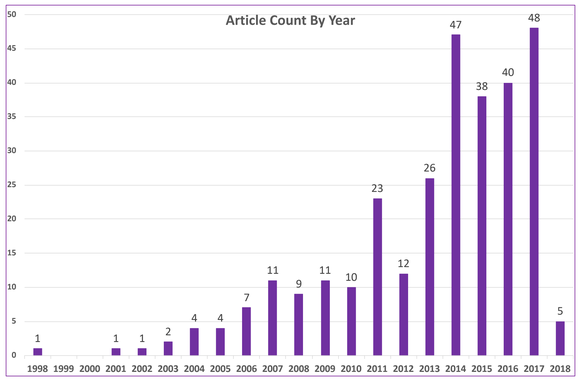
# Big Data Analytics for Historical Document Processing

James Patrick Philips
Mentor: M. H. N. Tabrizi
Department of Computer Science

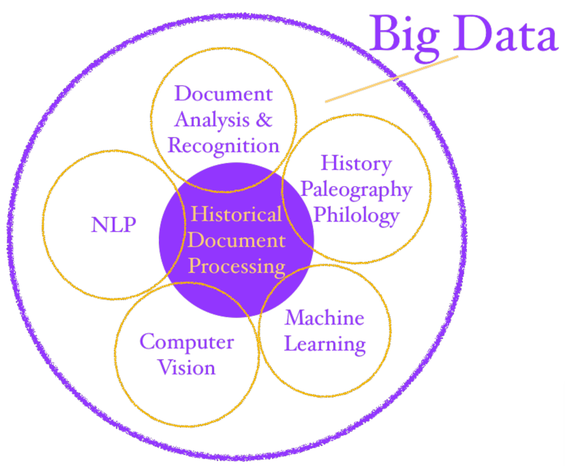
## INTRODUCTION

Historical Document Processing is the process of digitizing written material from the past for future use by historians and other scholars. It incorporates algorithms and software tools from various subfields of computer science, including computer vision, document analysis and recognition, natural language processing, and machine learning, to convert images of ancient manuscripts, letters, diaries, and early printed texts automatically into a digital format usable in information retrieval systems. Within the past twenty years, as libraries, museums, and other cultural heritage institutions have scanned an increasing volume of their historical document archives, the need to transcribe the full text from these collections has become acute. Big Data Analytics and infrastructure will be an essential tool in this field. This study compares performance analysis of two OCR systems, discusses HDP workflow, and highlights the role of OCR software in a RESTful API for HDPaaS.

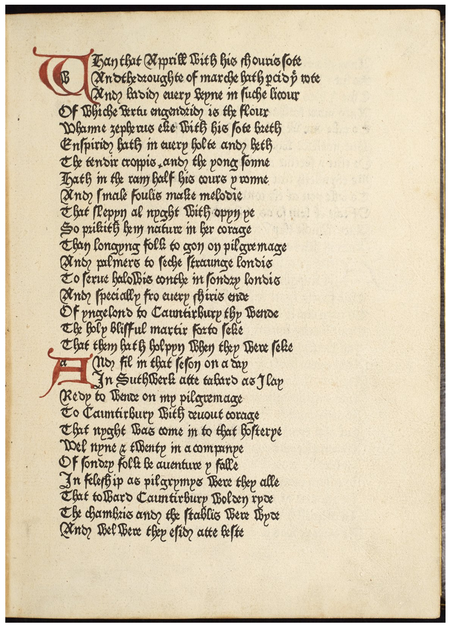
## LITERATURE REVIEW



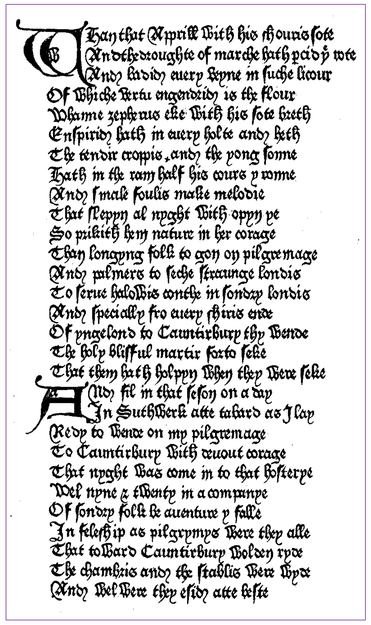

## BIG DATA HDP: A HYBRID FIELD



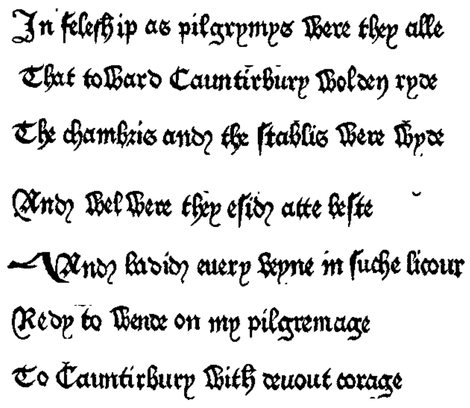

## HISTORIC DOCUMENT PROCESSING WORKFLOW



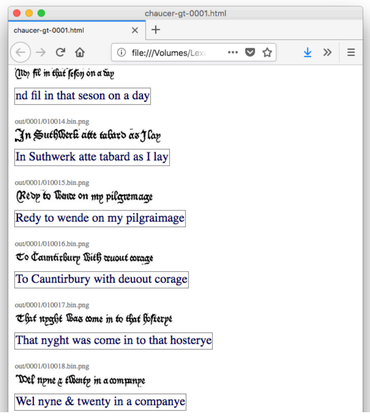
1. ORIGINAL IMAGE



2. BINARIZED IMAGE



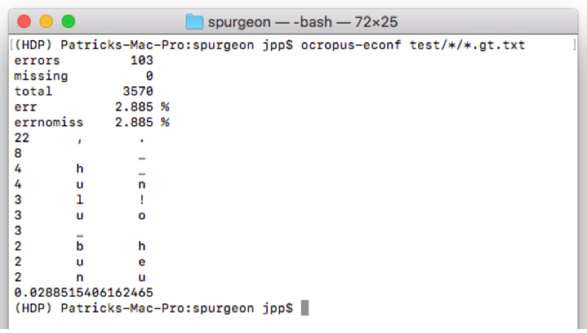

3. LINE SEGMENTATION




4. GROUND TRUTH ANNOTATION



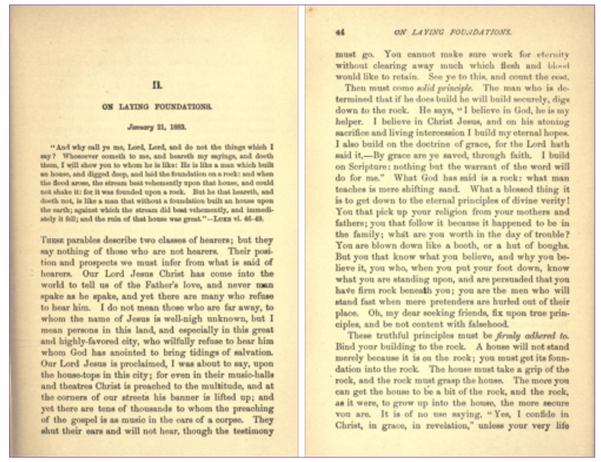

5. TRAINING




6. ANALYTICS

## CASE STUDY

- Used images from the sermons of C.H. Spurgeon
- Compared OCR performance of OCRopy and Ocular OCR software
- OCRopy uses BLSTM neural network & Ocular uses unsupervised machine learning with a multiple models.



II.

ON LAYING FOUNDATIONS.

January 31, 1883.

"And why call ye me, Lord, Lord, and do not the things which I say? Whosoever cometh to me, and heareth my sayings, and doeth them, I will shew you to whom he is like: He is like a man which built an house, and digged deep, and laid the foundation on a rock: and when the flood arose, the stream beat vehemently upon that house, and could not shake it: for it was founded upon a rock. But he that heareth, and doeth not, is like a man that without a foundation built an house upon the earth; against which the stream did beat vehemently, and immediately it fell; and the ruin of that house was great."—Luke vi. 46-49.

These parables describe two classes of hearers; but they say nothing of those who are not hearers. Their position and prospects we must infer from what is said of hearers. Our Lord Jesus Christ has come into the world to tell us of the Father's love, and never man spake as he spake, and yet there are many who refuse to hear him. I do not mean those who are far away, to whom the name of Jesus is well-nigh unknown, but I mean persons in this land, and especially in this great and highly-favored city, who wilfully refuse to hear him whom God has anointed to bring tidings of salvation. Our Lord Jesus is proclaimed, I was about to say, upon the house-tops in this city; for even in their music-halls and theatres Christ is preached to the multitude, and at the corners of our streets his banner is lifted up; and yet there are tens of thousands to whom the preaching of the gospel is as music in the ears of a corpse. They shut their ears and will not hear, though the testimony

44 ON LAYING FOUNDATIONS.

must go. You cannot make sure work for eternity without clearing away much which flesh and blood would like to retain. See ye to this, and count the cost.

Then must come *solid principle*. The man who is determined that if he does build he will build securely, digs down to the rock. He says, "I believe in God, he is my helper. I believe in Christ Jesus, and on his atoning sacrifice and living intercession I build my eternal hopes. I also build on the doctrine of grace, for the Lord hath said it,—By grace are ye saved, through faith. I build on Scripture; nothing but the warrant of the word will do for me." What God has said is a rock: what man teaches is mere shifting sand. What a blessed thing it is to get down to the eternal principles of divine verity! You that pick up your religion from your mothers and fathers; you that follow it because it happened to be in the family; what are you worth in the day of trouble? You are blown down like a booth, or a hut of boughs. But you that know what you believe, and why you believe it, you who, when you put your foot down, know what you are standing upon, and are persuaded that you have firm rock beneath you; you are the men who will stand fast when mere pretenders are hurled out of their place. Oh, my dear seeking friends, fix upon true principles, and be not content with falsehood.

These truthful principles must be *firmly adhered to*. Bind your building to the rock. A house will not stand merely because it is on the rock; you must get its foundation into the rock. The house must take a grip of the rock, and the rock must grasp the house. The more you can get the house to be a bit of the rock, and the rock, as it were, to grow up into the house, the more secure you are. It is of no use saying, "Yes, I confide in Christ, in grace, in revelation," unless your very life

**Results Comparison**

| | # Training Lines | # Test Lines | Training Time | CER |
|---|---|---|---|---|
| OCRopus (trained) | 2998* | 62 | 2 hours | 33.109% |
| OCRopus (default model) | NA | 3570 | NA | 2.885% |
| Ocular | 1785 | 1785 | 4 hours, 43 minutes** | 76.813% |

* *Some lines duplicated in training data per practitioner recommendations. A model was saved every 1000 lines*

** *Another training session with a training set 2x larger lasted 17 hours, 33 minutes*

- Based on my case study, I have found OCRopy to be a superior OCR system to Ocular due to accuracy and performance metrics
- If OCR software were used in a high performance cluster computing environment, HDP could be implemented with a RESTful API as a cloud service: HDPaaS.
- The extensive quantity of archival data in libraries necessitates a solution using Big Data analytics.

## ACKNOWLEDGEMENTS

James Patrick Philips wishes to thank his mentor, Dr. M.H.N. Tabrizi for his inspiration and guidance on this project. He also thanks Rebecca Stamilio-Ehret, Trey Cherry, and Timothy Boyd of Edgecombe Community College for their encouragement and support.