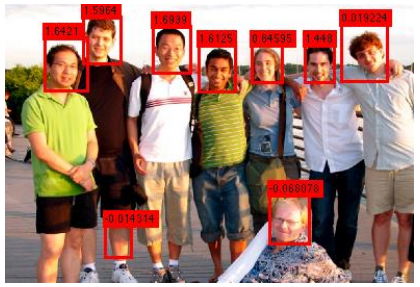
Progress through competition: linking humanities data with computer science research

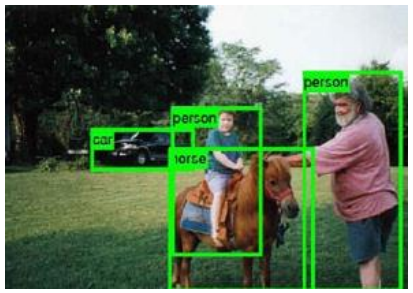
Florian Kleber, Wien, 02.11.2017

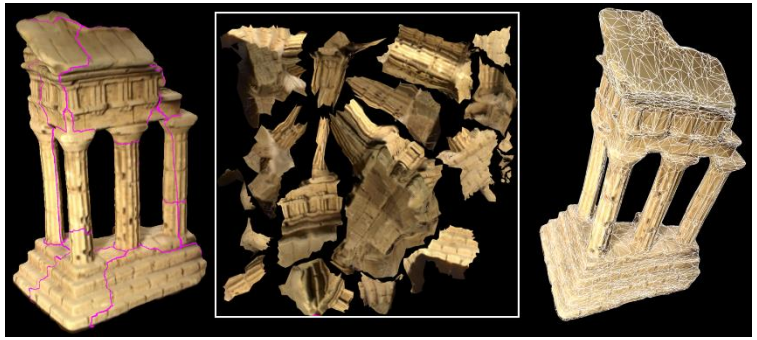
Why do we need DATASETS?

Computer Vision Tasks

- Computer Vision
  - Recognition
    - Detection
    - Classification
  - Segmentation
  - Reconstruction



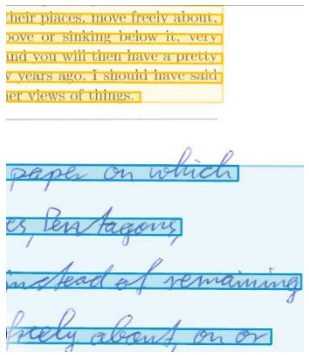




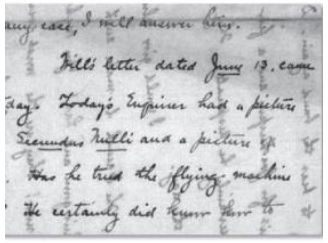
Document Analysis Tasks

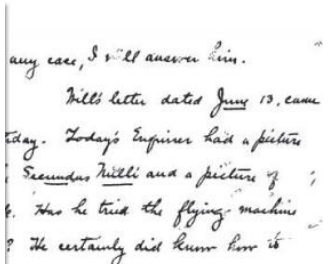
- Computer Vision
  - Recognition
    - Detection
    - Classification
  - Segmentation
  - Reconstruction

Optical Character Recognition (OCR)
Handwritten Text Recognition (HTR)
Text Classification (handwritten, printed)

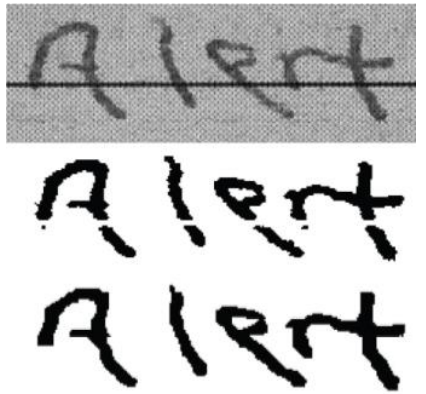


Binarization



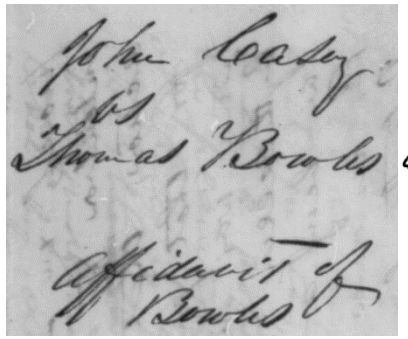


Character Reconstruction

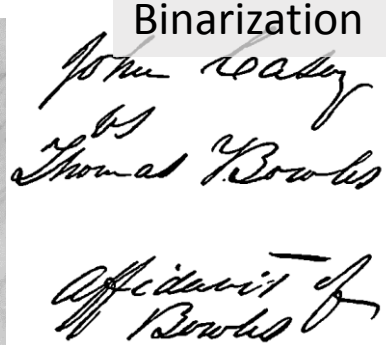


Datasets

- Different tasks require different images, Ground Truth (GT), dataset size



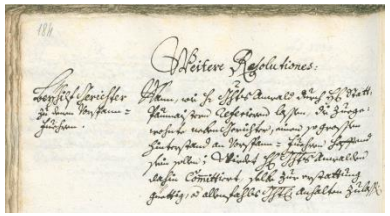
Binarization

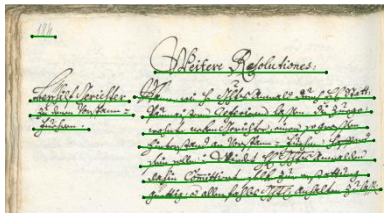


Skew

-11.66°

Baseline

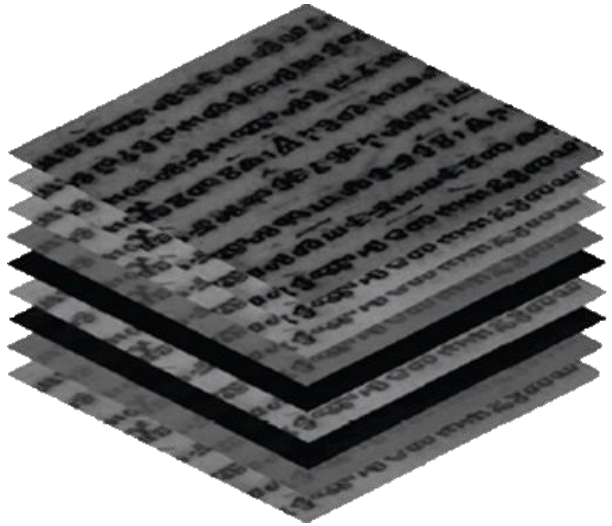


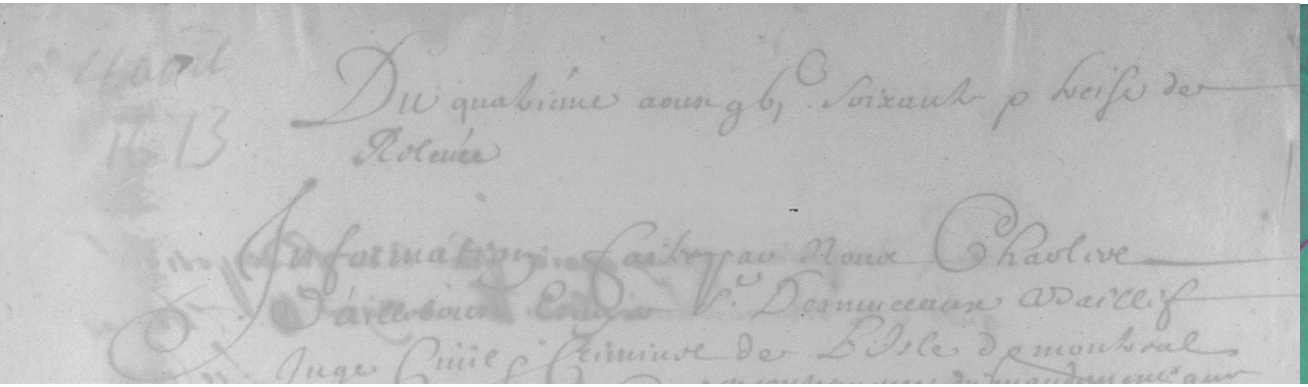


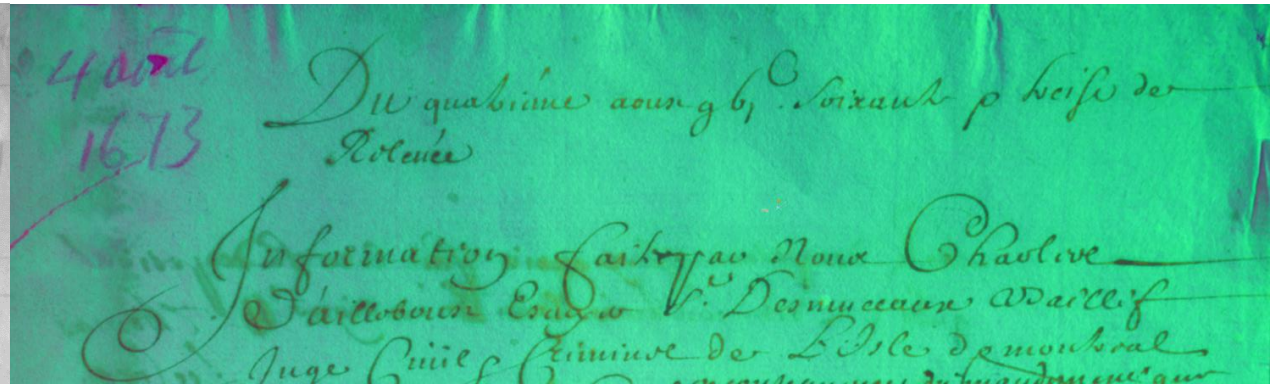
- Complexity of GT has a huge variation
  - Time consuming to create
  - Synthetic vs. real images
  - Synthetic images often do not represent the reality

MultiSpectral Text Extraction Contest (MSTex)

- Extract (binarize) only text written with a specific ink (ICDAR 2015 competition)
- Multispectral images are needed
- Specific requirements on the image acquisition



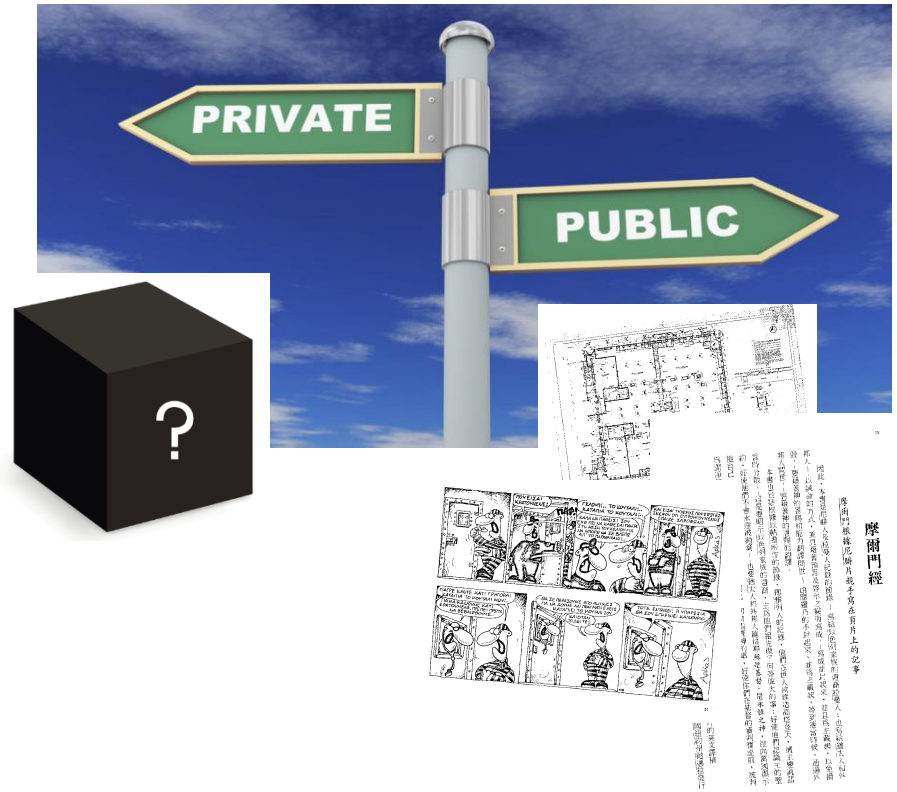




Why do we need PUBLIC datasets?

Evaluation of existing methods

- Allows comparison of different methods using the same data
- Public vs. private (not comparable)
- Specified splitting of training and test
- GT is time consuming/expensive



Author	Dataset	Size	Precision
Imade [ITW93]	own	unknown	82.0%
Kuhnke [KSK95]	NIST	≈ 1/3 page [1,068 chars]	78.5%
Fan [FWT98]	own	≈ 12 pages [50 text blocks]	86.0%
Pal [PC01]	own	≈ 150 pages [600 images]	98.6%
Zheng [ZLD04]	own	94 pages	96.0%
Kavallieratou [KSA04]	IAM-DB/GRUHD	50 pages	98.2%
Kandan [Kann+07]	own	150 pages	93.2%
Koyama [Kum+11]	ETL	≈ 1 page [1,000 chars]	97.0%
Chanda [CFP10]	own	≈ 73 pages [39,190 words]	95.9%
Pinson [PB11]	NIST	360 pages	84.6%
Zemouri	IAM-DB	21 pages	
Zagoris [Zag+12]	IAM-DB	103 pages	98.9%
Zhang [ZL12]	IAM-DB	50 pages	99.9%

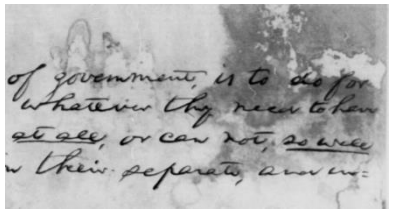
Example: Datasets used for Text Classification

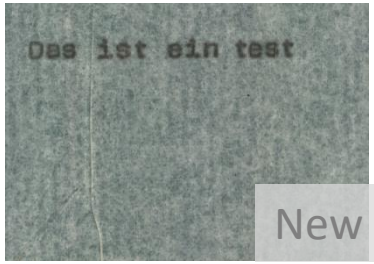
Why do we need NEW public datasets?

- Continous demand of new datasets
  - No longer challenging due to size, enhanced methods and technology, different needs
- Avoid specialization to a specific dataset
- E.g. certain type of problems are not represented
- Example Writer Identification

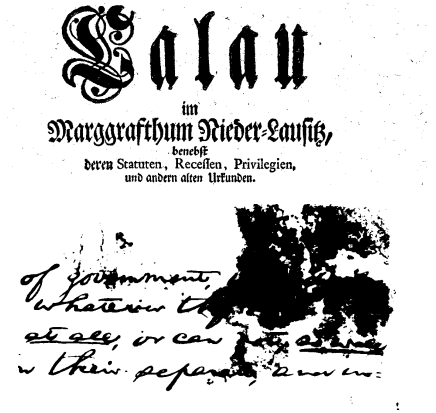
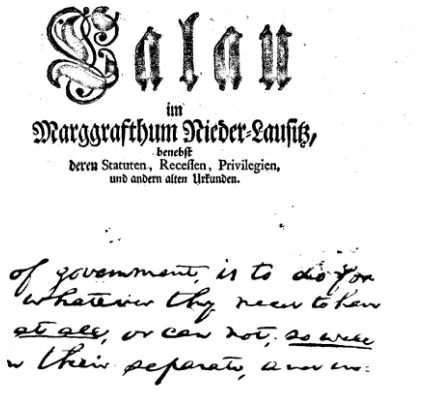
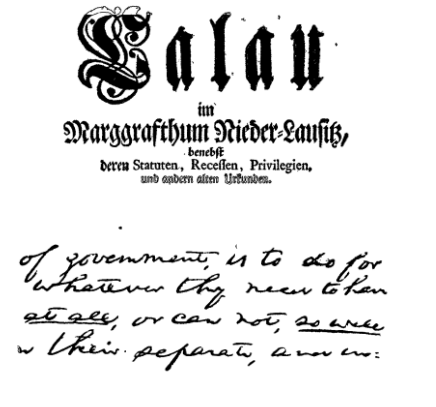
	2011	2012	2013	2014	2015
ICDAR 2011 (cropped) 26 Writer, 208 pages	90,9% (19 errors)		93,8% (11 errors)	94,7% (9 errors)	
ICDAR 2013 250 Writer, 1000 pages			90,9% (91 errors)	97,1% (29 errors)	99,9% (1 error)

Binarization Results DIBCO 2009





New type of noise

Otsu	Sauvola	Wolf	Su
			

Why do we need competitions?

Competitions

- Objective comparison of different methods on the same data (also same splitting of training and test set)
- Test set or GT is not public
- Avoid adjusting of parameters to the specific dataset (overfitting)
- Standardized evaluation metrics
  - Competitions can establish standards for evaluation
  - Results of new approaches are compared with results of competition methods
  - Evaluation of other methods on *private* datasets is often not possible since an implementation (binary) or detailed parameter set is not provided.

Choice of Metric

- Example Binarization: binary result is subjected to OCR – corresponding result is evaluated with respect to character or word accuracy

Akt 3VII-DK

Scientific Puzzle Solving: Current Techniques

Keywords: Puzzle, Reconstruction, Ancient Documents

Introduction

Finding solutiòns for puzzles need a well-defined definition of

Akt 3VII-DK

Scientific Puzzle Solving: Current Techniques

Keywords: Puzzle, Reconstruction, Ancient Documents

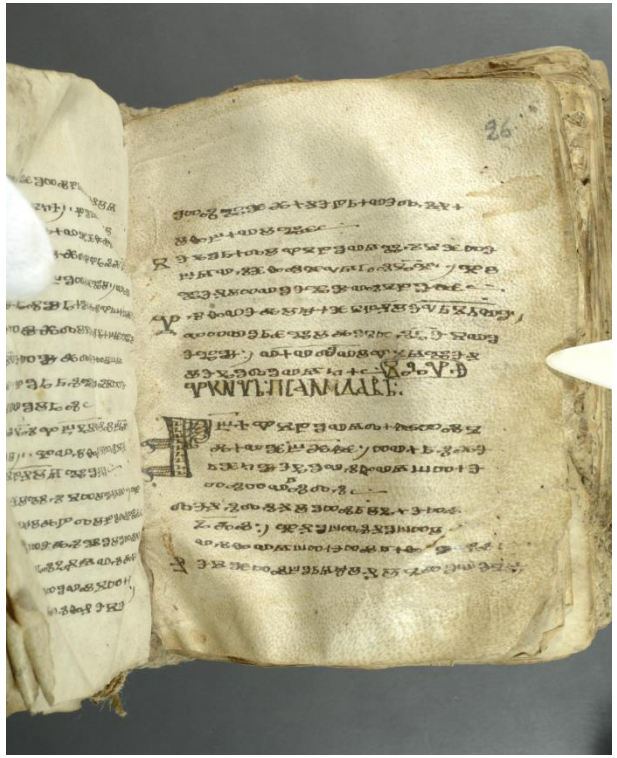
Introduction

Finding solutiòns for puzzles need a well-defined definition of

Adobe Acrobat OCR:

Ak t l:~ u-01(

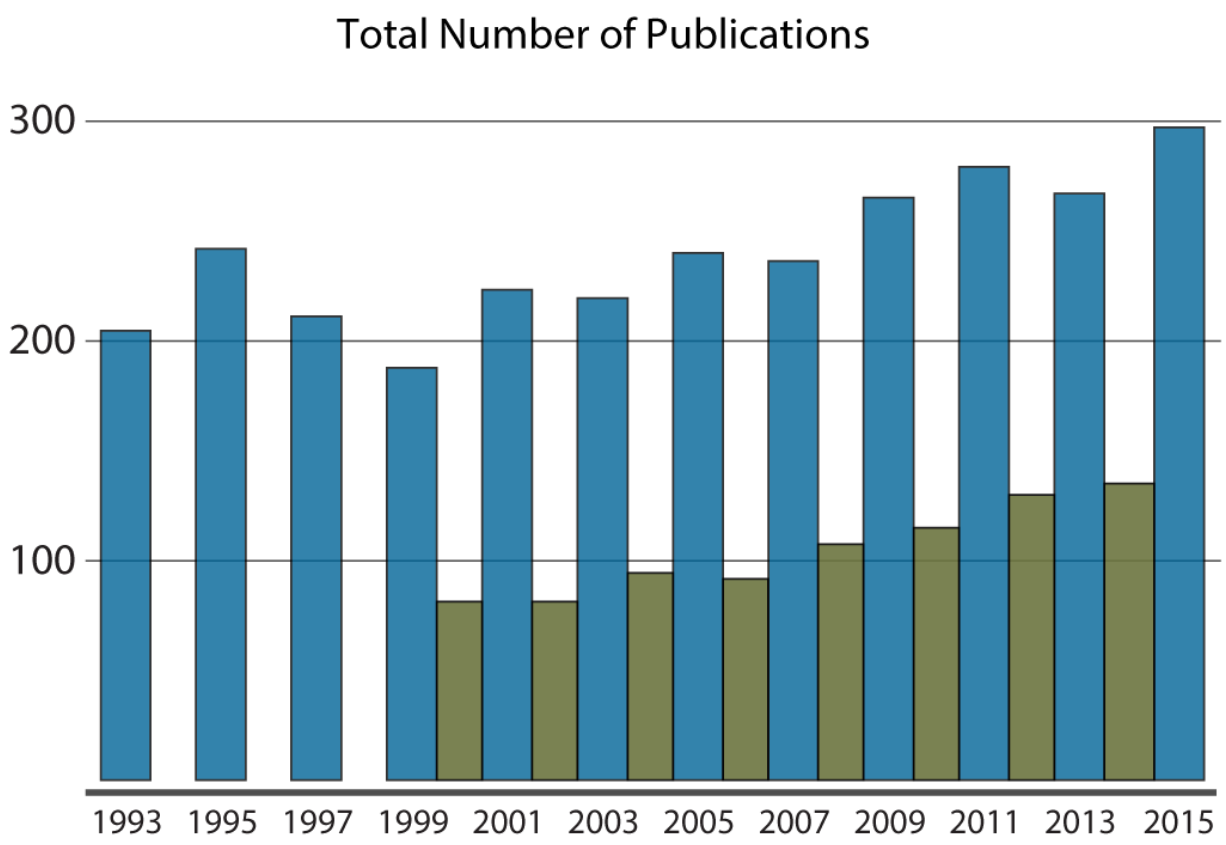
Scientific Puzzle Solylng: Current Techniques
Keywords: Puzzle, Reconstruction. Ancient Documents
Introduction
Finding solutlins fur pUlzles need I well-defined definition of

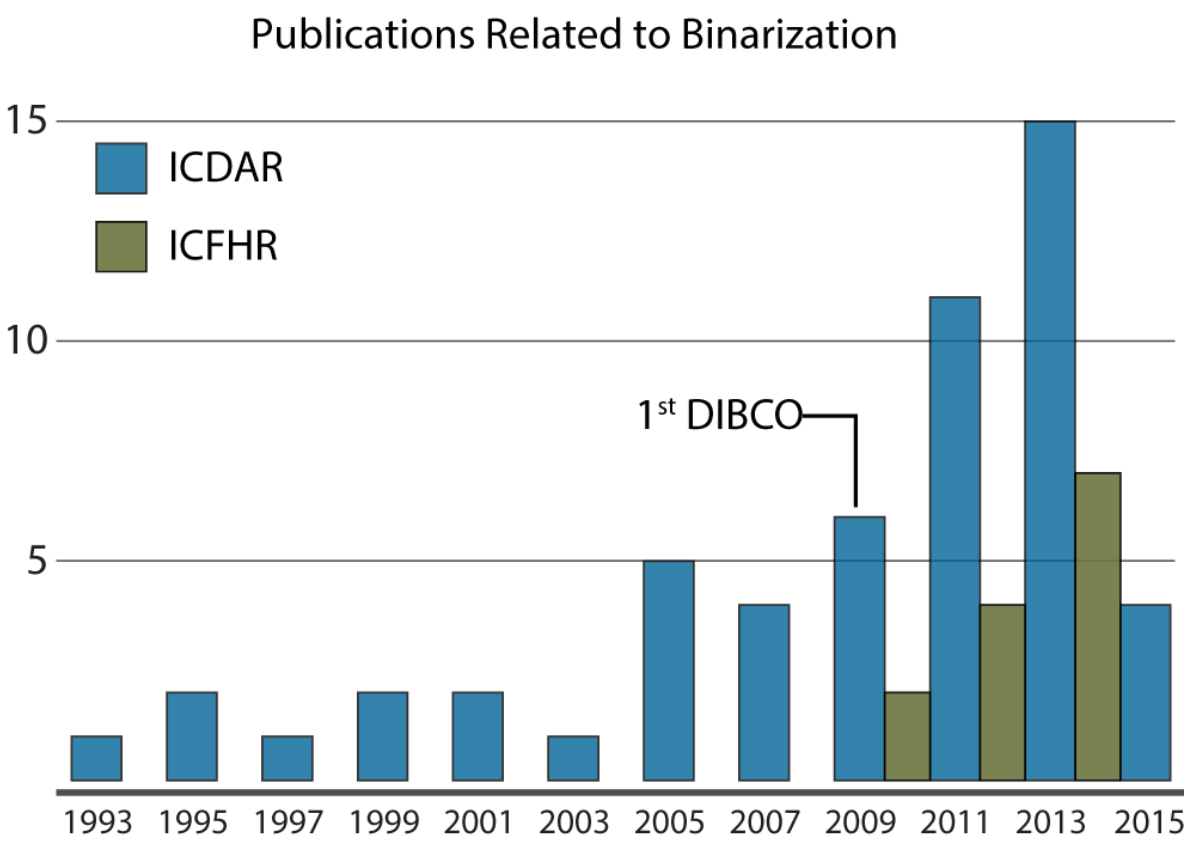


Evaluation Standards

Method	# pages	Precision (word)	Precision (text line)	F-Measure
Kavallieratou [KSA04]	50		98.2	
Zemouri [ZC11]	7	98.3		
Zagoris [Zag+12]	103			98.9
Zhang [ZL12]	50			99.9
Proposed no Seg	1,534	97.9	99.6	99.8

Impact of Competitions





Current Competitions

Current Competitions in Document Analysis

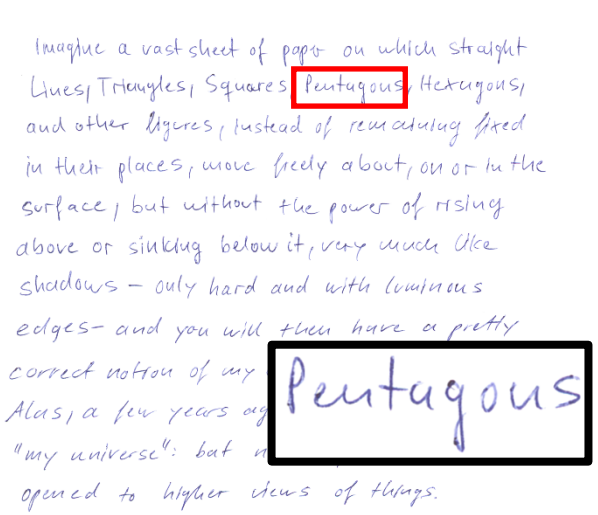
- ICFHR 2015
  - Recognition of Documents with Complex Layout
  - Robust Reading
  - Smartphone Document Capture and OCR
  - Handwritten Text Recognition on the tranScriptorium Dataset
  - Text Line Detection in Historical Documents
  - Keyword Spotting for Handwritten Documents
  - MultiSpectral Text Extraction Contest
  - Signature Verification and Writer Identification Competitions
  - Multi-Script Writer Identification and Gender Classification
  - Video Script Identification
  - Text Image Super-Resolution

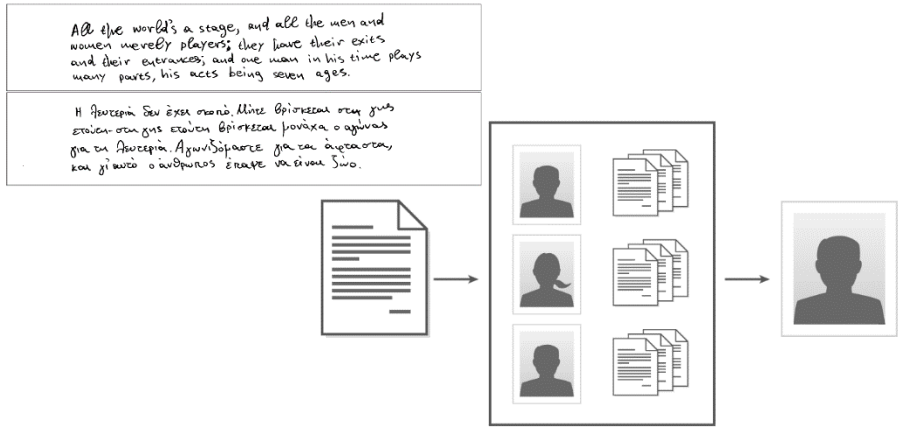
Current Competitions in Document Analysis

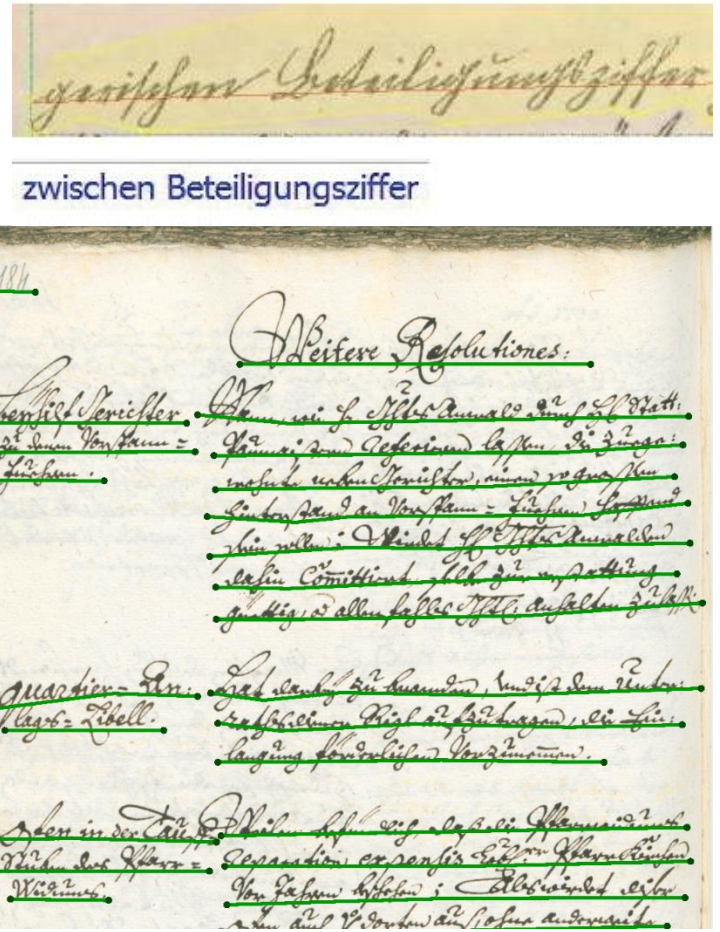
- ICDAR 2016
  - Competition on the Classification of Medieval Handwritings in Latin Script
  - Competition Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts
  - Competition on Arabic Online Text Recognition using Online-KHATT Database
  - Competition on Multi-script Writer Demographics Classification using „QUWI“ Database
  - Competition on Recognition of Handwritten Mathematical Expressions
  - Handwritten Keyword Spotting Competition
  - Handwritten Document Image Binarization Competition
  - Competition on Local Attribute Detection for Improving Handwriting Recognition
  - Competition on Handwritten Text Recognition on the READ Dataset

Planned READ competitions

- Handwritten Text Recognition (continuation of 2016)
- Layout Analysis with different tasks
- Writer Identification competitions
- Keyword Spotting







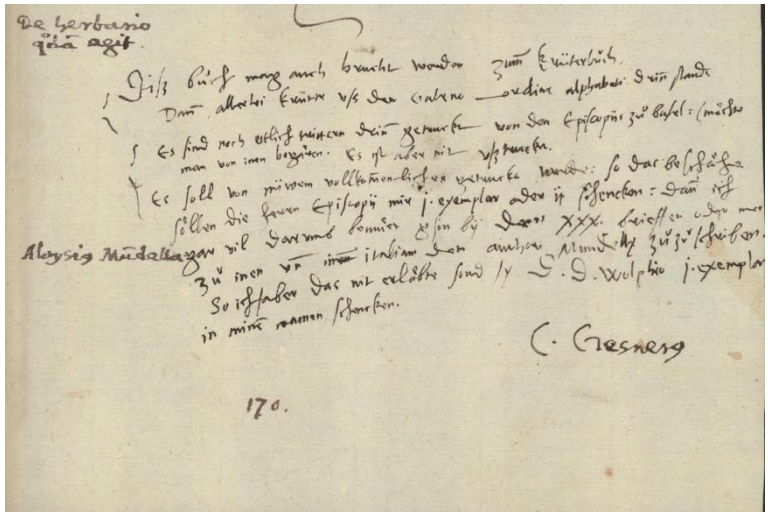
Baseline detection

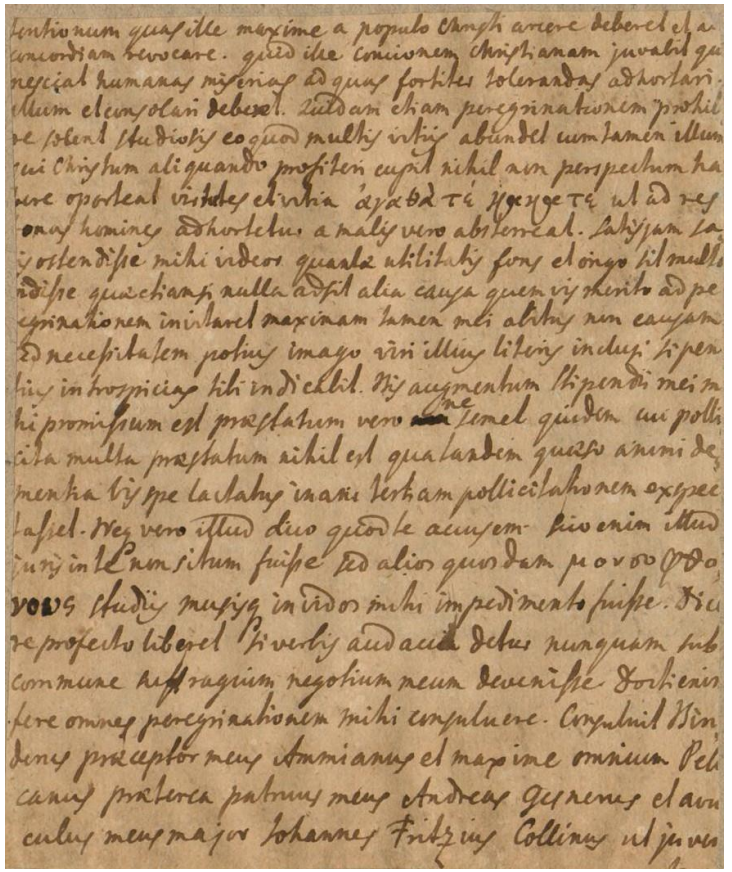
- Goal is to find the baseline of text lines
- Needed for Handwritten Text Recognition
- Also empty pages need to be detected
- Different collections to cover a variety of document classes
- 2250 pages

- 9 Collections are used from
  - Archive Bistum Passau
  - Brabant Historisch Informatie Centrum
  - Stadtarchiv Bozen
  - EPFL - Venice Time Machine
  - Humboldt Universität Berlin
  - National Archive Finland
  - Staatsarchiv Marburg
  - University College London
  - Universitätsbibliothek Basel

Writer Identification Competition

- Dataset from Universitätsbibliothek Basel
- First dataset on historic documents
- 720 writers
- 5 pages per writer
- Cropped manually





Conclusion

- Scientifc communities need public datasets and
- well defined metrics
- Allow objective evaluation
- Datasets and competitions can establish standards
- Competitions show the ongoing research achievements

Thank you