

# Automatic assessment of OCR quality in historical documents

**A. Gupta<sup>1</sup>, R. Gutierrez-Osuna<sup>1</sup>, M. Christy<sup>2</sup>, B. Capitanu<sup>3</sup>  
L. Auvil<sup>3</sup>, L. Grumbach<sup>2</sup>, R. Furuta<sup>1</sup>, and L. Mandell<sup>2</sup>**

<sup>1</sup>Computer Science and Engineering, Texas A&M University

<sup>2</sup>Initiative for Digital Humanities, Texas A&M University

<sup>3</sup>Illinois Informatics Institute, University of Illinois at Urbana-Champaign



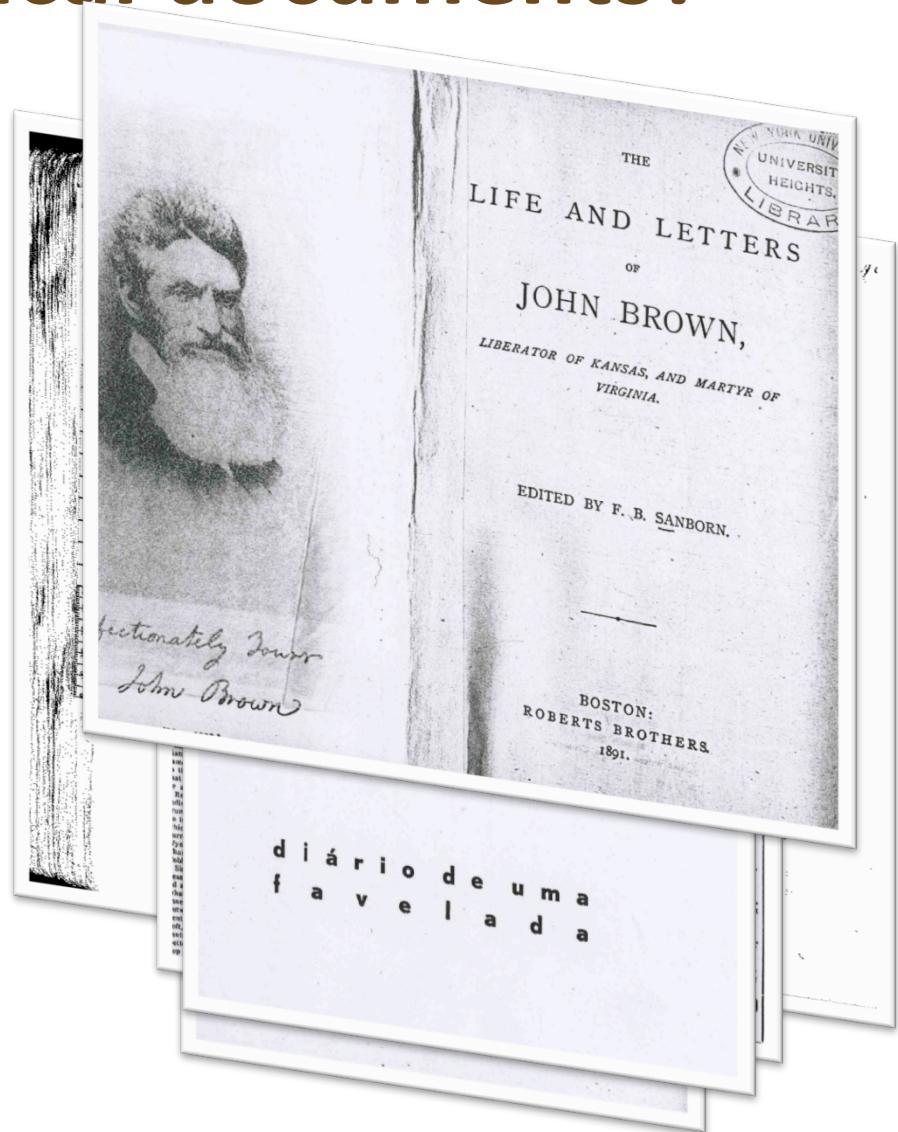
**COMPUTER SCIENCE  
& ENGINEERING**  
TEXAS A&M UNIVERSITY



perception  
sensing  
instrumentation

# What are historical documents?

- Correspondence
- Diaries
- Newspapers
- Government Documents
- Books



# Digitizing historical documents

## Why?

- Historical records are in analog form
- Due to their fragility, most of them are not accessible

## How to make them accessible?

- Digital text transcription

## Ways of digitization

- Hand transcribe each book
  - Resource intensive
- OCR: optical character recognition
  - high-error in text transcription

## Mass digitization projects

The screenshot shows the EEBO (Early English Books Online) search interface. At the top, there's a yellow header with the EEBO logo and navigation links for HOME, MARKED LIST, and SEARCH HISTORY. Below the header is a search bar with fields for 'Search using' (Variant spellings checked), 'Variant forms' (unchecked), and 'KEYWORD(s)', which includes a link to 'Select from a list' and a 'Check for variants' button. There are also fields for 'LIMIT TO', 'AUTHOR KEYWORD(s)', 'TITLE KEYWORD(s)', 'SUBJECT KEYWORD(s)', 'BIBLIOGRAPHIC NUMBER', 'LIMIT BY DATE' (From: 1473, To: 1900), and sorting options ('Sort results: Alphabetically by author' and 'Display: 10 results per page'). A 'Clear search' button and a large blue 'Search' button are at the bottom.



# Early modern OCR project (eMOP)

## Goal

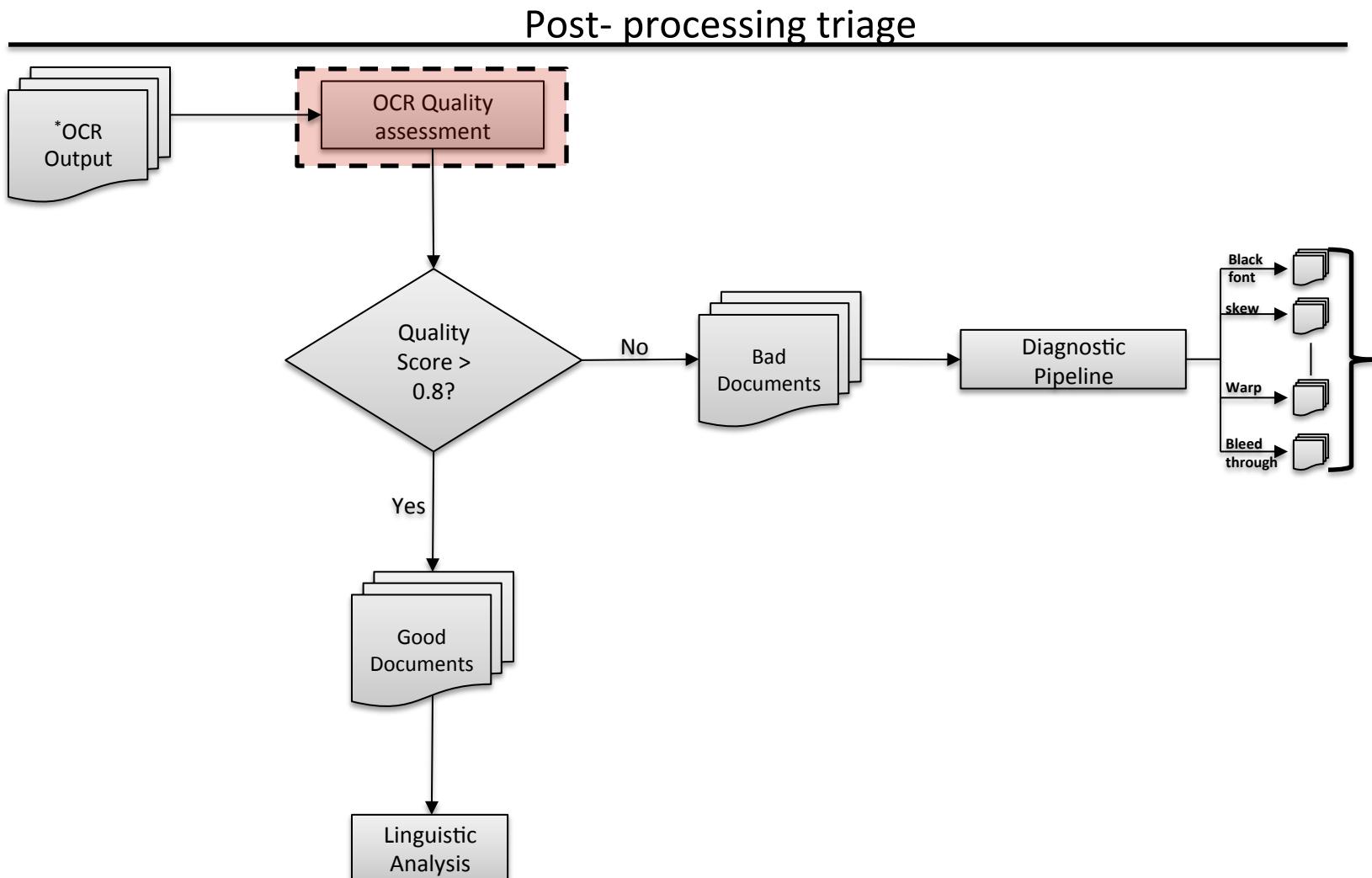
- Improve OCR accuracy for early modern texts
  - 300k documents, 45M pages
- Open source OCR tools

## Challenges

- Early modern printing
- Document image problems



# Why measure OCR quality?



# Our approach

## Post-process OCR output

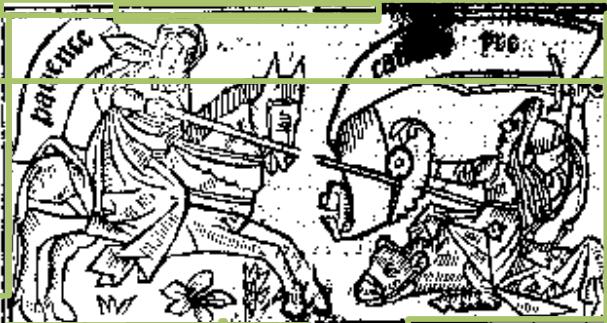
- Page segmentation result such as bounding box (BB) coordinates
- OCR word confidence

## Build ML models to remove noise

- Binary classification: classify each BB either as text or noise

**Quality $\downarrow$ OCR  $\propto$  1% noise BBs**

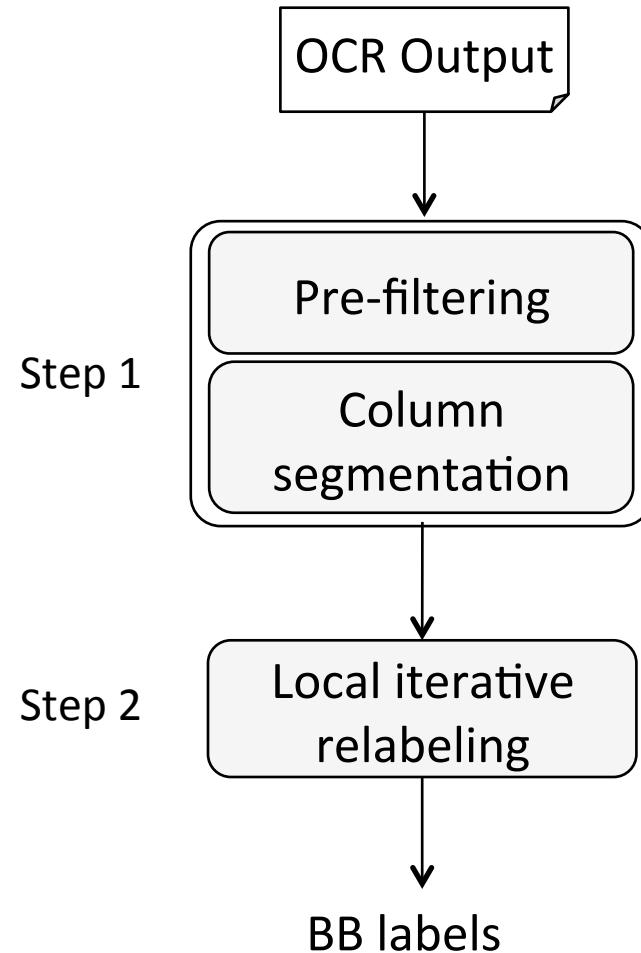
# Language-agnostic approach



20 I say whan þere is sike allaire.  
On thy shal come a tyrant dangerous  
Whose name is Ise withouten faute  
And al þis yere and despitus.  
And bide betwix alwape contaryous  
The which in leuauntes both abouide  
þe may well lyp that he is curios  
Whom he his dyce doch not confounde  
  
Cruelte betwix his bayes  
Telenye is his chyef champion  
þecuerlyte is his postere  
Madnes regnydeth his bouncon  
Wuced murde that rade belon  
At his hous is as it set capitayn  
Here is a cutted telengon  
A hym that foloweth theri trayne  
  
Cheþreþ of Trede the Driftell  
þeneþ thy force and thy purfaunce  
Call vnto the debonaþenne

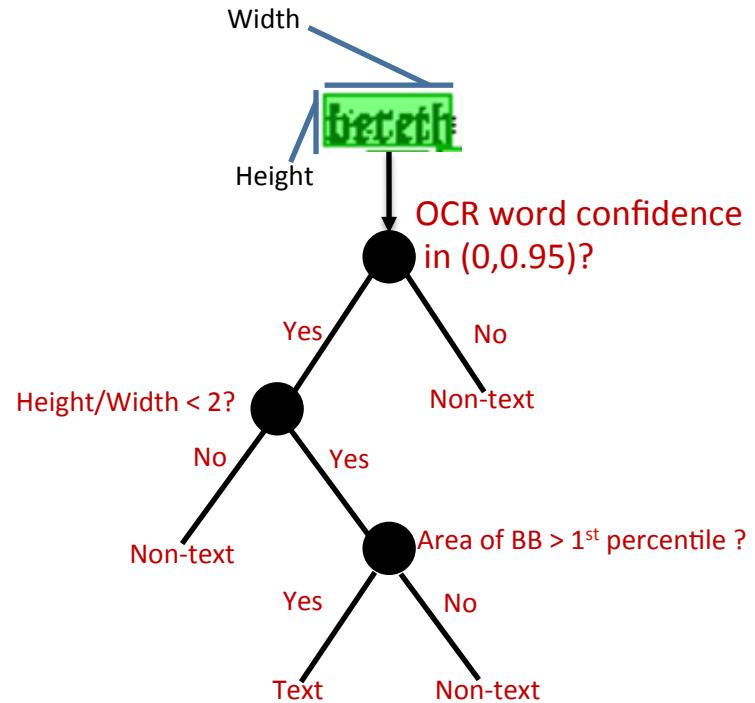
Item a 12 yere it shall sayg by lawe ice  
Whch her shall com fforre suffraunce  
Pacayence is chper with dyscrecyon  
In suffraunce with attemperaunce  
Sobydynge the vnde correccyon  
He hath neyther meret nor pety  
On man nor wooman here lyuyng  
But eadown arrayed ful cruelly  
Cunnewyd to deas and to waerte accordyng  
Constreynt of ech byce semynge  
Whose furoun meyest mannes het  
Robiche or his counsell are leyyng  
Wherfore thj ryght cheffis dyuerl  
It is impossibyl that a man troug  
Any vnde got no good seruice  
Therfore is alwyng ryght daungerous  
Whch comys without knyfys  
It is a ryght ded myghte upre  
Whiche often dothe ryght dete damage  
Sowþ thou art warched by thos myghte  
Lyft that he do to thre outrag  
Shewe thy force and myghtam  
Call unto the force with noblenesse  
Pray pacayence to be thy lassance  
And cheff shall this ryght lyuyng onysell  
Whom汝 is com vnewys shall her dresse  
On every lyde waye to ryghte  
Adicte or all byce ista re mervellously  
A strongi creautur delvyng ryght

# Quality assessment algorithm



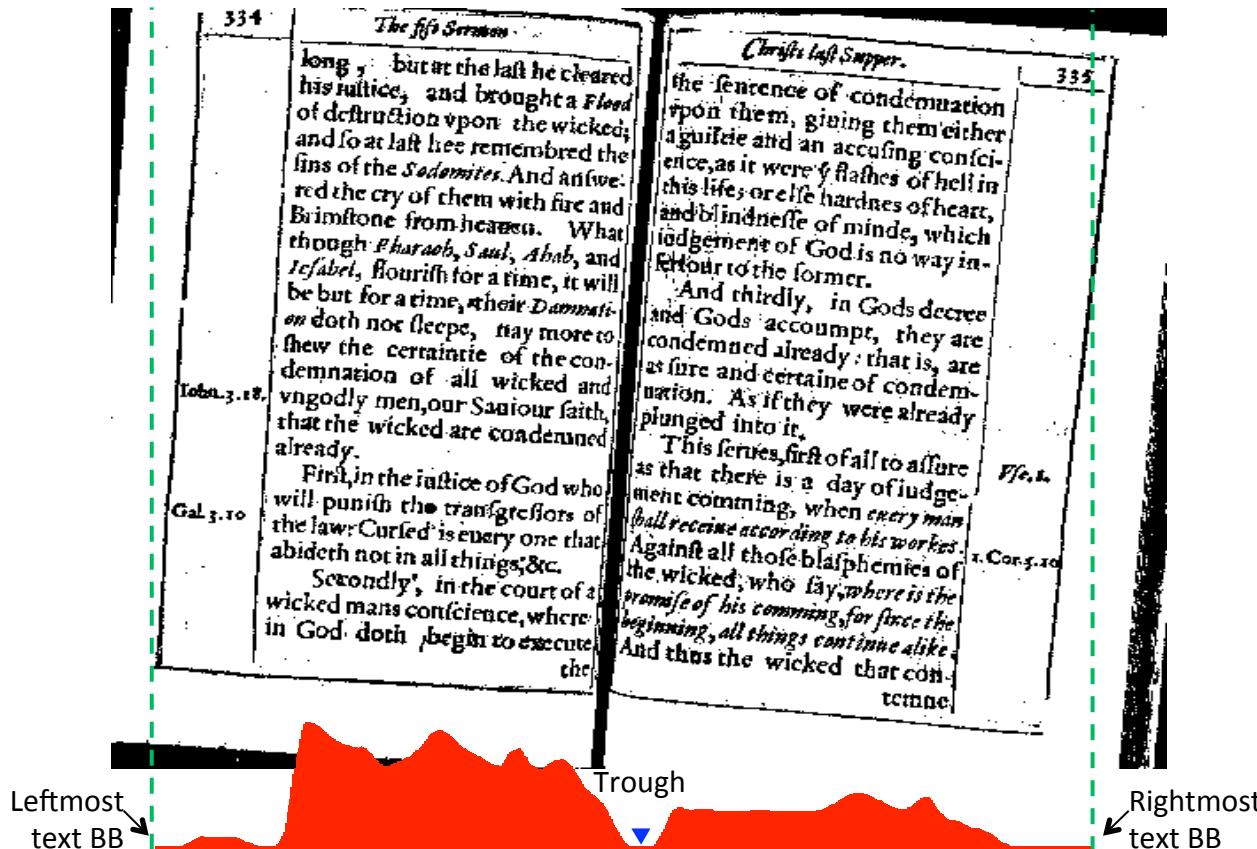
## Prefiltering

- Provides initial labels to be refined in later stages
- Rule based classifier
  - BB properties and OCR word confidence.
  - Conjunction of rules
- Problems
  - Many text BBs classified as noise
  - Need a way to recover lost text BBs



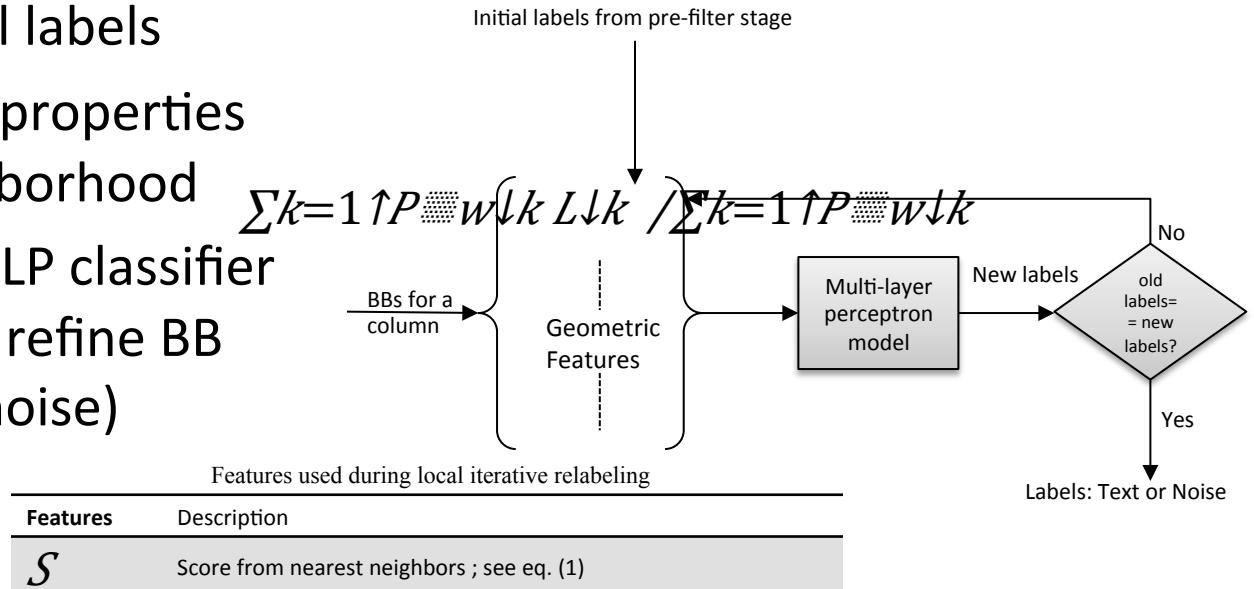
# Column extraction

- Extract individual column and then process each column



# Local iterative relabeling

- Refines initial labels
  - Based on BB properties and its neighborhood  $\Sigma$
  - Applies an MLP classifier iteratively to refine BB labels (text/noise)



*GLOC* OCR word confidence\*

R

$H/W$  Height-to-width ratio of BB\*

*A* Area of BB\*

$$H_{rm} \downarrow no \quad \text{Normalized height: } H_{rm} \downarrow norm = (H_{rm} \downarrow med - H_{rm} \downarrow med) / H_{rm} \downarrow IQR$$

# Final output



# Results

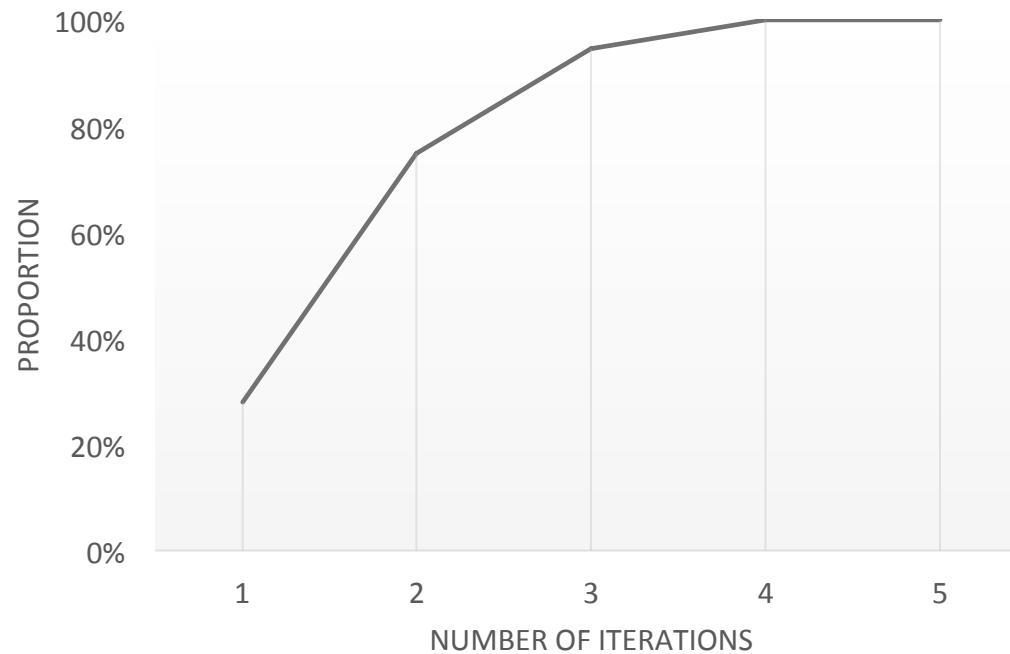
## Dataset refinement: local iterative relabeling.

- Binarized page images
- Image

- Mu  
pri

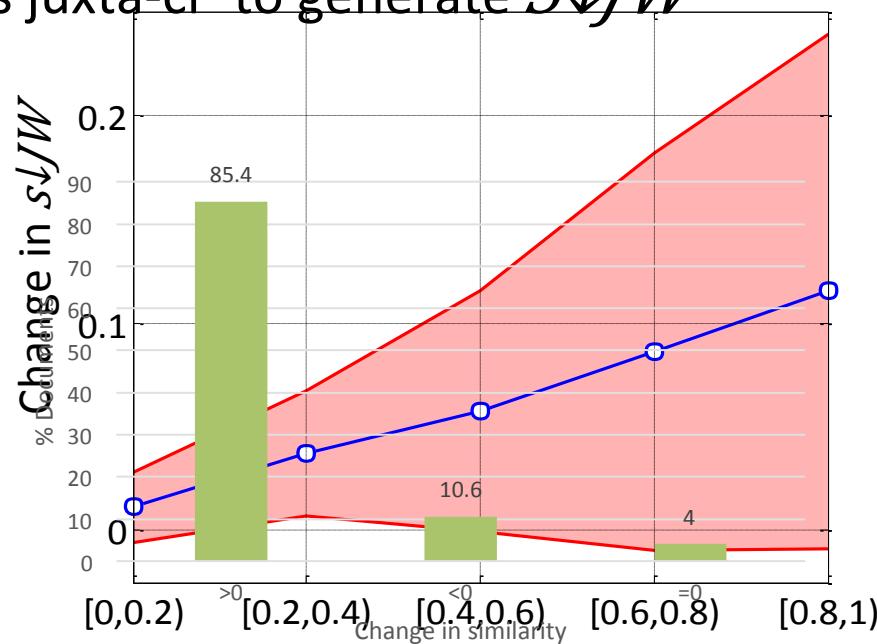
## Label creation

- Each
- 1: t $\epsilon$
- 72,3



## Filtering quality

- 6,775 test documents with ground truth text
- $S\downarrow JW$  similarity b/w OCR output and ground truth
- eMOP uses `juxta-cl*` to generate  $S\downarrow JW$



*BB↓noise*

# Discussion

## Summary

- Non-text OCR outputs suffice to
  - Identify text and noise in a document image
  - Estimate the document's overall quality
  - Improve OCR transcription performance when image processing based preprocessing is prohibitive

## Future work

- Diagnostic pipeline based on active learning
- Linguistic features can be explored

# Questions

THE  
ANDREW W.

MELLON  
FOUNDATION