



# Outlier detection via Topological Data Analysis (TDA)

Diego Jaramillo  
Agustina Blanco  
Nathan Rutherford  
Xiangyu Jin

T.A.: Edgar Ortiz  
Advisor: Matthew Graham

The logo for the LA SERENA SCHOOL FOR DATA SCIENCE 2022, featuring a dark background with a grid of colorful squares (green, yellow, blue) and the text 'LA SERENA SCHOOL FOR DATA SCIENCE 2022' in white. Below the main text, it says 'Applied Tools for Data-driven Sciences' and 'August 1-12, 2022' in a smaller font.

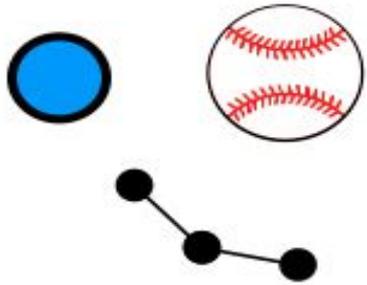
LA SERENA SCHOOL  
FOR DATA SCIENCE 2022  
Applied Tools for  
Data-driven Sciences  
August 1-12, 2022

# What is TDA?

- TDA uses the shape of the data to analyze a dataset. E.g., outlier detection and inference.
- A common approach to TDA is **Persistent Homology**

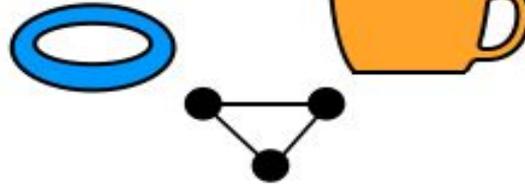


# What is TDA? (Cont'd)



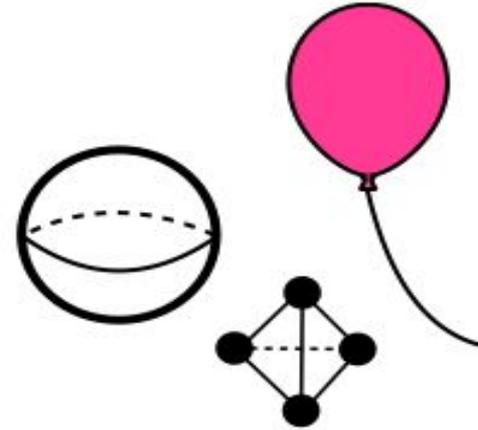
**Connected  
Components**

$$H_0$$



**Holes**

$$H_1$$

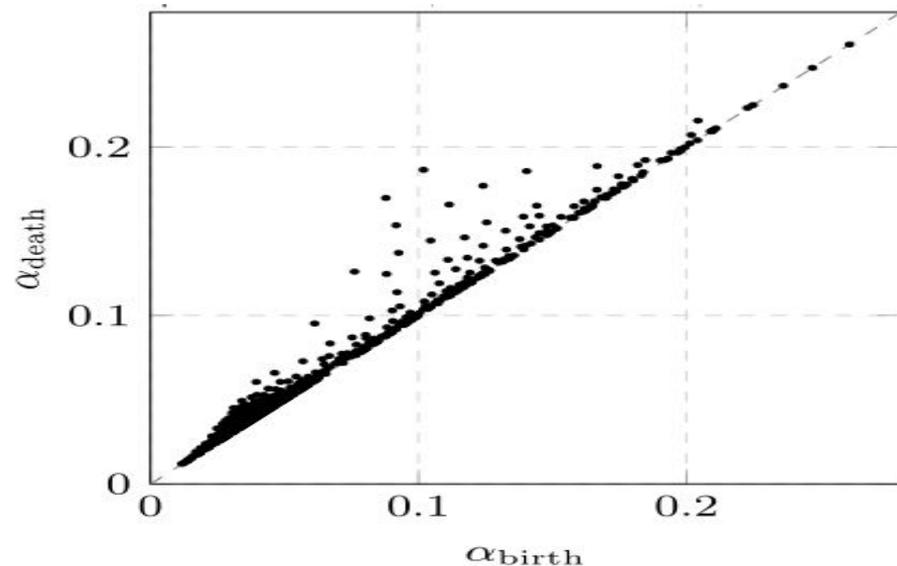
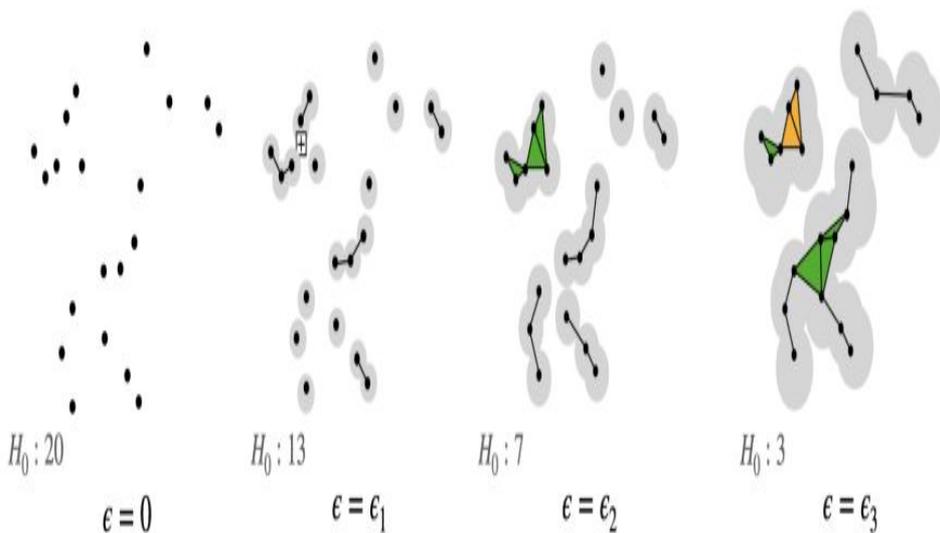


**Cavities**

$$H_2$$

# Persistent Homology

- “Only the most persistent holes survive.”- Shawhin Talebi
- Using these circles and their radii, we can keep track of their “births” and “deaths” to make a **persistence diagram**.



# Persistence Diagrams for Time Series

- For time series, blowing up circles doesn't make sense, instead we think of this as sweeping a horizontal line up the entire signal.

## Challenge 1

Get familiar with the notions of TDA for the analysis of time series.

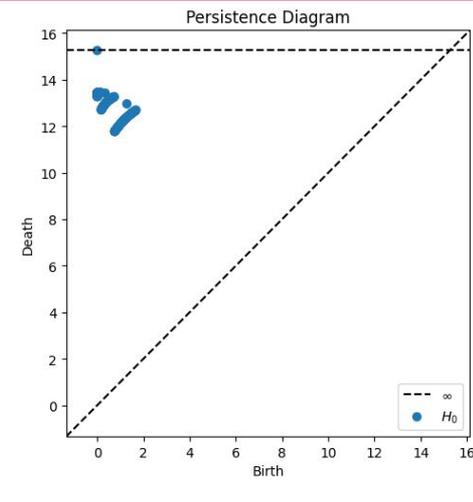
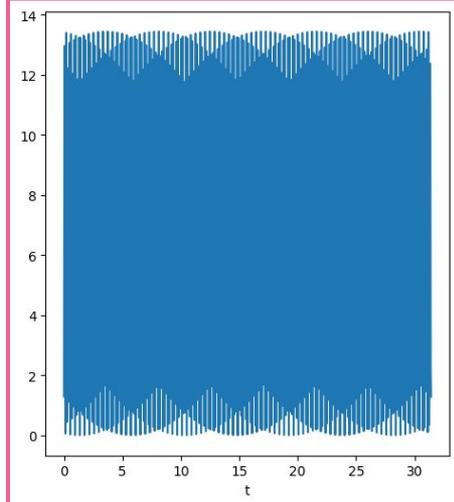
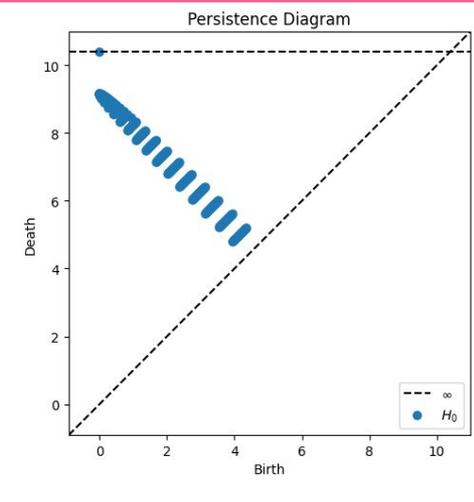
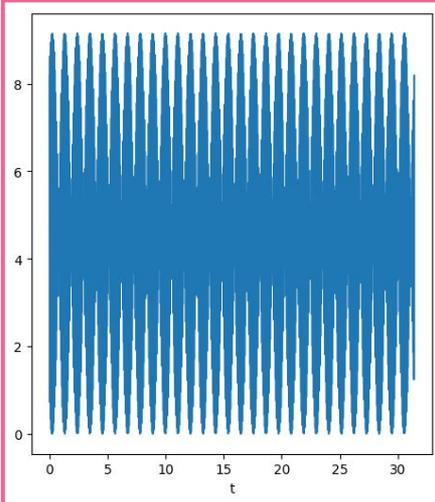
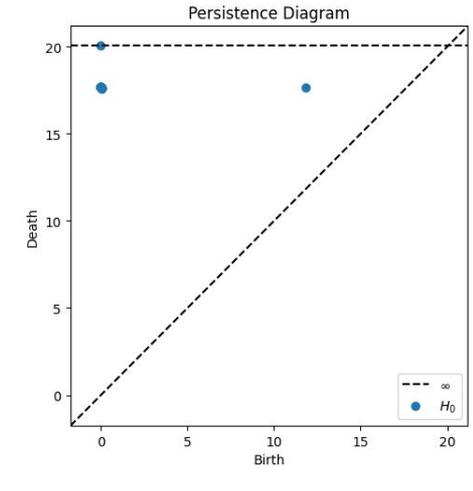
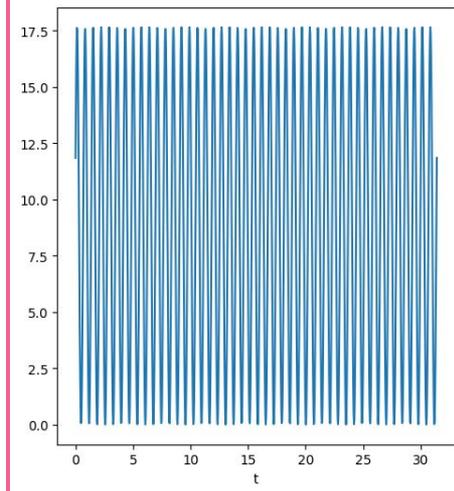
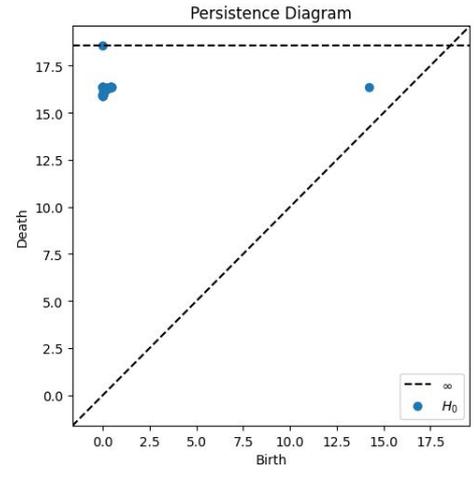
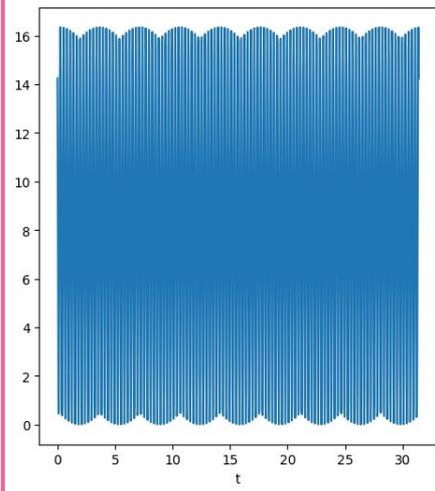
## Challenge 2

Set up code that computes and plots the PD for a time series.

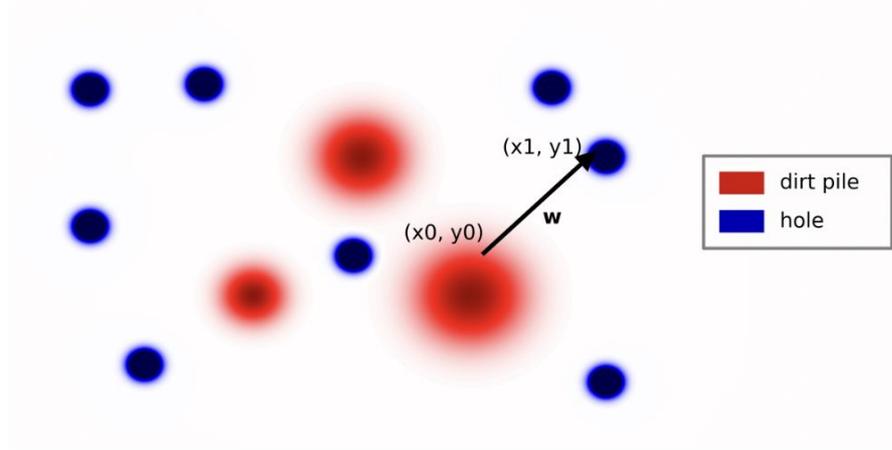
```
import persim  
from ripser import ripser  
from persim import plot_diagrams
```

## Challenge 3

Test the behavior of PD for different time series.



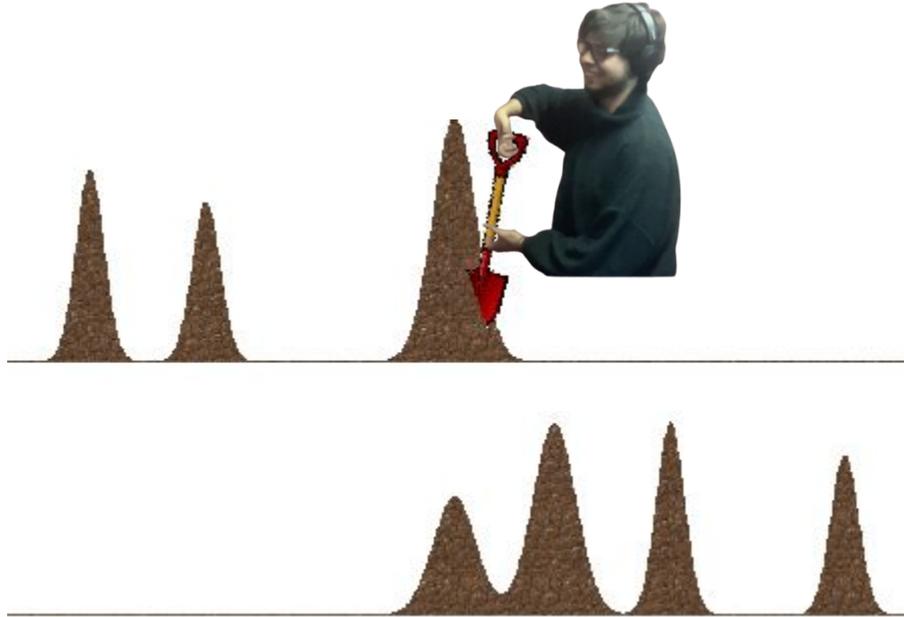
# Wasserstein Distance (Earth-mover's Distance)



**Example transport path.** The arrow schematizes  $w$  units of dirt being transported from location  $(x_0, y_0)$  to  $(x_1, y_1)$ . A complete transport plan specifies transport paths like this over all pairs of locations.

The dirt's image is from [Codewars](#).

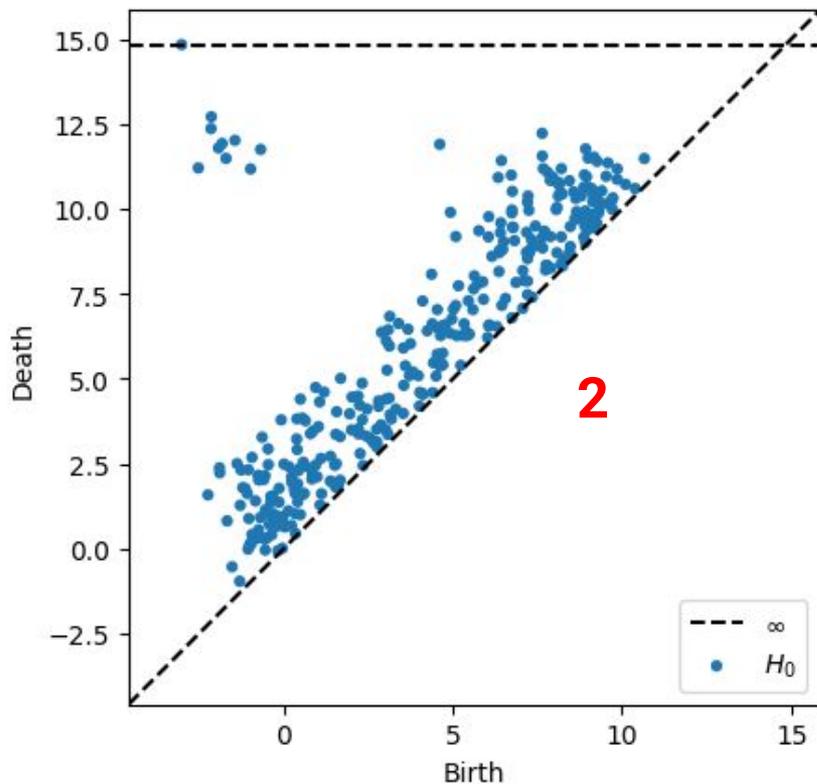
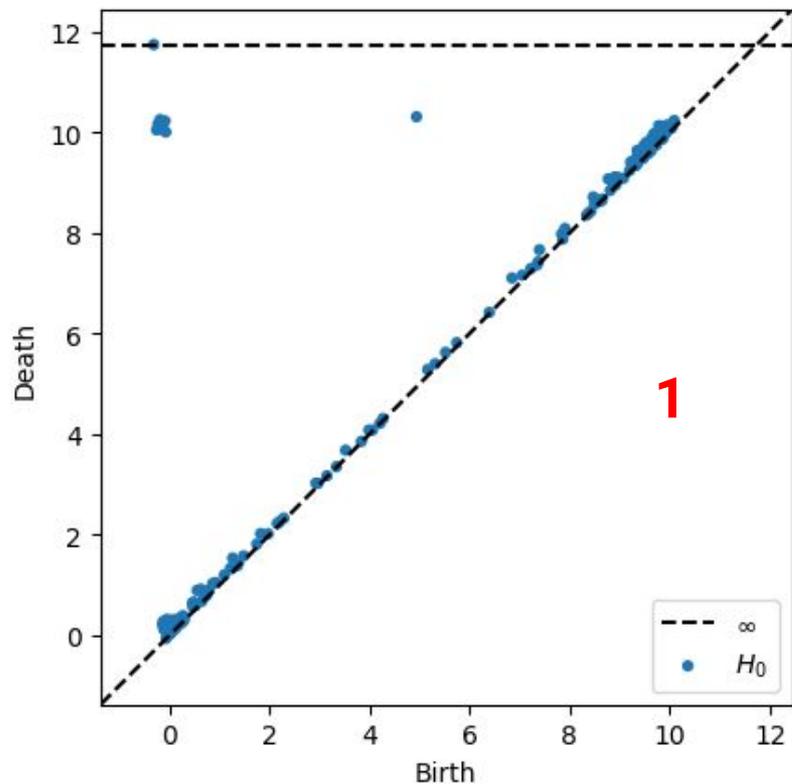
# Wasserstein Distance (Earth-mover's Distance)



Diego making a fool of himself.

# Wasserstein Distance (Earth-mover's Distance)

For example, between these 2 persistence diagrams.



# Distance matrix example

$d(\text{series}_1, \text{series}_1)$

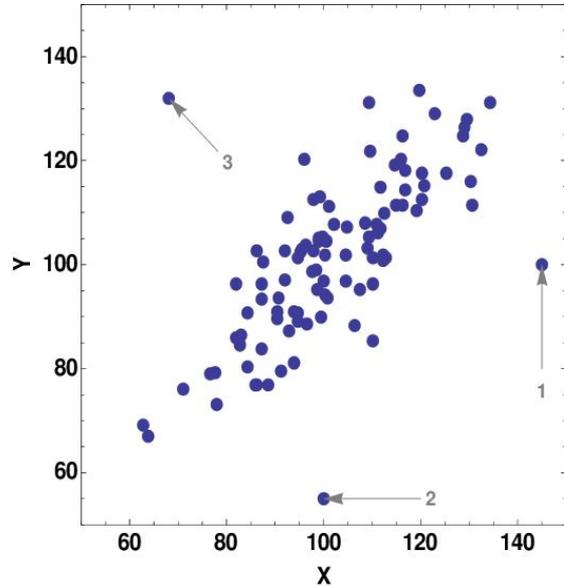
$d(\text{series}_1, \text{series}_2)$

0	1	3	1	4	6
1	0	5	2	2	7
3	5	0	3	2	7
1	2	3	0	6	7
4	2	2	6	0	7
6	7	7	7	7	0

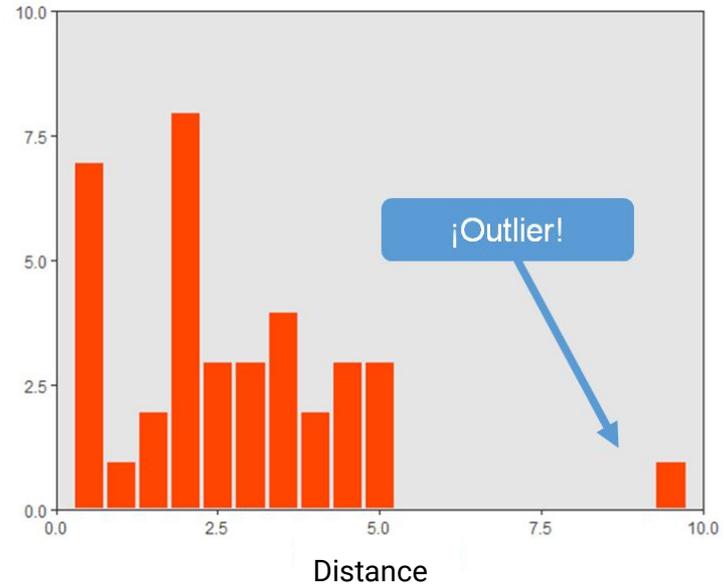
$d(\text{series}_2, \text{series}_1)$

Matrix made with [Matrixcalc](#).

# How to identify an outlier



Outlier diagram from [Denis Cousineau](#).

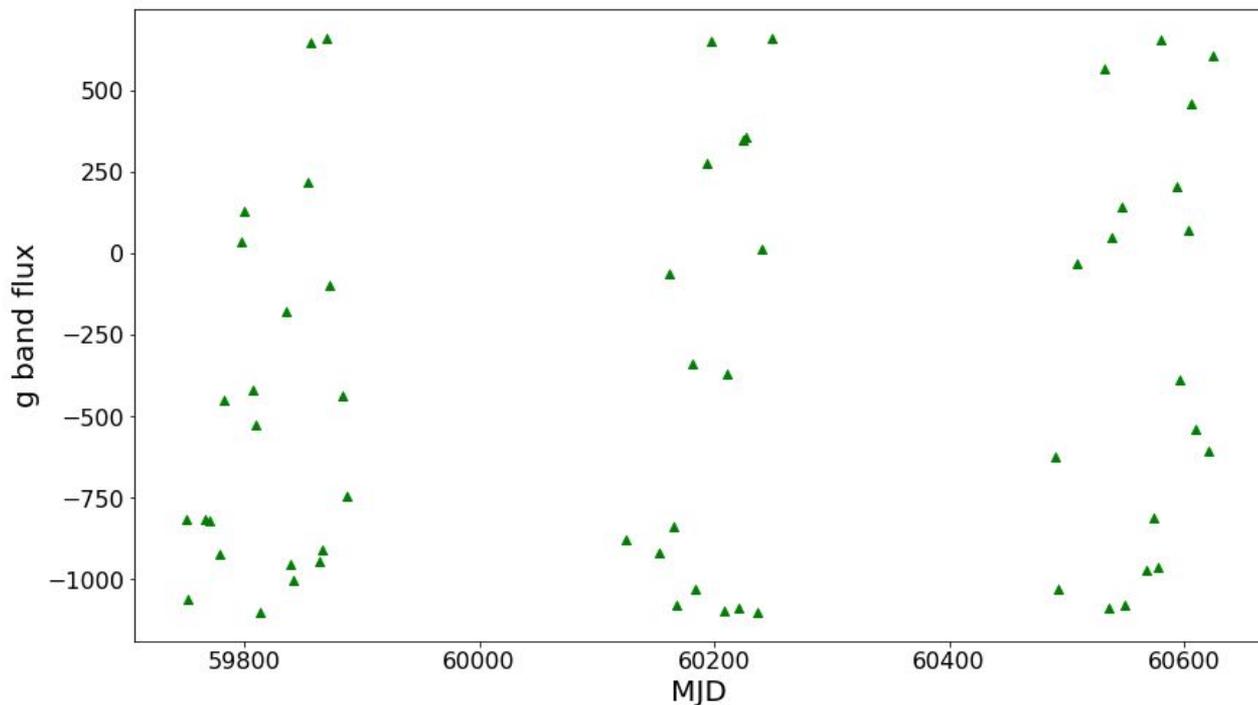


Distance Histogram from [Fernandoblancopsy](#).

# PLAsTiCC

## Photometric LSST Astronomical Time Series Classification Challenge

Simulated lightcurves in *ugrizy* bands, containing 14 classes of astronomical objects

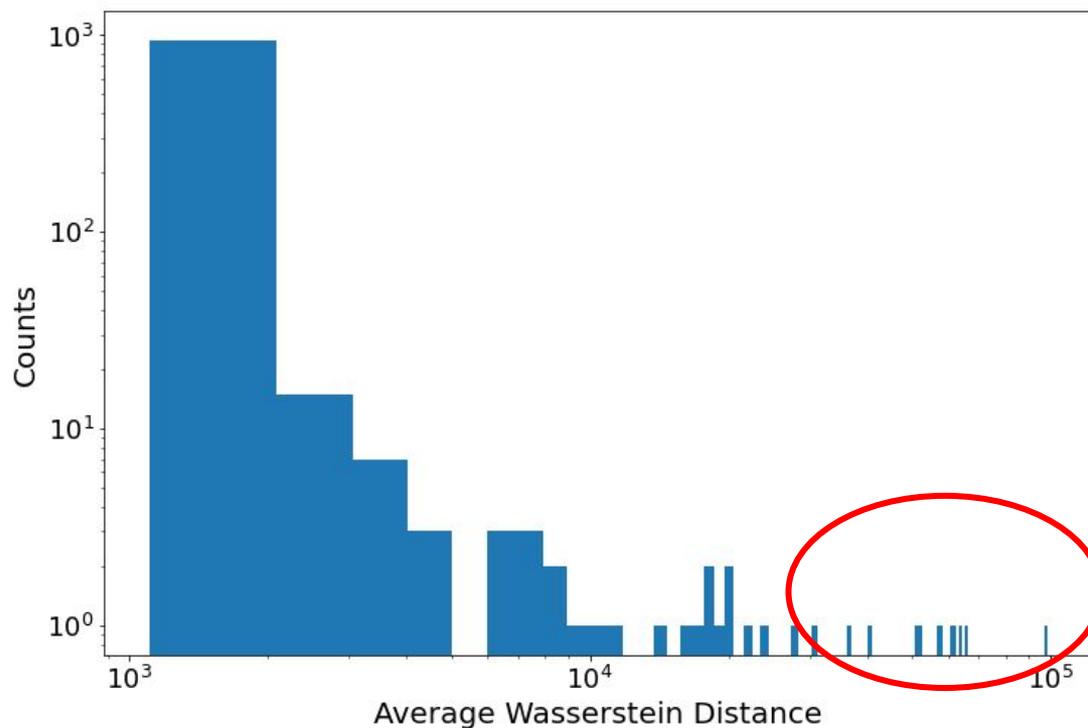


Negative fluxes are due to sky fluctuations

In our analysis, we set all negative fluxes as 0.

# The Distance Matrix of 1000 PLAsTiCC Lightcurves

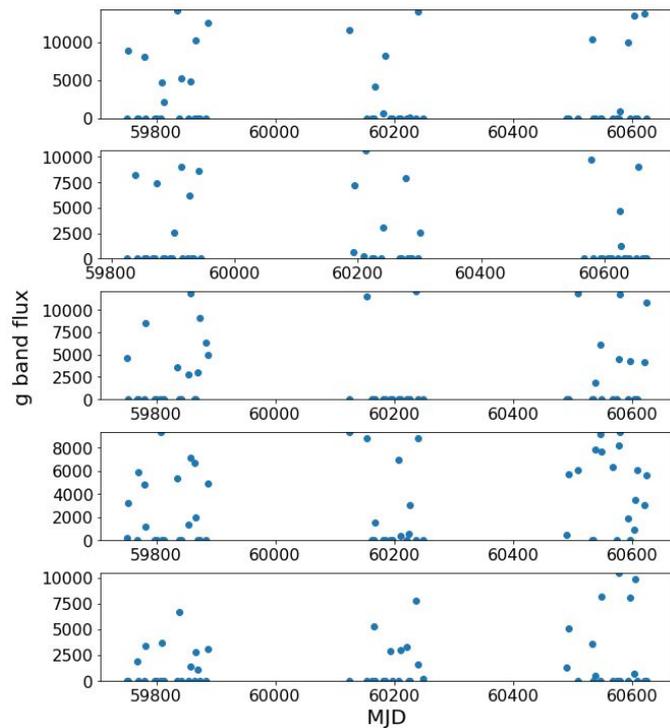
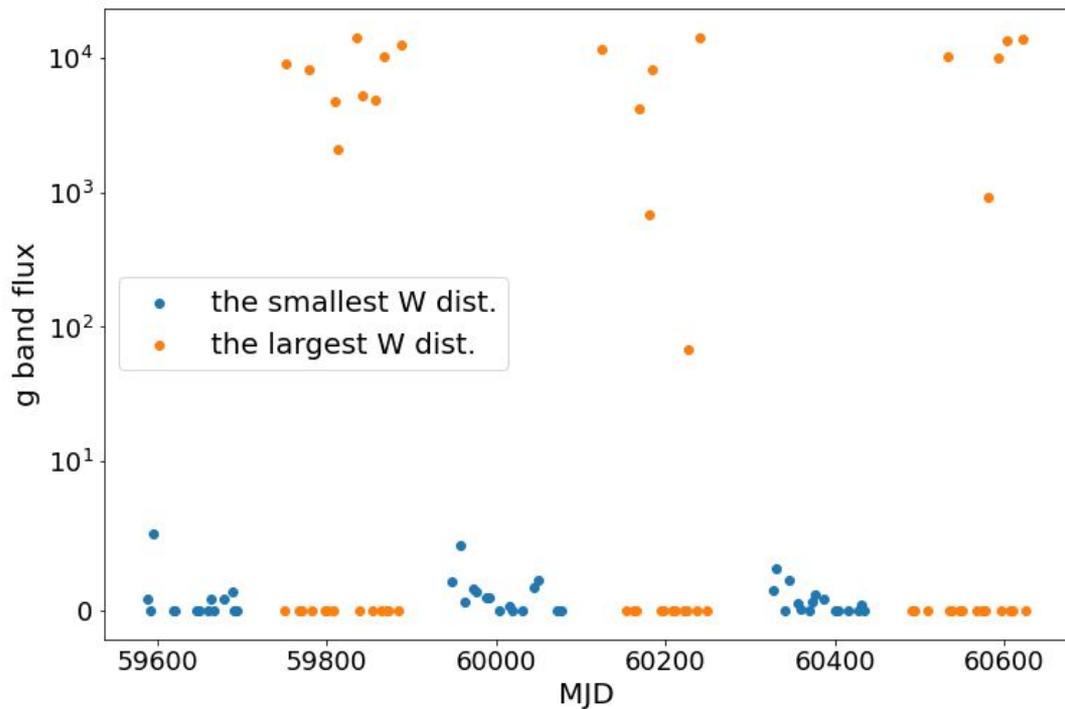
Average Wasserstein distance can differ by 2 order of magnitudes



**Outliers?**

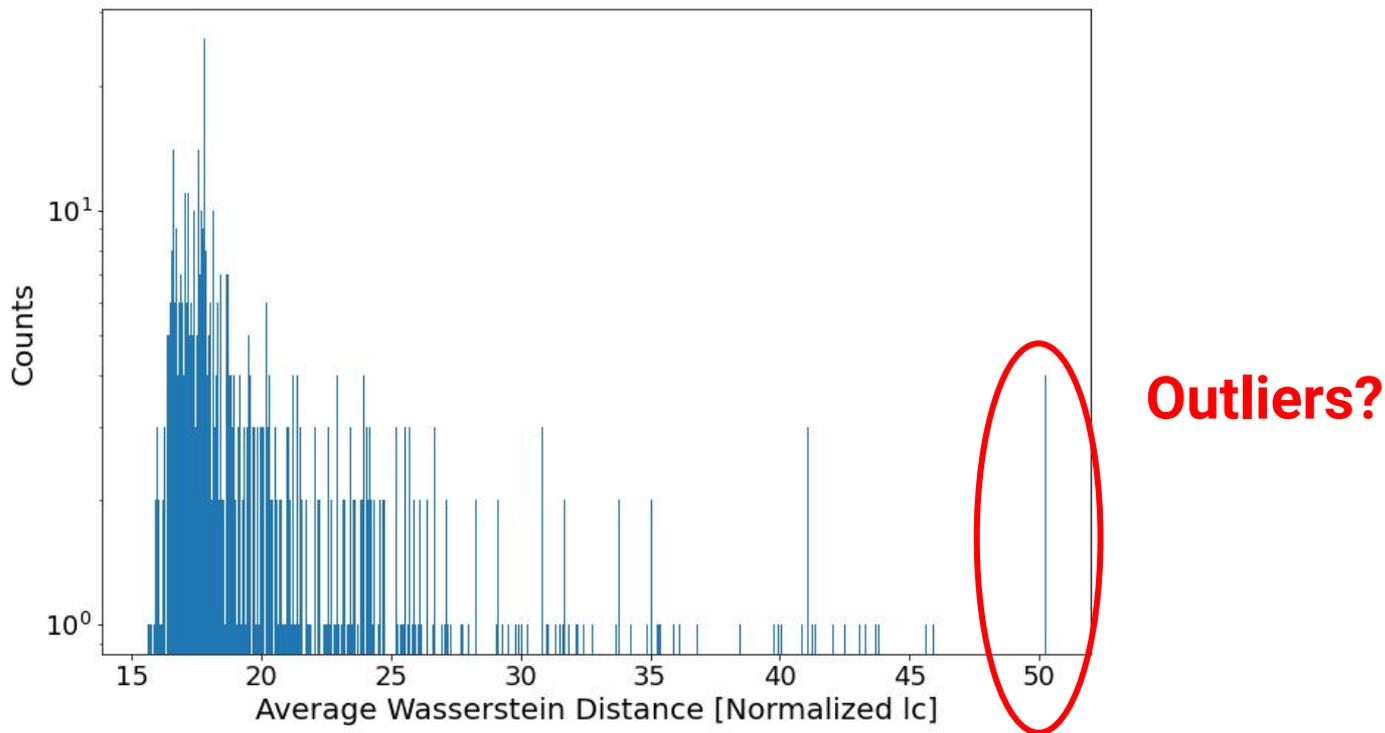
# Top 5 Outliers - Bright lightcurves

Amplitudes of lightcurves can influence Wasserstein distance



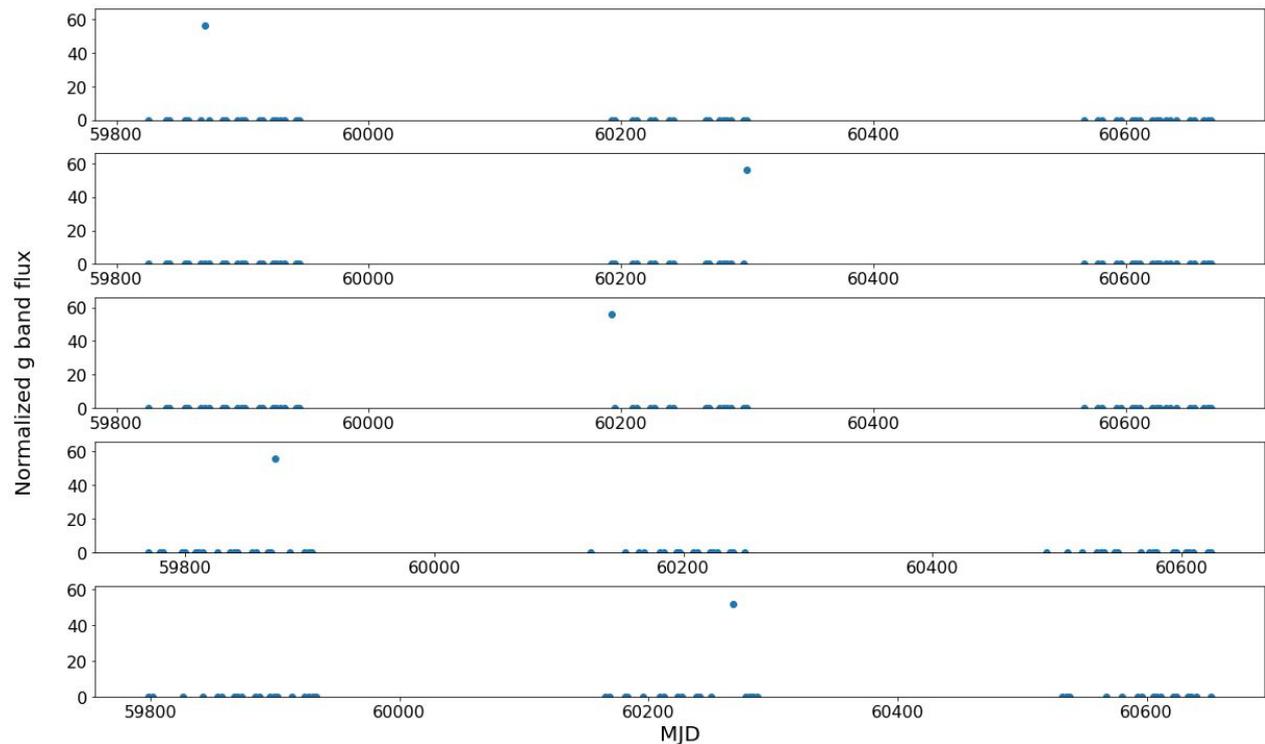
# Normalizing lightcurves by average brightness

Resulting in a much **narrower** distribution than the original distribution



# Top 5 Outliers of Normalized Lightcurves

Transients!



M dwarf flares

M dwarf flares

M dwarf flares

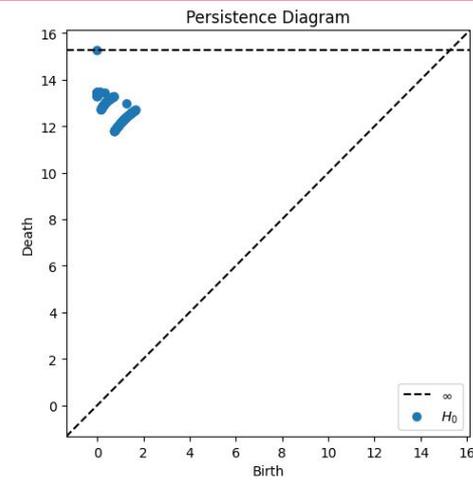
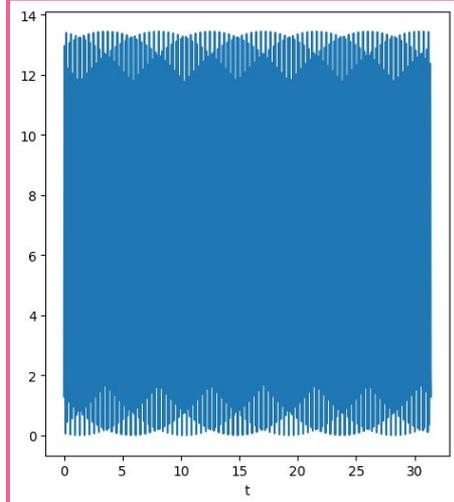
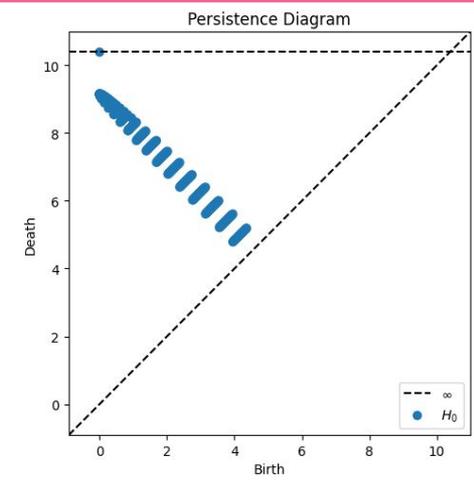
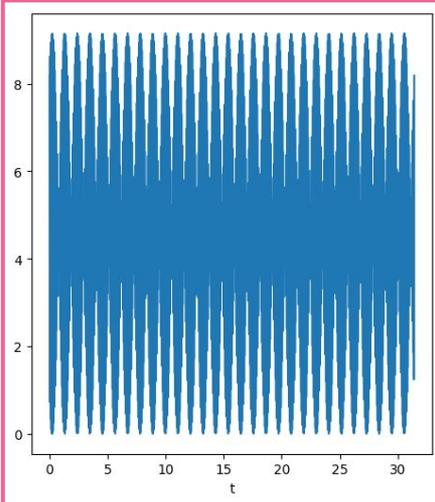
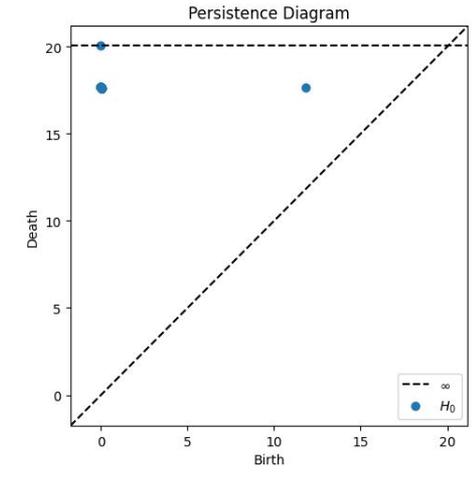
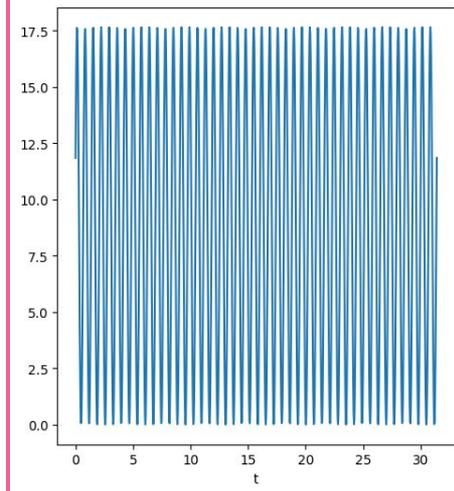
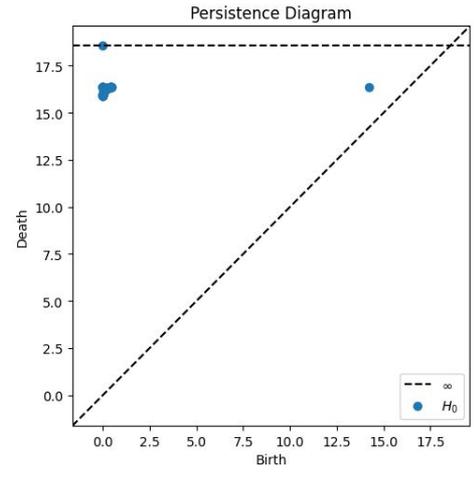
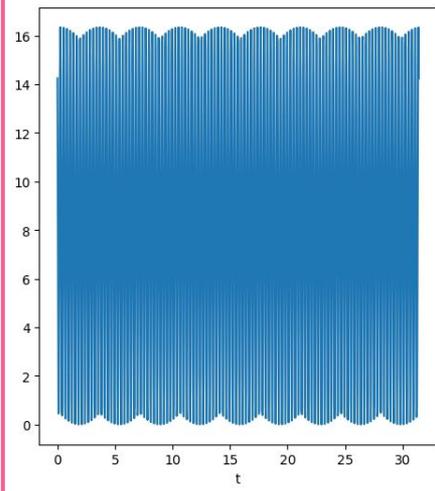
AGN

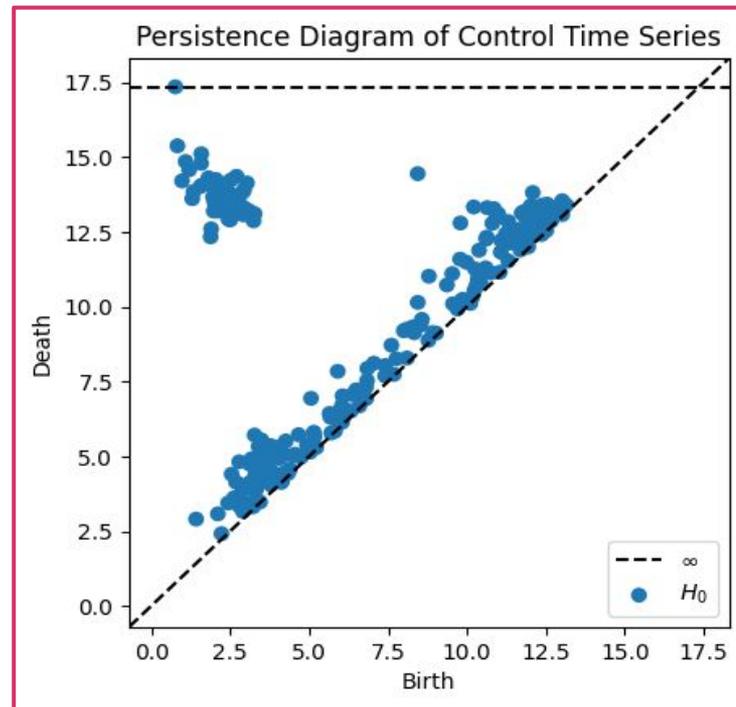
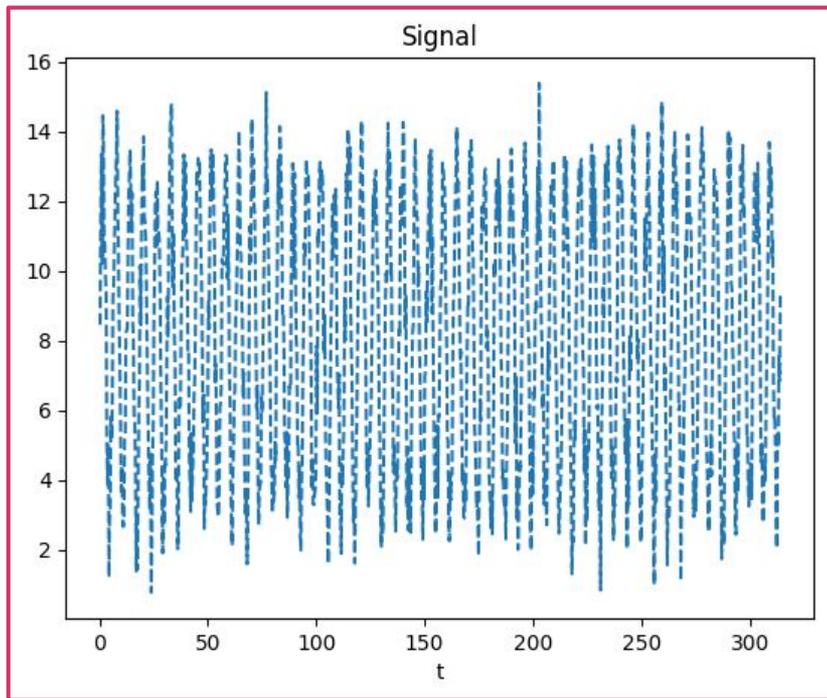
M dwarf flares



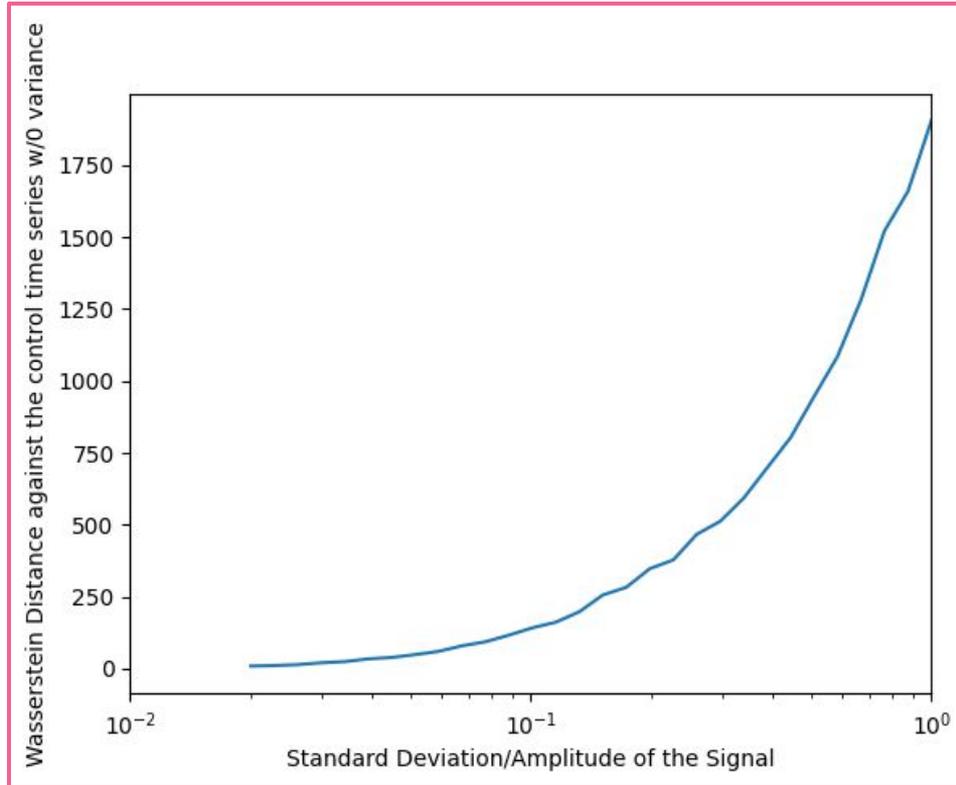
## Remark 1

**What is the effect of noise on the PD of a time series?**





## Wasserstein distance vs. N/S of sinusoidal functions



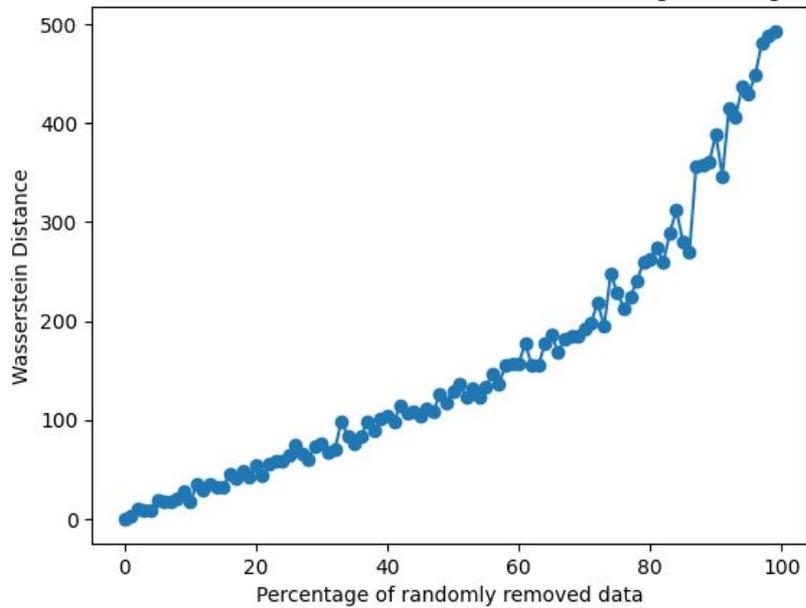
## Remark 1

**What is the effect of noise on the PD of a time series?**

## Remark 2

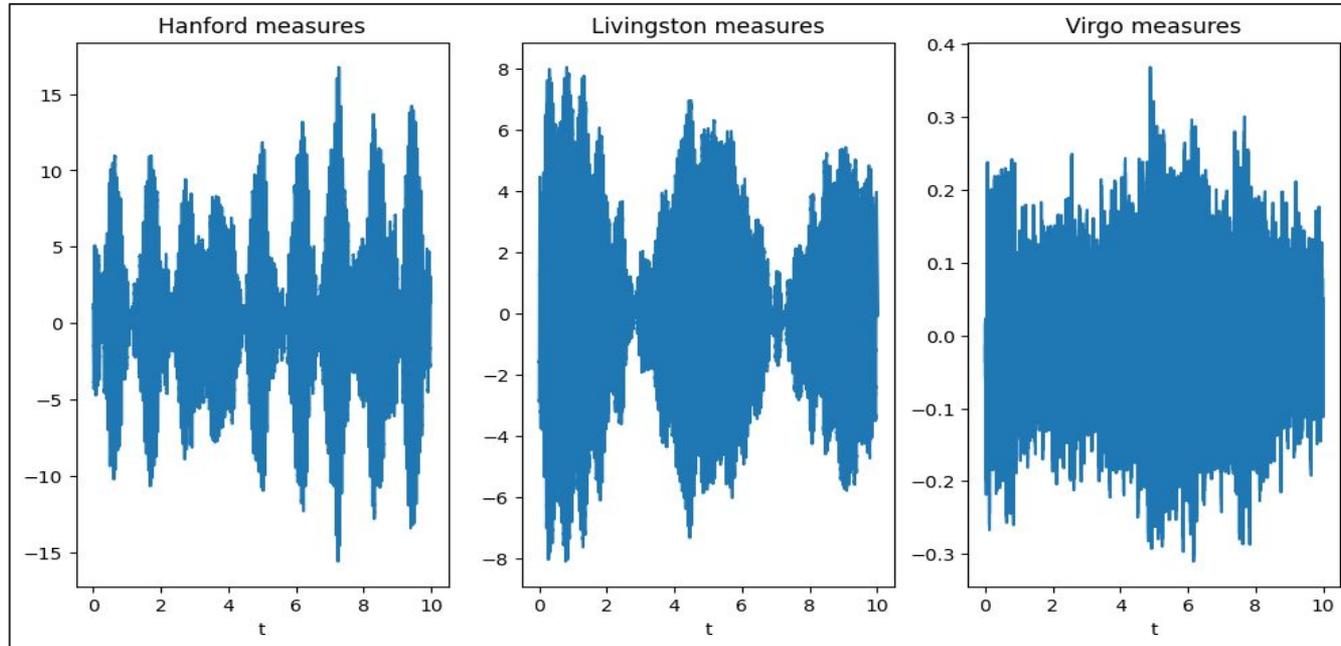
**What happens with the PD of a time series when we do irregular sampling on it?**

Wasserstein Distances for different observational cadence against original time series



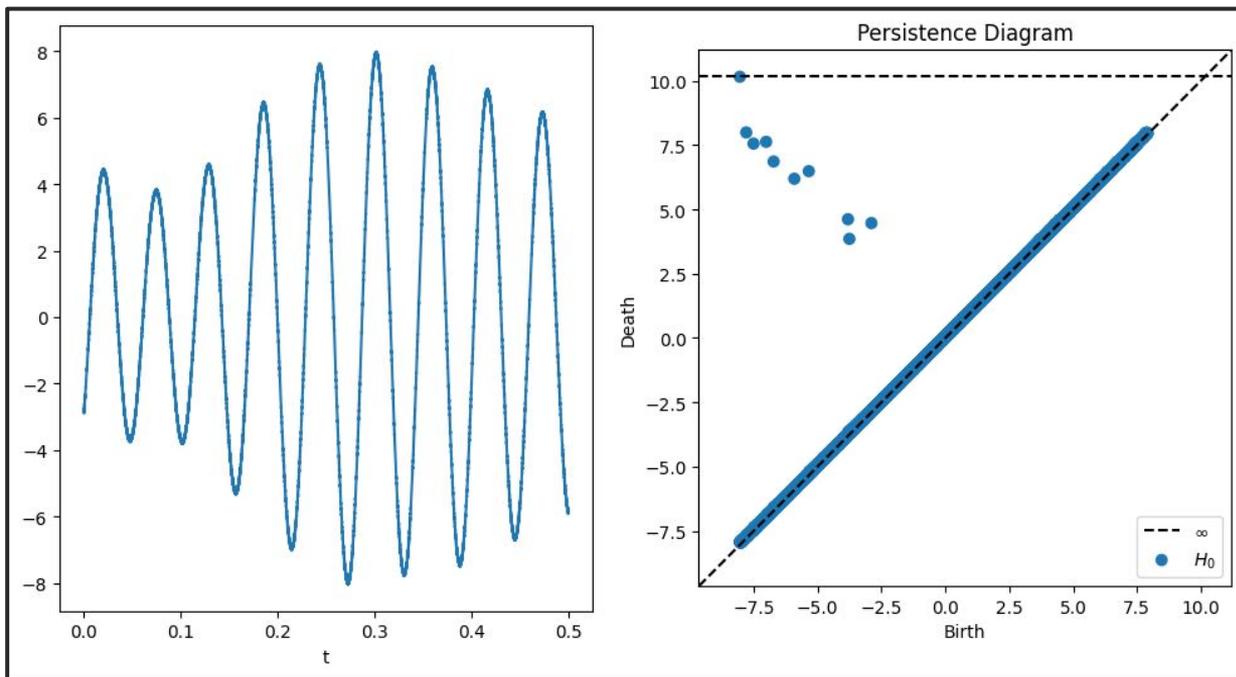
# Laser Interferometer Gravitational-Wave Observatory (LIGO) data

- We fetched the data associated with the *Big Dog Event*, a blind-injection test designed to measure the response of the instrument and the survey team to a potential signal.



# Laser Interferometer Gravitational-Wave Observatory (LIGO) data

- We performed segmentation of the time series in 20 chunks of 8192 points each.
- First chunk from Livingston dataset and its PD.

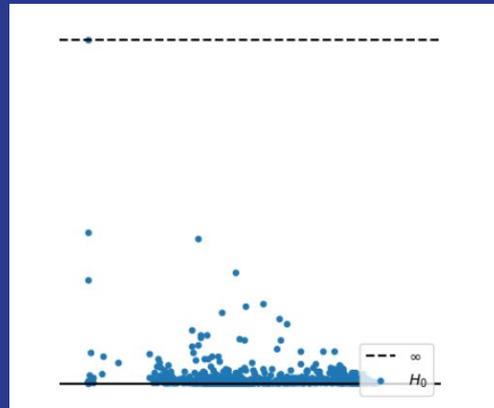


# Conclusions

- ★ **Persistent homology** provides a method of characterizing the *shape* of data.
- ★ The key strength of this approach is extracting robust topological features from data, **insensitive to noise**.
- ★ We were able to **create an outlier detection pipeline using TDA** and **prove the concept that TDA**, i.e. persistent homology, **can be used for outlier detection** in PLAsTiCC simulated time series.
- ★ We further tested the influence of **noise and observational cadence** on the Wasserstein distance using our simulated time series. We found they could **significantly increase** Wasserstein distance in some cases. We will investigate those factors before applying this method to real astronomical datasets.



Thank you! A special thanks to our TA Edgar Ortiz and our advisor Matthew Graham!



LA SERENA SCHOOL  
FOR DATA SCIENCE 2022  
Applied Tools for  
Data-driven Sciences  
August 1–12, 2022

# Links to references:

- ★ <https://towardsdatascience.com/persistent-homology-with-examples-1974d4b9c3d0>
- ★ <https://medium.datadriveninvestor.com/persistent-homology-f22789d753c4>
- ★ <https://ripser.scikit-tda.org/en/latest/notebooks/Lower%20Star%20Time%20Series.html>
- ★ [https://www.astroml.org/user\\_guide/datasets.html#time-domain-data](https://www.astroml.org/user_guide/datasets.html#time-domain-data)
- ★ <https://plasticc.org/>
- ★ <https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full>
- ★ <https://en.wikipedia.org/wiki/Topology>