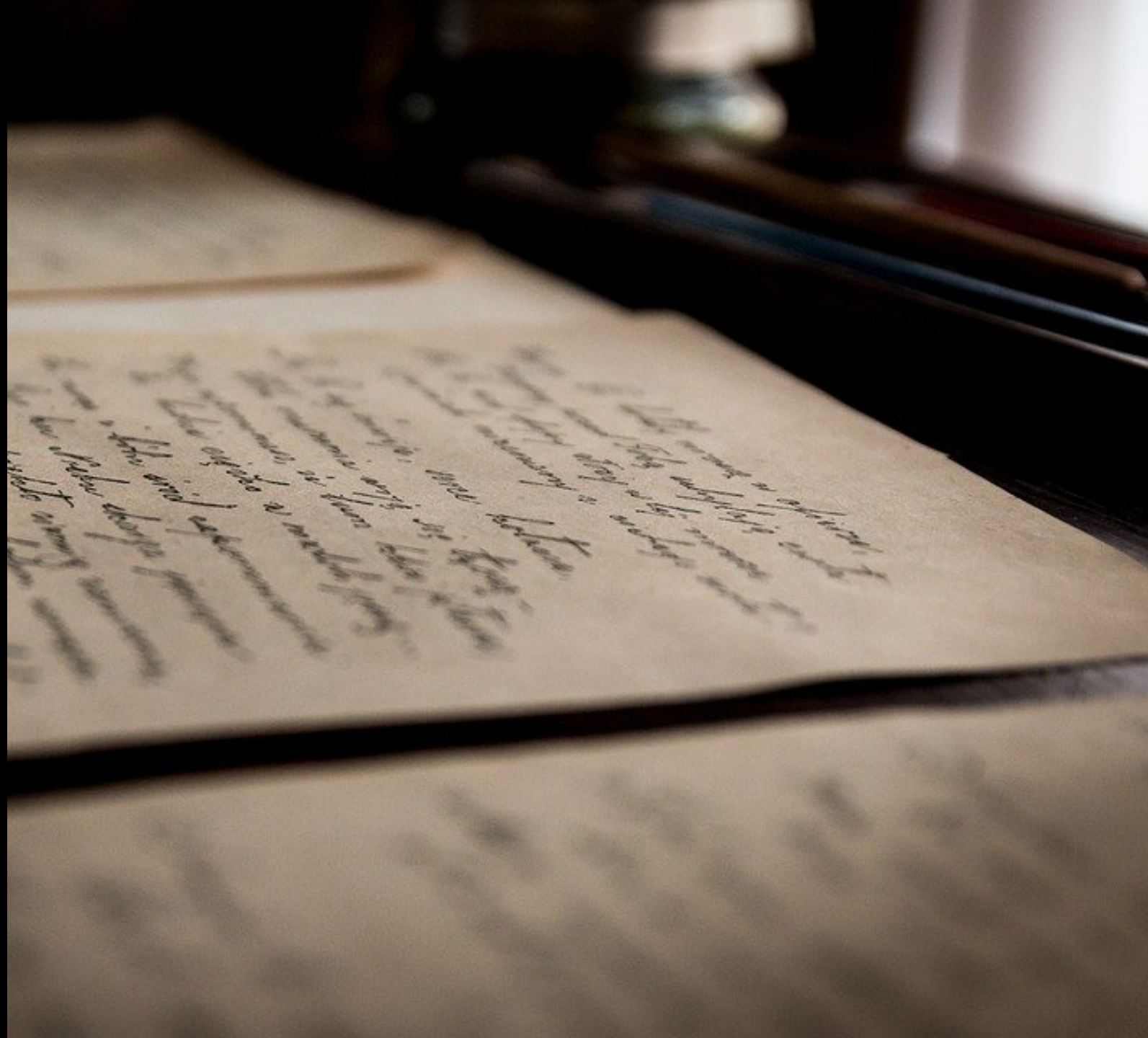
Cursive & Recursive:

Generating Transcriptions of Archival Documents Using Machine Learning



Week 1

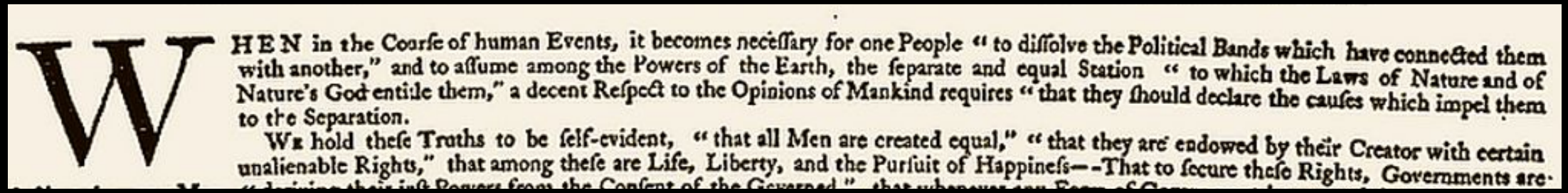
Buchanan Fellowship, Spring 2020

Vanderbilt University Library

Why bother with transcriptions anyway?

Make texts easier to read

- *Paleography* is the study of old writing to decipher, read, and date historical manuscripts
- Enable text to speech programs (screen readers) to help vision-impaired
- Enable machine translation



WHEN in the Courſe of human Events, it becomes neceſſary for one People "to diſſolve the Political Bands which have connected them with another," and to aſſume among the Powers of the Earth, the ſeparate and equal Station "to which the Laws of Nature and of Nature's God entitle them," a decent Reſpect to the Opinions of Mankind requires "that they ſhould declare the cauſes which impel them to the Separation.

We hold theſe Truths to be ſelf-evident, "that all Men are created equal," "that they are endowed by their Creator with certain unalienable Rights," that among theſe are Life, Liberty, and the Purſuit of Happineſs—That to ſecure theſe Rights, Governments are

Enable computer processing of content

- Digital humanities
- Data analysis

SALOON, CABIN, AND STEERAGE ALIENS MUST BE COMPLETELY MANIFESTED.

LIST OR MANIFEST OF ALIEN PASSENGERS FOR THE UNITED

...uired by the regulations of the Secretary of Commerce and Labor of the United States, under Act of Congress approved February 20, 1907, to be delivered

S. S. MAJESTIC. sailing from SOUTHAMPTON. June 16th, 1909

Enable computer processing of content

From a collection of ship manifests . . .

. . .To a searchable database

Trans-Atlantic Slave Trade - Database

Year range | Ship, nation, owner | Itinerary | Enslaved people | Dates | Captain and crew | Outcome | Source

Results | Summary statistics | Tables | Data visualization | Timeline | Maps | Timelapse

Showing 1 to 15 of 36,108 entries

Configure columns | Show 15 rows | Download

Voyage ID	Vessel name	Place where voyage began IMP	Principal place of purchase IMP	Principal place of slave landing IMP	Year arrived with slaves IMP	Slaves arrived 1st port	Captain's name
1	Pastora de Lima	Rio de Janeiro	Mozambique	Bahia, port unspecified	1817	290	Dias, Manoel José
2	Tibério	Bahia, port unspecified	Mozambique	Bahia, port unspecified	1817	223	Mata, José Maria da
3	Paquete Real	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	350	Ferreira, José dos Santos
4	Bom Caminho	Bahia, port unspecified	Quilimane	Bahia, port unspecified	1817	342	Dias, Domingos Francisco
5	Benigretta	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	516	
6	Voador	Bahia, port unspecified	Mozambique	Bahia, port unspecified	1817	515	
7	Formiga	Bahia, port unspecified	Malembo	Bahia, port unspecified	1817	204	Viana, Isidoro Antônio
8	Vigilante Africano	Pernambuco, port unspecified	Luanda	Bahia, port unspecified	1817	374	Amorim, José Gomes de
9	Constante	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	345	Narciso, Antônio
10	Comerciante	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	478	Braga, Isidoro Martins
11	Diligente	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	180	
12	Bonfim	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	542	Coelho, João Batista
13	Triunfo	Bahia, port unspecified	Luanda	Bahia, port unspecified	1817	503	
14	S Lourenço	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	471	Vilasboas, João da Silveira
15	Paquete da Bahia	Bahia, port unspecified	Cabinda	Bahia, port unspecified	1817	478	Almeida, Manoel Joaquim de

Previous | 1 | 2 | 3 | 4 | 5 | ... | 2408 | Next

https://slavevoyages.org

Transcription Options

Manual

- High accuracy
- Very slow and tedious
- Crowdsourcing can improve speed, but accuracy suffers

Automated

- High speed
- Lower accuracy
- Requires training sets (i.e., manual transcription of a subset)

Academic use cases

smartphone photo of letters from an archive

scanned historical documents

pdf download from online library

Commercial use cases

check deposit

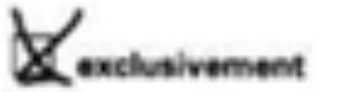
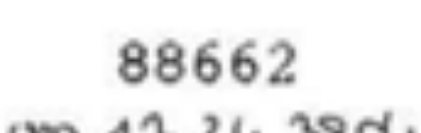
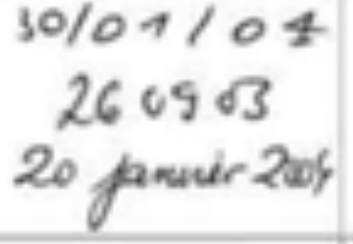
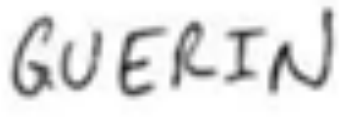
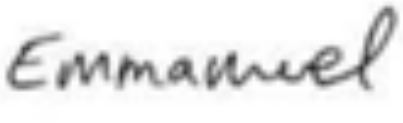
travel reimbursement

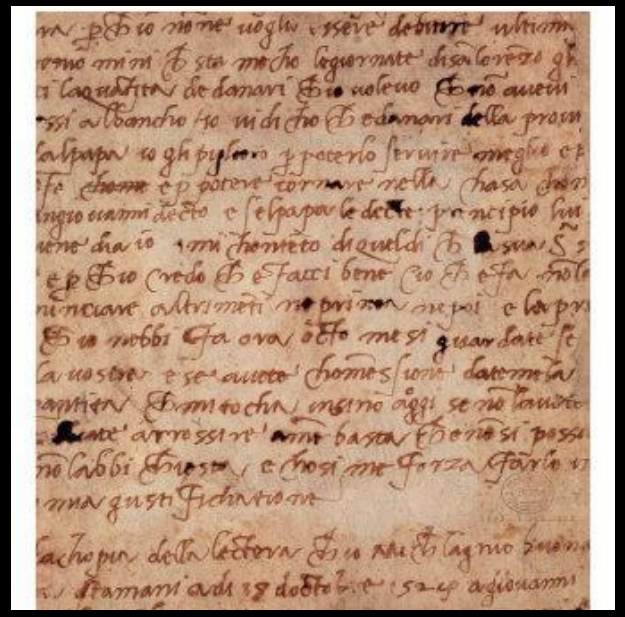
extract data from tax returns

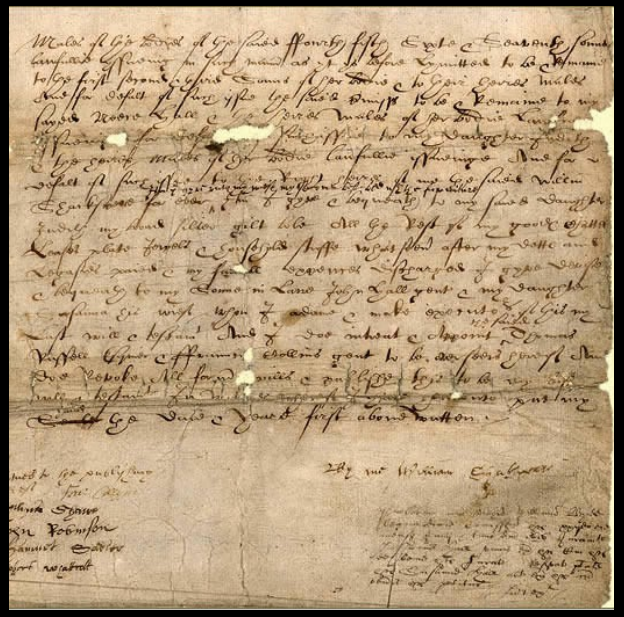
legal discovery

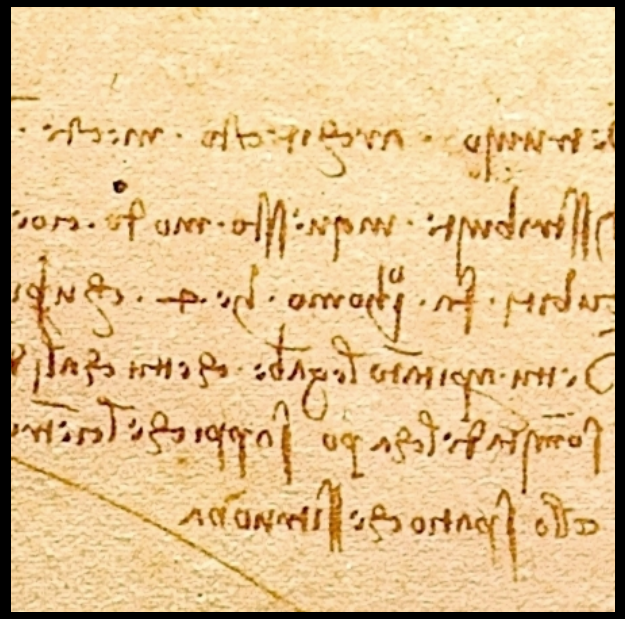
post office sorter

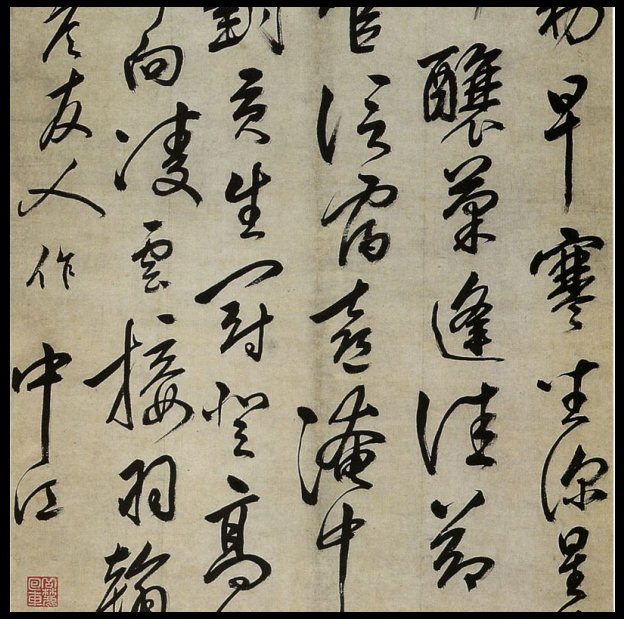
Tools must handle scenarios like these

		Millet Christine		
Box to tick	Overwriting detection	Machine print	Combs	Pre-box
				
Signature detection	Numeric (currency, postcode, phone…)	Dates	Unconstrained handprint	Cursive or freehand









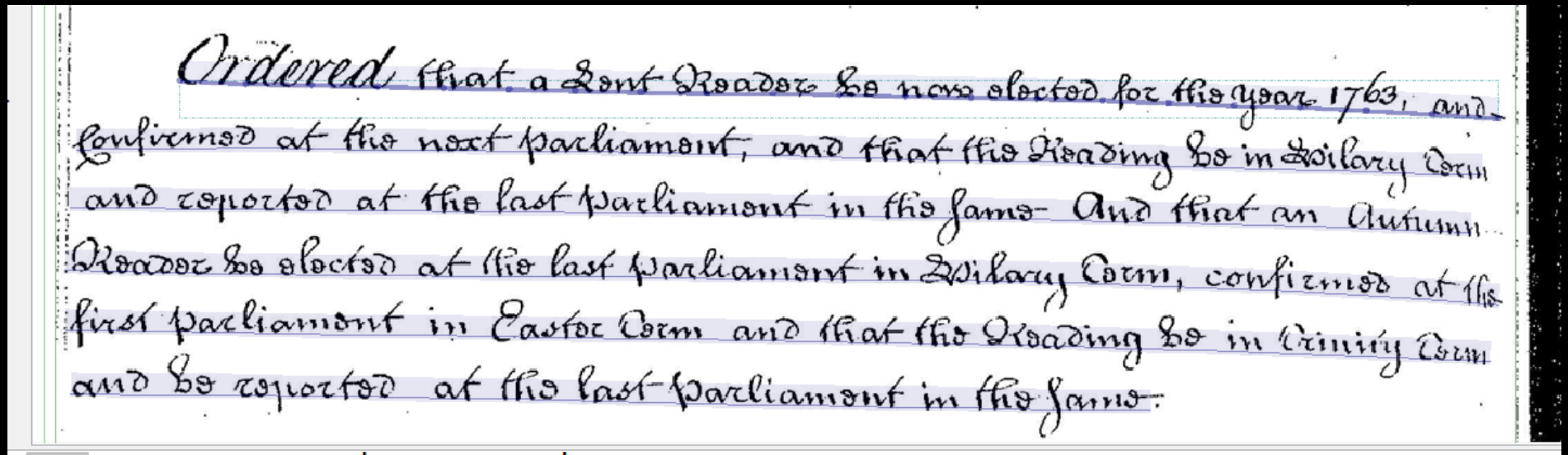
And these

Clockwise from upper-left

1. Michelangelo letter with ink splotches
2. Shakespeare's will with holes
3. Non-Latin scripts
4. Leonardo Da Vinci notebooks in mirror-image cursive

What we hope for

Automated transcription of a page from the Middle Temple records



Ordered that a Lent Reader be now elected for the year 1763, and confirmed at the next parliament, and that the Reading be in Hilary Term and reported at the last Parliament in the same. And that an Autumn Reader be elected at the last Parliament in Hilary Term, confirmed at the first parliament in Easter Term and that the Reading be in Trinity Term and be reported at the last Parliament in the same.

2-4 Ordered· that· a· Lent· Reader· be· now· elected· for· the· year· 1763,· and-↵
2-5 confirmed· at· the· next· Parliament,· and· that· the· Reading· be· in· Hilary· Term↵
2-6 and· reported· at· the· last· Parliament· in· the· same· And· that· an· Autumn↵
2-7 Reader· be· elected· at· the· last· Parliament· in· Hilary· Term,· confirmed· at· the↵
2-8 first· Parliament· in· Easter· Term· and· that· the· Reading· be· in· Trinity· Term↵

https://read.transkribus.eu

How it works …

- Preprocessing
  - Deskew
  - Convert to black and white or greyscale
  - Despeckle, remove noise
  - Identify layout blocks
- Character analysis
  - Identify character via pattern matching or glyph via feature extraction
  - Check word with dictionary
- Postprocessing cleanup

… and how it fails

- Problems with original
  - Poor printing – too much or too little ink, damaged metal type
  - Age – fading ink, yellow paper, stains, marginalia
  - Unusual fonts or handwriting
- Problems with the scan
  - B/W or greyscale at 300 – 600 dpi
  - Higher isn’t always better because picks up too much noise
  - Guttering, skewing
  - Light settings
  - Preferred format .tiff > .png > .jpg

It’s even hard for humans sometimes

Number	1
Lowercase L	l
Exclamation Point	!
Capital I	I
Pipe (vertical line)	|

= A very annoying WiFi password

Machine Learning Basics

- Give the computer lots of data, identify patterns, then predict outcomes based on previously learned patterns
- It's about statistical probabilities, not explicit rules

- Statistics
  - Type I errors (false positives)
  - Type II errors (false negatives)
- Machine learning algorithms attempt to reduce the number of one or both types of errors

Agenda

- Scanning best practices
- OCR tools for dealing with print
- Machine learning tools for handwriting recognition

- Full course schedule and hub: https://github.com/HeardLibrary/ocr
- Zotero library: assigned and supplemental readings