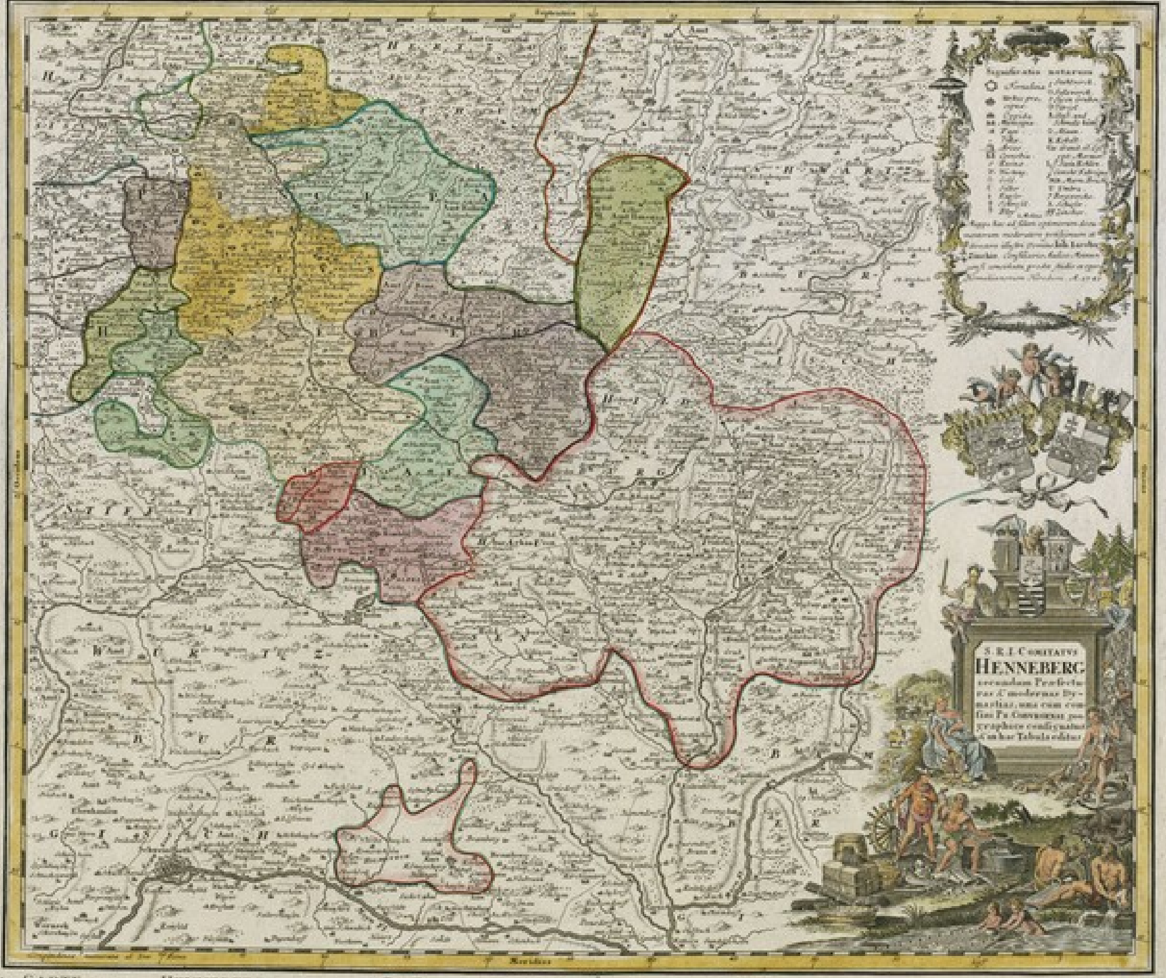
Efficient Algorithms and User Interaction for Metadata Extraction from Historical Maps

Benedikt Budig, Universität Würzburg

Overview

- Historical Maps: what and why?
- Sketch of a Pipeline
  - from bitmap image to georeferenced metadata
- Module 1: Locate Map Elements
- Module 2: Match Labels and Place Markers
- Open Questions & Future Work

Karte der Gefürsteten GRAFSCHAFT HENNEBERG, mit dem angränzenden FVRSTENTH. COBVRG u. andern GRAENZLAENDERN, nach authentischen Documenten und Nachrichten verfertiget, und in ihre Aemter eingetheilet. Herausgegeben von Homannischen Erben in Nürnberg A. 1743



La CARTE du Comté de HENNEBERG avec les pais confinés de la Principauté de COBVRG, le tout subdivisé en ses Baillages et dressé selon les memoires les plus authentiques
A Nuremberg chez les Heritiers de Homann A. 1743.

ut subdivisé en ses Ba

Homan. A. 1743.

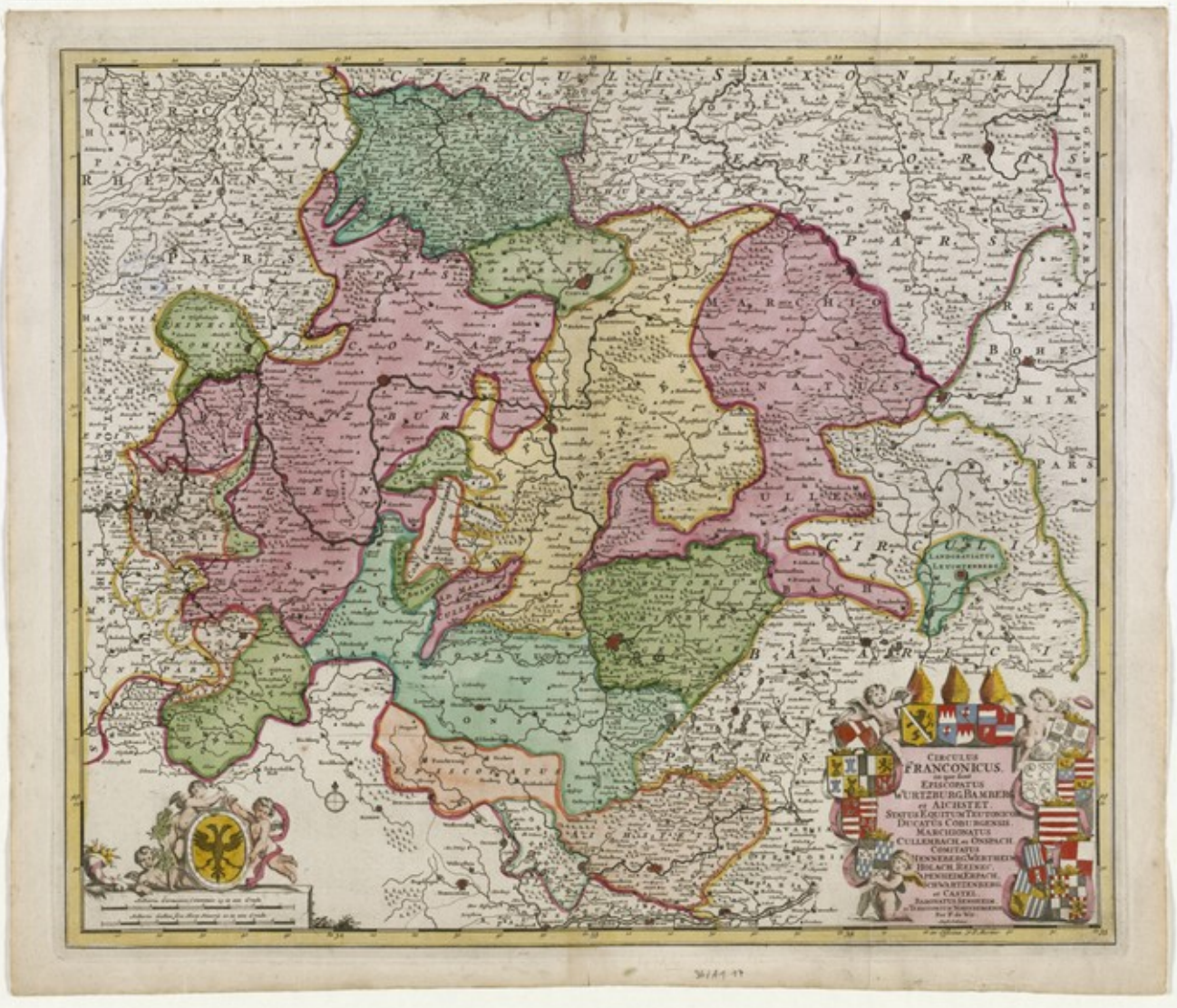

Selig
Leutendorff
S. R. I. COMITATVS
HENNEBERG
secundum Præfectu-
ras & modernas Dy-
nastias, una cum con-
fini PR: COBVRGENSI geo-
graphice consignatus
& in hac Tabula editus.

hausen
Ebertshausen
Albrechts
Hofnung
CH
Dolmarberg
VR
S
AE
Amt Suhla
C.
Suhla
Schwartza
Heinrichs
Rotwinderhof
Metzels
W. Ringles
Walbach
Soltz
Walldorf
Breuberg
N
E
B
Amt
E
Mebendorf
Spital
Brücken
Utendorf
Kündorff
Melckers
Welckershausen
Kündorff
Dietzhausen
Landsberg
Helba
Wichtshausen
W. Debertshausen
Meinungen
Altendambach
Langenbach
Eichenberg
Breitenbach
Bischofrod
Keulrod
Lange Bahn
Hasel Fl.
Dillstatt
Amt
Rohr
Clost. Rohr
Smeheim
W. Schneebach
Grube
Ahlstätt
Fischbach
Haselbach
Ober Massfeld
Elingshausen
Marisfeld
Amt
Der Stillberg
Grimenthal
Oberstatt
Kemete
Einhausen
Lengsfeld
Amt
U. Masfeld
Eichel Berg
Massfeld
Capel
Bitthausen
Themar
Sultzfeld
Ritschenhausen
Tachbach
Sophien Lust
Werra
Osterburg
Belrieth
Vachdorf
Leutersdorf
Bauerbach
Cl. Vessra
Zollbrück
Themar
Wölffershausen
Henfsstett
Bauerbach


Significatio notarum
Fortalitia
Urbes præcipuæ.
Oppida
Municipia.
Pagi.
Villæ.
Arces.
Coenobia.
Ruinæ.
W. Wüstung.
Gold.
Silber.
Kupfer.
Stahlwerck.
Salzwerck.
Eisen Gruben.
Vitriol.
Glaß. und Schmelz hütte.
Alaun.
K Kobalt.
Gr. Granit. od. Egypt: Marmor.
L. f. Stein Kohlen.
Gewehr Fabrique
MB. Marm. Bruch.
U. Umbra.
Bergwercke.
Milwitz
Geutersdorf
Marckau
Schala
45
40


Altendam
bach
Friedtero
Stuten
haus
Fraue
Hirschbach
Erlau
Breitenbach
Leupoldshoff
Amt
Wilcoken
Ahlstatt
Fischbach
hoff
Silbach
Schleusi
Neuncr
S. Kil
ian
Steinbach
Schl
eusingen
Rasen
Ziegelhof
Hinternahe
Langobach
Schönau


Kohnbach
Wiedershausen
Leimbeck
Grundmühl
Amt
Herpff
Aumühle
Heringen
Abterode
Frauense
Wölfertshausen
Fotzerode
H E S
Thalhausen
see
Knethof
Lengers Hof
Amt
Greynberg
Humelhausen
Karnrode
Ob. Cella
Kirchhof
Tif
fenor
Philipsthal
Unt Cella
Kieselbach
Tiefen
Nipmühl
Röhrig
S I S C H
Vacha Hof
Mercker
Yocker
Unter
Maur
Amt
Dorndorff
Kayser
30


CIRCULUS
FRANCONICUS.
EPISCOPATUS
WURTZEBURG BAMBERG
et AICHSTET.
STATUS EQUITUM TEUTONICOR.
DUCATUS COBURGENSIS.
MARCHIONATUS
CULLENBACH et ONSPACH.
COMITATUS
HENNEBERG WERTHEIM
HOLACH. REINEC.
SPANHEIM ERPACH.
CASTELL.
CIRCULUS
SAXONIÆ
SUPERIOR
REGNI
BOHEMIÆ
PARS


Freinsdorf
Puttenhaids
Haslach
1
Schlusselfeldt
Ebrach Flu
Schluselay
Hirshaid
Pomers
felt
Tisch
Flu
Reich
Elsendorf
Adelsdorf
Neuses
Mulhausen
AD LIMPURG
M
Hallersdorf
Neuhaus
Hausen
Forchain
Hoch stett
Ober
Schainfelt
Kerspach
Clepach
Rotenbach
1
Marck
Pibrach
Hanburg
Paiers
dorf
Marckschain
felt
Morendorf
Lamach
Dachsbach
Erlang
Buck


Das Francken Landt
Francken
Spessart
Das gepirg
In der Püchen


Grünsfelt
Haldinsfeld
Rinderfelt
Geigersheim
Prübach
Külsay
Elfelstat
Wirtzpurg
Wirtzburg
Rotendorff
Selgenstat
Tzel
Düttelbrun
Remling
Holtzkirchen
Werthaim
Hoenberg
Hochheim
Leinach
Tzelling
Reybach
Pirckenfelt
Trisenstain
Karbach

Study historical maps: why?

- Many libraries have large collections of historical maps
- Relevant for the (digital) humanities
  - History of cartography
  - General history
  - Specific example: onomastics

What happens with historical maps?

- Stored in a library basement
  - Retrievable by bibliographic information
- High-quality bitmap scans, online catalogue
  - Browsable by bibliographic information
- Useful queries?
  - In actual research practice
  - By interested laypeople

Metadata: what?

- Contained settlements
- Landscape topography
- Geopolitical features
- ...

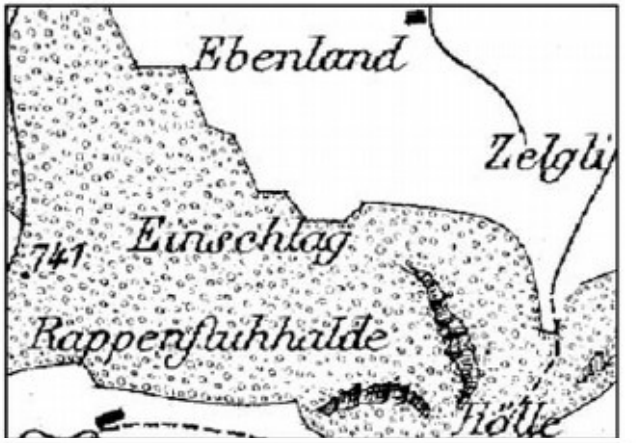
- Ideally: georeferenced

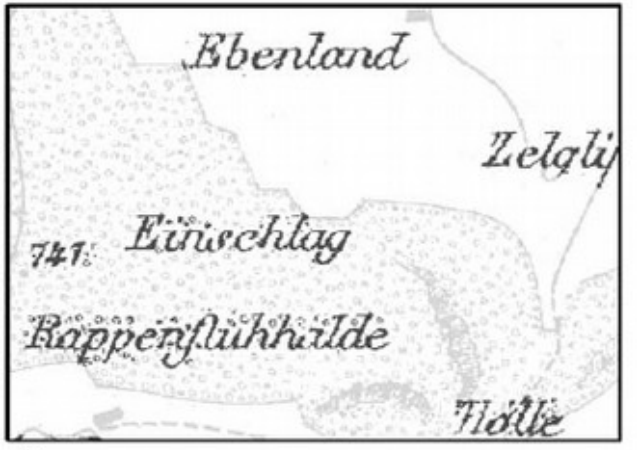
Metadata: how?

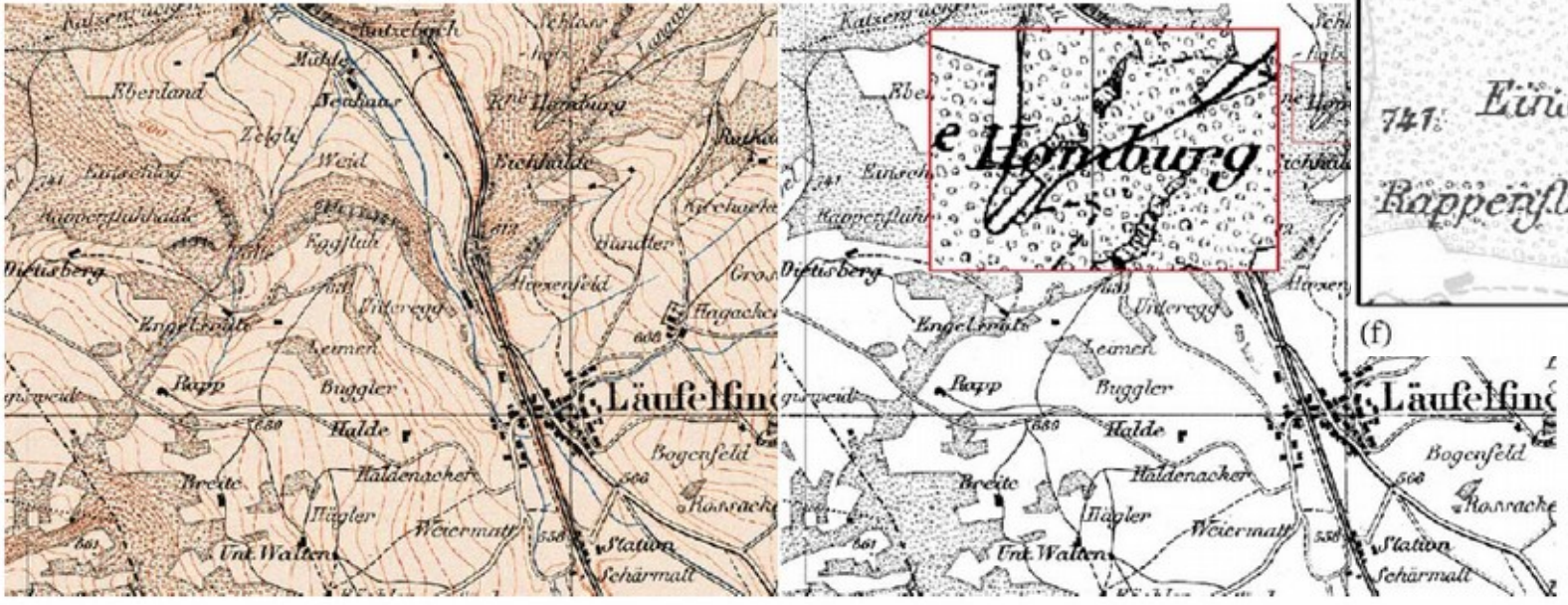
- Do it by hand
- Software: usability improvements
  - Gains in efficiency are limited
- Software: computer vision
  - No panacea, but can work well for restricted corpora
  - Significant custom R&D effort every time

For example...

- Forest-cover analysis of the "Siegfried Map"
  [Leyk, Boesch, Weibel]
- 6000 sheets, produced 1870 to 1922







(f)

Our scope

- We consider maps from early modern period forward
- Unique graphical styles
- Different fonts, handwriting
- Different cartographic conventions

So what now?

- Split problem into smaller goals
- Design a modular pipeline

Segmentation

Clustering and Matching

Optical Character Recognition

Georeferencing

Segmentation

Clustering
Matchin

Segmentation

- Smaller goals
- Look for one particular element on one map

  [Budig and van Dijk 2015]


Leibroc
Bamberg
Eich
Ludwach
Güspach
Gradelhoven
Scheslitz
Katelsdorff
Watterdorff
Tzapfendorf
Ebelffelt

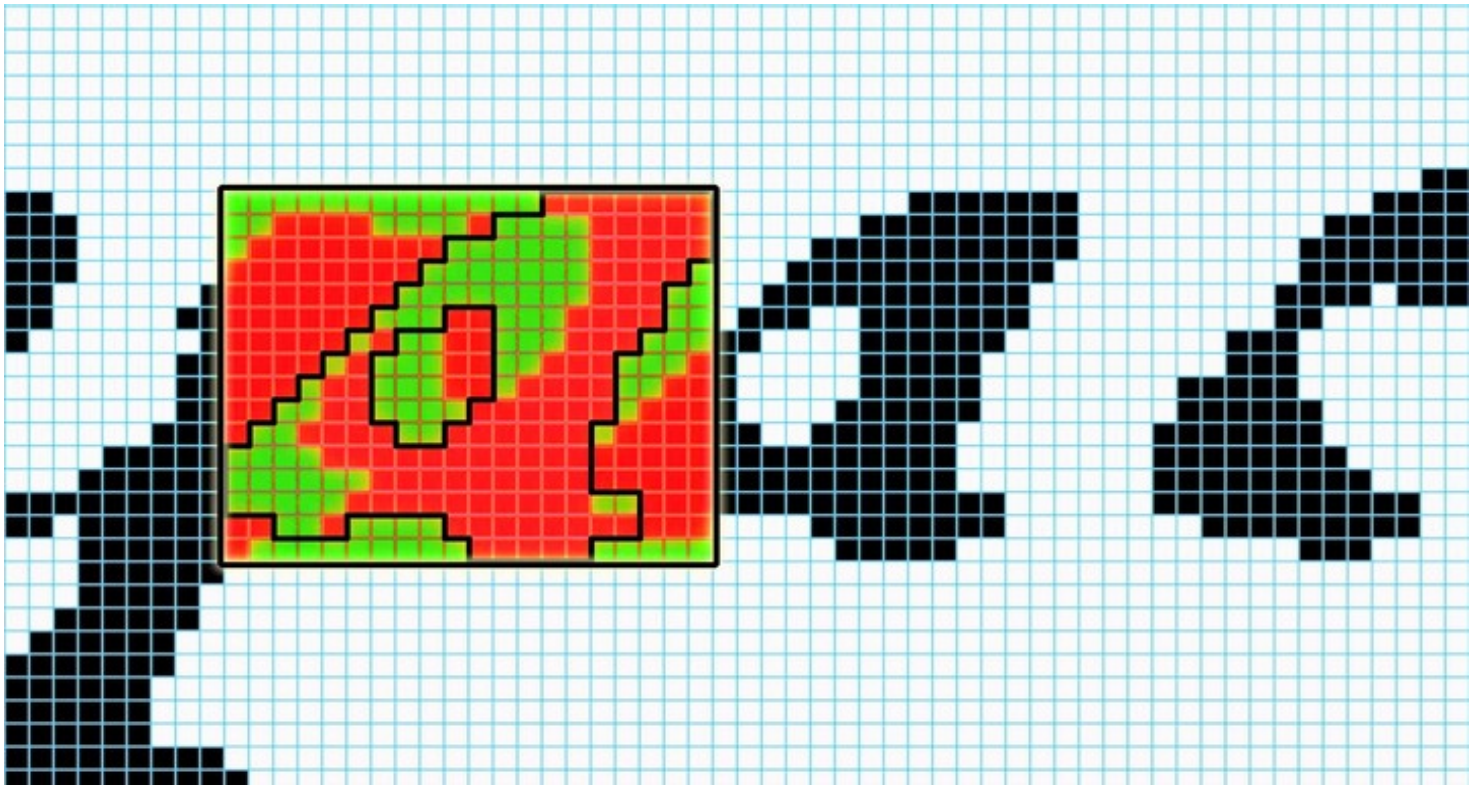
[Demo]

Ingredient 1: Template Matching

Template matching

- Find approximate repeat-occurrences of an example image
- Here:
  - 1-Bit black-and-white
  - Only considers translation
  - Percentage-correct score

Matching score

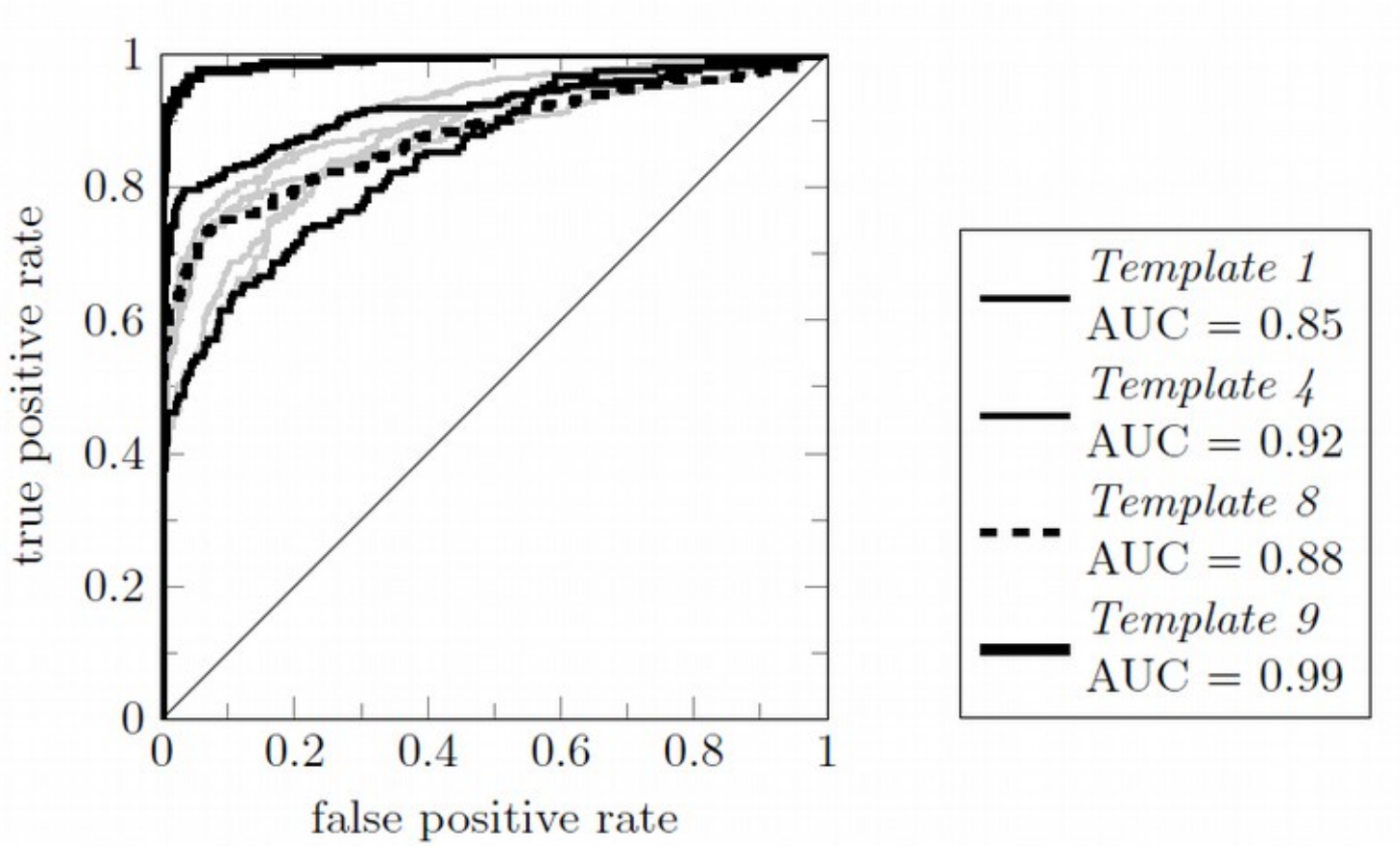


Ingredient 2: Active Learning

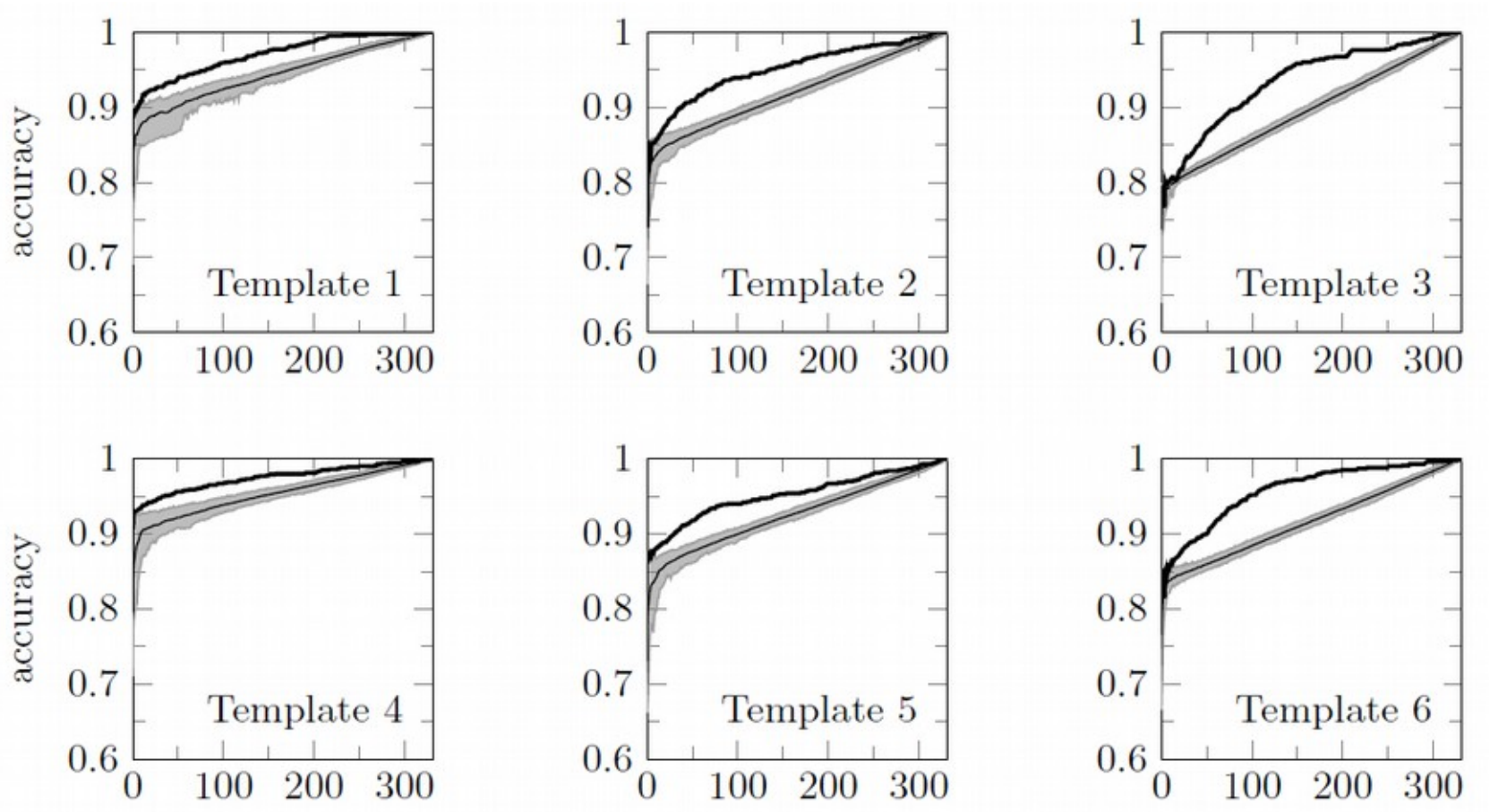
Data sets

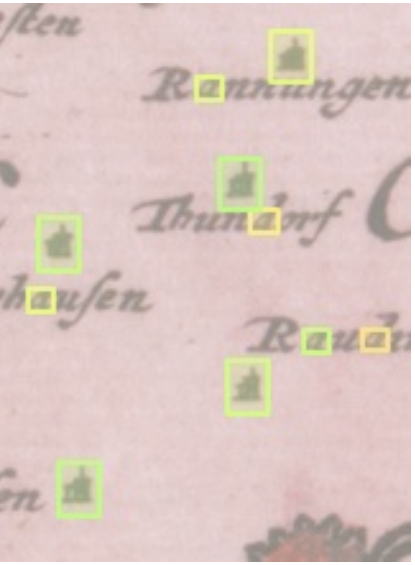
	Historical Map	Template	Accepted
1	*Carte Topo. D'Allemagne* (1787)		b, h
2	*Franciae Orientalis* (1570)		a, g, d
3	*Franciae Orientalis* (1570)		e
4	*Circulus Franconicus*, De Wit (1706)		a, g, d
5	*Das Franckenlandt* (1533)		a, g
6	*SRI Comitatus Henneberg* (1743)		n, m, h
7	*SRI Comitatus Henneberg* (1743)		e
8	*Circulus Franconicus*, De Wit (1706)		
9	*Circulus Franconicus*, Seutter (1731)		

ROC

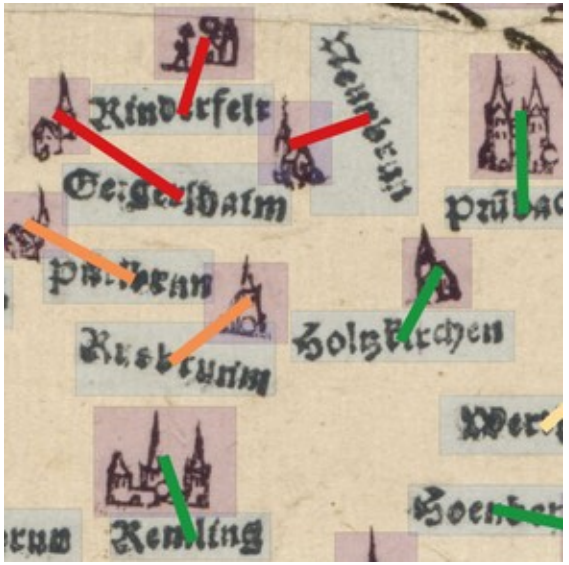

true positive rate
false positive rate
0
0.2
0.4
0.6
0.8
1
Template 1
AUC = 0.85
Template 4
AUC = 0.92
Template 8
AUC = 0.88
Template 9
AUC = 0.99

Learning curves





gmentation

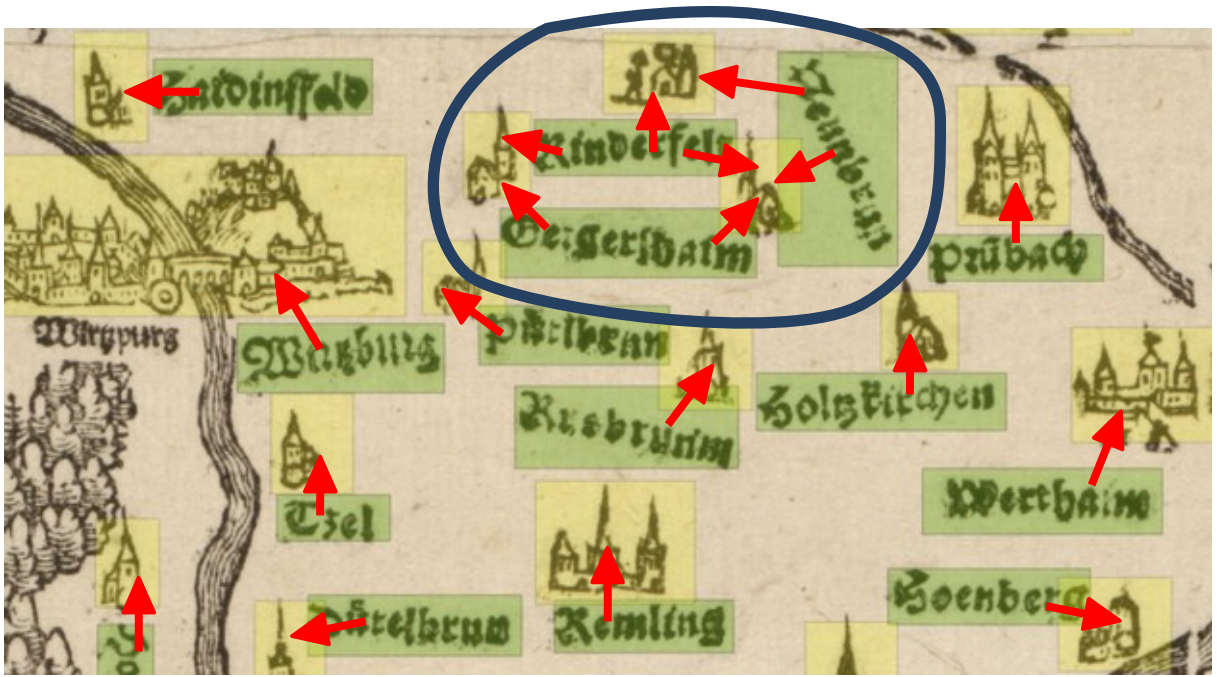


Clustering and Matching

Optical Char Recogniti

Matching Labels and Place Markers

- Assumption: labels and markers already detected
- Match the corresponding ones *[Budig, van Dijk, Wolff, 2014]*



Wanted: a Matching

- Find a *matching* of labels and place markers
- No 1-to-1 assignment possible
- Basic assumption: labels are *near* their corresponding markers
- Greedy strategy?

  → does not work well!

Experimental Results

- Franckenlandt (1533)
  - 539 markers, 524 labels
  - our algorithm: error rate 3.5%
  - greedy algorithm: error rate 17.8%

- Circulus Franconicus (1706)
  - 1663 markers, 1669 labels
  - our algorithm: error rate 1.3%
  - greedy algorithm: error rate 5.9%

What now?

- Error rates in experiments: 1.3% and 3.5%
- Unclear situations:

- Manual verification or correction necessary

Sensitivity

- Only show assignments our algorithms is *uncertain* about
- Calculate sensitivity analysis for the matching

→ For each assignment,

calculate the best matching that does *not* use it

and see how much worse it is.

Conclusion

- Historical maps are cool, but hard to search
- Modular pipeline is reasonable
- Human effort is necessary → smart interactions!

- Template matching & active learning work well
- Sensitivity analysis for efficient interactions

Open Questions & Future Work

- Solve more small goals from the pipeline, then integrate
  - Cluster template matches (e.g. into labels)
  - Use already collected information for OCR
  - Georeferencing, ...
- Should the pipeline really be sequential?
- Crowdsourcing?

Segmentation

Clustering and Matching

Optical Character Recognition

Georeferencing

Smartphone

Open Questions & Future Work

- Develop remaining modules in extraction pipeline
  - Cluster template matches (e.g. into labels)
  - Use already collected information for OCR
  - Georeferencing, ...
- Should the pipeline really be sequential?
- Crowdsourcing! Yes, but how exactly?
- What other algorithmically-guided user interactions?