Economics 2355: Introduction

Melissa Dell

Harvard University

January 2021

Outline

The Importance of Data Curation

- This course examines how to use revolutionary methods in deep learning to curate data at scale
- Data curation is central to quantitative social science because it fundamentally shapes the questions that we think to pose, as well as the questions that are feasible to answer.
- As a relatively young field, empirical social science has barely begun to scratch the surface in terms of illuminating the questions that are most important to people's lives and to the well-functioning of societies.

Economics Before Modern Computing

- While social scientists have of course long been disciplined by empirical observation, empirical analysis in the way the term is commonly used today requires computing power.
- Back in the days when simple regressions took many hours to run on mainframe computers, by and large the only information that could serve as inputs for empirical analyses were macro data points.
- Hence, quantitative researchers spent much of their time thinking about the sliver of questions that could be answered with highly aggregated data.
- Questions that these data could not answer were largely considered outside the realm of quantitative social science, left to qualitative analysis or ignored altogether, no matter how important they were to people's lives. We simply didn't have much to say.

The First Empirical Revolution

- During the 1990s, the massive growth in the power of personal computing revolutionized the quantitative social sciences, unleashing the first empirical revolution.
- Personal computing changed the types of data that could be collected and the types of data academics could process.
- This in turn fundamentally influenced the questions that we asked. Nowhere was this more true than in economics.

The First Empirical Revolution

- At the eve of the first empirical revolution, economics was primarily theoretical in its orientation. Empirical perspectives emphasized macroeconomic, correlational facts.
- The aggregated nature of the comparisons that tended to be made no doubt were a turn off to many in the qualitative social sciences, who often focused on questions that relied on disaggregated, local information to answer.
- The computing revolution made possible the creation of a great deal of knowledge about the micro foundations of important economic phenomena. Today, even macroeconomists make very extensive use of firm level data to understand the microeconomic origins of macroeconomic phenomena.

The First Empirical Revolution

- Causal identification moreover marched hand in hand with advances in computing, as more disaggregated data - as well as lower costs for collecting data - unleashed many more opportunities for isolating causal effects.
- This happened through having the computing power needed to analyze data from natural experiments, but also because computing facilitated the collection of new data, through efforts like household surveys and randomized control trials.
- It is hard to imagine the field of development economics developing the insights that it has without the advances in computing that spurred the first empirical revolution.

We are now experiencing the beginnings of a second empirical revolution

- As with the empirical revolution of 25 years ago, the current empirical revolution has been spurred by monumental advances in computing.
- In particular, these advances have made deep learning based methods feasible.
- Deep learning offers many potential applications to social science.

Raw data in social science

- Raw information in the quantitative social sciences often takes the form of image scans of historical documents: i.e. scans of tables, firm level reports, government documents, newspapers, directories, etc. The scans are often of poor quality, and the layouts of the documents can be highly complex.
- It is also oftentimes contained in reams of text.
- It can also be contained in photographs, videos, audio files, satellite imagery, and a variety of other formats.

The Promise of the Second Empirical Revolution

- Deep learning has the potential to unlock traditional types of data on a large scale, in contexts where manual digitization is unfeasible.
- It will also change the types of information that can be converted into computable data, a prerequisite for use in empirical analyses.
- For example, take the old adage that a picture is worth a thousand words. Historians talk all the time about iconic images that appeared in media. Deep learning based methods allow us to track the dissemination of photographs across millions of pages of historical media, tag what specific images contain, etc.

The Promise of the Second Empirical Revolution

- Systematically understanding fundamental questions about how societies change and grow over time requires highly granular data on the trajectories of individuals, firms, and communities.
- Many important questions are nearly impossible to shed light on with aggregate data, which often aggregates away the relevant variation.
- Rich disaggregated data - stretching back decades and even centuries - do exist. However, they are often either scattered throughout reams of text or trapped in images or hard copy documents. Manually converting these sources into computable data can be prohibitively costly.

The Methods Exist but Need to be Tailored to Our Use Cases

- Computing and the general methods to curate data at scale already exist.
- However, without fine tuning them to our applications - which requires methodological innovations, tailored interfaces, and an active research community - these approaches will tend to fail dismally.
- If we limit ourselves to the methods developed by other fields for data curation, we will also severely limit the questions that we can ask.

Outline

The Data Curation Pipeline

In general, the data curation pipeline has the following steps:

1. Detect content regions using document layout analysis
2. Optical character recognition
3. Post-processing and database assembly
4. Convert information to computable format

While not all projects will require all steps, many will require most of them, especially when working with historical data.

Step 1: detecting document structures

- In order to convert hard copy publications or image scans into data that can be used in quantitative analyses to elucidate important questions, just digitizing the words and numbers contained in the document scans is not sufficient.
- We also need to extract the text **structures.**
- For example, table headers, rows, columns, and footnotes in the case of a financial publication or headlines, articles, images, advertisements, captions, etc. in the case of historical newspapers.

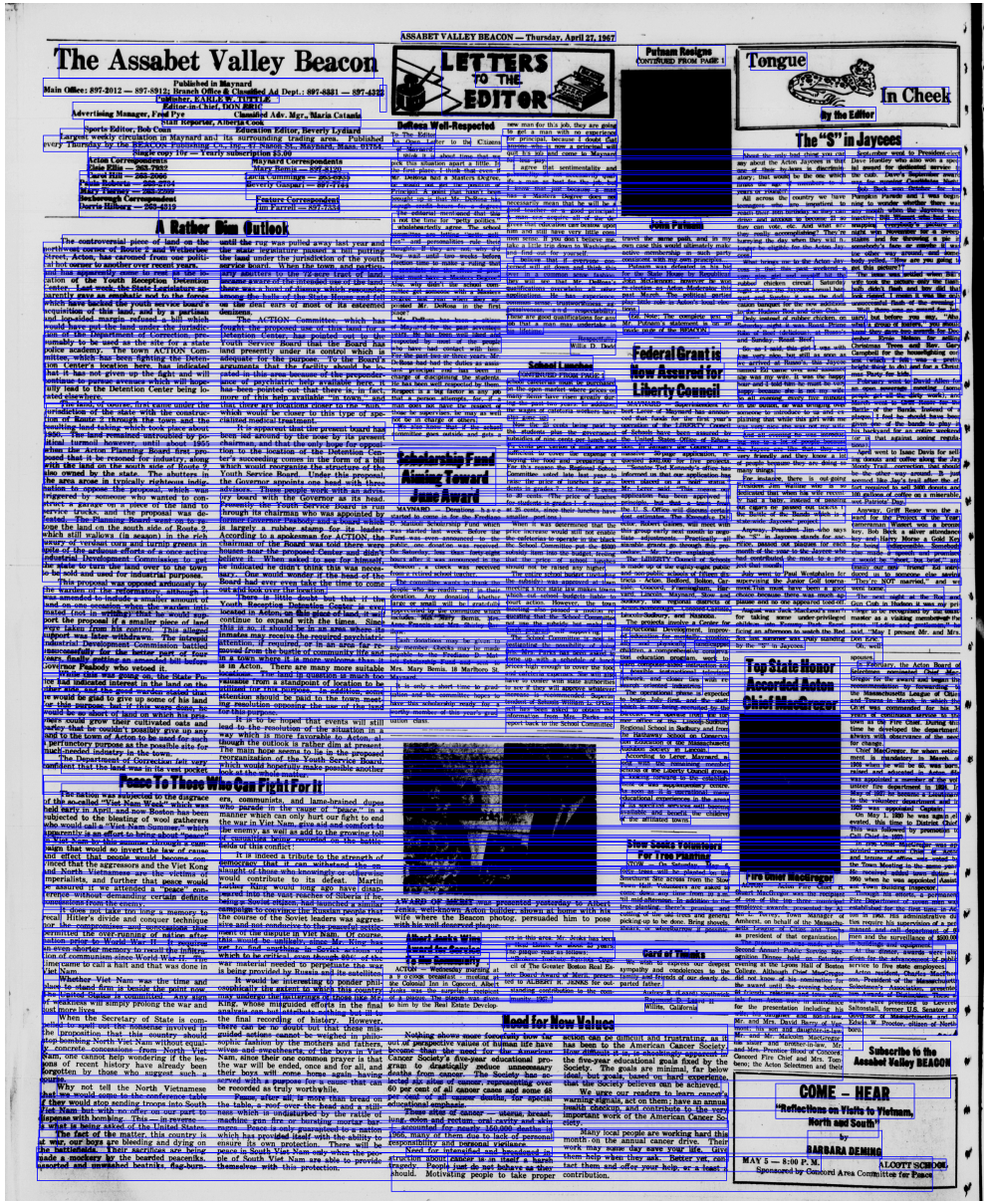
Step 1: detecting document structures

- Unfortunately, off-the-shelf tools are rarely capable of auto-detecting text structures when the typesetting is complicated, as is the case in many social science documents.
- Moreover, they do not utilize the layouts as a hint for estimating the texts. In other words, they are incapable of **layout understanding**.
- Commercial OCR softwares have been optimized for single column books and perform poorly on texts – especially noisy historical scans – that have complex layouts.

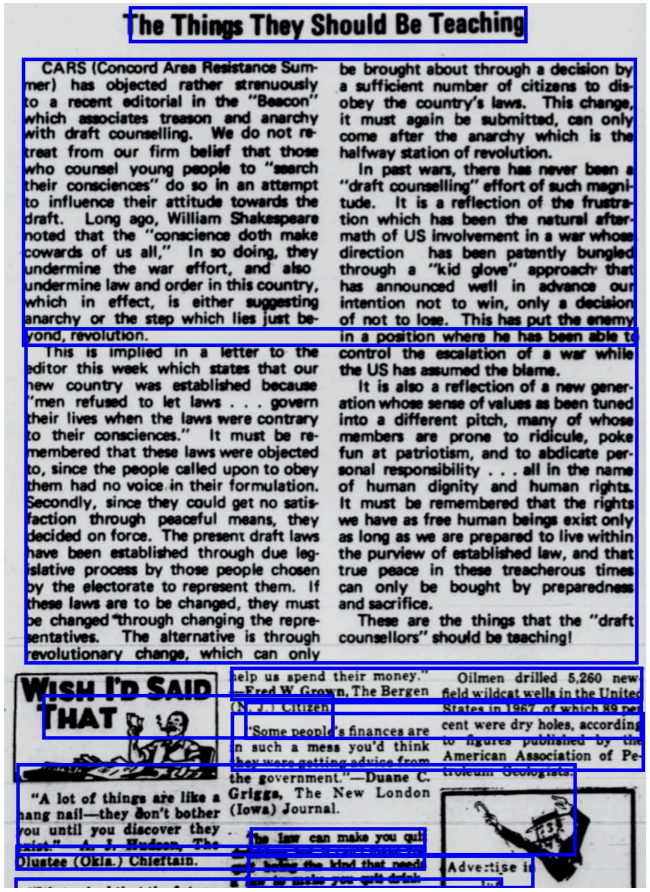
OCR Examples

- Here are a couple of typical examples of output from Google Cloud Vision (GCV), the leading commercial OCR software, on histrical newspaper scans.
- The blue boxes draw the GCV-detected paragraph bounding boxes.
- In the first image, GCV wrongly combines text from seven different columns, reading it from left to right as if it were a single column book. It fails to detect some text altogether, and it cannot distinguish between headlines, articles, and advertisements.

Example Layout Detection



Example Layout Detection



GCV wrongly combines the two columns at the top of the image, and cannot identify the paragraph breaks. For the three-column region at the bottom, GCV's bounding box detection results are totally disrupted.

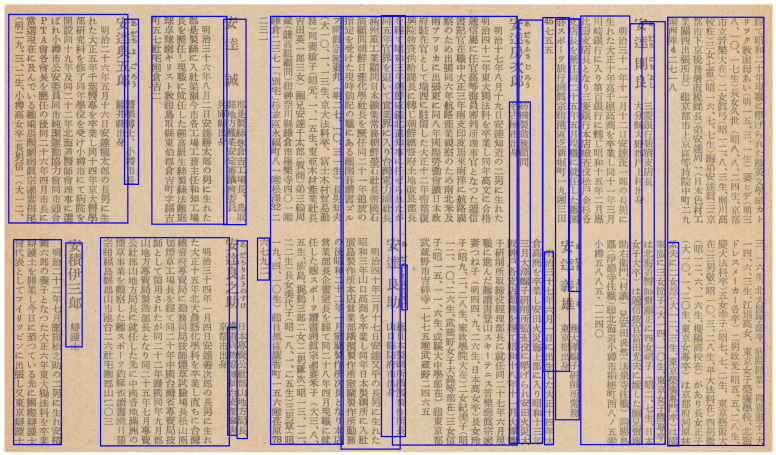
This matters for the questions that the literature tackles

- Existing historical newspaper datasets hence typically limit the user to analyzing the individual words present in a newspaper page scan and do not allow for analyses that take as inputs headlines, captions, sentences, paragraphs, or full articles.
- The limitations of the available data make it difficult to capture sentiment, to tag topics that use complex language, and to trace the dissemination of content - often abridged or reproduced without bylines - through national news markets.

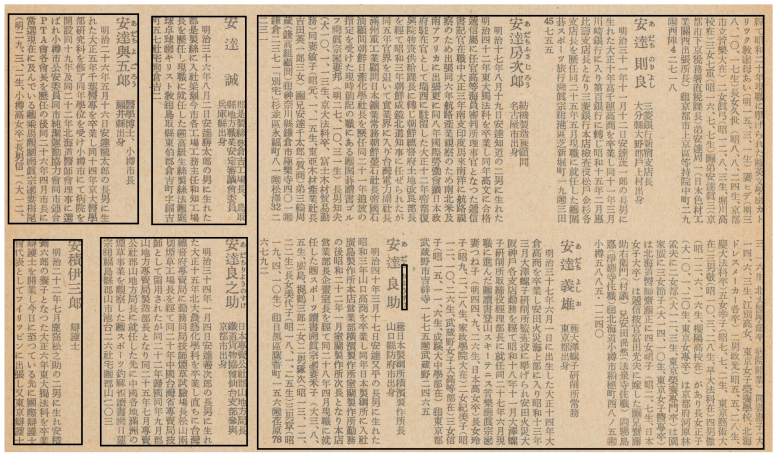
Another Example

- Figure 3 provides an additional example from Japanese biographical documents: this page is from a 1953 volume that contains detailed biographies for over 50,000 prominent Japanese citizens.
- Each biography contains a header for the person's name, a sub-header for their position, and a biography block.
- The paragraph detection blocks from GCV are unusable.

Example Layout Detection



(a) Blue boxes are GCV Detected Paragraph Blocks



(b) Black boxes are GCV Detected Text Blocks

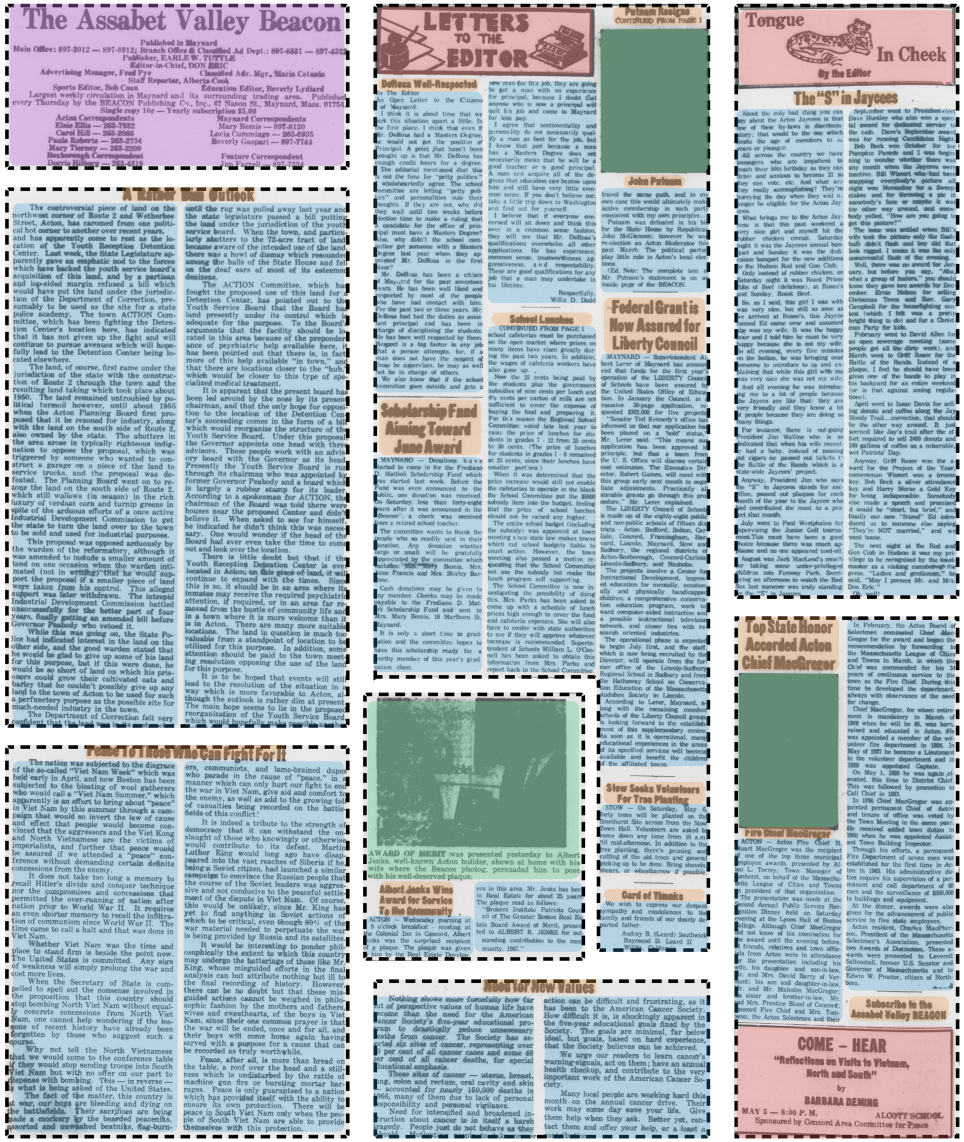
The Importance of Layout Analysis

- Importantly, the failure to detect layouts also leads to a (oftentimes substantial) deterioration in the OCR quality of individual words, with some texts failing to be detected altogether.
- While OCR software allows the user to manually draw bounding boxes on each page to specify the layout, this is an expensive, labor intensive process that has not been performed on the vast majority of the hundreds of millions of existing scans of historical documents - such as the newspapers above - that researchers might wish to use in their work.

Detecting document classes

- Beyond accurately separating text regions, we would also like tools that can automatically classify the **different types** of text regions.
- Documents that are of interest to quantitative researchers typically contain a variety of content regions. For example, a table will at a minimum contain a title, column headers, row headers, cell values, and a caption. Tables can also be significantly more complex, as is the case with historical firm financial reports that we have been curating.
- Alternatively, consider historical print newspapers, which will often contain a page header, headline, article text, images, captions, advertisements, etc.

Document Layout Example



Section Block Heading Title Column Box Figure Advertisement/Graphic Heading

Object detection for document layout analysis

- The first step to converting such documents into a format that can be analyzed is to detect content regions through document layout analysis.
- Each content region is defined by its **coordinates** as well as a **class** (i.e. headline, article, image, etc.)
- This is fundamentally an object detection task, just as detecting cats in a photograph is an object detection task.
- Object detection models can be fine-tuned to our contexts.

Step 2: Optical character recognition

- If the raw data are document image scans, after detecting the content regions you will need to send them for optical character recognition (OCR).
- It may be necessary to optimally stack individual layout regions to achieve the best results, as complex or non-standard layouts often lead to a deterioration in OCR performance.
- Image pre-processing can also significantly enhance OCR accuracy.
- If off-the-shelf solutions do not achieve an acceptable level of accuracy, it may be necessary to design your own character recognition engine.

Step 3: post-processing and database assembly

- Typically, some amount of additional post-processing and database assembly will be necessary before proceeding to further data analysis.
- In tables, you need to know, for example, which cells are associated with specific rows or columns.
- If text is wrapped within the table, you need to combine cells together in the appropriate order.
- Different types of tables in the same document may associate text differently, underscoring the importance of detecting not just individual layout elements but also different layout sections contained in the document.
- One would also want to check that rows or columns add up to the totals provided, and if not manually correct the discrepancies against the original documents.

Newspaper Post-Processing

- In the case of historical newspaper data, you need to associate headlines with articles and images within their captions. This is often (though not always) feasible solely with the layout coordinates and classes detected by document layout analysis (step 1).
- Moreover, articles may be wrapped in complex ways across columns, or appear on multiple pages, and these article blocks may need to be associated with each other to create full article texts. Here, NLP comes into play.

Newspaper Post-Processing

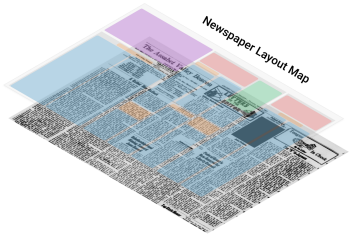
- The OCR engine typically makes many punctuation errors, interchanging periods and commas, missing punctuation altogether, or misdetecting background noise as punctuation.
- This will wreak havoc on downstream NLP tasks, which require coherent sentence structures.
- Fortunately, post-processing can help a lot in removing punctuation errors.

Step 4: Converting information into computable formats

- If you are working with a numerical table, you may be done after post-processing. However, many types of data require additional processing to convert to a computable format that can be analyzed statistically.
- Suppose you are working with newspaper articles. An article text is not computable. Instead, natural language processing tools need to be used to embed the article text - which is high dimensional and sparse - into a denser, lower dimensional object suitable for downstream analyses.
- These downstream tasks include retrieval, topic classification, and sentiment analysis, which will all receive extensive attention in the course.
- As with document layout analysis, existing state-of-the-art models can be fine-tuned to perform well on these tasks.

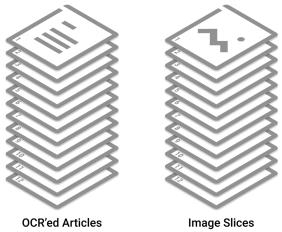
An Example Pipeline

Here is the pipeline for a project that I am working on with historical newspapers:



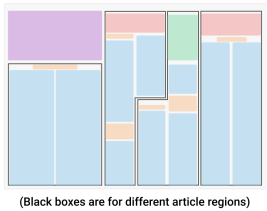
1. Newspaper Layouts Analysis

A layout mask will be extracted for each scan denoting the content segments and type.



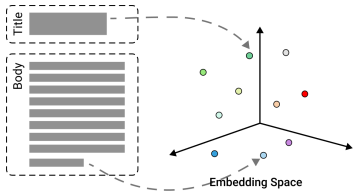
2. Content Digitization

For each text region, we OCR the text. And we store non-text image slices.



3. Reading Order Prediction

With a combination of textual and positional information, we associate invididual text regions and construct the articles.



4. Embeddings And Topic Generation

We use BERT to project texts into embedding vectors and compute their topics and sentiment.

General philosophy

In general, there are three methodological approaches that could underlie the above pipeline:

1. **Manual labor**: Often, manual labor cannot be avoided. But obviously our aim is to do as little manual labor as possible.
2. **Rule based methods**: These are effectively the status quo in computer vision based methods for document layout analysis, as well as in post-processing. Many social scientists often use rules (such as keyword search) for NLP.
3. **Deep learning:** Deep learning can underlie every step in the above analysis.

Automated Approaches

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.

Rule-based approaches

- Rule-based approaches have the advantage of being easy to understand. Most of us are used to interacting with computers via rule-based approaches.
- They certainly have their place, but we have found that they often perform poorly on historical documents and NLP alike.
- Complexity and noise are the enemy of rules, and social science documents tend to be rife with complexity and noise. Hence, as a rule (no pun intended), I am not a huge fan of rules.
- Next lecture will discuss various examples of rule-based approaches and why the results tend to be disappointing relative to deep learning.

The power of deep learning

Deep learning is a powerful approach that can play a central role in every step of the data curation pipeline. For example, in the newspaper pipeline shown above, deep learning:

- Underlies the layout analysis and image pre-processing
- Underlies the OCR
- Underlies assembling structured data (i.e. predicting the order of article text segments) and post-processing (i.e. to fix OCR punctuation errors)
- Underlies natural language processing

The power of deep learning

- As you can see, deep learning is incredibly versatile.
- There are many parallels between the methods underlying the detection of content regions, OCR, and natural language processing, even though to someone unaccustomed to the methods they may seem like very distinct problems.

Outline

Isn't there an app that will do this?

- When I tell people about what I've been working on lately, the most common reaction is "but isn't there an app (or other commercial product) that does this?"
- Unfortunately, there are not off-the-shelf solutions for our data curation aims (in either computer vision or NLP), and in fact existing commercial solutions don't get anywhere close to acceptable accuracy.

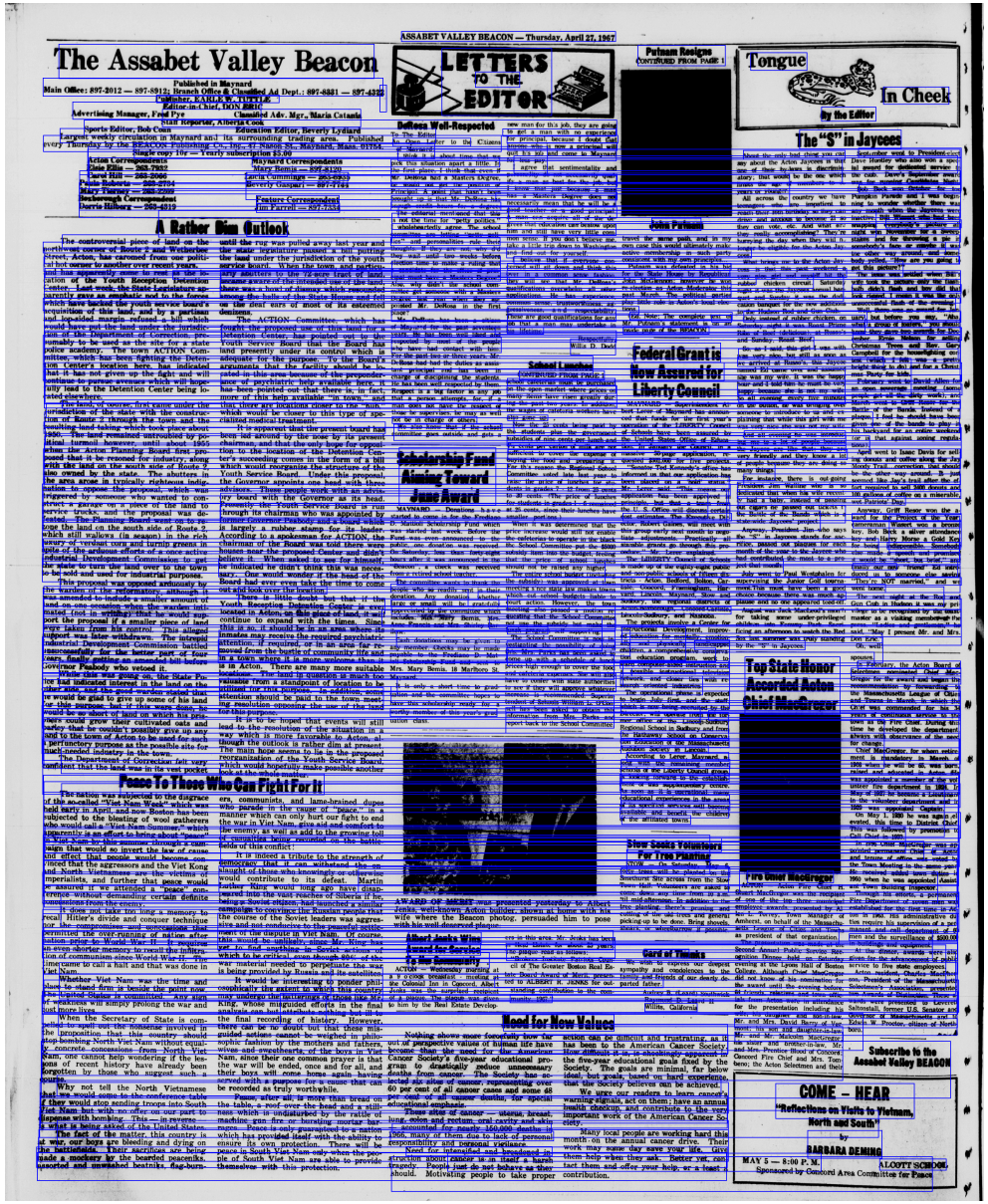
Commercial OCR softwares fail on complex layouts

- There are many challenges to converting information into computable data.
- The first and foremost challenge that we have faced is that commercial OCR softwares and other off-the-shelf models are incapable of auto-detecting the complex layouts that tend to characterize most quantitative historical documents.

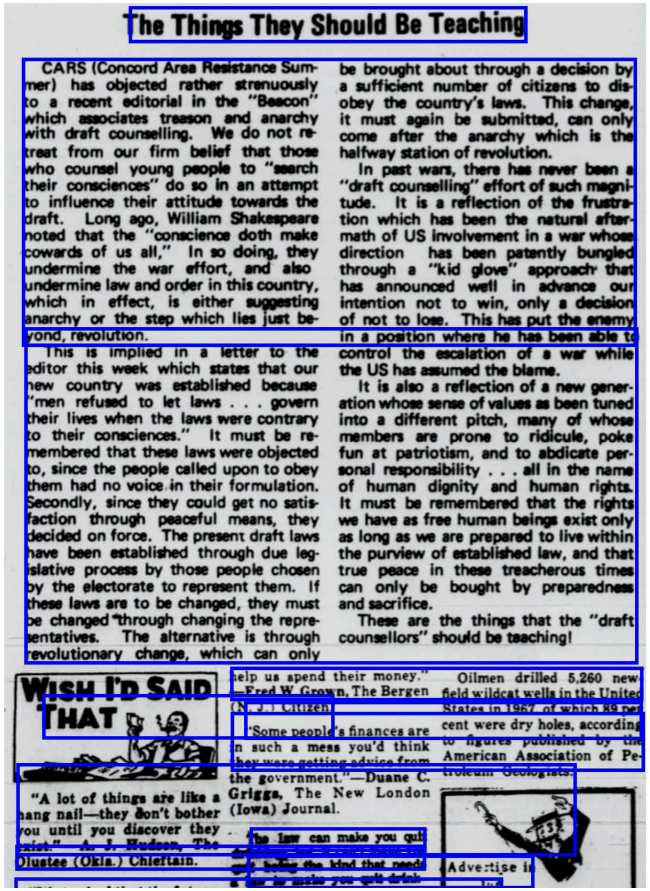
Commercial OCR softwares fail on complex layouts

- Commercial OCR softwares are primarily trained on **clean, modern documents with simple layouts**, like single column books.
- In contrast, we are interested primarily in noisy, historical documents, many of which have highly complex layouts.
- We are a long ways from achieving Artificial General Intelligence (AGI). If softwares haven't been trained on documents that look like the documents you wish to process, they will perform poorly.

Example Layout Detection



Example Layout Detection



Commercial OCR Solutions

- We did an exhaustive testing of off-the-shelf tools on our documents and found Google Cloud Vision was the best out there.
- Other off-the-shelf tools perform as poorly, or worse. For example, PubLayNet, the state-of-the-art pre-trained open source model for layout understanding, completely fails to detect the layouts of historical newspaper scans.
- Right now, document layout analysis relies heavily on supervised learning, and there simply are not labeled datasets that are representative of the vast diversity of documents that are of interest to social scientists.
- Hence, an off-the-shelf solution isn't going to work in general.

The failure to detect layouts leads to a deterioration in OCR quality

- The failure to detect layouts also leads to a deterioration in the OCR quality of individual words, with some text failing to be detected altogether.
- Many statistical tables have even more complex layouts than the newspapers shown above, and we find that often entire regions of the table are undetected, even when glaringly obvious to the human eye.
- When they do detect, characters from clearly distinct sections of the tables often end up scrambled together.

Why do off-the-shelf solutions perform poorly

Let's delve a bit more into why these tools are poorly suited towards our documents, as it highlights the innovations that need to be made to ensure better performance. Generally speaking, there are several features of our documents that are important:

1. Complex layouts, with sparsity
2. Noisy backgrounds
3. Different/noisy fonts
4. Noisy scan technology

Existing commercial products are trained on dense documents with simple layouts

- The deep learning based methods that underlie commercial OCR engines are data hungry.
- They can only learn about data that they have been exposed to. These products are trained on corpuses consisting primarily of modern documents with dense, clean layouts.
- There are some labeled datasets for historical documents, but these datasets are quite small, as annotating layouts on historical documents is extremely labor intensive.

Commercial solutions are trained on documents with clean, modern fonts

- Again, it is worth reiterating that the computer can only recognize characters that are reasonably similar to those that is has been exposed to in training.
- There are some historical fonts that are not present in modern document corpuses.
- Even if similar fonts are present today, historical printing is a much noisier technology than modern printing, with stroke widths varying depending on the amount of ink left in the printing press.

Commercial solutions are trained on documents with clean, modern fonts

- For example, some of our historical Japanese documents used a font for numbers in a balance sheet that is flatter than modern fonts.
- Even though it is trivial for the human eye to read those numbers, commercial OCR softwares completely failed to detect them as characters, returning no content for these regions.
- Thus, we had no choice but to train our own OCR engine to recognize those characters.

Commercial products are trained on documents with clean backgrounds

- **Text bleed** is a major issue in historical documents, due to the types of inks and papers used, the aging of the documents, and scanner settings. Text from one side of the page bleeds through to the other side.
- This creates a problem that we have termed "ghost 1's". The OCR engine detects characters that are not actually there due to the text bleed. Most frequently, the detected characters are "1's". If you are digitizing tables, this can create a major headache.
- More generally, text bleed leads to a deterioration in the performance of both layout analysis and OCR.
- Deep learning is an incredibly powerful tool for pre-processing images for text bleed

Commercial products are trained on clean scans

- **Distortions**. a physical document is a two dimensional object that nevertheless lives in three dimensional space unless the book binding was removed to scan it.
- **Scanner settings**. Sometimes the scanner settings used on historical documents leave a great deal to be desired.
- Again, deep learning-based pre-processing can help.

The data to train an OCR engine on complex social science documents don't exist

- Even if a commercial product wanted to integrate training data from historical documents, which are incredibly varied, into its product, such data do not currently exist at scale.
- Creating such data, given current technology, would not be commercially lucrative, given the variation in historical documents and the size of the market interested in digitizing such documents.

Tools to create annotated datasets cheaply

- I have been working on developing tools that allow researchers to create annotated datasets as cheaply as possible
- We have developed an open source package (Layout Parser) that can integrate labeled datasets, making it easy for researchers to detect layouts on similar documents once the annotations are created. Open sourcing annotated datasets will have substantial externalities for other researchers.
- We will be releasing an open source annotation tool soon

Are there hacks that can make off-the-shelf tools work better?

- You've come to the right place with that question.
- We are pretty sure we've spent more time than almost anyone else on the planet trying to hack Google Cloud Vision.
- We suspect these hacks may also improve the quality of other commercial OCR engines.

Creating computable measures goes beyond layout recognition and OCR

- Even if a product existed that could recognize layout regions and accurately digitize the texts within them, the data still would not be in a format that could be analyzed.
- Importantly, document layout regions need to be correctly associated with each other. How these regions are associated differs vastly across documents, and it is hard to imagine an off-the-shelf commercial product being able to infer such a wide range of associations anytime in the near future.
- Structured information extracted from the documents may still not be computable. For example, raw text requires NLP tools to convert to a computable format.

Off-the-shelf NLP usually doesn't work either

- Like for document layout analysis, off-the-shelf NLP models need to be fine-tuned to specific applications
- There is some hype around models that work without fine-tuning, and we will examine the most promising advances
- However, the need to fine-tune is not going to go away

NLP and Fine-Tuning

- For example, Open AI has hailed their SOTA NLP model - GPT 3 - as capable of doing as well as other leading models that have been fine-tuned, without fine tuning
- This is often not the case when departing from benchmark tasks
- They sold the (extremely large) model to Microsoft, and it is a proprietary black box - this does not bode well for academic applications

Are we close to a commercial solution?

- There is near infinite diversity in the types of documents and texts that social science researchers want to process.
- We don't think you'll be able to hit a single button and out pops the structured data or meaningful computable text measures anytime soon.
- We do think OCR will improve a ton in the next 5-10 years, and that there will ultimately be better interfaces for inputting training data into commercial engines.

Outline

Why not just send data off for manual entry? They promise 99.5% accuracy...

- After asking "isn't there an app that will do this," the next most common question I am asked about is "why not just send the data off for manual entry?" "Why spend so much of your time developing methods for data curation when you could outsource this?"
- I've sent data off for manual entry and have also keyed in lots and lots of data myself.
- Ultimately, I found this experience unsatisfactory, for a variety of reasons. Due to time constraints, I'm only going to discuss the top seven

Reason 1: Double entry is not a panacea and hence you may need to spend a lot of time reconciling errors

- If you read the marketing literature from data entry firms, you might expect to send data off to be manually digitized with double entry, resulting in very high ($>$99%) accuracy
- Researchers who advocate manual entry often repeat these numbers. Unfortunately, we have found accuracy to usually be much lower, even from highly recommended firms.
- When we carefully checked through double entered data, we found far more errors, including very serious ones - i.e. if the numbers in a table are all moved down one relative to the row headers (i.e. let's suppose the headers are municipality names and the values are average income in the municipality), each number will be associated with the wrong observation.

Why doesn't double entry reconcile most of these errors, as the companies claim?

- We think it is because data entry companies will use OCR as a first pass, manually annotating layouts by drawing boxes and then correcting flagrant errors.
- If you have two people working from the same base dataset to correct errors and neither is being very attentive, many errors will remain.
- If layout boxes have not been drawn correctly, this will include errors that make nearly every observation incorrect, i.e. shifting cell values down by a row relative to row headers.

The Perils of Double Entry

- An essential step in checking for errors is seeing whether the numbers add up appropriately in tables that have total rows.
- Of course, an observant person entering the data may also notice a total row, and use an excel formula to generate it, meaning that this check no longer identifies errors.
- We knew something was up when all the numbers checked out, and had to make it clear when contracting for data entry that we would not accept any output that used excel functions to enter data.

The Perils of Double Entry

- In short, I have spent a lot of time fixing problems and entering data myself when I didn't have resources to pay for data entry or the quality just wasn't there.
- If I'm going to be spending eons of time on data curation, I want to be intellectually stimulated and contribute to the public good.
- Sometimes tedious work cannot be avoided - its an inevitable, frequent part of doing research - but data curation is central enough to my research that I wanted to invest in a better solution.

Reason 2: Quality declines as the size of the dataset increases

- Not only does the quality of manual data entry usually leave something to be desired. It also declines as the size of the dataset increases.
- Data entry firms typically work by hiring freelancers.
- They will use their better freelancers first. As the size of the job increases, they may have to tap into freelancers who are not so careful in order to meet demand.
- Hence, for larger datasets, in our experience manual entry was particularly problematic.

Reason 3: Many datasets required to answer pressing questions are too large for manual entry

- In recent years, questions related to topics such as inequality, misallocation, and networks have fundamentally influenced economics and social science more generally.
- These questions, which are very important to the lives of many millions of people, are typically impossible to study with aggregated data.
- Moreover, taking a random sample of microdata may lead to biases. This is well-known in the networks literature but also appears in other contexts (i.e. for studying inequality, having data points from the tails is important).
- Yet, typically microeconomic datasets are too large to digitize manually.

This leaves a few alternatives:

1. Focus on recently-compiled datasets, which are more likely to be in machine-readable format already
2. Focus on a narrow context - i.e. study firms in a single city
3. Choose questions that can be examined with more aggregated data or a random sample
4. Develop methods that can automate data curation with high accuracy

These are all valid strategies, and all things I've done at one time or another. However, there are enough questions I care about that cannot be answered using 1), 2), or 3) that investing in 4) seems worthwhile.

Reason 4: Delving into data curation methods can change what we see as data

- Once you start thinking about how data can be generated, you start to see data that can be used to answer questions in places you never before thought to look.
- This is in large part what has motivated me to teach this course. Even for those who may never intend to implement deep learning methods, understanding what modern methods are capable of can expand the range of questions that can be formulated as testable hypotheses.

Reason 5: Automated curation can open understudied contexts to research, yielding important insights

- Most off-the-shelf historical datasets, particularly large-scale ones, focus on rich countries like the U.S. or the U.K.
- Few datasets have been digitized for lower income countries, even if they exist in hard copy.
- Studying more diverse settings has the potential to enrich our understanding of so many important questions, ensuring that **our research responds to the concerns of all people** and is not so heavily skewed towards groups that are easier to study.

Reason 6: Manual data entry tends to make the field fragmented

- Manual entry is costly, and the costs increase with the size of the dataset.
- Using deep learning is the opposite, it is costly, but once you pay those fixed costs, it scales well.
- Say I was digitizing some novel data to use in an RD. If I'm typing those data in manually, you can bet I'm only going to digitize the observations right around the threshold that I need for the RD.
- Whereas if I'm using deep learning to digitize the data, I might as well process the entire dataset. It wouldn't be much more costly and - even just considering my private returns - could be used for external validity checks, etc.

Reason 6: Manual data entry tends to make the field fragmented

- Now suppose some other researcher would like to use those data for a different question, with a different identification strategy.
- If I've just digitized around the threshold that is of interest to me, these data are not going to be of much use.
- Whereas if I've digitized the entire dataset, it is now much easier for the other research team to answer their question.

Reason 7: Automated curation can democratize access to data

- Sending data off for manual entry requires monetary resources. Not all students or faculty have access to such resources, and in general funding for these types of projects is hard to come by.
- The monetary resources required to run a deep learning model, while not zero (i.e. you need compute time and usually need to either annotate yourself or hire someone to do it), will tend to be more modest.
- Automated curation methods are more intensive in human capital and less intensive in research funding than manual digitization, and this is more pronounced the larger the size of the dataset.

Reason 7: Automated curation can democratize access to data

- Automation can make projects that require large-scale data curation more accessible to students and faculty who lack extensive funding but are willing to work hard to master and innovate the design of automated curation tools.
- This should be particularly beneficial in expanding the range of projects that are feasible for students to pursue.
- To the extent that these methods increase the amount of data curated and the data are placed in the public domain, they will democratize access to data more generally.

Outline

Why do we need novel tools for the digital social sciences?

- Now that we've discussed why there isn't an app to automate data curation at scale, and why manual data entry can be unsatisfactory, you may still be wondering why innovations being made in the digital humanities or computer science can't in general be directly applied to documents of interest to quantitative social scientists.
- We really wish we could just apply off-the-shelf methods and off-the-shelf training data sets, but unfortunately these tools by and large were not developed with the aim of answering social science questions.
- While this work is made possible by groundbreaking innovations in computer science, the questions - and hence the data - that are of interest to social scientists differ in fundamental ways from those that are typically of interest to computer scientists and scholars in the humanities.

Why do we need novel tools for the digital social sciences?

- There have been recent promising innovations in unsupervised learning, and we will cover them in this course
- Yet for many of the things we want to do, there is no general purpose AI technology that can solve the problem off-the-shelf without additional inputs
- Most centrally, implementing an automated curation pipeline for social science data requires building labeled datasets that resemble the information we want to process.
- It also sometimes requires making some modifications to existing methods to tailor them to our objectives (hence understanding how these methods work under the hood is important).

Existing methods power our methods but do not address many issues in our real world documents

Without the paradigm shifting advances in deep learning - applied to object detection and natural language processing tasks - there would be no potential to make substantial progress on automated data curation. Yet, our real world applications are different than those that have driven the deep learning literature thus far. Specifically:

1. Much of the literature on object detection - which underlies document layout analysis - focuses on natural images (like photographs of plants, cats, etc), not document images.
2. The subset of the literature examining documents primarily focuses on clean, modern documents and other highly tailored use cases, that are different from most documents of interest to us.
3. NLP on data with OCR noise, common in our applications, is also fairly niche in the literature

Will the computer science literature develop off-the-shelf methods to process our documents in the near future?

- I don't think this is likely. Importantly, our documents are quite diverse.
- We are working to both encourage greater fluency in DL methods and to give researchers open source tools that make them easier and more efficient to apply to social science documents.
- In the future, when more automated data curation pipelines have been implemented in the social sciences, it is more likely that there will be an off-the-shelf model that works on any given dataset.
- We aim to compile such models for layout analysis and OCR into a user-friendly data curation platform called Layout Parser (still very beta)

Will the computer science literature develop off-the-shelf methods to process our documents in the near future?

- Recent NLP models (i.e. GPT 3) have touted as being able to match performance of other models that are fine-tuned on task specific data without fine-tuning.
- This is, though, often not true outside the benchmark tasks.
- GPT 3 (and other future general purpose models like it) are huge. Open AI sold the rights to Microsoft, so all the details are proprietary. Black box NLP tools are unlikely to be well suited to academic research.

Why do I need to learn these methods? Can't I just collaborate with a computer scientist?

- Collaborations are great but often not possible.
- The incentives in economics and computer science are somewhat different.
- As with all interdisciplinary partnerships, this can lead to frictions that make interdisciplinary collaboration fairly uncommon.

Incentives in Economics

- In economics, we have a quite limited number of journals, and with a few notable exceptions, there has not been a substantial expansion in publishing outlets as the field has grown.
- Researchers typically pay to submit to these journals, rather than to publish their paper if it is accepted, so the journals do not have a financial incentive to expand the number of papers.
- Papers are long, and there are huge incentives to "go big or go home," so to say, including with data curation.
- The discipline largely disincentivizes incremental contributions, which can be difficult to publish in the limited number of outlets available.

Incentives in Computer Science

- In computer science, it is common and professionally rewarded to publish in conference proceedings.
- These papers are short, and conferences may accept hundreds or even thousands of papers. At least for the conferences I'm familiar with, you pay to attend the conference - not to submit - so there are incentives to accept lots of papers (although top conferences are still very selective).
- As a result, computer scientists publish many short papers, most making incremental contributions. It is not uncommon for pre-doctoral fellows applying to grad school in computer science to have more published papers than a tenured professor in economics.

Incentives in Computer Science

- For the bulk of contributions that are not revolutionary advances, the computer science publishing model described above tends to favor writing many incremental papers quickly.
- When seeking applications for new methods, it is very common to use pre-existing benchmark datasets, which are well-suited to this sort of incremental progress.

Benchmark Datasets and Computer Science Research

- For example, in layout analysis, a common benchmark dataset is PubLayNet, which consists of image scans of modern journal articles and their accompanying layout annotations, extracted from the pdf metadata.
- There is no need for humans to annotate the layout regions because they can be extracted from the pdf metadata. It is hence nearly costless to produce a very, very large dataset of scans and associated content region annotations to use to test incremental improvements in algorithm design.
- Moreover, referees know exactly what they are looking at when reviewing papers that evaluate methodological improvements on benchmark datasets like PubLayNet.

Computer Science and Benchmark Datasets

- Throughout this course, you will see the landmark papers that shattered benchmarks by a wide margin, with transformative methods
- Yet incremental progress on benchmarks - much of the literature - is usually not of much relevance to our applications, as it is tailored towards benchmarks that don't resemble our tasks.
- What we need is a cheap way to generate training data and an incremental tailoring of methods towards the types of datasets we would like to curate. The latter is difficult until we have the former, which is a job in which other fields are unlikely to take an interest.

What are the differences with the digital humanities?

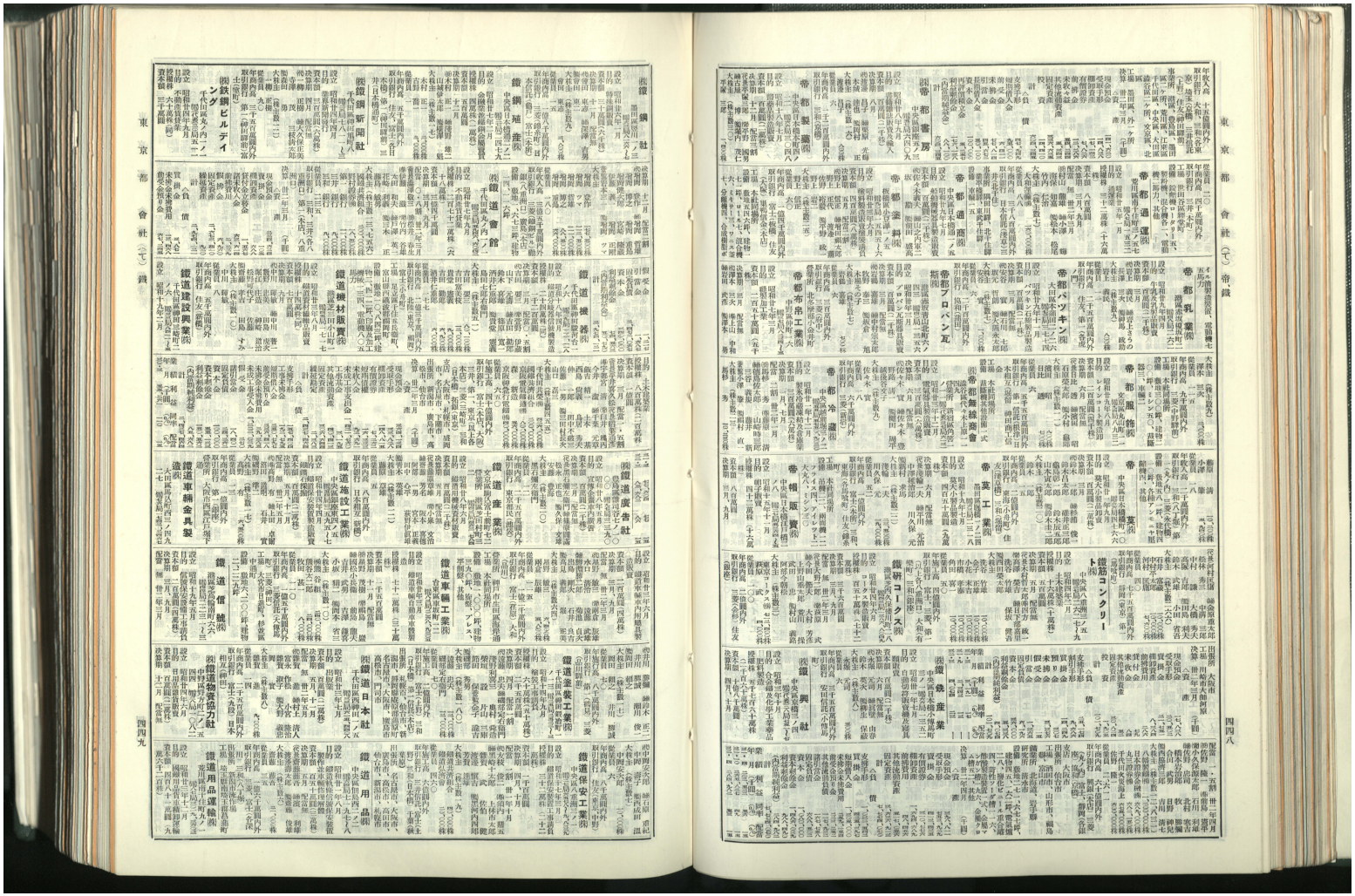
- The questions we are working on have a lot in common with the digital humanities, who like us are primarily interested in noisy, real world documents.
- Notably, though, the downstream analyses in the digital humanities tend to be very different than those in the quantitative social sciences.
- The fact that their questions are different means that their documents are systematically different, and the types of noise that will invalidate their downstream analyses are also often quite distinct.

How are humanities documents different?

1. Their documents have different layouts
2. Errors tend to be more catastrophic and harder to fix with post-processing in quantitative documents
3. The layouts of quantitative documents can make OCR errors more likely

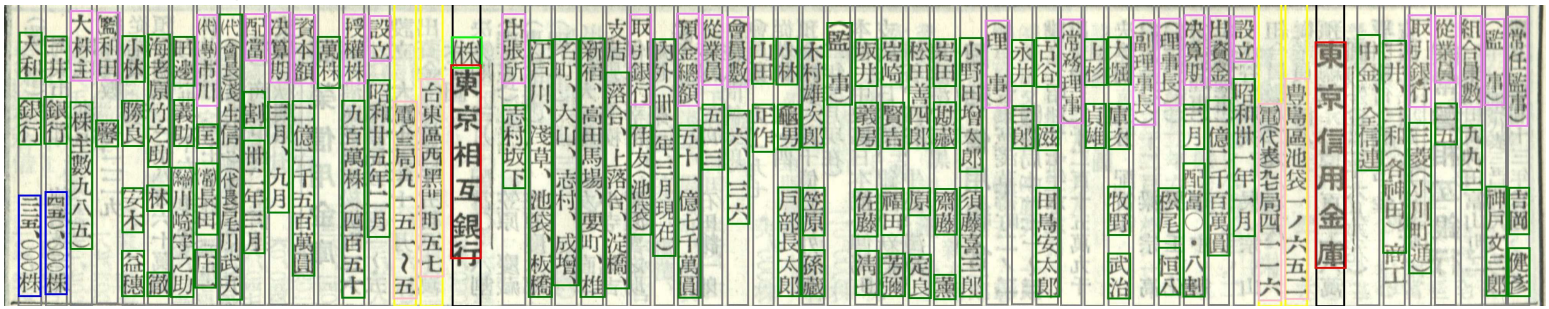
Our Layouts Are Distinct

The below image provides an example scan - of which there are tens of thousands in total - from historical firm level financial reports that we are processing.



Our Layouts Are Distinct

Each scan contains over a thousand layout elements: i.e. company titles, addresses, specific variable names and their values, section headers. The bounding boxes and classes of each of these individual layout elements needs to be accurately detected.



Then, the elements need to be correctly associated with each other to create structured datasets. Layouts and word wrapping rules differ within different sections of each firm's table, and the information provided varies from firm to firm.

Our Layouts Are Distinct

- This contrasts to a single column book, where there might be a single layout element (i.e. one big text block) on a page.
- Books with straightforward layouts are much more likely to be of central interest in the humanities than to quantitative researchers, where complex tables are paramount.
- Hence, the focus in the digital humanities tends to be more on downstream tasks, like recognizing ancient characters or natural language processing. Without detecting layouts accurately, we cannot get to these downstream tasks. The project dies with the failure of layout analysis.

When layouts are complicated, the OCR quality tends to deteriorate:

1. Often, text regions fail to detect altogether, because if the OCR software does not recognize a layout element on that portion of the page, it will not attempt to recognize what characters are present.
2. On the other hand, text bleed may lead to the detection of characters that aren't actually there. We term these characters "ghost 1's", as text bleed can lead 1's to be detected where there should be blank space.

Text bleed is more of a problem:

1. **When there is more blank space on a page.** This is less common in a book with dense text, and much more common in tables where white space is essential to demarcating complex layouts.
2. **When downstream analyses are quantitative.** A few stray 1's may do little harm or be easy to remove in post-processing from a book that should primarily contain letters and not numbers. However, if they appear in a table, they can make numbers off by an order of magnitude, fundamentally altering downstream statistical analyses.

Complex layouts and low tolerance for errors make our problems hard

- Fortunately the concepts (i.e. math) that underlie deep learning will be very familiar to economists, and we are well-posed to deal with these significant challenges.
- Of course, there are other features - like ancient languages - that make digital humanities problems hard. The point here is simply that humanities and the social sciences are different.