Machine learning-supported Historical Handwritten Recognition: Tools and Trends

27/06/24

Sara Ferro, Post. Doc. @CCHT

Mail: Sara.Ferro@iit.it

@PyDataVenice

Università Ca'Foscari Venezia

iit ISTITUTO ITALIANO DI TECNOLOGIA CENTER FOR CULTURAL HERITAGE TECHNOLOGY

A Brief Overview of CCHT

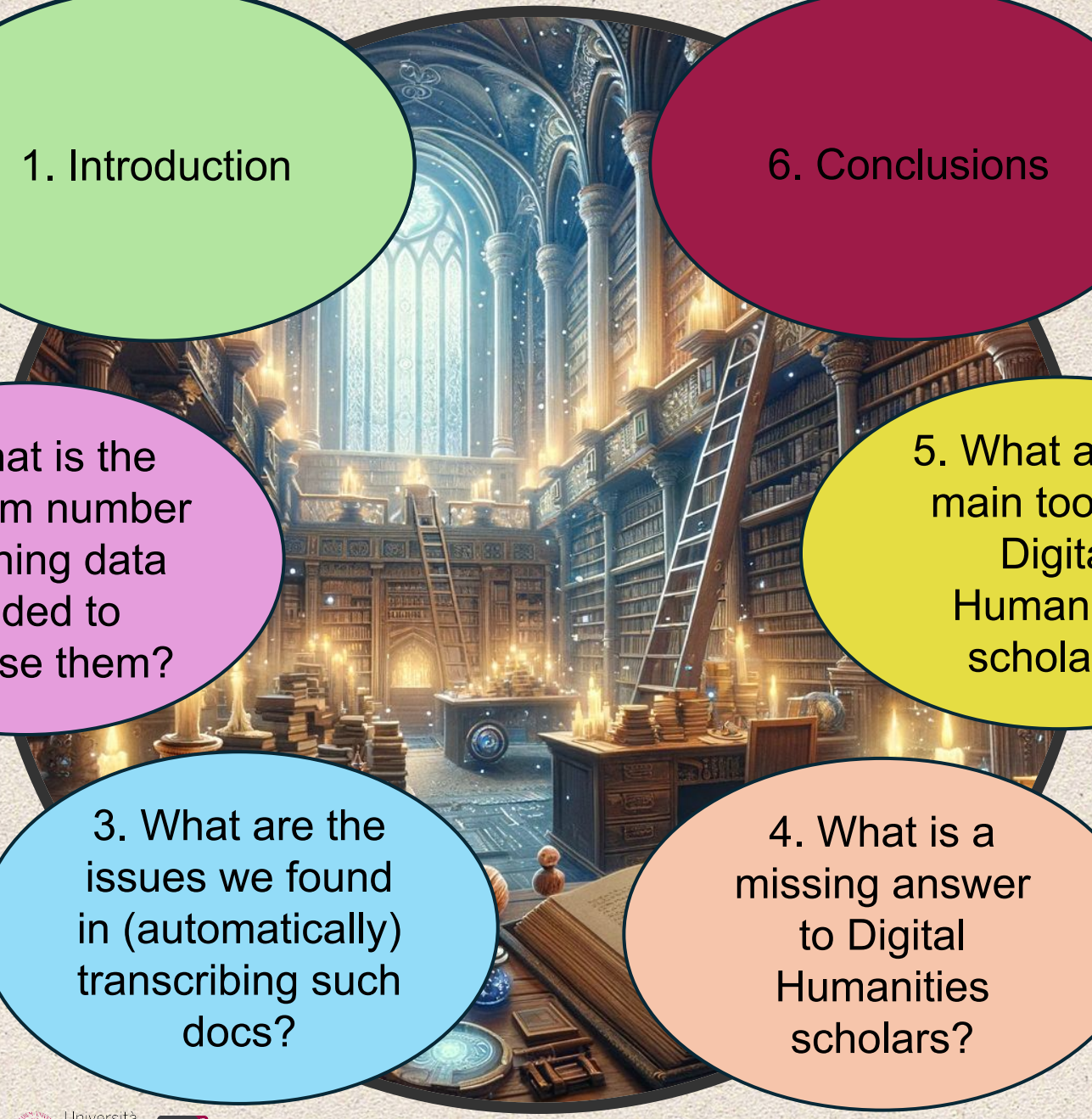
Center for Cultural Heritage Technologies (CCHT)

Three branches:

1. *Machine Learning*
2. *Chemistry*
3. *Robotics*

applied to cultural heritage data

RITHMS
Research, Intelligence and Technology for Heritage and Market Security



1. Introduction

2. What is the minimum number of training data needed to digitalise them?

3. What are the issues we found in (automatically) transcribing such docs?

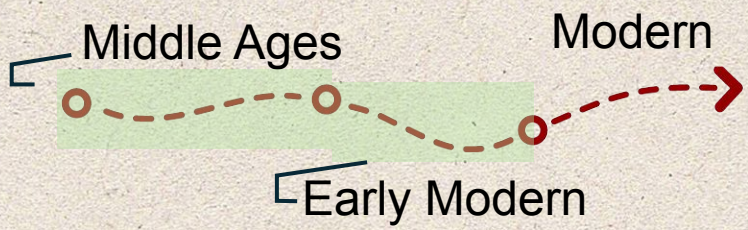
4. What is a missing answer to Digital Humanities scholars?

5. What are the main tools for Digital Humanities scholars?

6. Conclusions

1.a. Historical Documents

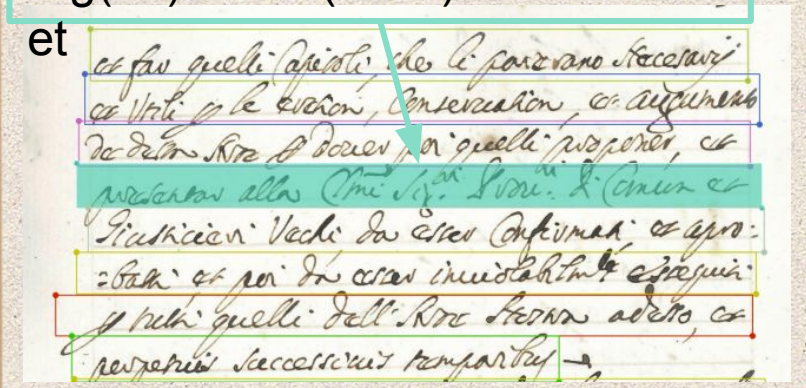
- *Western languages* written in *Latin characters*



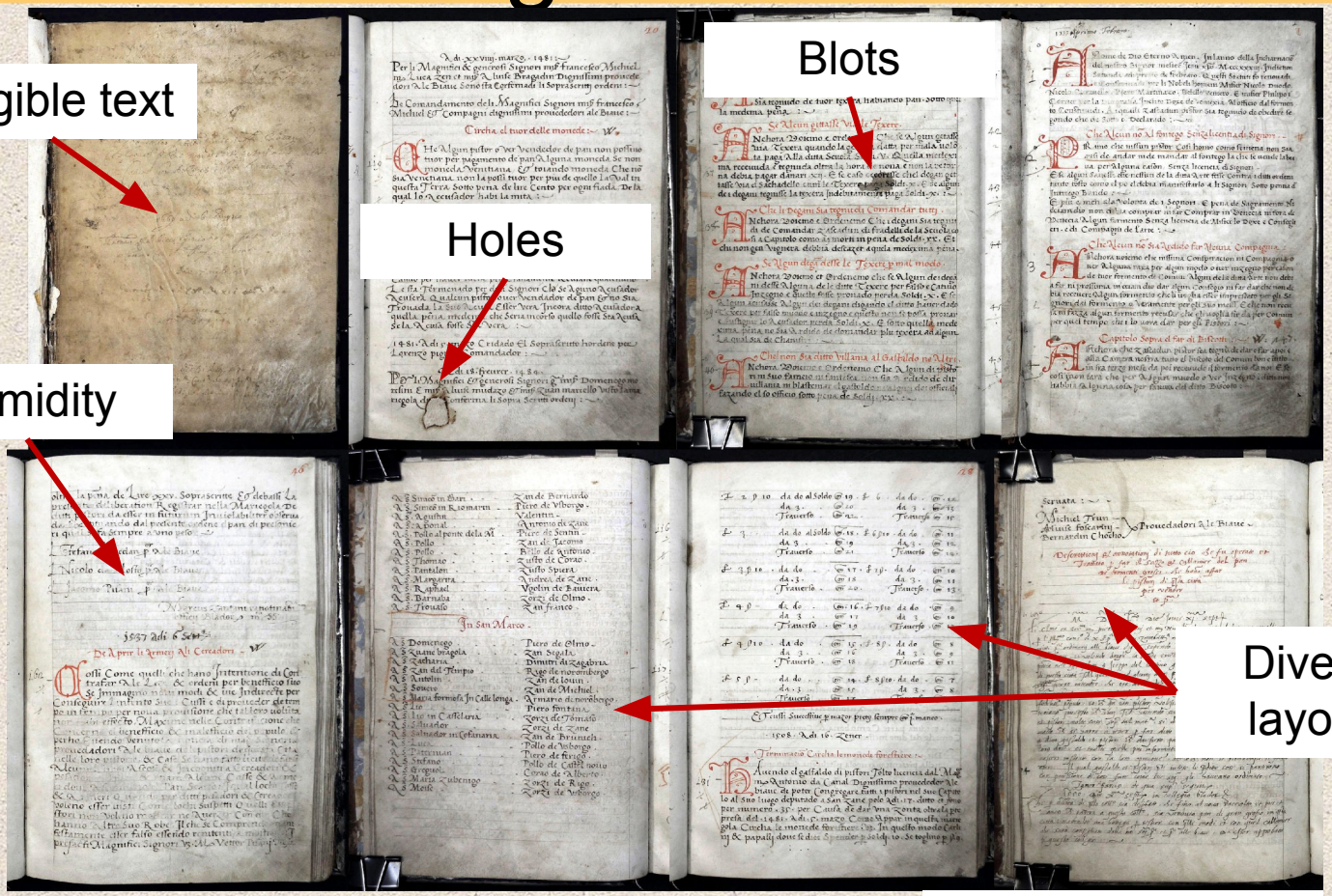
- Introduce a *new dataset* of handwritten pages written in a *vernacular Italian*
- Use of *open-access* historical datasets

- Most models have been released for line-level datasets

presentar alla Cl(arissi)mi Sig(no)ri Prov(edito)ri di Comun et



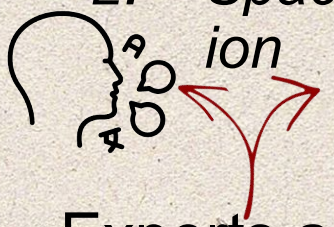
1.a. Historical Documents – complications in transcribing



Images from the Correr Museum – Mariegola CI IV 005 - Pistori

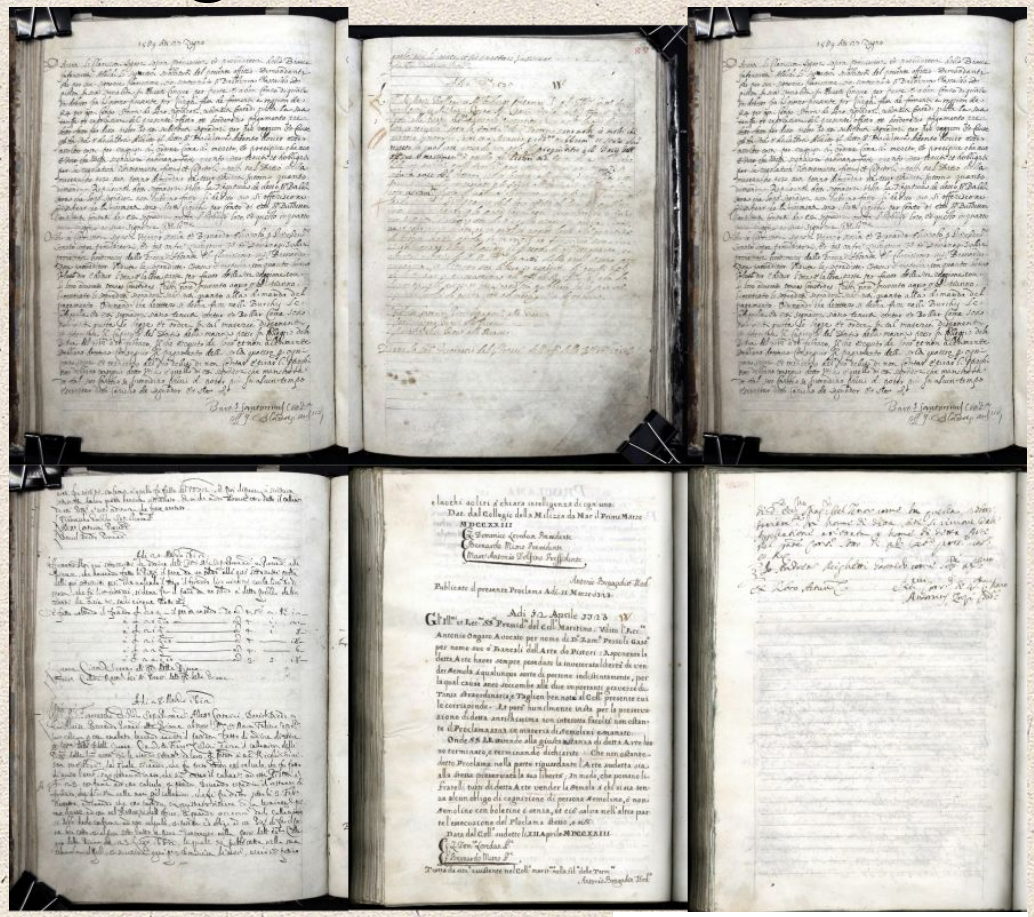
1.a. Historical Documents – complications in transcribing

- *Diverse handwriting, varying over:*
  1. *Time*
  2. *Space/region*



- Experts are needed to label

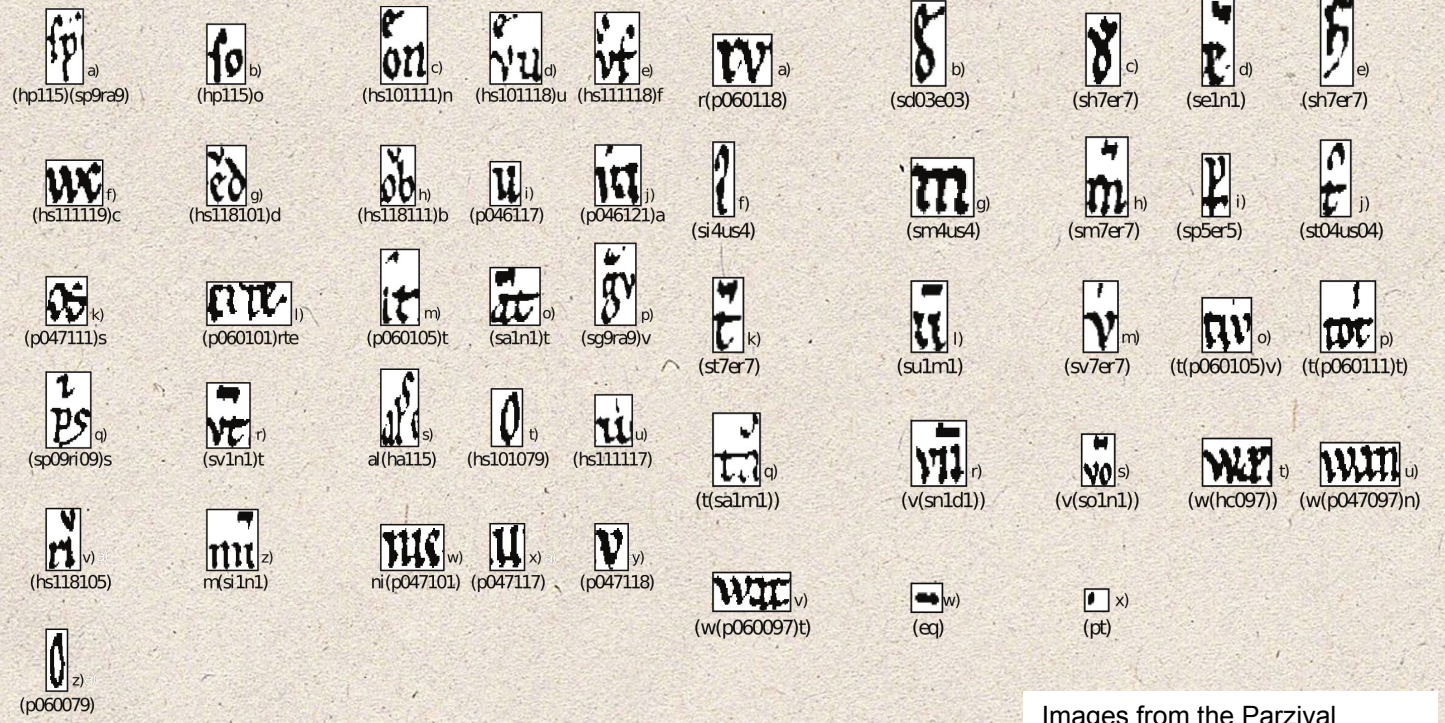
Few samples



Images from the Correr Museum – Mariegola Cl IV 005 - Pistori

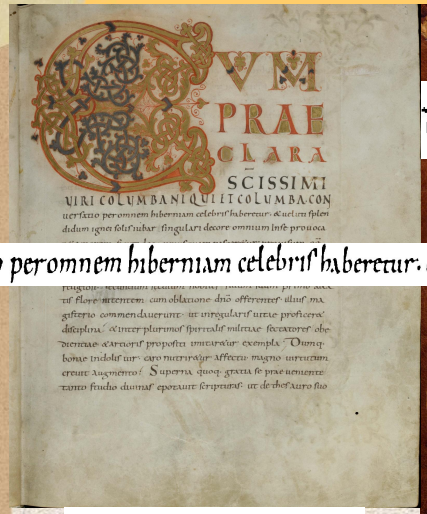
1.a. Historical Documents – complications in transcribing

- *Feature many special symbols (e.g., abbreviations)*



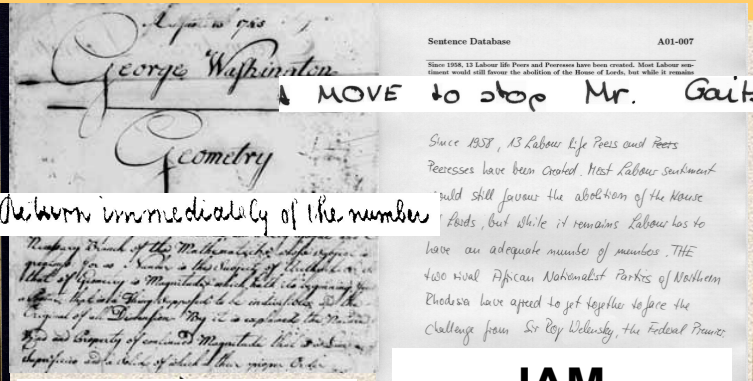
Images from the Parzival dataset – IAM-HistDB

1.a. Historical Documents – Used Datasets



Saint Gall
Latin, 9th c.,
1 scribe

Parzival
German, 13th c.,
3 scribes



Washington
English, 18th c.,
2 writers

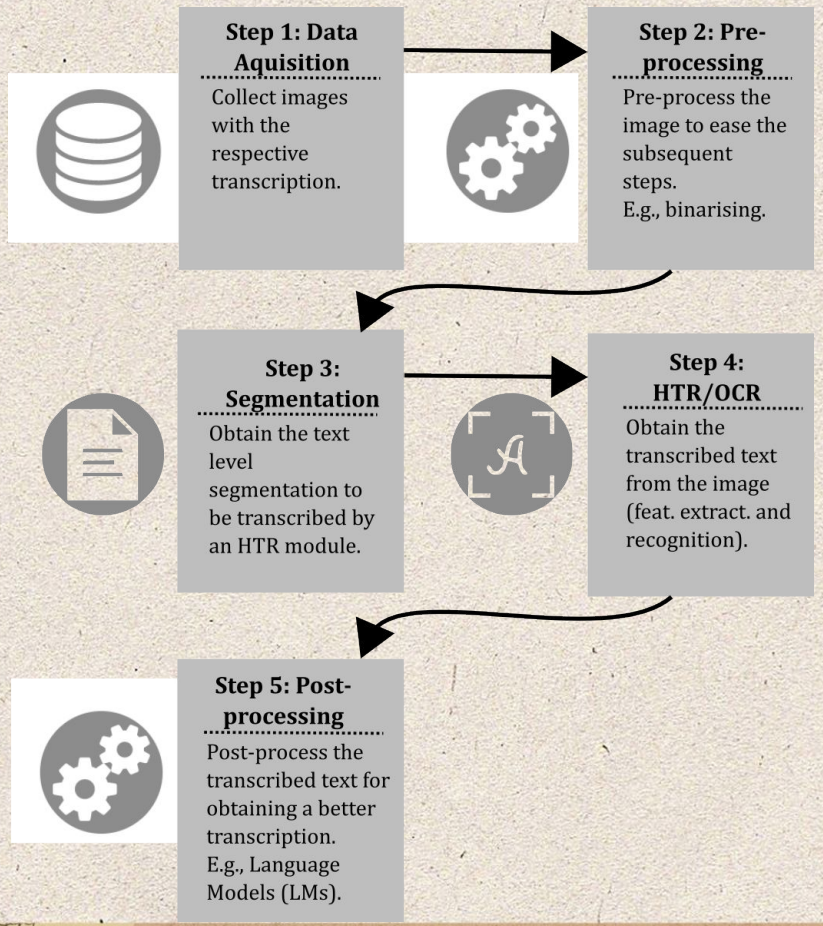
IAM
English,
657 writers

Min.

Max.

Name	Train.	Val.	Test	N. chars
Saint Gall	468	235	707	**49**
Parzival	2237	912	1328	**96**
Washington	**325**	**168**	**163**	83
IAM	**6482**	**976**	**2915**	79

1.a. Historical Handwritten Text Recognition – Generic Schema



Preprocessing

1. Cleanse
2. Refine/enhance
3. **Binarise** (e.g., Otsu)
4. **Normalise** (many meanings)

Goal: *eliminate elements from the image* that are *not informative* to (automatically) *transcribe* the image

1.b. Historical Handwritten Text Recognition – Different Levels

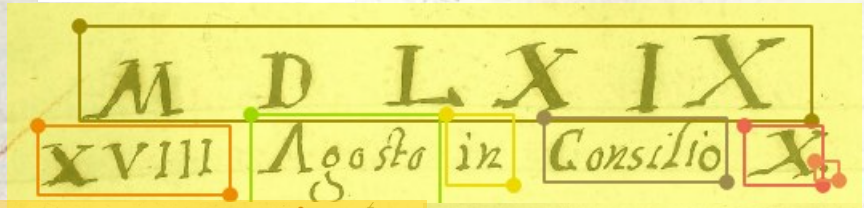
Character-level

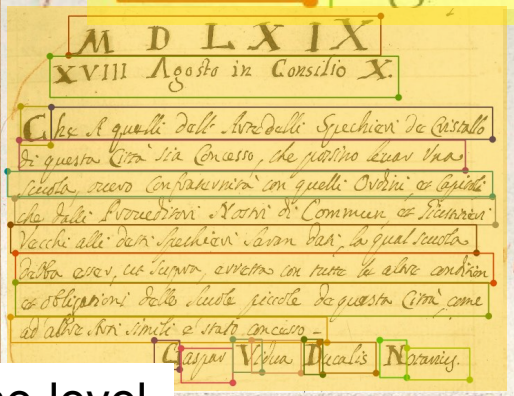
Word-level

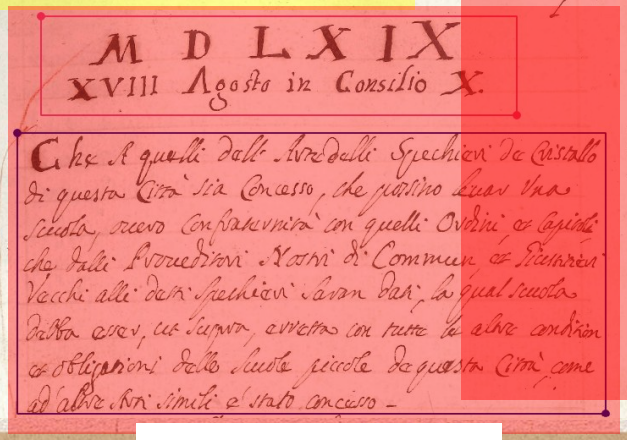
Line-level

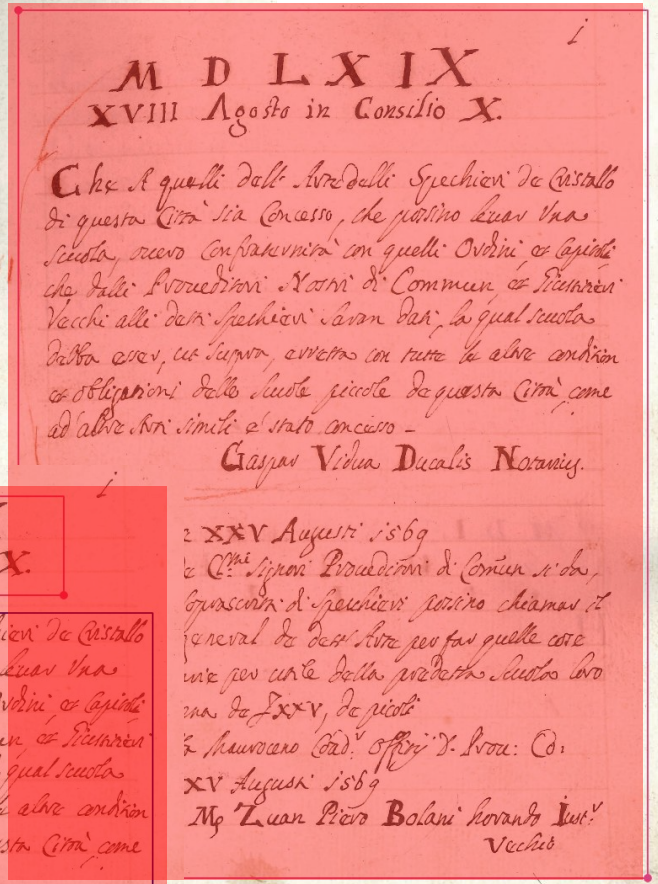
Page-level

Paragraph-leve







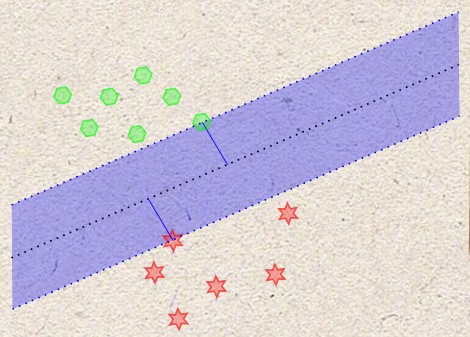


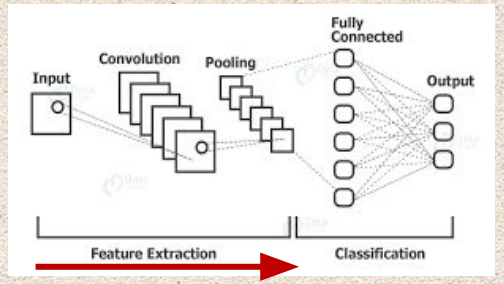
1.b. Character-level Models

- **Hp.**: the characters are already *well segmented*
- **Goal**: recognise/classify them

1. Handcrafted features have been defined

   (+ dimensionality reduction)
2. Any classification algorithm can be used (MLP, SVM, etc.)
3. Characteristics obtained through DNNs permit better results



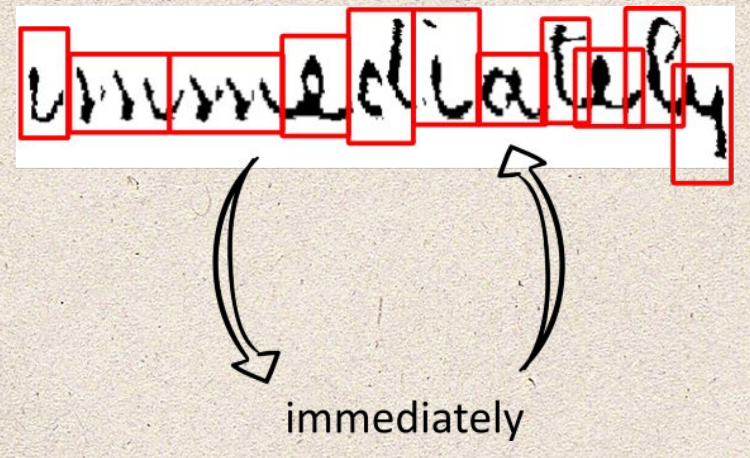


From basic features to more complex ones

1.b. Word-level Models

Option: break-down the word image into characters and recognise the characters themselves

Go to “Character-level Models”

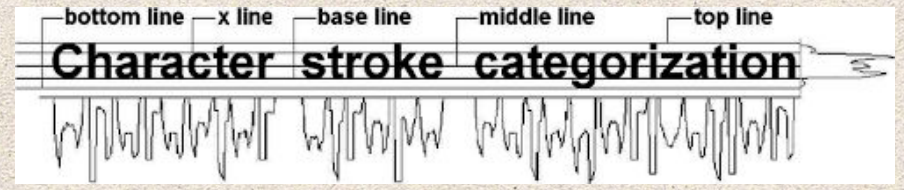


“A cursively written word cannot be recognised without being segmented and cannot be segmented without being recognized.” [Sayre’s paradox]

1.b. Word-level Models

Option: Segmentation till the character-level

1. *Projection profiles* (+ refinements such as closed character detection and pixel count)



2. *Oversegmentation of words*: the segmentation problem becomes a recognition problem of the "cut points"

+ refinements

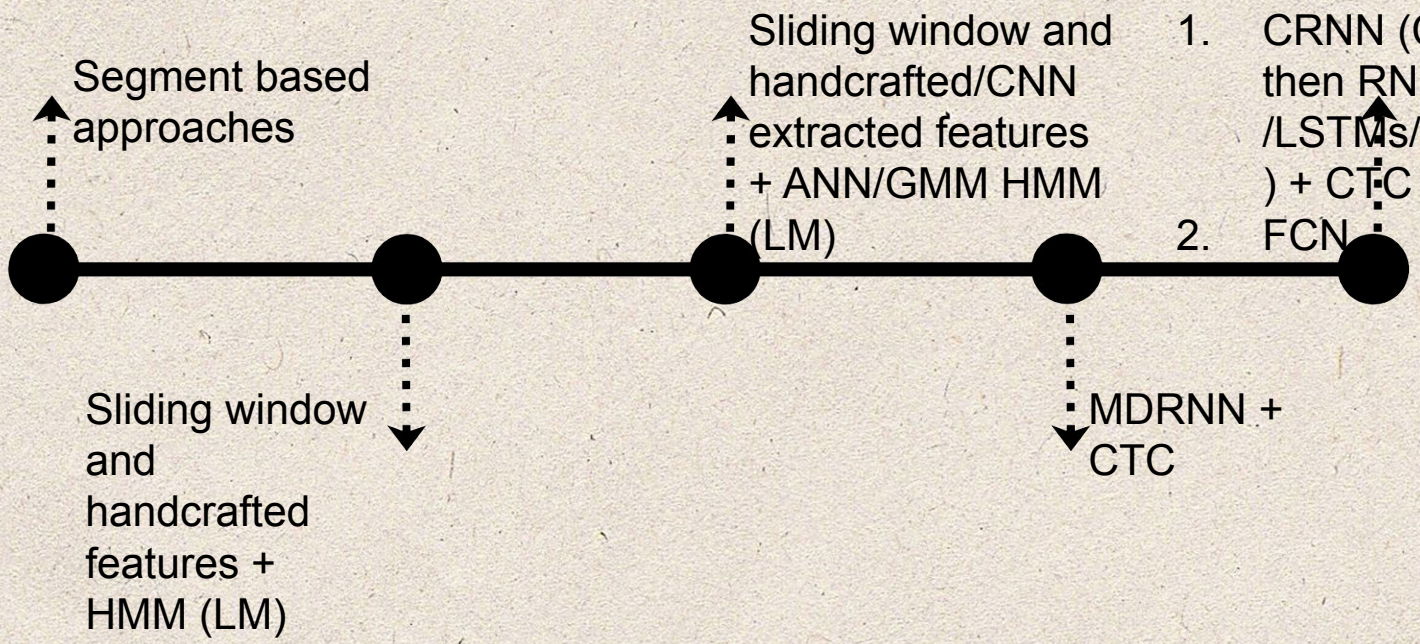
These methods fail when:

1. Characters touching multiple tim
2. Warped characters

Samples from two different Marigolds (Mariegole)

1.b. Word-level Models

- Toledo et al. (2017) pretrained on a PHOCNet
- Connectionist Temporal Classification

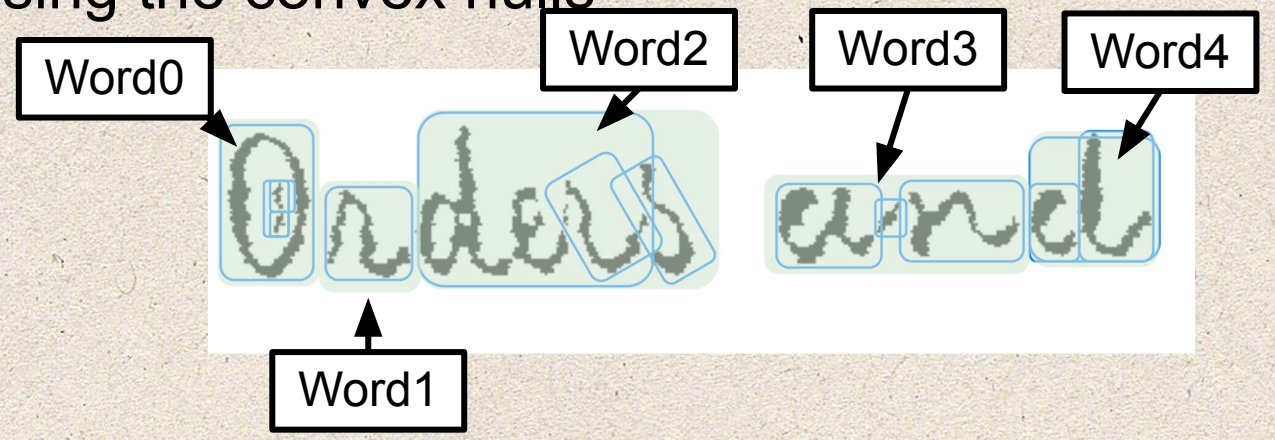


1.b. Line-level Model

Deterministic and unsupervised methods segment into words or characters

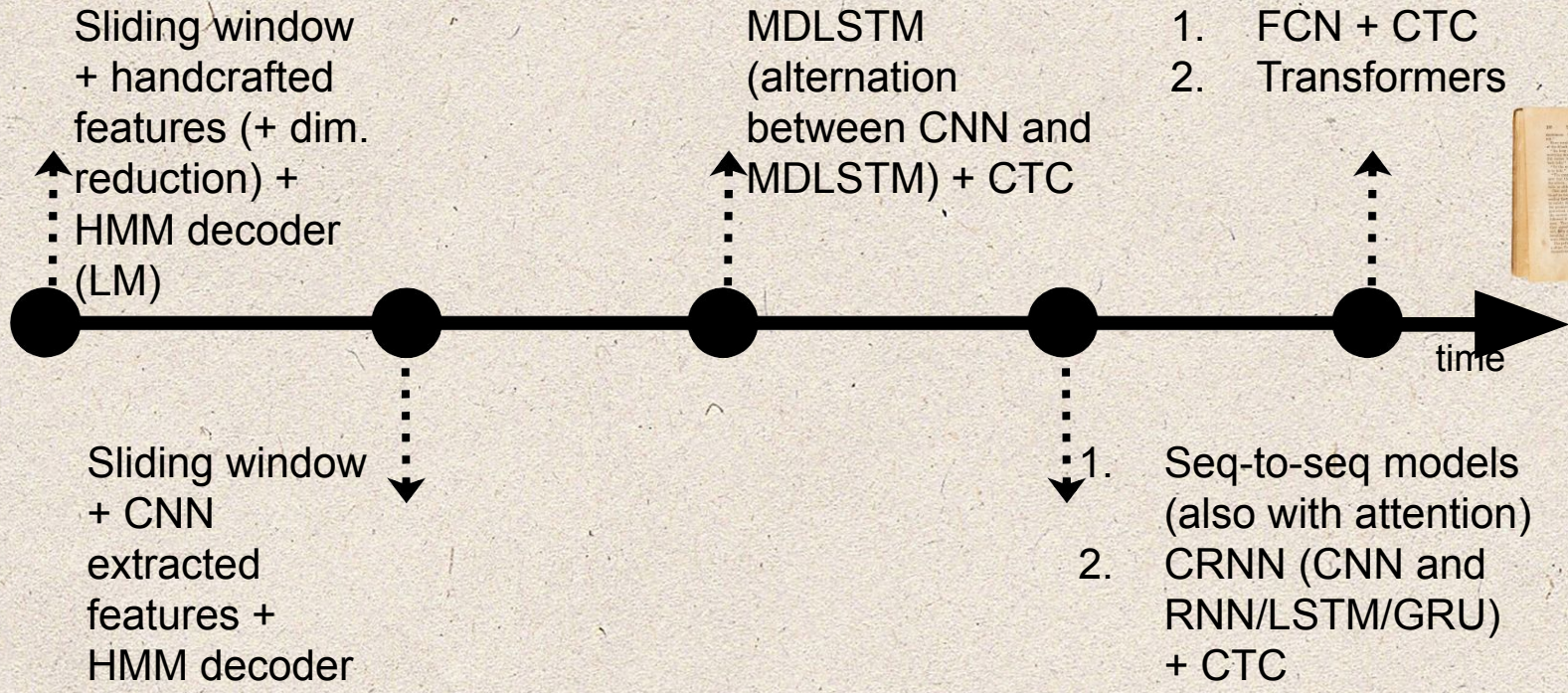
Example (deterministic)

Using the convex hulls



Prone to **errors**…though there are many possible refinements

1.b. Line-level Model



1.b. Whole-page-level Transcription

- Very few methods can work at page-level
- Many works need also the segmentation of the text lines

Projection Profiles

Recursive XY cutting alg.

Seam carving alg.

Possible Issue	Description
1. Overlapping lines ✗	One line partially overlaps the subsequent
2. Horizontally Adjac… Lines ✗	Multiple columns and side notes
3. Skewed Lines ✓	Not horizontal lines, but at an angle
4. Curved Lines ✓	Not horizontal lines, but with multiple angles
5. Warped Lines ✓	Deformed lines

1.b. Whole-page-level Transcription

DNN-based works:

OrigamiNet

Document Attention Network

Possible Issue	Description
1. Overlapping lines ✓	One line partially overlaps the subsequent
2. Horizontally Adjacent Lines ✗ ✓	Multiple columns and side notes
3. Skewed Lines ✓	Not horizontal lines, but at an angle
4. Curved Lines ✗ ✓	Not horizontal lines, but with multiple angles
5. Warped Lines	Deformed lines

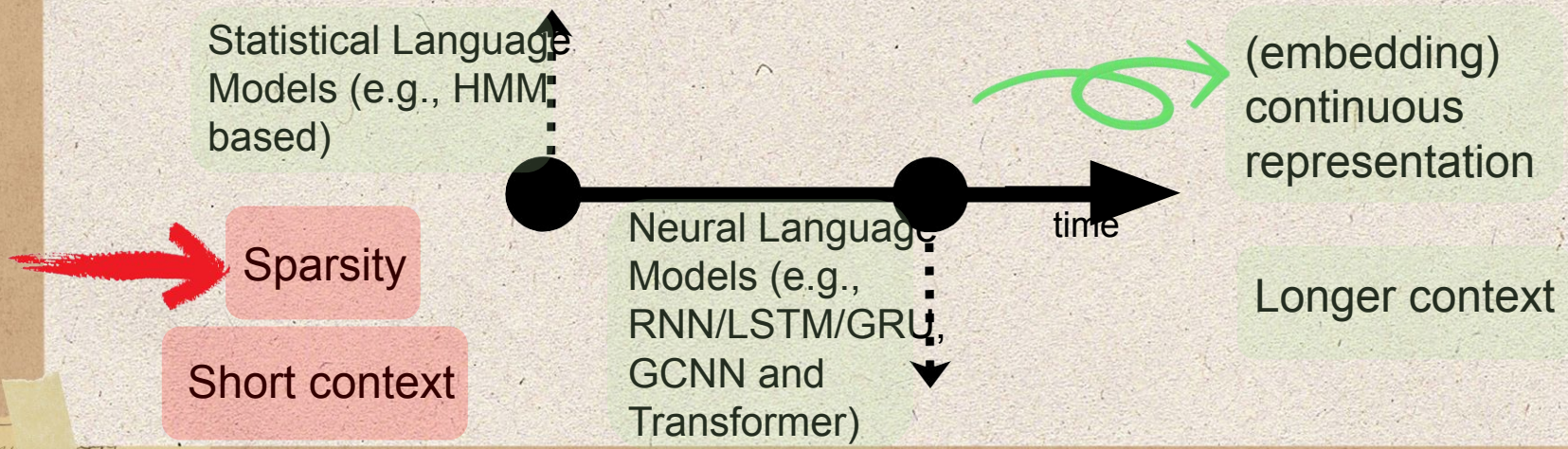
1.b. Post-processing Methods

- Language Models (LMs) can be utilised in the:

1. Post-processing (refinement) of text digitalisation
2. Decoding component of the recognition modules

Word at pos. i

$$P(w_0, \dots, w_N) = P(w_0)\Pi_{i=1}^{N} P(w_i | w_0, \dots, w_{i-1})$$



2. Quantity of Data Needed to Digitalise Historical Documents

- Expert palaeographers are needed
- Decrease effort to create such datasets:
  1. Data augmentation
  2. Transfer Learning (TL) and fine-tuning
  3. Active Learning
  4. (…)

Minimum amount of training data is needed to effectively transcribe historical documents?

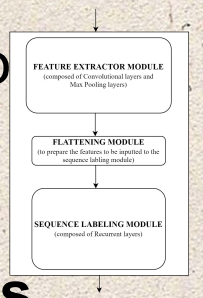
3. Quantity of Data Needed to Digitalise Historical Documents

CRNN

1. Data augmentatio
2. Fine-tuning

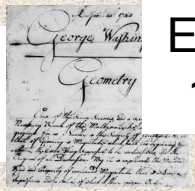
10M params

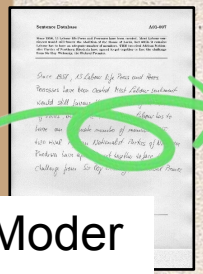
Few hundred lines is enough



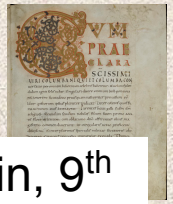
German, 13th c.

English, 18th c.





Modern English



Latin, 9th c.

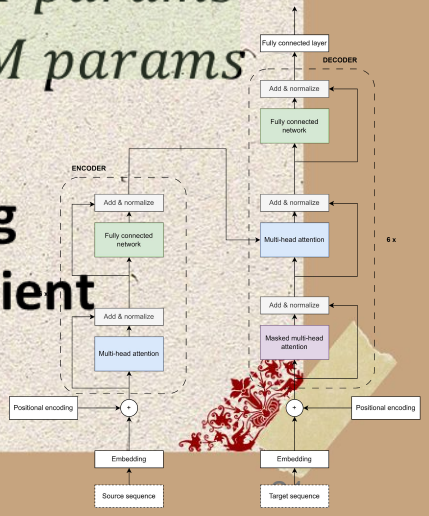
Venetian, 18th c.

Transformer (TrOCR released by Microsoft)

Fine-tuning (already trained over Ms modern English text-lines)

1. $TrOCR_{small}$ 62M *params*
2. $TROCR_{base}$ 334M *params*
3. $TrOCR_{large}$ 558M *params*

The original training data can be insufficient



2. Quantity of Data Needed to Digitalise Historical Documents

Dataset Name	Training Set	'Evaluation' Set
Saint Gall dataset	468	942
Parzival dataset	2'237	2'240
Washington dataset	325	331
Specchieri MarVen dataset Style 0	635	424
Specchieri MarVen dataset Style 1	1672	719

Saint Gall dataset

N. Training Samples	CER
100	19,53%
200	10,82%
300	8,13%

Parzival dataset

N. Training Samples	CER
100	96,14%
200	23,48%
300	15,4%

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	8,12%	9,54%
200	5,68%	6,63%
300	4,98%	5,58%

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	96,38%	65,48%
200	6,46%	6,94%
300	2,97%	4,43%

2. Quantity of Data Needed to Digitalise Historical Documents

N. Training Samples	CER
100	69,39%
200	28,44%
300	23,33%

Washington dataset

N. Training Samples	CER
100	67,73%
200	34,64%
300	23,62%

Specchieri MarVen Style 0 dataset

N. Training Samples	CER
100	65,97%
200	32,94%
300	24,77%

Specchieri MarVen Style 1 dataset

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	24,57%	46,35%
200	9,20%	9,65%

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	44,37%	30,8%
200	15,34%	12,27%

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	31,90%	27,09%
200	14,31%	12,46%
300	11,42%	10,56%

2. Quantity of Data Needed to Digitalise Historical Documents

- 687.3 M images of printed English
- 17.9 M of handwritten text lines in English

1. $TrOCR_{small}$ $DeiT$ encoder and $MiniLM$ as decoder
2. $TrOCR_{base}$ $BEiT$ encoder $RoBERTa$ decoder

$TrOCR_{small}$ results

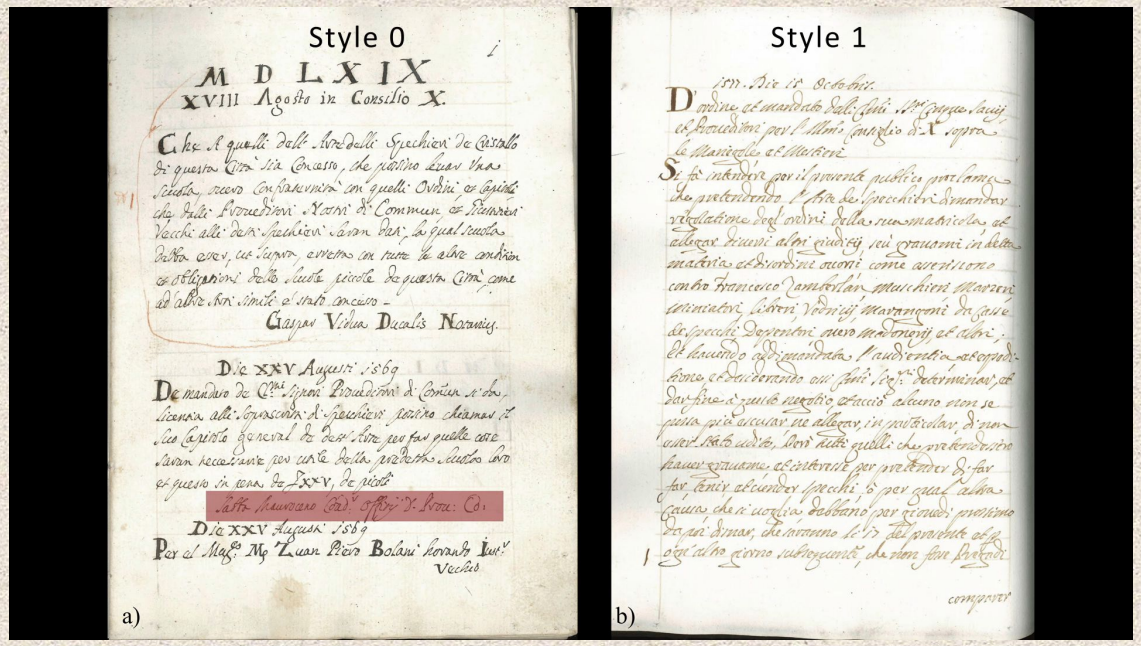
Dataset	Test CER	Test WER
Parzival	73,74%	100%
Saint Gall	80,12%	100%
Washington	79,59%	100%
Specchieri MarVen - General	80,22%	100%

$TrOCR_{base}$ results

Dataset	Test CER	Test WER
Parzival	1,98%	8,15%
Saint Gall	6,72%	37,85%
Washington	8,21%	19,25%
Specchieri MarVen - General	4,53%	15,72%

3. A Dataset we Want to Digitalise

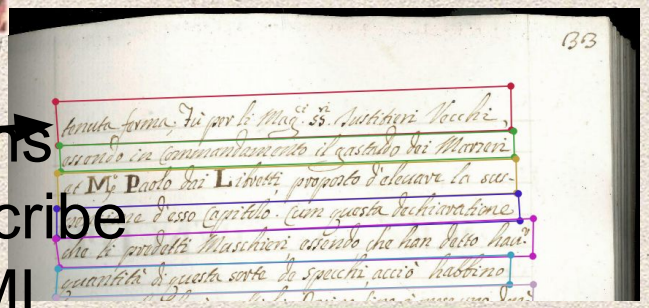
"The Marigold books contain the statutes – i.e., the regulations – of the devotional brotherhoods, associations or corporations of the arts and crafts in Venice during the Middle Ages and the Early Modern Period"



3. A Dataset we Want to Digitalise

Numerous abbreviations → Transcription challenge

labelme



- Rectangles and polygons
- No unique way to transcribe

Historical documents for ML applications:

1. "diplomatic transcription"
2. Removing the expanded abbreviations
3. **Normalised transcription**

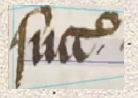
e mandato de Cl:mi Signori Proveditori di Com_un si da ,

e mandato de Clmi Signori Proveditori di Comun si da ,

e mandato de Cl(arissi)mi Signori Proveditori di Com(m)un si da ,

4. Abbreviations Expansion – A Missing Answer

- HTR models to recognize and transcribe abbreviations *while digitalizing the text*
- *Abbreviations:*
  1. *Optimize available writing space*
  2. *Standardized symbols* with *variations across regions*

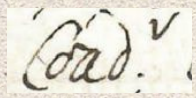


Succ.

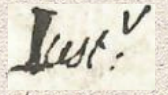
Succ(essoris)

Succ(essoribus)

r



Coad(iuto)r

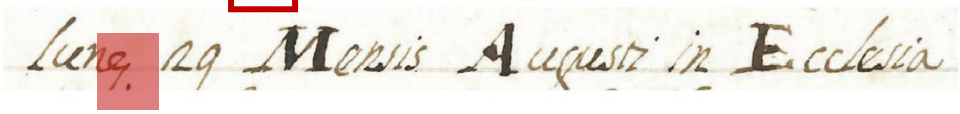


Iust(itie)r

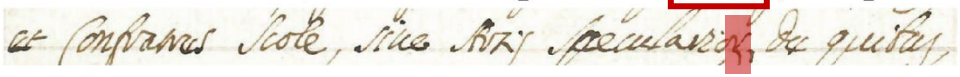
4. Types of Abbreviations

Six types of abbreviations

lun(ae) 29 Mensis Augusti in Ecclesia



et Confratres Scole , sine Artis Speculario(rum) , de quibus ,

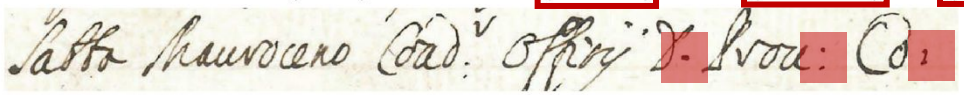


come (per) fede appar fatte sotto di 25 del pr(ese)nte mese , et

Class 0 (Symbol)

- Same level main text
- ~~Its expansion~~ is after/inplace

Sabba Mauroceno Coad(iuto)r Offitii D(omini) Prov(isorum) Co(mmun)

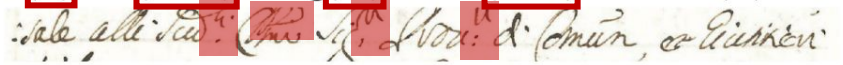


Class 1 (Suffix)

- Same level main text
- Its expansion ~~is after/inplace~~

4. Types of Abbreviations

= sale alli Sud(et)ti Cl(arissi)mi Sig(no)ri Prov(edito)ri di Com(m)un , et Giustitieri

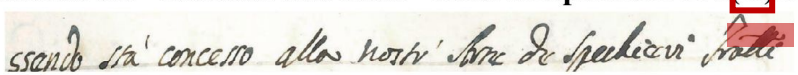


Class 2 (Ending right apex)

- Above the main text
- Its expansion is after/inplace

er el Mag(nifi)co M(e)s(sier) Zuan Pie(t)ro Bolani horando Iust(itie)r

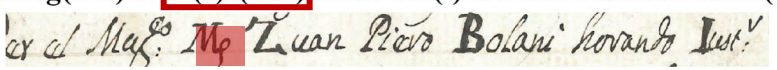
ssendo sta ' concesso alla nostr ' Arte de Spechieri Fra(te)lli



Class 3 (One grapheme above the word)

- Above the main text
- Its expansion is after/inplace

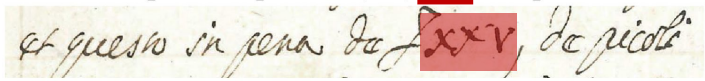
er el Mag(nifi)co M(e)s(sier) Zuan Pie(t)ro Bolani horando Iust(itie)r



Class 4 (Internal crasis)

- Same level as the main text
- Its expansion is before and after

et questo in pena de (lire) (25), de picoli



Class 5 (Numbers)

- Same level as the main text
- Translation of Roman numerals to base ten numbers

5. Main Tools

Both originated by EU Projects:

1. *eScriptorium*

(based on *Kraken*)
https://test2.fondue.unige.ch/
FREE

2. *Transkribus*

(the most used sw)
https://www.transkribus.org/
SUBSCRIPTION

Conclusions

1. **Line-level** models: best decision between the *effort needed* to create a training dataset and the *accuracy* of transcription
2. For historical documents it is important to **securing "normalised" transcriptions** and **develop models to obtain such a transcription**
3. Quite old topic, but it still needs a lot of effort to automate the digitalisation
4. **Collaboration** between **CS** and **Humanities scholars** is important

Thanks!

*We have **open positions**:*

1. ***Research Fellow in Digital Twins***
2. ***Post Doc in ML for Earth Observations***

More…

*Are you a student willing to do a **master thesis** with us?*

We are willing to follow you in projects about:

1. ***NLP***
2. ***Clustering***

Get in contact!
Sara.Ferro@iit.it