An Information Theory Approach on Deciding Spectroscopic Follow Ups

JAVIERA ASTUDILLO
PAVLOS PROTOPAPAS
KARIM PICHARA
PABLO HUIJSE

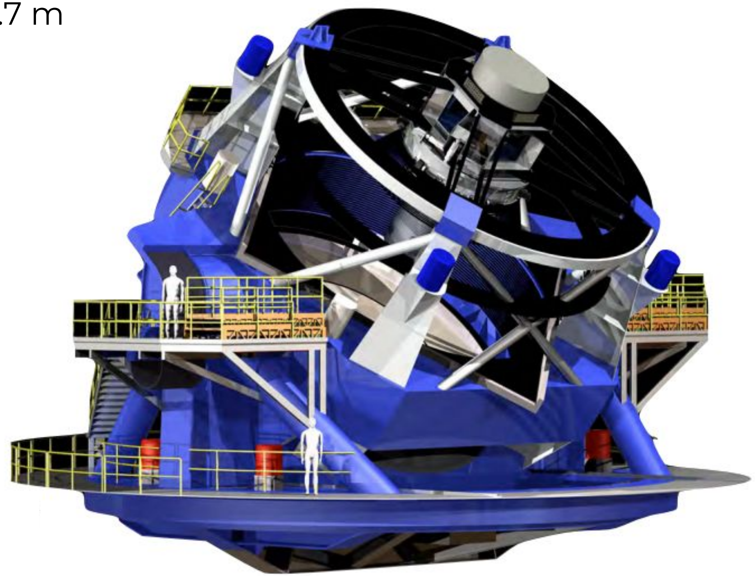
OUTLINE

1

LSST

- Effective aperture of 6.7 m
- FoV 9.6 deg^2
- Large *etendue* (Collecting area x FoV)

1



Wide-Deep-Fast

Cover large swaths of sky to faint magnitudes in a short amount of time.

2022-2032

1

BACK OF THE ENVELOPE LSST

Sky area: 41252.96 deg^2, LSST field of view (FoV): 9.6 deg^2

→ southern sky area ~ 20000 deg^2 ~ 2000 LSST FoV

Night length ~ 8 hr, LSST exposure: 30 + 7 sec

→ ~800 LSST exposures / night

Survey length ~ 10 yr (-10%) in 6 bands (ugrizy)

→ 365 x 9 x 800 ~ 2.5M LSST exposures
→ ~2.5M visits / 2000 FoV = 1250 visits per FoV
→ average of one visit every 3 days per FoV
→ ~250 visits per region per band
→ average of one visit every 15 days per FoV per band

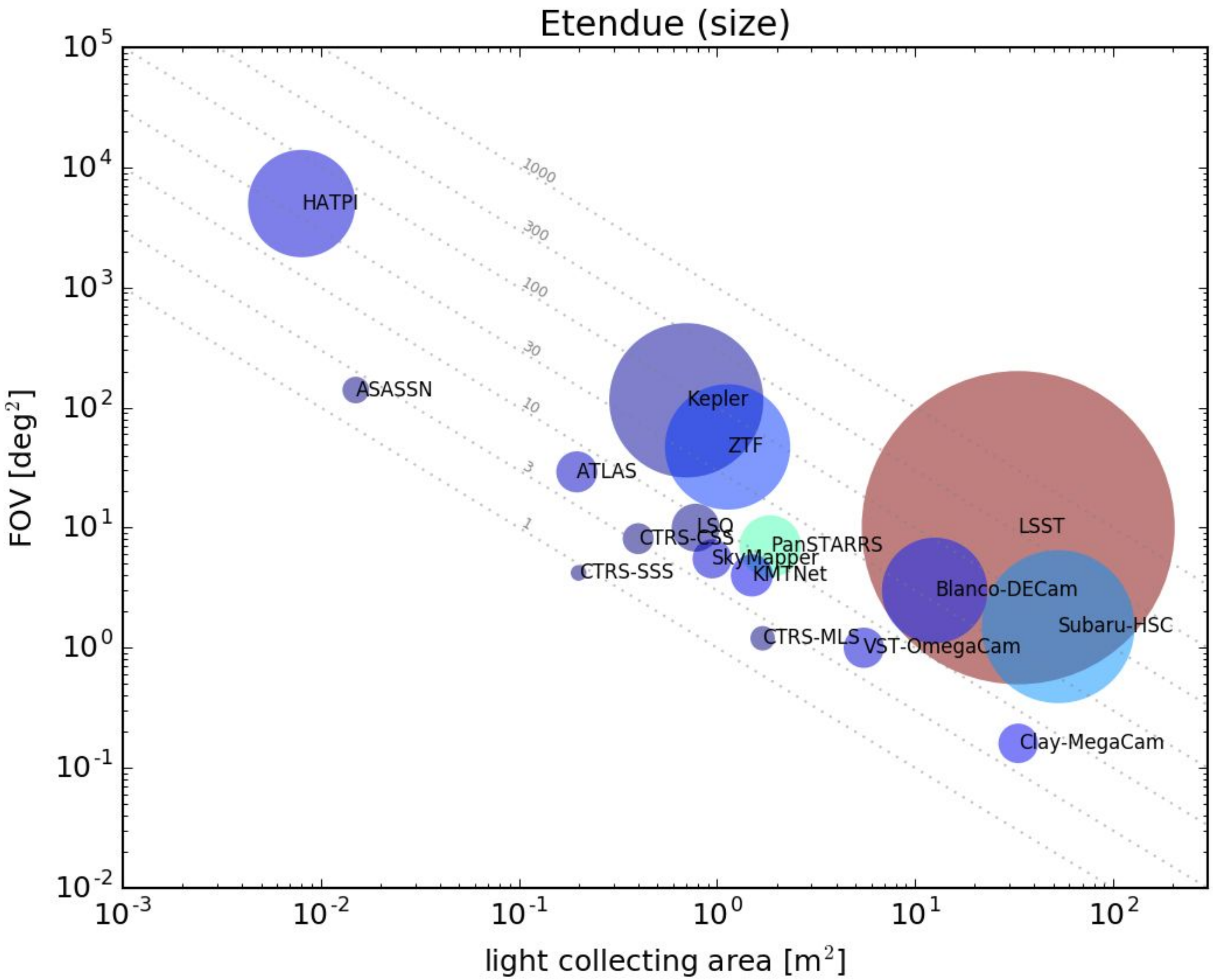
Single epoch depth in r band ~ 24.5 mag

→ Coadded depth of 24.5 + 2.5 $\log_{10}(\mathrm{sqrt}(250))$ ~ 27.5

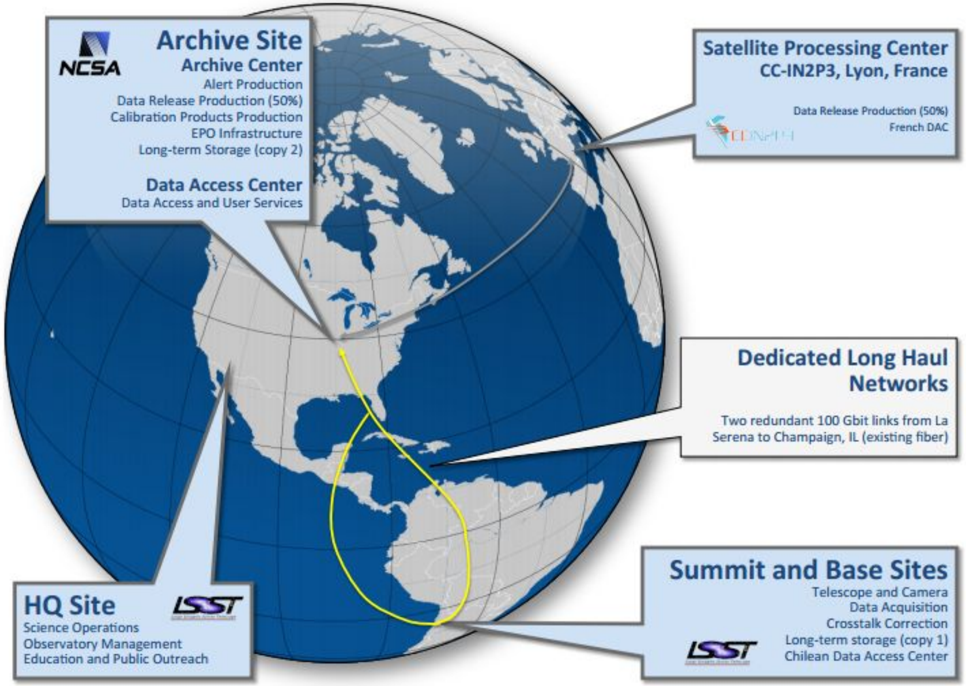
LSST's goal is to have 1000 visits to every region of the southern sky.

Exact distribution of visits TBD: different science goals can have complementary/conflicting optimal strategies (e.g. proper motion studies vs supernova studies, a.k.a. *cadence wars*).

1


Etendue (size)
FOV [deg^2]
light collecting area [m^2]
1000
300
100
30
10
3
1
HATPI
ASASSN
Kepler
ZTF
ATLAS
LSQ
CTRS-CSS
PanSTARRS
SkyMapper
KMTNet
CTRS-SSS
LSST
Blanco-DECam
Subaru-HSC
CTRS-MLS
VST-OmegaCam
Clay-MegaCam

1



Archive Site
NCSA
Archive Center
Alert Production
Data Release Production (50%)
Calibration Products Production
EPO Infrastructure
Long-term Storage (copy 2)
Data Access Center
Data Access and User Services
Satellite Processing Center
CC-IN2P3, Lyon, France
Data Release Production (50%)
French DAC
Dedicated Long Haul Networks
Two redundant 100 Gbit links from La Serena to Champaign, IL (existing fiber)
Summit and Base Sites
Telescope and Camera
Data Acquisition
Crosstalk Correction
Long-term storage (copy 1)
Chilean Data Access Center
HQ Site
Science Operations
Observatory Management
Education and Public Outreach

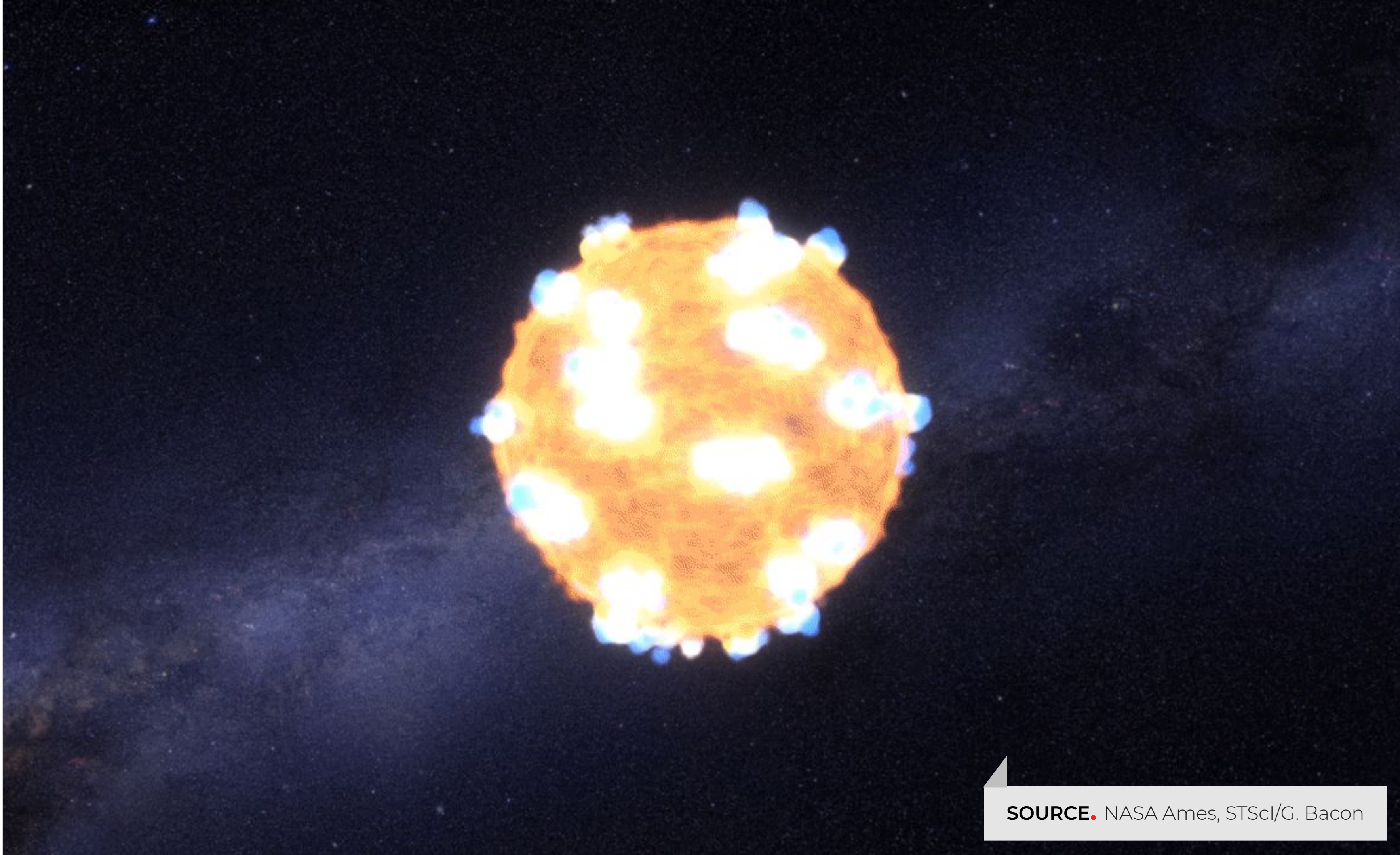
1


Veracity
Capture
Curation
Storage
Search
Velocity
BIG DATA
Sharing
Variety
Analysis
Volume
Visualization
Transfer

2

VARIABLE OBJECTS

2



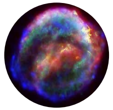
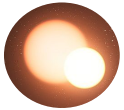
SOURCE. NASA Ames, STScI/G. Bacon

2

SOURCE. NASA Ames, STScI/G. Bacon

VARIABLE STARS

2

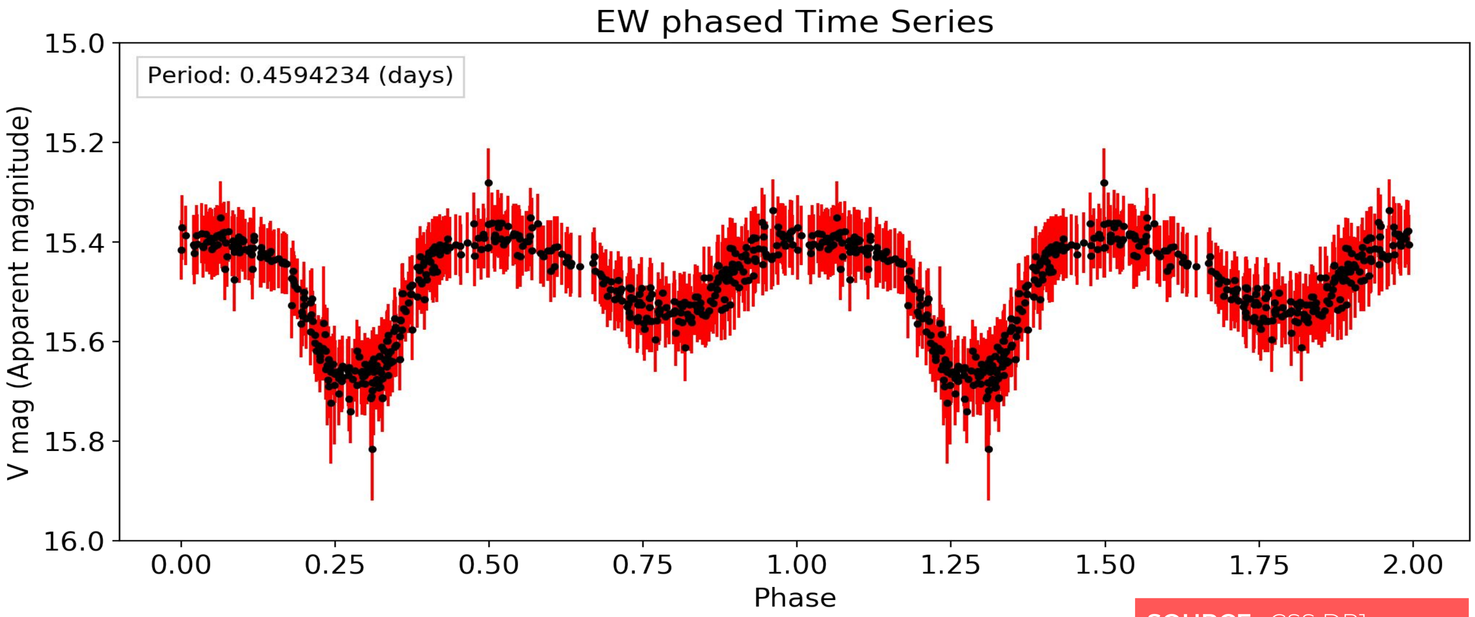
- VARIABLE STARS
  - INTRINSIC
    - PULSATING
      - Cepheids
      - RR Lyrae
      - RV Tauri
      - Long-period variables 
    - ERUPTIVE
      - Supernovae 
      - Novae
      - Recurrent novae
  - EXTRINSIC
    - ECLIPSING BINARIES 
    - ROTATING VARIABLES

3

TYPES OF DATA

TIME SERIES

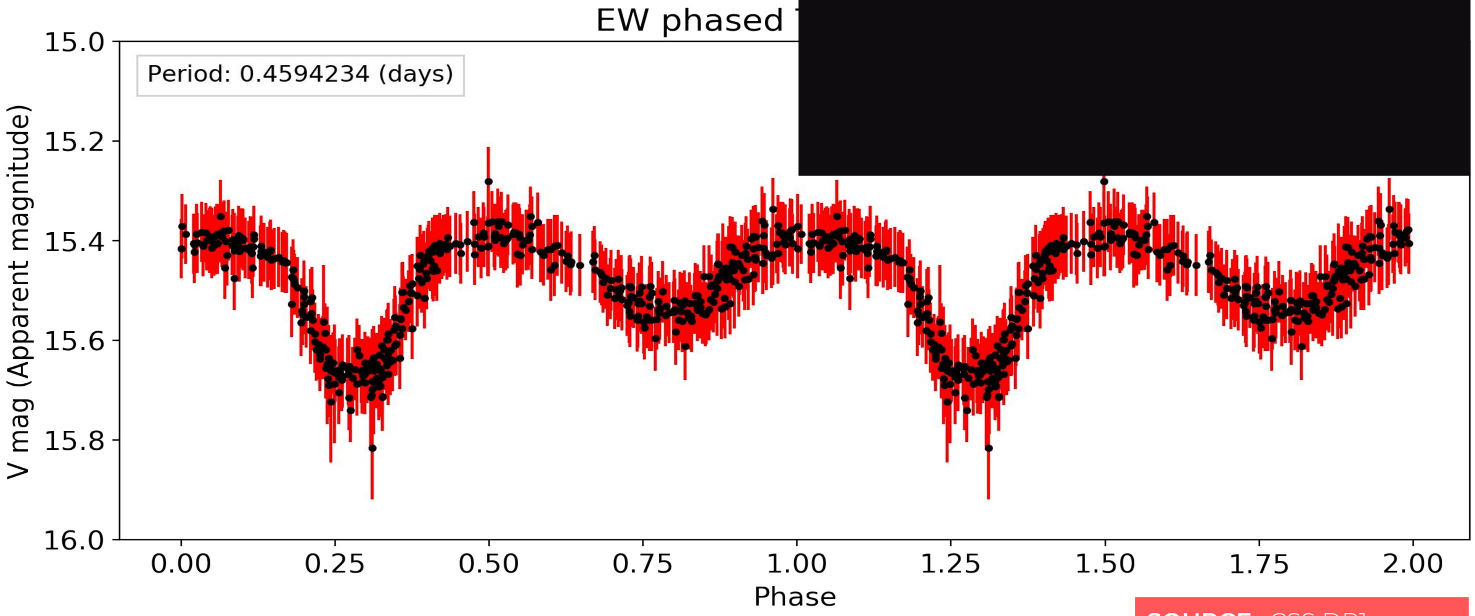
3



SOURCE. CSS DR1

3

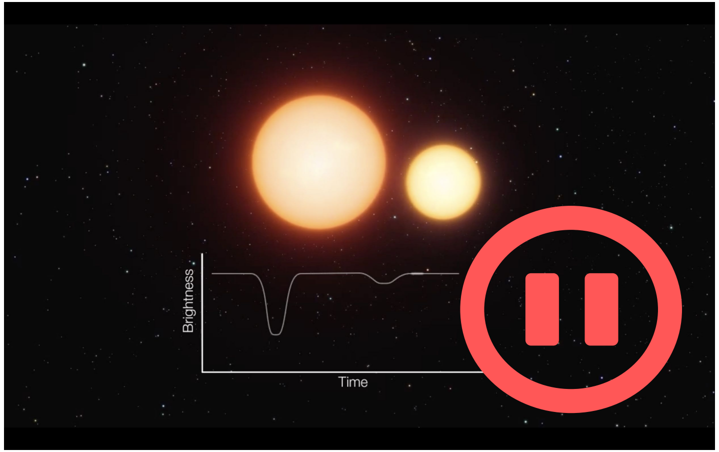
TIME SERIES

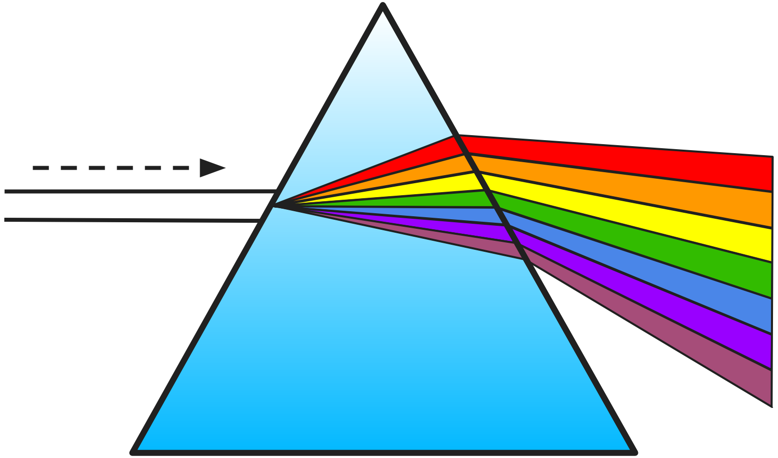


SOURCE. CSS DR1

3

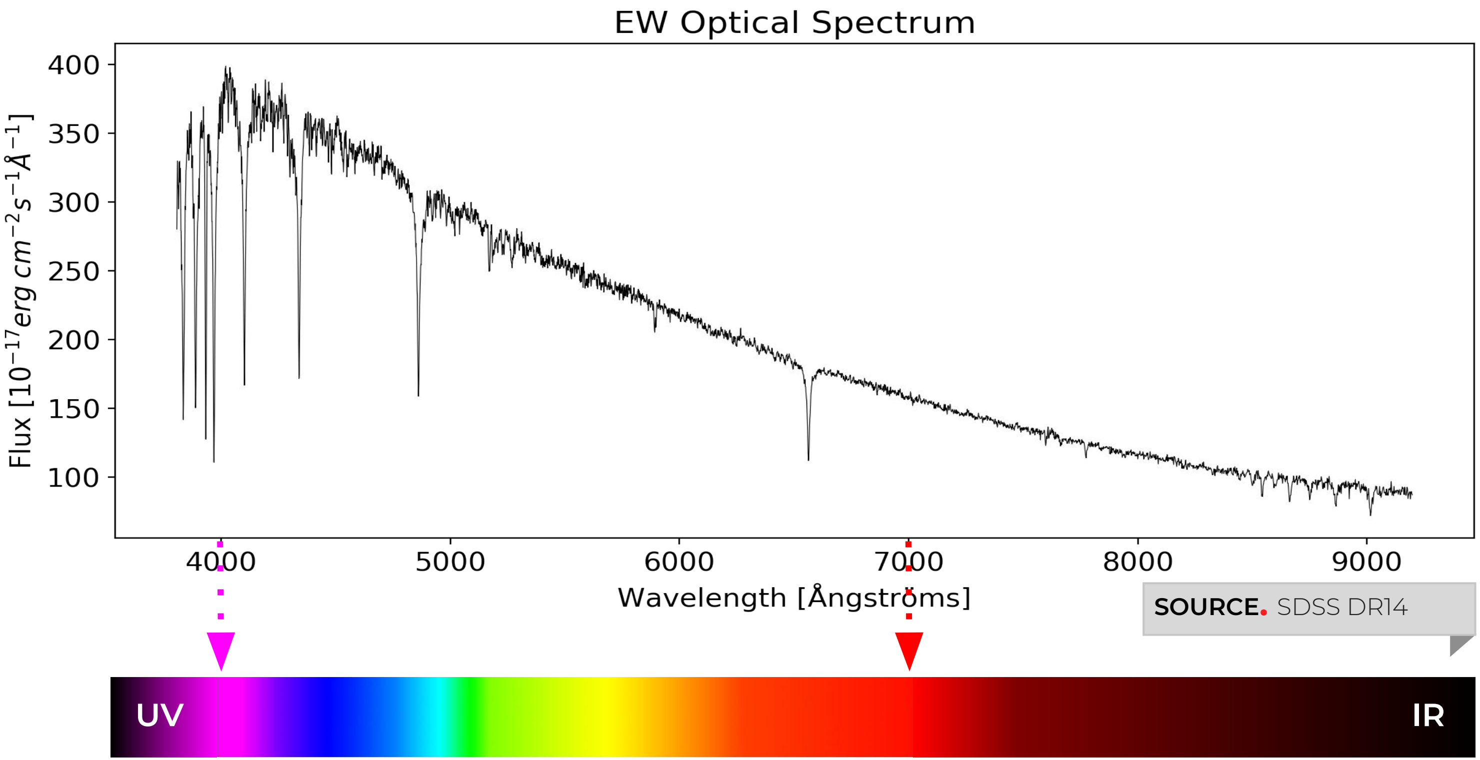
SPECTRUM





3

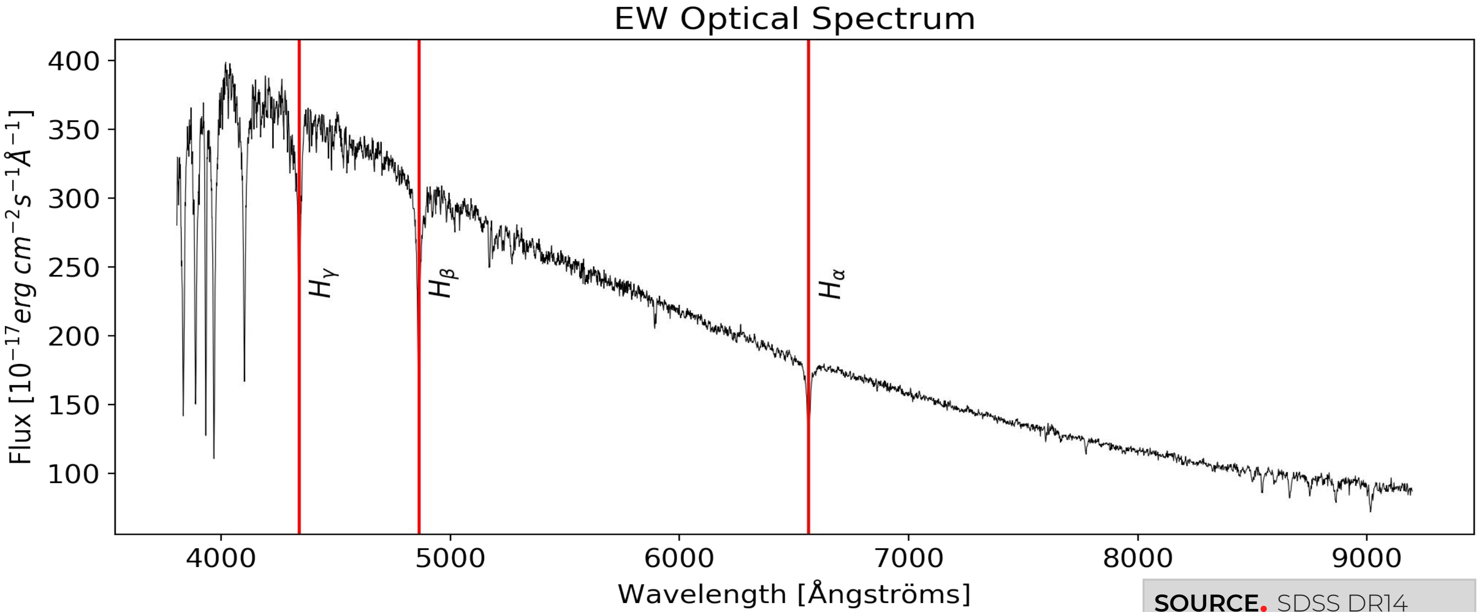
SPECTRUM



SOURCE. SDSS DR14

3

SPECTRUM



SOURCE. SDSS DR14

3

DATA COLLECTION

TIME SERIES

30 s./Exposure
Many

4 EXPOSURES
200 OBSERVATIONS
400 MIN

CSS DR1

>>

SPECTRUM

15 m./Exposure
Few

3 EXPOSURES
1 OBSERVATION
45 MINUTES

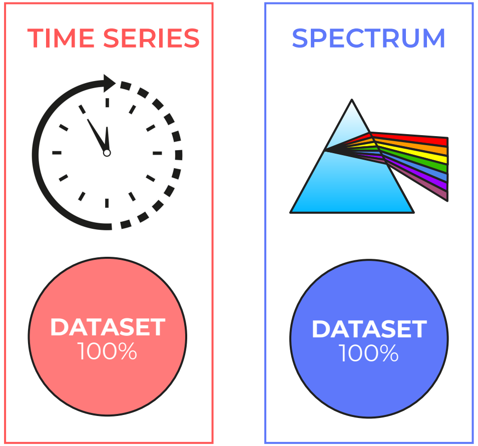
SDSS DR14

SDSS

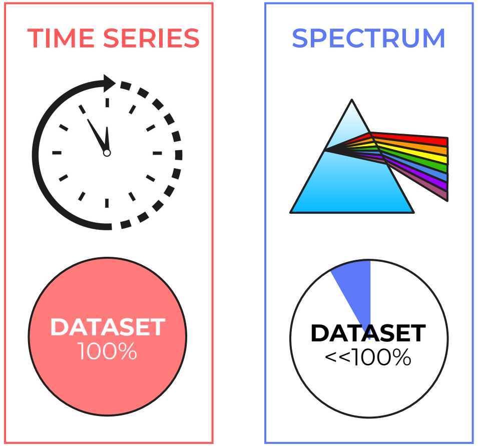
19

3

IDEAL SCENARIO



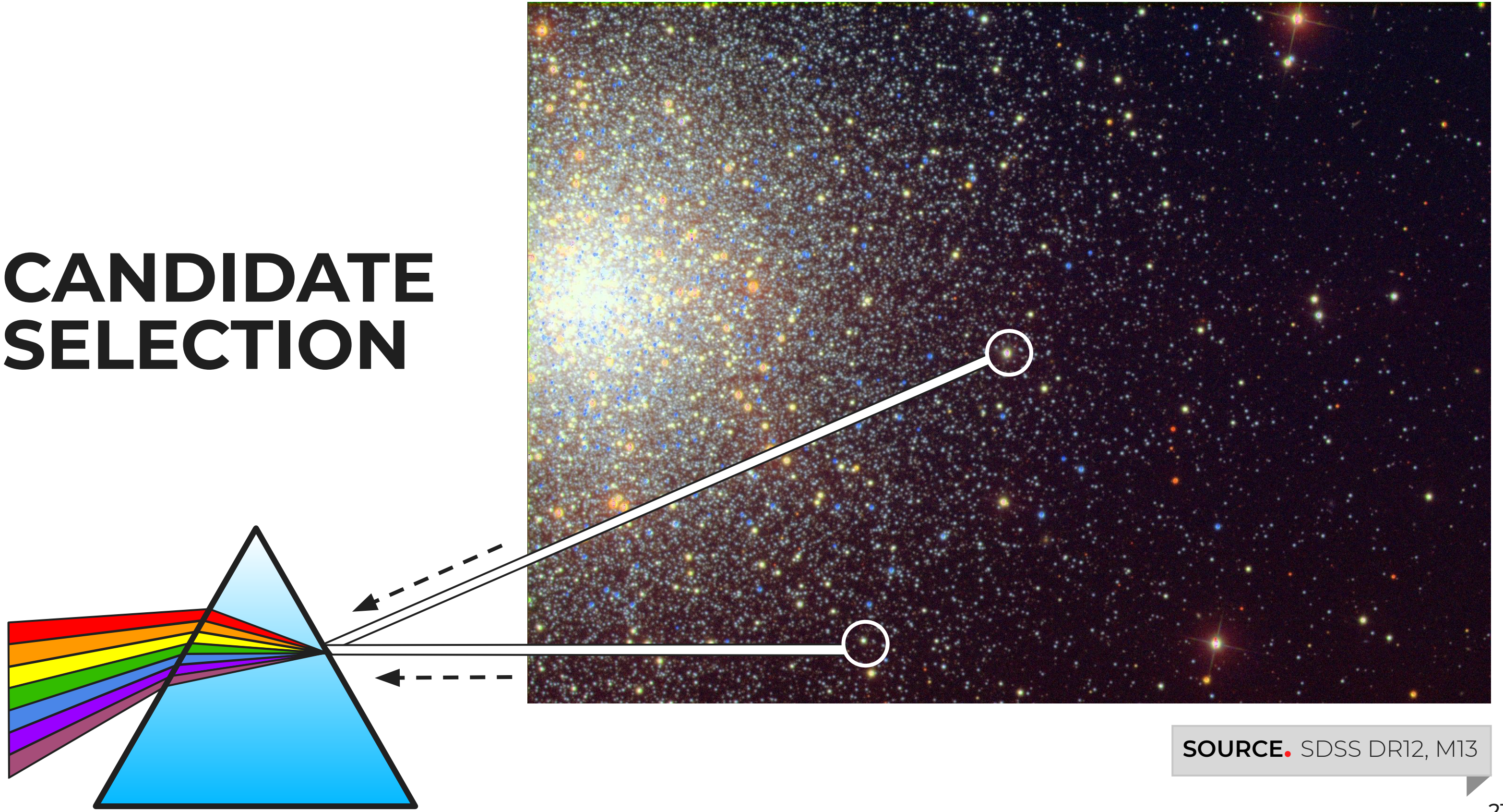
REALITY



20

3

CANDIDATE SELECTION



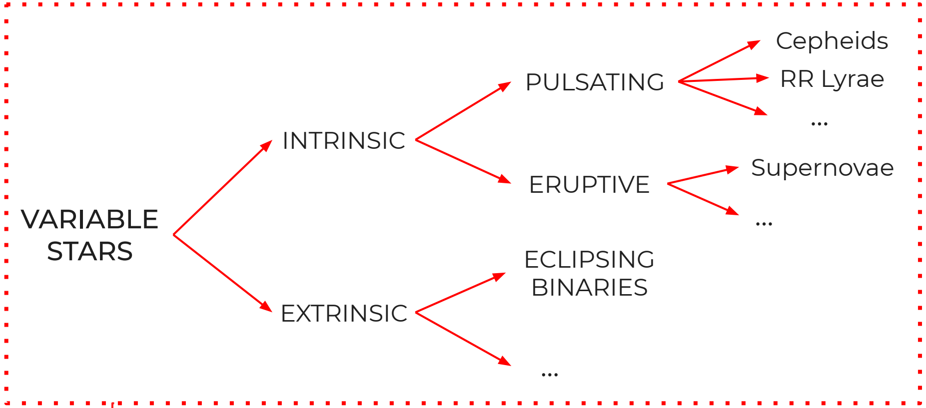
SOURCE. SDSS DR12, M13

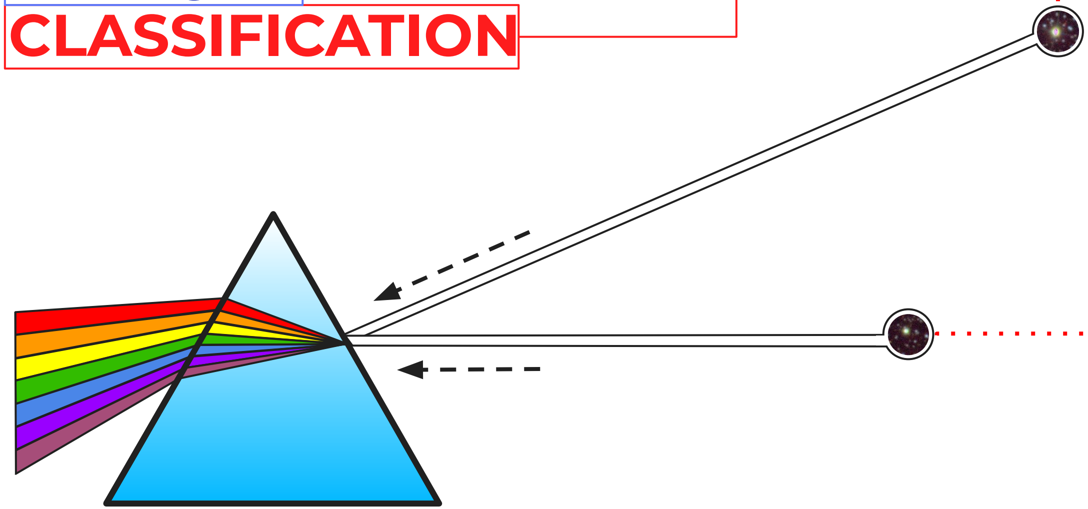
3

+ HIGHER ACCURACY

+ HIGHER CONFIDENCE IN THE CLASSIFICATION

IMPROVE CLASSIFICATION





3

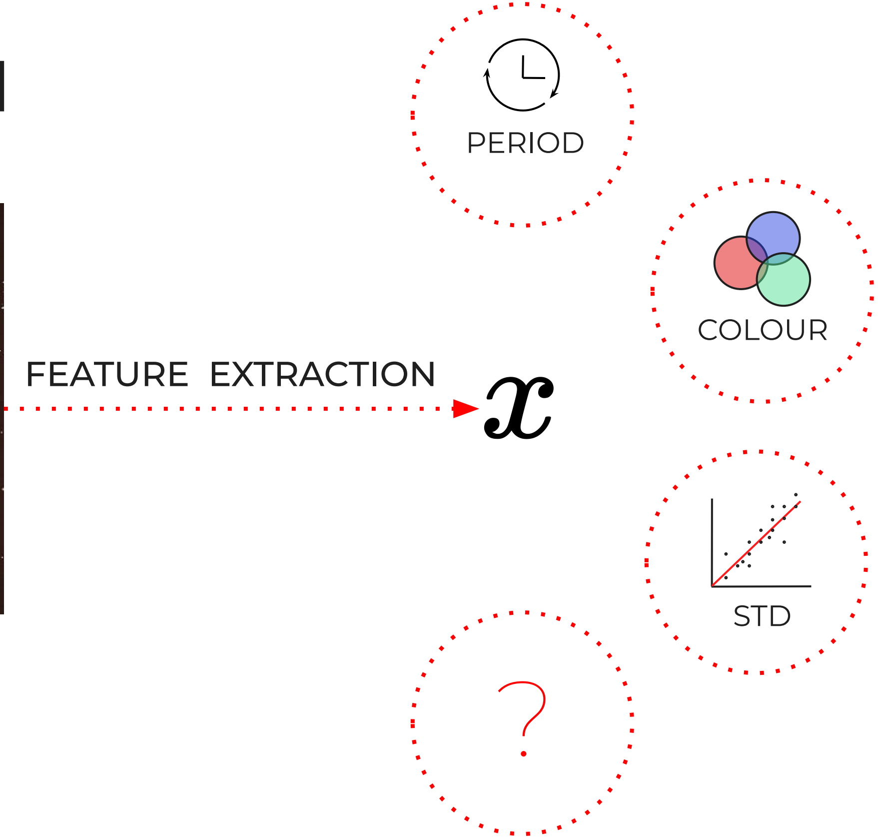
IMPROVE CLASSIFICATION

4

BACKGROUND THEORY

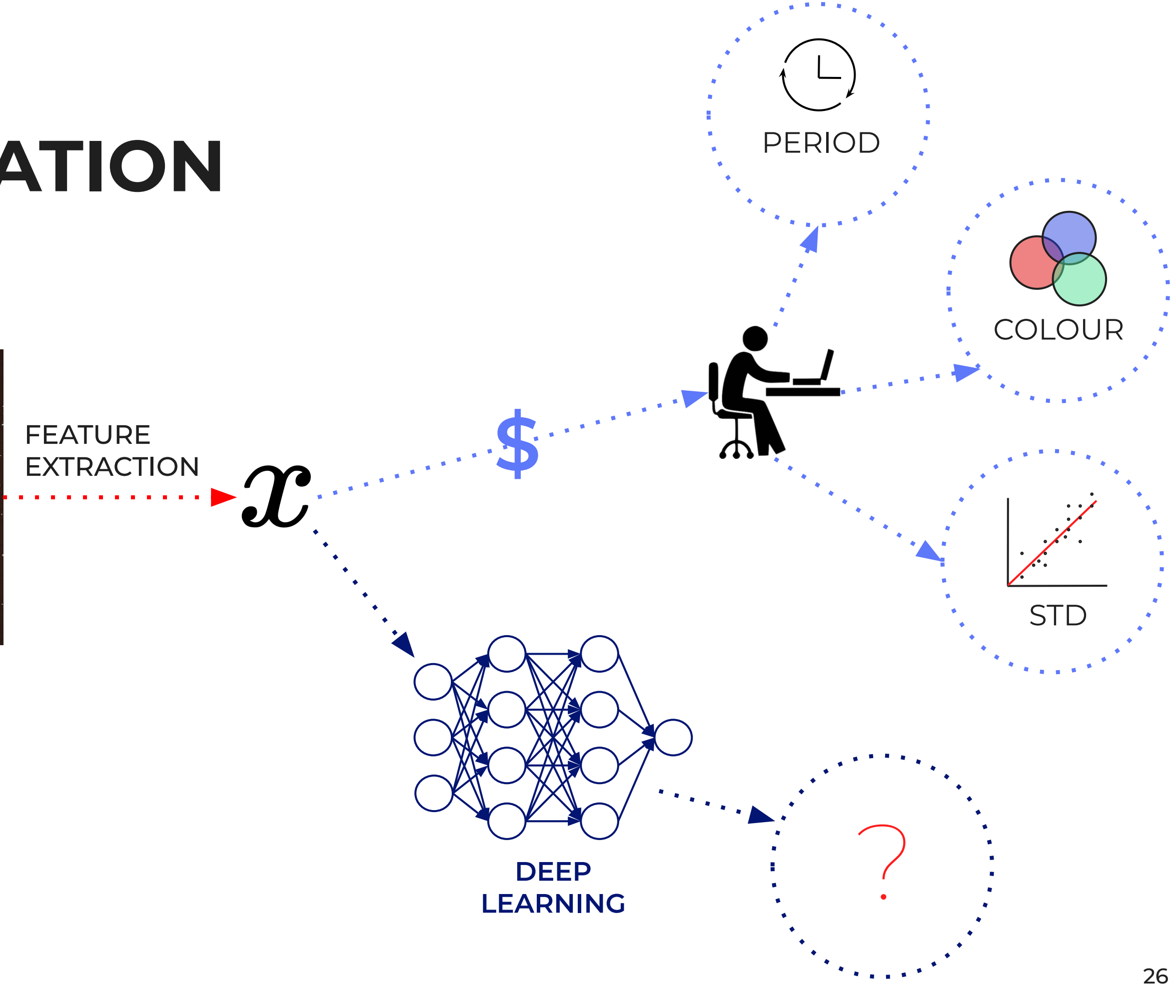
4

CLASSIFICATION



4

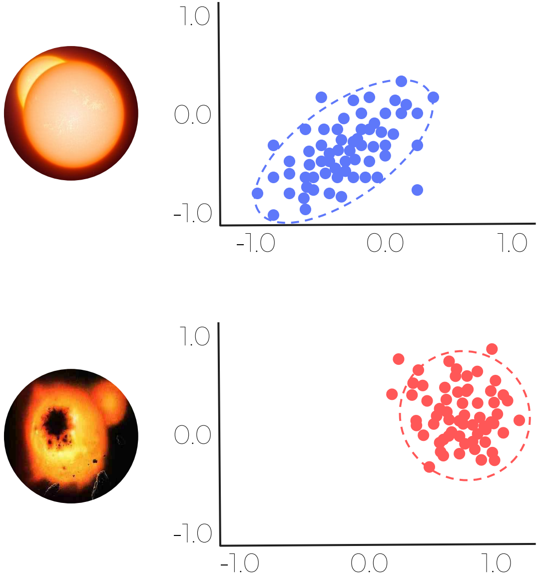
CLASSIFICATION

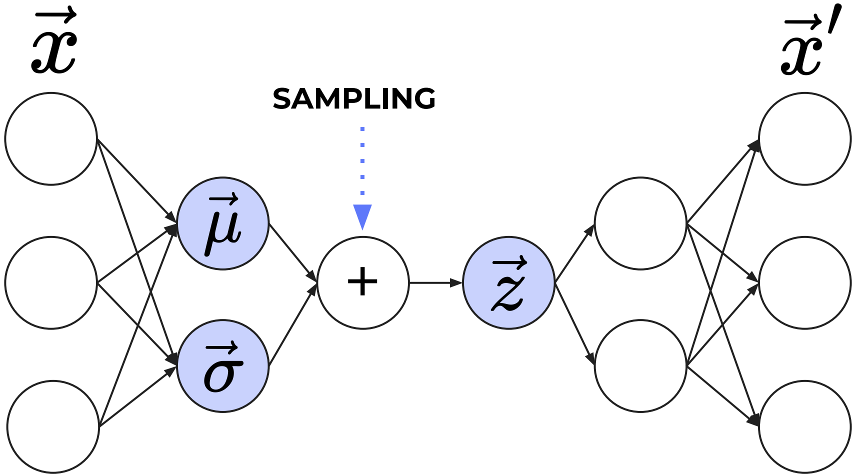


4

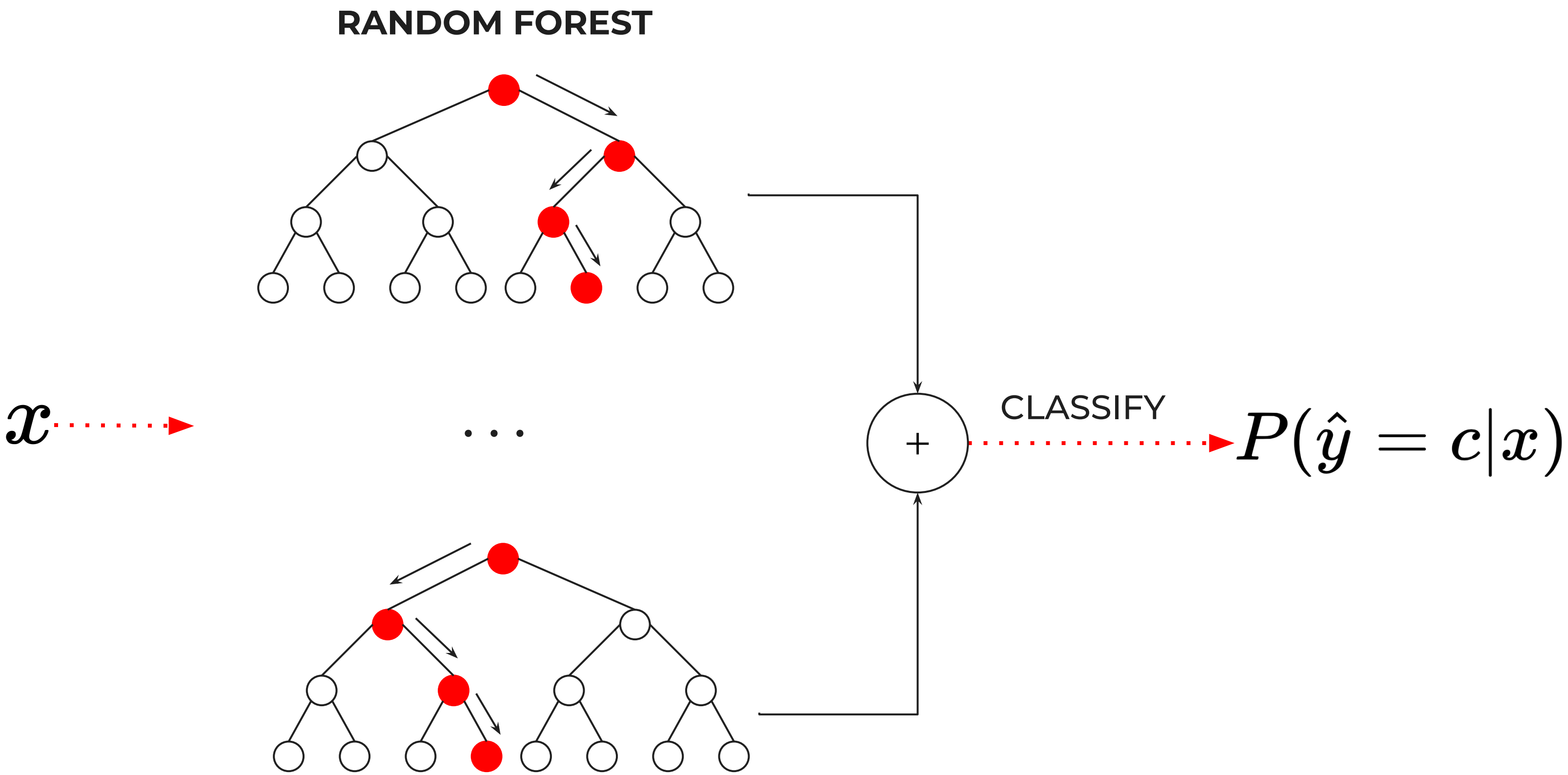
FEATURE EXTRACTION

$\vec{z}$: ENCODING SPACE



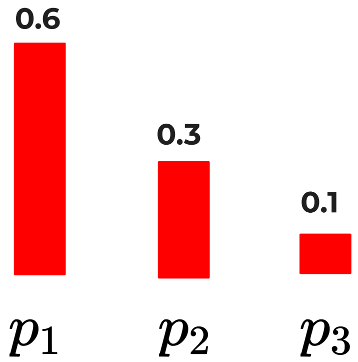


4



4

ENTROPY



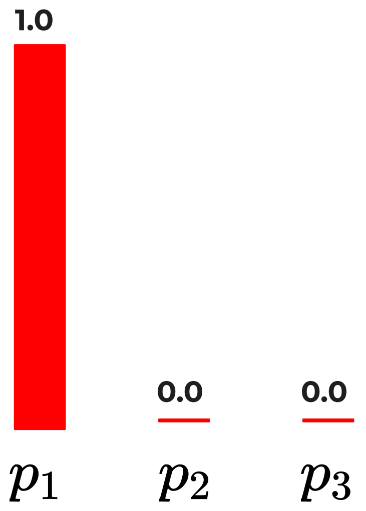
$= 1$

$$H = -\sum_{i=1}^{n} p_i \log\ p_i$$

$$H = 0.389$$

4

ENTROPY

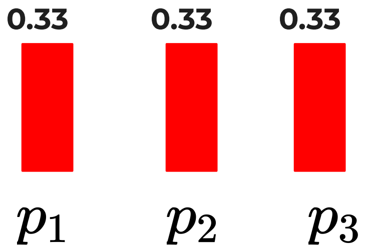


LEAST CONFUSED!

$$H = 0$$

4

ENTROPY



MOST CONFUSED!

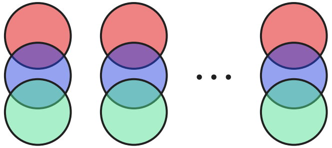
$$H = 0.477$$

5

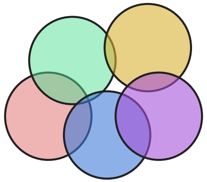
RELATED WORK

5

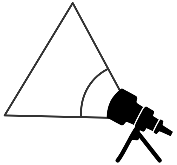
RELATED WORK



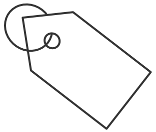
PETERS ET AL.
(2015)



YANG
(2015)



XIA
(2016)



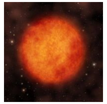
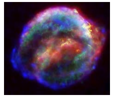
ISHIDA
(2019)

6

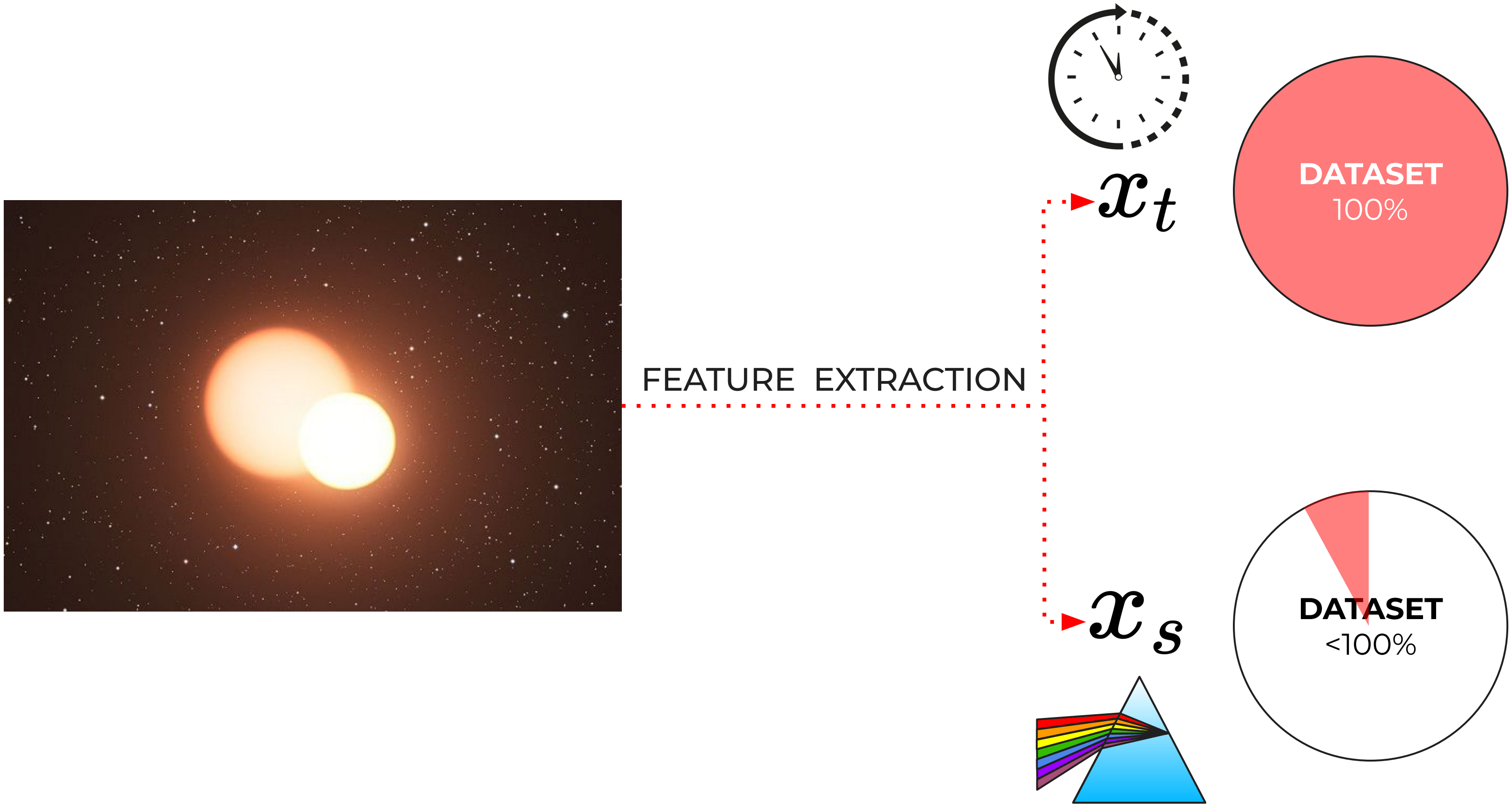
PROBLEM DESCRIPTION

6

CLASSIFICATION

- **VARIABLE STARS**
  - **INTRINSIC**
    - **PULSATING**
      - Cepheids
      - RR Lyrae
      - RV Tauri
      - Long-period variables 
    - **ERUPTIVE**
      - Supernovae 
      - Novae
      - Recurrent novae
  - **EXTRINSIC**
    - **ECLIPSING BINARIES** 
    - **ROTATING VARIABLES**

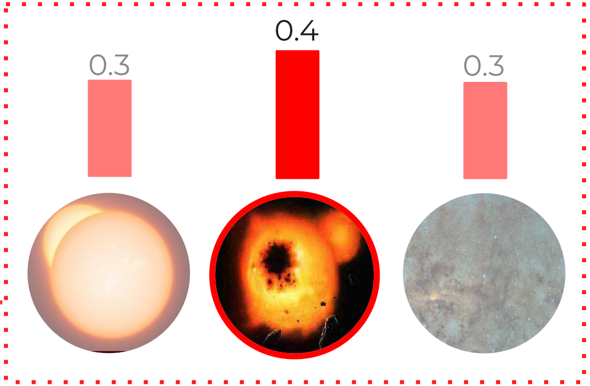
6



6

RANDOM FOREST **N°1**

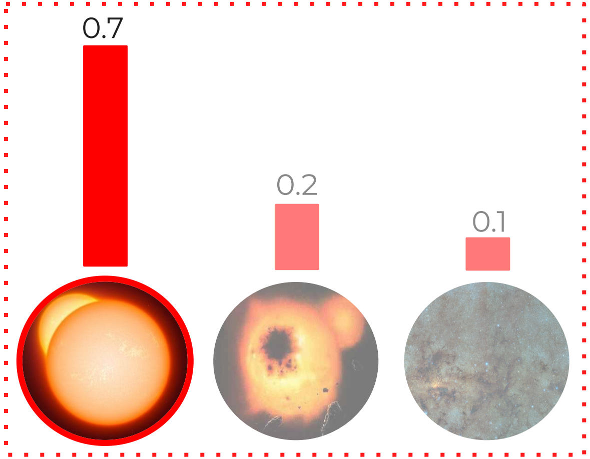
x_t CLASSIFY $\rightarrow P_t(\hat{y} = c|x_t)$



RANDOM FOREST **N°2**

x_t + x_s CLASSIFY $\rightarrow P_{ts}(\hat{y} = c|x_t, x_s)$

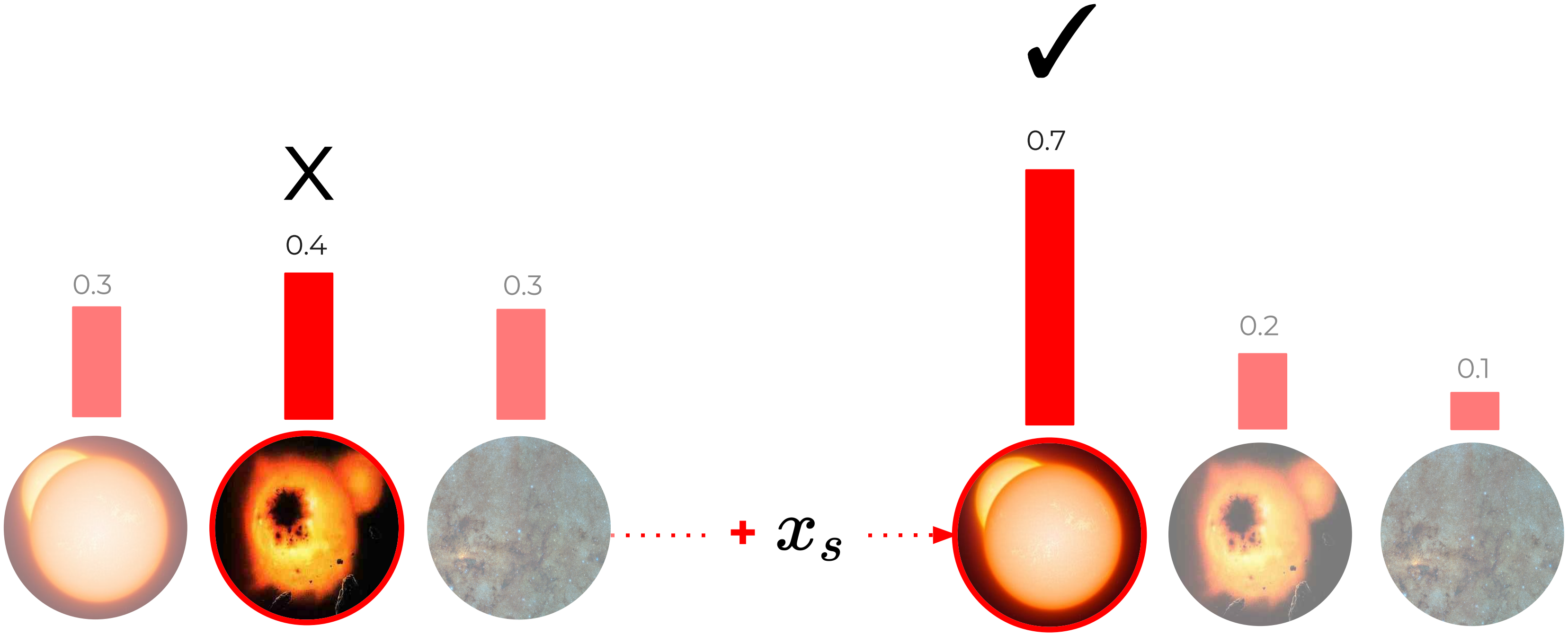
WHICH ONES?



6

WHICH ONES?
CASE 1
ACCURACY

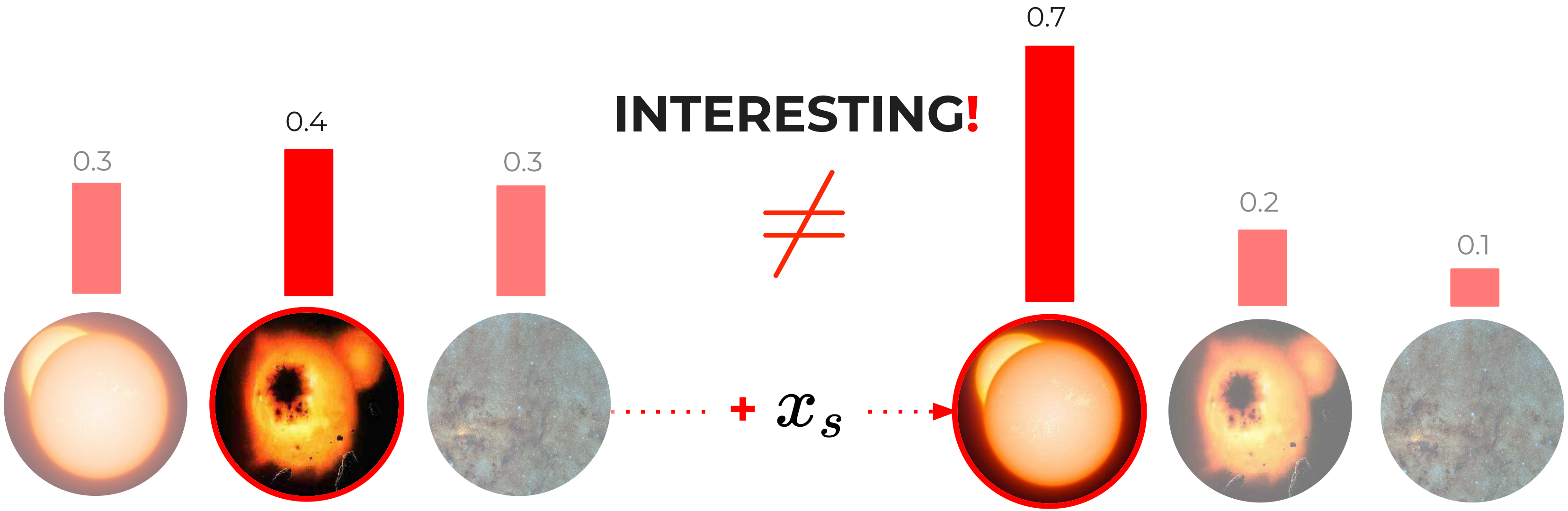
6



$$P_t(\hat{y} = c | x_t)$$

$$P_{ts}(\hat{y} = c | x_t, x_s)$$

6

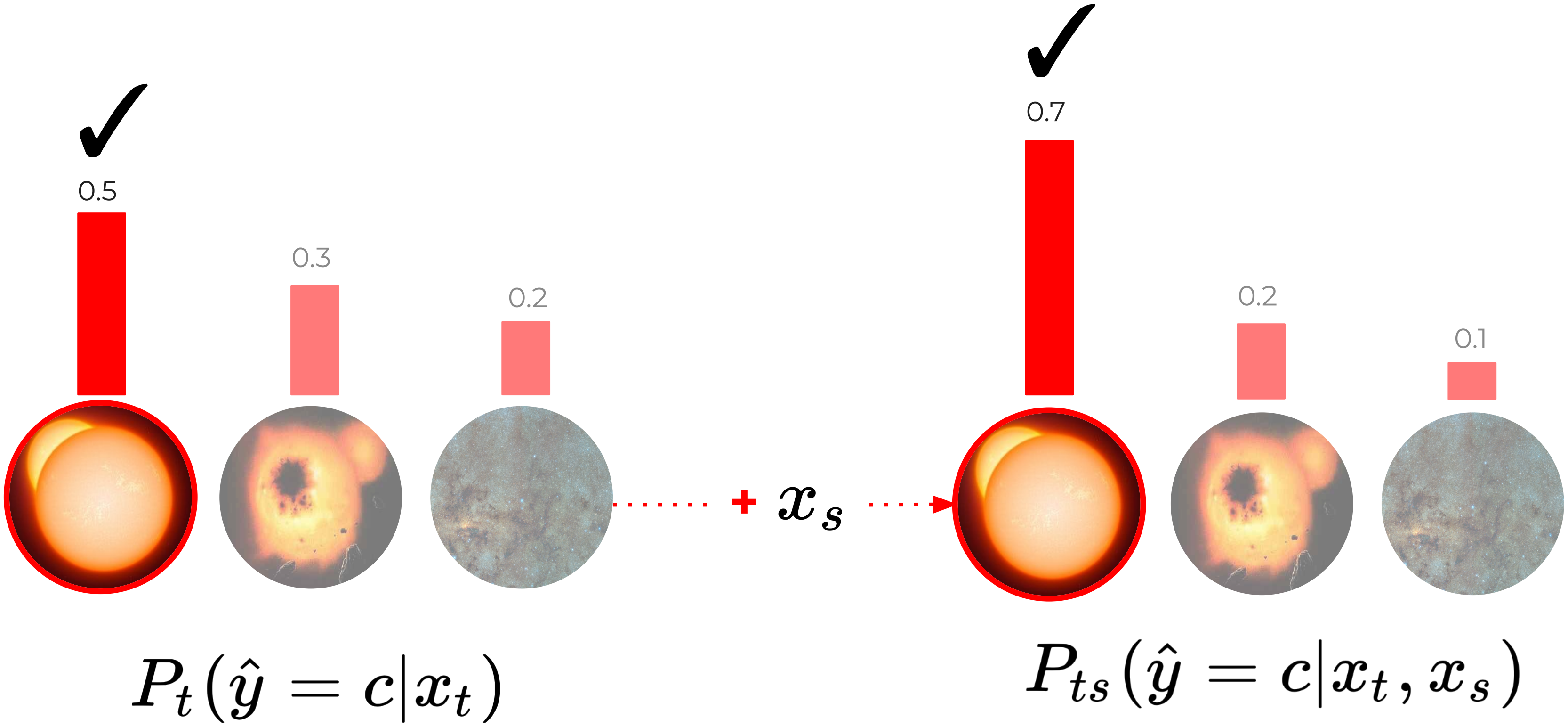


$$\Delta\hat{y}(x_t, x_s) = 1$$

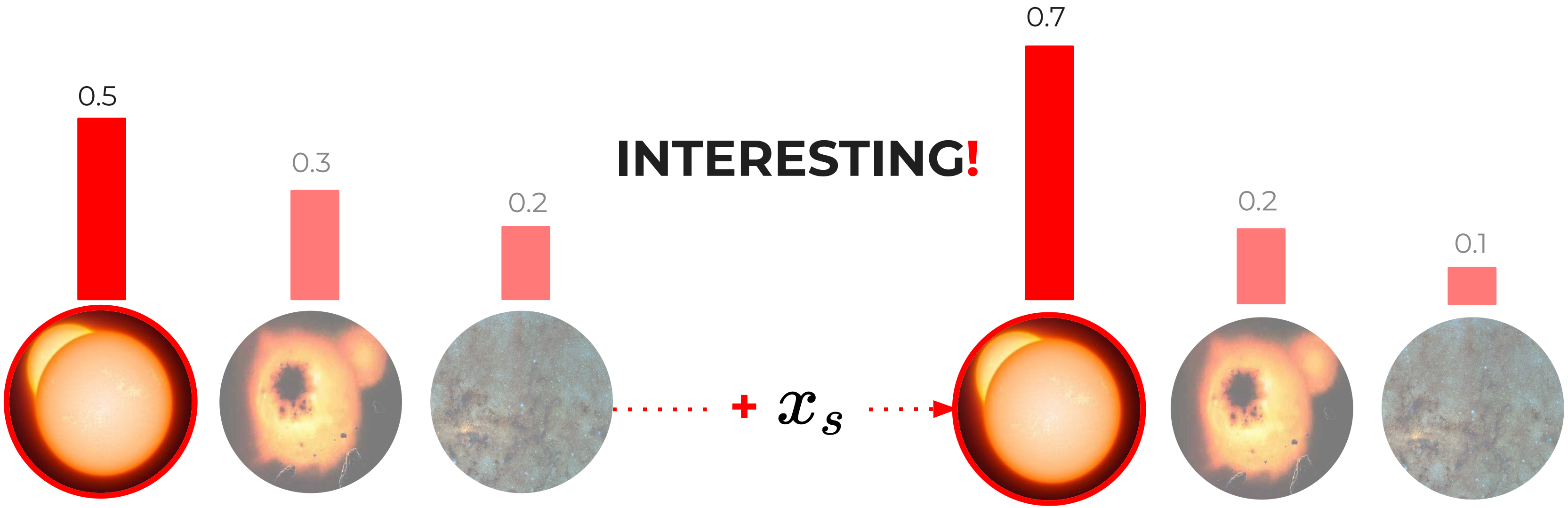
6

WHICH ONES?
CASE 2
CONFIDENCE

6



6



$$H_t(\hat{y}|x_t) > H_{ts}(\hat{y}|x_t, x_s)$$

$$IG(x_t, x_s) = H_t(\hat{y}|x_t) - H_{ts}(\hat{y}|x_t, x_s) > 0$$

6

METRICS

CASE 1

ACCURACY
Formerly wrongly classified

$$\Delta\hat{y}(x_t, x_s) = 1$$

CASE 2

CONFIDENCE
Less confused outcome

$$IG(x_t, x_s) > 0$$

7

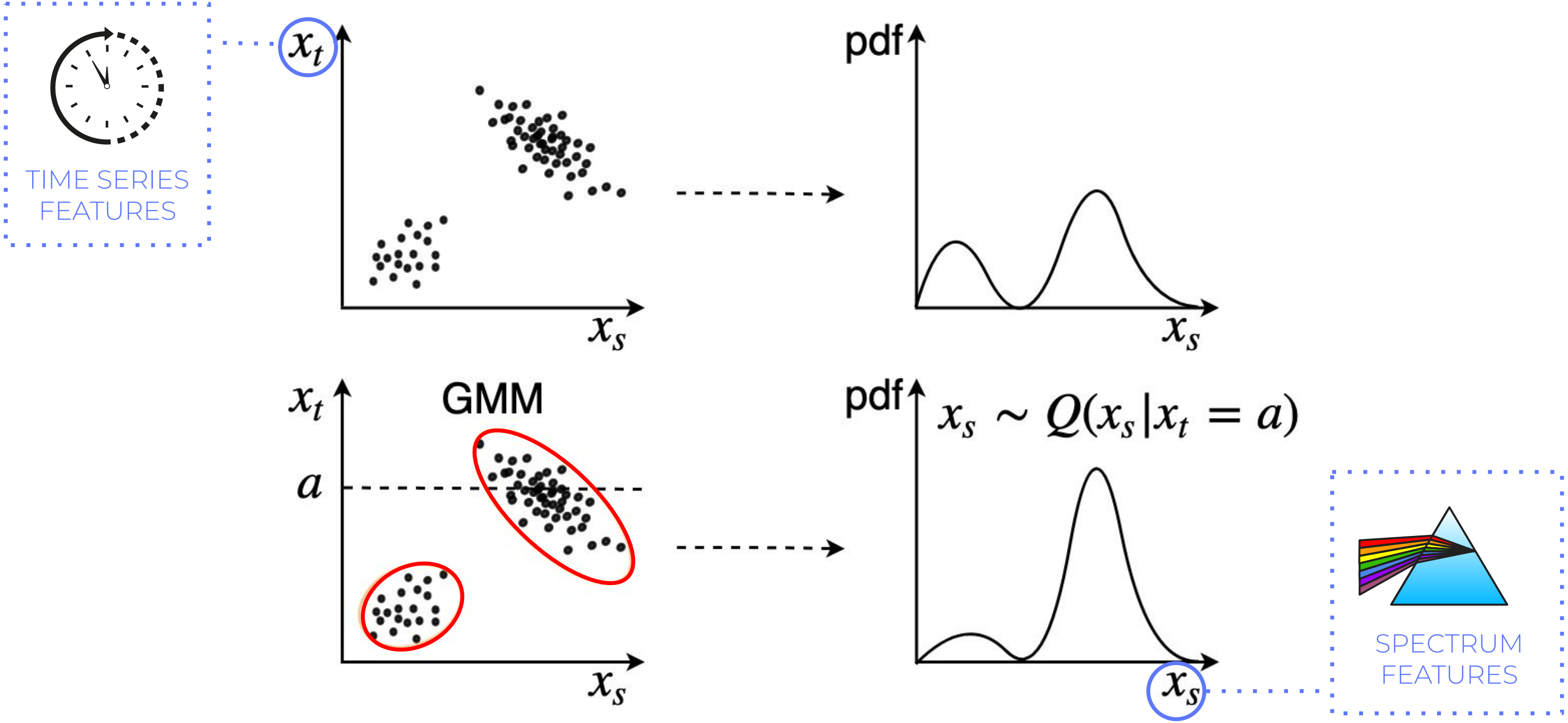
METHODOLOGY

7

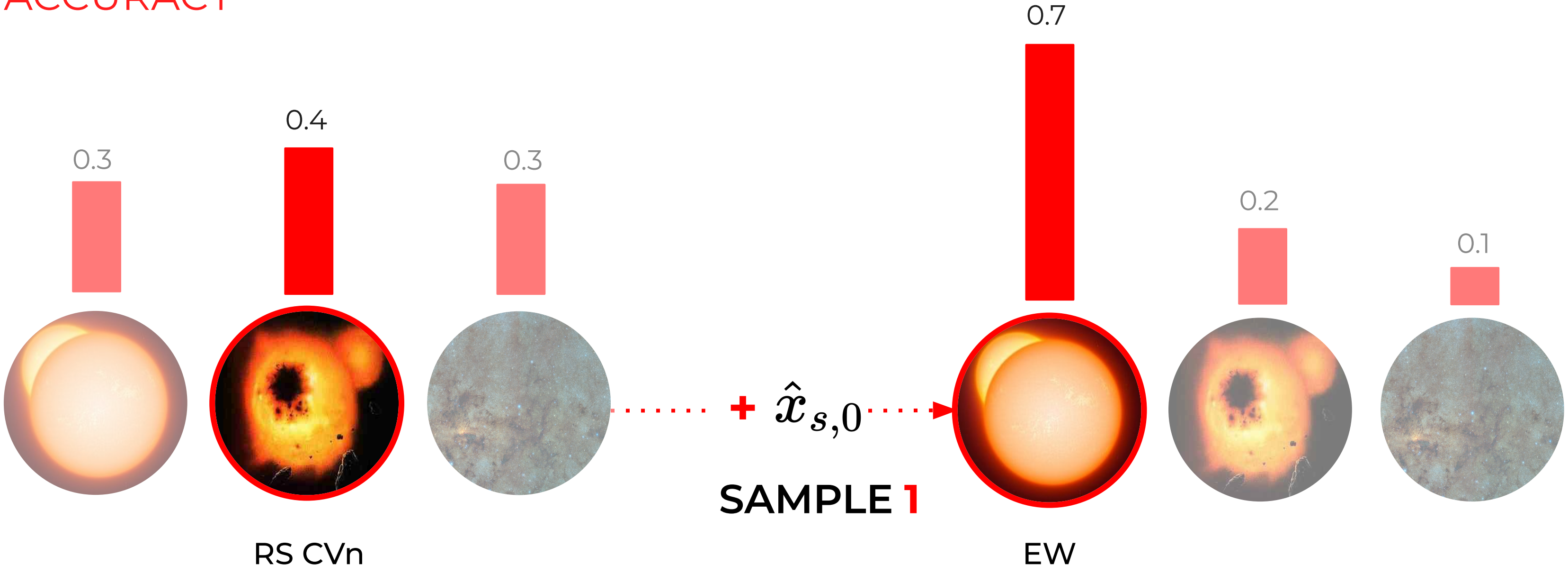
$$x_t \xrightarrow{\text{ESTIMATION}} x_s$$

$$Q(x_s | x_t)$$

7

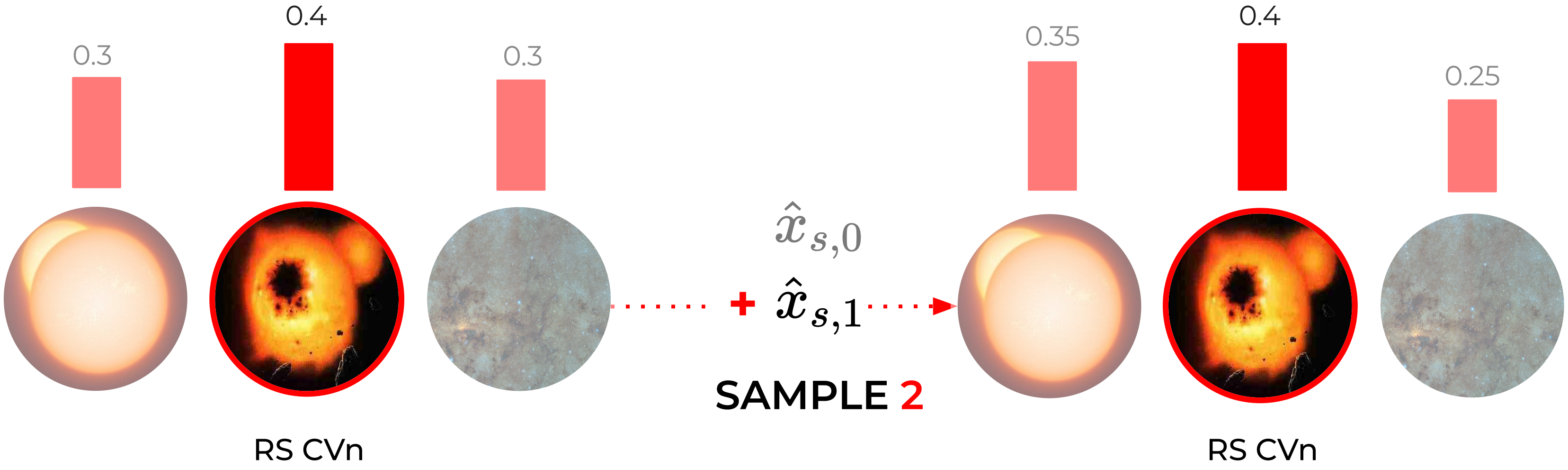

TIME SERIES FEATURES
x_t
x_s
pdf
x_s
x_t
GMM
a
x_s
pdf
$x_s \sim Q(x_s | x_t = a)$
x_s
SPECTRUM FEATURES

CASE 1
ACCURACY



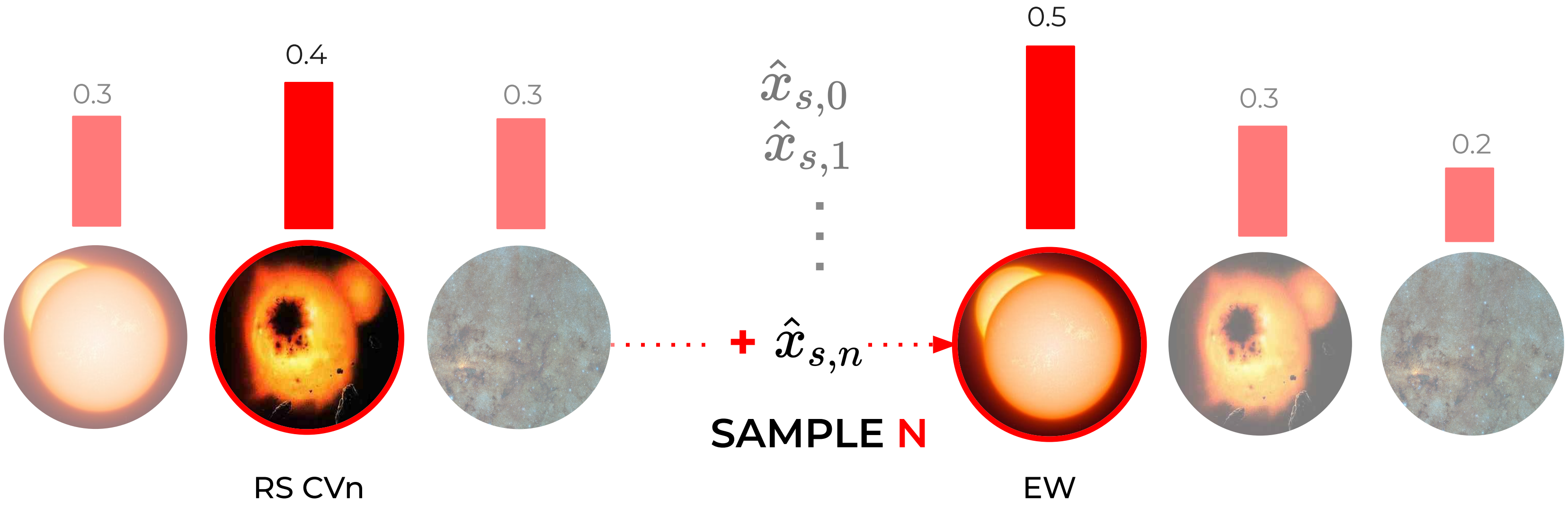
CASE 1
ACCURACY

7



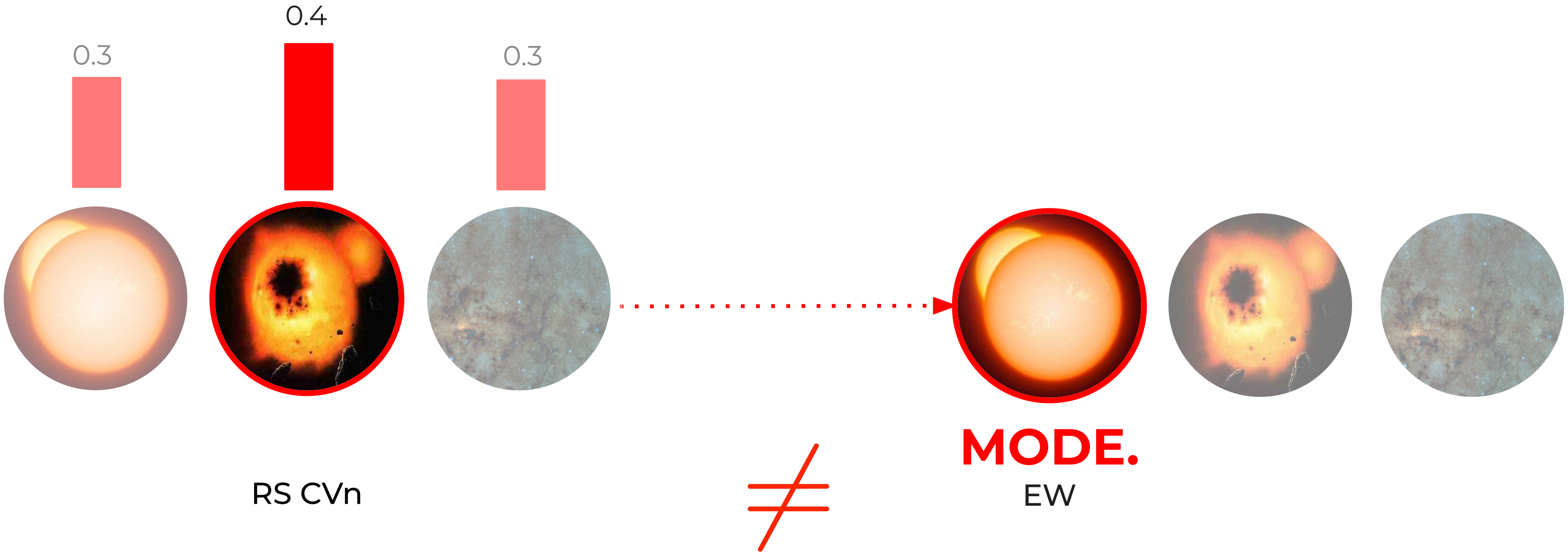
CASE 1
ACCURACY

7



7

CASE 1
ACCURACY

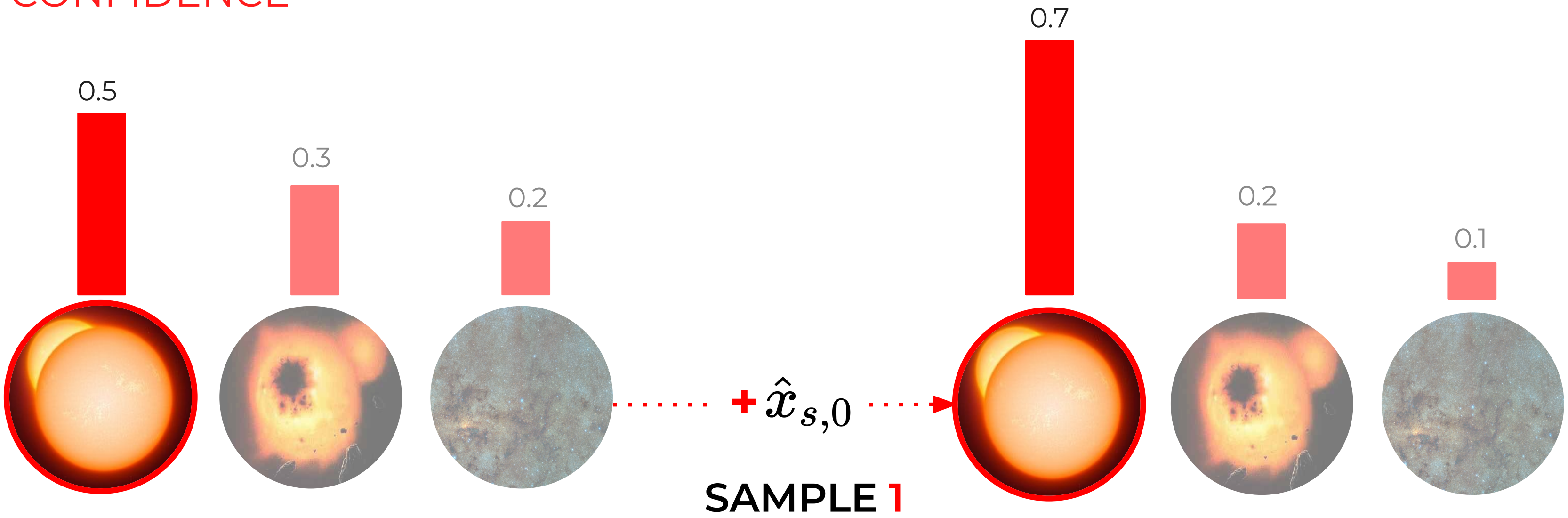


INTERESTING!

$$\Delta\hat{y}(x_t) = 1$$

7

CASE 2
CONFIDENCE

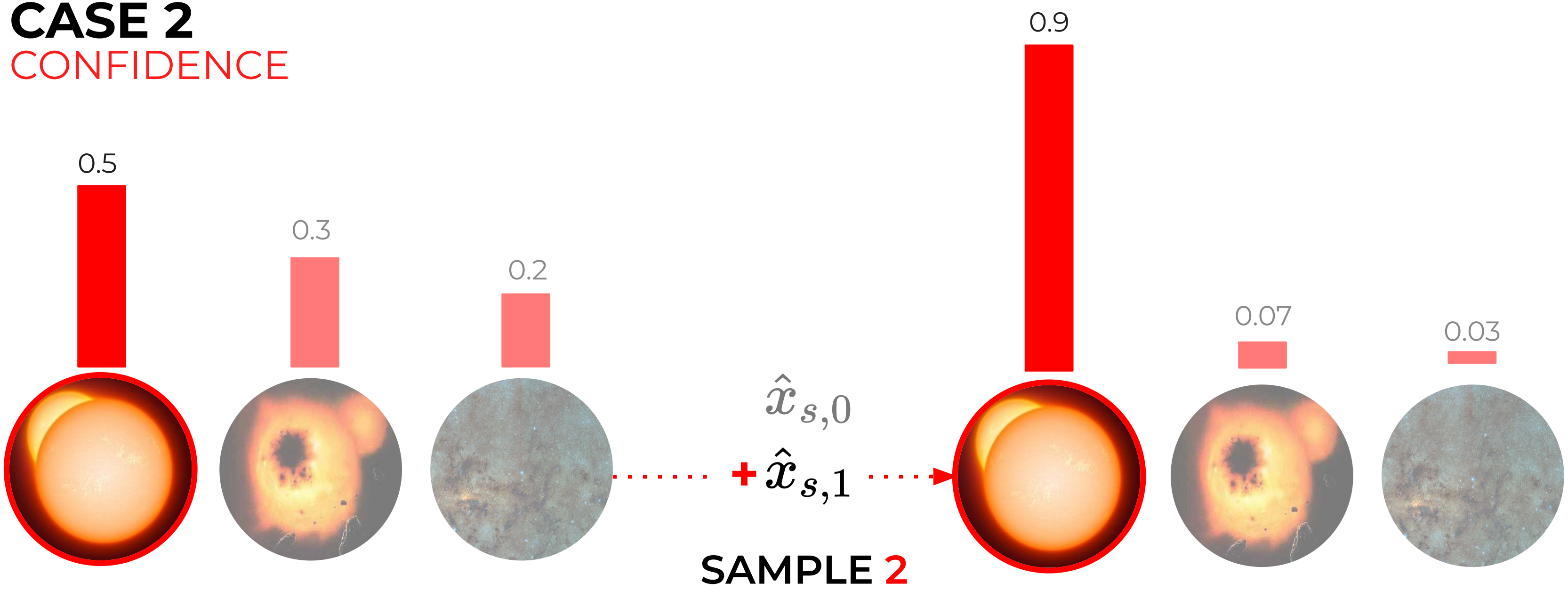


$$H_t(\hat{y}|x_t) = 0.45$$

$$H_{ts}(\hat{y}|x_t, \hat{x}_{s,0}) = 0.35$$

7

CASE 2
CONFIDENCE

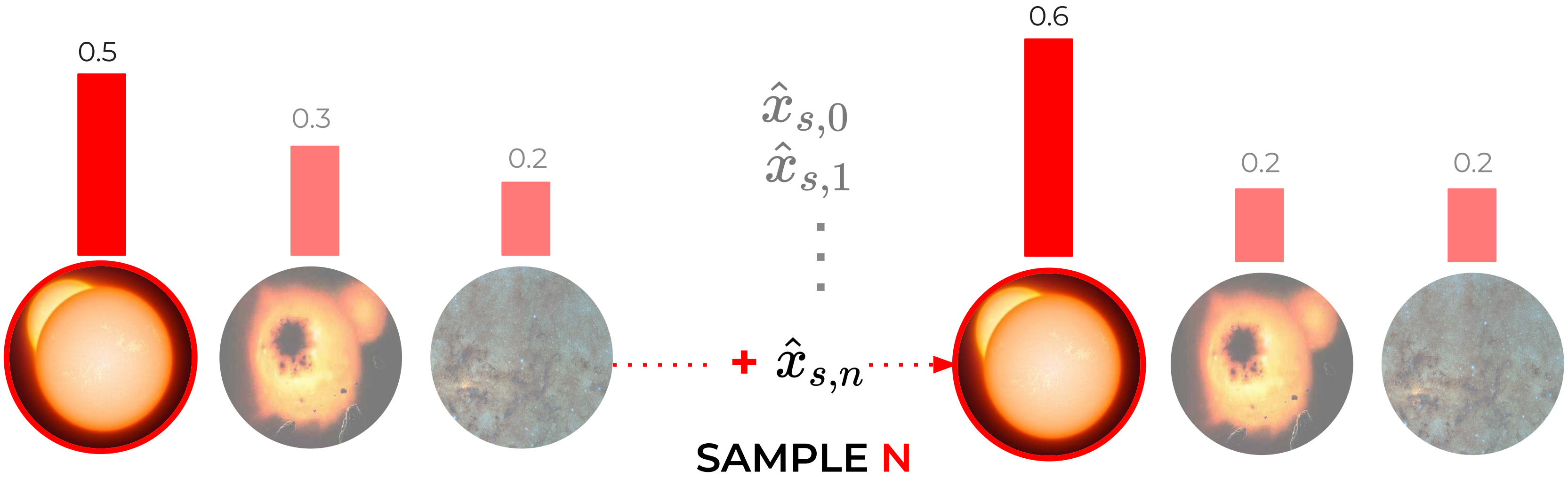


$$H_t(\hat{y}|x_t) = 0.45$$

$$H_{ts}(\hat{y}|x_t, \hat{x}_{s,1}) = 0.17$$

CASE 2
CONFIDENCE

7

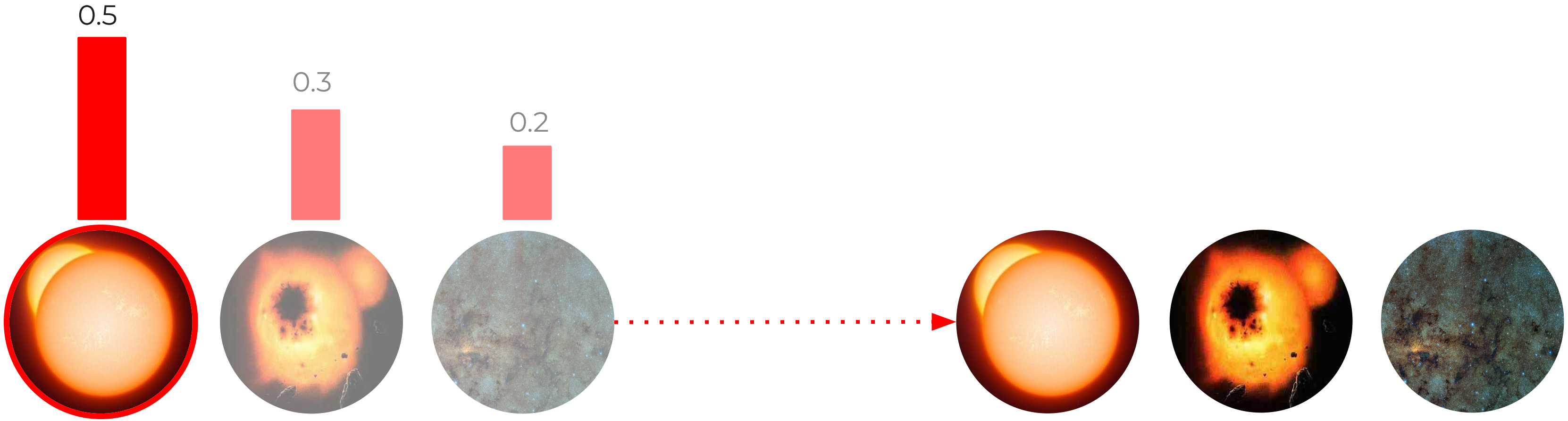


$$H_t(\hat{y}|x_t) = 0.45$$

$$H_{ts}(\hat{y}|x_t, \hat{x}_{s,n}) = 0.41$$

7

CASE 2
CONFIDENCE



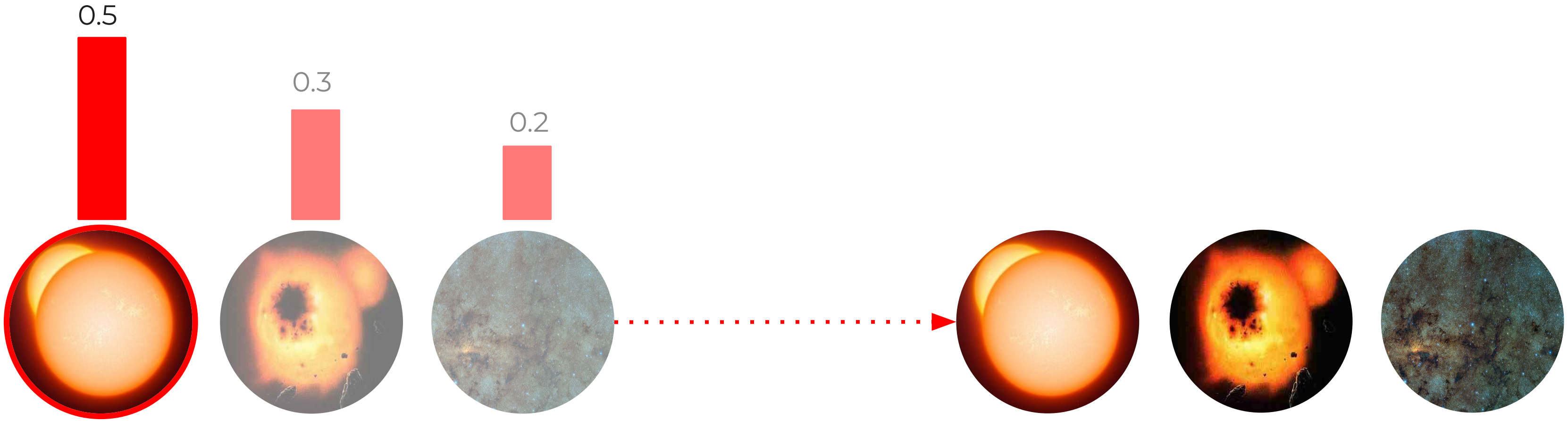
$$H_t(\hat{y}|x_t) = 0.45 \quad > \quad \overline{H}_{ts}(\hat{y}|x_t) = 0.31$$

AVERAGE

LESS CONFUSED
INTERESTING!

CASE 2
CONFIDENCE

7



AVERAGE

$$H_t(\hat{y}|x_t) = 0.45 \quad > \quad \overline{H}_{ts}(\hat{y}|x_t) = 0.31$$

$$IG(x_t) = H_t(\hat{y}|x_t) - \overline{H}_{ts}(\hat{y}|x_t)$$

7

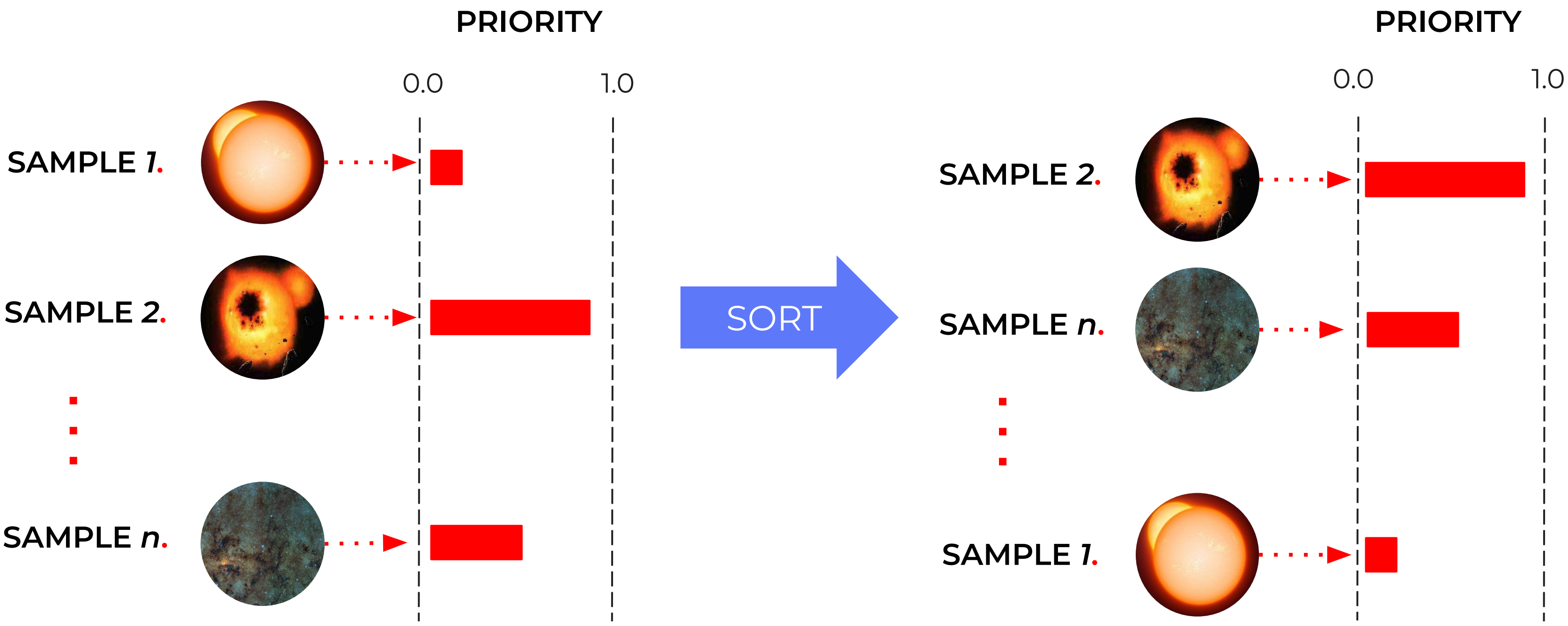
STRATEGIES

$p:$ priority function

$$p : \{IG(x_t), \Delta\hat{y}(x_t)\} \longrightarrow [0,1]$$

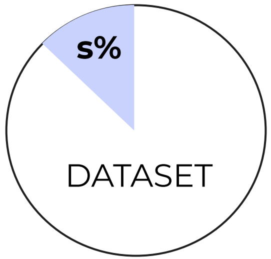
7

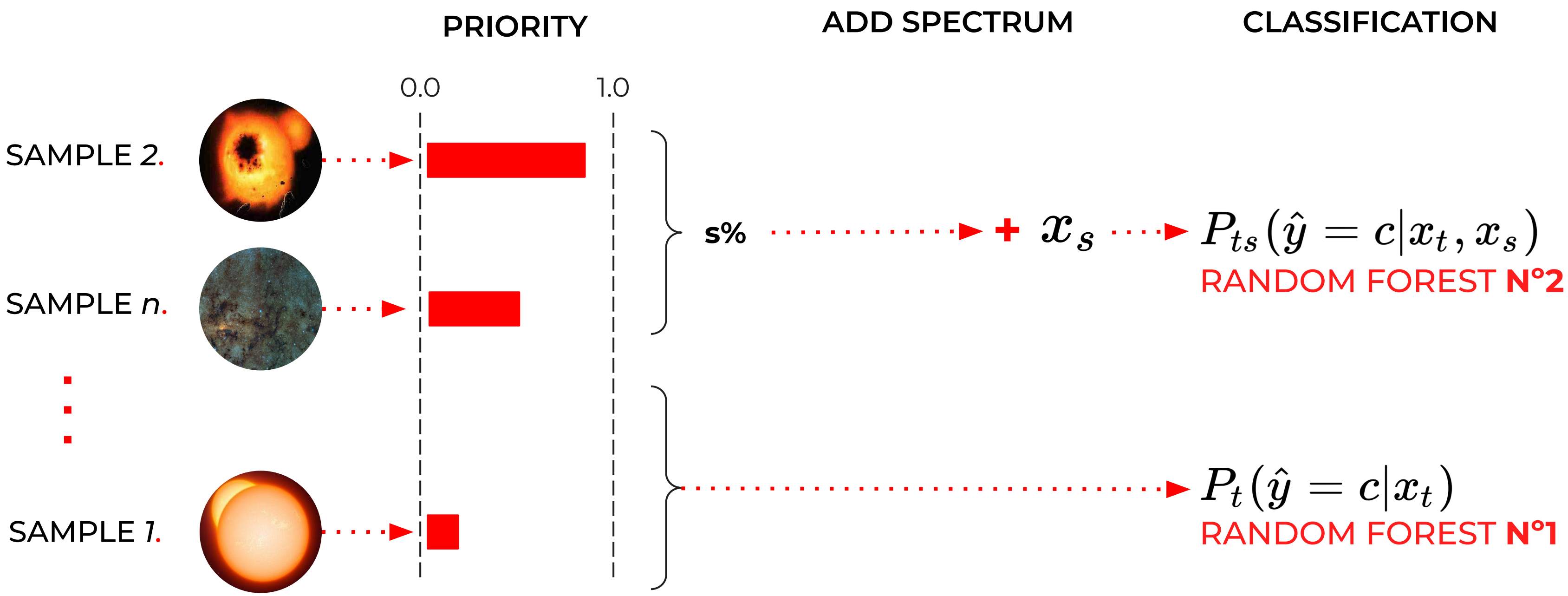
STRATEGIES



7

STRATEGIES





7

STRATEGIES

BASELINE	PROPOSED	FOR REFERENCE
RANDOM	IG	IDEAL SCENARIO
H_t	$IG + \Delta\hat{y}$	
	$H_t + \Delta\hat{y}$	

7

FEATURE EXTRACTION

FEATURE EXTRACTION

7

EXPERT FEATURES
FATS (Nun et al. 2015)

TIME SERIES → $\overrightarrow{fats} = fats_1 \; fats_2 \; \cdots \; fats_{44}$

LEARNED FEATURES
VAE (Kingma et al. 2013)

SPECTRUM → $\vec{\mu} = \mu_1 \; \mu_2 \cdots \mu_7$

7

OVERVIEW

OVERVIEW

1 TRAIN CLASSIFIERS

7

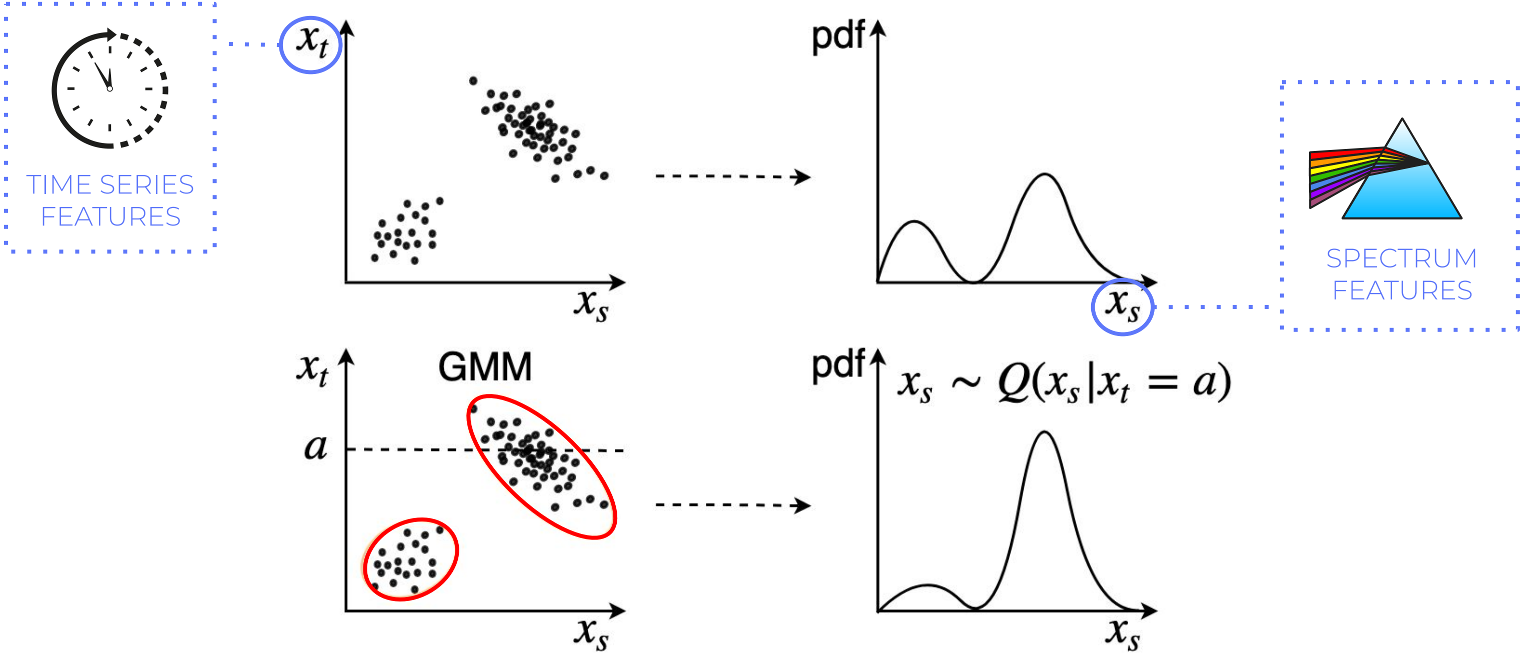
$x_t \cdots\cdots\blacktriangleright P_t(\hat{y} = c | x_t)$ .... RANDOM FOREST **N°1**

$\neq$

$x_t, x_s \cdots\cdots\blacktriangleright P_{ts}(\hat{y} = c | x_t, x_s)$ .... RANDOM FOREST **N°2**

OVERVIEW
2 MODEL FEATURE SPACE

7



7

OVERVIEW

3 METRICS APPROXIMATION

CASE 1

ACCURACY
Formerly wrongly classified

$$\Delta\hat{y}(x_t, x_s) = 1 \quad \overset{\approx}{\dashrightarrow} \quad \Delta\hat{y}(x_t)$$

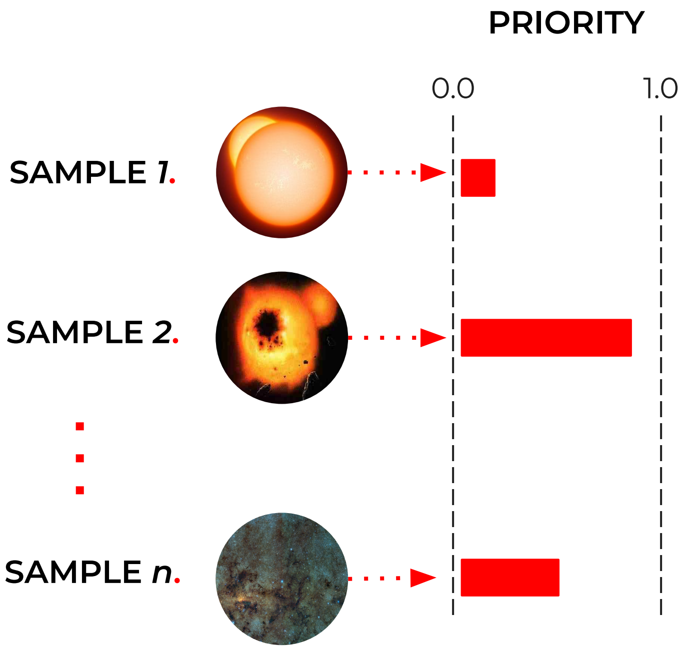
CASE 2

CONFIDENCE
Less confused outcome

$$IG(x_t, x_s) > 0 \quad \overset{\approx}{\dashrightarrow} \quad IG(x_t)$$

OVERVIEW
4 STRATEGIES

7



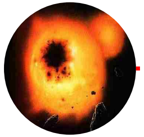
7

OVERVIEW

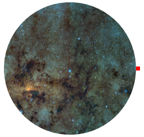
5 SELECTION AND EVALUATION

0.0 1.0

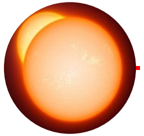
SAMPLE 2.



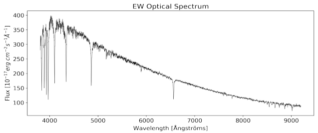
SAMPLE *n*.



SAMPLE 1.



s% + x_s → $P_{ts}(\hat{y} = c | x_t, x_s)$

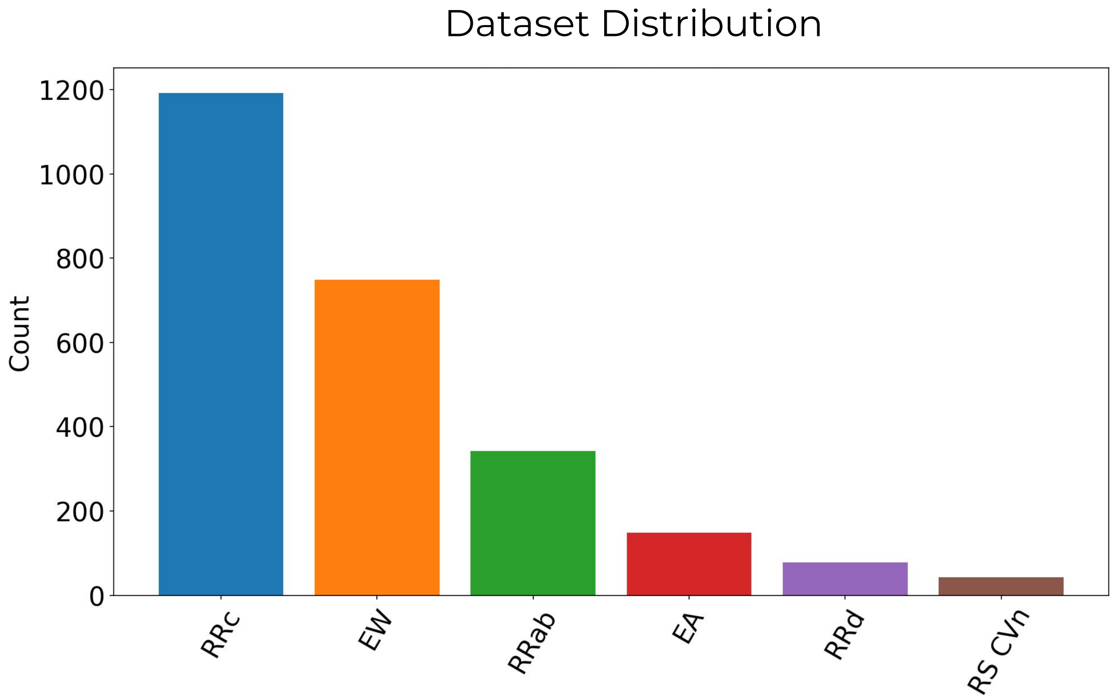


$P_t(\hat{y} = c | x_t)$

8

DATA

DATA.



2554 SAMPLES

8

RESULTS

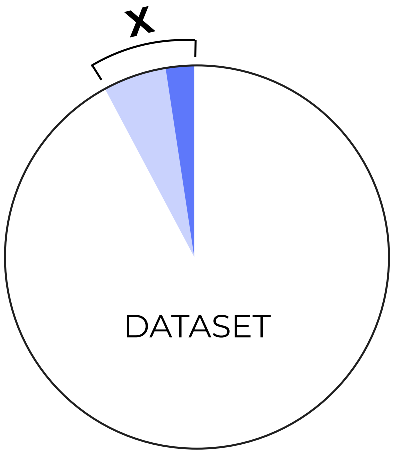
RESULTS
ACCURACY

8

$$\Delta\hat{y}(x_t) = 1$$

113 POSITIVE OBJECTS = 4% DATASET

43 ARE CORRECTABLE = 46% OF AVAILABLE CORRECTABLE



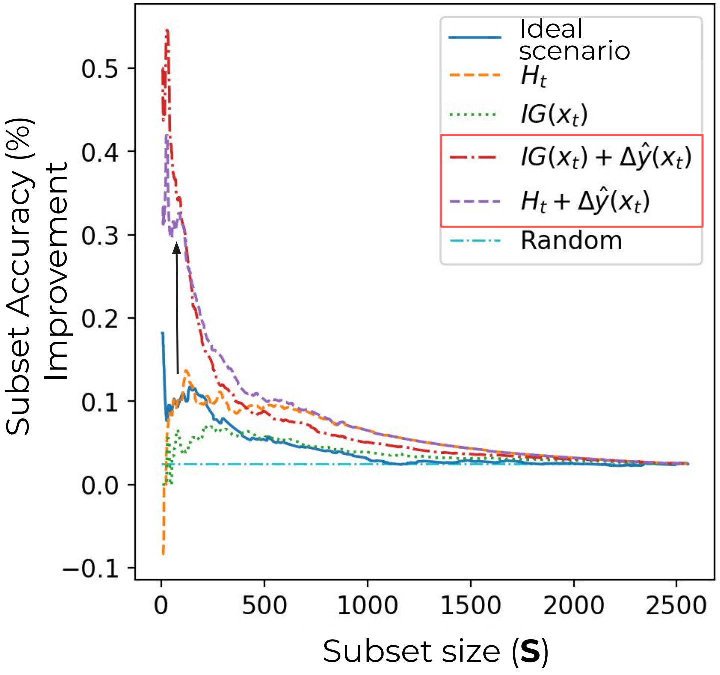
+ Wrong Time Series Prediction (15%)

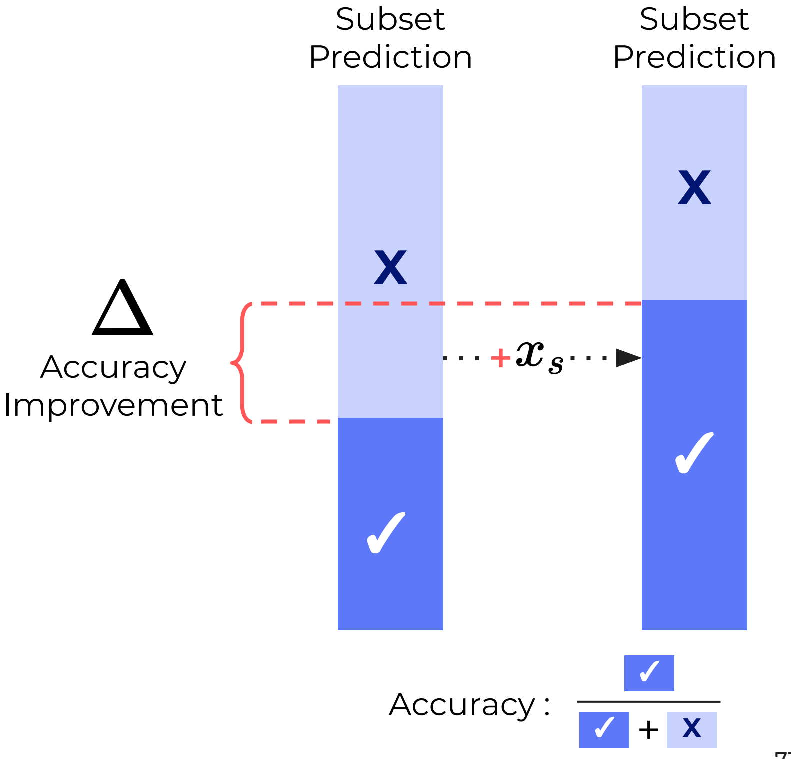
Correctable (4%)
X → ✓

8

RESULTS
ACCURACY

Subset Accuracy (%) Improvement

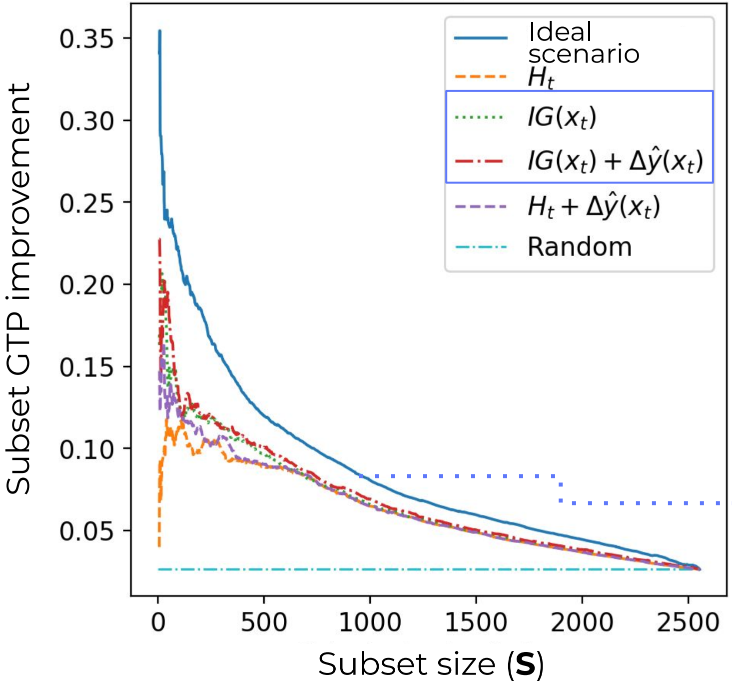




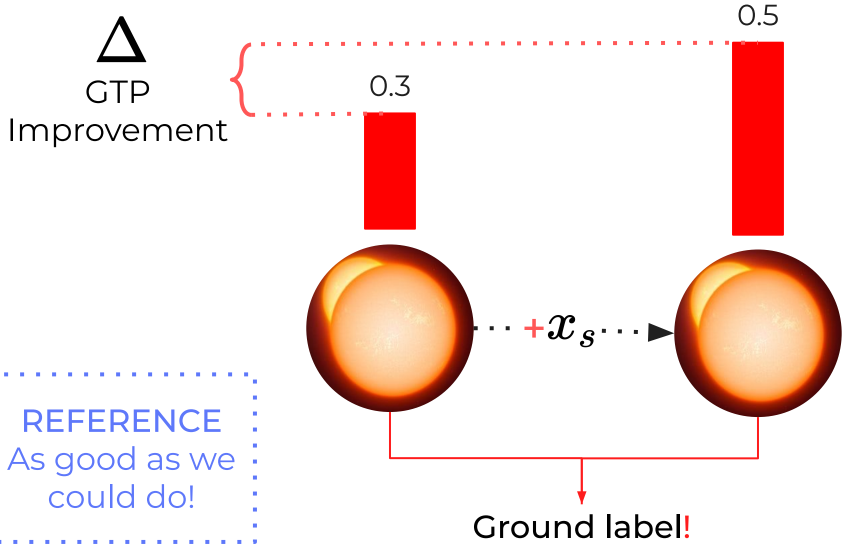
RESULTS
CONFIDENCE

8

GTP average improvement

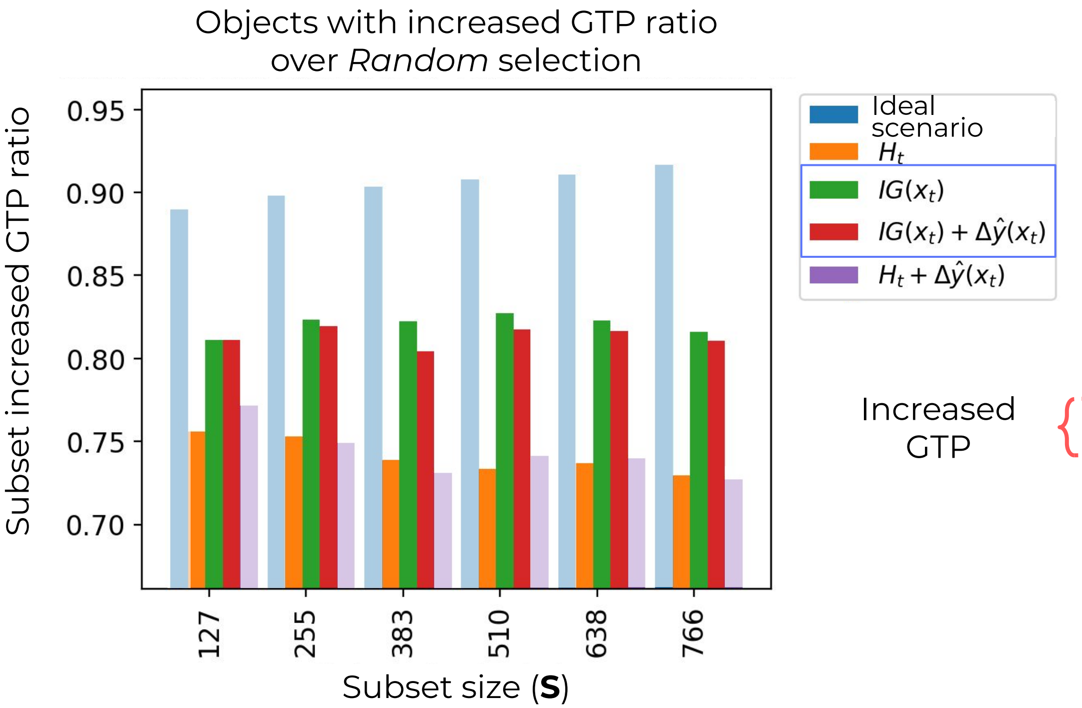


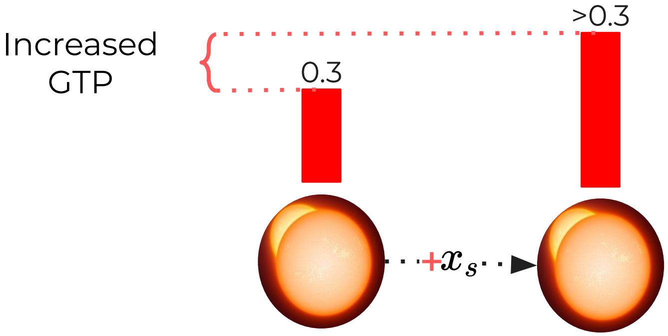
REFERENCE
As good as we could do!



RESULTS
CONFIDENCE

8

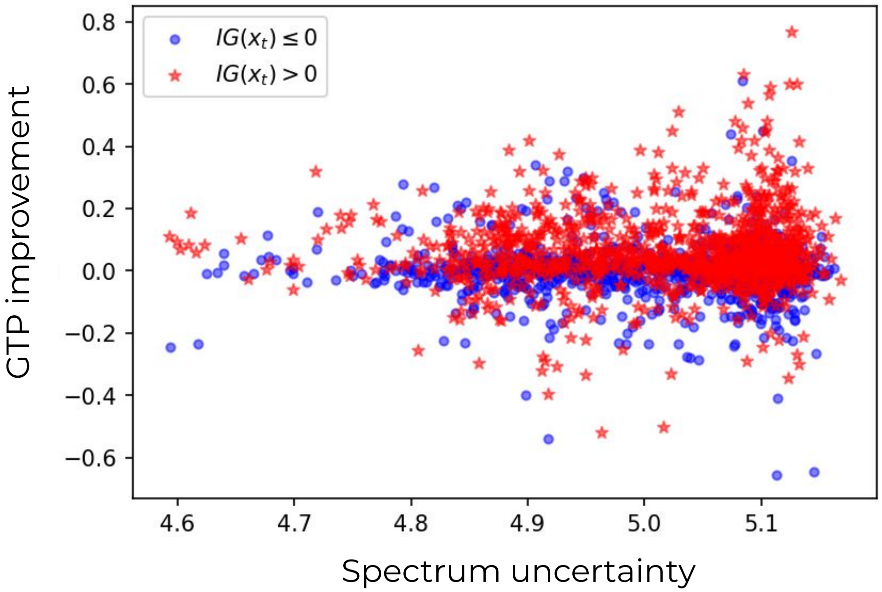




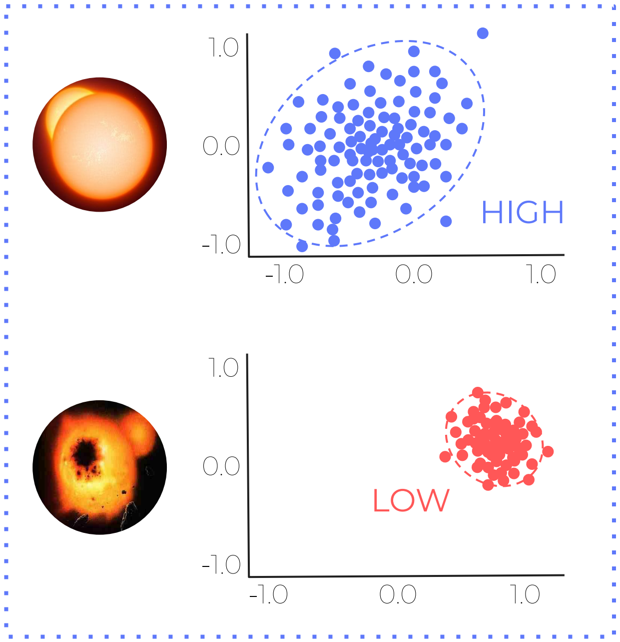
8

RESULTS
CONFIDENCE

GTP improvement v/s spectrum uncertainty



UNCERTAINTY



9

CONCLUSIONS AND FURTHER WORK

9

CONCLUSIONS

- $IG(x_t)$: SURPASSES BASE-LINE STRATEGIES **Δ**GTP
- $IG(x_t)$: IF FURTHER IMPROVED, MORE RESOURCES MAY BE SAVED
- $\Delta\hat{y}(x_t)$: DETECTION OF 46% OF CORRIGIBLE SAMPLES
- $\Delta\hat{y}(x_t)$: SURPASSES BASE-LINE STRATEGIES **Δ**ACCURACY
- WHICH STRATEGY? DEPENDS ON THE USE

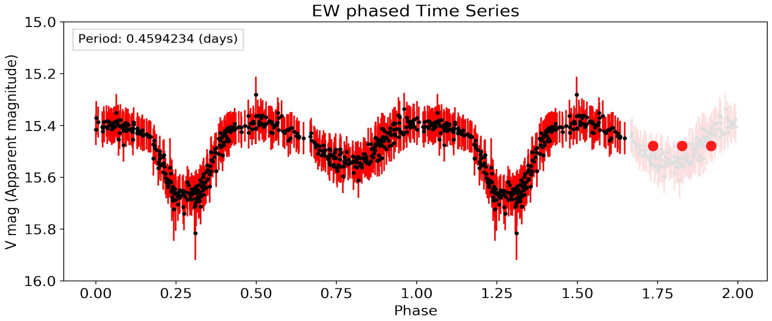
9

FURTHER WORK

- TIME SERIES FEATURE LEARNING
- UNBALANCED CLASSES
- **ONLINE SELECTION**
- $Q(x_t, x_s)$
- OPTIMUM NUMBER OF POINTS TO BE QUERIED
- OTHER METRICS

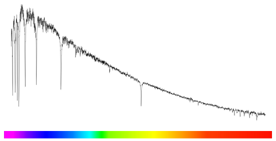
8

ONLINE SELECTION



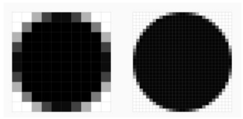
OPTION **1**
MORE EPOCHS

OPTION **3**
SPECTRUM



OPTION **2**
MORE COLOR BANDS

OPTION **4**
SPECTRUM RESOLUTION



9

QUESTIONS?

RESULTS
ACCURACY

8