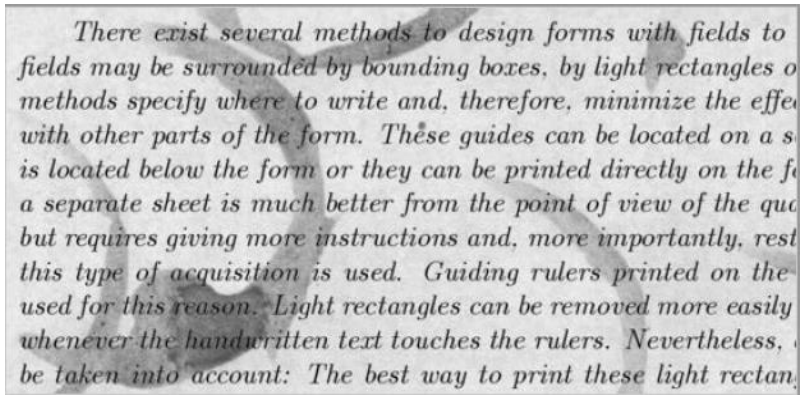
Introduction

- OCR – Optical Character Recognition
- Technique for converting printed text into machine-encoded text.

Text Image

OCR Output

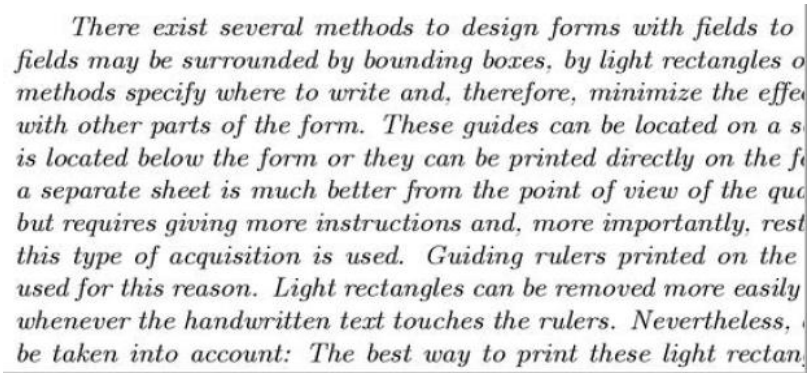
Noisy Image



There exist several methods to design forms with fields to
fields may be surrounded by bounding boxes, by light rectangles o
methods specify where to write and, therefore, minimize the effe
with other parts of the form. These guides can be located on a s
is located below the form or they can be printed directly on the f
a separate sheet is much better from the point of view of the qua
but requires giving more instructions and, more importantly, rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten text touches the rulers. Nevertheless,
be taken into account: The best way to print these light rectan

There exist se·ueral 1nethod11- to design forms with fields to '
fields 1nay be s·urrounded by bounding boxes. by light rectangles o
methods specify whe (o write and. therelo re. 11iini1nize the effe
with other parts of the Jonn. These guides can be located on a s
is located below the Jonn or they can be printed directly on the f<
a separate sheet is 1nuch better fro1n the point of view of tJie quc
but requires giving 1nore 'instructions and. 1nore i1 nportantly. rest
this type of a -isitio11 is used. Gitiding rulers printed on the
used jdr this Light rectangles can be re1noved 1nore easily
whene el tluJ ·ilten text touches tJie rulers. Ne·uertheless
be taken into ccount: The best iuay to print these light rectan:

Clean Image

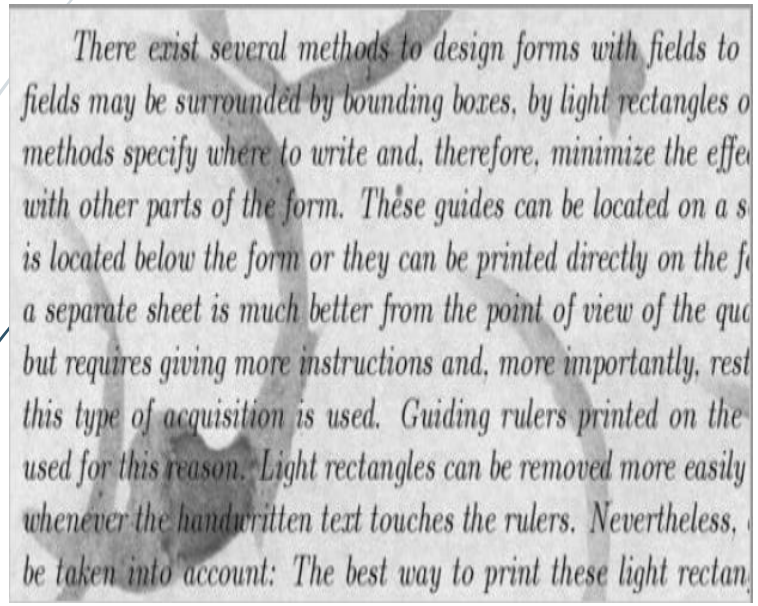


There exist several methods to design forms with fields to
fields may be surrounded by bounding boxes, by light rectangles o
methods specify where to write and, therefore, minimize the effe
with other parts of the form. These guides can be located on a s
is located below the form or they can be printed directly on the f
a separate sheet is much better from the point of view of the qua
but requires giving more instructions and, more importantly, rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten text touches the rulers. Nevertheless,
be taken into account: The best way to print these light rectan

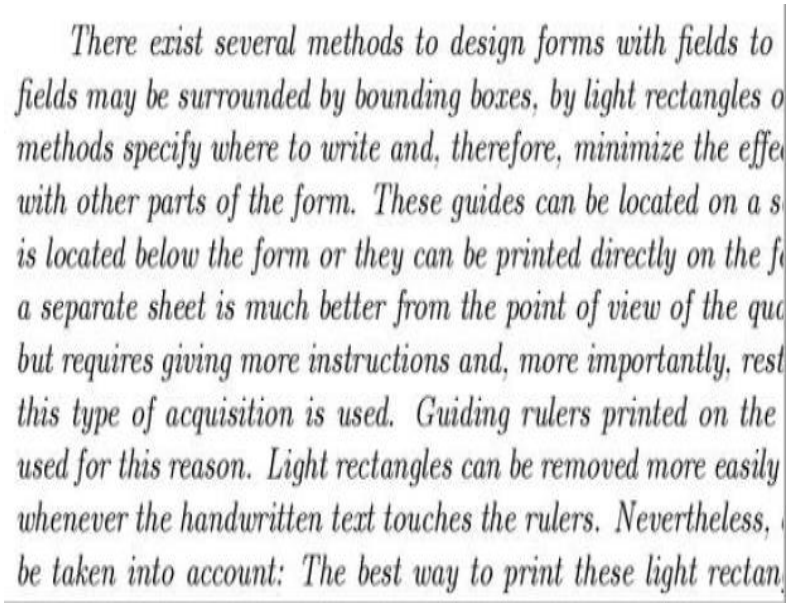
There exist several methods to design forms with fields to
fields 1nay be surrounded by bounding boxes, by light rectangles o
methods specify where to write and. therefore, mini1nize the effe,
with other parts of the for1n. These guides can be Located on as,
is Located below the form or they can be printed directly on the f<
a separate sheet is much better from the point of view of the quc
but requires giving 1nore instructions and, 1nore i1nportantly. rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten te:i:t touches the rulers. Nevertheless,
be taken into account: The best way to print these light rectan:

Objective

Using image processing and machine learning approaches create a predictive algorithm to clean up the noisy images of text



Noisy Image



Clean Image

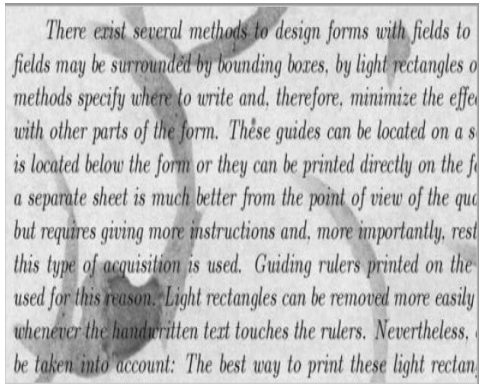
Business Use Case

Lot of old and fragile documents can be digitized in a readable format.
Will improve OCR technique in accurately converting text image to live text.

Data set

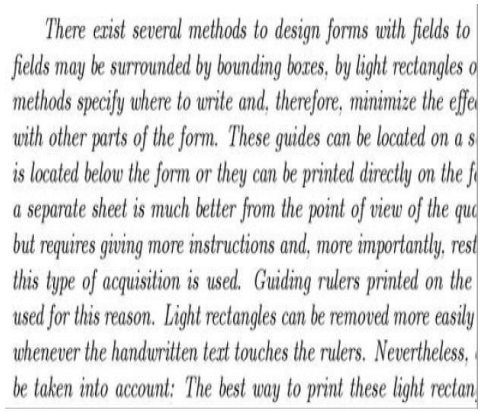
Dirty images

- Grayscale images
- 8 bit images
- Contain Noise

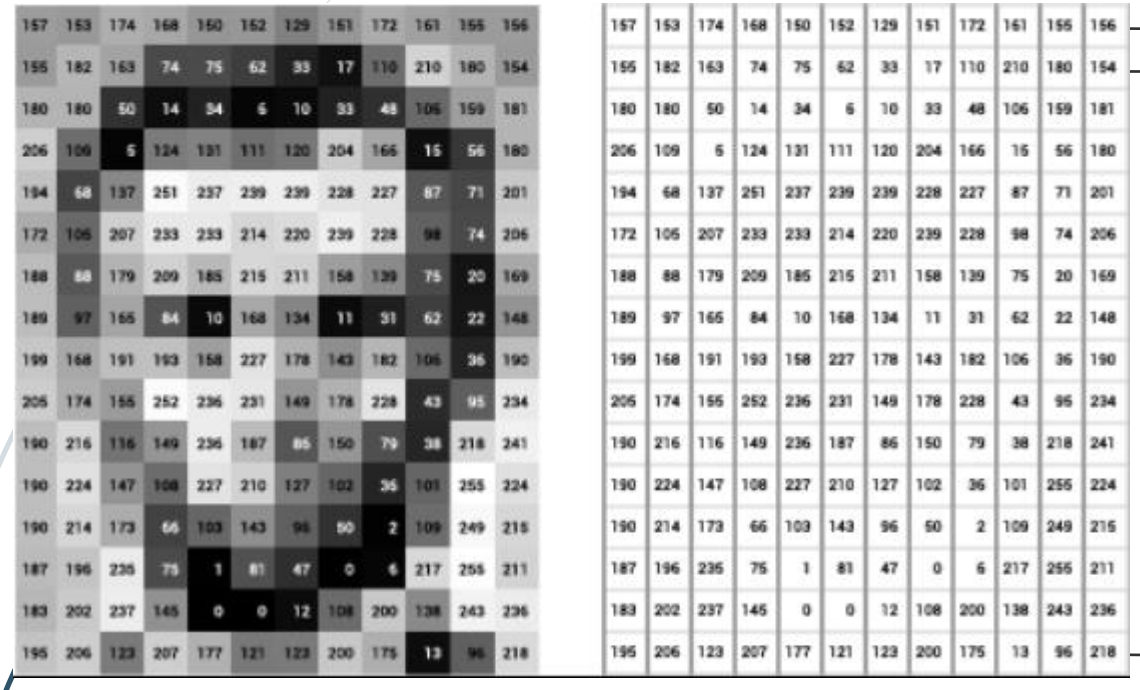


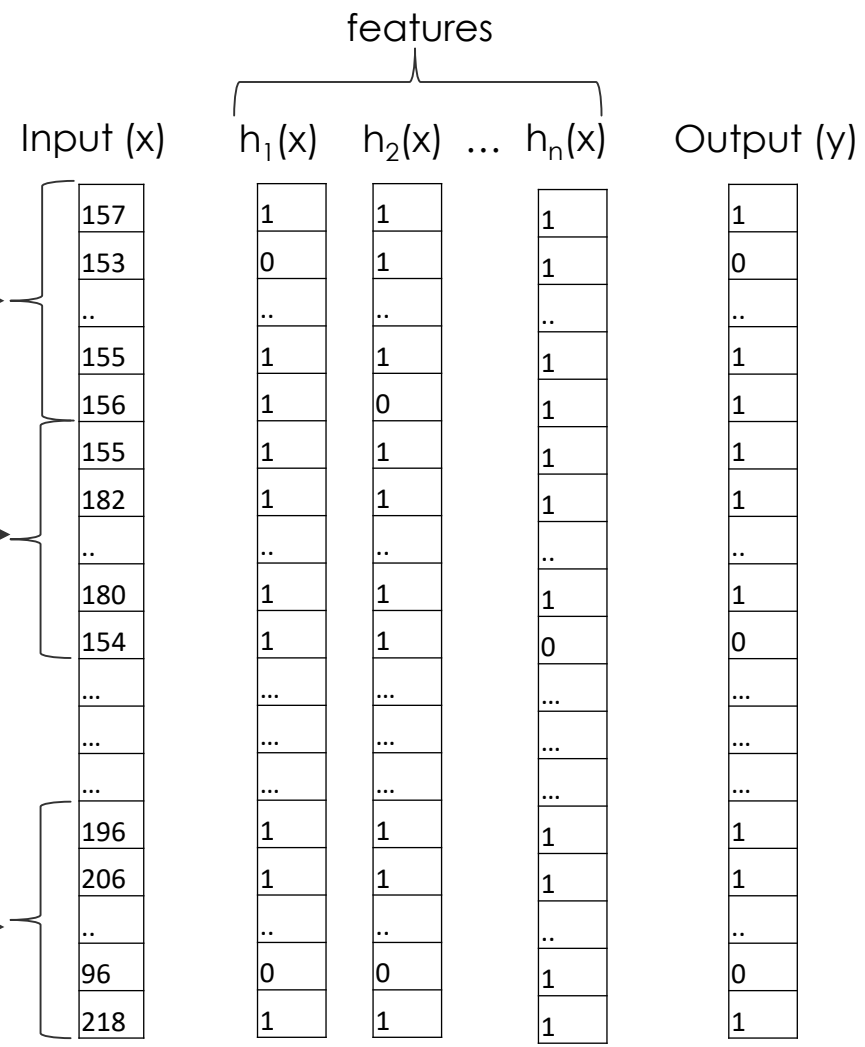
Clean images

- B&W images
- 1 bit images
- No Noise



Approach

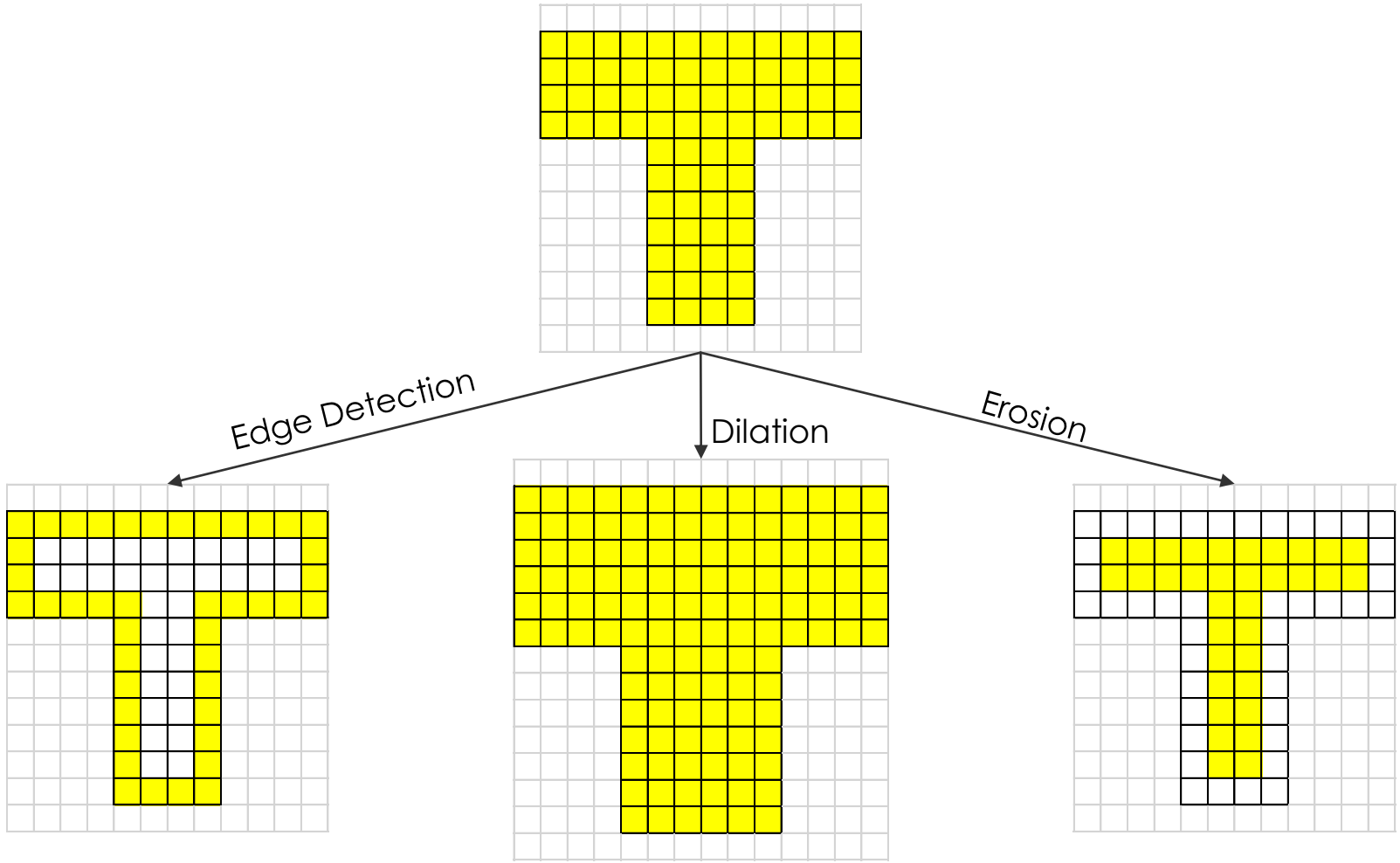




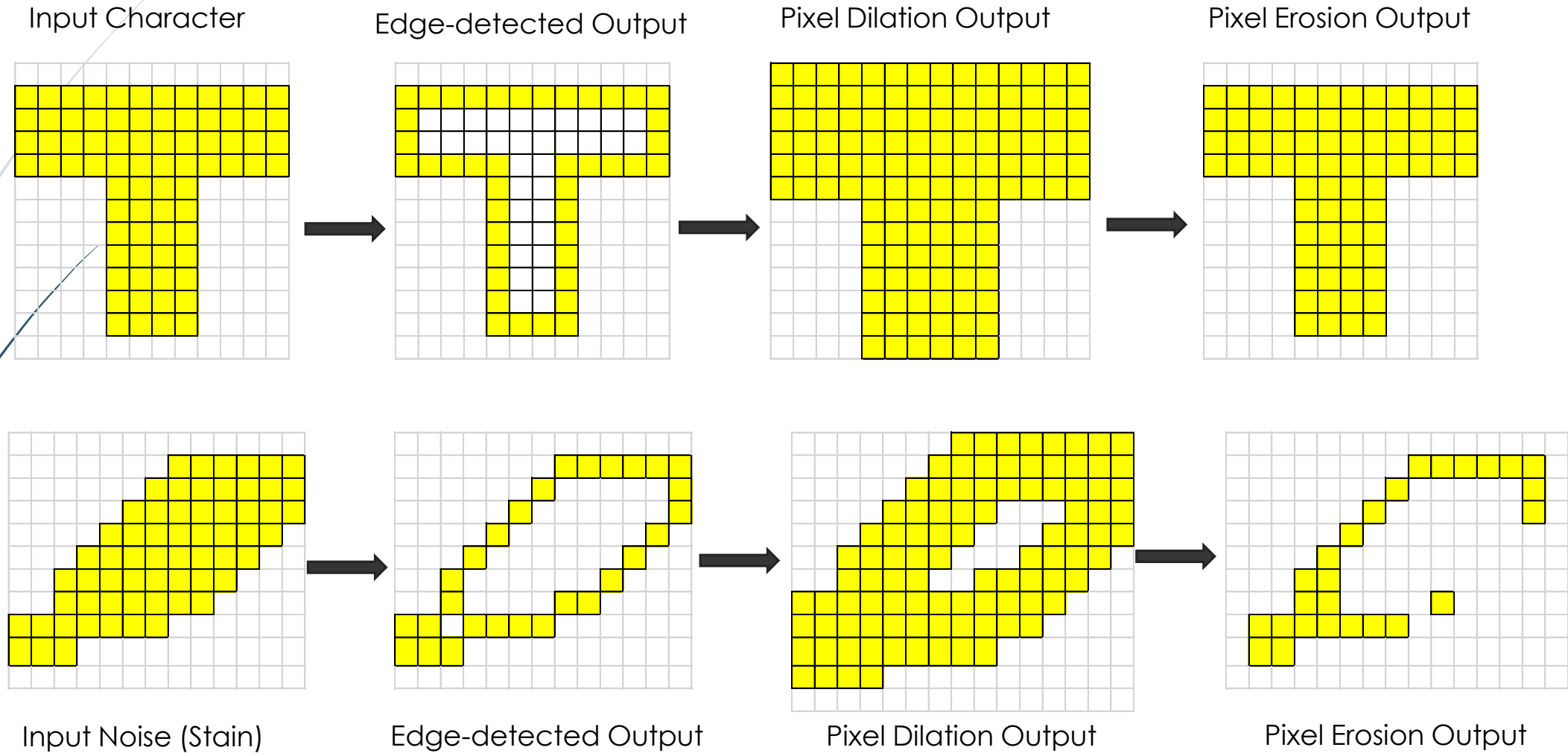
Step1 : Use image processing techniques to create features – $h_1(x)$, $h_2(x)$...
Step 2: Use machine learning techniques to find relationship b/w y and input variables
$y = f(x, h_1(x), h_2(x)\ldots\ldots, h_n(x)) + e$

Edge Detection and Morphology

Image Processing Technique 1:
Edge Detection and Morphology



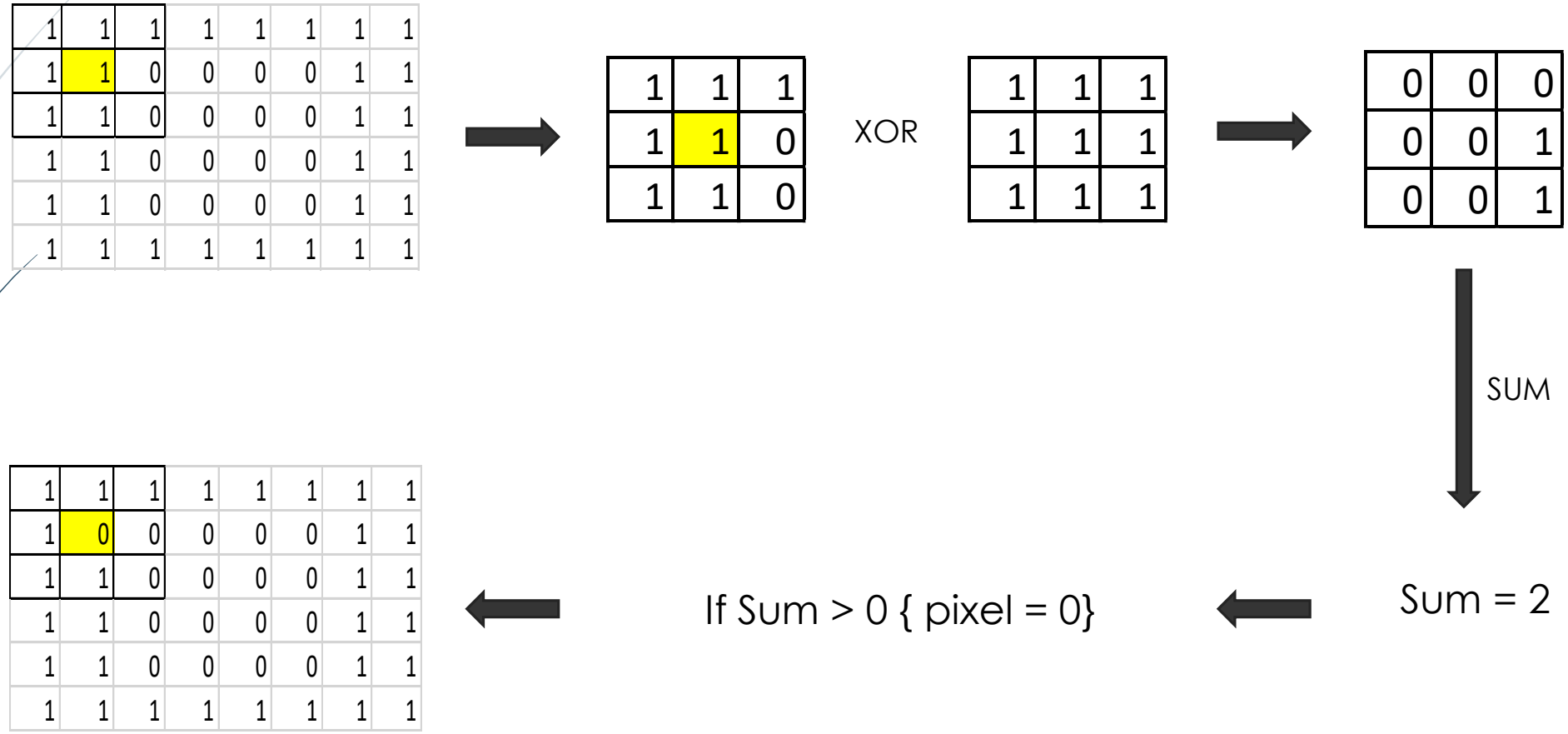
Motivation


Input Character
Edge-detected Output
Pixel Dilation Output
Pixel Erosion Output
Input Noise (Stain)
Edge-detected Output
Pixel Dilation Output
Pixel Erosion Output

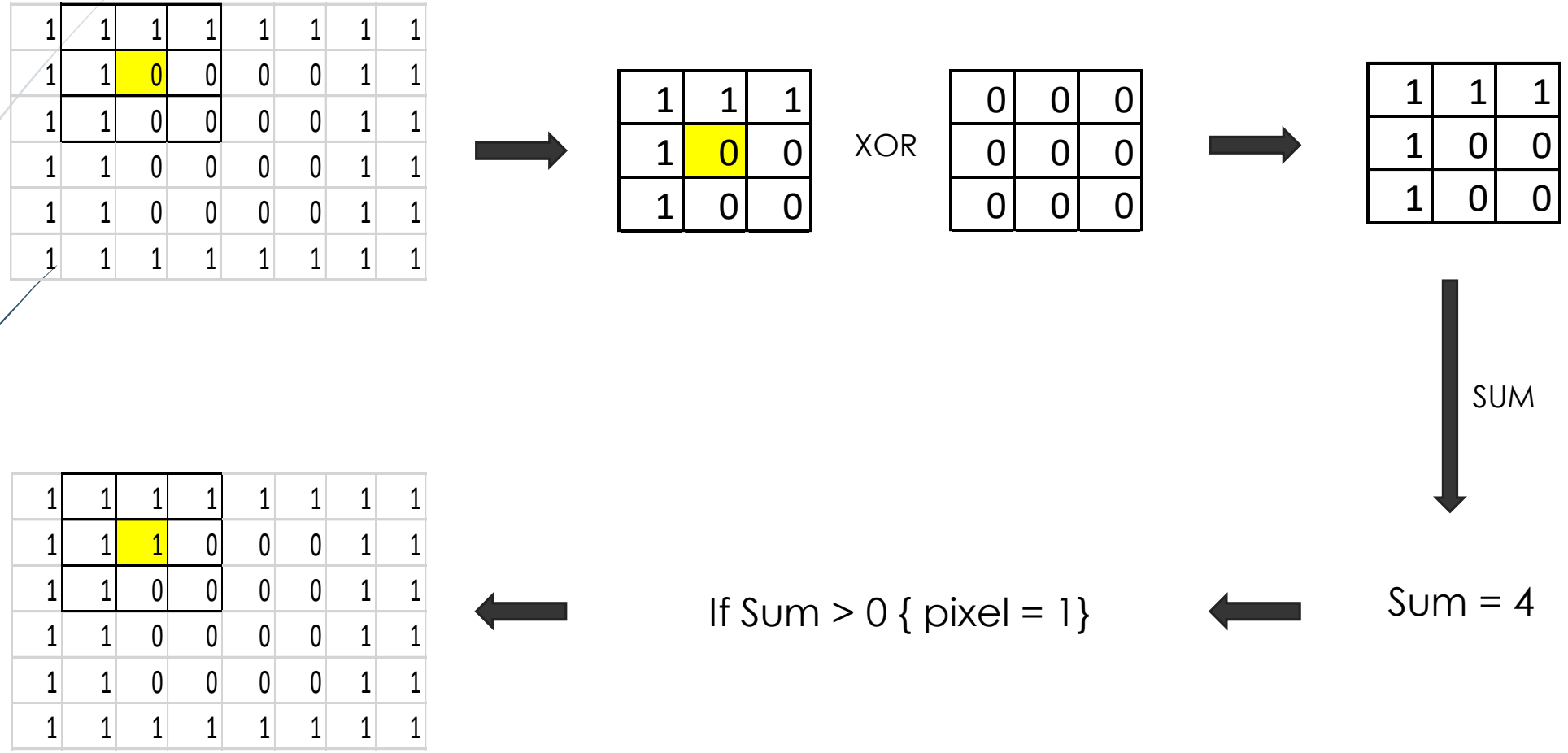
Algorithms Used

- Edge Detection Algorithm
  - Used biOps package
  - imgCanny() function does edge detection using the Canny algorithm
- Pixel Dilation
  - Created a function pixelDilation()
- Pixel Erosion
  - Created a function pixelErosion()

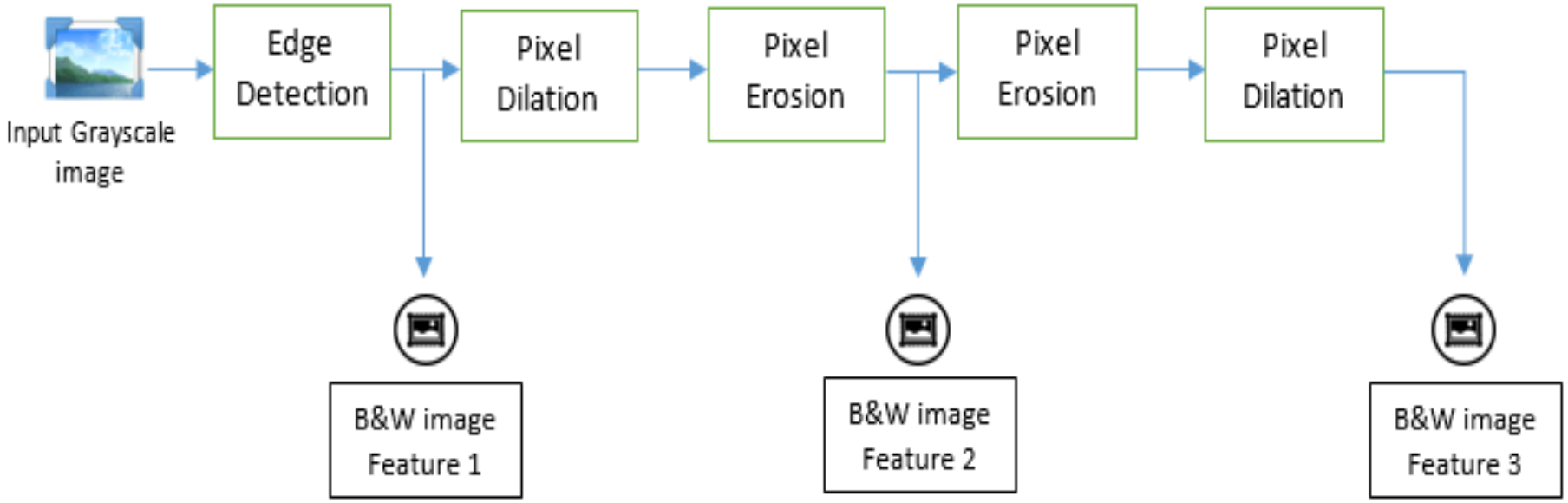
Pixel Dilation Algorithm



Pixel Erosion Algorithm



End-to-End Block Diagram



Adaptive Thresholding

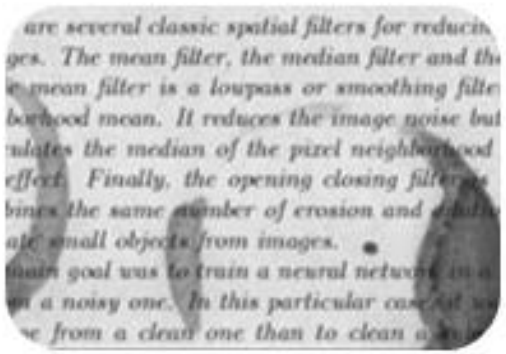
Adaptive Thresholding

- Global thresholding works if background is relatively uniform
- Partition the original image into several sub images and utilize global thresholding techniques for each sub image

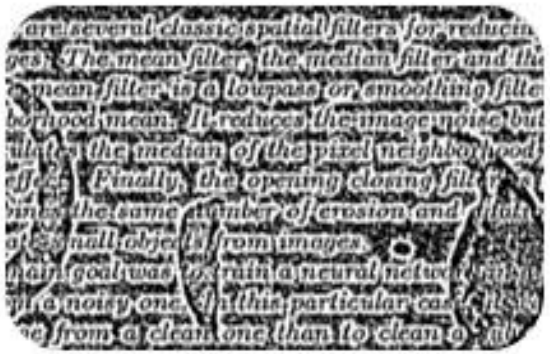
Adaptive Thresholding technique	R package (function)
Empirical Bayes thresholding	EbayesThresh (ebayesthresh)
Tree based thresholding	treethresh (treethresh)
wavelet thresholding	treethresh (threshold)
Local adaptive	EBImage (thresh)

Adaptive Thresholding

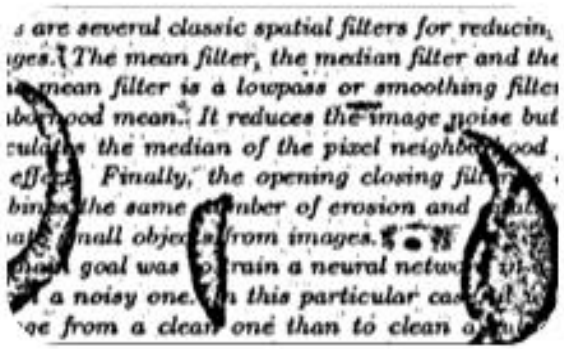
thresh(x, w=5, h=5, offset=0.01)
x: Image object,
w, h: width and height of moving rectangular window.
Offset: Thresholding object from the average value



Original



w=2 , h=2



w = 10, h =10

Median Filtering

Median Filter Implementation

- Nonlinear digital filtering technique, often used to remove noise
- An image filter that replaces a pixel with the median value of the pixels surrounding it
- This technique wipes out small features, but maintains broad features
- The resultant image is the 'background' of the image

$$\frac{MED + IAN}{2}$$

2-D Median Filter Algorithm

- Consider the following matrix $\begin{bmatrix} 5 & 4 & 8 \\ 2 & 1 & 9 \\ 13 & 3 & 11 \end{bmatrix}$

```
img = matrix(c(5,2,13,4,1,3,8,9,11),nrow=3,ncol = 3)
img
```

```
> img
     [,1] [,2] [,3]
[1,]    5    4    8
[2,]    2    1    9
[3,]   13    3   11
```

- Create an empty output matrix of the same size as that of the input matrix (in this case, a 3*3 matrix) as follows

```
#Create an empty output matrix of the same size as the input image
Y = matrix(0,nrow(img),ncol(img))
Y
```

```
> Y
     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0
>
```

2-D Median Filter Algorithm

- Pad the input matrix with zeros on all sides (based on the median filter width)

```
k = 3
n = floor(k/2)
#Modify the input image matrix by padding 0s outside the input image matrix
#to make it size having an additional n rows and n columns
img_modify = matrix(0,nrow(img)+2*n,ncol(img)+2*n)
img_modify
```

```
> img_modify
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
[4,]    0    0    0    0    0
[5,]    0    0    0    0    0
```

- Place the input matrix (img), in the center of this newly created padded matrix as follows

```
#Copy the Original matrix to the zero/padded matrix
row_seq = seq(nrow(img))
col_seq = seq(ncol(img))

for (x in row_seq)
{
  for (y in col_seq)
  {
    img_modify[x+n,y+n] = img[x,y]
  }
}
img_modify
```

```
> img_modify
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    5    4    8    0
[3,]    0    2    1    9    0
[4,]    0   13    3   11    0
[5,]    0    0    0    0    0
>
```

2-D Median Filter Algorithm

- Consider a window of size 3 by 3 from the input matrix
- window = $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 5 & 4 \\ 0 & 2 & 1 \end{bmatrix}$
- Sort of the window = $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 2 & 4 & 5 \end{bmatrix}$ (Median is 0)
- Extend the algorithm to iterate through all combinations of a 3*3 window and record the output
- Final Output Y

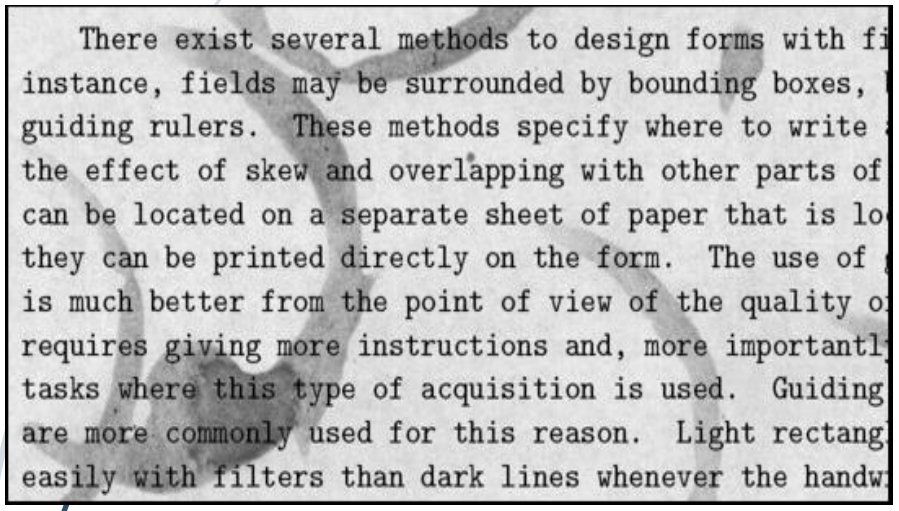
```
> img_modify
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    5    4    8    0
[3,]    0    2    1    9    0
[4,]    0   13    3   11    0
[5,]    0    0    0    0    0
>
```

```
> Y
     [,1] [,2] [,3]
[1,]    0    2    0
[2,]    2    5    0
[3,]    0    0    0
>
```

```
for (i in seq(nrow(img_modify)-k))
{
  for (j in seq(ncol(img_modify)-k))
  {
    window = matrix(0,k*k,1)
    c = 1
    for (x in seq(k))
    {
      for (y in seq(k))
      {
        window[c] = img_modify[i+x-1,j+y-1]
        c = c + 1
      }
    }

    med = sort(window)

    Y[i,j] = med[((k*k)+1)/2]
  }
}
```

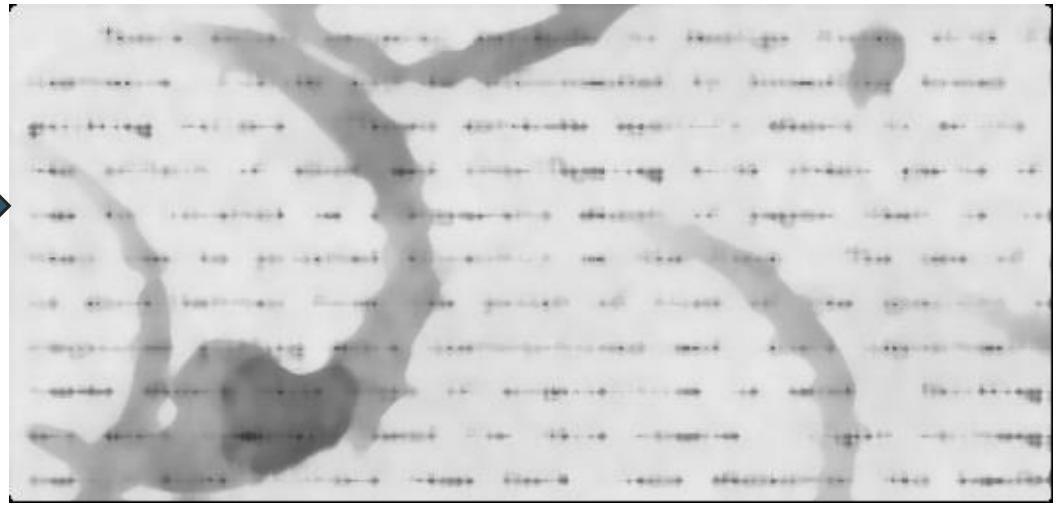
Median Filter – Image Processing

Input Image (with noise)
median



Median Filter

Background Image (Applying the filter – window size of 9)



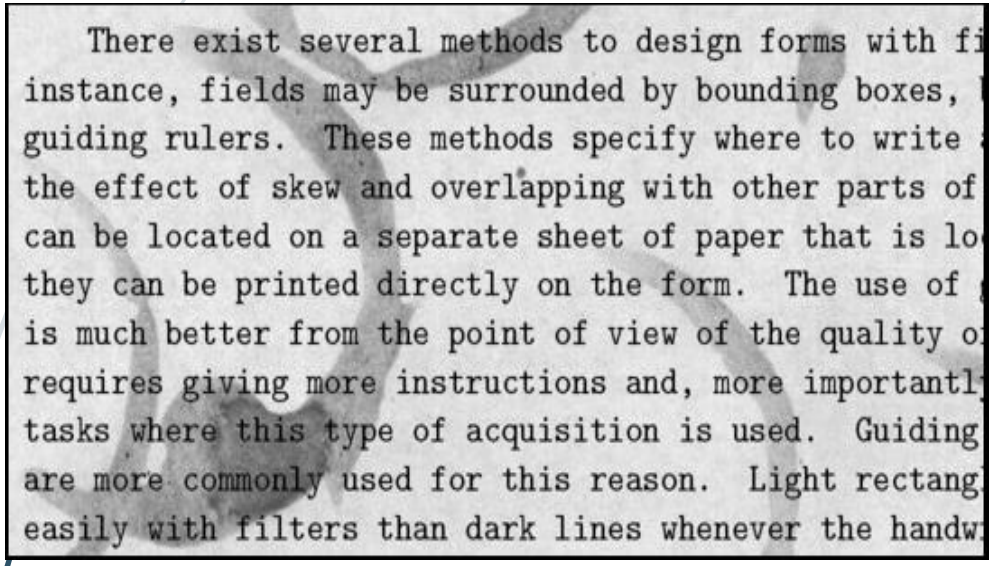
Extract the foreground from the background

- Subtract the background from the original image and normalize the output image to have values between [0 1]

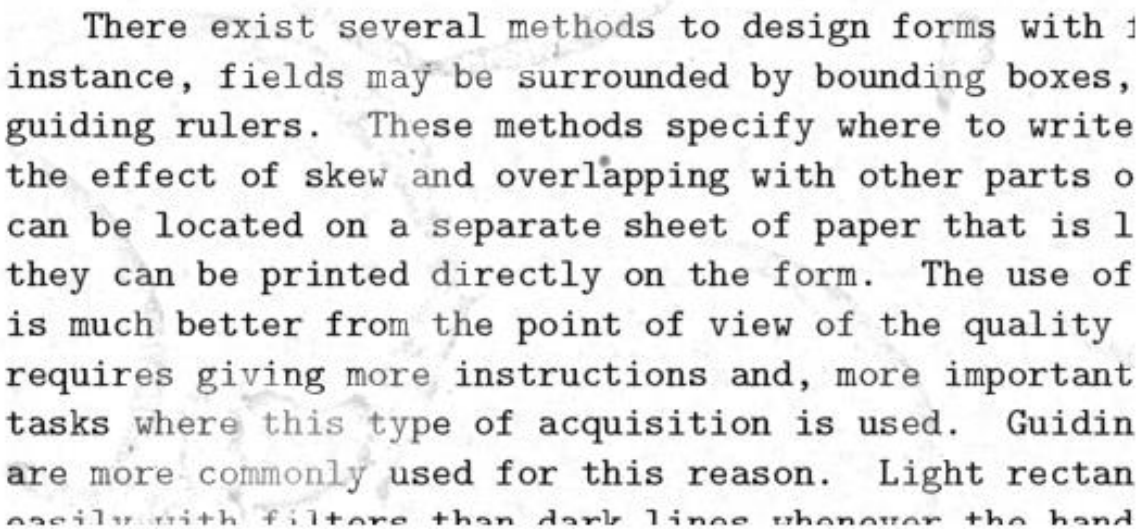
```
foreground = img - background
#In this case, we know that the writing is always darker than the background,
#so, our foreground should only show pixels that are darker than background
foreground[foreground > 0] = 0

#Normalizing the final results (pixels) to lie between 0-1
m1 = min(foreground)
m2 = max(foreground)

foreground = (foreground - m1) / (m2 - m1)
```

Processed Image

Input Image (with noise)

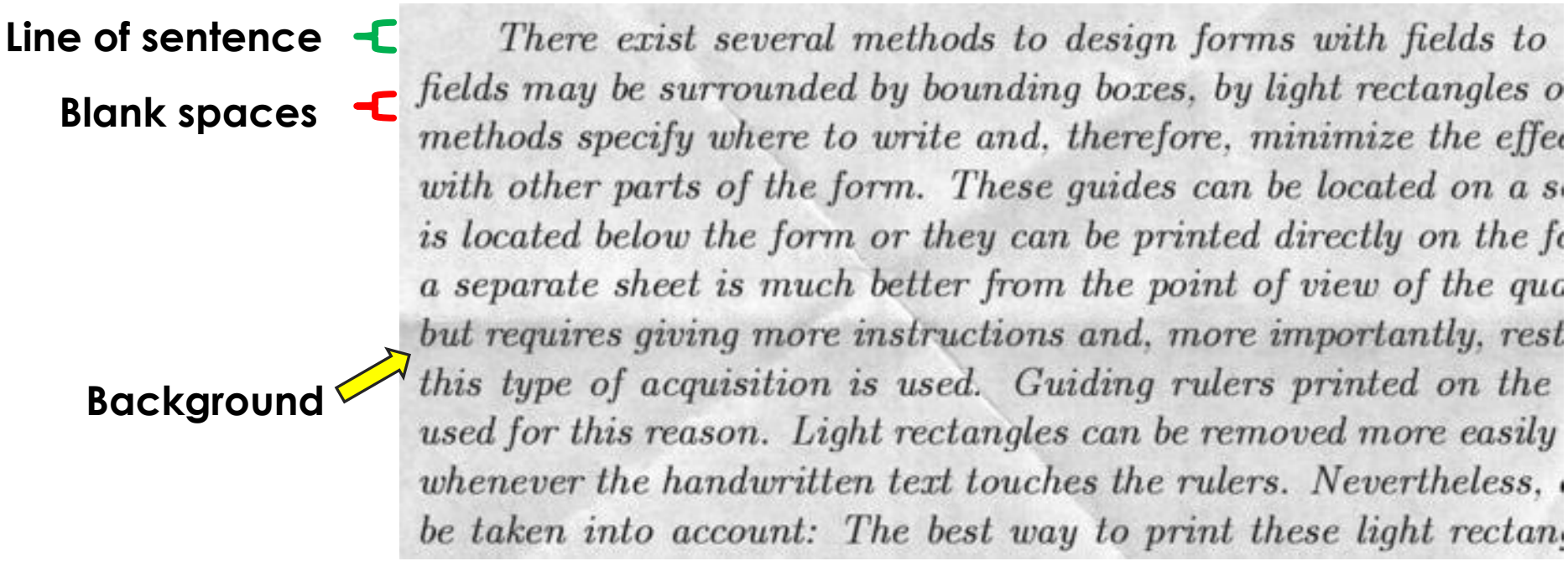


Cleansed Image

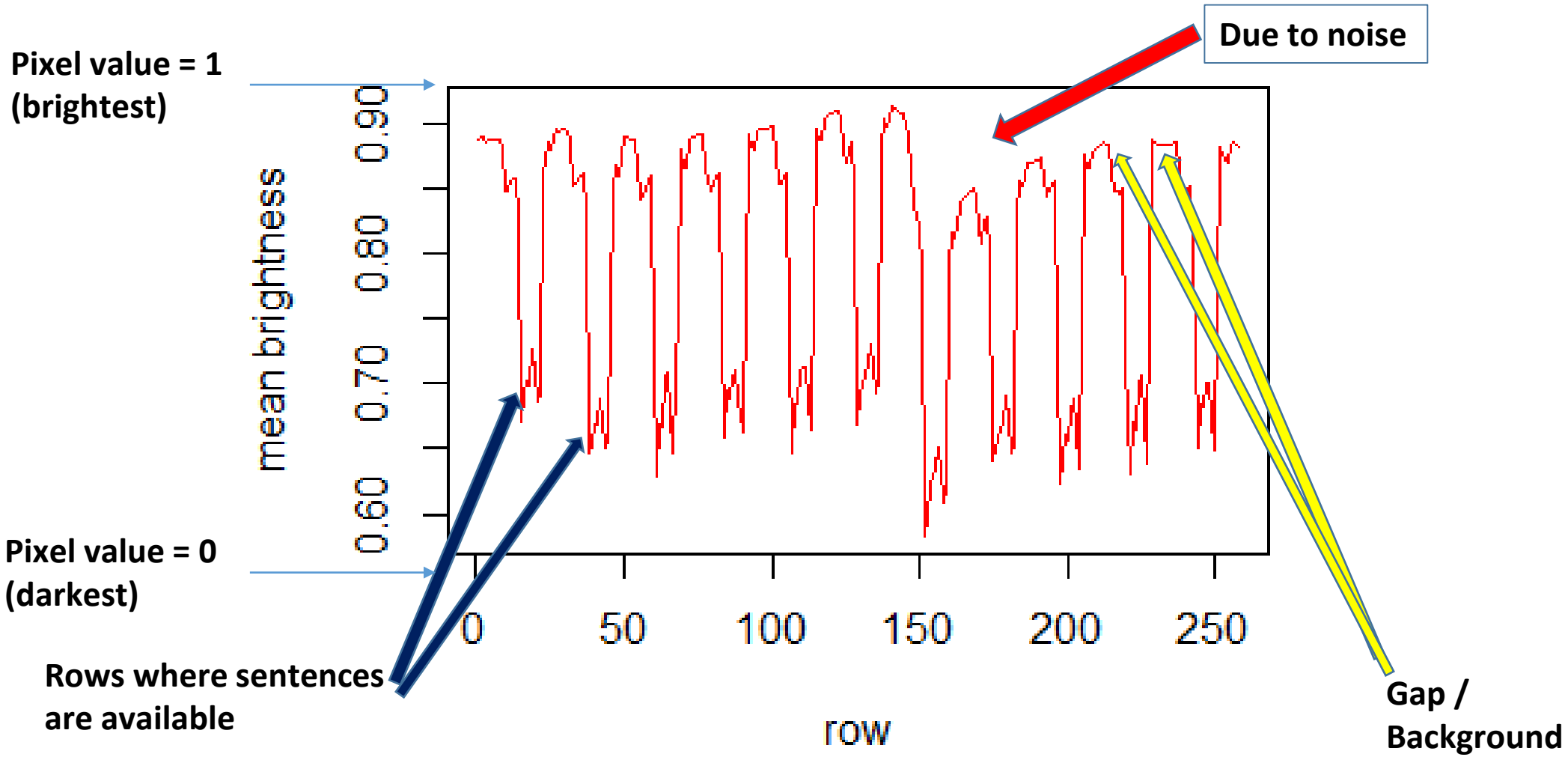


Identifying gaps between lines of text

The image with noise

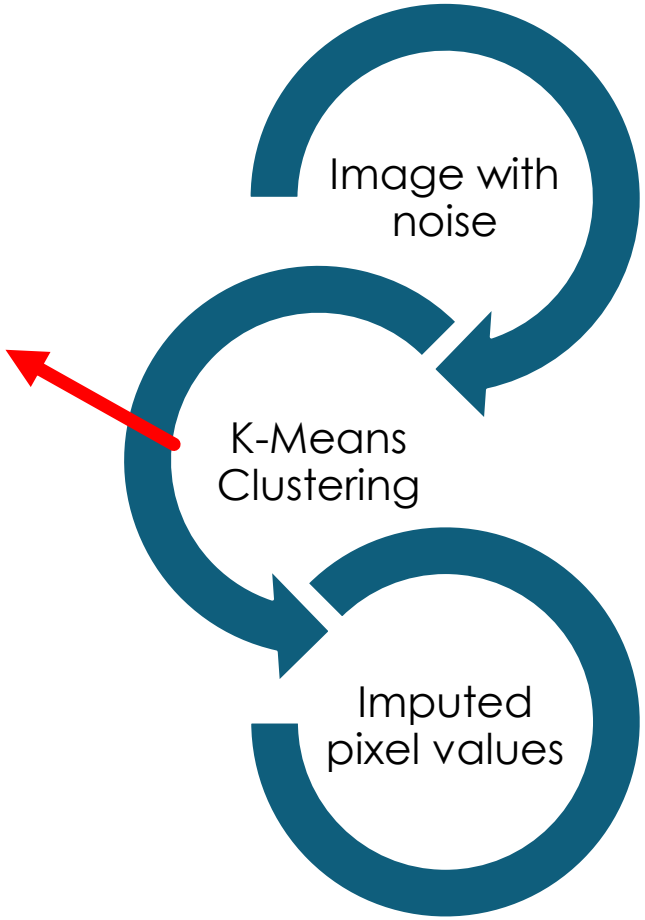


Identifying mean pixel value for each row



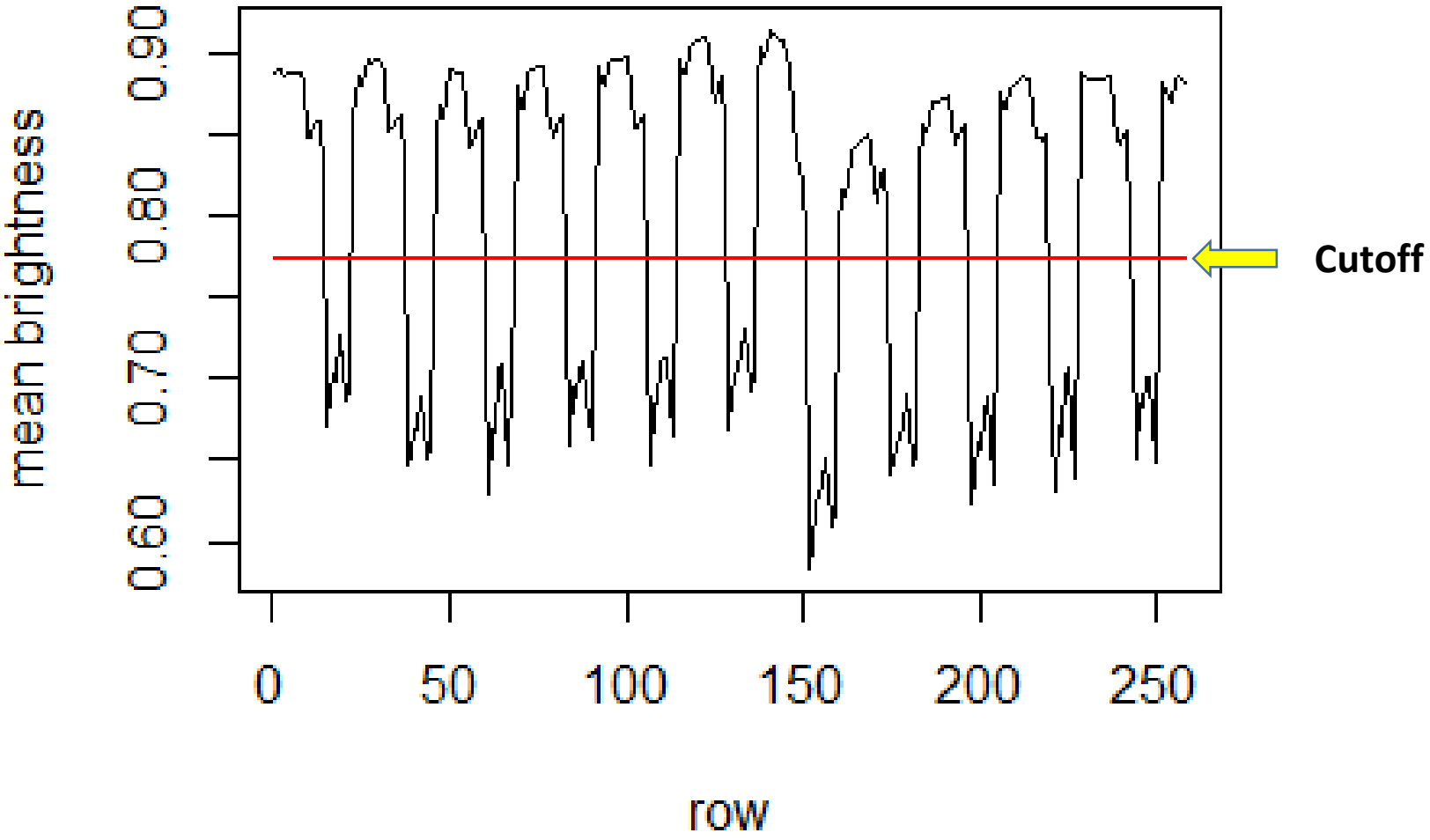
Approach

K-Means clustering:
- Cluster size: 2
- Lower cluster: Lower pixel values
- Upper cluster: Higher pixel values

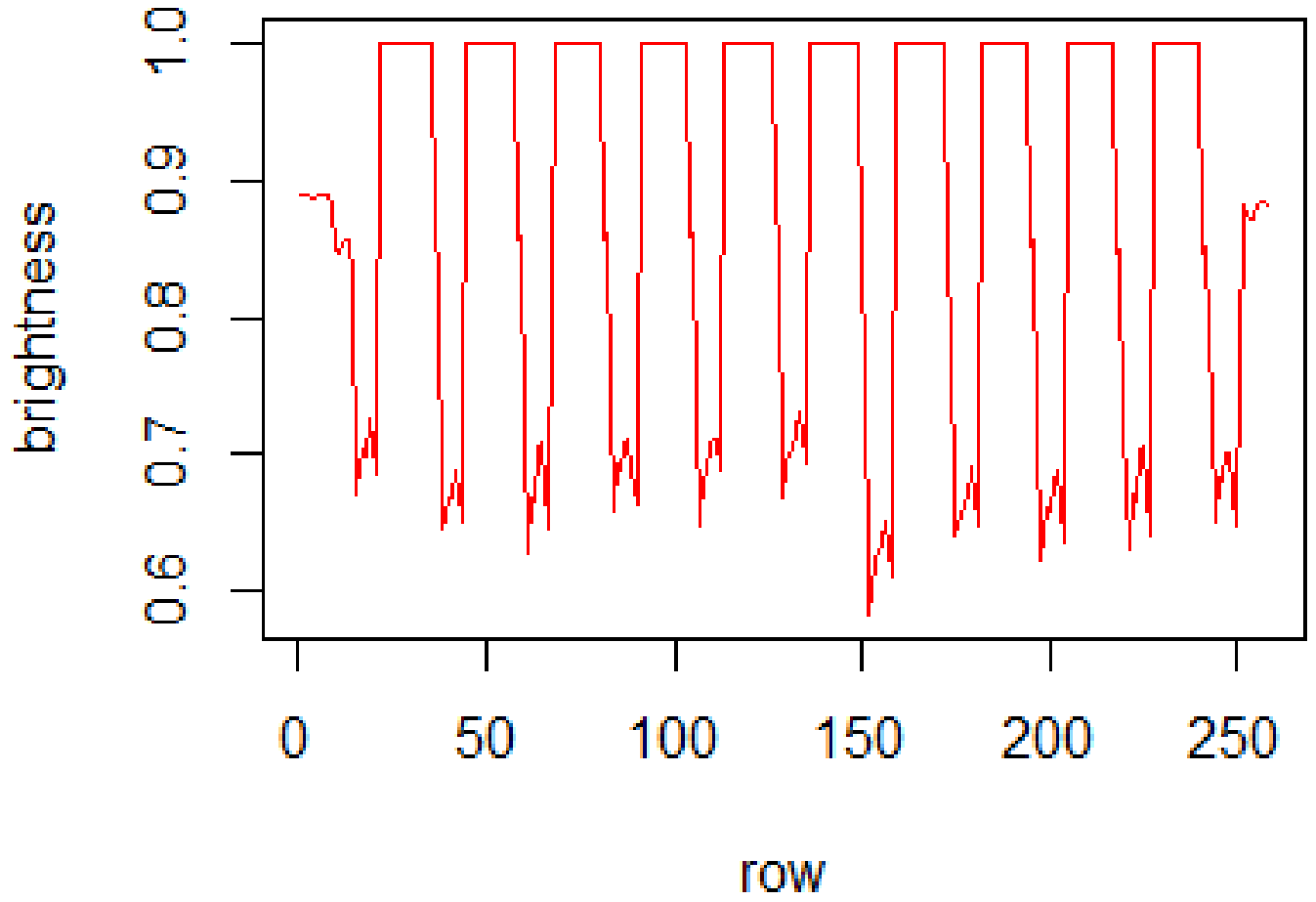


Identifying cutoff

Cutoff = **Mean of (Highest value of lower cluster) AND (Lowest value of upper cluster)**

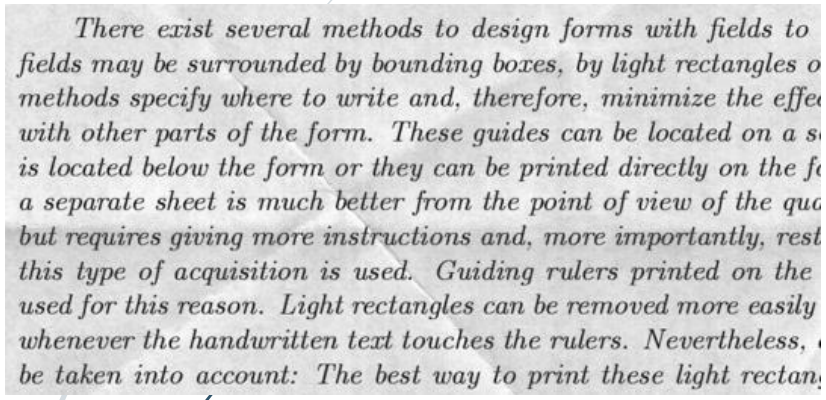


Imputing background

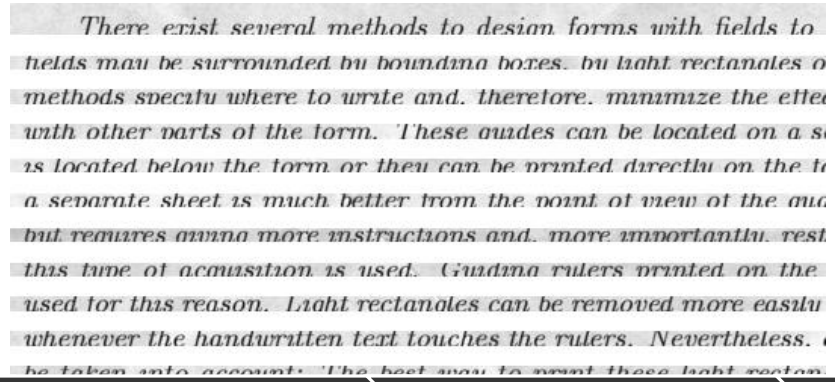


Plotting the skeleton – determining the gaps

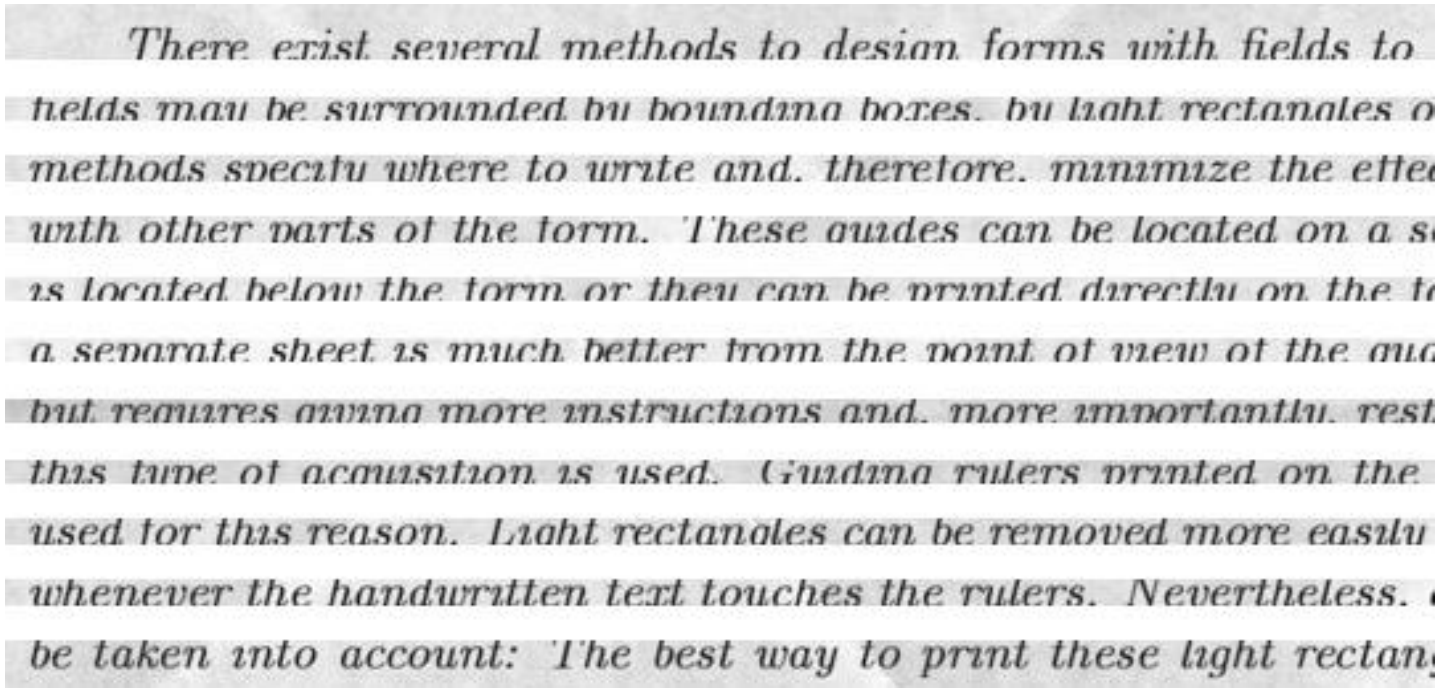
Final image


There exist several methods to design forms with fields to
fields may be surrounded by bounding boxes, by light rectangles o
methods specify where to write and, therefore, minimize the effe
with other parts of the form. These guides can be located on a s
is located below the form or they can be printed directly on the f
a separate sheet is much better from the point of view of the qu
but requires giving more instructions and, more importantly, rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten text touches the rulers. Nevertheless,
be taken into account: The best way to print these light rectan

+


There exist several methods to design forms with fields to
fields may be surrounded by bounding boxes, by light rectangles o
methods specify where to write and, therefore, minimize the effe
with other parts of the form. These guides can be located on a s
is located below the form or they can be printed directly on the f
a separate sheet is much better from the point of view of the qu
but requires giving more instructions and, more importantly, rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten text touches the rulers. Nevertheless,

Final Image

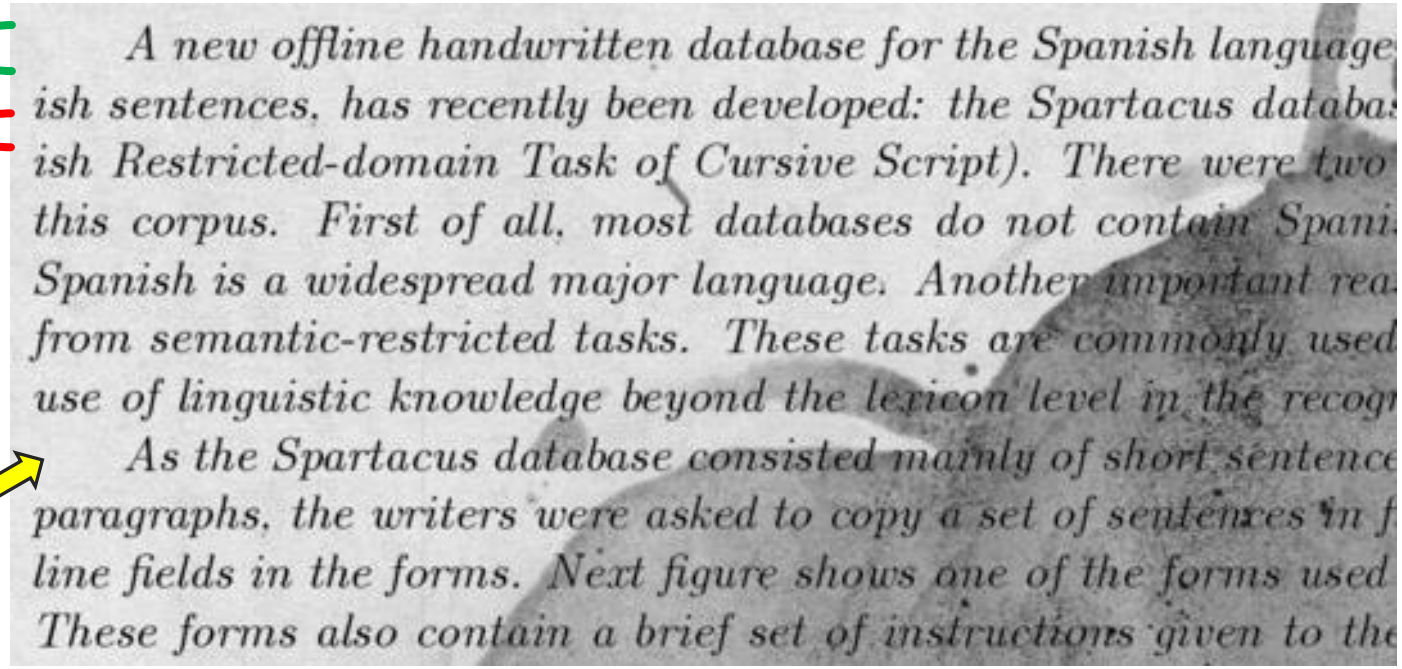

There exist several methods to design forms with fields to
fields may be surrounded by bounding boxes, by light rectangles o
methods specify where to write and, therefore, minimize the effe
with other parts of the form. These guides can be located on a s
is located below the form or they can be printed directly on the t
a separate sheet is much better from the point of view of the qua
but requires giving more instructions and, more importantly, rest
this type of acquisition is used. Guiding rulers printed on the
used for this reason. Light rectangles can be removed more easily
whenever the handwritten text touches the rulers. Nevertheless,
be taken into account: The best way to print these light rectan

The image with noise

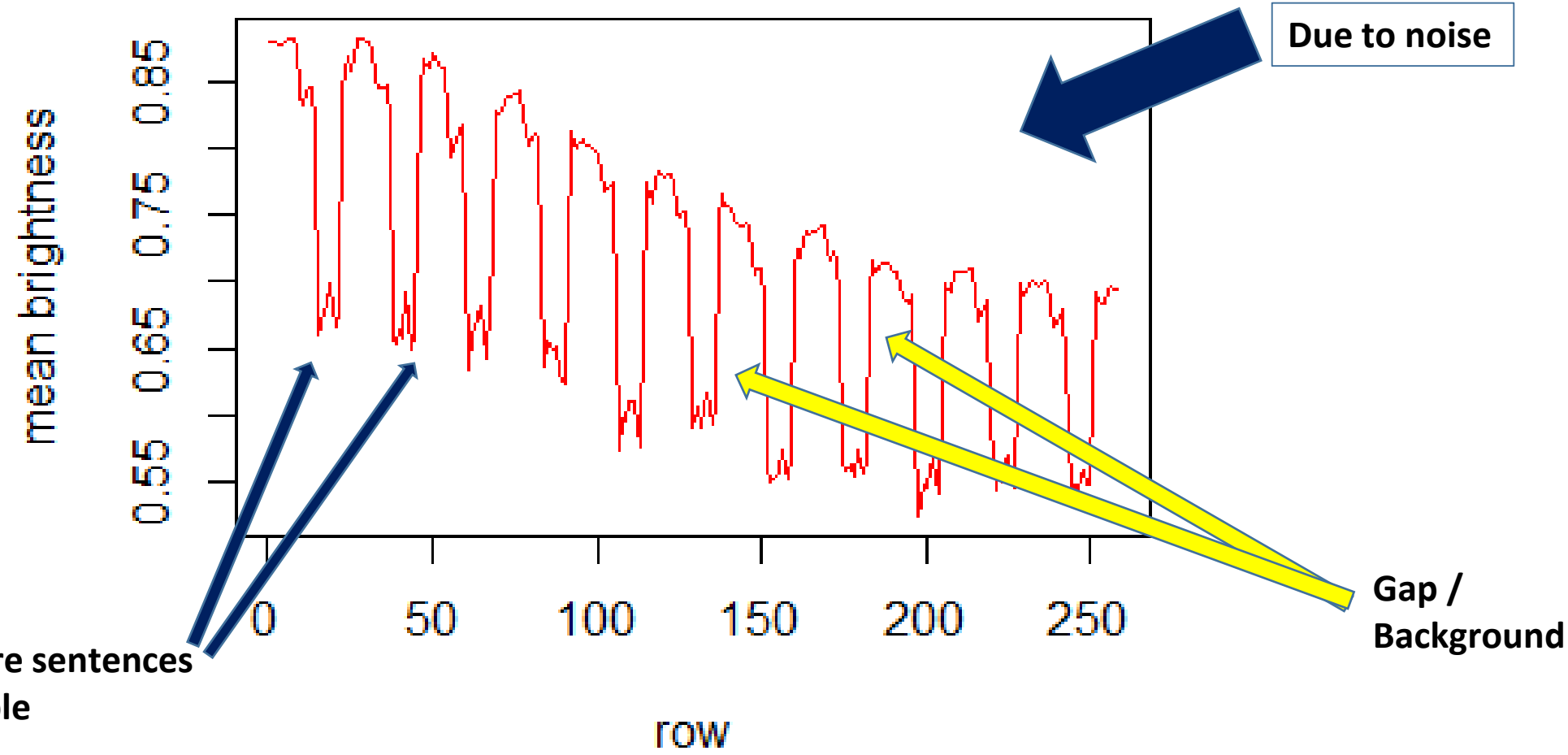
Line of sentence

Blank spaces

Background

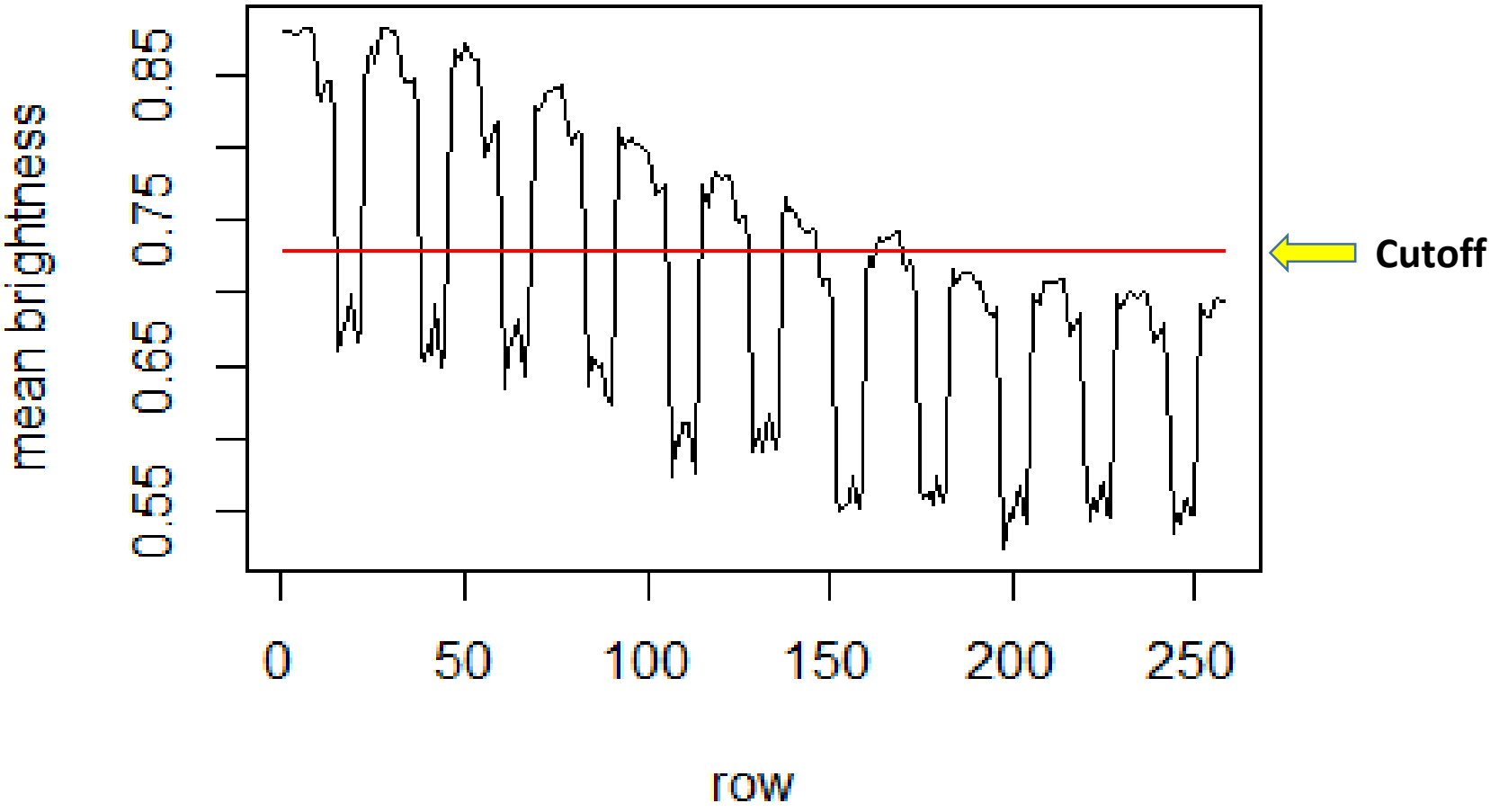


Identifying mean pixel value for each row

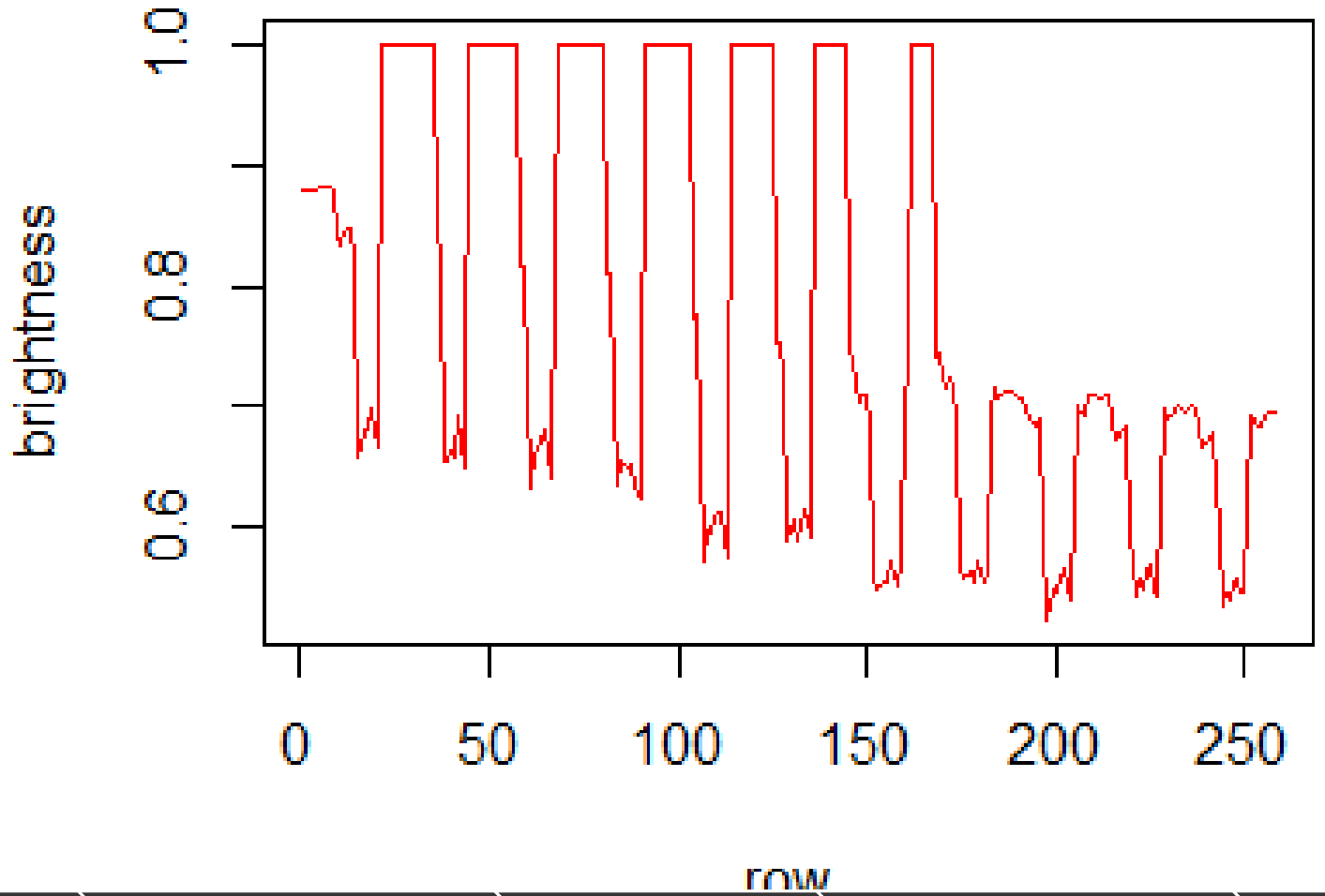


Identifying cutoff

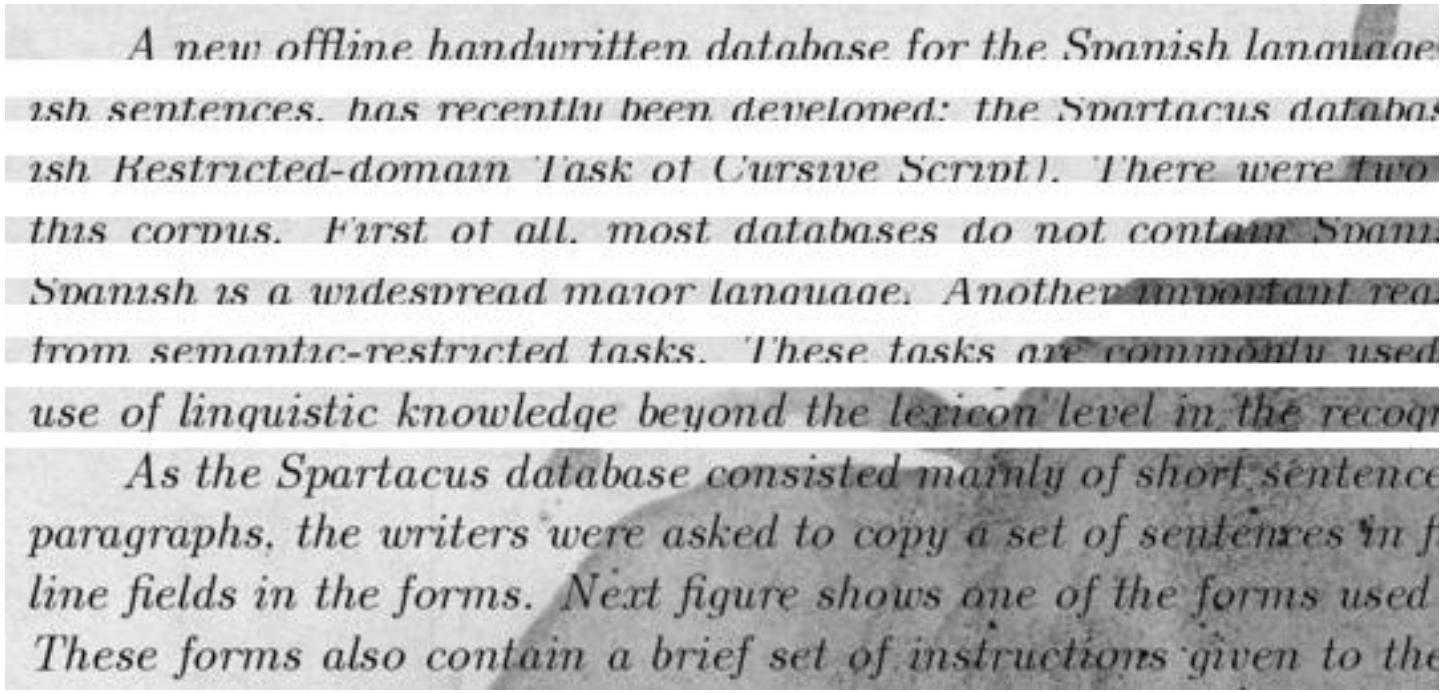
Cutoff = **Mean of (Highest value of lower cluster) AND (Lowest value of upper cluster)**



Imputing background



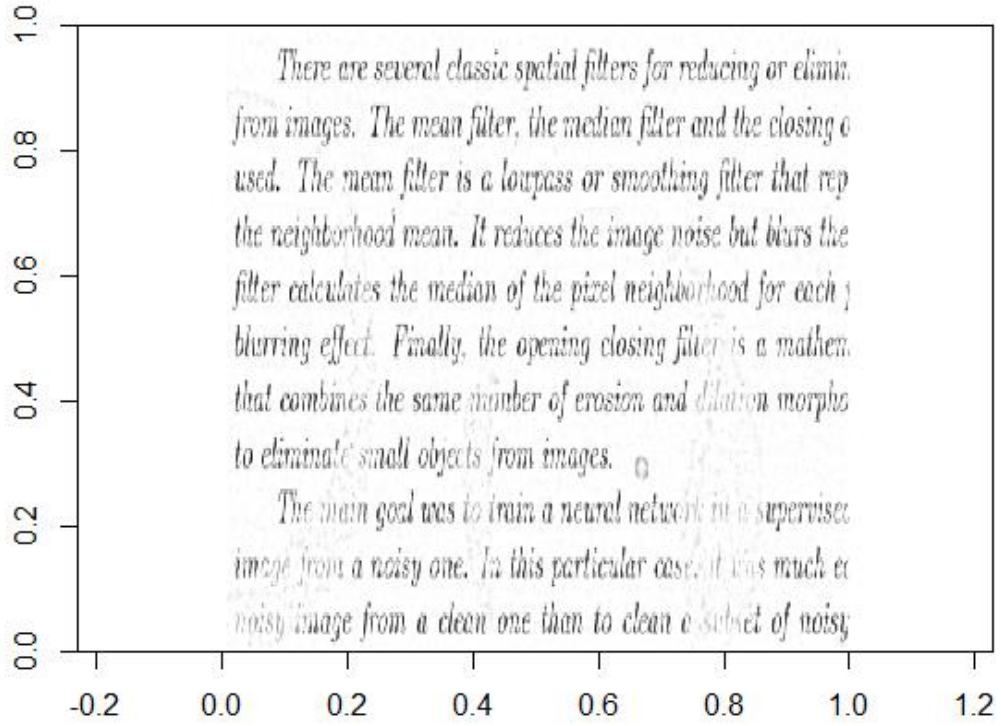
Final Image


A new offline handwritten database for the Spanish language
ish sentences, has recently been developed: the Spartacus databas
ish Restricted-domain Task of Cursive Script). There were two
this corpus. First of all, most databases do not contain Spani
Spanish is a widespread major language. Another important rea
from semantic-restricted tasks. These tasks are commonly used
use of linguistic knowledge beyond the lexicon level in the recogn
As the Spartacus database consisted mainly of short sentence
paragraphs, the writers were asked to copy a set of sentences in f
line fields in the forms. Next figure shows one of the forms used
These forms also contain a brief set of instructions given to the

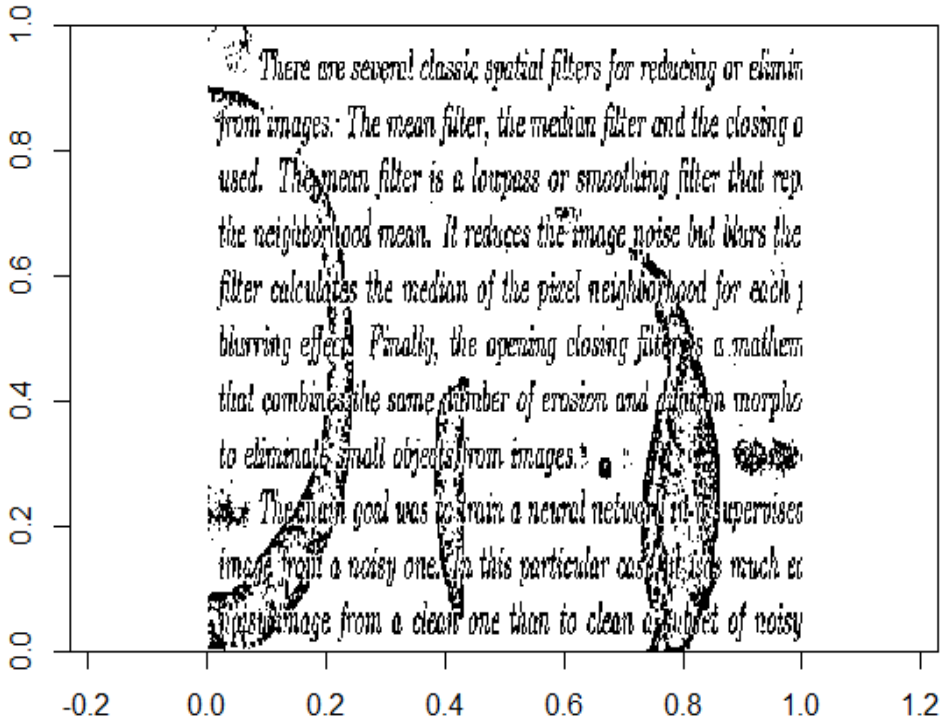
Ensembling the results of different techniques - XGBoost

Individual Image Cleaning Algorithms Might Not Produce Optimum Results

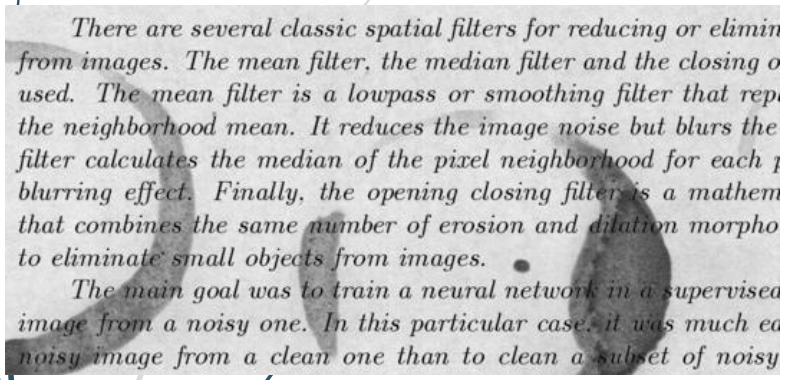
Output from a Median Filter



Output from a Adaptive Thresholding



Process



There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing o used. The mean filter is a lowpass or smoothing filter that rep the neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each blurring effect. Finally, the opening closing filter is a mathem that combines the same number of erosion and dilation morpho to eliminate small objects from images.

The main goal was to train a neural network in a supervise image from a noisy one. In this particular case, it was much e noisy image from a clean one than to clean a subset of noisy

Median Filter

Edge Detection

Adaptive Thresholding

Featurization

XGBoost

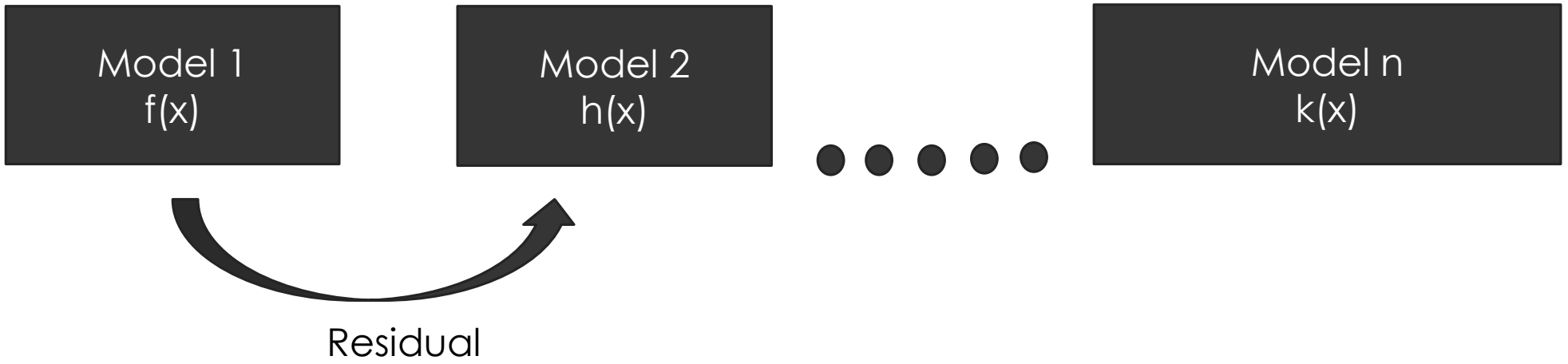
There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing o used. The mean filter is a lowpass or smoothing filter that rep the neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each blurring effect. Finally, the opening closing filter is a mathem that combines the same number of erosion and dilation morpho to eliminate small objects from images.

The main goal was to train a neural network in a supervise image from a noisy one. In this particular case, it was much e noisy image from a clean one than to clean a subset of noisy

Input | Predictors | Model | Target

Gradient Boosting

- A sequence of weak learners are used to produce more powerful predictions
- Later models focus on learning errors better



Algorithm

For each image in the data, do:

- Read dirty image & and convert it into a one column matrix
- Read clean image & convert it into a one column matrix
- Apply median filtering & convert the cleaned image into a one column matrix
- Apply feature detection filter & convert the cleaned image into a one column matrix
- Column bind all these features
- Append it to the dataframe containing these features for the previous images

	Original	Featurization	Median Filtering	Target
1	0.8941176	0.8941176	1.0000000	1
2	0.8823529	0.8823529	1.0000000	1
3	0.8980392	0.8980392	1.0000000	1
4	0.9137255	0.9137255	1.0000000	1
5	0.8941176	0.8941176	1.0000000	1
6	0.9215686	0.9215686	1.0000000	1
7	0.9098039	0.9098039	1.0000000	1
8	0.8980392	0.8980392	1.0000000	1
9	0.9098039	0.9098039	1.0000000	1
10	0.8941176	0.8941176	1.0000000	1
11	0.9137255	0.9137255	1.0000000	1
12	0.9215686	0.9215686	1.0000000	1
13	0.9294118	0.9294118	1.0000000	1
14	0.9215686	0.9215686	1.0000000	1
15	0.9176471	0.9176471	1.0000000	1
16	0.9098039	0.9098039	1.0000000	1
17	0.9176471	0.9176471	1.0000000	1
18	0.9176471	0.9176471	1.0000000	1
19	0.9137255	0.9137255	1.0000000	1
20	0.9176471	0.9176471	1.0000000	1

Algorithm contd...

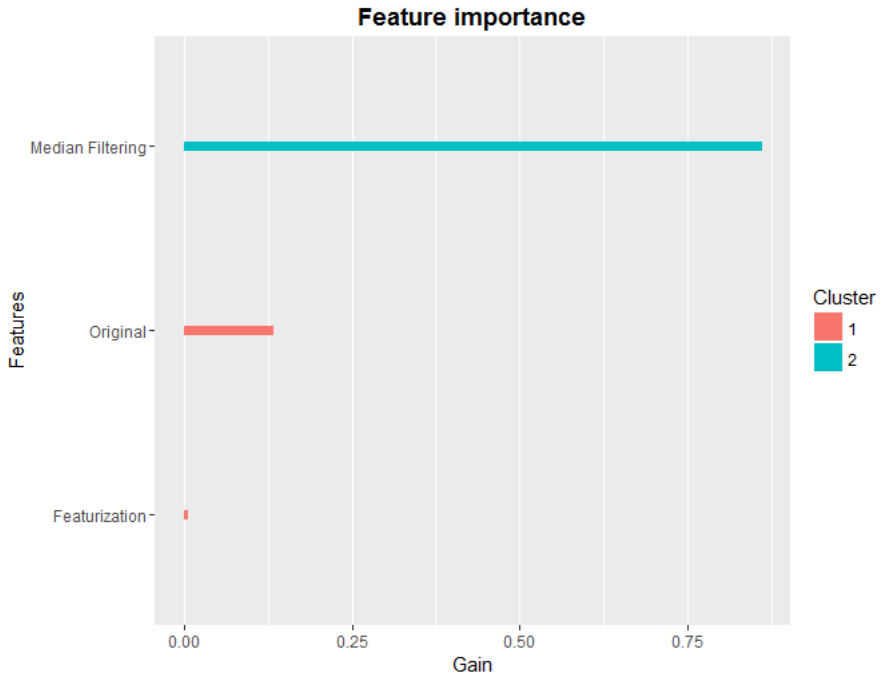
For each image in the data, do:

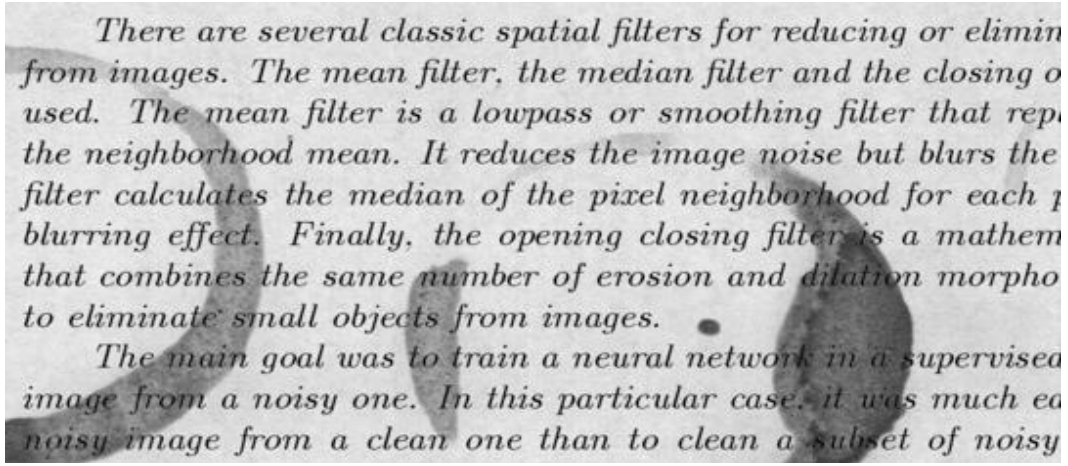
- Take a randomly selected sample from the data(250000 rows were selected)
- Convert the data into a dense matrix with label being the cleaned image
- Using cross validation, determine the optimum number of rounds for the xgboost model
- Create the model with RMSE as the evaluation parameter
- Predict for test images using this model
- Convert the result into a matrix of the same dimensions as the original image

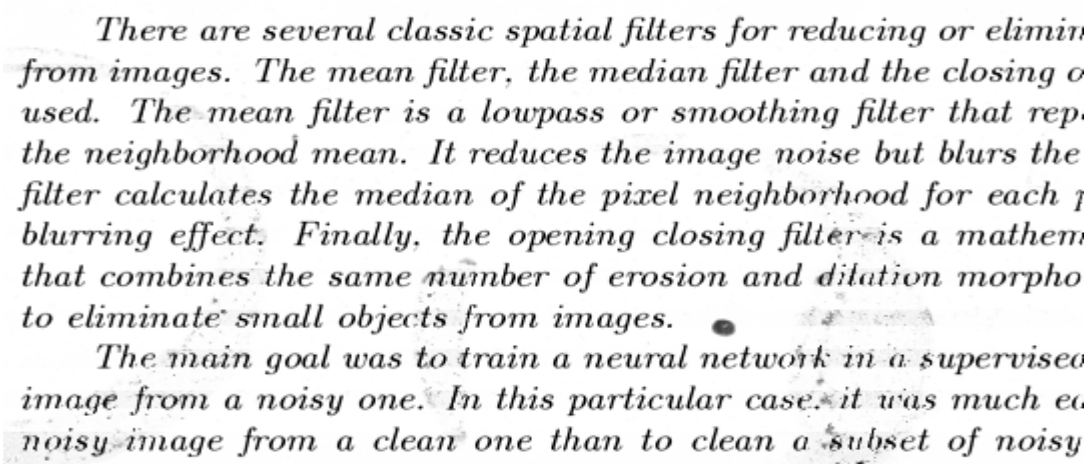
	pred
1	0.9979517
2	0.9981245
3	0.9977194
4	0.9977194
5	0.9974357
6	0.9974357
7	0.9974357
8	0.9979517
9	0.9981074
10	0.9981245
11	0.9981074
12	0.9982852
13	0.9981245
14	0.9981245
15	0.9981074

Conclusion

- RMSE:
  - Training: 0.029384
  - Test: 0.032019




There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
blurring effect. Finally, the opening closing filter is a mathem
that combines the same number of erosion and dilation morpho
to eliminate small objects from images.
The main goal was to train a neural network in a supervise
image from a noisy one. In this particular case, it was much e
noisy image from a clean one than to clean a subset of noisy


There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
blurring effect. Finally, the opening closing filter is a mathem
that combines the same number of erosion and dilation morpho
to eliminate small objects from images.
The main goal was to train a neural network in a supervise
image from a noisy one. In this particular case, it was much e
noisy image from a clean one than to clean a subset of noisy