Intelligent lossy compression for DUNE data

LDRD Project

Emma Weiler

PI : Amit Bashyal - Mentors : Peter Van Gemmeren, Zelimir Djurcic, Sheng Di

Contents

1. Context of the project
2. DUNE experiment
3. SZ3 algorithm
4. Different inputs and results :
    1. Synthetic simple waveforms
    2. Synthetic signal like waveforms
    3. ProtoDUNE data

1. Context

Lossless compression of raw data for HEP experiments is standard practice.

- no information lost

Lossy compression outside of HEP experiments showed promise in terms of compression ratio and data fidelity (atmospheric studies and climate models).

Compressed raw data could be used to **accompany** the reconstruction of the particle interactions; complimentary data , the goal is **not** to replace the data with compressed data.

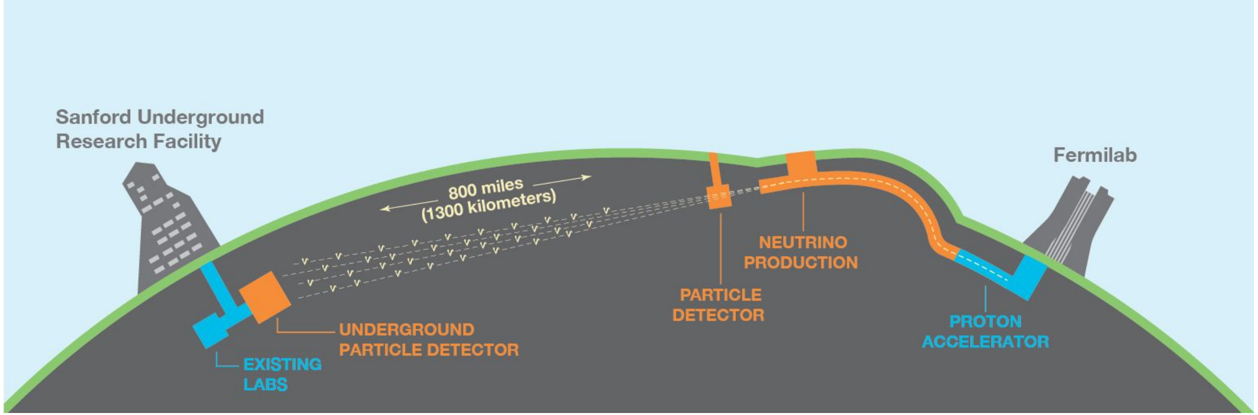
Advantages :

- When issues arise in the reconstruction, able to check against raw data;
- Practical access : instantaneous, easier to handle in different softwares;

> Assess how well the HEP experimental raw data can be compressed using intelligent lossy compression algorithms :

- MGARD; IDEALEM; **SZ3 (developer in Argonne)**
- **DUNE : ideal candidate**

2. DUNE : The experiment

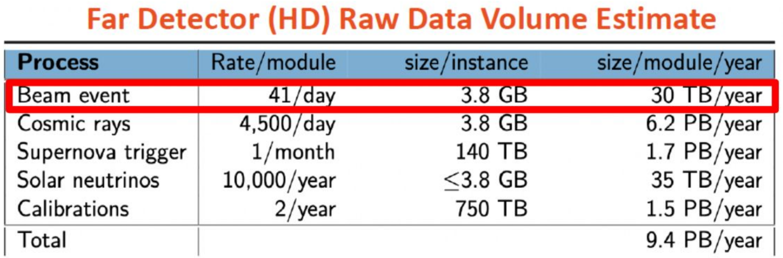


DUNE will measure neutrino oscillations for insights on neutrino mass ordering, matter antimatter asymmetry, astrophysical events.

Consists of two detectors : near and far.

- Near detector will characterize the beam of neutrinos.
- Far detector will measure the neutrino oscillations.

2. DUNE : The Far Detector



Far Detector (HD) Raw Data Volume Estimate

Process	Rate/module	size/instance	size/module/year
Beam event	41/day	3.8 GB	30 TB/year
Cosmic rays	4,500/day	3.8 GB	6.2 PB/year
Supernova trigger	1/month	140 TB	1.7 PB/year
Solar neutrinos	10,000/year	≤3.8 GB	35 TB/year
Calibrations	2/year	750 TB	1.5 PB/year
Total			9.4 PB/year

Charged particles produced by neutrinos ionize Argon > Each electronic signal is recorded by the channels

Each channel reads the electronic signal as a waveform.

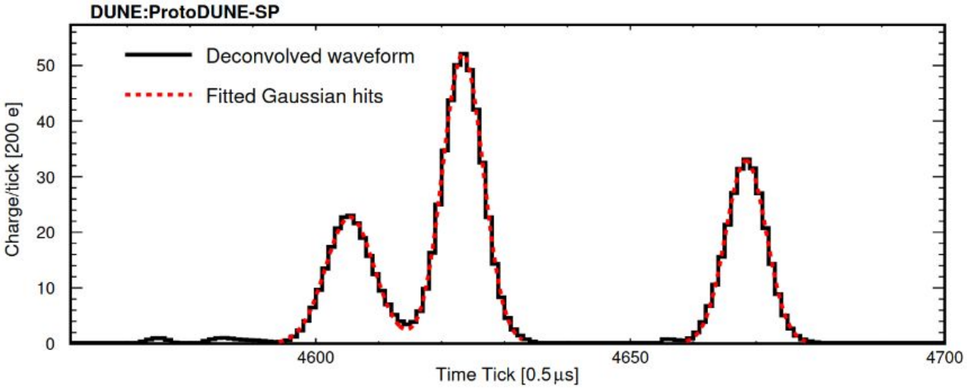
Homogeneous detector : filled with liquid Argon

4 modules each with 17 kT of Argon

Readout channels : ~ million of them

A single event will be of few GBs.

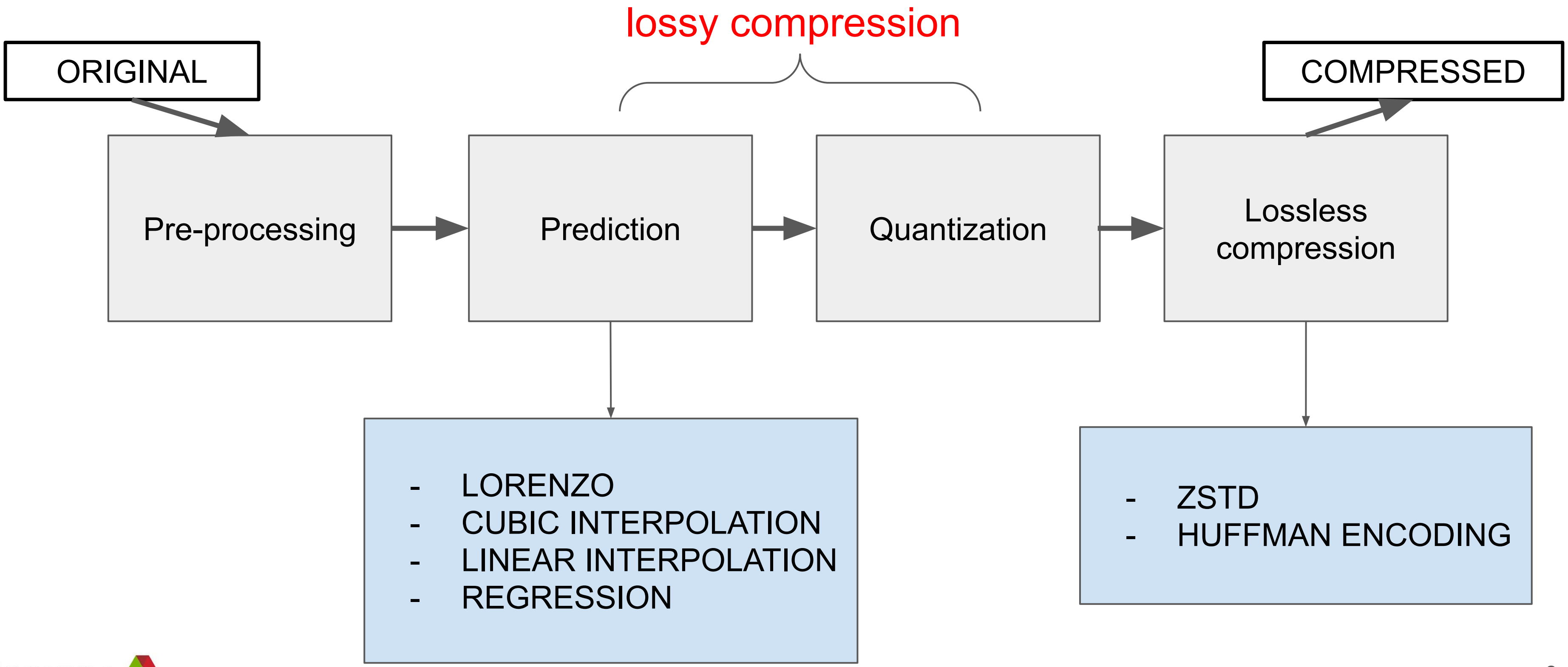
Raw data will be stored as HDF5 files.



In short, DUNE is a good candidate for the project because of heavy data and simple waveform.

3. SZ3 : Understanding the algorithm

tunable algorithm; supports compression with different kinds of error bounds and different predictors depending on the input data type. https://github.com/szcompressor/SZ3



lossy compression

ORIGINAL

Pre-processing

Prediction

Quantization

Lossless compression

COMPRESSED

- LORENZO
- CUBIC INTERPOLATION
- LINEAR INTERPOLATION
- REGRESSION

- ZSTD
- HUFFMAN ENCODING

3. SZ3 : Important metrics

Error Bounds (EB) :

quantify the deviation of compressed data from original data : compressed data will take value within a range of : original data ± error bound.

Depending on user's needs, SZ3 can support absolute (ABS), relative (REL), normal (NORM), peak signal to noise ratio (PSNR) type of error bounds.

Most of the analysis was restricted to the absolute error bound.

Compression ratio (CR):

The CR is approximately : size of original data set / size of compressed data.

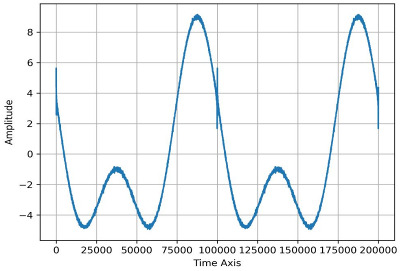
3. SZ3 : The different tested configurations

Predictors	Interpolation cubic	Interpolation linear	Lorenzo + Regression	Interpolation + Lorenzo
Absolute Error Bound				

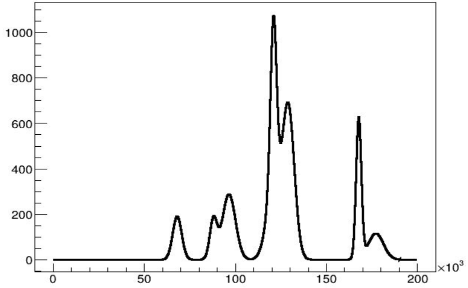
For combinations of different predictors : the algorithm “chooses” the best configuration for the whole data set.

4. Different inputs and results

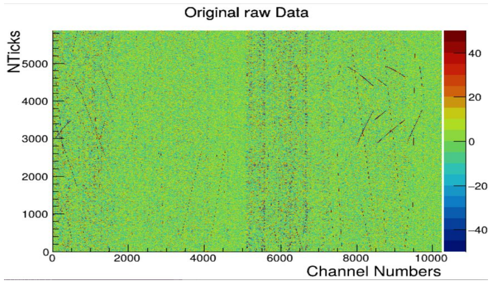
1. Synthetic data : basic periodic waveforms

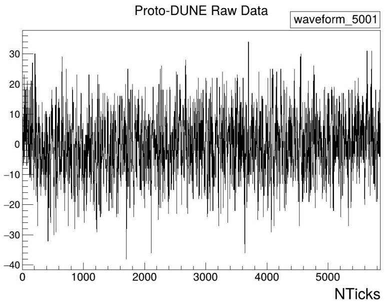


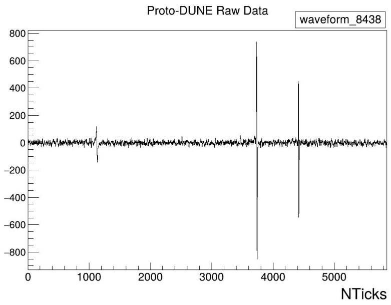
2. Synthetic data waveforms with signal like peaks



3. ProtoDUNE Data







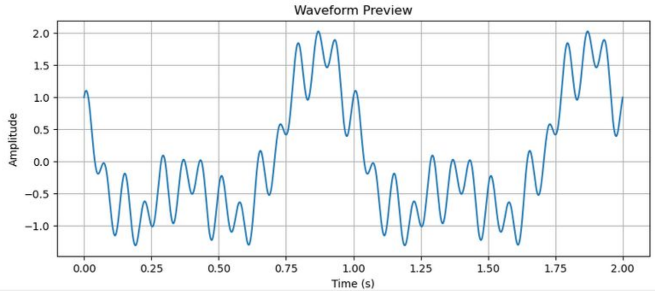
4.1. Synthetic inputs - basic waveforms

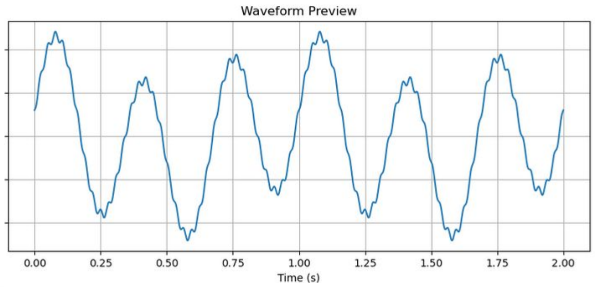
HDF5 file containing multiple waveforms.

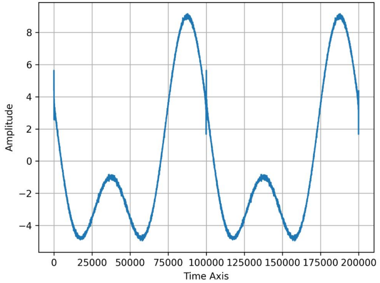
Different waveforms were compressed : noise, amplitude, number of peaks and size of the waveforms were varied.

Ran the algorithm for different error bound values, tested all different configurations.

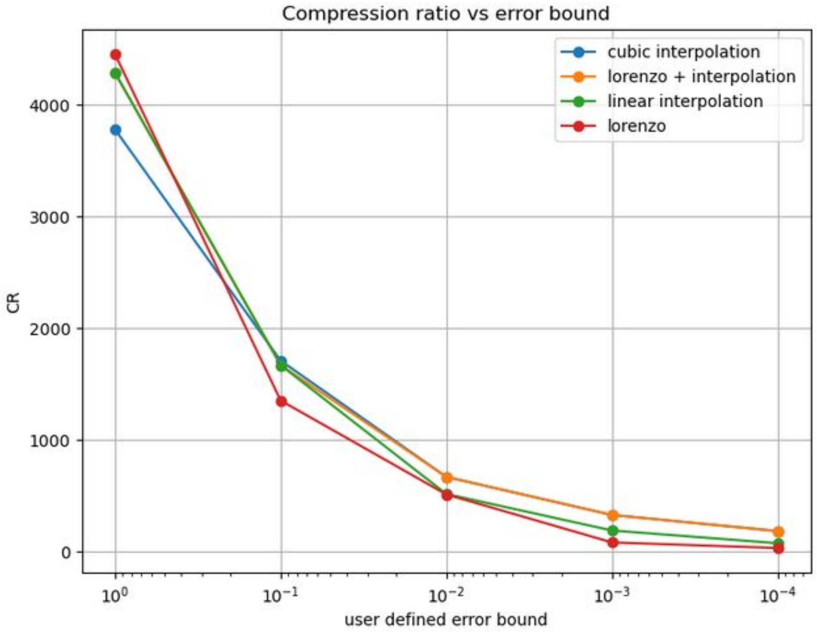
Measured the compression ratio and observed the similarity of original and decompressed data.

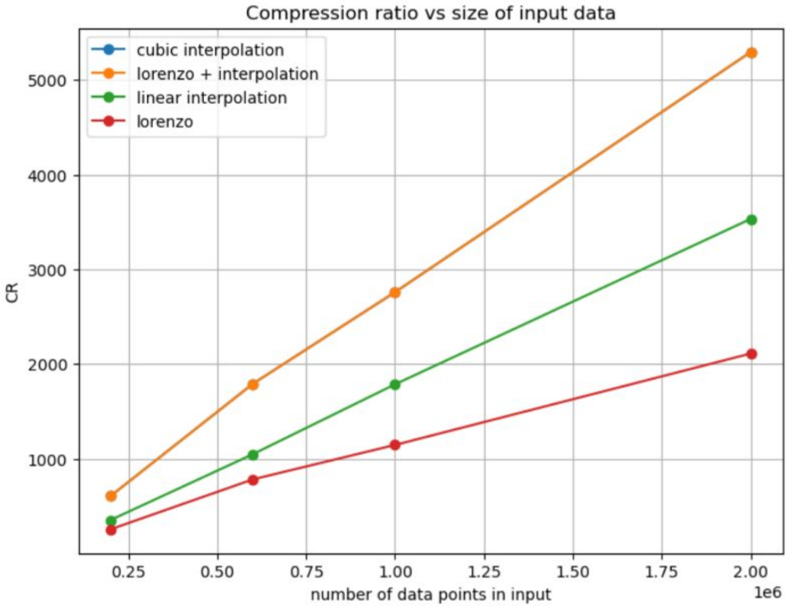






4.1 Basic waveforms - Main results

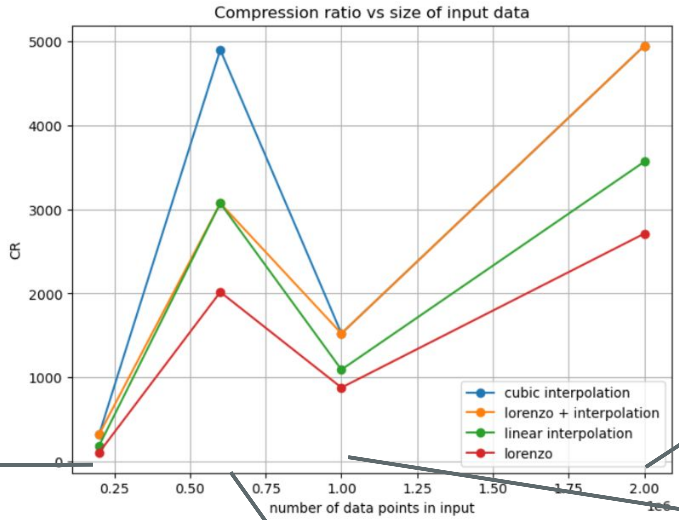


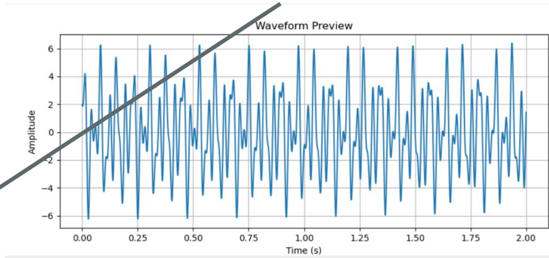


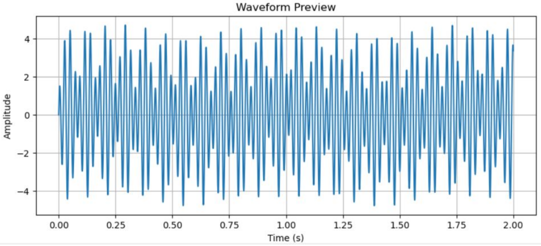
All configurations follow a similar trend :

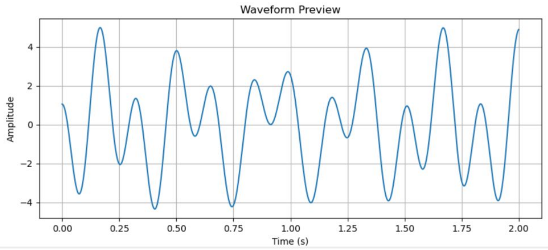
- A higher error bound yields a higher compression ratio but a larger loss of information
- A larger data size yields a higher compression ratio

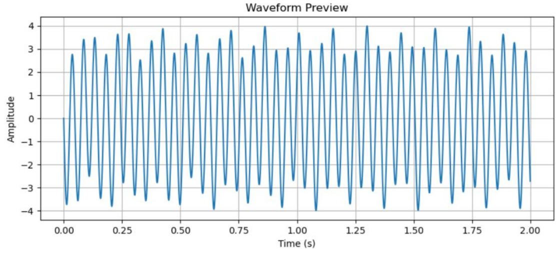
4.1. Basic waveforms - Main results











4.1. Basic waveforms - Main results

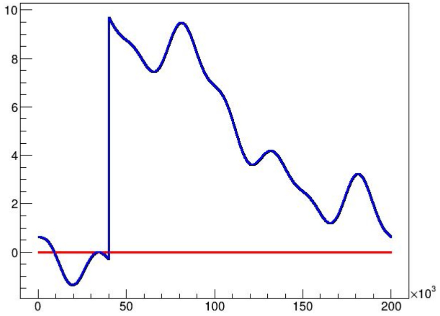
- All SZ3 configurations behave in the same way when they depend on error bounds or size of the input data.

- shape of the input is very important :
  - well defined function (less noise) > higher compression ratio

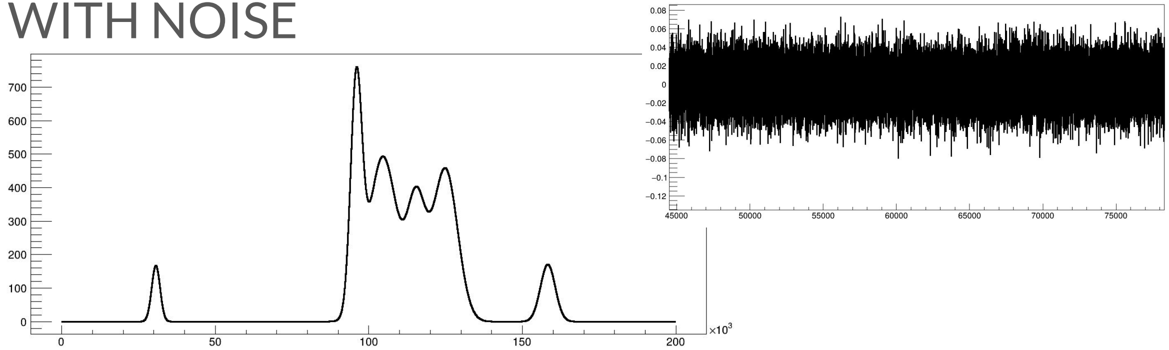
- the smaller the error bound the smaller the compression ratio (because less leeway)

4.2. Synthetic inputs - Signal like peaks

- LANDAU DISTRIBUTION



- GAUSSIAN PEAKS WITH NOISE



4.2. Signal like peaks - Main results

Landau Distribution :

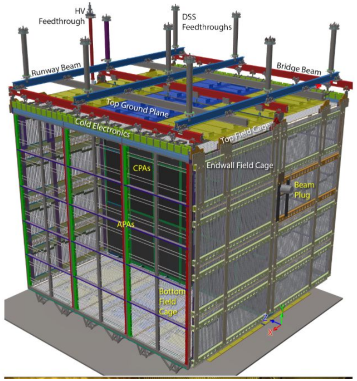
- **well defined function**
- very high compression ratio : 1700< CR < 2600
  for EB = 0.01

Gaussian peaks:

- **noise/jitter**
- low(er) compression ratio :
  20<CR< 30
  for EB = 0.01

>still no combination that provides higher compression ratio with higher data fidelity

4.3. ProtoDUNE



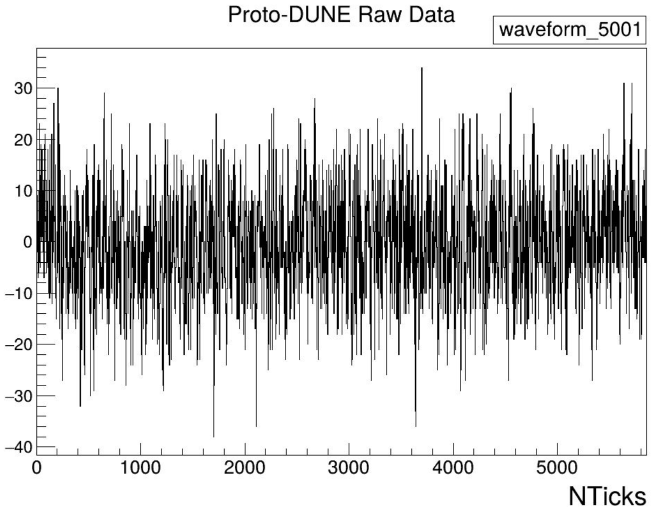
Argon filled cryostat detector in CERN:

- demonstrator experiment for DUNE far detector:
    - Cryostat made of 4 Anode Plane Assemblies (APA) each with 2560 channels.

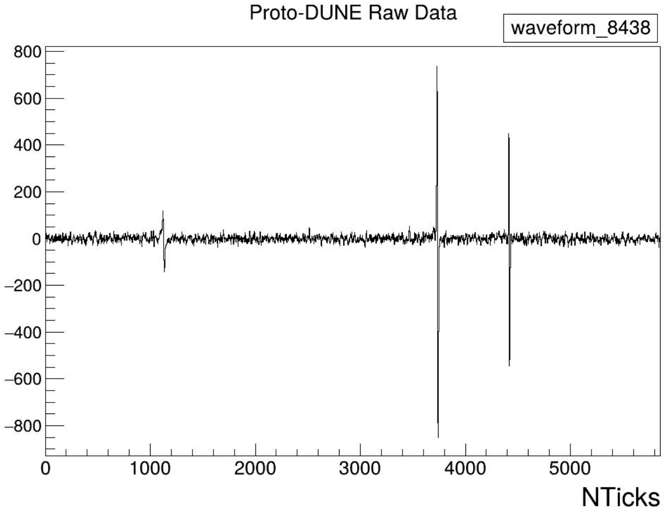
- main differences with expected DUNE data :
    - less channels than DUNE, size of the files is smaller
    - different noise : electronics and cosmic radiation

- collecting data since 2021 (B. Abi et al 2020 : https://arxiv.org/pdf/2007.06722)

4.3. ProtoDUNE - Experimental Inputs : channel readouts



channel with no signal



channel with signals

4.3. ProtoDUNE - Results for channel readouts: 1D waveforms

-> Algorithm performs better for waveforms with no signal peaks (just noise).

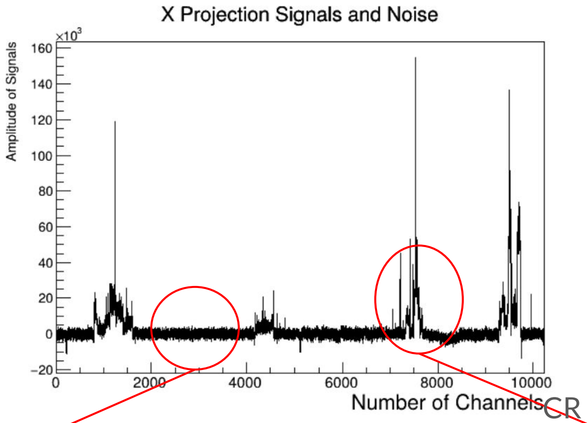
-> Ranges of CR for different error bounds :

- 20 < CR < 30 for EB = 4 (4 is 10% of observed extreme value for waveform with no signal)
- 30 < CR < 40 for EB = 8 (20%)
- 60 < CR < 89 for EB = 20 (50%)
- 90 < CR < 285 for EB = 40 (100% ; not realistic value but to see how high the CR could be)

4.3. ProtoDUNE - Compression Ratio for each channel

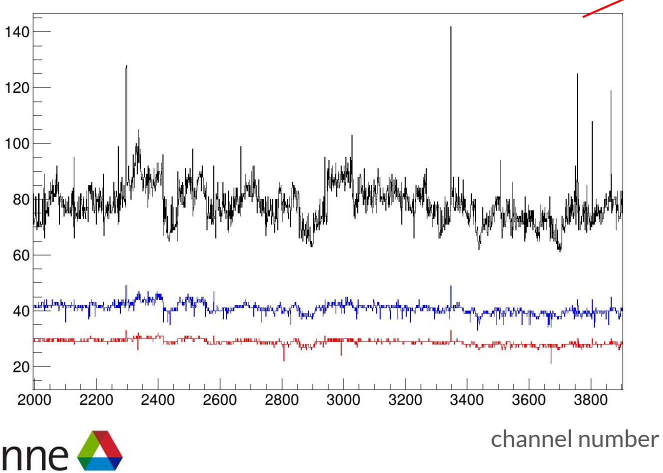
Average compression ratio:
sum of CR of all channels / number of channels

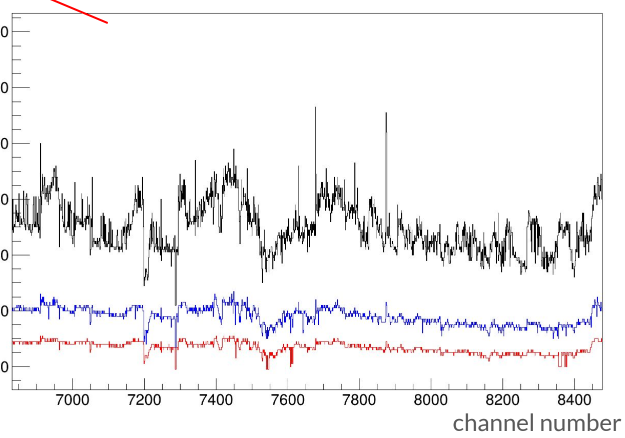
10% → CR = ~27
20% → CR = ~38
50% →CR = ~72



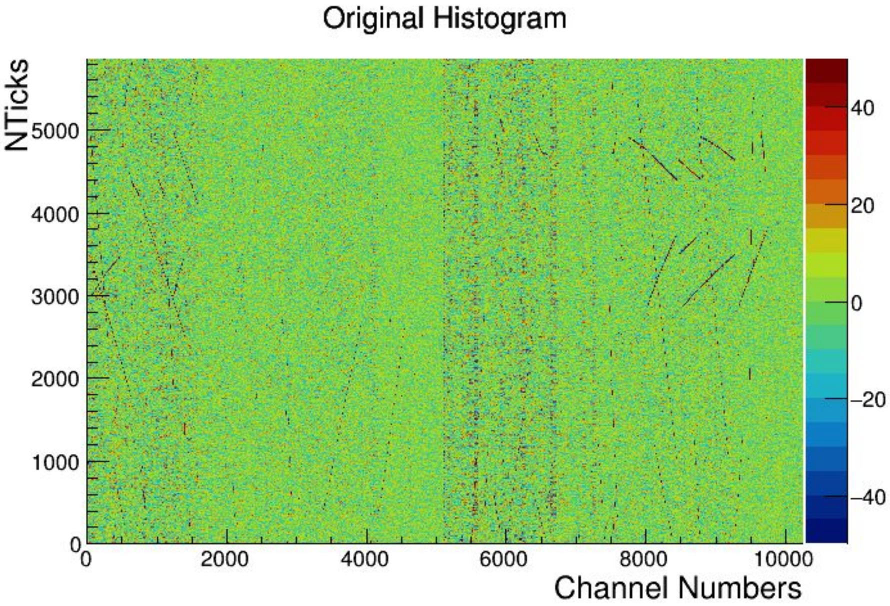
Lorenzo + Interpolation

- eb 50 %
- eb 20%
- eb 10%





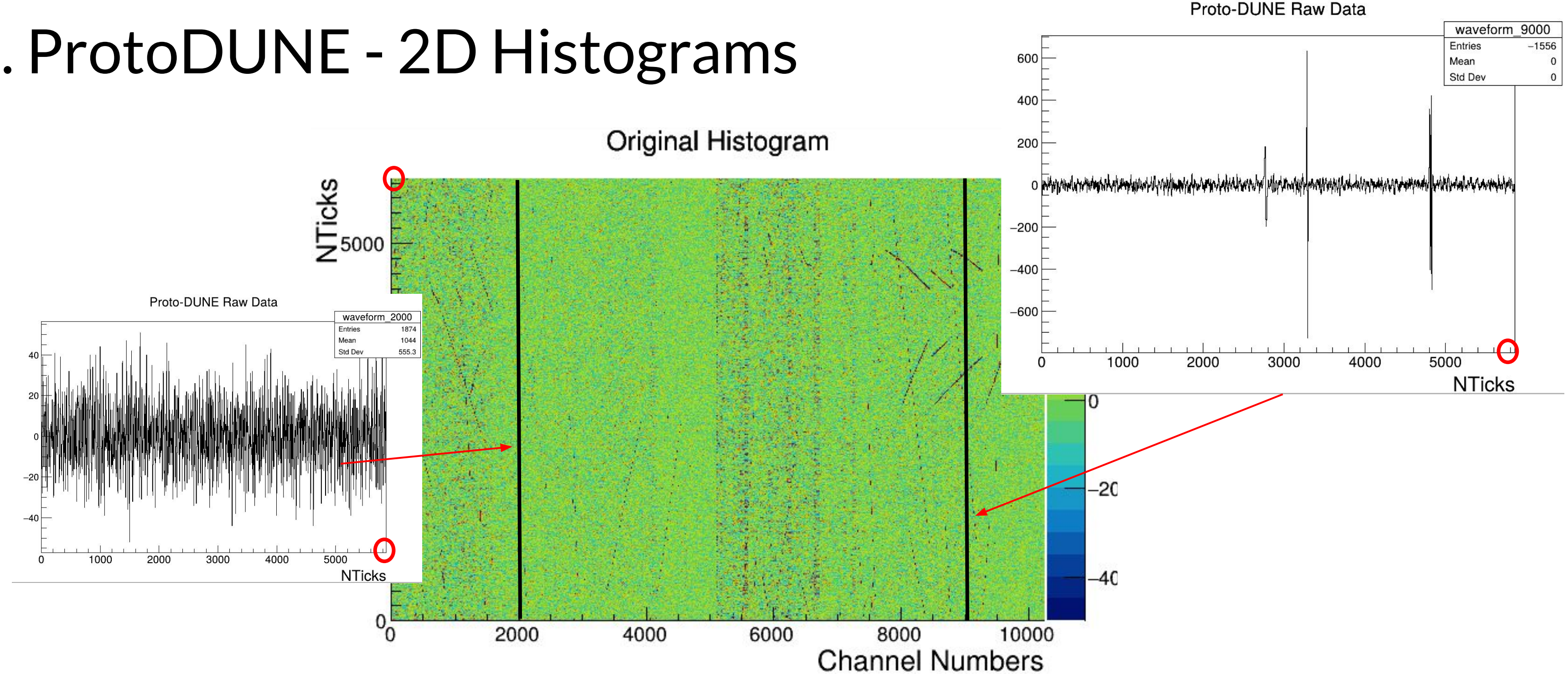
4.3. ProtoDUNE - 2D Histograms



number of ticks is the size of the 1D waveform, it is proportional to the total recorded time.

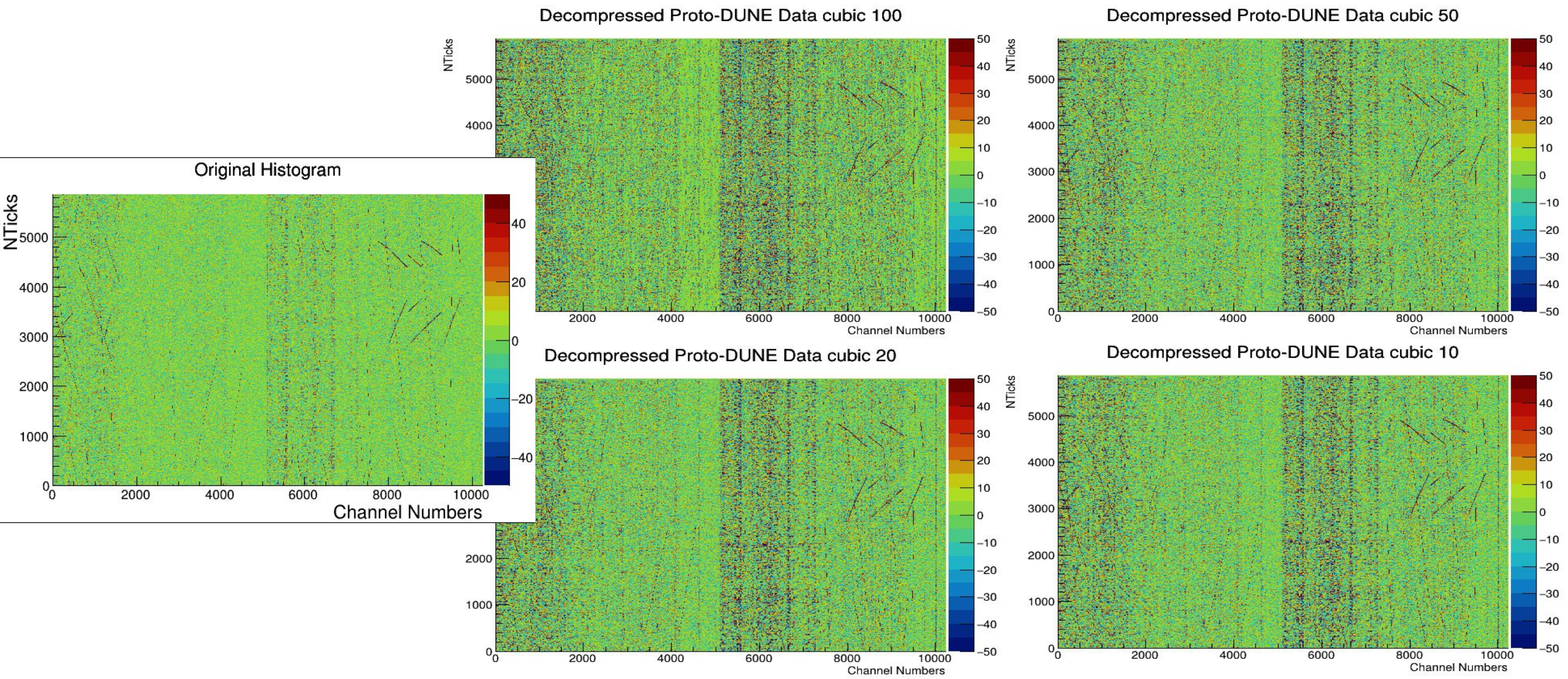
From Barnali Chowdhury

4.3. ProtoDUNE - 2D Histograms



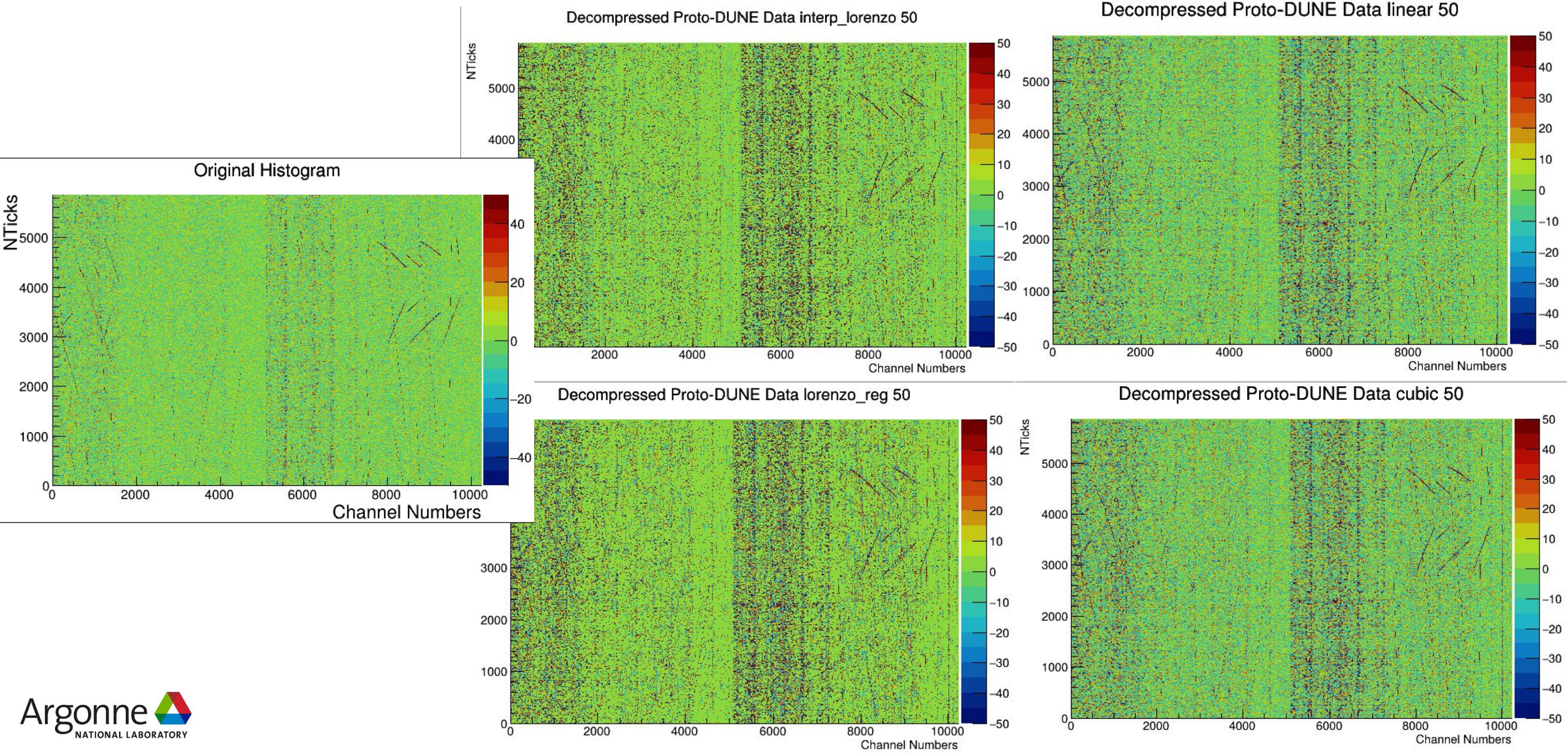
number of ticks is the size of the 1D waveform, it is proportional to the total recorded time.

4.3. ProtoDUNE - 2D histograms results



-> The compression seems to introduce some noise which needs to be investigated.

4.3. ProtoDUNE - 2D histograms results



Argonne NATIONAL LABORATORY

4.3. ProtoDUNE - compressing all channels at once

All data points are stored in a 1D waveform instead of individual channels (one channel after the other).

When compressing this long 1D waveform CR is different :

10% → CR=30
20% → CR = 50
50% → CR = 115

The CR in this case are higher than the average of CR for individual waveform compression.

Expected results -> For heavier input (long 1D vector) the CR is higher.

In the future could consider compressing 2D data (maybe using different algorithms)

4.3. ProtoDUNE - Main results

-> Main features like the signal peaks are conserved.

-> Compression ratio is significant : even with small error bounds of 10%, 20 <CR <30.

-> No big difference observed in the histograms.

-> Lorenzo prediction gives highest CR (but at what cost ?)

ProtoDUNE to DUNE :

In DUNE, we are expecting slightly different data :

Heavier waveforms (longer acquisition time) and less noise (no cosmic radiation and better resolution)

According to the observations made along the project :

- SZ3 gives higher CR for heavier distributions
- SZ3 gives higher CR for better defined functions (less noise)

We would expect SZ3 to perform even better with DUNE Data and give even more important CR.

Thank you !

APPENDIX

A. ERROR BOUNDS : DIFFERENT TYPES
B. ERROR BOUNDS : WHEN TOO IMPORTANT, WHEN TOO SMALL
C. LORENZO AND INTERPOLATION : EXPLANATION
D. LORENZO AND INTERPOLATION : BEHAVIOUR FOR PROTODUNE DATA
E. COMPRESSION RATIO PER CHANNEL; COMPLETE GRAPH

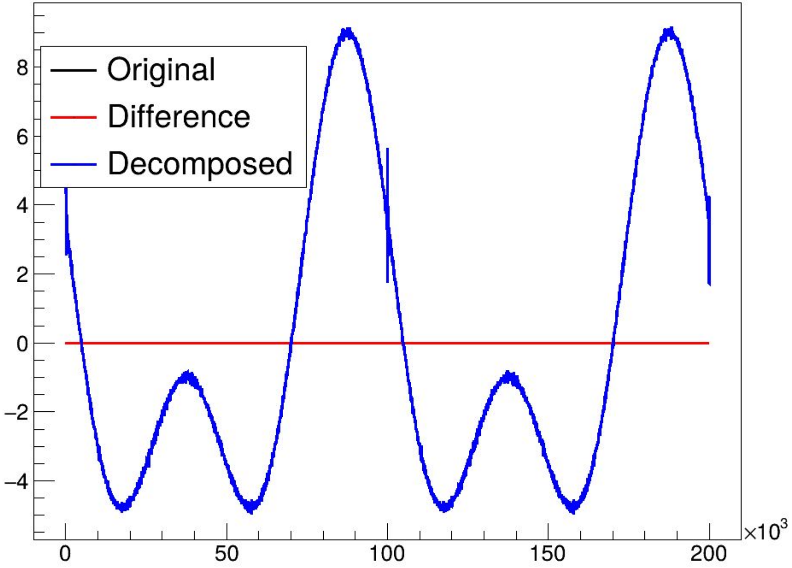
A . Error bounds

Different error bound types supported by SZ3 :

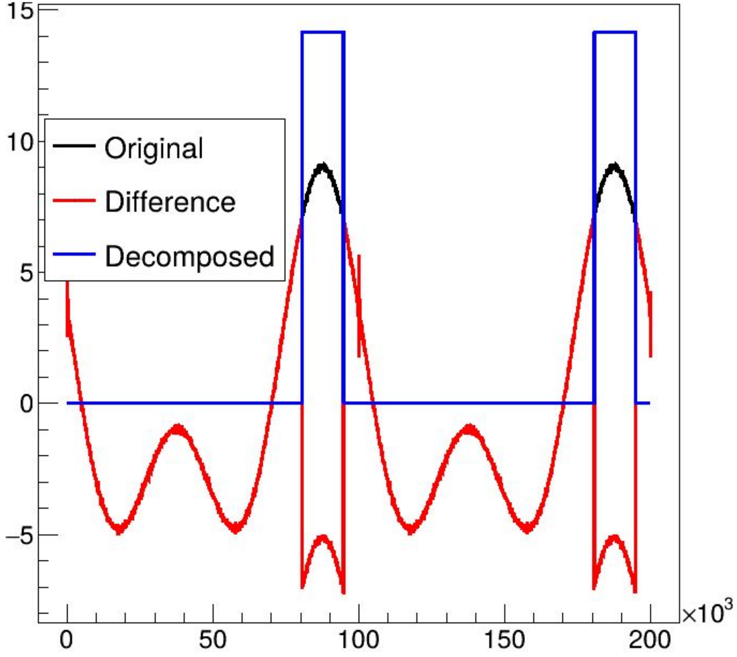
- absolute (ABS): $x - x'$
- relative (REL): $\frac{x-x'}{x} \cdot Range\ of\ data$
- normal (NORM): L2 = $\sqrt{\sum_{i=1}^{N}(y_i - y_i^{pred})^2}$
- peak signal to noise ratio (PSNR) $PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$

B . ERROR BOUNDS : WHEN TOO IMPORTANT, WHEN TOO SMALL

Extreme case EB = 0, CR = 1



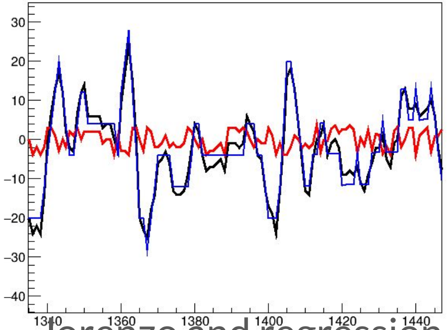
Lorenzo and regression , EB =14.14, CR = 7655



C. LORENZO AND INTERPOLATION : EXPLANATION

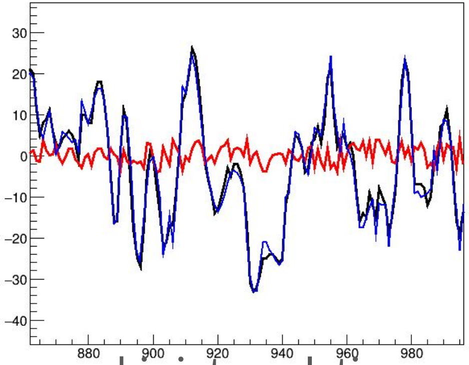
Lorenzo : prediction based on the difference of previous data points.

Start from one data point, tries to find the next data point in quantised levels of the error bound.

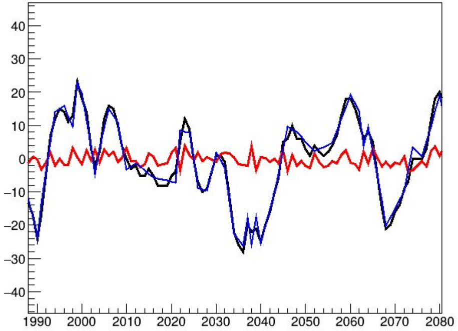


lorenzo and regression

Interpolation : prediction based on all data points (more like a fitting method).



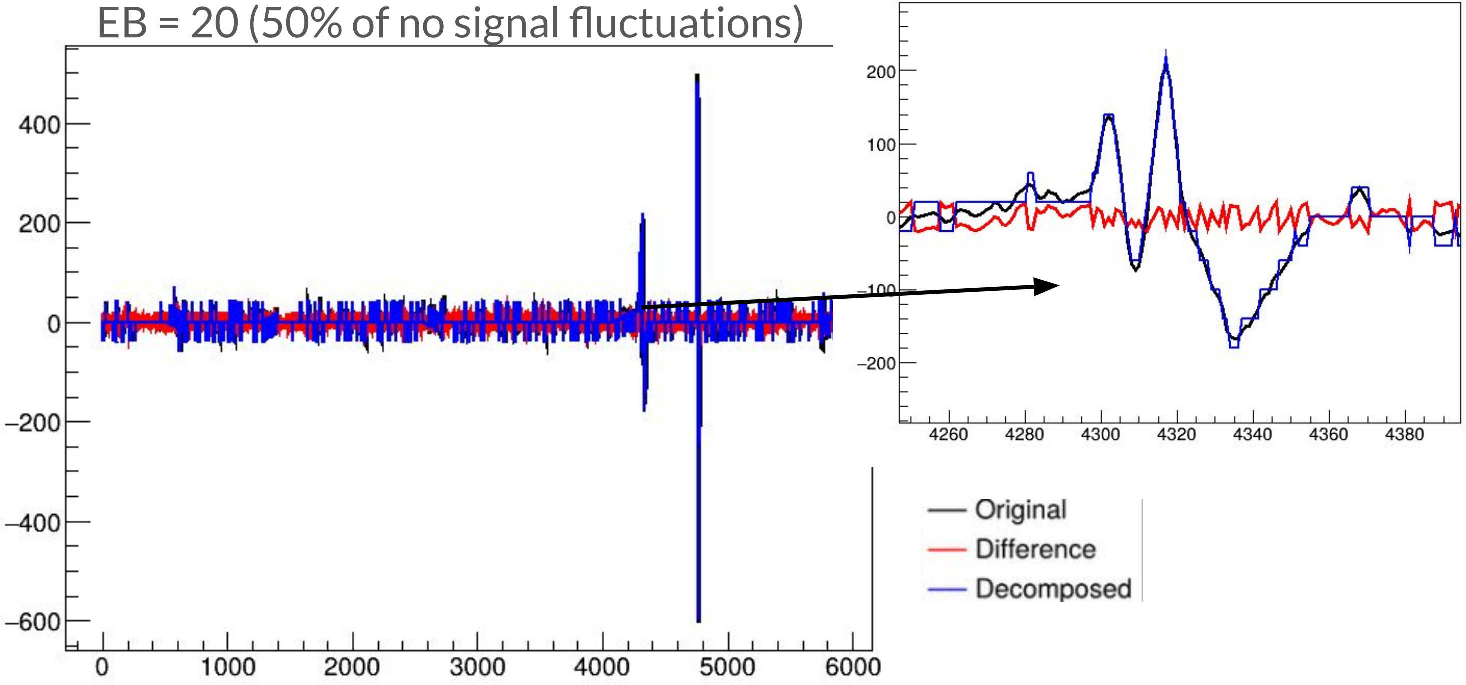
cubic interpolation



linear interpolation

D. LORENZO AND INTERPOLATION : BEHAVIOUR FOR PROTODUNE DATA

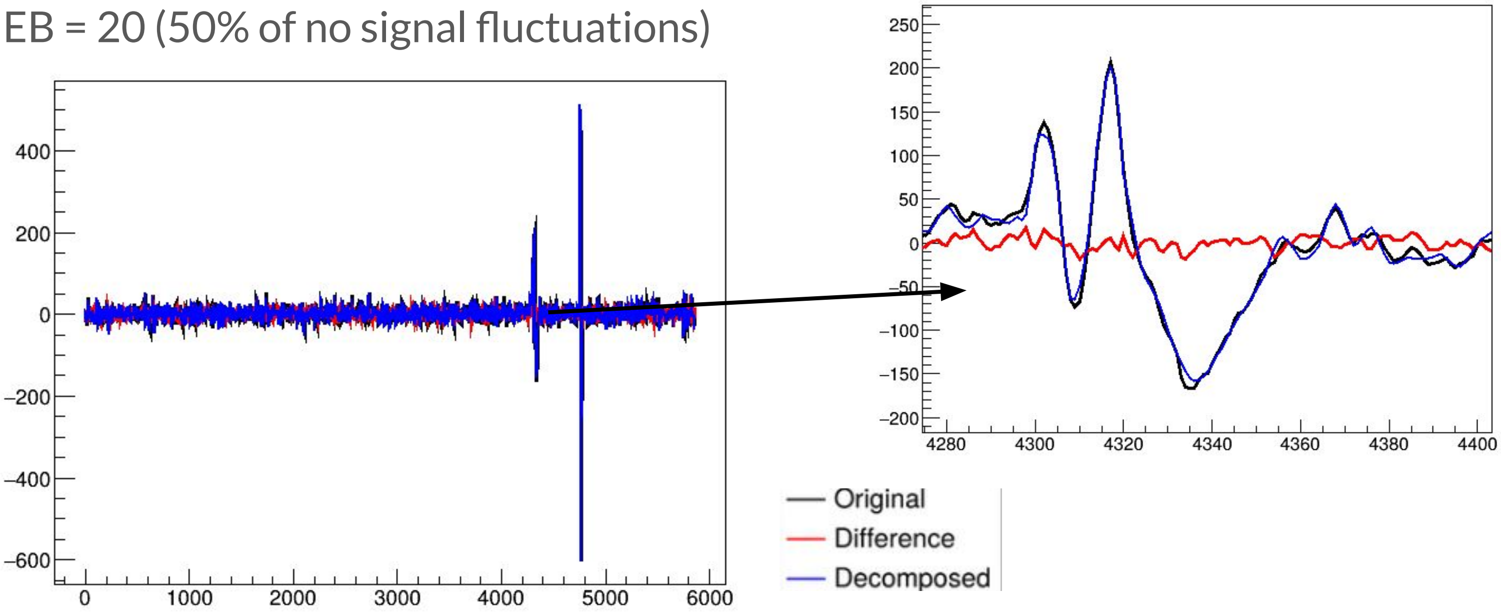
EB = 20 (50% of no signal fluctuations)



Lorenzo predictions

D. LORENZO AND INTERPOLATION : BEHAVIOUR FOR PROTODUNE DATA

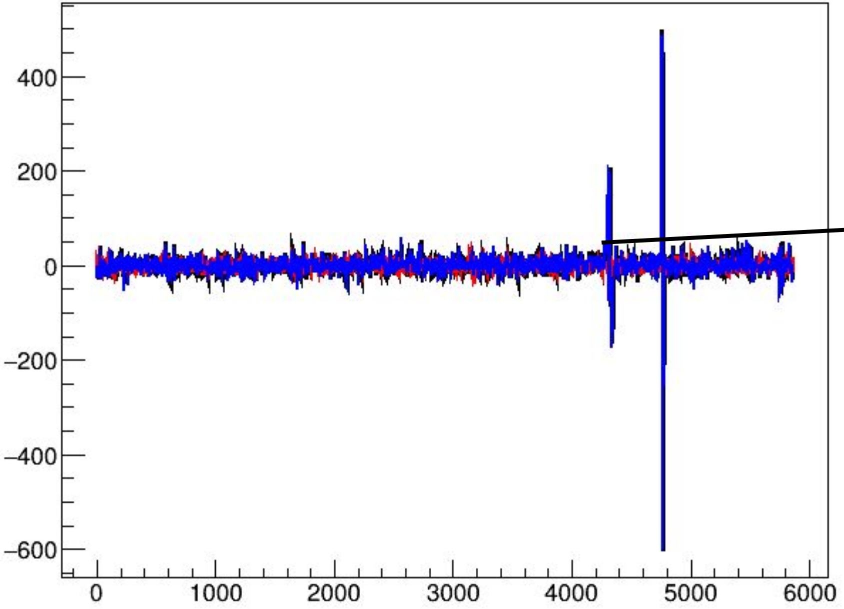
EB = 20 (50% of no signal fluctuations)

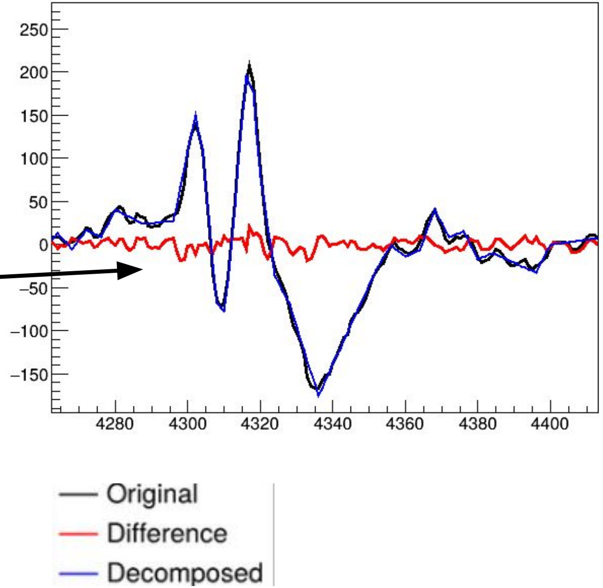


Cubic interpolation

D. LORENZO AND INTERPOLATION : BEHAVIOUR FOR PROTODUNE DATA

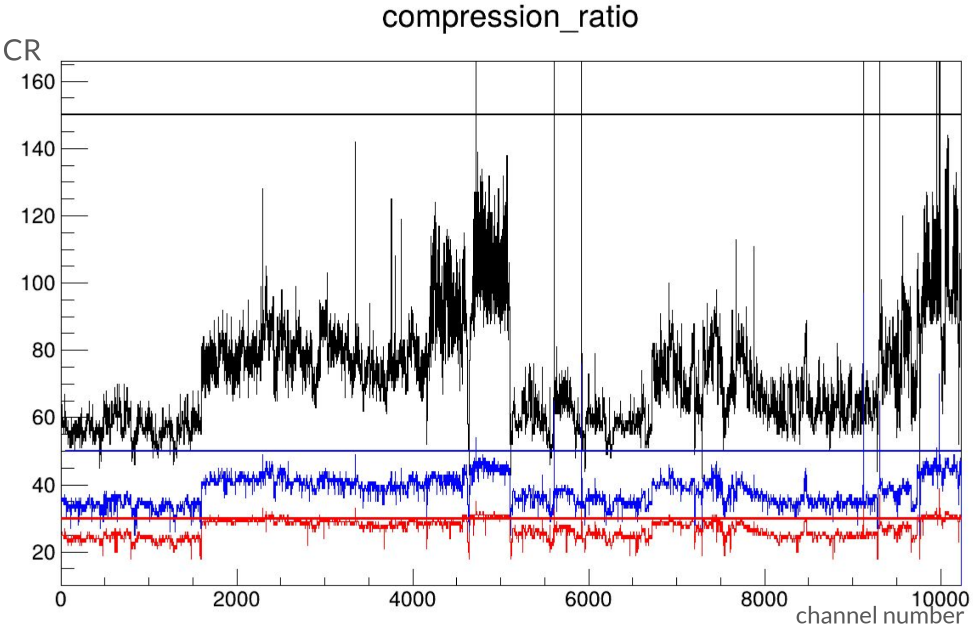
EB = 20 (50% of no signal fluctuations)





Linear interpolation

E. COMPRESSION RATIO PER CHANNEL; COMPLETE GRAPH


compression_ratio
CR
channel number