Universität Rostock
Traditio et Innovatio

CITlab
Computational Intelligence Technology

READ

Text2Image matching

How to use existing images and transcripts to produce Automated Text Recognition

Tobias Hodel
Gundram Leifert

State Archives of Zurich
University of Rostock

Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions

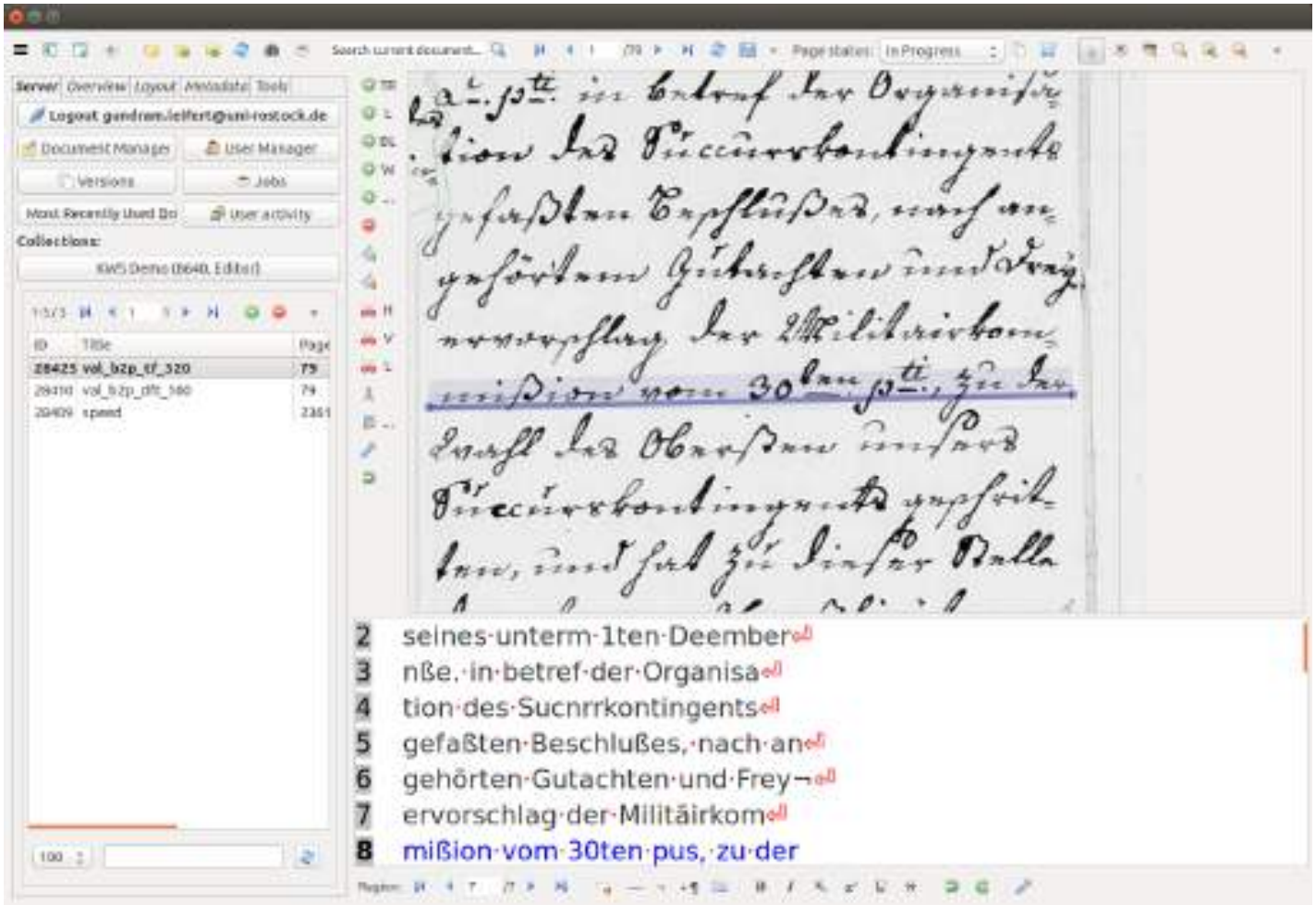
Another perspective to train an ATR

Workflow

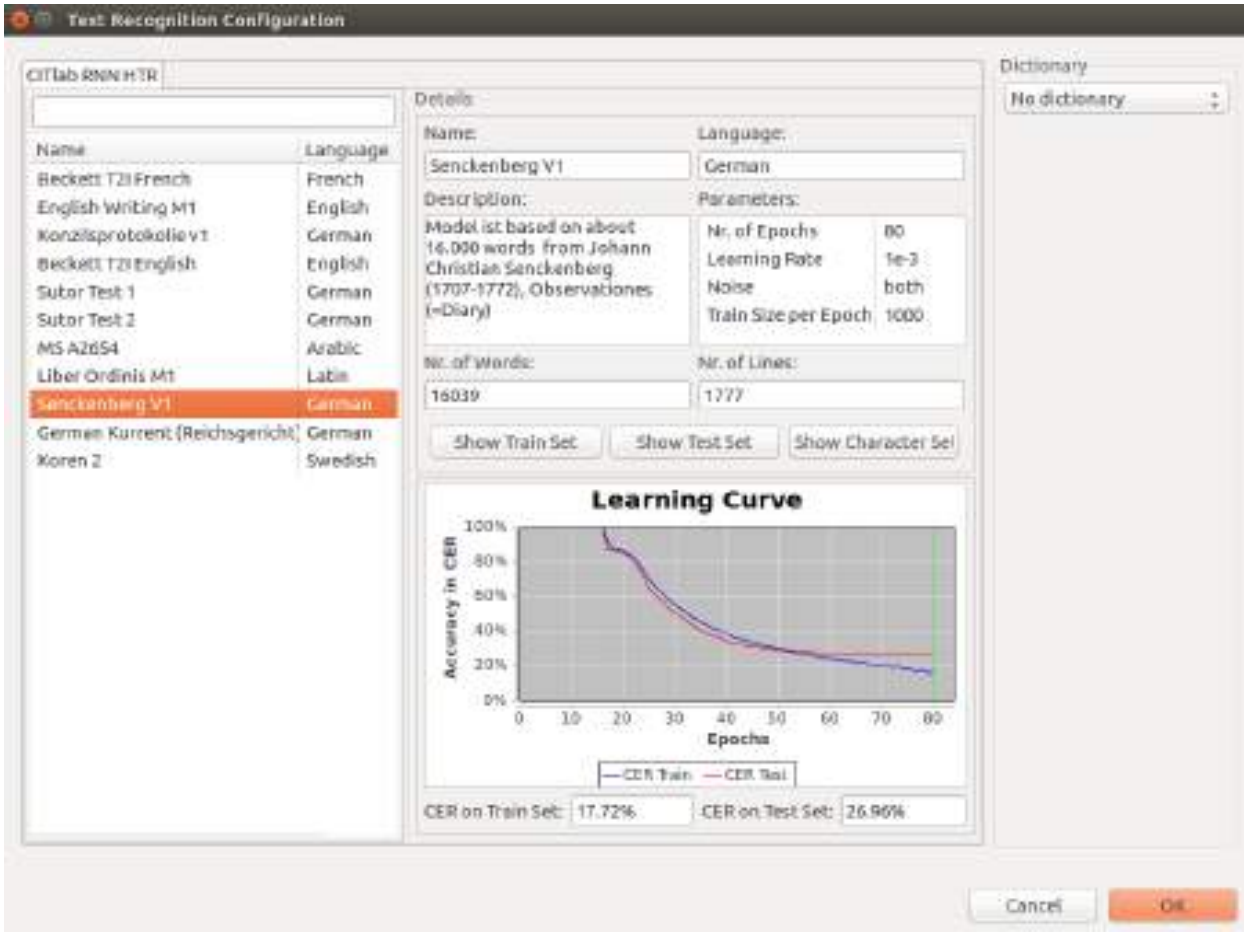
Text2Image on StAZH images

More demos and impressions

Another perspective to train an ATR



Another perspective to train an ATR



Another perspective to train an ATR

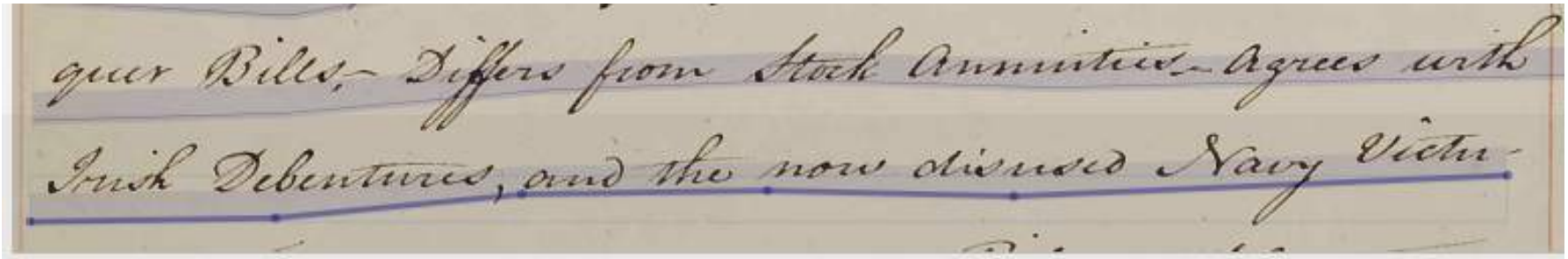
We want to have a good Automated Text Recognition (ATR)!

Another perspective to train an ATR

We want to have a good Automated Text Recognition (ATR)!
We know:

- the more training data, the better the ATR
- classical training data production is expensive

ground truth (GT) – training sample



quer Bills,_ Differs from Stock Anuunties_-Agrees with↵
Irrish Debentures, and the now disused Navy Victu-↵

Another perspective to train an ATR

We want to have a good Automated Text Recognition (ATR)!
We know:

- the more training data, the better the ATR
- classical training data production is expensive

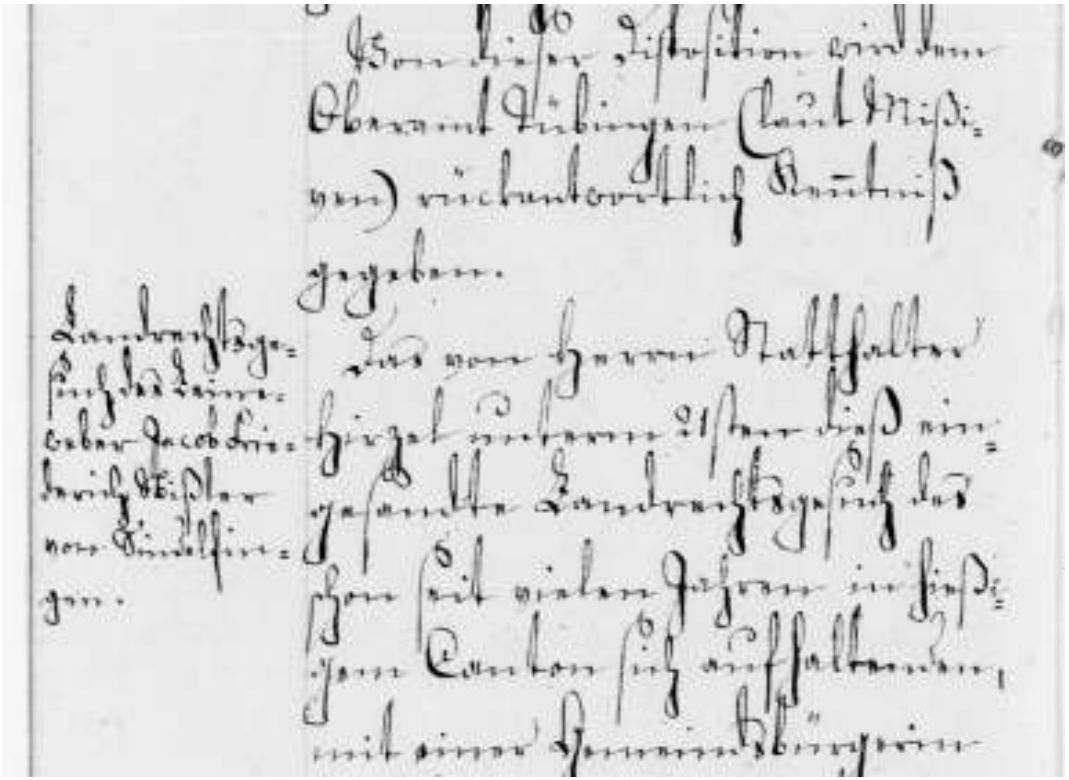
Another option

we know: for many documents the transcripts are already available

but: the transcripts are not alligned to the text lines of the images

goal: create training data from these images and transcripts for classical training

Regierungsratsprotokolle StAZH (large scale demonstrator)



Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen. Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(a) Image

(b) Transcripts

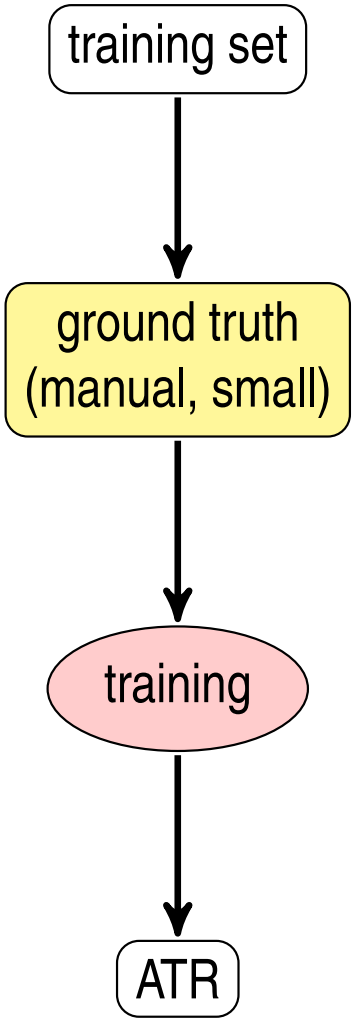
Another perspective to train an ATR

Workflow

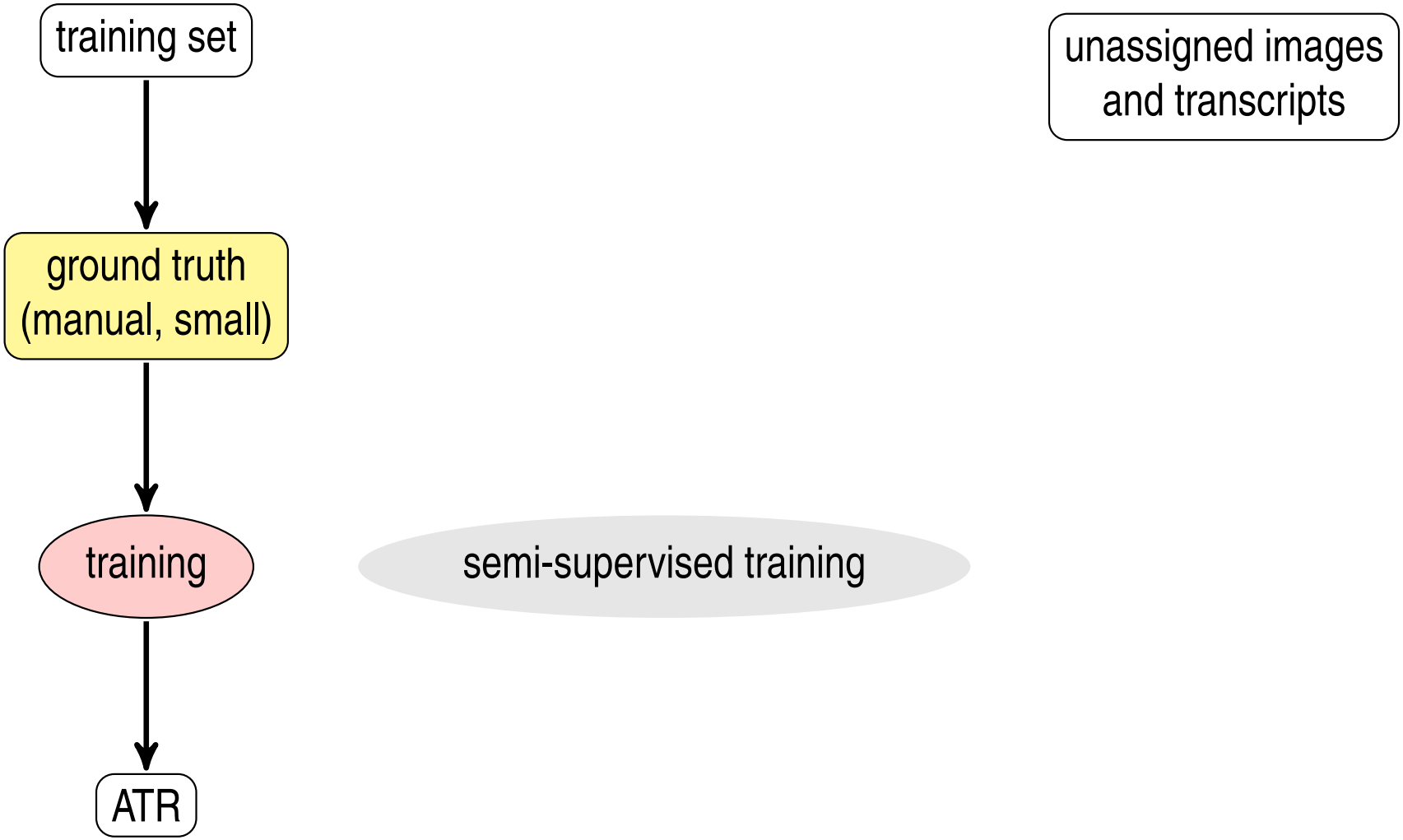
Text2Image on StAZH images

More demos and impressions

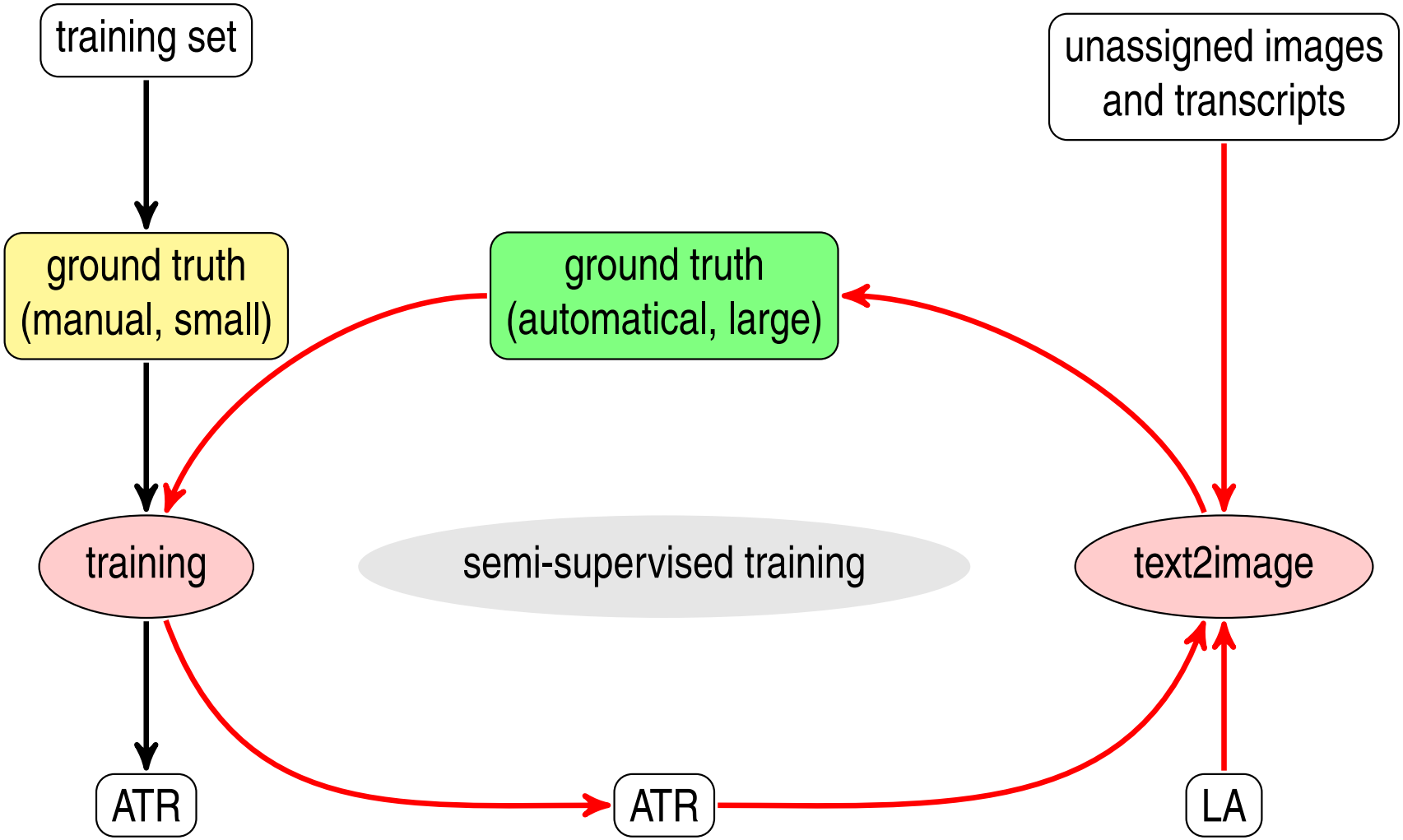
classical training workflow



semi-supervised training workflow

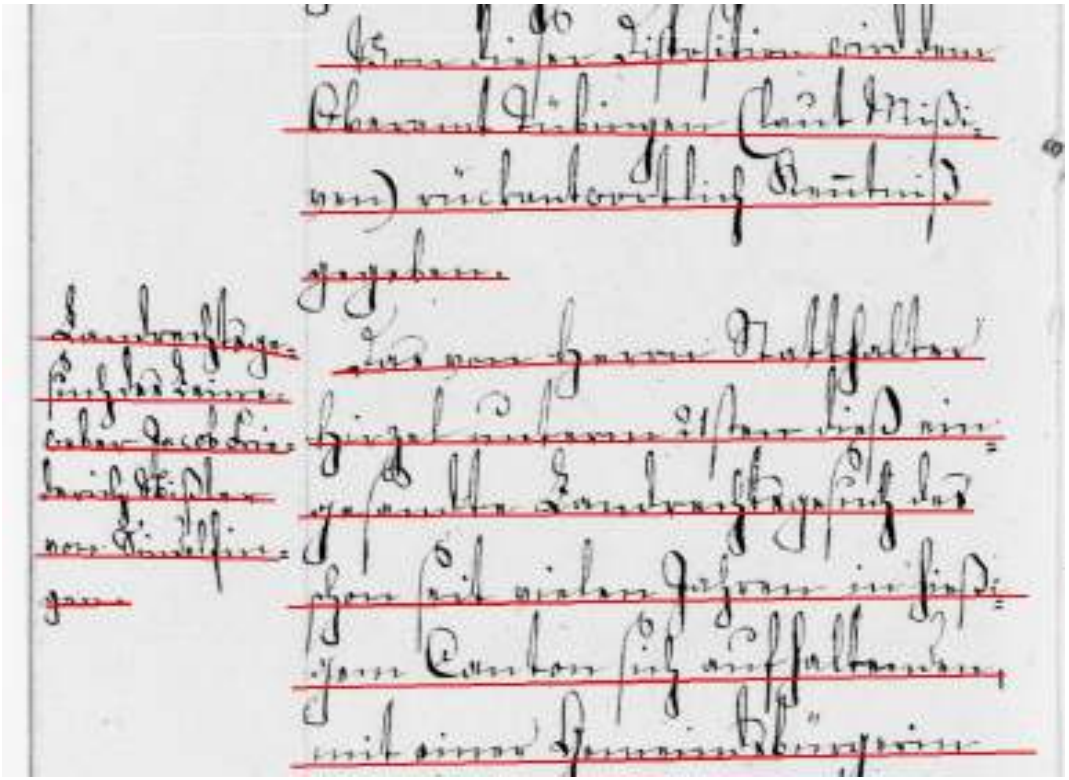


semi-supervised training workflow



workflow for Text2Image

transcript with corresponding image with baselines



Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen. Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(a) Image

(b) Transcripts

Another perspective to train an ATR

Workflow

Text2Image on StAZH images

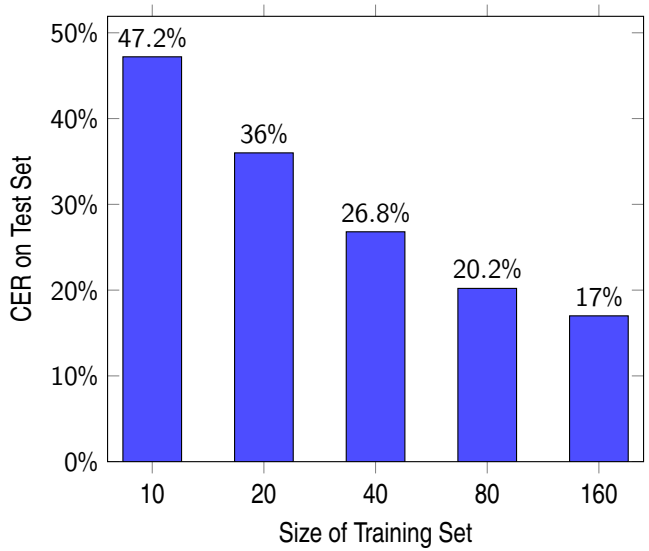
More demos and impressions

Text2Image on StAZH images

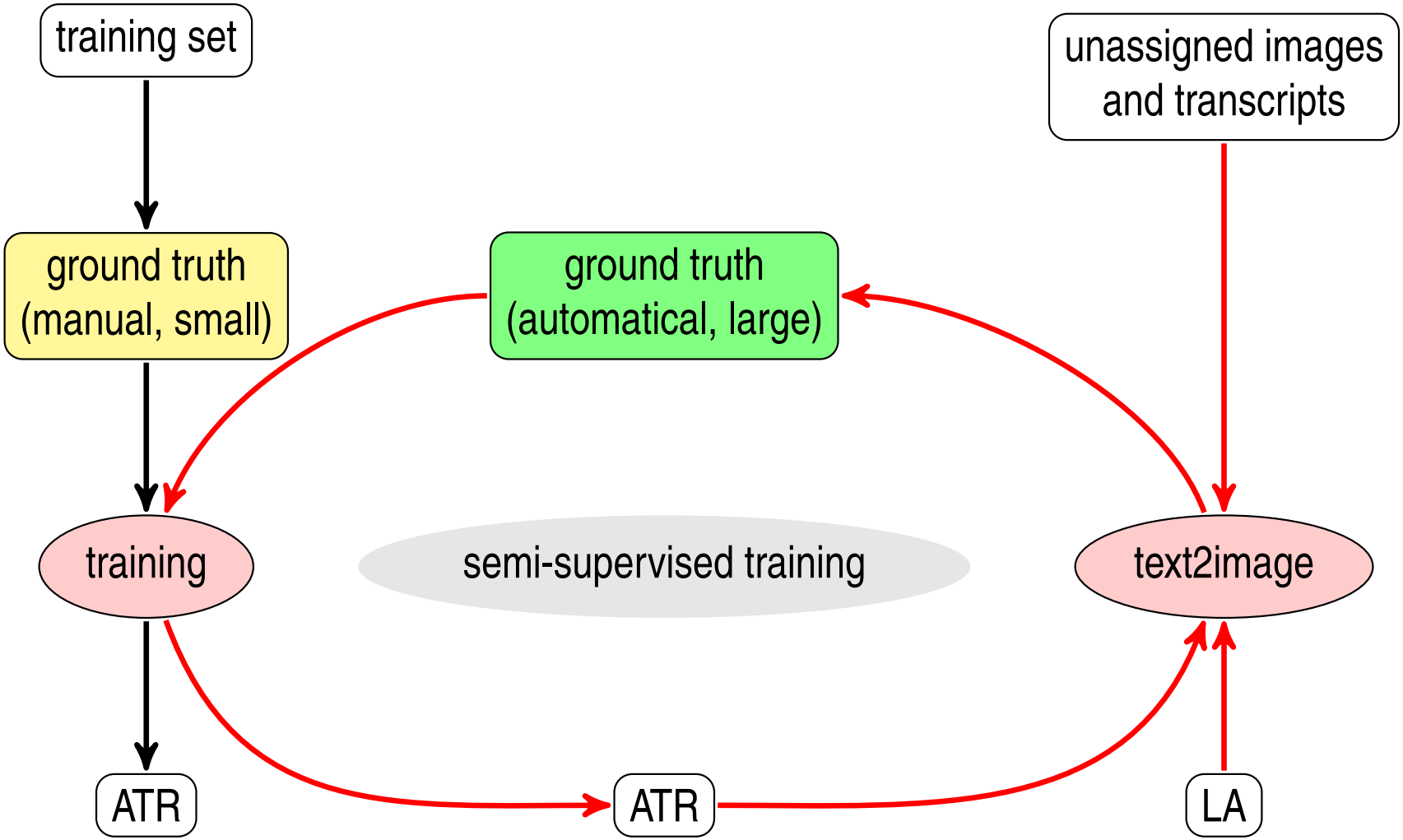
StAZH test collection (part)

Size of collection	pages
training set	160
test set	40

Character Error Rate on test set

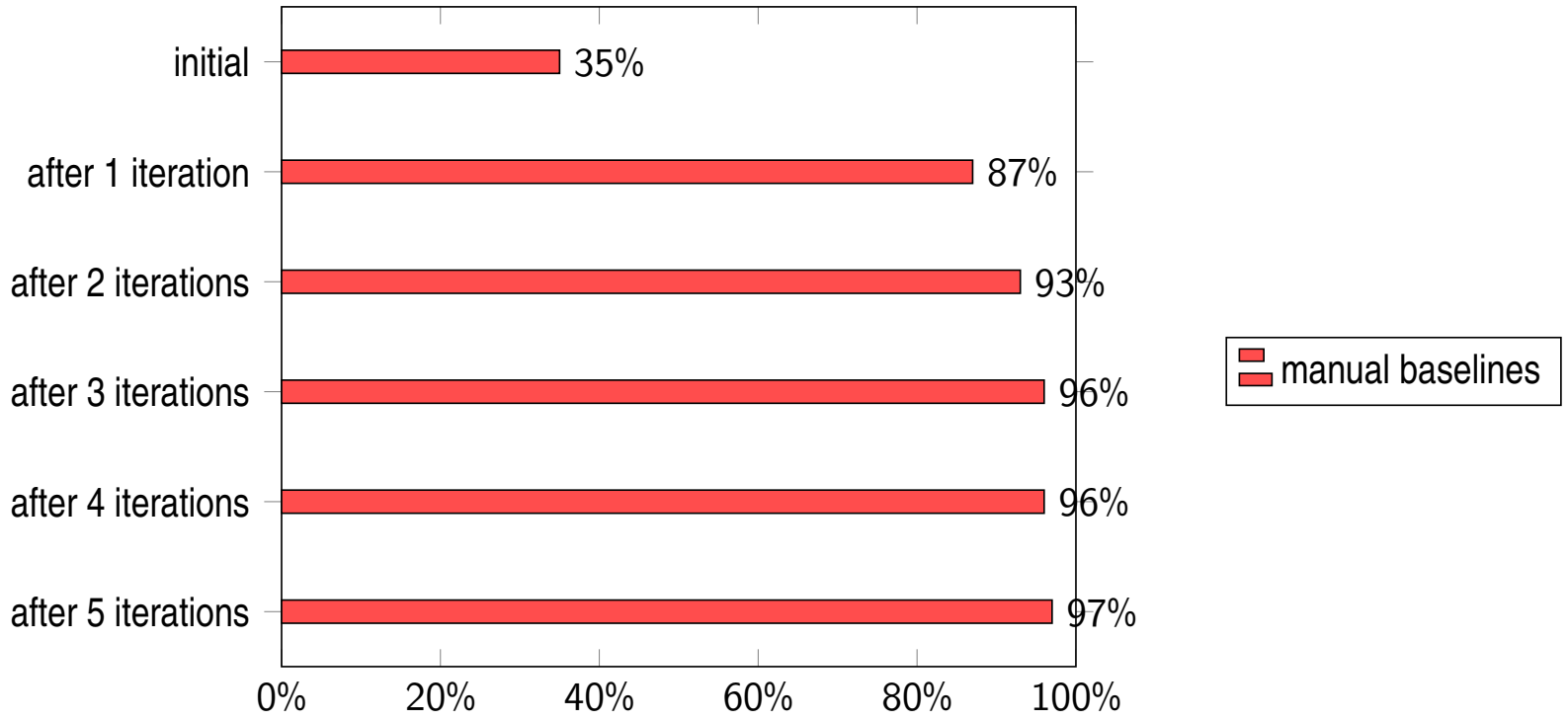


semi-supervised training workflow



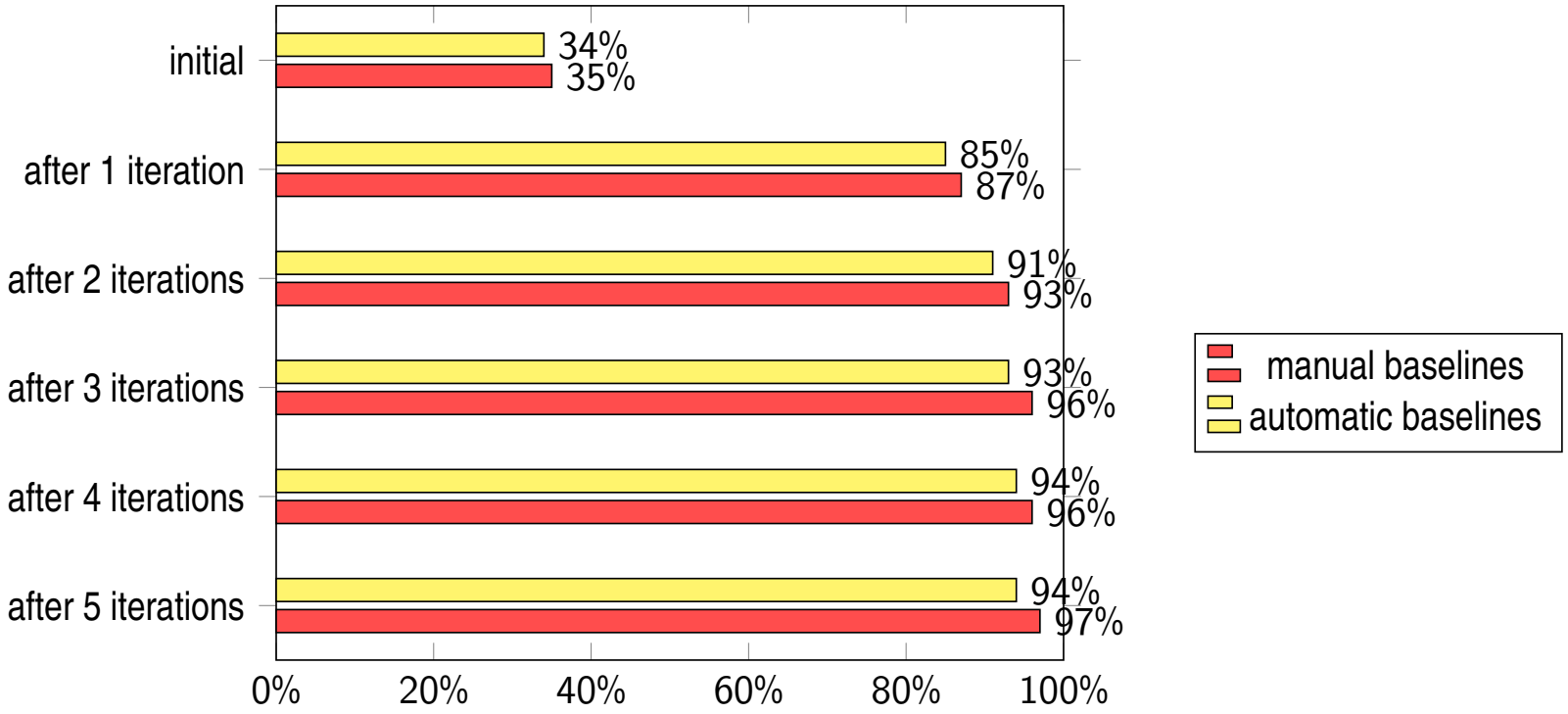
Text2Image on StAZH images

Percentage of alligned Groundtruth



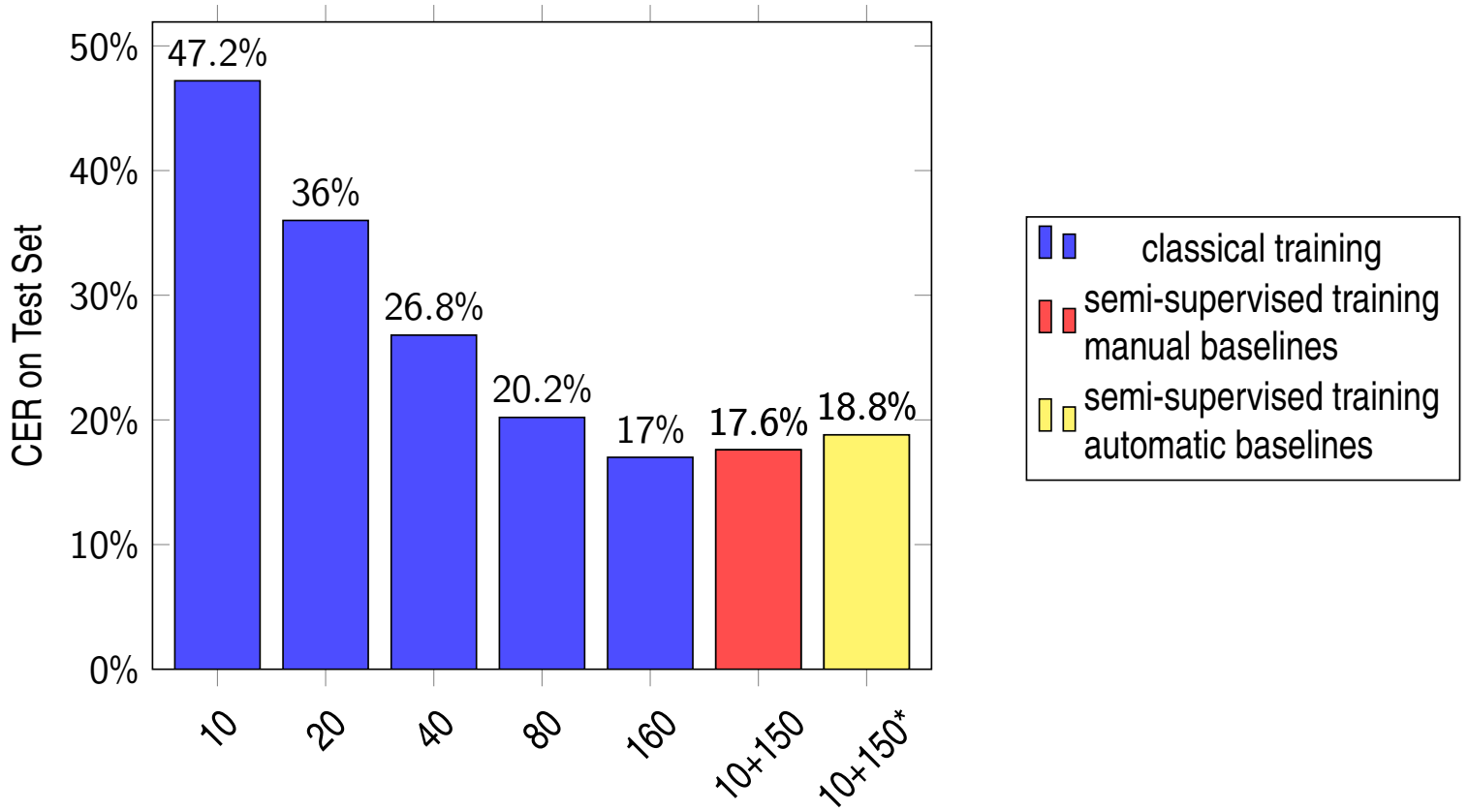
Text2Image on StAZH images

Percentage of alligned Groundtruth



Text2Image on StAZH images

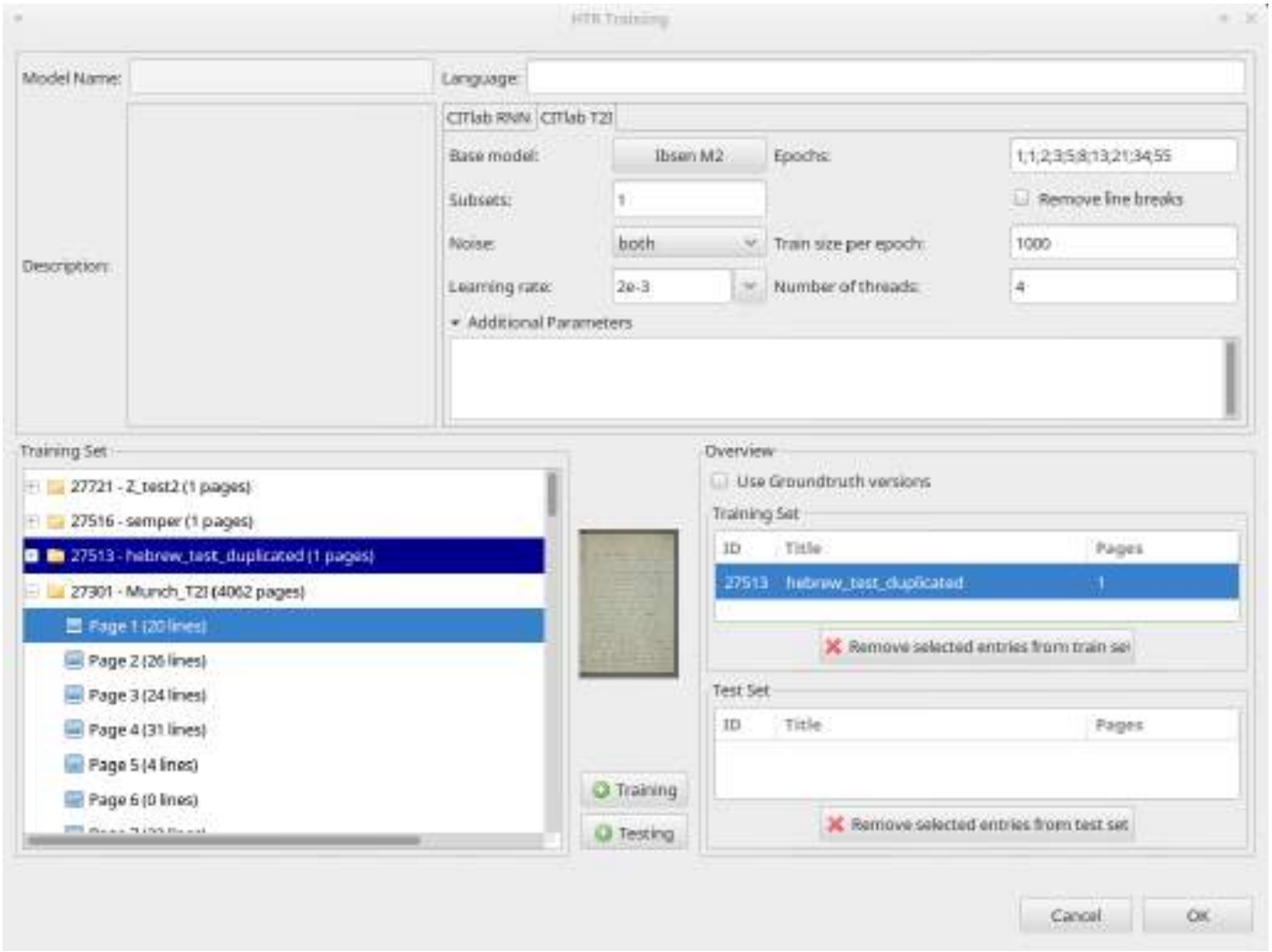
Character Error Rate on Testset



What can we handle?

- unknown/new characters in transcripts
- missing line breaks in transcripts
- missing hyphenations in transcripts
- missing baselines and transcripts
- wrong baselines
- wrong order between baselines and transcripts
- right-to-left languages
- **hard:** abbreviations in transcripts

Text2Image in Transkribus



Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions

Kanton Zürich
Direktion der Justiz und des Innern
Staatsarchiv

Project READ

Text2Image

Demo

Direktion der Justiz
und des Innern

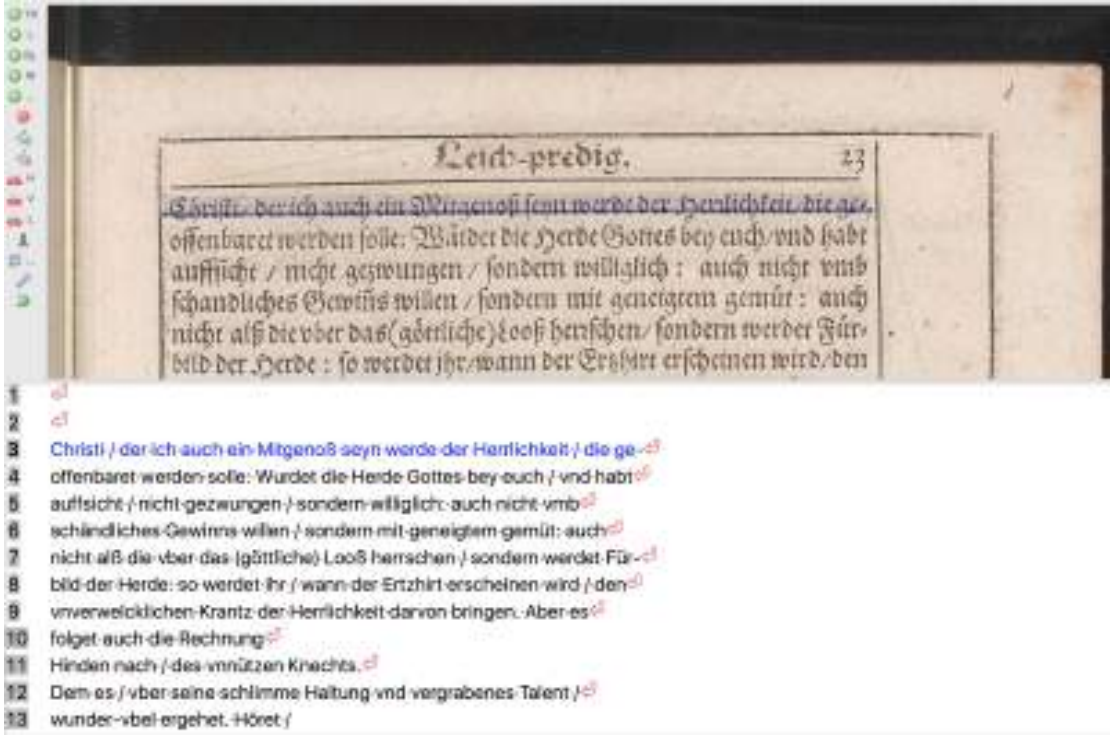
READ

Medieval Text

28
29
30 so habt sich an die blanken,
31
32 der mit steten gedanken
33 Diz fligende bispel,
34
35 ist tumben liuten gar zesnél,
36
37
38 sine mugens niht erdenken,

- Transcription
- Transkribus t2i with „generic model“ (about 20% CER, no GT)
- Identification of ~50% of lines
- 230 pages
- Trained HTR: 13% CER (on test set)

Printed Text (16th century)



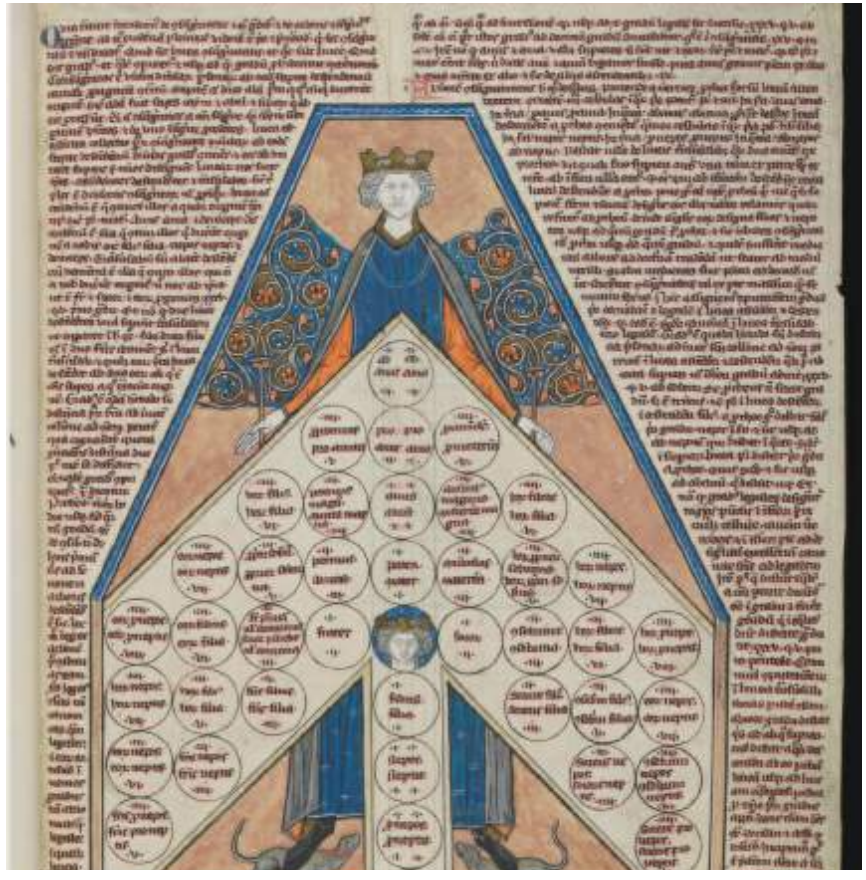
- Transcription
- Transkribus t2i with „fraktur model“ (about 20% CER, no GT)
- Identification of ~90% of lines
- 36 pages
- Trained HTR: 5% CER

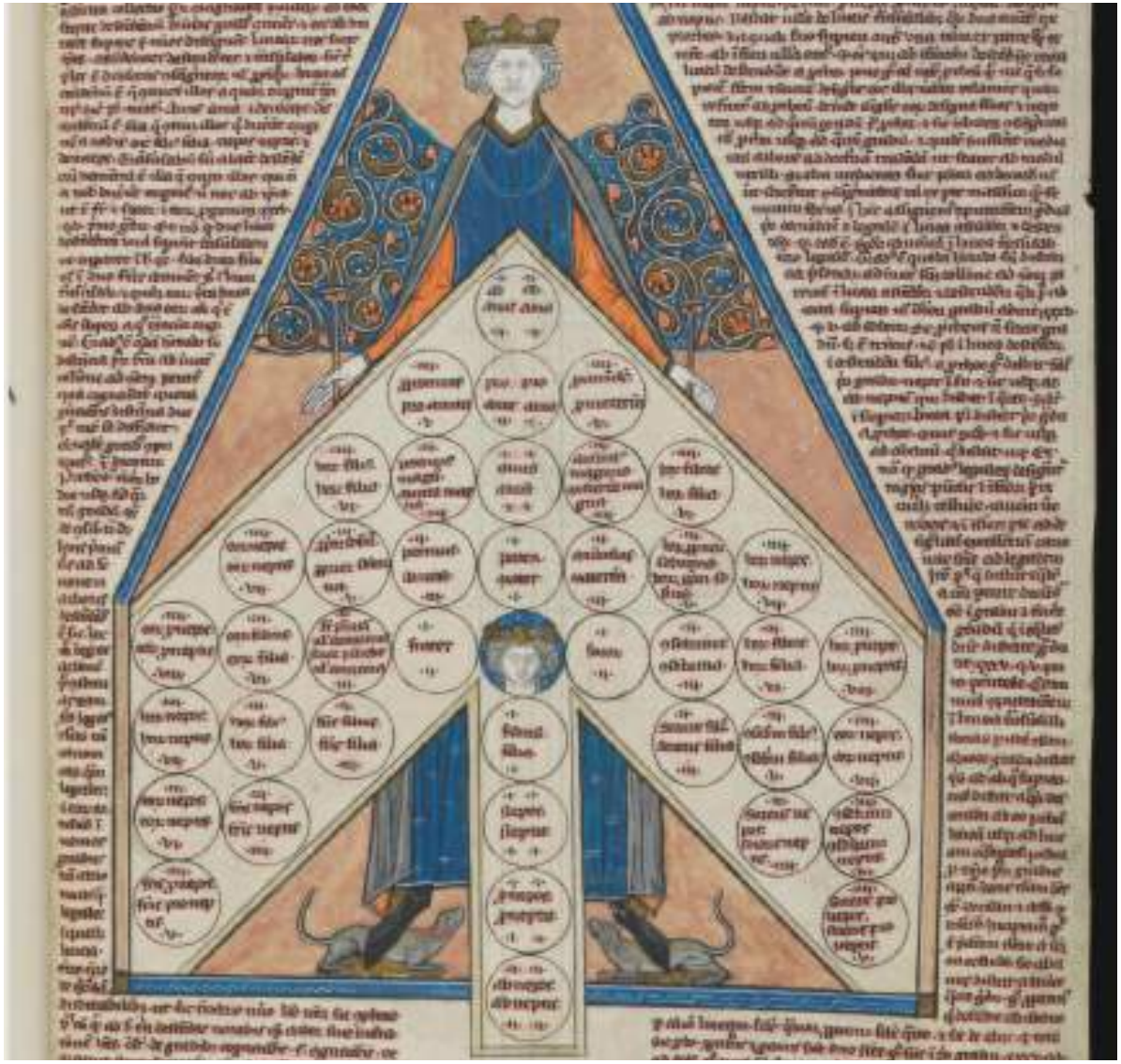
Direktion der Justiz
und des Innern

READ

Input

- 2 Images of pages
- 2 Text files (one txt file per pages; also TEI/Word-Doc)





Direktion der Justiz
und des Innern

READ

THANKS …

CITlab Group

CITlab
Computational Intelligence Technology

MoU Partner SME

PLANET artificial intelligence GmbH

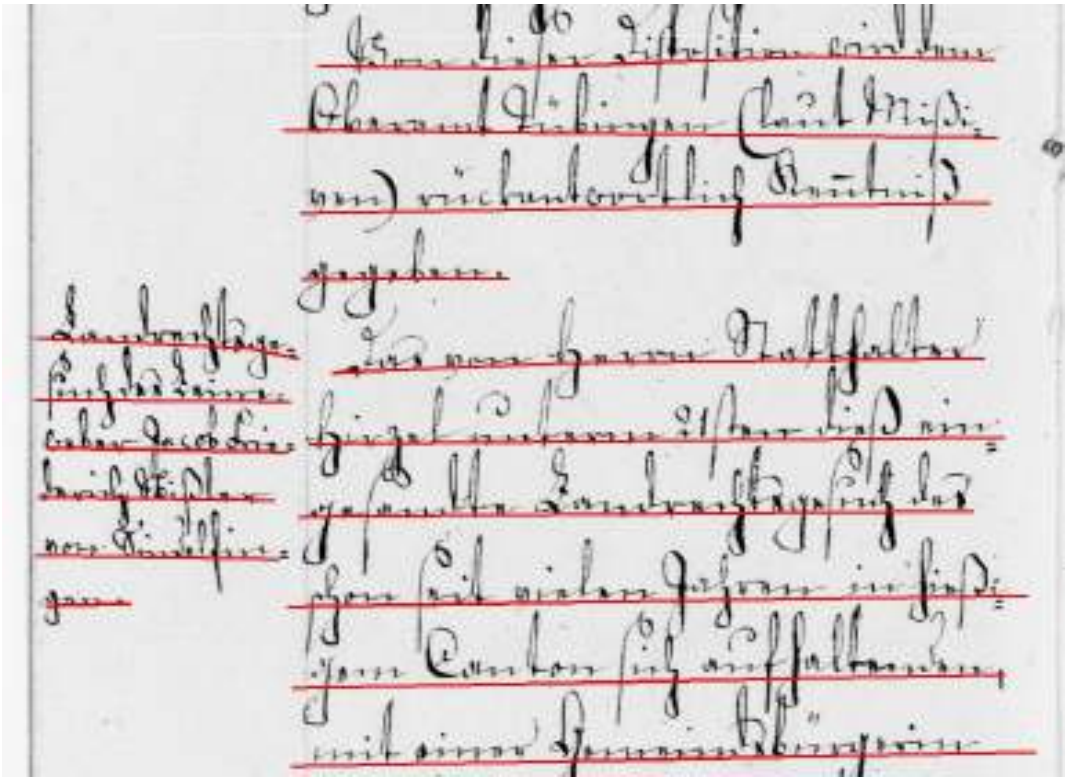
EU HORIZON2020 Grant

READ Recognition and Enrichment of Archival Documents

… for your kind attention!

workflow for Text2Image

transcript with linebreaks



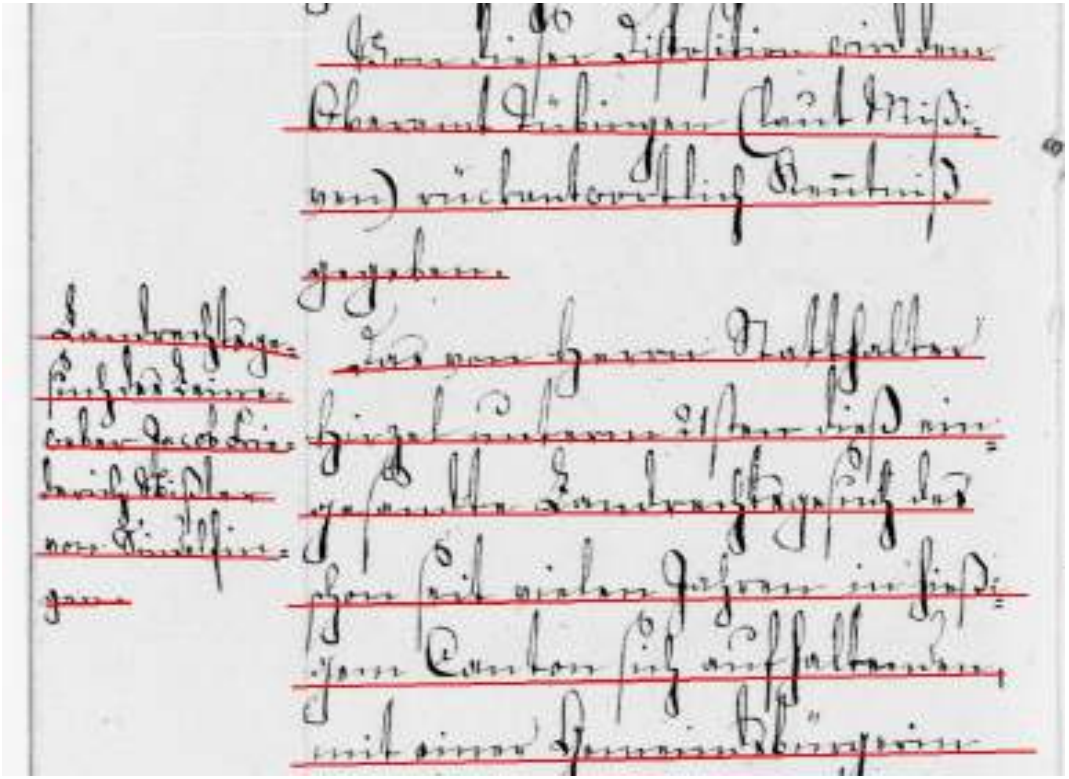
(a) Image

Von dieser Disposition wird dem
Oberamt Tübingen (laut Mißi¬
ven) rückantwortlich Kenntniß
gegeben.
Landrechtsge¬
such des Leine¬
weber Jacob Frie¬
derich Nißler
von Sindelfin¬
gen.
Das von Herrn Statthalter
Hirzel unterm 21sten dieß ein¬
gesandte Landrechtsgesuch des
schon seit vielen Jahren in hießi¬
gem Canton sich aufhaltenden,
mit einer Gemeindsbürgerin

(b) Transcripts

workflow for Text2Image

transcript without linebreaks

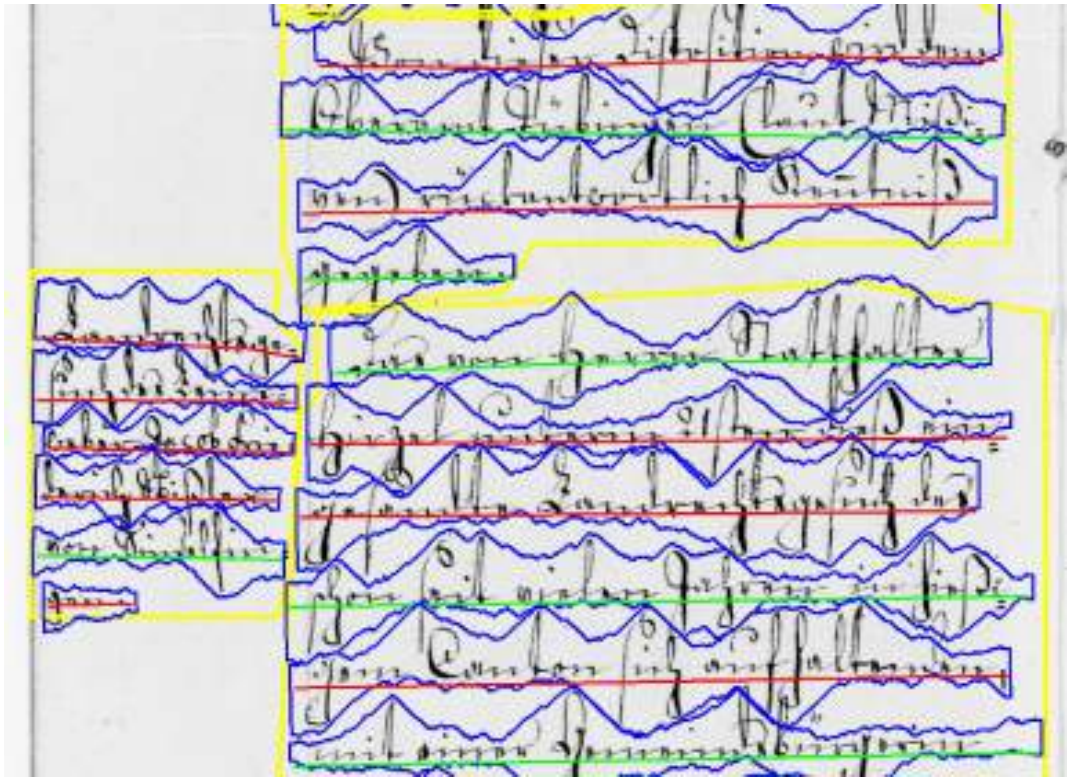


(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen. Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(b) Transcripts

Matching lines

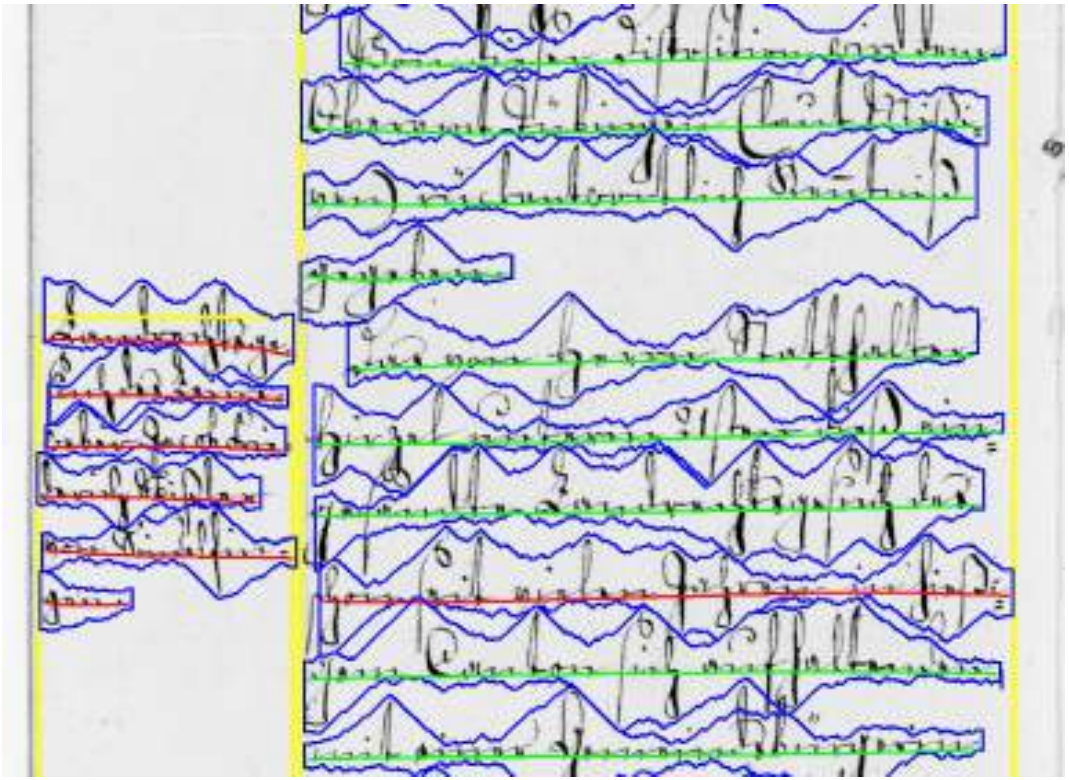


(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen. Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(b) Transcripts

Matching lines



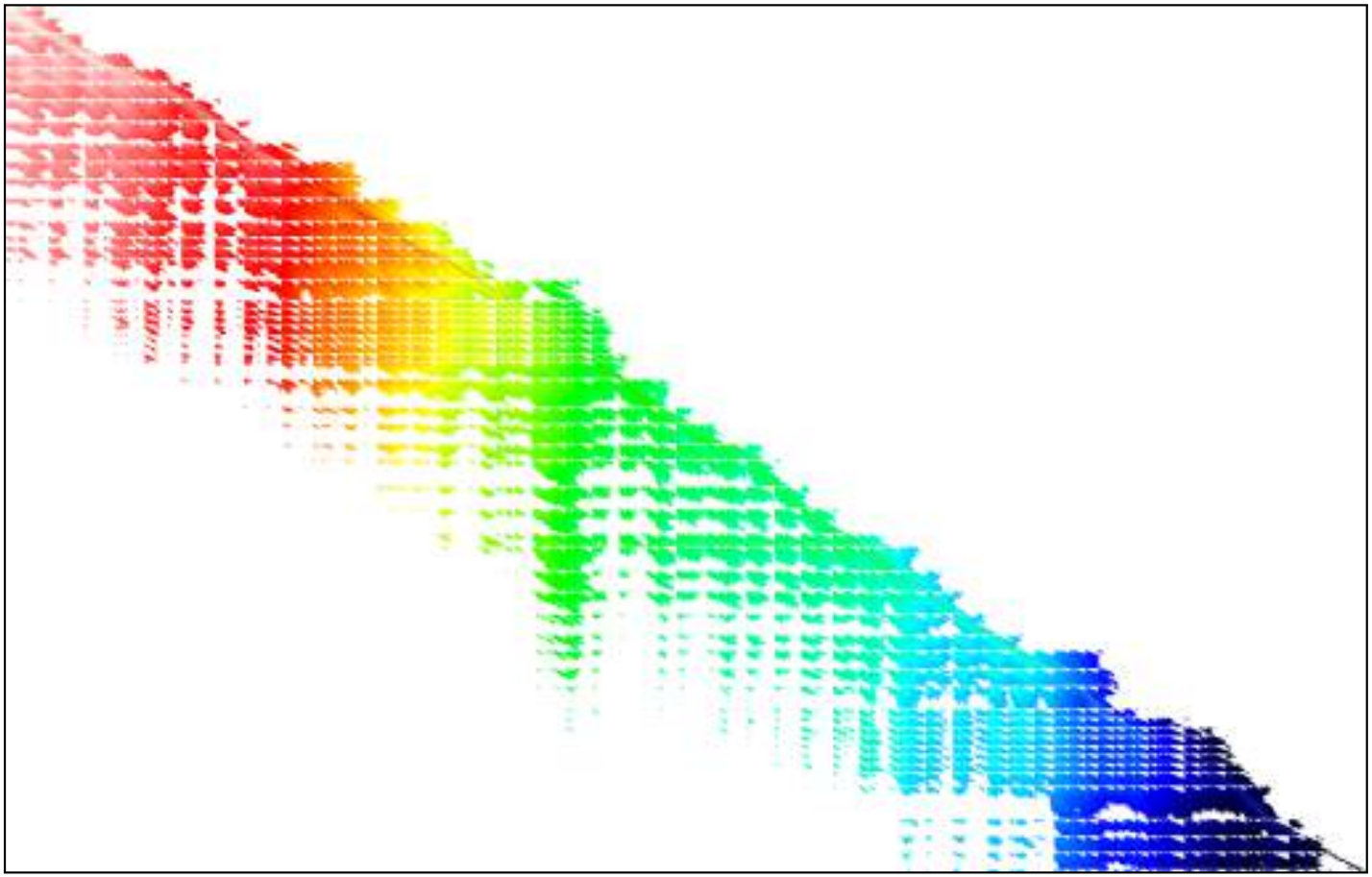
(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen. Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(b) Transcripts

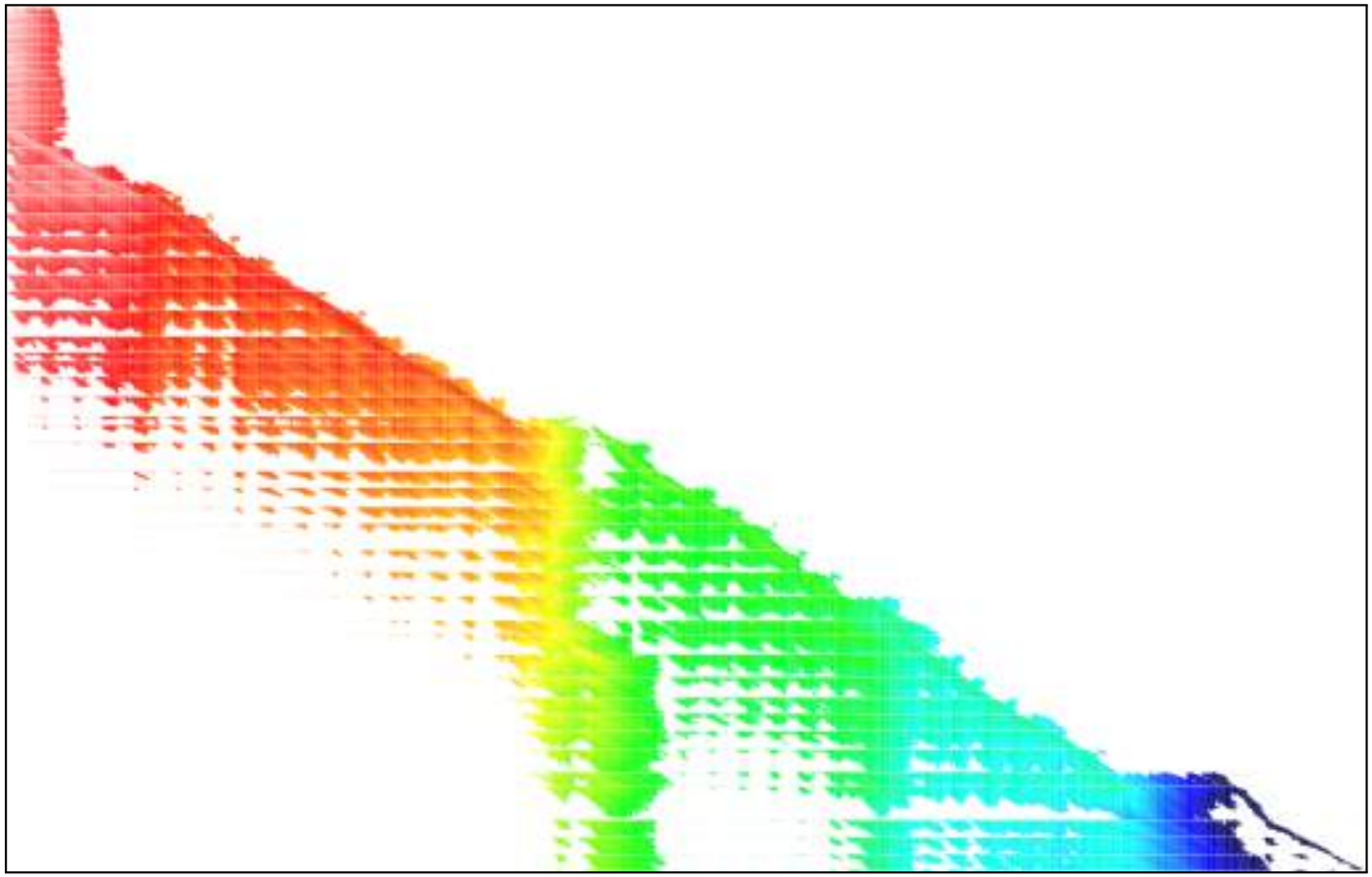
workflow for Text2Image

Some nice images from the assignment precedure...



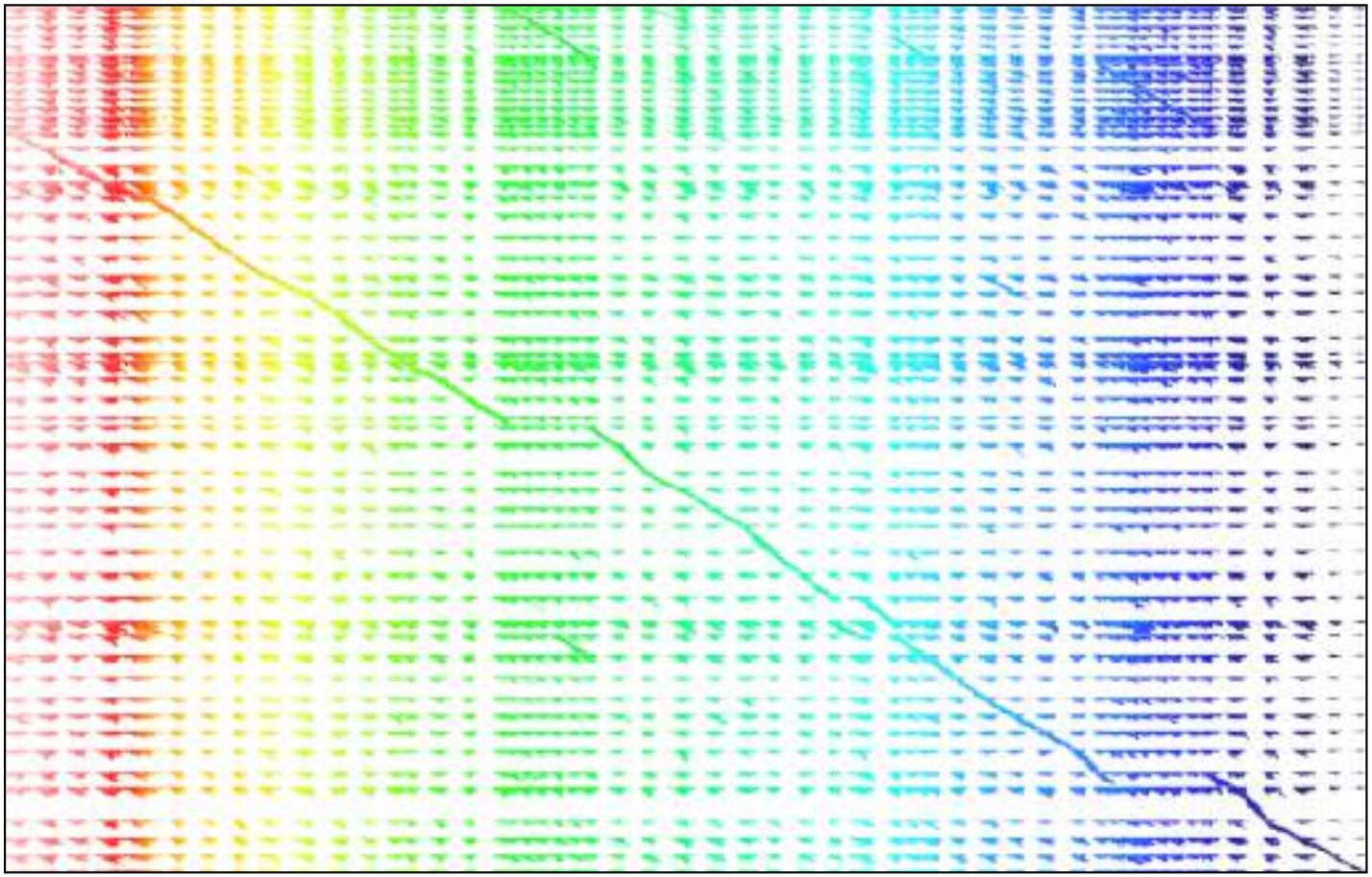
workflow for Text2Image

Some nice images from the assignment precedure...



workflow for Text2Image

Some nice images from the assignment precedure...



Text2Image matching

We want to have a good Automated Text Recognition (ATR)!

We know:

- the more training data, the better the ATR
- classical training data production is expensive

Another option

- for many documents the transcripts are already available
- so far, the transcripts are not alligned to the text lines of the images
- we can create training data from these images and transcripts to train an ATR