



Observatório
Nacional

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO



Inteligência Artificial e a era de Big Data em Astronomia

Lilianne Nakazono, PhD
Observatório Nacional

Escola de Inverno em Astrofísica 2025

Sobre mim

Bacharela em Estatística (IME-USP)

Doutora em Astronomia (IAG-USP)

- Doutorado sanduíche na University of Washington

Colaborações internacionais:

- S-PLUS
- J-PAS
- COIN

Coordenação do Grupo de
Astroinformática

*Por favor não gravar ou
tirar fotos em slides que
tiverem esse ícone!*





Grupos de Pesquisa



Astrofísica
Computacional



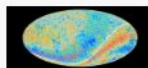
Astrofísica Estelar e
Galáctica



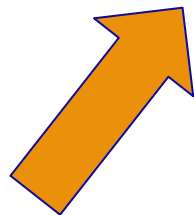
Astrofísica Extragaláctica



Ciências Planetárias



Cosmologia e
Astropartículas





Conteúdo

01

Introdução

O que é inteligência artificial?

02

IA na Astronomia

Como se deu o crescimento de IA na Astronomia?

03

Fundamentos

Uma breve descrição matemática

04

Aplicações

Um pouco do meu trabalho

Conteúdos extras (se der tempo!): iniciativas brasileiras e ferramentas úteis



01

Introdução



Inteligência Artificial

O que vem à sua mente quando
falamos de inteligência artificial?

Respostas dos participantes da Escola de Inverno do ON 2025:



Inteligência Artificial

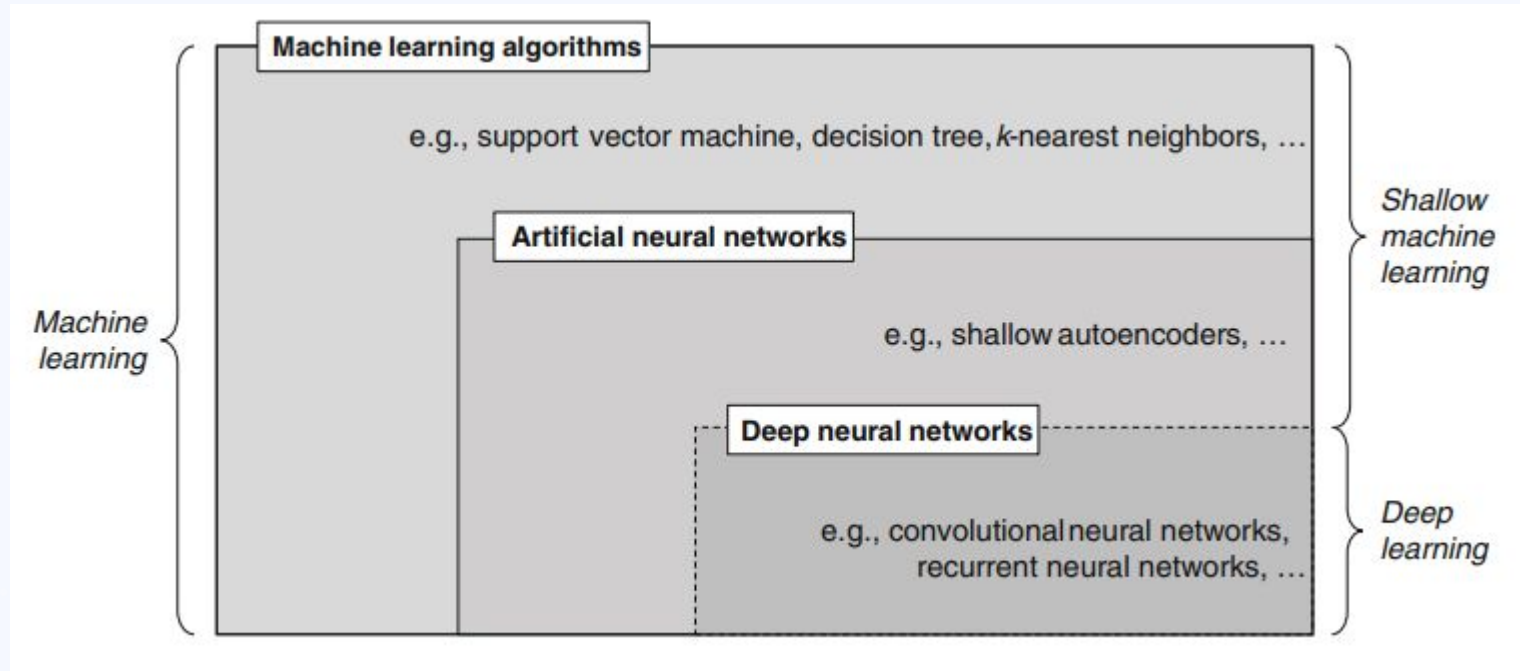


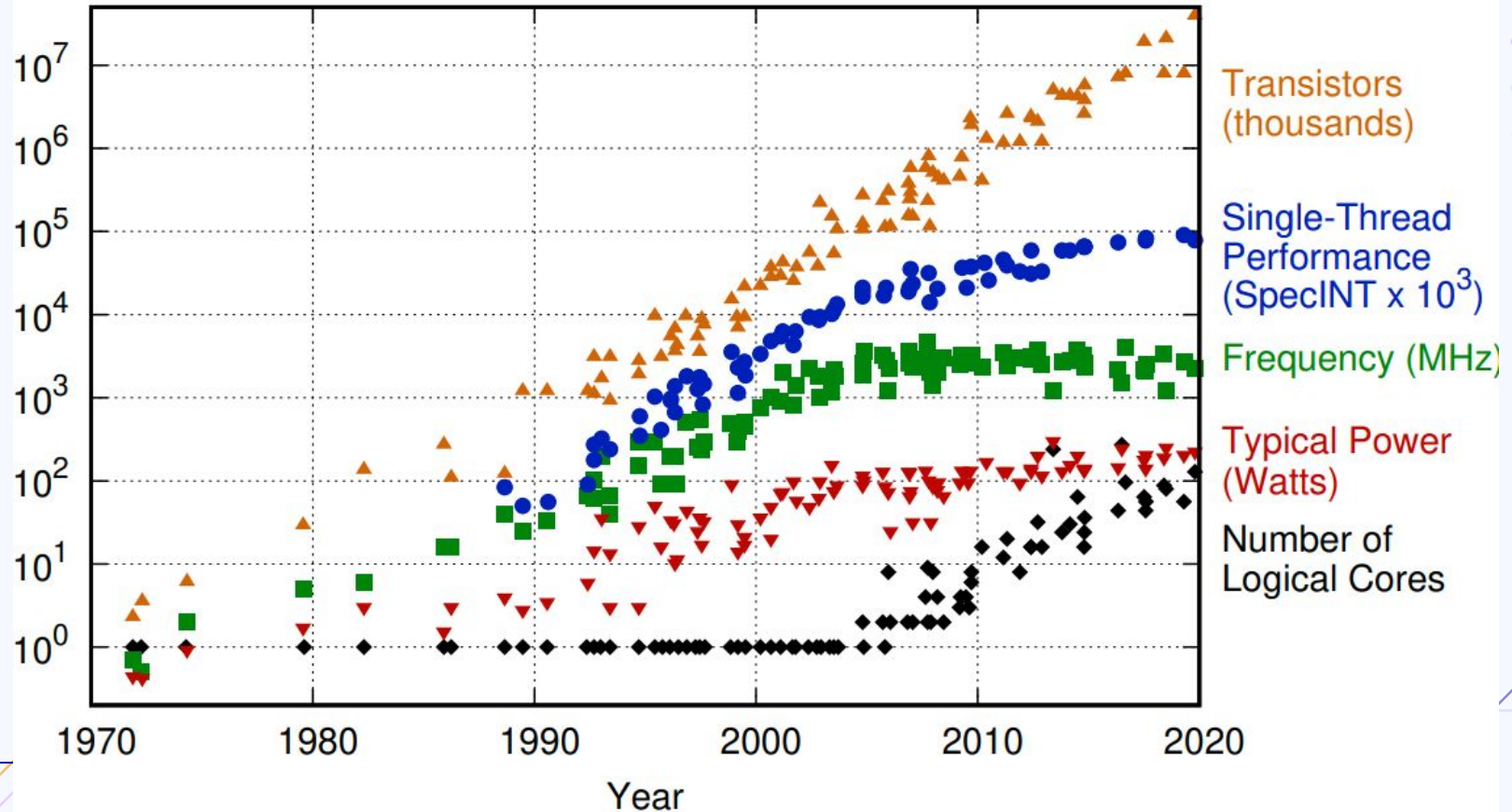
Diagram of Machine learning algorithms learning concepts and classes (inspired by Goodfellow et al. 2016, p. 9). Source: Janiesch, Zschech & Heinrich, 2021

Não é um conceito novo!

- O termo “Inteligência Artificial” surge em 1956 em uma conferência
- Proposição de redes neurais artificiais (Rosenblatt, 1957)
- Primeiro uso de redes neurais em artigo na Astronomia (Jeffrey & Rosner, 1986)

Por que o “boom” da IA só aconteceu recentemente?

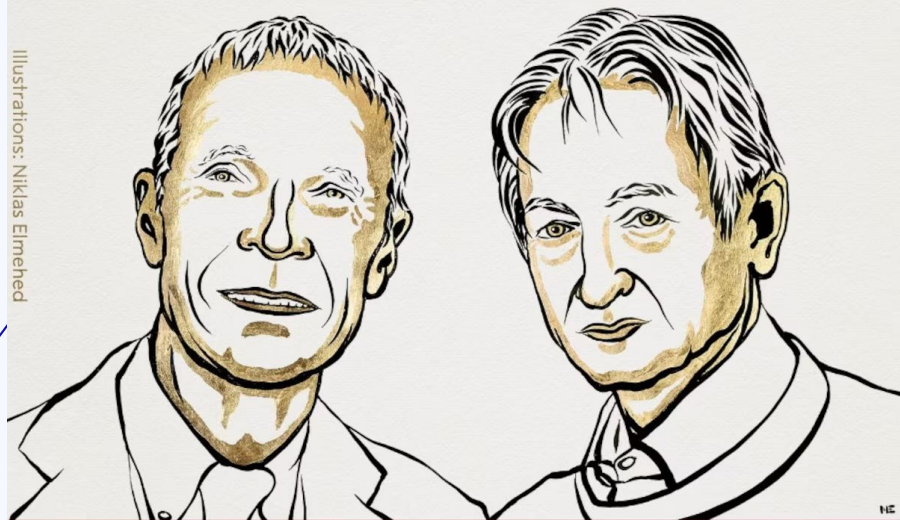
48 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp

THE NOBEL PRIZE IN PHYSICS 2024

Illustrations: Niklas Elmehed



John J. Hopfield

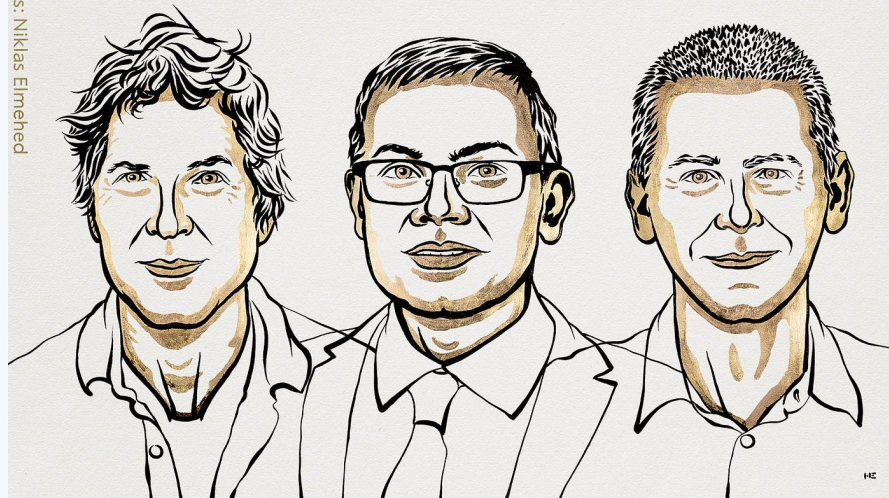
Geoffrey E. Hinton

"for foundational discoveries and inventions
that enable machine learning
with artificial neural networks"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

THE NOBEL PRIZE IN CHEMISTRY 2024

Illustrations: Niklas Elmehed



**David
Baker**

**Demis
Hassabis**

**John M.
Jumper**

"for computational
protein design"

"for protein structure prediction"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

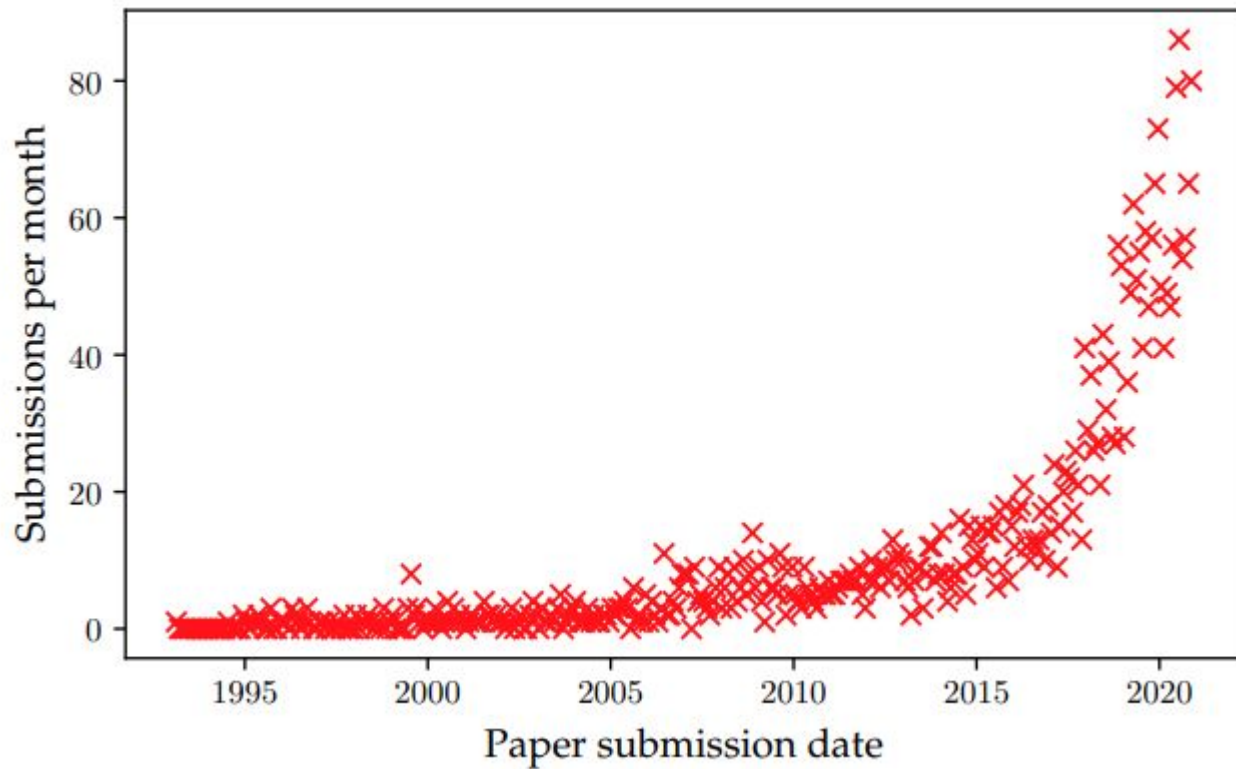


02

IA na Astronomia

Um contexto histórico





Quantidade de submissões por MÊS no arXiv:astro-ph cujos resumos contêm os termos "machine learning", "ML", "artificial intelligence", "AI", "deep learning", ou "neural network" (Smith & Geach, 2023)

IA na Astronomia

(Smith & Geach)



1ª onda 1990~2010

Surgem os primeiros levantamentos do céu (ex: SDSS)



2ª onda 2010~

Ascensão das redes neurais convolucionais e redes neurais recorrentes



3ª onda hoje

Deep learning auto-supervisionado e generativo

IA na Astronomia

(Smith & Geach)



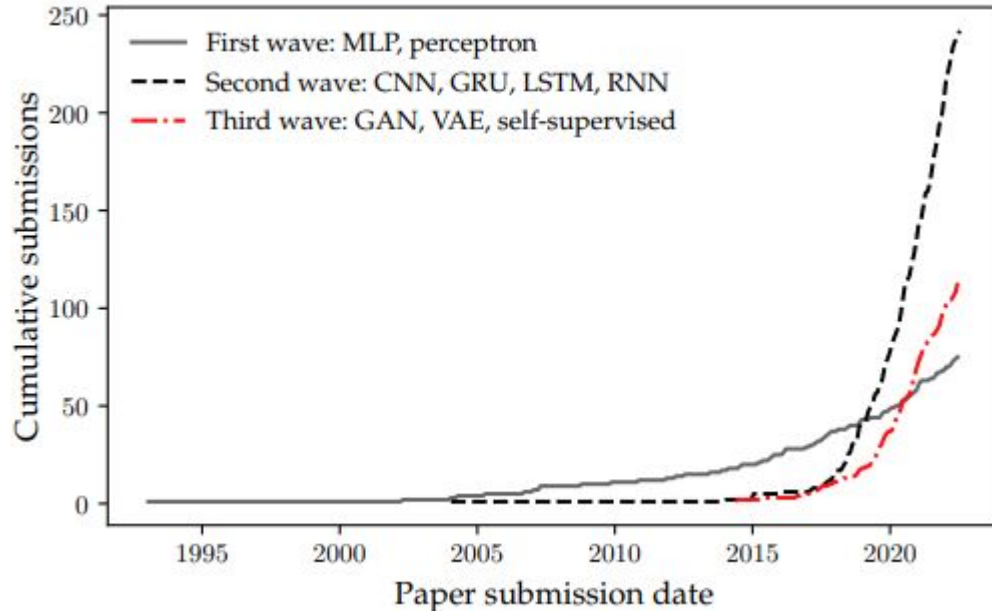
1ª onda

Surgem
levantam
(ex: SDS



3ª onda

Deep lea
auto-su
generat



is e

IA na Astronomia

(Smith & Geach)



1ª onda 1990~2010

Surgem os primeiros levantamentos do céu (ex: SDSS)



3ª onda hoje

Deep learning auto-supervisionado e generativo



2ª onda 2010~

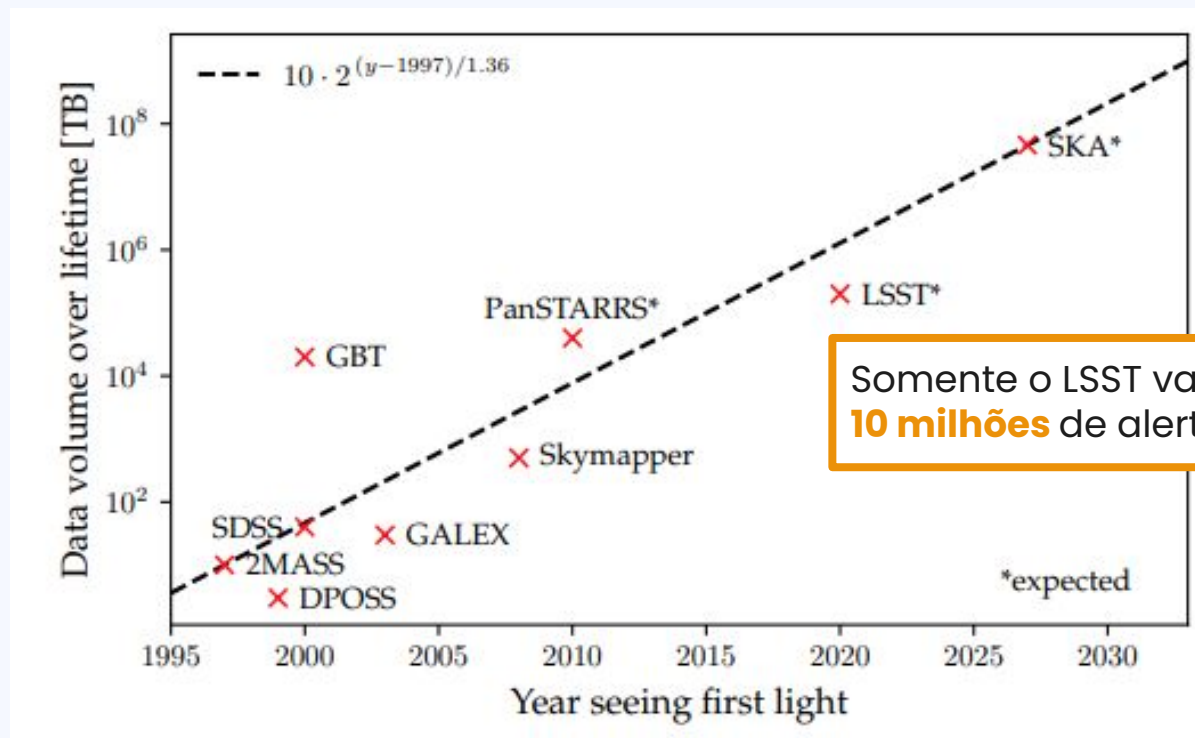
Ascensão das redes neurais convolucionais e redes neurais recorrentes



... próxima onda?

Foundation models (ex: GPT)

E será que estamos preparados?



Volume total de dados de alguns levantamentos do céu (Smith & Geach, 2023)

03

Fundamentos

Mostrando o fundamento matemático por trás do conceito de “aprendizado” algorítmico em poucos slides...

Linear Regression

Consider an one-dimensional input and output, i.e., $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in \mathbb{R}^2, i = 1, 2, \dots, N$.

We want to find a function f that relates each $\mathbf{x}^{(i)}$ to $\mathbf{y}^{(i)}$ in the **best way possible**:

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

A family of hypothetical functions $h: \mathbb{R} \rightarrow \mathbb{R}$ can be written as:

$$h_{\mathbf{w}}(x) = w_0 + w_1 x$$

How can we find w_0 and w_1 ?

Answer: by minimizing a loss function \mathcal{L}

$$\mathcal{L}(w_0, w_1) = \sum_{i=1}^N (w_0 + w_1 x^{(i)} - y^{(i)})^2$$

Linear Regression: Analytical Solution (OLS)

Minimizing a loss function \mathcal{L} means taking its derivative:

$$\hat{w} = \arg \min_w \mathcal{L}(w_0, w_1)$$

For a 2-dimensional case:

$$\begin{aligned}\hat{w}_0 &= \bar{Y} - w_1 \bar{X} \\ \hat{w}_1 &= \frac{\sum_{i=1}^N (x^{(i)} - \bar{X})(y^{(i)} - \bar{Y})}{\sum_{i=1}^N (x^{(i)} - \bar{X})^2}\end{aligned}$$

Analytical solution for any size of w (Ordinary Least Squares estimator):

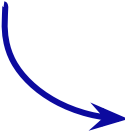
$$\hat{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

There are some advantages of having this analytical solution, but I won't dive into these details here!

Numerical solution (Gradient descent)

Why do we care about a numerical solution for linear regression if we have an analytical one?

$$\hat{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



inverting $\mathbf{X}^T \mathbf{X}$ can be very computationally expensive! OLS works well for small datasets

Plus, OLS only works for linear models. If you have very large dataset, many features and wants to fit a nonlinear model, you need a numerical solution method.

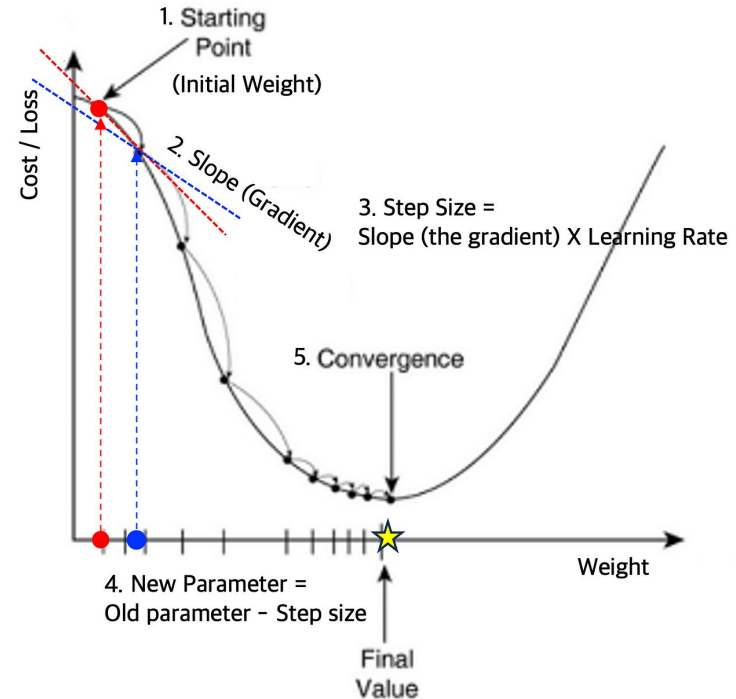
This will be very important for neural networks!

Visualization of Gradient Descent (1-dimensional)

Imagine that the curve in this illustration is the loss function

$$\mathcal{L}(w_0, w_1) = \sum_{i=1}^N (w_0 + w_1 x^{(i)} - y^{(i)})^2$$

The idea of applying a numerical solution method is to find the minimum by iterative calculations



Numerical solution (Gradient descent)

Iterative algorithm:

- Initialize parameters $\mathbf{w}(0)$ with randoms
- Compute the gradient of \mathcal{L} , i.e. $\nabla \mathcal{L}(\mathbf{w}(0))$
- Update the parameters \mathbf{w} as follows:

$$\mathbf{w}(1) = \mathbf{w}(0) - \eta \nabla \mathcal{L}(\mathbf{w}(0))$$

↖ learning rate: arbitrary value

- Repeat until convergence

Numerical solution (Gradient descent)

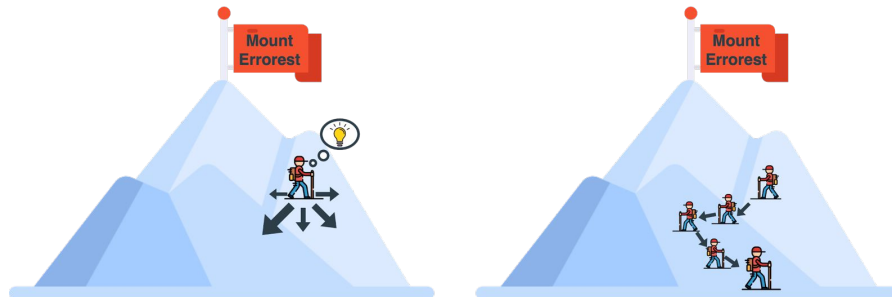
Iterative algorithm:

- Initialize parameters $\mathbf{w}(0)$ with randoms
- Compute the gradient of \mathcal{L} , i.e. $\nabla \mathcal{L}(\mathbf{w}(0))$
- Update the parameters \mathbf{w} as follows:

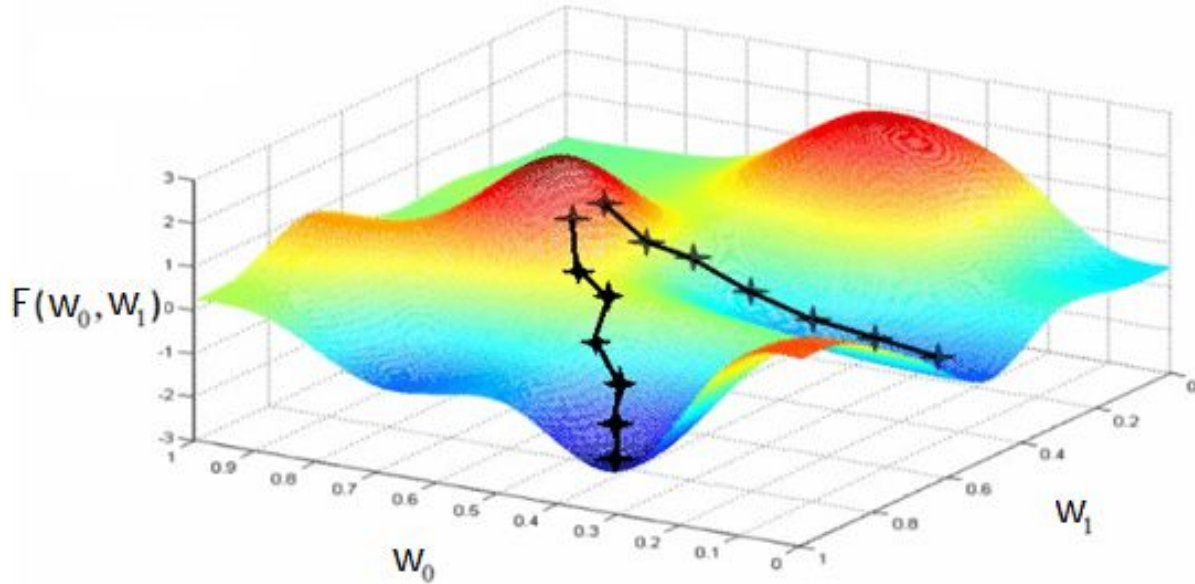
$$\mathbf{w}(1) = \mathbf{w}(0) - \eta \nabla \mathcal{L}(\mathbf{w}(0))$$

↖ learning rate: arbitrary value

- Repeat until convergence



Visualization of Gradient Descent (2-dimensional)



Source:

https://www.researchgate.net/publication/322203555_Memristive_crossbar_arrays_for_machine_learning_systems

Em resumo...

Algoritmos de machine learning e deep learning são, em essência:

- Minimização de uma função de perda
- Operações matriciais (e tensoriais) em alta dimensão

GPUs são ótimas pra isso!



04

Aplicações

Um pouco do meu trabalho

Tipos de problemas

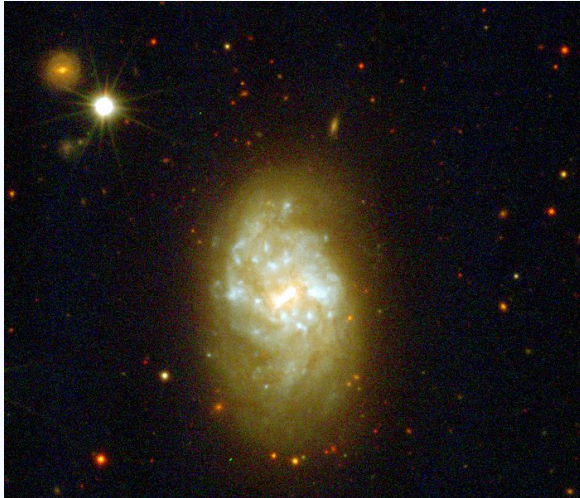
Classificação

Regressão

Tipos de problemas

Classificação

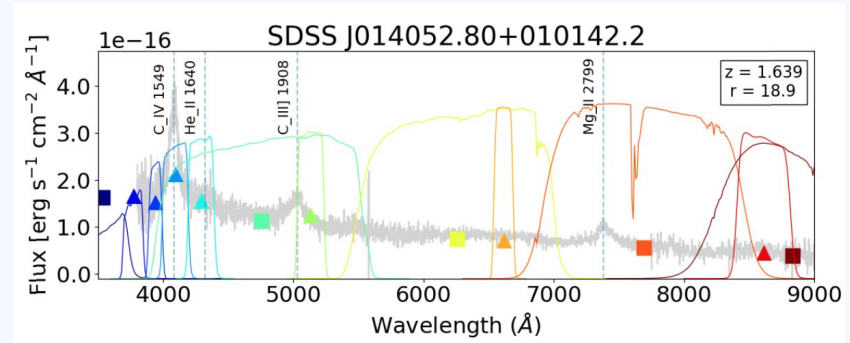
Exemplo: Que objeto celeste é este?



$Y = \{\text{star, galaxy, quasar, ...}\}$
Valores discretos

Regressão

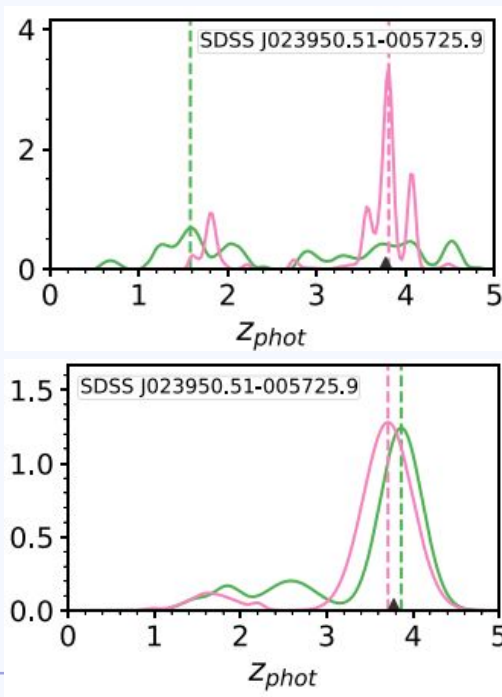
Exemplo: Conseguimos estimar redshift apenas com dados fotométricos?



$Y = \mathbb{R}$
Valores contínuos

Regressão

- Estimativa de redshift fotométrico de quasares usando redes neurais bayesianas e FlexCoDE



Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **531**, 327–339 (2024)
Advance Access publication 2024 April 9

<https://doi.org/10.1093/mnras/stae971>

The Quasar Catalogue for *S-PLUS* DR4 (QuCatS) and the estimation of photometric redshifts

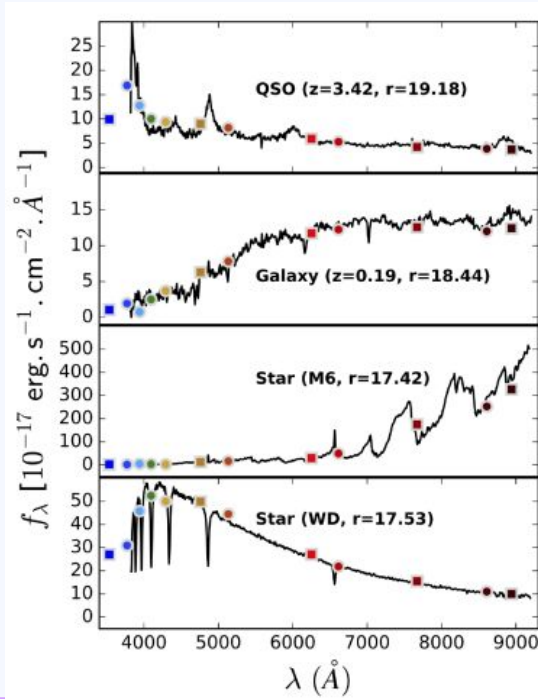
L. Nakazono^{1,2,3}, R. R. Valença¹, G. Soares³, R. Izbic³, Ž. Ivezić², E. V. R. Lima¹, N. S. T. Hirata⁴, L. Sodré Jr¹, R. Overzier⁵, F. Almeida-Fernandes^{1,6}, G. B. Oliveira Schwarz⁷, W. Schoenell⁸, A. Kanaan⁹, T. Ribeiro¹⁰ and C. Mendes de Oliveira¹

¹Universidade de São Paulo, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Departamento de Astronomia, São Paulo, SP 05508-090, Brazil
²Department of Astronomy and DiRAC Institute, University of Washington, Box 351580, Seattle, WA 98195, USA
³Universidade Federal de São Carlos, Departamento de Estatística, São Carlos, SP 13565-905, Brazil
⁴Universidade de São Paulo, Instituto de Matemática e Estatística, Departamento de Ciência da Computação, SP 05508-090, São Paulo, Brazil
⁵Observatório Nacional / MCTIC, Rua General José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil
⁶NSF's NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA
⁷Universidade Presbiteriana Mackenzie, R. da Consolação 930 - Consolação, São Paulo, SP 01302-907, Brazil
⁸GMTO Corporation 465 N. Halstead Street, Suite 250 Pasadena, CA 91107, USA
⁹Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC 88040-900, Brazil
¹⁰Rubin Observatory Project Office, 950 N. Cherry Ave., Tucson, AZ 85719, USA

Accepted 2024 March 26. Received 2024 March 25; in original form 2023 September 27

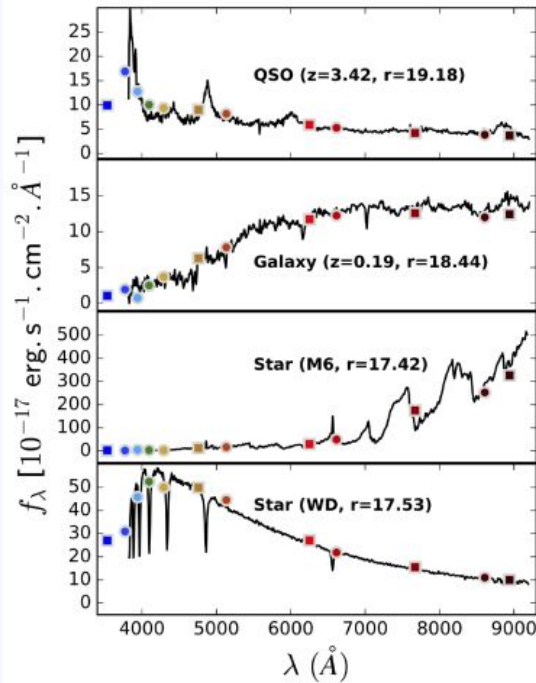
Classificação

- Estrela/quasar/galáxia para o **S-PLUS** desde 2020 usando Random Forest



Classificação

- Estrela/quasar/galáxia para o **S-PLUS** desde 2020 usando Random Forest



Monthly Notices

ROYAL ASTRONOMICAL SOCIETY

MNRAS **507**, 5847–5868 (2021)

Advance Access publication 2021 July 12

<https://doi.org/10.1093/mnras/stab1835>

On the discovery of stars, quasars, and galaxies in the Southern Hemisphere with S-PLUS DR2

L. Nakazono^{1,★}, C. Mendes de Oliveira,¹ N. S. T. Hirata,² S. Jeram,³ C. Queiroz,⁴ Stephen S. Eikenberry,³ A. H. Gonzalez,³ R. Abramo,⁴ R. Overzier,⁵ M. Espadoto,² A. Martinazzo,² L. Sampedro,¹ F. R. Herpich,¹ F. Almeida-Fernandes,¹ A. Werle,^{1,6} C. E. Barbosa,¹ L. Sodré Jr.,¹ E. V. Lima,¹ M. L. Buzzo,¹ A. Cortesi,⁷ K. Menéndez-Delmeire,⁷ S. Akras,⁸ Alvaro Alvarez-Candal,^{5,9,10} A. R. Lopes,⁵ E. Telles,⁵ W. Schoenell,¹¹ A. Kanaan¹² and T. Ribeiro¹³

¹Instituto de Astronomia, Geofísica e Ciências Atmosféricas da U. de São Paulo, Cidade Universitária, 05508-900 São Paulo, SP, Brazil

²Departamento de Ciência da Computação, Instituto de Matemática e Estatística da USP, Cidade Universitária, 05508-090 São Paulo, SP, Brazil

³Department of Astronomy, University of Florida, 211 Bryant Space Center, Gainesville, FL 32611, USA

⁴Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, SP, Rua do Matão 1371 São Paulo, Brazil

⁵Observatório Nacional / MCTIC, Rua General José Cristino 77, 20921-400 Rio de Janeiro, RJ, Brazil

⁶INAF – Osservatorio Astronomico di Padova, Vicolo Osservatorio 5, I-35122 Padova, Italy

⁷Osservatorio do Valongo, Universidade Federal do Rio de Janeiro, Ladeira Pedro Antônio 43, Saide, CEP 20080-090 Rio de Janeiro, RJ, Brazil

⁸Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Penteli GR 15236, Greece

⁹Instituto de Astrofísica de Andalucía, CSIC, Apt 3004, E-18080 Granada, Spain

¹⁰IUFACyT, Universidad de Alicante, San Vicent del Raspeig, E-03080 Alicante, Spain

¹¹GMTO Corporation, 465 N. Halstead Street, Suite 250 Pasadena, CA 91107, USA

¹²Departamento de Física, Universidade Federal de Santa Catarina, 88040-900 Florianópolis, SC, Brazil

¹³NOAO, P.O. Box 26732, Tucson, AZ 85726, USA

Accepted 2021 June 18. Received 2021 June 16; in original form 2020 July 29

Relembrando um dos primeiros slides...

IA na Astronomia

(Smith & Geach)



1ª onda 1980~2000

Surgem os primeiros levantamentos do céu (ex: SDSS)



3ª onda hoje

Deep learning auto-supervisionado e generativo



2ª onda 2000~

Ascensão das redes neurais convolucionais e redes neurais recorrentes



... próxima onda?

Foundation models (ex: GPT)

Classificação

Testamos se **Deep Learning auto-supervisionado** treinado com as imagens nos daria uma performance melhor que o modelo de Random Forest para o nosso problema

| Experimento | Acurácia (%) | F-Score (%) |
|-------------|-----------------------------------|-----------------------------------|
| NW_RF | 96.87 \pm 0.05 | 95.38 \pm 0.13 |
| WW_RF | 98.44 \pm0.03 | 97.78 \pm 0.05 |
| UF_RF | 98.43 \pm 0.04 | 97.79 \pm0.08 |
| → CNN | 95.45 \pm 0.18 | 94.01 \pm 0.24 |

... e concluímos que não.

Combinação de Dados Tabulares e Imagens para a Classificação de Objetos Astronômicos

G. Jacob Perin*, L. Nakazono[†], C. Mendes de Oliveira[†] e N. S. T. Hirata*

*Instituto de Matemática e Estatística

[†]Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Universidade de São Paulo

Resumo—No contexto do S-PLUS (Southern Photometric Local Universe Survey), um projeto de imageamento do céu do hemisfério sul em 12 bandas fotométricas, foram desenvolvidos anteriormente métodos baseados em aprendizado de máquina para a classificação de três tipos de objetos (estrelas, galáxias e quasares). Um dos trabalhos utilizou dados de catálogo, incluindo objetos com dados faltantes, enquanto o outro utilizou imagens, sem objetos com dados faltantes. Neste trabalho apresentamos dois avanços: avaliamos os dois métodos em condições de igualdade e propomos a utilização de técnicas de *ensemble* para combinar os dois tipos de classificadores. Experimentos realizados com o quarto *Data Release* do S-PLUS mostram que o *ensemble* proposto supera ambos os métodos anteriores com respeito aos objetos mais difíceis de serem classificados.

Abstract—In the context of S-PLUS (Southern Photometric Local Universe Survey), a 12-band photometric sky survey of the southern hemisphere sky, two machine learning based methods were previously developed for the classification of three types of objects (stars, galaxies and quasars). One of these works has used catalog data, including objects with missing information, while the other has used images, removing objects with missing information. In this work we present two advances: we evaluate the two methods under equal conditions and we propose the use of *ensemble* techniques to combine the two types of classifiers. Experiments performed with the fourth *Data Release* of the S-PLUS show that the proposed *ensemble* outperforms both previous methods with respect to the most difficult objects to be classified.

I. INTRODUÇÃO

porém utilizando uma seleção de dados mais conservadora e restritiva [3].

Neste trabalho, comparamos ambos os métodos propostos em condições de igualdade e propomos um *ensemble* que combina classificadores que utilizam dados tabulares e classificadores que utilizam imagens. Os resultados indicam que classificadores que utilizam dados de catálogo são superiores aos que utilizam imagens, e que o *ensemble* apresenta desempenho superior para os objetos mais difíceis de serem classificados do S-PLUS¹.

Nas seções a seguir, explicamos inicialmente alguns conceitos de astronomia. Em seguida, descrevemos a metodologia adotada (seleção dos dados e treinamento dos modelos) e os resultados obtidos. Por fim, apresentamos as conclusões.

II. CONCEITOS DE ASTRONOMIA

Diversos tipos de objetos podem ser observados no céu. Nesse trabalho, três tipos de objetos são considerados para a classificação: **estrelas**, **galáxias** e **quasares**.

Nesse contexto, duas formas de se realizar medições desses objetos são chamadas de **espectroscopia** - que gera um espectro relacionando fluxo luminoso com comprimento de onda - e **fotometria**, que consiste em capturar imagens de comprimentos de onda específicos. Enquanto a primeira carrega mais informações e é mais demorada de ser realizada, a segunda é mais rápida, porém torna a classificação dos objetos mais

<https://doi.org/10.5753/sibgrapi.est.2023.27472>
Conference on Graphics, Patterns and Images
(SIBGRAPI)

Deep Learning **nem sempre** **é** a melhor alternativa!



Às vezes é como matar uma formiga com uma bazuca!

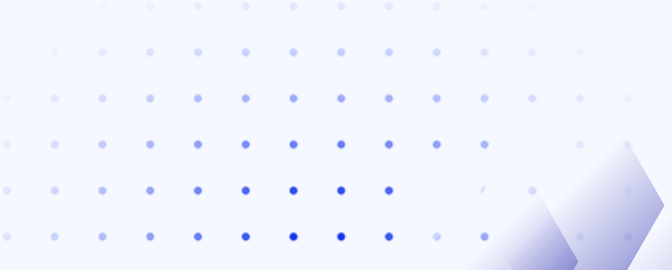
Exige um bom conhecimento do problema e dos dados para escolher algoritmos ou métodos de análise

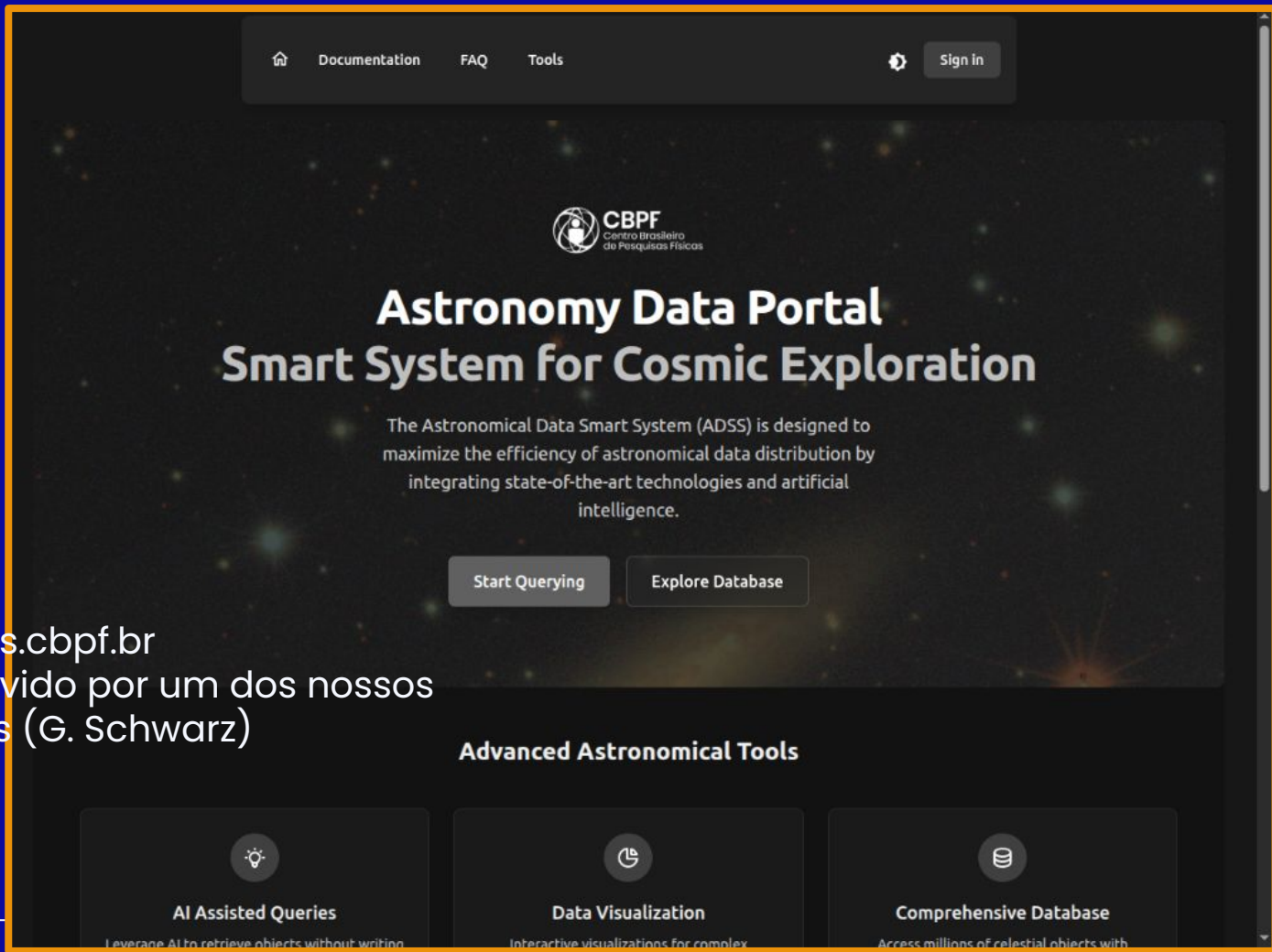
“Domain expertise” **sempre** vai ser necessário



Conteúdo extra Iniciativas

Grupos de pesquisas interdisciplinares no Brasil com ênfase
em Astronomia





www.adss.cbpf.br

Desenvolvido por um dos nossos
membros (G. Schwarz)



[COIN](#)[Projects](#)[Residence Programs ▾](#)[Focus ▾](#)[Community](#)[Highlights](#)

The Cosmostatistics Initiative

The Cosmostatistics Initiative (COIN) is a worldwide endeavor aimed to create an interdisciplinary community around data-driven problems in Astronomy.

It was designed to promote innovation in all aspects of academic scientific research.

The group is co-lead by **Rafael S. de Souza** (University of Hertfordshire, UK), **Emille E. O. Ishida** (Universite Clermont Auvergne, France) and **Alberto Krone-Martins** (University of California Irvine, USA).

Researchers willing to join are welcome to contact any of the chairs!

<https://cosmostatistics-initiative.org/>

ELEPHANT: Extragalactic aLErt Pipeline for Hostless AstroNomical Transients

P. J. Pessi¹*, R. Durgesh², L. Nakazono³, E. E. Hayes⁴, R. A. P. Oliveira⁵, E. E. O. Ishida⁶,
A. Moitinho⁷, A. Krone-Martins⁸, B. Moews^{9,10}, R. S. de Souza¹¹, R. Beck¹²,
M. A. Kuhn¹¹, K. Nowak¹¹, and S. Vaughan¹³ (for the COIN collaboration)

¹ The Oskar Klein Centre, Department of Astronomy, Stockholm University, AlbaNova 106 91, Stockholm, Sweden

² Independent Researcher, Ingolstadt, Germany

³ Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP, 05508-900, São Paulo, Brazil

⁴ Institute of Astronomy and Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK

⁵ Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland

⁶ Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France

⁷ CENTRA, Universidade de Lisboa, FCUL, Campo Grande, Edif. C8, 1749-016 Lisboa, Portugal

⁸ Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA

⁹ Business School, University of Edinburgh, 29 Buccleuch Pl, Edinburgh, EH8 9JS, UK

¹⁰ Centre for Statistics, University of Edinburgh, Peter Guthrie Tait Rd, Edinburgh, EH9 3FD, UK

¹¹ Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK

¹² Independent Researcher, Budapest, Hungary

¹³ School of Mathematical and Physical Sciences, Macquarie University, NSW 2109, Australia



Curso de IA em Astronomia e Geofísica
Quando: início de 2026

Fiquem de olho nos anúncios :)

CIAC-ON

Esforço da diretoria, da divisão de tecnologia da informação e de pesquisadores da Astronomia e Geofísica computacional para a criação de um **centro de inteligência artificial do Observatório Nacional**





Conteúdo extra

Ferramentas úteis

Programming Languages



Up to date, it is the most used programming language for data science in Astronomy



In Astronomy, we use **ADQL** (*Astronomical Data Query Language*) to query structured data. Syntax is very similar to SQL

Python packages

Astronomy-specific



Data Visualization



Data manipulation



Mathematical computing



Image manipulation



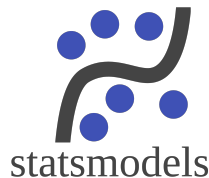
OpenCV



Machine Learning



Statistical Modeling



Deep Learning



Python packages

Astronomy-specific



Data manipulation



Image manipulation



For distributed and parallel computing:
(essential for dealing with big data!)



dask

matplotlib



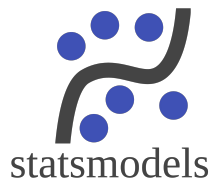
seaborn



Machine Learning



Statistical Modeling



Deep Learning



Tools: code version control



Web-based platforms that hosts Git repositories:



Tools: code editor

My personal recommendation:



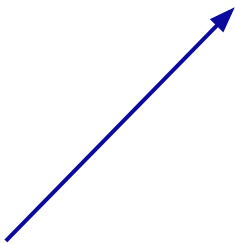
Visual Studio Code

With these extensions installed: Python, Jupyter, Remote - SSH, GitHub Copilot (plus others... Those make your programming life much much easier, trust me!)

Programming best practices

Your data science project starts with setting up code environment (e.g. pyenv, conda) and properly organizing your code files!

One of my repositories



```
> config
> data
> img
> logs
✓ src
  > __pycache__
  > evaluation
  > experiments
  > models
  > notebooks
  > preprocess
  > production
  > scripts
  > utils
  📄 __init__.py
  📄 .gitignore
  ! environment.yml
  📄 LICENSE
  ⓘ README.md
  📄 setup.py
  📄 stilts.jar
```



Obrigada!

Perguntas?

Email: liliannenakazono@on.br

Website: marixko.github.io

GitHub: <https://github.com/marixko>

CREDITS: This presentation template was created by **Slidesgo**, and includes icons, infographics & images by **Freepik**

