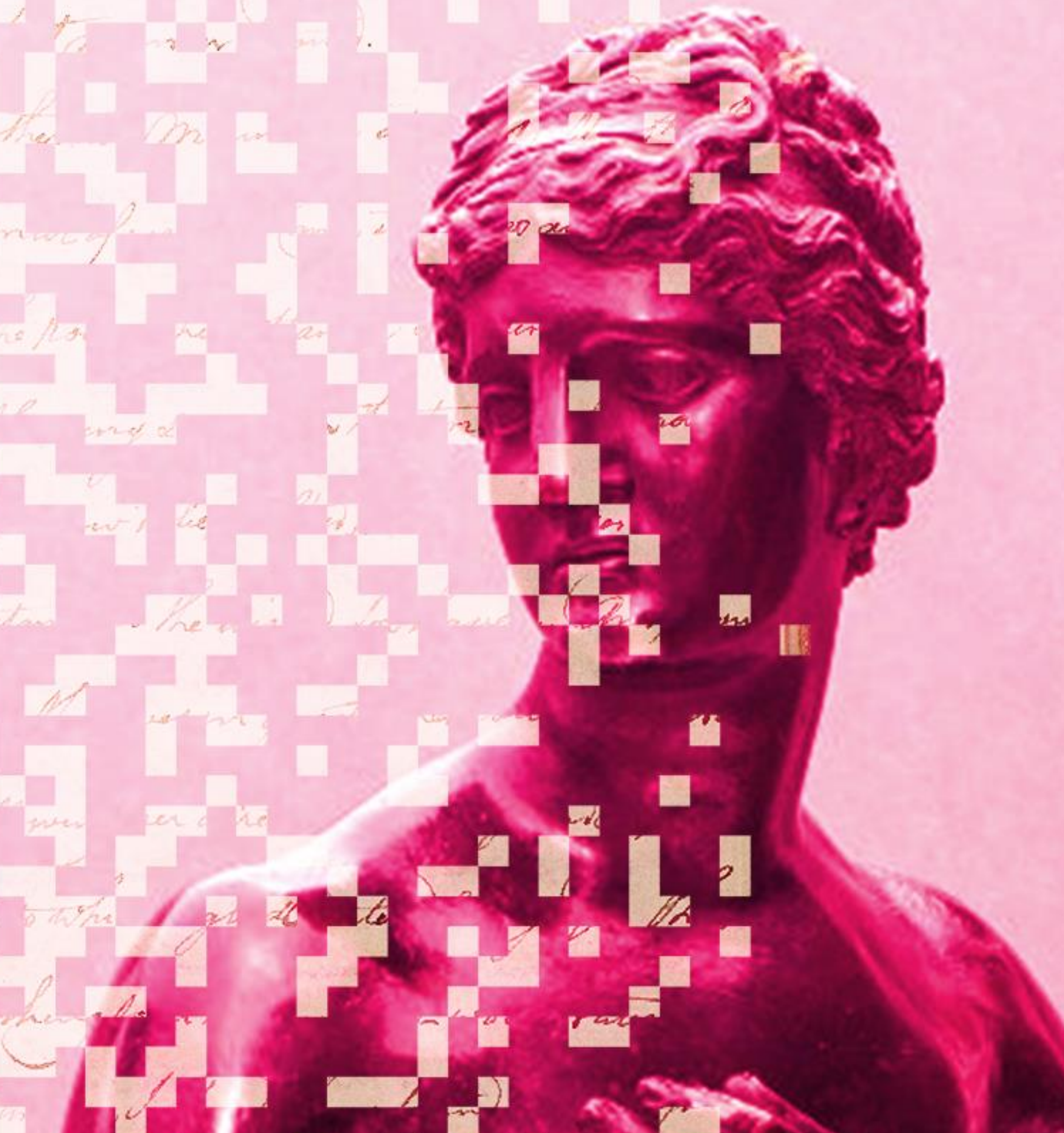
THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society

DATA
CULTURE
SOCIETY

Data Culture & society



@edCDCS

THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society

Working with OCR

9 & 16 Oct 2025

@edCDCS

DATA CULTURE SOCIETY

Overview

Day 1 (9 Oct.)

- Introduction to OCR
- Challenges of OCR in practice
- Activity: Ready-made OCR tools

Day 2 (16 Oct.)

- Review of programming basics
- Introduction to OCR packages
- Working with OCR in Python & R

THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society

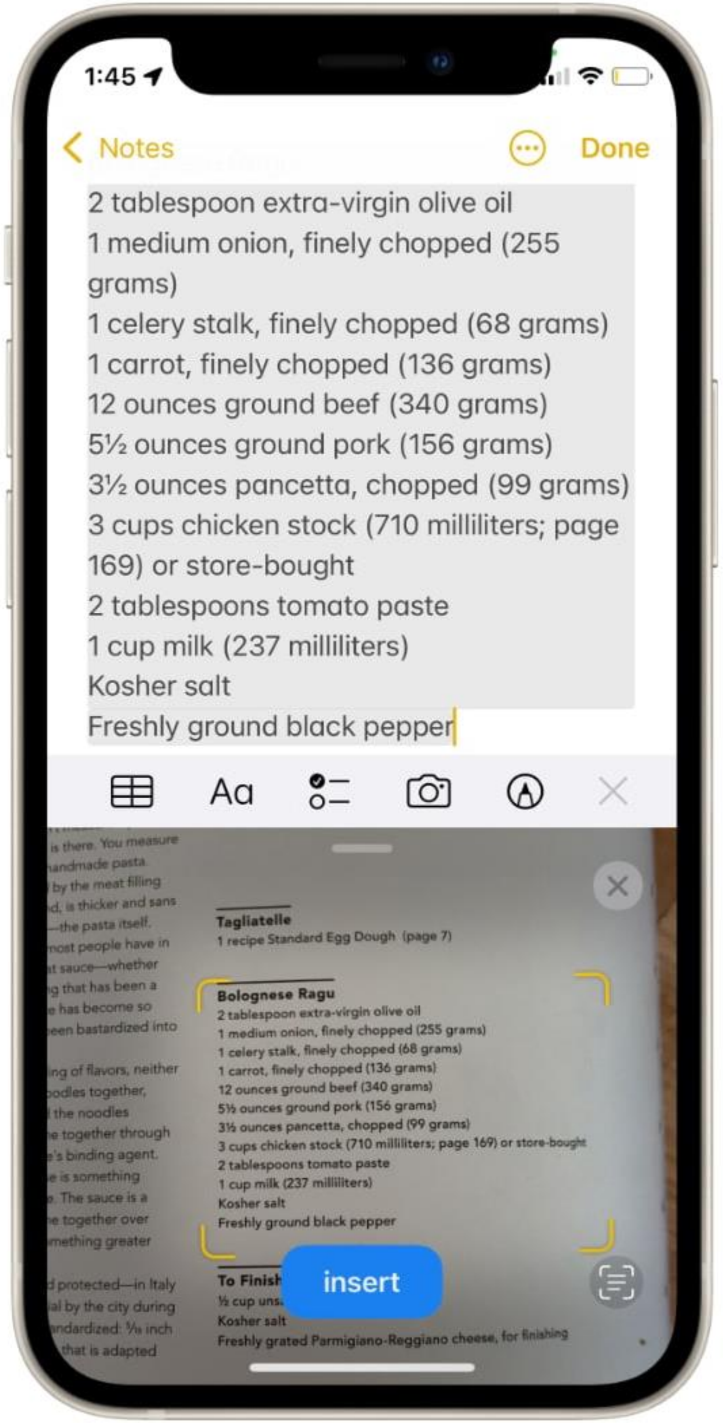
Course Material

What Is OCR?

- **O**ptical **C**haracter **R**ecognition
- OCR is the technique to process images of text, such as written or printed documents, and produce machine-readable documents.
- Machine-readable documents are encoded in formats that computers can process, allowing the text to be searched, edited and analysed.

OCR In the Wild
Real-World Examples

- Scanning your passport at the airport.
- Generate machine-readable text for text-to-speech technology.
- Making digitalised physical archives searchable.
- Creating a dataset of for text mining or text analysis.



2 tablespoon extra-virgin olive oil
1 medium onion, finely chopped (255 grams)
1 celery stalk, finely chopped (68 grams)
1 carrot, finely chopped (136 grams)
12 ounces ground beef (340 grams)
5½ ounces ground pork (156 grams)
3½ ounces pancetta, chopped (99 grams)
3 cups chicken stock (710 milliliters; page 169) or store-bought
2 tablespoons tomato paste
1 cup milk (237 milliliters)
Kosher salt
Freshly ground black pepper

Tagliatelle
1 recipe Standard Egg Dough (page 7)

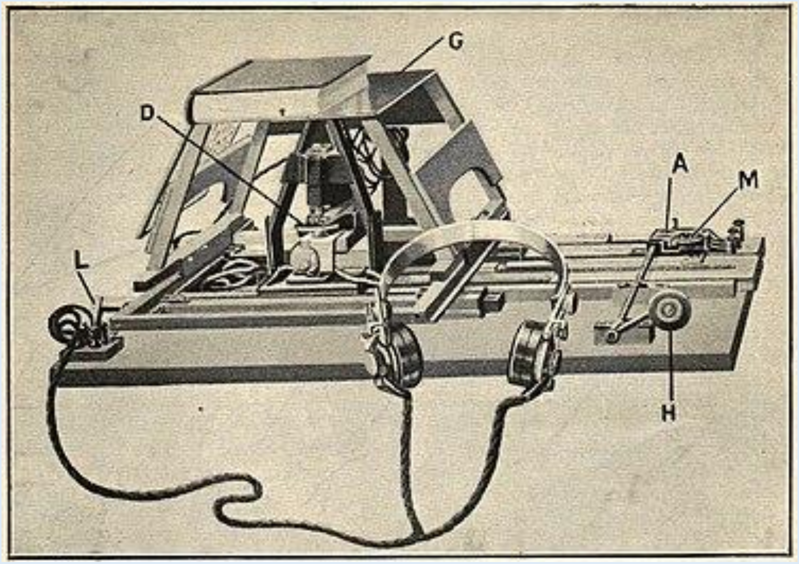
Bolognese Ragu
2 tablespoon extra-virgin olive oil
1 medium onion, finely chopped (255 grams)
1 celery stalk, finely chopped (68 grams)
1 carrot, finely chopped (136 grams)
12 ounces ground beef (340 grams)
5½ ounces ground pork (156 grams)
3½ ounces pancetta, chopped (99 grams)
3 cups chicken stock (710 milliliters; page 169) or store-bought
2 tablespoons tomato paste
1 cup milk (237 milliliters)
Kosher salt
Freshly ground black pepper

To Finish
½ cup uns
Kosher salt
Freshly grated Parmigiano-Reggiano cheese, for finishing

OCR Timeline: A Brief History

Early Foundations

1914: Emanuel Goldberg invents a machine that converts text into telegraph code; Edmund Fournier d’Albe invents the Optophone, which converts text to audio tones matched with letters



1931: Goldberg invents the “Statistical Machine,” which searches archives through optical code recognition

Mid-Century Machinery

1974: Ray Kurzweil develops the “Reading Machine,” an early form of OCR technology compatible with text-to-speech software

1976: Kurzweil sells to Xerox

Kurzweil Reading Machine
(circa 1978)

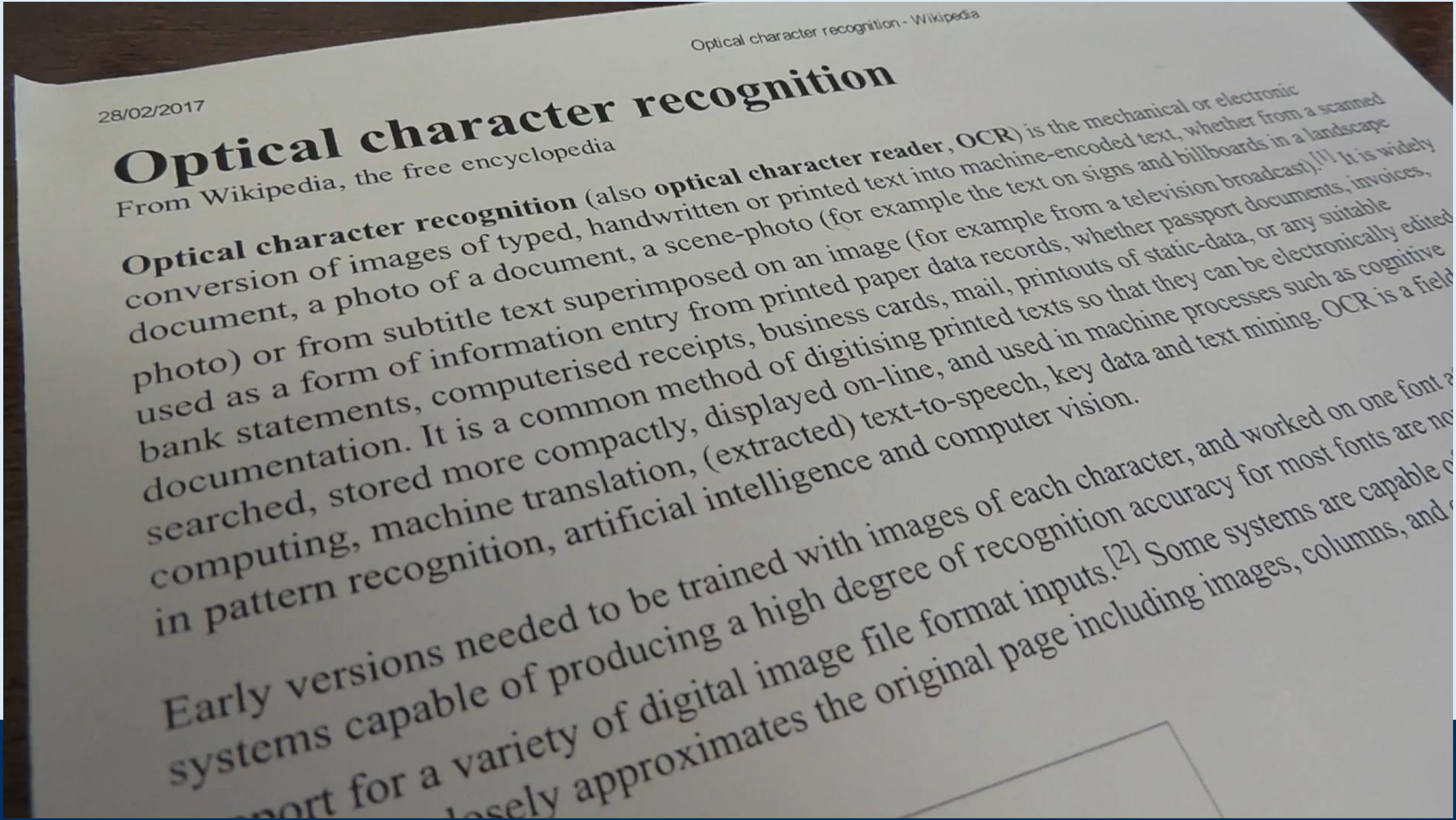
The Internet Era

1990s-00s: OCR goes mainstream; improved accuracy; mass digitization projects

00s-present: increased accessibility and accuracy

OCR workflow

1. Select the images to be scanned
2. Scan the images with OCR software
3. Inspect and 'clean' the processed files
4. Save the results for further use



Optical character recognition

From Wikipedia, the free encyclopedia

Optical character recognition (also **optical character reader**, **OCR**) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).[1] It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitising printed texts so that they can be electronically edited searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font a
systems capable of producing a high degree of recognition accuracy for most fonts are no
port for a variety of digital image file format inputs.[2] Some systems are capable o
losely approximates the original page including images, columns, and

Activity: OCR Workflow

1. Identify a dataset (images of text) that you might use in your research
2. Write the steps to obtained encoded text from your dataset. (e.g. software, OCR module)
3. Share your dataset, workflow, and plan with your small group.
4. Identify potential issue in each step. And discuss how you might address them.

Challenges of OCR

Accuracy depends on **dataset quality**, **visual complexity** and **software capability**. **Some common sources of error include:**

- Human errors and typos
- Age and damage (stained or blurry)
- Mixed text and images, or multiple languages
- Cursive handwriting

Challenges of OCR

Possible solutions:

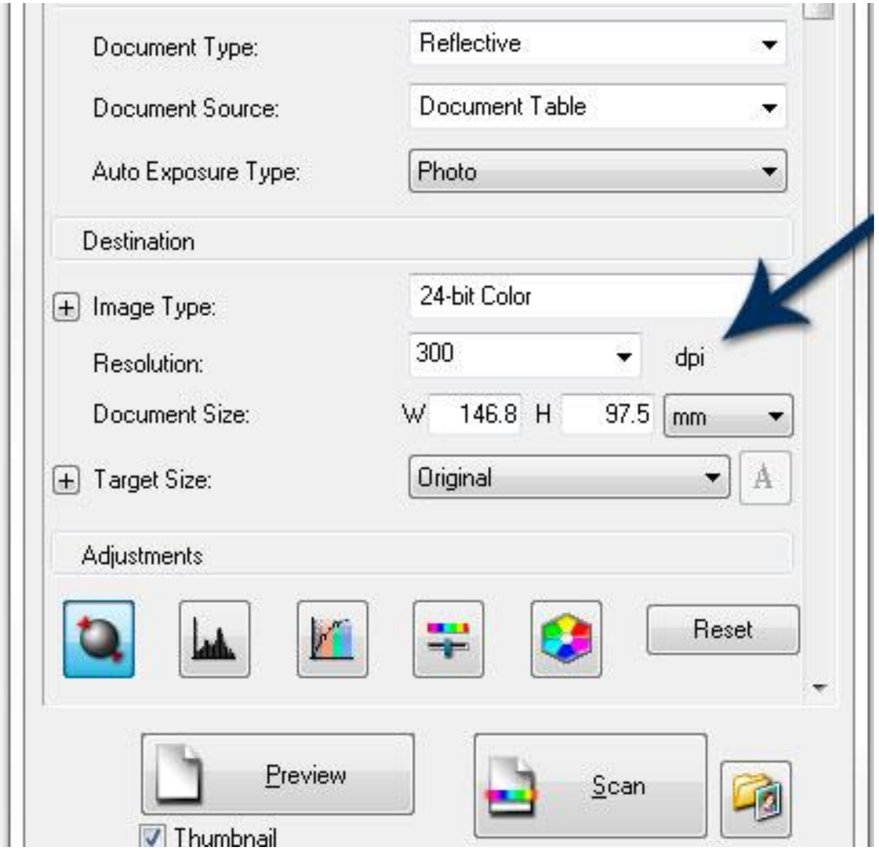
- Select good quality dataset to begin with.
- Pre-process your dataset to improve its quality
- Correct errors in OCR-produced files, if they are predictable.
- Improve OCR engine capability

Data Selection

Image resolution

300 DPI is often used as a benchmark for good quality printing reproducibility for photographs, but this may vary

Ex. National archive



Document Type: Reflective

Document Source: Document Table

Auto Exposure Type: Photo

Destination

Image Type: 24-bit Color

Resolution: 300 dpi

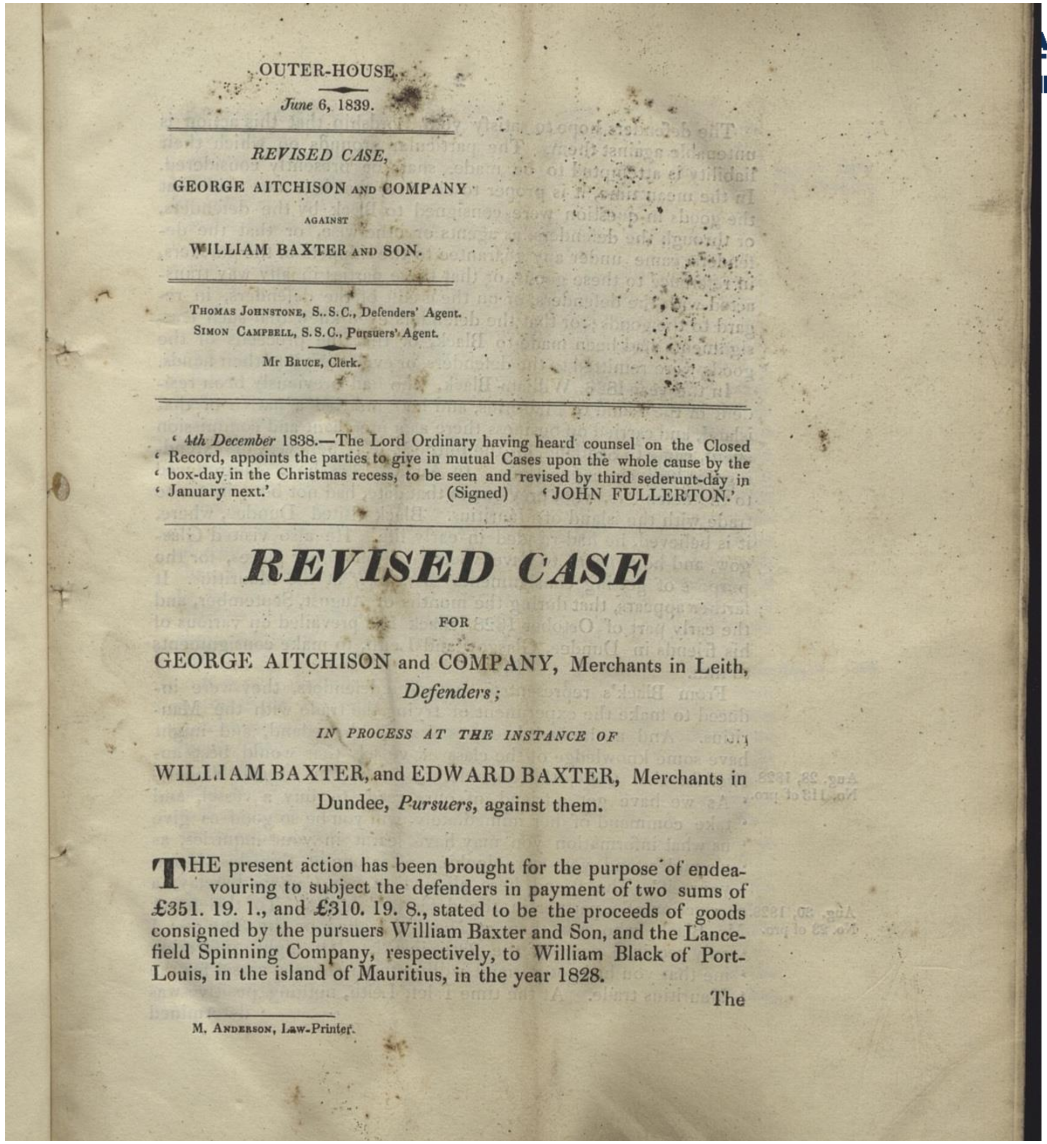
Document Size: W 146.8 H 97.5 mm

Target Size: Original

Adjustments

Data Selection

Manual inspection



OUTER-HOUSE.

June 6, 1839.

REVISED CASE,

GEORGE AITCHISON AND COMPANY

AGAINST

WILLIAM BAXTER AND SON.

Thomas Johnstone, S..S.C., Defenders' Agent.
Simon Campbell, S.S.C., Pursuers' Agent.

Mr Bruce, Clerk.

'*4th December* 1838.—The Lord Ordinary having heard counsel on the Closed 'Record, appoints the parties to give in mutual Cases upon the whole cause by the 'box-day in the Christmas recess, to be seen and revised by third sederunt-day in 'January next.' (Signed) 'JOHN FULLERTON.'

REVISED CASE

FOR

GEORGE AITCHISON and COMPANY, Merchants in Leith,
Defenders;

IN PROCESS AT THE INSTANCE OF

WILLIAM BAXTER, and EDWARD BAXTER, Merchants in
Dundee, *Pursuers*, against them.

THE present action has been brought for the purpose of endeavouring to subject the defenders in payment of two sums of £351. 19. 1., and £310. 19. 8., stated to be the proceeds of goods consigned by the pursuers William Baxter and Son, and the Lancefield Spinning Company, respectively, to William Black of Port-Louis, in the island of Mauritius, in the year 1828.

The

M. Anderson, Law-Printer.

Activity:
Inspect the Images

- Identify issues you might encounter when processing the following images.
- Are there any steps you could take to preprocessing the document that might improve the output accuracy?

Pre-processing

- **Manually**: Adjust the colour, contrast, or formats of the files.
- **OCR tools**: Some OCR software have built-in correction functions such as patches.

Scanning

Pay attention to the limitations of the software

- **File Size**
- **File Format**
- **Text orientation**
- **Languages**

Activity: Scanning

- https://tools.pdf24.org/en/ocr-pdf
- https://www.onlineocr.net/
- https://www.sodapdf.com/ocr-pdf/
- https://www.sejda.com/ocr-pdf
- https://ocr.space/
- https://avepdf.com/pdf-ocr

1. Identify a picture or scanned pdf
2. Scan the documents with OCR software
3. Compare the results and discuss what are the advantages and limitations of these options?

(if you can't find one, please try the pictures in the previous slides, or you can also try the Edinburgh archive database:

https://archives.collections.ed.ac.uk/)

Cleaning

- Manually remove errors

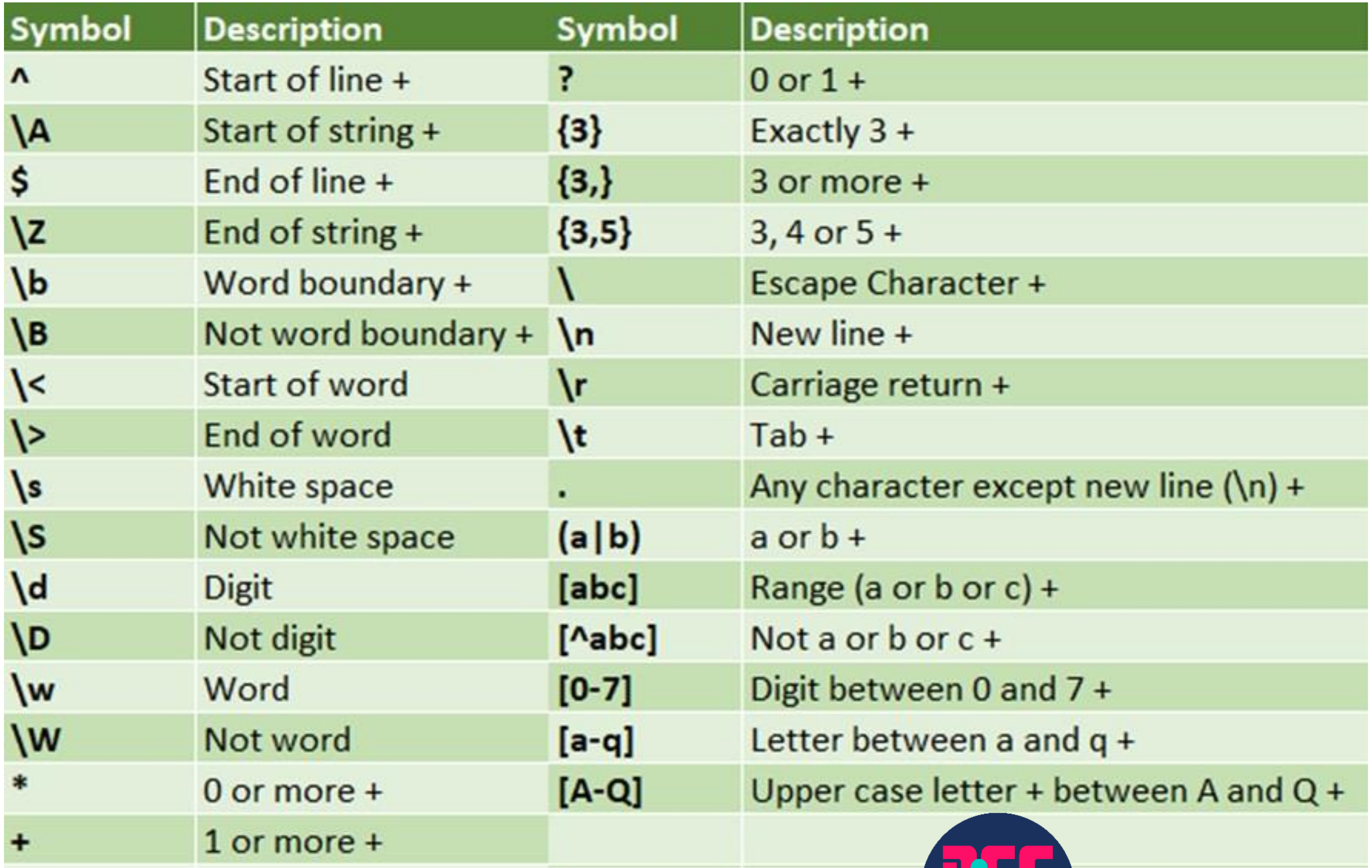
  Predictable errors can be fixed with codes , such as Regex.

Regex

https://devhumor.com/media/the-nightmare

- **Concept** used in many **different programming environments** for **pattern matching**.
- Powerful tool **to find, manage, and transform data and files.**
- Use a sequence of characters to define a **search to match strings**
  - **Match on types of characters** (e.g. ‘upper case letters’, ‘digits’, ‘spaces’, etc.).
  - **Match patterns** that repeat any number of times.

Regex



Symbol	Description	Symbol	Description
^	Start of line +	?	0 or 1 +
\A	Start of string +	{3}	Exactly 3 +
$	End of line +	{3,}	3 or more +
\Z	End of string +	{3,5}	3, 4 or 5 +
\b	Word boundary +	\	Escape Character +
\B	Not word boundary +	\n	New line +
\<	Start of word	\r	Carriage return +
\>	End of word	\t	Tab +
\s	White space	.	Any character except new line (\n) +
\S	Not white space	(a|b)	a or b +
\d	Digit	[abc]	Range (a or b or c) +
\D	Not digit	[^abc]	Not a or b or c +
\w	Word	[0-7]	Digit between 0 and 7 +
\W	Not word	[a-q]	Letter between a and q +
*	0 or more +	[A-Q]	Upper case letter + between A and Q +
+	1 or more +		

- https://programminghistorian.org/en/lessons/cleaning-ocrd-text-with-regular-expressions
- https://programminghistorian.org/en/lessons/understanding-regular-expressions
- https://librarycarpentry.org/lc-data-intro/01-regular-expressions/

Printed Text Recognition using Python pytesseract

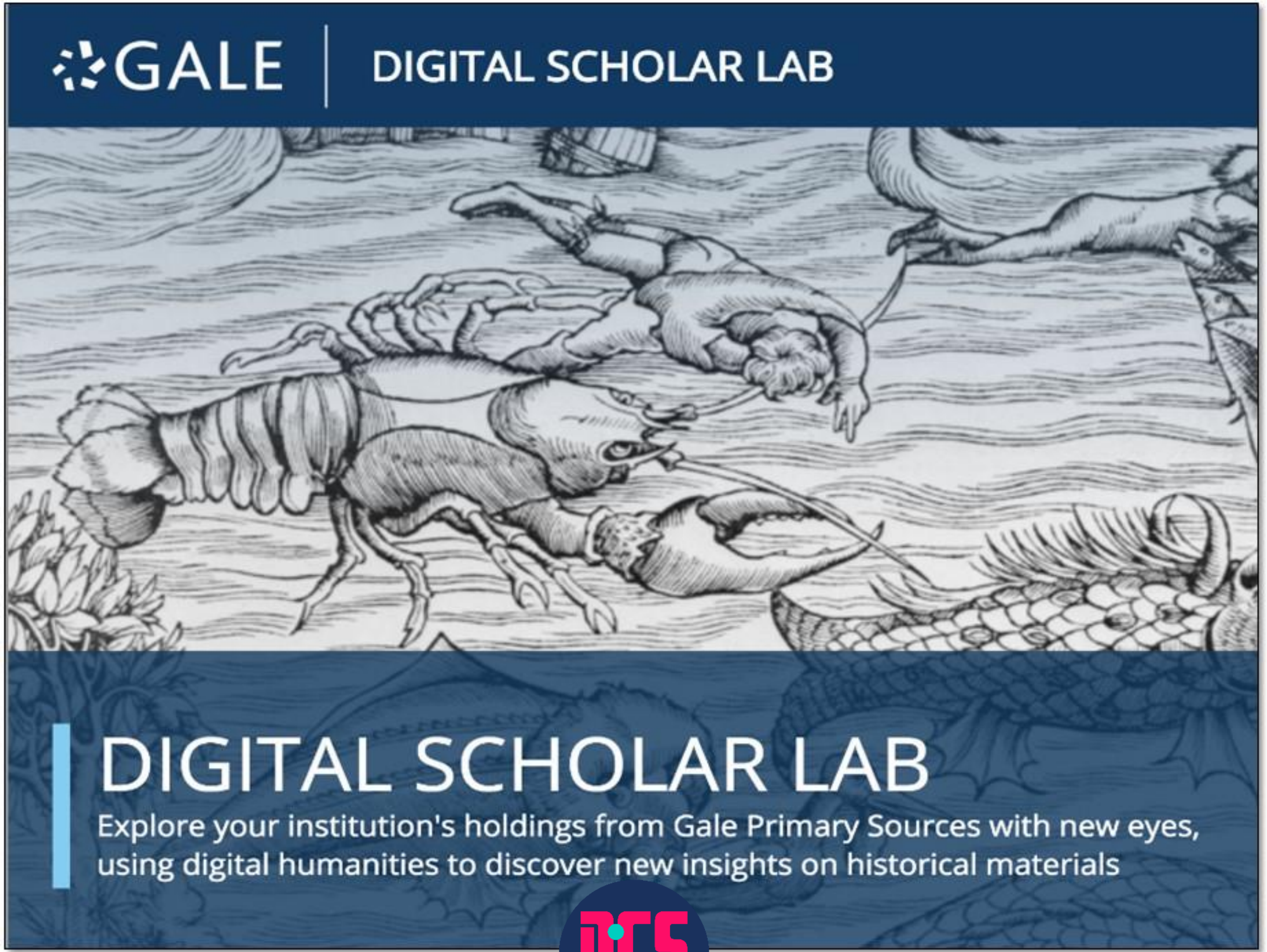
Tesseract is an OCR engine developed for various operating systems

- It has become open-source since 2005
- It is of the most accurate open-source OCR engines available
- Originally only support English, but more languages have been added
- Can run from command line interface or embedded in main coding languages (R & Python)

Handwritten Text Recognition with Python trOCR

- Select good quality dataset to begin with
- Pre-process your dataset to improve its quality
- Correct errors in OCR-produced files, if they are predictable.
- Improve OCR engine capability

[Digital Scholar Lab]



GALE | DIGITAL SCHOLAR LAB

DIGITAL SCHOLAR LAB

Explore your institution's holdings from Gale Primary Sources with new eyes, using digital humanities to discover new insights on historical materials