Introduction to the Data Challenge

Filip Morawski
NCAC PAS

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Data Science School, Braga 25-27.03.2019

LG2NET

Link to notebooks

https://cernbox.cern.ch/index.php/s/VSDpUpsavpmZR4A

Link to the data

https://owncloud.ego-gw.it/index.php/s/nHXFIJrCvAoDWob

Virtual Environment

```
cd
mkdir -p .pythons/3.5
cd .pythons
wget https://www.python.org/ftp/python/3.5.3/Python-3.5.3.tgz
tar zxvf Python-3.5.3.tgz
cd Python-3.5.3
make clean
./configure --prefix=$HOME/.pythons/3.5
make -j4
make install
cd ..
rm -rf Python-3.5.3*
```

Virtual Environment

```
cd $HOME
mkdir python-virtualenvs
cd python-virtualenvs
~/.pythons/3.5/bin/pyvenv myenv3.5

source ~/python-virtualenvs/myenv3.5/bin/activate
```

Virtual Environment

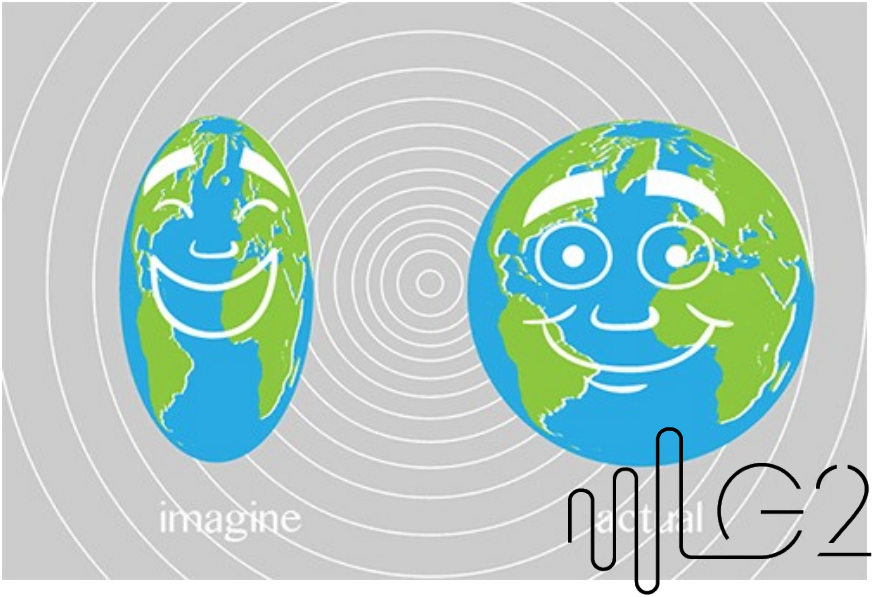
```
pip install numpy
pip install scipy
pip install matplotlib
pip install pandas
pip install sklearn
pip install h5py
pip install tensorflow
pip install keras
pip install astropy
pip install gwpy
```

LIGO/Virgo data

Main information in the LIGO/Virgo data is “strain”

Strain – relative change in distance

$$h = \frac{\delta L}{L}$$



LIGO/Virgo data

Data is stored in various formats but we focus only on the hdf5.

What we can find inside hdf5 tree?

1. meta – meta-data of the file like GPS times covered, which instrument etc…

2. quality – quality of each second of data (not important for you)

3. strain – main data collected by the interferometer

Available data sets

You can find data at https://www.gw-openscience.org
For the sake of the challenge we already downloaded and prepared the data for you. Stay tuned… :)

Two different types of data release:

1. Gravitational wave data surrounding discoveries

2. Data taken during a whole observation run

Available data sets

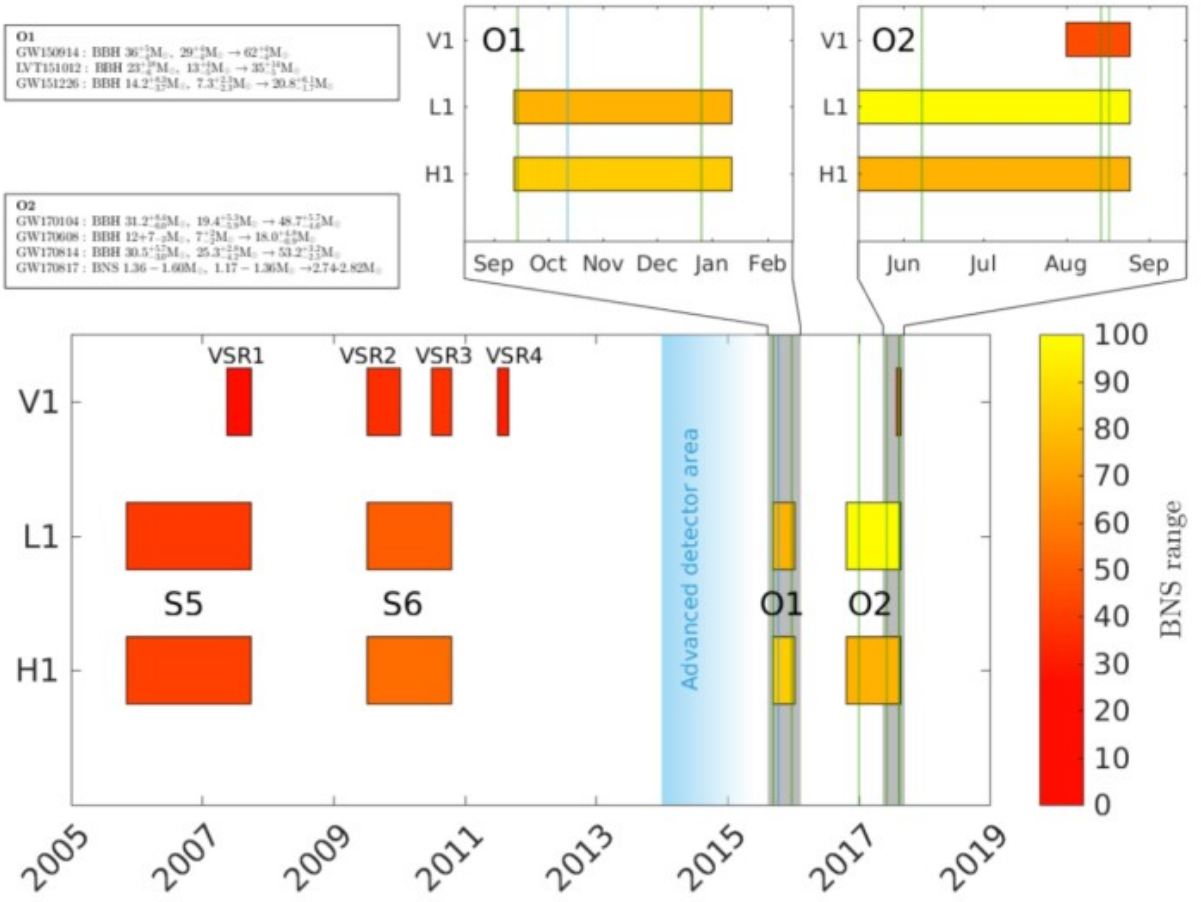


Image Credit: Courtesy Valentin Frey

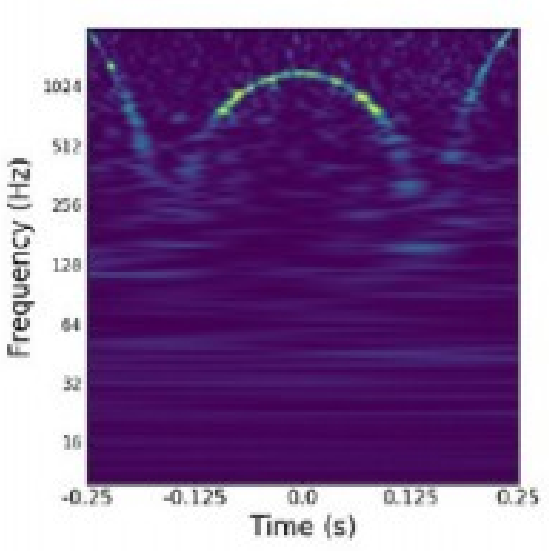
Data from scientific runs

In our challenge we are going to focus on the data from O1 – one of the two newest scientific runs.

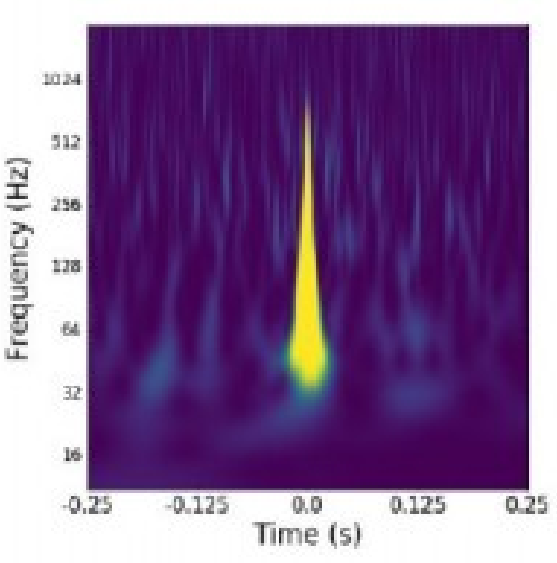
What we can find inside beside astrophysical signal? Glitches!

Glitches

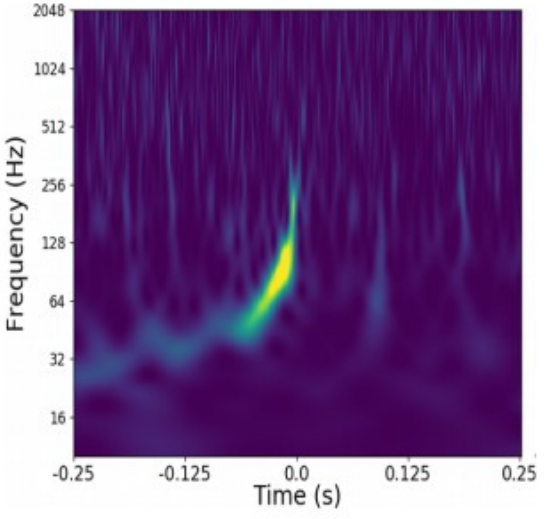
Glitch is a transient noise event that might mimic real astrophysical signal and/or influence the quality of data.



Whistle



Blip



Gravitational Wave

Your challenge is to classify various glitches buried in the detector noise

Data content

- 6667 labeled glitches from O1
- 22 classes
- unbalanced dataset (number of instances for each glitch varies)
- metadata + time-series

Let’s have a look into the data

→ IntroductionData notebook

The Challenge(s)

We offer two challenges:

1. Classification of glitches using metadata

2. Classification of glitches using strain in the form of time-series or images/spectrograms

Classification of glitches using metadata

Create a classifier using only metadata:

- Input: 5 metaparameters (or more - feature engineering)
- Output: 22 classes
- Algorithms: neural networks, random trees, xgboost

Classification of glitches using time-series/images

Create a classifier using time-series/images:

- Input: 1D time series or spectrograms (plus metadata ?)
- Output: 22 classes
- Algorithms: convolutional neural networks (1D or 2D), xgboost

What is GWpy?

A python package for gravitational-waves analysis:
http://gwpy.github.io

Depends on numpy, scipy, astropy and matplotlib.

Provides methods to access the data, process and visualize them.

Signal processing with GWpy

The data you get at the beginning is just the **raw** signal. It might not be the best representation depending on the analysis you want to follow. You might consider processing it into more suitable form.

GWpy provides few signal processing methods like:

- bandpass, lowpass and highpass filtering,
- whitening.

Data preprocessing

How can you prepare the data for the second challenge?

→ DataPreprocessing Notebook

Good luck!