

# Toward a **Big Data** Analysis System for **Historical Newspaper Collections** Research

Sandeep Puthanveetil Satheesan, Bhavya, Adam Davies, Alan B. Craig, Yu Zhang, and ChengXiang Zhai

*Platform for Advanced Scientific Computing (PASC)  
Conference, Basel, Switzerland - June 27, 2022*



UNIVERSITY OF  
**ILLINOIS**  
URBANA - CHAMPAIGN

**XSEDE**

Extreme Science and Engineering  
Discovery Environment

**FRESNO**  **STATE**

# Introduction

- Large digitized historical newspaper collections are now available
  - *Chronicling America*
  - *ProQuest Historical Newspapers™*
  - *Europeana Newspapers*
  - *The British Newspaper Archive*
  - ...
- We now can ask complex and broad **questions of historical relevance** as depicted in the **newspapers**
- Previously, one had to visit a library or an archive and perform **extensive** and **tedious manual research** to access this type of information
- Gap in availability of **historical newspaper collections** vs. **big data systems** for processing and analyzing such collections, and extracting relevant information
- We propose a scalable and customizable **big data analysis system** using **HPC** and **AI analysis tools** for images and text



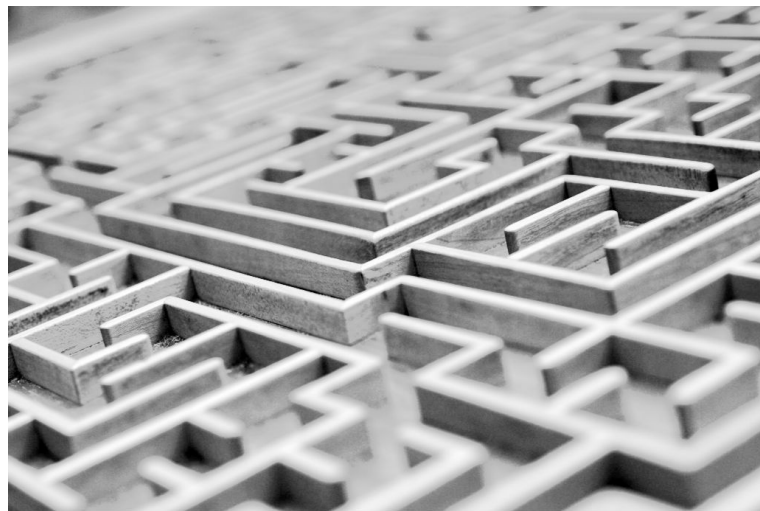
# Motivating Question

- How has the concept of *juvenile delinquency* been **socially constructed** over time in news media?
  - Our system has been developed in collaboration with Dr. Yu Zhang's research on this topic
  - Recorded rates of juvenile delinquency in the US and UK increased substantially in the early/mid 19th century. Was this due to a(n)...
    - **objective** increase in criminal juvenile **behavior** or resulting harm?
    - **subjective** change in **social construction** of juvenile criminality?
  - Millions of historical news articles could provide novel insights
    - Only feasible to analyze this volume of data with **HPC** and **AI analysis tools** for images and text



# Challenges

- Designing efficient, intelligent algorithms that can quickly process **millions of newspaper records** (hundreds of terabytes of data)
- **Segmenting historical newspaper images** that comes with variations and imperfections, **differentiating different content types** (e.g. headlines, articles, ads), and **extracting text** through OCR/re-OCR
- Retrieving **relevant information** from heterogeneous newspaper collections
- Customizing **image** and **text analysis** tools to support researchers in **interactively analyzing** and **visualizing** information in real-time



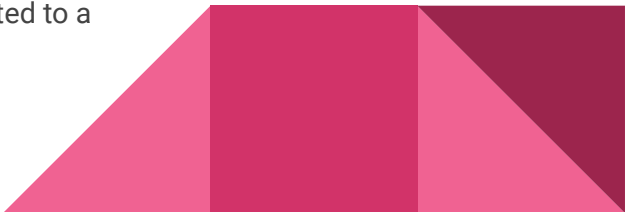
# Our Contributions

- Investigate **design principles** for developing a big data analysis system for newspaper collections research
- **Qualitatively evaluate** current AI techniques related to processing newspaper images and texts
- Use flexible text analysis operators to **support diverse analysis workflows**, enabling social scientists to investigate a wide variety of **novel research questions**
- Facilitate **human-in-the-loop collaboration** between social scientists and AI systems



# Related Work

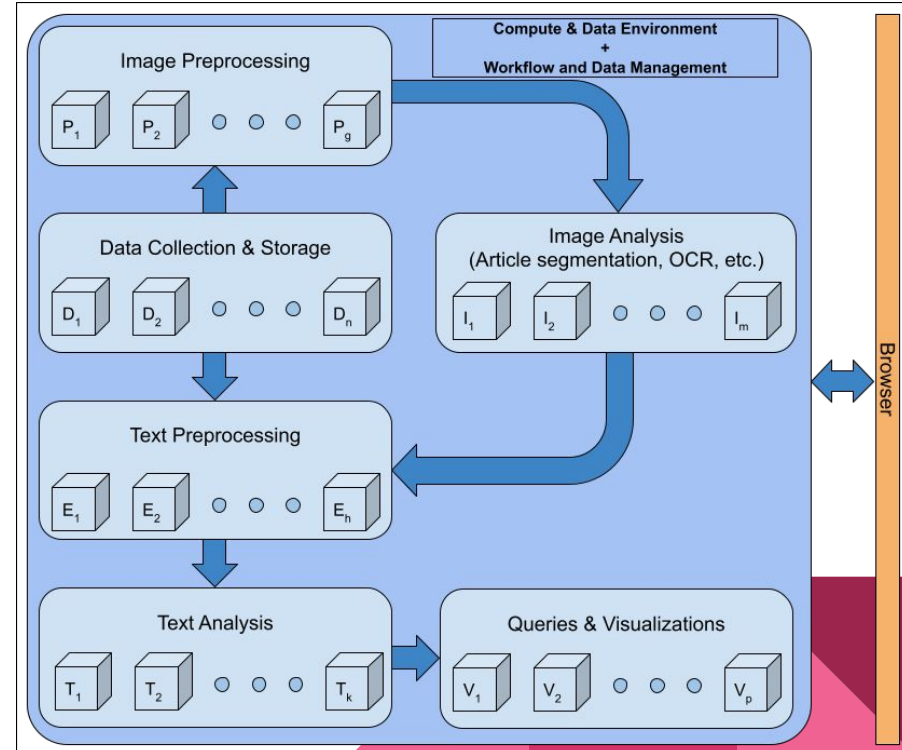
- **The Valley of the Shadow** (<https://valley.lib.virginia.edu/>)
  - Digital archive containing thousands of original letters and diaries, newspapers and speeches, census and church records from two U.S. counties in the Civil War era
  - Outcome of tireless manual effort of collecting, curating, and transcribing documents
- **Newspaper Navigator** (<https://news-navigator.labs.loc.gov/>)
  - Public dataset containing extracted headlines and visual content processed on 16.3 million historical newspaper pages on the LOC-CA
  - Multiple data products, a scalable pipeline, and a fine-tuned deep learning model to recognize content like photographs, cartoons, headlines, advertisement, etc.
  - Our system has many similar components, but also includes a variety of interactive text analysis operators and visualizations (e.g. temporal semantic analysis, human-in-the-loop text classification, topic modeling, etc.)
- **ShiCo** ([Kenter et al., 2015](#); [Martinez-Ortiz et al., 2016](#))
  - Uses historical word embeddings to find and visualize changing vocabulary related to a concept over time
  - We extend this tool and integrate it with other text analysis operators



# System Design & High-level Architecture

- Design Principles

- **Reuse** existing software
- **Modularize** system
- **Move compute to data** via application of Docker containers
- **Integrate** innovative and existing text analysis operators that users can **flexibly** combine
- Support **human-in-the-loop** AI/human collaboration



# Data Collection

- Collect data from relevant sources and download it into suitable data and compute resources
- Library of Congress Chronicling America (LOC-CA)
  - Historical American newspapers in the public domain from **1777 to 1963** – more than **19.6 million pages** of digitized U.S. newspapers
  - Downloaded **18.4 million text files** and metadata files and **500,000 scanned newspaper images** LOC-CA Bulk Data API
- ProQuest Historical Newspapers™
  - Digital archive offering full-text and full-image articles for newspapers dating back to the **18<sup>th</sup> Century**
  - Digitized newspaper data that are not in the public domain - due to restrictions on the data, we had to take additional steps to ensure compliance
  - Using about **58 million documents** from this collection





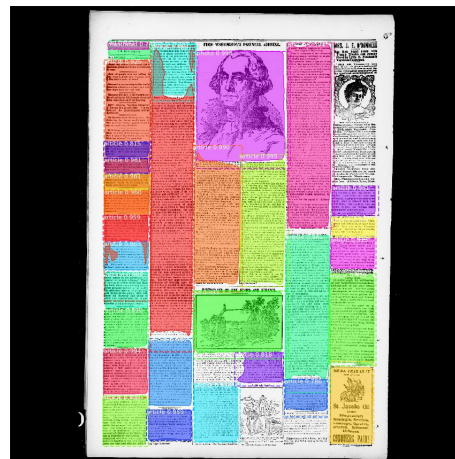
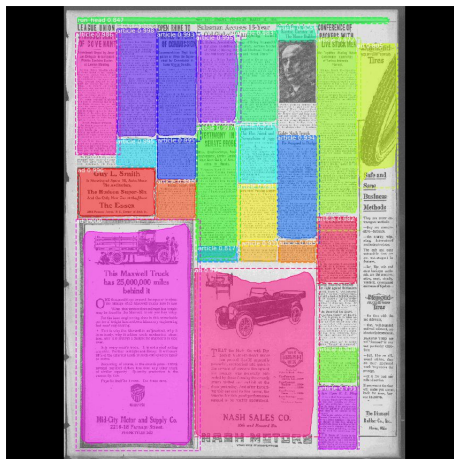
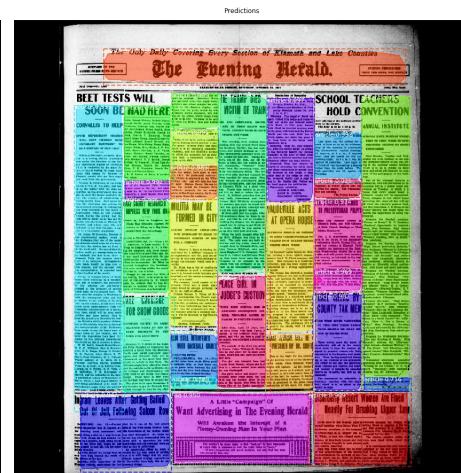
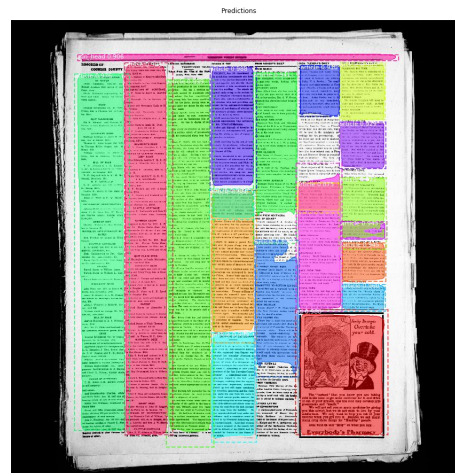
# Image Analysis

- Methods used to perform OCR or newspaper segmentation or any other operation performed on the image to retrieve some useful information from it
- Analyzed OCR metadata to understand the data and to recreate the reading order and segment articles from text files ([Satheesan et al., 2019](#)). Later focussed on newspaper segmentation of the LOC-CA collection and re-OCR
- Newspaper region segmentation - Mask-RCNN model
  - ResNet-101 backbone
  - Pre-trained on MS COCO dataset
  - **Training set: 60 newspaper images**; 1728 manually annotated regions, **validation set: 25 newspaper images**; 801 manually annotated regions (latest count)
  - Classes: **Advertisement, article, banner, masthead, photo, and run head**
  - With data augmentation - horizontal flipping, rotation, added brightness and blurring pixels



# Image Analysis: Newspaper Segmentation

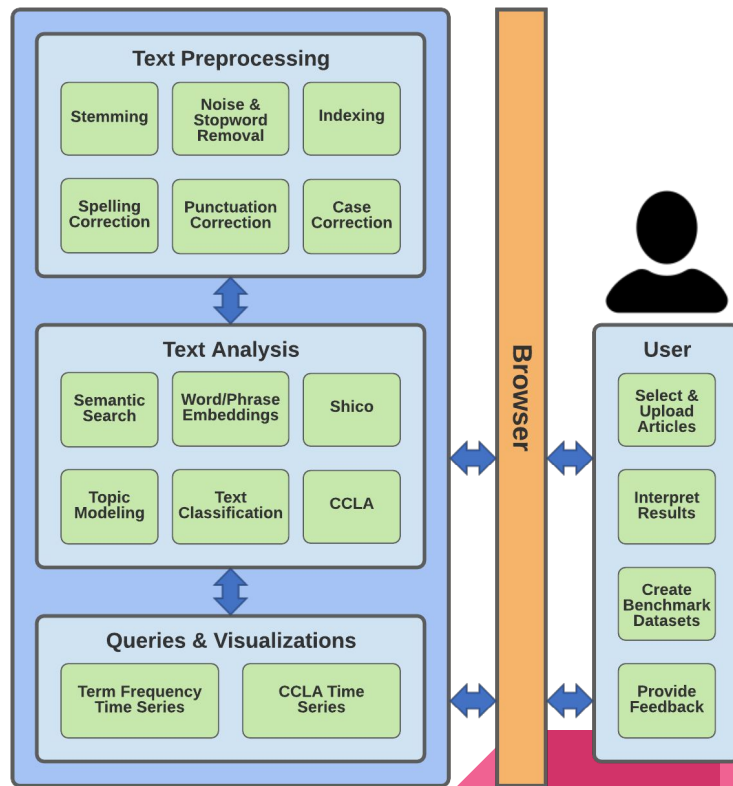
Output of our current  
best performing  
Mask-RCNN model on  
some example  
newspaper images  
showing different  
segmented regions





# Text Analysis Architecture

- Primary Modules:
  - Text Preprocessing
  - Text Analysis
  - Queries and Visualizations
- Submodules implemented as Python Flask microservices
- Human-in-the-loop (HITL) approach to provide user feedback



# Document-level Text Analysis

- Identify articles about a given concept
  - Unsupervised approach: automatically **discover** related terms and topics
    - **Query expansion** with Word2vec embeddings
    - **Topic modeling** with LDA
  - Online/supervised approach: **human-in-the-loop** document classification

Word: juvenile\_delinquency

Models: 1990.w2v

Number of words: 10

GET THE SIMILAR WORDS

Model: 1990.w2v

Word: juvenile\_delinquency

- delinquency 0.7948
- drug addiction 0.79210
- drug abuse 0.7775
- teen pregnancy 0.7653

GENERATE THE QUERY

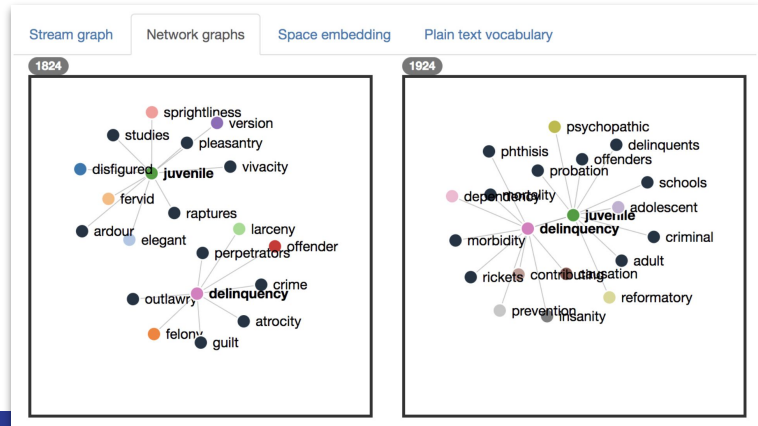
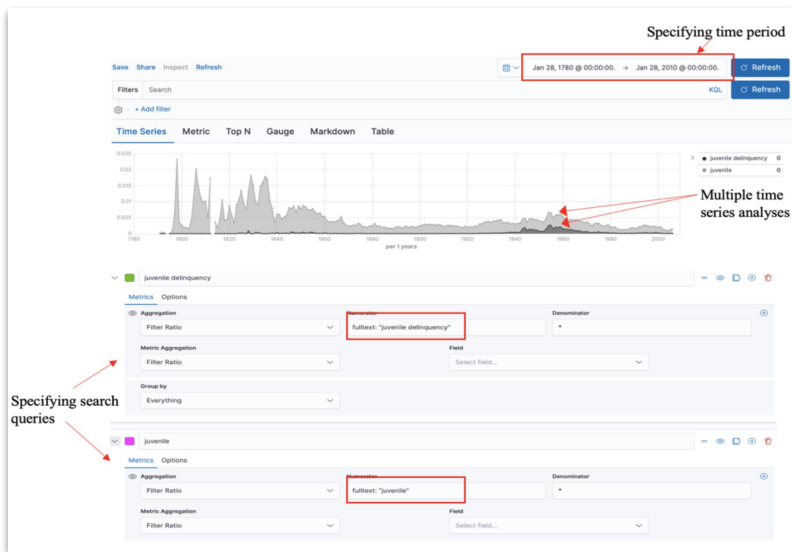
## LOCAL & AND PROVINCIAL INTELLIGENCE

LOCAL AND PROVINCIAL INTELLIGENCE. On the 1st of the month which bears the above  
inspiring title and which by the way has not been so free from dirt for number  
of year as it has been on this occasion was held on Tuesday last in Chapel street  
Saloon of the city there was great abundance and they were almost freely given  
away very few purchasers being turned back if they any thing in the shape of  
money. There was also numerous show of both cattle and horses but prices were  
remarkably low of horses the quality was the worst we ever recollect to have  
seen. To see was very numerous and imposing display of shows for the juvenile  
class of fair visitors comprising dwarfs and giants of every height and amongst  
other most wonderful curiosities all to be seen for one penny for there was not  
single show keeper that charged more for sight of his exhibition we suppose to  
accommodate them to the depression of the times were either three or four  
animal to be sold before they were not all however of one kind as one of them

Not Relevant  Relevant

# Word-level Text Analysis

- Characterize historical semantic dynamics
  - **Term-frequency analysis:** find proportion of documents that use given terms each year
  - Word embedding (Word2vec) analyses:
    - **Shico:** automatically find and visualize related terms over time
    - **Cross-Context Lexical Analysis (CCLA):** quantify semantic consistency over time to detect periods of stability or change



# Preliminary Experiments

- **Recall *Motivating Question***: How has the concept of ***juvenile delinquency*** been **socially constructed** over time in news media?
  - Start by...
    - **retrieving** news articles describing *juvenile delinquency*
    - **detecting** when social construction of *juvenile delinquency* changes
- Experiments:
  - **Human-in-the-loop (HITL) Text Classification**: manually annotate news articles that describe *juvenile delinquency*, then train machine learning classifiers to find more
  - **Cross-Context Lexical Analysis (CCLA)**: quantify periods of semantic change in *juvenile delinquency*-related terms, and look for periods of rapid change

# HITL Text Classification

	Accuracy	Precision	Recall	F1
Random	0.466±0.05	0.136±0.02	0.38 ±0.12	0.199±0.04
Stratified Random	0.688±0.05	0.175±0.1	0.184±0.1	0.176±0.09
Logistic Regression	0.842±0.04	0.631±0.34	0.276±0.19	0.354±0.23
XGBoost	<b>0.868±0.02</b>	<b>0.754±0.14</b>	<b>0.442±0.12</b>	<b>0.541±0.1</b>

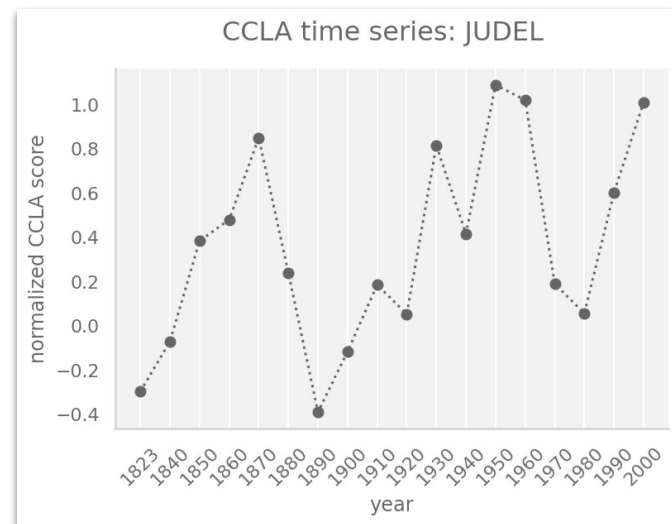
**Human-in-the-loop** text classification helps identify relevant articles

- Unsupervised search method identified 237/92k articles (1790-1830)
- 19/237 manually annotated as “relevant”
- Trained machine learning models on the annotated dataset
- Optimize Effort vs. Accuracy tradeoff




# CCLA (Cross-Context Lexical Analysis)


- How have terms denoting *juvenile delinquency* changed over time?
  - **Aggregate** terms that describe concept of *juvenile delinquency*
  - **Train** Word2vec models (for each decade)
  - **Compute** (normalized) CCLA score for each successive pair of decades
  - **Visualize** CCLA time series
    - Peak/plateau: (relative) semantic **stability**
    - Dip: (relative) semantic **change**



# Computational Performance and Scalability (1/2)

- **Downloading Datasets (LOC-CA):** developed specialized program to download data batches in parallel using the LOC-CA Bulk Data API
    - Used 1 compute node from Pittsburgh Supercomputing Center Bridges-2 (AMD EPYC™7742 - 128 cores and 256GB RAM per node)
    - 128 parallel download tasks with 0.01 seconds sleep time after each HTTP request
    - Full download took ~15.3K core-hours
  - **Image Analysis:** fine-tuned Mask R-CNN for newspaper article segmentation using LOC-CA
    - Used 1 compute node from NCSA Hardware Accelerated Learning (HAL) Cluster (NVIDIA® Tesla® V100 - 1 GPU, 16 CPU cores per node, 1.2GB RAM per CPU core)
    - Avg. image resolution is 5595 x 7653 px; rescaled to 800 x 1024 px during training
    - Training time ≈ 8-10 hours
- 

# Computational Performance and Scalability (2/2)

- **Text Analysis:** preprocessed and analyzed text of ProQuest Historical Newspapers™
    - Used *in-house* Intel Xeon Silver 4210 CPU (*due to licensing restrictions*)
    - **Offline operations:**
      - Indexed 175 GB data took ~1wk on 2 cores
      - Trained LDA topic model in ~2d (on 8 cores), word embedding models in ~2d (on 4 cores), and text classifiers in ≤30min (on 1 core)
    - **Online operations:**
      - Text classifiers and CCLA operate almost instantaneously (<10ms on 1 core)
      - Kibana or ShiCo can take up to 4 minutes for some complex queries (e.g. containing many terms or search operators)
- 

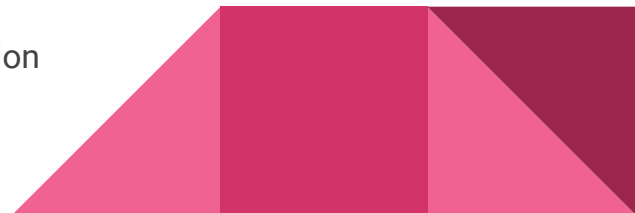
# Conclusion & Future Work

- Presented the **design, implementation, and preliminary evaluation** of a big data analysis system for historical newspaper collections
- Applied AI analysis tools to historical newspapers to enable **investigation of novel social science research questions**
- *Remaining Challenges:*
  - **Support many users:** improve efficiency and parallelization of online analysis operators to support queries, visualizations, and feedback from many users simultaneously
  - **Integrate system:** unify all modules into single system, use results of downstream analyses to improve upstream preprocessing/analysis modules, etc.
  - **Improve robustness:** adapt term-level text analysis operators (e.g. ShiCo, CCLA) to be less sensitive to imperfect inputs (e.g. errors in OCR and word segmentation)
  - **Study social construction:** apply our system to study the historical social construction of *juvenile delinquency*



# Acknowledgements

- **Extreme Science and Engineering Discovery Environment (XSEDE) and Bridges System at Pittsburgh Supercomputing Center** - Primary data storage and compute resources
- **XSEDE Extended Collaborative Support Service (ECSS)** - Research software development and consulting expertise
- **Illinois Campus Cluster** - Additional data storage and compute resources
- **Hardware Accelerated Cluster (HAL) at NCSA** - Deep Learning model training and evaluation
- **Cline Center for Advanced Social Research and University of Illinois at Urbana-Champaign** - Access to ProQuest Historical Newspapers™
- **Benjamin Meier** - Correspondence regarding newspaper article segmentation



# Questions?

Feel free to contact  
{sandeeps, bhavya2, adavies4}@illinois.edu

