

Information Retrieval Over Historical Scans with Non-trivial Layouts

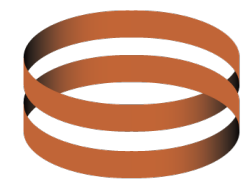
Zejiang Shen

Data Science Fellow at IQSS at Harvard University

zejiang_shen@fas.harvard.edu | www.szj.io



HARVARD
Faculty of Arts and Sciences



The Institute for
Quantitative Social Science

Contents

1. Background & Overview

Why historical data matters

2. Document Layout Analysis

A combination of traditional and deep learning approach

3. Structural Information Extraction

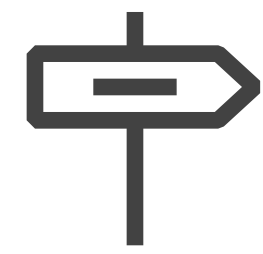
Work with noisy OCR outputs

4. Conclusion

Thinking from a Data Science Perspective

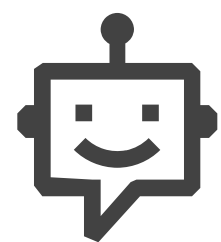
Why historical documents matter?

Why historical documents matter?



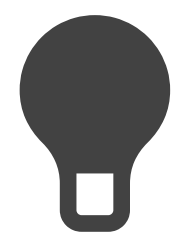
It's about the nature of our society

- ▶ Many problems today can be traced back to the past



It's about understanding the future

- ▶ Using past observation to predict the future

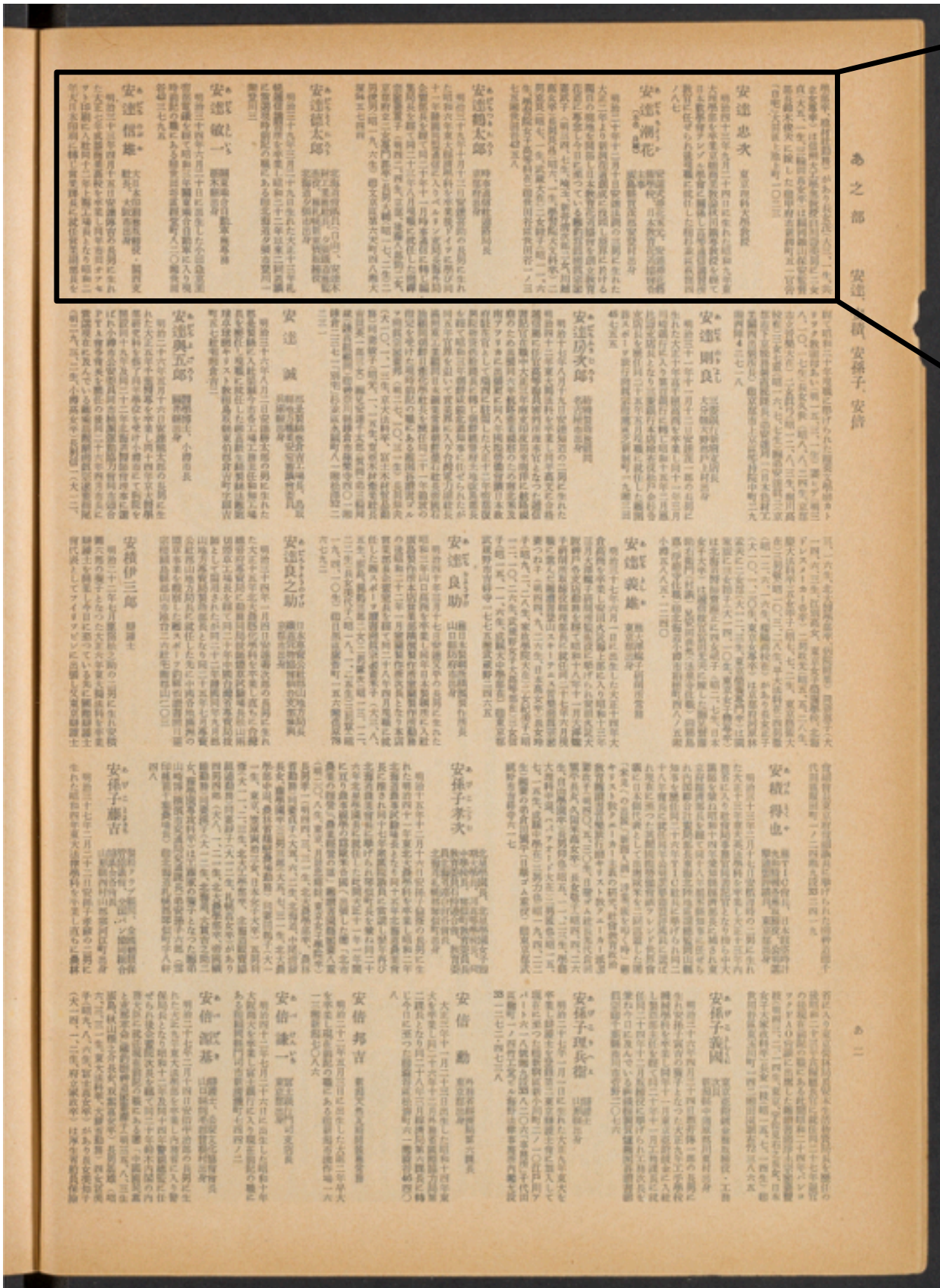


And it's also challenging to analyze

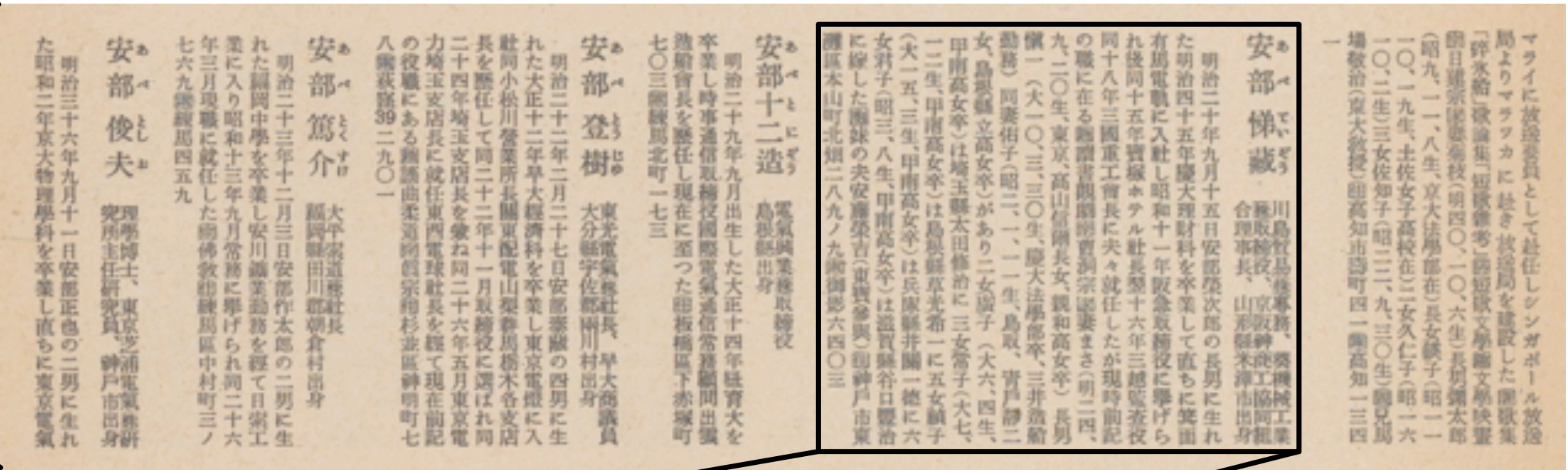
- ▶ Noise in the text, the old printing technology, paper wearing, and the scan technology

Overview of our problem

- We want to analyze a specific type of publication, the reference books, in Japan around the mid-20th century.



(a)



(b)

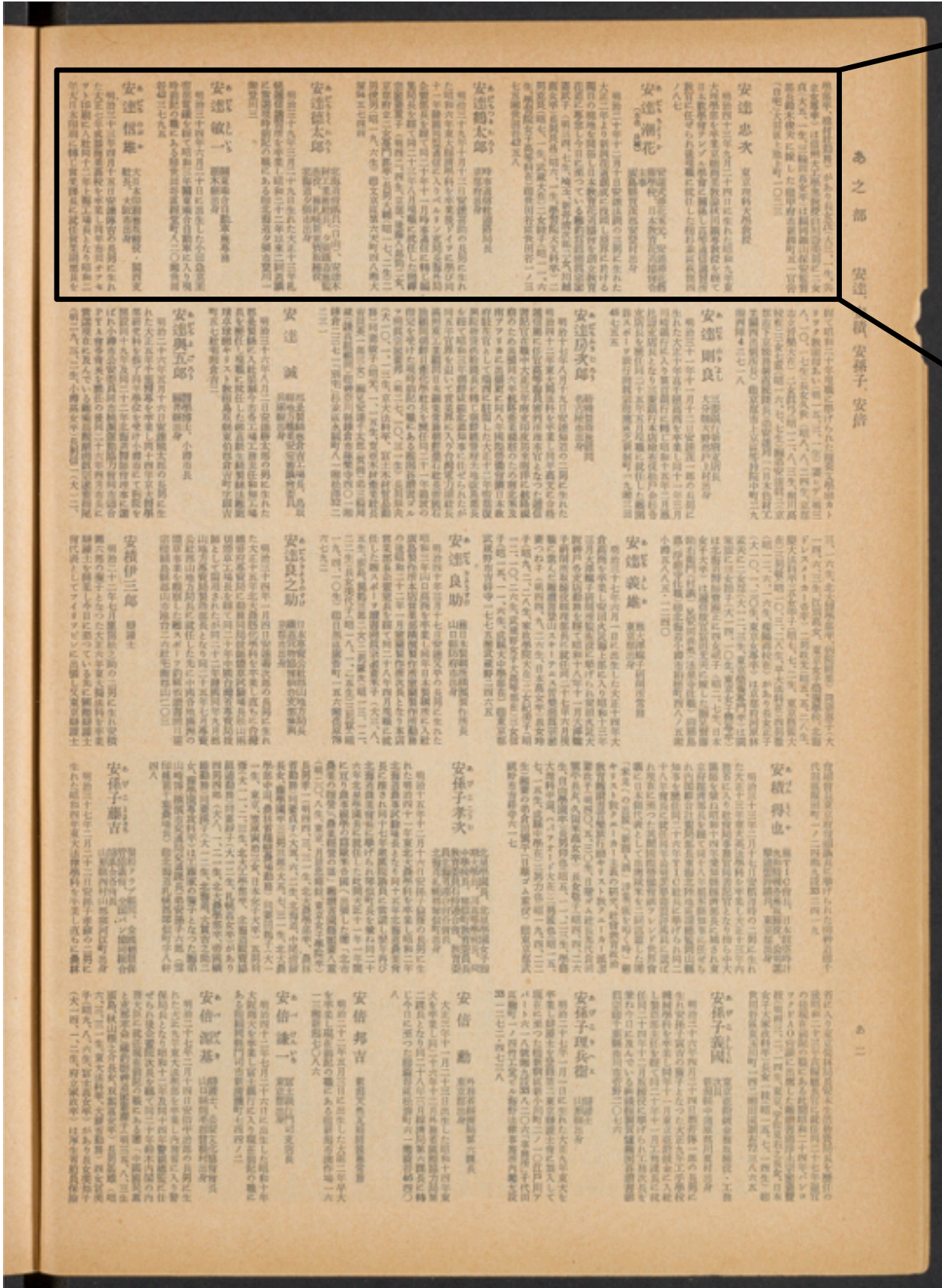


(c)

- Name
- Short Biography
- Position

Overview of our problem

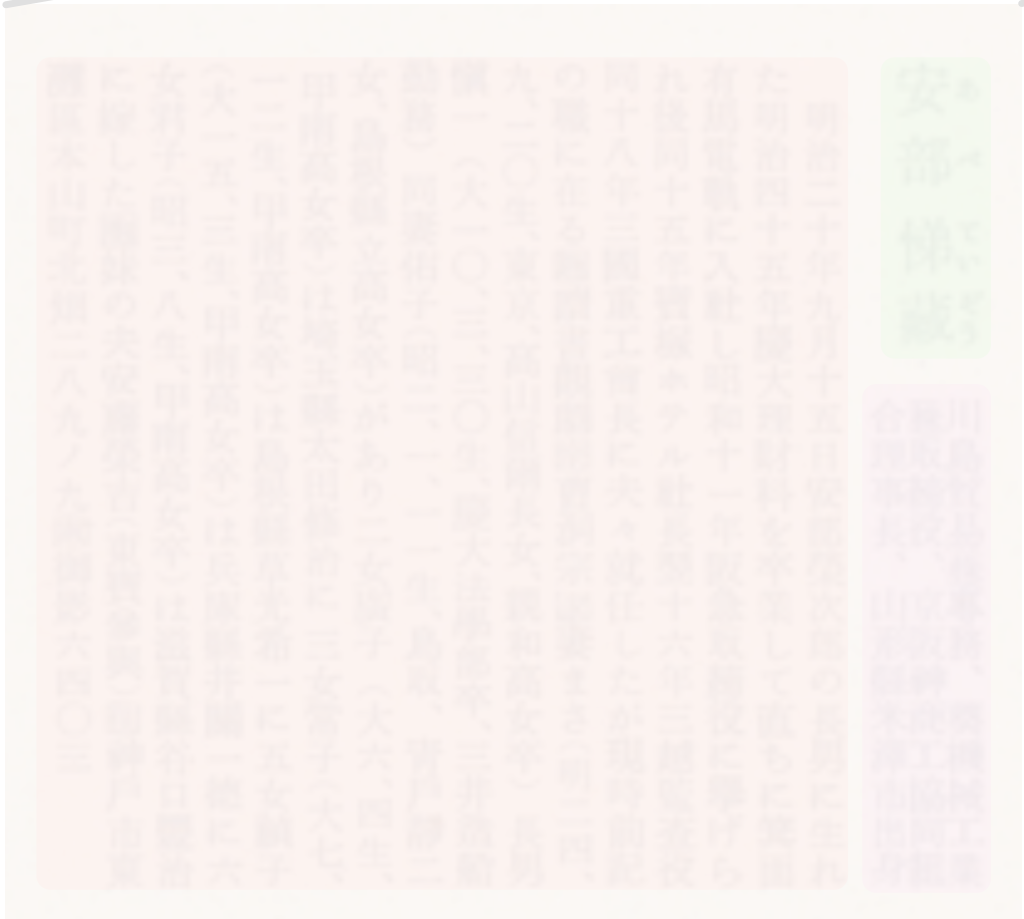
- We want to analyze a specific type of publication, the reference books, in Japan around the mid-20th century.



(a)



(b)




(c)

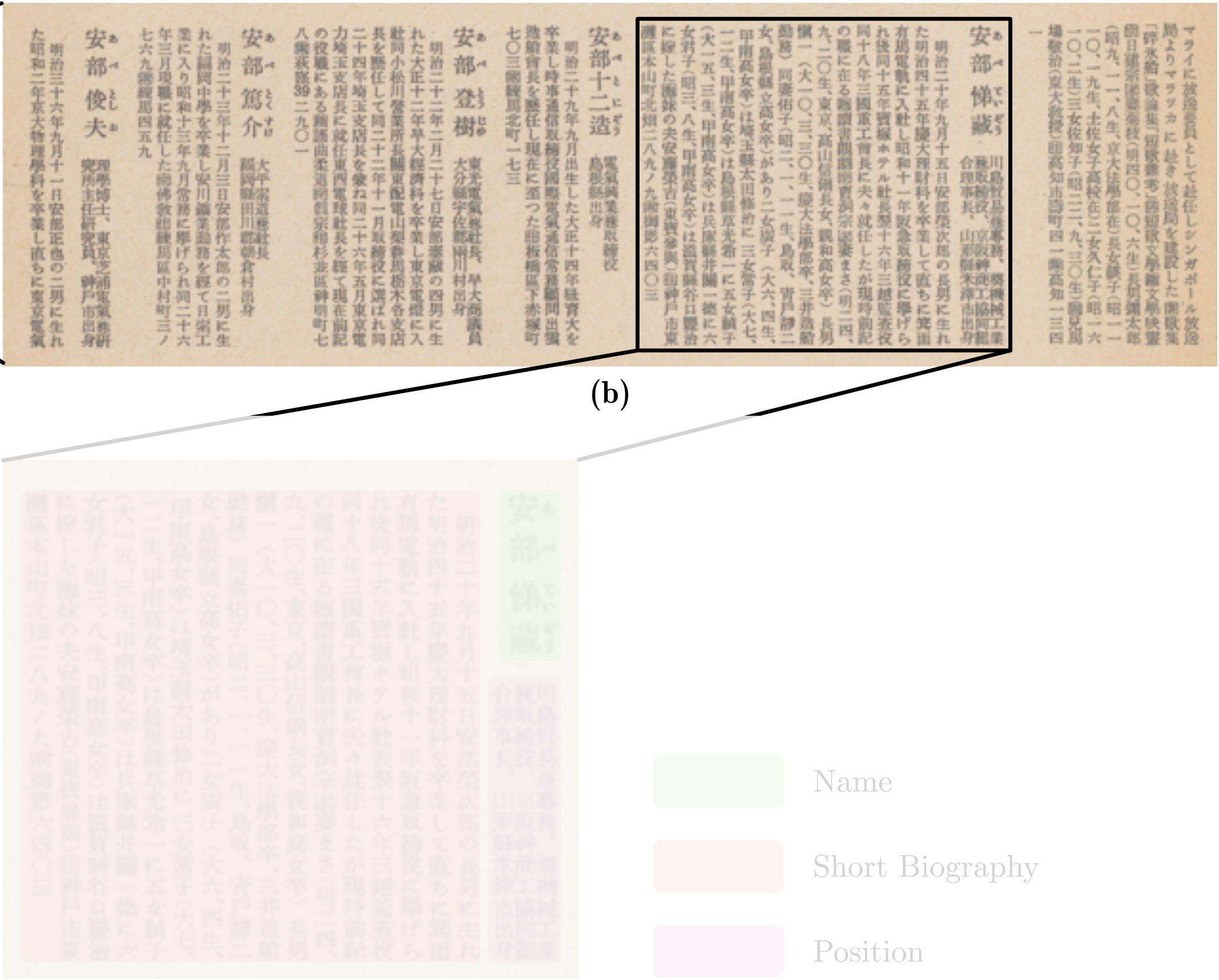
- Name
- Short Biography
- Position

Overview of our problem

- We want to analyze a specific type of publication, the reference books, in Japan around the mid-20th century.



(a)



(b)

Name


Short Biography

Position


(c)

Overview of our problem


- We want to analyze a specific type of publication, the reference books, in Japan around the mid-20th century.



(a)



(b)



(c)

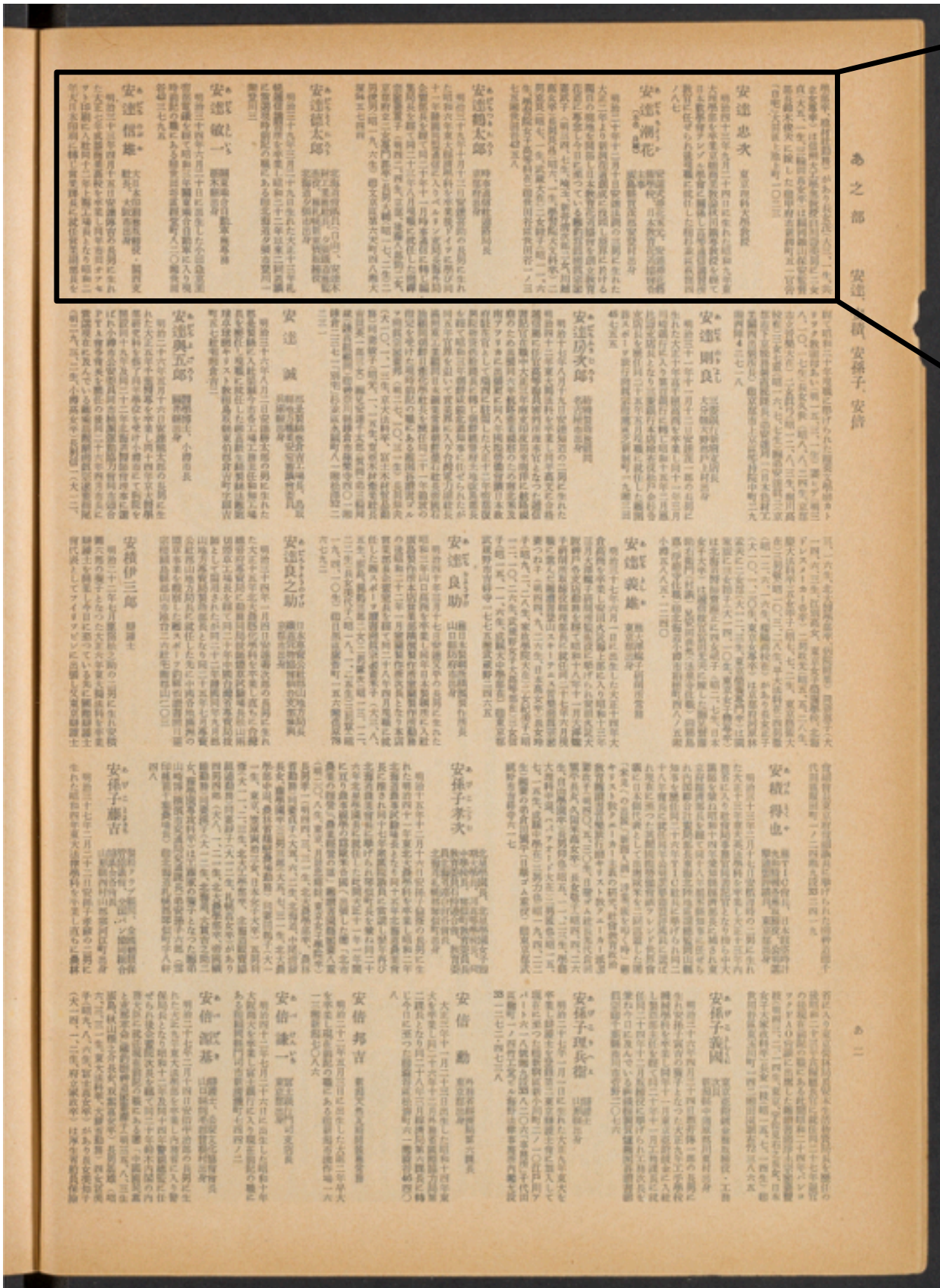
Name

Short Biography

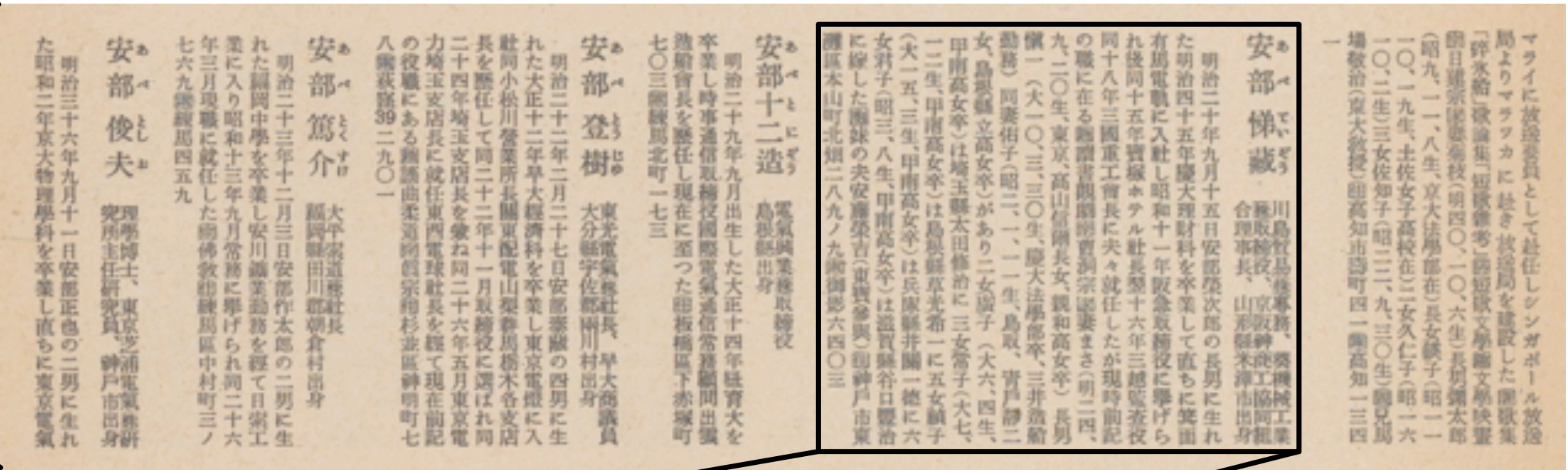
Position

Overview of our problem

- We want to analyze a specific type of publication, the reference books, in Japan around the mid-20th century.
- It's about important individuals in Japan Society at that time
- Let's call it the *who's who* book



(a)



(b)

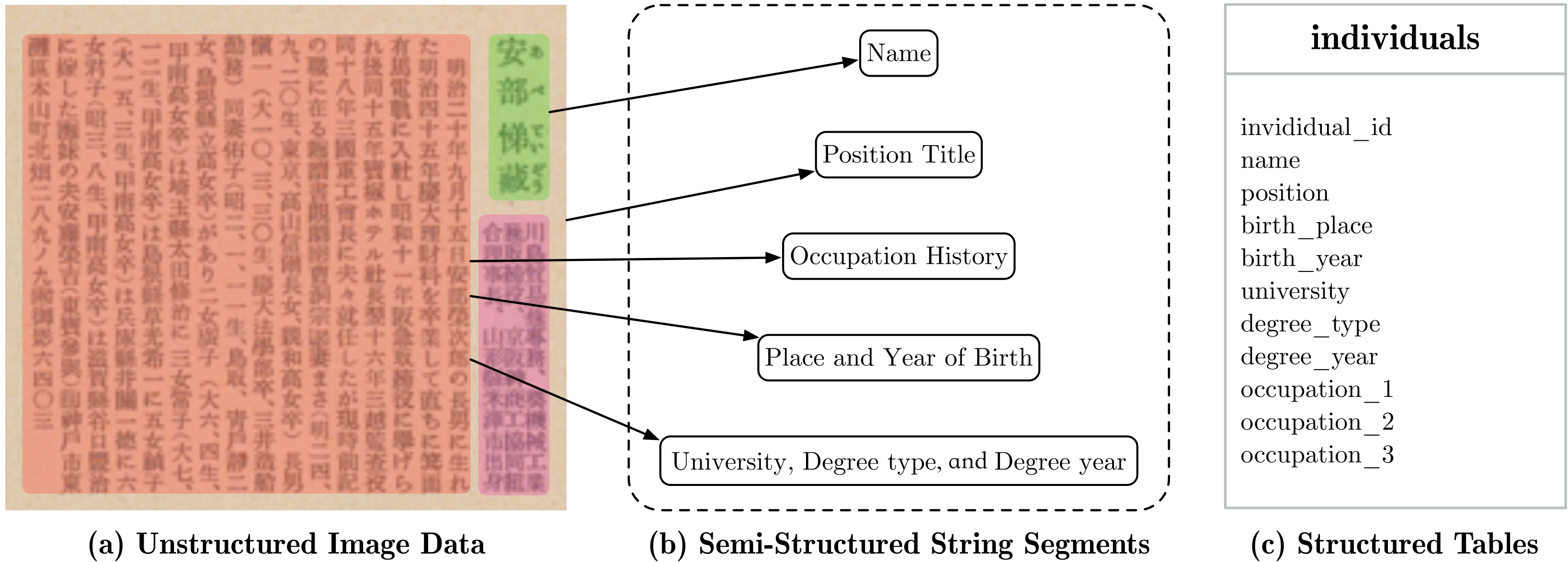


(c)

- Name
- Short Biography
- Position

Overview of our problem

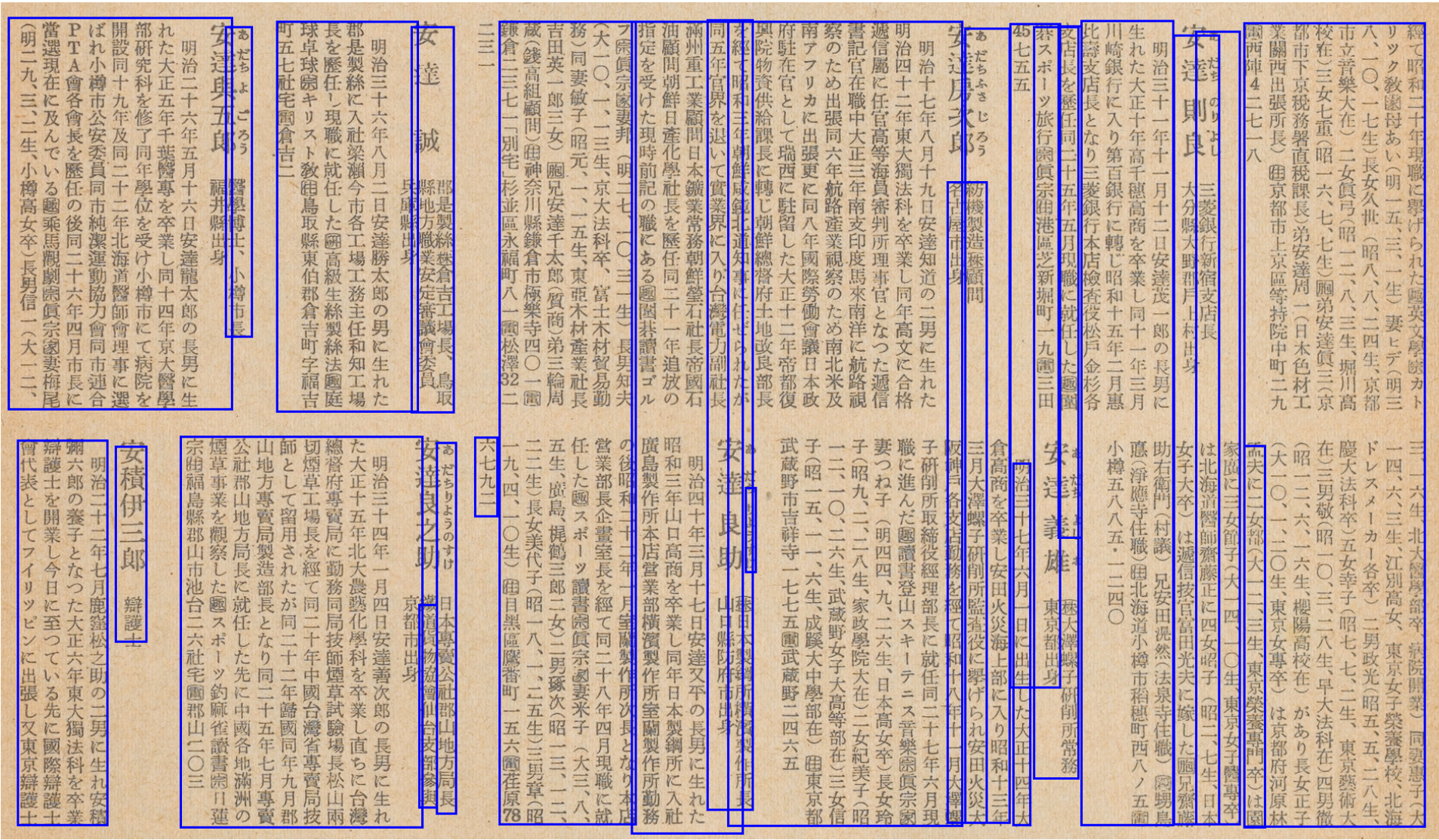
- Our objective is to transform the unstructured scan images into structured feature tables
- The tables contains the important information about the individuals in the book



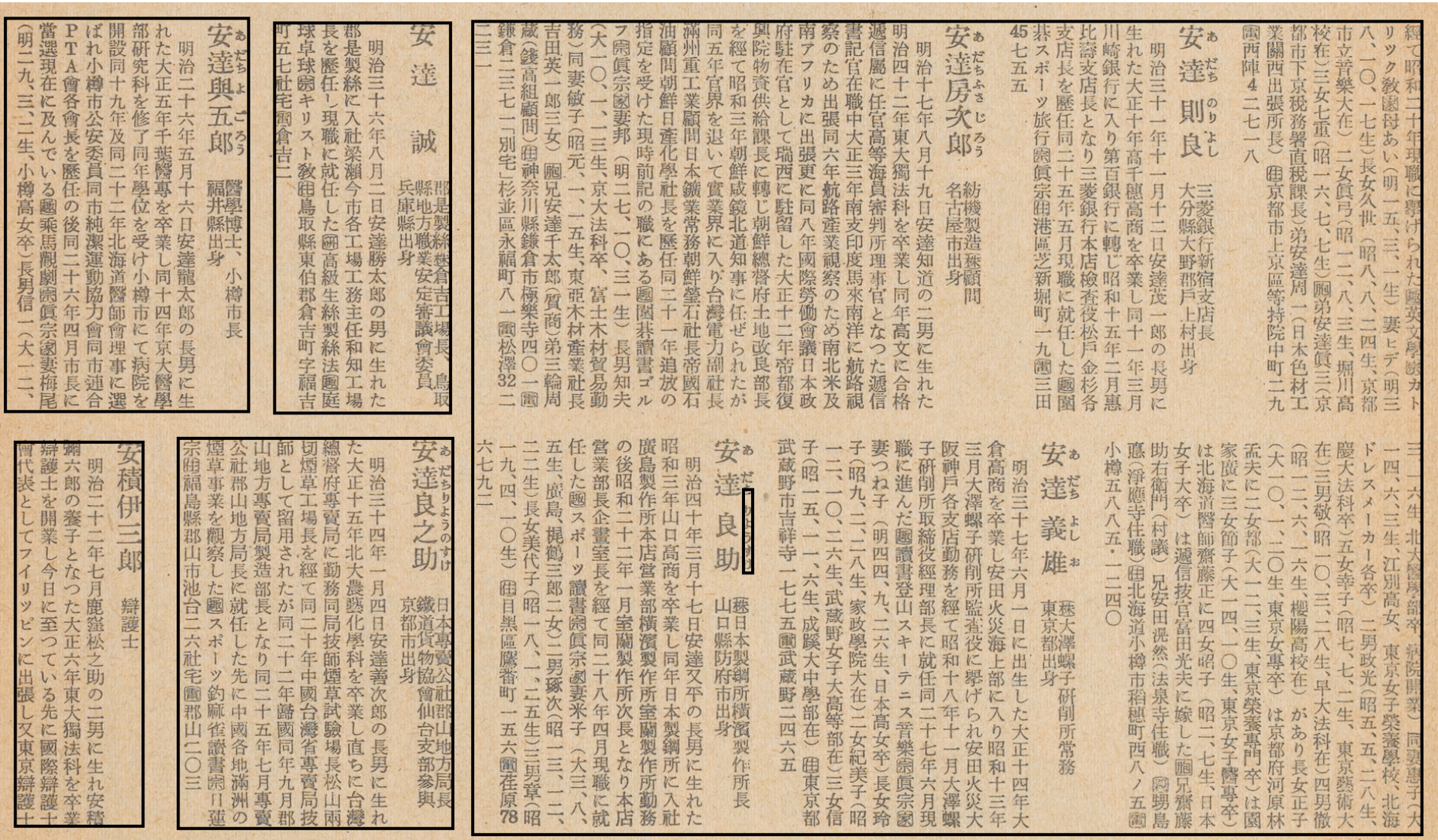
The challenge of complex layouts

- Currently, the Swiss-knife tool for dealing with Document Images is Optical Character Recognition (OCR).
- However, when the organization of documents becomes complex, the method could hardly work.
- For example, let's check Google Cloud Vision (GCV)'s output on our dataset.

The challenge of complex layouts



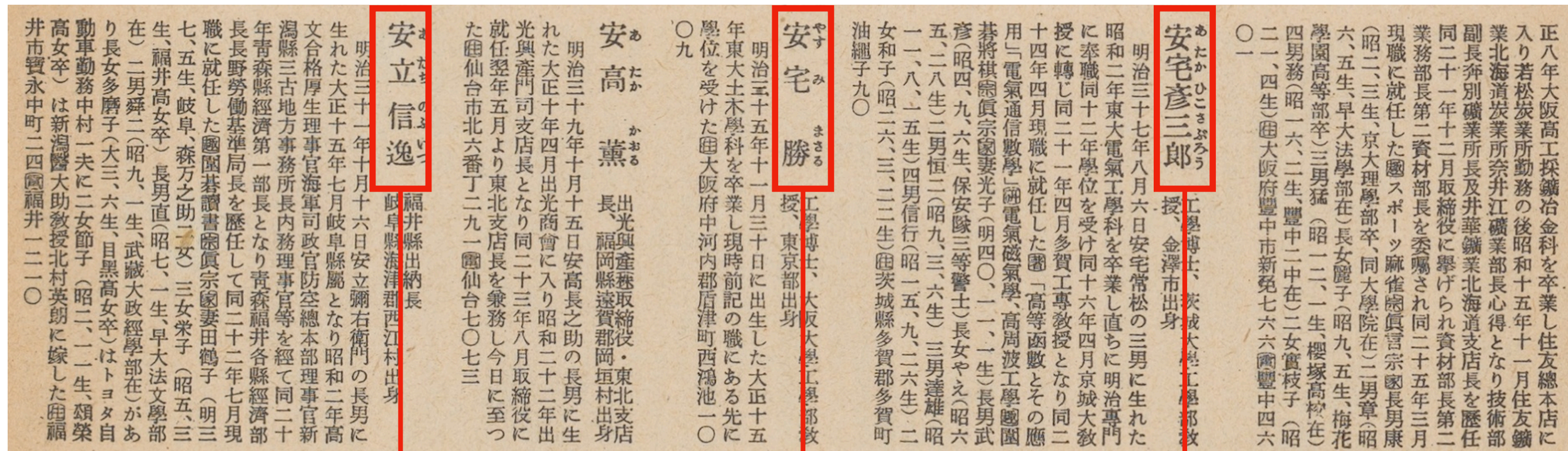
(a) Blue boxes are GCV Detected Paragraph Blocks



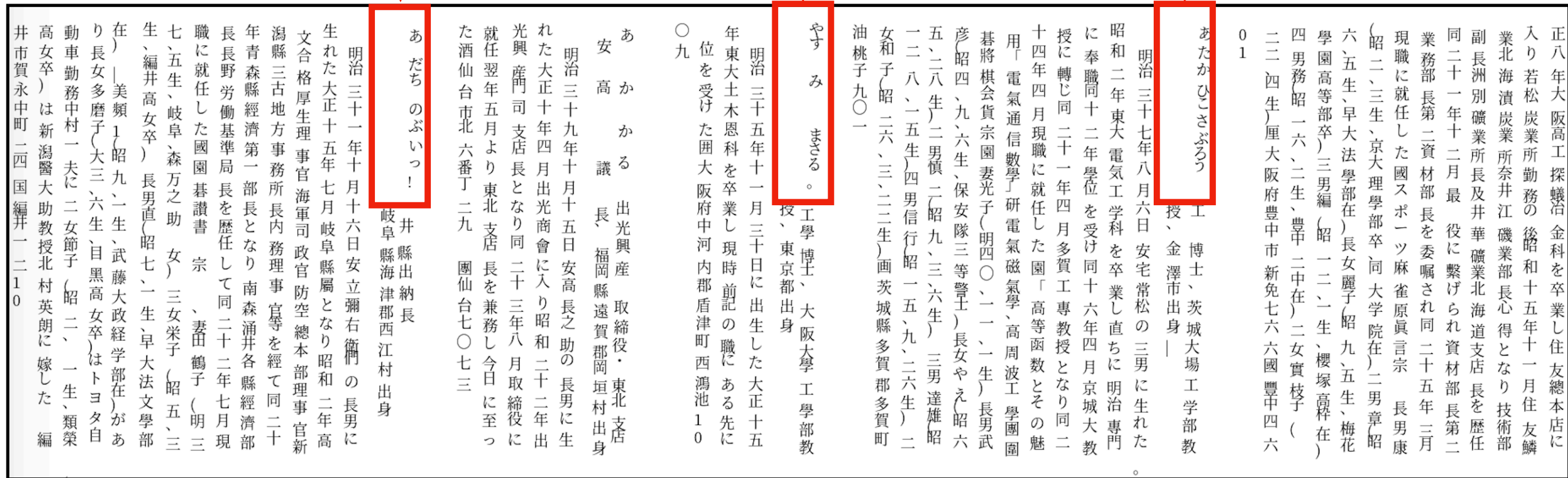
(b) Black boxes are GCV Detected Text Blocks

- GCV combines text according to the text and paragraph blocks.
- The inaccurate text block extraction means the output text will be messy.
 - It will incorrectly connect text in different blocks

The challenge of complex layouts



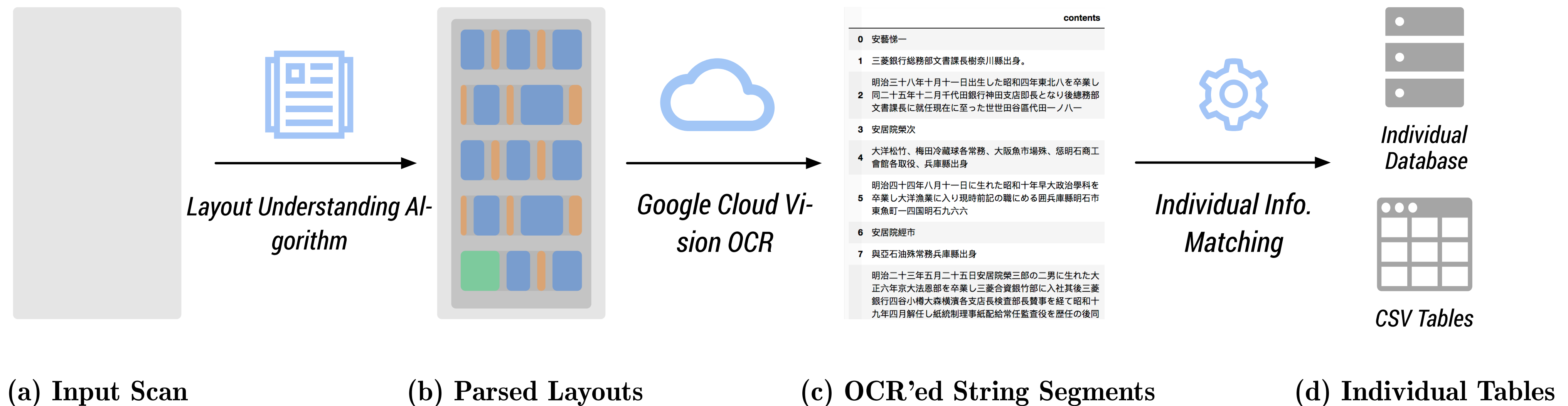
(a) The Input Page Scan of a Row



(b) The Reconstructed GCV output based on coordinates

- GCV cannot detect the large fonts
- But if we send in simple layouts, it can do a great job!

The solution



- Build a layout understanding algorithm to identify the text blocks
- Crop the images based on the layouts and use GCV for character recognition
- And build the additional text understanding algorithms for generating the individual information tables

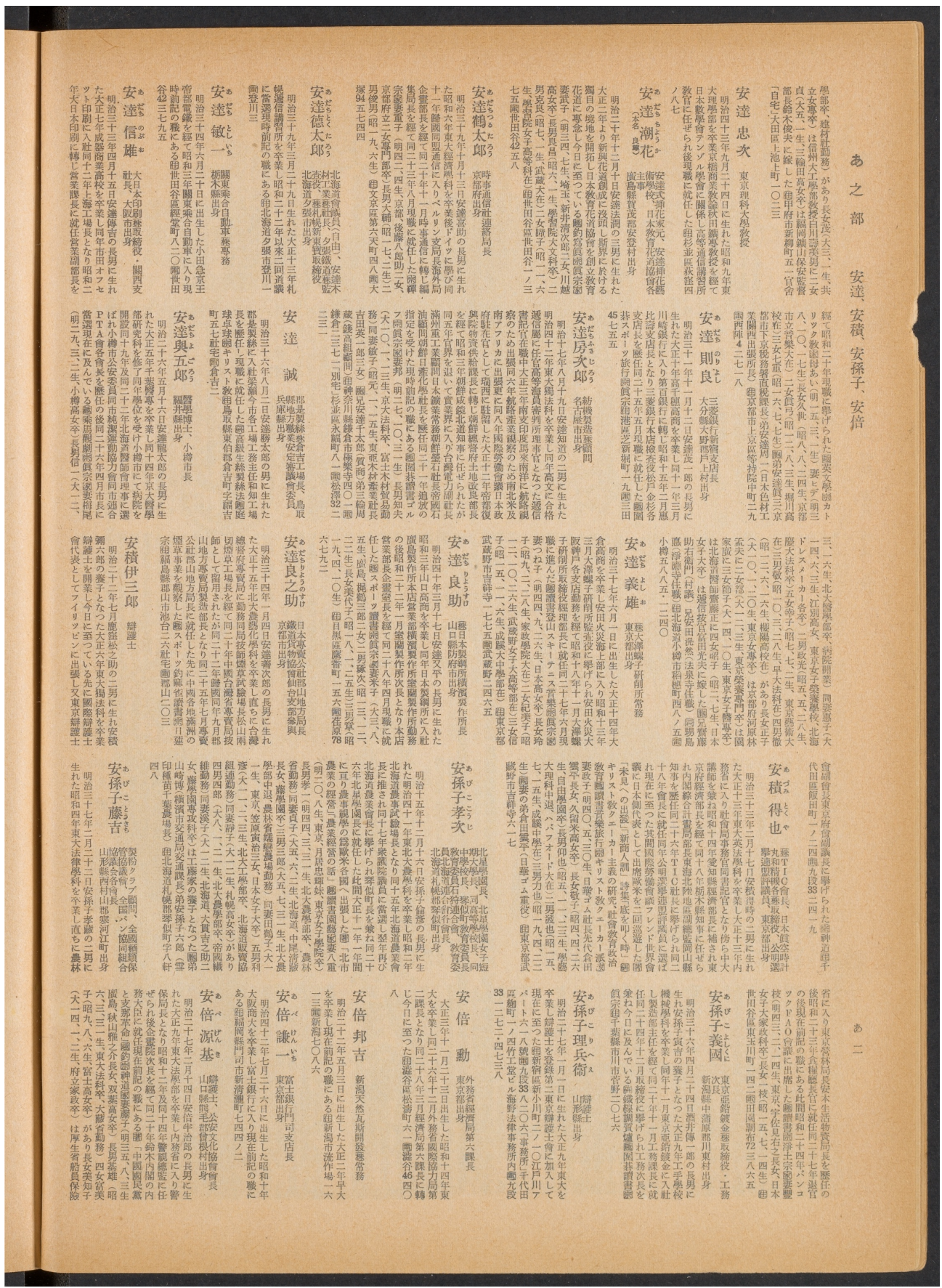
Our Team - The Automated History Archive (AHA) Group

- PI:
 - Prof. Melissa Dell, Department of Economics, Harvard University
- Current Team Members:
 - Ed Jee, Predoctoral Fellows Economics, IQSS
 - Yukako Kitamura, Predoctoral Fellows, NBER
 - Krishna Prasad, Predoctoral Fellows Economics, IQSS
 - Zejiang Shen, Predoctoral Fellows, IQSS
 - Kaixuan Zhang, Predoctoral Fellows, IQSS
- Collaborators:
 - Sahar Parsa, Department of Economic, New York University

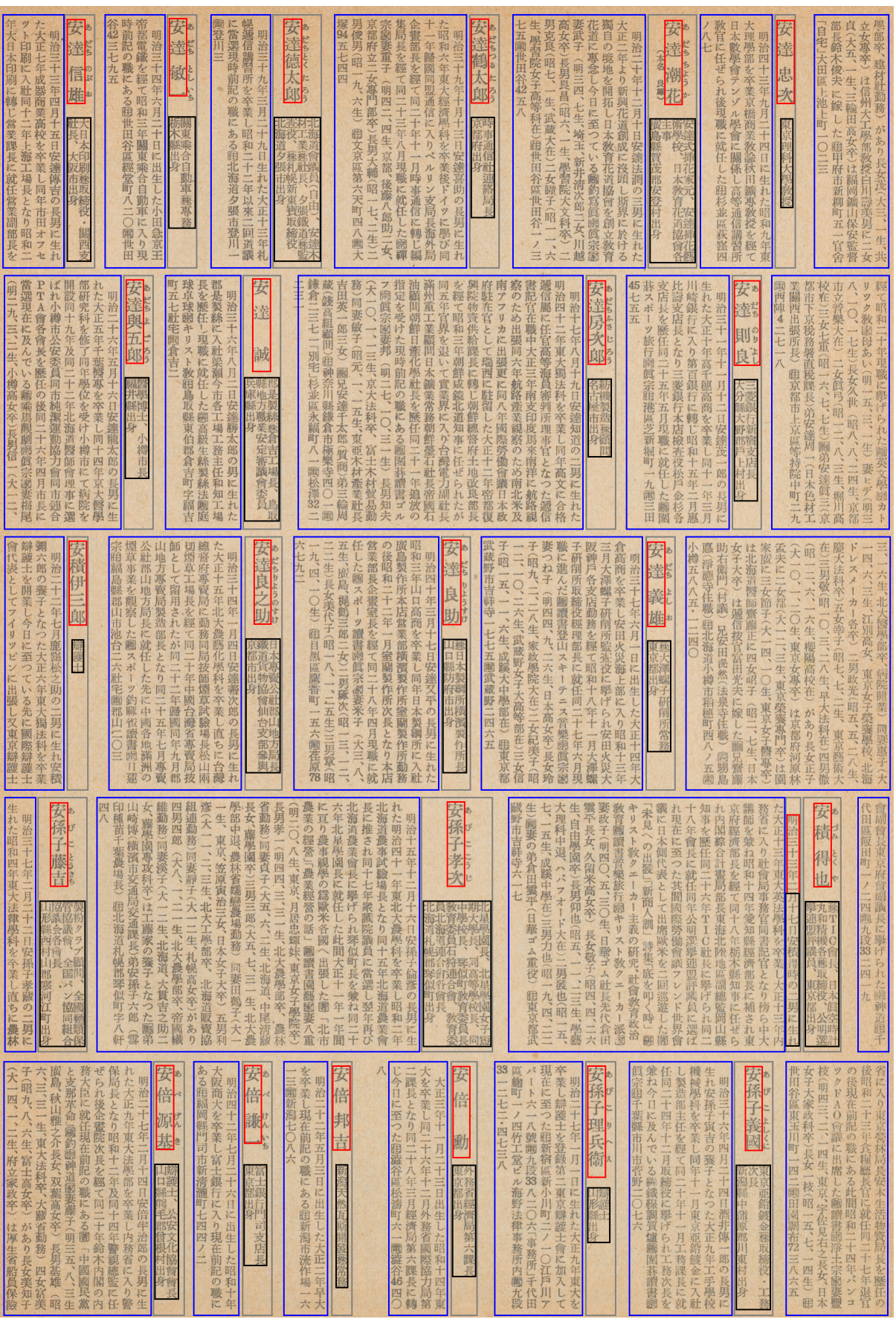
Document Layout Analysis

Objective

- We want to design an algorithm that takes the raw scan (a) as an input.
- It can detect the text region and extract text block region and types accordingly (b)

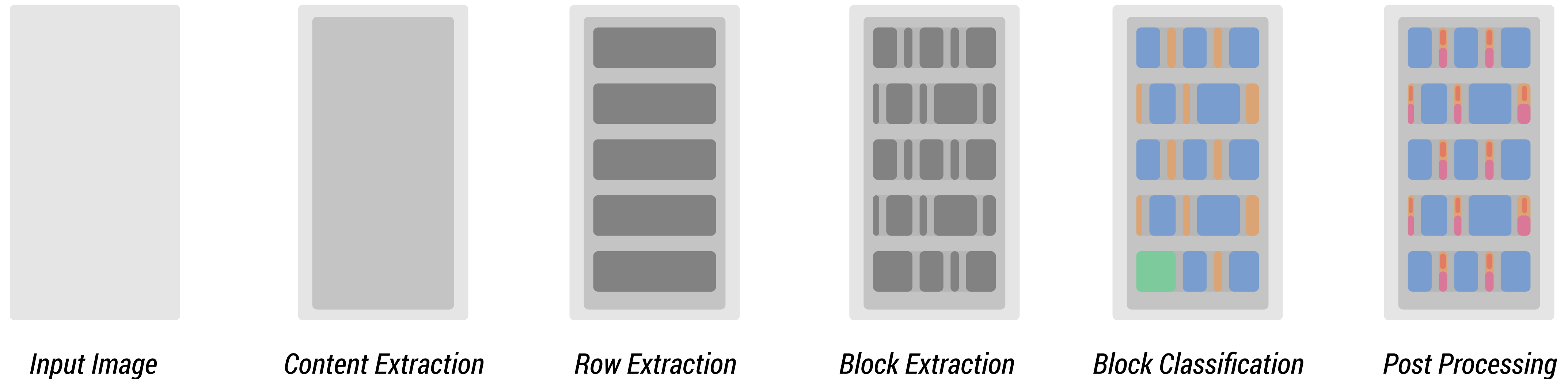


(a) A Input Raw Scan



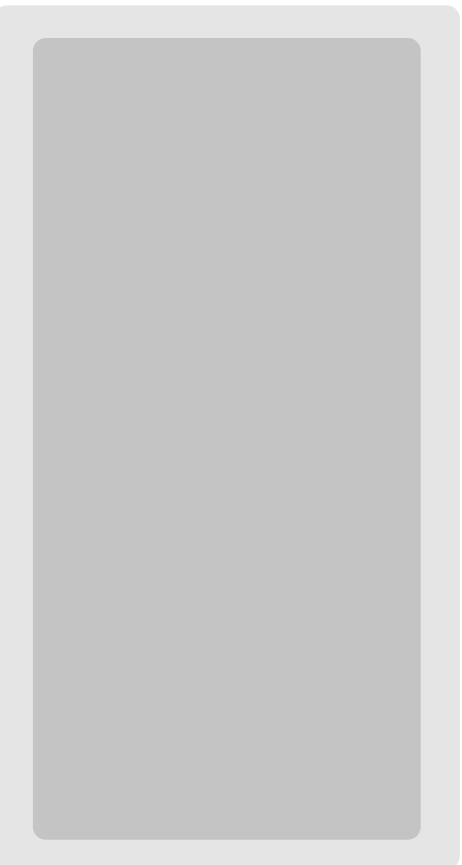
(b) Scan with Extracted Block Boxes and Types

Layout Analysis Method



- The extraction works in a procedural fashion
- It combines rule-based traditional computer vision algorithm and deep learning based methods

Layout Analysis Method



Content Extraction



Row Extraction

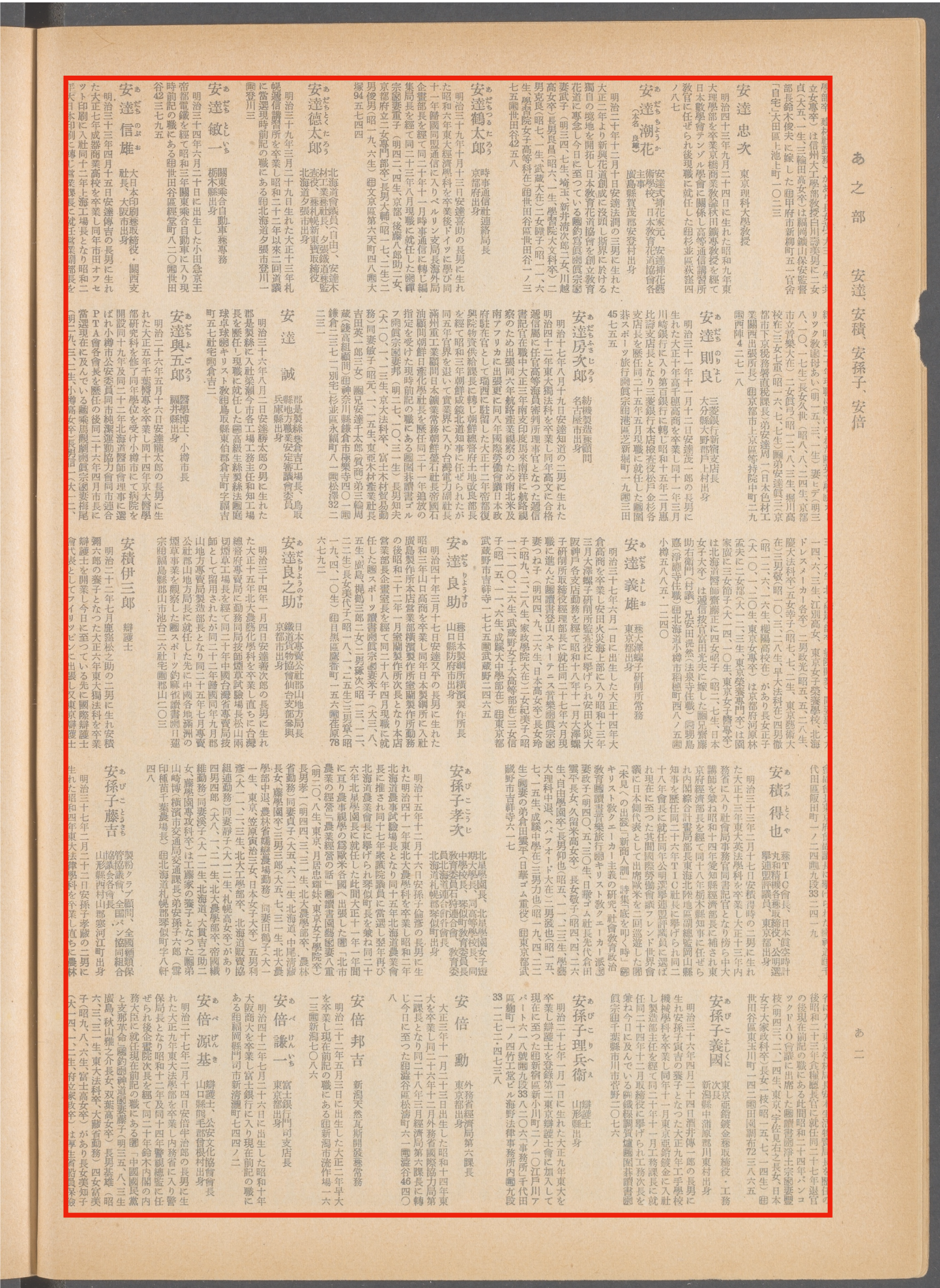


Block Extraction

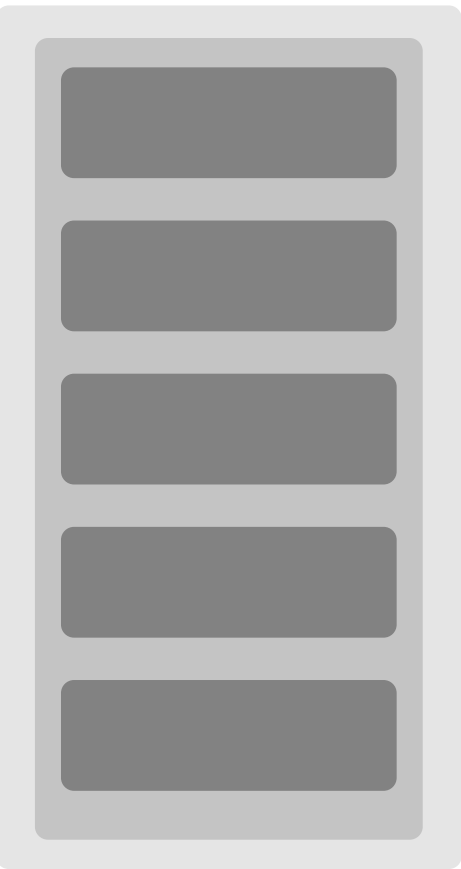


Block Classification

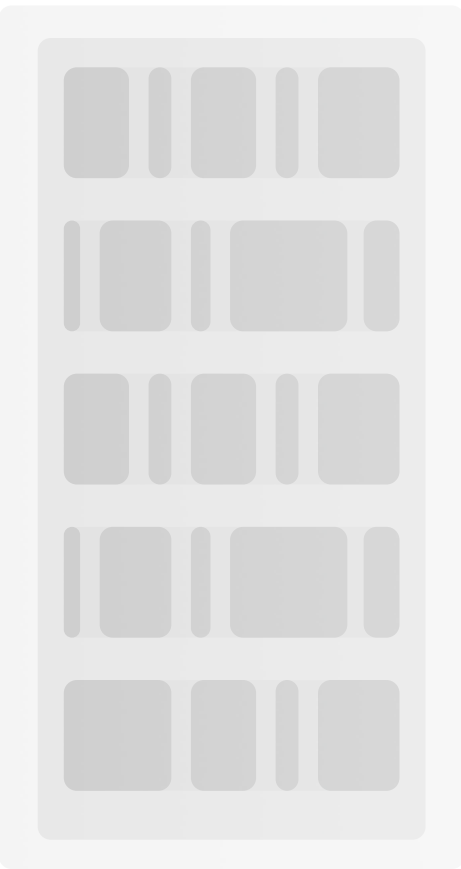
- It extracts the content box (or page frame) in the input scan, crops out irrelevant areas, and estimates an affine transformation to correct the skewness
- It needs to be robust to the rotation of the scans



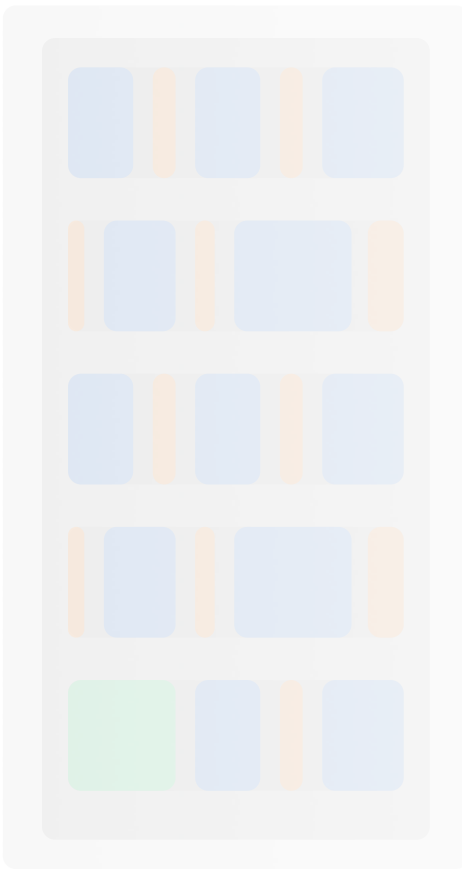
Layout Analysis Method



Row Extraction



Block Extraction

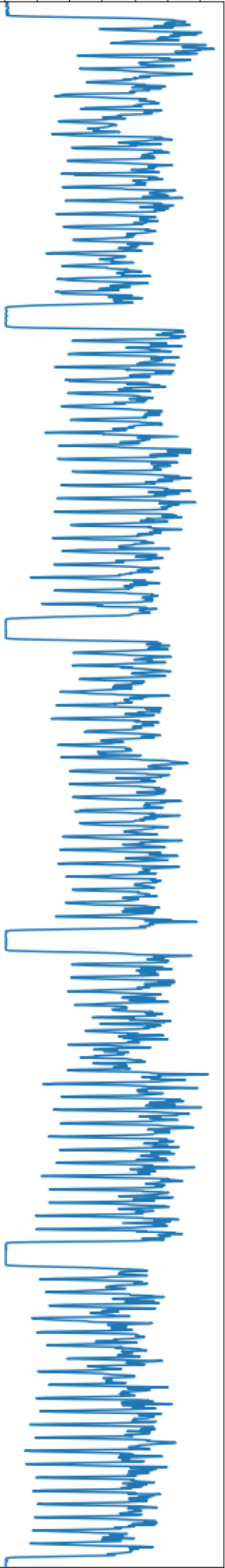
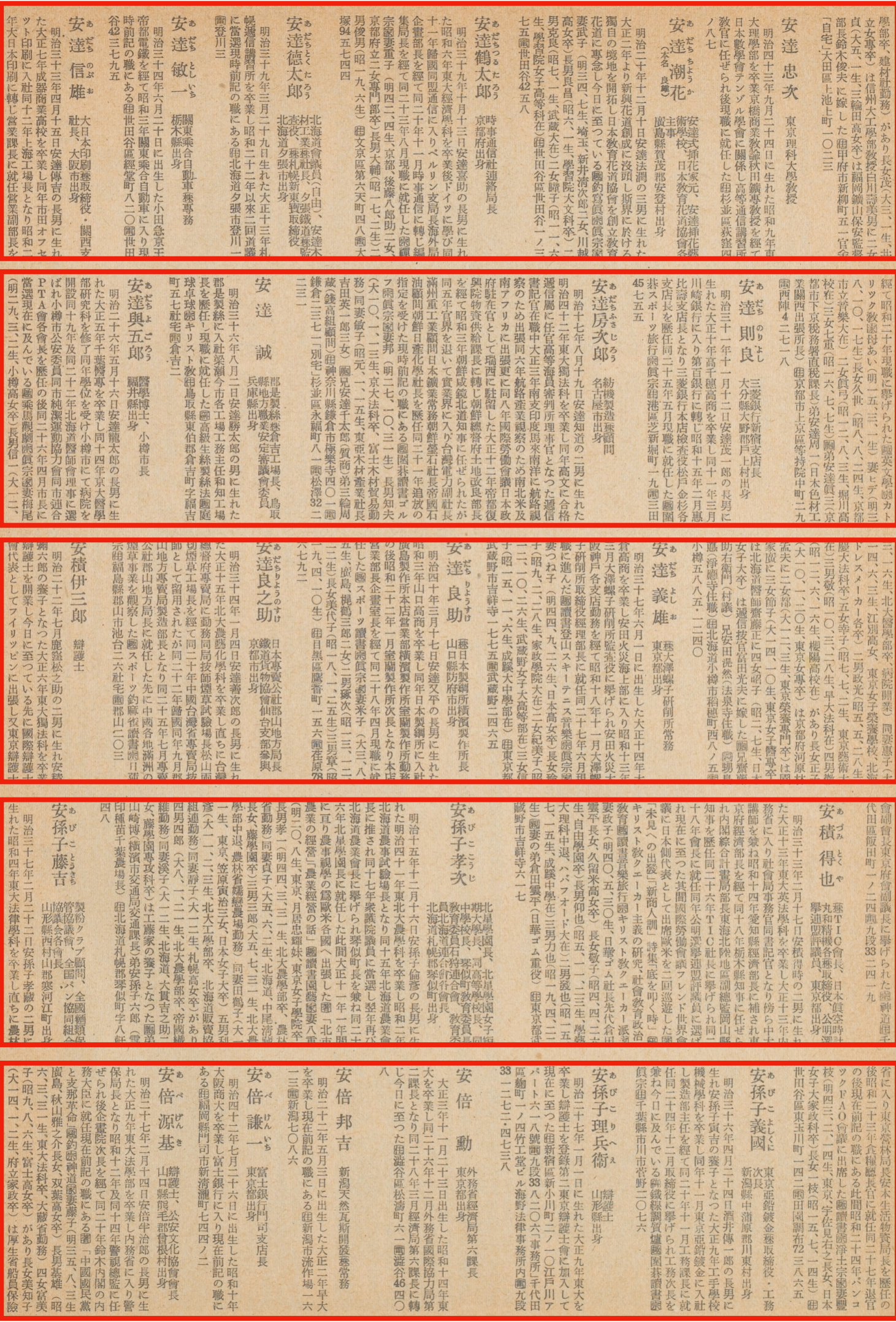


Block Classification

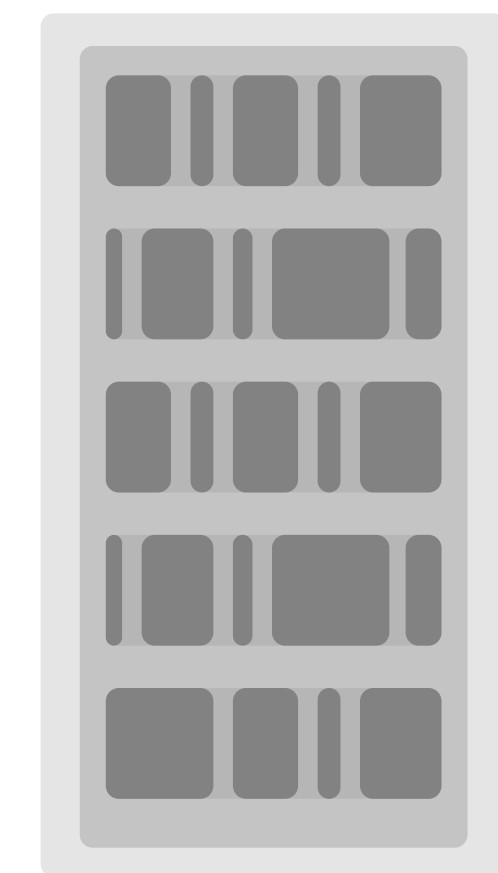


Post Processing

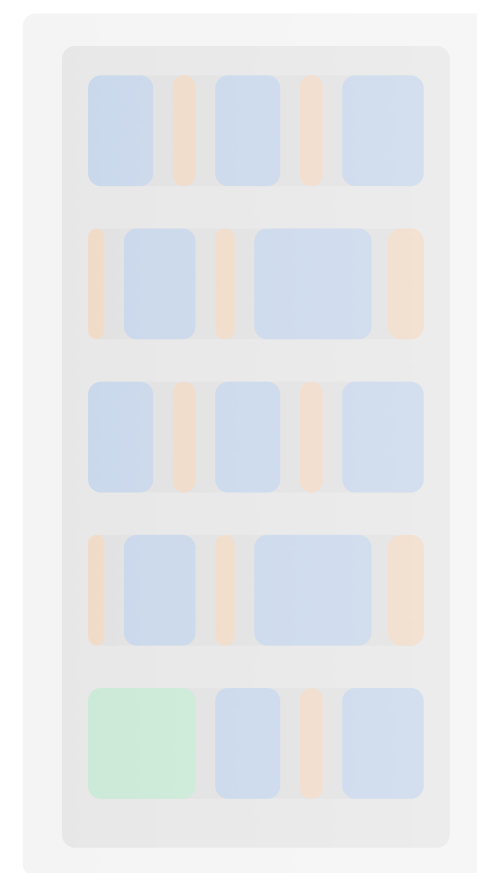
- Based on the clean page image, it segments the rows utilizing the large horizontal gaps between them



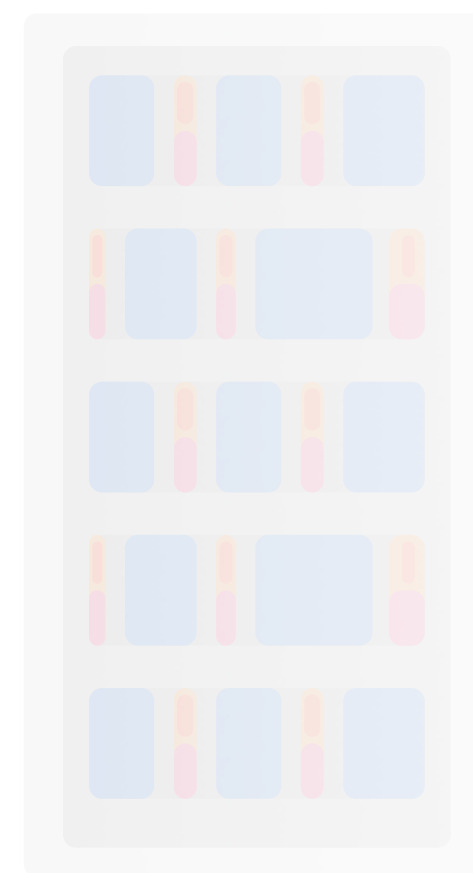
Layout Analysis Method



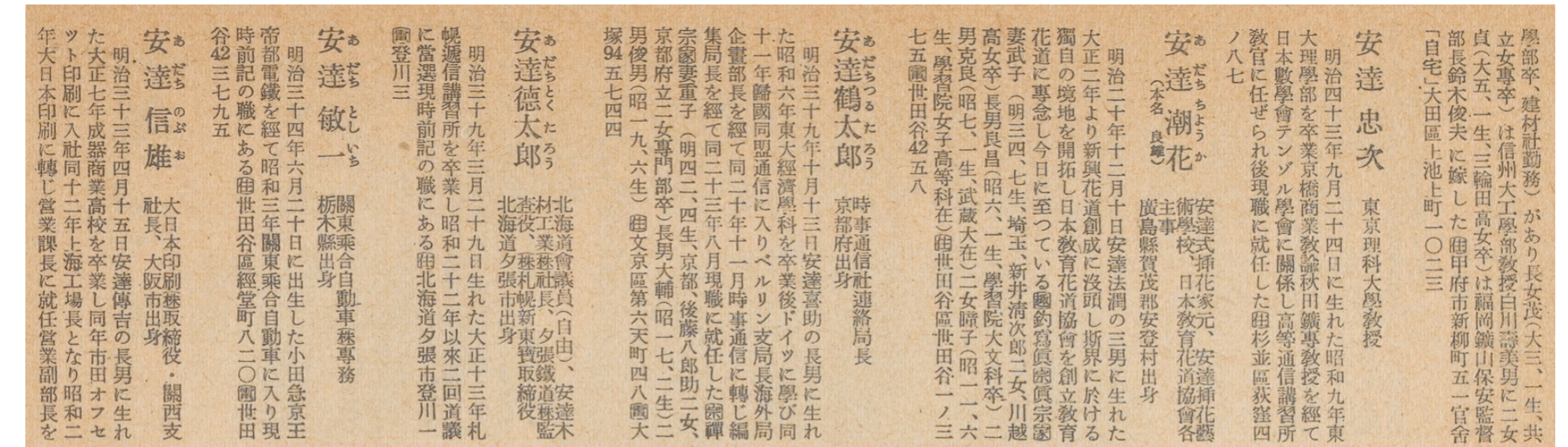
Block Extraction



Block Classification

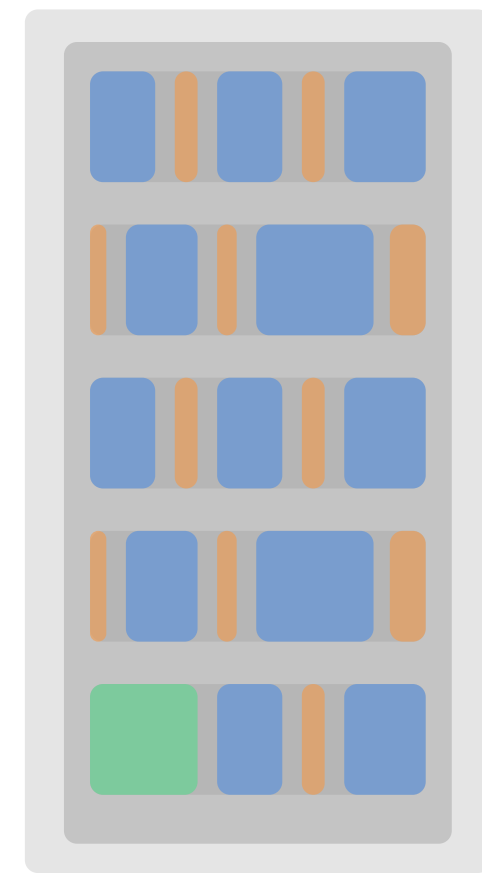
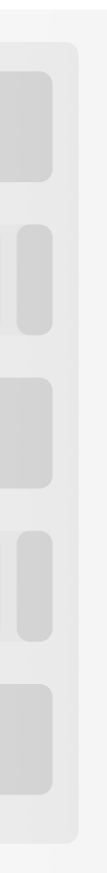


Post Processing

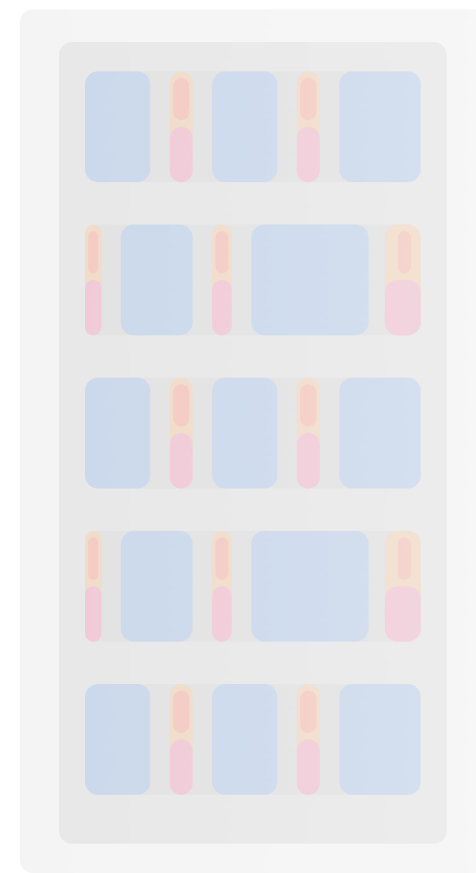


- For each extracted row region, the algorithm uses the vertical gaps to separate the different blocks
 - The Run Length Smoothing Algorithm (RLSA) and connected component analysis algorithm is used
 - Some parameters are hard-coded: they constitute human-designed rules

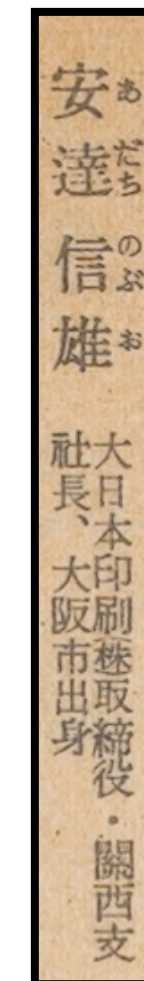
Layout Analysis Method



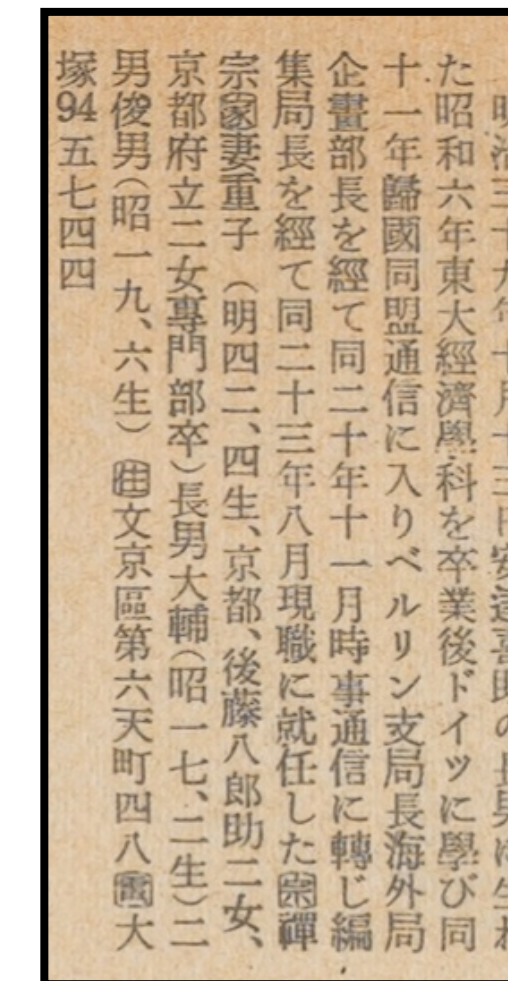
Block Classification



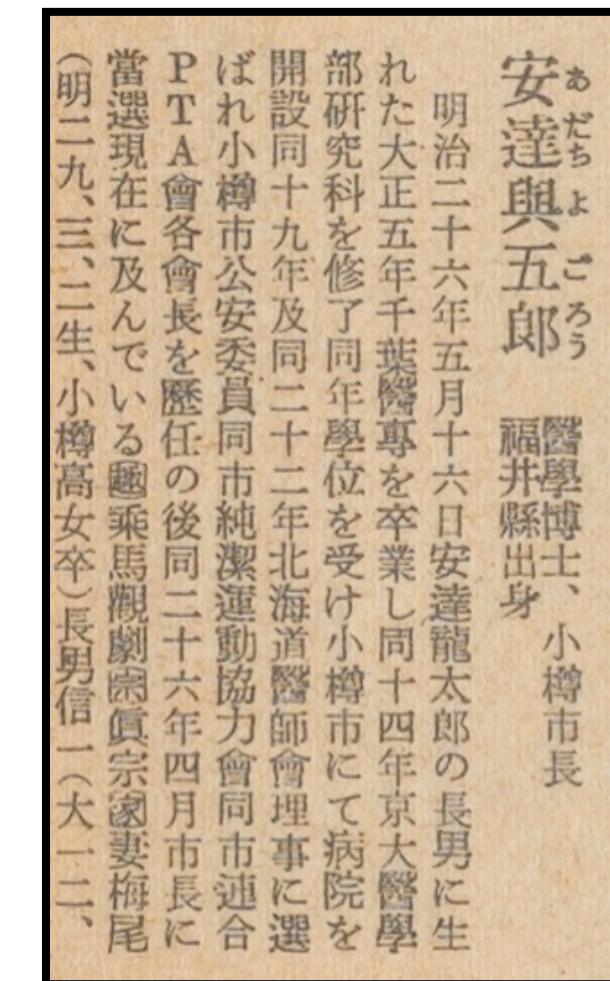
Post Processing



Title Region



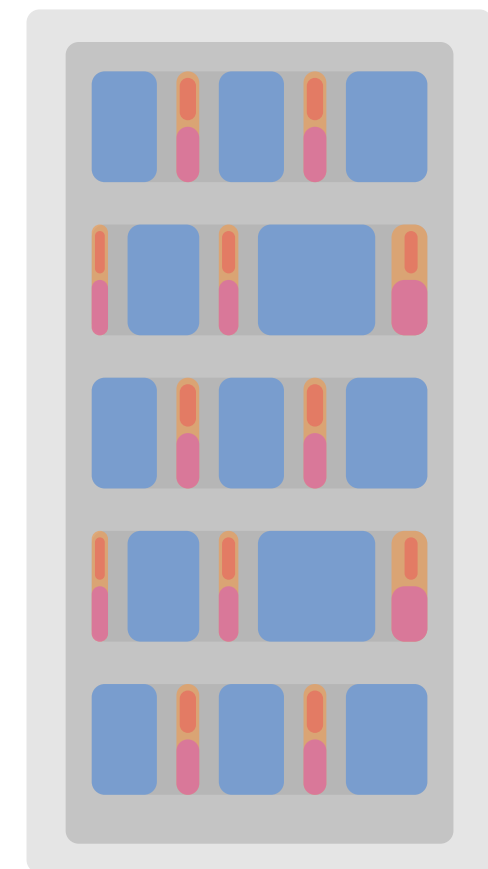
Bio Region



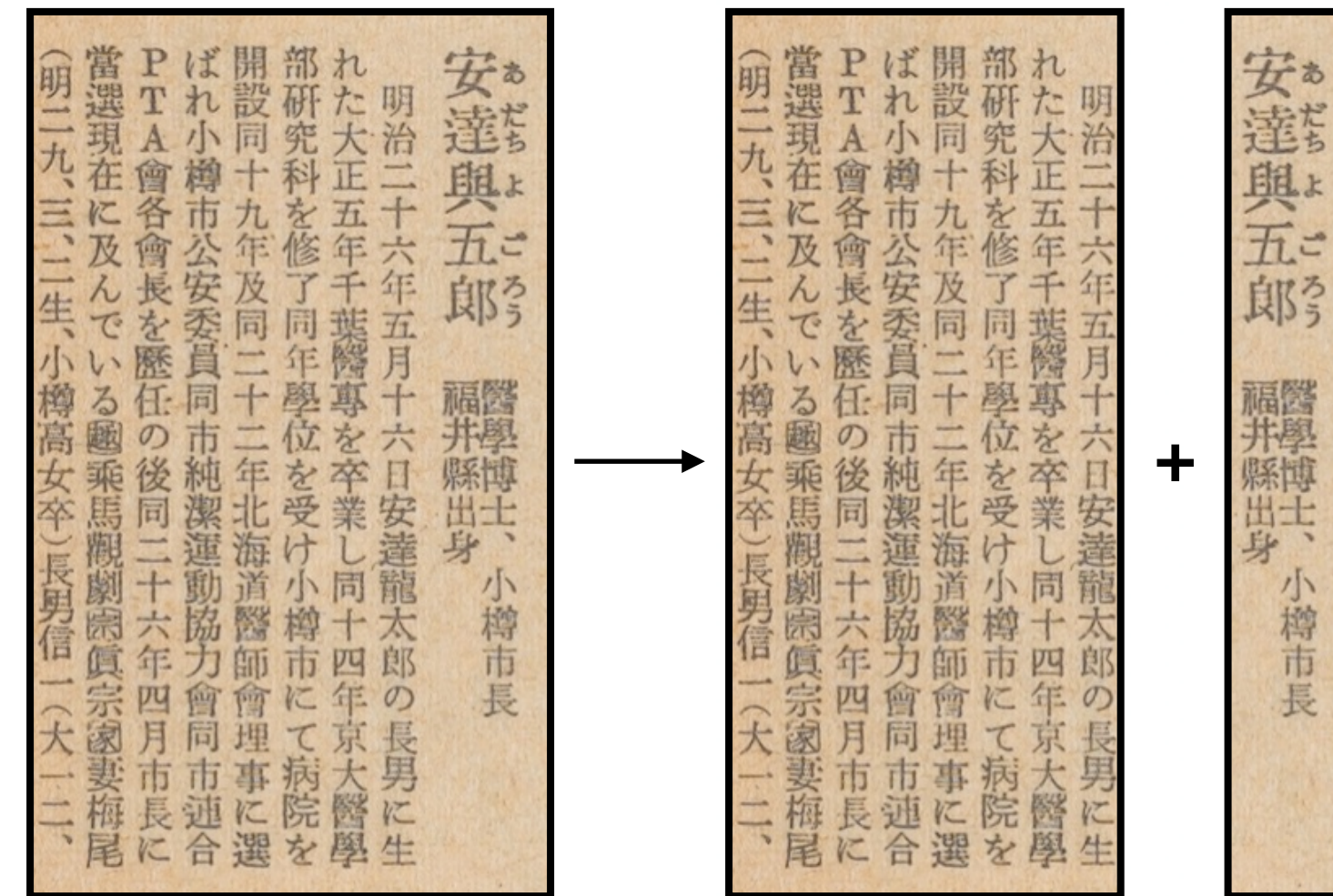
Problematic Segmentation

- A deep convolutional network (CNN) is used for classifying different types of blocks.
 - In addition to the biography and title region blocks, there's an additional 'Problematic' class for the wrongly segmented blocks.
 - Sometimes, the algorithm fails to split the biography region and the title region, and the classifier helps identify this problem.

Layout Analysis Method



Post Processing



Problematic Segmentation Postprocessing

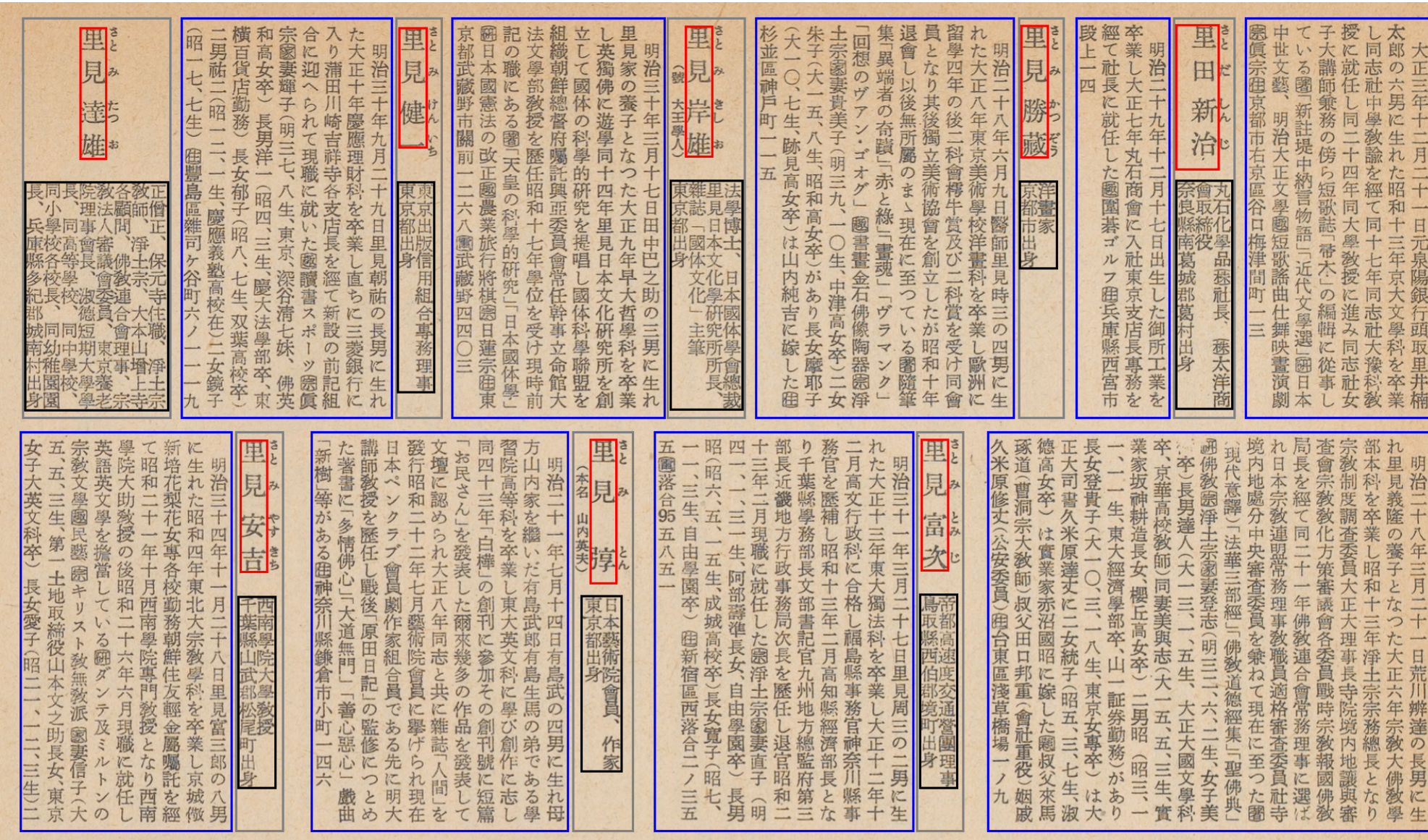
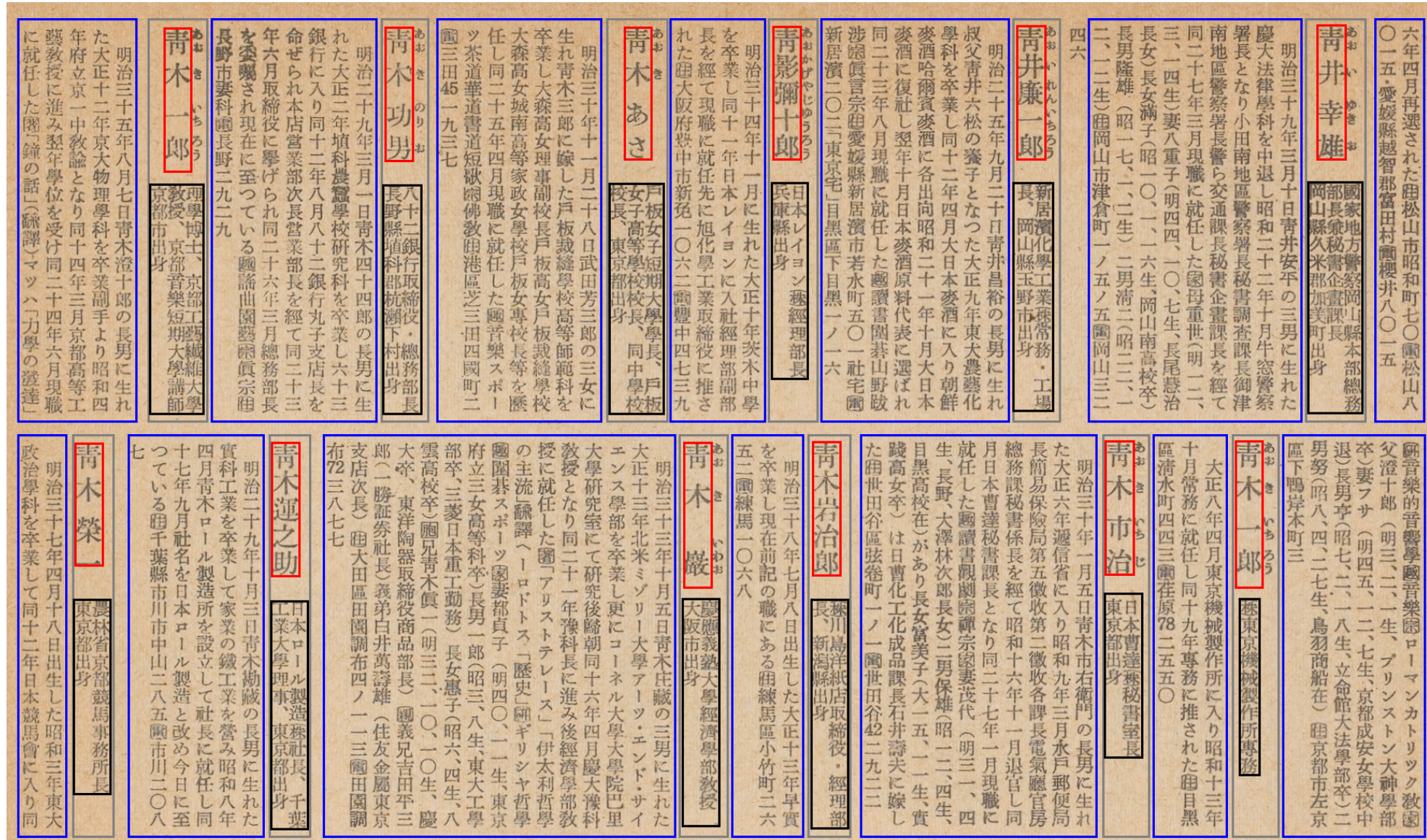


Title Region Postprocessing

- Finally, depending on the classification results, the blocks are further processed.
 - The problematic ones are sent for re-segmentation
 - The title region is split into name and position blocks

Our Results

- The results of our layout understanding algorithm
 - The block boxes are marked in different colors:
 - Red for name block, black for title block, blue for description block, and grey for name region.

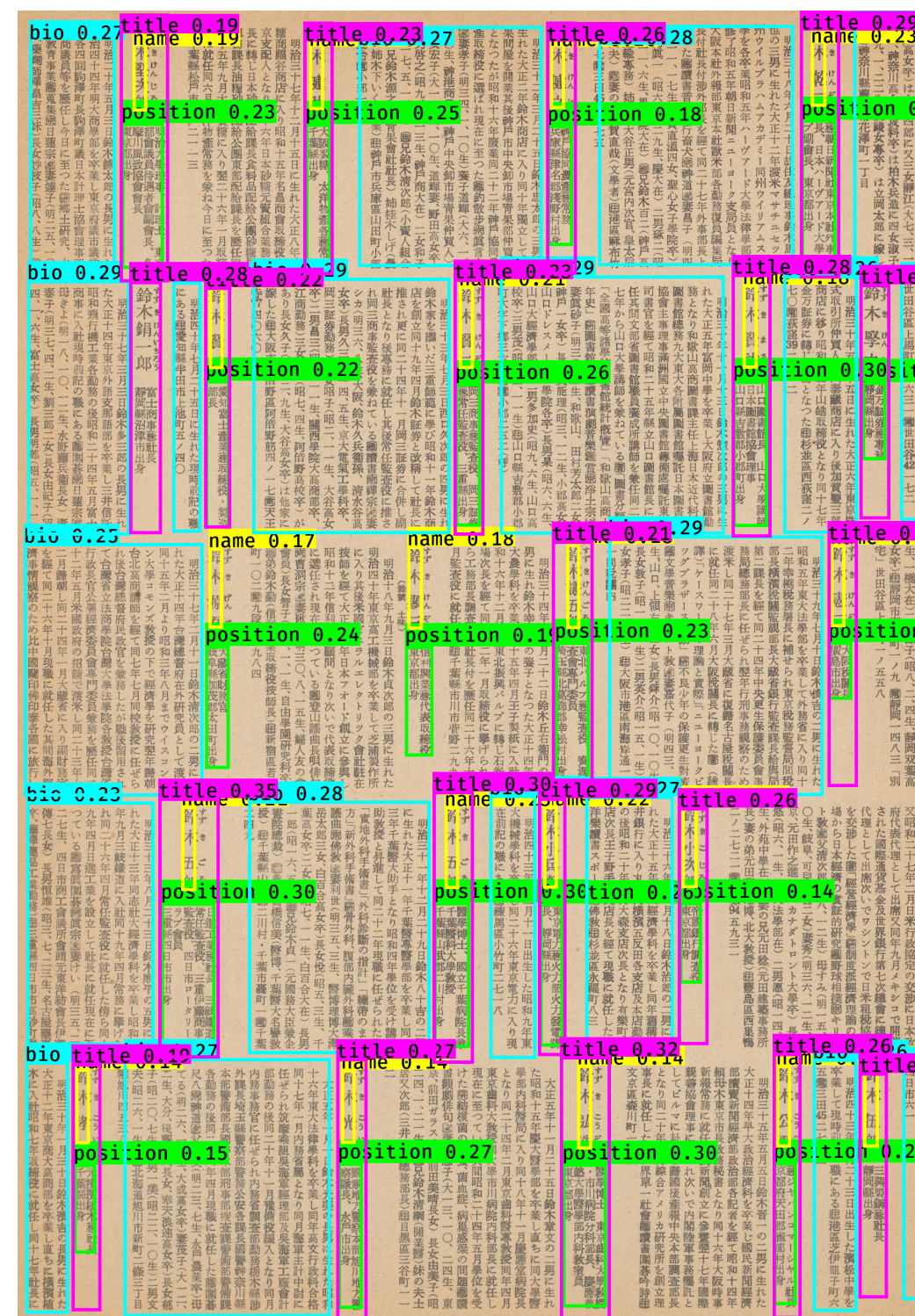


Why not Object Detection?

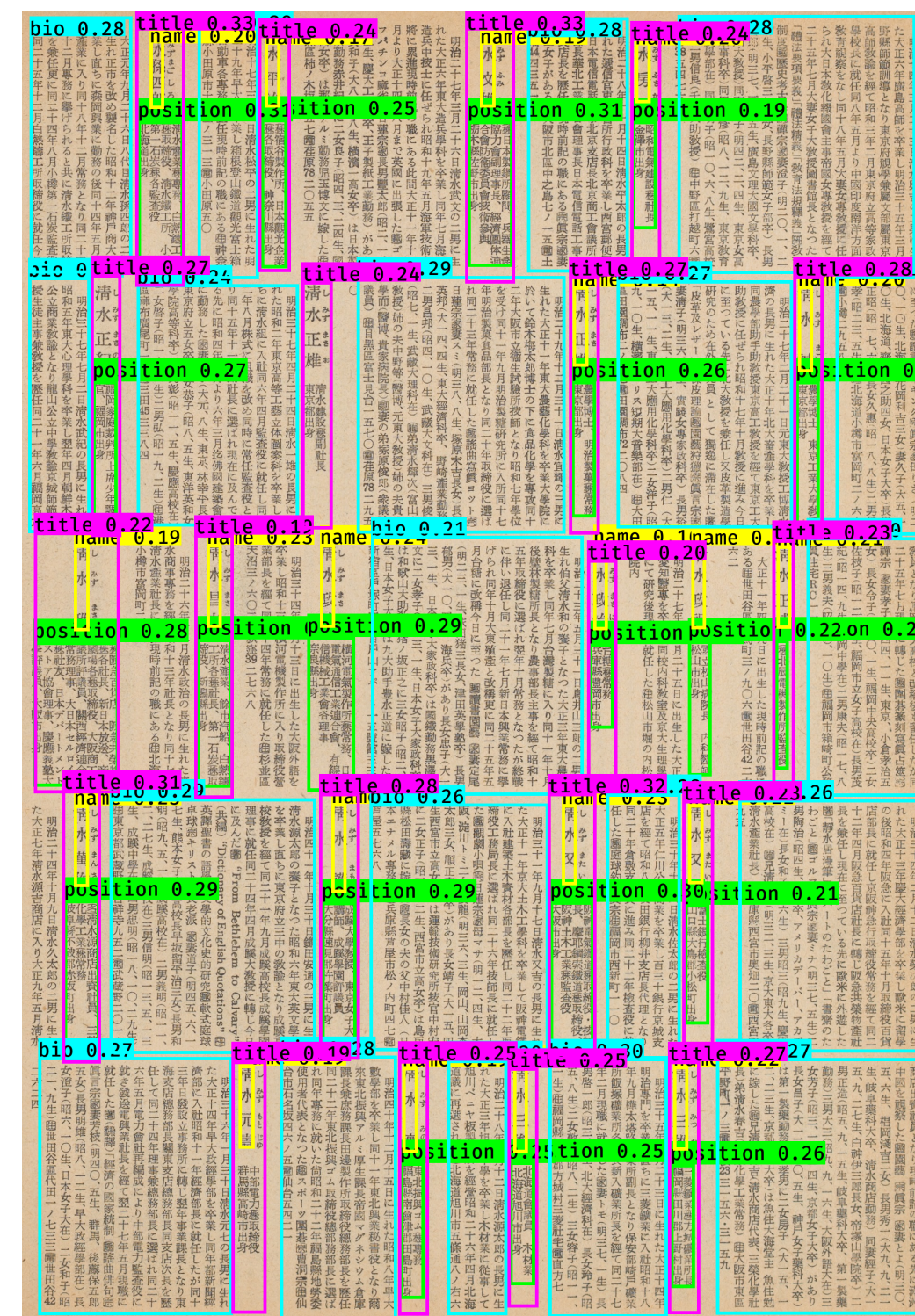
- We didn't have the ground-truth in the beginning.
- No similar models to turn to - our datasets are very unique.
 - Hard for transfer learning using ImageNet weights.

Why not Object Detection?

- After obtaining the text blocks using our algorithm, we did the experiments using YOLOv3.



Yolo Example 1

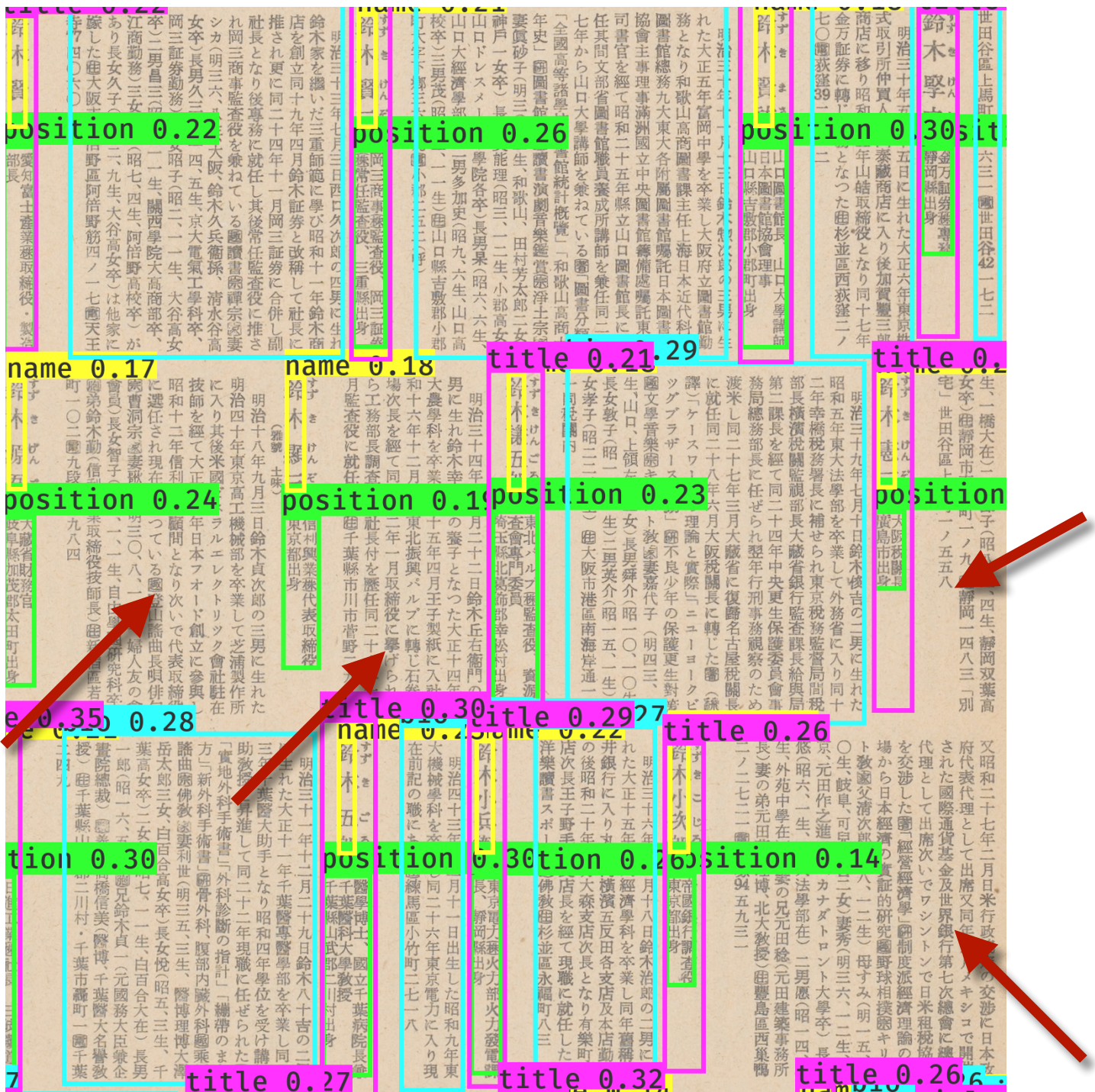


Yolo Example 2

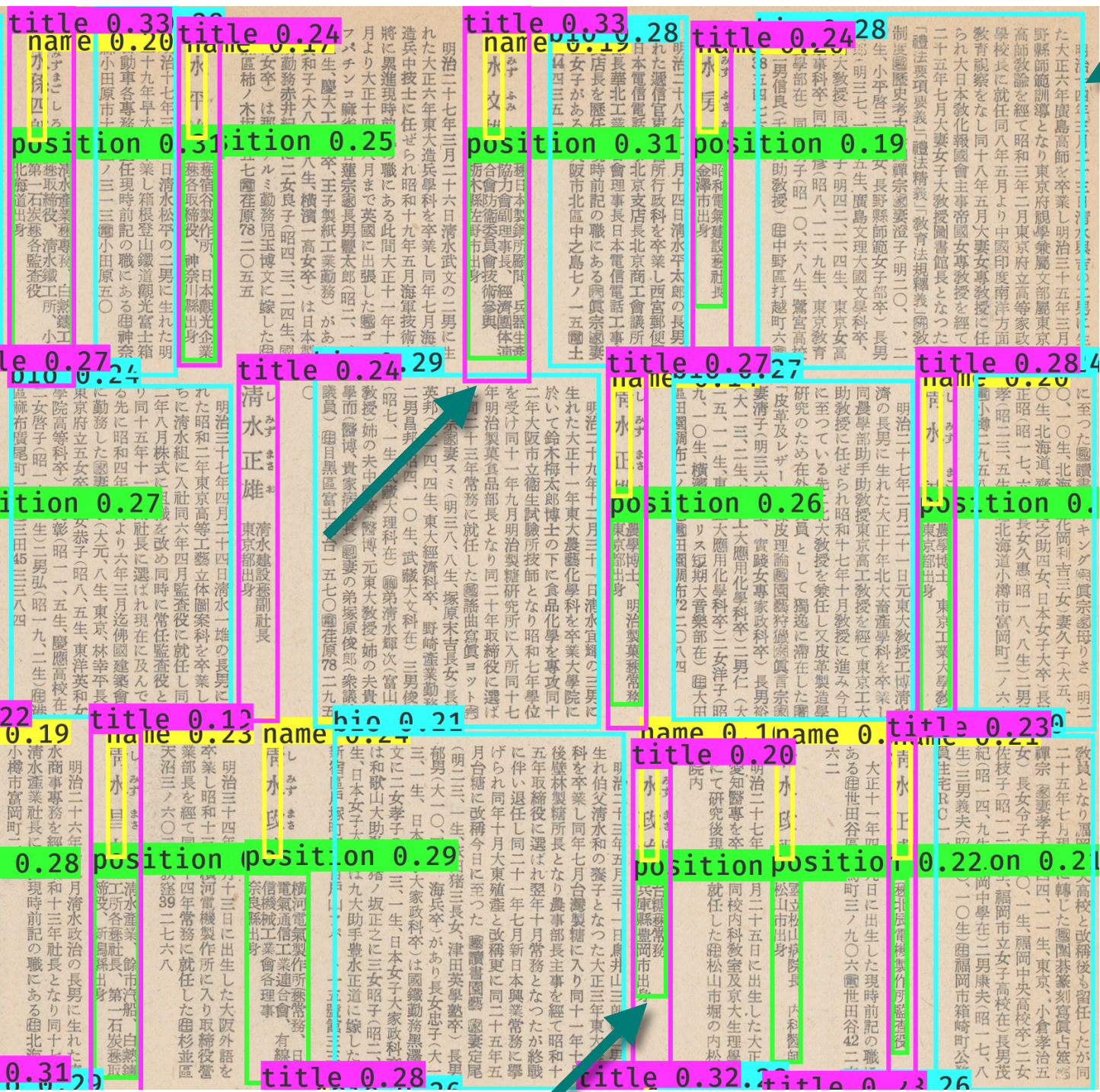
- The yolo input size has a similar aspect ratio as the original scans but is ~4x smaller. (864,576)
- The training batch size is 1, and there are ~900 training samples. It only trains the top layers for the first 50 epochs while the whole model for the last 50 epochs.
- The implementation is based on <https://github.com/qqwweee/keras-yolo3>

Why not Object Detection?

- The the loss converges in the end, the detection is not accurate: missed blocks and inaccurate boundaries



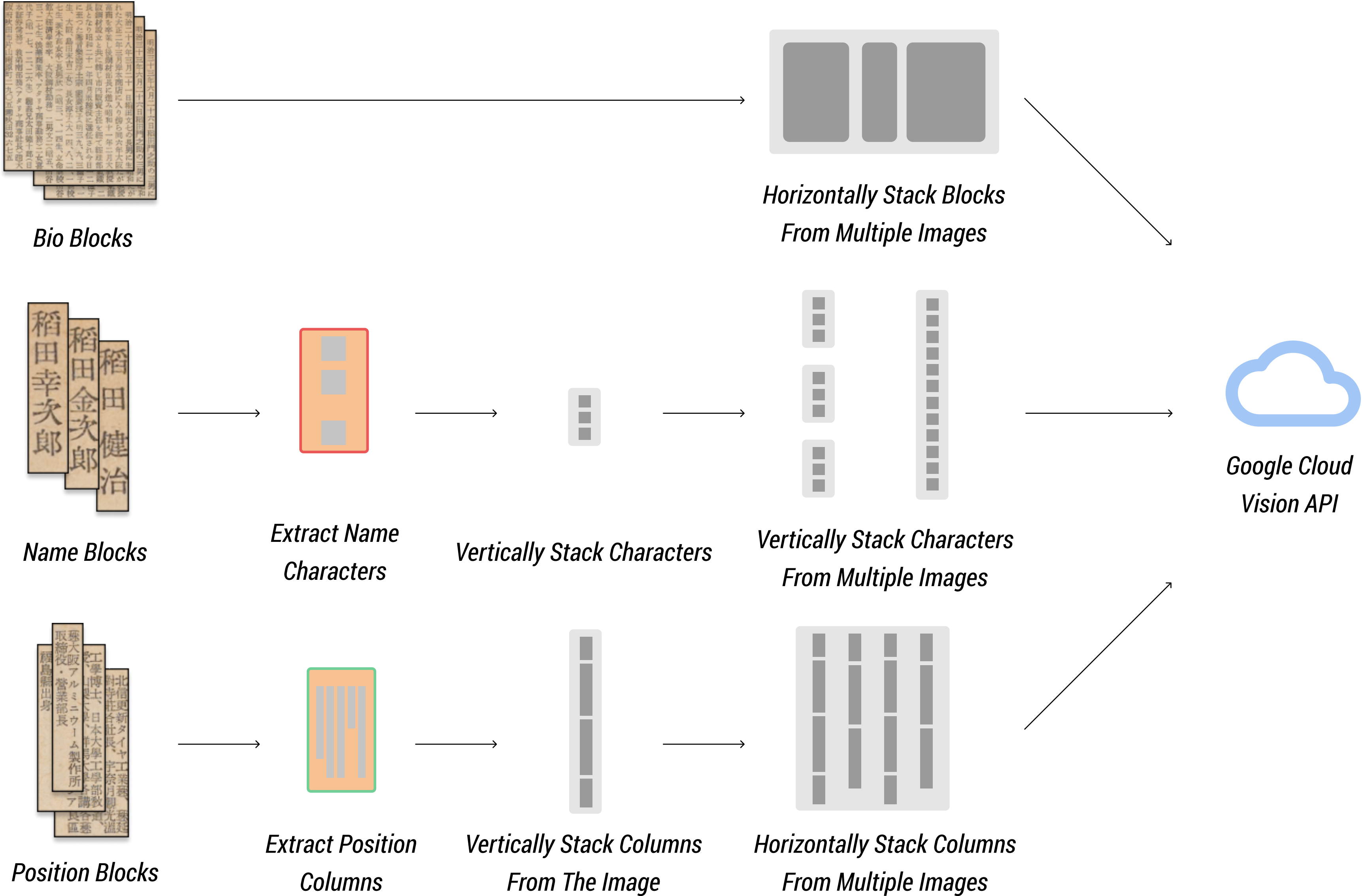
Yolo Example 1

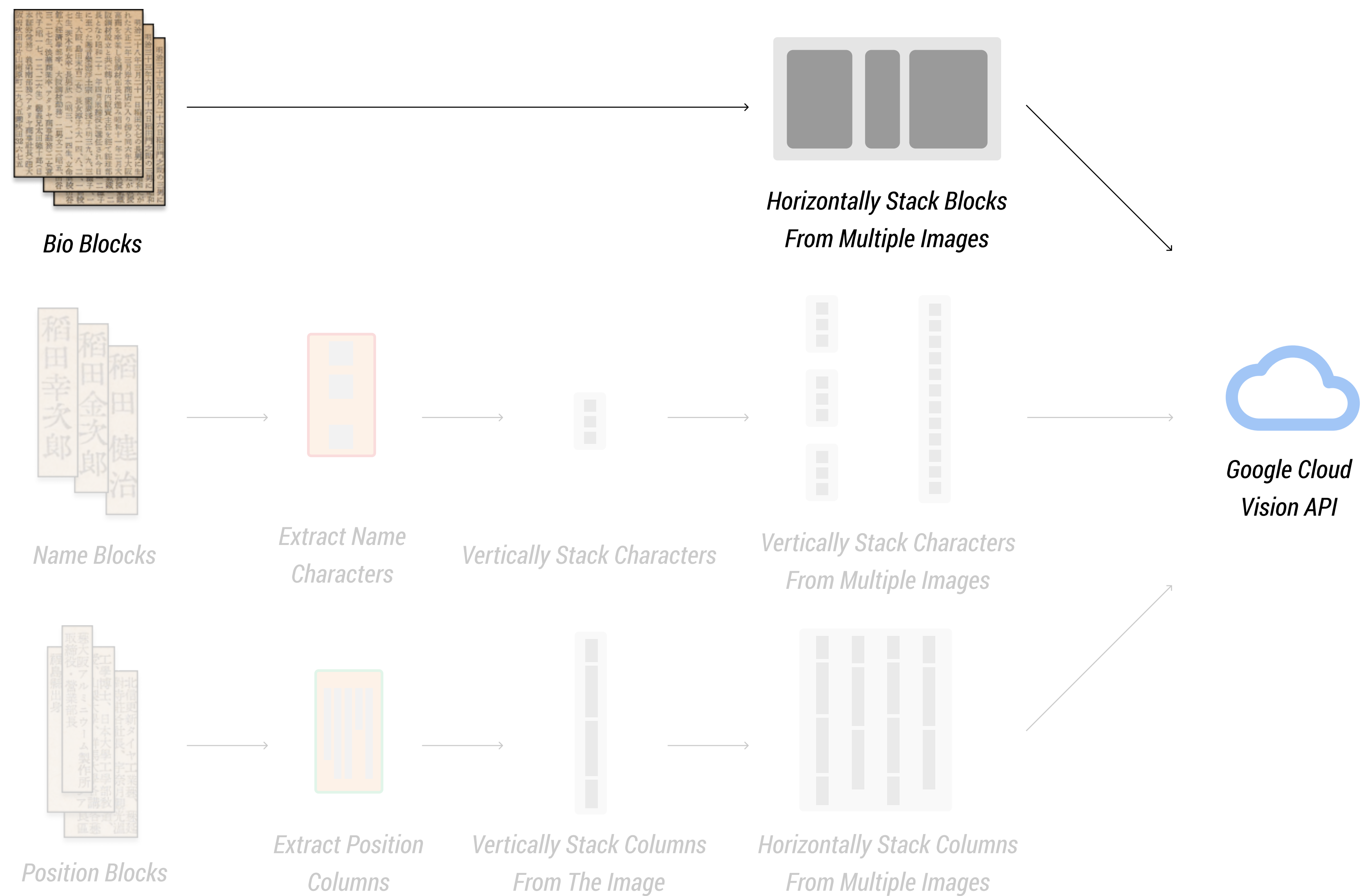


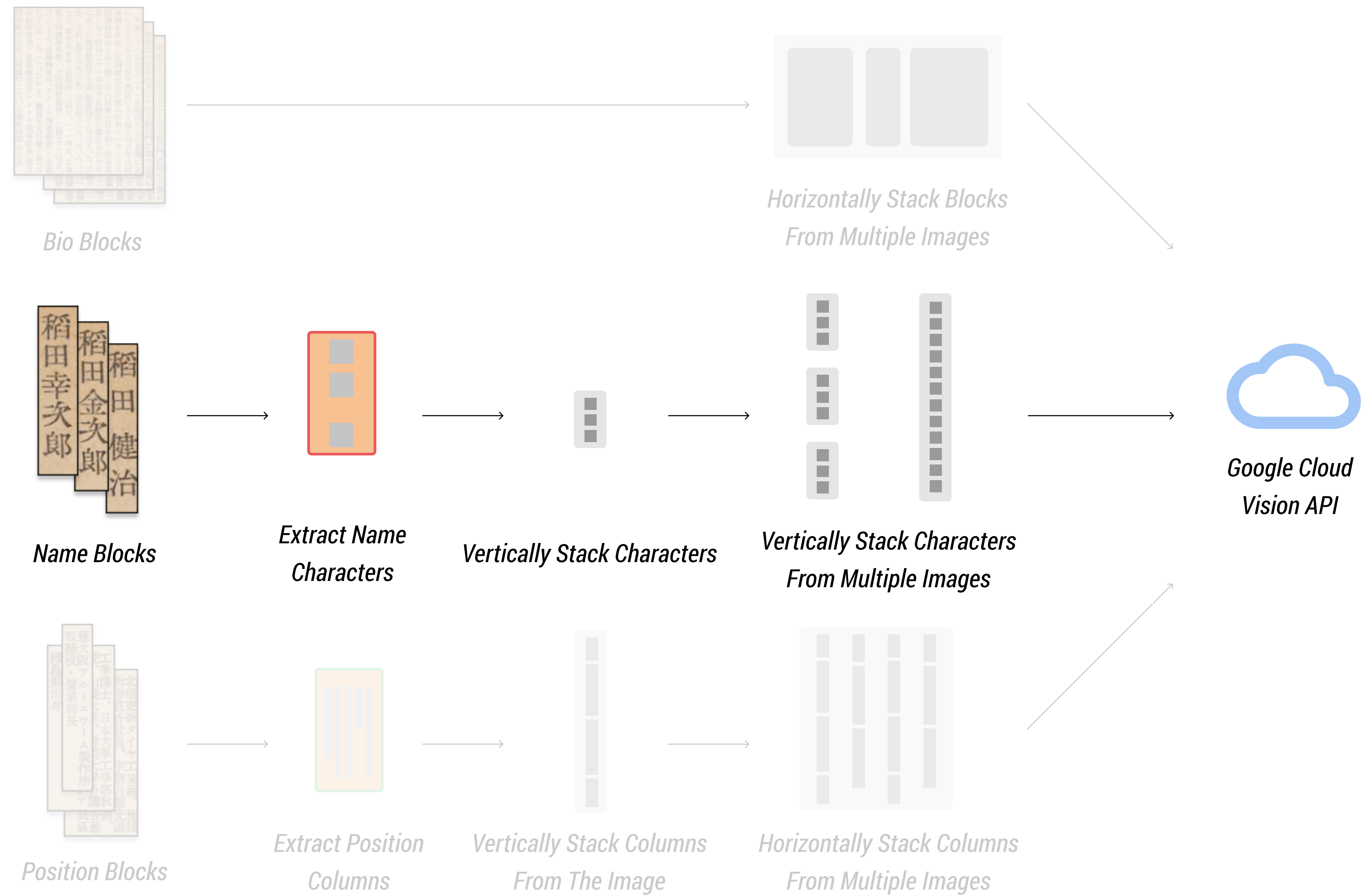
Yolo Example 2

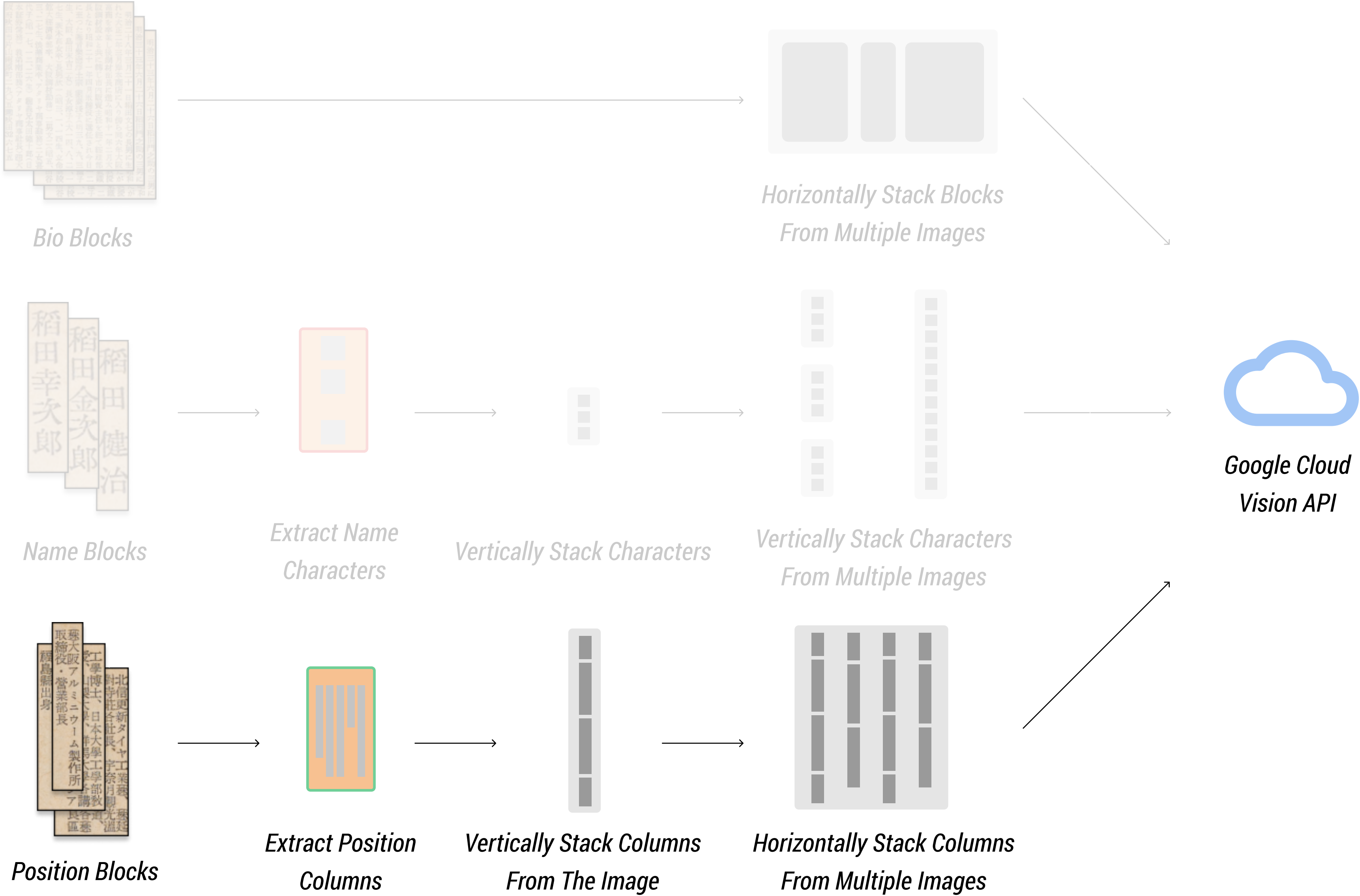
Postprocessing for OCR

- One important observation: OCR softwares work well on simple layouts.
 - That's what they are designed for.
- We use the layout information to post process the raws scan into images of simple layouts.
 - Crop -> Recombine -> OCR -> Decompose





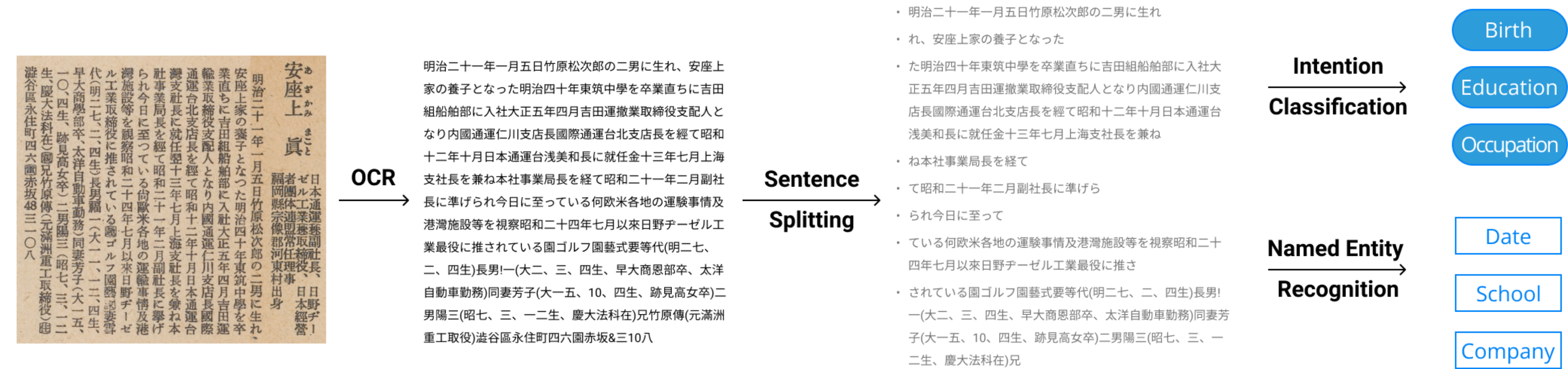




Structura Information Extraction

Work in Progress

Structura Information Extraction



- The OCR will generate a long text string for each biography region
 - The long text contains multiple information pieces -> Split the string into sentence segments
 - Each segment has different types of information -> Intention classification
 - Named Entity Recognition for each sentence segments

Conclusion

- We propose several approaches to convert the unstructured and noisy historical document scans into structural tables for further analysis.
 - A layout analysis algorithm that facilitates the OCR
 - An NLP method that extracts specific text information
- Thoughts from a data science perspective
 - Real-world data are challenging
 - Creative solutions are important when you don't have the labels

Relevant Materials

- Kaggle Bengali.AI Handwritten Grapheme Classification Competition
<https://www.kaggle.com/c/bengaliai-cv19>
- Kaggle Kuzushiji Recognition Competition
<https://www.kaggle.com/c/kuzushiji-recognition>
- International Conference on Document Analysis and Recognition
<http://icdar2019.org>

Thank you!