Characterizing Systematics in (Astro)Physical data:
Insights from Current Experiments and Implications for LiteBIRD

Emile CARINOS, IRAP, 1st year PhD student
emile.carinos@irap.omp.eu

References :
1- Zachary R. Claytor et al. 2021 :Recovery of TESS Stellar Rotation Periods Using Deep Learning
2-Jade Powell et al. 2016 : Classification methods for noise transients in advanced gravitational-wave detectors II: performance tests on Advanced LIGO data.
3-Beryl Hovis-Afflerbach et al. 2020 : Identifying and Repairing Catastrophic Errors in Galaxy Properties Using Dimensionality Reduction
4-Olaf Ronneberger et al. 2015 : U-Net: Convolutional Networks for Biomedical Image Segmentation

I-Introduction
Systematics, but what kind?

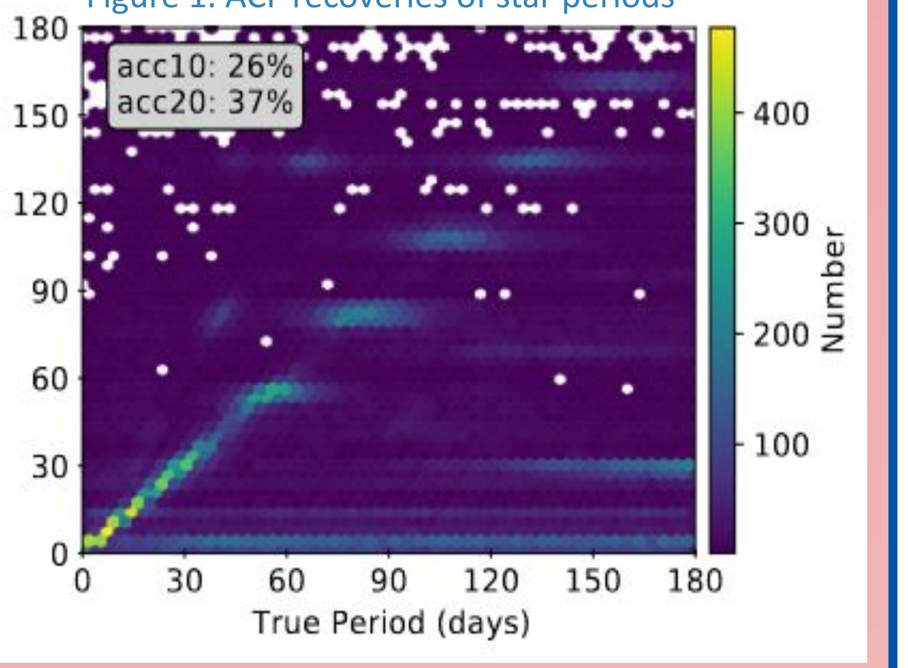
→ Noise in measurements : **statistical** and **systematic** noise.
→ Statistical noise has its origin in random processes
→ **Systematic noise** is caused by **factors that consistently influence** the measurements in a **particular, predictable direction**.
→ **Unexpected systematic errors can occur if the factors are not well known**

What can we do with such unpredicted, hard to characterize systematics? Is it possible to go beyond these limitations?

→ Example of systematic noise : Periodic systematics parasitizing TESS (Transiting Exoplanet Survey Satellite) data of rotating stars.

→ In order to predict a star's rotation period, the autocorrelation power spectrum of its measured light curve is computed.

Figure 1. ACF recoveries of star periods



→ TESS has a **14 days** rotation period around the earth and a **27 days** segmented observation strategy
→ These systematics are directly seen on figure 1. as well as their aliases.

→**LiteBIRD** in this story : while the scientific objective is very different, the question of the mission's observation strategy and the impact of systematics on the measured data is essential.
The LiteBIRD simulation framework (https://github.com/litebird/litebird_sim) could provide Time-Ordered Data (TOD) and maps taking into account **realistic noise** and different **scanning strategies** to **test different systematics characterization methods.**

II - Tracking systematics :
a Classification Problem

→ Identifying and characterizing a systematic error in data can be seen as a **classification problem**.
→ Classification is all about **balancing discriminativeness and invariance**
→Very **general** way to formulate this problem : Insights from various approaches to this problem can help
→ **Key problem: finding an appropriate space to represent the data in.**

The following examples explore possible methodologies that use the strengths of computer vision to solve classification problems.

III - Augment dimensionality to better reduce it
(trust me, it works)

1

Going back to the TESS example, one solution explored to **go beyond the limitations set by the systematics** was to :
→**Transform** the TOD into a **time-period representation** using the Continuous Wavelet Transform (CWT).
→Apply a Convolutional Neural Network (**CNN**) to the obtained 2D "**image**", having previously trained the CNN on realistic simulations.

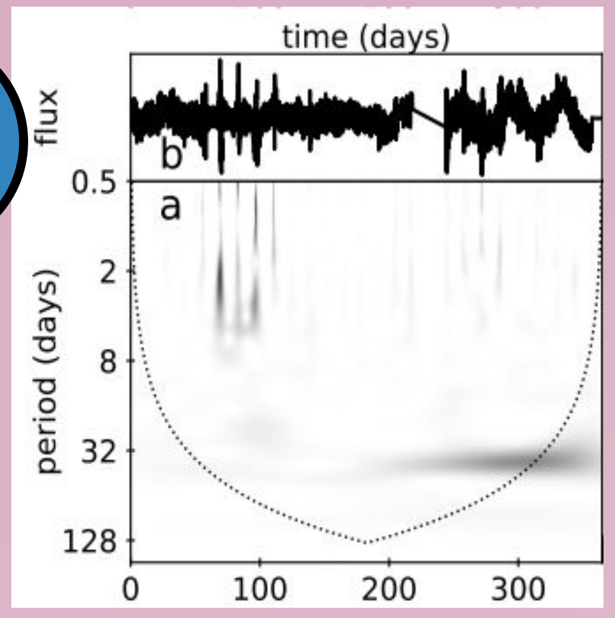


Figure 2. Time-Period representation of the TOD, from its CWT

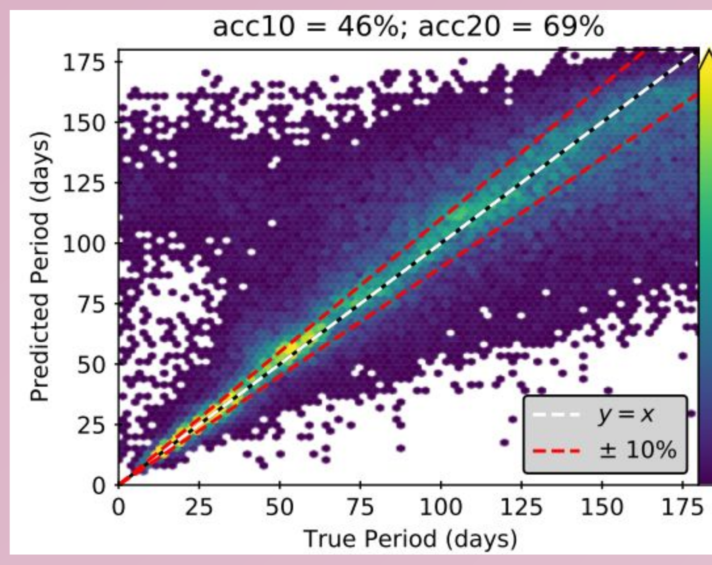


Figure 3. CNN recoveries of star periods

→ The CNN performs better at this task than traditional methods
→ **Recognition of Morphological properties** : exactly what computer vision excels at

Gravitational Wave data analysis : classification of transients

2

→ **Hunt for transients** : noise or relevant data?
→ Different approaches, some ML based some not

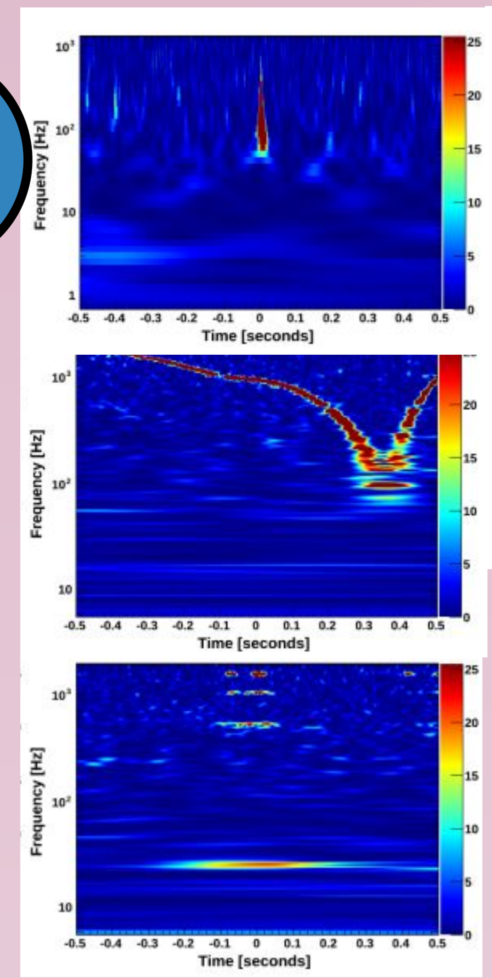


Figure 4. Spectrograms of transients typically found in aLIGO data

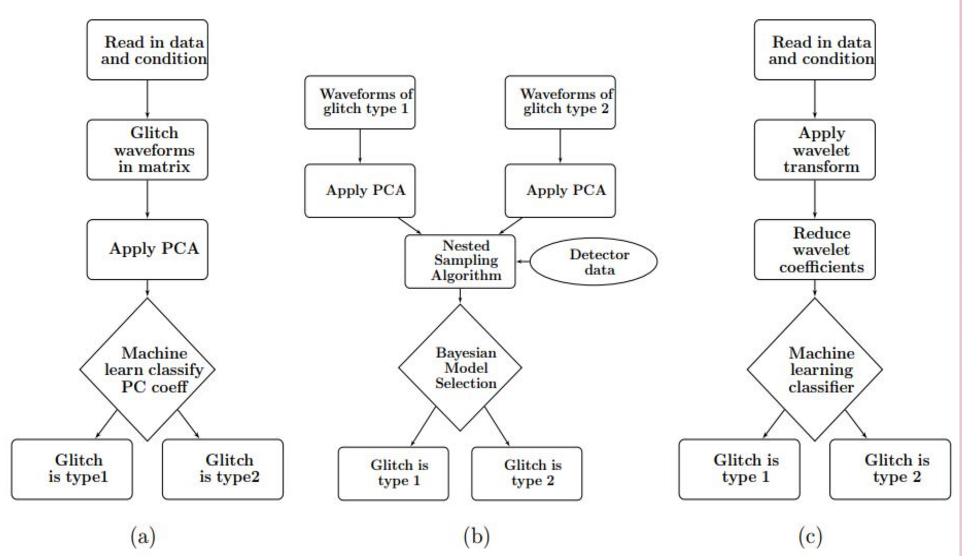


Figure 5. Classification procedures for the three different methods used in this study.
(a) PCAT applies **principal component analysis** to all transients detected in a stretch of data and then applies a **machine learning classifier** to the principal component coefficients.
(b) PC-LIB uses a combination of **principal component analysis** and **Bayesian model selection** to determine the glitch type.
(c) WDF-ML applies a **machine learning classifier** to wavelet coefficients obtained by applying a **wavelet transform** to the transients in the data.

IV - Identify and correct Outliers in data :
Highlight Invariance to allow Discrimination

3

→ Data with **high dimensionality** : discriminativeness is strongly reduced
→ To mitigate this, **dimension-reduction** methods such as t-distributed Stochastic Neighbor Embedding (**t-SNE**) or ML-based **Self Organizing Maps** can be used to plot data points that have very close characteristics as **neighbors in a 2D map**. This can be used to **identify and correct outliers** in measurements, such as in this example of photometric redshift measurements of galaxies.

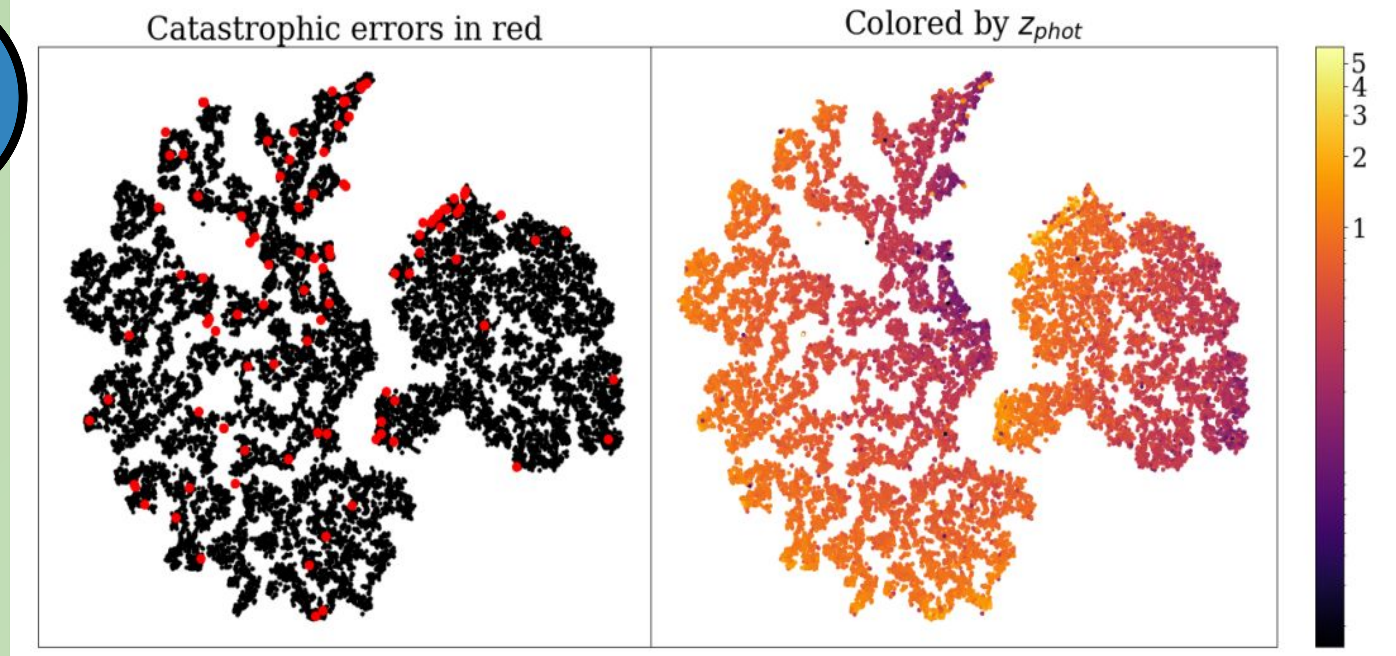


Figure 6. t-SNE maps of "neighbour" galaxies, based on their photometric measurements

→ Whatever the dimension, we can think of ways to **transform** the data we're working with to an **appropriately classifying 2D space**
→ One issue lies in the **interpretability** as well as the **adaptability** to unexpected systematics of the methods shown here.
→ Do this with algorithms that show what they actually discriminate (**U-net** in fig 7). From there, we could even think about using such "blind" discrimination to **define new classes of unexpected systematics.**

4

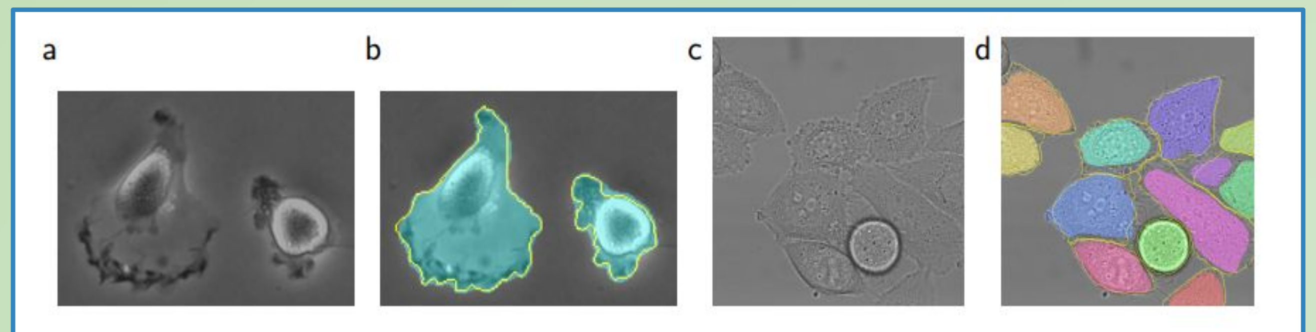


Fig. 7 Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

V - Conclusion

→ The **characterization of systematics**, seen as a classification problem, can be **successfully investigated via the approach of computer** vision.
→ These **snapshots** from different fields, amongst many others, will be sources of **inspiration for my future work on LiteBIRD systematics**.
→ Bridging the gap between the particularities of this experiment and the other showcased here is a challenging yet rewarding challenge I am very excited to be working on for the next 3 years.