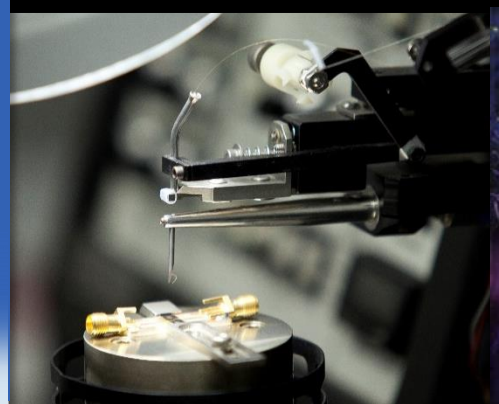
Centro Brasileiro de Pesquisas Físicas

Métodos para Análise de grande volume de dados e Astroinformática

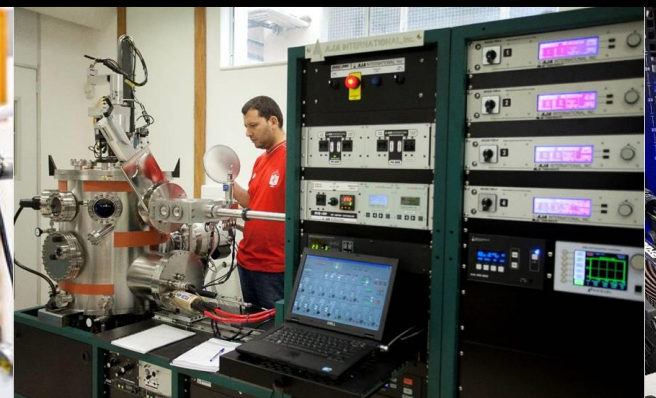
Clécio Roque De Bom – debom@cbpf.br

clearnightsrthebest.com

CBPF








DataFrame Operations

```
persons = pd.read_csv("persons.csv")
persons
```

	age	height	name	sex
0	23	156	Alice	female
1	21	181	Bob	male
2	27	176	Charlie	male
3	24	167	Eve	female
4	19	172	Frances	female
5	31	191	George	female

A Pandas DataFrame containing some persons

Adapted from: https://towardsdatascience.com/spark-vs-pandas-part-1-pandas-10d768b979f5

DataFrame Projection

```
persons[["age", "height"]]
```

	age	height
0	23	156
1	21	181
2	27	176
3	24	167
4	19	172
5	31	191

Adapted from: https://towardsdatascience.com/spark-vs-pandas-part-1-pandas-10d768b979f5

DataFrame Operations - Filtering

```
persons[persons["age"] > 21]
```

	age	height	name	sex
0	23	156	Alice	female
2	27	176	Charlie	male
3	24	167	Eve	female
5	31	191	George	female

Filtering people, which are older than 21 years.

DataFrame Operations - Join

```
addresses = pd.read_csv("addresses.csv")
addresses
```

	name	city
0	Alice	Hamburg
1	Bob	Frankfurt
2	Henry	Berlin

```
addresses = addresses.set_index("name")
addresses
```

name	city
Alice	Hamburg
Bob	Frankfurt
Henry	Berlin

DataFrame Operations - Join

```
addresses = addresses.set_index("name")
addresses
```

name	city
Alice	Hamburg
Bob	Frankfurt
Henry	Berlin

```
df = persons.join(addresses, on="name")
df
```

	age	height	name	sex	city
0	23	156	Alice	female	Hamburg
1	21	181	Bob	male	Frankfurt
2	27	176	Charlie	male	NaN
3	24	167	Eve	female	NaN
4	19	172	Frances	female	NaN
5	31	191	George	female	NaN

DataFrame Operations - Concatenate

```
persons2 = pd.read_csv("persons2.csv")
persons2
```

	age	height	name	sex
0	28	178	Harold	male
1	43	172	Irene	female
2	37	181	Joe	male

```
pd.concat([persons, persons2])
```

	age	height	name	sex
0	23	156	Alice	female
1	21	181	Bob	male
2	27	176	Charlie	male
3	24	167	Eve	female
4	19	172	Frances	female
5	31	191	George	female
0	28	178	Harold	male
1	43	172	Irene	female
2	37	181	Joe	male

DataFrame Operations - Agregation

```
df = pd.read_csv(os.path.join(basedir,"csse_covid_19_time_series/time_series_covid19_confirmed_global.csv"))
df.head()
```

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	10/8/20	10/9/20	10/10/20	10/11/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	39616	39693	39703	39799
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	14899	15066	15231	15399
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	52658	52804	52940	53072
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	2568	2696	2696	2696
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	5958	6031	6246	6366

5 rows × 274 columns

DataFrame Operations - Agregation

```
df = pd.read_csv(os.p
df.head()
```

	Province/State	Country
0	NaN	Afg
1	NaN	
2	NaN	
3	NaN	
4	NaN	

5 rows × 274 columns

```
# Project to case numbers
covid_cases = covid[covid.columns[4:]]
# Calculate new cases by differencing
covid_new_cases = covid_cases.diff(axis=1)
# Calculate average increase
covid_new_casses_avg = covid_new_cases.mean(axis=1)
# Create new DataFrame with state/country columns
pd.concat([
    covid[covid.columns[:4]],
    covid_new_casses_avg
    ],
    axis=1
)
```

10/11/20
39799
15399
53072
2696
6366

Pandas Final comments

- Conventional Pandas is single threaded and cannot be distributed to different machines.
- Pandas implements a so called *eager* execution model. It runs immediately, no optimization plan.
- In conventional PANDAS, data should fit in your computer.
- You can have only local files, no server required.
- Pandas itself is written in Python and Cython.
- If your dataset scales.. You will need spark, Hadoop, etc…

Notebook exemplo – Legacy Survey

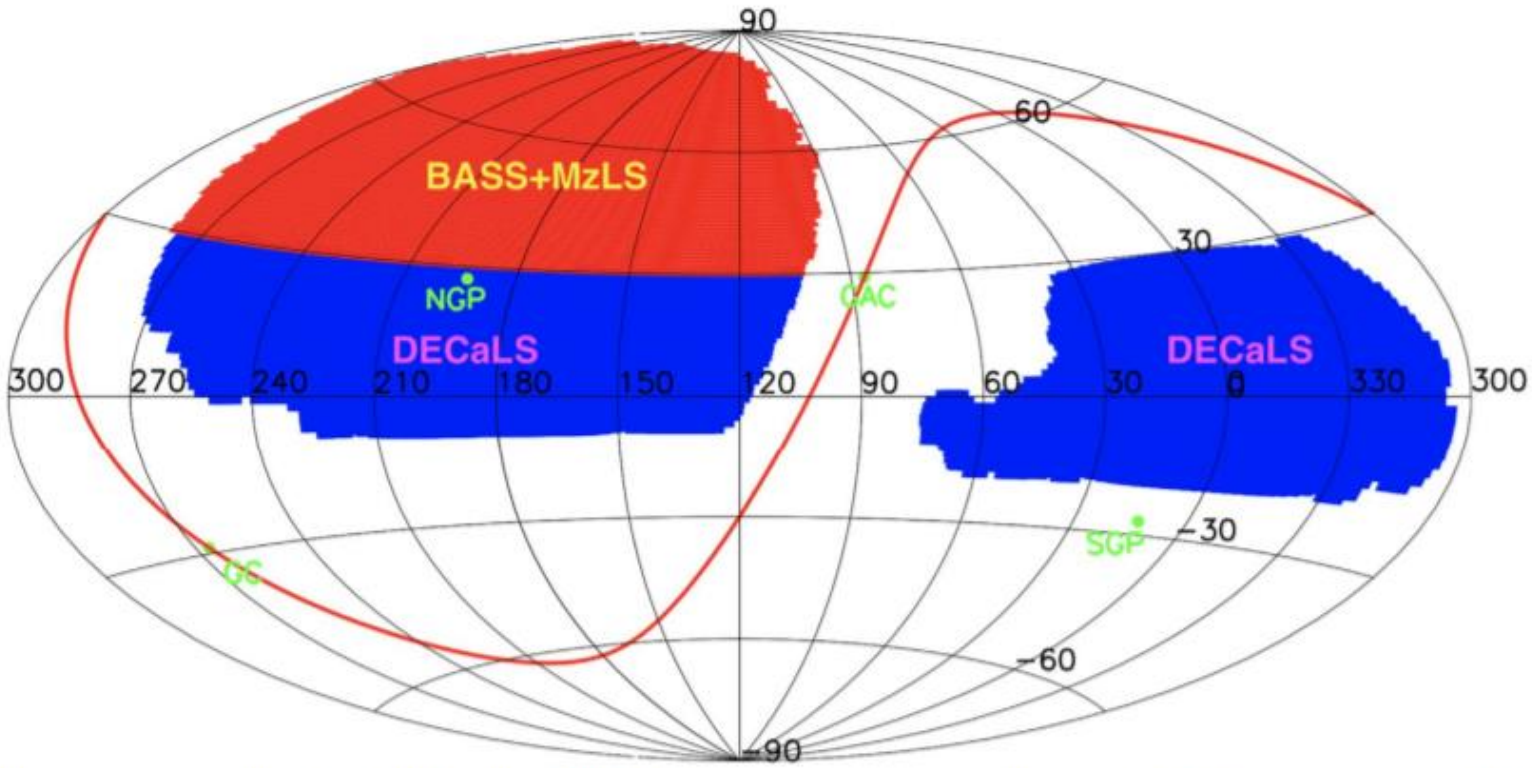


Figure 1. Survey footprints of BASS (red), MzLS (red), and DECaLS (blue). The red curve represents the Galactic plane. The green points with labels show the positions of the Galactic center (GC), Galactic anticenter (GAC), north Galactic pole (NGP), and south Galactic pole (SGP).

Table 1
Imaging Surveys for DESI Targeting

Survey	Telescope	FoV (degree)	Filters	Area (deg^2)	Depth (AB mag)
BASS	2.3 m Bok	1.0	g, r	5400	$g = 24.0$, $r = 23.5$
DECaLS	4 m Blanco	3.0	g, r, z	9000	$g = 24.5$, $r = 23.9$, $z = 22.9$
MzLS	4 m Mayall	0.6	z	5400	$z = 22.9$
unWISE	0.4 m *WISE*	0.78	$W1$, $W2$	full sky	$W1 = 20.4$, $W2 = 19.5$

Notebook exemplo – Legacy Survey

Daframe 1

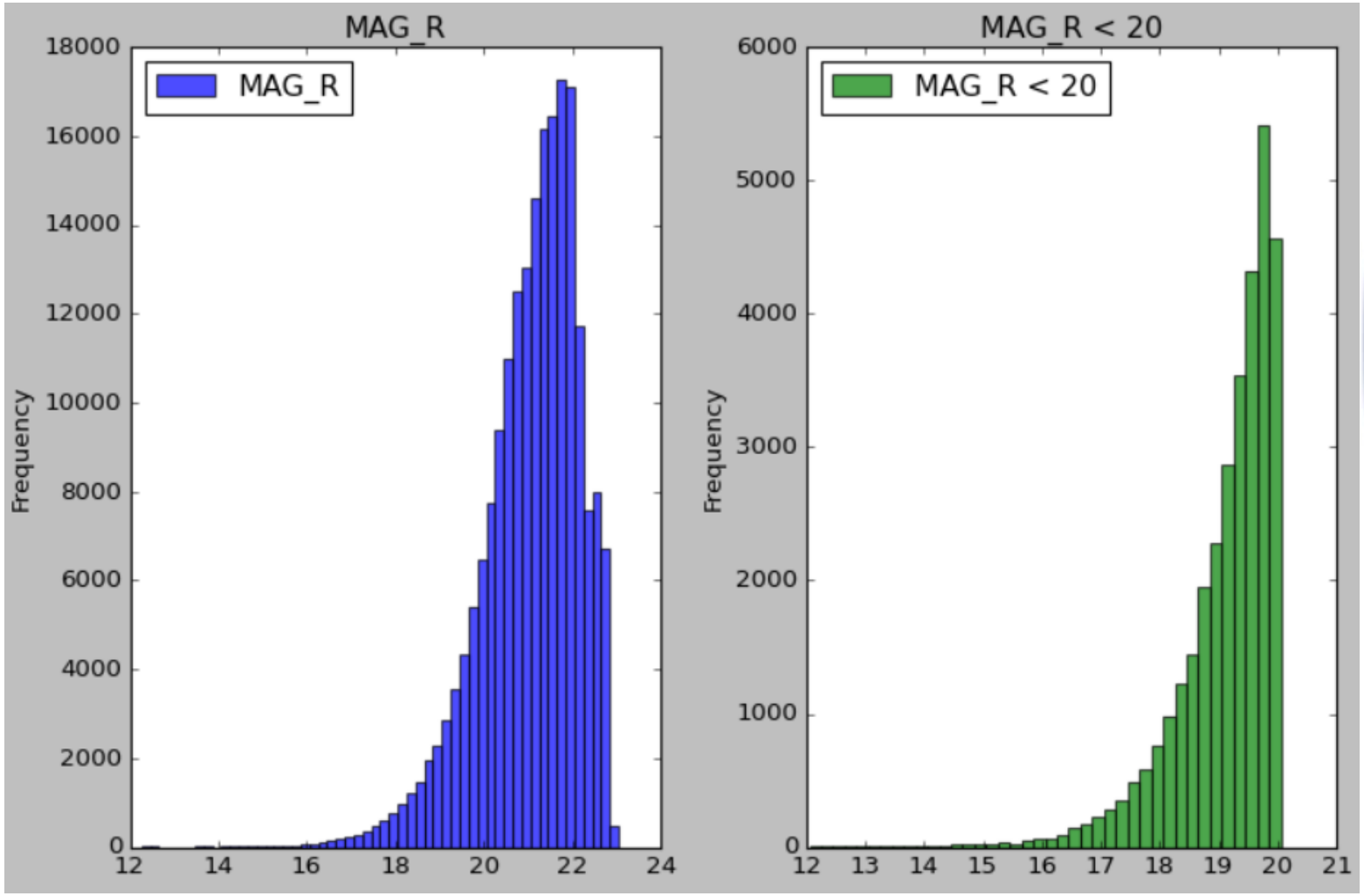
	_id	MAG_R	...	MAGErr_W3	MAGErr_W4
0	219.082895179545033 4.62503563674176	22.280201	...	0.114776	-0.075900
1	219.01868665390734.62537369835652	20.950979	...	0.706567	-0.799649
2	218.976383127798134.626465741964694	18.890739	...	-9.551555	-0.682852
3	218.9246198813485234.62559613136096	20.606371	...	4.317370	-0.843075
4	219.1042252828217234.62584794653823	20.372391	...	1.281699	1.788509
...	...	...	...	...	...
203479	221.384626937292535.03130360684637	20.289848	...	18.655874	-0.504265
203480	221.2671058669764435.03265425222566	22.278854	...	-0.115952	0.318850
203481	221.3579308393770835.0319557953791	21.256090	...	0.324572	0.561515
203482	221.1400346667740635.031834870454304	20.928772	...	-1.286056	-0.460461
203483	221.3061151788022335.032600651235455	22.024387	...	-0.767880	-0.541523

[203484 rows x 15 columns]

['_id', 'MAG_R', 'MAG_G', 'MAG_Z', 'MAG_W1', 'MAG_W2', 'MAG_W3', 'MAG_W4', 'MAGErr_R', 'MAGErr_G', 'MAGErr_Z', 'MAGErr_W1', 'MAGErr_W2', 'MAGErr_W3', 'MAGErr_W4']

Notebook exemplo

- Como definir a Magnitude Limite do Survey?



Exercício: Magnitude Limite

$$m - m_0 = -2.5 \log_{10} f$$

If we consider a linear relation
between the Signal S in charge units and some flux units $f = C_0 S$

$$C_1 \equiv m_0 + \log_{10} C_0$$

$$m \pm \delta m = C_1 - 2.5 \log_{10} (S \pm N)$$

$$m \pm \delta m = C_1 - 2.5 \log [S(1 \pm \frac{N}{S})]$$

Exercício: Magnitude Limite

$$m \pm \delta m = C_1 - 2.5 \log_{10} (S \pm N)$$

$$m \pm \delta m = C_1 - 2.5 \log [S(1 \pm \frac{N}{S})]$$

$$m \pm \delta m = C_1 - 2.5 \log(S) - 2.5 \log(1 + \frac{N}{S})$$

$$\delta m = \pm 2.5 \log(1 + \frac{1}{S/N})$$

Exercício: Magnitude Limite

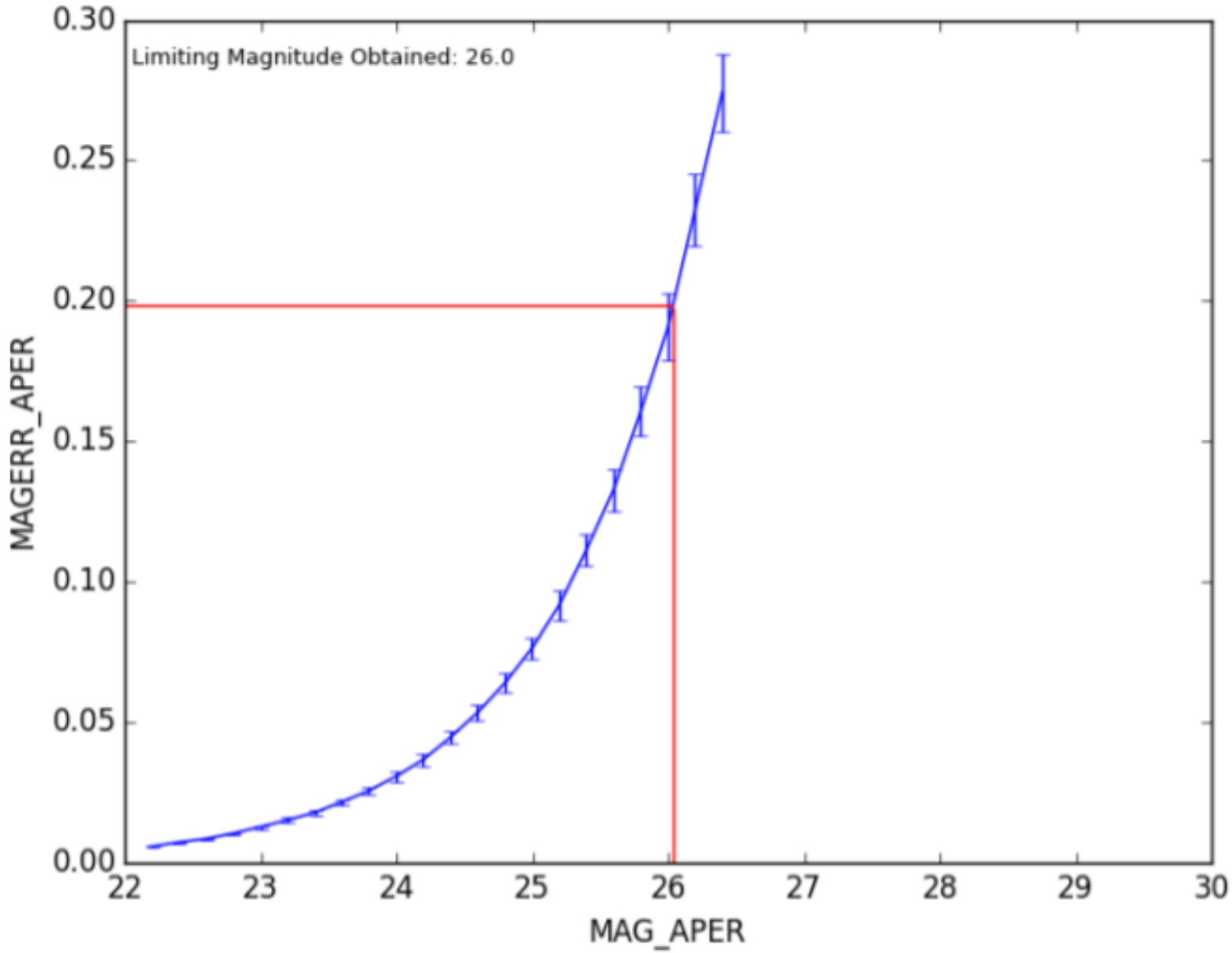
$$\delta m = \pm 2.5 \log(1 + \frac{1}{S/N})$$

which may be expanded to

$$\delta m = \pm 1.0875 \left(\frac{1}{SNR} - \frac{1}{2} \left(\frac{1}{SNR} \right)^2 + \mathcal{O}(3) \right)$$

$$\delta m \approx \pm \frac{1.0875}{SNR}.$$

Exercício: Magnitude Limite



Sobre Imagem Digital

- Imagem é um **sinal digital (2D)** de suporte a informação (Teoria dos Sinais)
- Uma imagem digital é uma função discreta de posição (2D ou 3D, tempo e banda espectral) e níveis de cinza. Cada coordenada da imagem contem uma informação de luminância (ou crominância).

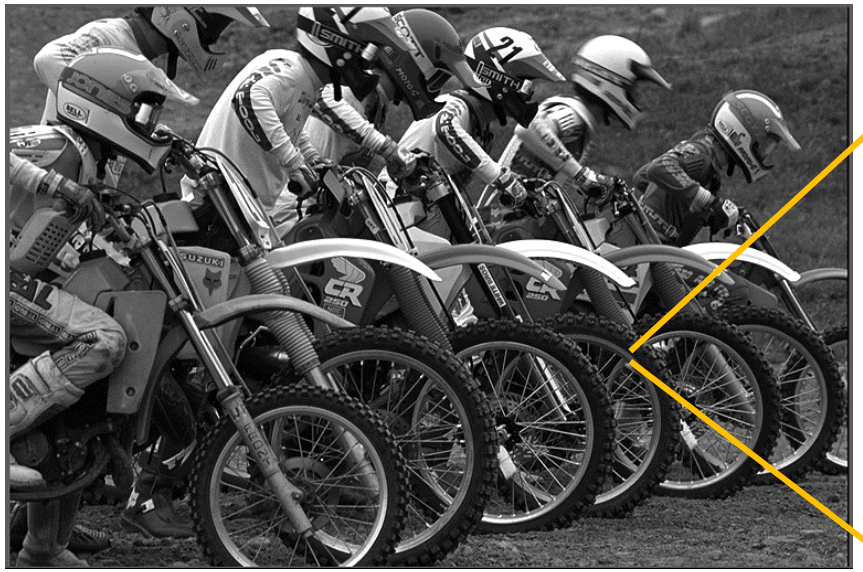
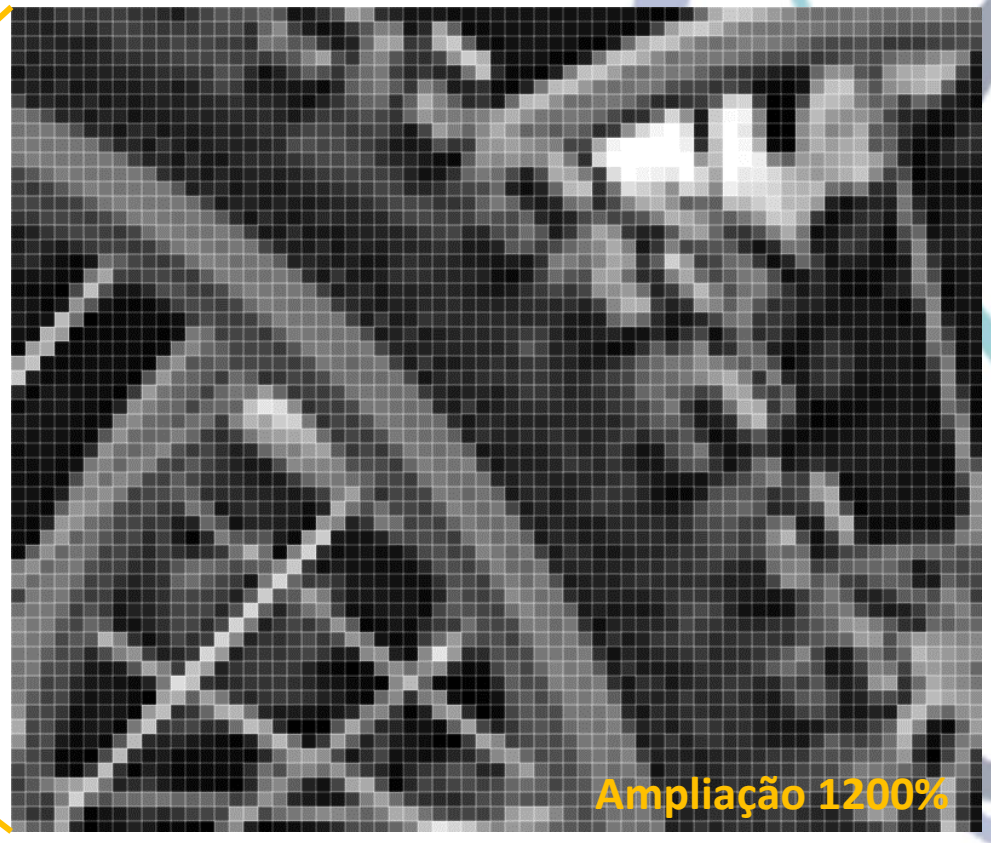
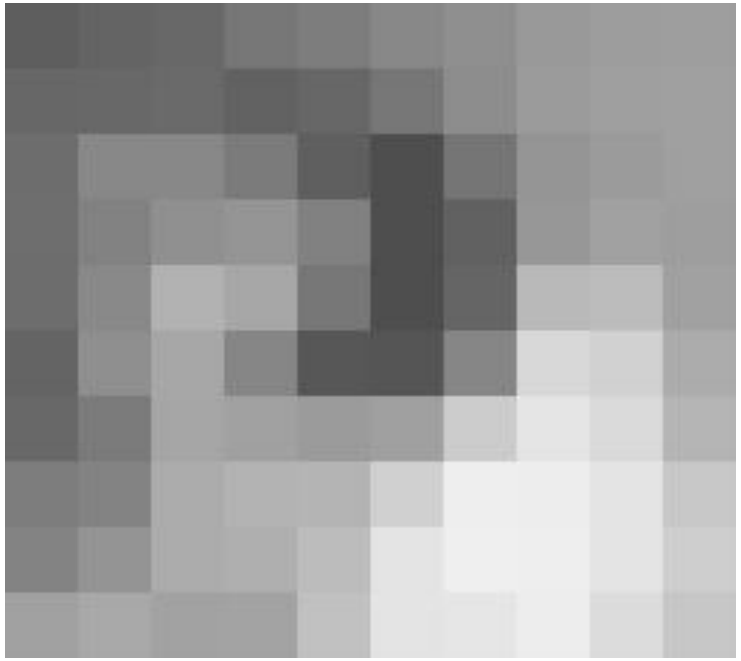


Imagem digital (KODAK – Free)



Sobre Imagem Digital

Uma imagem digital pode ser vista como uma matriz de níveis de cinza, ou valores de intensidade luminosa.



Ampliação

94	100	104	119	125	136	143	153	157	158
103	104	106	98	103	119	141	155	159	160
109	136	136	123	95	78	117	149	155	160
110	130	144	149	129	78	97	151	161	158
109	137	178	167	119	78	101	185	188	161
100	143	167	134	87	85	134	216	209	172
104	123	166	161	155	160	205	229	218	181
125	131	172	179	180	208	238	237	228	200
131	148	172	175	188	228	239	238	228	206
161	169	162	163	193	228	230	237	220	199

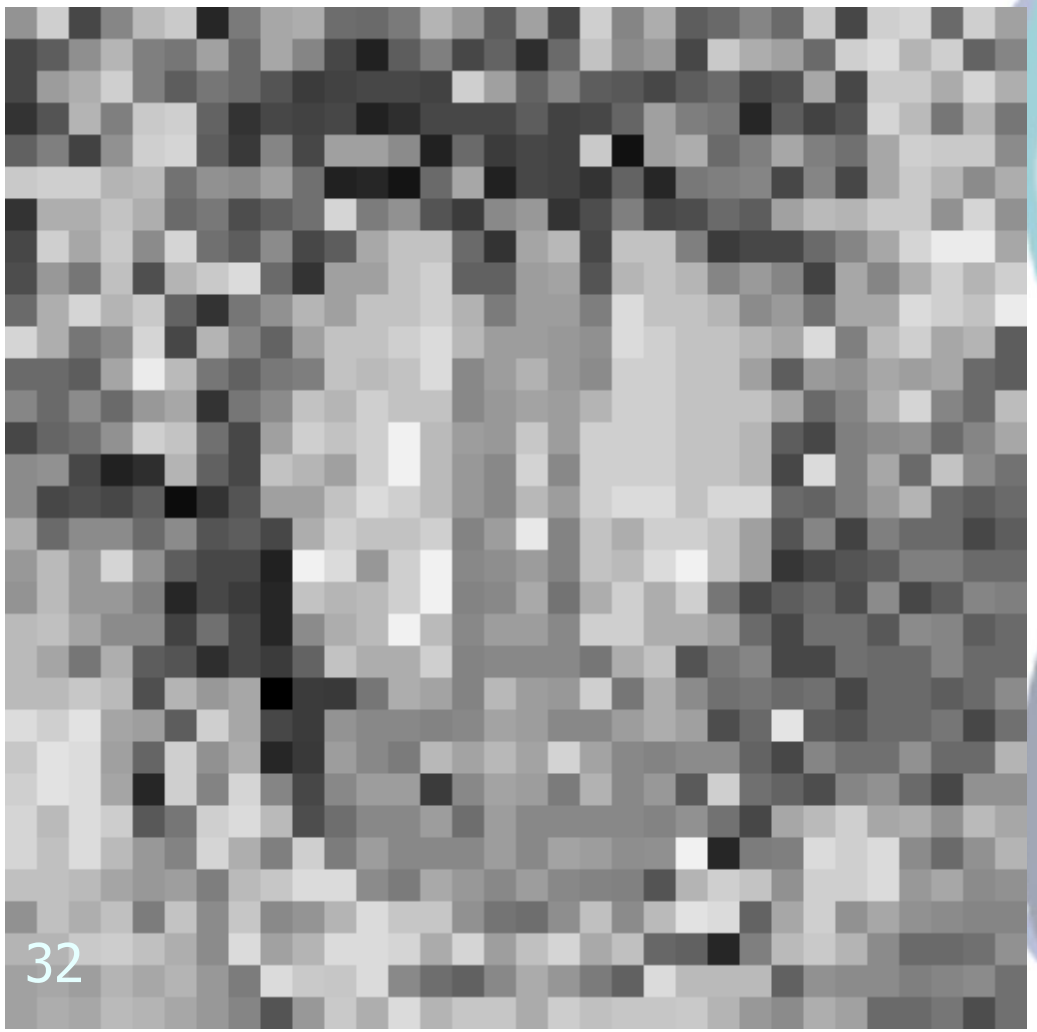
Valores de intensidade luminosa (8 bits)
Níveis de Cinza

Sobre Imagem Digital

Resolução da Imagem

- **DPI – “dots per inch”**
  - Scanners (variável)
- **Número de pixels**
  - Vídeo (fixo)
- **Exemplo simples**
  - Foto de 5x5 cm – 2x2 in.
  - Resolução: 300 dpi
  - Tamanho: 600x600 pixels
  - **Filme Fotográfico:** 5000x5000 dpi

Unidade de comprimento

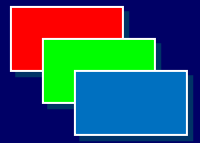


Sobre Imagem Digital (CORES)

- A cor é definida como uma "sensação" na percepção humana.
- Do ponto de vista da Física, a cor é o resultado da incidência de uma onda eletromagnética na retina. Esta tem um comprimento de onda entre 400 a 700nm.

SISTEMA DE CORES

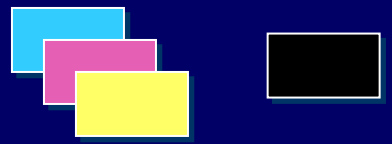
√ RGB (Red Green Blue)



É um Sistema Aditivo/Emissão. A proporção de cada uma das cores primarias (*Red, Green and Blue*) são a base para todas as outras cores, quando somadas. A implementação é feita por meio de circuitos eletrônicos (e.g. em televisão, câmeras, sistemas de computação gráfica, etc.).

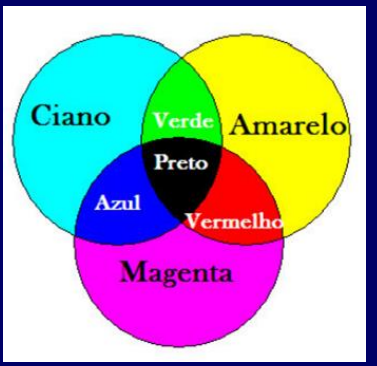
Cores primárias de emissão

CMY(K) (Cyan Magenta Yellow (Black))



√

É um Sistema de cores Subtrativo/Absorção. Utilizado normalmente por dispositivos de impressão e/ou fotográficos. Estes sistemas incluem normalmente uma 4a. cor (preto), para reduzir "*custos*" para produzir todas as cores.



Cores primárias de absorção

Sobre Imagem Digital (CORES)

Escala de 0 a 255

Escala em %

R = 234 G = 212 B = 20	Amarelo	C = 10% M = 11% K=1% Y = 94%
R = 83 G = 12 B = 64	ROXO	C = 57% M = 98% K=32% Y = 22%
R = 20 G = 202 B = 114	Verde Aqua	C = 77% M = 0% K=0% Y = 71%

HSL → OUTRO MODELO: Hue (Matiz), Saturação, Luminosidade

Centro Brasileiro de Pesquisas Físicas

Métodos para Análise de grande volume de dados e Astroinformática

Clécio Roque De Bom – debom@cbpf.br

clearnightsrthebest.com

CBPF