Economics 2355: Why Deep Learning?

Melissa Dell

Harvard University

January 2021

Outline

The two general approaches to automated data curation

Generally speaking, there are two distinct approaches to automated data curation:

- You can write a set of instructions that tells the computer how to process the data - by defining a series of rules
- You can let the computer learn how to process the data from empirical examples, using deep learning.

Rules have their place - and can make data curation easier and more efficient when used judiciously - but in many cases we have found that deep learning leads to more satisfactory, more usable results.

Rules are the status quo approach to document curation

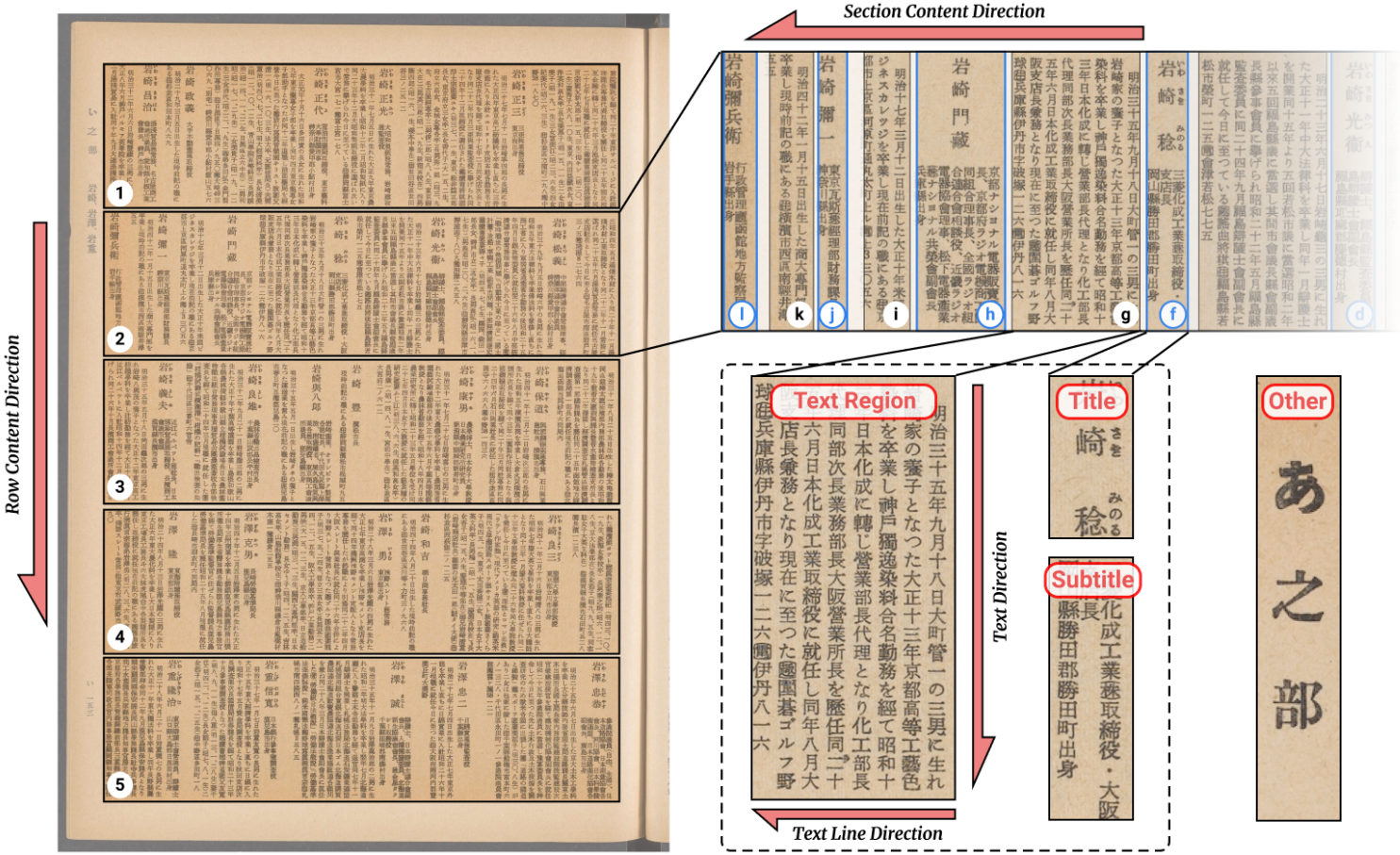
- For most researchers, myself included, the intuitive reaction to seeing a document you'd like to automatically process is to search for a set of rules.
- By rules, I mean user-defined parameters that govern how information is converted to computable format.

Rules are the quo approach to document curation

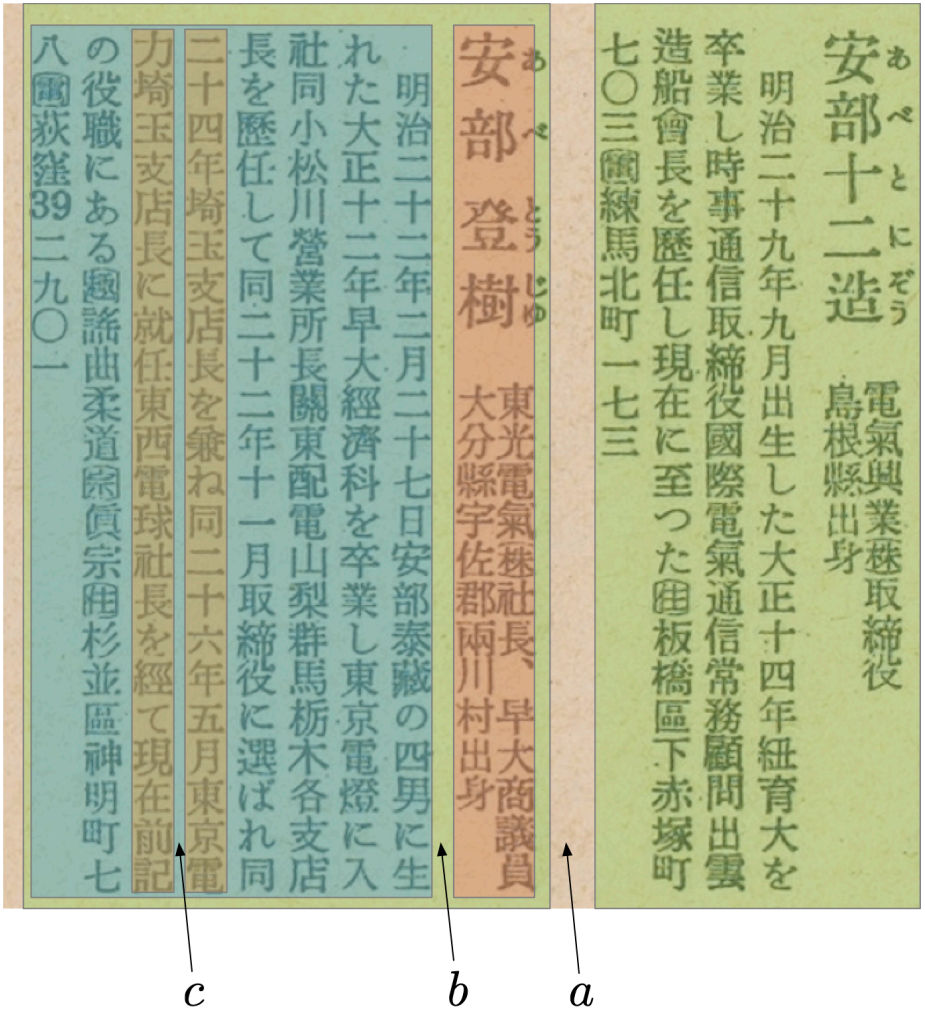
- Most of the off-the-shelf computer vision methods for document layout analysis are based on rules.
- Likewise, most tools used in post-processing implement rules.
- Methods commonly used by social scientists for text analysis, such as searching for a particular keyword, also use rules.

An Example

Consider this example of using rules to detect document layouts. The example is taken from a compendium of Japanese biographies published in 1953.



An Example


安部十二造（あべとにぞう）　電氣興業㈱取締役　島根縣出身
明治二十九年九月出生した大正十四年紐育大を卒業し時事通信取締役國際電氣通信常務顧問出雲造船會長を歴任し現在に至つた㊐板橋區下赤塚町七〇三㊐練馬北町一七三
安部登樹（あべとうじゅ）　東光電氣㈱社長、早大商議員　大分縣宇佐郡兩川村出身
明治二十二年二月二十七日安部泰藏の四男に生れた大正十二年早大經濟科を卒業し東京電燈に入社同小松川營業所長關東配電山梨群馬栃木各支店長を歴任して同二十二年十一月取締役に選ばれ同二十四年埼玉支店長を兼ね同二十六年五月東京電力埼玉支店長に就任東西電球社長を經て現在前記の役職にある㊐講道館柔道㊐眞宗㊐杉並區神明町七八㊐荻窪39二九〇一
c
b
a

A Naive Rule-Based Approach

Recall:

- a is the space between biography blocks for *different individuals*
- b is the space between the title region and the biography text region *within* an individual's biography block
- c is the space between lines within a biography text region

If it is always the case that $a > b > c$, we could segment the documents using thresholds for these parameters. In other words, the user would set parameter values for a, b, and c to segment the text accordingly.

A Naive Rule-Based Approach

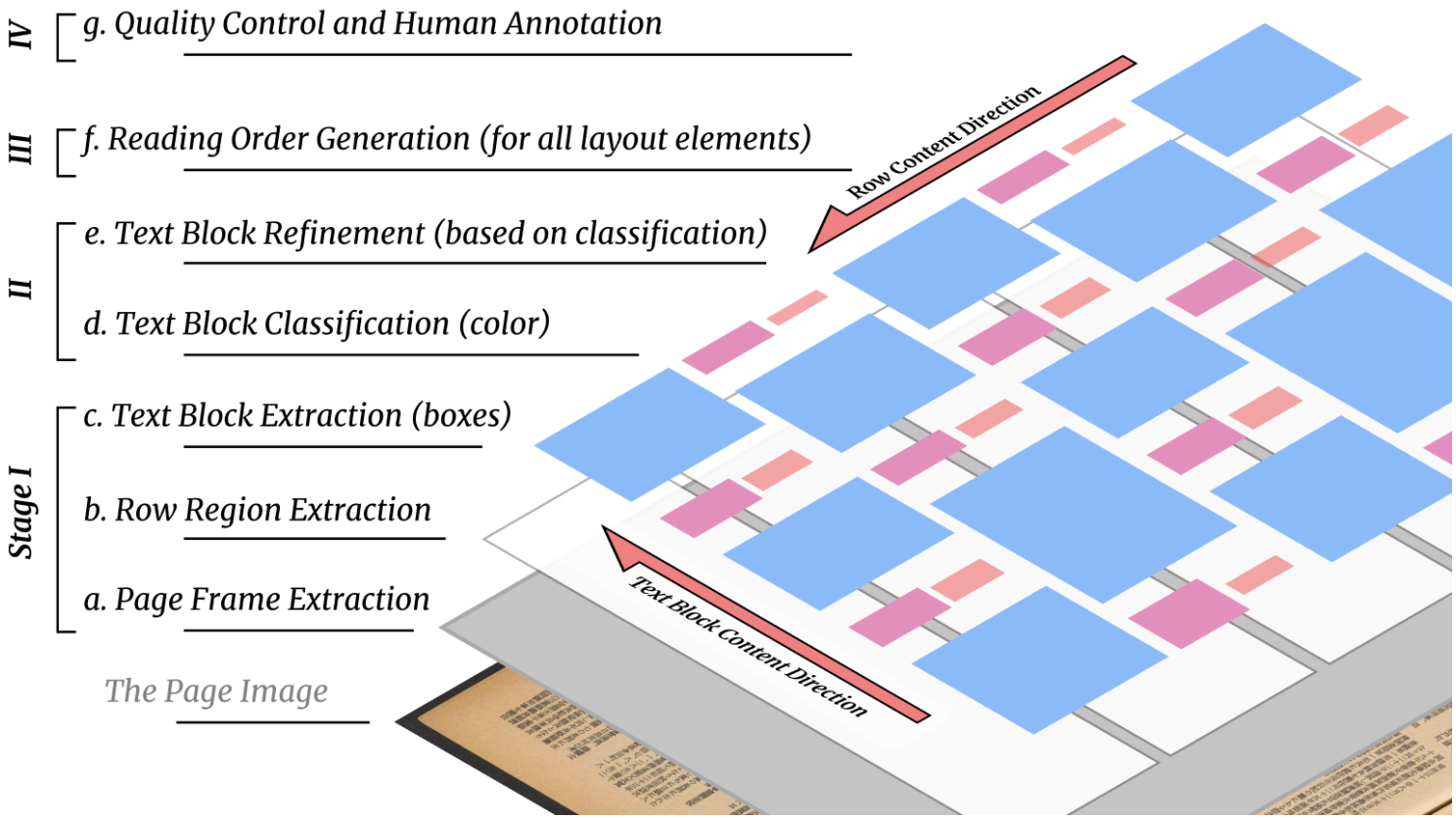
This approach has the advantage of being very straightforward to grasp. However, it also has two substantial disadvantages:

1. If it is not always the case that $a > b > c$, the approach will fail
2. It only works for a specific document. A different document would require the user to set different rules. In other words **rules don't usually scale well, because they require a lot of very domain specific user-inputs.** What you learn from one document doesn't help you to process another document.

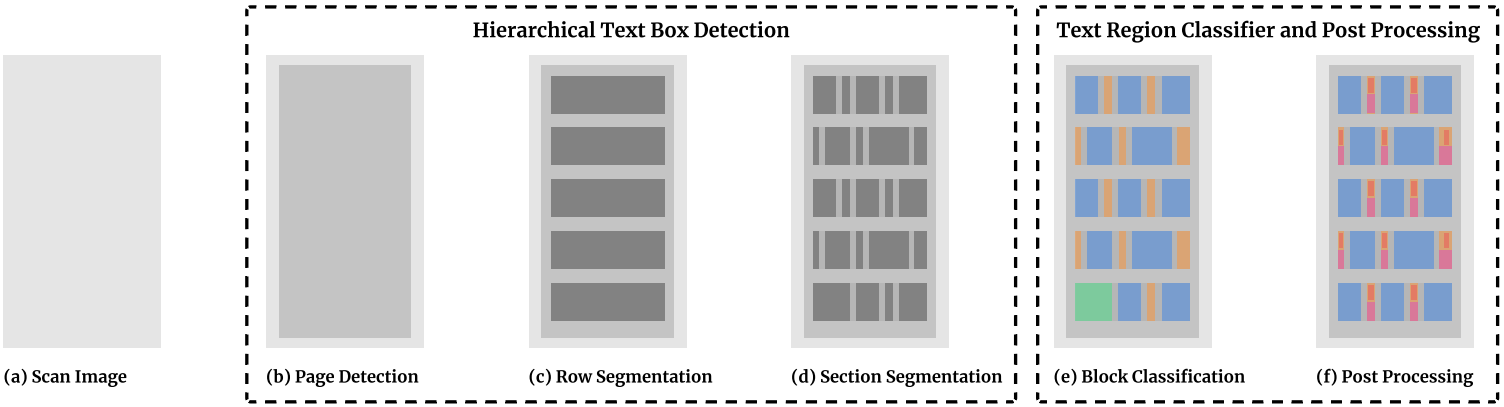
A Rule-Based Approach

- If this were a modern document made with a word processing template, it would likely be the case that $a > b > c$ for all biography blocks. Of course, if the document were made using a modern word processor - and you had that file - you wouldn't need to digitize it.
- Unfortunately, for a historical document (including the one above), more often than not there will be exceptions to the rules. In the case of the above document, there are many exceptions to rules.
- Strict adherence to document formatting rules is by and large the result of modern computing. During the mid 20th century, when the above document was created, humans manually laid out the printing press. Unlike a Microsoft Word template, **humans usually did not adhere strictly to the rules**.

A more realistic rule-based approach



A more realistic rule-based approach



Step 1: Detect the Page Frame

1. Estimate the page frame box using contour detection (groups pixels with similar visual properties like color or intensity)
2. The largest intensity contour denotes the page boundary
3. Estimate the four coordinates of the circumscribed quadrilateral for this contour as the page box
4. Convert the page image inside the quadrilateral to a rectangle with a warp affine transformation

Step 2: Segment the rows and text/title within a row

- Connected Component Labeling (CCL) and Run Length Smoothing Algorithm (RLSA) are used for splitting the five rows of contents vertically inside the page frame.
- As we apply the RSLA algorithm horizontally, each row is connected, and CCL can be applied to differentiate the rows.

Rule-based segmentation

- Binarize and invert the image
- Change white pixels to black pixels if the number of adjacent white pixels is less than c
- This links neighboring black areas that are separated by less than c pixels
- Can be applied horizontally or vertically
- In splitting the page, for example, we know to look for the four largest connect components of black pixels, which delineates the white space between columns in the inverted, binarized image

Text Region Classification

- A three-class CNN classifier is trained to identify the text, title, and wrongly-segmented regions
- If it is classified as mis-segmented, a CCL-based method is applied to split it into text and title region
- Title regions are further broken down into more refined title and subtitle segments

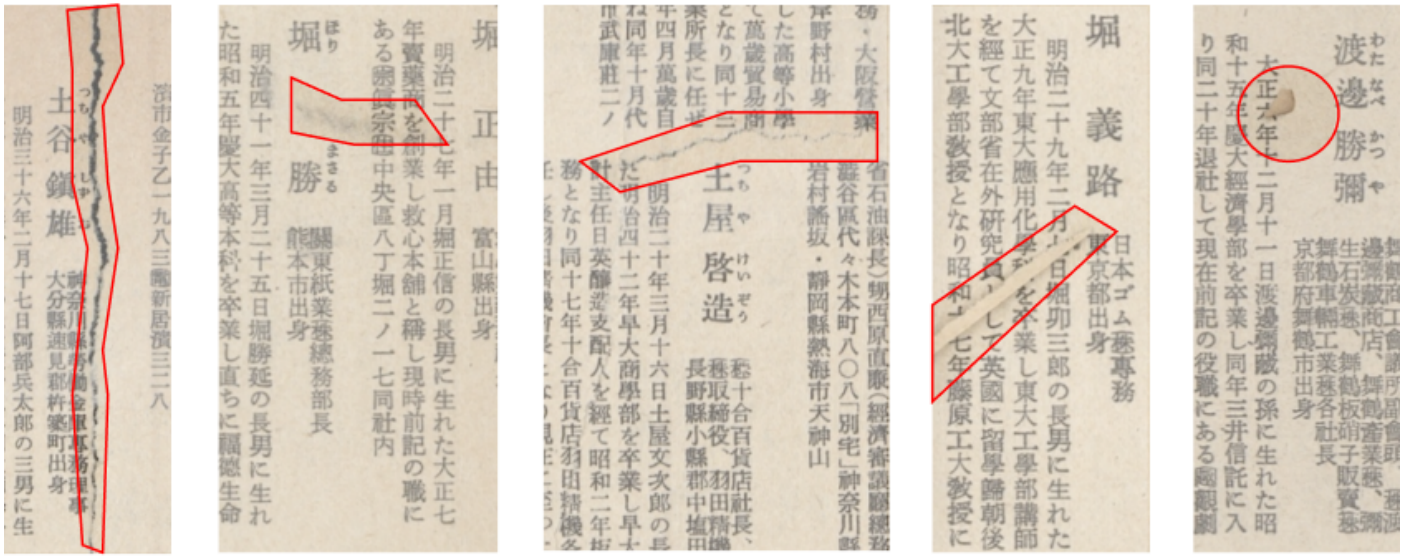
Text Region Classification

- The classifier is trained on 1,200 hand-labeled samples and tested on 100 samples.
- As mis-segmentation rarely appears (only 3 in 1000 samples), we re-balance the dataset distribution by manually creating 250 mis-segmented images.
- The input images are rescaled to the same size of 200 height and 522 width.
- We achieve a final test accuracy of 0.99.

Remaining Steps

- Reading orders need to be predicted (using rules) and errors assessed and corrected
- We examine statistics about blocks and pages. As the main pages are densely printed, we find the number of layout elements remains consistent across pages, and blocks in a row are usually evenly spaced.
- Hence, by filtering layout elements that are significantly different in these statistics, we obtain a limited number of misdetection candidates.

Noise in Page Scans



Rules v. Deep Learning

- This was a pretty good case scenario for a historical document scan - the layouts are quite clean and it was a very high quality scan (the library sent it to a company specialized in preservation scanning of historical books that removed the binding for scanning)
- Yet, there was still a fair amount of manual correction involved due to exceptions to the rules
- The motivation was to use rules on this volume to create labeled data, and then use that data to fine-tune a layout detection model that could be quickly further fine-tuned to other volumes of the same publication
- This could be an efficient way to get labels if you have one document that is fairly clean that resembles other documents you'd like to process; paper published in CVPR

Rules v. Deep Learning

- It is not obvious ex post that this was any easier than just hand annotating for deep learning, but the general idea could apply if you have a clean scan of a volume from a multi-volume publication
- This is a big part of our motivation to develop tools and resources that make implementing a deep learning-based pipeline easier

OpenCV

- Most rule-based CV methods can be implemented through OpenCV.
- If you want to do rule-based computer vision, look up the documentation for this package and go from there (lots of resources)

Another example

- In the social sciences, we also often frequently use rules for text analysis.
- For example, we might measure the number of times particular phrases appear in newspapers and use this to proxy some outcome or treatment of interest.

Another Example: NLP

- As with document layout analysis, both noise and complexity will disrupt the results of a rule-based analysis.
- Noise may be rife due to OCR errors, and these errors may take many different forms, making them difficult to filter out.
- Language complexity can also pose challenges to a rule-based approach. To the extent that there are many different ways to say the same thing, a rule-based approach like keyword search will struggle to tag the relevant information.

The Rules Trap

- You look initially at a document and think it can be processed using rules.
- However, you have not understood the full complexity/noise present in the document, and later discover that there are many exceptions to the rules that you defined.
- But you think: "oh, I could fix these exceptions with more rules." Only to later discover that there are also exceptions to the exceptions...
- Even if high accuracy is achieved at the end of the process, this approach takes a lot of user inputs (and hence user time) that are very tailored towards a specific application.
- It will not scale well, and may fall apart even on documents that are just slightly different.

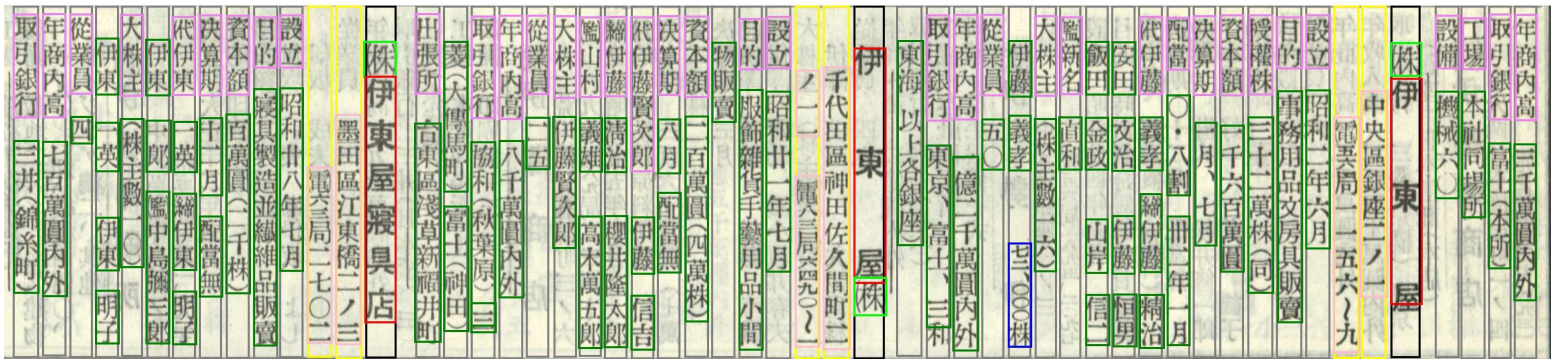
The Rules Trap

- This is not to say that the rules trap always emerges.
- Sometimes documents do follow clear, simple, exploitable rules.
- I simply urge everyone to proceed cautiously when rules are involved and to ask yourself whether it may be worth implementing a more scalable solution...

An example from historical tables

- In some cases, though, a judiciously chosen rule can distinguish information in post-processing very accurately, that would have taken quite a bit of labeled data to distinguish via deep learning.
- Here is an example from a historical publication of firm level reports

Rules



A simple, inflexible rule

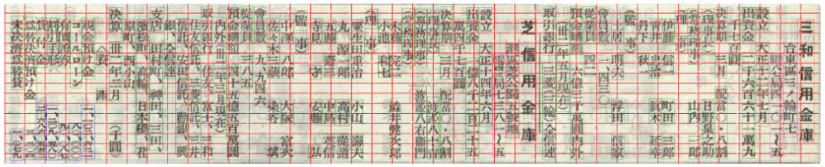
- This is an example of a steadfast rule, that through skimming images we realized holds very strongly throughout the publication, despite some noise.
- We could try to distinguish the pink boxes from the green ones simply through feeding the deep learning-based layout analysis model enough labeled data.
- Or we could implement this rule in post-processing to extract structure.

A simple, inflexible rule

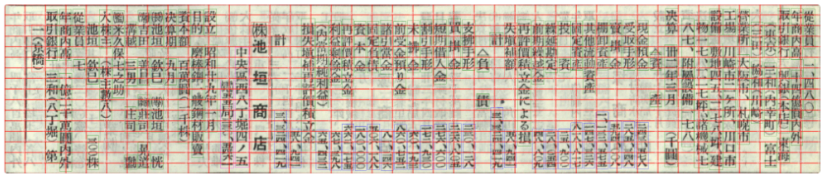
- During the layout analysis phase, both the pink and green segments are labeled as variable text. The layout model detects the coordinates of their bounding box.
- To detect variable names in post-processing, it is best to use relative coordinates, rather than hard coding values.
- Based on the aspect ratio, we divide each column into a "16 x 40" grid.
- In the text columns - even when the column images are quite skewed - variable names always start in the top row whereas variable values do not.

The layout grid

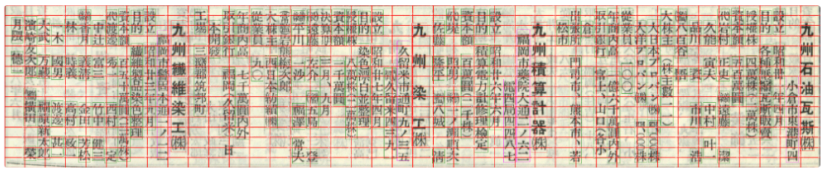
tk1957_0018_5_0-row4



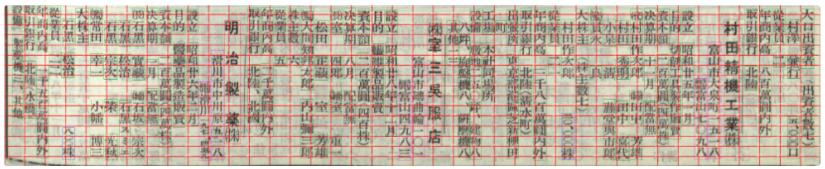
tk1957_0024_3_0-row6



tk1957_0290_5_0-row3



tk1957_0249_2_0-row7



No hard and fast rules about when to use rules

- My key takeaway is that the human brain is remarkably good at filtering out noise and heterogeneity when looking at images, on a subconscious level.
- If you have never processed images before, you may have never stopped to think about just how good we are at this.
- So when deciding to use rules, just look very carefully, to make sure the rule is really:
  1. Universal enough in the original documents that there aren't too many special cases and exceptions (and exceptions to the exceptions...)
  2. Strong enough that it won't be swamped by noise that results from document scanning, or other processes such as document aging, text bleed, etc.

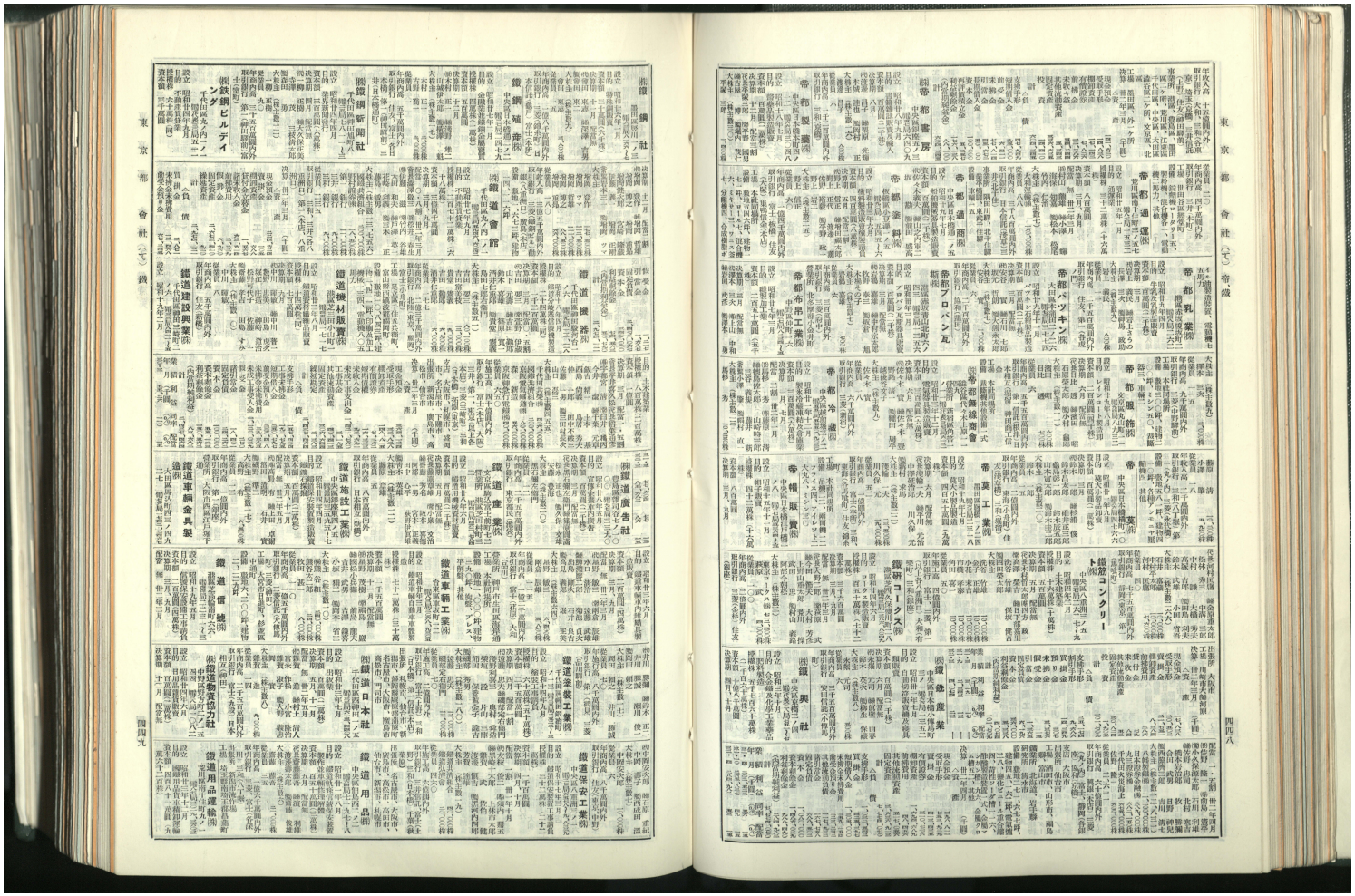
Why Rules Fail

There are two fundamental reasons why rules fail:

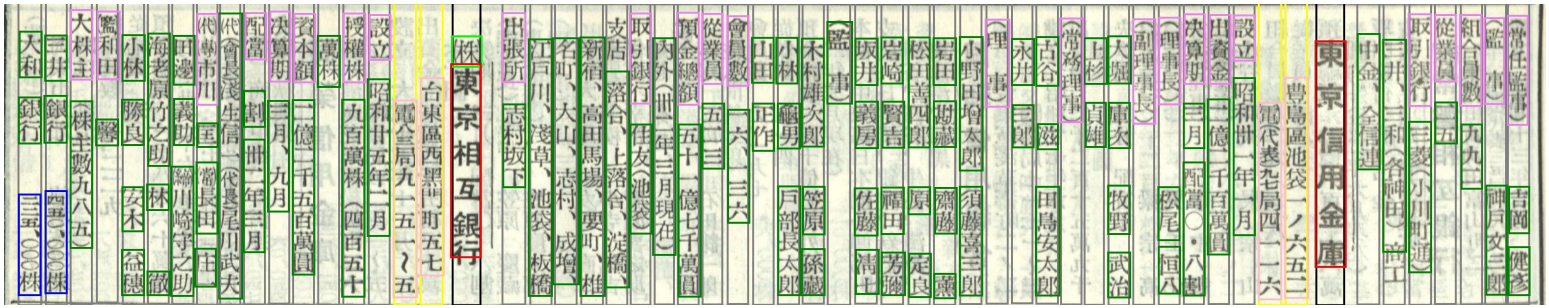
1. Complexity and rules do not mix
2. Noise and rules do not mix

In some sense, of course, these are the same thing - noise adds complexity to what would otherwise be a simple structure.

Rules failing because of complexity



Rules failing because of complexity



Rules failing because of complexity

- Such documents are orders of magnitude more complex than a single column book.
- In another project, we are detecting layouts of historical newspapers. Each paper had its own conventions, which could also vary across page types and time.
- Extracting by rules would be impossible.

Complexity is not limited to layout detection tasks

- Language is inherently complex.
- There are 60,000 words in the English language, with many ways to convey the same thing, and subtly different ways to convey very different sentiments.
- In commercial applications, modern NLP dwarfs the performance of rules like keyword search

Historical documents are inherently noisy

- Unfortunately, nice clean documents that perfectly follow templates tend to be an artifact of modern computing.
- The human brain is pretty good at filtering out this noise when looking at historical document scans (with the exception of the really old or handwritten documents, which sometimes even take a professional historian to decipher!) Rules, not so much...

Noise enters historical documents in many ways

- Historical typesetting did not allow for the precision of modern printing, meaning spacing is rarely regular throughout a book.
- Over and under inking of characters are common due to imprecision in historical printing presses.
- Old papers were often thin and led to significant text bleed. When combined with under and over inking, it can be difficult to remove with rule-based methods.
- If you don't remove the book binding to scan, there will be distortions from projecting a 3D object into 2D space
- Many scans that can be readily downloaded are poor quality, especially if binarized
- Old documents can have all sorts of other irregularities, from yellowing pages to pencil marks.

Legibility

- This makes me think of James Scott's discussion about legibility in *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*.
- To recap, Scott's brilliantly insightful thesis is that social planners look at some complex reality, fail to understand all the subtleties that make the complex order function, and then use authoritarian power to impose a radically simple, orderly, rule-based vision that in turn fails miserably because reality is more complicated.
- For example, an orderly, monocrop forest will eventually die because it lacks the biological diversity and robustness that makes the complex forests we observe in nature thrive.

Seeing Like a Researcher

- I might paraphrase this to the current application as "Seeing Like a Researcher: How Your Rule Based Scheme to Process Certain Information Will Fail."
- Unless your document truly is very simple and noiseless, a few sweeping rules will not capture its complexity.
- Deep learning is essentially about making millions of tiny adjustments to move towards a desired objective, essentially the antithesis of grand sweeping rules.

Outline

Deep Learning

- Fortunately there is now an alternative to rules: deep learning. Deep learning has the potential - across the data curation pipeline - to produce results that are more robust to both noise and complexity.
- It has already revolutionized the world by replacing rule-based approaches in many domains, none more so than in image, audio, video, and language processing.
- All these are domains with numerous applications to curating social science data.

Deep learning versus rules

The *Nature* review article on the reading list by Yann LeCun, Yoshua Bengio and Geoffrey Hinton - termed the "godfathers of deep learning" - provides a succinct summary:

> *Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level... The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.*

Deep learning requirements

1. A **lot** of labeled data
2. A **lot** of computational power
3. The appropriate architecture

Deep learning requirements

- As social scientists, we don't have to solve problems 2) and 3). Cloud services put previously unimaginable computing power at our fingertips (or if deep learning is integral to your work, it may be more economical to buy your own GPU, whose price has fallen dramatically in recent years).
- The underlying architectures have already been studied extensively by computer scientists.
- Very smart people at organizations like Facebook Research or Huggingface have produced open source code that works remarkably well.

Deep learning requirements

- However, having enough data to train models geared towards our incredibly diverse applications is the major stumbling block to automated data curation.
- Still, we're not starting from scratch. Orgs like Google and Facebook have pre-trained NLP and object detection models on massive amounts of text or images using monumental amounts of compute. But these models need to be fine-tuned to our contexts.
- We will discuss how to make creating labeled data more efficient and feasible. The course also aims to provide enough knowledge to be able to set up the problem and fine-tune models to your own applications.

A restatement

In discussing at a conference why deep learning did not take off in the 1980s, Geoffrey Hinton put it bluntly:

1. Our labeled datasets were thousands of times too small
2. Our computers were millions of times too slow
3. We initialized the weights in a stupid way
4. We used the wrong type of non-linearity

Approaches when labeled data are limited

- **Use an off-the-shelf pre-trained model without additional fine tuning.** This may occasionally work, but at present deep learning models have rarely been applied to social science data curation, so usually you can't just run the model off-the-shelf and expect good results. This is particularly true when it comes to layout analysis.
- **Annotate samples by hand, using an active learning framework geared towards social science documents**
- **Use data augmentation to generate samples that simulate the data you want curate**, or some form of self-supervised learning. The existing methods don't necessarily translate well to the types of data that we are likely to need to simulate.

Outline

I have papers to write and rule-based approaches seem easier. Does the noise from rule-based approaches really matter?

- It depends. How much and what sorts of noise are introduced? How could these types of noise bias downstream analyses? There is no one size fits all approach to setting an acceptable threshold for noise.
- Getting things right upstream can often save you a lot of time in post-processing downstream and reduce the probability that your analyses produce noisy mush.
- Data curation can be incredibly time consuming, regardless of the method pursued, but it can be worth it if it helps you to answer important questions and do research that is meaningful to you.

Easier isn't obvious

- There are fewer startup costs but it may cost you time down the road.
- It may come down to spending time manually correcting the problems of rule-based approaches, versus making the investments to understand and implement the deep learning based approaches, and ultimately facing a different ceiling on what can be achieved.
- There is no escape from manual tediousness in empirical research. Deep learning will often require you to create labeled data, which can be the very definition of tedious. However, by judiciously using the method most suited to the task - and thinking carefully about how to do it efficiently - the tediousness will be less than it would otherwise be.