

# Machine learning-supported Historical Handwritten Recognition: Tools and Trends

27/06/24

**Sara Ferro, Post. Doc. @CCHT**

**Mail: [Sara.Ferro@iit.it](mailto:Sara.Ferro@iit.it)**

**@PyDataVenice**



# 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



Research, Intelligence and Technology for  
Heritage and Market Security

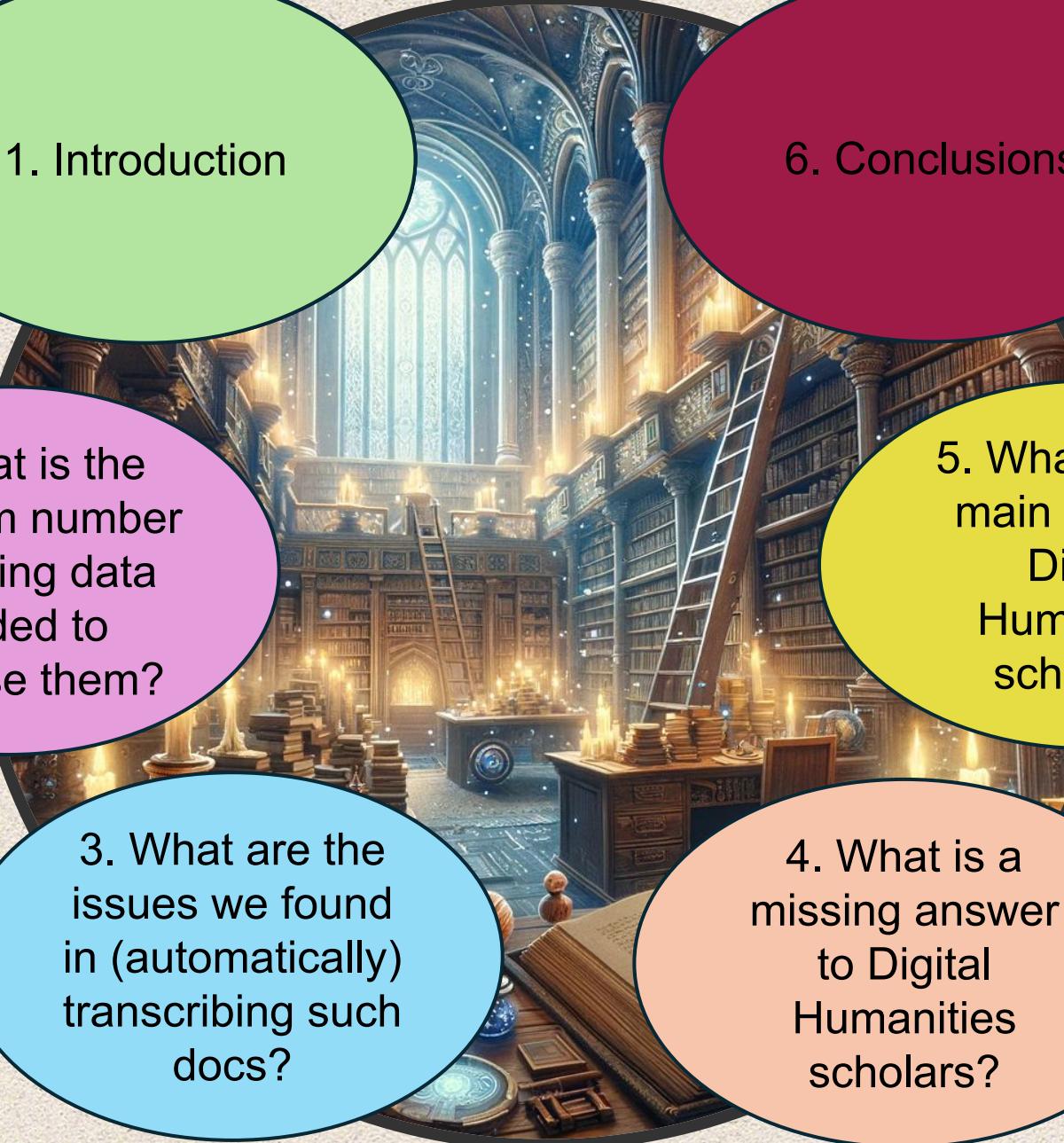


Università  
Ca' Foscari  
Venezia



ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY





1. Introduction

6. Conclusions

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?



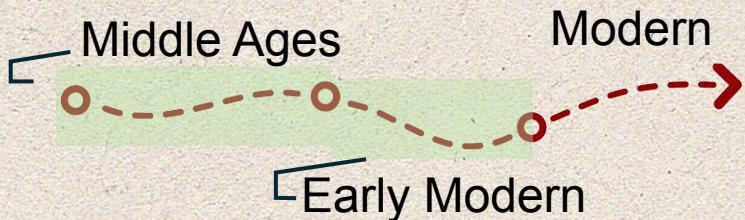
Università  
Ca' Foscari  
Venezia



ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

# 1.a. Historical Documents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

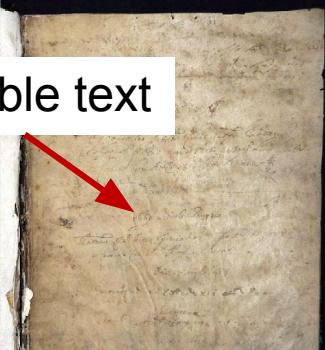
presentar alla CI(arissi)mi  
Sig(no)ri Prov(edito)ri di Comun

et

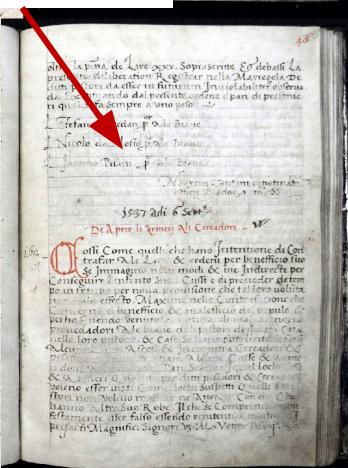
ce fai quelli aperti che li parevano necessari  
ca Vtli y le evron, conservation, o augmentare  
de dem. Rro d'ocres, o i quelli poveri, ce  
mezzano alla Cm. Vtli. S. & Cm. et  
Giustizian. Vecchi da esser confirmati o ap-  
paltati et poi da esser invidiabilmente eseguiti  
y tutti quelli dell' Rro Stora adatto, ce  
perpetui successori temporali,

# 1.a. Historical Documents – complications in transcribing

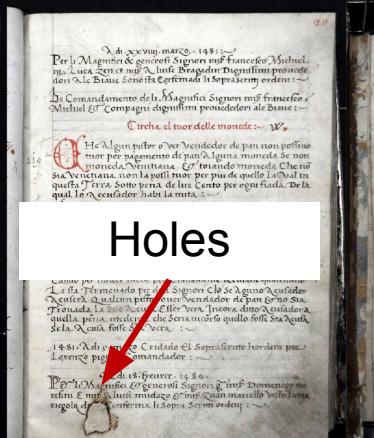
Illegible text



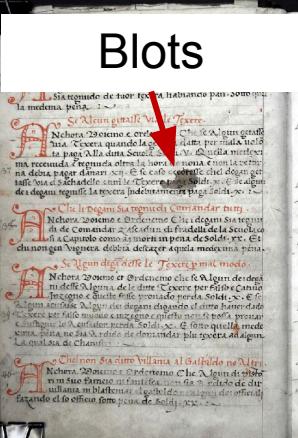
Humidity



Holes



Blots



Diverse layouts



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



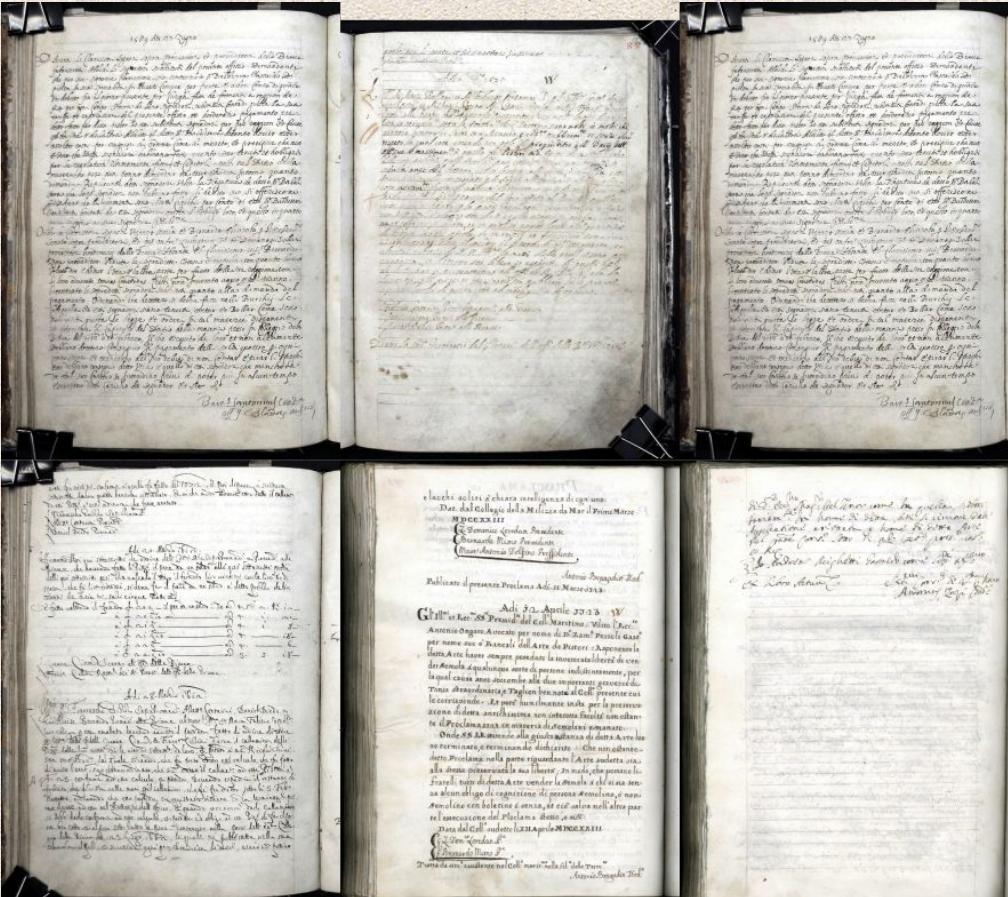
Few samples



Università  
Ca' Foscari  
Venezia



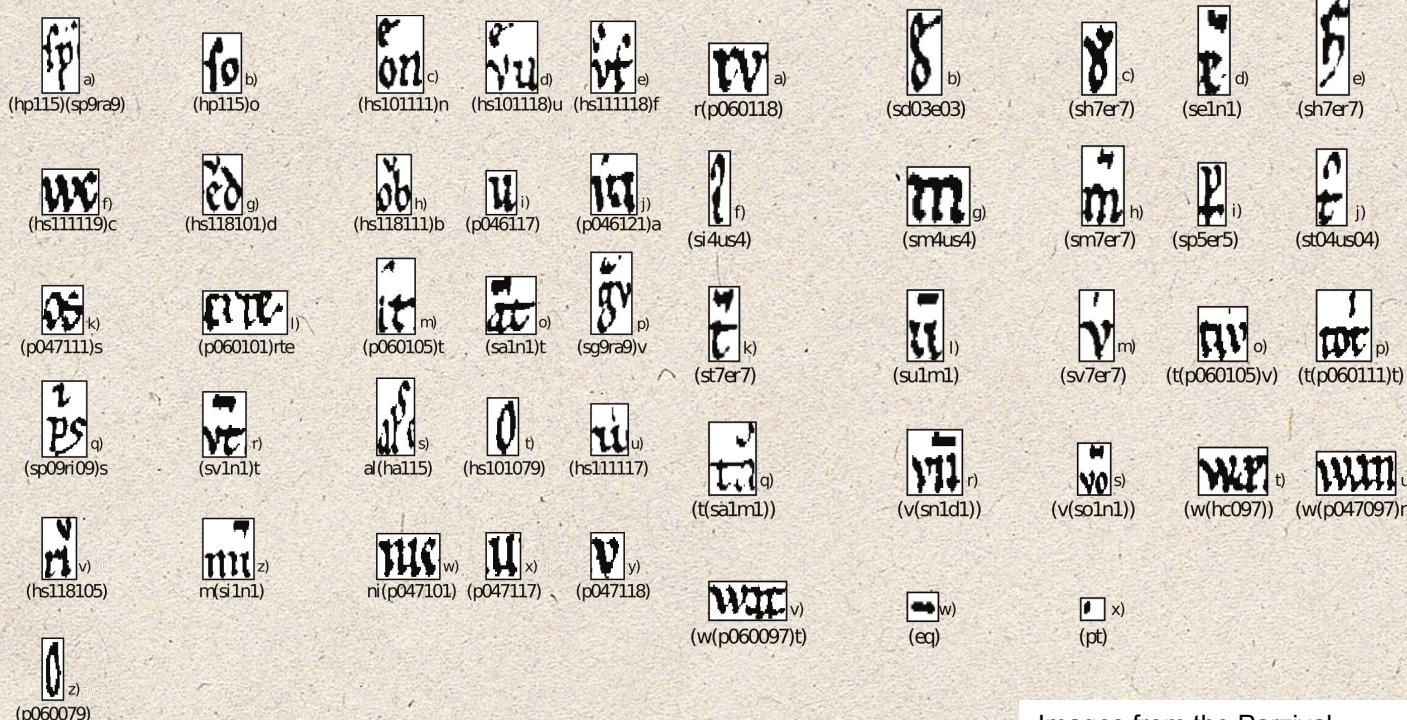
ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY



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

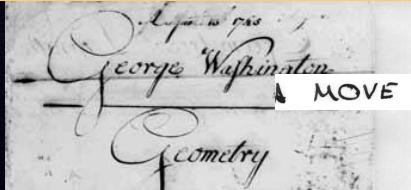


ueratio per omnem hiberniam celebris haberetur. & ueluti splen  
rimum exire cum oblatione dno offerebatur. sicut ma-  
gister commendauerit. ut irregulariter uice proficeret  
disciplina. & inter plurimos biremis multas sectatores  
dicitur. & auctor proposita unicus exemplis. Domi-  
nus iridit in eis. caro nesciatur. affectus magno uitium  
crevit. Augmento! Sapientia quoq; grecis se praemittente  
timo studio diuina episcopatu scripta. ut de doct' auro suo

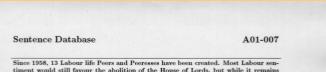
**Saint Gall**  
Latin, 9<sup>th</sup> c.,  
1 scribe



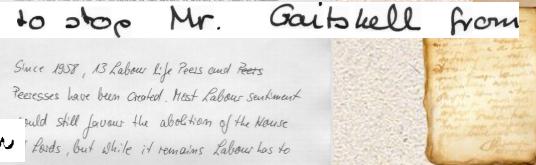
**Parzival**  
German, 13<sup>th</sup> c.,  
3 scribes



**Washington**  
English, 18<sup>th</sup> c.,  
2 writers



Since 1958, 13 Labour Life Poets and Poets  
Peacessers have been created. Most Labour sentmant  
ould still favour the abolition of the House  
of Lords, but while it remains Labour has to  
have an adequate number of members. THE  
two rival African Nationalist Parties of Northern  
Rhodesia have agreed to get together before the  
Challenge from Sir Roy Welensky, the Federal Presi-



**IAM**  
English,  
657 writers

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

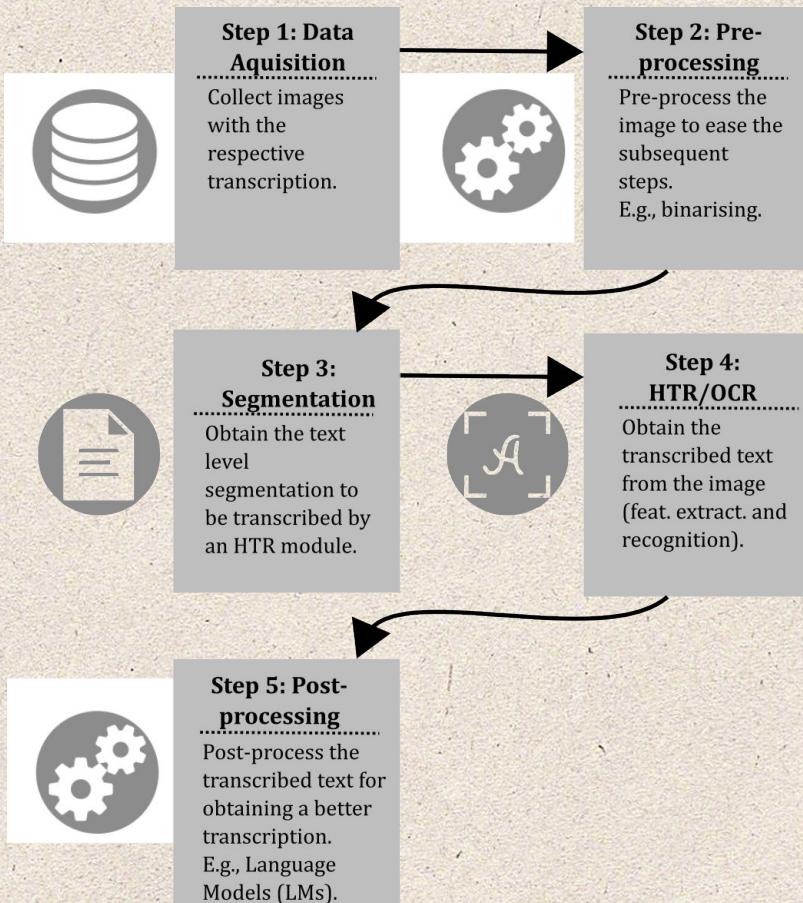
Max.

Min.  
Max.



Università  
Ca' Foscari  
Venezia

# 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

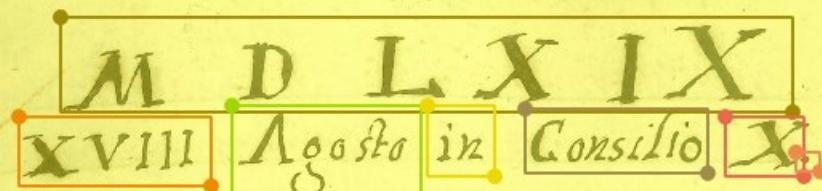


# 1.b. Historical Handwritten Text Recognition – Different Levels

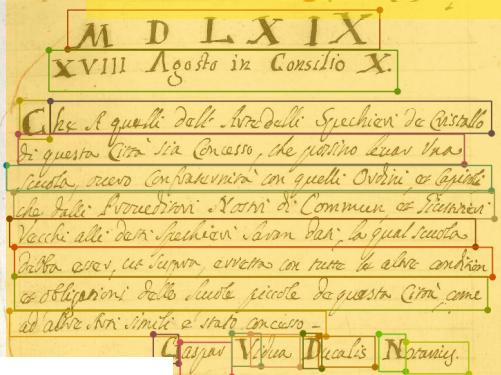
Character-level



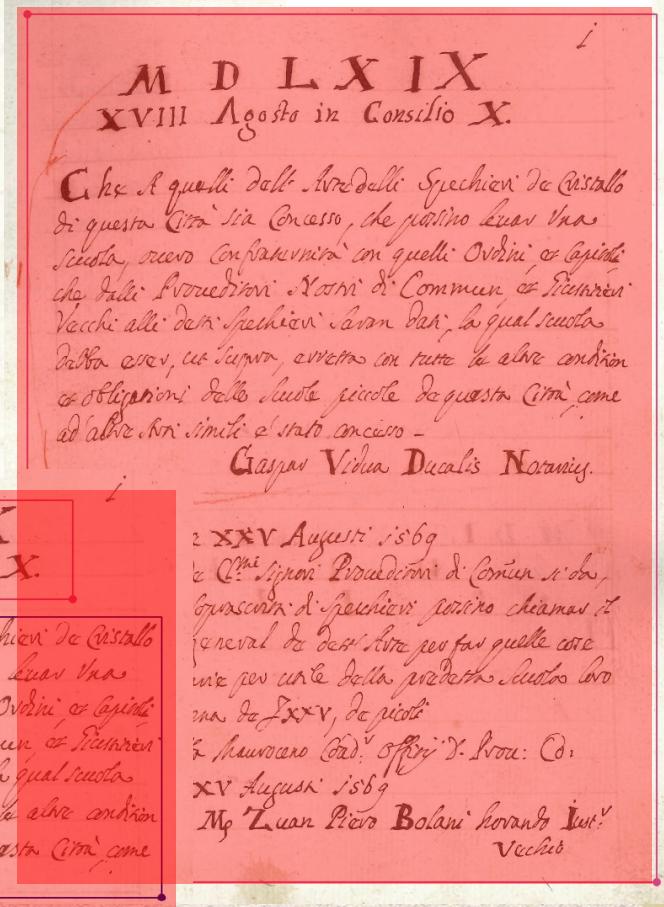
Word-level



Line-level



Page-level



Paragraph-level

Line-level

Francesco  
Guerri  
Neri

Costume



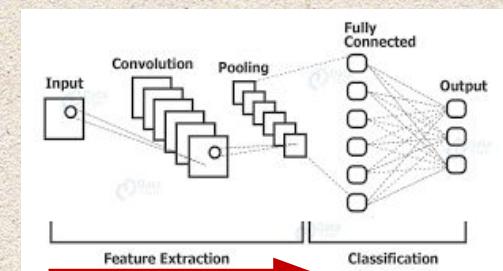
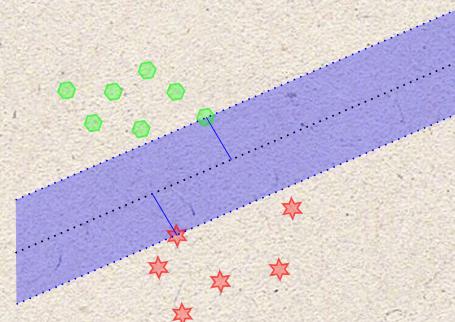
Università  
Ca' Foscari  
Venezia



ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

# 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



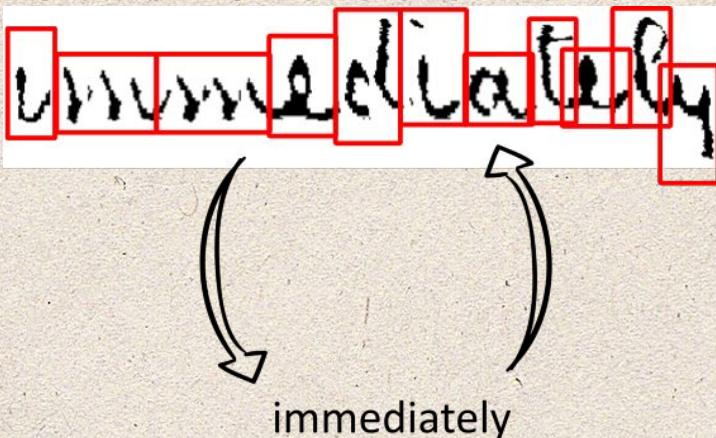
From basic features to more complex ones

# 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]



Università  
Ca' Foscari  
Venezia

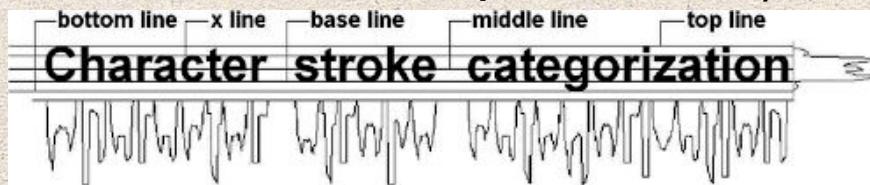


ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

# 1.b. Word-level Models

**Option:** Segmentation till the character-level

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

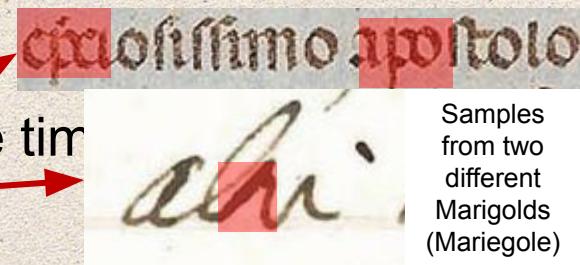


2. *Oversegmentation of words*: the segmentation problem becomes a recognition problem of the “cut points”

+ refinements

These methods fail when:

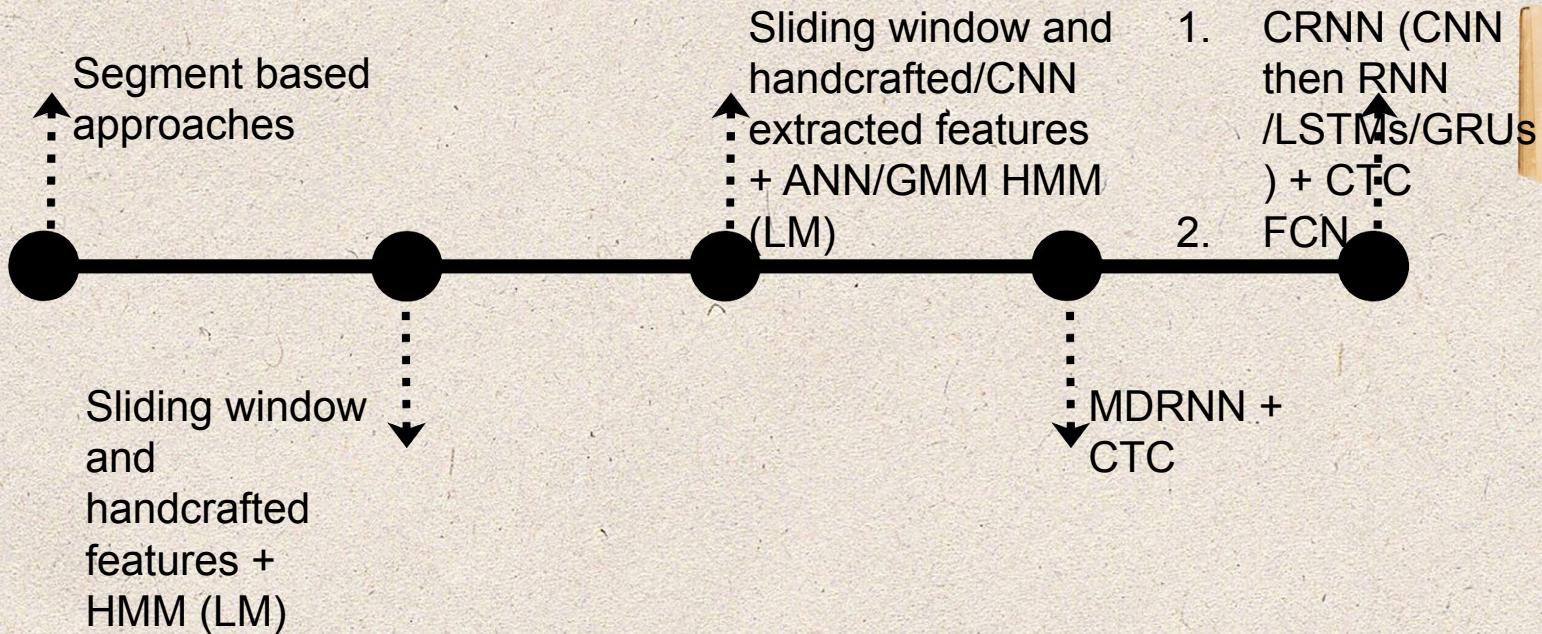
1. Characters touching multiple times
2. Warped characters



Samples  
from two  
different  
Marigolds  
(Mariegole)

# 1.b. Word-level Models

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

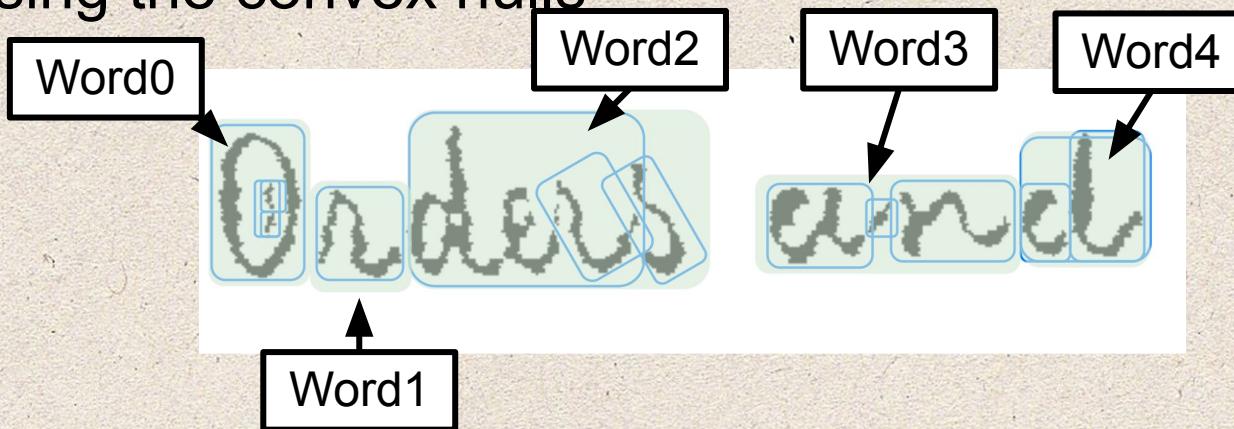


# 1.b. Line-level Model

Deterministic and unsupervised methods  
segment into words or characters

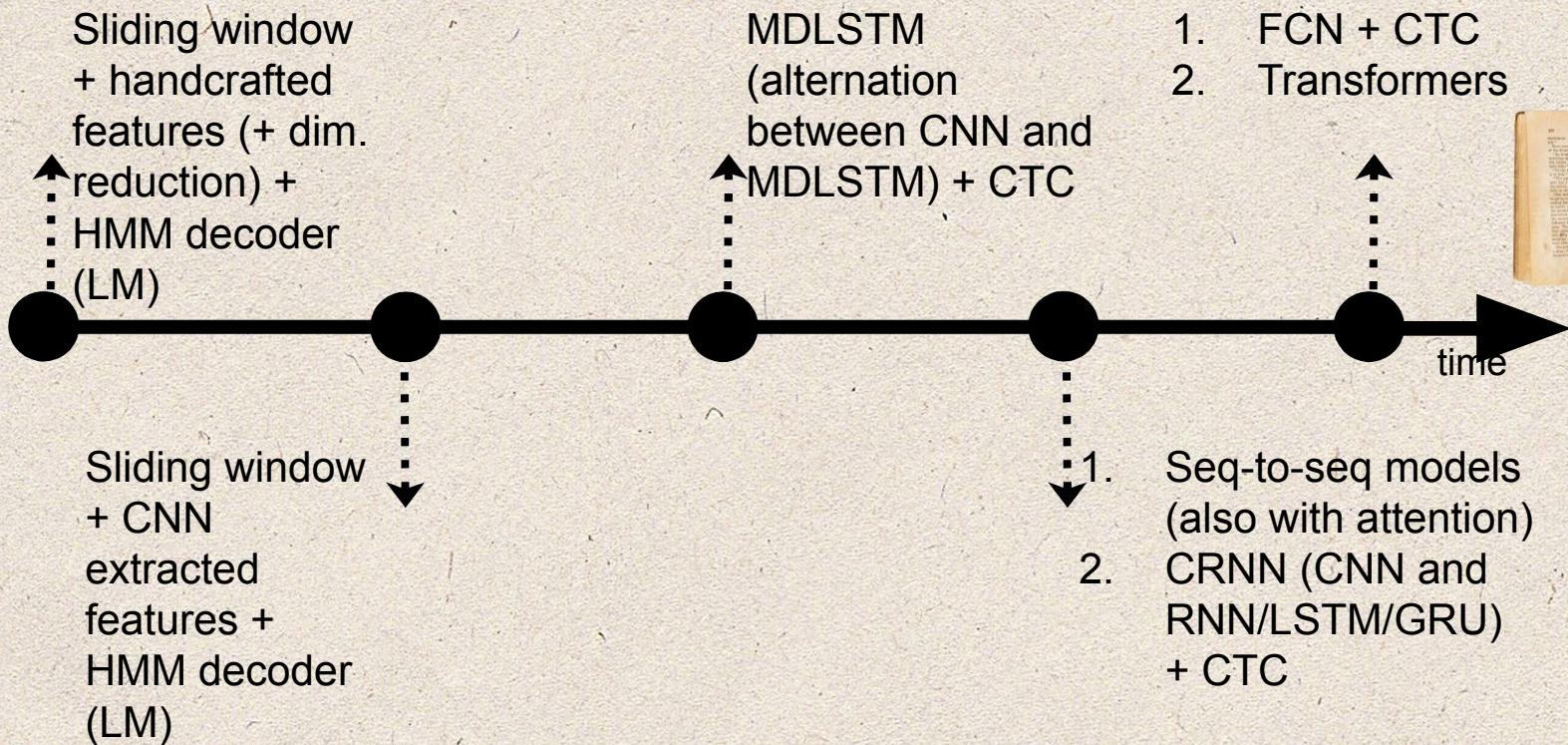
**Example (deterministic)**

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 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. 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

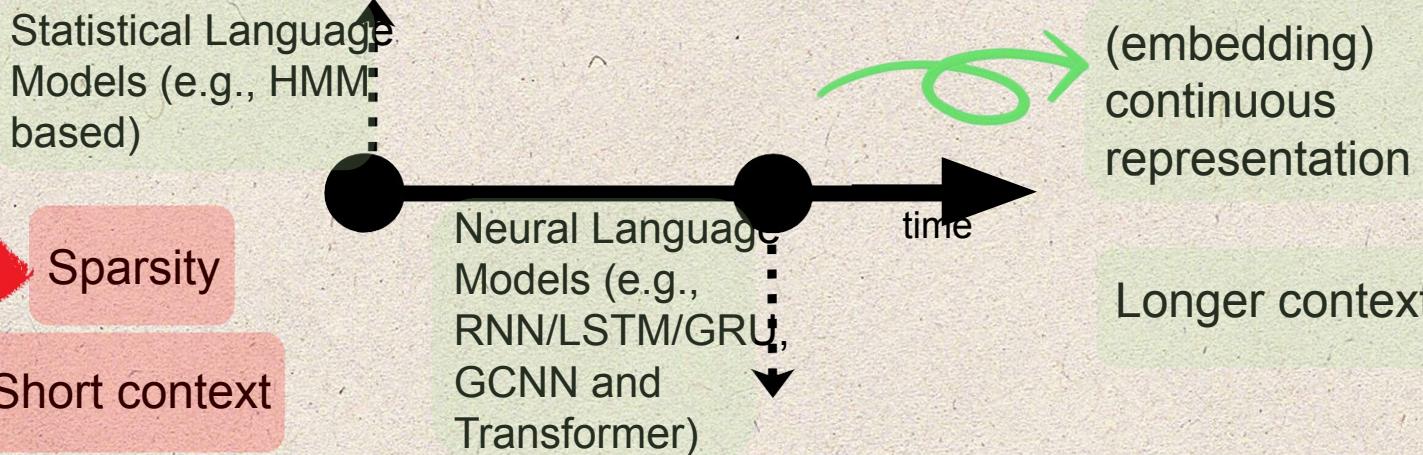
# 1.b. Post-processing Methods

Language Models (LMs) can be utilised in the:

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

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

Word at pos.



Università  
Ca' Foscari  
Venezia



ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

## 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

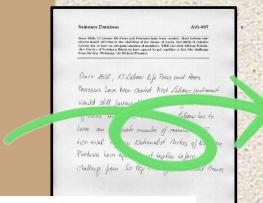


# CRNN

1. Data augmentation
  2. Fine-tuning

*10M params*

# Few hundred lines is enough

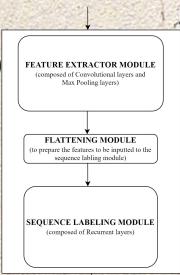


# Moder n English



## Latin, 9<sup>th</sup>

C.



# English, 18<sup>th</sup> c.



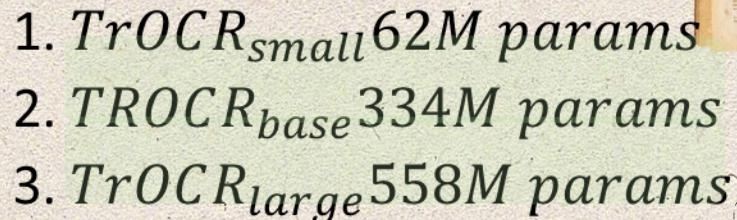
# Venetian, 18<sup>th</sup> c.



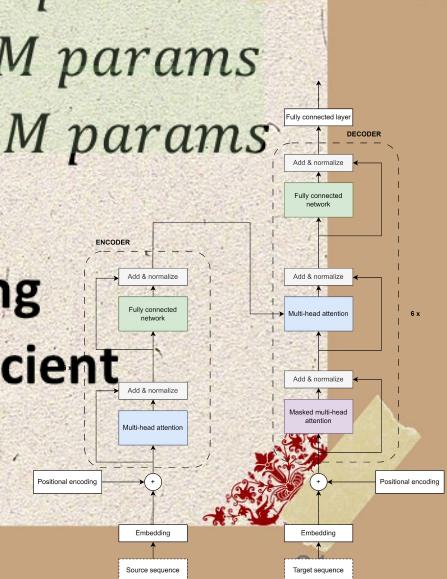
10

# Transformer (TrOCR released by Microsoft)

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



## 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
Specchiери MarVen dataset Style 0	635	424
Specchierie 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%



Università  
Ca' Foscari  
Venezia



ISTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

# 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<sub>small</sub>* *DeiT* encoder and *MiniLM* as decoder
2. *TrOCR<sub>base</sub>* *BEiT* encoder *RoBERTa* decoder

*TrOCR<sub>small</sub>* 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<sub>base</sub>* 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”



**Style 0**

M D L X I X  
xviii Agosto in Consilio x.

Che il quelli dell' Accademia Spechier de Castello di questa Città sia Concesso, che perche sieno vissute  
sesta, novena e decima, con quelli Ovini, et Capisti  
che fanno Associazioni, Nazari, et Communi, et Poveri  
Vecchi alle due Accademie, siano datte, la qual rientra  
della ex, et super, etmetre in tutto le altre antiche  
e consigliarie delle Sante piecole de questa Città, come  
di altre, non simile, o stato ancora -  
Grazie Venerabili Decali Noviss.

Dic xxv Augusti 1569  
Demando de Quo' Signo' Presidente d' Accademia Spechier  
Scavano alle Spechie, et peccati, fatto chiamare al  
Capitolo General de Genova per fare quella cose  
suan necessario, con sorte della sorte, tutto lo  
et questo in anno de LXXV, de recto

Alto numero F. offr. V. Rom. Co.

Dic xxv Augusti 1569  
Per il Mag. Mo' Zuan Piero Bolani honorabili Int.  
Vecchi

**Style 1**

D'ordine et mandato del Signor M. D. G. da Venezia, la cui  
et Provvidione per l' Ultimo Consiglio d' Accademia  
e Marigolda et Martini.

Si fa intendere per il presente nostro parlamento  
da potentissimo Signore, che perciò dimanda  
ratificazione del sortire della sua marigola, et  
alzare di tutti altri quelli, sei gravoni, in detta  
marigola, et svariane svariane, come avvenne  
con Francesco Lombolan, marchese Martini  
governatore, banchi, Venezi, Maranoni, da quale  
de peccati Dottorato aveva mandato et fatto.  
Et facendo appurata, l'autenticatio, ratificando  
come et consentendo omni fide, et determinato, et  
d' acciò è questo nostro statuto alcuno non se  
perita, ne' accusar, ne' allegar, in contradicere, et per  
non poter credere, dov' nella quale che potrebbe, non  
far cosa gravona, et contraria per credere, et far  
far, tenir, et candere peccati, & per qual altra  
causa, da n' uoglio delibano, per quanto possibile  
d' un d' un, che non ha, et per quel preche ab  
ogni alio giorno successante, de non far peccati

a)

b)



# 3. A Dataset we Want to Digitalise

Numerous abbreviations  
challenge → Transcription

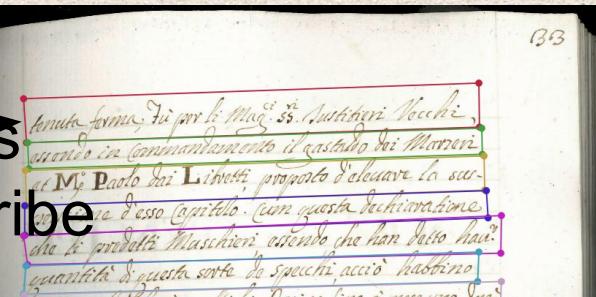
labelme



- 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<sup>mi</sup> Signori Proveditori di Com<sub>un</sub> si da ,  
e mandato de Cl<sub>m</sub>i Signori Proveditori di Com<sub>un</sub> si da ,  
e mandato de Cl(arissi)mi Signori Proveditori di Com(m)un si da ,



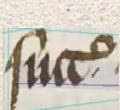
Università  
Ca' Foscari  
Venezia

**iit**  
INSTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

# 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  
)



Coad(v)  
Coad(iuto)r



## **lust(itie)n**

# 4. Types of Abbreviations

## Six types of abbreviations

lun(ae) 29 Mensis Augusti in Ecclesia

*Lun(ae) 29 Mensis Augusti in Ecclesia*

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

*et confratres scole, sine artis speculario, de quibus,*

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

*come si stile appar fatte sotto d'as' del p'nto mese, et*

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

*Sabba Mauroceno Coad(iuto)r Offitii D(omi)ni Prov(isorum) Co(mmun)*

### Class 0 (Symbol)

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

### 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  
*:ale alli Sd. Cl. Sg. d'Comun, et Giustieri.*

## **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  
*er al Mag. M. Zuan Piero Bolani horando Iust.*

## **Class 3 (One grapheme above the word)**

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

ssendo sta ' concesso alla nostr ' Arte de Spechieri Fra[te]lli

*ssendo sta' concesso alla nostr' Arte de spechieri fratelli*

## **Class 4 (Internal crasis)**

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

## **Class 5 (Numbers)**

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

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

*et questo in pena da XXV, de picoli*

# 5. Main Tools

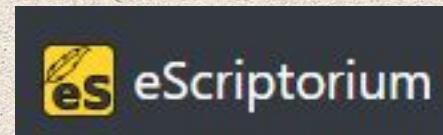
Both originated by EU Projects:

1. eScriptorium

(based on Kraken)

<https://test2.fondue.uniae.ch/>

 FREE

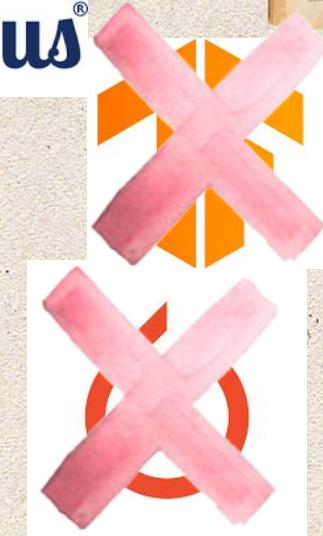


2. Transkribus

(the most used sw)

<https://www.transkribus.org/>

 SUBSCRIPTION



Università  
Ca' Foscari  
Venezia



INSTITUTO ITALIANO  
DI TECNOLOGIA  
CENTER FOR CULTURAL  
HERITAGE TECHNOLOGY

A robotic arm, with a metallic and mechanical appearance, is shown holding an open book. The setting is a grand, ornate library with tall, dark wood bookshelves filled with books. The lighting is dramatic, coming from a large arched window at the top of the stairs, casting light through the shelves and illuminating the pages of the book held by the robot.

# 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](mailto:Sara.Ferro@iit.it)