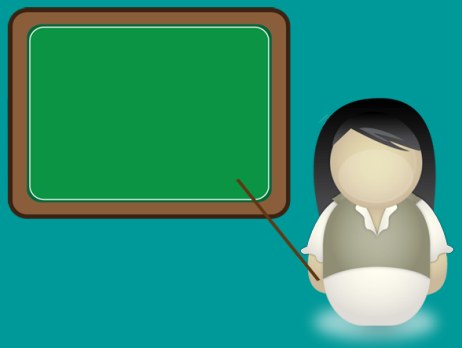
Machine Learning for Gravitational Wave signals classification in LIGO and Virgo

Elena Cuoco
European Gravitational Observatory
www.elenacuoco.com
@elenacuoco

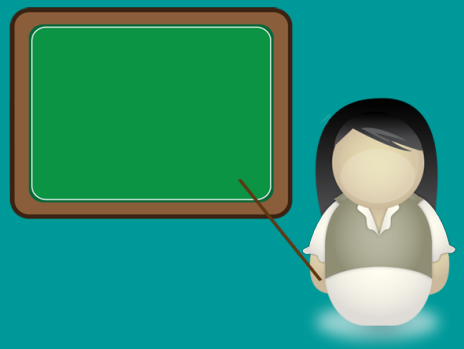
About me

- In Virgo collaboration since 1995. EGO staff since 2004.
- Working as Data Analyst
- Noise and data preprocessing
- Machine Learning ‘challenger’
- ML *promoter* in LIGO/Virgo collaboration
- Data analysis passionate (more on www.elenacuoco.com)

Contents

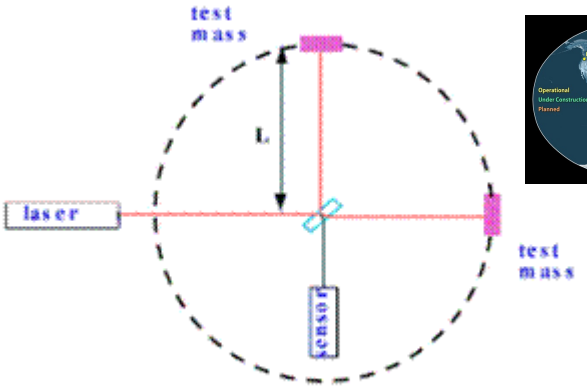


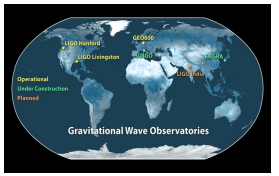
- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next



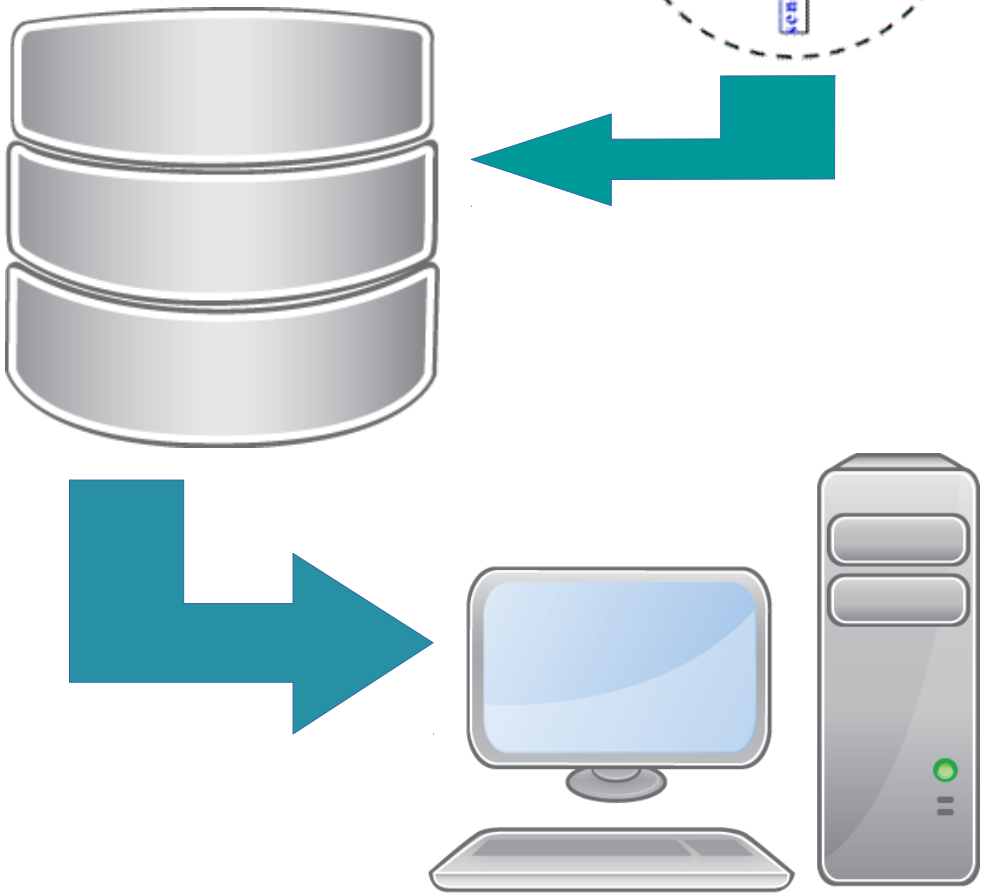
- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

The Data

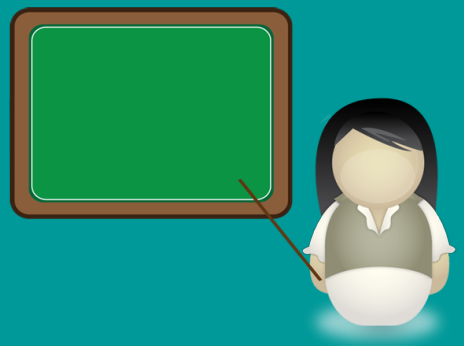




The Data

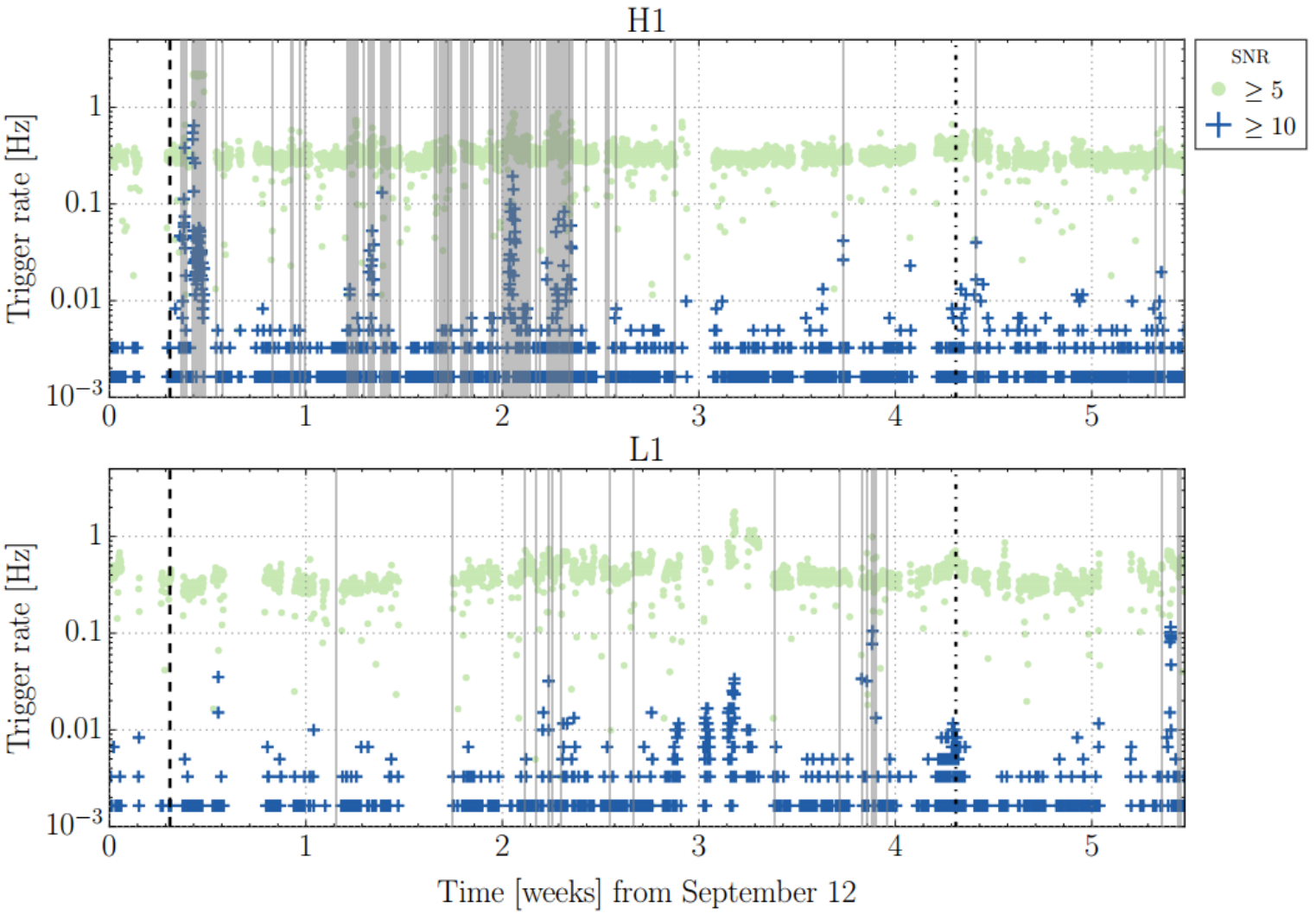


- Time series…
  - ***Noisy*** time series
- Many channels recorded. Data Flux 40 MB/s
- O1 run GW channel: 1.3TB
- GW Event to be identified in 2-3 mins



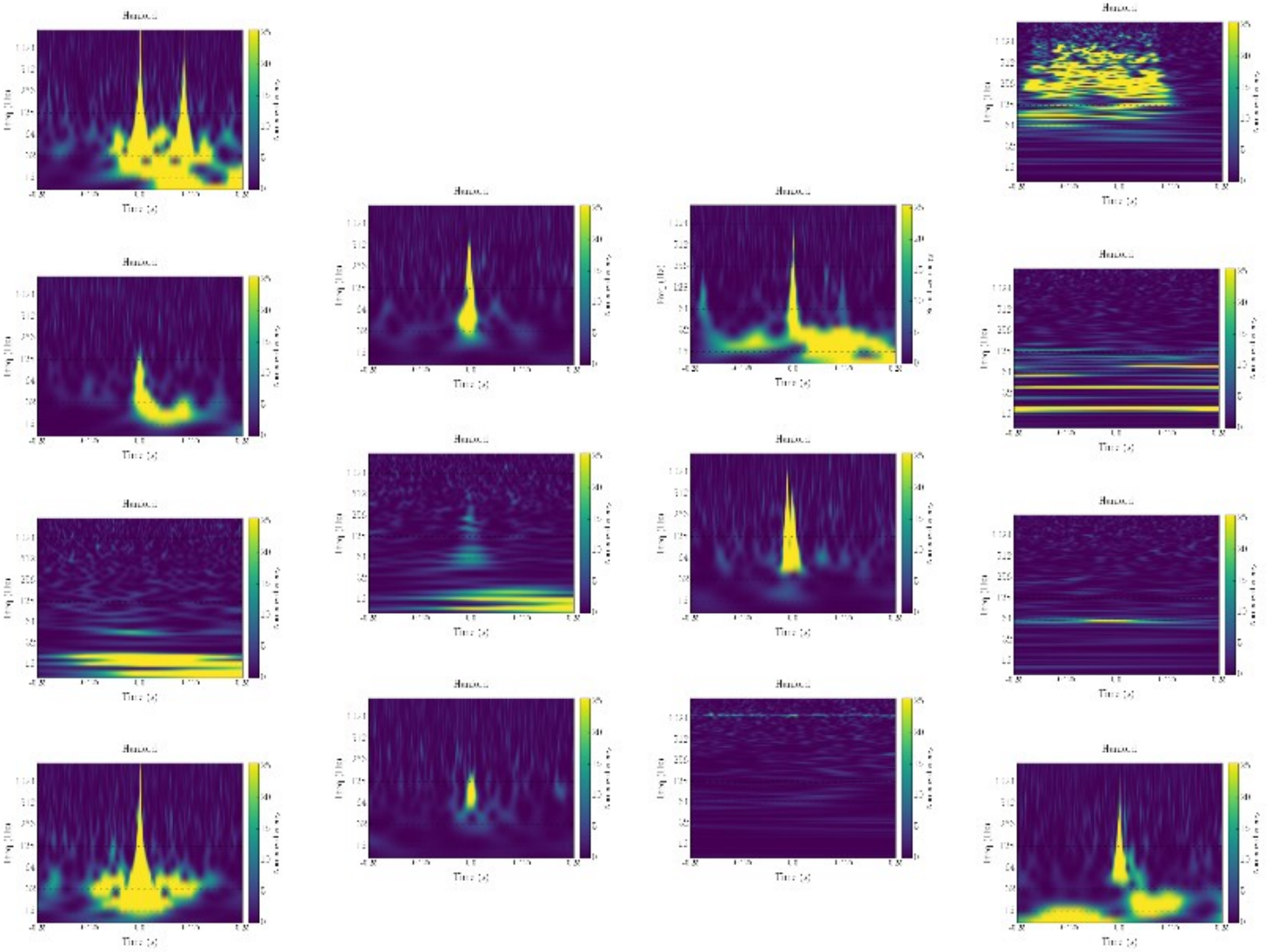
- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

How many “trash” events?



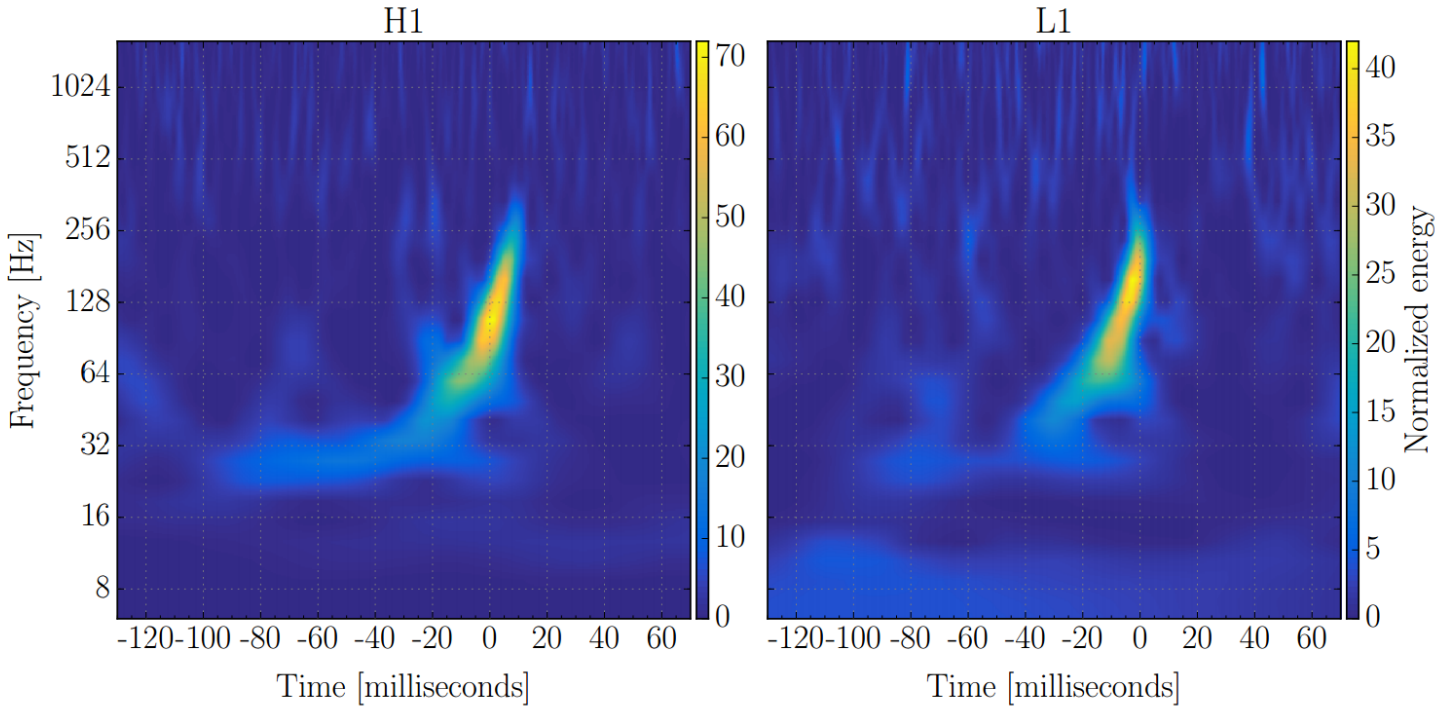
LIGO L1 and H1 triggers rates

That trash is our **Glitches zoo**



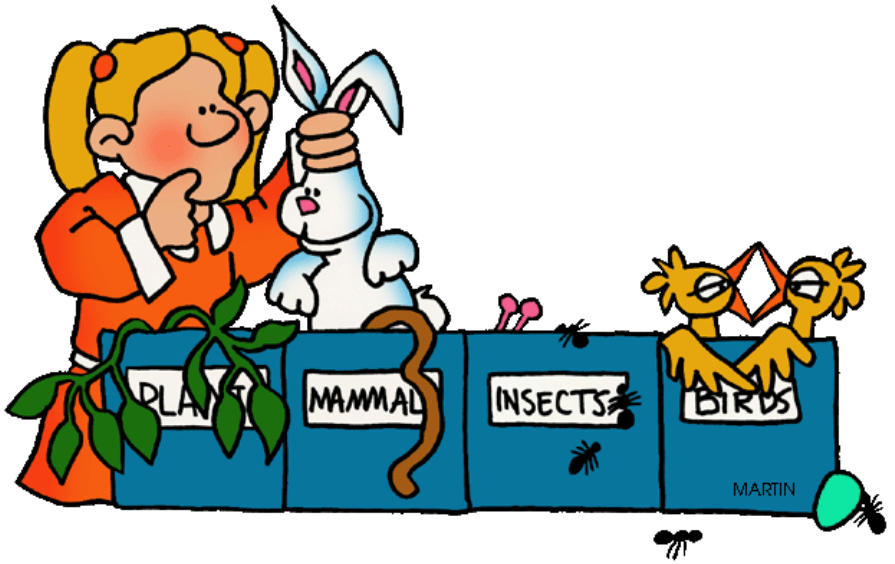
Time Frequency images

Gravitational Waves Signal

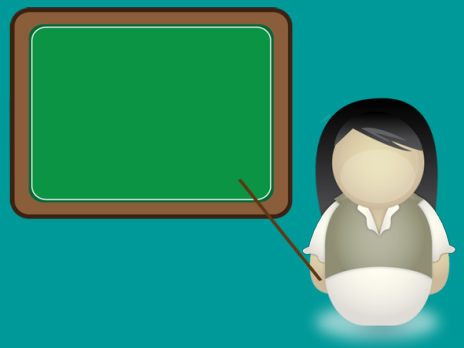


Why Signals Classification ?

- In our time series most of the signals are due to noise events.
- If we are able to classify the noise events, we can clean the data in a fast and clear way.

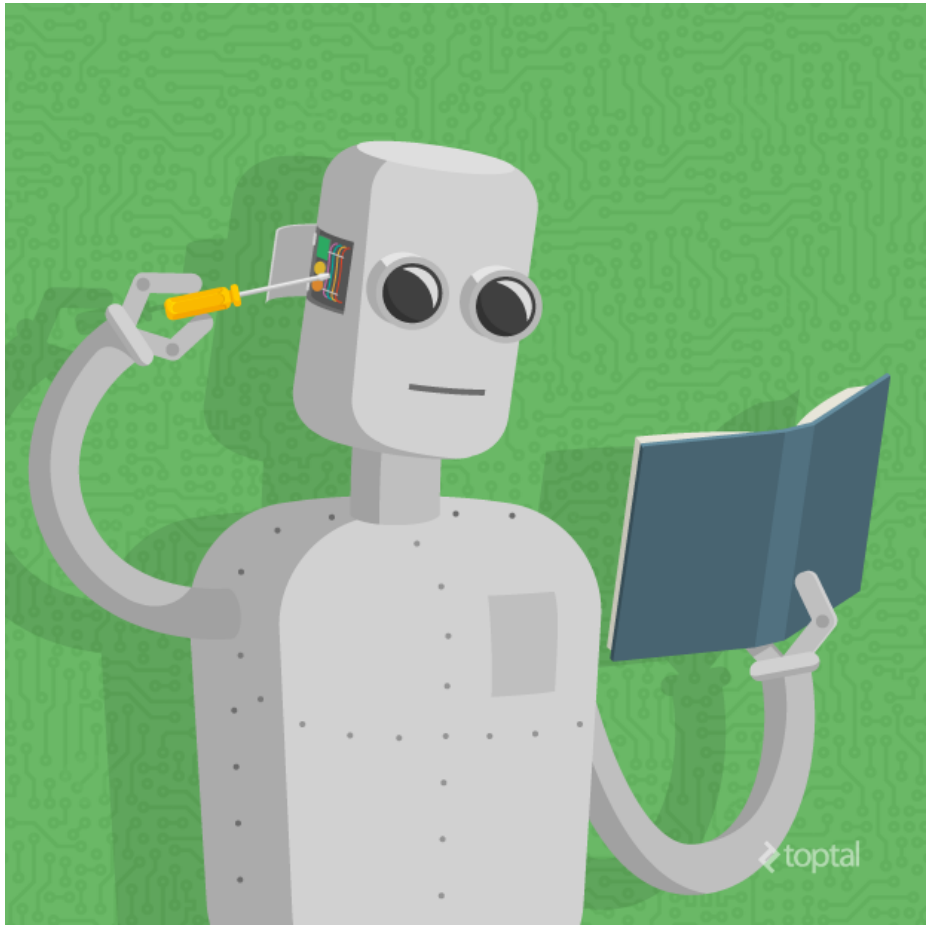


Machine learning approach for classification



- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

What is Machine Learning?

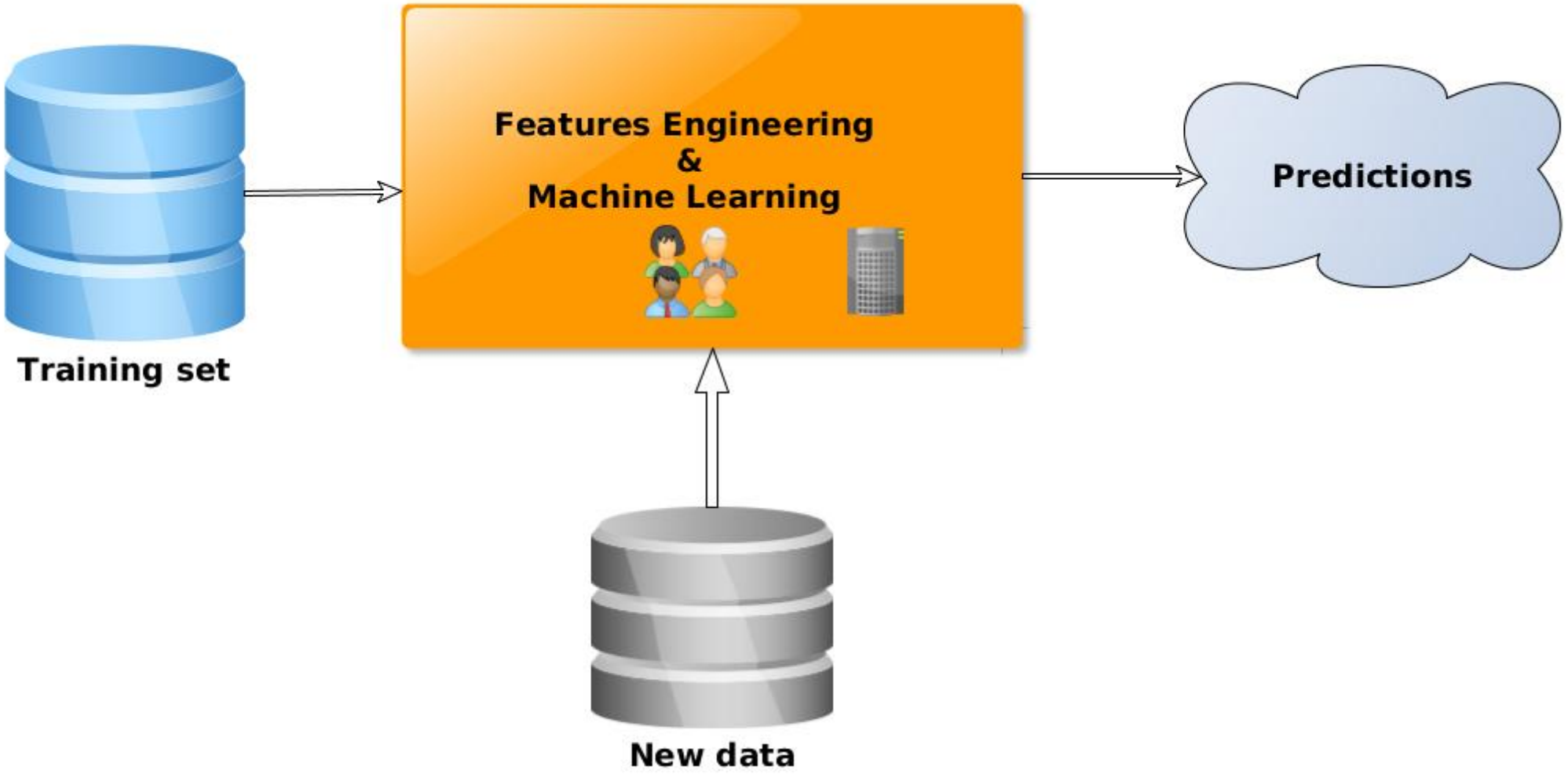


Arthur Samuel in 1959: "[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed."

Machine Learning is on all our day by day lives:

- Google search
- Facebook
- Images recognition
- Bank accounting
- Shopping
- Travels
- …and much more

Machine Learning Flow



Machine Learning approaches

Supervised

- Training Data set with labeled inputs.
- Let your algorithm learn from that
- Do predictions

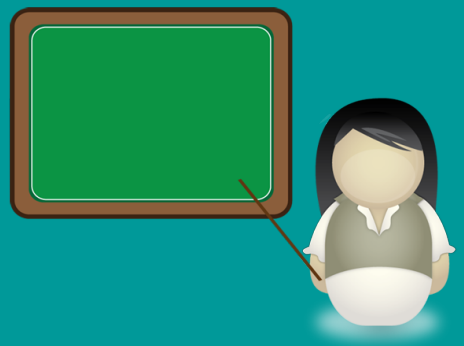
Example: SVM, RandomForest,

Neural Net, etc..

Unsupervised

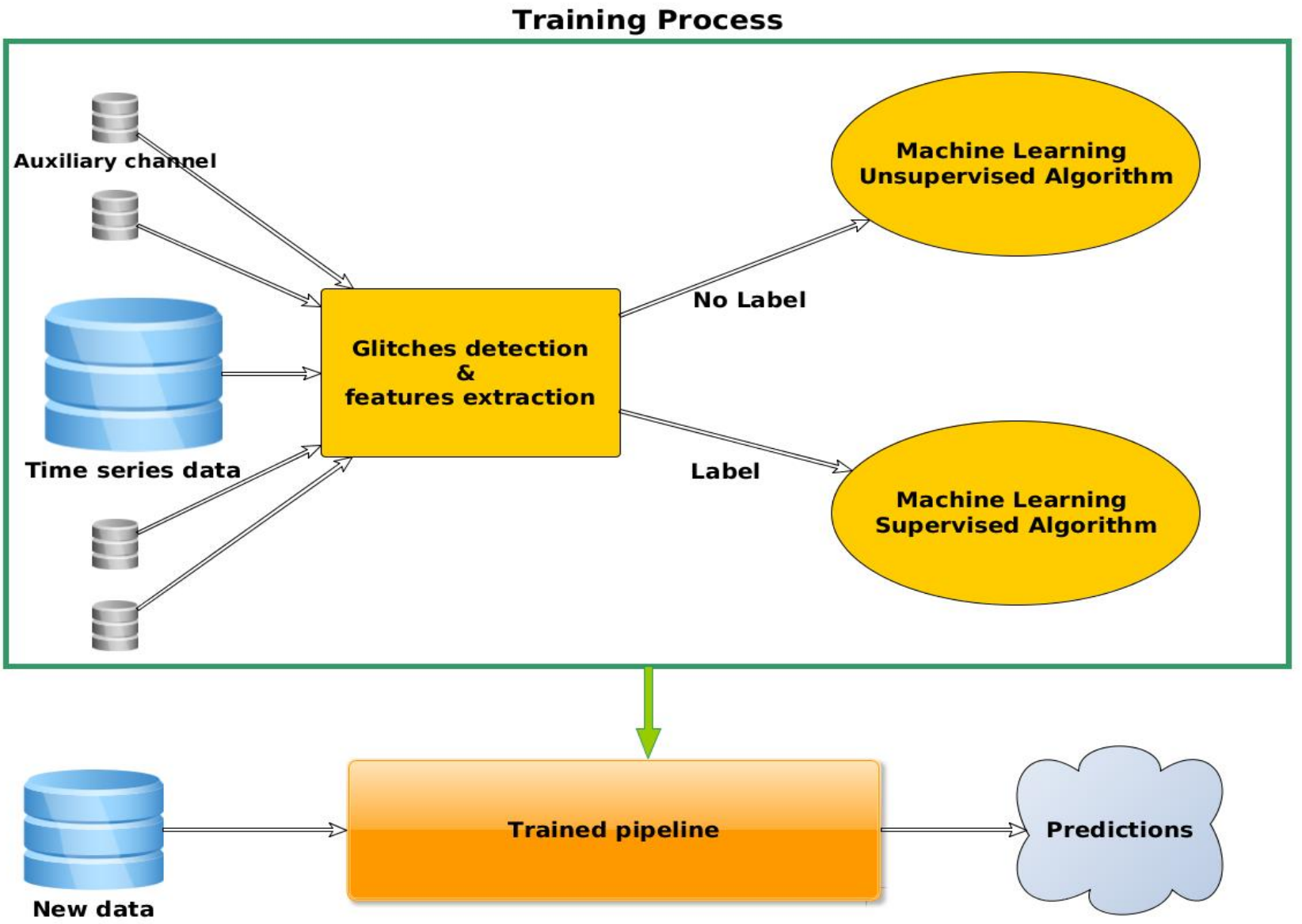
- Training Data set not labeled
- Let your algorithm identify main pattern inside the data
- Do predictions

Example: Clustering, K-Means,etc…



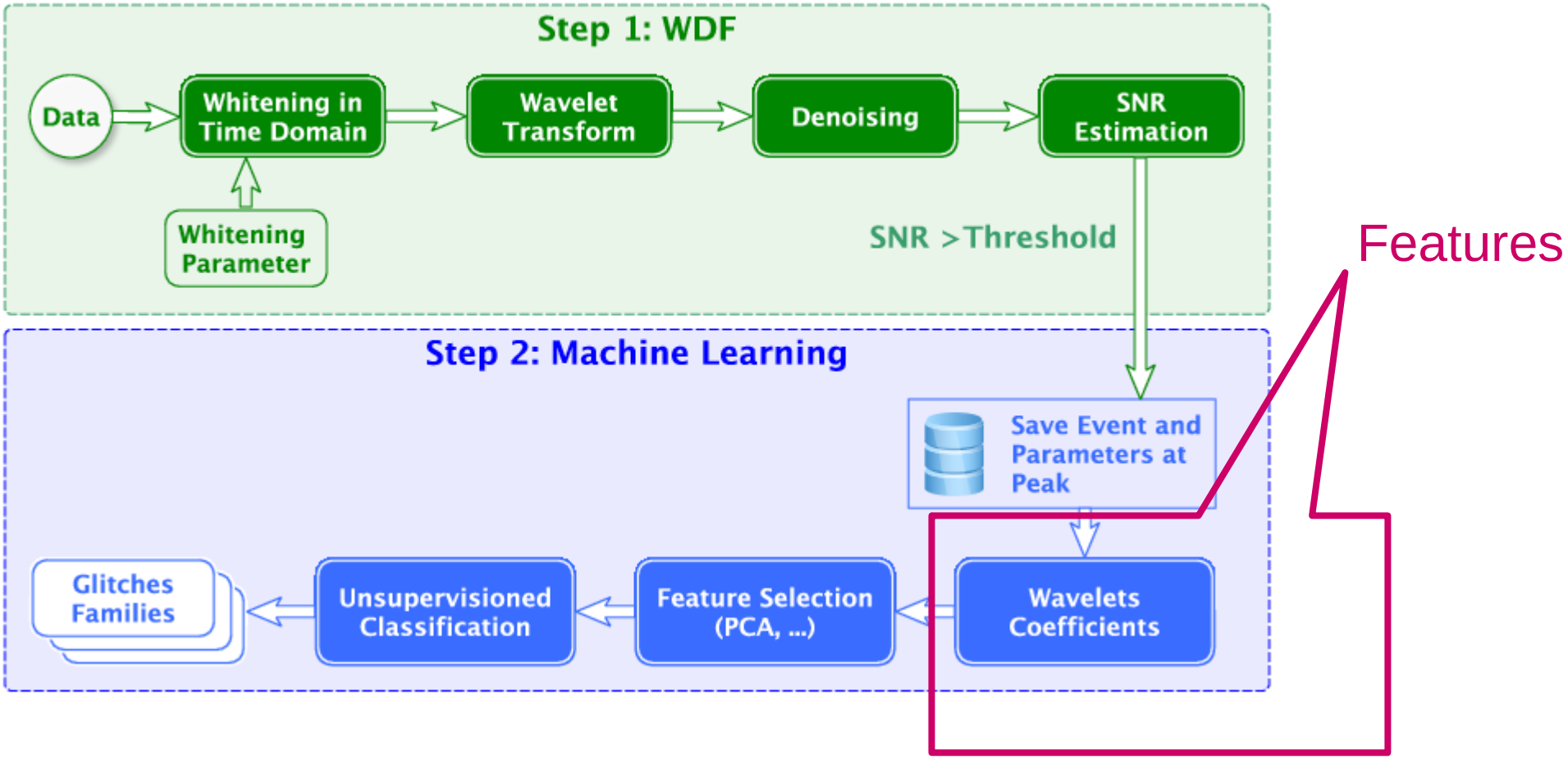
- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

Glitches Classification Strategy



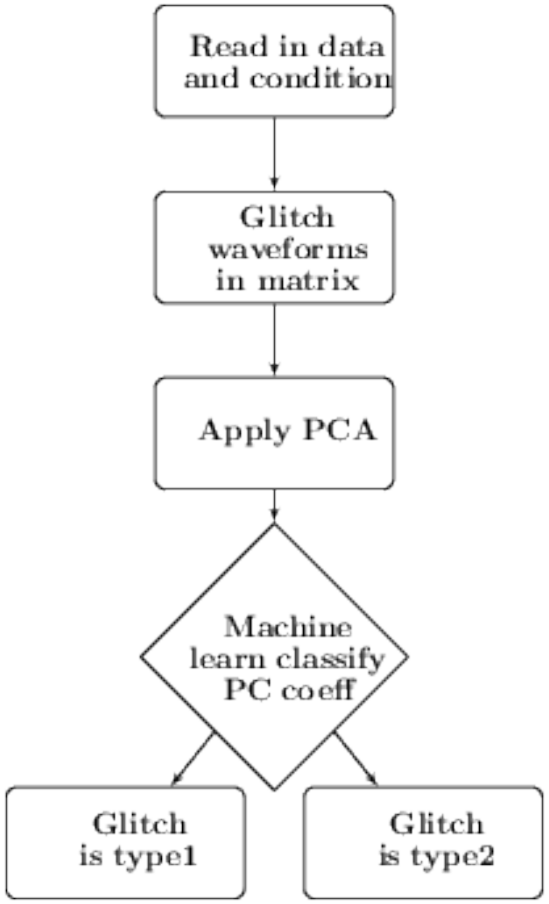
EXAMPLE OF GLITCHES DETECTION AND ML PIPELINE

Wavelet Detection Filter and ML

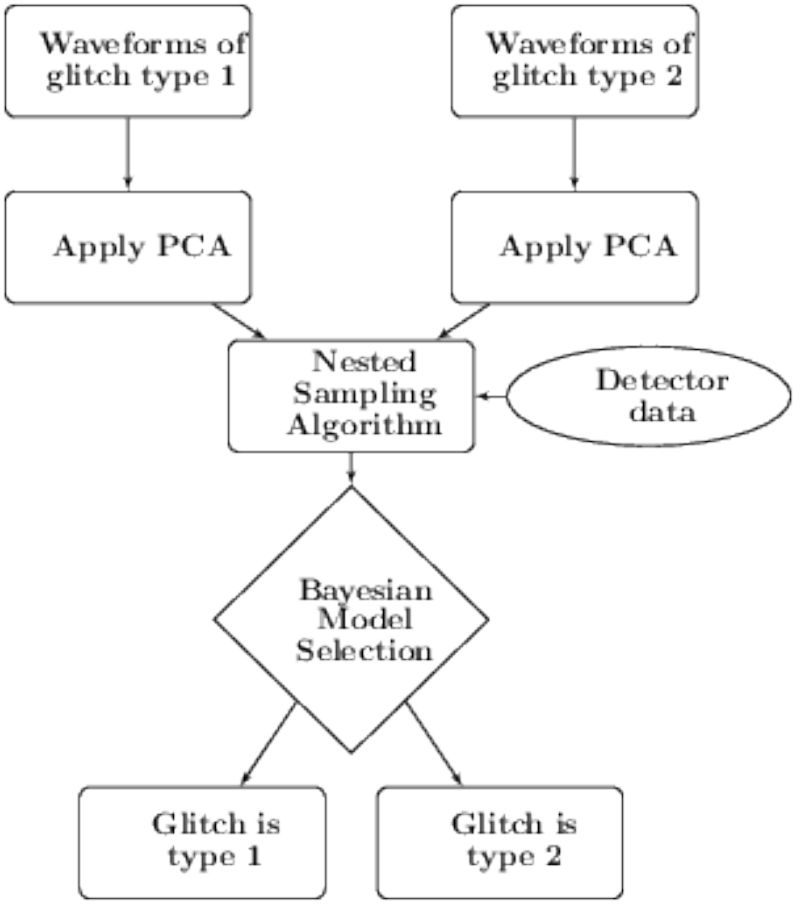


Two more pipeline: PCAT and PC-LIB, LIGO collaboration

PCAT



PC-LIB

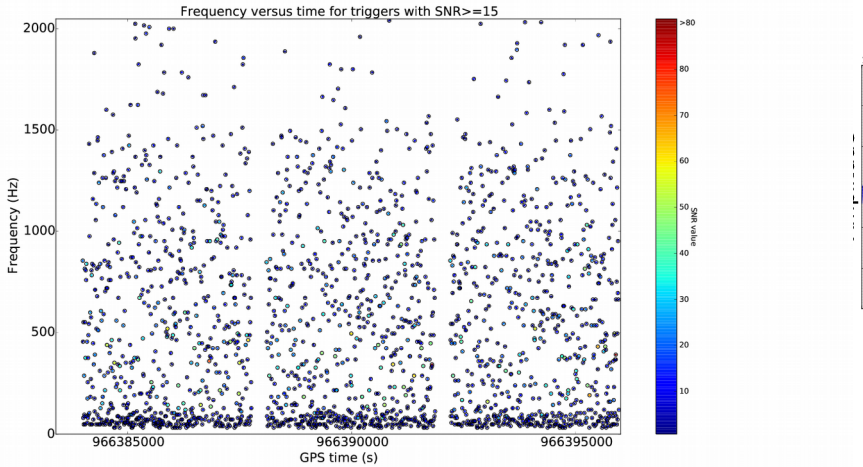


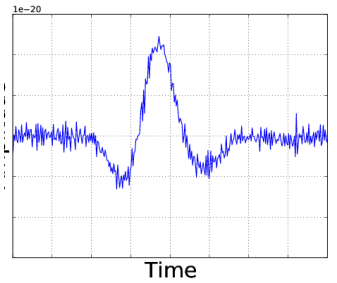
Collaboration with Glasgow and Missisipi University
J.Powell, D. Trifirò, M. Cavaglia, E. Cuoco, H.K. Siong

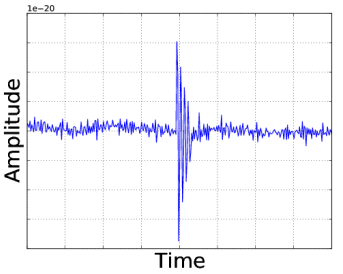
Test on simulated data

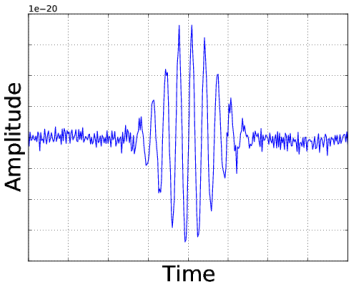
A snapshot of simulation

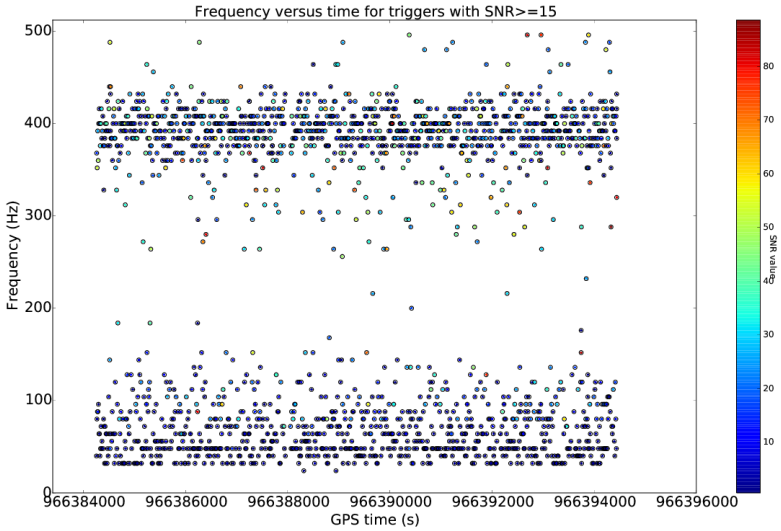
aLIGO-like simulated noise with transient signals injected

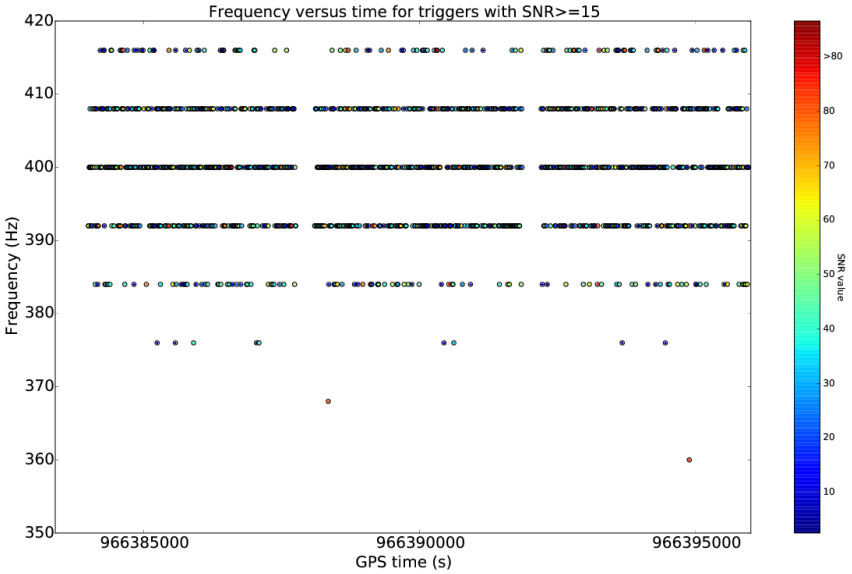




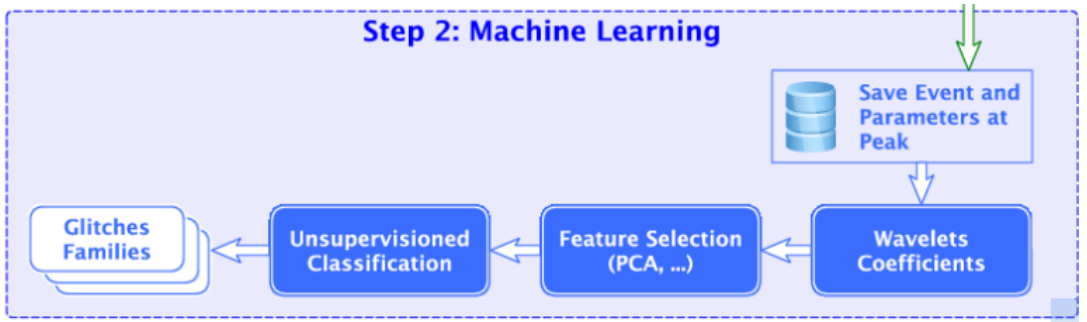




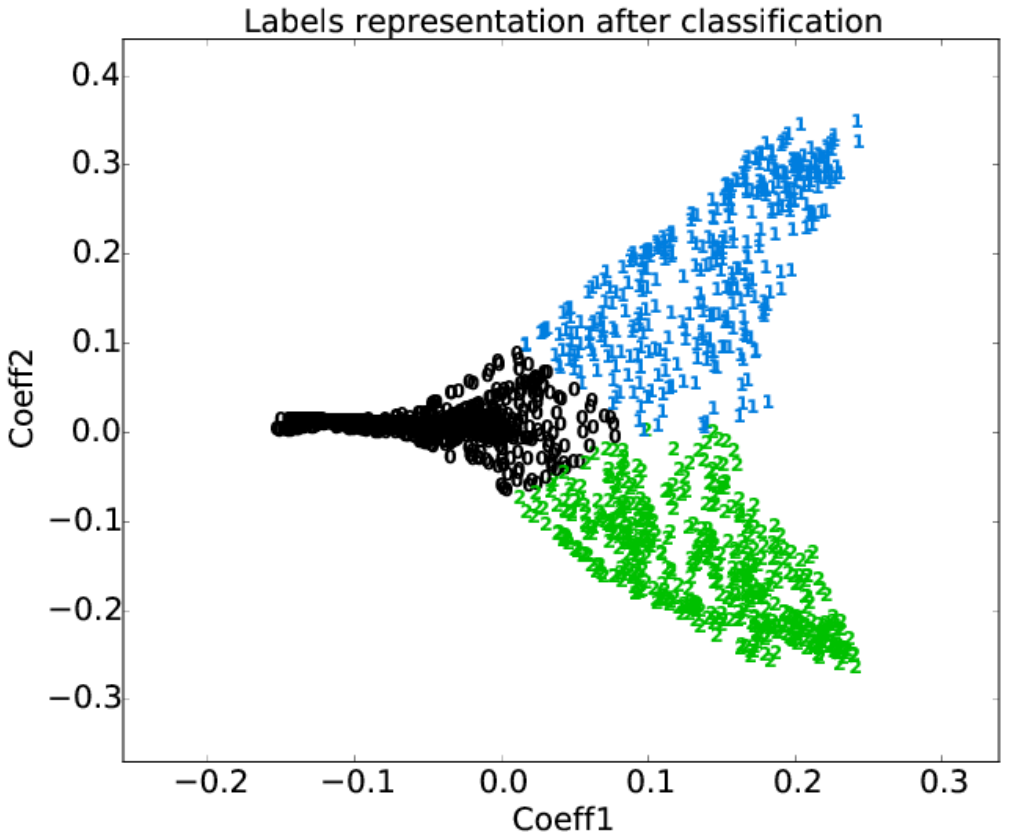


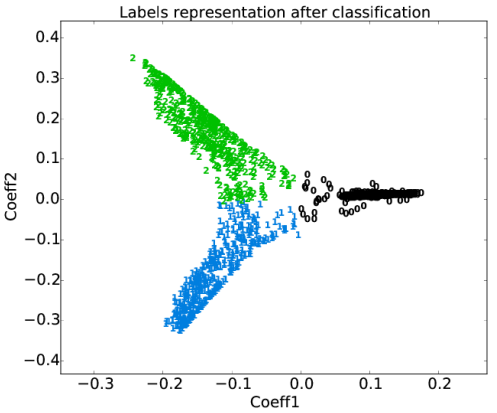


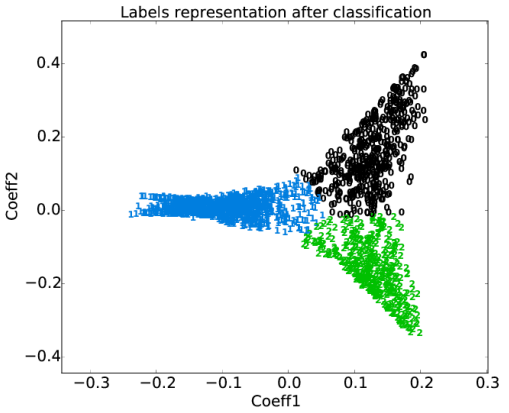
WDF-ML is able to cluster the 3 main classes



Reduced features projections after labeling







PCAT, PC-LIB, WDF-ML comparison

	SG	G
PCAT Type 1	99%	0%
PCAT Type 2	1%	100%
LIB Type 1	99.9%	5%
LIB Type 2	0.1%	95%
WDF Type 0	99.5%	2.4%
WDF Type 1	0.3%	46.1%
WDF Type 2	0.2%	51.5%

	SG	RD
PCAT Type 1	1.1%	97.4%
PCAT Type 2	98.9%	2.5%
LIB Type 1	97.8%	4.8%
LIB Type 2	2.2%	95.2%
WDF-ML Type 0	8.7%	100%
WDF-ML Type 1	48.0%	0%
WDF-ML Type 2	43.3%	0%

	SG	G	RD
PCAT Type 1	15.5%	0%	13.6%
PCAT Type 2	36.8%	0%	41.4%
PCAT Type 3	14.2%	0%	13.0%
PCAT Type 4	9.1%	0%	13.0%
PCAT Type 5	0.8%	0%	0.3%
PCAT Type 6	21.8%	0%	17.2%
PCAT Type 7	1.8%	100%	1.5%
LIB Type 1	39.5%	4.9%	23.8%
LIB Type 2	17.3%	88.3%	23.2%
LIB Type 3	43.3%	6.8%	53.0%
WDF-ML Type 0	89.5%	9.6%	86.9%
WDF-ML Type 1	5.9%	49.7%	7.0%
WDF-ML Type 2	4.6%	40.7%	6.1%

Classification methods for noise transients

in advanced gravitational-wave detectors

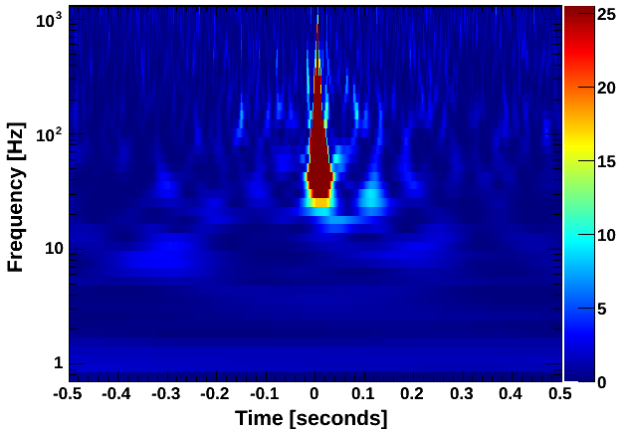
Class. Quant. Grav., 32 (21), pp. 215012, 2015

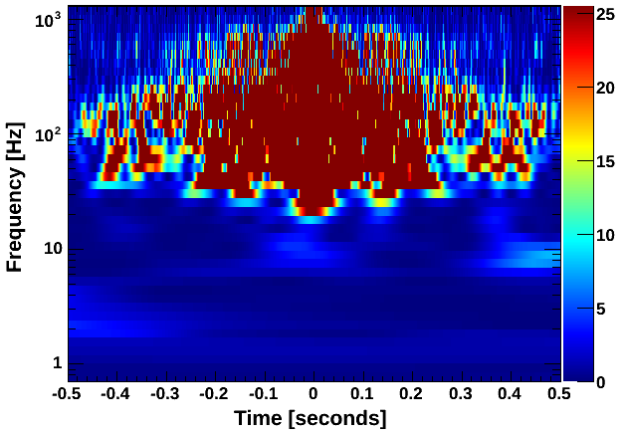
Test on real data

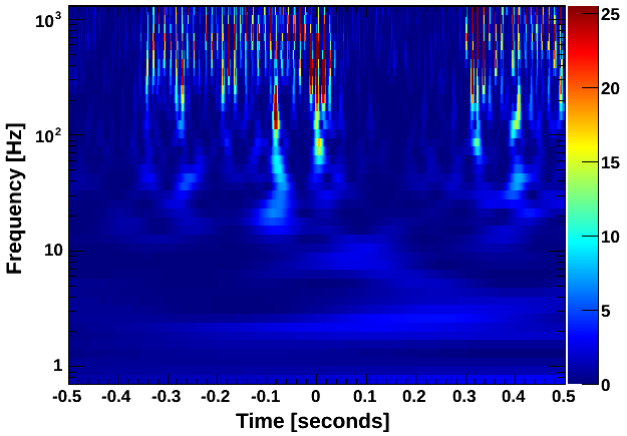
ER7 LIGO engineering run

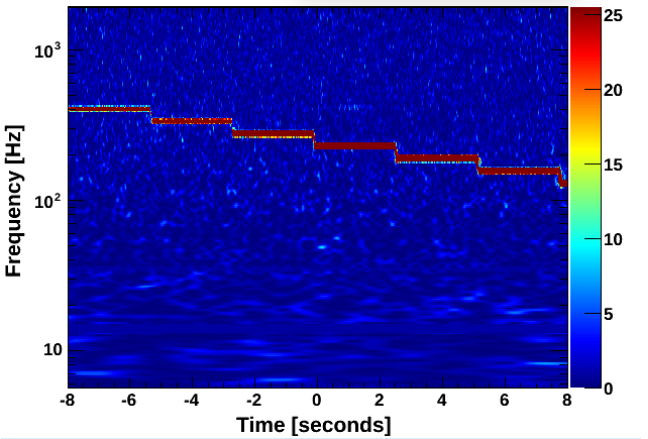
- Data from the 7th aLIGO engineering run (ER7), which began on the 3rd of June 2015 and finished on the 14th of June 2015. The average binary neutron star inspiral range for both Hanford and Livingston detectors in data analysis mode during ER7 was 50-60 Mpc.
- The total length of Livingston data analysed is about 87 hours.
- The total length of Hanford data analysed is about 141 hours.

ER7 LIGO Hanford

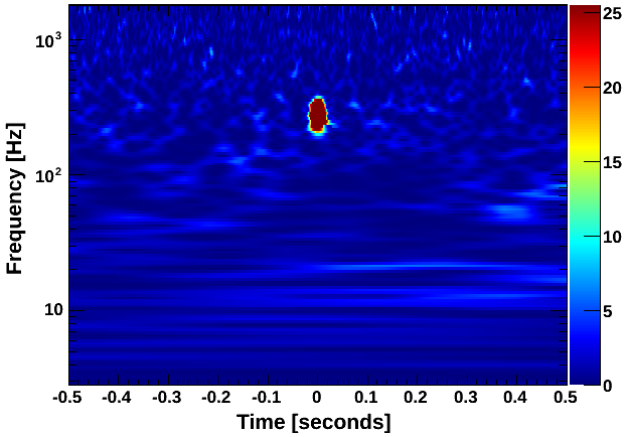


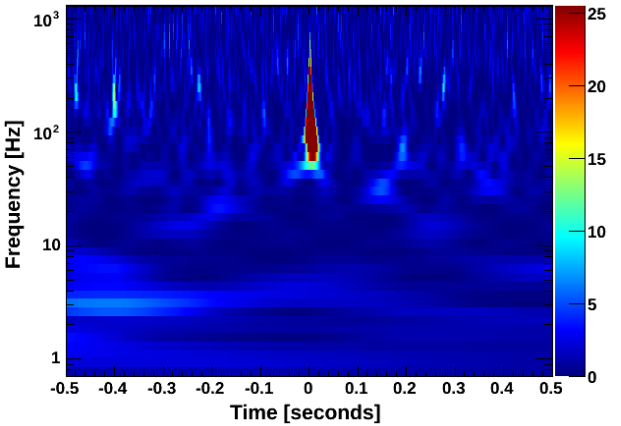


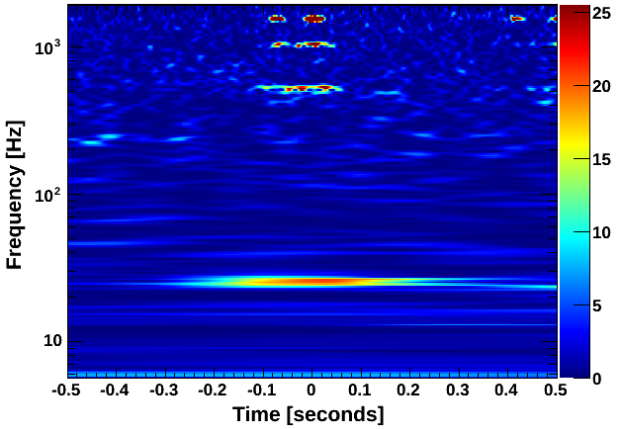


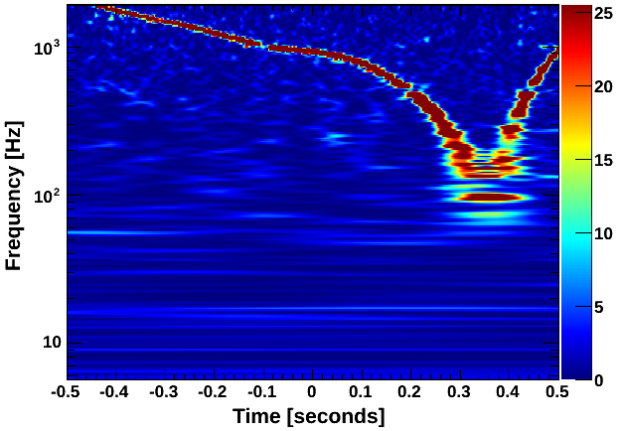


ER7 LIGO Livingston









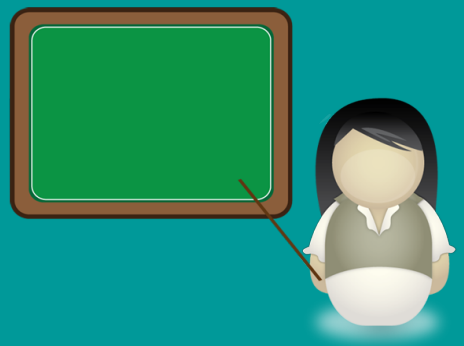
Results: Paper II

LIGO Hanford		
Pipeline	Correct classification	Missed triggers
PCAT	99%	120
PC-LIB	95%	6
WDF-ML	92%	0

LIGO Livingston		
Pipeline	Correct classification	Missed triggers
PCAT	95%	90
PC-LIB	98%	33
WDF-ML	97%	0

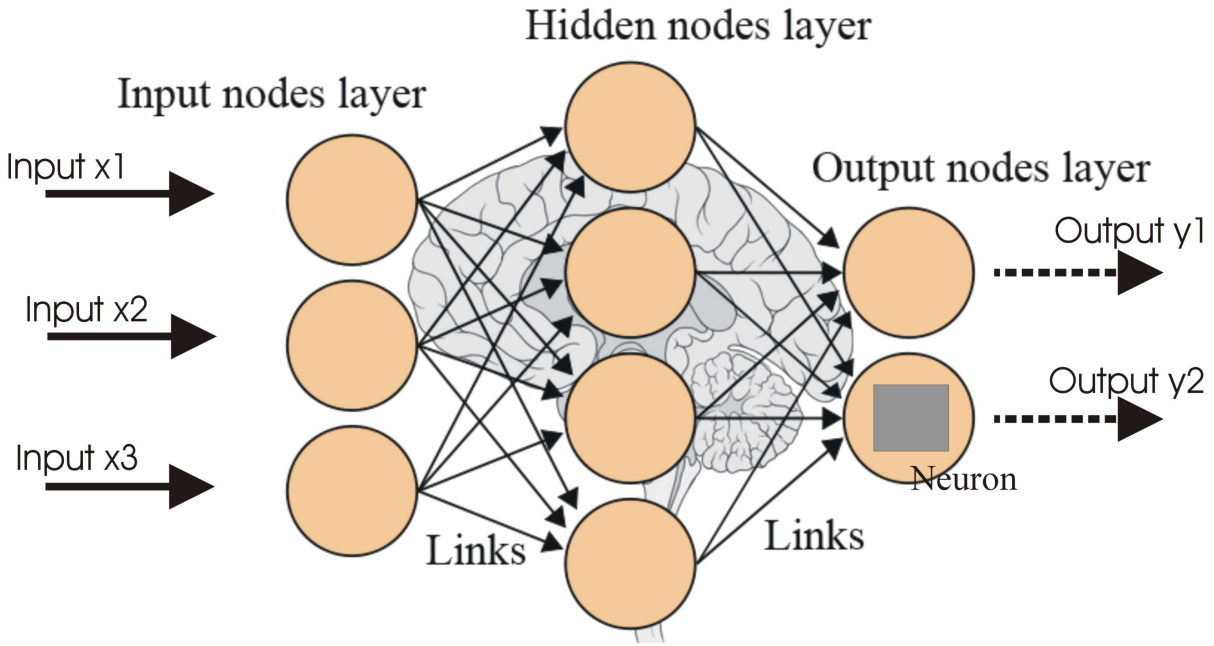
We conclude that our methods have a high efficiency in real non-stationary and non-Gaussian detector noise

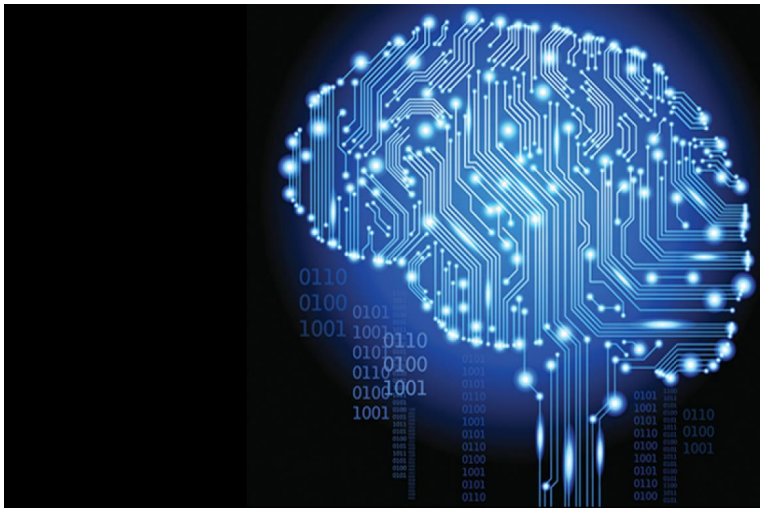
Classification methods for noise transients in advanced gravitational-wave detectors II: performance tests on Advanced LIGO data Class. and Quant. Grav, 34 (3) 2017



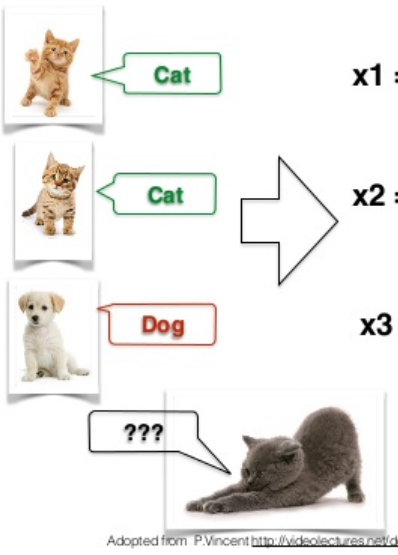
- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

What is Deep Learning?



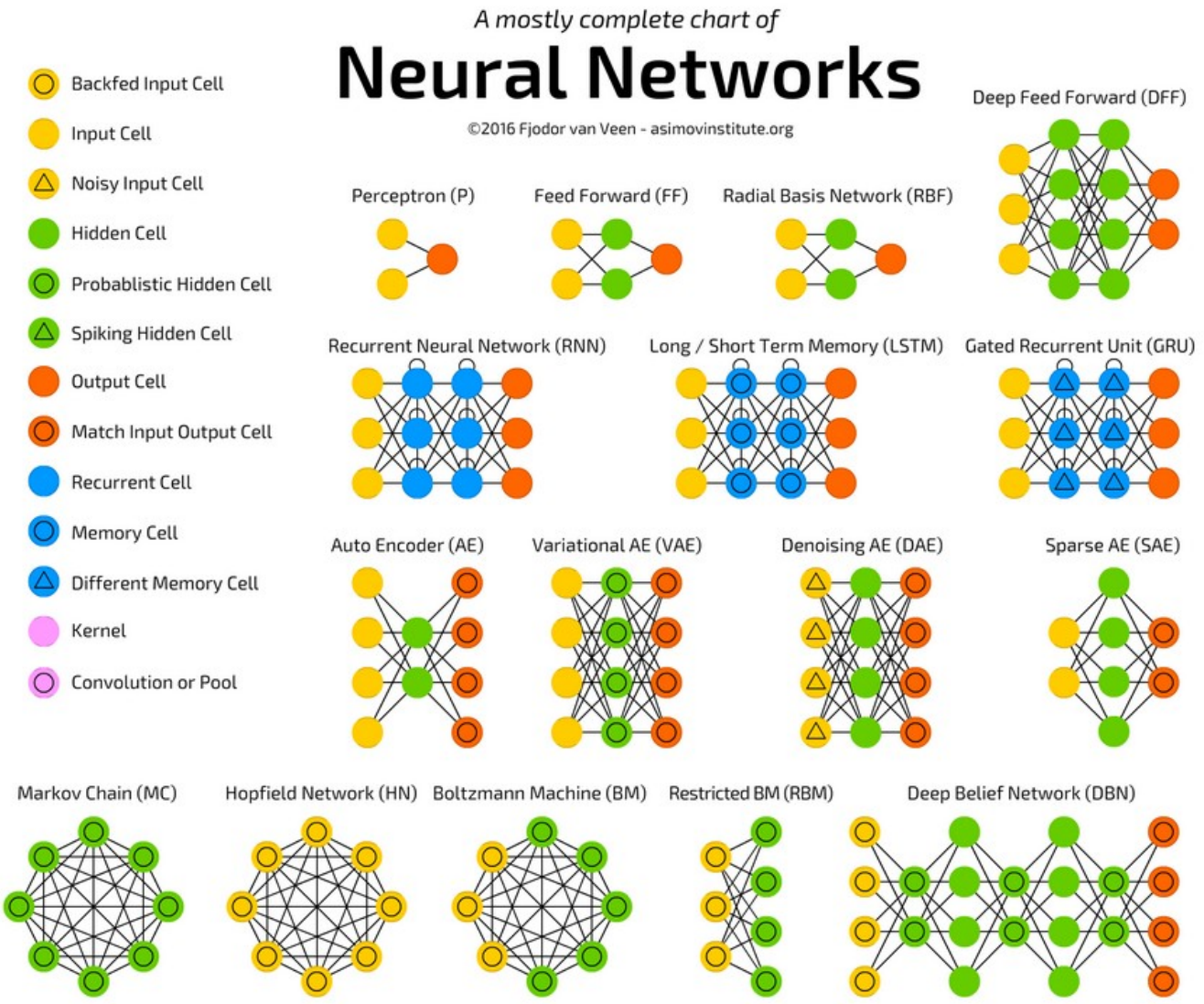


Supervised Learning

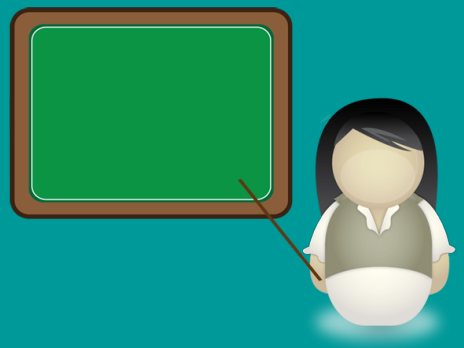


Adopted from P.Vincent http://videolectures.net/deeplearning2015_vincent_machine_learning/

Neural networks



http://www.asimovinstitute.org/neural-network-zoo/

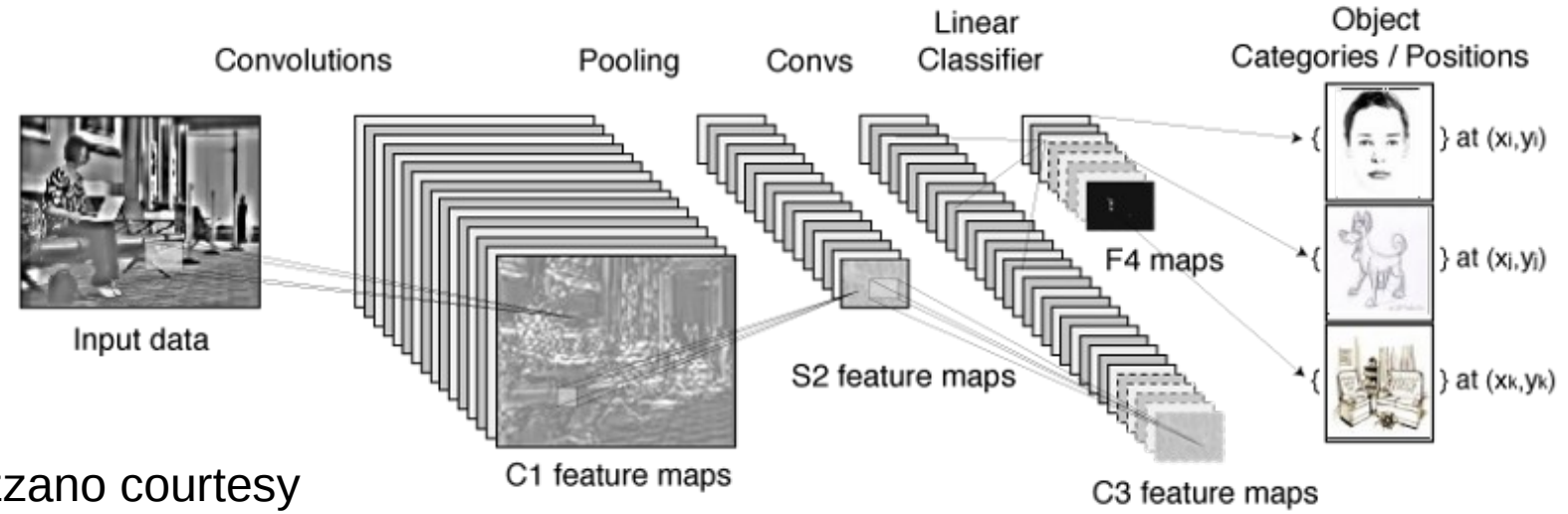


- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

Image Based Glitches Classification

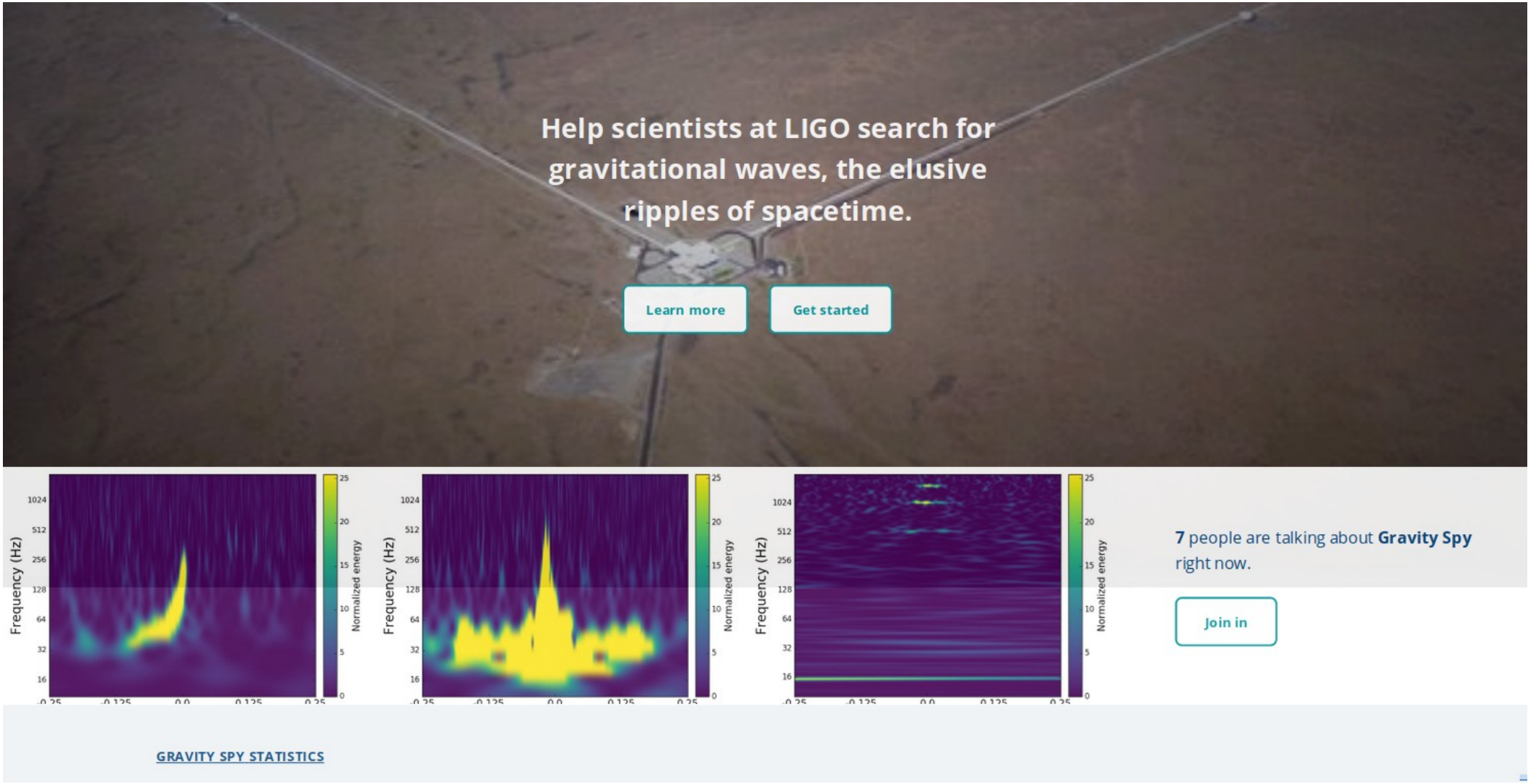
In collaboration with Massimiliano Razzano (Pisa University)

- Many approaches to data: we choose image classification of time frequency images
- The architecture is based on Convolutional deep Neural Networks (CNNs).
- CNNs are more complex than simple NNs but are optimized to catch features in images, so they are the best choice for image classification



M. Razzano courtesy

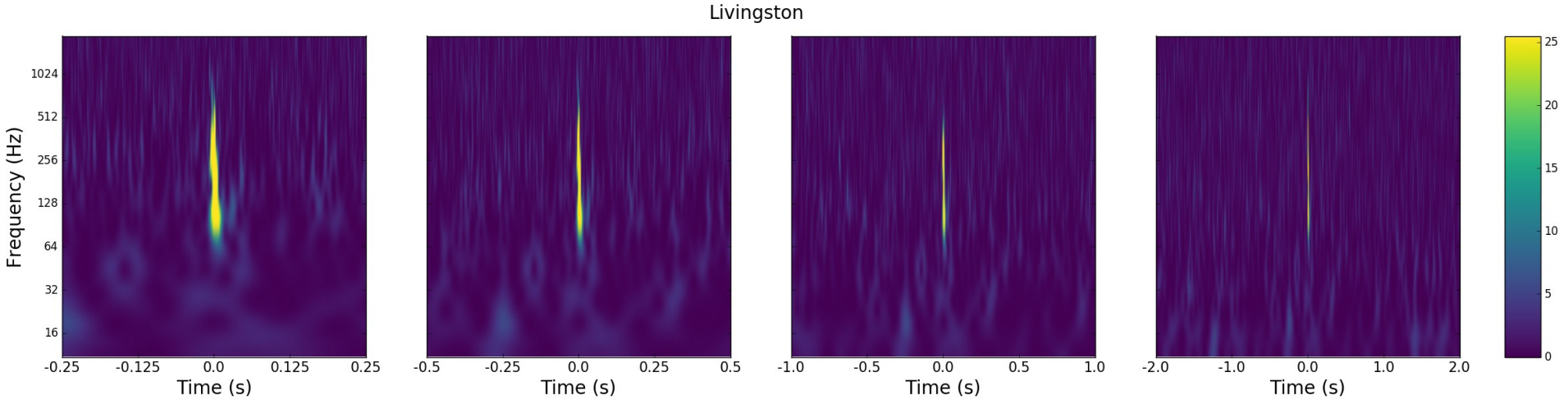
Gravity spy project



https://www.zooniverse.org/projects/zooniverse/gravity-spy

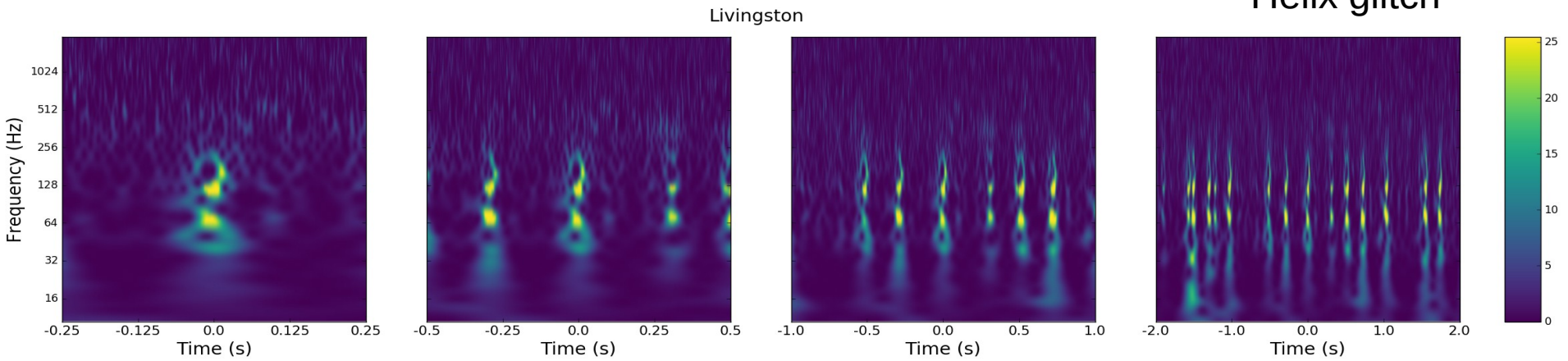
Tests on Time-Frequency plot images

Example: Blip glitches

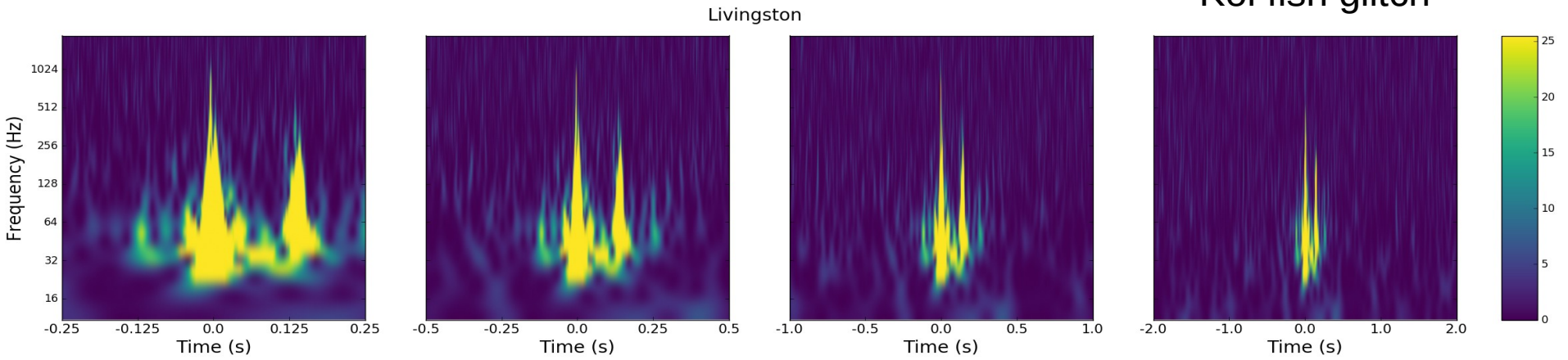


Glitches Gallery

Helix glitch

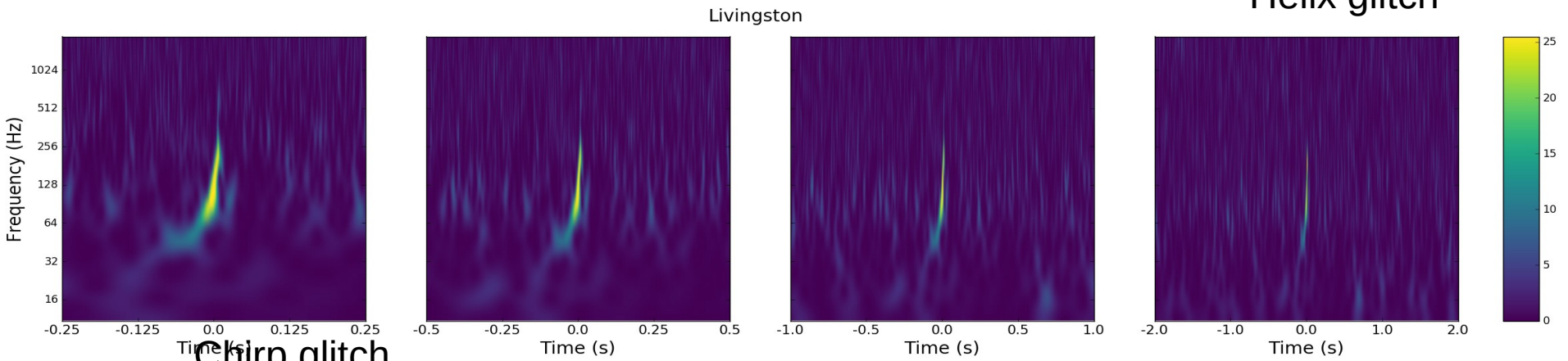


Koi fish glitch



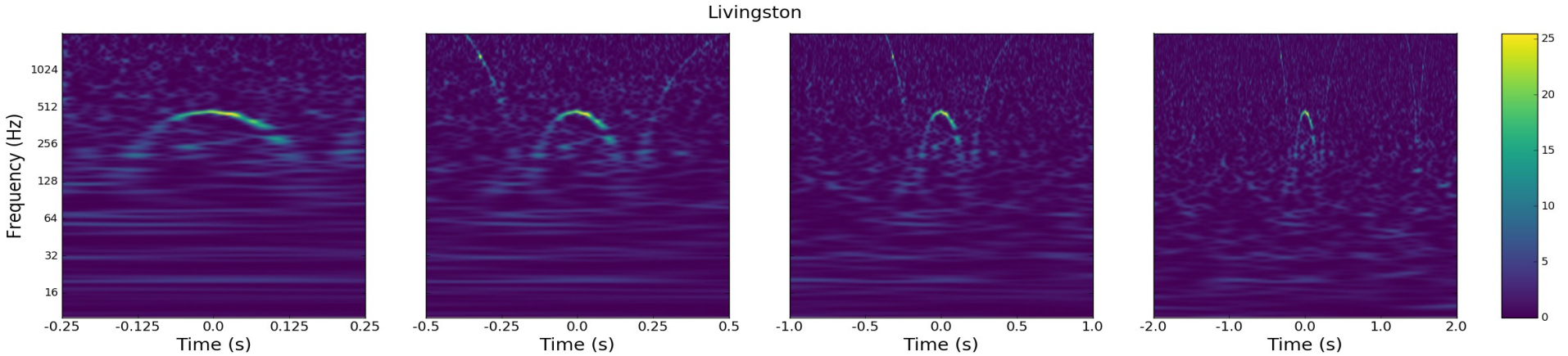
Glitches Gallery

Helix glitch



Chirp glitch

Wistle glitch



O1 data set

The O1 run spanned September 2015 through January 2016.
It produced two detections that were reported by the LIGO-Virgo collaboration.

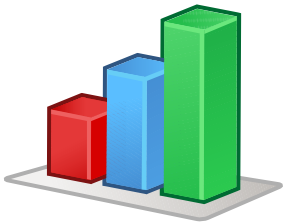
Glitch name	# in H1	# in L1
Air compressor	55	3
Blip	1495	374
Chirp	34	32
Extremely Loud	266	188
Helix	3	276
Koi fish	580	250
Light Modulation	568	5
Low_frequency_burst	184	473
Low_frequency_lines	82	371
No_Glitch	117	64
None_of_the_above	57	31

Glitch name	# in H1	# in L1
Paired doves	27	-
Power_line	274	179
Repeating blips	249	36
Scattered_light	393	66
Scratchy	95	259
Tomte	70	46
Violin_mode	179	-
Wandering_line	44	-
Whistle	2	303

Hardware details

GPU Nvidia GeForce 780GTX ti (2.8k cores, 3 Gb RAM)

- Input layer: RGB images (32x32 or 64x64)
- 4 convolutional layers + 2 pooling layers + 1 fully connected layer
- Output layer: N-sized layer of probability to belong to a class (N= number of classes)
- About 6.8 M parameters to fit
- Developed in Python + CUDA-optimized libraries
- Training phase depends on the number of images in the training datasets (here is ~hours on a desktop+GPU)
- Training is done once → then classification is very fast (~1-10ms per image)

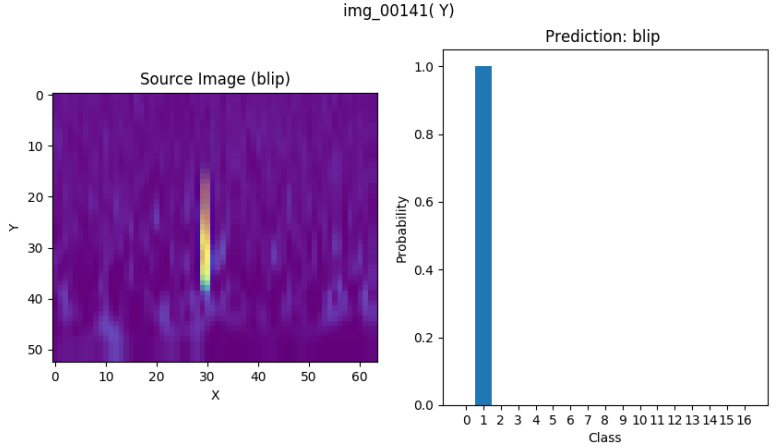


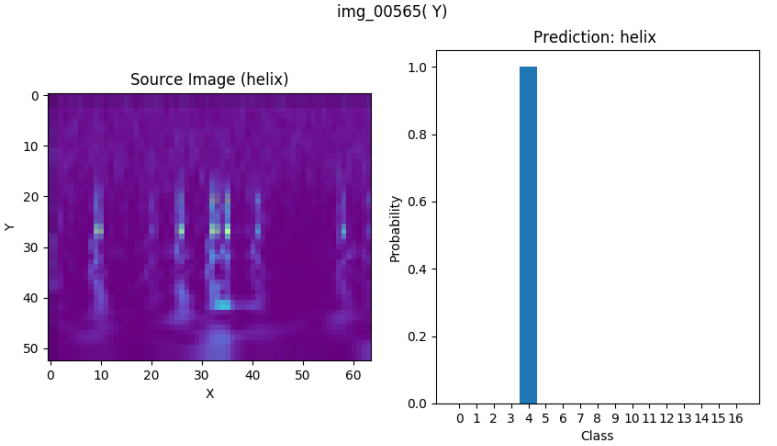
Results

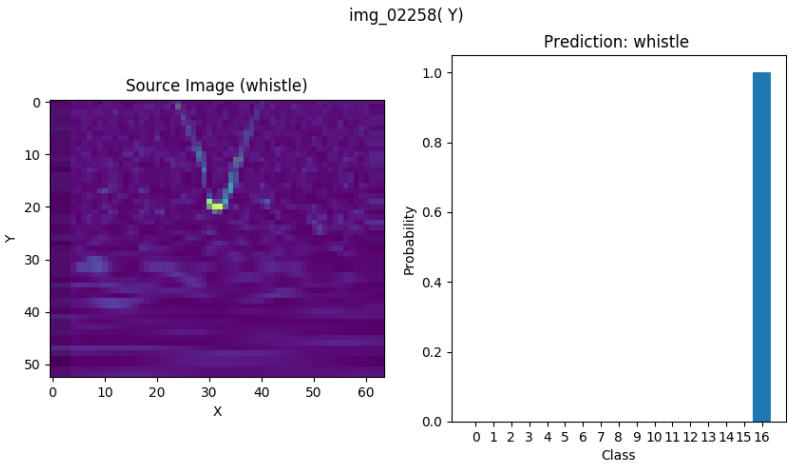
- Classes not balanced (i.e. not the same number of images per each class) --> Possible bias introduced
- It's a well known problem in DL, we cured it with 2 strategies to make balanced classes:
  - Image duplication
  - Image augmentation (i.e. duplicate images introducing small distortions)

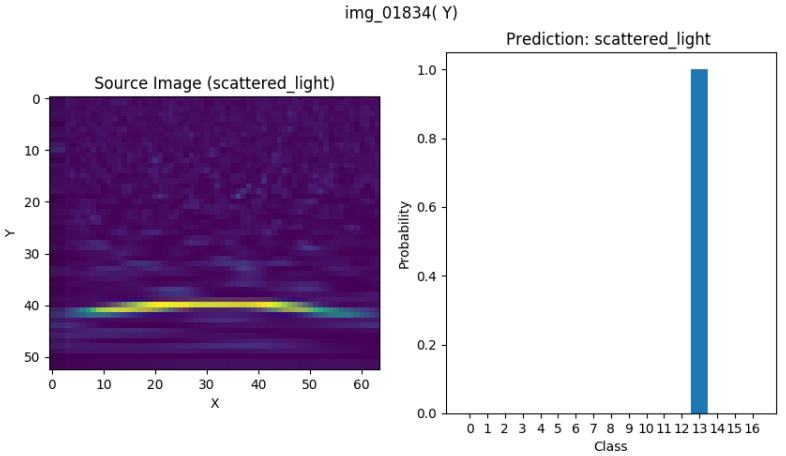
Accuracy in the order of ≈98-99% on multiclass classification

Sample results



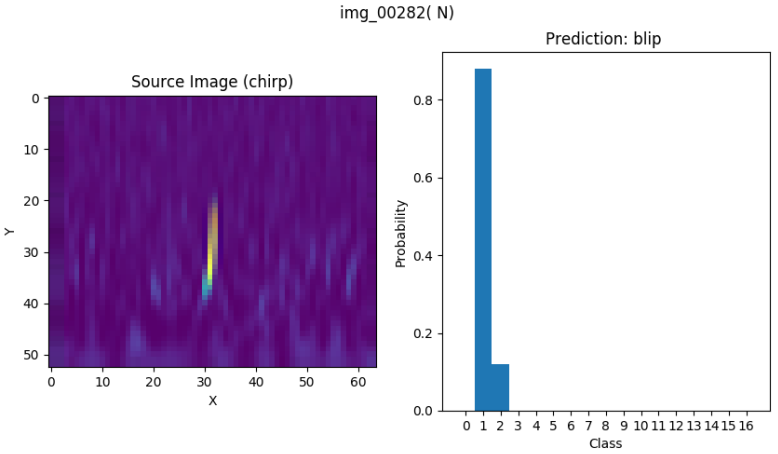


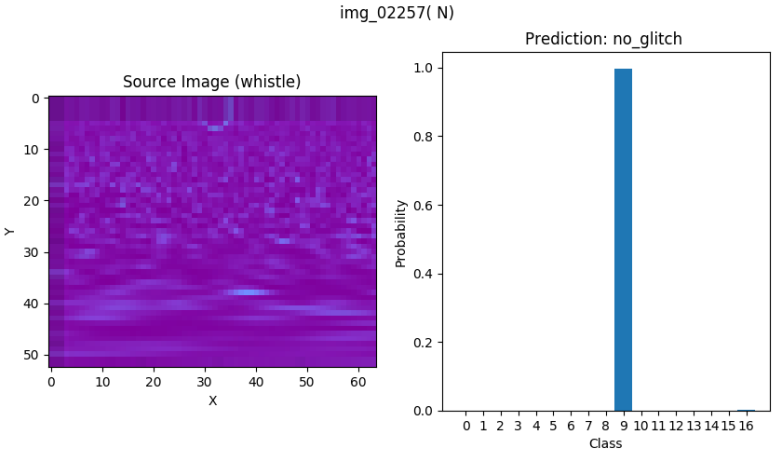




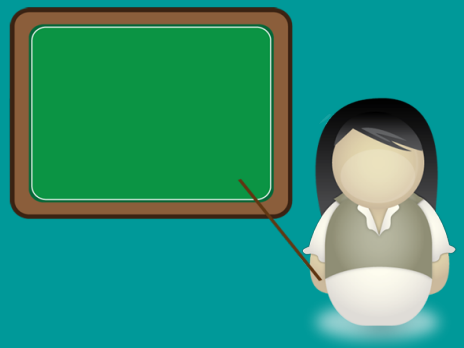
Sample results (wrong classification)

Here the problem is the zoom on the image





Here the problem is the poor contrast

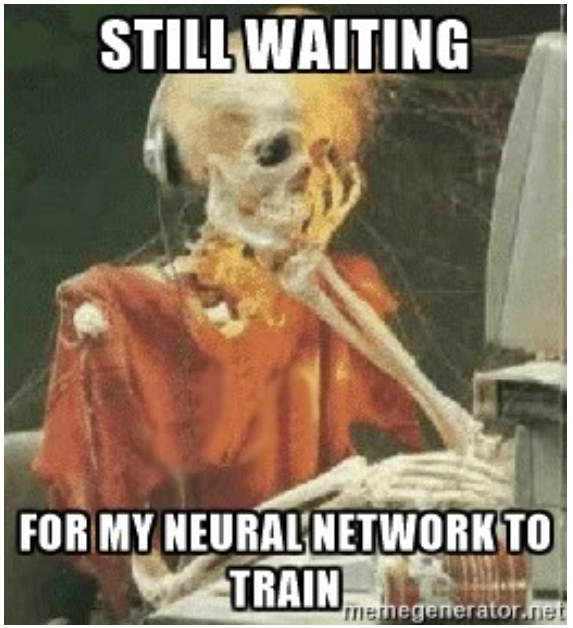


- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

Hardware and Software

Hardware

- CPUs
- Clusters
- GPUs



Software(c++, python, java, scala)

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikit learn

mxnet

theano

dmlc XGBoost

TensorFlow

cuDNN

It depends on the problem we have to face

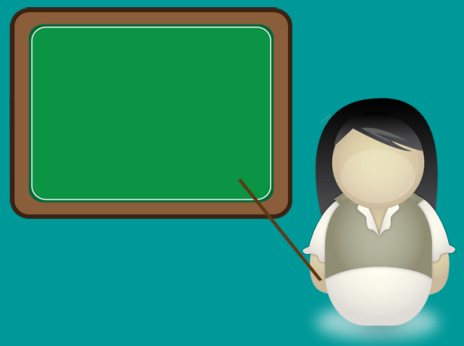
Why GPUs?

GPU-accelerated computing is the use of a graphics processing unit (GPU) together with a CPU to accelerate deep learning, analytics, and engineering application (from NVIDIA website) https://developer.nvidia.com/

- Image analysis is high computer power demanding. GPU are optimal for matrix computation
- A lot of new python library and easy to use tutorial
- @PI department

  GeForce 780GTX

- @EGO TeslaK40C 12GB (thanks to G. Attardi)



- LIGO and Virgo Data
- Our problem: Signals classification
- Machine Learning
  - Our problem: Tests and Results
- Deep Learning
  - Our problem: Tests and Results
- Software & Hardware solution
- What next

How to deal with big data

What if we have many glitches to classify?

I participate to **kaggle** competition (http://www.kaggle.com)
Personal experience : First approach to apache spark and pyspark

Data

Matrix:
1183747 X 4265
~14.3Gb

Hardware

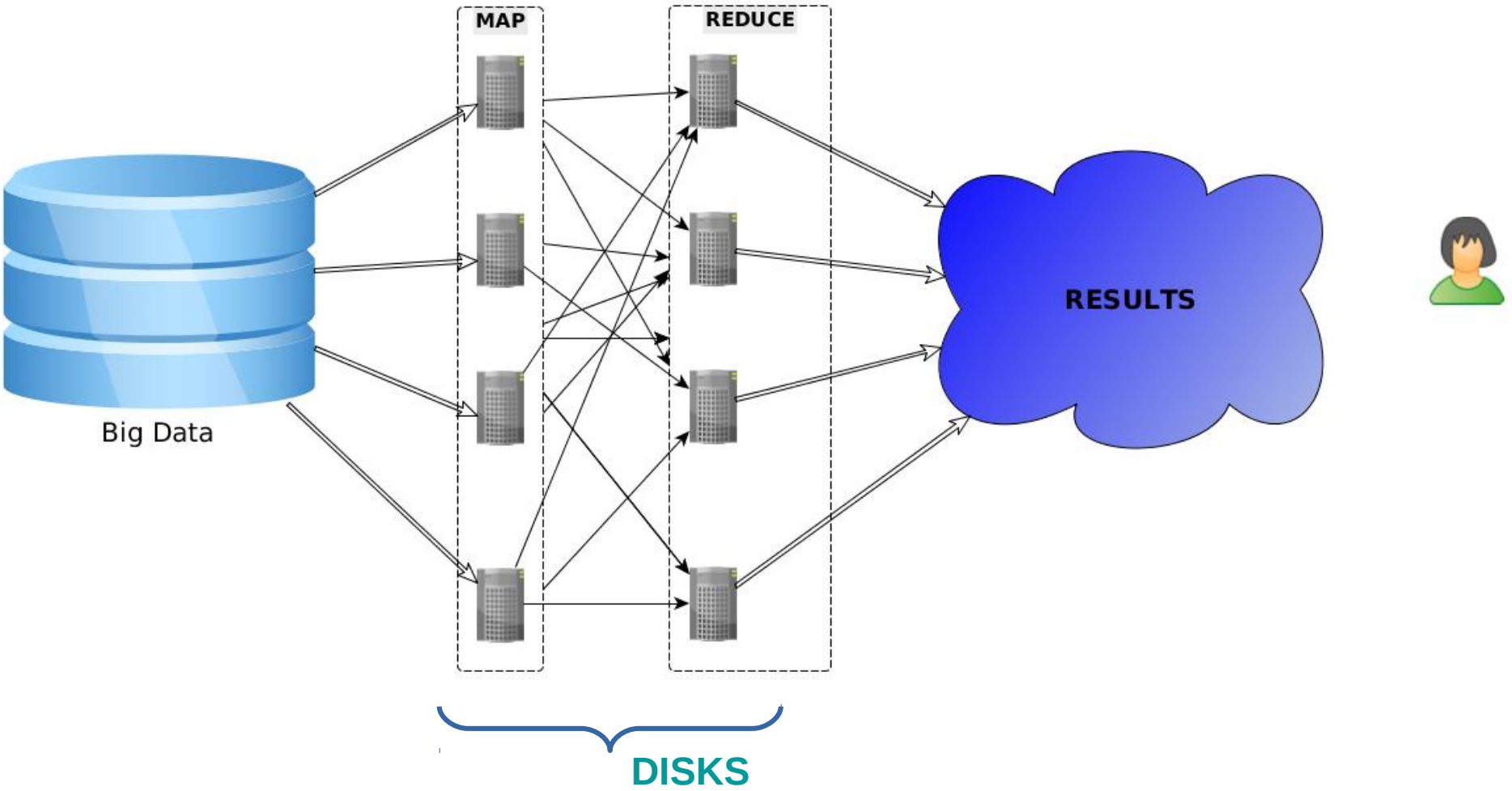
I7 core with 16Gb RAM

Software

Using Pandas: PC completely freezed
...having read about Apache Spark...I gave a trial
PySpark: Full data processed in few minutes!

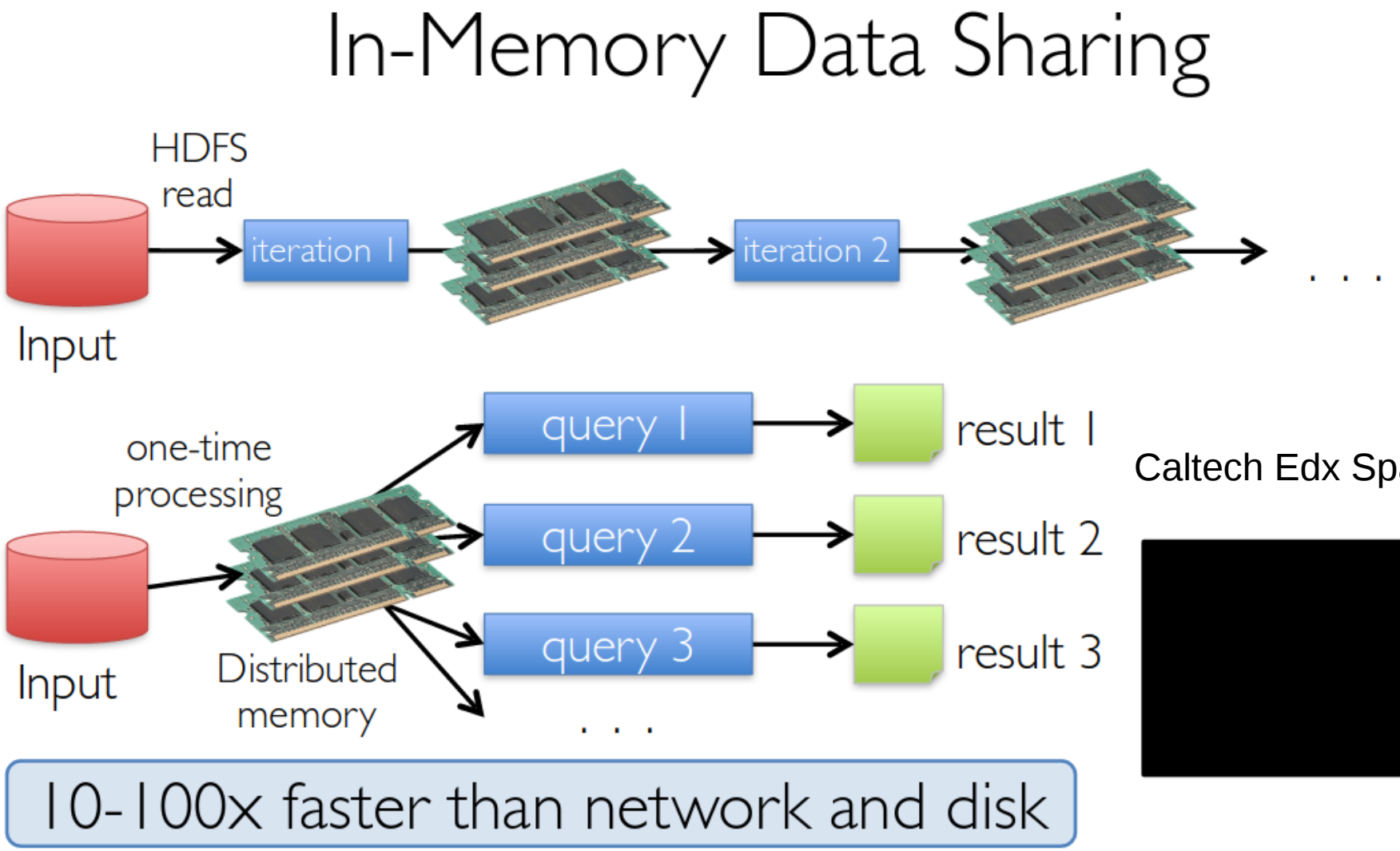
- Pandas DataFrame is not well suited for large data set, while very good for data study
- Spark DataFrame (from Spark 2.0 on) are very easy to use and very efficiently manage memory
- ML spark library contains many ML algorithm
- Spark can be used standalone mode or cluster mode

Google Map Reduce framework (Hadoop)



How to deal with Big Data?

Apache Spark: use memory instead of disk and lazy computation



Caltech Edx Spark course

Scalable, efficient analysis of Big Data

Spark Overview

- Apache Spark is a fast and general-purpose cluster computing system.
- It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs.
- It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

How to deal with Big Data?

Benchmarks

Spark wins CloudSort Benchmark as the most efficient engine
Apache Spark won the 2016 CloudSort Benchmark

First Public Cloud Petabyte Sort (2014)

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Cluster disk throughput	3150 GB/s (est.)	618 GB/s	570 GB/s
Sort Benchmark Daytona Rules	Yes	Yes	No
Network	dedicated data center, 10Gbps	virtualized (EC2) 10Gbps network	virtualized (EC2) 10Gbps network
Sort rate	**1.42 TB/min**	**4.27 TB/min**	**4.27 TB/min**
Sort rate/node	**0.67 GB/min**	**20.7 GB/min**	**22.5 GB/min**

Daytona Gray 100 TB sort benchmark record (tied for 1st place)

http://databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html

Machine Learning Algorithms (MLA) LIGO Virgo collaboration informal group

MLA LIGO Virgo collaborations

Activities & what we talk about…

- Transversal group (Detchar, Detection, Control,…)
- We started regular meetings each 2 weeks.
- F2F meeting planned at next LVC

- Genetic programming
- Images classification
- DeepLearning
- Dictionary learning
- Continous waves analysis approach

What next

- Test of these architectures (Distributed Deep Learning with Apache Spark and Keras) (*with L. Rei*)
- Setup of in-time machine for glitches classification
- ML pipeline based on spark
- Work with new labeled data set for Virgo as benchmark

Thanks for you attentions!

Copyright for the images and icons

- https://img.clipartfest.com/
- https://www.toptal.com/
- http://www.asimovinstitute.org/nreural-network-zoo
- https://www.edx.org/school/caltechx