



PROJECTOR:  
"WHERE IS  
WALDO?"





# OUTLIERS

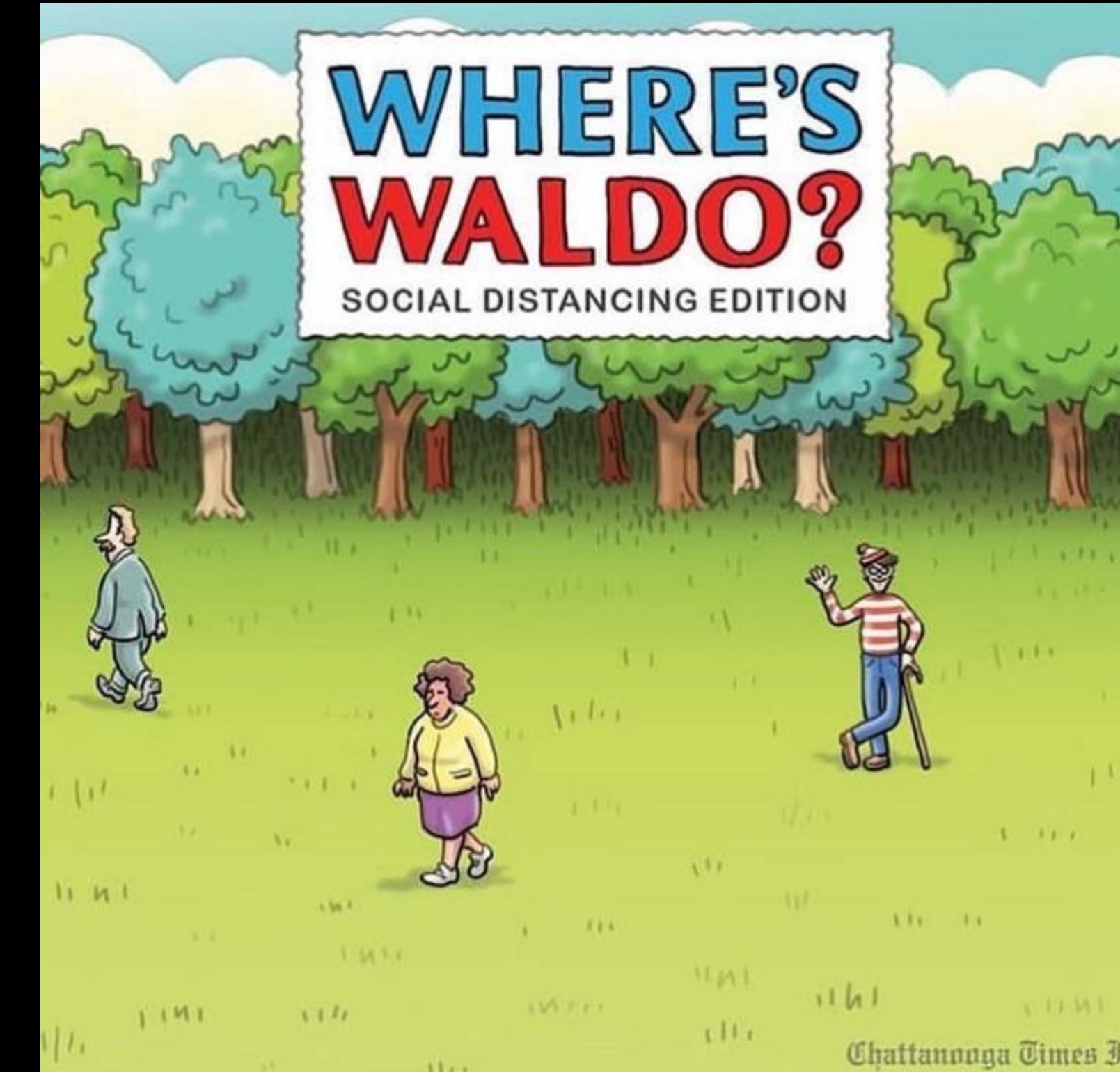
VALORES DISCREPANTES EM CONJUNTOS DE DADOS SÃO DÚBIOS: ELES PODEM, EM PROCESSOS DE REGRESSÃO OU CLASSIFICAÇÃO, RESULTAR EM AJUSTE INADEQUADO E PODER PREDITIVO RUIM PARA O MODELO. MAS, NO REINO DA ASTRONOMIA, UM OUTLIER PODE SIGNIFICAR UM OBJETO RARO OU MESMO UM FENÔMENO NUNCA ANTES DETECTADO!



## OBJETIVO

USAR TÉCNICAS DE DETECÇÃO DE OUTLIERS PARA LOCALIZAR OBJETOS RAROS, COMO GALÁXIAS EM PROCESSO DE FUSÃO, GALÁXIAS IRREGULARES E TALVEZ UM UNICÓRNIO!

COM A AJUDA  
DO PYTHON,  
TOPCATE  
ALADIN,  
ESPERAMOS  
QUE A  
DIFICULDADE  
DIMINUA...



## PASSO 01



OBTER UM CATÁLOGO CELESTE DO S-PLUS CLOUD.  
OPTEI POR UM CONTENDO 100.000 ENTRADAS EM 138 COLUNAS DA PRIMEIRA VERSÃO DE DADOS DISPONIBILIZADOS, COM AS SEGUINTESSPECIFICAÇÕES:  
A - MAG\_R < 19  
B - MAGR\_ERR < 0.1



# CATÁLOGO A SER USADO NO TOPCAT E ALADIN

ADQL Query

Query Results

Examples HERE

Schema	Table	Column
dr1	all_dr1	A
dr2		Aperture
dr2_vacs		B
dr3		Chi2
ivoa		CLASS
		Dec
		eF378_aper
		eF378_auto

ADQL Query

```
1 SELECT TOP 100000 * FROM dr1.all_dr1 WHERE (er_auto < 0.1 AND r_auto < 19)
```

Add example to query editor

Cone Search

Upload VOTable Crossmatch

Joining all tables

Format fits Execution Mode Async Upload Table ->□ Procurar... Nenhum arquivo selecionado. Submit

# CATÁLOGO A SER USADO NA BUSCA POR OUTLIERS

ADQL Query

Query Results

Examples HERE

Schema

- dr1
- dr2
- dr2\_vacs
- dr3
- ivoa

Table

- all\_dr1

Column

- A
- Aperture
- B
- Chi2
- CLASS
- Dec
- eF378\_aper
- eF378\_auto

ADQL Query

```
1 SELECT TOP 100000 uJAVA_auto, g_auto, r_auto, i_auto, z_auto, F515_auto, F660_auto, F861_auto  
FROM dr1.all_dr1 WHERE (er_auto < 0.1 AND r_auto < 19)
```

Add example to query editor

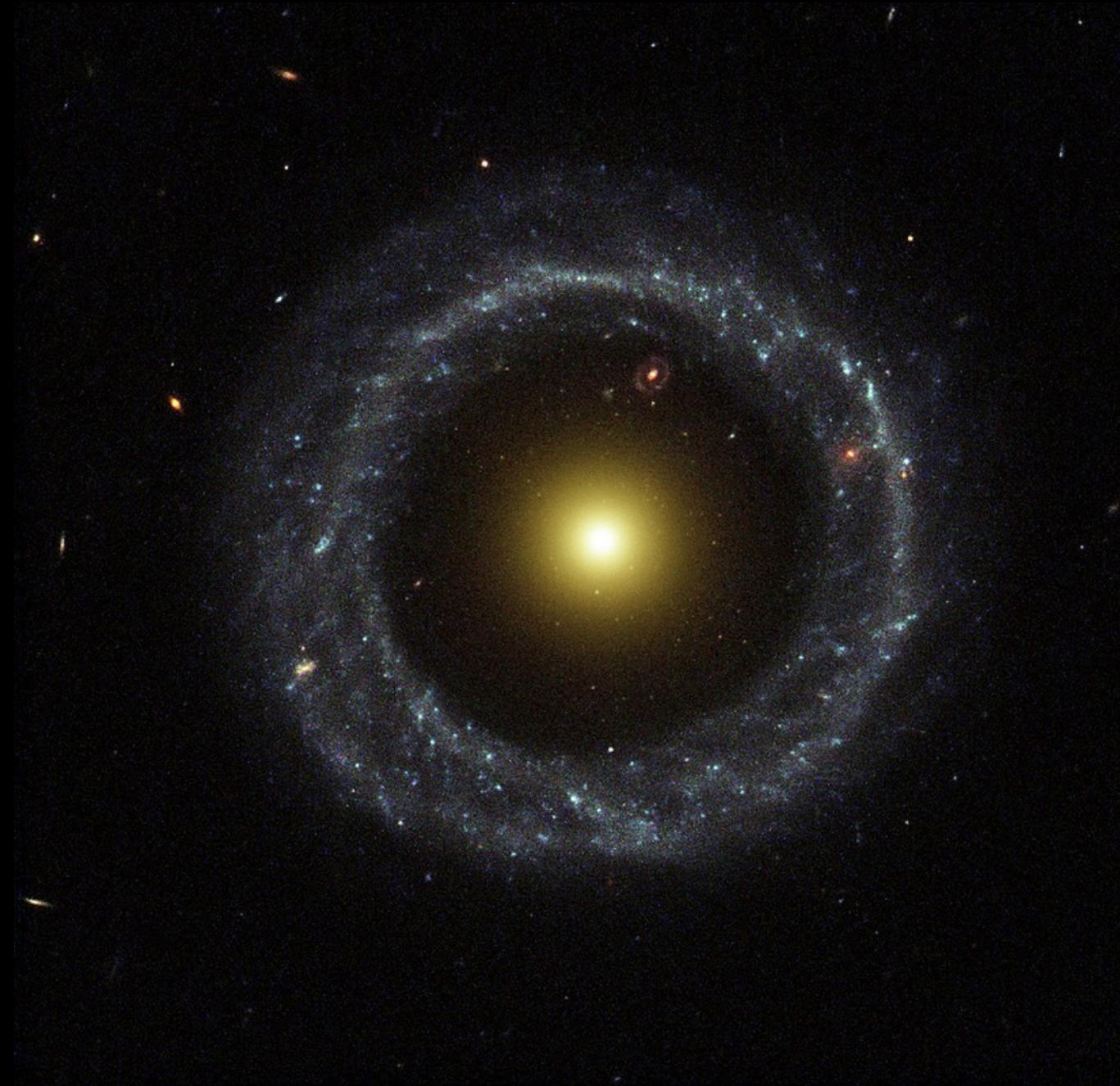
Cone Search

Upload VOTable Crossmatch

Joining all tables

Format fits ▾ Execution Mode Async ▾ Upload Table ->□ Procurar... Nenhum arquivo selecionado. Submit

# CATÁLOGO A SER USADO NA BUSCA POR OUTLIERS



100000 OBJETOS COM  
DETECÇÃO NAS  
SEGUINTE BANDAS DO  
S-PLUS: U, G, R, I, Z, FS15,  
F660, F861.

## PASSO 02



PRECISAMOS DETECTAR E ISOLAR OUTLIERS. MÉTODOS DE DETECÇÃO AUTOMÁTICA PODEM SER USADOS DURANTE O PIPELINE DE MODELAGEM EM PYTHON, COMO QUALQUER OUTRA TRANSFORMAÇÃO QUE SERÁ APLICADA AO CONJUNTO DE DADOS. NOSSO OBJETIVO É FORNECER MODELOS DE DETECÇÃO AUTOMÁTICA DE OUTLIERS COMO MÉTODOS ALTERNATIVOS ÀS TÉCNICAS ESTATÍSTICAS, E QUE SUPORTEM UM GRANDE NÚMERO DE VARIÁVEIS DE ENTRADA QUE APRESENTAM INTER-RELAÇÕES COMPLEXAS E DESCONHECIDAS.

## PALAVRAS CHAVE:

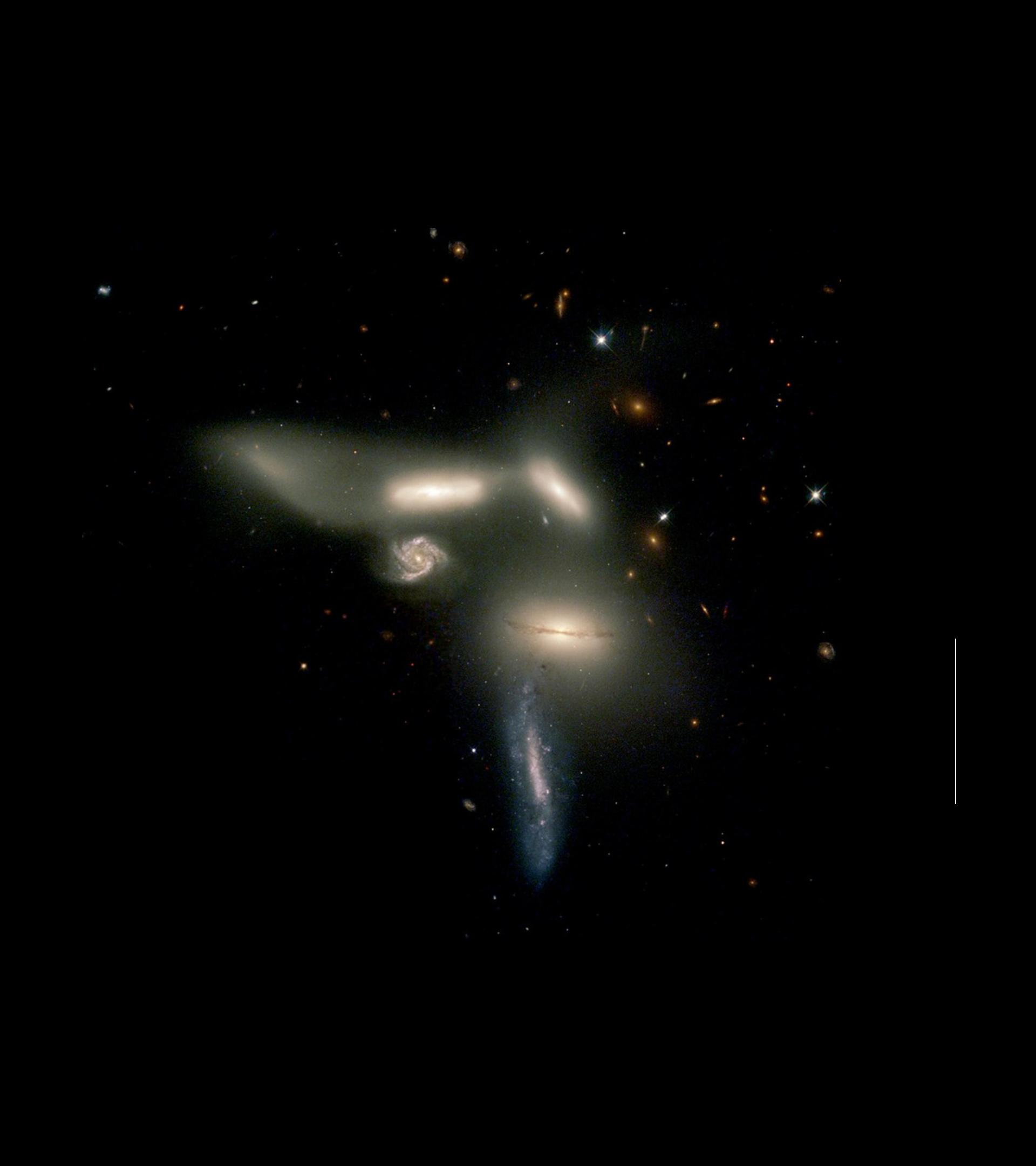
- MACHINE LEARNING:
- TREINAMENTO DE MODELO:
- EXCLUSÃO DE OUTLIERS:
- SCIKIT-LEARN LIBRARY:
- CAIXA-PRETA!
- DADOS EM QUANTIDADE!

## MÉTODO 01: ISOLATION FOREST



ISOLATION FOREST, OU IFOREST PARA ABREVIAR, É UM ALGORITMO DE DETECÇÃO DE ANOMALIAS BASEADO EM ÁRVORES DE DECISÃO.

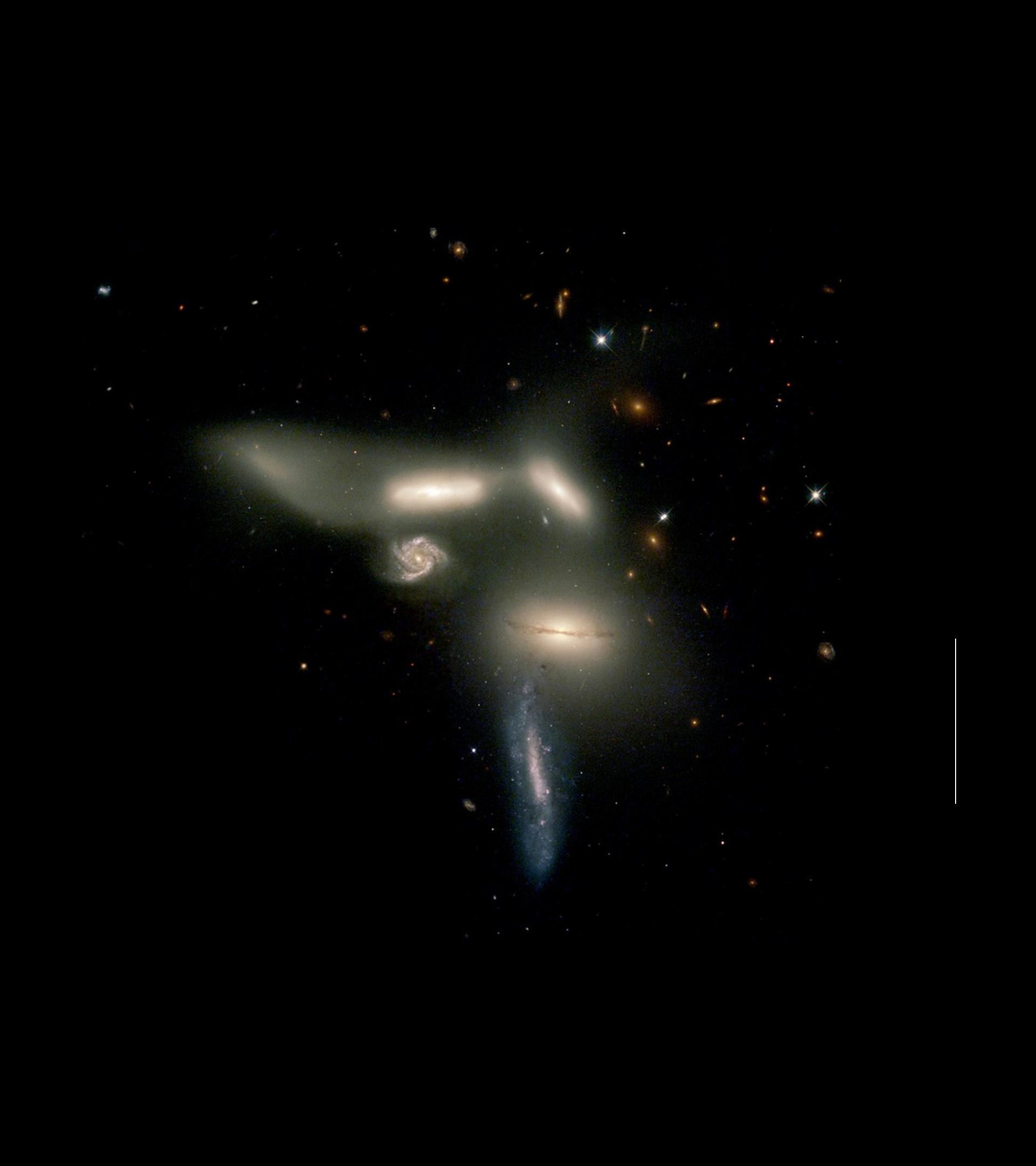
"...OUR PROPOSED METHOD TAKES ADVANTAGE OF TWO ANOMALIES' QUANTITATIVE PROPERTIES: I) THEY ARE THE MINORITY CONSISTING OF FEWER INSTANCES AND II) THEY HAVE ATTRIBUTE-VALUES THAT ARE VERY DIFFERENT FROM THOSE OF NORMAL INSTANCES."



## MÉTODO 02: MINIMUM COVARIANCE DETERMINANT

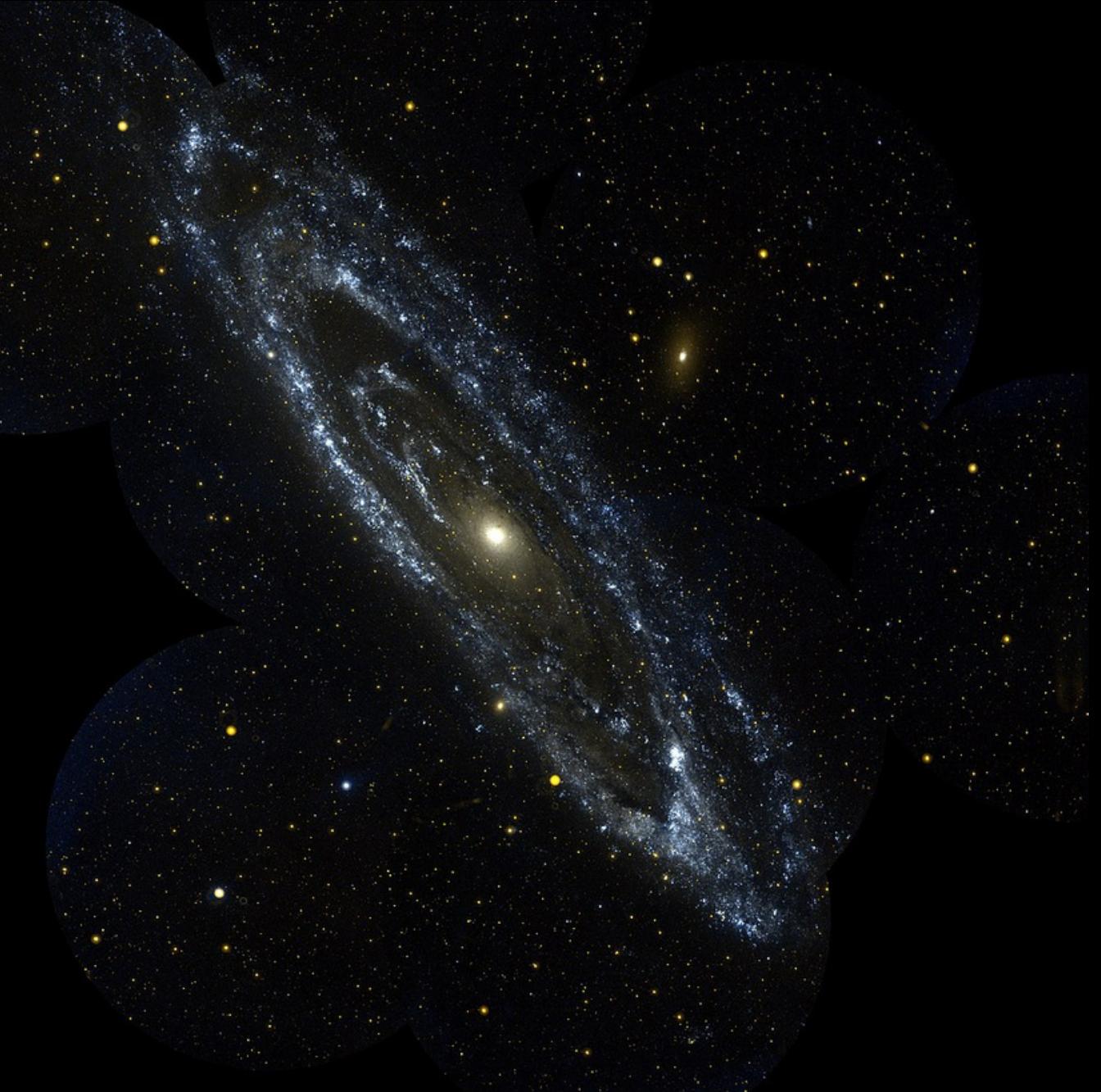
SE AS VARIÁVEIS DE ENTRADA TIVEREM UMA DISTRIBUIÇÃO GAUSSIANA, MÉTODOS ESTATÍSTICOS SIMPLES PODEM SER USADOS PARA DETECTAR OUTLIERS.

POR EXEMPLO, SE O CONJUNTO DE DADOS TEM DUAS VARIÁVEIS DE ENTRADA E AMBAS SÃO GAUSSIANAS, O ESPAÇO DE CARACTERÍSTICAS FORMA UMA GAUSSIANA MULTIDIMENSIONAL E O CONHECIMENTO DESSA DISTRIBUIÇÃO PODE SER USADO PARA IDENTIFICAR VALORES DISTANTES DA DISTRIBUIÇÃO.



## MÉTODO 02: MINIMUM COVARIANCE DETERMINANT

ESSA ABORDAGEM PODE SER GENERALIZADA DEFININDO UMA HIPERESFERA (ELIPSÓIDE) QUE COBRE OS DADOS NORMAIS, E OS DADOS QUE FICAM FORA DESSA FORMA SÃO CONSIDERADOS DISCREPANTES. UMA IMPLEMENTAÇÃO EFICIENTE DESTA TÉCNICA PARA DADOS MULTIVARIADOS É CONHECIDA COMO DETERMINANTE DE COVARIÂNCIA MÍNIMA, OU MCD.



## MÉTODO 03: LOCAL OUTLIER FACTOR

O LOCAL OUTLIER FACTOR, OU LOF PARA ABREVIAR, É UMA TÉCNICA QUE TENTA APROVEITAR A IDEIA DE VIZINHOS MAIS PRÓXIMOS PARA DETECCÇÃO DE VALORES DISCREPANTES. CADA EXEMPLO RECEBE UMA PONTUAÇÃO BASEADO EM QUÃO ISOLADO ESTÁ, OU NA PROBABILIDADE DE SER UM VALOR DISCREPANTE COM BASE NO TAMANHO DE SUA VIZINHANÇA LOCAL. OS EXEMPLOS COM A MAIOR PONTUAÇÃO TÊM MAIOR CHANCE DE SEREM OUTLIERS.

## MÉTODO 04: ONE-CLASS SVM



A SUPPORT VECTOR MACHINE, OU SVM, FOI UM ALGORITMO DESENVOLVIDO INICIALMENTE PARA CLASSIFICAÇÃO BINÁRIA, PODENDO SER USADO PARA CLASSIFICAÇÃO DE UMA CLASSE.

AO MODELAR UMA CLASSE, O ALGORITMO CAPTURA A DENSIDADE DE PROBABILIDADE DA CLASSE MAJORITÁRIA E CLASSIFICA OS EXEMPLOS NOS EXTREMOS DA FUNÇÃO DE DENSIDADE DE PROBABILIDADE COMO OUTLIERS. ESTA MODIFICAÇÃO DO SVM É CONHECIDA COMO ONE-CLASS SVM.

## PASSO 03: AÇÃO!

1 - FAZER O DOWNLOAD DO CATÁLOGO  
SIMPLIFICADO A PARTIR DO  
REPOSITÓRIO/DROPBOX.

2 - RODAR UM ALGORITMO DE REGRESSÃO  
LINEAR SOBRE OS DADOS E OBTER O MEAN  
ABSOLUTE ERROR (MAE).

## PASSO 03: AÇÃO!

3 - RODAR CADA MÉTODO SOBRE O CATÁLOGO, OBTENDO O MAE (PARA FINS DE MENSURAÇÃO DA PERFORMANCE) E OS OUTLIERS MAIS EXPRESSIVOS DO CONJUNTO TREINO E DO CONJUNTO TESTE.

4 - BUSCAR OS OUTLIERS USANDO AS FERRAMENTAS TOPCAT E ALADIN.

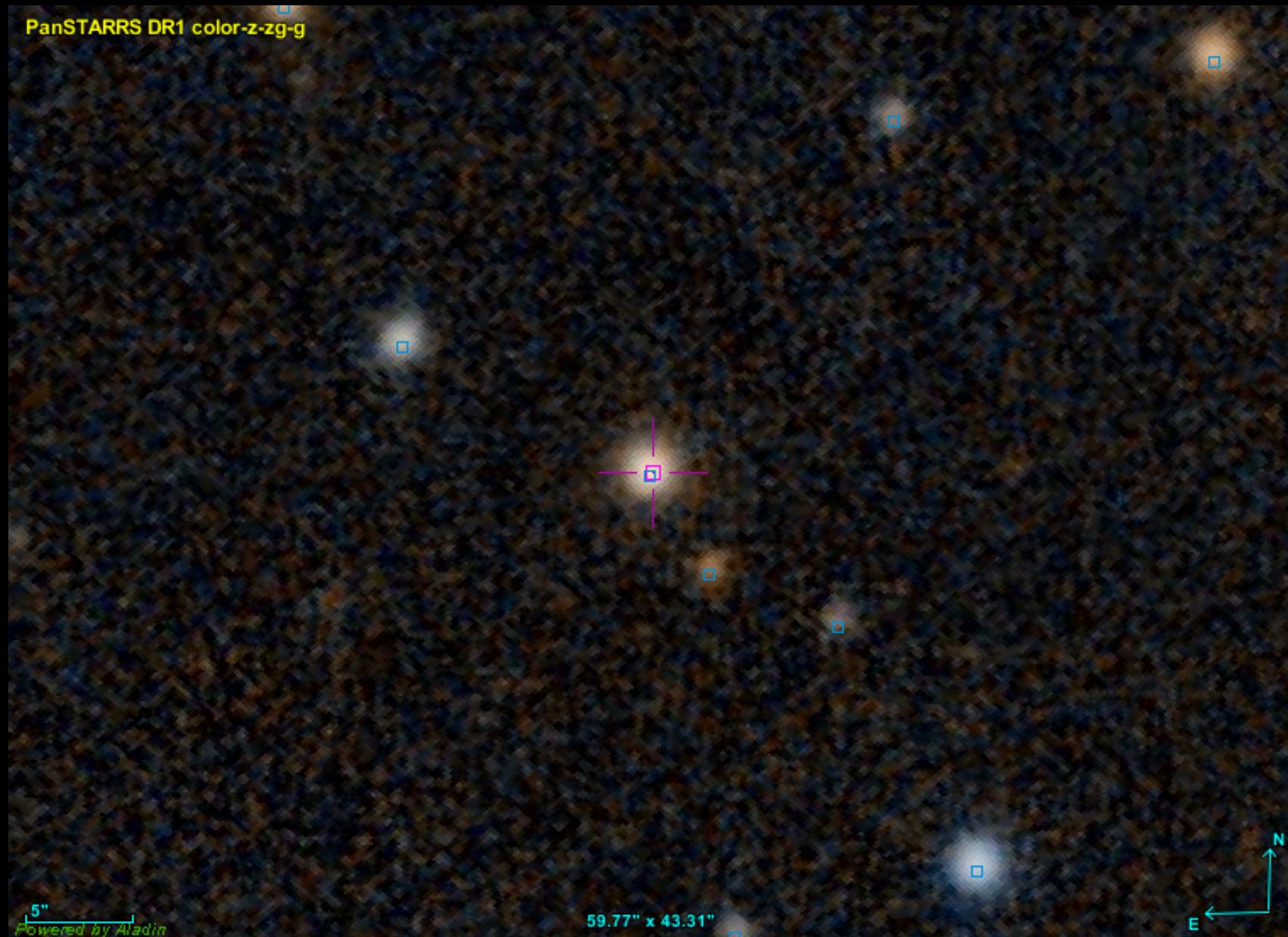
**RODADA N° 1**

	MAE	N° DE OUTLIERS RETIRADOS
REGRESSÃO LINEAR	0.775	---
ISOLATION FOREST	0.665	6700
MINIMUM COVARIANCE DETERMINANT	0.669	670
LOCAL OUTLIER FACTOR	0.758	2350
ONE-CLASS SVM	0.732	670

**RODADA N° 2**

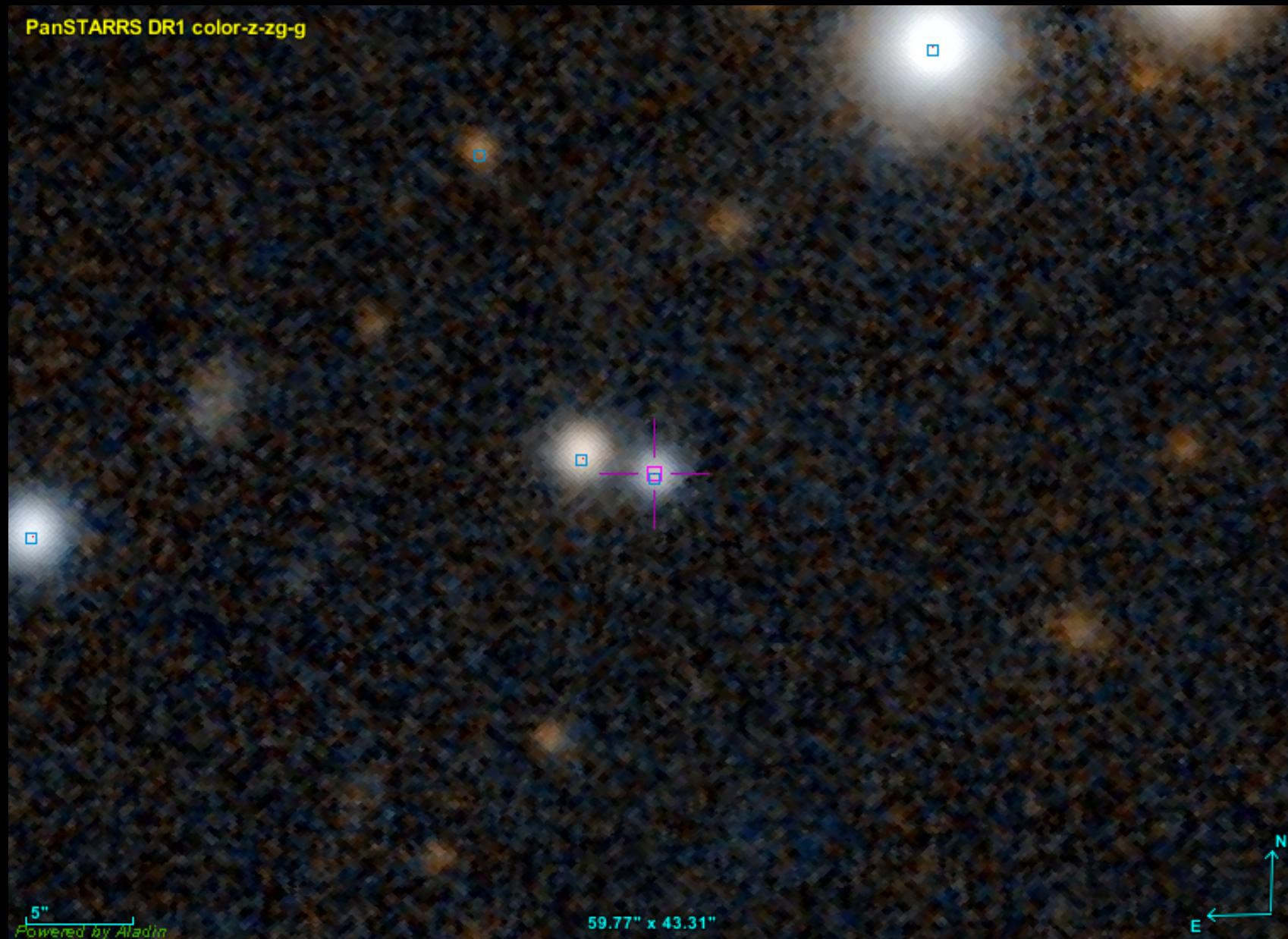
	MAE	N° DE OUTLIERS RETIRADOS
REGRESSÃO LINEAR	0.775	---
ISOLATION FOREST	0.889	6700
MINIMUM COVARIANCE DETERMINANT	0.745	670
LOCAL OUTLIER FACTOR	0.747	2318
ONE-CLASS SVM	0.731	669

# ISOLATION FOREST



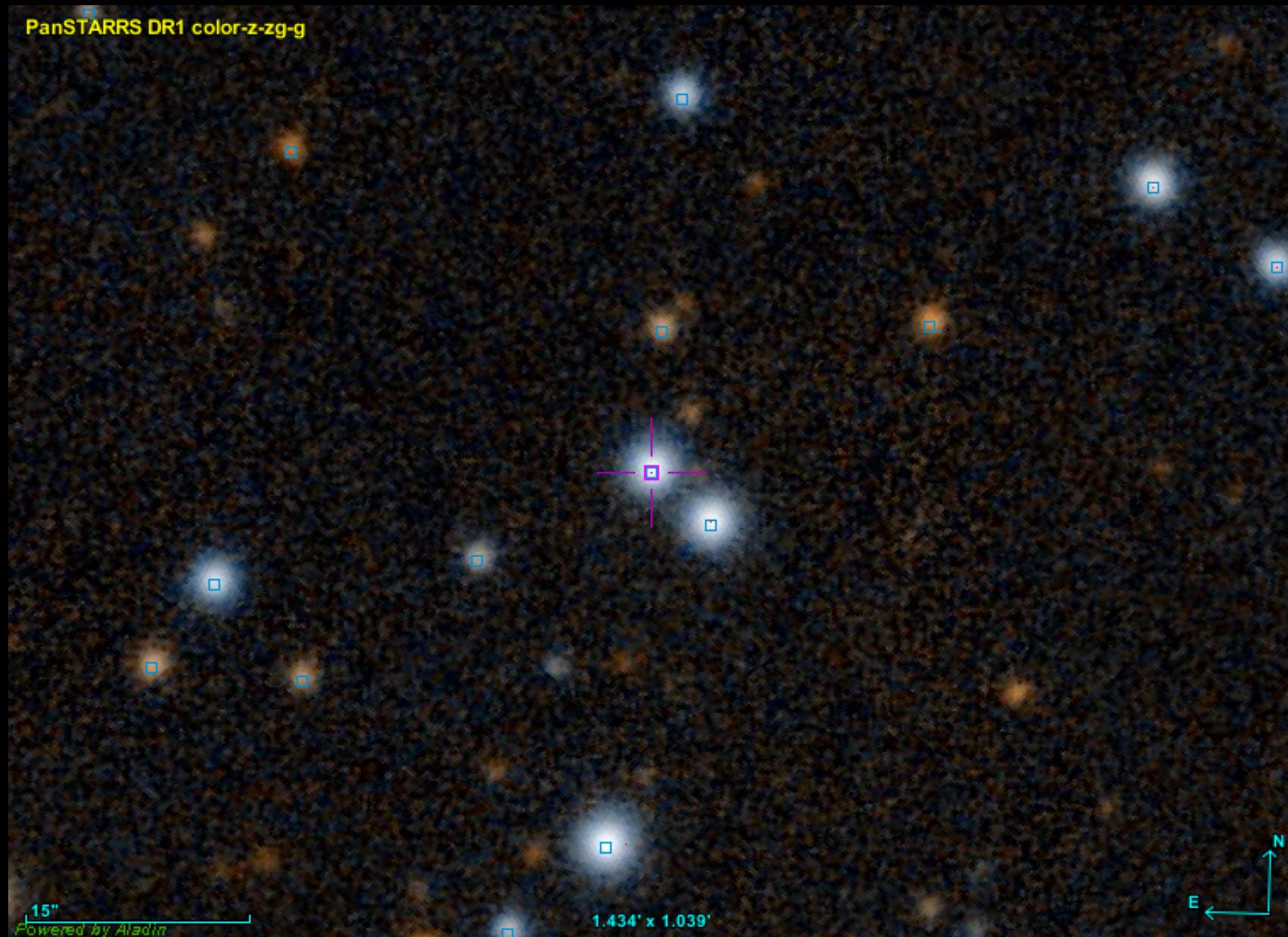
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: 2MASS J20225537-0113490  
TYPE: IRS (INFRARED SOURCE)

# ISOLATION FOREST



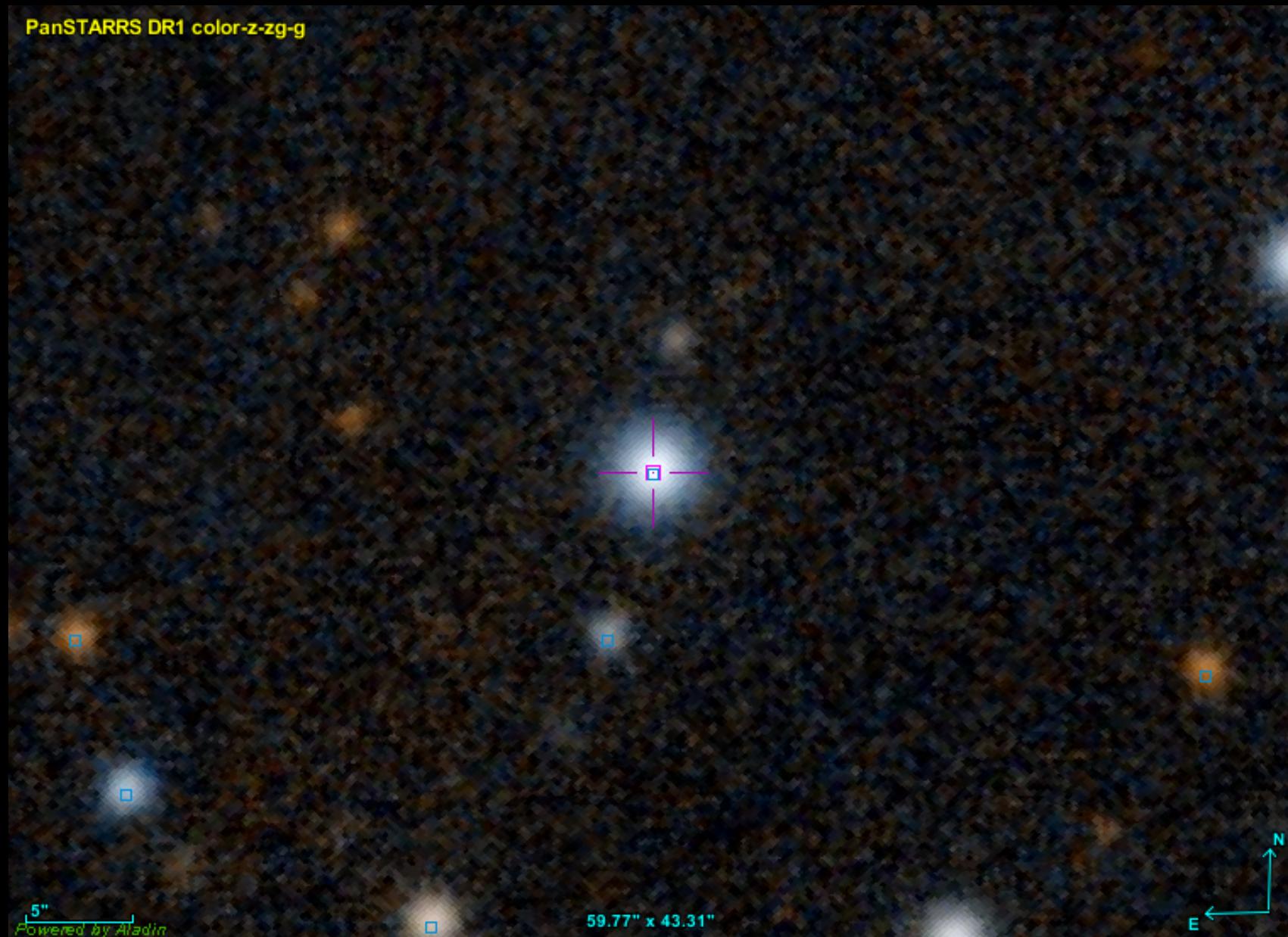
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: NÃO CATALOGADA  
TYPE: ---

# ISOLATION FOREST



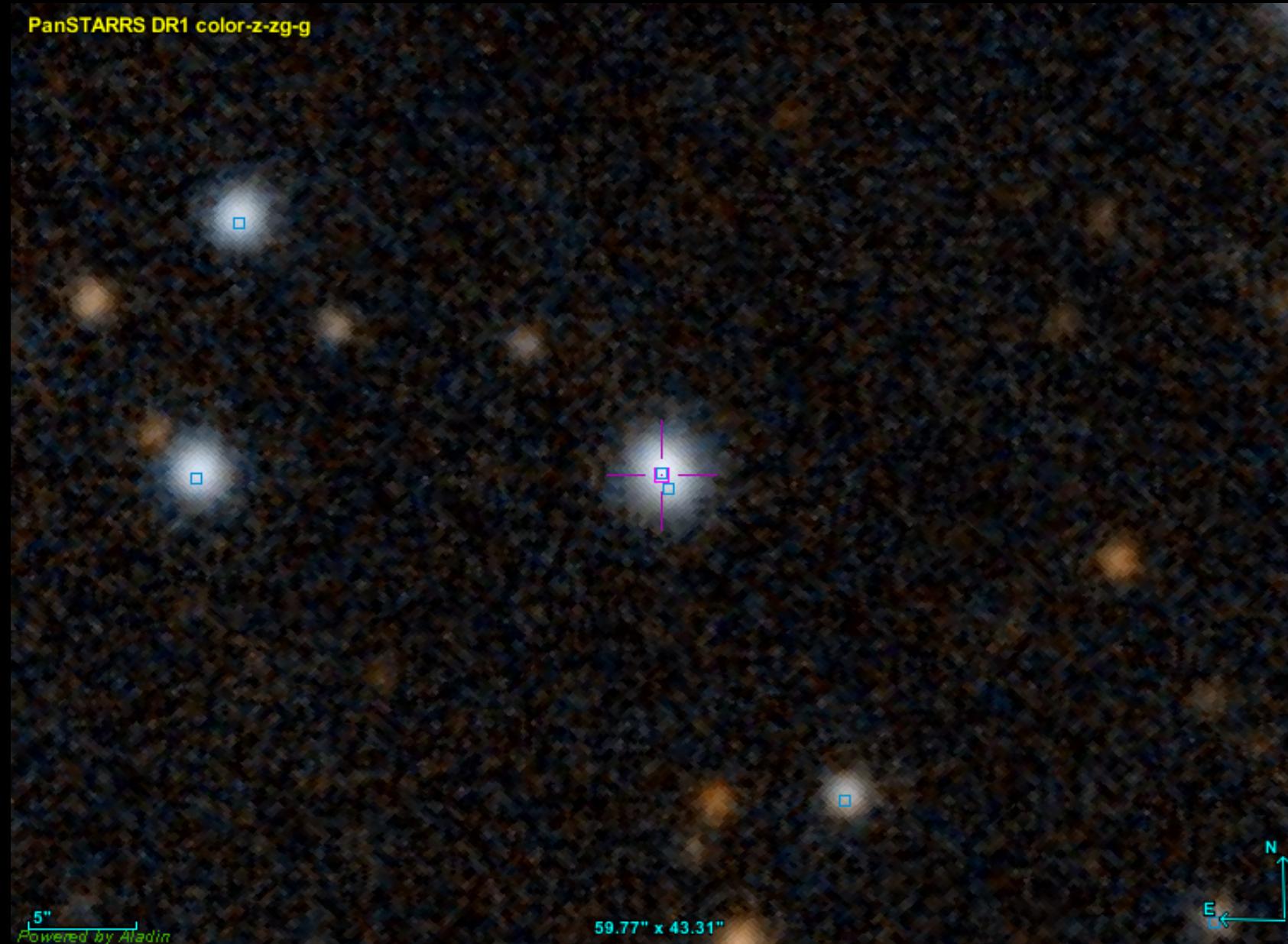
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: 2MASS J20192496-0105256  
TYPE: IRS (INFRARED SOURCE)

# ISOLATION FOREST



NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J202333.98-010910.8  
TYPE: UVS (ULTRAVIOLET SOURCE)

# MINIMUM COVARIANCE DETERMINANT



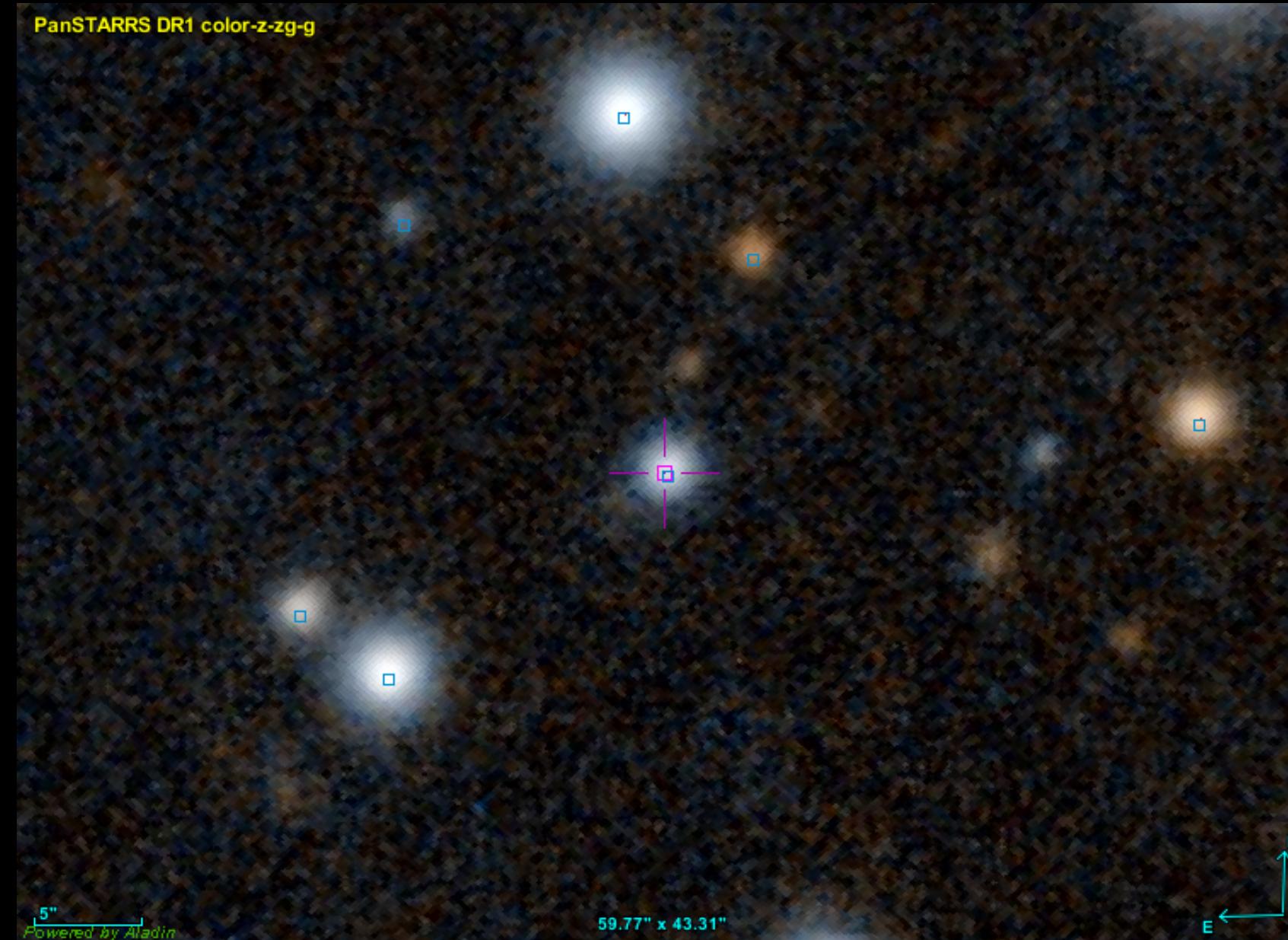
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: 2MASS J20183579+0041509  
TYPE: IRS (INFRARED SOURCE)

# MINIMUM COVARIANCE DETERMINANT



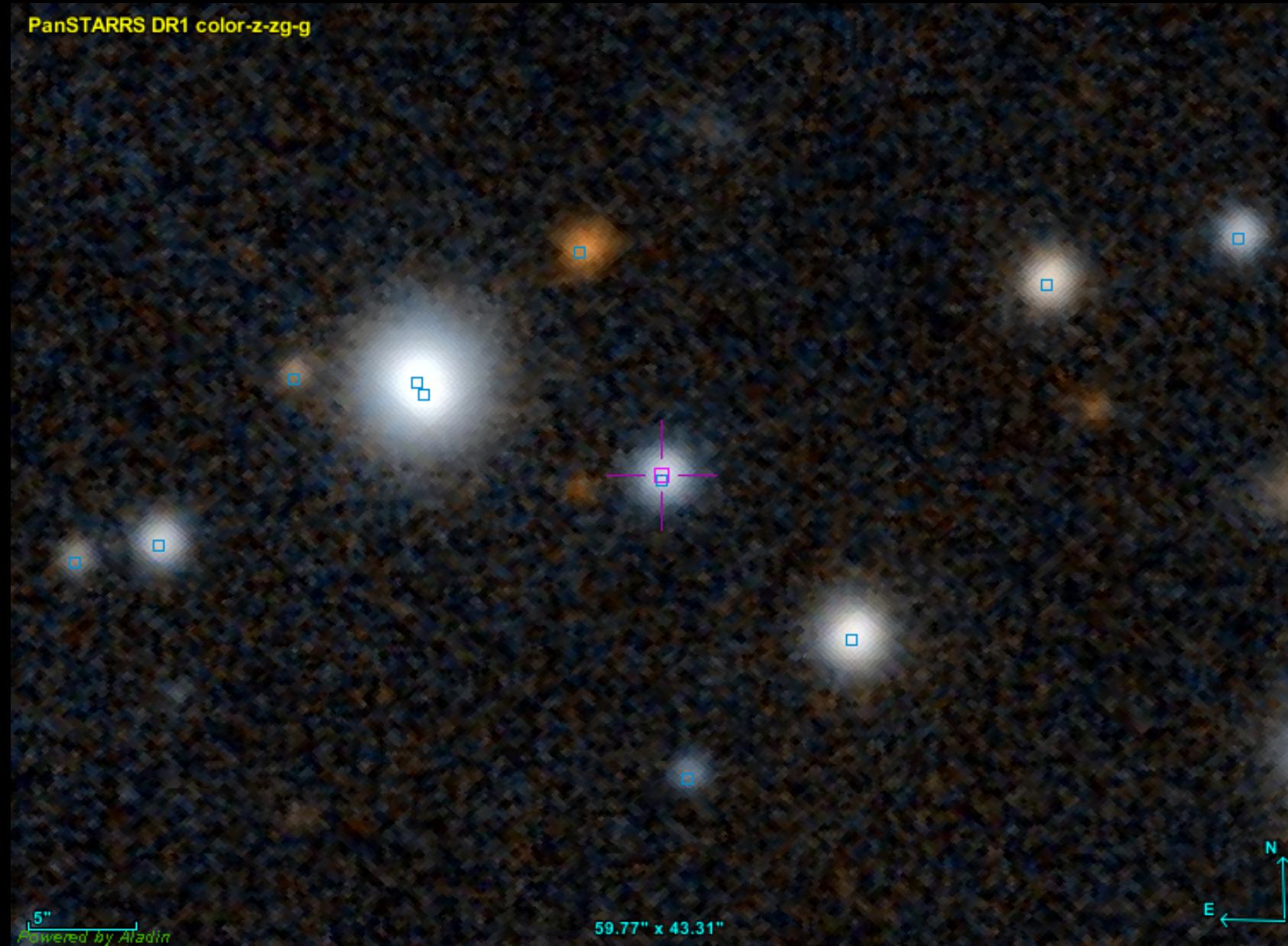
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J201558.34-000513.S  
TYPE: IRS (INFRARED SOURCE)

# MINIMUM COVARIANCE DETERMINANT



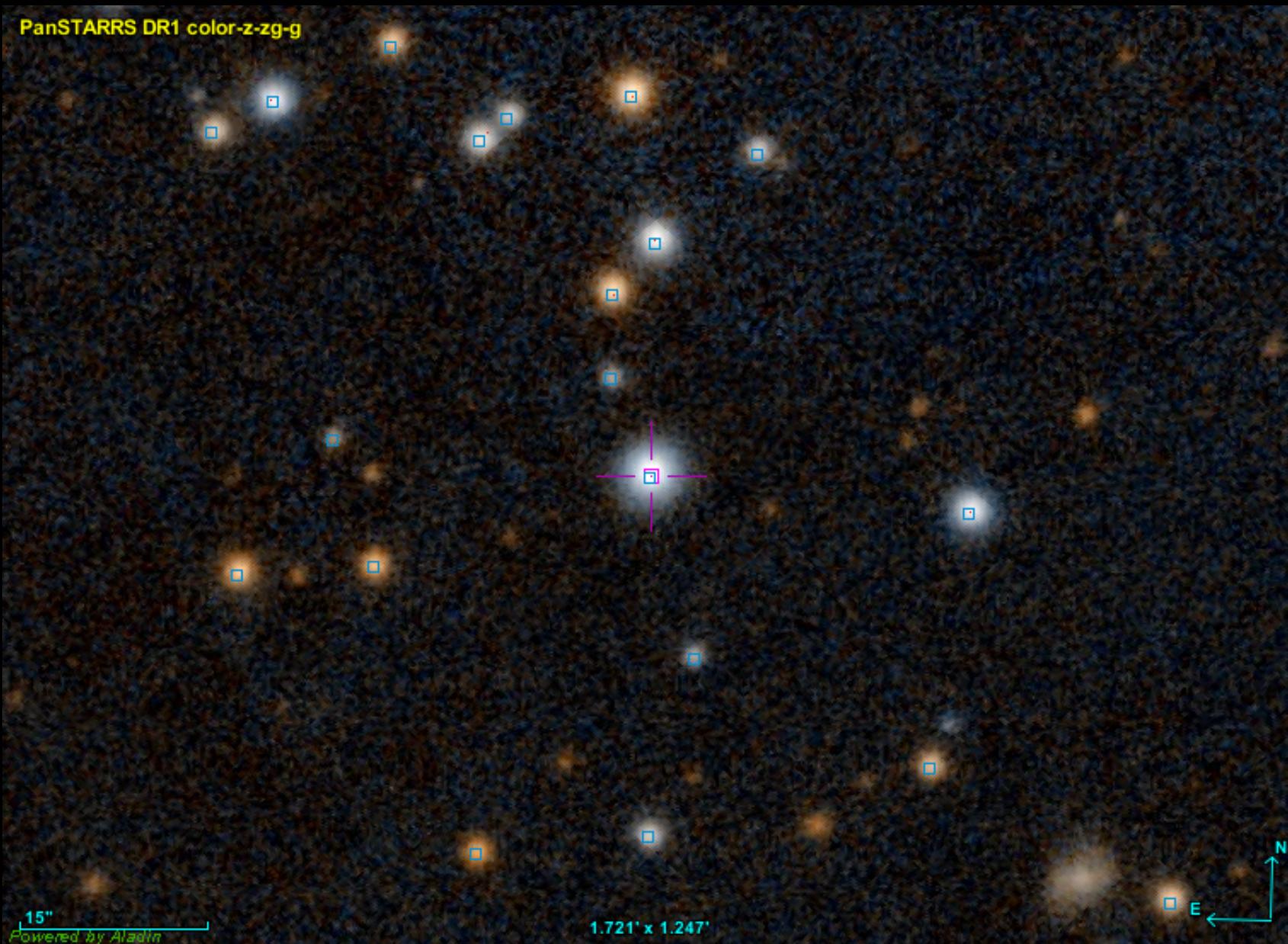
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J202348.81-010118.7  
TYPE: IRS (INFRARED SOURCE)

# MINIMUM COVARIANCE DETERMINANT



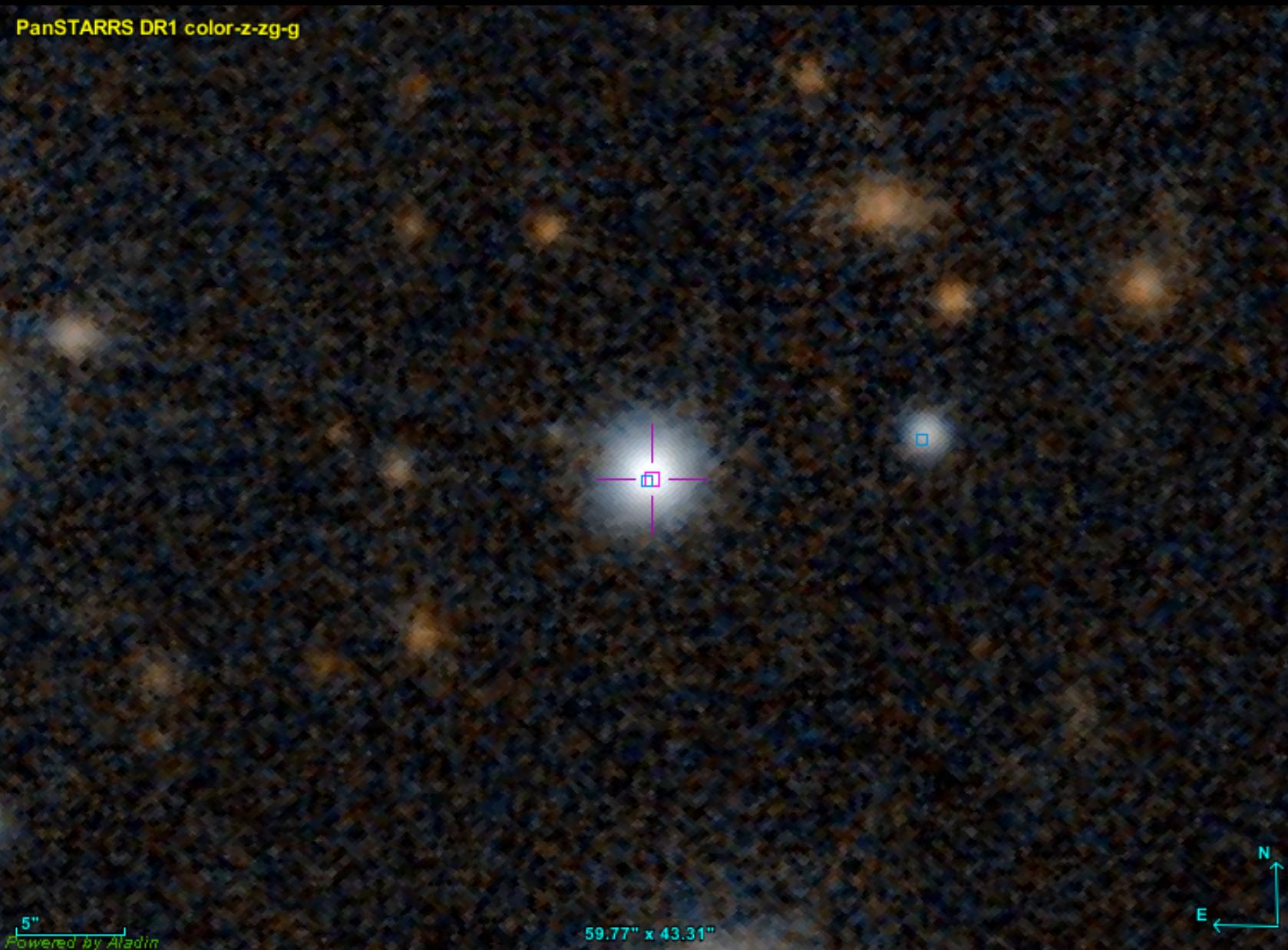
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: NÃO CATALOGADA  
TYPE: ---

# LOCAL OUTLIER FACTOR



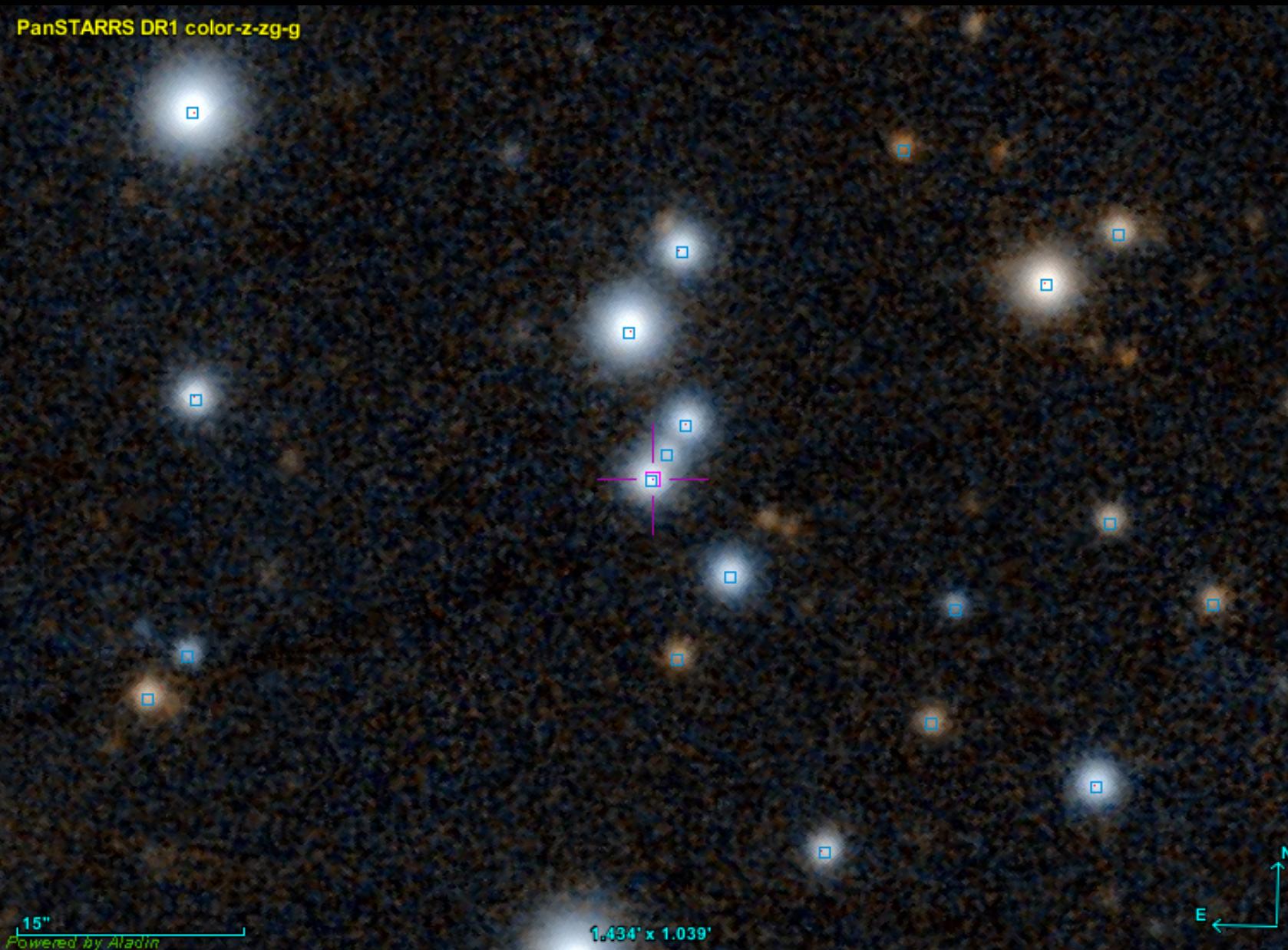
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J202512.41-000801.9  
TYPE: IRS (INFRARED SOURCE)

# LOCAL OUTLIER FACTOR



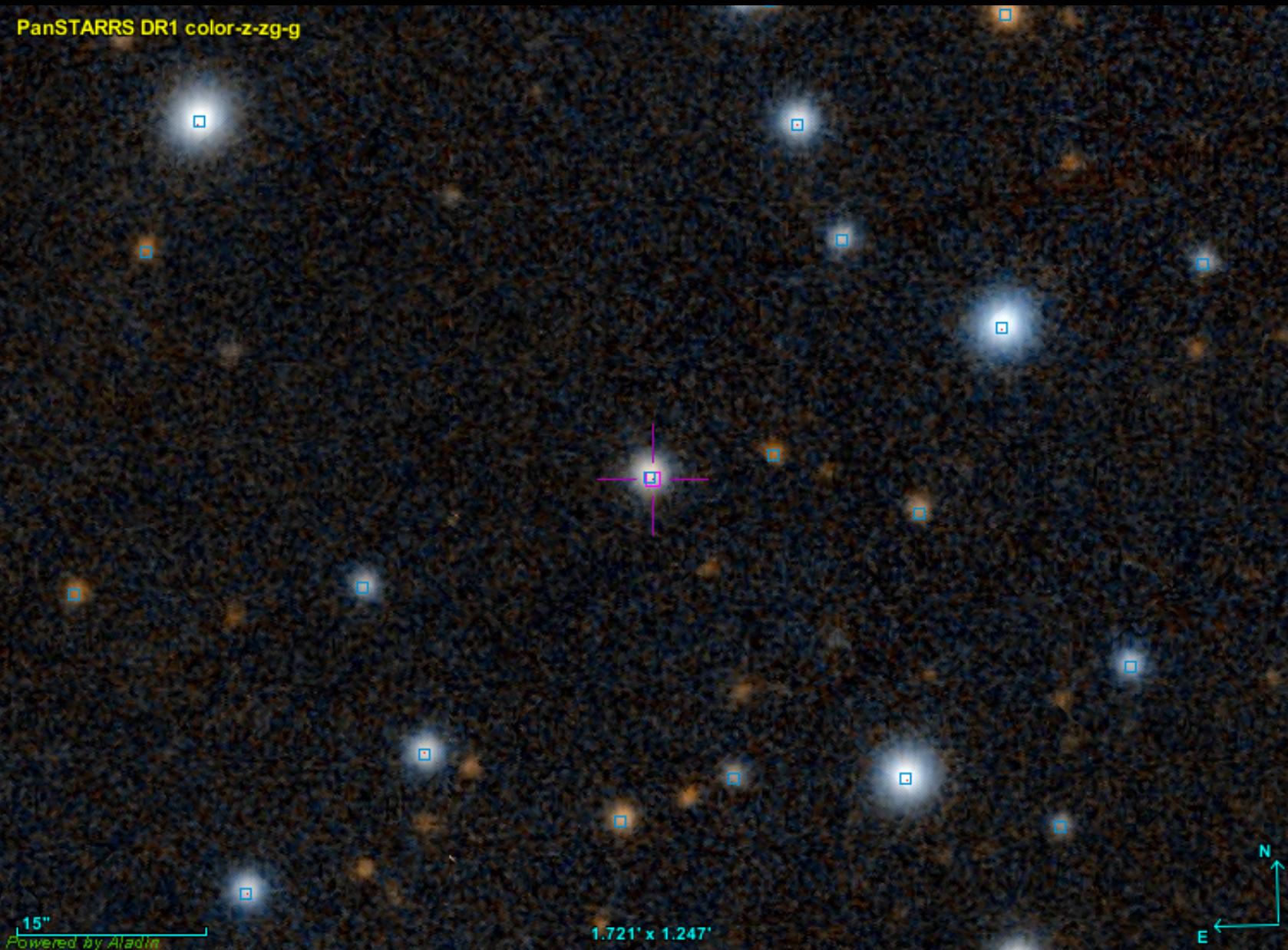
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J202018.13+001118.5  
TYPE: IRS (INFRARED SOURCE)

# ONE-CLASS SVM



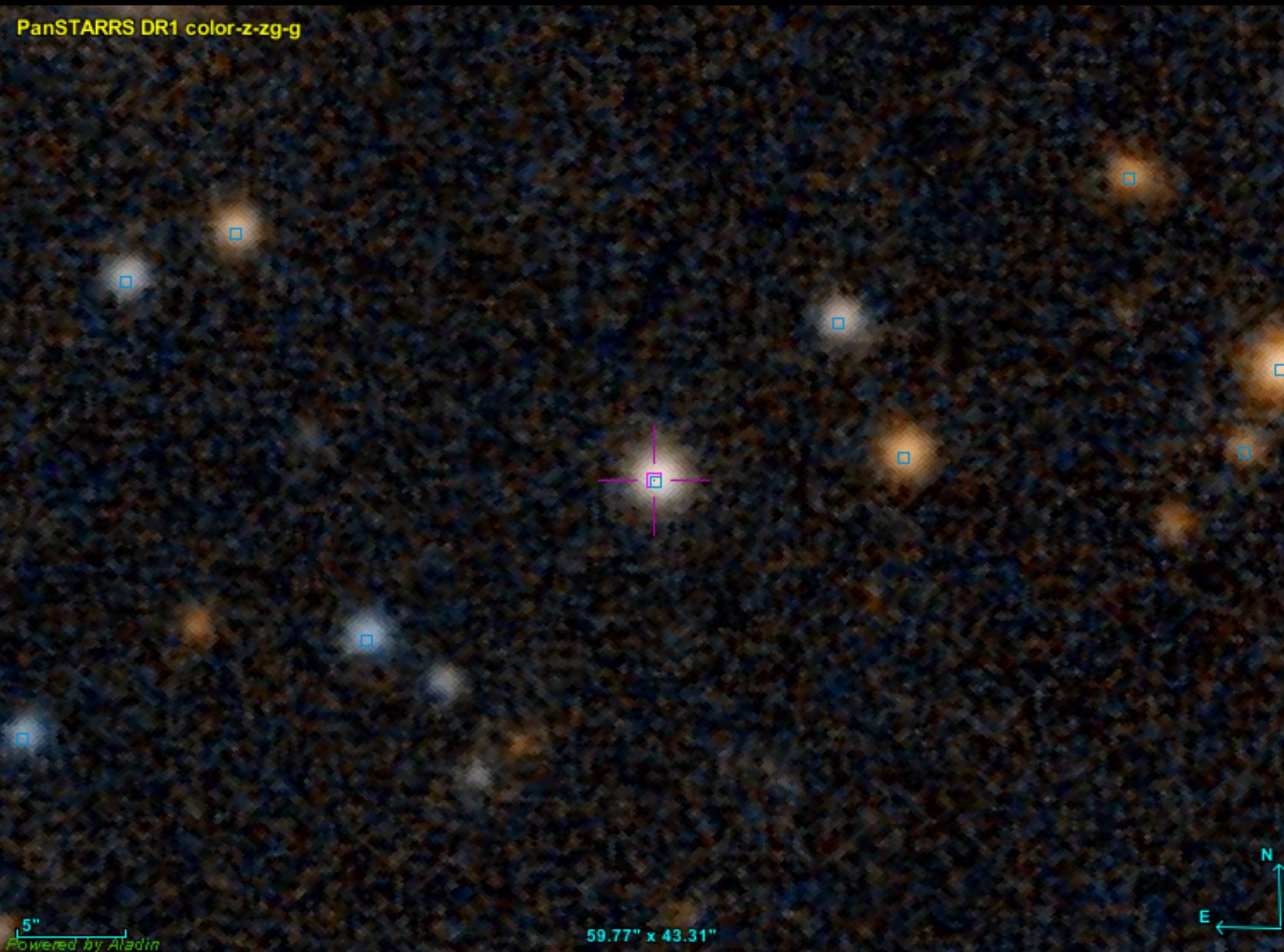
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: 2MASS J20162230-0026066  
TYPE: IRS (INFRARED SOURCE)

# ONE-CLASS SVM



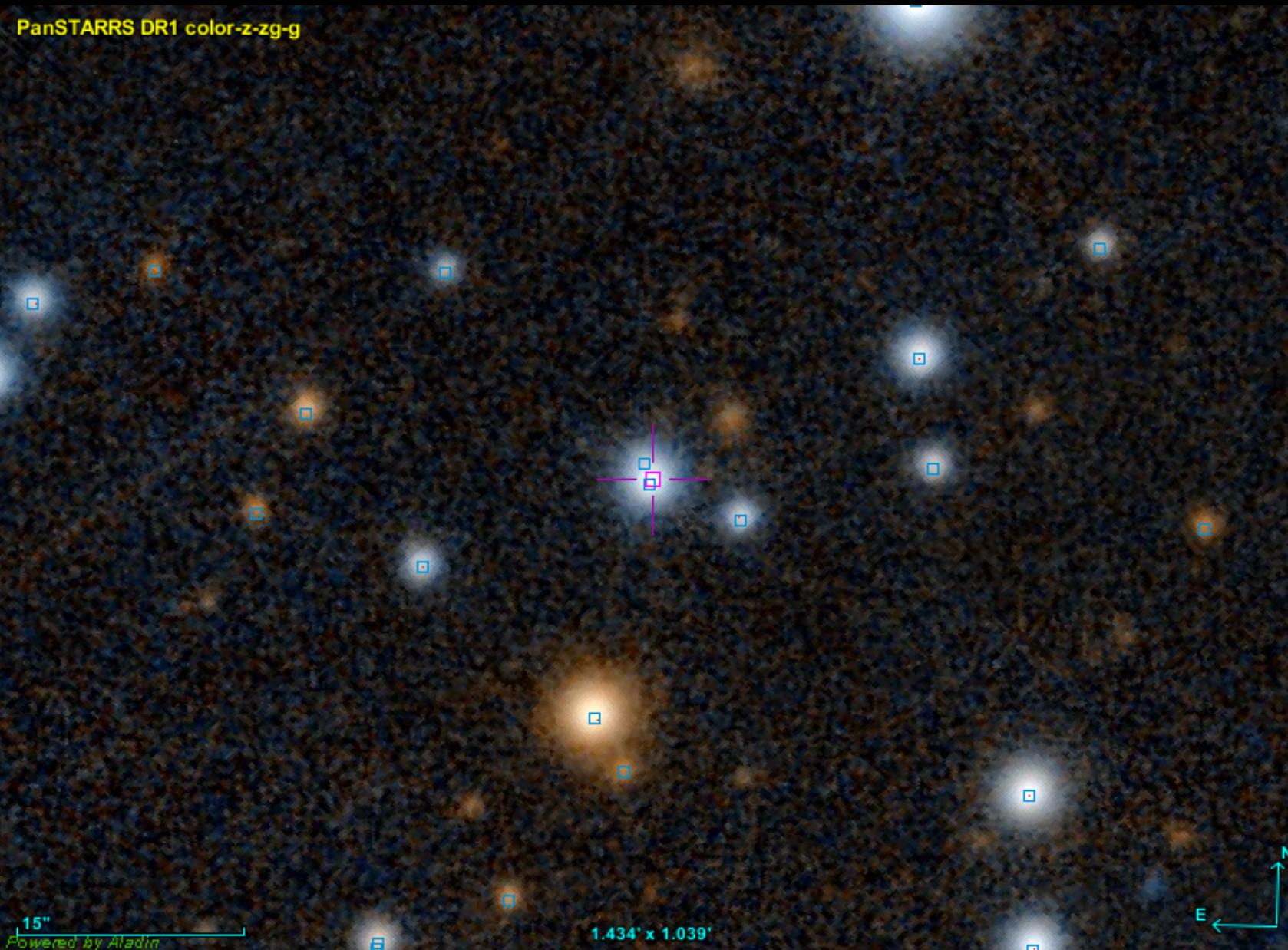
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J202715.11+001235.5  
TYPE: IRS (INFRARED SOURCE)

# ONE-CLASS SVM

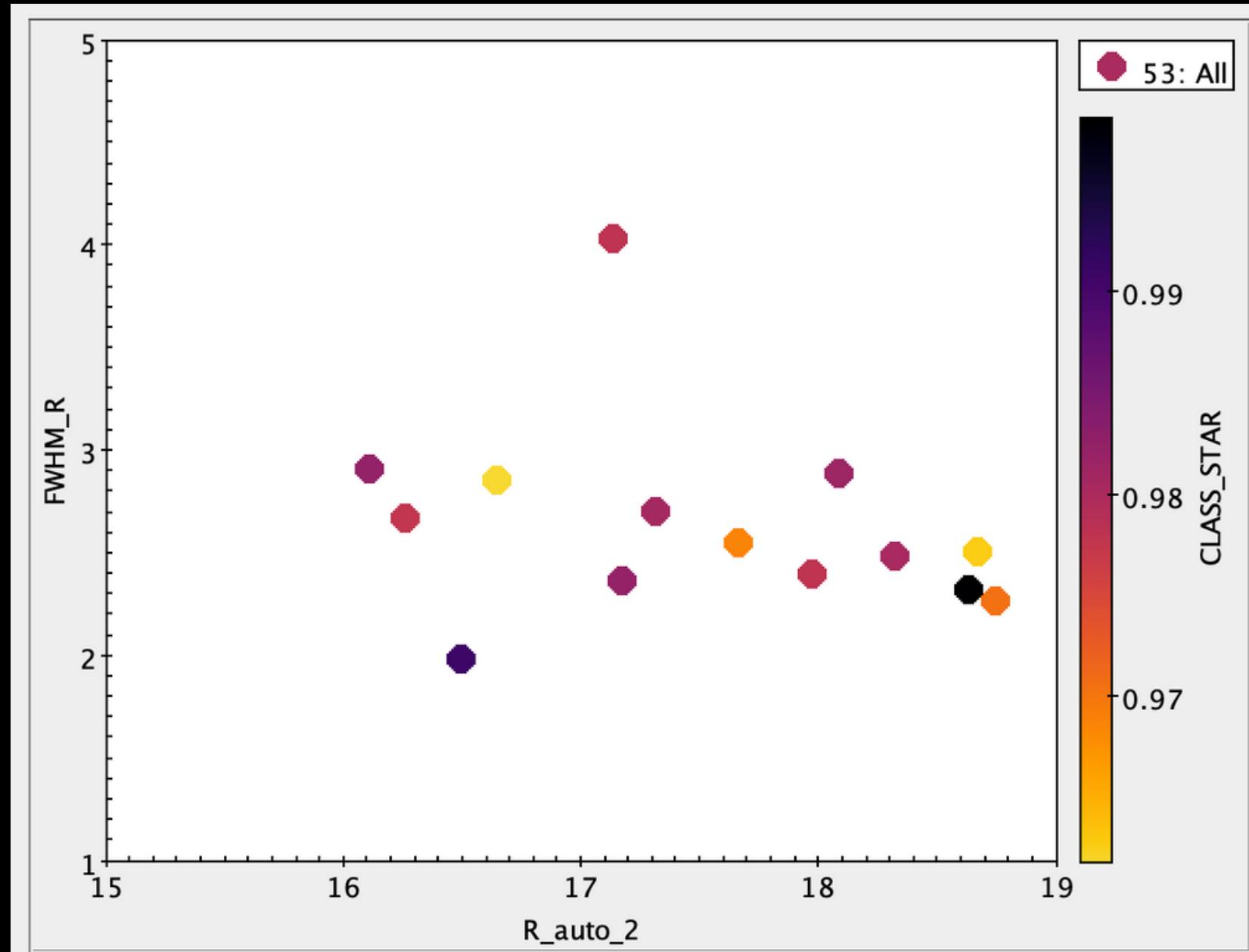


NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: WISEA J201608.63+003936.5  
TYPE: IRS (INFRARED SOURCE)

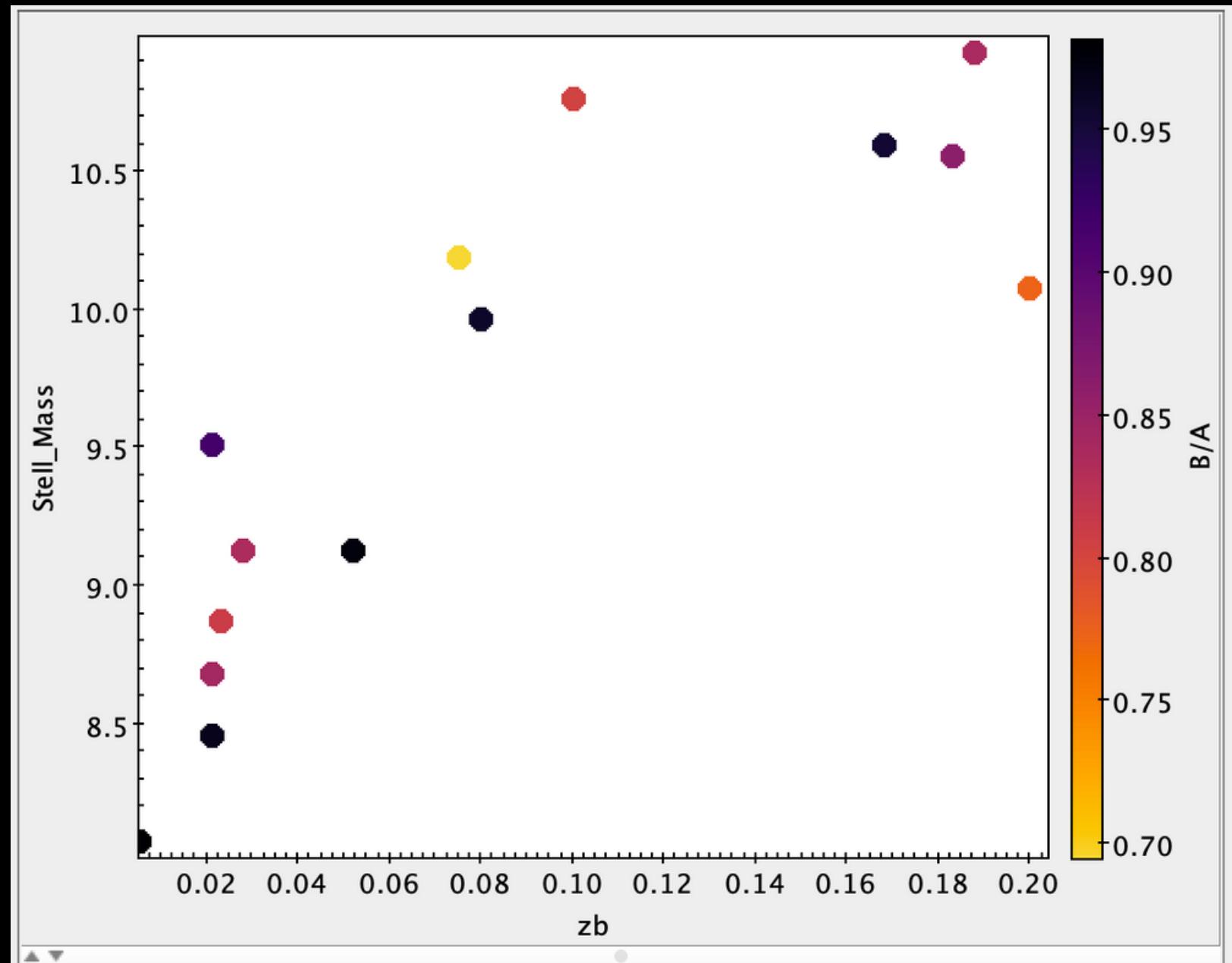
# ONE-CLASS SVM



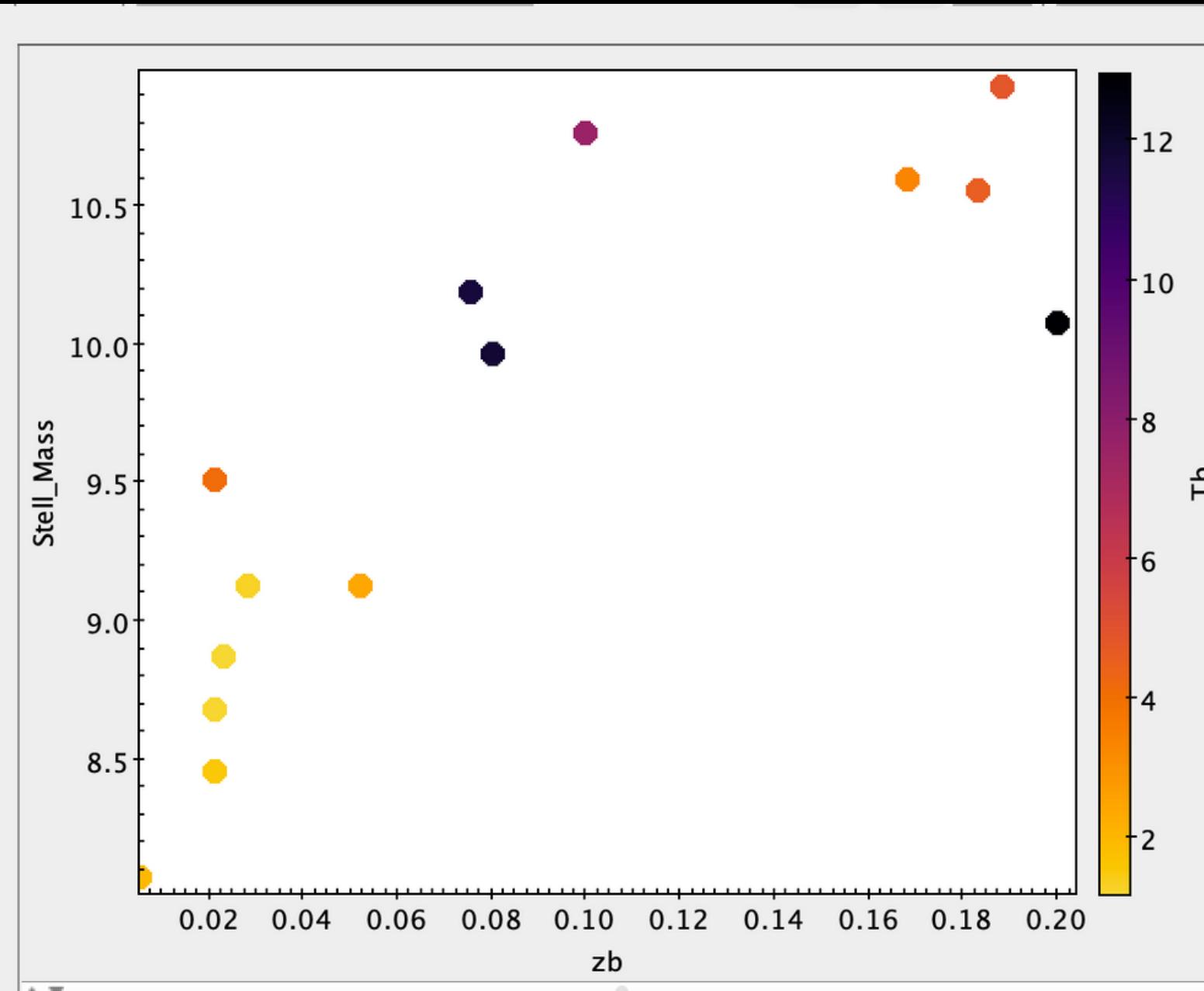
NED - NASA/IPAC EXTRAGALACTIC DATABASE  
OBJECT NAME: 2MASS J20181520-0013351  
TYPE: IRS (INFRARED SOURCE)



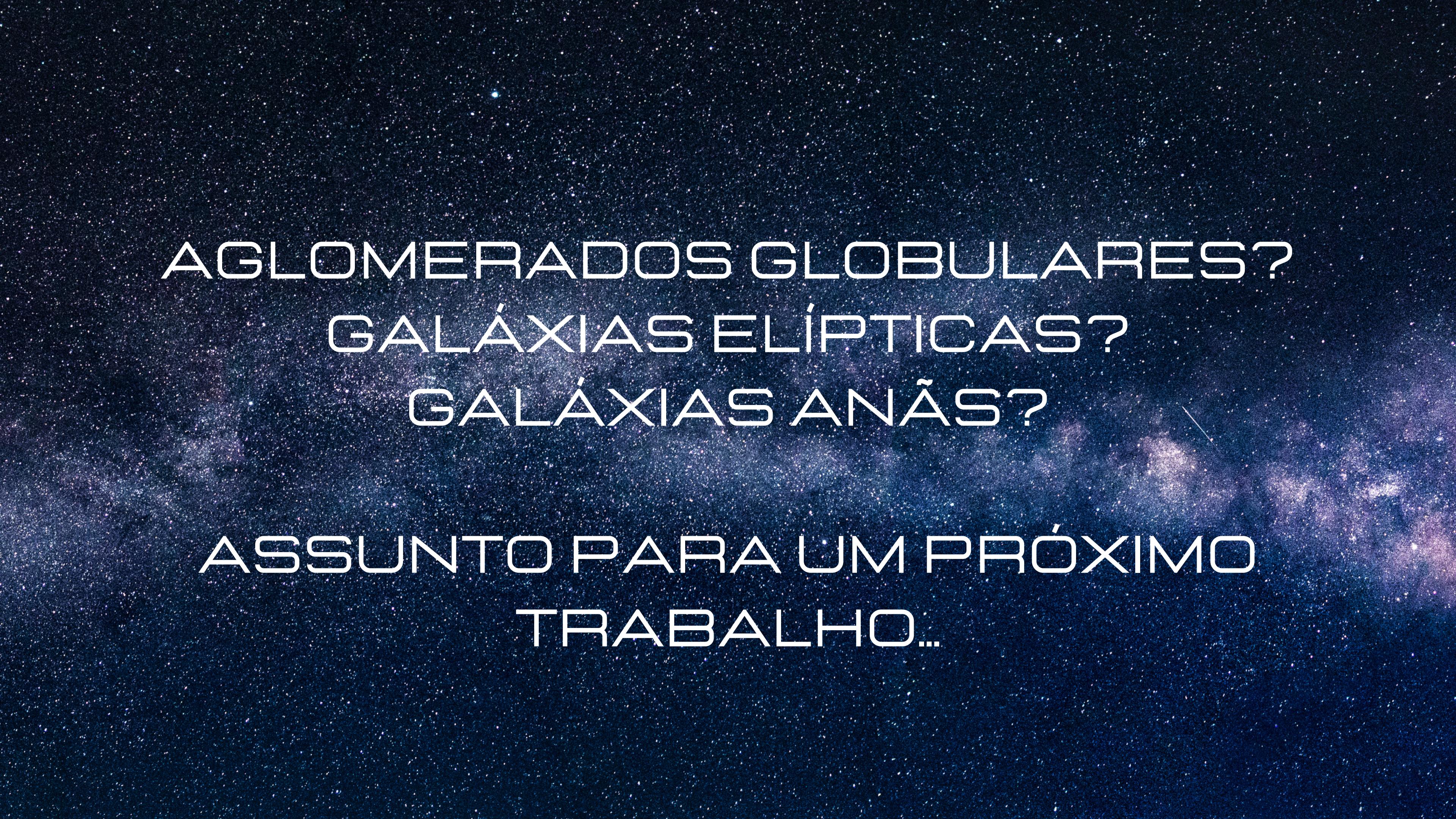
PROBABILIDADE DE SEREM ESTRELAS (CLASS\_STAR DEVERIA SER  
PRÓXIMO A 1.0)?  
O FATO QUE SIMBAD NÃO AS RECONHECE NÃO CONFIRMA ISSO...  
SÃO OBJETOS "COMPACTOS" NO CÉU.



A MAIORIA SÃO FACE-ON, I.E.  $B/A \sim 1$ , SÃO "REDONDOS".

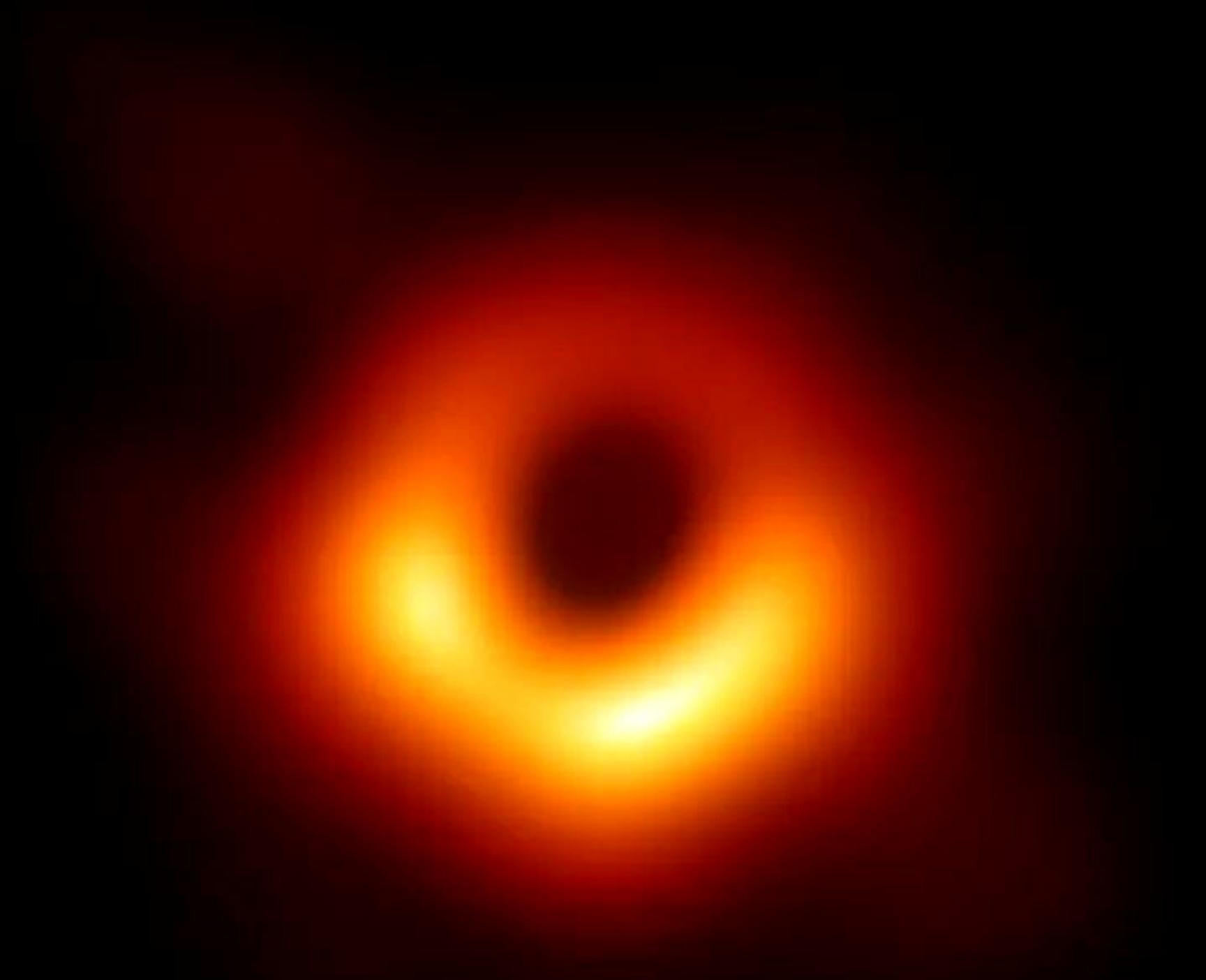


SÃO MASSIVOS (STELL\_MASS > 8)! A MAIORIA DELES NÃO FORMAM ESTRELAS: TB (BAYESIAN SPECTRAL TYPE) < 6. SÃO EARLY TYPES.



AGLOMERADOS GLOBULARES?  
GALÁXIAS ELÍPTICAS?  
GALÁXIAS ANÃS?

ASSUNTO PARA UM PRÓXIMO  
TRABALHO...



OBRIGADO A TODOS!  
UM AGRADECIMENTO ESPECIAL A  
CLÉCIO DE BOM E ARIANNA CORTESI.