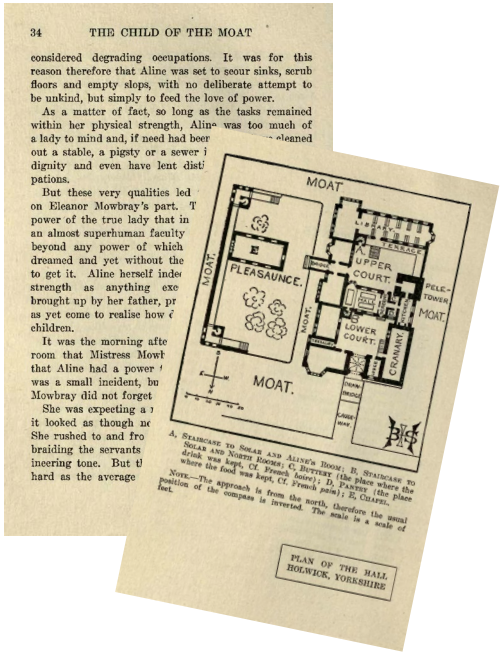
Challenges of Optical Character Recognition

Orville Anderson

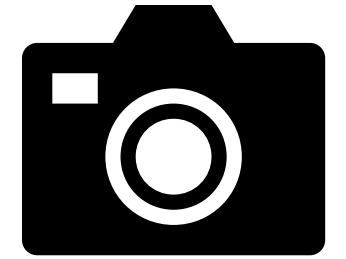
University of Minnesota Morris, Senior Seminar, Fall 2025

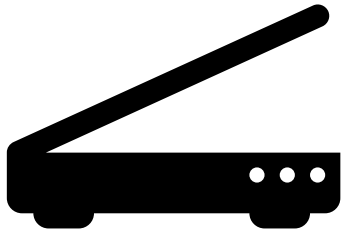
Process

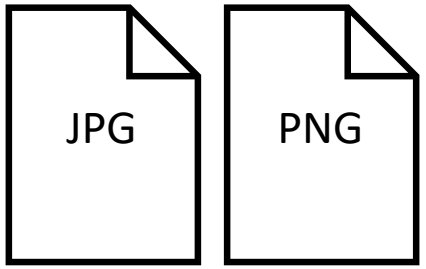
Limitations

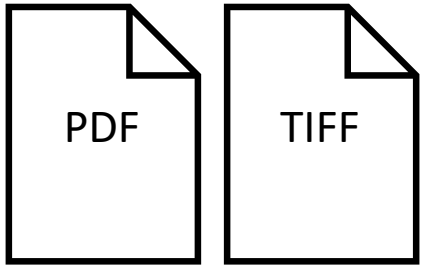


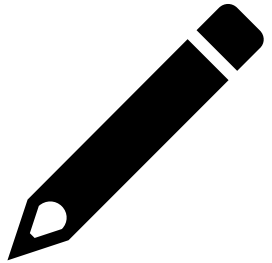
Hegghammer, 2022











How to make a Scanned Document

What Is Optical Character Recognition (OCR)?

- Defined as: the process of extracting text from images
- OCR model is something performing this process
- Useful to digitize scanned documents (for research, digital accessibility, etc.)
- Reduced accuracy on certain documents

Background

- What is OCR?
- Stages of OCR
- Methods

Case Study

- Introduction
- Traits
- Results

Conclusion

- Case Study Importance
- Larger Picture

OCR Stages

Input

Image (Scanned Doc)

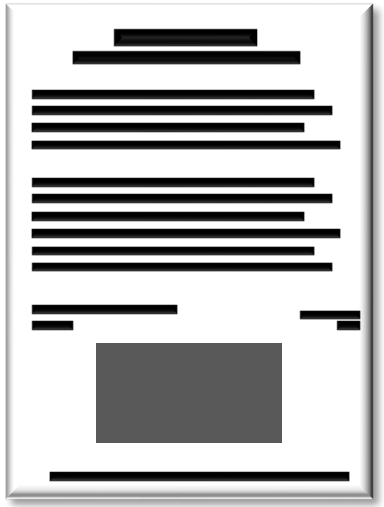
Document Layout Analysis

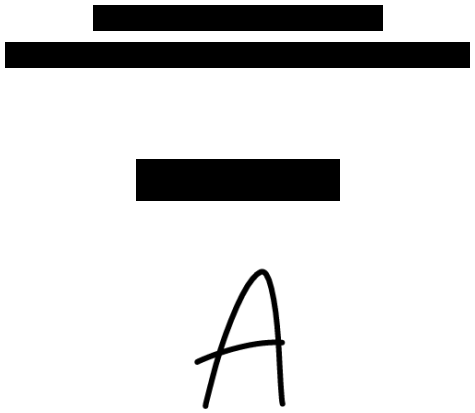
Text Line Detection

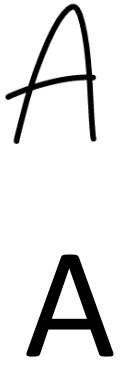
Recogniton

Output

Editable Text



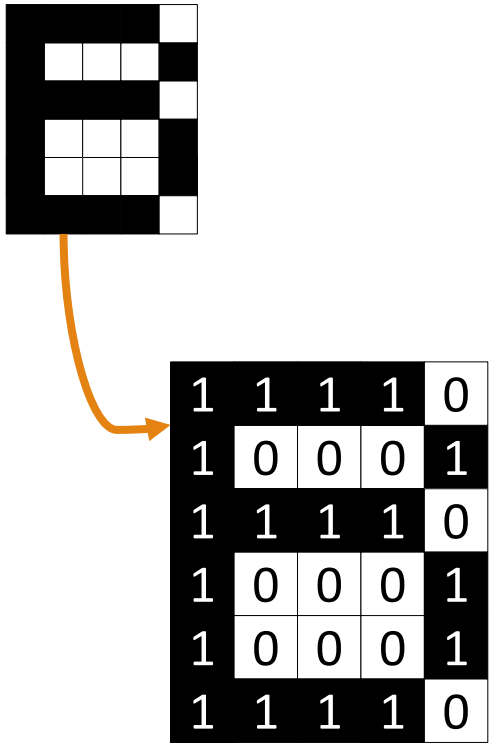


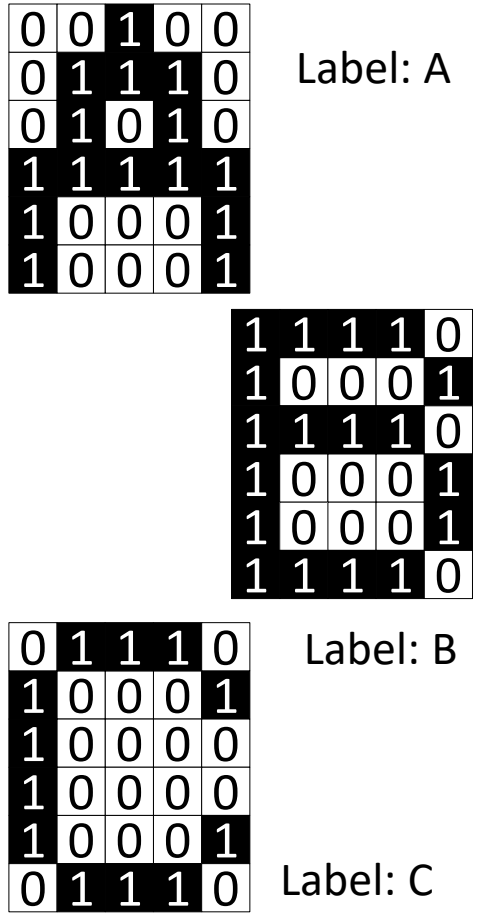


Classification Method – Matrix Matching

- Foundational classification method
- Patented in early 1930s
- In use into the 1990s

Input	Processing	Output
From TLD Stage	Compare to templates	





A: 17 mismatched points

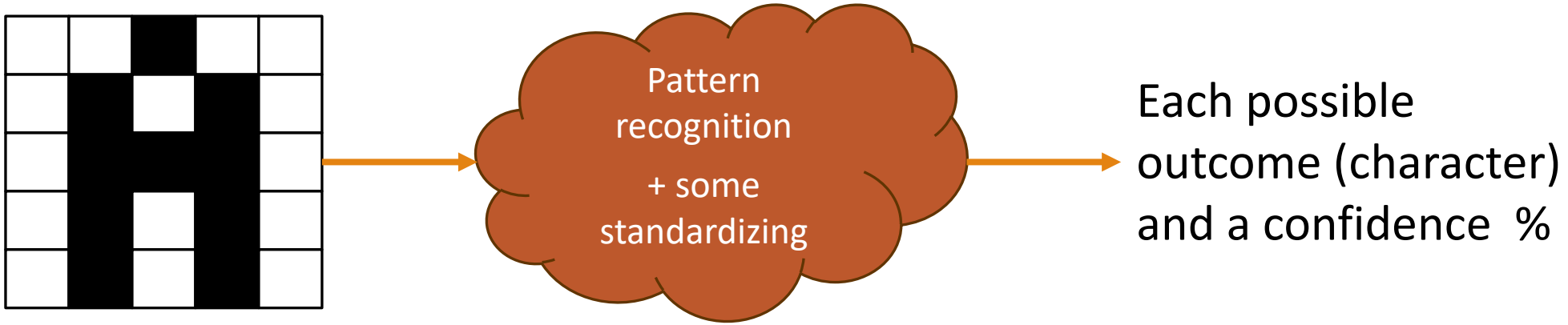
B: 0…

C: 6…

Diagram inspired by LI

Classification Method – Neural Network

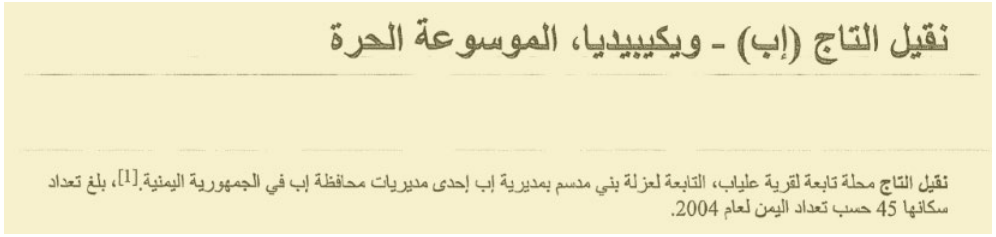
- A more complex version of Matrix Matching
  - Matrix Matching – each pixel has uniform impact
  - Neural Networks – each pixel and outcome has an associated weight
- Easier to update collection of known characters
- Allows us to identify full lines of text at once



Summary

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
0123456789
.,!@#$%^&*()

Limited to “known characters” and fonts


نقيل التاج (إب) - ويكيبيديا، الموسوعة الحرة

نقيل التاج محلة تابعة لقرية علياب، التابعة لعزلة بني مدسم بمديرية إب إحدى مديريات محافظة إب في الجمهورية اليمنية.[1]، بلغ تعداد سكانها 45 حسب تعداد اليمن لعام 2004.

Hegghammer, 2022

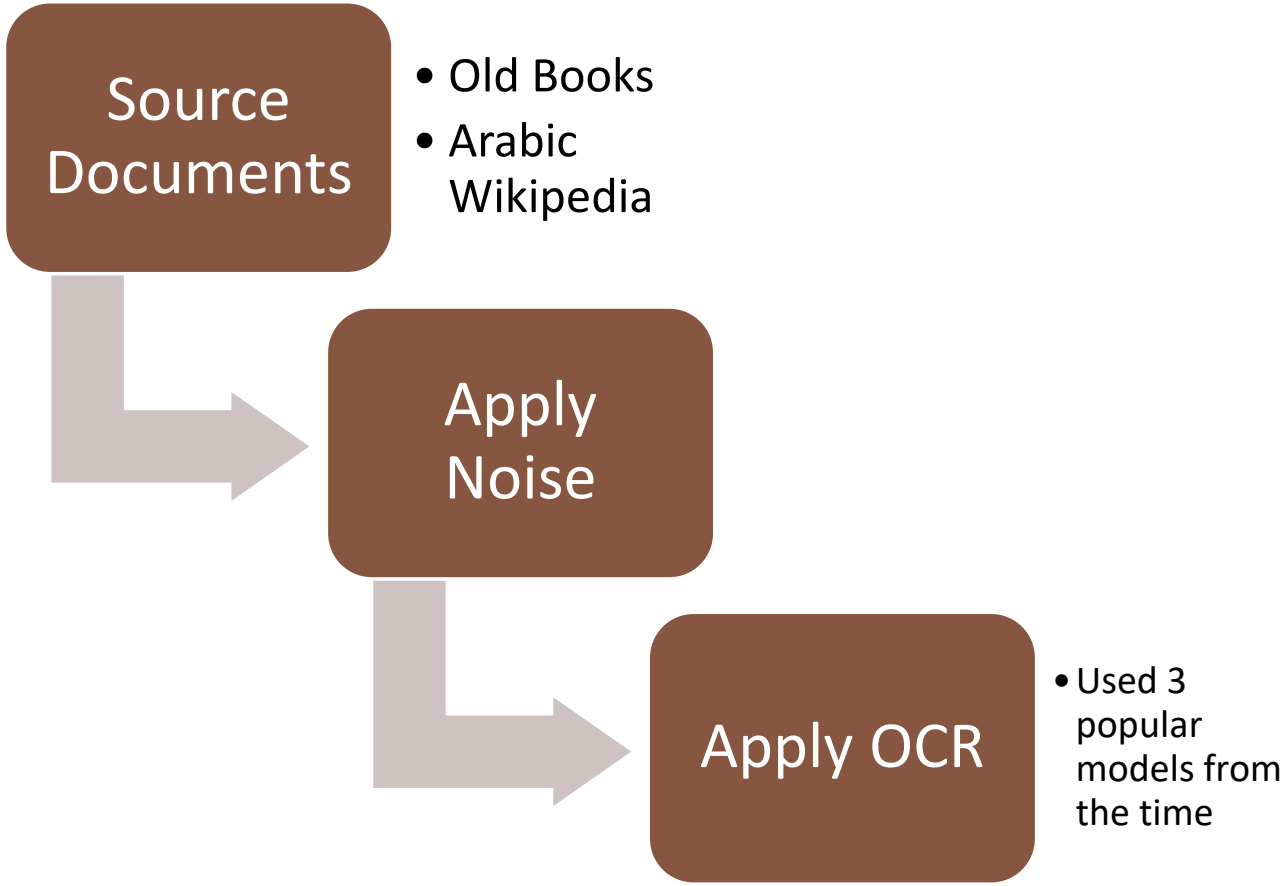
Need examples of more characters and fonts

Case study is a document collection to evaluate accuracy, but can be used to train

Case Study

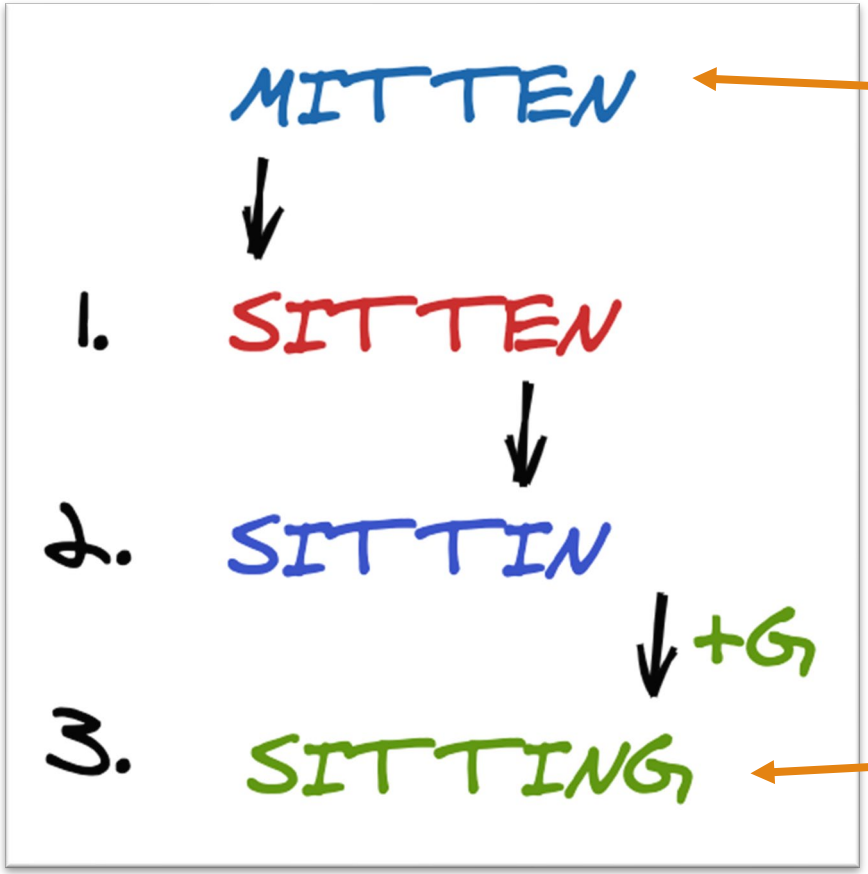
- Thomas Hegghammer, 2021
- Historian who works with Arabic texts
- Collection is called Noisy OCR Dataset (NOD)
- Focus on noise, layout, and language

How accurate are general OCR models on documents found in social studies and humanities research?



Comparison

- Many ways to compare OCR models
- For accuracy:
  - Need OCR output
  - Need ground truth
- Judge accuracy by individual characters



[algodaily.com]

considered degrading occupations. It was for this reason therefore that Aline was set to scour sinks, scrub floors and empty slops, with no deliberate attempt to be unkind, but simply to feed the love of power.

As a matter of fact, so long as the tasks remained within her physical strength, Aline was too much of a lady to mind and, if need had been, would have cleaned out a stable, a pigsty or a sewer itself, with grace and dignity and even have lent distinction to such occupations.

But these very qualities led to further antagonism on Eleanor Mowbray's part. They were part of that power of the true lady that in Aline was developed to an almost superhuman faculty and which went entirely beyond any power of which Mistress Mowbray even dreamed and yet without the child making any effort to get it. Aline herself indeed was unconscious of her strength as anything exceptional. She had been brought up by her father, practically alone and had not as yet come to realise how different she was from other children.

It was the morning after the discovery of the secret room that Mistress Mowbray had the first indication that Aline had a power that might rival her own. It was a small incident, but it sank deeply and Eleanor Mowbray did not forget it.

She was expecting a number of guests to dinner and it looked as though nothing would be ready in time. She rushed to and fro from the hall to the kitchen upbraiding the servants and talking in a loud and domineering tone. But the servants, who were working as hard as the average of their class, became sullen and

at the peg, pull the entire strand thru the thumb and forefinger to prevent twisting, and pull the end down thru the hole on the opposite parallel rail next to the corner hole, and then up thru the hole next to it. See that the right side of the cane is out on the

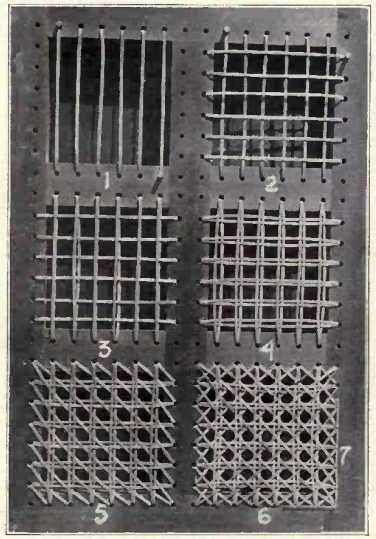


FIG. 4. THE SEVEN STEPS IN CANING.

underside of the frame as on the top. Pull the cane reasonably taut, and fasten with a peg to prevent the strand from slipping back and becoming loose. Draw the cane thru the thumb and forefinger again; pull it across the frame and down thru the hole next to the peg and up thru the hole next to it. Pull taut and fasten with the

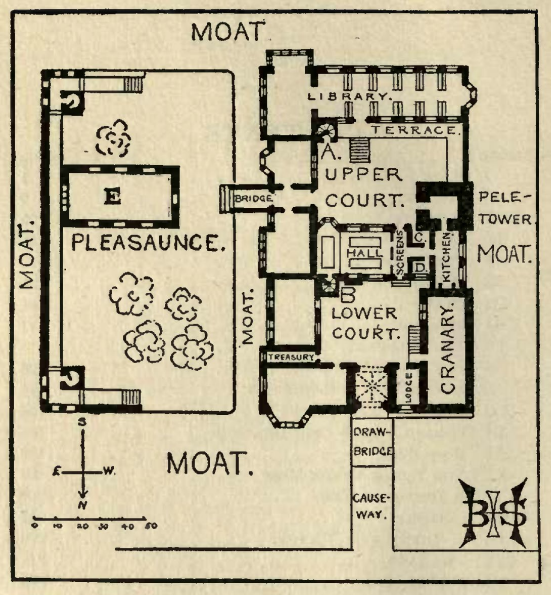
may take the place of the awl. A pair of dividers and rule are necessary for marking. Several wood pegs are needed. These may be classed with the tools. They are made from a ¼ in. dowel rod, or the equivalent. Cut them about 4 in. long and point them as you would a lead pencil. The amateur is inclined to use a number of pegs. Four should prove amply sufficient.

Beginning the Operation.—Fig. 2 is a photograph of an upholstered leg rest with caned sides. This rest will be used for our initial work in cane weaving, inasmuch as the area for caning is rectangular. It is not advisable for the beginner to have his initial experience on a chair seat, for the area is usually of an odd shape, and arms, legs, and back interfere. However, any rectangular area on which there are no projections to bother may be used for the first trial.

FIG. 2. LEG REST.

It is assumed that the sides of the rest have been fitted. The rails and stiles are then assembled with glue, without the posts. When the glue has set the proper length of time, and the frame is cleaned and sanded, the rails and stiles are ready to dimension.

Draw pencil lines entirely around the inner sides of the rails and stiles, ½ in. from the edges. This distance remains constant, usually, on all areas and with canes the various widths. With a pair of dividers set at ½ in. space off points on the pencil lines, starting from the intersection of the extended lines on each rail. Fig. 3 is a working drawing of a corner, dimensioned as suggested. It will make clearer the directions. It is fundamental that the spacing be done in the same direction on parallel rails, for at times



A, STAIRCASE TO SOLAR AND ALINE'S ROOM; B, STAIRCASE TO SOLAR AND NORTH ROOMS; C, BUTTERY (the place where the drink was kept, Cf. French *boire*); D, PANTRY (the place where the food was kept, Cf. French *pain*); E, CHAPEL.

NOTE.—The approach is from the north, therefore the usual position of the compass is inverted. The scale is a scale of feet.

PLAN OF THE HALL
HOLWICK, YORKSHIRE

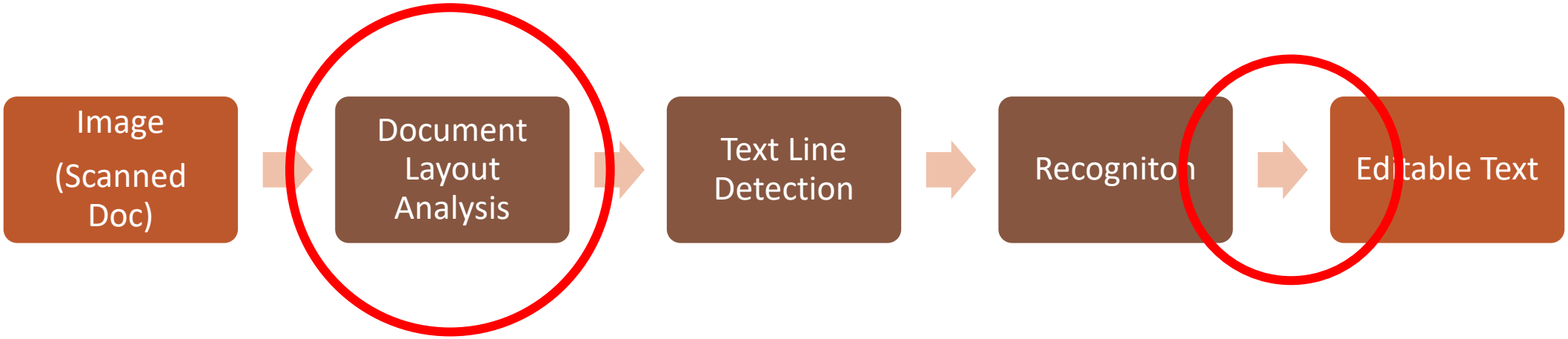
Layout

Examples are of English old books [Hegghammer, 2022]

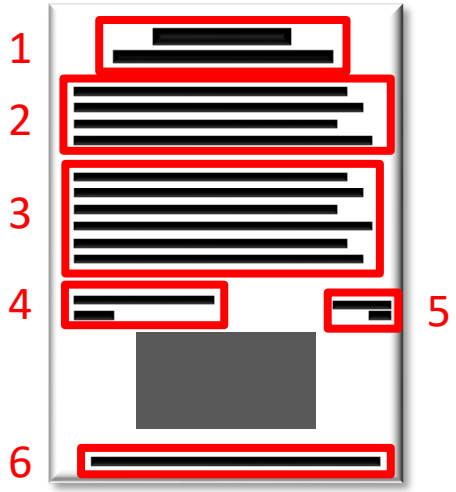
Arabic Wikipedia articles had limited layout variation

Does not include two column layouts and tables

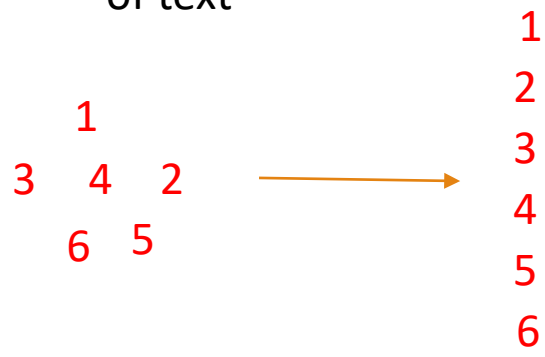
Preserving Layout



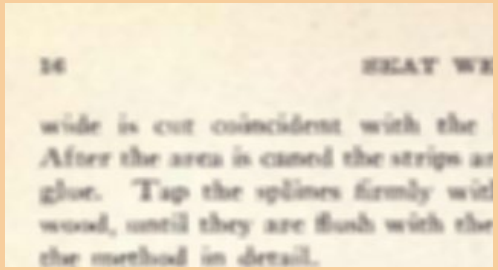
Record order of text



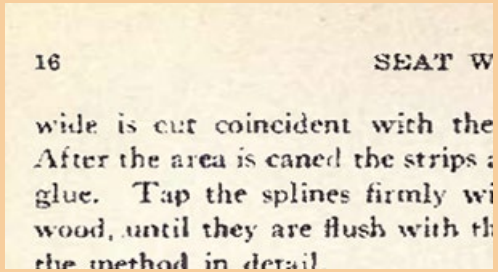
Reconstruct order of text



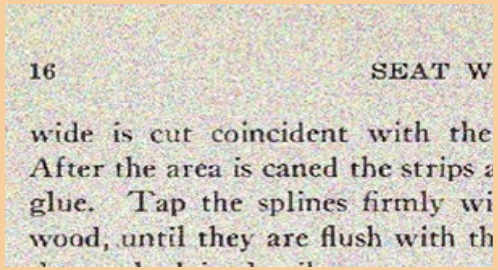
Integrated



Blur

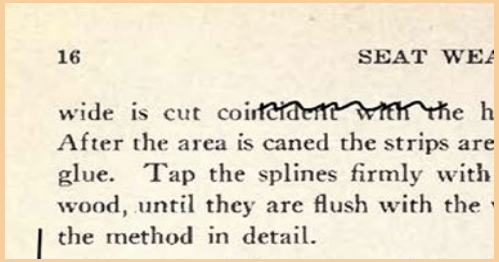


Weak Ink

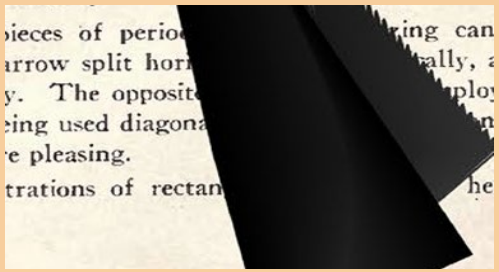


Salt and Pepper

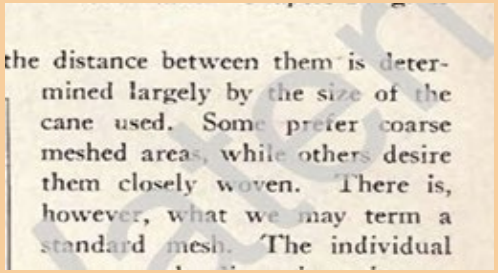
Superimposed



Scribbles



Ink Stain



Watermark

Noise

Picked because they are the most common noise found in historical documents

Falls into two categories:

- Integrated: built into page
- Superimposed: added on top of text

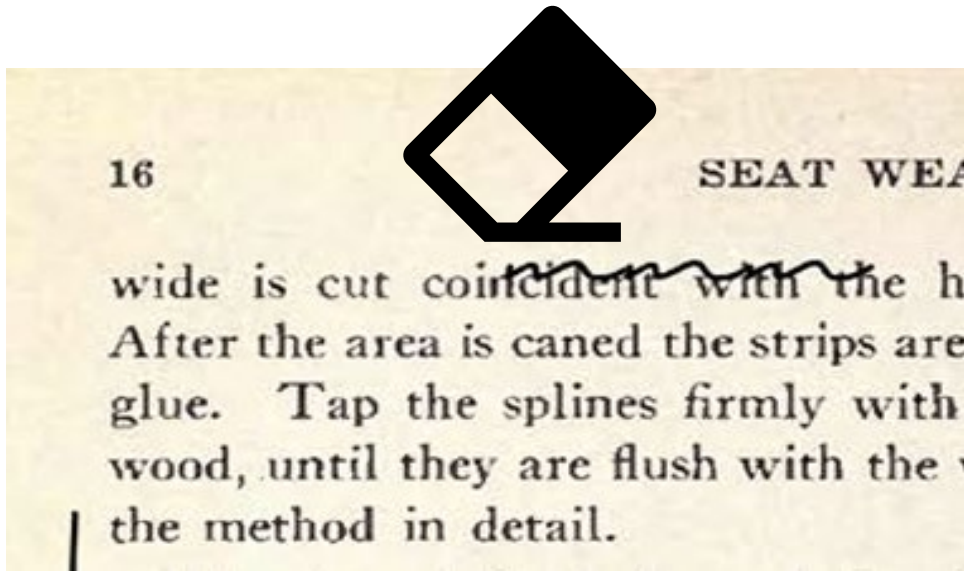
Noise can obscure and/or add characters

Examples from Hegghammer, 2022

Impact of Noise

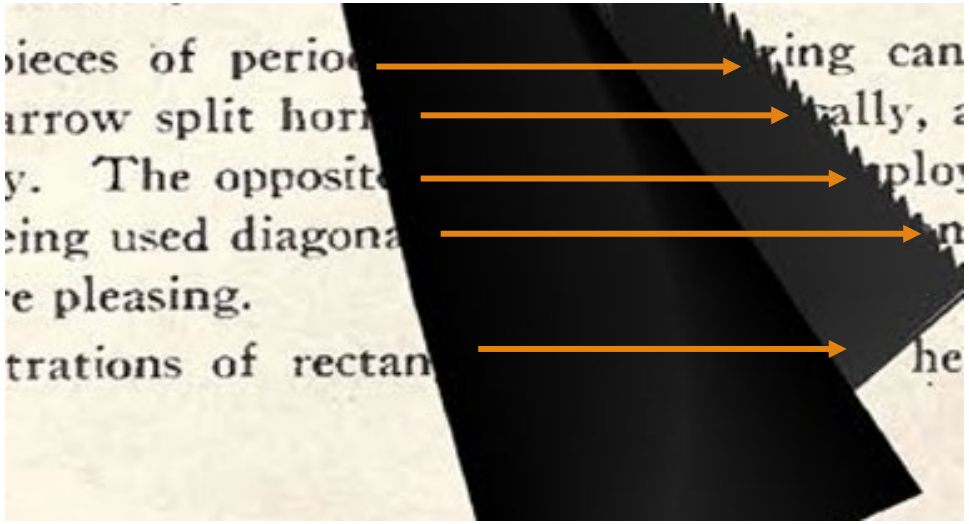
Mistaken for text:

1. Remove noise (pre-processing)
2. Spellcheck (post-processing)



Obscured Text:

OCR needs to know what to do with the unreadable space (related to layout struggles)



Hegghammer, 2022

Writing Systems

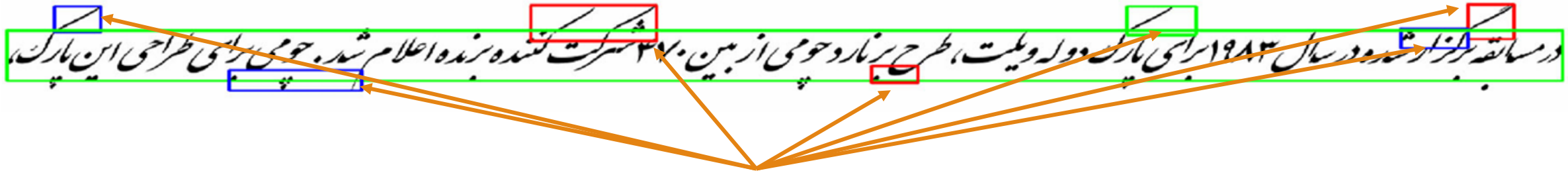
- Most OCR are made for Latin based written languages: English, Spanish, German, etc.
- Latin is the most widely used writing system
- OCR Recognition is limited to characters it has been exposed to previously
- Arabic, the third most widely used writing system is very different from Latin
  - Diacritics
  - Cursive or connected characters

أبجدية رومانية

Why is Writing System a Challenge

- Need to adjust Text Line Detection box sizes
- Noise prevention techniques can remove diacritics

Example of output from a model which uses full lines of text



Areas with low confidence

Fateh, 2023

Results – Models, Writing Systems, and Layout

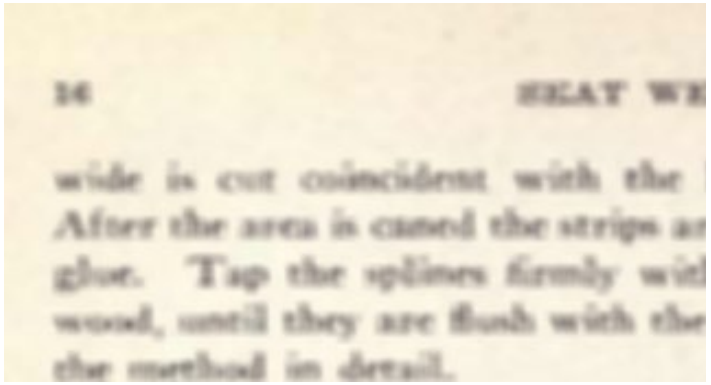
- Some OCR models cost money
- Some were easier to use
- Some had limited support for languages
- Layout variety was minimal – a limitation

Model	Paid?	User Interface?	Arabic?
Google Document AI	Yes	Yes	Yes
Amazon Textract	Yes	Yes	No
Tesseract	No	No	Yes

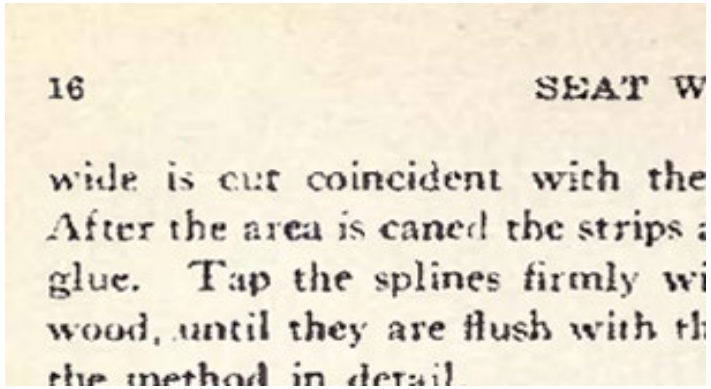
Results – Noise

- Documents with multiple types of noise added, had lower accuracy
- Some noise types had a larger impact on accuracy

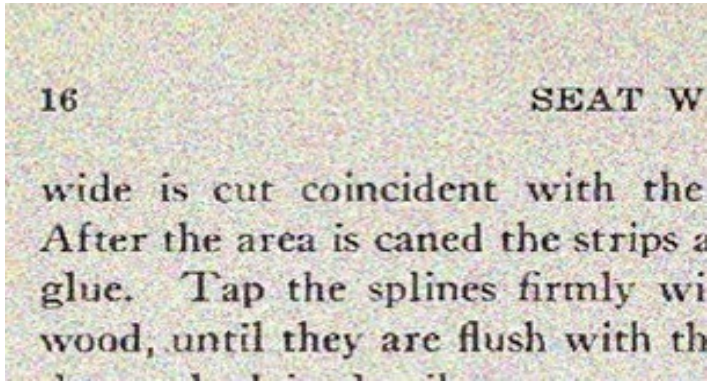
Blur



Weak Ink



Salt and Pepper

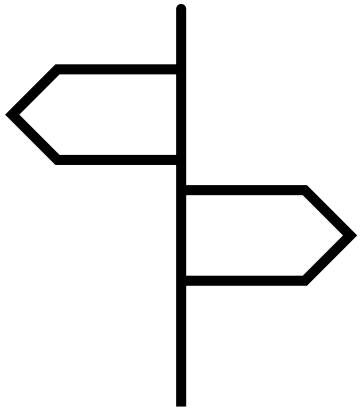


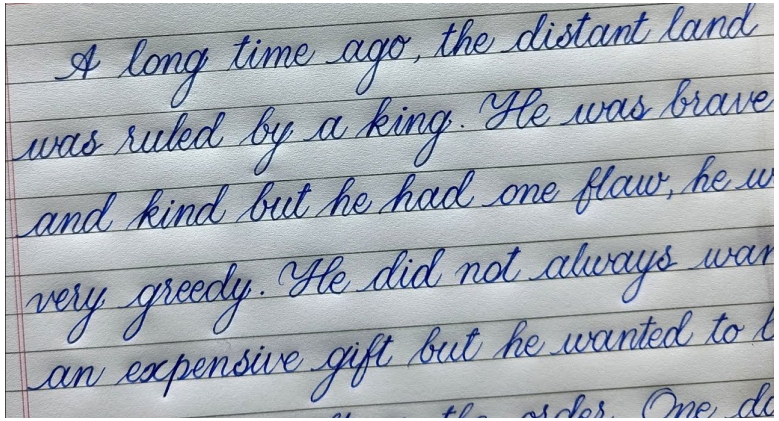
Hegghammer, 2022

Case Study Significance

- Insight on impact of noise
- Commentary for other academics
- Document collection and data is publicly available
- Noise generator is publicly available

[hegghammer.net]





[Palash Calligraphy]

Larger Picture

- Guides future research
- Increases number of datasets and related tools
- Addresses specific needs
- Addresses broader needs

Questions?

Sources

Hegghammer, T. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Comput Soc Sc* **5**, 861–882 (2022). https://doi.org/10.1007/s42001-021-00149-1

Fateh, Amirreza, Mansoor Fateh, and Vahid Abolghasemi. “Enhancing Optical Character Recognition: Efficient Techniques for Document Layout Analysis and Text Line Detection.” *Engineering Reports* 6, no. 9 (December 14, 2023). https://doi.org/10.1002/eng2.12832.

Vaughan, Don. 2025. The World’s 5 Most Commonly Used Writing Systems. https://www.britannica.com/list/the-worlds-5-most-commonly-used-writing-systems

Wikipedia contributors. 2025. Arabic diacritics — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Arabic_diacritics&oldid=1317677291 [Online; accessed 26-October-2025].

Li, Ning. ‘An Implementation of OCR System Based on Skeleton Matching’. University of Kent, Canterbury, UK: University of Kent, Computing Laboratory, March 1993. https://kar.kent.ac.uk/21129/.

Schantz, Herbert F. *The history of OCR, optical character recognition*. Manchester Center, Vt: Recognition Technologies Users Association, 1982.

Unless otherwise specified, all images are Adobe PowerPoint graphics or made by the Author