OPTICAL CHARACTER RECOGNITION

Motivating Example: OCR

Optical Character Recognition:

- image → text
- closely related to object recognition, autonomous driving, speech recognition, ...
- much better statistical understanding of...
  - class structure
  - noise
  - ground truth
  - syntactic structure / language modeling

Applications

- surprisingly, a hot topic!
- wanted: information extraction from unstructed documents
- NLP technology didn't use to be up to it
- with new LLMs, that has changed dramatically

Motivating Example: OCR

NEW MEMBERS Qualified

OBITUARIES

Philip G. Johnson

Lorraine J. D. Frazer

Daniel C. Green

Archie Frank Fanger

DO YOU USE WIRE CLOTH STRAINERS?

Michigan Wire Cloth Company can assist you in selecting the correct grade of wire cloth for any type of strainer, and can help in the development of proper strainer design.

Complete manufacturing facilities for drawing the wire - weaving the cloth - and fabricating the assembly all in one continuous production line.

Send Michigan Your Inquiries

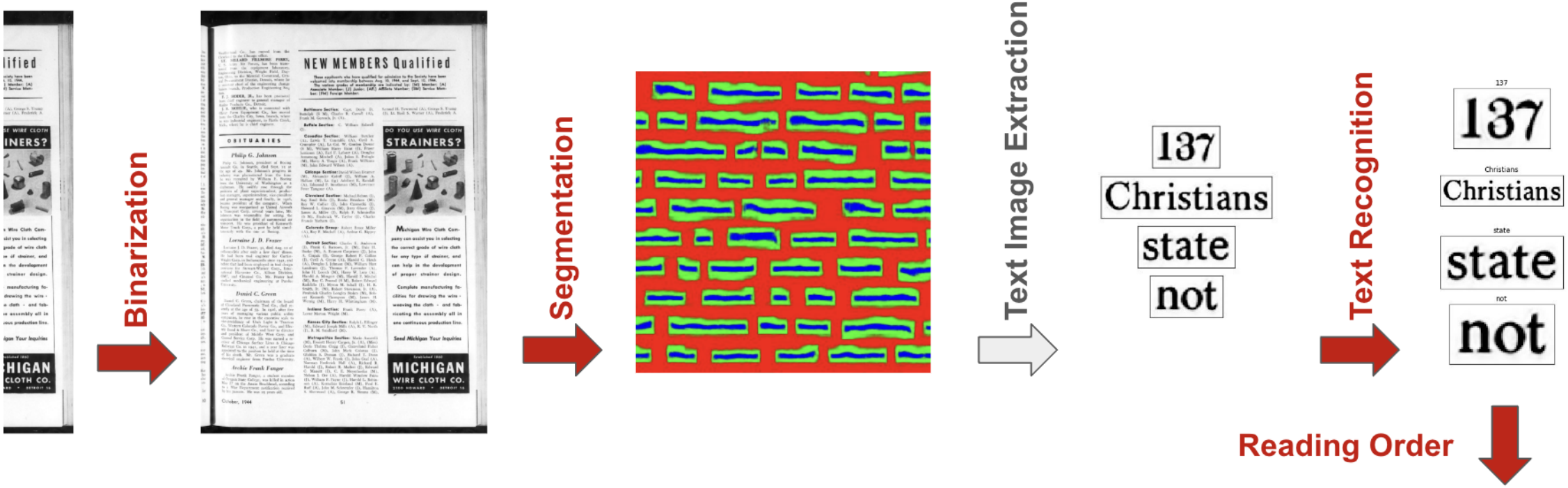
MICHIGAN WIRE CLOTH CO.

October, 1944 51

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
 <head>
  <title></title>
  <meta http-equiv="Content-Type" content="text/html;charset=utf-8" />
  <meta name='ocr-system' content='tesseract 4.1.1-rc2-25-g9707' />
  <meta name='ocr-capabilities' content='ocr_page ocr_carea ocr_par ocr_line ocrx_word ocrp_wconf' />
 </head>
 <body>
  <div class='ocr_page' id='page_1' title='image "sample.jpg"; bbox 0 0 2130 3433; ppageno 0'>
   <div class='ocr_carea' id='block_1_1' title='bbox 115 224 1804 284'>
    <p class='ocr_par' id='par_1_1' lang='eng' title='bbox 115 224 1804 284'>
     <span class='ocr_line' id='line_1_1' title='bbox 115 224 1804 284; baseline -0.001 -11; x_size 51.408451; x_descenders 6.4084506; x_ascenders 10'>
      <span class='ocrx_word' id='word_1_1' title='bbox 115 224 220 284; x_wconf 75'>1654)</span>
      <span class='ocrx_word' id='word_1_2' title='bbox 598 228 824 278; x_wconf 93'>Prerogative</span>
      <span class='ocrx_word' id='word_1_3' title='bbox 867 231 979 275; x_wconf 71'>Court</span>
      <span class='ocrx_word' id='word_1_4' title='bbox 1022 232 1059 271; x_wconf 96'>of</span>
      <span class='ocrx_word' id='word_1_5' title='bbox 1103 231 1343 279; x_wconf 73'>Canterbury.</span>
      <span class='ocrx_word' id='word_1_6' title='bbox 1738 236 1804 271; x_wconf 87'>273</span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_2' title='bbox 118 292 1792 301'>
    <p class='ocr_par' id='par_1_2' lang='eng' title='bbox 118 292 1792 301'>
     <span class='ocr_line' id='line_1_2' title='bbox 118 292 1792 301; baseline 0 3132; x_size 20; x_descenders 5; x_ascenders 5'>
      <span class='ocrx_word' id='word_1_7' title='bbox 118 292 1792 301; x_wconf 95'> </span>
     </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_3' title='bbox 116 350 1805 1098'>
    <p class='ocr_par' id='par_1_3' lang='eng' title='bbox 116 350 1805 1098'>
     <span class='ocr_line' id='line_1_3' title='bbox 116 350 1797 397; baseline -0.003 -12; x_size 41; x_descenders 8; x_ascenders 12'>
      <span class='ocrx_word' id='word_1_8' title='bbox 116 352 376 389; x_wconf 90'>SLADDEN,</span>
      <span class='ocrx_word' id='word_1_9' title='bbox 402 353 540 395; x_wconf 94'>James,</span>
      <span class='ocrx_word' id='word_1_10' title='bbox 566 352 606 386; x_wconf 93'>of</span>
      <span class='ocrx_word' id='word_1_11' title='bbox 626 364 668 396; x_wconf 91'>p.</span>
      <span class='ocrx_word' id='word_1_12' title='bbox 689 353 742 386; x_wconf 96'>St.</span>
      <span class='ocrx_word' id='word_1_13' title='bbox 770 353 980 397; x_wconf 96'>Margaret,</span>
      <span class='ocrx_word' id='word_1_14' title='bbox 1006 364 1095 386; x_wconf 96'>near</span>
      <span class='ocrx_word' id='word_1_15' title='bbox 1116 351 1339 389; x_wconf 96'>Rochester,</span>
      <span class='ocrx_word' id='word_1_16' title='bbox 1366 351 1478 388; x_wconf 96'>Kent,</span>
      <span class='ocrx_word' id='word_1_17' title='bbox 1504 350 1586 394; x_wconf 96'>July</span>
      <span class='ocrx_word' id='word_1_18' title='bbox 1612 360 1667 393; x_wconf 96'>27,</span>
      <span class='ocrx_word' id='word_1_19' title='bbox 1700 350 1797 392; x_wconf 96'>1653.</span>
     </span>
```

image $\rightarrow$ text, markup

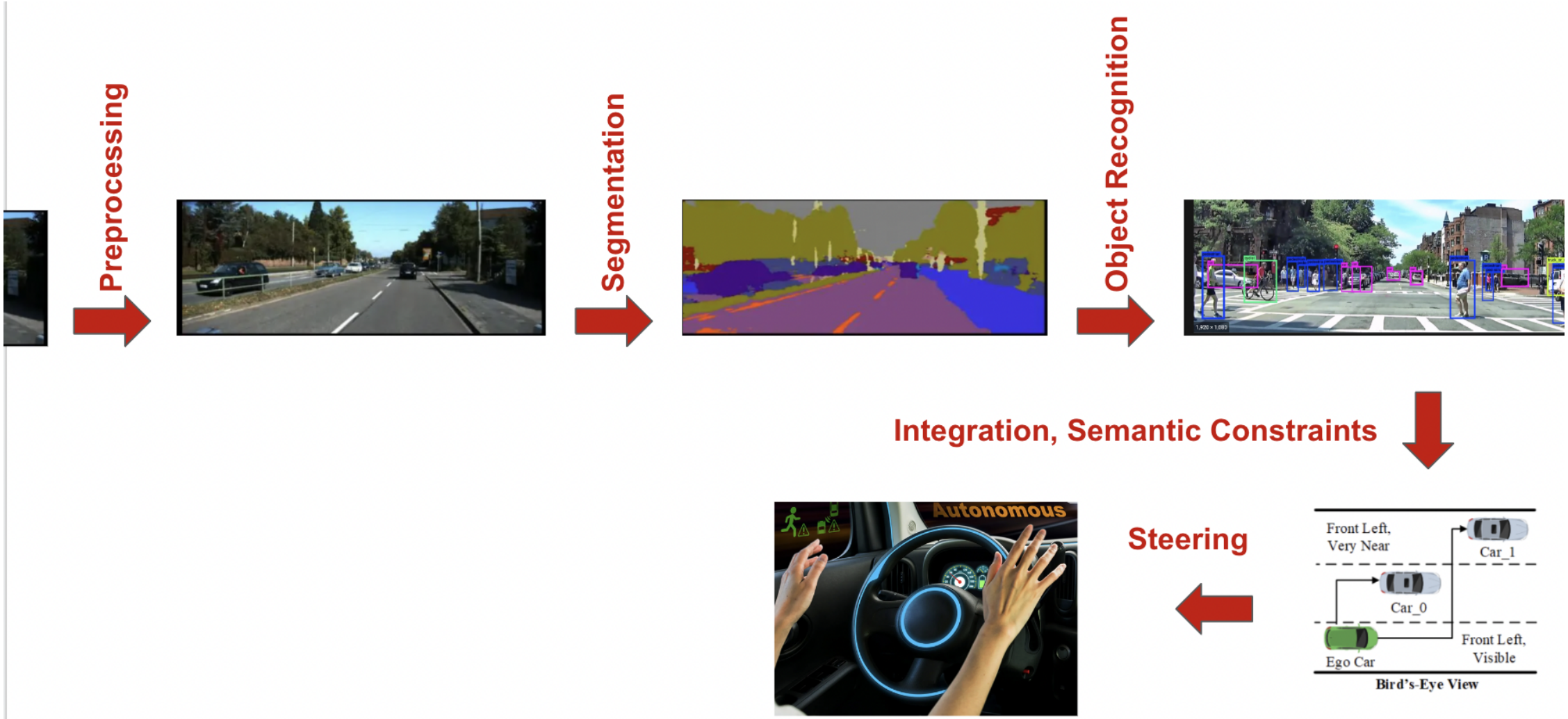
Motivating Example: OCR



1500s, when an unknown printer took a Bonorum et Malorum" by Cicero are also galley of type and scrambled it to make a reproduced in their exact original form, type specimen book. It has survived not accompanied by English versions from

"1500s, when an unknown printer took a galley of type and scrambled it to make a type…"

OCR vs AV Pipeline



Motivating Example: OCR

We need to train multiple models:

- preprocessing
- segmentation
- recognition
- syntactic model (language model)
- geometric model (reading order model)

This is common to many computer vision tasks using deep learning.

OCR and Semi-Supervised Learning

Available data:

- thousands of scanned pages with manual segmentation and text
- tens of thousands of scanned pages with approximate text
- millions of scanned pages without text
- large amounts of text without document images
- the ability to generate new texts and printed documents

Typical unsupervised / semi-supervised learning problem.

First Step: Supervised Training

Train supervised models on manually labeled / transcribed data:

- segmentation model
- image $\rightarrow$ text model

Yields good performance *on data similar to training data*.

Unlabeled training data primarily helps with generalization to new datasets.

Using Millions of Untranscribed Pages

Idea:

- Use our or other OCR systems to transcribe those pages and use the output as training data ("pseudolabel")

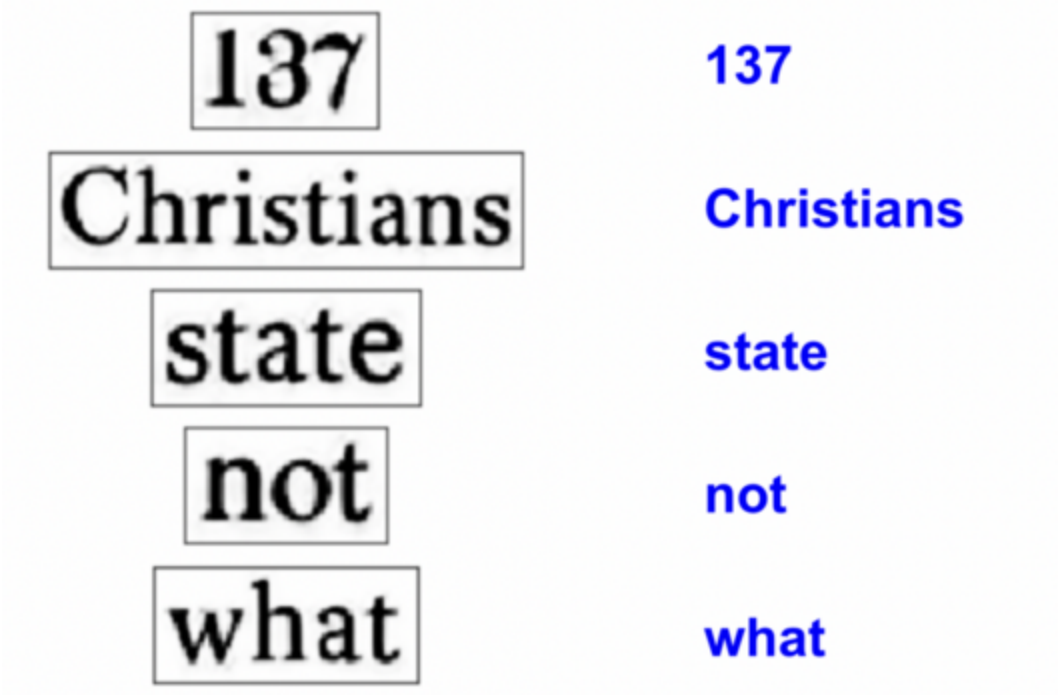
Questions:

- How can using a worse OCR system work for training a better OCR system?
- Is there anything we can do to improve this?

SEMI-SUPERVISED TEXT RECOGNITION

Example: Word Recognition Problem

Let's focus just on recognizing words for the following examples (forget about the rest of the OCR system):



Using Millions of Untranscribed Pages

Idea:

- run the existing OCR system, giving word images and corresponding text
- construct a new training set by...
  - rejecting any word that the OCR system has a low confidence in
  - rejecting any word that is not found in the dictionary
- this way, we obtain a new training set that contains "mostly good" training samples
- iterate this multiple times

Why does this work?

Network estimates $P(c|x)$ (x: image, c: class)

For OCR, we know:

- true $P(c|x)$ is approximately 0 or 1 (no ambiguities)
- any uncertainty in classifier output is due to mislabeled training data
- e.g. if 20% of training data mislabeled: $\tilde{P}(c|x) = 0.8$ for true class c
- if we use pseudolabel $\arg\max_c \tilde{P}(c|x)$, many accidentally mislabled training labels will actually be correctly labeled
- as a result, the posterior probability will be estimated higher on the next training round and the model improves

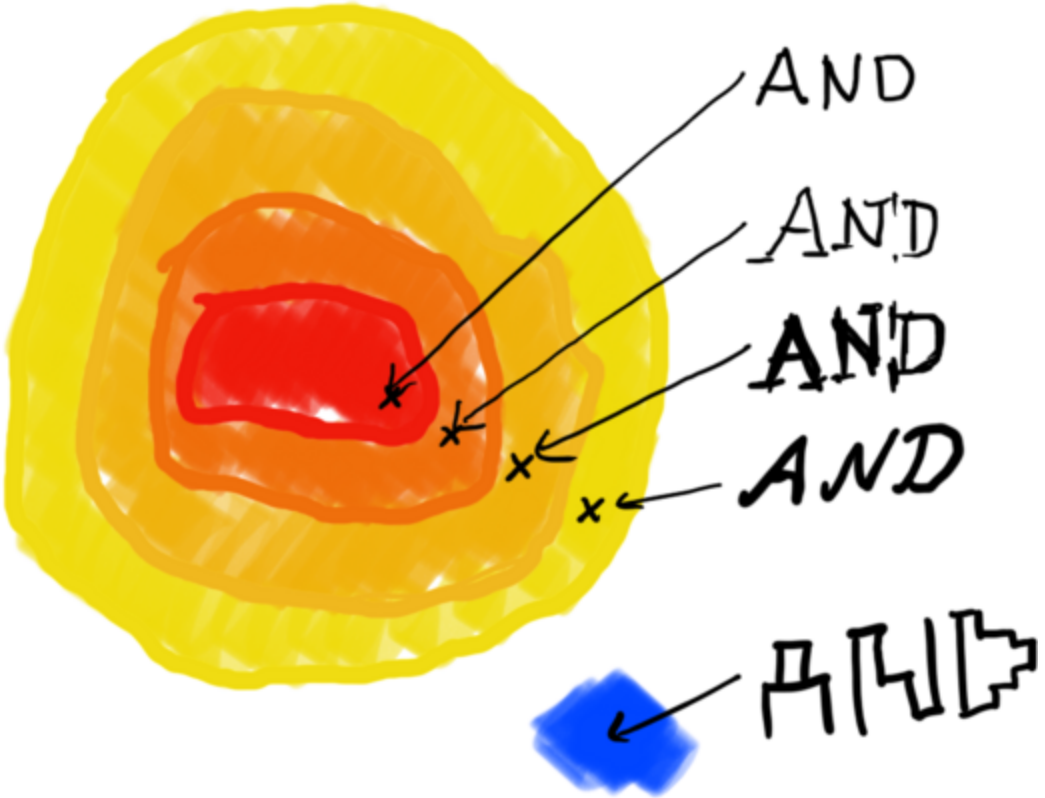
EM Algorithms

- latent variables in semi-supervised OCR
  - labels for unlabeled portion of training set
  - outlier status for unlabeled portion of training set
- EM algorithms recover latent variables by...
  - "making a best guess" given the current model
  - retraining the model as if that guess is correct

Unsupervised Learning

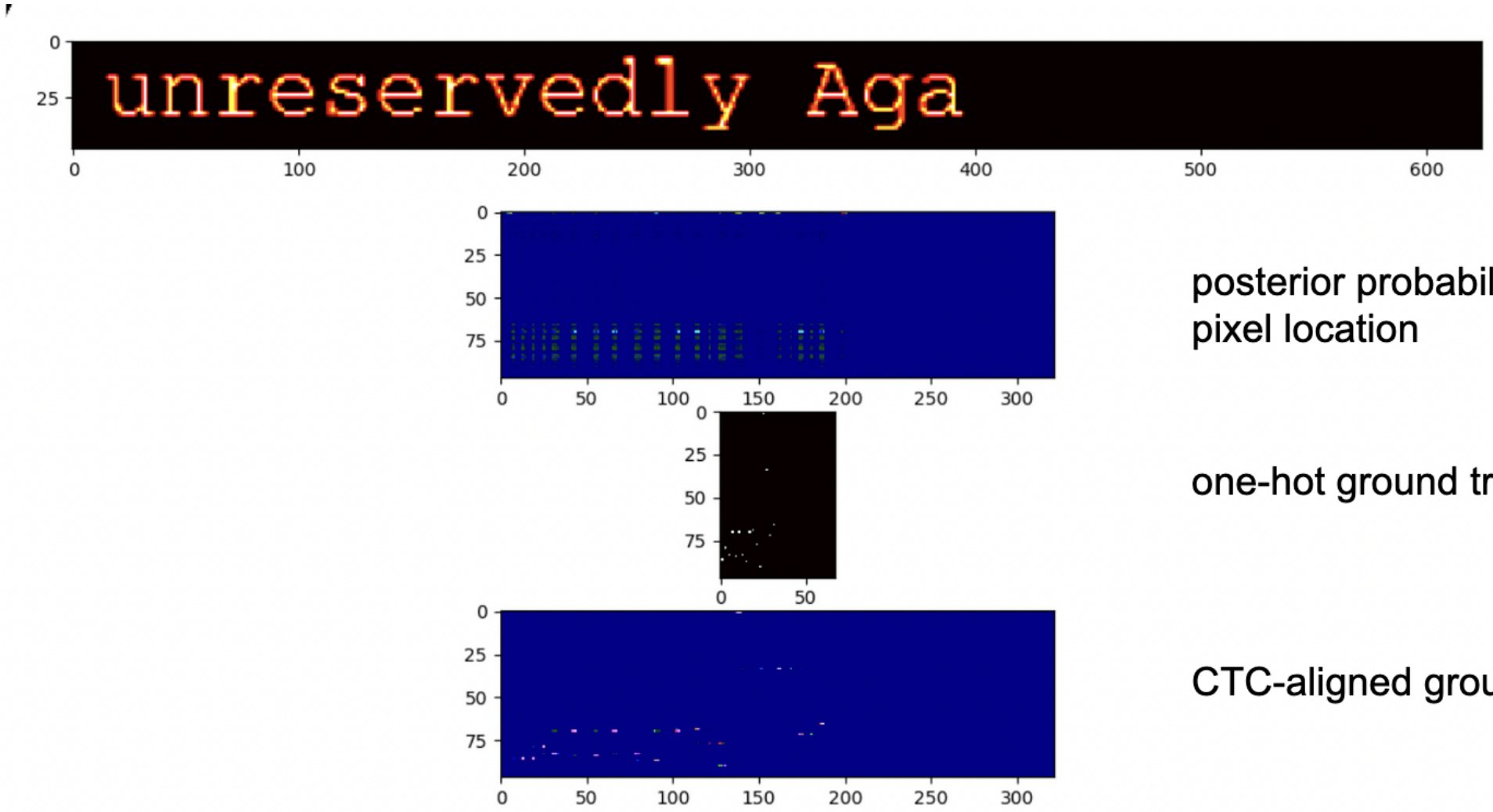
- identify the *latent variables* that are being recovered
- identify the *prior assumptions / inductive biases* in the model
- identify the *EM algorithm implementation*

Iterated Recognition / Dataset Construction



WEAKLY SUPERVISED TEXT RECOGNITION

EM Training with Language Models (aka CTC)



EM Training with LSTM/Conv + CTC

- ideally, the LSTM/Conv model outputs the correct class only where the character occurs
- outputs ϵ (no character) everywhere else
- our transcript does not contain the character positions
- CTC performs *alignment* between classifier output and transcript
- CTC estimates the most likely positions of characters given the current model

CTC Decoding

EM Training without Any Transcripts

Normal EM-Training:

- image + transcript, EM-training recovers alignment

Language Model-Based EM-Training:

- image + language model, EM-training recovers text + alignment

EM Training, Information Theory

There always has to be *some* source of information:

- unigram perplexity: 1000 ⟵ what the classifier outputs
- bigram perplexity: 200
- trigram perplexity: 100 ⟵ what the language model imposes
- full transcript perplexity: 1 ⟵ fully supervised training

Additionally: several bits per character for character location.

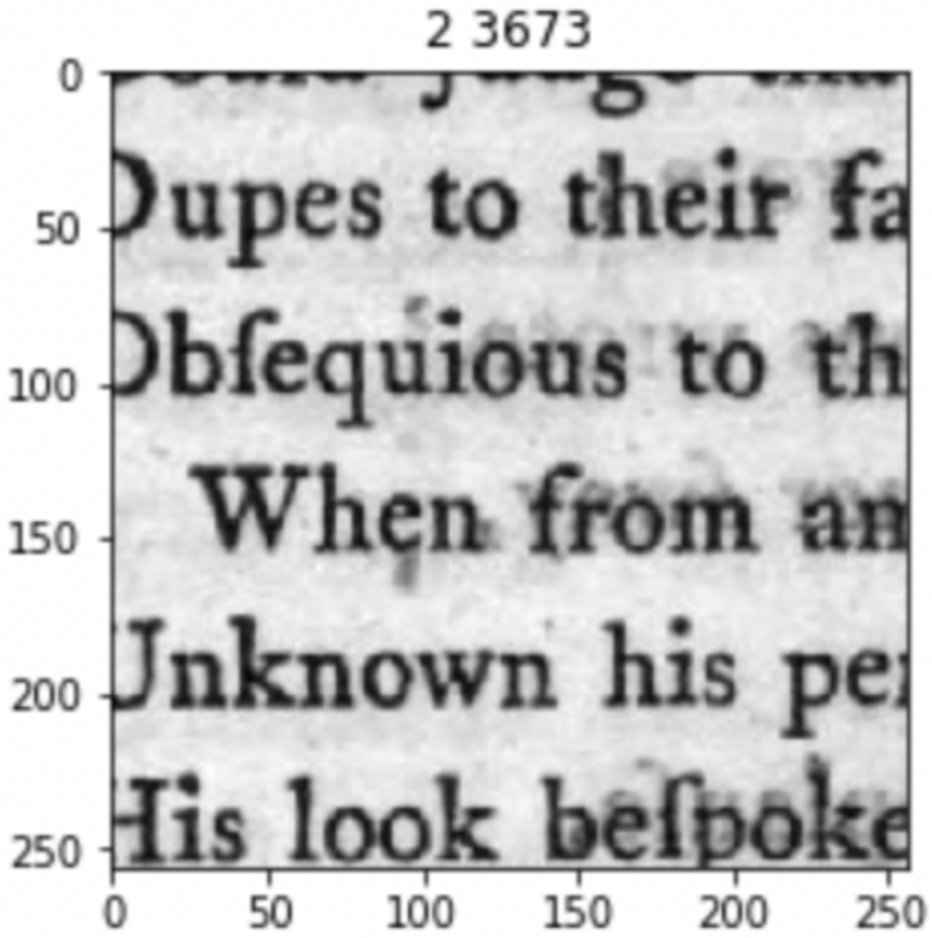
Different Forms of Weak-/Unsupervised Learning

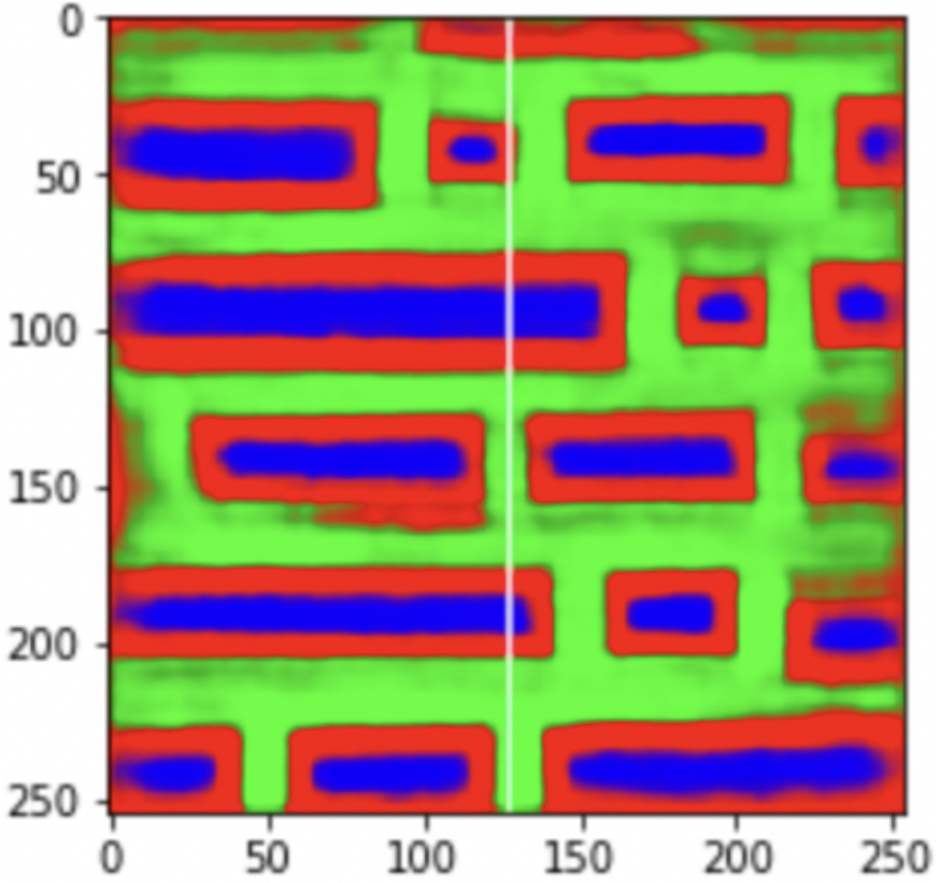
We have seen three different forms of weak/unsupervised training for text → image training:

- using image + transcript, lacking just the alignment/segmentation
- using image + language model only
- using just pseudolabels from weak classifier

SELF-SUPERVISED PAGE SEGMENTATION

Self-Supervision of Page Segmentation





Self-Supervision of Page Segmentation

Assume we start with errorful segmenter from small labeled training data.

Two kinds of errors:

- segmenter returns word for non-words
- segmenter misses actual words

Approach:

- validate each returned word using OCR (no OCR = not a word)
- mark everything that is not identified as a word as "don't know" and exclude from training

Self-Supervision

Approaches to self-supervision usually incorporate prior knowledge and require some manual design:

- text recognizer
  - reject non-words from "soft labels"
  - use language models as part of EM training
- page segmentation
  - validate segmentation via OCR
  - introduce "don't know" mask during training

Self-Supervision

Prior knowledge about the task is required to choose meaningful self-supervision tasks.

- word recognition: pseudolabels, language-model based rejection
- segmentation: verification via OCR, introduction of "don't care" regions
- object reconition, natural image segmentation: *later*

ACTIVE LEARNING

Active Learning

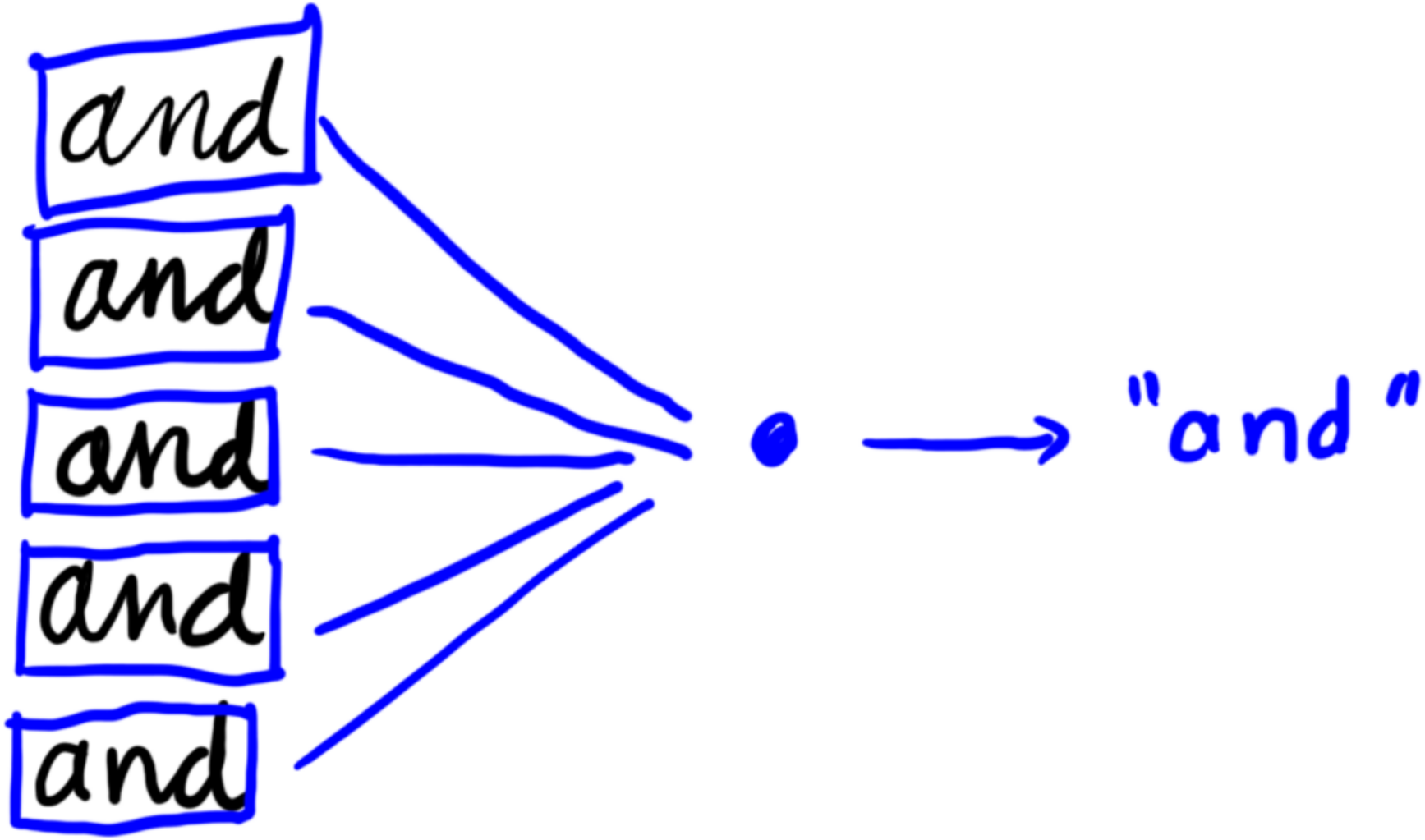
Request Help from Oracle:

- run OCR as above and send low confidence outputs to manual transcribers
  - output near decision boundaries
  - output that can maximally help the classifier improve

Clustering:

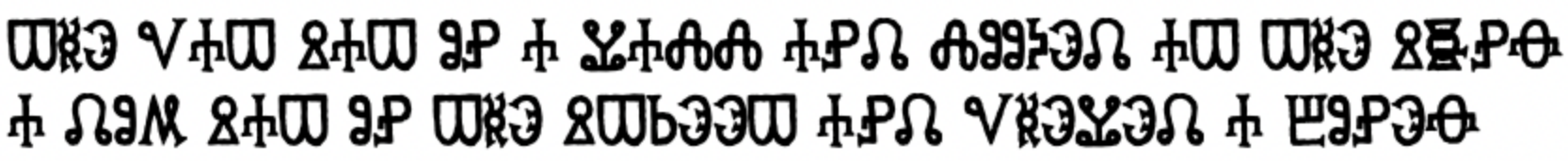
- perform clustering on character or word images and manually transcribe each cluster

Transcription with Clustering



Question: what similarity measure do we use for clustering?

OCR by Solving a Cryptogram



- cluster characters by shape
- use frequency and patterns to infer character identity
  - e.g.: first word is likely "THE"
    - letter frequencies
    - word frequencies
  - single letter word is likely "A"
- explicit form of many EM-based recognition algorithms

OCR by Solving a Cryptogram

THE CAT SAT ON A WALL AND LOOKED AT THE SUN.
A DOG SAT ON THE STREET AND CHEWED A BONE.

DATA SOURCE MODELS

Hierarchical Bayes

- clustering is based on a simple hierarchical Bayesian source model
- latent variables can be recovered by EM or Bayesian methods
- simple clustering assumes $\Sigma = 1$
- both c and μ are latent; need oracle to recover actual class labels

Channel View

$c \in \mathbb{Z}_k \xrightarrow{\text{render}} \xi \xrightarrow{\text{noisy channel}} x \xrightarrow{\text{classifier}} P(c|x)$

OCR: render = digital typeset; noisy channel: printing + scanning

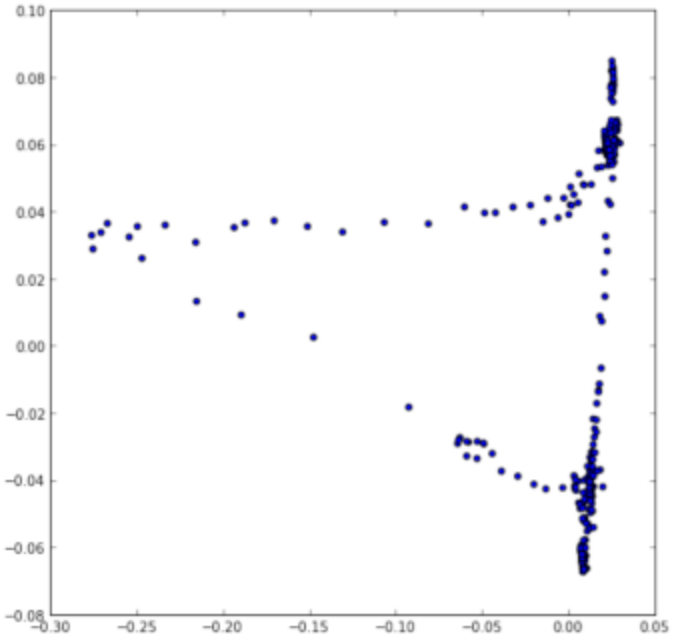
Vision: render = choice of pose ; noisy channel: lighting, camera

Rendering can involve transformation parameters, giving rise to *view manifolds*.

Real View Manifold

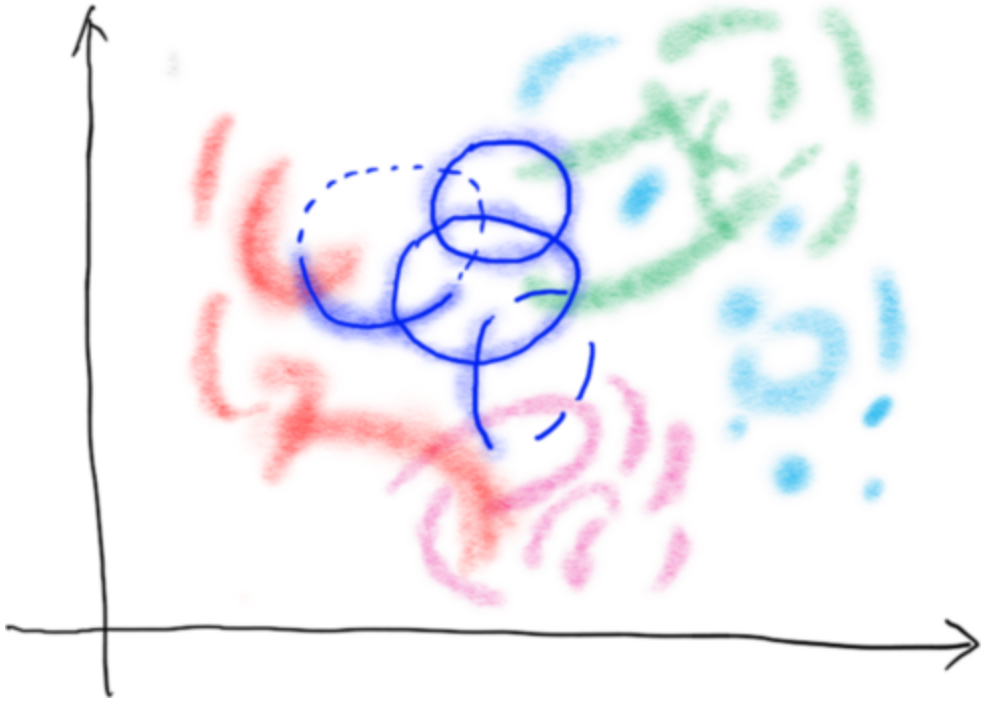
5 second motion sequence of object rotating back and forth

```
from sklearn.decomposition import RandomizedPCA
pca = RandomizedPCA(20)
lo = pca.fit_transform(data)
mds = manifold.LocallyLinearEmbedding()
vl = mds.fit_transform(lo)
scatter(vl[:,0],vl[:,1])
```



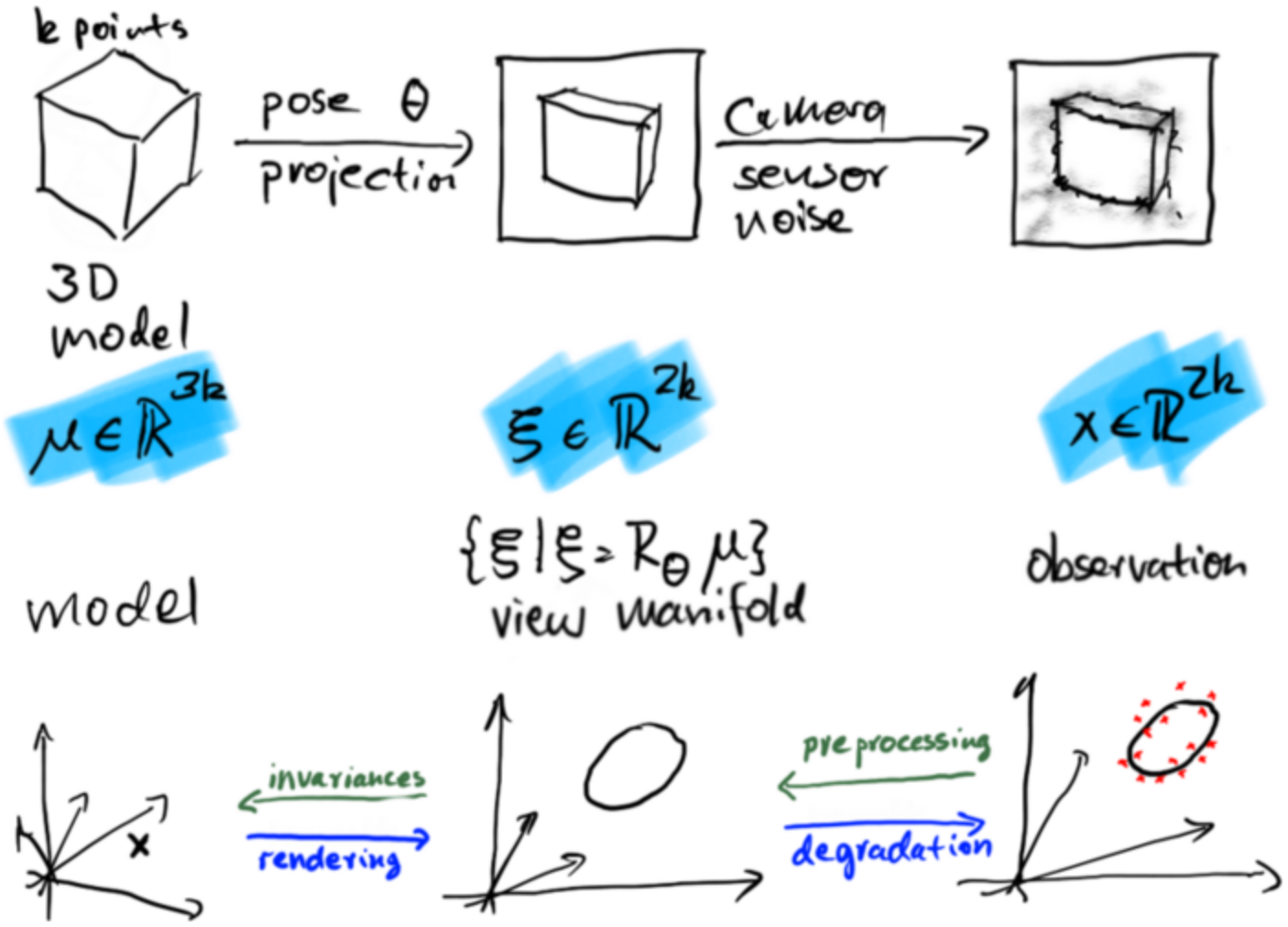
Dimensionality reduction in pixel space: $\mathbb{R}^{256} \rightarrow \mathbb{R}^2$. Video: rotating.mp4

Manifolds Learning



- think of real data as clouds surrounding...
- union of manifolds with boundary (only shown for blue class)
- "manifold learning" = union-of-manifold-s-with-boundary learning

Model, View Manifold, and Noise



Unsupervised Learning

Given or Learned:

- rendering
- degradation

Constructed or Learned (inverse problems):

- invariances
- preprocessing / image cleanup

CycleGAN can learn both directions simultaneously with no supervision.

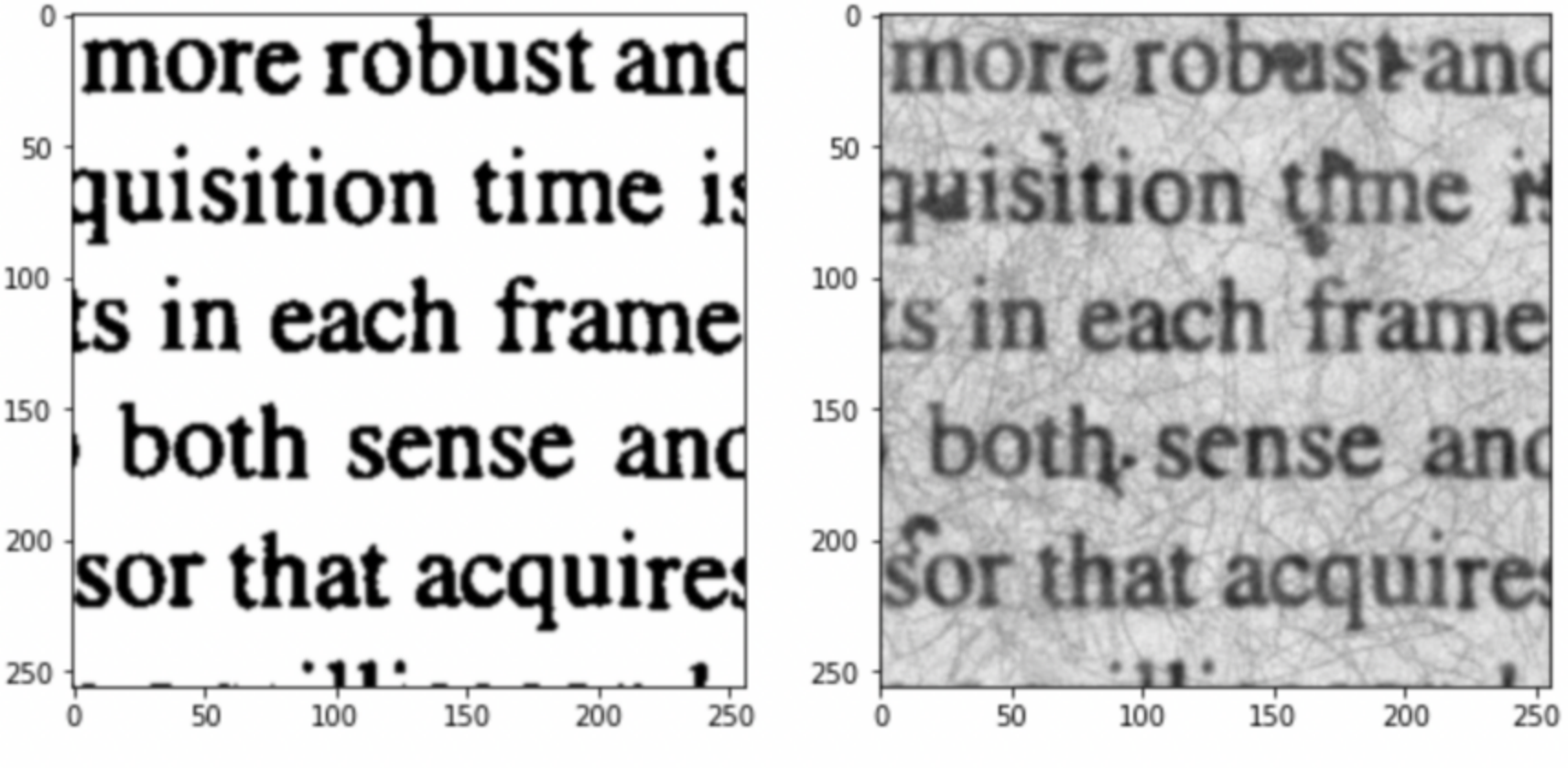
TRAINING DATA GENERATION / DEGRADATION

Artificial Data Generation

This is the "forward path" in the channel view of recognition:

- digital typesetting = perfect "artificial" document generation
- take any text, generate pages of perfect text
- for OCR, we need "degraded images"
  - printing, scanning, photographing, photocopying, ...
- physical processes for document image degradataion are well known

Artificial Data Generation



Document Image Restoration

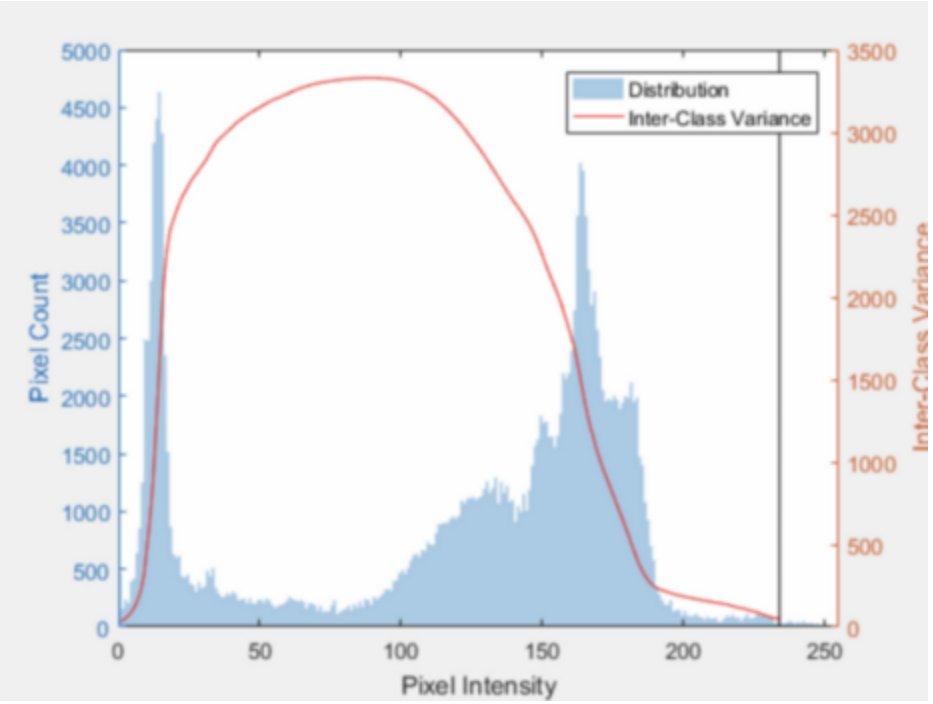
- recognition is easier if documents are not degraded
- can we restore the "clean" image via unsupervised learning?

AV/object recognition: image translation prior to recognition

Classical Document Image Restoration

- binarization
  - optimal thresholding
  - optimal linear filtering (deconvolution, etc.)
  - clustering
  - performance-based dynamic thresholding
- deep learning
  - supervised restoration (LSTM, pix2pix)
  - unsupervised restoration via CycleGAN

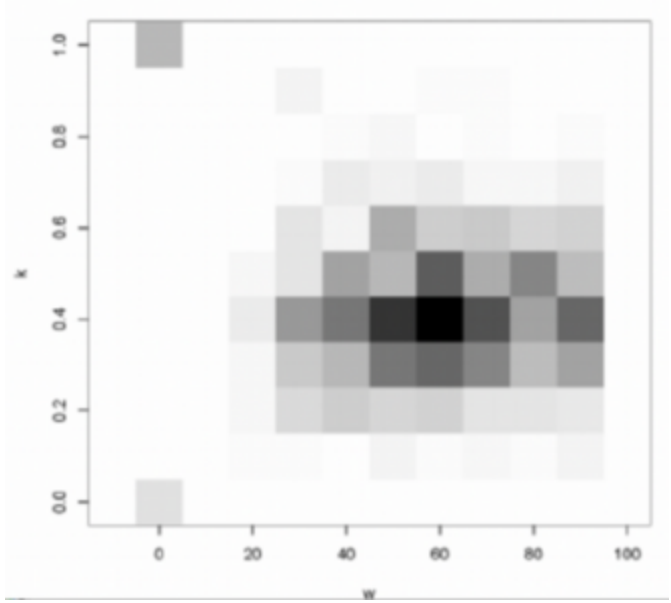
Otsu's Method: "Optimal" Thresholding



Otsu's method:

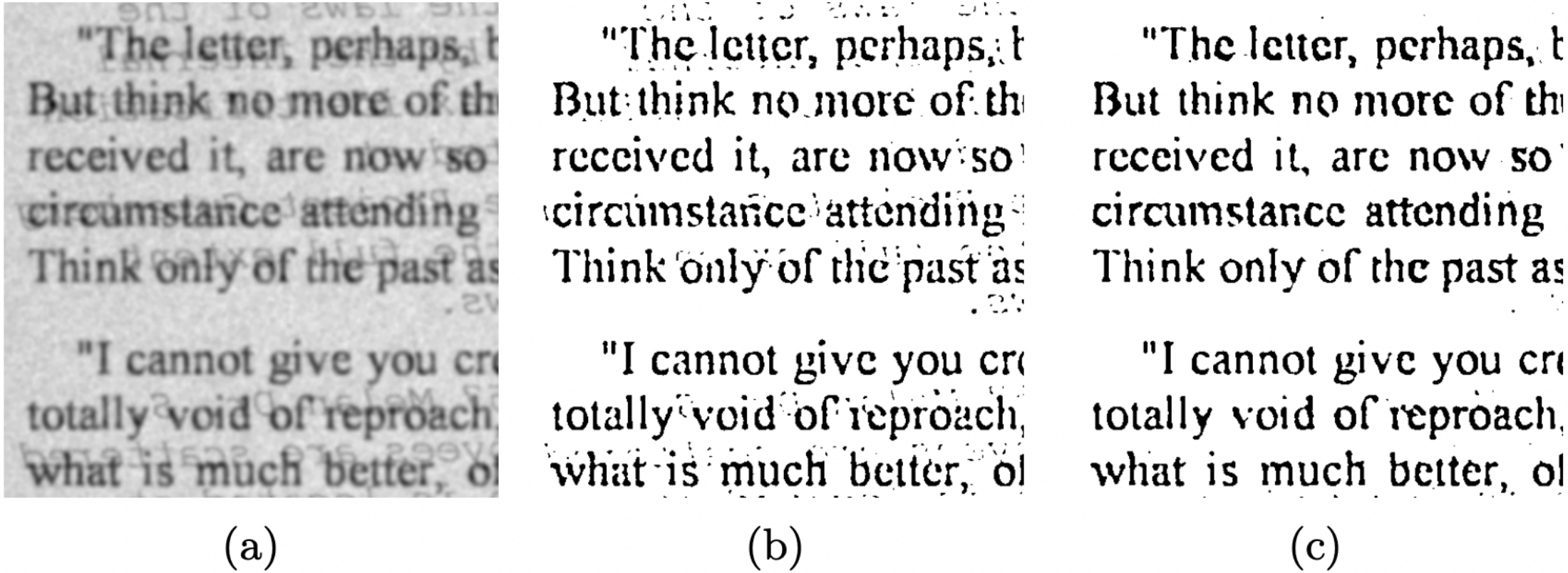
- manually constructed inverse degradation model based on prior knowledge/assumptions
- assume some kind of mixture model of image pixel generation
- maximize inter-class variance

Performance-Based Thresholding



- run thresholding with many parameters
- pick the parameters that yield the best OCR output
- no transcript, so use proxy
  - use statistics/classifier, or #words in dictionary
  - closely related to GAN methods
- assumption/prior: local thresholding = good way of inverting degradation model

Binarization by Supervised Deep Learning



(a) original, (b) Sauvola, (c) LSTM-based binarization

Self-Supervised Training for Binarization

1. generate clean images, degrade with degradation model
    - takes advantage of prior knowledge of degradataion models
2. take degraded images, use performance-based thresholding
    - takes advantage of knowledge of statistical properties of output

CycleGAN for Document Preprocessing

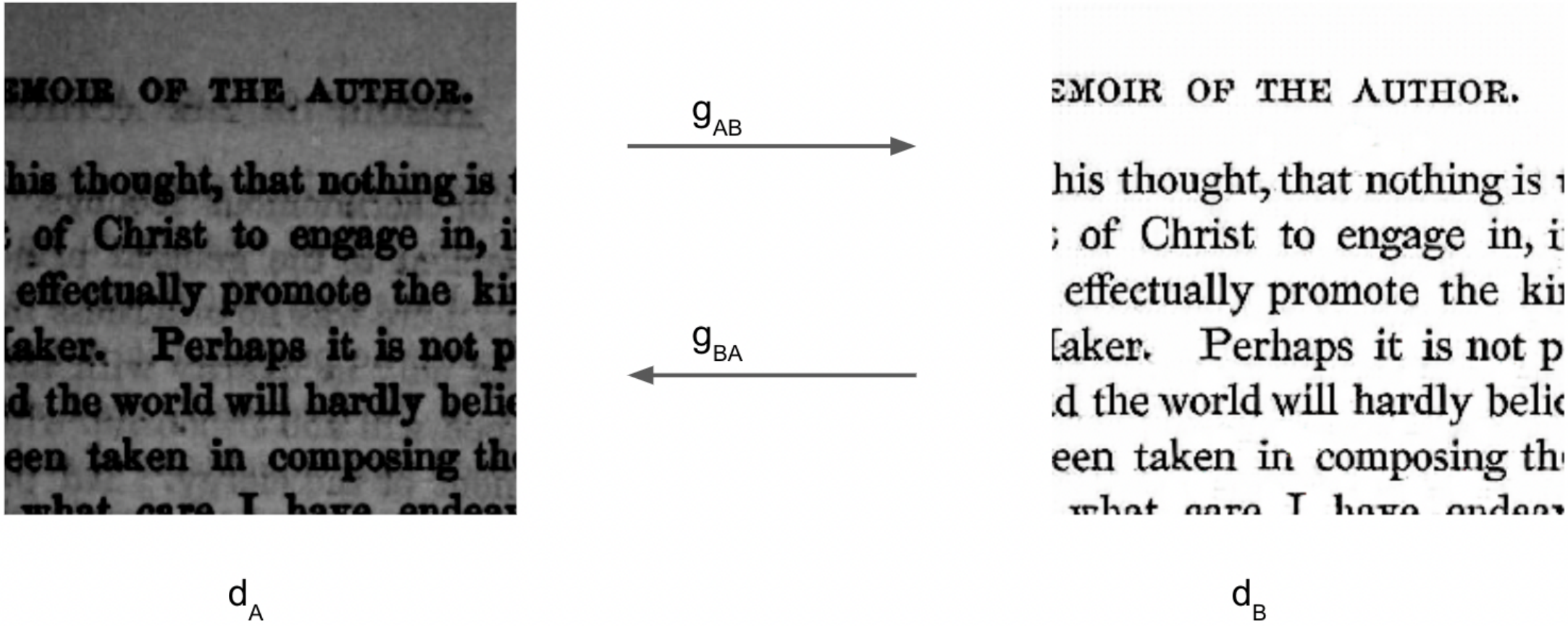
CycleGAN replaces all those components with trainable networks:

- clean image $\rightarrow$ degraded image
- degraded image $\rightarrow$ clean image
- clean image detector
- degraded image detector

These correspond to the forward and backward arrows in our channel model.

CycleGAN can be trained end-to-end without any labeled data or (significant) prior assumptions.

CycleGAN for Document Preprocessing



End-to-End OCR

NEW MEMBERS Qualified

OBITUARIES

Philip G. Johnson

Lorraine J. D. Frazer

Daniel C. Green

Archie Frank Fanger

DO YOU USE WIRE CLOTH STRAINERS?

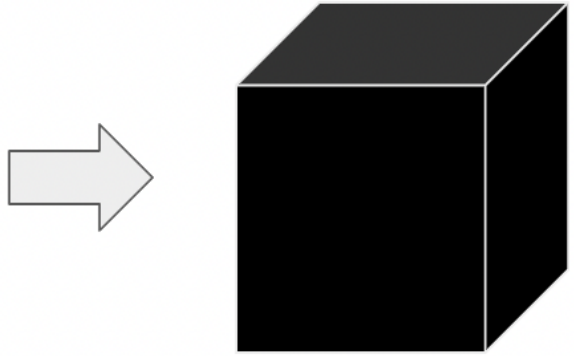
Michigan Wire Cloth Company can assist you in selecting the correct grade of wire cloth for any type of strainer, and can help in the development of proper strainer design.

Complete manufacturing facilities for drawing the wire - weaving the cloth - and fabricating the assembly all in one continuous production line.

Send Michigan Your Inquiries

MICHIGAN WIRE CLOTH CO.

October, 1944 51



Weatherhead Co., has moved from the Cleveland to the Chicago office.
LT. MILLARD FILLMORE PERRY, U. S. Army Air Forces, has been transferred from the equipment laboratory, Engineering Division, Wright Field, Dayson, Ohio, to the Materiel Command, Central Procurement Division, Detroit, where he is assistant chief of the engineering change branch
…

End-to-End OCR

Recent Developments

- transformer models permit full end-to-end training for image-to-text transcriptions
- do not need intermediate segmentation, text-line recognition

LayTR Multicolun Recognition

Image-to-HTML tags.

LayTR Table Recognition

tweezed	hydroxylase	preinvasive	filices	unlabialising
annalists	implodes	lachrymation	soir	scleroprotein
yashiro	zibeth	vespering	crotchets	ascescency
hundrodponny	trogonoid	aquatical	limburgito	prododication
similarize	rhizocephala	brooklike	supervalued	contrary
milligramage	cativo	yorkshireism	protension	geleem
midnightly	pedatifid	scienced	bacteriosis	germanomania

bayberries	cyanidin	nicotinise	stercorary	incogitant	talepyet
unlethal	snuffle	poltroonery	scimiterpod	abasedness	boomers
sediments	mulctatory	spindlewise	timocratical	renunciatory	standardizer

fleahopper gallivanting
oncoming heathenize
breakback celebrators
overwrestle sampler

midianitish	elastomeric
rodding	polymelian
cottonwick	verbena
smouser	gingalls
shippound	verifiably
melter	lurch

Figure 3: Four samples (a)-(d) featured in the LayTr-Tables dataset.

Image-to-HTML tags.

Self-Supervised Training

- apply the same principles
- train an initial model using supervised data
- compute output on unlabeled data ("soft labels")
- correct the output using language models
- retrain
- possibly use auxiliary tasks for pretraining (later)

Or has it been solved already...

Summary

In the OCR example, we have seen most of the major concepts of unsupervised and semi-supervised training:

- EM training (aka soft labeling)
- use of language modeling as data source
- clustering
- unsupervised preprocessing / image enhancement