Outlier detection via Topological Data Analysis (TDA)

Diego Jaramillo
Agustina Blanco
Nathan Rutherford
Xiangyu Jin

T.A.: Edgar Ortiz
Advisor: Matthew Graham

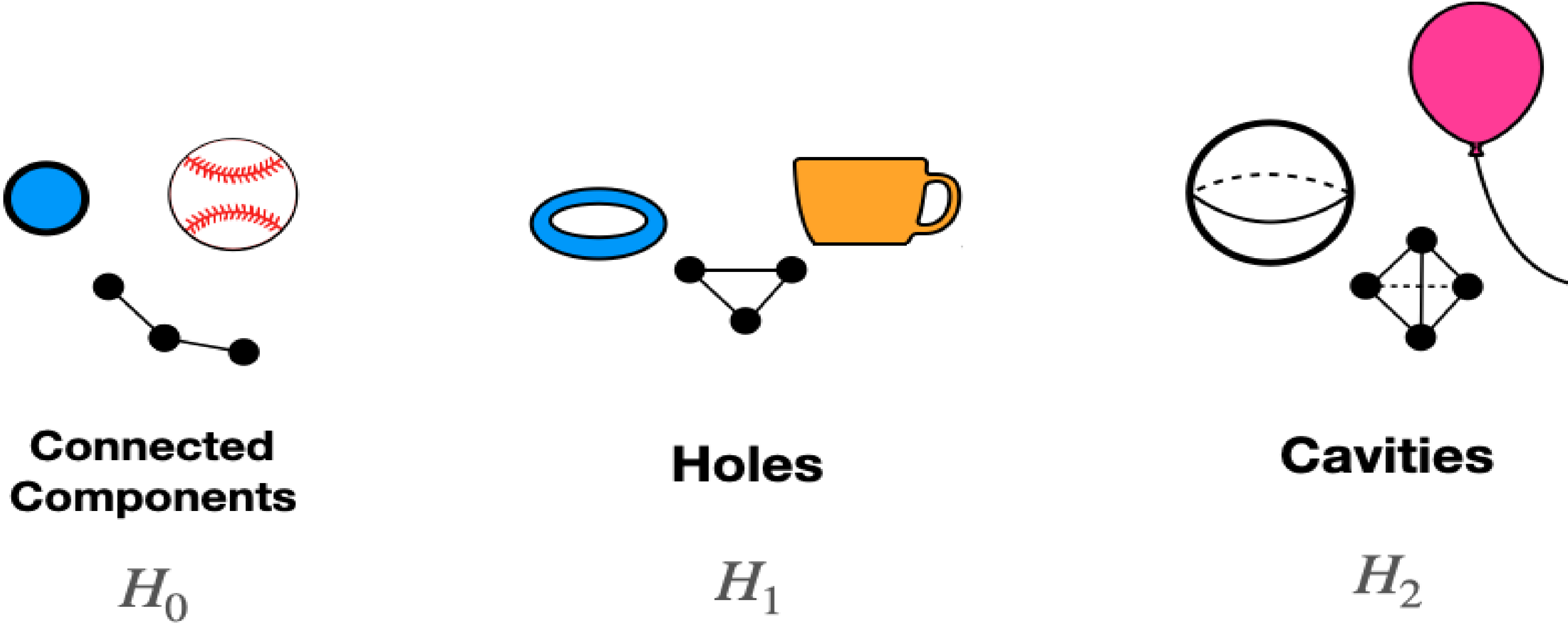
What is TDA?

- 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**

Crane, Keenan (2016) *Keenan's 3D Model Repository*

Segerman,Henry (2015)

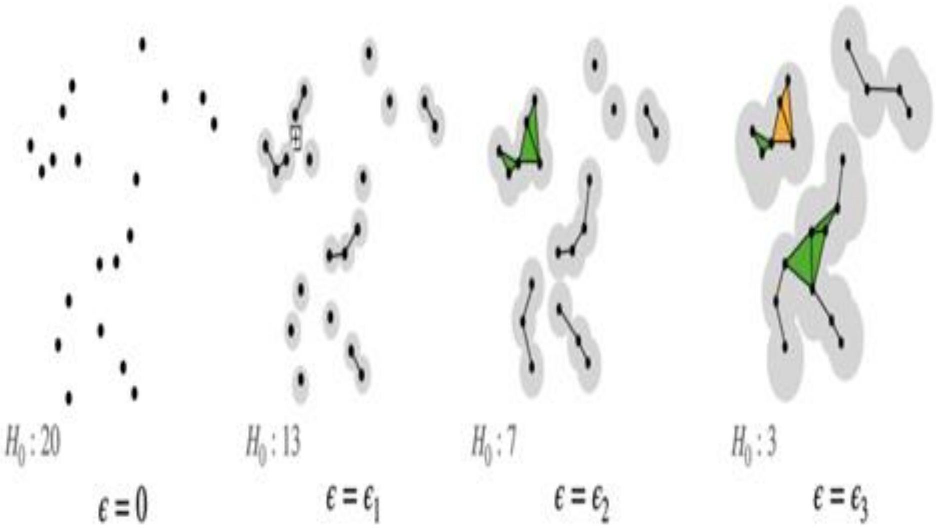
What is TDA? (Cont'd)

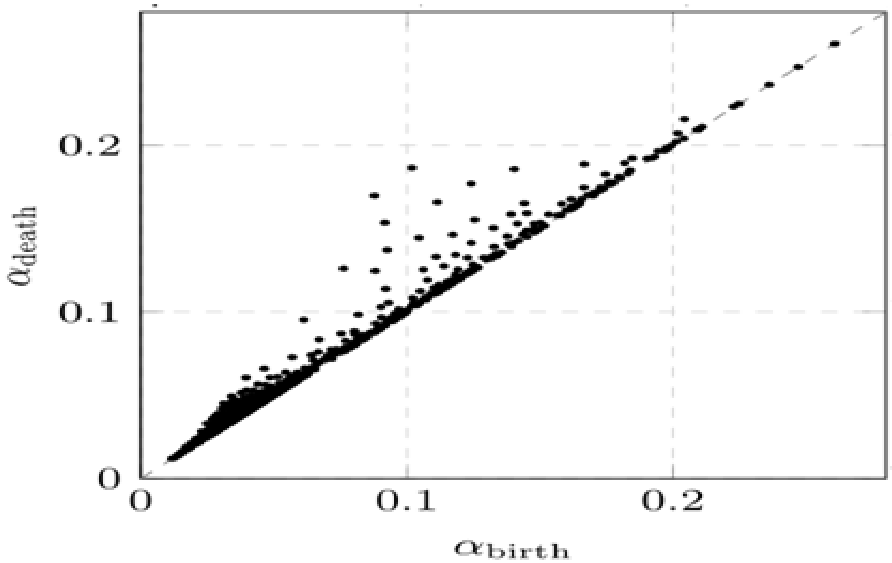


Talebi, Shawhin (2022) *Persistent Homology*

Persistent Homology

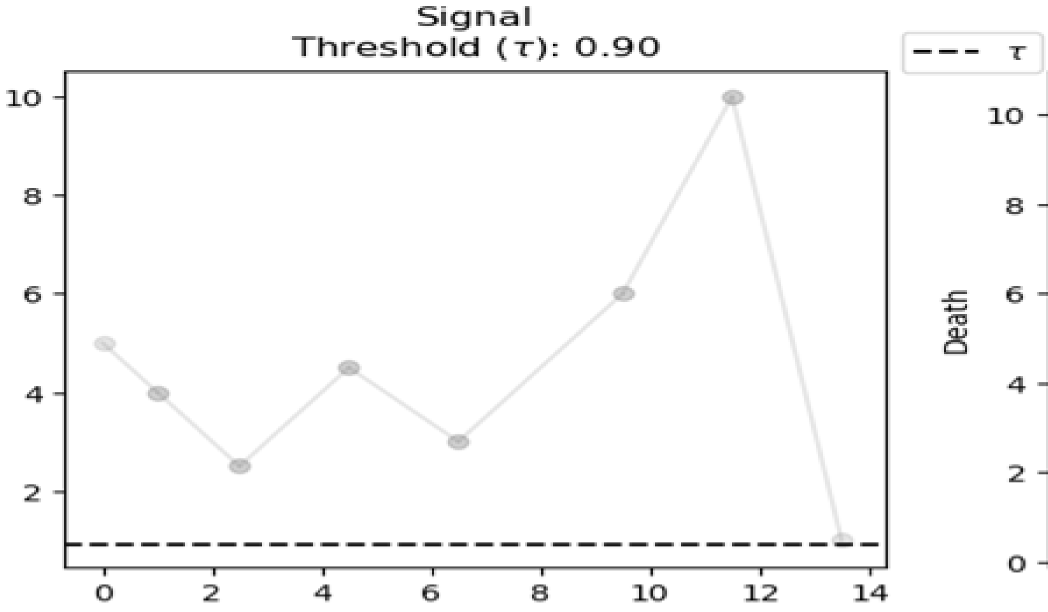
- *“Only the most persistent holes survive.”*- Shawhin Talebi
- Using these circles and their radii, we can keep track of their ``births” and “deaths” to make a **persistence diagram.**

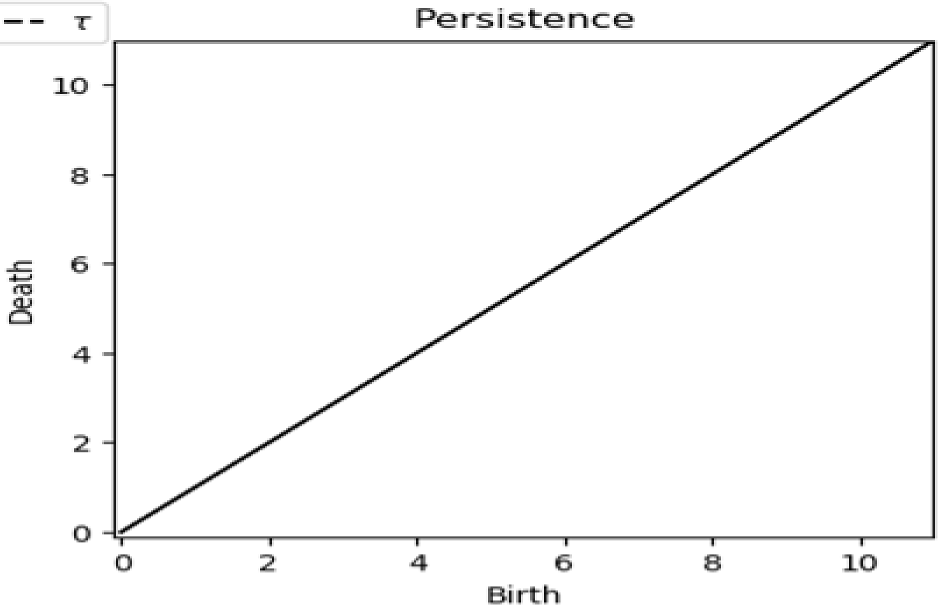




Persistence Diagrams for Time Series

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





Koplik, Gary (2019) *Persistent Homology: A Non-Mathy Introduction with Examples*

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.

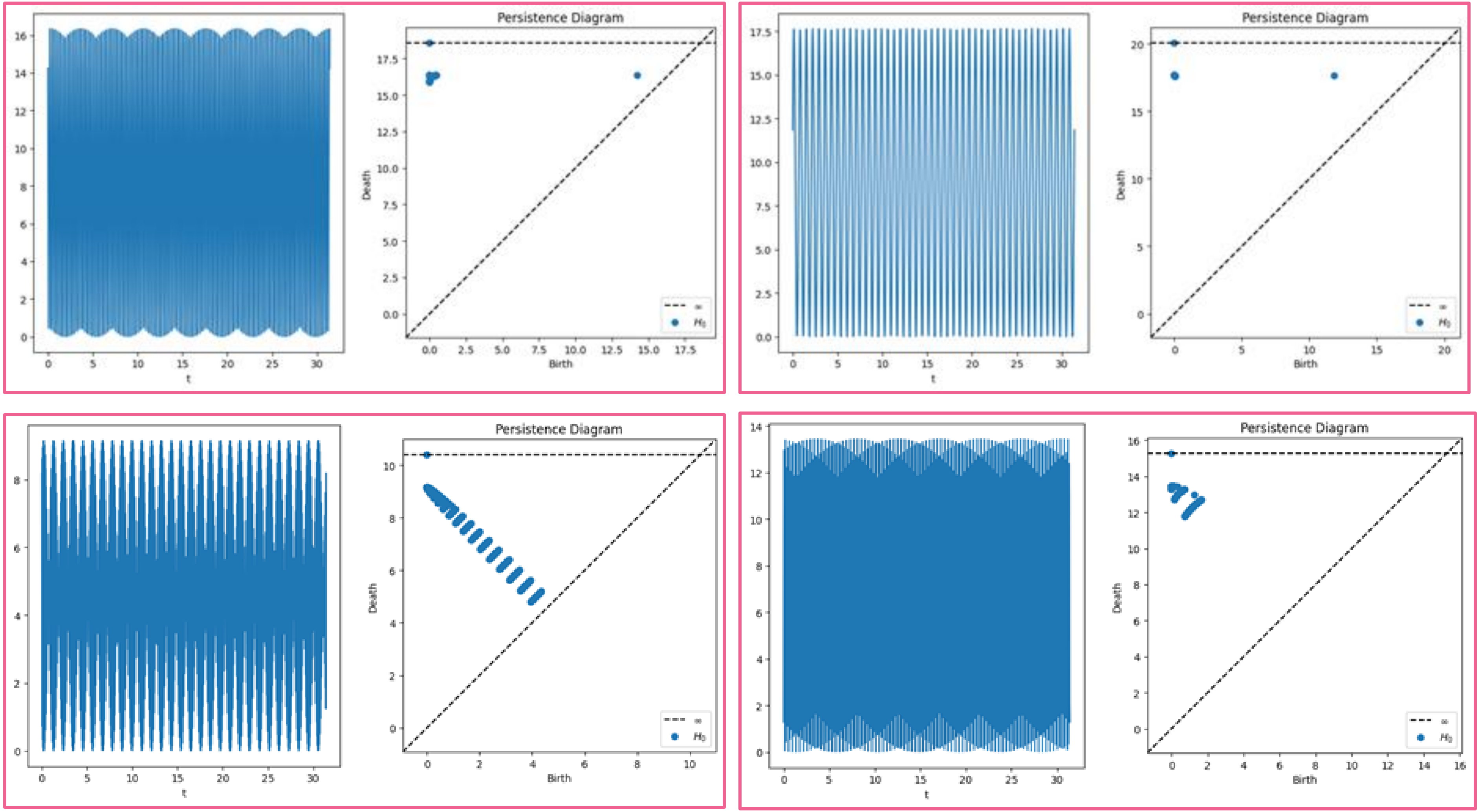
```
import persim

from ripser import ripser

from persim import plot_diagrams
```

Challenge 3

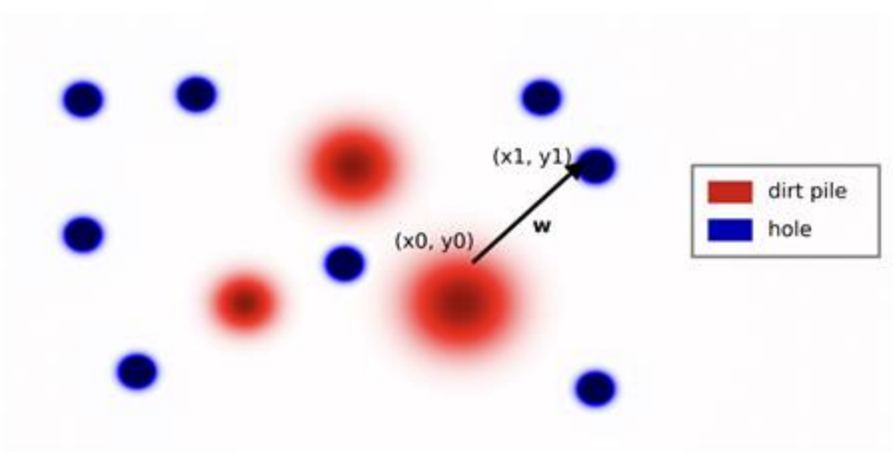
Test the behavior of PD for different time series.



Persistence Diagram
t
Death
Birth
∞
H_0
Persistence Diagram
t
Death
Birth
∞
H_0
Persistence Diagram
t
Death
Birth
∞
H_0
Persistence Diagram
t
Death
Birth
∞
H_0

Williams, A 2020

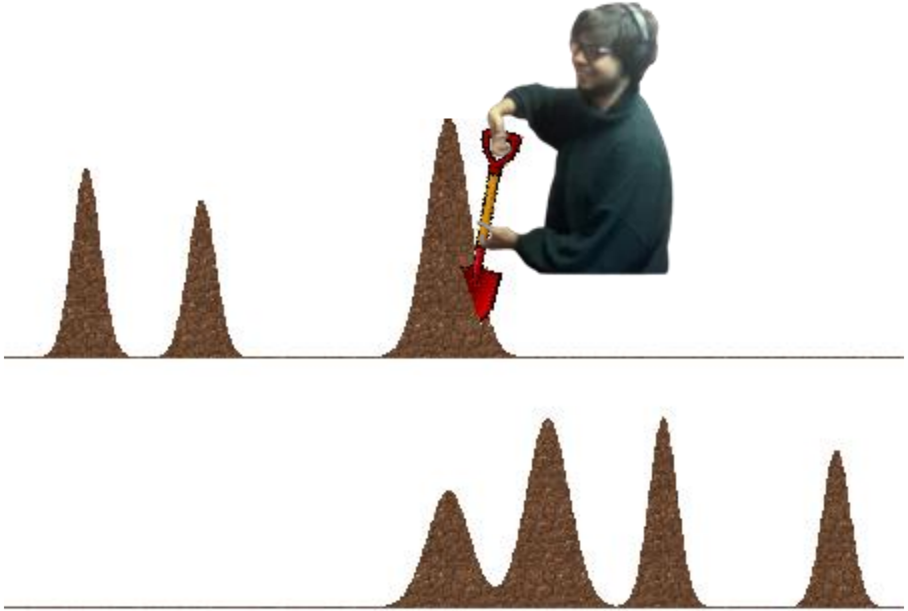
Wasserstein Distance (Earth-mover's Distance)



Example transport path. The arrow schematizes w units of dirt being transported from location (x0, y0) to (x1, y1). A complete transport plan specifies transport paths like this over all pairs of locations.

The dirt's image is from Codewars.

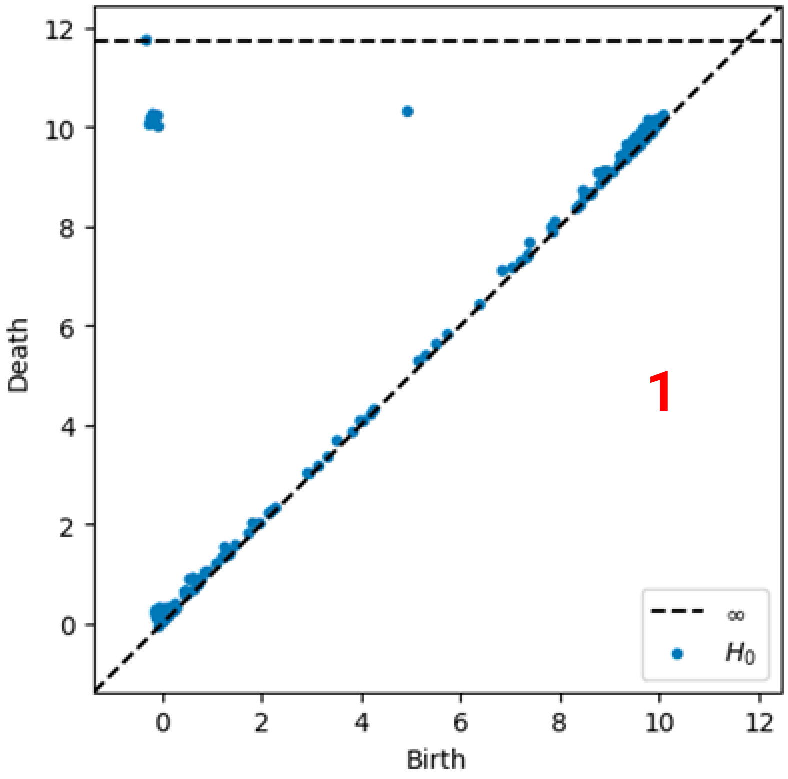
Wasserstein Distance (Earth-mover's Distance)

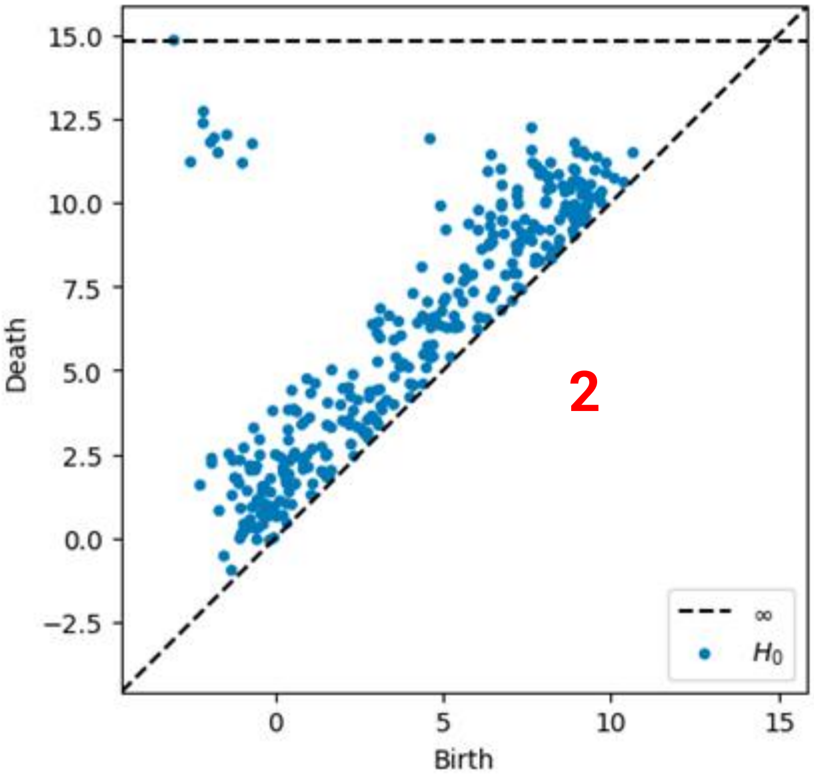


Diego making a fool of himself.

Wasserstein Distance (Earth-mover's Distance)

For example, between these 2 persistence diagrams.





Distance matrix example

d(series_1,series_2)

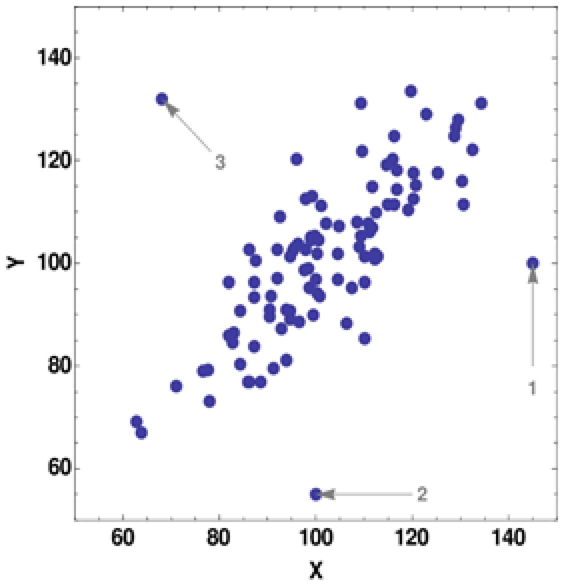
d(series_1, series_1)

d(series_2,series_1)

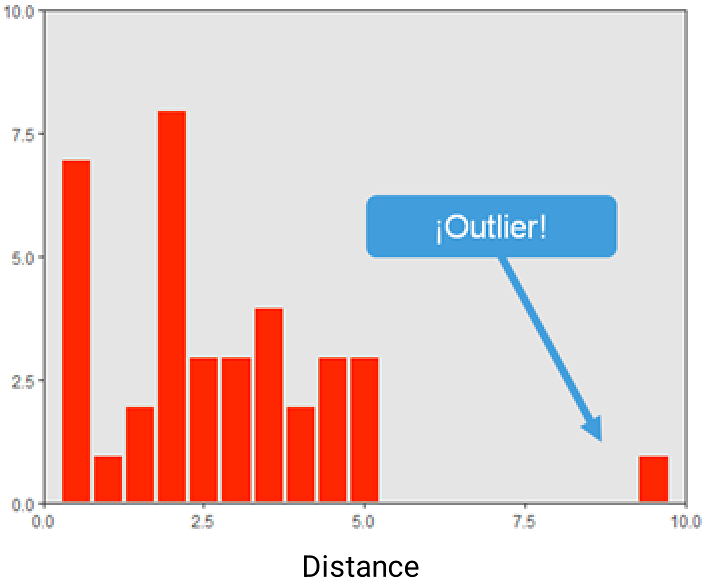
$$
\begin{pmatrix}
0 & 1 & 3 & 1 & 4 & 6 \\
1 & 0 & 5 & 2 & 2 & 7 \\
3 & 5 & 0 & 3 & 2 & 7 \\
1 & 2 & 3 & 0 & 6 & 7 \\
4 & 2 & 2 & 6 & 0 & 7 \\
6 & 7 & 7 & 7 & 7 & 0
\end{pmatrix}
$$

Matrix made with Matrixcalc .

How to identify an outlier



Outlier diagram from Denis Cousineau.

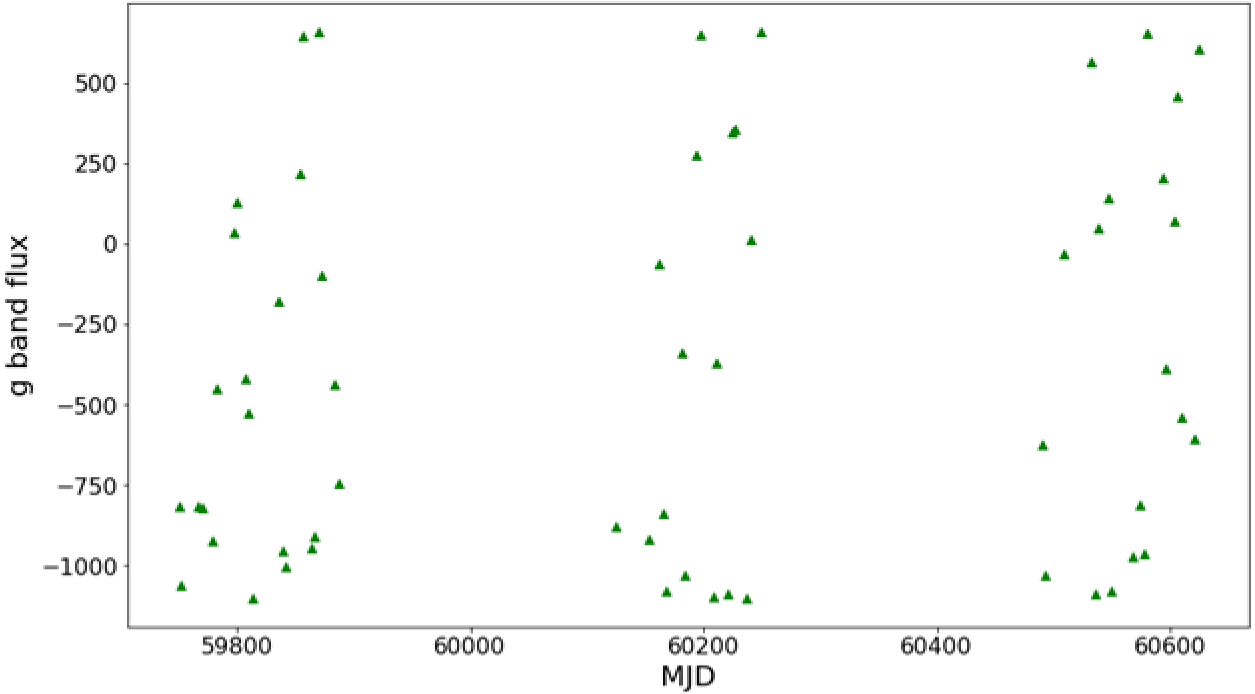


Distance Histogram from Fernandoblancopsy.

PLAsTiCC

Photometric LSST Astronomical Time Series Classification Challenge

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

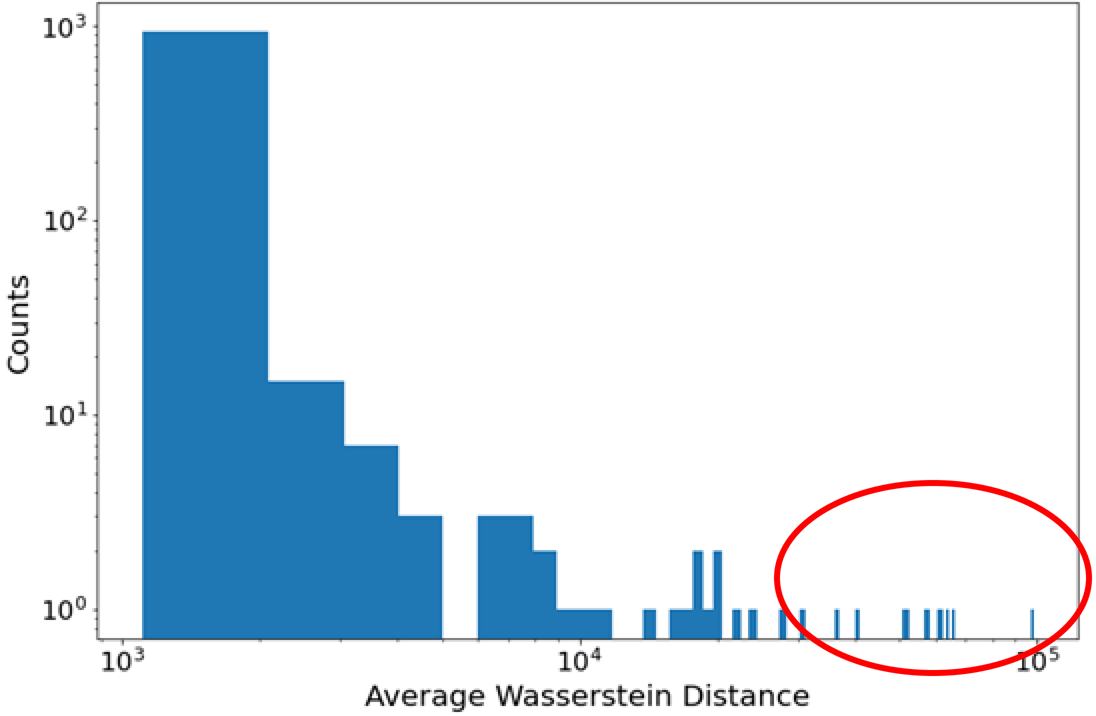


Negative fluxes are due to sky fluctuations

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

The Distance Matrix of 1000 PLAsTiCC Lightcurves

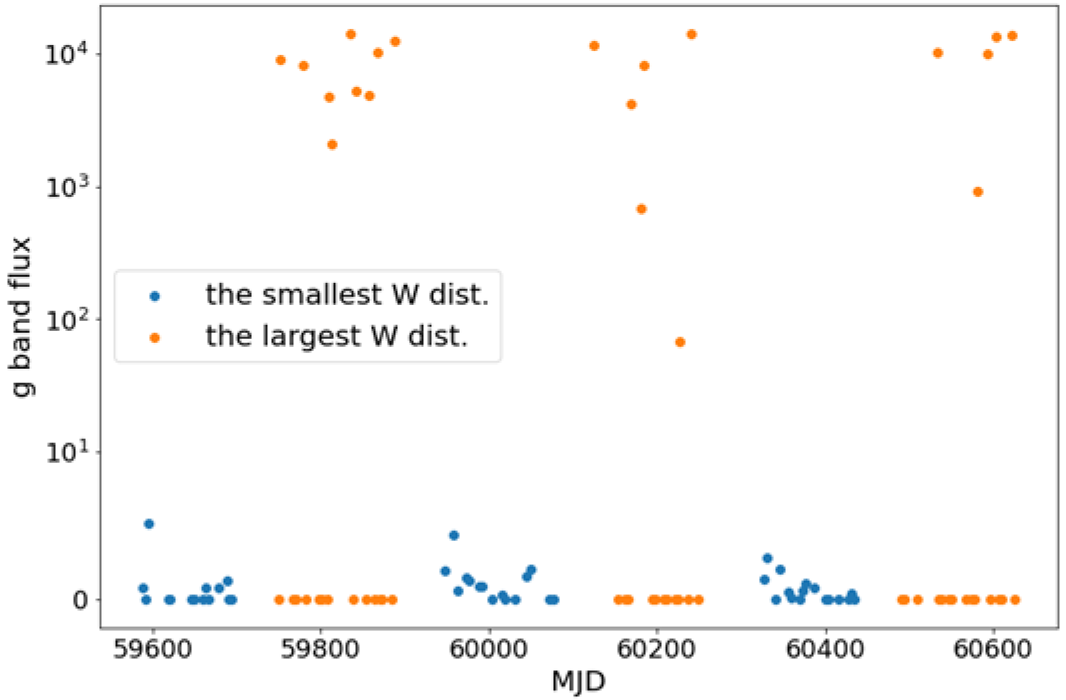
Average Wasserstein distance can differ by 2 order of magnitudes

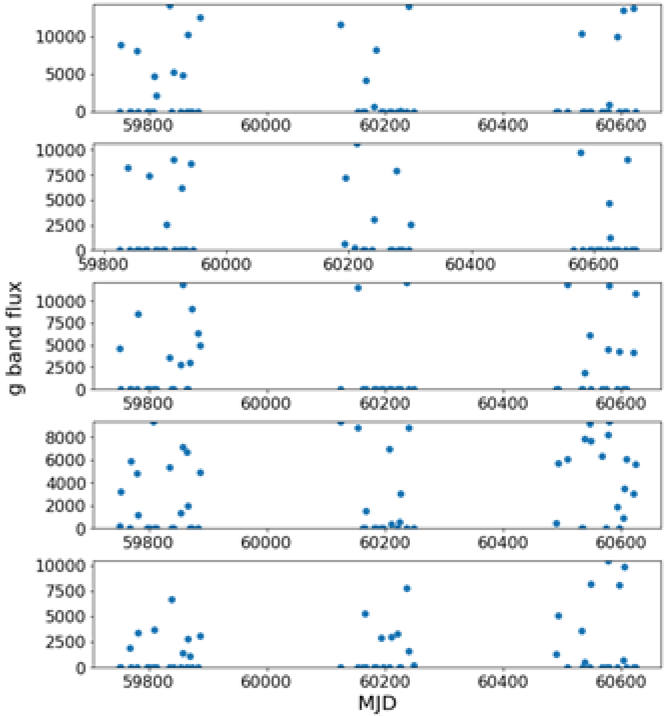


Outliers?

Top 5 Outliers - Bright lightcurves

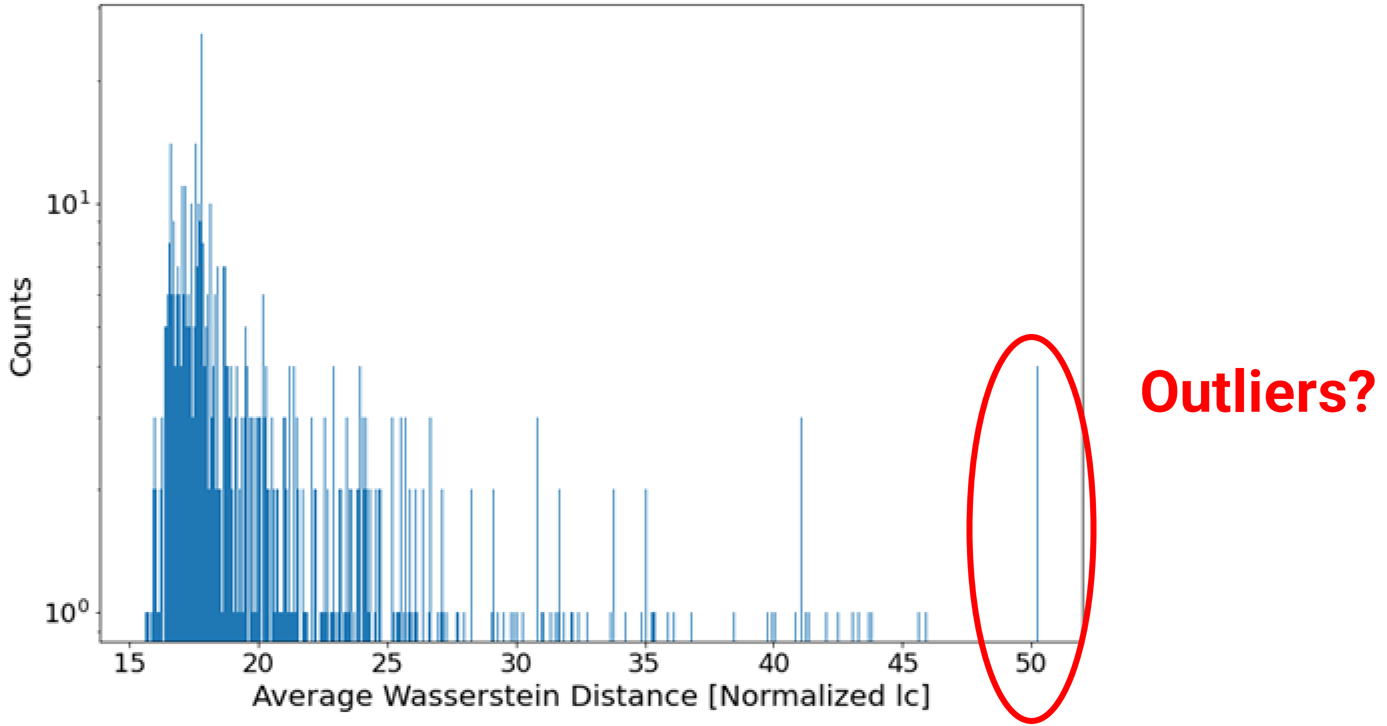
Amplitudes of lightcurves can influence Wasserstein distance





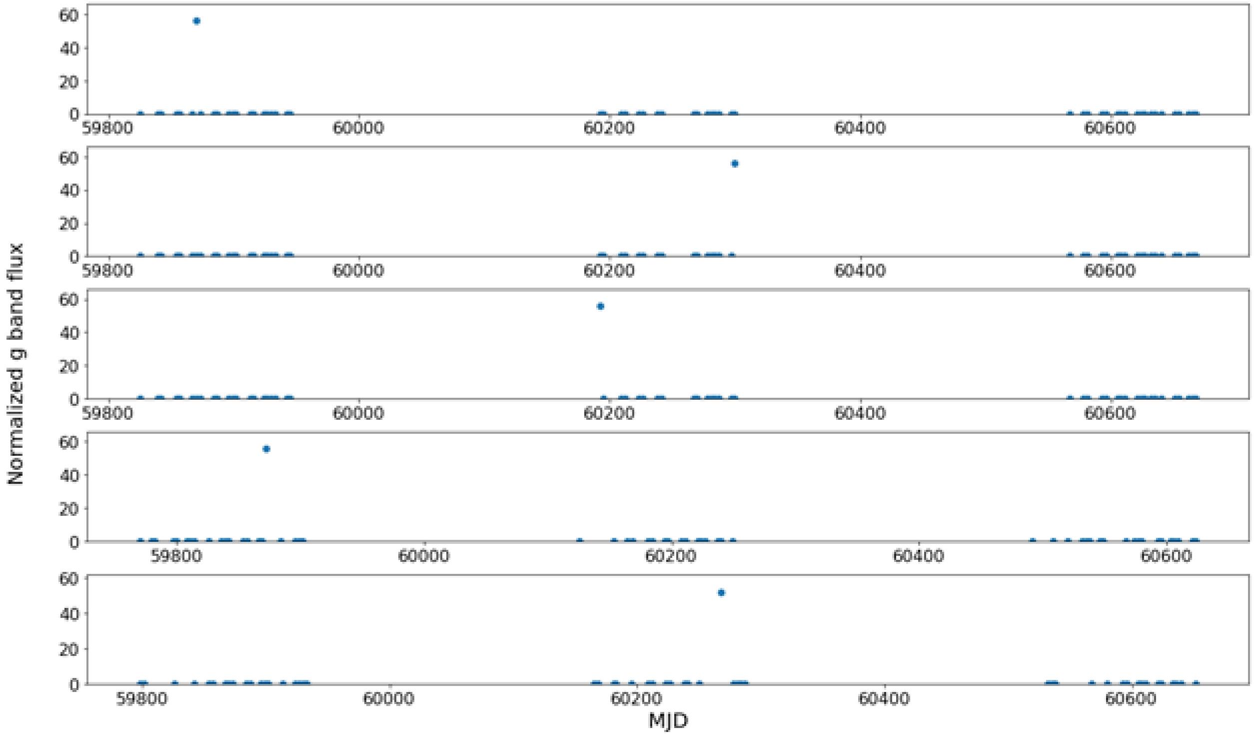
Normalizing lightcurves by average brightness

Resulting in a much narrower distribution than the original distribution



Top 5 Outliers of Normalized Lightcurves

Transients!



M dwarf flares

M dwarf flares

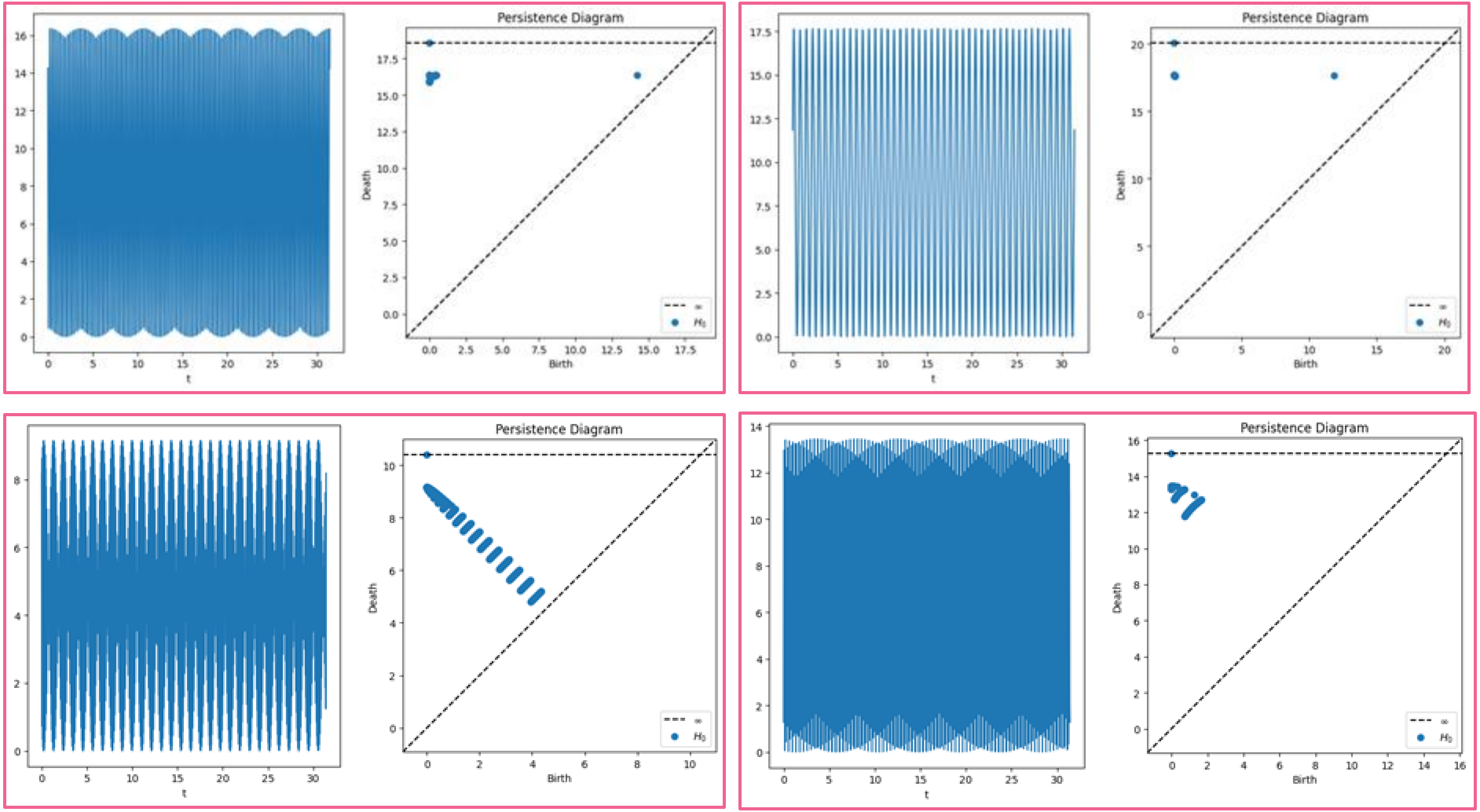
M dwarf flares

AGN

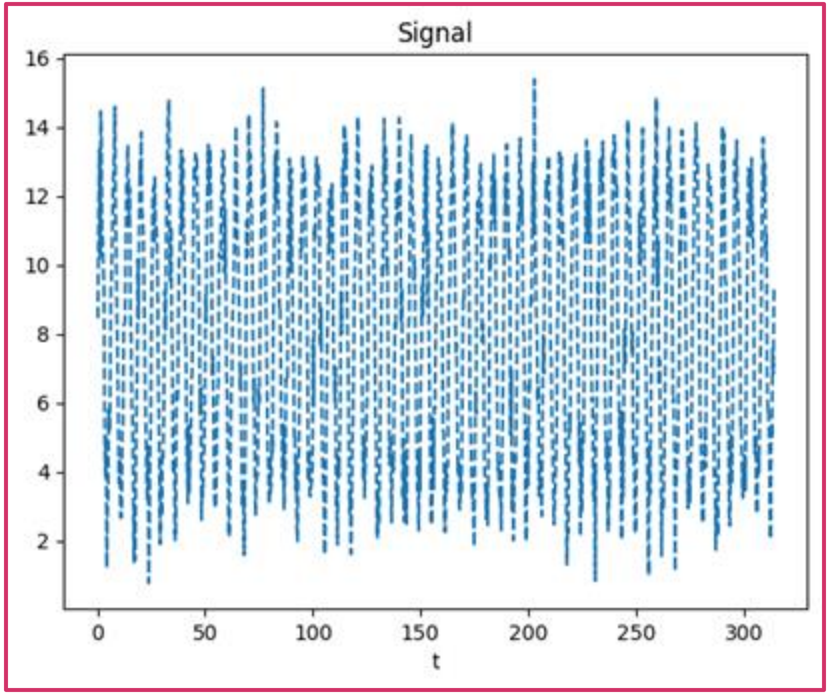
M dwarf flares

Remark 1

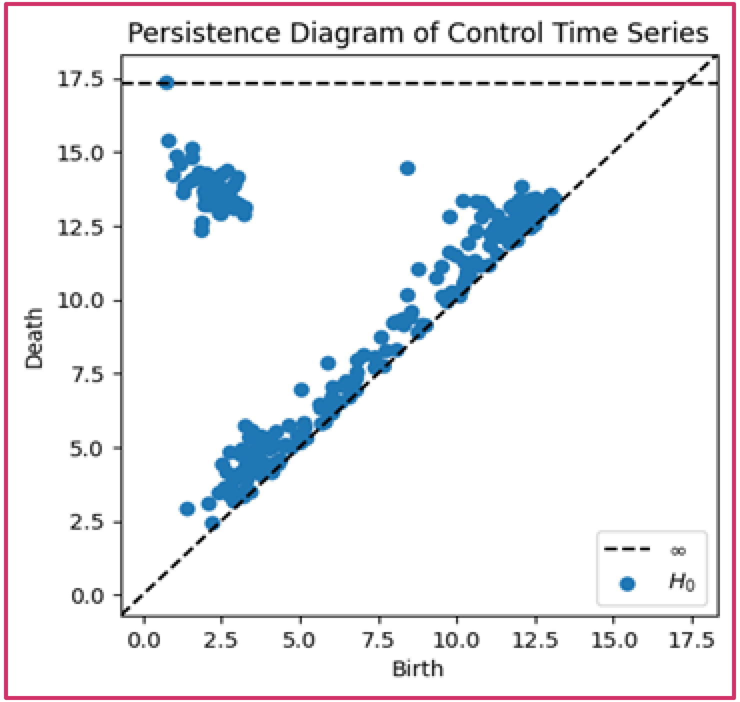
What is the effect of noise on the PD of a time series?



Persistence Diagram
Death
Birth
t
∞
H0
Persistence Diagram
Death
Birth
t
∞
H0
Persistence Diagram
Death
Birth
t
∞
H0
Persistence Diagram
Death
Birth
t
∞
H0

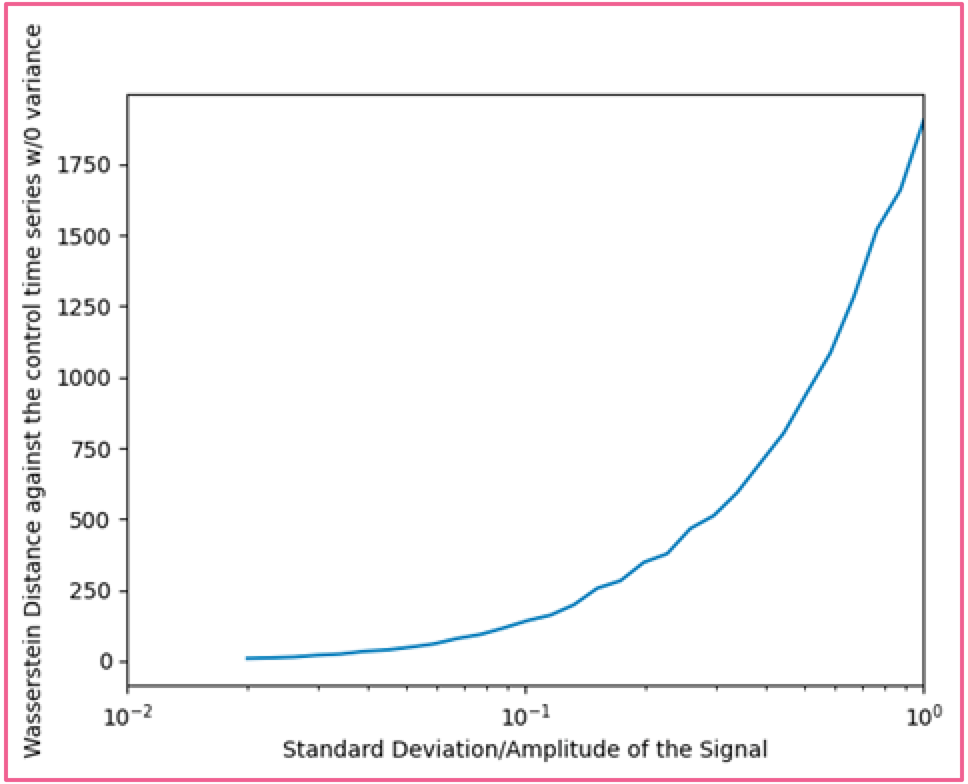


Signal
16
14
12
10
8
6
4
2
0
50
100
150
200
250
300
t



Persistence Diagram of Control Time Series
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Death
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Birth
∞
H_0

Wasserstein distance vs. N/S of sinusoidal functions

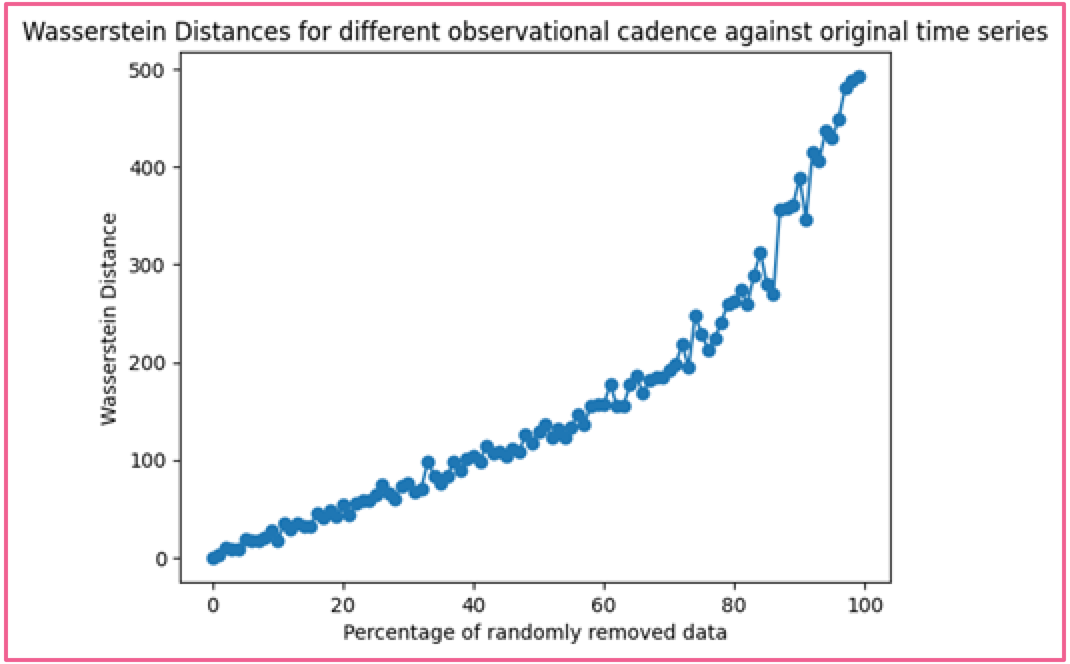


Remark 1

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

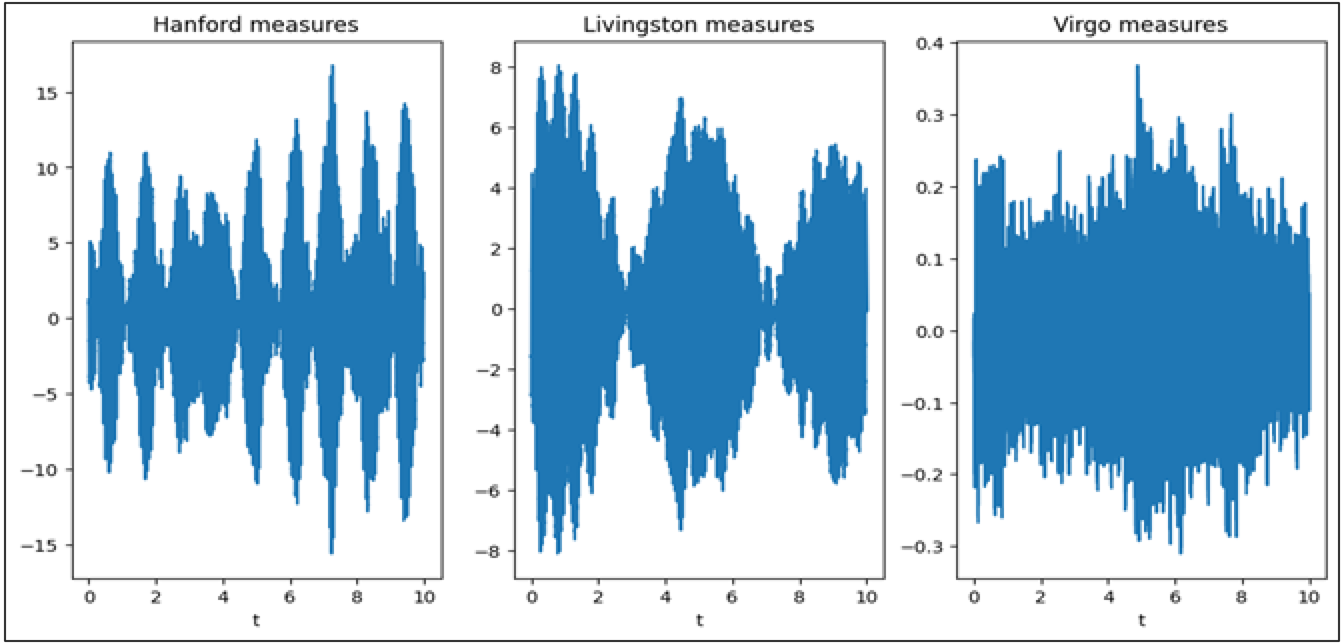
Remark 2

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
Wasserstein Distance
Percentage of randomly removed data
0
100
200
300
400
500
0
20
40
60
80
100

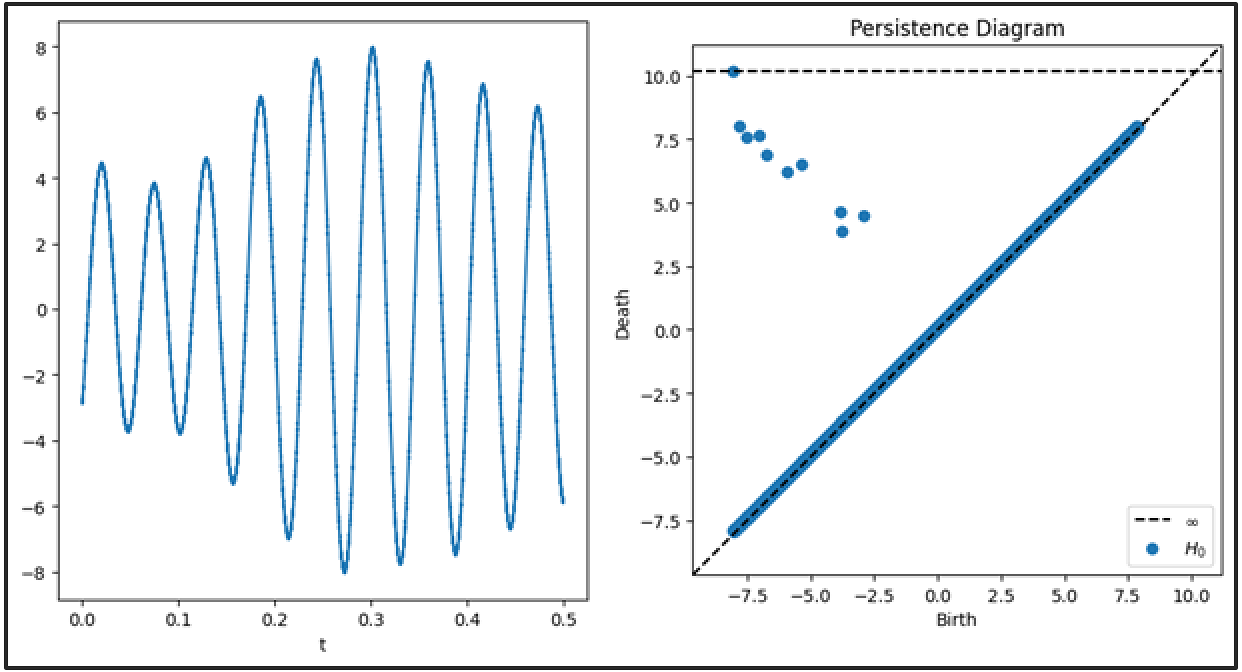
Laser Interferometer Gravitational-Wave Observatory (LIGO) data

- 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

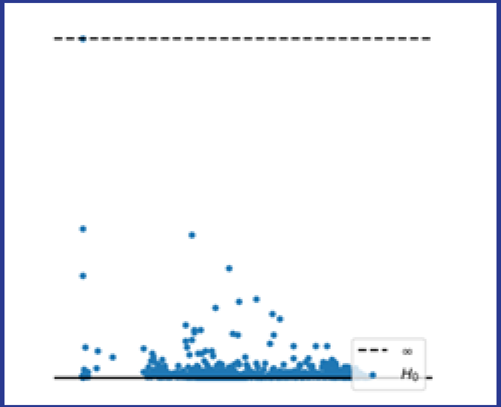
- We performed segmentation of the time series in 20 chunks of 8192 points each.
- First chunk from Livingston dataset and its PD.



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!



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://plasticc.org/
★ https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full
★ https://en.wikipedia.org/wiki/Topology