

# Análise de texto de fontes desestruturadas e Web

## Aula 03

Maciel Calebe Vidal

[macielcv@insper.edu.br](mailto:macielcv@insper.edu.br)

# Motivação

- Era **digital** vs **papel**
  - Governos
  - Empresas
  - Universidades

# Motivação

- Provedores de identidade
- Extrair informações relevantes de imagens



# Objetivo

Digitalizar

Automaticamente

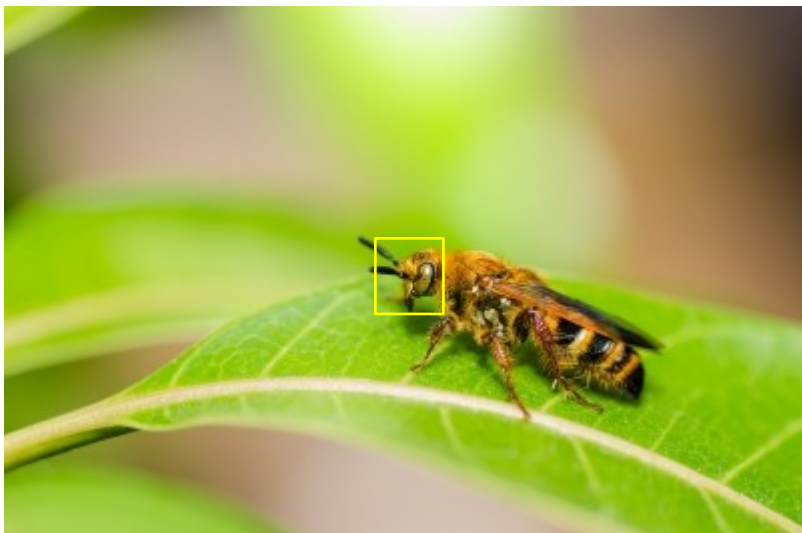
Extrair informações  
de interesse

Armazenar

Analisar

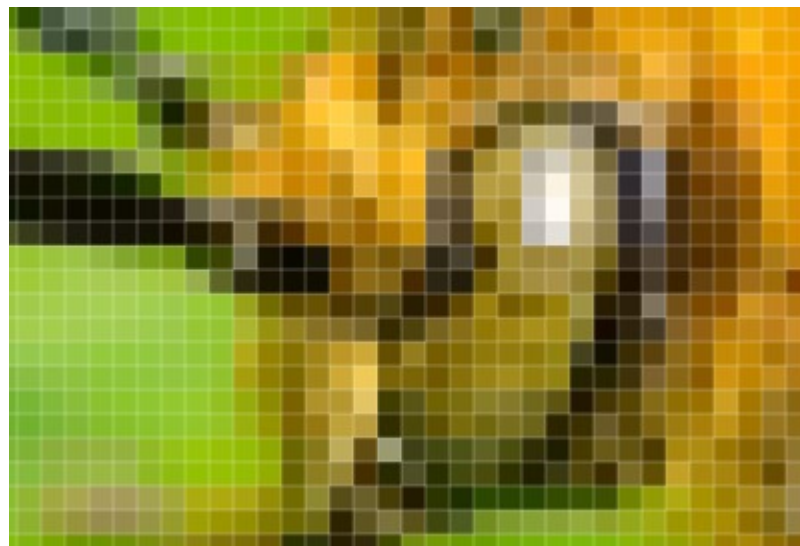
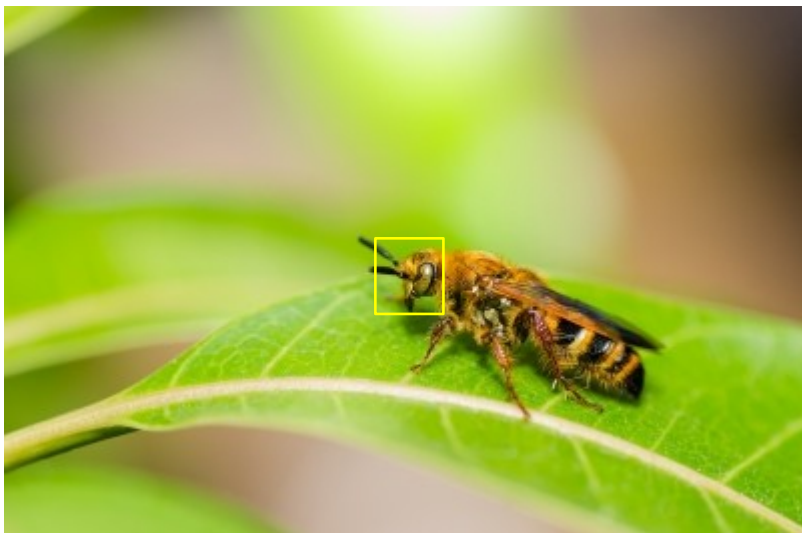
Obter valor

# Como imagens funcionam?



<https://www.desenhoonline.com/site/saiba-como-e-formada-uma-imagem-em-bitmap-raster/>

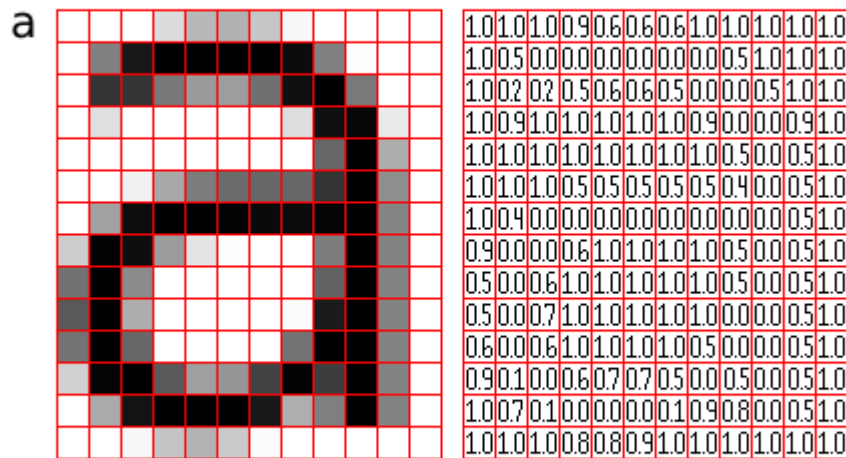
# Como imagens funcionam?



<https://www.desenhoonline.com/site/saiba-como-e-formada-uma-imagem-em-bitmap-raster/>

# Como imagens funcionam

Figure 1.3. Raster image




A rasterized form of the letter 'a' magnified 16 times using pixel doubling

[http://pippin.gimp.org/image\\_processing/chap\\_dir.html](http://pippin.gimp.org/image_processing/chap_dir.html)

# Padrões de cores


- Acesse [https://www.w3schools.com/colors/colors\\_picker.asp](https://www.w3schools.com/colors/colors_picker.asp)

Pick a Color:

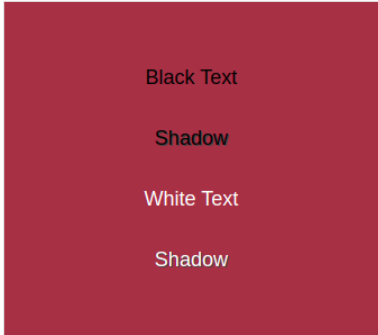


Or Enter a Color:

Or Use HTML5:



Selected Color:



Black Text

Shadow

White Text

Shadow

**#964545**  
rgb(150, 69, 69)  
hsl(0, 37%, 43%)

Lighter / Darker:

100%	#ffffff
95%	#f7eeee
90%	#efdcdc
85%	#e7cbcb
80%	#dfb9b9
75%	#d7a8a8
70%	#cf9696
65%	#c78585
60%	#bf7373
55%	#b76262
50%	#af5050
45%	#9d4848
43%	<b>#964545</b>
40%	#8c4040
35%	#7a3838
30%	#693030
25%	#572828
20%	#462020
15%	#341818
10%	#231010
5%	#110808
0%	#000000



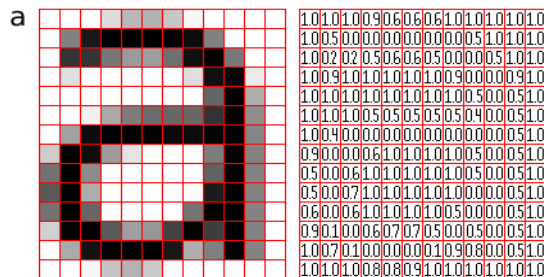
# Recuperação de informação

Busque por tweets que falem de Bitcoin

Vs

Extraia o CPF da imagem

Figure 1.3. Raster image



A rasterized form of the letter 'a' magnified 16 times using pixel doubling


[http://pippin.gimp.org/image\\_processing/chap\\_dir.html](http://pippin.gimp.org/image_processing/chap_dir.html)

# Optical Character Recognition (OCR)

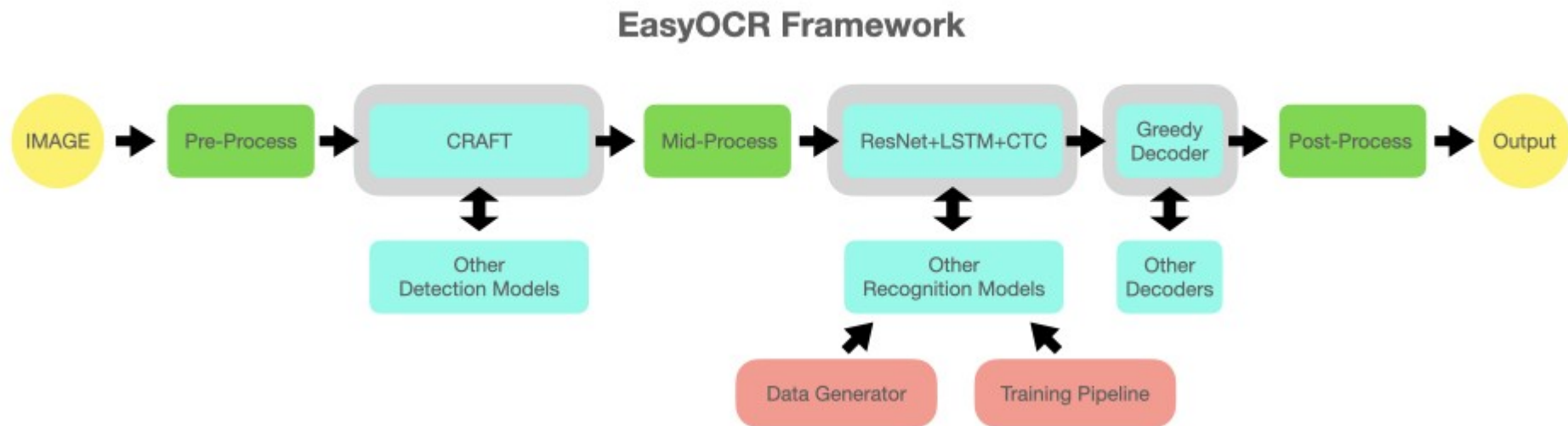
- O que é?
  - Digitalização de documentos
  - Conversão de **bitmap** em **texto** editável
- Como fazer?

# EasyOCR

- OCR de uso facilitado
- 80+ linguagens
- Deep learning PyTorch

image	
result	<p>'Reduce your risk of coronavirus infection:', 'Clean hands with soap and water', 'or alcohol based hand rub', 'Cover nose and mouth when coughing and', 'sneezing with tissue or flexed elbow', 'Avoid close contact with anyone with', 'cold or flu like symptoms', 'Thoroughly cook meat and eggs', 'No unprotected contact with live wild', 'or farm animals', 'World Health', 'organization'</p> <p>'เส้นทางลัด', 'เพชรบุรี'</p> <p>'du l"', 'Mairie', 'Palais du', 'LOUVRE', 'LES ARTS DÉCORATIFS', 'Musée du LOUVRE', 'Théâtre', 'du PALAIS-ROYAL'</p>

# EasyOCR



# OCR – Aplicações práticas

- Diminuir falhas
- Ganho de produtividade
- Melhoria de processos

# OCR – Aplicações práticas

- Converter documento inteiro
  - Livro
- Extrair informações específicas de um documento
  - NFe
- Processar grandes volumes de documentos
  - Seguradora

 **BRASIL**

**CORONAVÍRUS (COVID-19)**

Simplifique!

Participe

Acesso à Informação

Legislação

Canais 

Conheça a NF-e · Serviços · Legislação · Documentos · Downloads · Área Restrita · Documentos e outros

**NOTA FISCAL ELETRÔNICA**

 **Serviços** · **Legislação** · **Documentos** · **Downloads**

Você está aqui: [Página Principal](#) > [Serviços](#) > [Consultar NF-e](#)

**Consultar NF-e**

Chave de Acesso da NF-e

☐ Não sou um robô   
reCAPTCHA  
Privacidade · Termos

Continuar

Limpar

**Estatísticas da NF-e**  
**NF-e Autorizadas**  
26,614 bilhões  
**Número de Emissores**  
1,818 milhões  
[... saiba mais](#)

**Buscar**

 **Área Restrita**

 **Central de Atendimento**

 **Perguntas Frequentes**

 **Portais e Secretarias**

**Portais Estaduais da NF-e**  

Selecione ▼

**Secretarias de Fazenda**  

Selecione ▼



# Dúvidas?

Obrigado pela participação!