



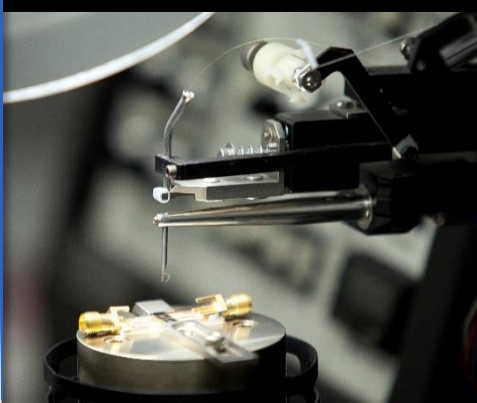
Centro Brasileiro de Pesquisas Físicas



Métodos para Análise de grande volume de dados e Astroinformática

Clécio Roque De Bom – debom@cbpf.br

clearnightsrthebest.com



Métodos para Análise de grande volume de dados e Astroinformática

O objetivo é dar ao aluno uma introdução prática aos conceitos e métodos de análise de dados e principais métodos computacionais utilizados em astronomia utilizando, majoritariamente linguagem python

Créditos: 4



Métodos para Análise Astroinformática

O objetivo é dar ao aluno um
e principais métodos computacionais
linguagem python

Créditos: 4

Programa:

1. Introdução à Python e bibliotecas de manipulação de dados (PANDAS);
2. Estatística descritiva e distribuições de probabilidades em catálogos;
3. Acesso a base de dados Astronômicas; queries
4. Estrutura de dados; fits; hdf; npy, json
5. Análise e visualização Catálogos e imagens
6. Introdução a análise e processamento de imagens
7. Matching de dados astronômicos
8. Análise de series temporais
9. O método da máxima verossimilhança;
10. Inferência Bayesiana;
11. Simulações de Monte Carlo e MCMC;
12. Testes de hipóteses;
13. Análise de correlação;
14. Comparação de modelos;
15. Introdução ao Aprendizado de Máquina (ML);
16. Métodos não-supervisionados em ML;
17. Regressão e Classificação com ML;

15



Informações Gerais

Seminário apresentando resultados ao fim da disciplina.

Código disponível no github (público ou privado) com notebook até 3/12.

Ferramentas recomendadas:

Python (anaconda) 3.X

Tensorflow (Keras), preferencialmente 2.X

Tensorflow Probability

Jupyter

Pandas

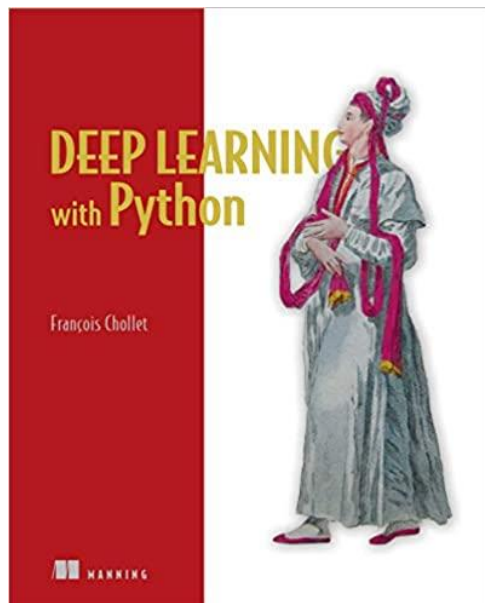
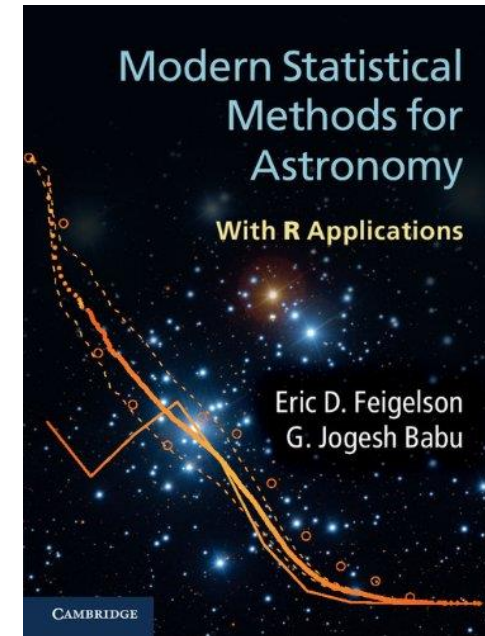
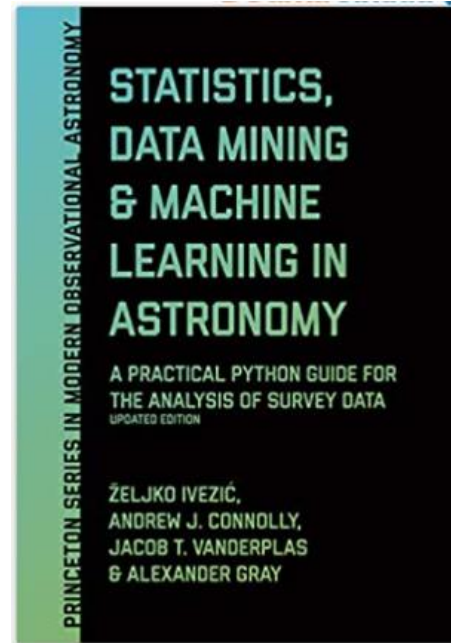
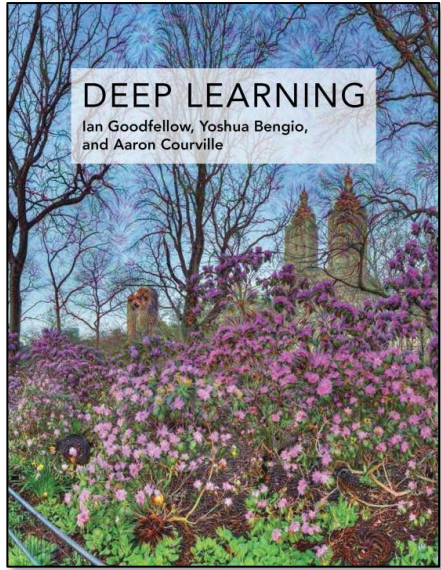
Google Colab

Github

Astropy

Scikit-learn

Bibliografia



- Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey - Željko Ivezić
- Goodfellow, I., et al., "**Deep Learning**", 2016. MIT press. <http://www.deeplearningbook.org>.
- LeCun, Y., Bengio, Y. and Hinton, G., "**Deep learning**." Nature 521.7553 (2015): 436.
- Feigelson, Eric D., and G. Jogesh Babu. Modern statistical methods for astronomy: with R applications. Cambridge University Press, 2012.
- CHOLLET, Francois. Deep Learning with Python and Keras. MITP-Verlags GmbH & Co. KG, 2018.

Big Data

- É o termo de TIC que trata sobre a manipulação de grandes quantidades de dados que precisam ser **processados, transferidos e/ou armazenados**.
- O conceito de “**grande**” é sempre relativo. Um conjunto de dados que é considerado grande hoje quase certamente será considerado pequeno amanhã.
- Big data representa **a condição no qual existem mais dados dos que as técnicas tradicionais podem processar**.
- Alguns **projetos definem que o “Big data” não é uma função da quantidade de dados**, mas da sua complexidade. Consequentemente, é o grau de relacionamento no conjunto de dados que define o que é Big Data.
- **Outra definição → Big-Data e os 3Vs: Volume, Velocidade e Variedade; +V (4Vs): Veracidade, (confiança e incerteza).**

Data Science

- Estatística é a ciência capaz de fazer inferências e tomar decisões onde existe alguma incerteza. É uma **ferramenta** cada vez mais relevante **devido à ampla quantidade e disponibilidade de dados e dos recursos computacionais atuais.**
- A **necessidade de processar e gerenciar grandes quantidades de dados** tornou-se uma característica fundamental das ferramentas estatísticas modernas e é comumente chamada de **Ciência de Dados.**

Inteligência

A controvérsia do QI

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

- **Qualidade Mental**
 - Aprender a partir de alguma experiência anterior
 - Resolver problemas
 - Se adaptar a situações novas

Inteligências Múltiplas (R. Sternbnerg – psicólogo)

Analítica:

Habilidade para pensar de forma abstrata e resolver problemas

Criativa/Sintética:

Habilidade se adaptar a alguma nova situação ou gerar novas ideias

Prática:

Habilidade contextual ou de adaptação as condições do ambiente

Emocional:

Habilidade de perceber, entender, gerenciar e usar emoções em suas interações com os outros

Inteligência

QI → Inteligência

???

Inteligências Múltiplas (R. Sternberg – psicólogo)

Analítica:

Habilidade para pensar de forma abstrata e resolver problemas

Criativa/Sintética:

Habilidade se adaptar a alguma nova situação ou gerar novas ideias

Prática:

Habilidade contextual ou de adaptação as condições do ambiente

Emocional:

Habilidade de perceber, entender, gerenciar e usar emoções em suas interações com os outros

Cristalizada:

Habilidade em acumular conhecimento, habilidades verbais

Fluída:

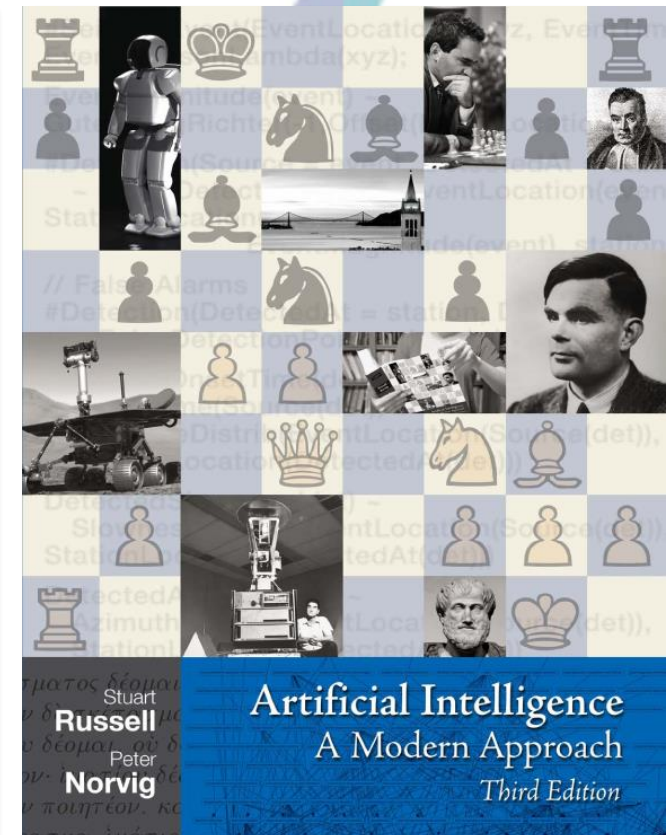
Habilidade abstração → capacidade de raciocínio

IA – Inteligência Artificial

- Inteligência Artificial (IA) é o termo frequentemente usado para descrever **máquinas ou computadores que imitam funções cognitivas** de seres humanos, que estão associados a *mente humana*, como **“aprendizado”** ou a **capacidade de “resolver problemas”**.



199
6



<http://aima.cs.berkeley.edu/>

200
9

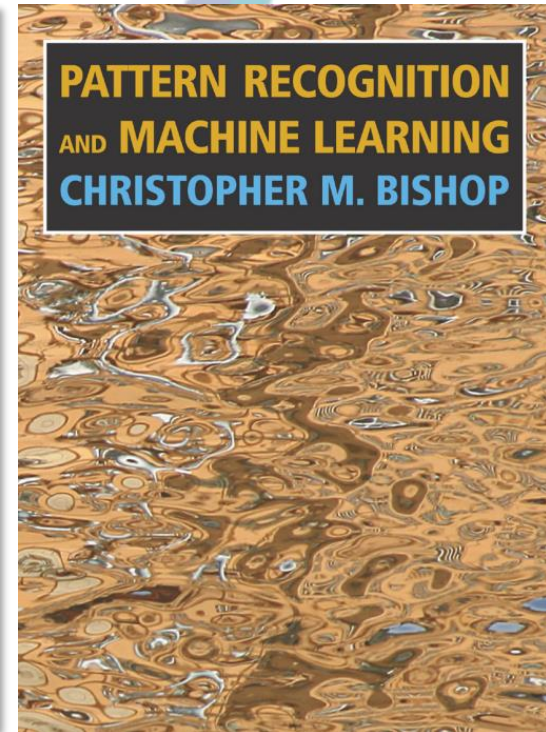
A convergência entre Big Data e IA

- Em vez de depender de dados representativos, os **“Cientistas de Dados” têm atuado cada vez mais a partir dos dados reais**. É devido a essa característica que muitas instituições passaram de uma **abordagem** baseada em hipóteses para uma abordagem/decisão baseada em “dados reais”.
- O **processamento de grandes quantidades de dados** auxiliam na **tomada de decisão** das instituições. A análise de grandes quantidades de dados (Big Data) incentiva a descoberta de padrões por meio de algoritmos iterativos. Como resultado, as instituições podem rapidamente, **experimentalizar mais e aprender mais**.

<https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/>

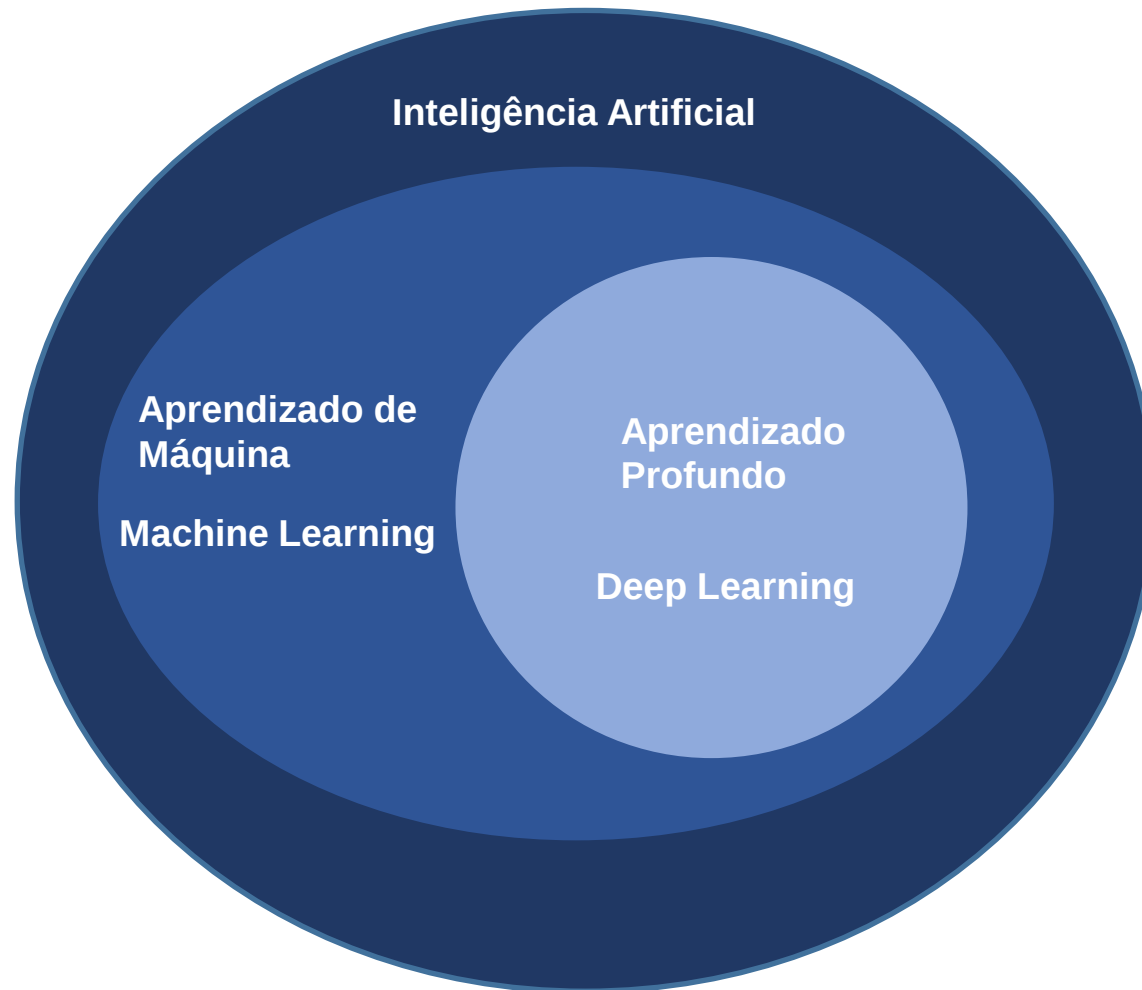
Aprendizado de Máquina

- **Machine learning (ML):** estudo científico de algoritmos e modelos estatísticos por meio de sistemas computacionais para executar uma tarefa específica de maneira eficaz, sem usar instruções explícitas, confiando em padrões e inferências. É uma área da IA.
- **Algoritmos de ML constroem um modelo matemático** baseado em um conjunto amostra, conhecido como "dados de treinamento", para fazer previsões ou permitir decisões sem ser explicitamente programado para executar a tarefa.

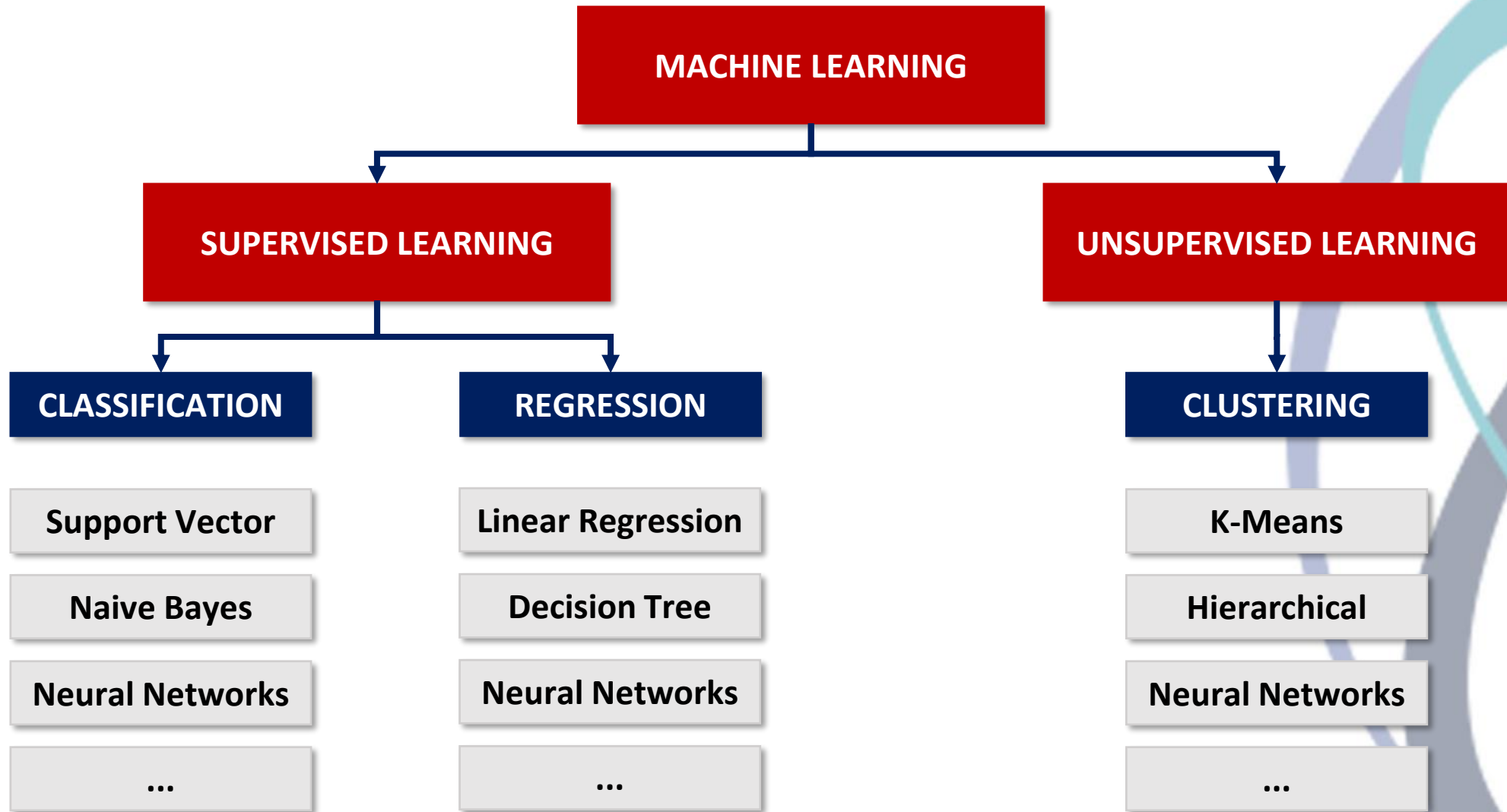


<https://www.springer.com/gp/book/9780387310732>

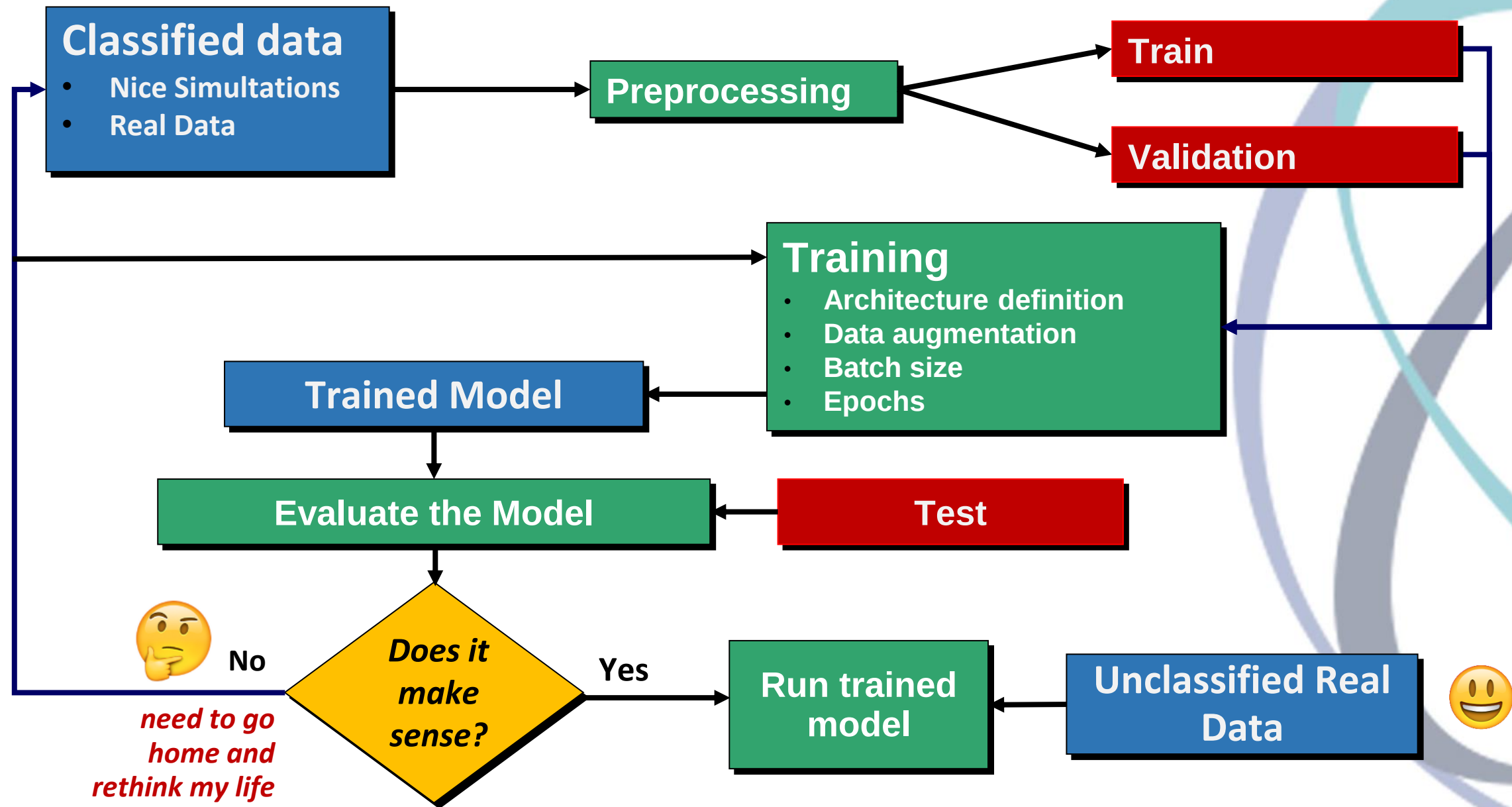
IA – Inteligência Artificial



ML: Supervised & Unsupervised

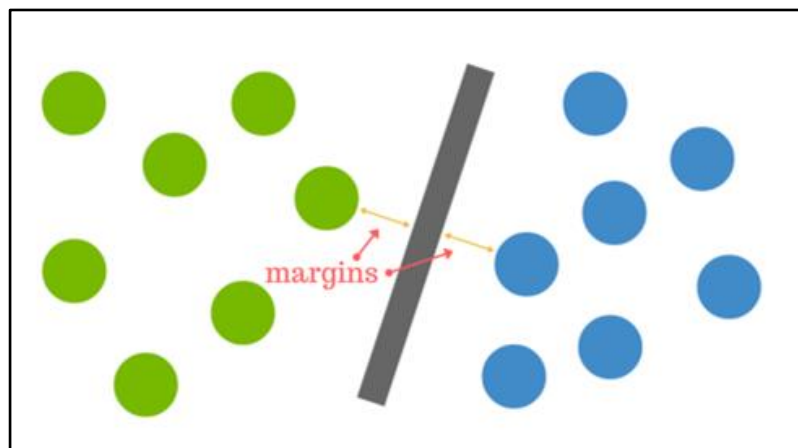
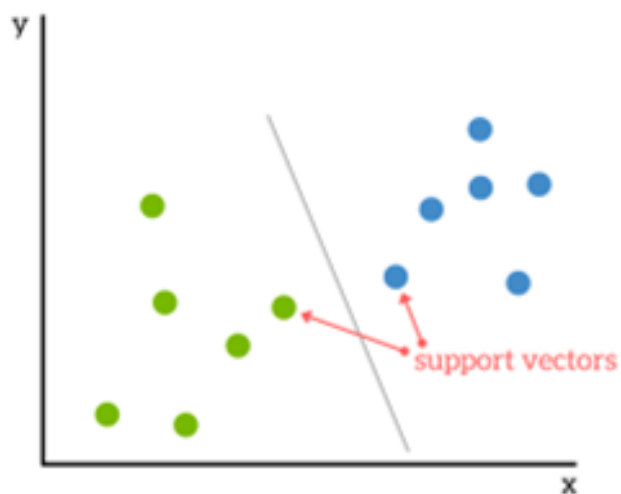


ML: Workflow Supervised

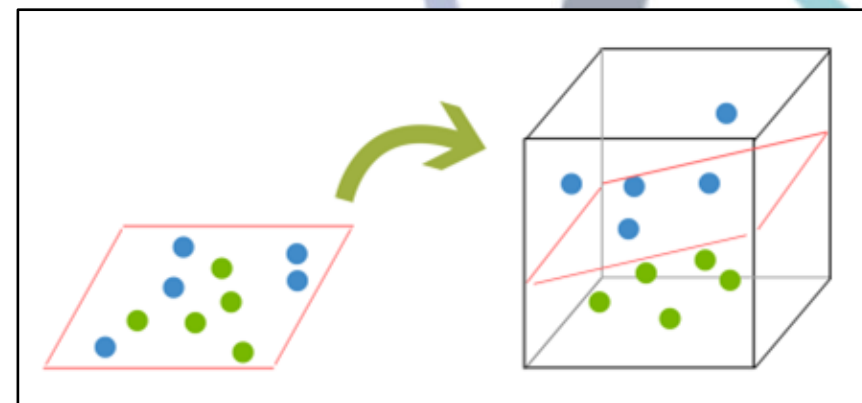
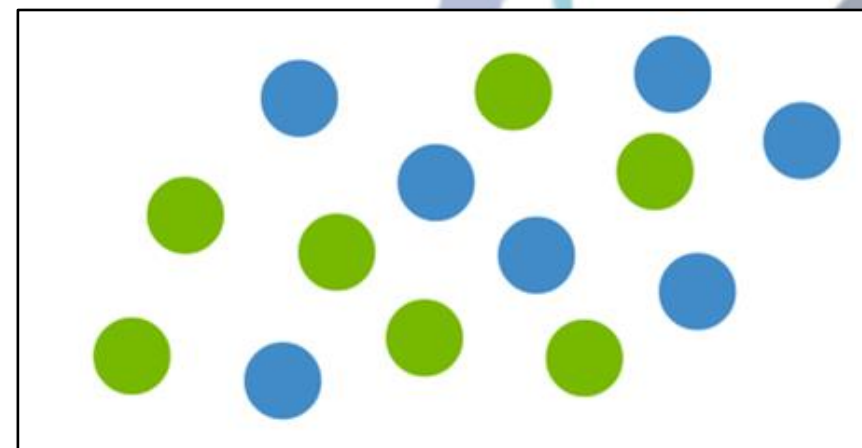


ML: Support Vector Machines - SVM

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes.

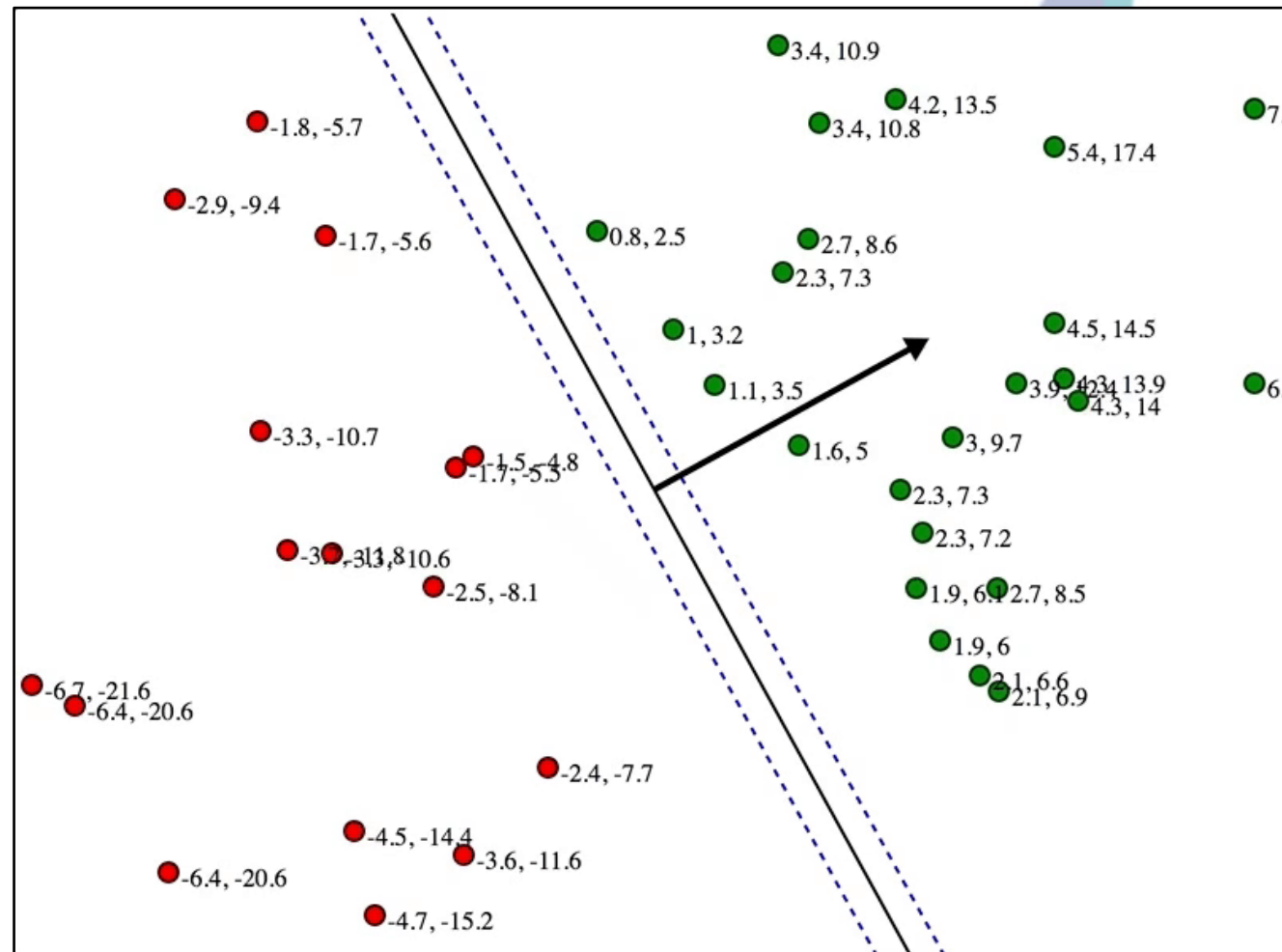
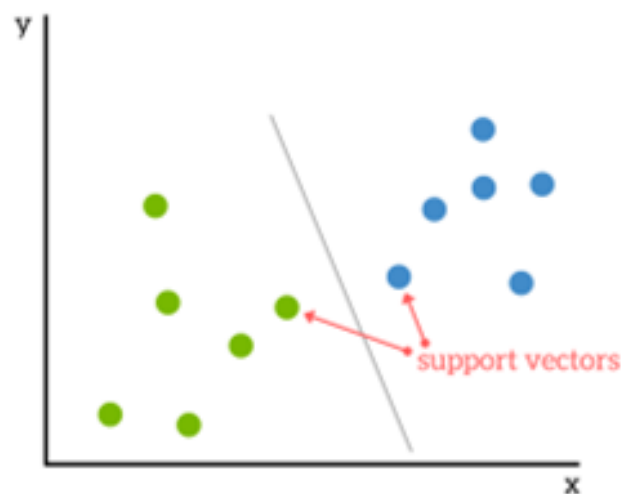


The goal is to choose a **hyperplane** with the **greatest possible margin** between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

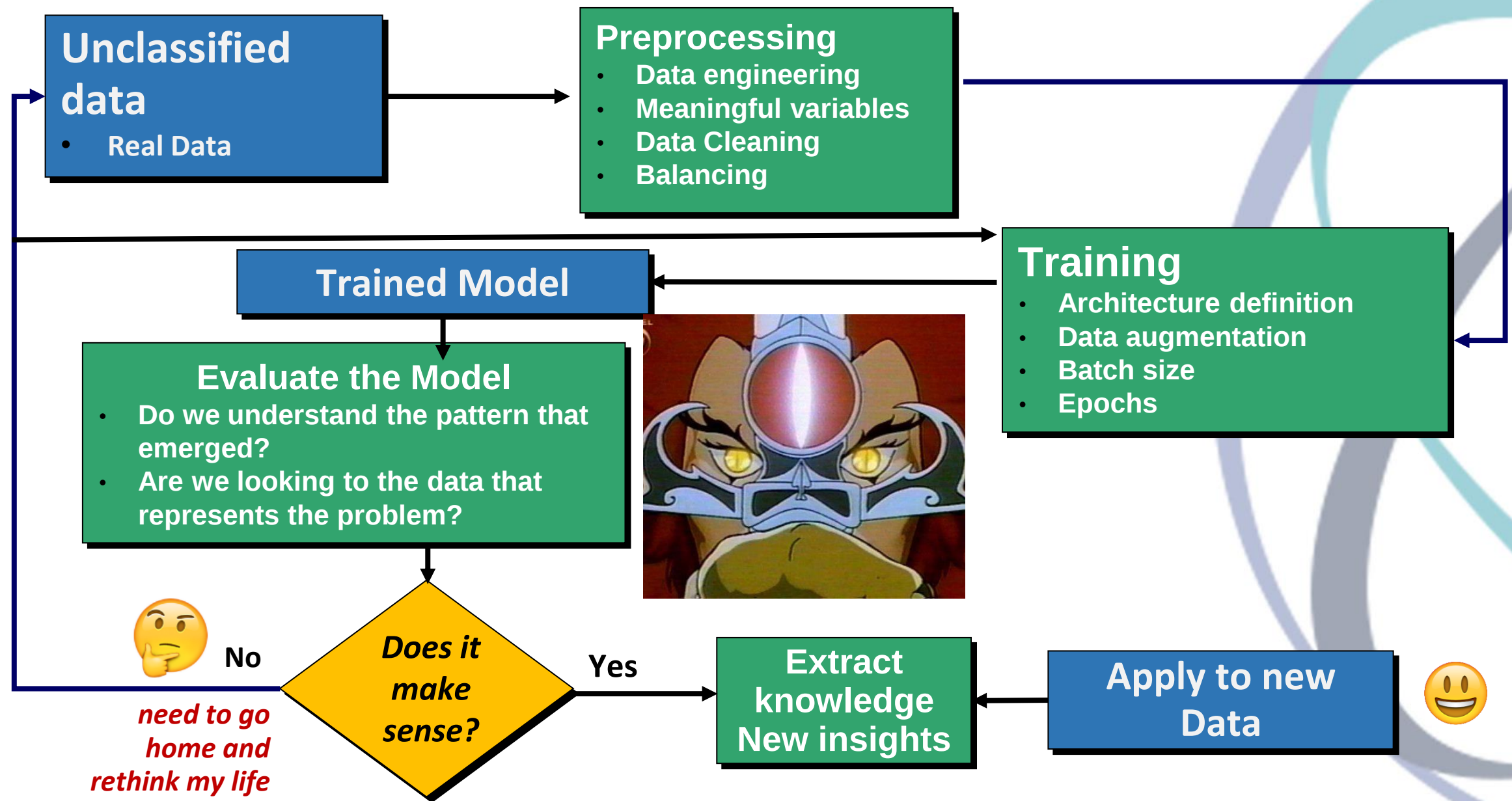


ML: Support Vector Machines - SVM

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes.



ML: Workflow unSupervised



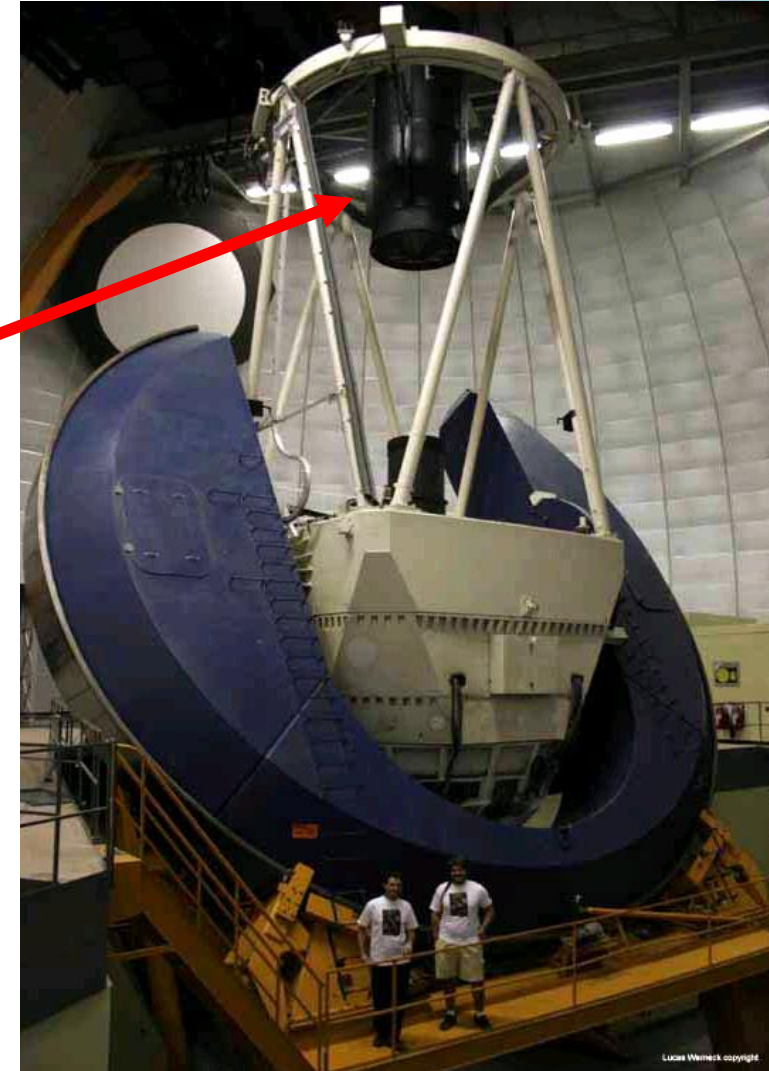
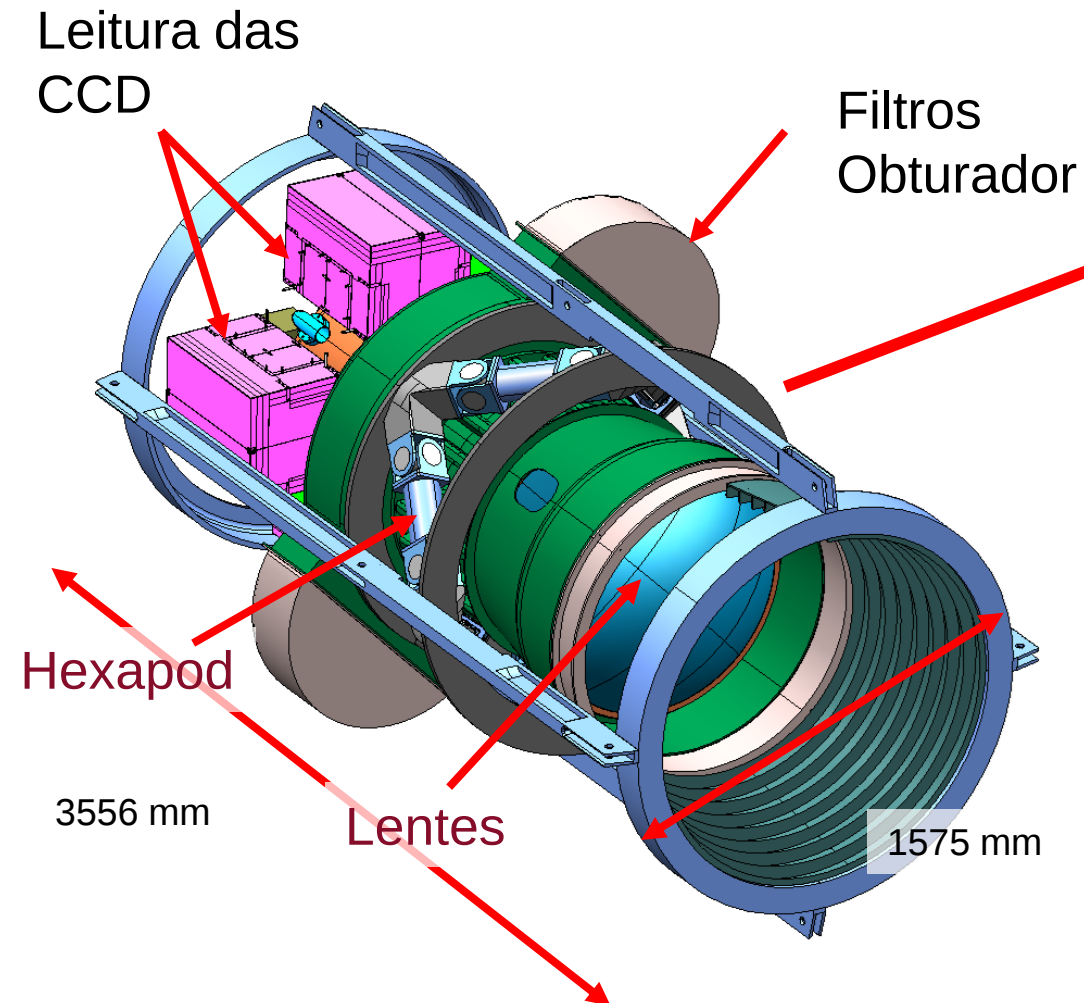
ML: K-Means (Unsupervised)

- K-Means partition the space into K classes.
- Each point belongs to the cluster with the nearest mean
- Here “nearest” is based on some norm (e.g. Euclidean norm)

4 Classes



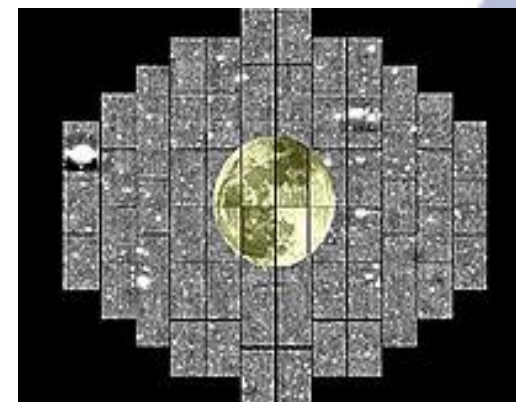
Where are the Big data astronomy?



Where are the Big data astronomy?

• *Dark Energy Survey*

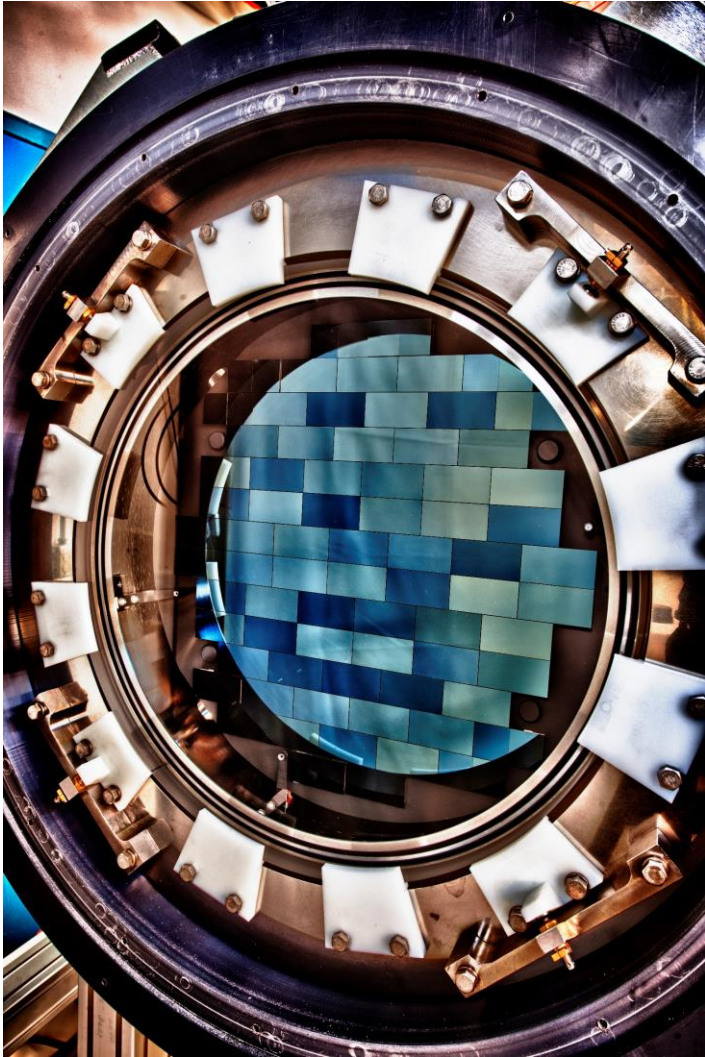
- *Blanco CTIO 4m*
- DECam:
 - 500 Megapíxeis,
 - $\sim 3 \text{ deg}^2$ fov
 - CCDs sensíveis no vermelho
- 30% of CTIO in 5 yr (525 noites)
- $g, r, i, Z, Y = 24.6, 24.1, 24.3 (23.9)$



62 2kx4k Image CCDs: 520 MPix
8 2kx2k focus, alignment CCDs
4 2kx2k guide CCDs

Melhor instrumento óptico para a cosmologia de sua época

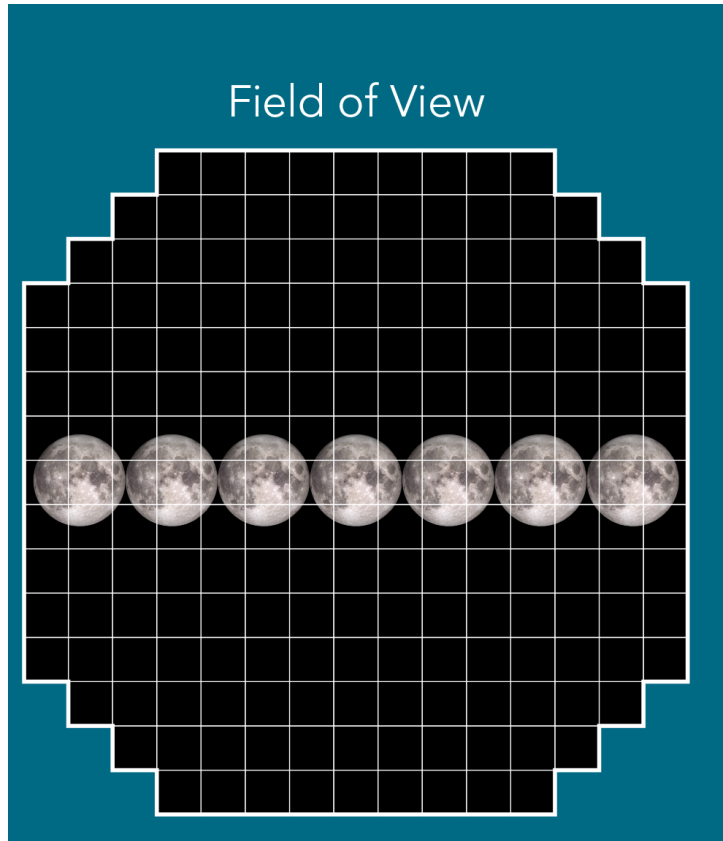
Where are the Big data astronomy?



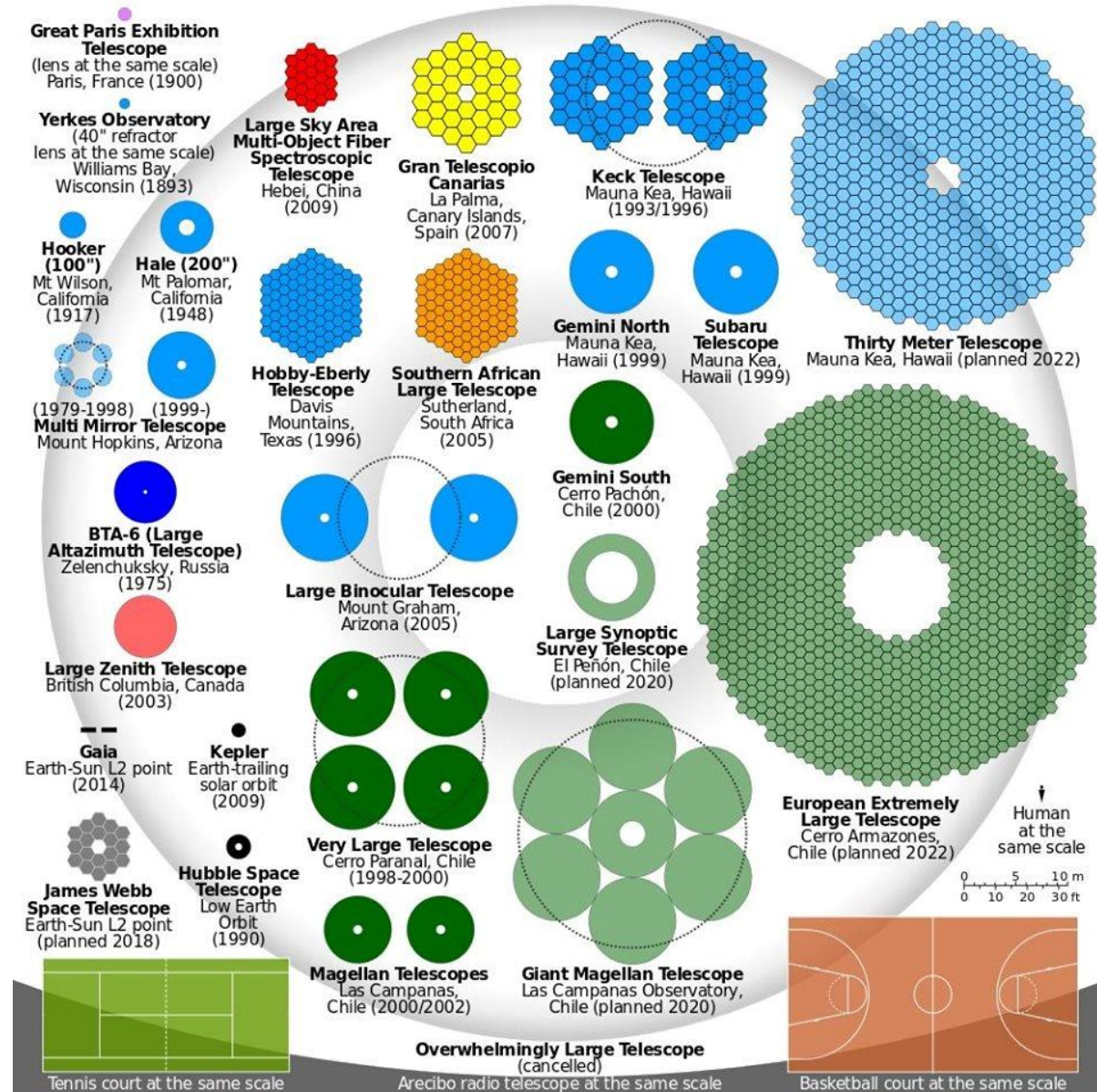
a 570-megapixel CCD camera,
a low-noise electronic readout system,
a wide-field optical corrector,
a combination shutter-filter system, and
a hexapod adjustor to provide stability.

Where are the Big data astronomy?

LSST (in Vera Rubin Observatory) will generate about 15 terabytes (15 trillion bytes) of raw data per night and 30 petabytes over its 10-year survey life. After processing, that's still a 15 PB (15,000 TB) store.

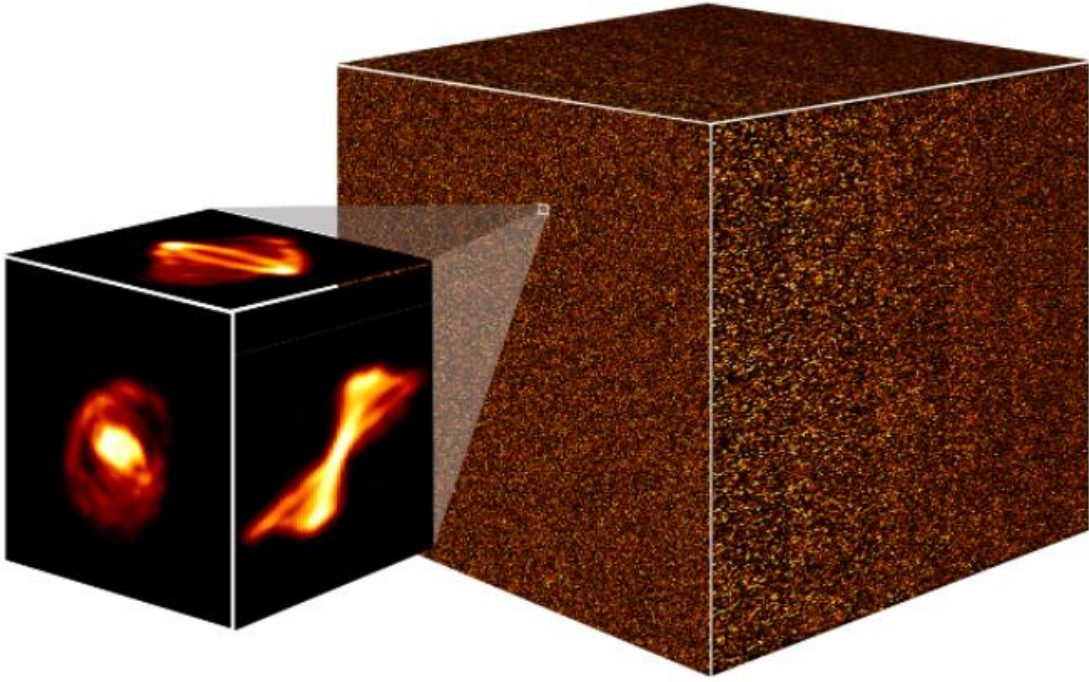


Where are the Big data astronomy?



SKA Challenge

Data cube ~ 4 TB



NVIDIA RTX3090 24GB (2021)



NVIDIA Ampere 24GB (2021)

Radio Galaxy Zoo: ClaRAN — a deep learning classifier for radio morphologies

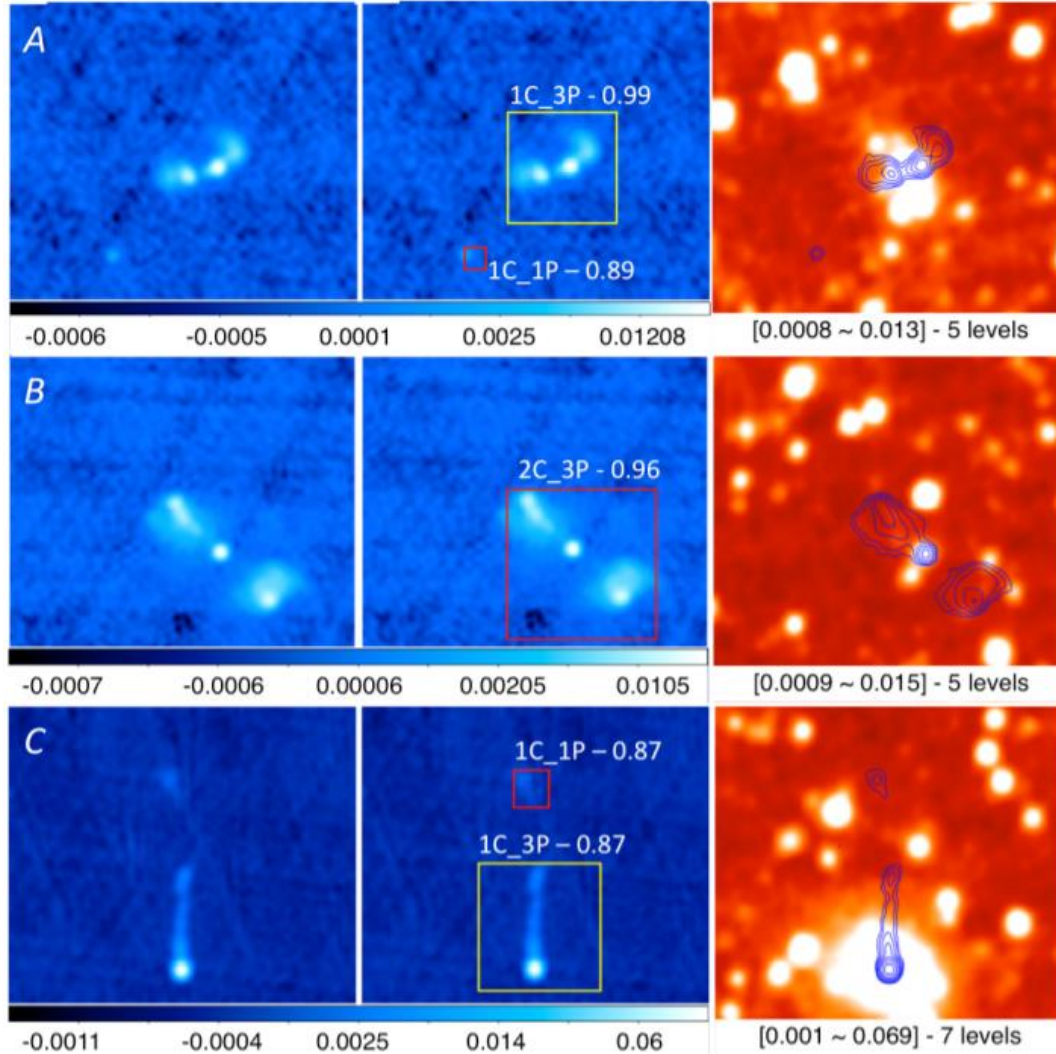
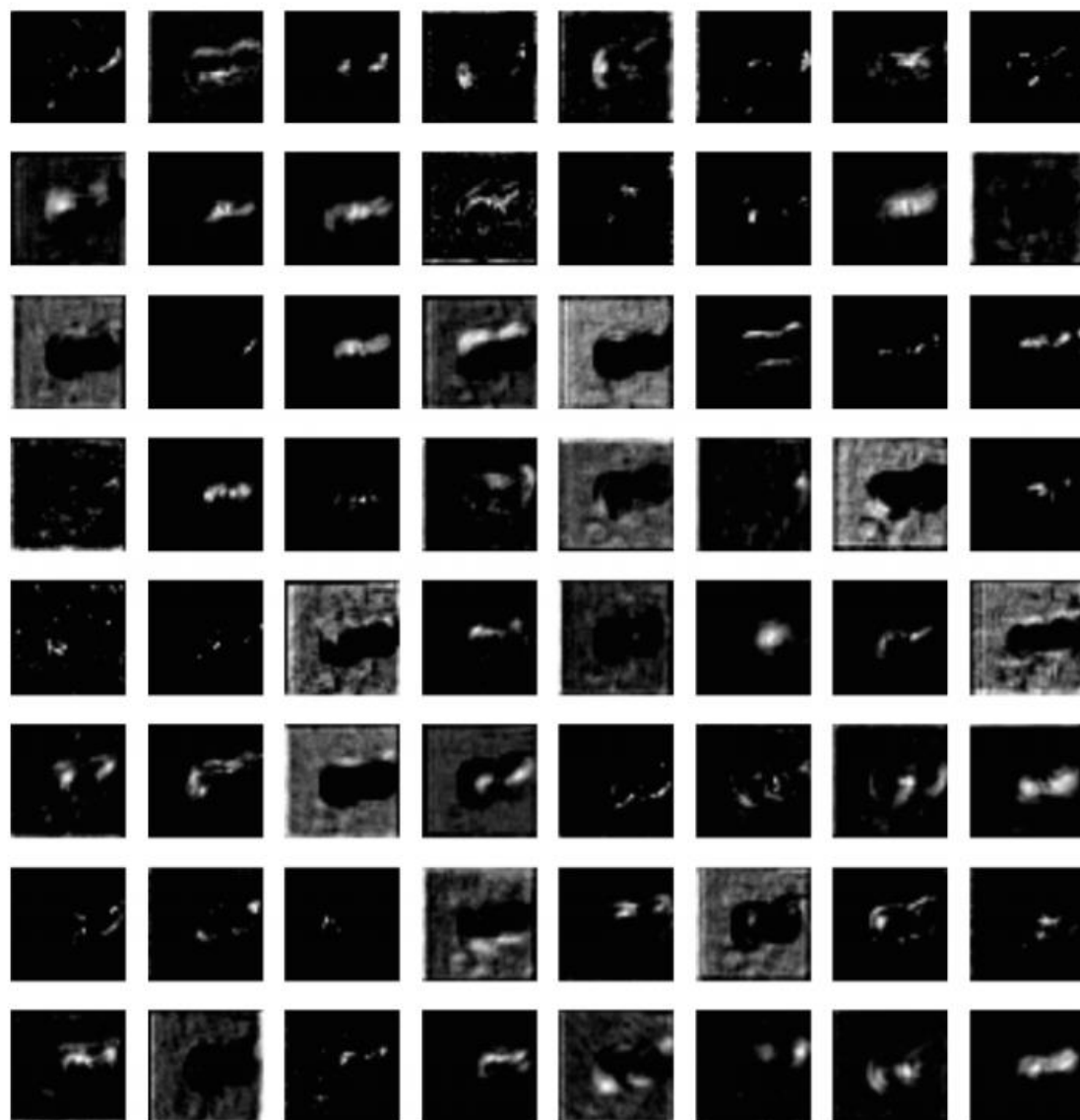
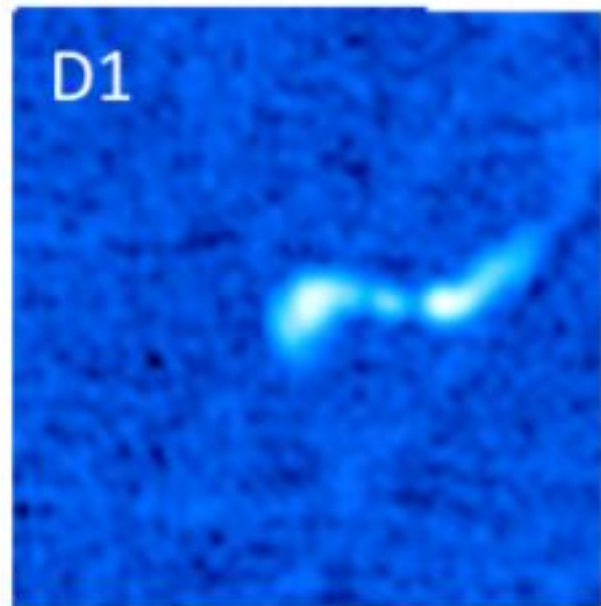
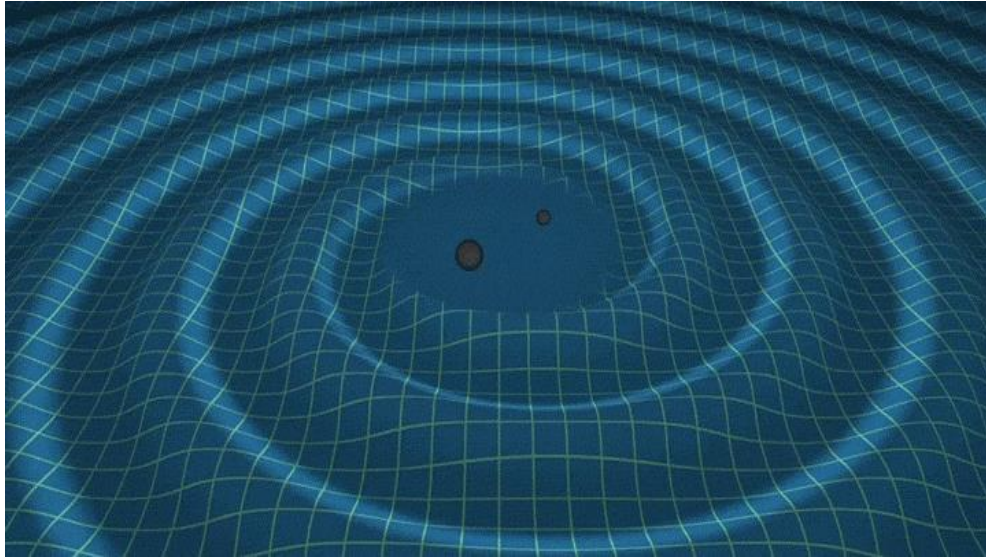


Figure 1. Three classification examples (*A*, *B*, and *C*) on RGZ subjects — each of them $3' \times 3'$ in size — FIRST J081700.6+571626, FIRST J070822.2+414905, and FIRST J083915.7+285125. The first column shows the FIRST radio emission. The second column shows the CLARAN output — a box encompassing each identified source, and its morphology is labelled as iC_jP , where i and j denotes the number of radio components and the number of radio peaks respectively. Each morphology label is associated with a score between 0 and 1, indicating the probability of the quoted morphology class. The first two columns share the same color bar at the bottom, denoting flux density values in Jy/beam. The last column shows the corresponding WISE infrared image overlaid with 5σ radio contours. The contour levels (Jy/beam) are shown at the bottom of each infrared image.

Radio Galaxy Zoo: ClaRAN — a deep learning classifier for radio morphologies



A Busca das Kilonovas e contrapartida de ondas Gravitacionais



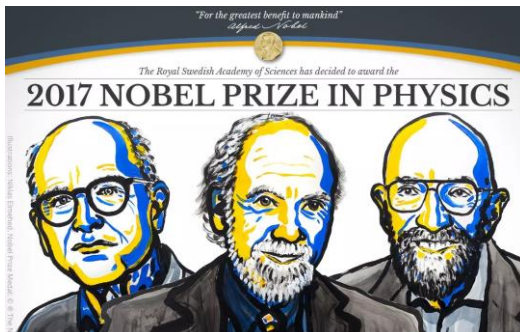
Colisão de estrelas de nêutrons geram **ondas gravitacionais**.

Estes eventos estão entre os mais energéticos no Universo.

Quando uma onda acontece, não sabemos de onde exatamente ela veio

Ajudam a obter medidas fundamentais para o modelo de Universo.

Maior parte do ouro no Universo e outros elementos são gerados nesses eventos.



Rainer Weiss, Barry Barish e Kip Thorne foram laureados com o Prêmio Nobel de Física de 2017 pela descoberta das ondas gravitacionais.

A Busca das Kilonovas e contrapartida de ondas Gravitacionais

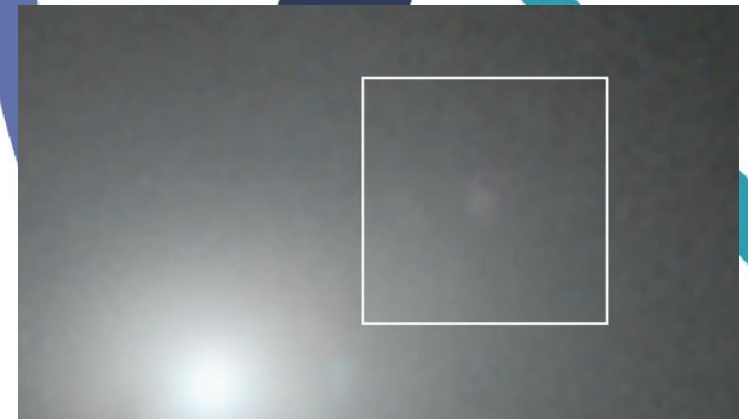


A Busca das Kilonovas e contrapartida de ondas Gravitacionais



N. Risinger (skysurvey.org)

A Busca das Kilonovas e contrapartida de ondas Gravitacionais



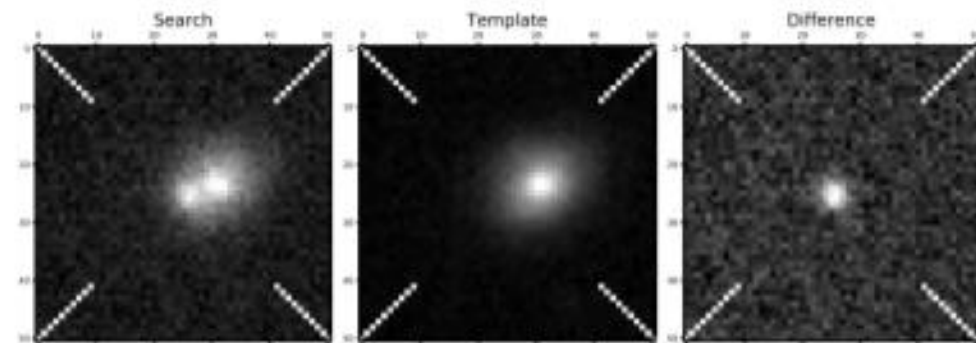
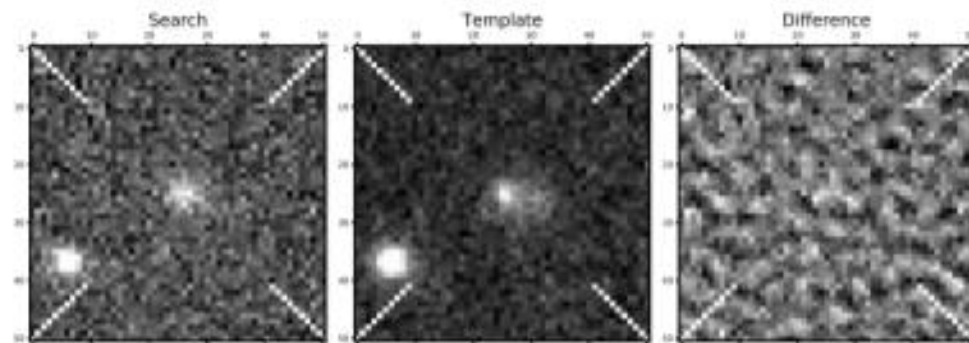
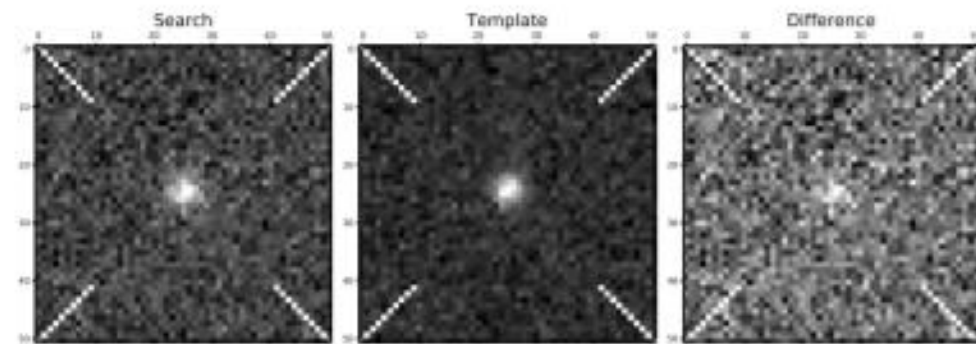
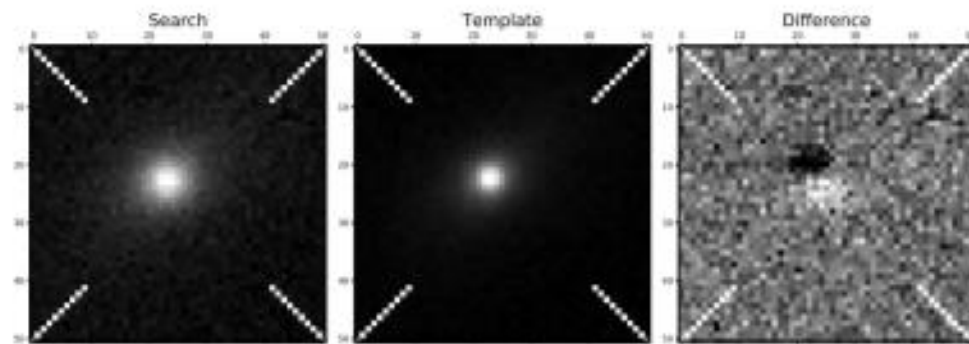


Figure 1. The upper left panel shows an example of a “bad subtraction” image. The dipole-like structure present in the difference image is an indicative feature of this type of artifact. The upper right panel shows an example of a “pre-existing point source” image. The lower left panel shows an example of a “no obvious transient” image. The lower right panel shows an example of a “transient + host” image where a smaller, point-like object is visible within a larger structure representing the host galaxy.

Fits Tables

Table Data as a Record Array

What is a Record Array?

A record array is an array which contains records (i.e., rows) of heterogeneous data types. Record arrays are available through the records module in the NumPy library.

Here is a sample record array:

```
>>> import numpy as np
>>> bright = np.rec.array([(1, 'Sirius', -1.45, 'A1V'),
...                        (2, 'Canopus', -0.73, 'F0Ib'),
...                        (3, 'Rigil Kent', -0.1, 'G2V')],
...                       formats='int16,a20,float32,a10',
...                       names='order,name,mag,Sp')
```

What do you find in astro tables?

Name
ra
dec
MAG_R
MAG_G
MAG_Z
MAGErr_R
MAGErr_G
MAGErr_Z

MAGErr_R
MAGErr_G
MAGErr_Z
MAGErr_W1
MAGErr_W2
MAGErr_W3
MAGErr_W4
z
zErr
zmin
zmax
z_spec

Fits Tables

```
>>> from astropy.io import fits
>>> fits_table_filename = fits.util.get_testdata_filepath('btable.fits')

>>> hdul = fits.open(fits_table_filename) # open a FITS file
>>> data = hdul[1].data # assume the first extension is a table
>>> # show the first two rows
>>> first_two_rows = data[:2]
>>> first_two_rows
[(1, 'Sirius', -1.45000005, 'A1V') (2, 'Canopus', -0.73000002, 'F0Ib')]
>>> # show the values in field "mag"
>>> magnitudes = data['mag']
>>> magnitudes
array([-1.45000005, -0.73000002, -0.1          ], dtype=float32)
>>> # columns can be referenced by index too
>>> names = data.field(1)
>>> names.tolist()
['Sirius', 'Canopus', 'Rigil Kent']
>>> hdul.close()
```

Exercicio : Crie a sua coluna cor

- Abra o fits table
- Faça um histograma das magnitudes
- Crie um array com cores e faça um plot
- Salve a nova coluna no fits table
- Documentação disponível em:
docs.astropy.org/en/stable/io/fits/usage/table.html

Manipulating fits tables

```
import os
import numpy as np
from astropy.table import Table
import matplotlib.pyplot as plt
import pandas as pd

def open_fits_catalog(fits_file):
    hdu_list=fits.open(fits_file, ignore_missing_end=True)
    #print hdu_list
    hdu = hdu_list[1]    # table extensions can't be the first extension, so there's a dummy image extension at 0
    #print hdu.header
    cat_table = Table(hdu.data)
    cols=hdu.columns
    return cat_table, cols
```

```
[ ] dados_dir = '/home/dados229/gabriel/escolacbpf/'
data_file = os.path.join(dados_dir, 'specz_catalog_ecbpf.fits')
data, _ = open_fits_catalog(data_file)
```

```
[ ] ### MAG PLOTS

mags = ['petroMag_'+band for band in 'ugriz']

key_list = mags
color_list = ['black', 'green', 'red', 'tomato', 'palevioletred']
hor = 2
ver = 3

plt.style.use('classic')
```



Centro Brasileiro de Pesquisas Físicas



Métodos para Análise de grande volume de dados e Astroinformática

Clécio Roque De Bom – debom@cbpf.br

clearnightsrthebest.com

