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

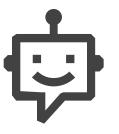
Contents

- Background & Overview 1. 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Why historical documents matter?

Why historical documents matter?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

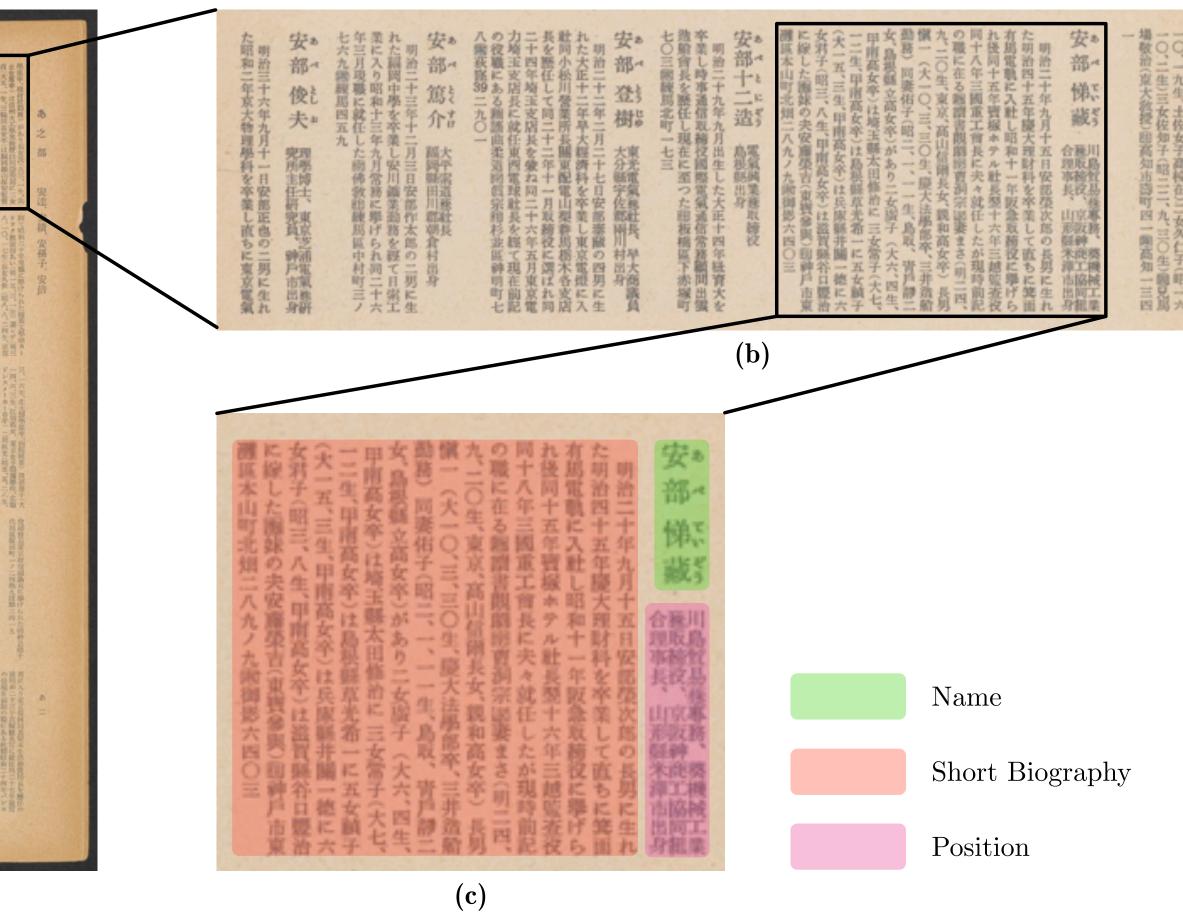
Noise in the text, the old printing technology, paper wearing, and the scan technology

We want to analyze

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

(a)





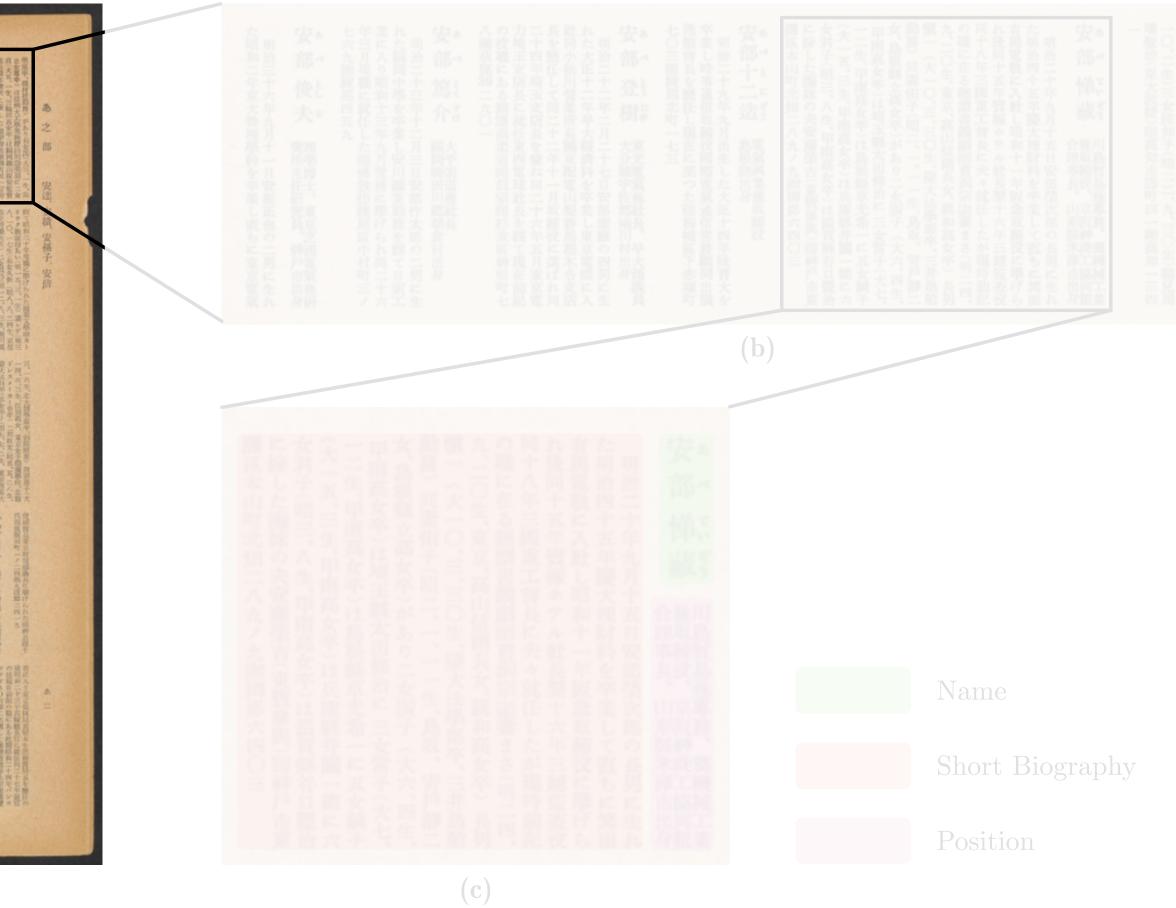


We want to analyze

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

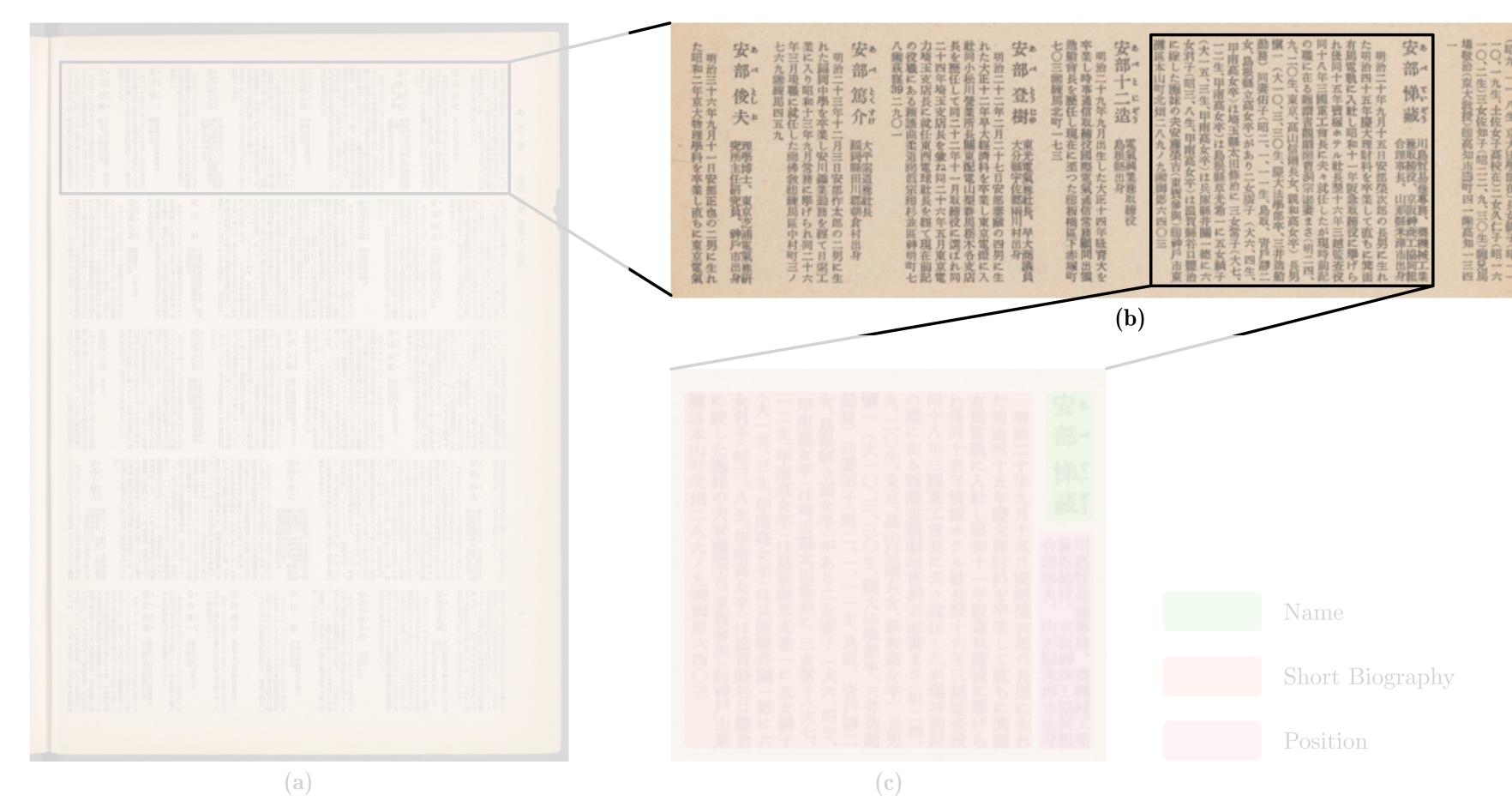
(a)



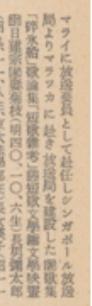


We want to analyze

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







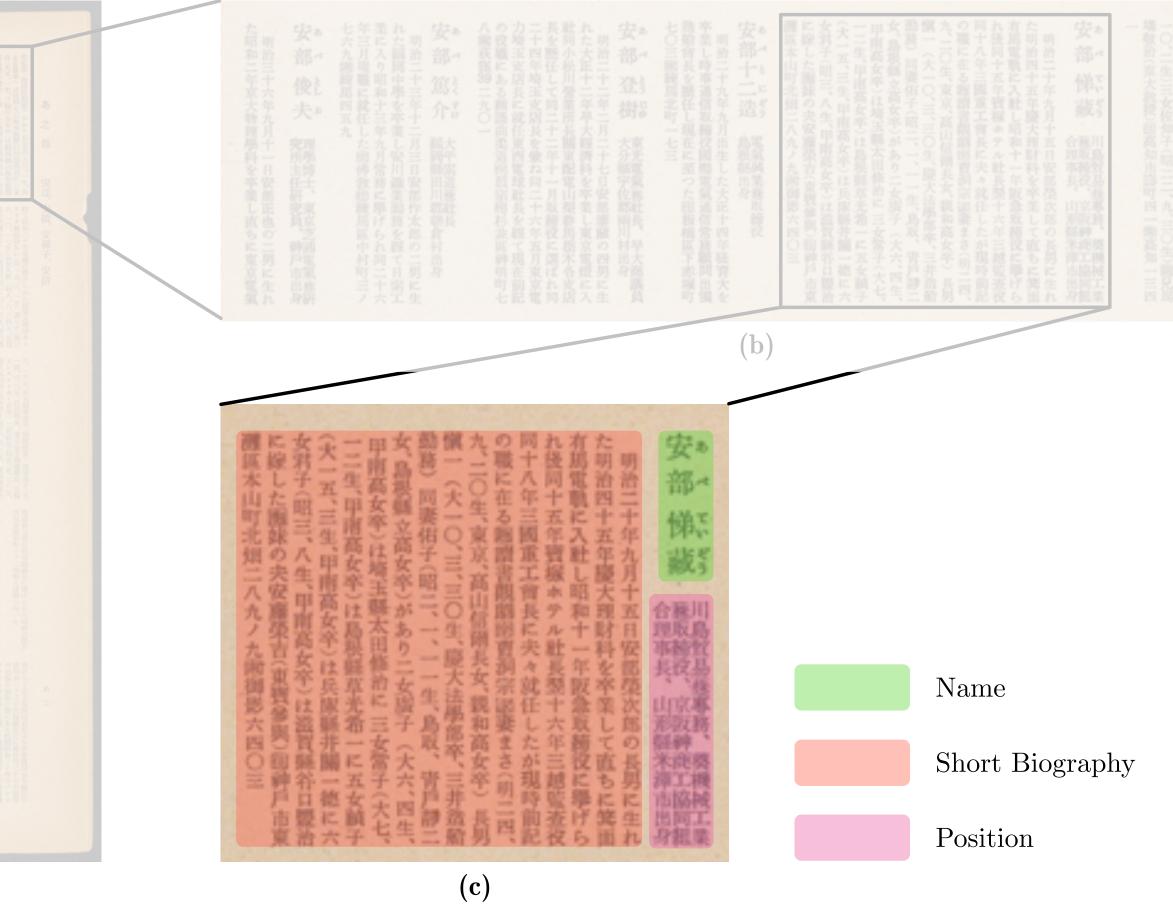
We want to analyze

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

And the second s	

(a)





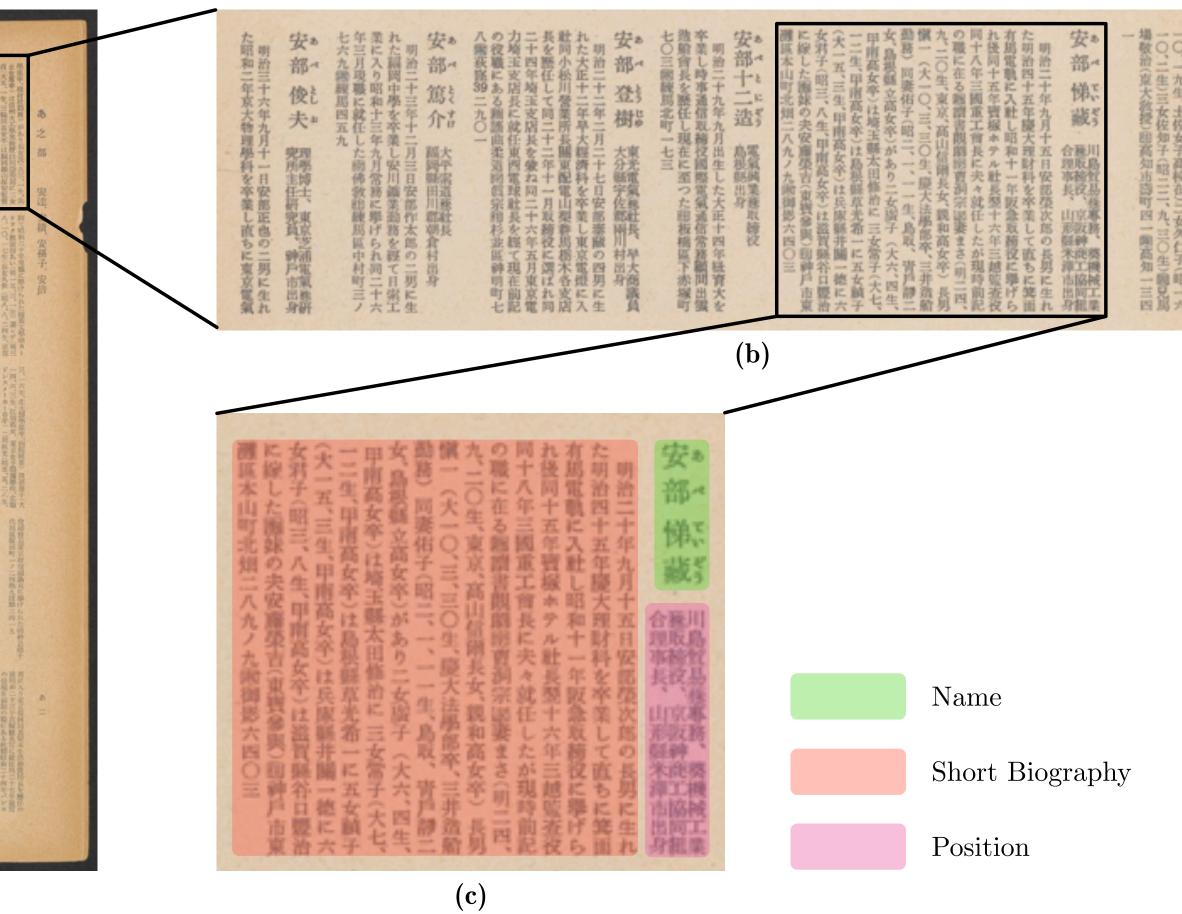


- We want to analyze

 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

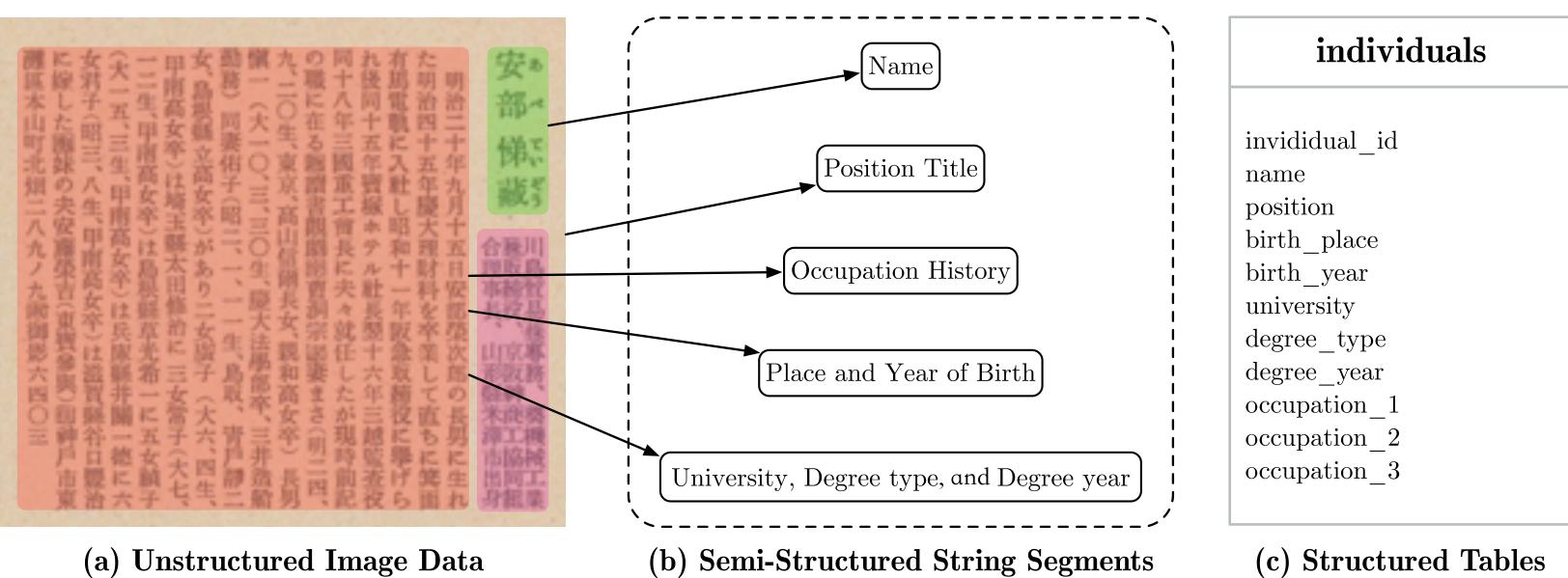
新大大明朝 安本 大明朝 建築	222		安全 義務		- St. 5	5.25.5 G	曹杰明 安	• 人發呈 点。	安美
ないた。 ないた。 ないた。 ないた。 のために、 したのでの たいた。 ないた。 ないた。 ないた。 たのでの たいた。 ないた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの たいた。 たのでの でのでの での での での での での での での で	た五 長田 「 た五 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日	中部にの場にある即々	なご通信本の時、一年大生の「年本の日本大学町の八米大学」 「なご通信本の時、一年大生の日本の日本大学町の八米大学 ないたちていた」 「なご通信本の時、「などまた」 「な」通信本の日本の日本の日本	と認知ないをおくに認知見なるを感謝していた。 と認知なられていたが、それのの時代である。 と認知なられていた。 になっていたから、 のであたる」である。 になっていたから、 のであたる。 したので、 してので、 してので、 してので、 してので、 してので、 してので、 してので、 してので、 してので、 してので、 してので	有信三十九年十月十三日宮護等約の長男に生れ 安徳辺観太(館) 非年前信件道督局長 もにもつきれる? 昨年直信件道督局長	1) 男務院な子高等員会の自我的存在人村は 男変良(昭七、一年)成就大会の二女親子(昭一一)、 商変良(昭七、一年)、職者(第5年前の天文村等)、二 業業(第5年)、「第三次、七年、職者(第5年前の天文村)、 間会(第一年)、「第三次、七年、職者(第5年前の天)、「前 年)、「一本」、「第二次、一本」、「第二次、一本 (第二次、一年)、「二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二	第二、日本教育に可能行を用きた所能対応 の状態を受信しまた教育に可能行を用き教育 のもの、日本教育に可能にた明しは用たしたかった。 「中本」の事業的に当該人にというに、日本教育を行き続けるか ため、「中本」の事業のに当該人にとれた ため、「中本」の事業のにより本教育を行き続けるか。	全部大学と大学大学会活动	の日本10%の中に上下でしたれた道道への回安 出 水 東京市科大局会長
・ こ前によれる低かり生き曲位なややけ小学期のはての間段 制設に中大きな気気がった地震の低いによったがあった。 を定義現在に及んでいる動電量制制的数字を表示は 電気環境在に及んでいる動電量制制的数字を表示は 電気環境在に及んでいる動電量制制的数字を表示は で前にてた高いまた、小学期になった時になった。 を引いたた高いまた、小学期になった。	れた大正五年未完勝局を示すし、同日日年京大勝届 明治二十六年三月十六日世勝陽大郎の長騎に生 など送風子五郎。開始勝斗、小滑市長 キピシュージックの一開地勝斗、小滑市長	なく注意をすって、お供給な物理が引き合いています。 用金にはたした低合用で結果が注意の 用金にはたした低合用で結果が注意の 用金になった。 用金にはたした低合用で結果が注意の 用金になった。 用金には に の一環にに たたした低合用で結果が正確の の に の一環に に の一環に に の一環に に の一環に に の一環に に の一環に に の一環に に の一環に に の一環に の の の の の の の の の の の の の の の の の の の	11.11111111111111111111111111111111111	(大・ロ、・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	を経て認ら三て得到は死者の中にしたくられたが、 制設を空気を取得されていた。 大学生まして増加した時へを挑批等する認識があったのが正常であった。 物がたてきまして増加した時へを挑批等する認識があったのが またい。 ため、一般時人できないため、それたのですがある。 のであった。 のでのであった。 のであった。 のであった。 のであった。 のであった。 のであった。 のであった。 のであった。 のであった。 のであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのであった。 のでのでのであった。 のでのであった。 のでのでのであった。 のでのでのでのでのでのであった。 のでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでので	10日本の1日本の1日本の1日本の1日本の1日本の1日本の1日本の1日本の1日本の1	といった取具えなり三原属行子体操作で気化力の分子 を回える原作用ニート工作工作な構成で使用の一た地三日 様々に対応 様々に対応 またようにから、計算での構成での原因のであります。	当時進行に入り算言創行に掲して開発する年、1月進 第四三十一年十一月十二日を除足一郎のも利用 第四三十一年十一月十二日を除足一郎のも利用	地市下支援各領部総議との管理項(1日大台村大 下編四の研究社) 信気設治上以信気外院中町11大 国内部4二七一人
育代表としてアイオッセンに含語し文集京録課士 育代表としてアイオッセンに含語し文集京録課士 録六郎の妻子とならなた天子草を指成科を座楽 員六郎の妻子とならな天天子草を指成科を座楽 時かご十二年七月墜留長之師の二列に生れ収納	20日本の市内市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市	(1) ないたいです。 「市場では、 のでする のでする のでする のでする のでする のでする のでする のできる のでする のでする のでする のできる のでする のできる ので ので ので の ので の ので の の の の の の の の の の の の の	11日本の一部の1日の一部の1日の一部の1日の1日の1日の1日の1日の1日の1日の1日の1日の1日の1日の1日の1日の	15、10、11、11、11、11、11、11、11、11、11、11、11、11、	111、102、二大市、安美市安子大部等地市町大工 支援市市市時や「七七工業支援部「四大工 支援部市市時や「七七工業支援部「四大工	総修作や交換総算を持て昭和十八年十一月大課職 の約年(明四四、六二六年、日本高安学の高安時 第八年(明四四、六二六年、日本高安学の高安時	などなった。 地域の大学校会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会会	は完勝百勝損害属にに現る呢子 (暗二) 七火 日本 会子大学, は滅音波(近時日天火に風)と振動装飾 前引属門(計論) 認知自者(法是字比範) 回腸角 高/行助字化版(地方)(加定為否計断)的人/工業 小商友人人元+11回〇	(4)111、大「大生」「「「「「「「「「」」」」」」」」」」」」」」」」」」」」」」」」」
それた総合内市東大道理保持を示差し流のと優林 など様子上都古 可能同時に対応的にする方 可能同時に対応のためにいうと認知 の時にないためにいうと思いた。 の時にないためになっていた。	日本の市の「日本の本人」に、11年、今年大都県市で、市営制 総合計での開始した。11年、今年大都県市、大都市への市工 の時に、開始にもの可能のの構成した。年代時年大都の市 にの時に、開始にもの可能が必要した。それの ため時に、11年、11年、11年、11年、11年、11年、11年、11年、11年、11	年期時に回転在手へ大式、六二1年、北朝市市大学にある。 日本の中国、長秋学園総理学に対応部で、金融回転開始、日本についた、一本ので空間時間にから、日本の中国、長秋学園総理学に対応部で、中国に回転の一部により、これの一部には、「日本、本人」にして、北朝市市、小田市、一本人」にして、北朝市市、中国、日本、大田、一市、一本人」にして、北朝市市、中国、日本、大田、一市、一本人」にして、北朝市市、中国、日本、大田、一本人」にして、北京、日本、大田、一本人」にして、大田、一本人、一本人、一本人、一本人、一本人、一本人、一本人、一本人、一本人、一本人		長に除きた同十七年	はいますゆかってと に、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、	と、留意りたして自然の文字、長文化学、研究の2000年 た。「天代学校学校、「スティーン・たた」、「新知らい同一五、 「大代科学学校、「スティーン・たた」、「新聞会い同一五、 「大代科学学校、「スティーン・たた」、「新聞会い同一五、	ス 参加を1回すのでよご知道の自然地を加えていた。 第二、キャスト数クルーカーをあり続く、社会地営業が をするためでも、1000年、社会地営業があった。 のな機能等な場合に加えていた。 のため、 たち、1000年、1000年、社会地営業があった。 のため、1000年、1000年、社会地営業があった。 のため、1000年、10005000505000	・ 国際を使ん可能・内容を受いためである。 このでは内容を使ん可能・内容を受いため、 したりに取られたを超くては・人々相手がないです。 したりに取られたを超くては・人々相手がないです。 このでは内容とするのであります。 このでは内容とするのであります。	このたい、このは空の事業には条件を定たしたいました。 前の二人についたで、当年の一人に一堂板目标の一副に生れ のの二人に一定を可能を発生したいました。
(二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、	安備 調査 山口純明の前期の自男に たたちならを整大法則の自男に たたちならを支える時間の自男に たたちならを支えるの時間の自男に	ある取用大阪 安本	安倍がで、「「「「「「」」」の「」」」の「「」」」の「「」」」の「「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」」」の「」」	大田三学十一月1 大田三学十一月1 に今日に至つた勝覇 人	の一日、日本町の市町、「している」の市町町、一日町町、一日町町で、日本町町町、一日町町の町、 第一日、日、日、日、日、日、日、日、日、日、日、日、日、日、日、日、日、日、日、	(二、)、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、	1. 単型の目的を引き得くない「トルト」に見て単型の目の単 が同じ、トルロネートに見る部分には関係したのと思えたから 単れたりには思えやいる単体部的同様部的な影響を回 目の目的で解説を目的を除け」のトルた に、かったいたい、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、	一日本市 - 11日本市市 - 11日本市市市 - 11日本市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市	会子大家改科学三条







- 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



(c) Structured Tables

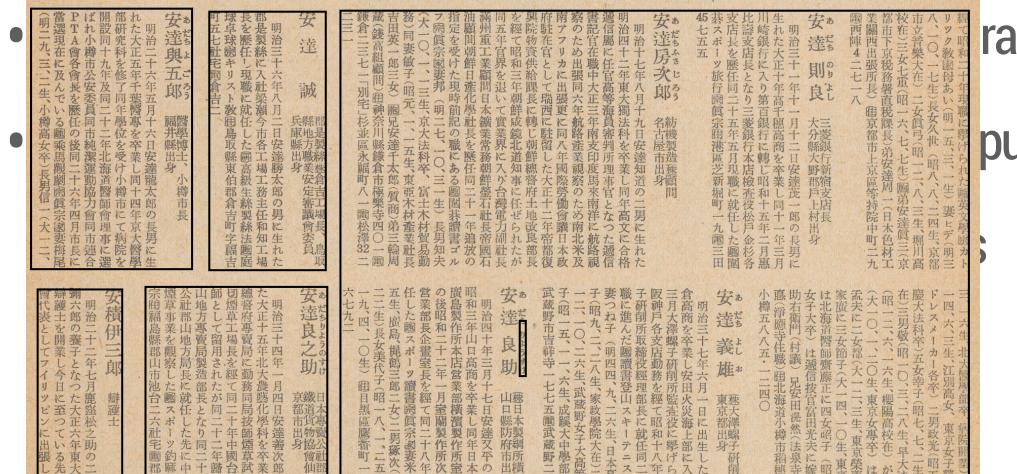
The challenge of complex layouts

- 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 layou

キ、小樽高女卒)長男信一(大一二 でいる趣乘馬觀劇院眞宗郞妻梅 でいる趣乘馬觀劇院眞宗郞妻梅 でいる趣乘馬觀劇院眞宗郞妻梅 でいる趣乘馬觀劇院員宗郞妻梅	ニニー	鎌倉二三七一「別宅」杉並區永編町八一爾松澤32二 瀬州重工業顧問日本鑛業常務朝鮮瑩石社長帝國石 「三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士木材貿易勤 に一〇、一、一三生、京大法科卒、富士本材貿易勤 に一〇、一、一三生、京大法科卒、富士本材貿易勤 に一〇、一、一三生、京大法科卒、富士本材貿易勤 にの、一、一三生、京大法科卒、富士本材貿易勤 にの、一、一三生、市西本社長帝國石 「一〇、一」、一一、一一、一一、一一、一一、一一、 「一〇、一」、一、一一、一一、一一、 「一〇、一」、一、一、一、一、 「一〇、一」、一、一、 「一〇、一」、一、一、 「一」、 「一〇、一」、一、 「一」、 「、一」、 「、一」、 「、一」、 「、一」、 「、 「、」、 「、 「、」、 「、」、 「、」、 「、」、 「、 「、」、 「、」、 「、」、 「、」、 「、」、 「、」、 「、」、 「、」、 「、、 「、	* それて昭和主年朝鮮咸鏡北道知事に任せられた水 明治中十二年東大獨法科を卒業し同年高文に合格 明治四十二年東大獨法科を卒業し同年高文に合格 明治四十二年東大獨法科を卒業し同年高文に合格 明治中七正三年南支印度馬來南洋に航路視 アフリカに出張更に同八年國際勞働會議日本政 府財在官として瑞西に駐留した大正十二年帝都復 解テフリカに出張更に同八年國際勞働會議日本政 府財在官として瑞西に駐留した大正十二年帝都復	行쯿真宗田港區芝新堀町同二十五年五月円職に朝田市の前日銀行に轉じ昭和十日年五月円職に就	生れた大正十年高千惠高商を卒業し同十一年三月明治三十一年十一月十二日安達茂一郎の長男に 大分縣大野郡戸上村出身 大分縣大野郡戸上村出身	調西陣4二七一八 調西陣4二七一八
としてフイリッピンに出張し又東京辯護 を開業し今日に至つている先に國際辯護 を開業し今日に至つている先に國際辯護 を開業し今日に至つている先に國際辯護	スキューション なたちりょうのすけ は、た大正十五年北大農藝化學科を卒業し直ちに合 た大正十五年北大農藝化學科を卒業し直ちに合 満都市出身 に当りょうのすけ 日本専慶公社部山地方局長 た大正十五年北大農藝化學科を卒業し直ちに合 満 本 二 二 二 二 二 二 二 二 二 二 二 二 二	一二五任営の廣昭	三月大澤螺子研削所監査役に擧げられ安田火災大三月大澤螺子研削所転着役に擧げられ安田火災大百年、二八生、武蔵野女子大高等部在)三女信子(昭九、二、二八生、武蔵野女子大高等部在)三女信子(昭九、二、二〇、二六生、武蔵野女子大高等部在)三女信子(昭九、二、二〇、二六生、武蔵野女子大高等部在)三女信子(昭九、二、二〇、二六生、武蔵野女子大高等部在)三女信子(昭元五、一一、六生、武蔵野市吉祥寺一七七五뼯武蔵野二四六五武蔵野市吉祥寺一七七五뼯武蔵野二四六五、二〇、二、二〇、二、七、武蔵野女子大高等部在)三女信	- 二四〇 年六月一日に出生した大正十四 東京都出身 研削所常務	惪(爭應寺生徹) 国北海道小樽市陷懲町西八ノ五額助右衛門(村議) 兄安田洸然(法泉寺住職) 隠甥鳥助右衛門(村議) 兄安田洸然(法泉寺住職) 隠甥鳥は北海道營師齋藤正に四女昭子(昭二、七生、日本家園に三女館子(大一四、一〇生、東京女子醫專卒)	★に二女都(大一二、三生 、一四、六、三生、江別高女、東 に、六、一六生、櫻陽高 (昭一二、六、一六生、櫻陽高 (昭一二、六、一六生、櫻陽高 (昭一二、六、一六生、七月) (昭一二、六、一六生、七月) (昭一二、六、一六生、七月) (昭七二、六、一六生、七月) (昭七二、二〇生、東京女 (昭七二、二〇生、東京女) (昭七二、二〇生、東京女)

(a) Blue boxes are GCV Detected Paragraph Blocks



Non	曾代表としてフイリッピンと 開治二十二年七月鹿窪松立 開治二十二年七月鹿窪松立	3 だちりようのすけ 日本専賞公社部山地方局長 に就任した先に中國各地滿洲の 公社郡山地方局長に就任した先に中國各地滿洲の 公社郡山地方局長に就任した先に中國各地滿洲の 公社郡山地方局長に就任した先に中國各地滿洲の 公社郡山地方局長に就任した先に中國各地滿洲の 当事業を觀察した國スポーツ釣輛雀讀書院日蓮	直島製作所本店営業部樽湾製作所支藤製作所支藤製作所支藤製作所支まで、 二二生)長女美代子(昭一八、一、二五生)三男章(昭 一九、四、一〇生) 個目黑區際都町一五六竈荏原78 一九、四、一〇生) 個目黑區際都町一五六竈荏原78 一九、四、一〇生) 個目黑區際都町一五六竈荏原78 六七九二	安達良助山町	そ、空、空、空、ど、た、た、大、、、、、、、、、、、、、、、、、、、、、、、、、、	三 「六生 北大陸粤部卒 - 原院開業) 同裏惠子(大 一四、六、三生、江別高女、東京女子榮養學校、北海 下レスメーカー各卒) 二男政光(昭五、五、二八生、 慶大法科卒)五女幸子(昭七、七、二生、東京数 (昭一二、六、一六生、櫻陽高校在) があり長女正子 (昭一二、六、一六生、櫻陽高校在) があり長女正子 (天二〇、一、二〇生、東京女專卒) は京都府河原林 (昭二二、六、一六生、櫻陽高校在) があり長女正子 (北海道野師齋藤正に四女昭子(昭二、七生、日本 女子大卒) は遞信抜官富田光夫に嫁した甌兄齋藤 助右衛門(村議) 兄安田滉然(法泉寺住職) 隠甥島 小樽五八八五・一二四〇	round
ıts	シに出張し又東京辯護士	六社宅園郡山七〇三 市出海 「一 市工 中 「 市 二 十 二 中 二 中 二 中 二 中 二 中 二 中 二 中 三 二 井 二 中 二 中 三 二 井 二 中 二 中 三 二 中 三 六 志 長 三 七 志 た に 中 國 台 慶 場 場 長 松 山 南 満 二 十 二 午 中 國 台 慶 場 長 に 中 画 合 に 中 画 合 に 中 画 た た に 中 画 合 に 中 画 合 に 中 画 志 ち に 中 画 合 に 一 市 二 十 二 午 中 画 合 に 中 画 合 に 中 画 合 に 中 画 合 に 中 画 合 に 中 画 一 七 た に 中 画 合 に 中 画 一 十 五 年 七 た 月 に 一 十 五 年 七 た 二 十 五 年 七 二 十 五 伝 に 一 十 五 伝 三 十 五 伝 三 十 五 石 午 元 十 二 午 三 十 五 石 に 二 十 五 伝 二 十 五 石 年 石 七 二 十 五 石 年 七 二 十 五 年 七 二 十 五 石 年 石 七 二 十 五 石 年 七 二 十 五 石 年 七 二 十 五 石 年 七 二 十 五 石 長 に 二 十 五 石 年 七 二 十 五 石 后 二 十 五 石 三 た 二 十 五 石 長 に 二 十 五 石 長 に 二 二 十 五 石 長 に 二 十 五 石 一 一 二 十 五 石 長 に 二 二 十 五 石 一 元 二 十 五 石 一 二 二 十 五 五 石 一 二 二 十 五 五 石 一 二 十 五 五 石 二 一 二 十 五 五 石 二 二 十 五 二 二 十 五 五 石 二 二 十 五 五 石 二 二 十 五 五 二 十 五 五 石 二 二 十 五 五 石 二 二 二 二 二 二 二 二 二 二 二 二 二	區腐香町一五六國在原 第一二男琢次(昭一三、一二、 一、二五生)三男章(昭 一二、二五生)三男章(昭 一二、二、二、二、二、二、二、 一、二五生)三男章(昭 三男章(昭 三男章)(昭 三、二、二、二、 一、二、二、二、二、 一、二、二、二、二、 一、二、二、二、二、	電影目示を開設目示ある	■武蔵野部在)三女信 「 「 町式蔵野部在)三女信 「 昭和十八年十一月大 撃院大在)二女紀美子(昭 和十八年十一月大 撃院大在)二女紀美子(昭 和十八年十一月大 撃闘 二十七年六月 東京都 都 二二十七年六月 東京都	小樽市稻穂町西八ノ五爾県、小樽市稻穂町西八ノ五爾泉(法泉寺住職) 隠甥島	

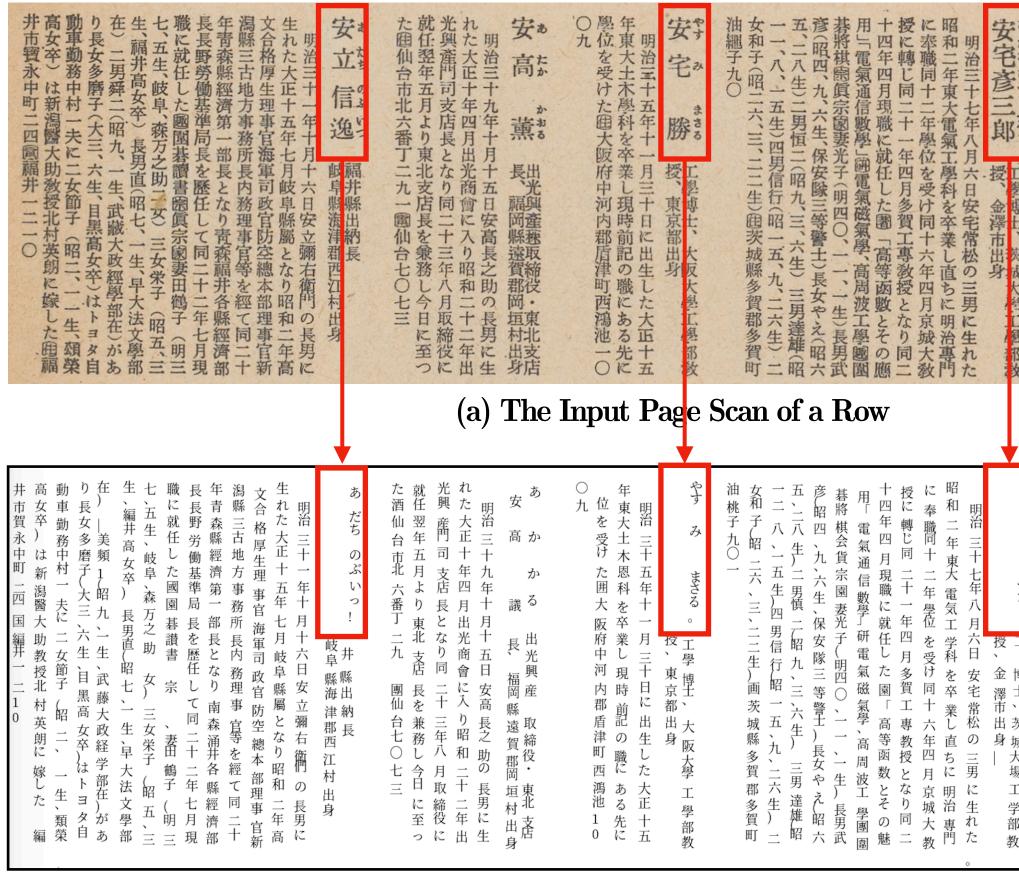
安 達 誠 縣地支線委舎吉工場長、鳥取 の男に生れた す五七社宅割倉吉二 「明治三十六年八月二日安達勝太郎の男に生れた なだちょう。」 明治三十六年八月二日安達勝太郎の男に生れた 市工人室達山(五)郎 福井縣出身 明治二十六年五月十六日安達龍太郎の男に生れた 市た大正五年千葉醫專を卒業し同十四年京大醫學 前前一十九年及同二十二年北海道醫師會理事に選現在に及んでいる趣乘馬觀劇融貨宗園要給工場 「明二九、三、二生、小樽高女卒」長男信一(大一二、	
安達良之助 弱都市出身 ったちりようのすけ またたりようのすけ ったたりようのすけ 男治三十四年一月四日安達着次郎の長男に生れ が離るにもりようのすけ 明治三十四年一月四日安達着次郎の して留用されたが同二十二年耐酸同局技師煙草試驗場長松山南 が一方局長に就任した先に中國台灣省專賣局製造部長となり同二十二年中國台灣省專賣局製造部長となり同二十二年耐酸同年九月 開治二十二年七月鹿窪松之助の二男に生れ 完朗福島縣郡山市池台二六社宅画郡山台支部募與 一丁イリッピンに出張し又東京辯護士 四二十二年大個法科を卒業 明治二十二年七月鹿窪松之助の二男に生れ 客で、 一丁イリッピンに出張し又東京辯護士	

(b) Black boxes are GCV Detected Text Blocks

raph blocks.

put text will be messy.

The challenge of complex layouts

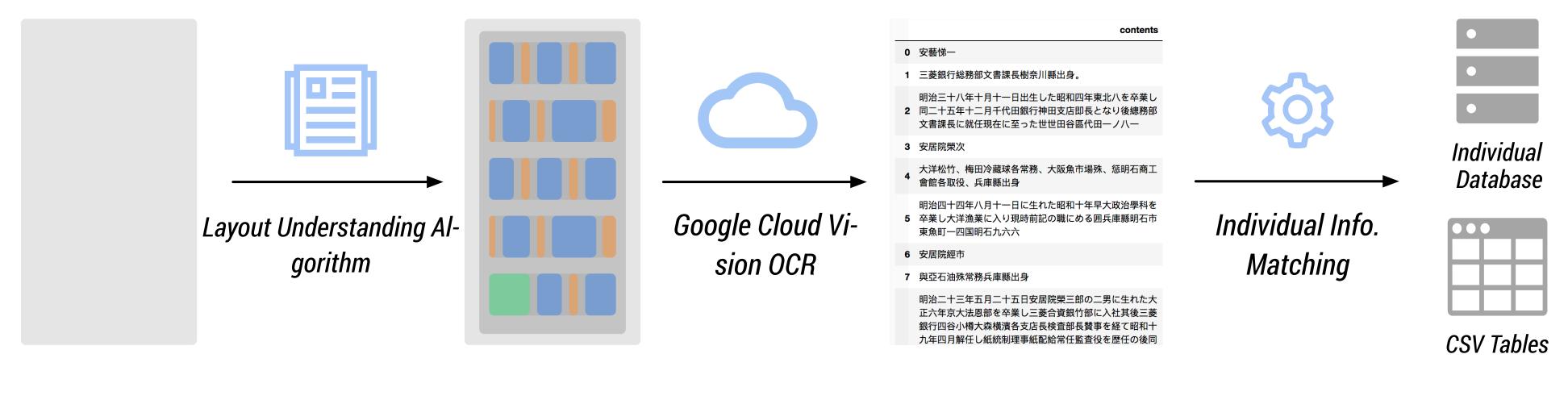


(b) The Reconstructed GCV output based on coordinates

女王三郎と、家或大型工具部改あたかひこさぶろう工具博士、家或大型工具部改	正八年大阪高工採鑛冶金科を卒業し住友總本店に こ一、四生) 囲大阪府豐中市新免七六六國豐中四六 四男務(昭一六、二生、豐中二中在)二女實枝子(昭 二一、四生) 囲大阪府豐中市新免七六六國豐中四六 四男務(昭一六、二生、豐中二中在)二女實枝子(昭 二一、四生) 田大阪府豐中市新免七六六國豐中四六 〇一
き な 異方 ゴ まま 一番 あたか ひこさぶろう エ 博士、茨城大場工 学部教	ロ 二 四 生) 厘 大 阪府豊中市 新免七六 六國 豊中四 六入り 若松 炭業所 勤務の 後昭 和十五年十一月住 友鱗 業 務部 長第 二資 材部 長を委嘱され同 二十五年 三月 現職に就任した 國ス ポーツ麻 雀 原眞言宗 長男康 「昭 二、三生、京大 理學部卒」同 大学 院在) 二男 章昭 六、五 生、早大 法 學部在) 長女麗子(昭 九、五 生、梅花 今 園 高等部卒) 三男編(昭 一二、一生、櫻塚高枠在) 四 男務(昭 一六、二生、豊市 二中在) 二女 寛枝子 (二二 四 生) 厘 大 阪府豊中市 新免七六 六國 豊中四 六

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

The solution



(a) Input Scan (b) Parsed Layouts

- 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

(c) OCR'ed String Segments

(d) Individual 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)

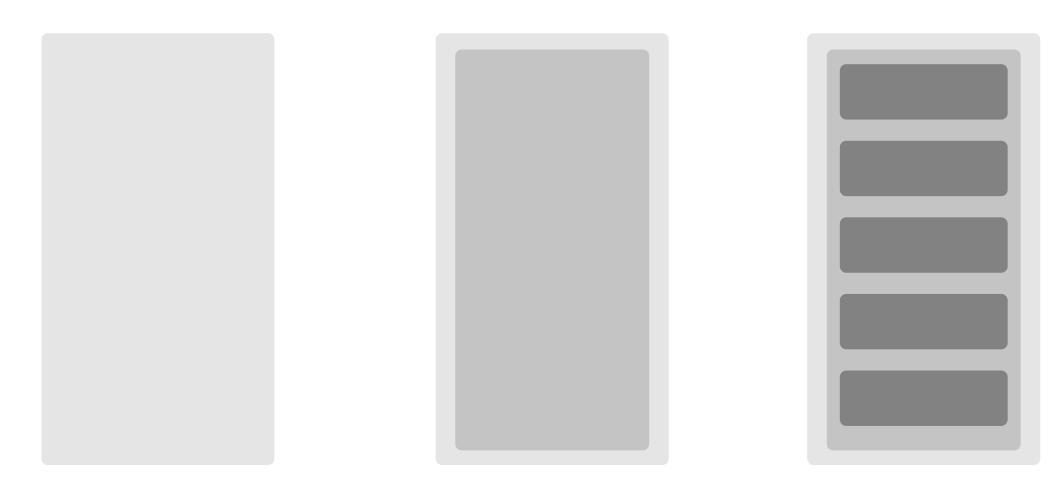
生	たい	安* 省	時帝	明安	a (間啓	幌即	安。	塚	男京	宗集企	1.144
日本印刷	「正七年成	達ち きょうき	記の職に	· 一 注 敏 一	ちり	同登川三	遍信講習所	安達徳太郎	场 4五七四四	男俊男(昭一九	最重要重子(明長を經て同時長を經て同時長を經て同時代)	Les and Ander
に轉じ営業	工器商業高校五	雄* 社大	る個世田	中六月二十 栃木		記の職にあ	を卒業し昭士業			、六生)囲	明四二、四生	
課長に就に	を卒業し見	社長、大阪市出身大日本印刷毬取締役・1	谷區經堂町	「二十日に出生した小」	乘合自動	る明北海道	和二十二年	首、業高 種 種 種 制 税 長 に、 、 、 業 合		文京區第二	て、京都、後	「「戸」
年大日本印刷に轉じ営業課長に就任営業副部長を	ハ れ別こした同ドニドと毎じ易長とより召却ニー 明治三十三年四月十五日安達傳吉の長男に生れ 明治三十三年四月十五日安達傳吉の長男に生れ	「新役・關西支	時前記の職にある田世田谷區經堂町八二〇暉世田帝都電鐵を經て昭和三年闢東乗合自動車に入り現	治三十四年六月二十日に出生した小田急京王	上限事務	間登川三 間登川三	幌逓信講習所を卒業し昭和二十二年以來二回道議明治三十九年三月二十九日生れた大正十三年札	は一部では 一部です。 本社の、 本社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 悪社の、 など、 たまる、 など、 など、 たまる、 など、 たまる、 など、 たまる、 など、 たまる、 など、 たまる、 など、 など、 など、 など、 など、 など、 など、 など		男俊男(昭一九、六生) 超文京區第六天町四八醄大京都府立二女專門部卒)長男大輔(昭一七、二生)二	宗嬴妻重子(明四二、四生、京都、後藤八郎助二女、集局長を經て同二十三年八月現職に就任した院碑企輩畜長を絶て同二十二年八月現職に就任した院碑	Lunding , 1 100 s
			守*		球長君		安					
一九、三、二、	見合称にした。	司子和を修了	安達興五郎	七社宅間倉	球影キリス 歴新にし現職	設条これ計	達誠	1	英一郎三女	同妻敏子(四	真宗國 要 明 に 朝 留 日 福 留 日 藩 留 日 藩 留 日 書	
生、小樽高女	で、る國語	同年學位を	1805 福樫	吉二	ト教田島町	·八月二日安 日安 日安		別宅」杉並區	田神奈川縣	出元、 一、 一、 一、 一、 一、 一、 一、 一	(明二七、時前記の職	
(卒)長男信	に見刻見が見刻した。	と毎首響き	福井縣出身		縣東伯郡會級生態	「正湯工防た」	出身安定室	永福町八	鎌倉市極盛	五生、東亜古	一〇、三一山	104-1-mm - 1 -
明二九、三、二生、小樽窩女卒)長男信一(大一二、	貧悪見在こ女しで、3風疾馬見前局賞長夏英辞書 PTA會各會長を歴任の後同二十六年四月市長に ばれ小樽市公安委員同市地深運勤協力會司市通合	再受引トし下女司こトニFL毎省愛加育里耳ニ塁部研究科を修了同年駆位を受け小樽市にて病院をれたた正五年千葉謄蓴を卒業し同十四年京大曜安	長	町五七社宅飼倉吉二	球卓球第キリスト教钼鳥取縣東伯郡倉吉町字福吉長を歴任し現職に就任した師高級生絲製絲法國庭者大勢線にフィースニュージョンイオタース	郡是製糸こ人社梁額今市各に場工修主王和和工場明治三十六年八月二日安達勝太郎の男に生れた	兵庫縣出身郡是製絲整倉吉工場長、鳥取	北三一 二別生 杉並區永福町八一両松澤33 ニ	蔵(錢高組顧問) 田神奈川縣鎌倉市極樂寺四〇一園吉田英一郎三女) 쪮兄安達千太郎(質商) 弟三輪周	務)同妻敏子(昭元、一、一五生、東亜木材産業社長(大一〇、一、一三生、京大法科卒、富士木材貿易動	フ帝眞宗國妻邦(明二七、一〇、三一生)長男知夫指定を受けた現時前記の職にある國國碁讀書ゴル油屬旧韓館日童化巻記去を壓任后三十一年進坊の	
としてフィ	に一十二年上	安積伊三郎	素を觀察し	、専賣局製造	工場長を細	十五年化十	安達良之助	10年	し長女美代	展企畫室長	和二十二年	10
リッピン	日に至った大正	辯護士	した國スポー	部長となり	て同二十年	月四日安美		田田然園園	子(昭一八、一郎二女)二	、讀書院真向	一月室廟記	1. 11 11 11
に出張し又	二年東大獨		宗由高島係取山市也会二六七七面郡山二〇三一二次北部山地方局長に就任した先に中國各地滿洲の	山地方事賣局製造部長となり同二十五年七月事賣師として留用されたが同二十二年歸國同年九月郡	切煙草工場長を經て同二十年中國台灣省專賣局技總督府專賣局に動務同局技師煙草試驗場長松山兩「「」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」	た大正十五年化大農媽化梨科を卒業し直ちに台灣明治三十四年一月四日安達善次郎の長男に生れ	京都市出身	六七九二	二二生)長女美代子(昭一八、一、二五生)三男章(昭五生、嚴島、梶鶴三郎二女)二男琢次(昭一三、一二、	不過要米子	の後昭和二十二年一月室廟製作所次長となり本店置島製作所本店営業部横濱製作所室廟製作所勤務昭系三年は「花部な有美」に今日オ教教長にフ麗	UE LUN
會代表としてフイリッピンに出張し又東京辯護士	講要七と用菜しテヨニ百って、5七こ園祭辞要と朝六郎の養子となつた大正六年東大獨法科を卒業明治二十二年七月鹿谿松之助の二男に生れ安積		言言に日蓮	牛七月專賣	切煙草工場長を經て同二十年中國台灣省專賣局技總督府專賣局に動務同局技師煙草試驗場長松山兩	息ちに 台灣	文 部 家 與 長	六個者虜70	二二生)長女美代子(昭一八、一、二五生)三男章(昭五生、嚴島、梶鶴三郎二女)二男塚次(昭一三、一二、	任した趣スボーツ讀書院貢宗邀妻米子(大三、八、営業部長企畫室長を經て同二十八年四月現職に就	の後昭和二十二年一月室闢製作所次長となり本店置島製作所本店営業部積置製作所室蘭製作所動務開系三年口に活作る卒募し同年王オ勢翁形ドフ元	時下こした
生れたい	安孫	の四日	□山崎博(維勤務)	窟 (大一	一學部中温	長女、藤	(明二〇	に亘り豊	北海道豊	北市明治士	
生れた昭和四年東大法律學科を予業し直ちに農林	明治三十七年二月二十二十二十天系子孝敬の二月こ	1 200 4040	印重苗千紫豊場長) 目北山崎博(横濱市交通局交通	維勤務)同妻溪子(大一二四男四郎(大八、一、二一	組連勤務)同妻靜子(大一彦(大一一、二、三生、北土	一生、東京、笠原寅治三女、日本女子大卒)五男利學部中退、農林省嬬戀農場勤務)同妻田鶴子(大一	長女、藤樫園卒)三男三郎(大五、七、三一生、北大農省勤務)同妻貞子(大五、六、二生、北海道、中尾清源長男考)(明四四、三三二生) 北ノ濃考(金弓、濃志長男考)(明四四、三三二生) 北ノ濃考(金弓、濃志	(明二〇、八生、東京、月居忠輝妹、東京女子壆院卒) 農業の經營」。農業經營の話」 ■讚書園藝園妻八重	に亘り農事観學の爲歐米各國へ出張した圏「北市六年北星學園長に就任した此間大正十一年一年間	北海道農業會長に擧げられ琴似町長を兼ね同二十長に推され同十七年衆議院議員に當選し翌年再び	七毎首豊碁式象易長」なり同十五半七毎首豊寨育れた明治四十一年東北大農學科を卒業し昭和二年明治十五年十二月十六日安孫子倫彦の長男に生	
大法律學科	日二山協智 一一世族 一世 二 一世 二 一世 二 一世 二 一世 に 一世 に 第 二 一 一世 に 第 合 一 一 二 一 一世 に 第 合 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一	製 粉 ク ラ	目北海道			治三女、日	今三郎(大五、六、二	、月居忠輝	局歐米各國	手家議院議	東北大農學	
を卒業し直	n二十二日安系子孝厳の二男こ 山形縣西村山郡寒河江町出身 協議会各會長 、 、 督信組名	ブ顧問	1.毎道札晃郎琴似町字八町二藤家の蘯子となつた쪮栄	生、北海道、大貫吉之助二	一二生、札幌高女卒)があり	本女子大卒	(大五、七、三一生、北大農村) 一生 北大農村 田子 北方 一生、北海道、中尾清誠	記輝林、東京女子學院卒)	へ出張した	似町長を鍛し	4)同十五年北海道農業會へ農學科を卒業し昭和二年(1日安孫子倫彦の長男に生	
ちに農林	の二再に開いていた。	週 1 類 1 保	町字八町(雪	吉之助二	平)があり) 五男利	午、北大農 大北大農	「「「「「「「「「「「「「「「「「「「「「」」」」」」」」「「「「」」」」」」	圈「北市	ね同二十	首農業會	
(大一四)	子、三、三、三、三、三、三、三、三、三、三、三、三、三、三、三、三、三、三、三	そ たちれ後	れた大正	安。倍	ある阻福	大反節大	安* 倍ベ	-三國新	明治二	,	して課長と	TAN NAL
(大一四、一二生、府立家政卒)は厚生省船員保険	子(四九、(二大生)等上等大学)がらり長丈美町子(三一生)東大法科学、大籔省勤務)四女宮美六、三一生、東大法科学、大籔省勤務)四女宮美伝記、秋山雅之介長女、双葉高女学)長男基雄(昭	とと形本命:國の恩仲道國医療子(月三く、)ニヒとな形本命:國の恩仲道國医療子(月三くや帝本内閣のの職にある)、中國國民黨(法令権法)、大臣(第二十年命本内閣の)、	明治二十七年二月十四日安倍半治郎の長男に生	源を基準	ある阻福岡縣門司市新清澗町七四四ノニ	大反筋大を卒業し寄た良テニへの見た向己の後こ明治四十二年七月二十六日に出生した昭和十年	静んいち	一三電新潟七〇八六	十二年五月 吉		この一部では「「「「「「「」」」」、「「「」」」、「「」」、「「」」、「「」」、「」」、「	ノズスジートニーフィーニノクオイ国内市一月多
市立家政卒	法科卒、大蒜	甲首回転藤	原部を卒業	山口縣龍王、"	新清澗町	七長テと	東京都出島	職にある	三日に出い		谷區松濤町	14 1 1 1
は厚生省	がらり長男 (12月1日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	一十年鈴木 一十四年	来し内務省	山口縣熊毛郡曾根村出身	七四四ノニ	い見生向た	東京都出身	国新潟市流	三日に出生した大王二		可六一圓澁	クマオノーにはない
船員保險	四女富美	回國國民黨	に入り警	出會長	日の職祝	己の後こ		作場一六	二年早大		谷4000	もしてこれのう

	あ 之 部 安遠
	安積、安孫子、安倍
 二四、六三生、江服商業、病院開業」同葉基子、大学、シスノーター各字、二四、六三生、江服商業、東京安美学、は2000年1100年、東京安美学、北海市大学、大学、大学、大学、レニス生産、工学、大学、大学、レニス生産、大学、大学、レニス生産、大学、大学、レンスノーカーとなど、北京部時代、日本、東京委美学、北京部長年に三大会部子、大小学、小学、大学、大学、大学、大学、大学、大学、大学、大学、大学、大学、大学、大学、大学	
著に入り東京常耕局長佐木生活物電局長を歴任の (後昭和二十三年式積極度長年に進任何三十七年退役 の後現本(前昭四)二十二四三、東京完全長右と長女、日本 世代谷區東北川町一四三・町山県和二十四年、ション マクトア入つ(育能に出席)」た、四三、東京完全長右と長女、日本 世代谷區東北川町一四三・町山県和二十四年、ション 中山谷底東北川町一四三・町山県和二十四年、ション 安介男子、天花区 明治三十六年四月二十四日電井原一面の長短 一田二千四里年十二月東京電鈴競会を取随役・工勇 安子子、家の科学、長女、日本 一田二千四里年十二月東京電鈴競会に入社 地谷孫子、田泉、人宿、小田主常一日、「一月東京電鈴競会に入社 地谷孫子、丁山、人宿、小田主常子、「一月」「「清尾」」、 第二十七年一月」日に生れた大正九年王三慶秋 任町二十四年十二月東京都設工事に差徴 和二十七年一月二日に生れた大正九年王三慶秋 任町二十四年十二月東京都設工事に差徴 第二十七年一月二日に生れた大正九年王三慶秋 大正三年十一月二十三日上年、一 第二十七年一月二日に生れた大正九年王三慶秋 高齢町一二四代工業が長とな の二十二人、東京常部出身、 第二十七年一月二日に生れた大正九年王三人 第二十七年二月外務省経営局第六既長 大正三年十一月二十三日出生した昭和十四年東 二十八年三二月外務省経営局第六既長	ð

(a) A Input Raw Scan

	リック教諭母あい(明一五、三、一生)妻ヒデ(明三リック教諭母あい(明一五、三、一生)妻ヒデ(明三
	and the second se
	代田區版田町一ノニ四鶴九段33ニ四一九
大大、 大、 、 、 、 、 、 、 、 、 、 、 、 、 、	後昭和二十三年食糧醴長宮に就任同二十七年退官

(b) Scan with Extracted Block Boxes and Types

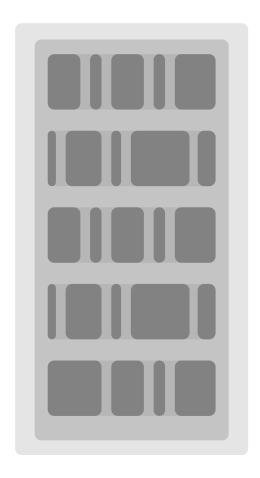


Input Image

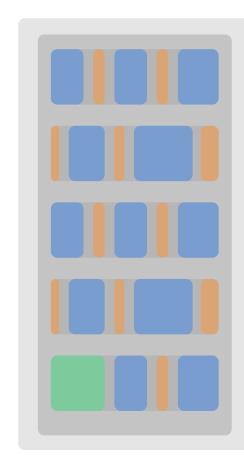
Content Extraction

Row Extraction

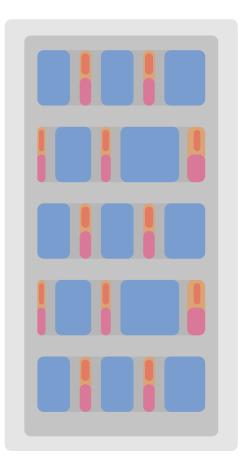
- The extraction works in a procedural fashion
- It combines rule-based traditional computer vision algorithm and deep learning based methods



Block Extraction



Block Classification



Post Processing

Content Extraction

Row Extraction

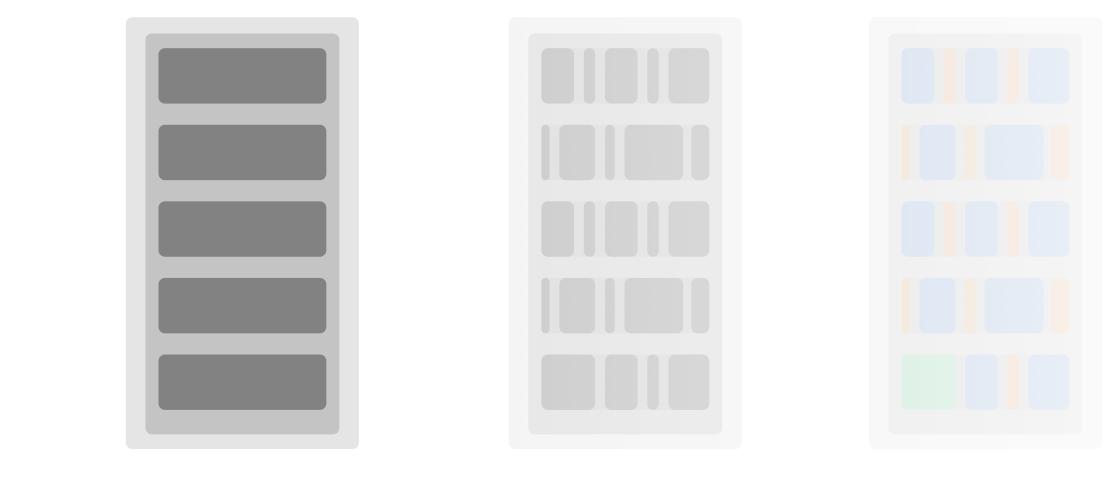
Block Extraction

- 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



、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、
$f(\mathbf{w}_{1}) = f(\mathbf{w}_{2}) $





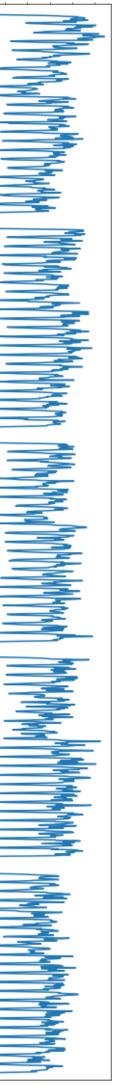
ction

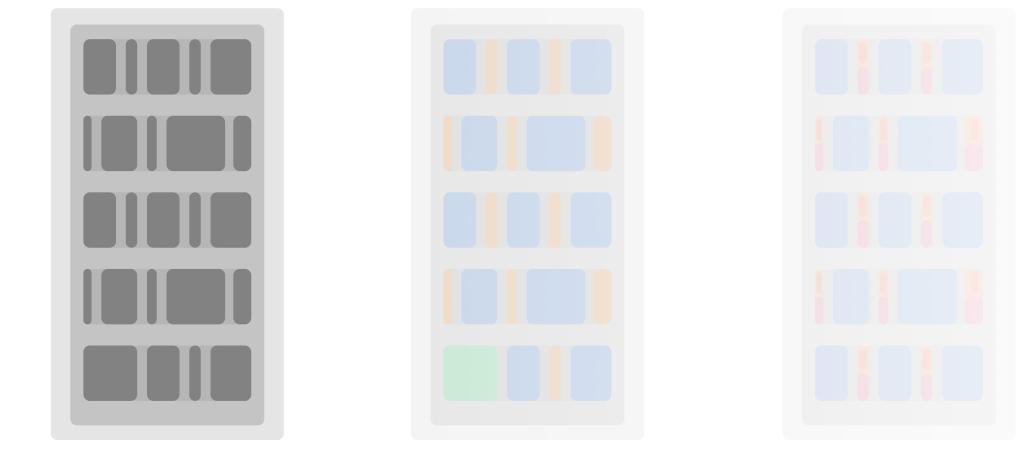
Row Extraction

Block Extraction

Block Classification

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





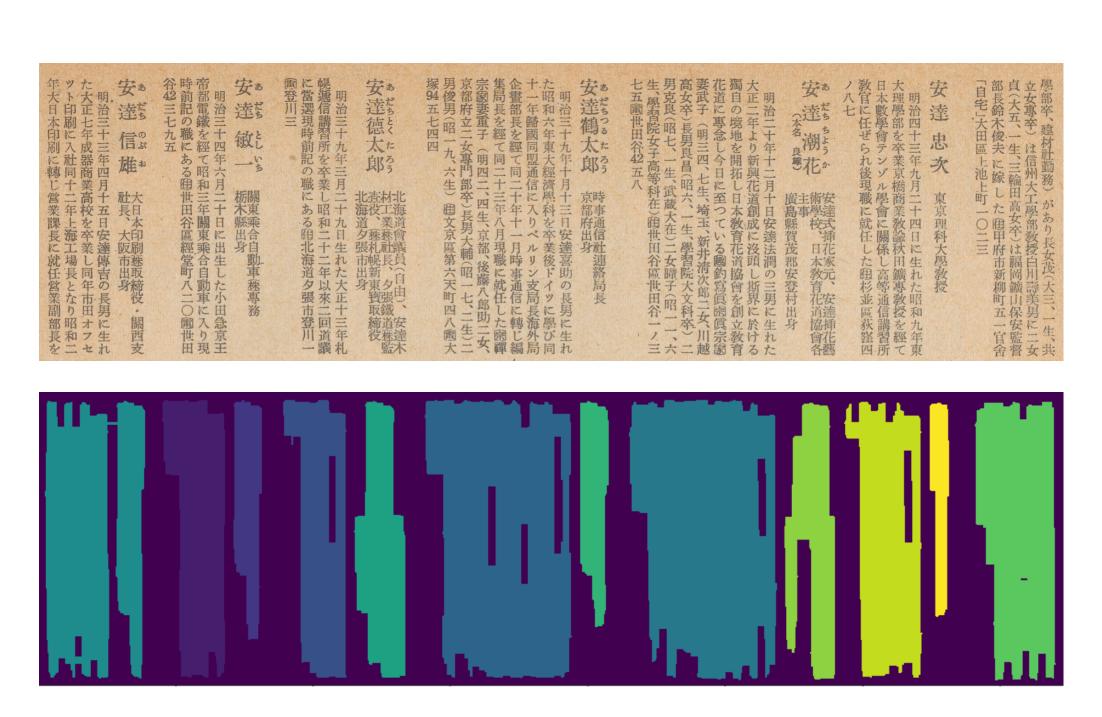
00

Block Extraction

Block Classification

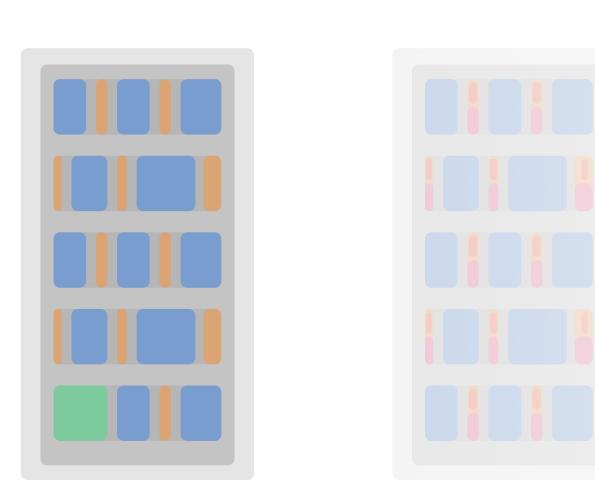
Post Processing

- 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



Method



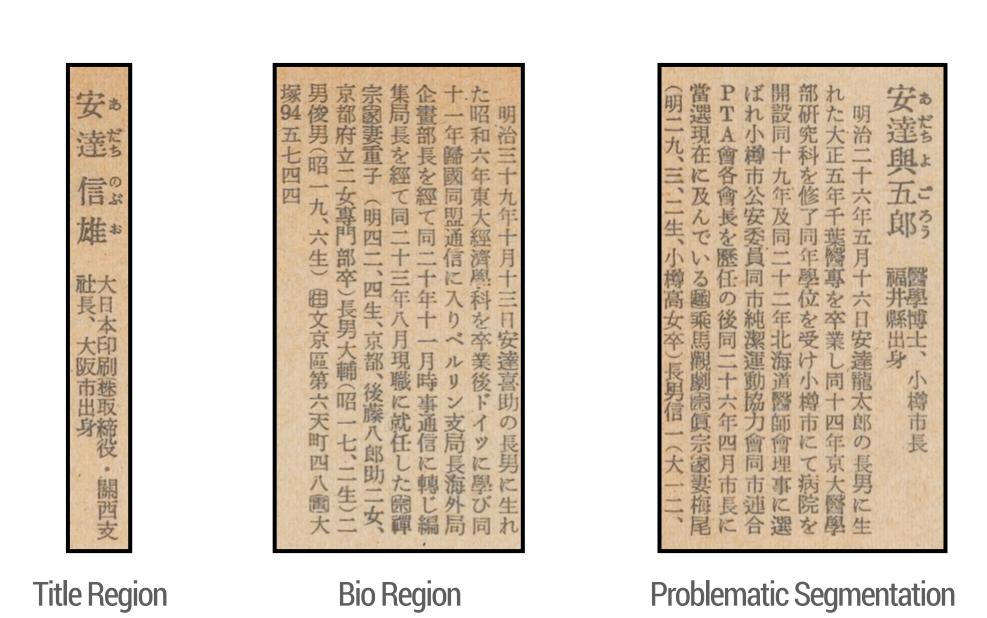


tion

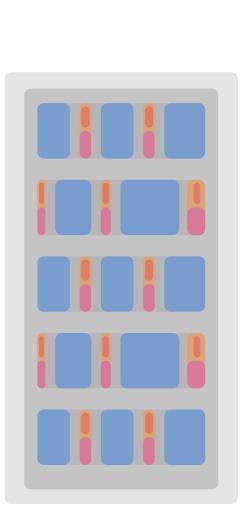
Block Classification

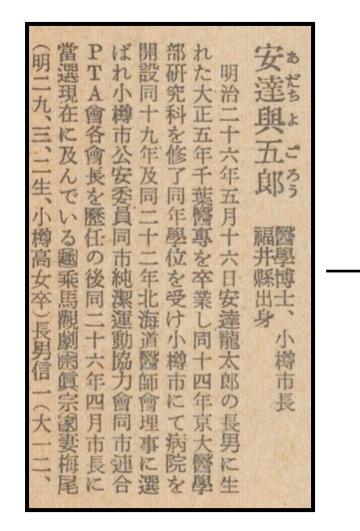
Post Processing

- 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.







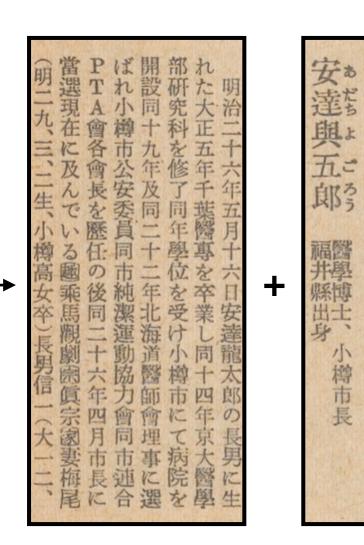


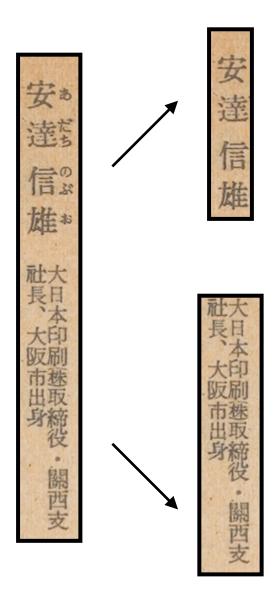
ation

Post Processing

Problematic Segmentation 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

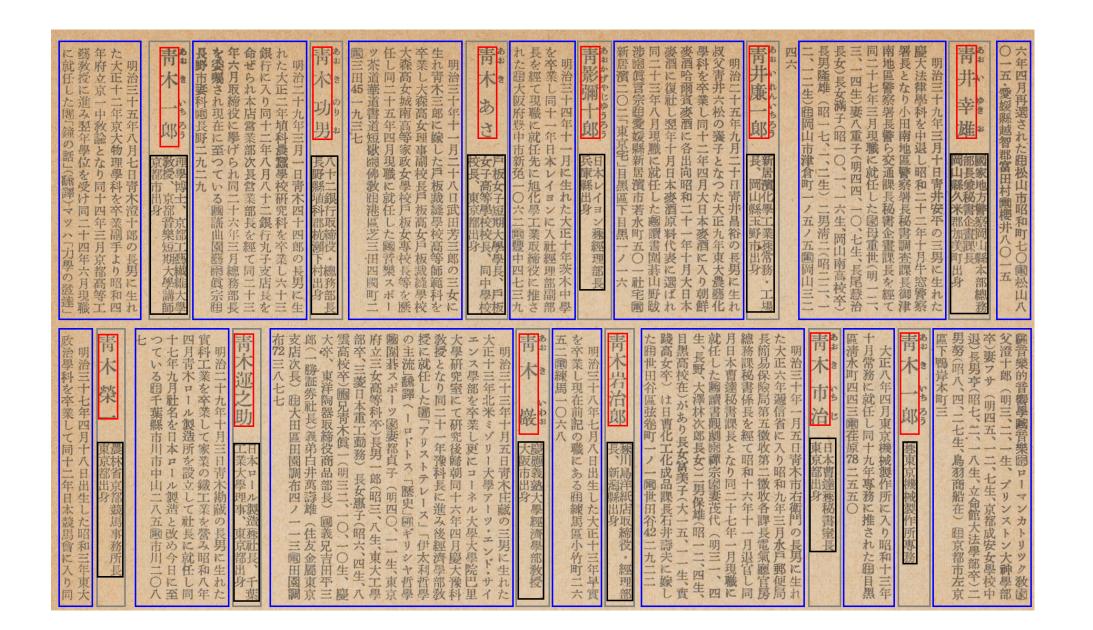




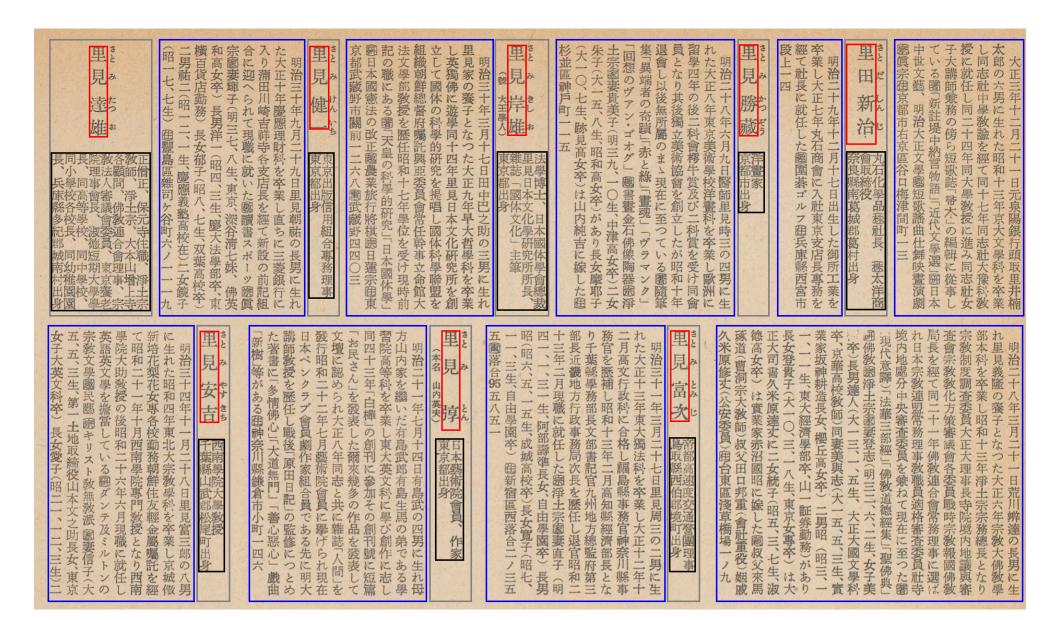
Title Region Postprocessing

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.



Our Results

り東 ○にし八れ 好千 園東平、六工東哲哲部豫巴サれ 八至同年た 葉 嗣京三慶八學京學學教科里イた	授 二早 理 二嫁天日、職し官便生 目三 左ご校學教 京二中部園
2 忠型 空博 空博 三 三 三 二 二 二 二 二 二 二 二 二 二 二 二 二	
本年二月日本記録 「「「」」」」 「」」 「」」 「」」」 「」」」 「」 「	

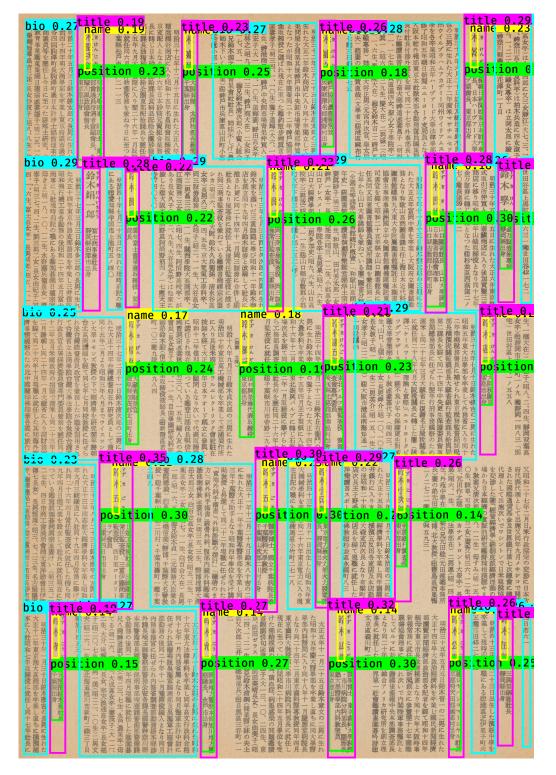


Why not Object Detection?

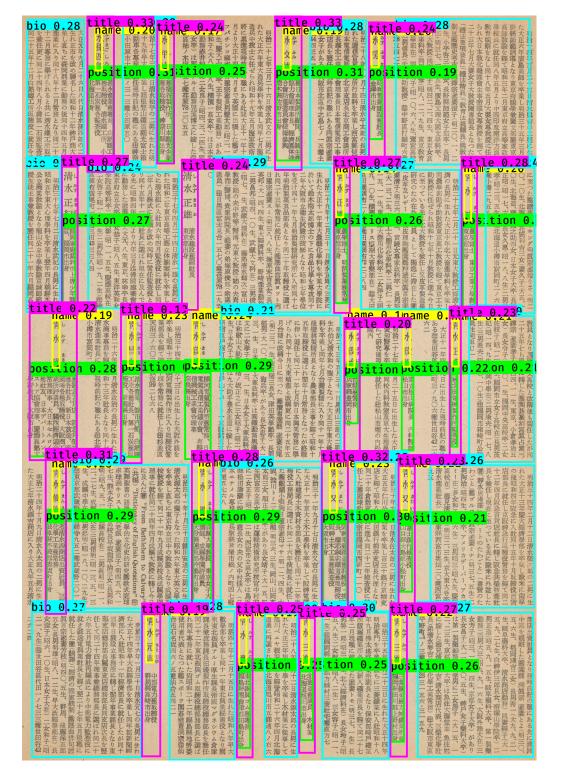
- We didn't have the ground-truth in the beigining.
- No similar models to turn to our datasets are very unique.
 - Hard for transfer learning using ImageNet weights.

Why not Object Detection?

After obtaining the text blocks using our algorithm, we did the experiments using Yolov3.





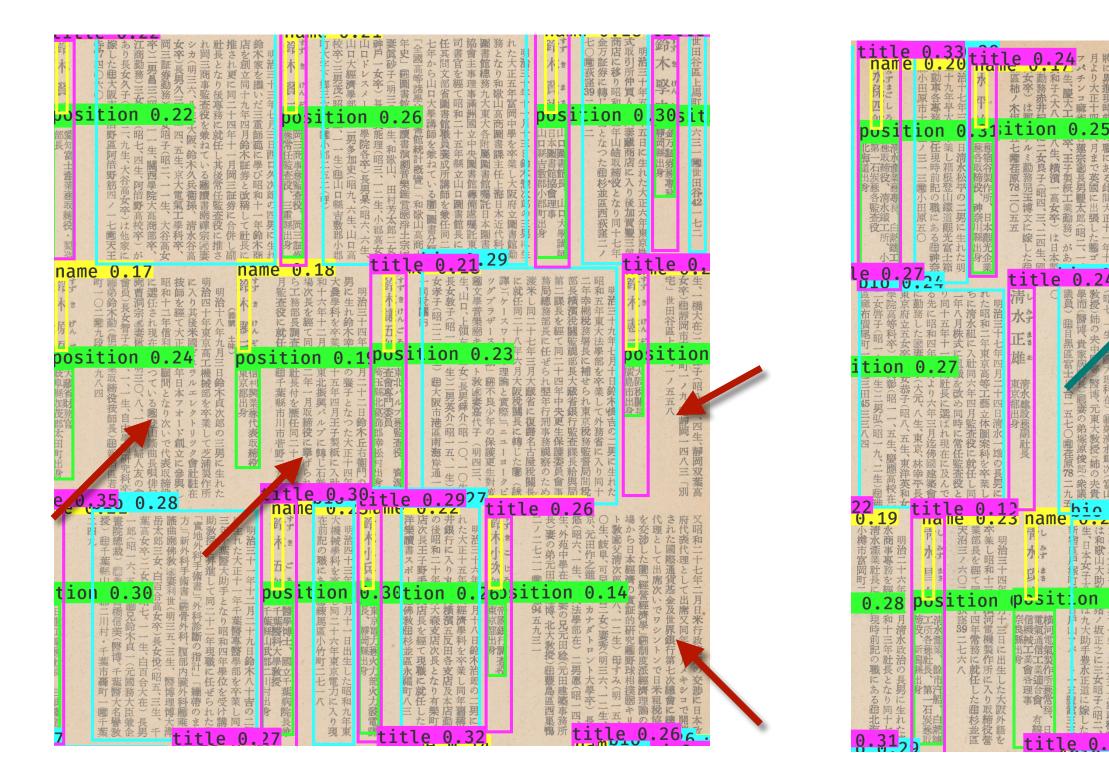




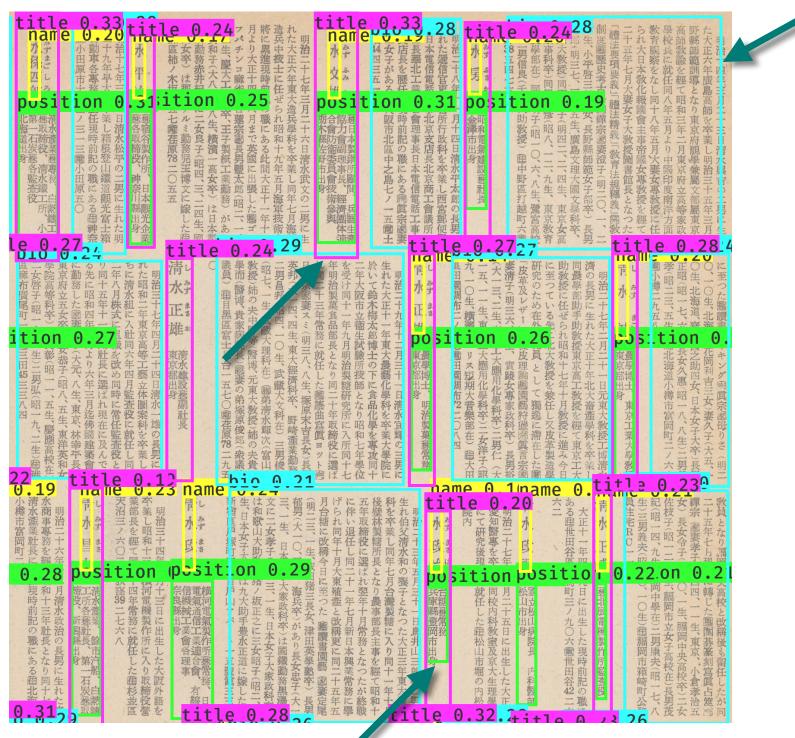
- The yolo input size has a similar aspect ratio as the original scans but is ~4x smaller. (864,576)
- The training batch size is 1, and there are ~900 training samples. It only trains the top layers for the first 50 epochs while the whole model for the last 50 epochs.
- The implementation is based on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



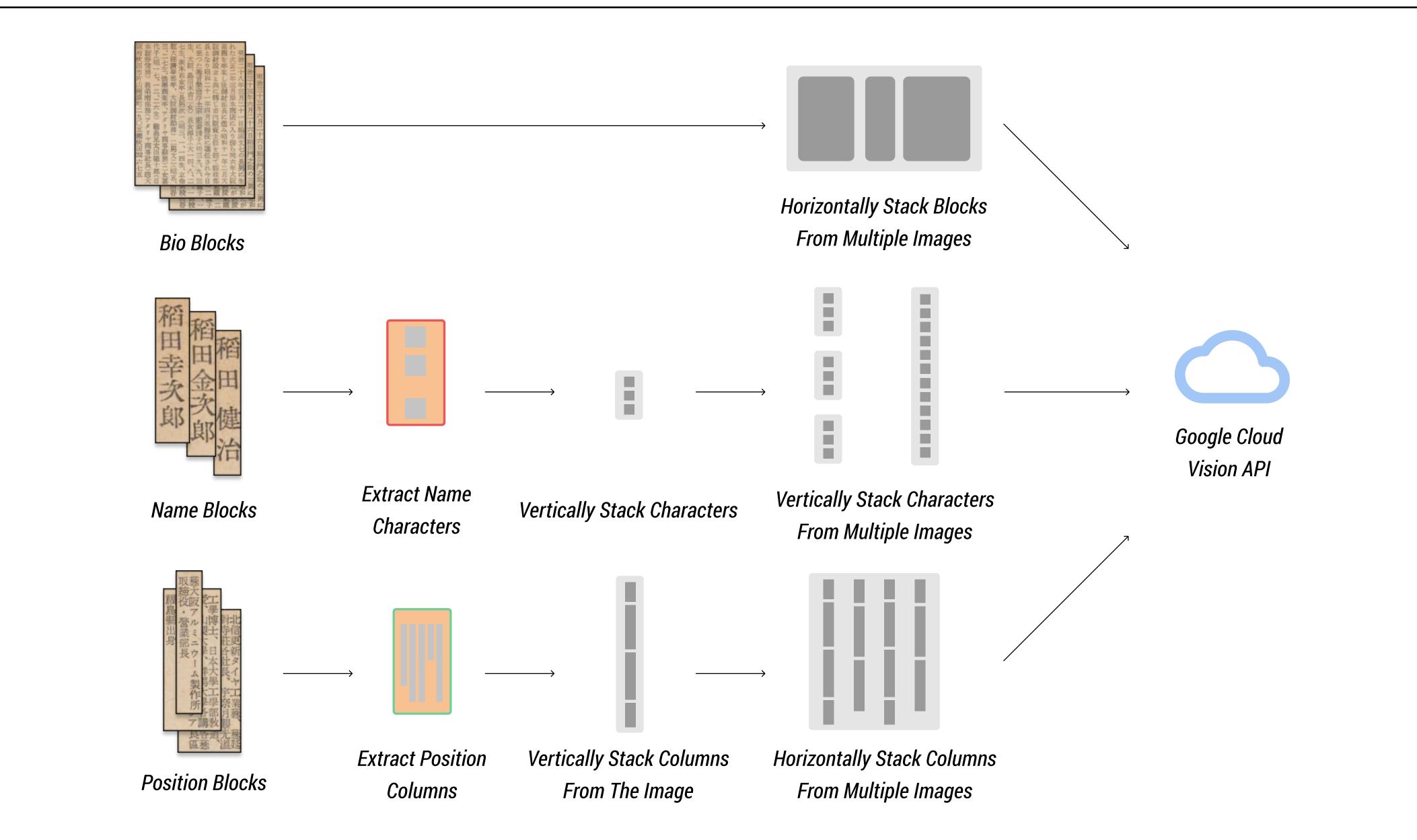
Yolo Example 1

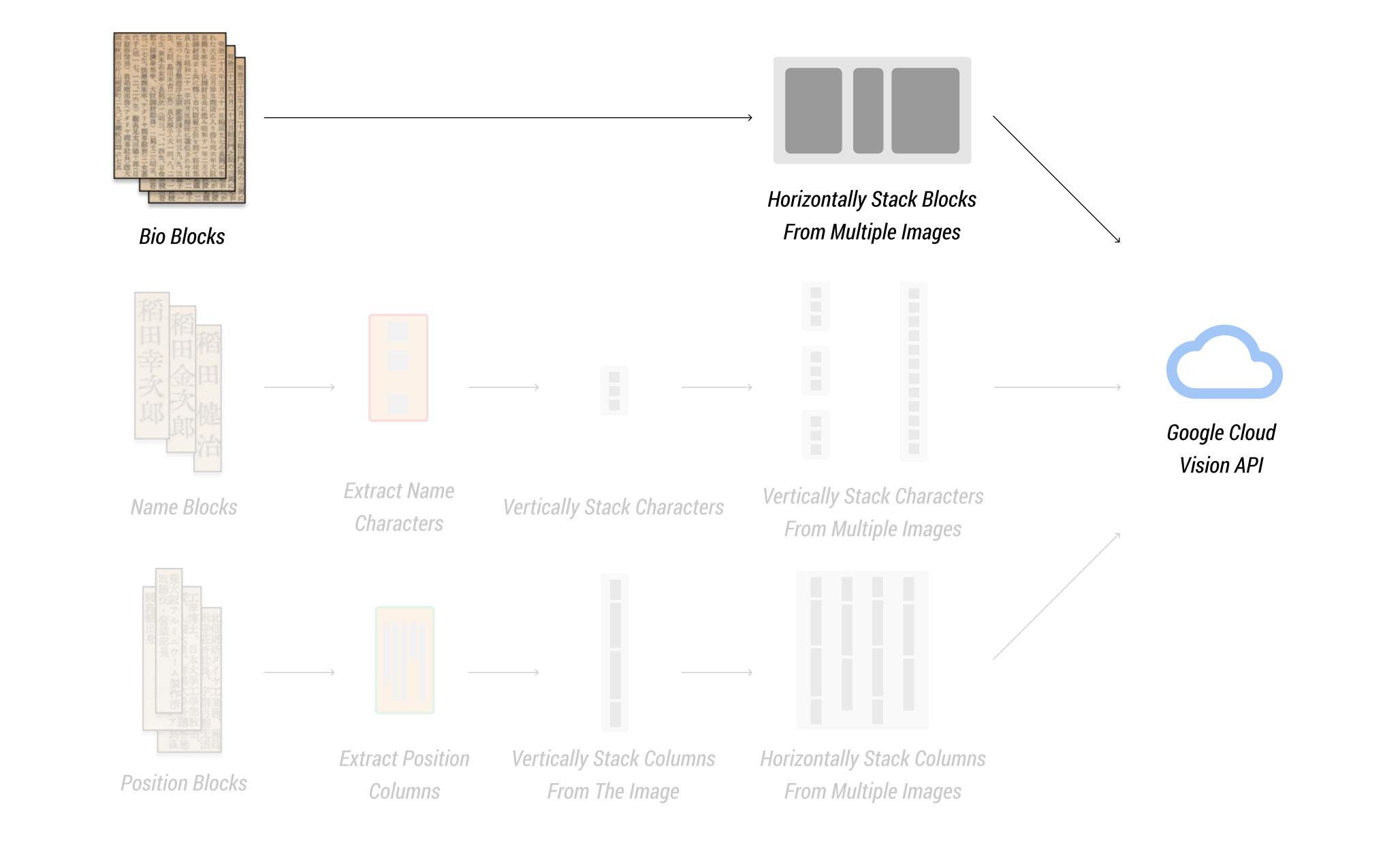


Yolo Example 2

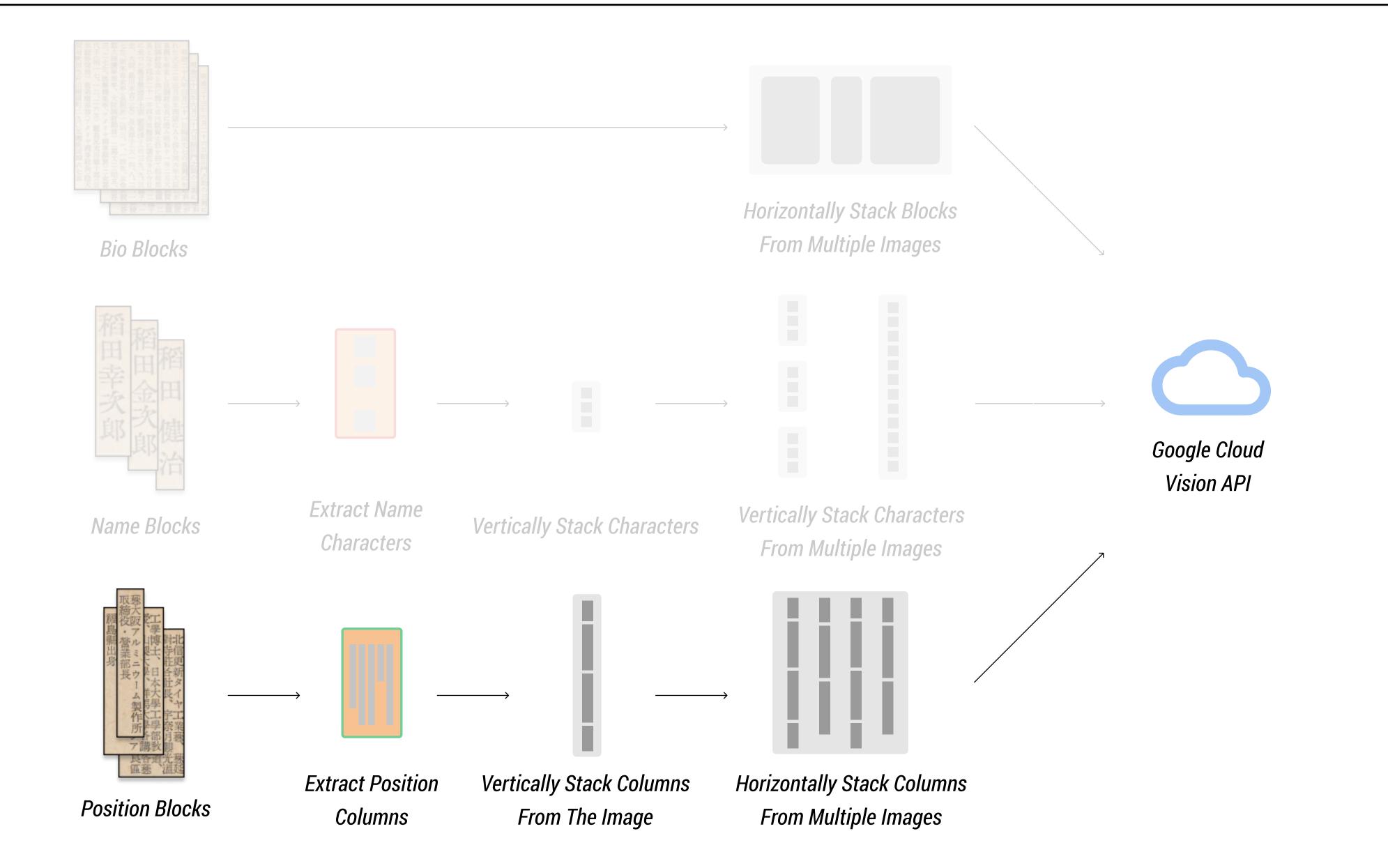
Postprocessing for OCR

- 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



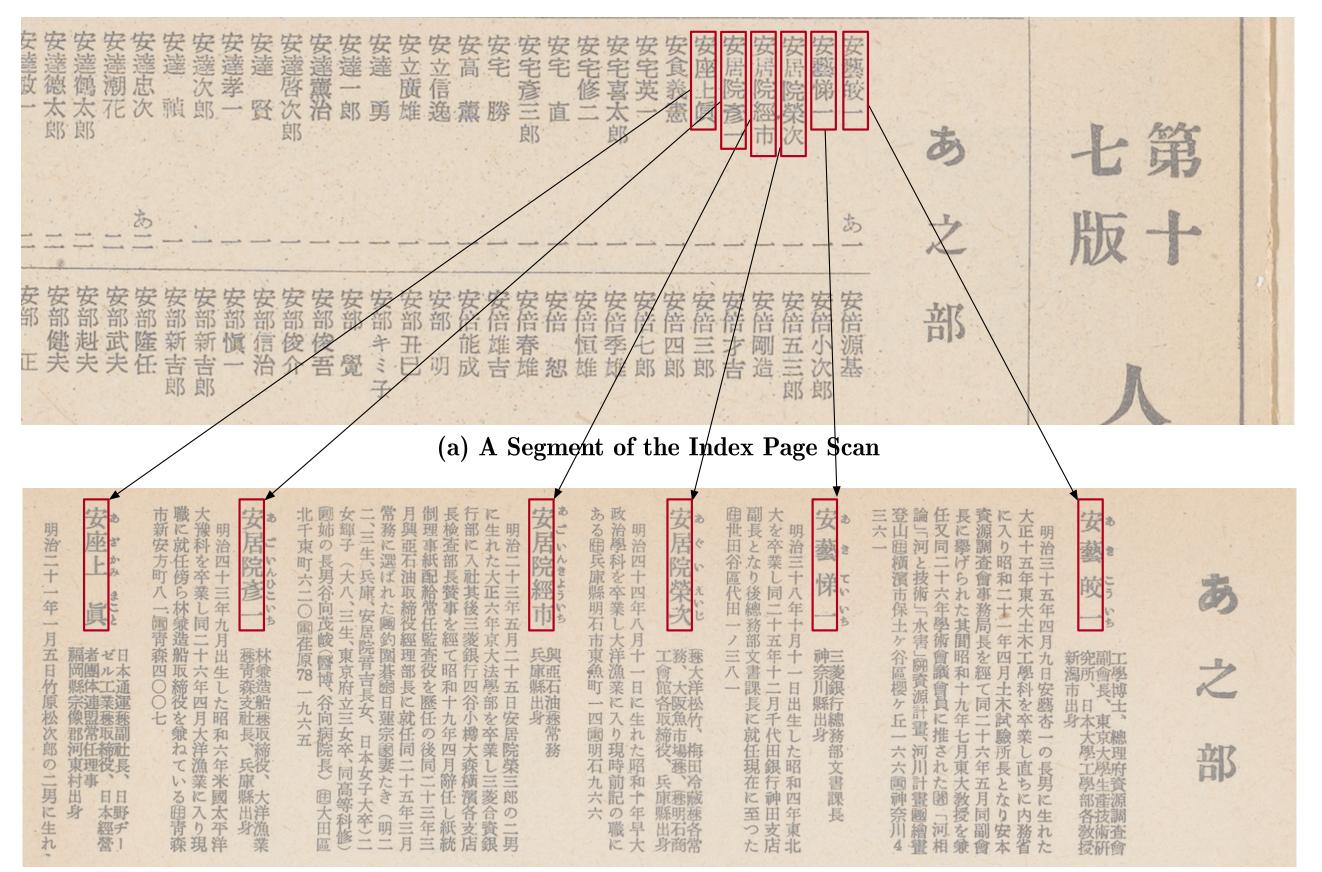






Improve the Name Detection with Index

- However, the OCR results are still sometimes inaccurate.
- We can use another source of information to improve recognition accuracy: the index.
- The names of the individuals (which are really important) appear twice:
 once in the index and in the main book.
- We can match the two sources of information and improve the accuracy.



(b) A Segment of the Main Page Scan

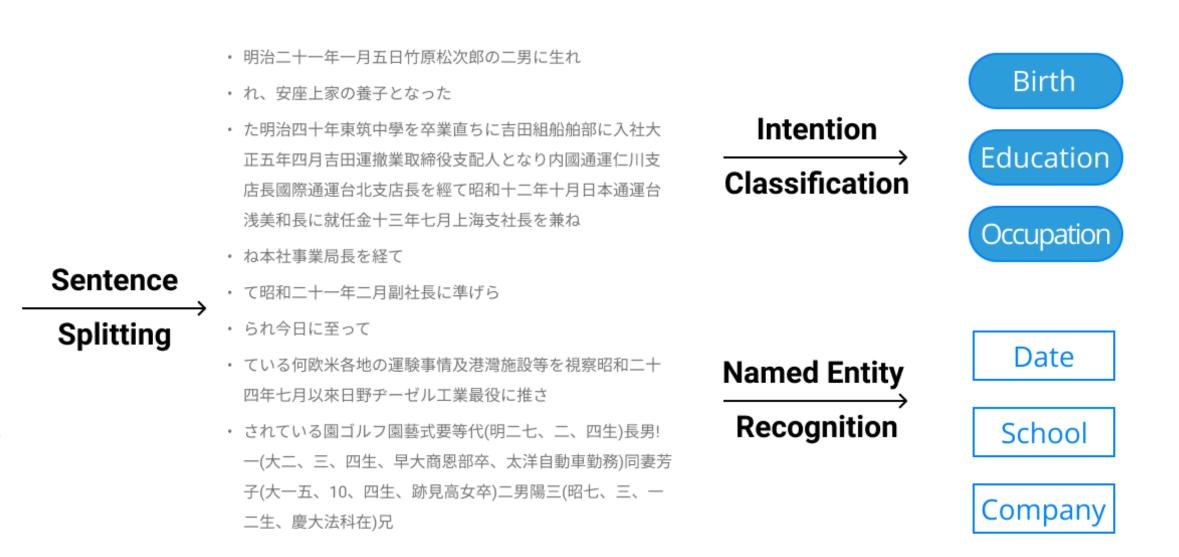
Structura Information Extraction *Work in Progress*

Structura Information Extraction



明治二十一年一月五日竹原松次郎の二男に生れ、安座上 家の養子となった明治四十年東筑中學を卒業直ちに吉田 組船舶部に入社大正五年四月吉田運撤業取締役支配人と なり内國通運仁川支店長國際通運台北支店長を經て昭和 十二年十月日本通運台浅美和長に就任金十三年七月上海 支社長を兼ね本社事業局長を経て昭和二十一年二月副社 長に準げられ今日に至っている何欧米各地の運験事情及 港灣施設等を視察昭和二十四年七月以來日野ヂーゼルエ 業最役に推されている園ゴルフ園藝式要等代(明二七、 二、四生)長男!一(大二、三、四生、早大商恩部卒、太洋 自動車勤務)同妻芳子(大一五、10、四生、跡見高女卒)二 男陽三(昭七、三、一二生、慶大法科在)兄竹原傳(元滿洲 重工取役)澁谷區永住町四六園赤坂&三10八

- 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.Al 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!