Automatic assessment of OCR quality in historical documents

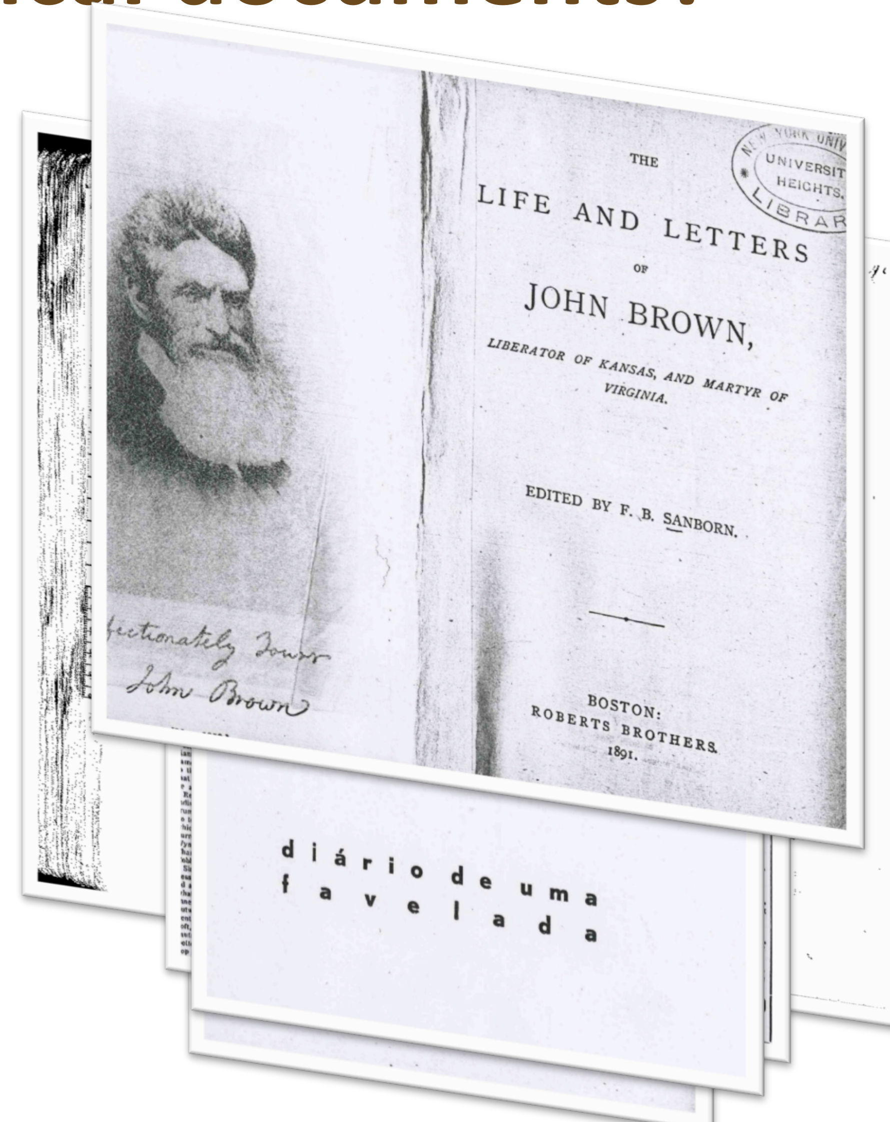
A. Gupta[1], R. Gutierrez-Osuna[1], M. Christy[2], B. Capitanu[3]
L. Auvil[3], L. Grumbach[2], R. Furuta[1], and L. Mandell[2]
[1]Computer Science and Engineering, Texas A&M University
[2]Initiative for Digital Humanities, Texas A&M University
[3]Illinois Informatics Institute, University of Illinois 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What are historical documents?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Digitizing historical documents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

Early modern OCR project (eMOP)

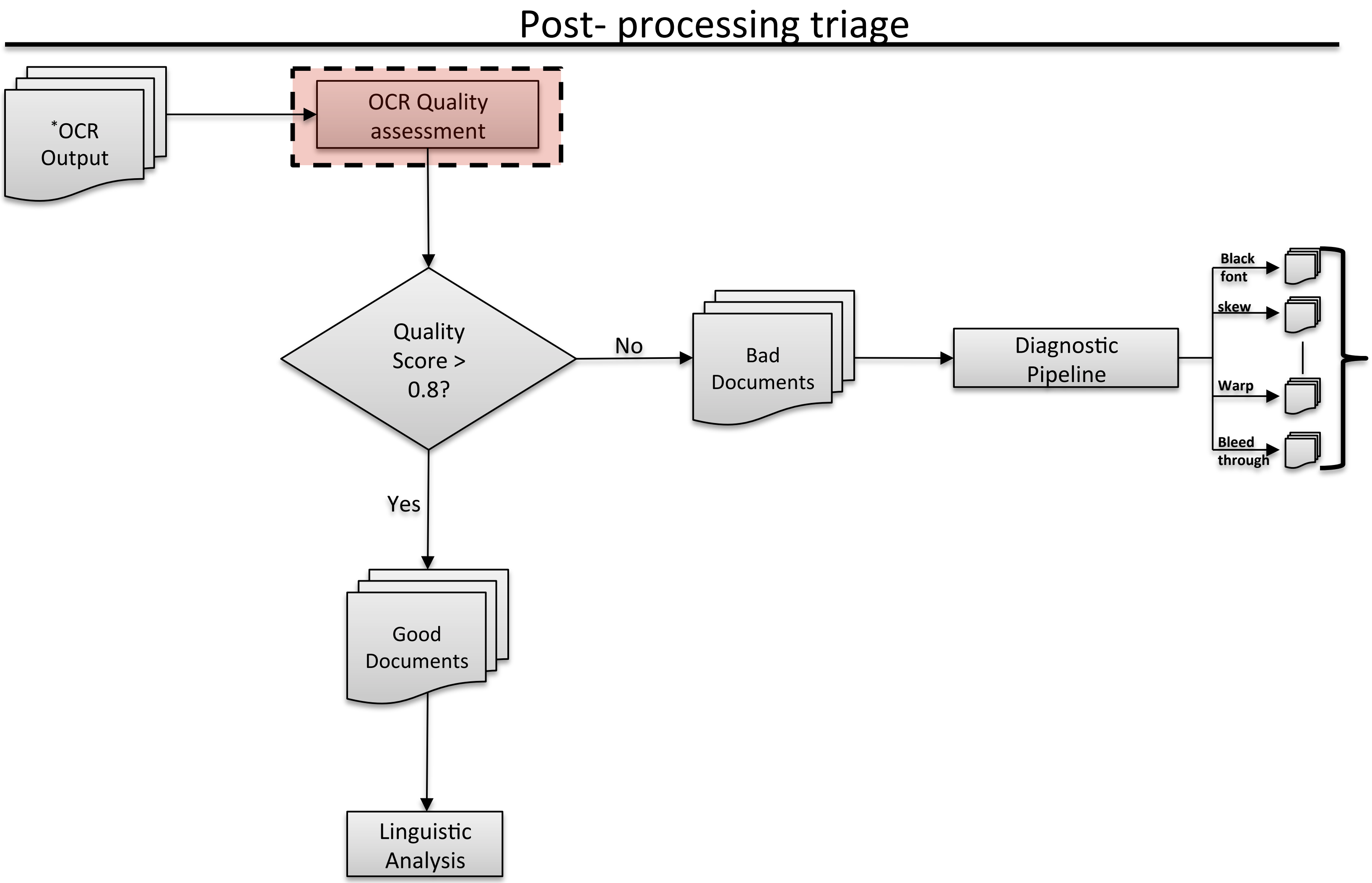
Goal

- 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?



*hOCR output from Tesseract OCR.

Our approach

Post-process OCR output

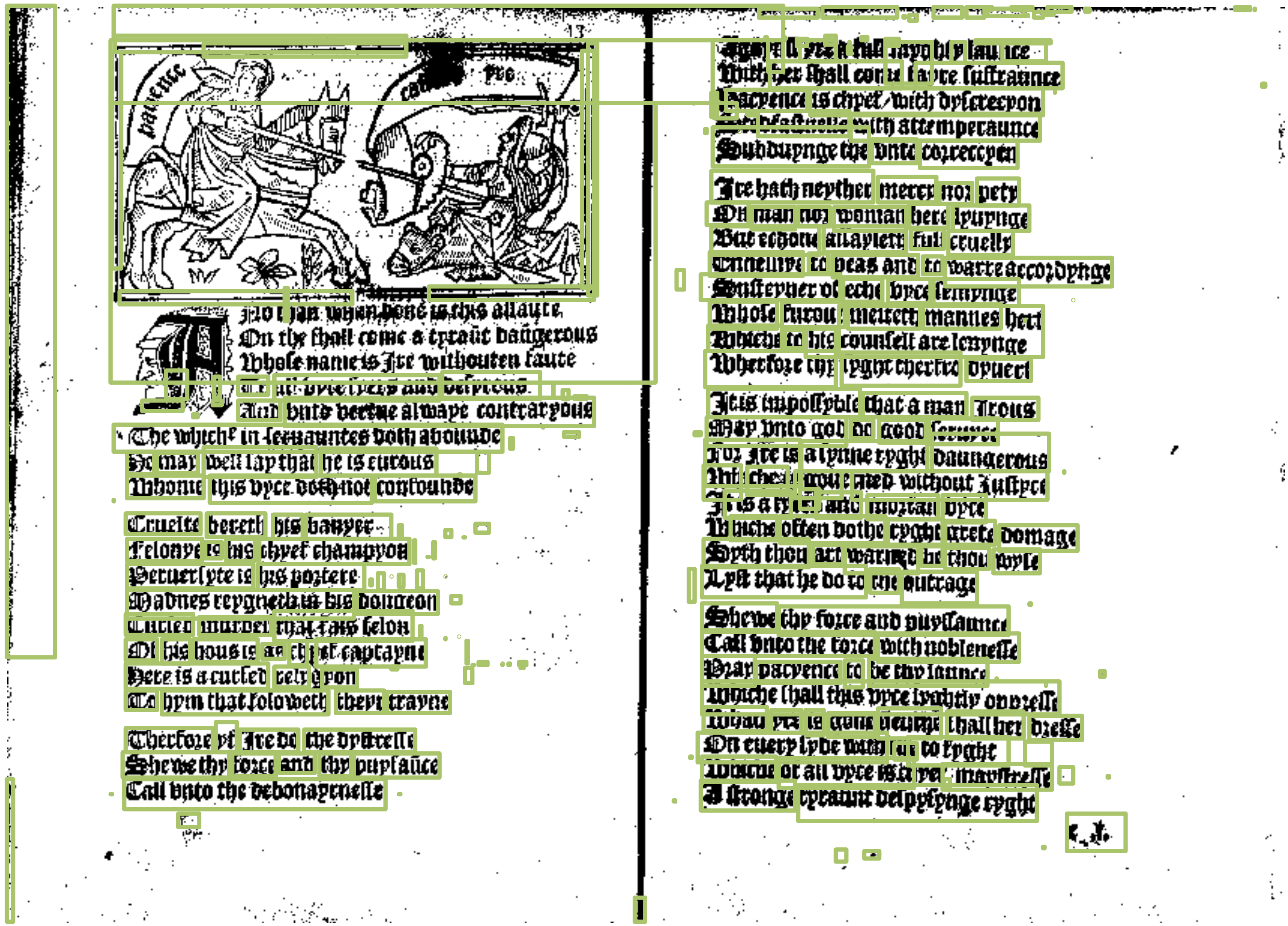
- Page segmentation result such as bounding box (BB) coordinates
- OCR word confidence

Build ML models to remove noise

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

$$\text{Quality}{\downarrow}\text{OCR} \propto 1/\%\ \text{noise BBs}$$

Language-agnostic approach



On the shall come a tyraũt daũgerous
Whose name is Ire withouten faute
And vnto vertue alwaye contraryous
The whiche in seruauntes doth abounde
He may well say that he is eurous
Whome this vyce doth not confounde

Cruelte bereth his banner
Felonye is his chyef champyon
Peruersyte is his portere
Madnes regneth in his donieon
Cursed murder that false felon
Of his hous is the chyef captayne
Here is a cursed relygyon
To hym that foloweth theyr trayne

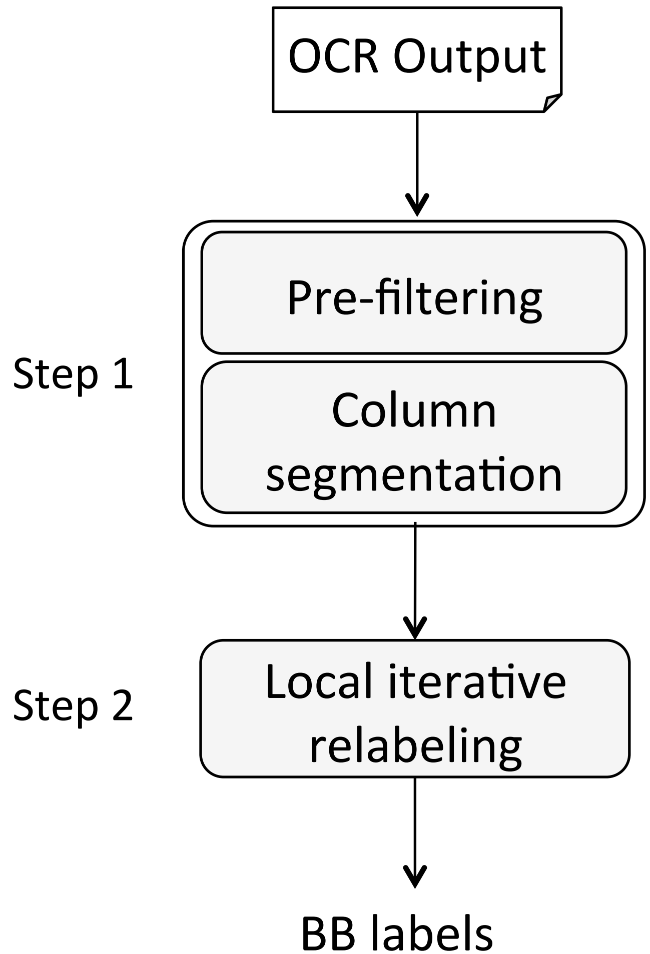
Therfore yf Ire do the dystresse
Shewe thy force and thy puyssaũce
Call vnto the debonayrnesse

Pacyence is chyef / with dyscrecyon
And mekenesse with attemperaunce
Subduynge the vnto correccyon

Ire hath neyther mercy nor pety
On man nor woman here lyuynge
But echone assayleth full cruelly

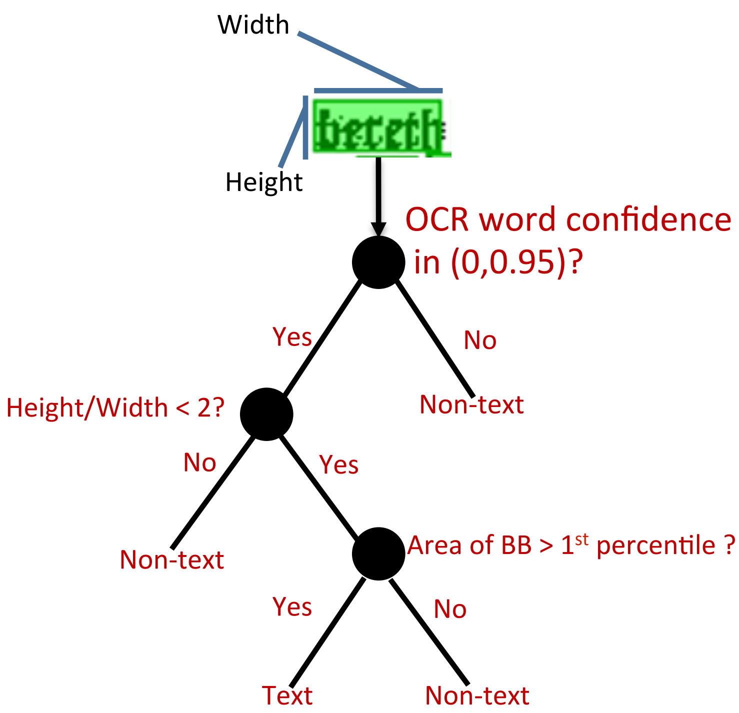
Shewe thy force and puyssaunce
Call vnto the force with noblenesse

Quality assessment algorithm



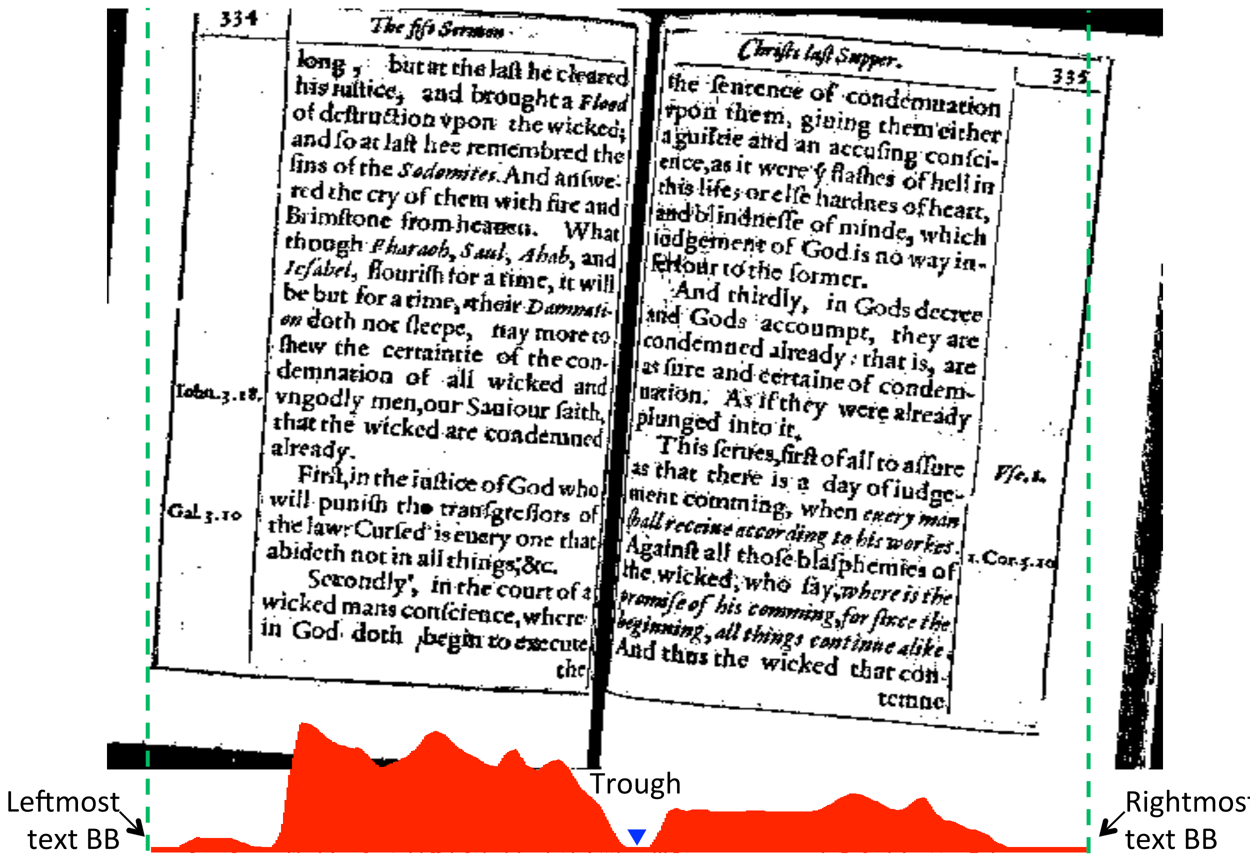
Prefiltering

- 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

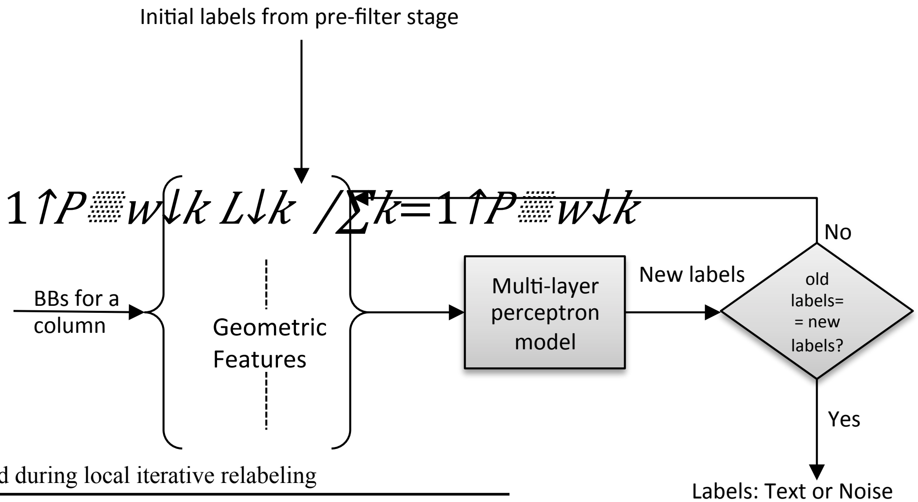
– Extract individual column and then process each column



Local iterative relabeling

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

$$\sum_{k=1}^{P} w_k L_k \Big/ \sum_{k=1}^{P} w_k$$



Features used during local iterative relabeling

Features	Description
S	Score from nearest neighbors ; see eq. (1)
C_{OCR}	OCR word confidence*
H/W	Height-to-width ratio of BB*
A	Area of BB*
H_{norm}	Normalized height: $H_{norm} = (H - H_{med})/H_{IQR}$

Final output

Results

Dataset

Label refinement: local iterative relabeling.

– Binarized page images

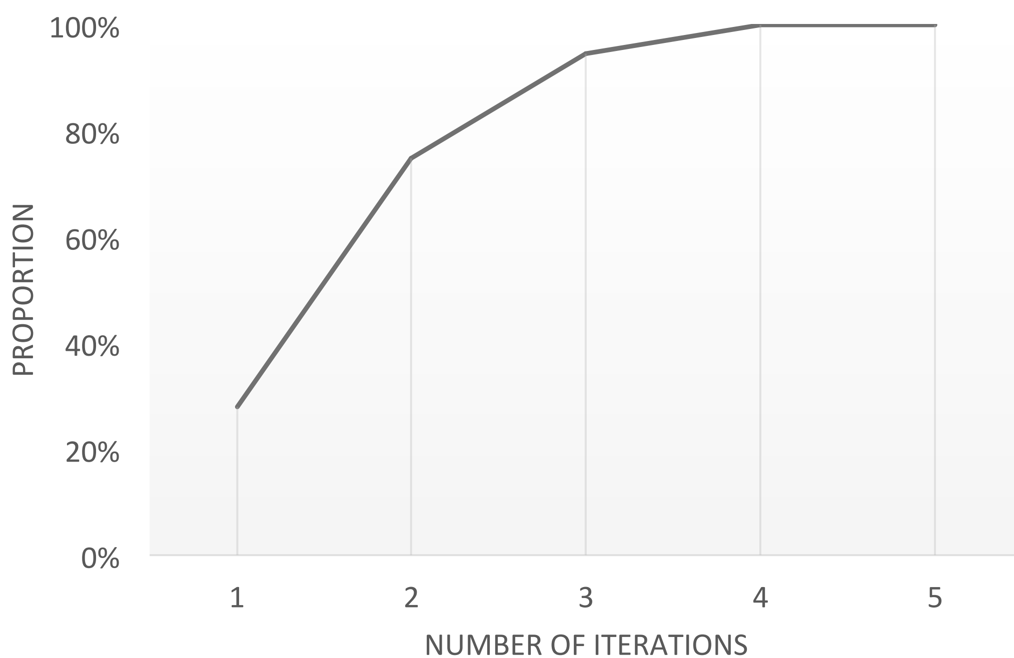
– Image … P corpora.

- Mu… kew; warping; pri…

Label cr…

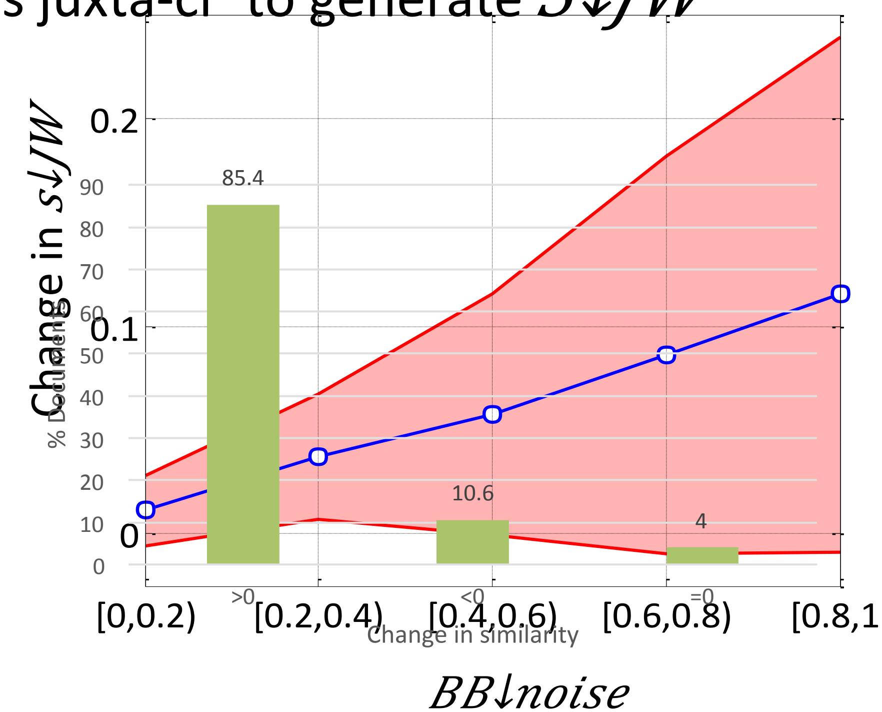
– Each … noise and 1: te…

– 72,3…



Filtering quality

- 6,775 test documents with ground truth text
- S_{JW} similarity b/w OCR output and ground truth
- eMOP uses juxta-cl* to generate S_{JW}



*implementations from juxtacommons.org

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