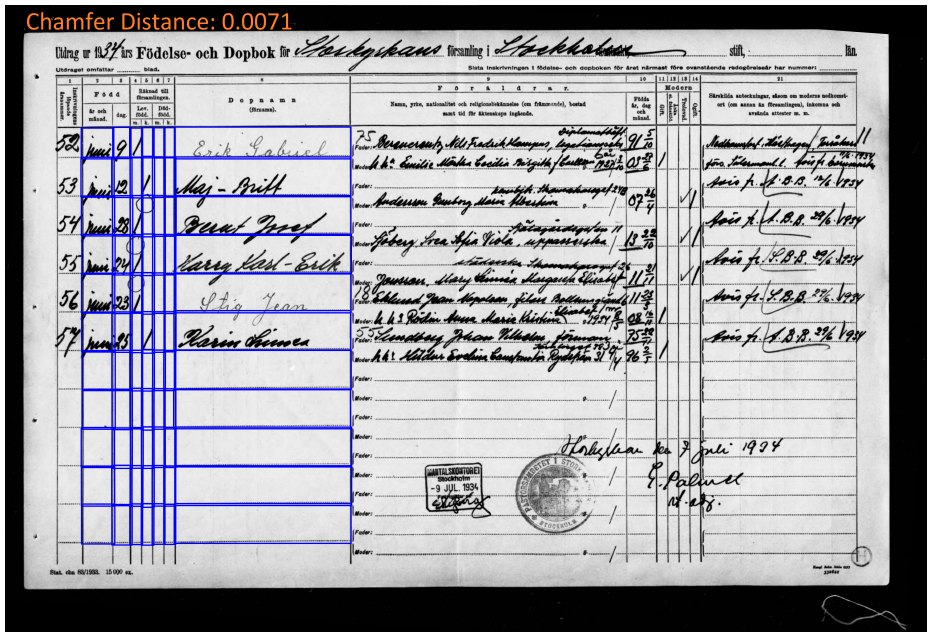
Sliding the Window

- Slide the window left to right and top to bottom.
- At every frame, do semantic segmentation.
- Compare point clouds using Chamfer Distance

$$d_{CD}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} ||x - y||_2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} ||y - x||_2.$$

Applying Cells Directly vs. After Transformation





A New Data Scientist (Personal Insights and Tips)

Disclaimer: These are personal suggestions and not going one route or the other does not mean you're worse off!

- Data Science and ML content and development is ever increasing.
- Learn how to learn!
- Fundamentals, fundamentals, fundamentals.
  - Within Programming
  - Within Math (Stats, Linear Algebra, Calculus, Probability)
- It's a marathon, not a sprint.
- Think: What could you see yourself working with in the future?

Mathematics (Personal Insights and Tips)

- Foundations
  - Statistics
  - Probability
  - Linear Algebra
  - Calculus
- Figure out whether you benefit most from reading or listening (or both).
- Check out 3Blue1Brown on Youtube for some amazing series.
- Apply what you learn!

Data Analysis (Personal Insights and Tips)

- What areas/domains are you passionate about?
  - Medicine (computer vision, signal processing, ...)
  - Environment.
  - News.
- Search for data sets.
- Do tiny analyses.
- Kaggle competitions.

Deep Learning (Personal Insights and Tips)

- Consider learning about Autograd
  - Implement a toy version yourself (just scalar)
- Learn partial derivation.
  - Also in the case of broadcasting!
  - Dead neurons.
  - Inefficient training.
- Why is weight initialization important, why do we need it?
- Art of stacking well-behaved differentiable blocks.
- Check out Andrej Karpathy's youtube.
- The annotated transformer: https://nlp.seas.harvard.edu/2018/04/03/attention.html
- Try and implement foundational papers.

Programming (Personal Insights and Tips)

- Don't get paralyzed by the "correct" way to code.
- Learn a language with an interpreter (Python, R, ...).
- Learn a language with a compiler (C/C++, Zig, Rust, Java, ...)
- Projects, projects, projects, don't get stuck in tutorial hell.
  - What interests you? Hobbies, scientific/non-scientific fields, etc.
  - See if you can combine it with a programming project.
  - If you're learning a new language: Program something you've done before.
- Why could low-level programming be important to me?
  - Where others hit a wall, you will not.
  - Less reliance on other departments.
  - Be able to create fast and *good* programs.
- If this doesn't interest you, no worries!

On ChatGPT and LLMs (Personal Insights and Tips)

- The GPS of critical thinking.
- Don't take everything it says at face value.
- A templating machine and enhanced google.
- It's not deterministic.

General Tips

- Start small, it's okay to get overwhelmed.
- Find your area of interest (will require a lot of experimenting)
- Find a mentor, many people gladly lend a hand because they likely got help too at some point.
- Track your time, where does it go?

Resources

ML Math:

- Mathematics for Machine Learning. (Deisenroth et al.)

Linear Algebra:

- Linear Algebra Done Right (Sheldon Axler)
- Linear Algebra and Learning from Data (Gilbert Strang)

Statistics / ML

- Introduction to Statistical Learning: https://www.statlearning.com/ (free)
- All of Statistics
- Probabilistic Machine Learning: An Introduction. (free) https://probml.github.io/pml-book/

Deep Learning (Field moves so fast)

- Dive Into Deep Learning. http://d2l.ai/index.html (free)

Q&A