Data life-cycle in particle physics

Alessandro Razeto – LNGS

Laboratori Nazionali del Gran Sasso - 30/11/23

Physics vs Metaphysics

Physics focus on measurable quantities.

Large part of metaphysics by the Scholasticists is now modeled by physics

Particle physics

Cosmogony

- **The matter is the result of the interaction of elementary particles**
- **Discovering the laws modeling the behavior of elementary particles**
- **The applications are all around us**
  - Semiconductors
  - Medicine
  - Cosmology
- **The models are point-to-point**
  - Not necessarily providing the tools to describe complex bodies

Theory vs reality

- **Theoretical physicists provide plenty of models**
  - Or models with plenty of degree of freedom
- **Experimentalists have to pin the reality to the right model**
  - Building experiments to probe the reality
- **Experiments are getting bigger and bigger**
  - More and more expensive and complicated

Water - Earth – Fire – Air

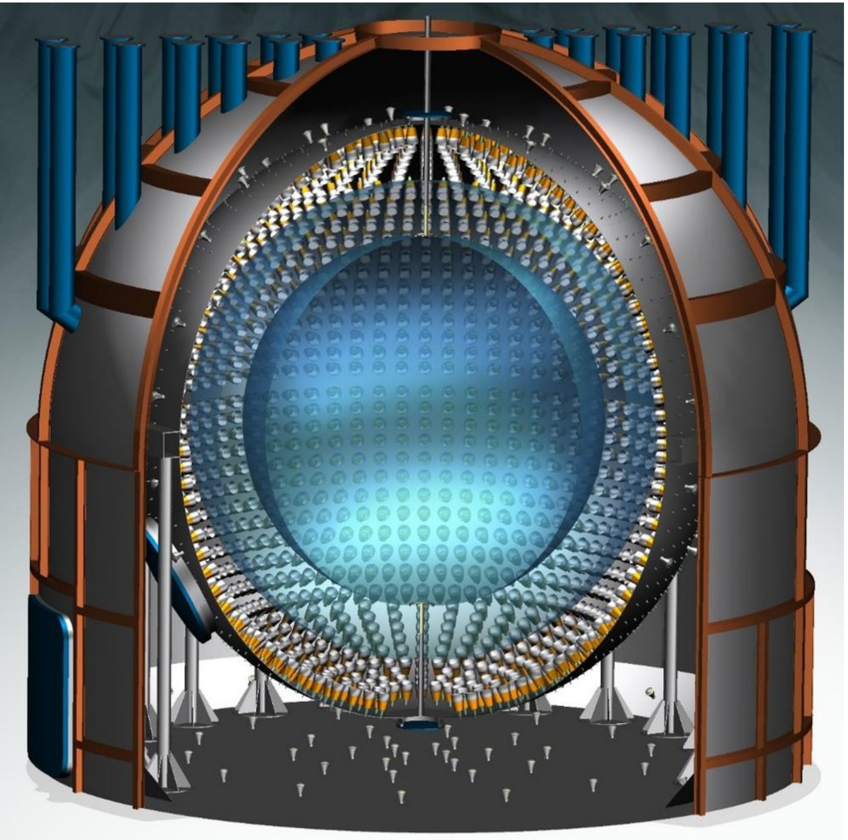
- **Elementary particles interact with matter producing measurable quantities**
  - Electrical Charge
  - Change of Temperature
  - Light Emission
- **Sensors are used to detect these quantities → into an electrical signal**
  - Photo-multiplier tube for light
  - Thermistors for temperature

Detectors

- **The detectors are designed to maximize the signal**
- **Using elements having special properties to detect particles/radiation**
  - Scintillators are materials that emit faint light pulses during interaction with particles
- **Using active and passive shielding to reject unwanted events**

LNGS Experiments

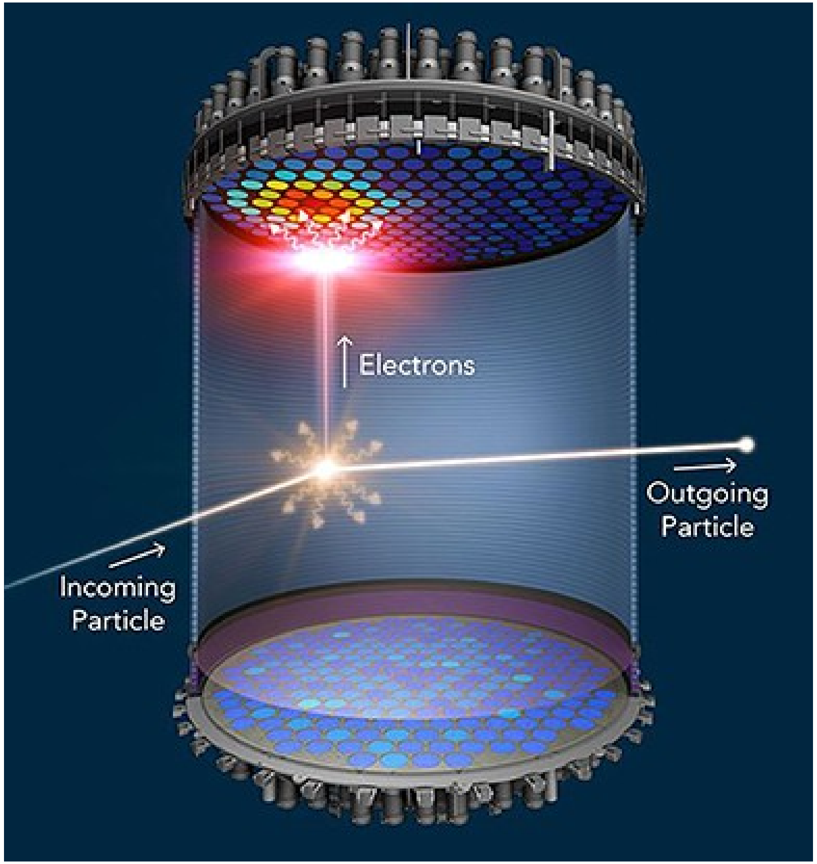
BOREXino since 2004



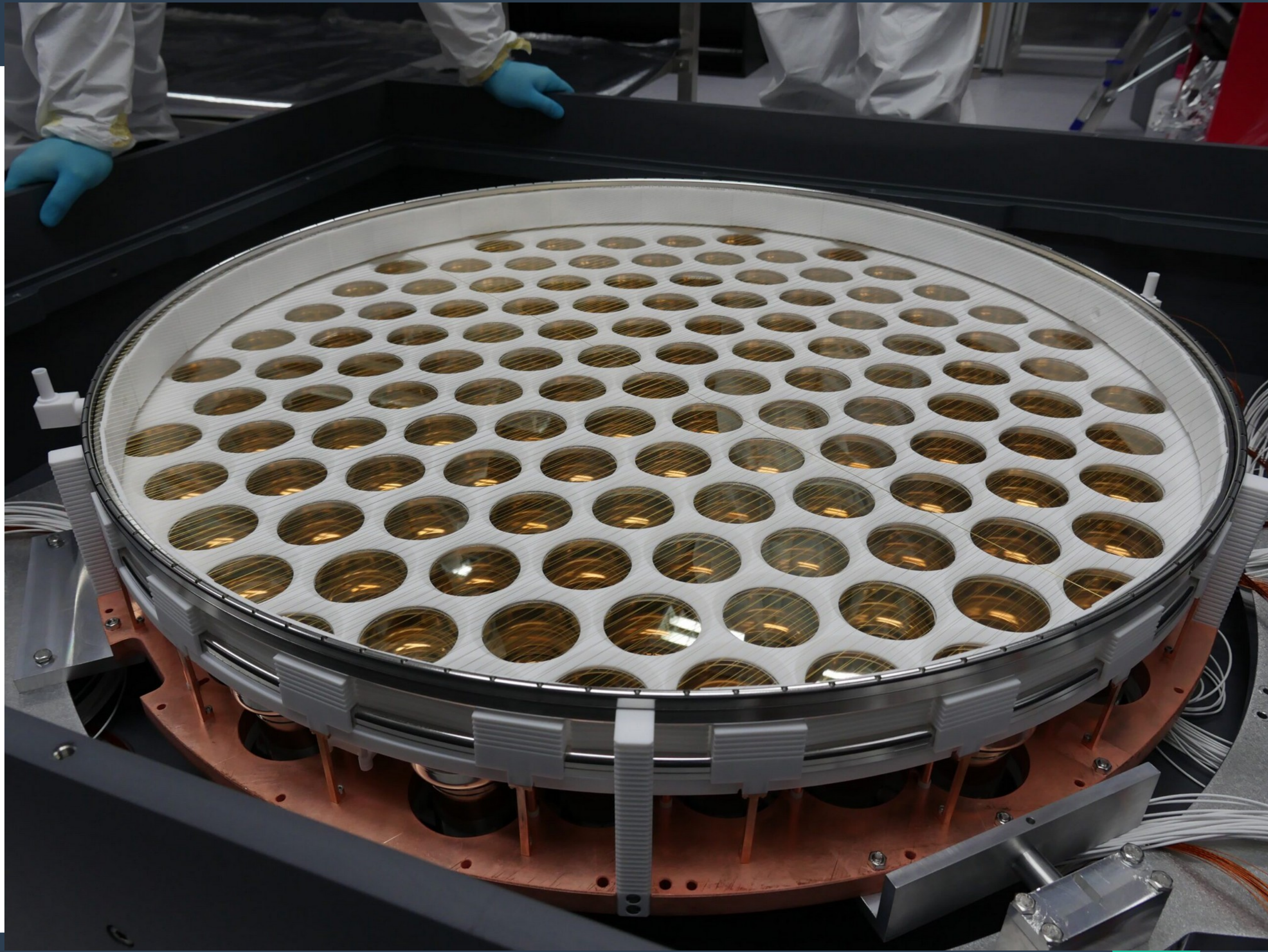
- **Detect neutrinos from the Sun**
- **Low background design**
  - Radio-pure materials
  - Active and passive shielding
- **Reached unprecedented contamination levels**
- **300 tons of liquid scintillators + 2000 photomultiplier tubes**

Borexino

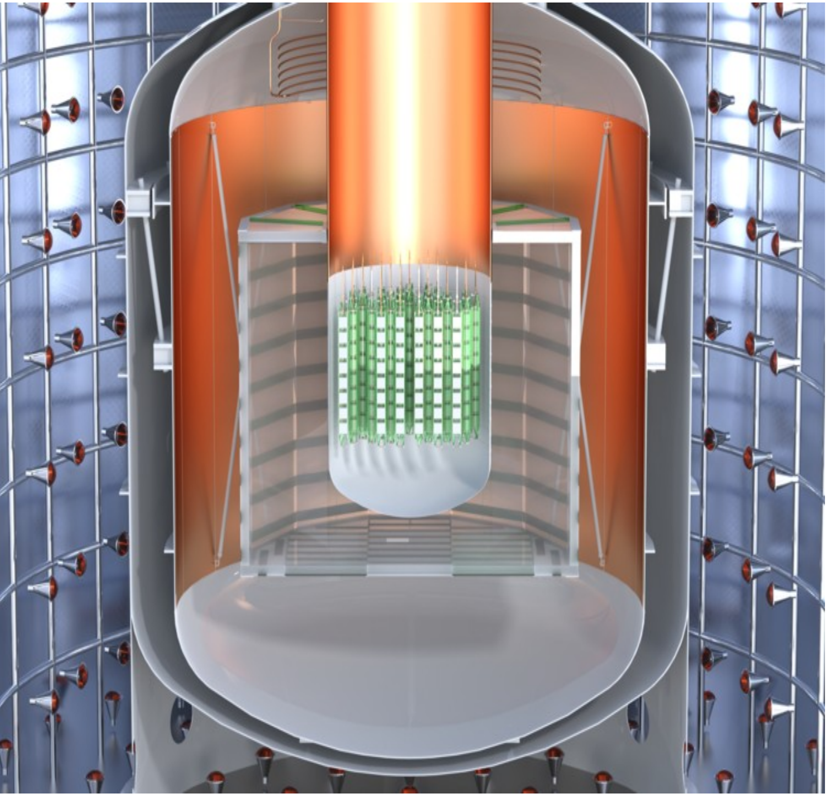
XENON-nT



- **Detect WIPMs as dark matter candidate**
- **XENON based experiments achieved the best sensitivity**
  - Negative result
- **Using the ultra-low background design (materials, vetoes)**
- **WIMPs interaction generate light and charge**
  - Light detected by ~500 PMTs
    - Between 3 to 100 photons
  - The charge is converted in a second light pulse

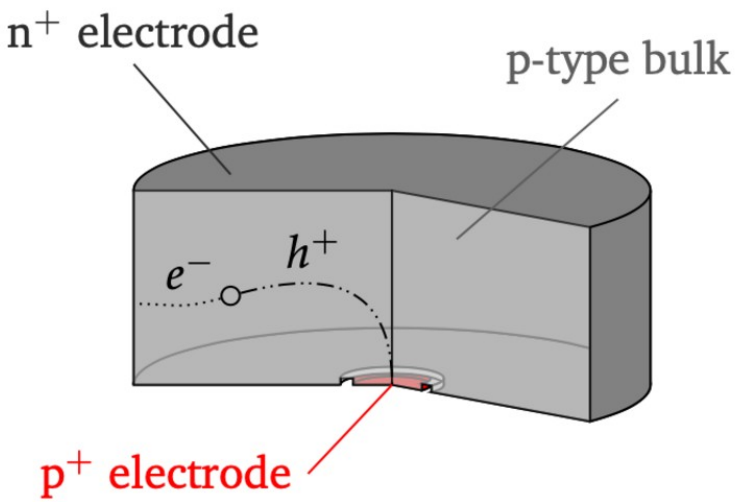


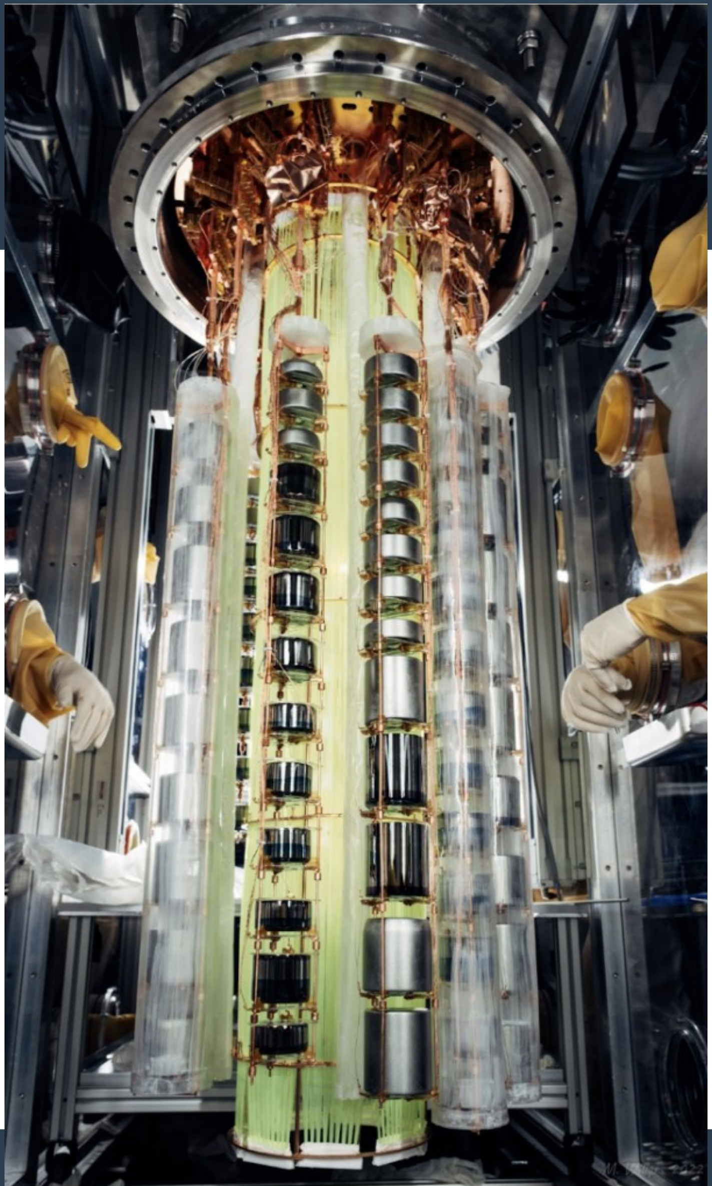
LEGEND



- **Searching for *forbidden* double beta decay in ^{76}Ge**
  - Current limit is larger than 2 10^{26} years
- **200 – 1000 kg of enriched Ge**
- **Using the ultra-low background design (materials, vetoes)**
- **Ge crystals are configured as large fully-depleted diodes**
  - Detect the ionized charge induced by the decay

LEGEND



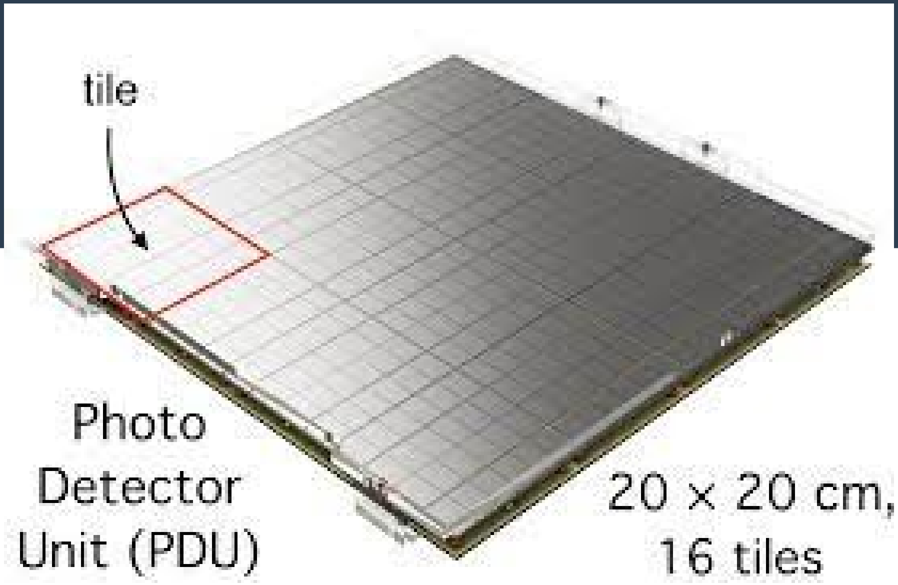


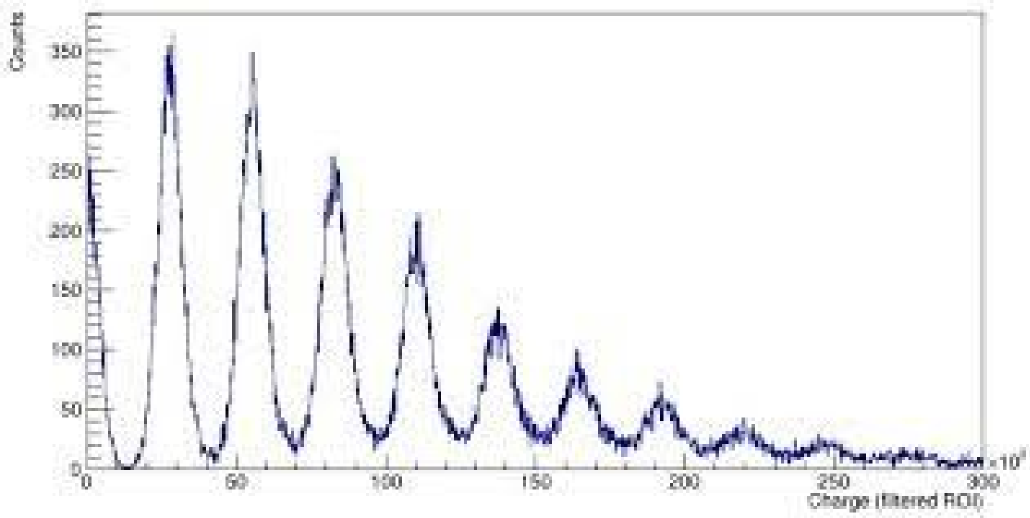
Signal extraction

Front-End Electronics

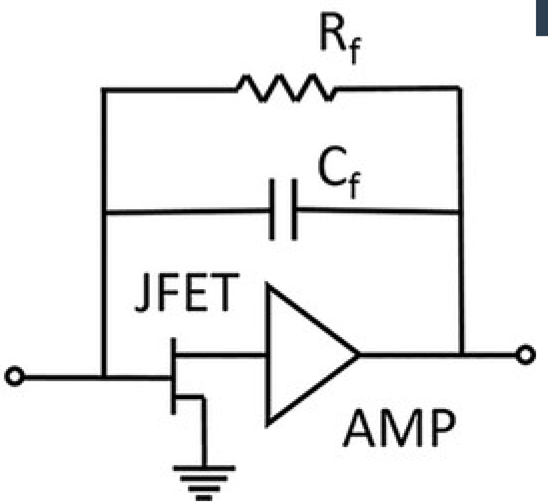
- **Typically the output of the sensors is very small**
  - For the signal acquisition
  - For the transmission on cables
- **Very low noise amplifiers are developed**
  - To optimize the signal integrity
- **The front-end can be at room temperature or in cryogenic environment**

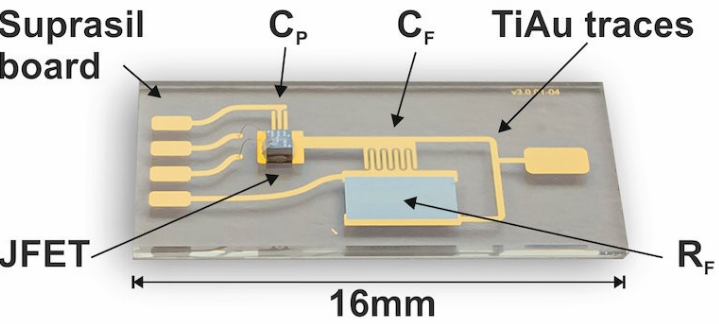
LNGS Examples

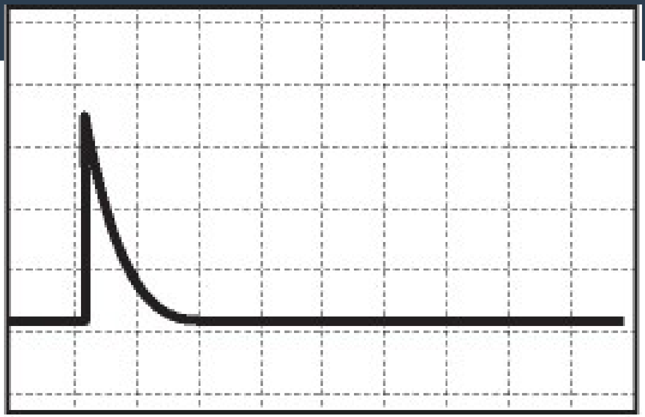




Charge sensitive amplifiers

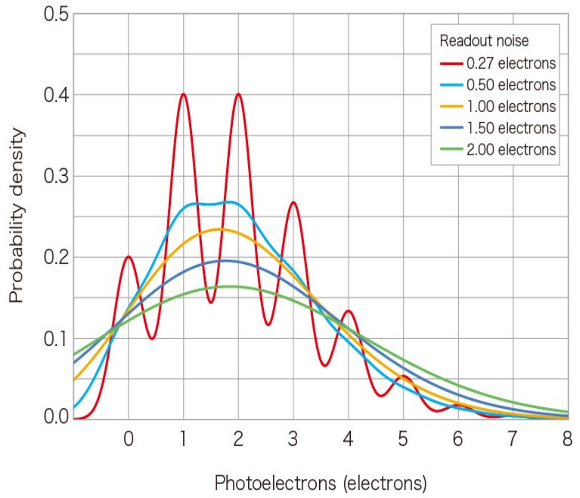




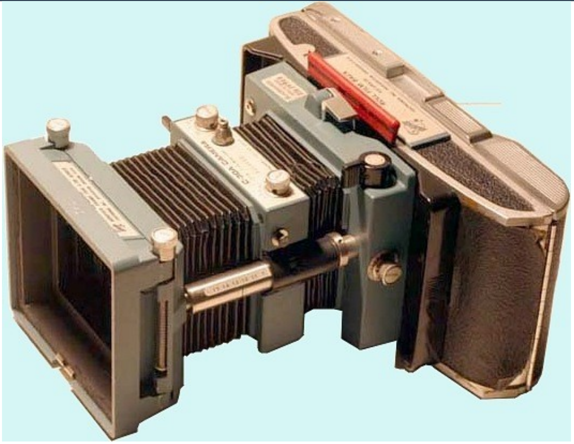


$i_d(t)$: detector current pulse: 10ns/div

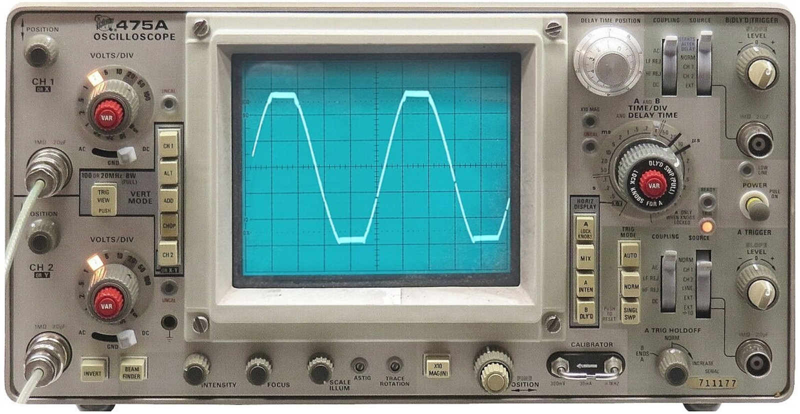
$V_{out}(t)$: CSP output pulse: 10ns/div

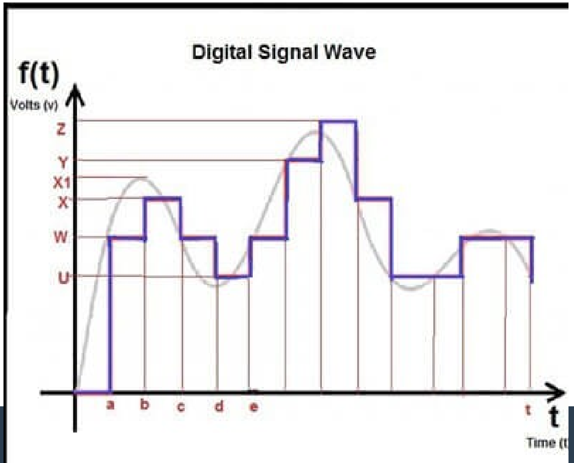


Waveform Digitization



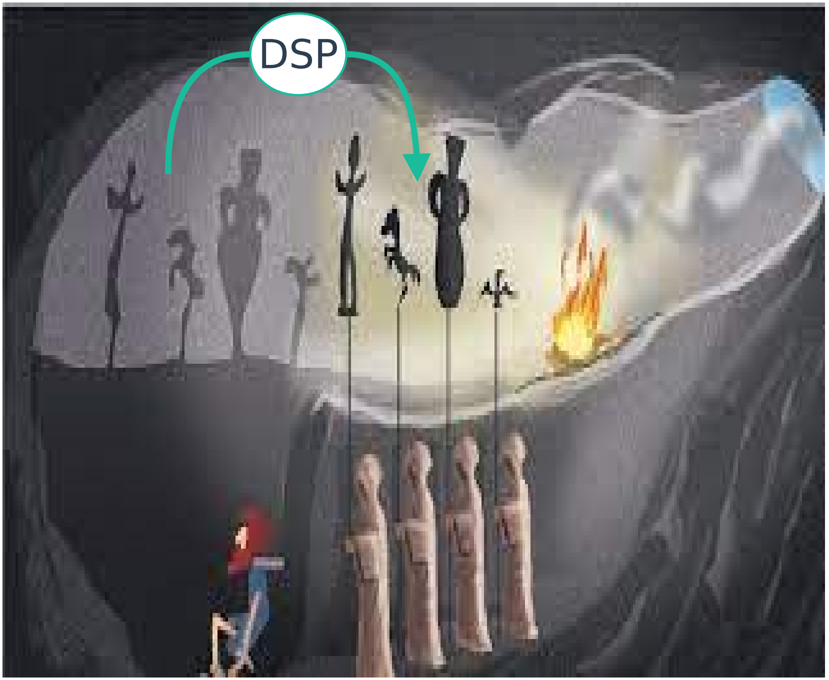
- **Before 1971 photos on oscilloscope waveforms**
- **Modern experiments are based of fast digitizers (>100 MHz) with 12-16 bits**





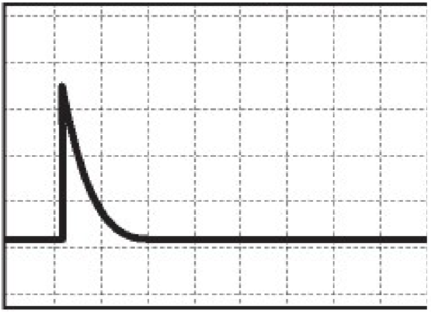
100 GS/s

Digital Signal Processing

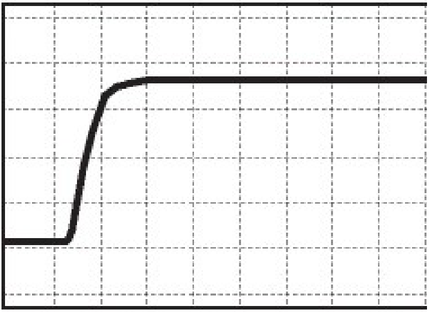


- **Digitized waveform can be further processed to maximize the signal to noise ratio**
  - Typically filtering is used to abate noise
    - At the expense of losing part of the signal
- **DSP use is ubiquitous in modern devices**
  - Multi-rate filtering ← iPOD (mp3)
  - RADAR

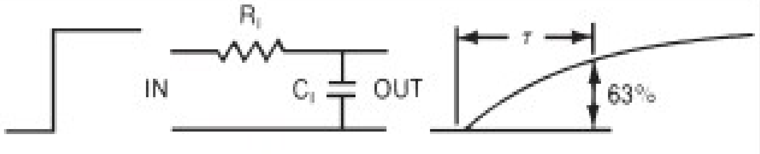
Spectroscopy Amplifiers



$i_d(t)$: detector current pulse: 10ns/di

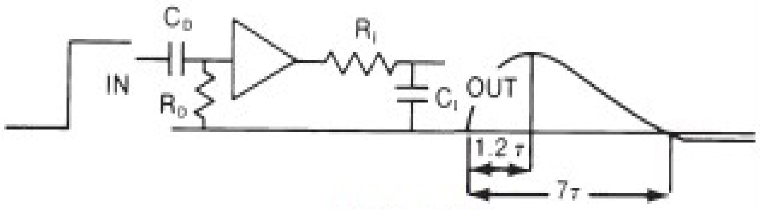


$V_{out}(t)$: CSP output pulse: 10ns/div



$\tau = R_I C_I$

Fig. 8. RC Integration.



$\tau = R_D C_D = R_I C_I$

Fig. 9. CR-RC Pulse Shaping.

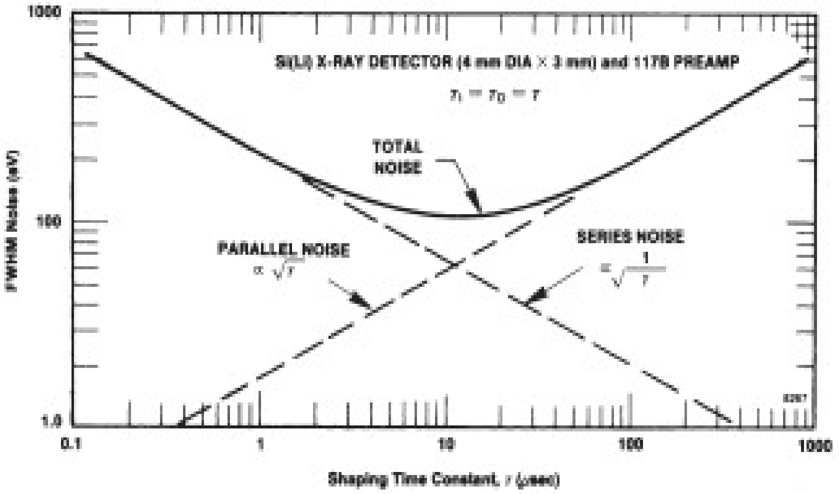
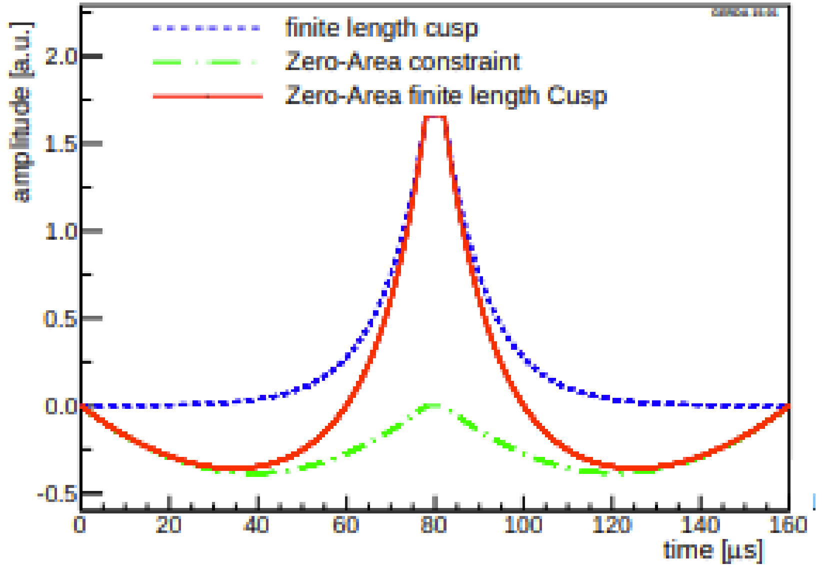
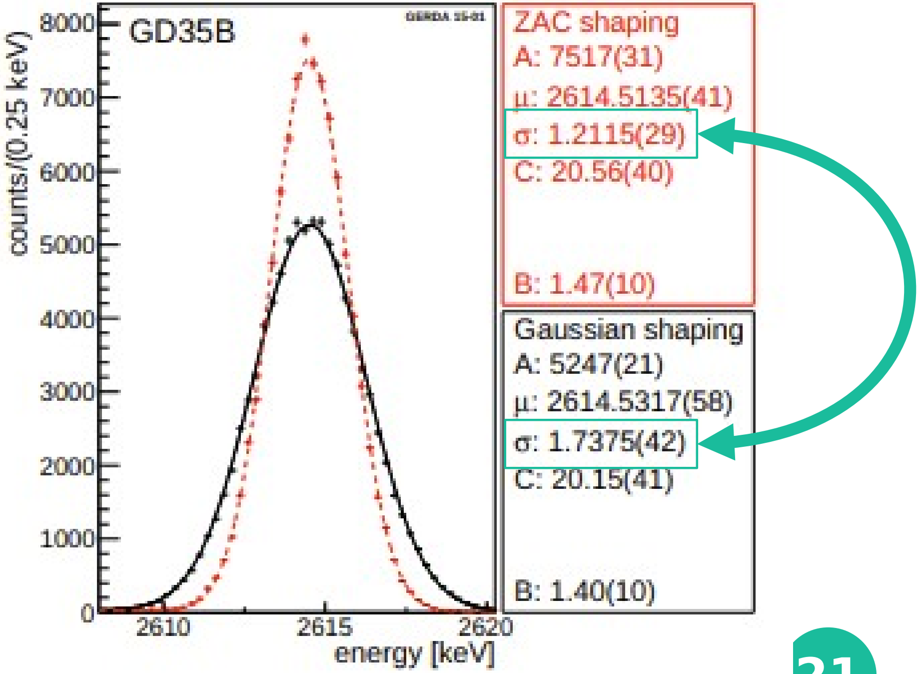


Fig. 10. The Dependence of the Preamplifier Noise Contribution on the Amplifier Shaping Time Constant.

DSP for improved filtering

In the digital domain the filtering can be much more advanced
CRC filters can be replaced by gaussian or other symmetric shapes





Matched filter

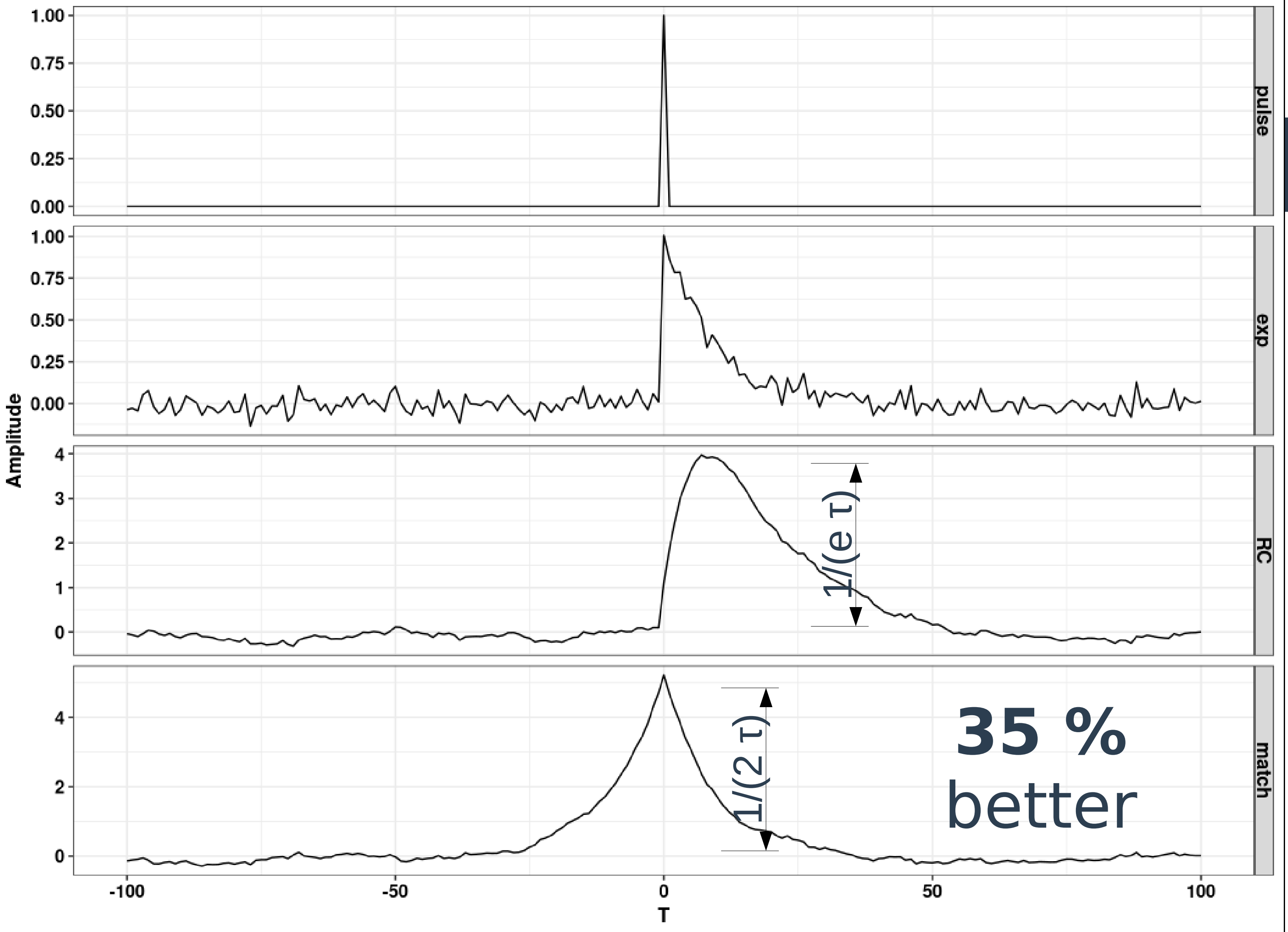
Turin 1960
Maximize SNR
(amplitude)

Introduced for RADAR

Correlate the signal
with the known ref

Minimize the phase
Dispersion

Can be implemented
only in digital
(anti-causal)

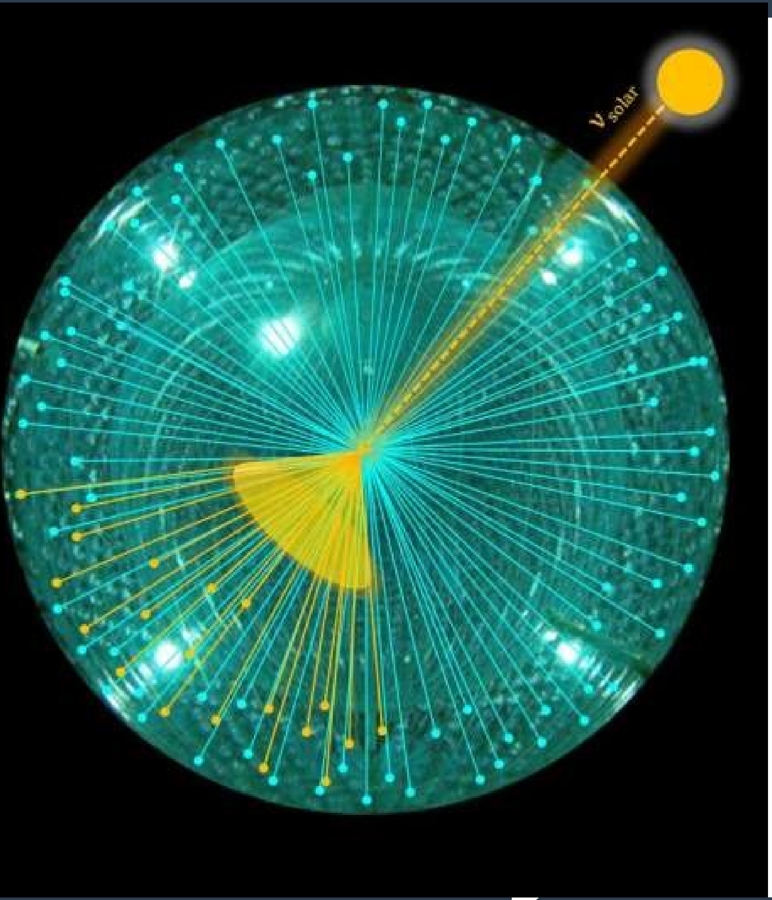


Low Level Analysis

- **Reconstruct the waveforms in high level data**
- **Include calibrations**
- **Position reconstruction**

- **In general the most complicated part of the analysis chain**
  - Knowledge of the detector details

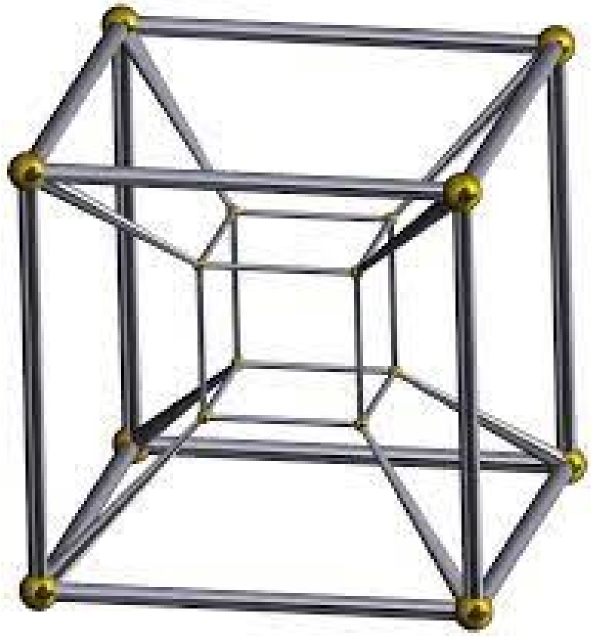
Borexino



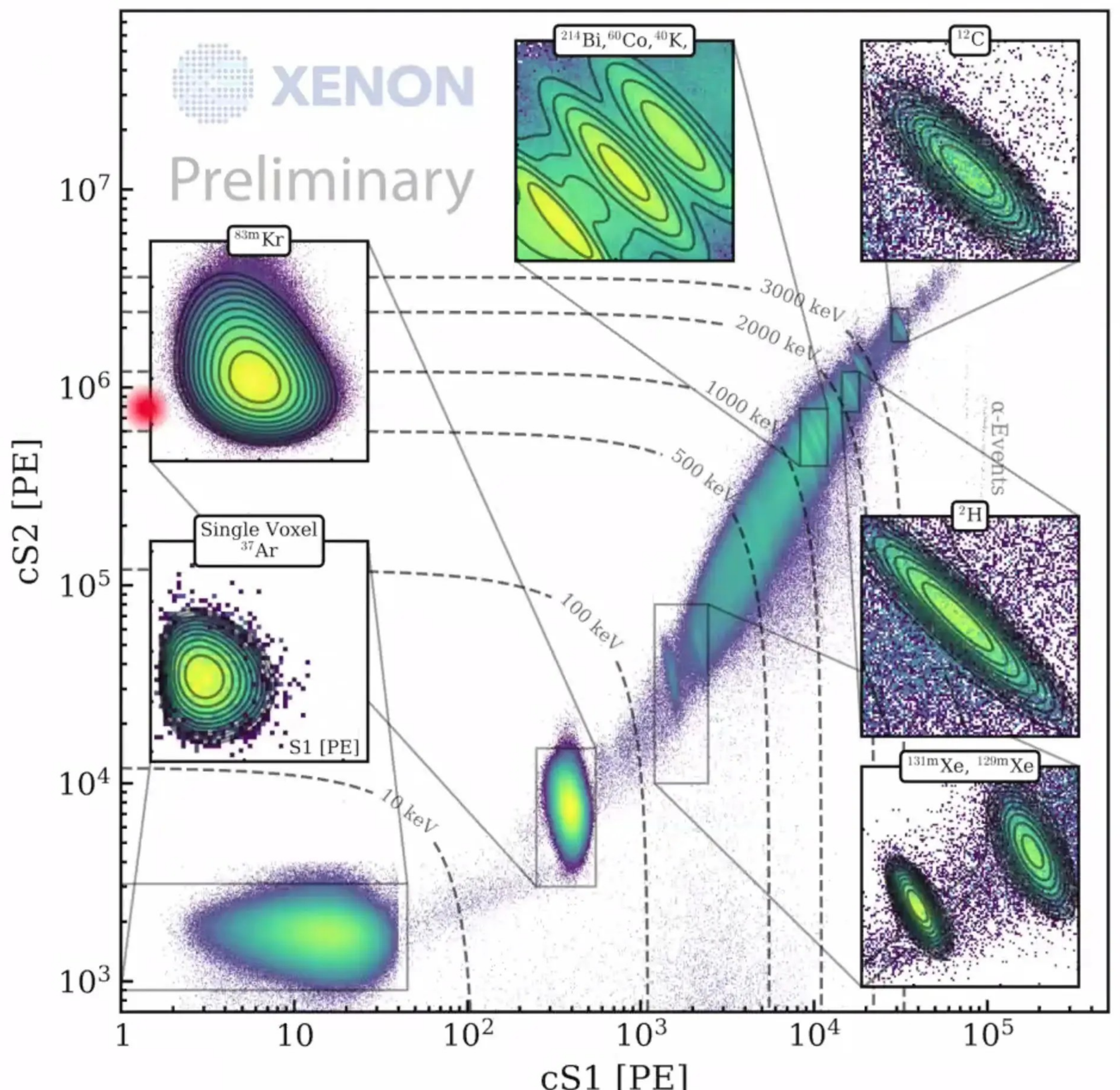
- **Interaction of solar neutrino generates about 500 photons**
  - ~ hundreds per day over a background of 1 million
- **Detected by the PMTs**
- **We record the time, amplitude and position of each signal**
- **We reconstruct**
  - Interaction energy
  - Pulse shape
  - Position in the detector
  - Status of the vetoes

High Level Analysis

Multi-Dimensional Data



- **High level data are nativelly highly-dimensional**
- **An *event* correspond to an interaction in the detector**
  - Several information are available
    - Primary data (energy, position, topology)
    - Nuisance data (veto, metadata)
- **However we typically do not employ multi-dimensional regression**



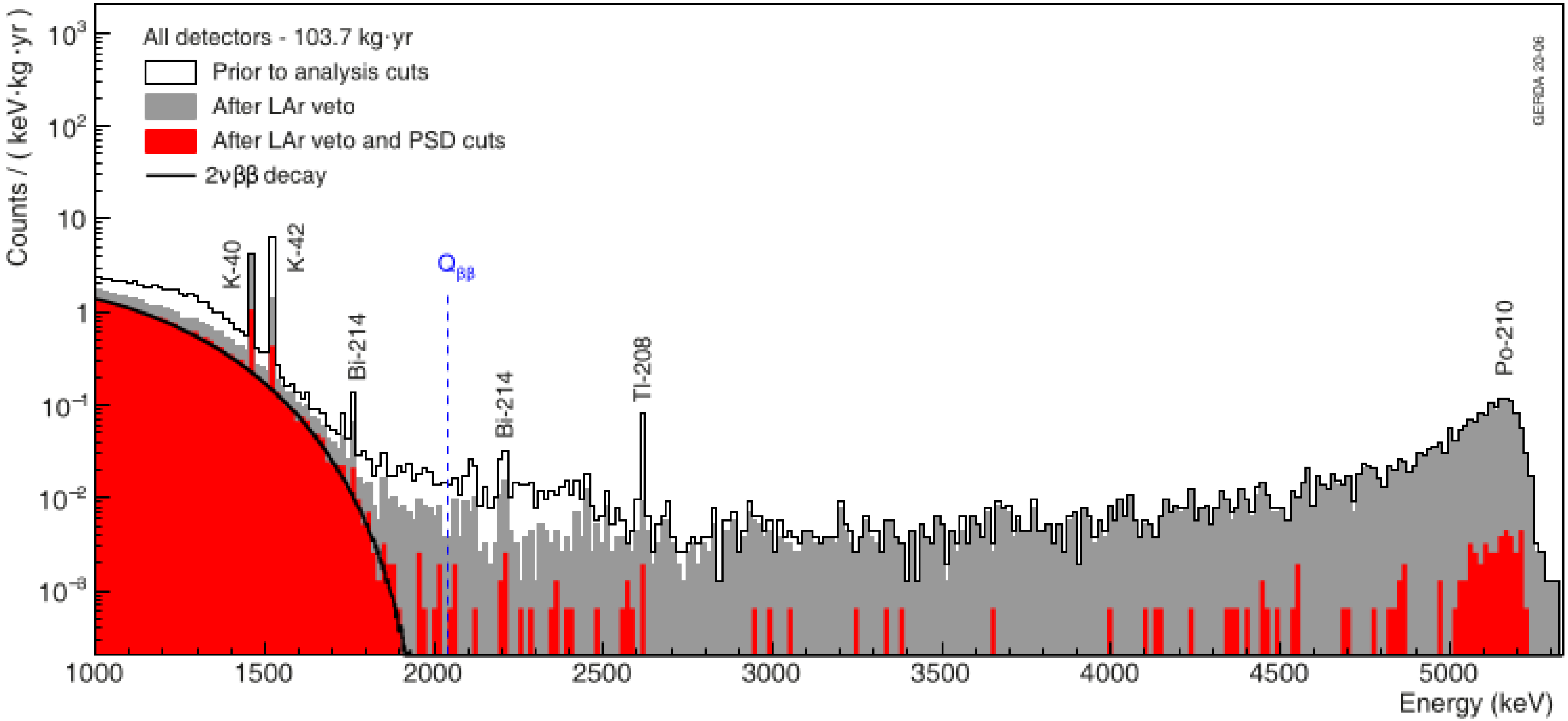
XENON
Preliminary
83mKr
Single Voxel 37Ar
S1 [PE]
214Bi, 60Co, 40K,
12C
2H
131mXe, 129mXe
3000 keV
2000 keV
1000 keV
500 keV
100 keV
10 keV
α-Events
cS2 [PE]
cS1 [PE]

Histogram

- **Typically we populate histograms with our primary data**
- **Search peaks (or other features) in these histograms**

- **Unbinned likelihood are not common**
- **Histogramming is often abused**
  - Even to calculate the with of a peak we typically use histograms

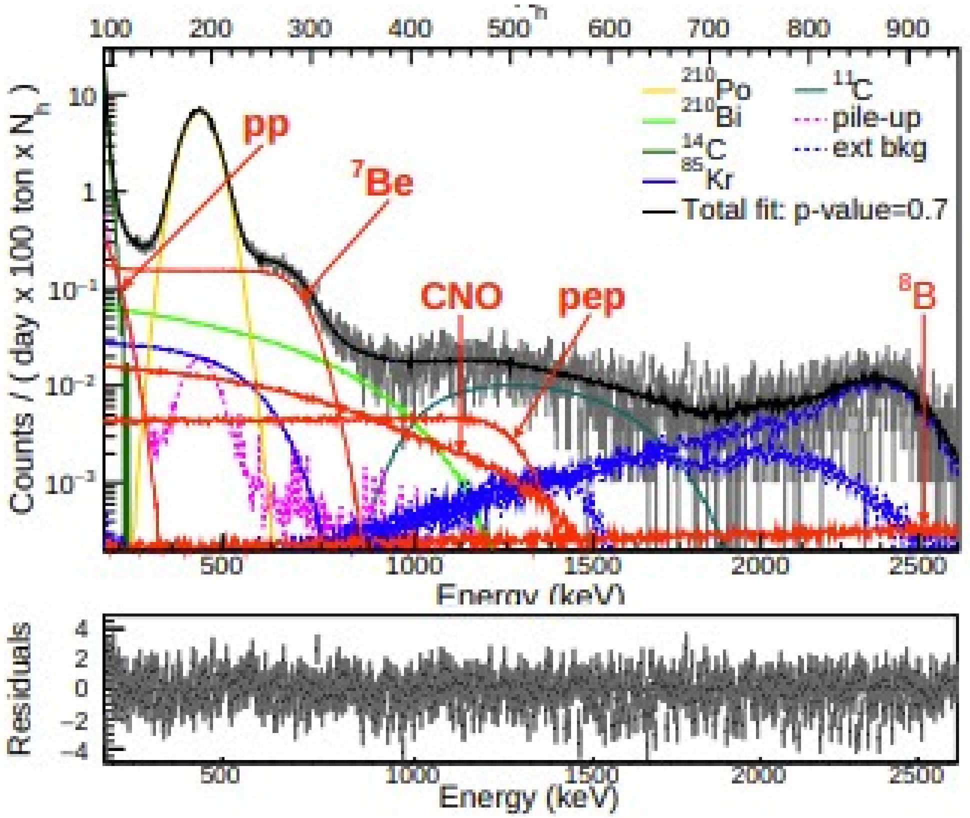
GERDA Final data

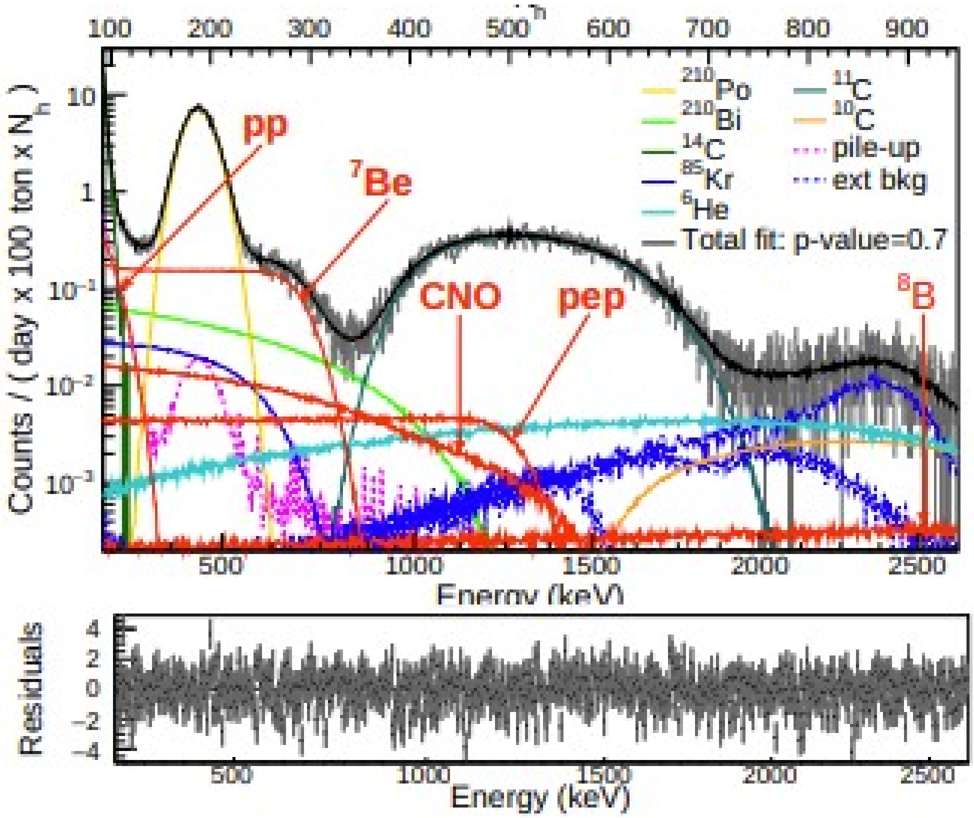


χ^2 test

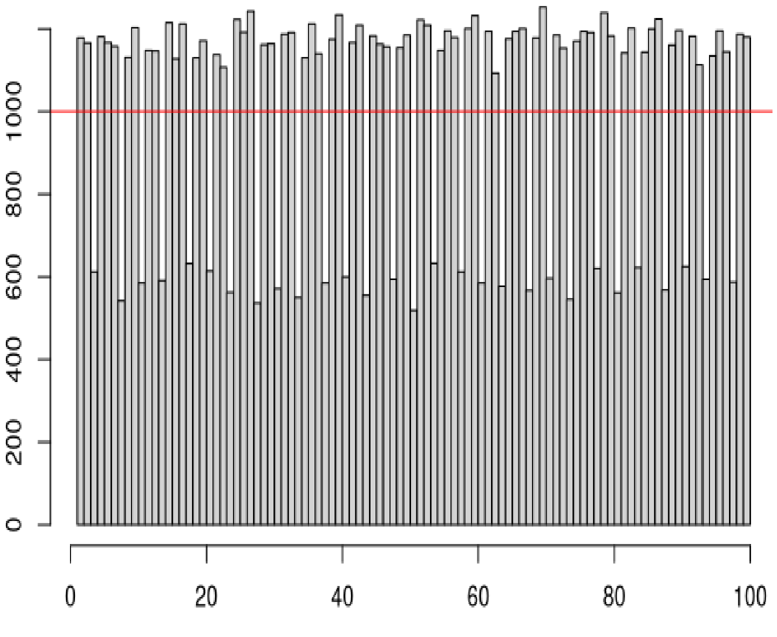
- **The research of features on histograms is based on the χ^2**
  - We assume each bin is an independent random variable
  - We use the Neyman χ^2
- **The minimization of the χ^2 provides the model parameters and the goodness of fit**
- **It is a very handy process**
  - Not immune to several quirks

Multivariate fit of Borexino data





χ^2 not always is benevolent

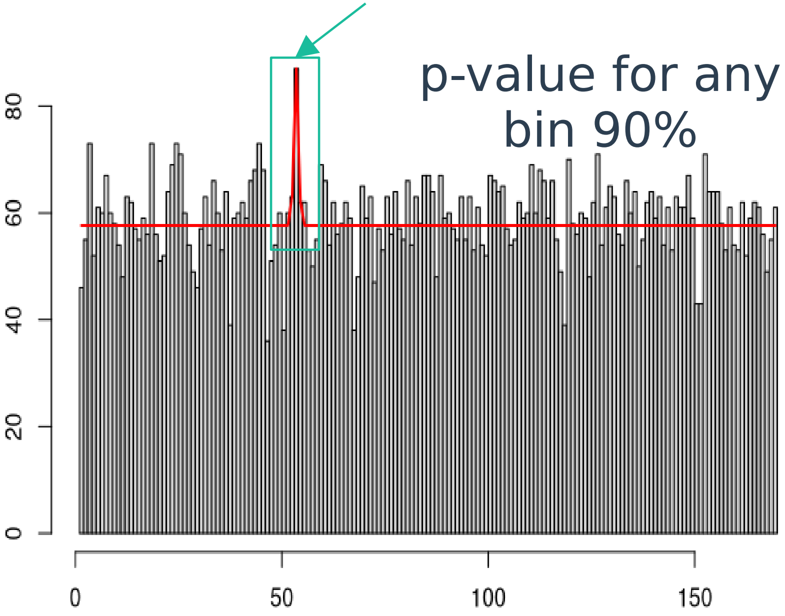


Rounding problems

- **Not always the bin contents are independent**
  - Or the binning is wrong
  - # of degree of freedom is overestimated
    - P-value wrong
- **Look elsewhere effect**
  - Random peaks can appear
- **…**
- **CERN requires 5 σ evidence**

χ² not always is benevolent

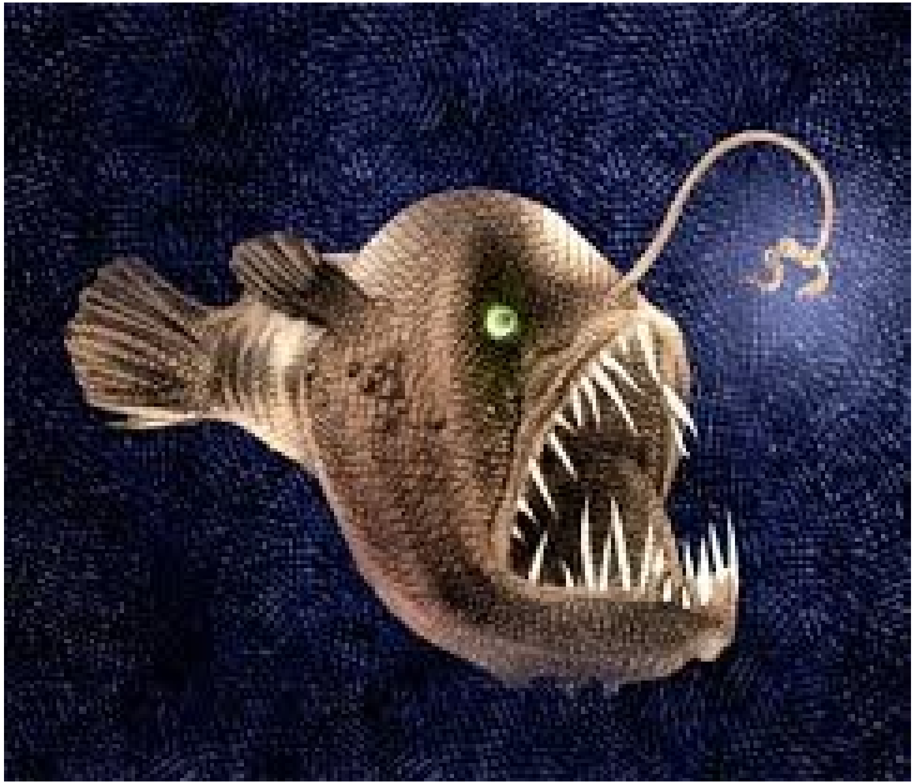
p-value of fluctuation .5%



Look elsewhere

- **Not always the bin contents are independent**
  - Or the binning is wrong
  - # of degree of freedom is overestimated
    - P-value wrong
- **Look elsewhere effect**
  - Random peaks can appear
- **...**
- **CERN requires 5 σ evidence**

Model Driven vs Data Driven

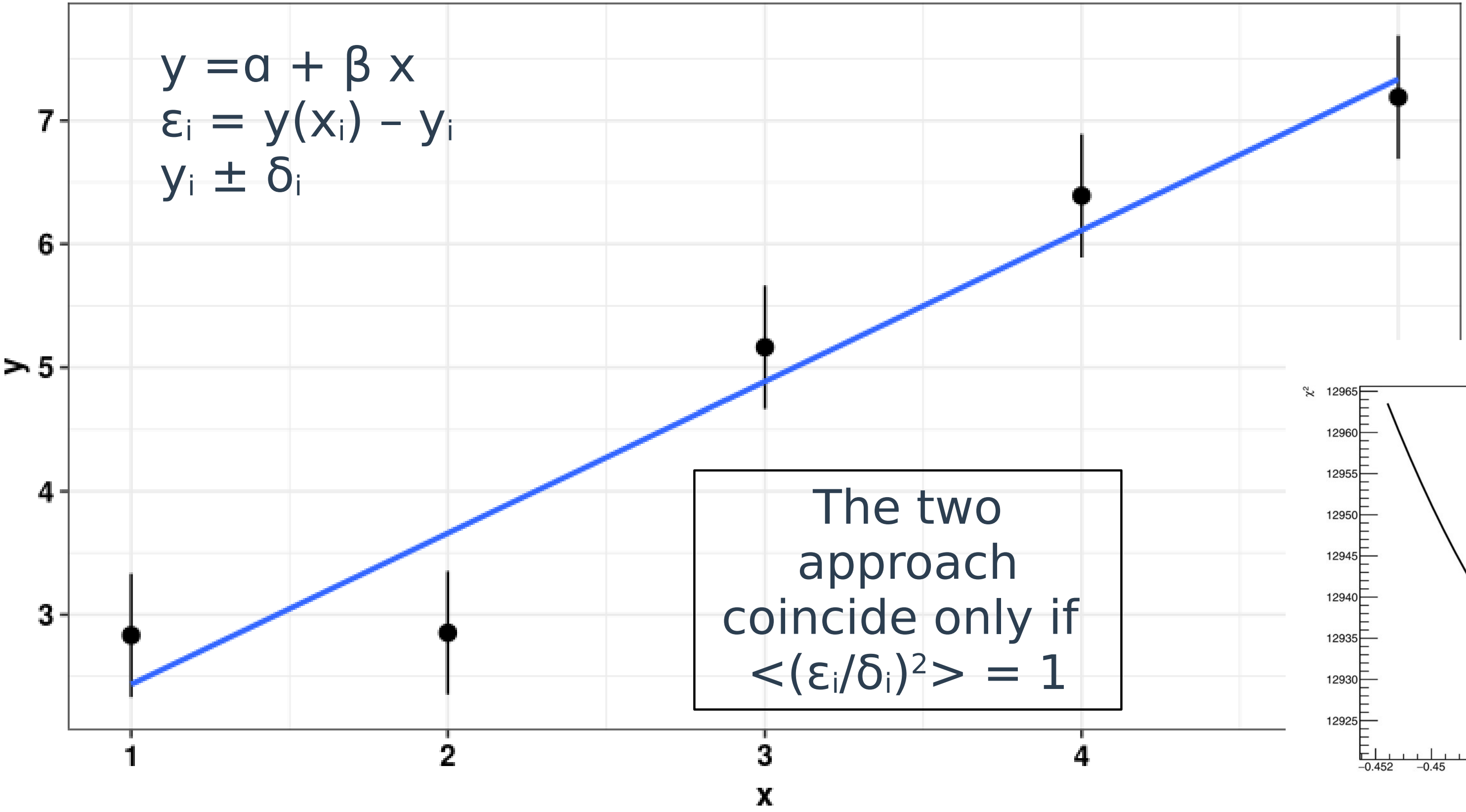


Interpretation of linear regression

$y = \alpha + \beta x$

$\varepsilon_i = y(x_i) - y_i$

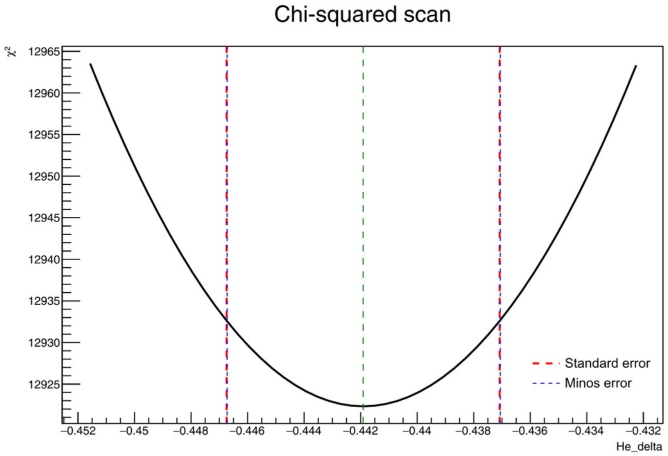
$y_i \pm \delta_i$



The two approach coincide only if $\langle(\varepsilon_i/\delta_i)^2\rangle = 1$

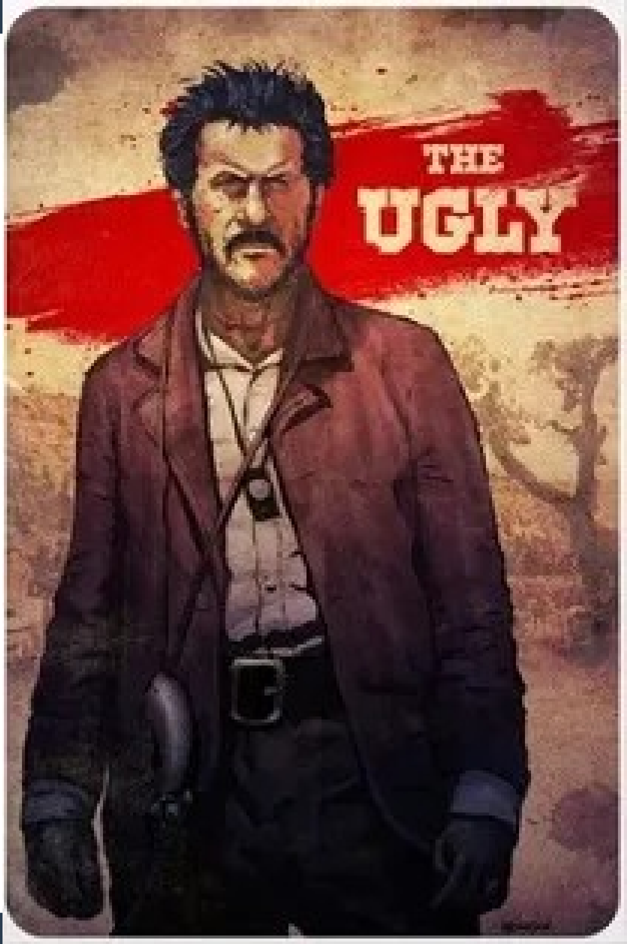
$$s_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$s_{\hat{\alpha}} = s_{\hat{\beta}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} =$$



Software Tools

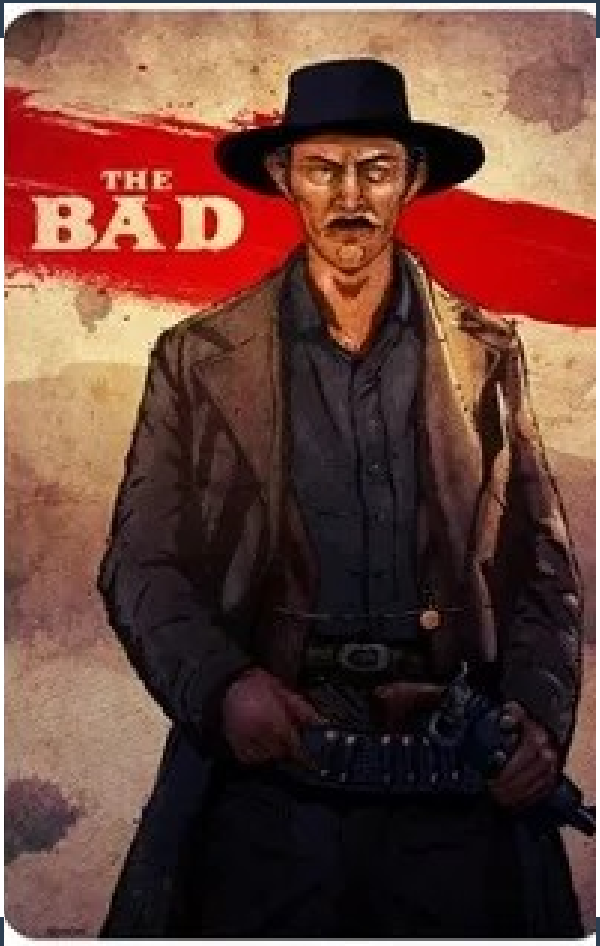
ROOT



- **ROOT is a software package developed at CERN**
  - With the main focus on particle physics
- **It has a nice data storage (columnar db-file)**
  - Allows to analyze large data set
    - Larger than memory size
- **It was written in C++**
  - Well before C++11
  - It uses a C++ interpreter (clang)

Its syntax is ugly

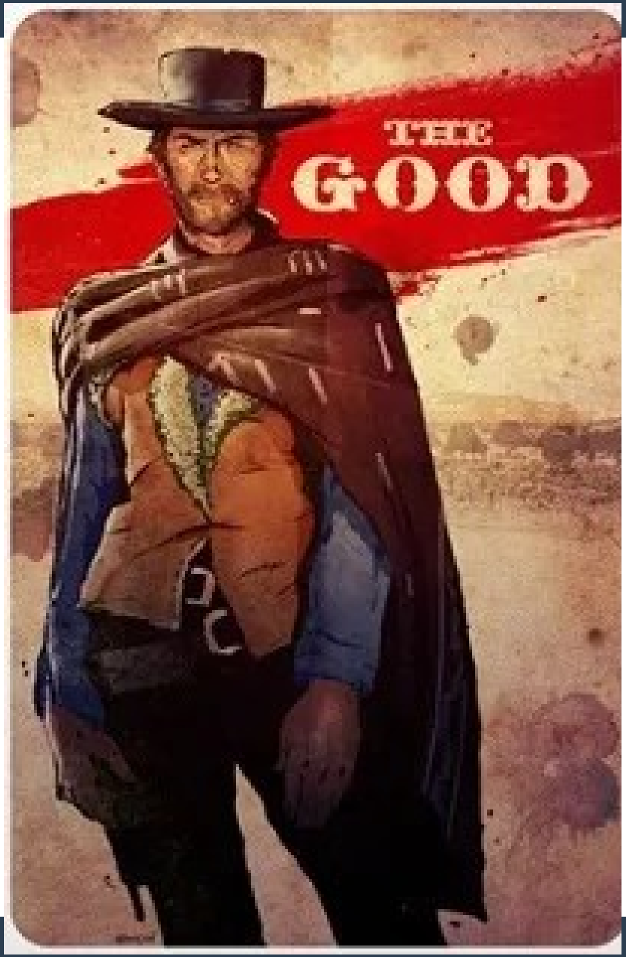
C++



- **After Fortran77 CERN pushed the scientific community for the adoption of C++**
- **C++ is fine for the DAQ**
  - And for the intensive low-level analysis
- **C++ is not a scripting language**
  - Complex (bloated) syntax
  - Not vectorized + external iterator
  - Handling string is painful
  - Missing syntactic sugar
  - **Prone to memory leaks**

"There are only two kinds of languages: the ones people complain about and the ones nobody uses"
Bjarne Stroustrup

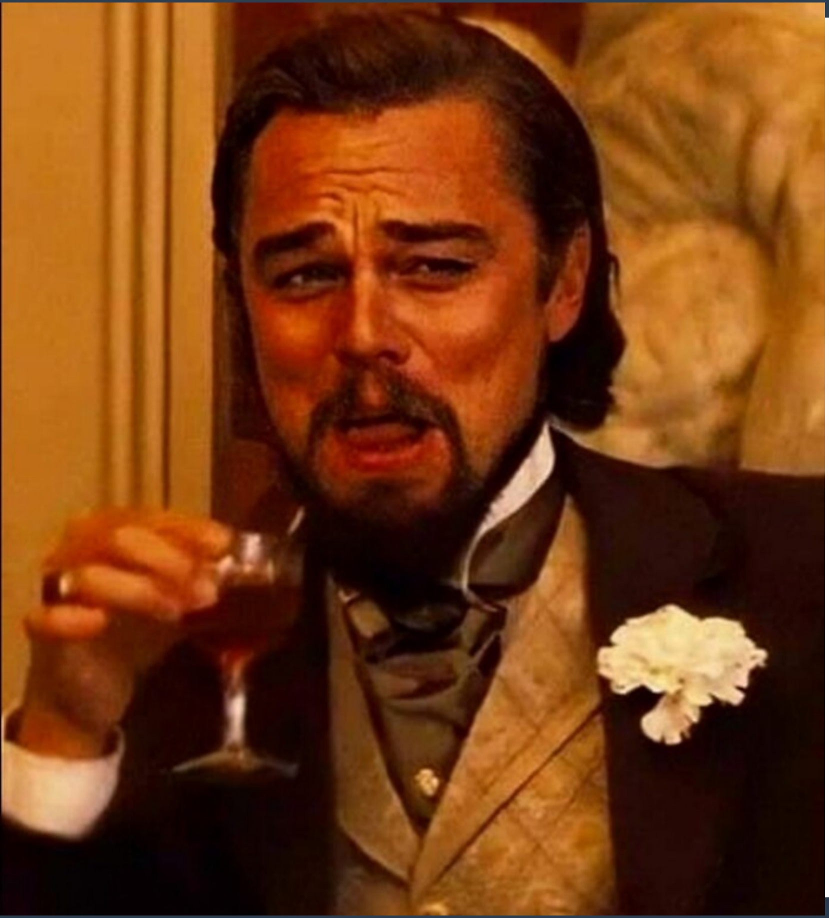
PYTHON



- **Python is acquiring popularity in particle physics**
  - It is easy to learn
  - Has advanced scientific libraries
    - ROOT library as well

- **It is much better than C++ but**
  - Missing large data-set operation
  - Many dislike the syntax
  - It is much slower than C++
    - Some compilation/optimization options available

R

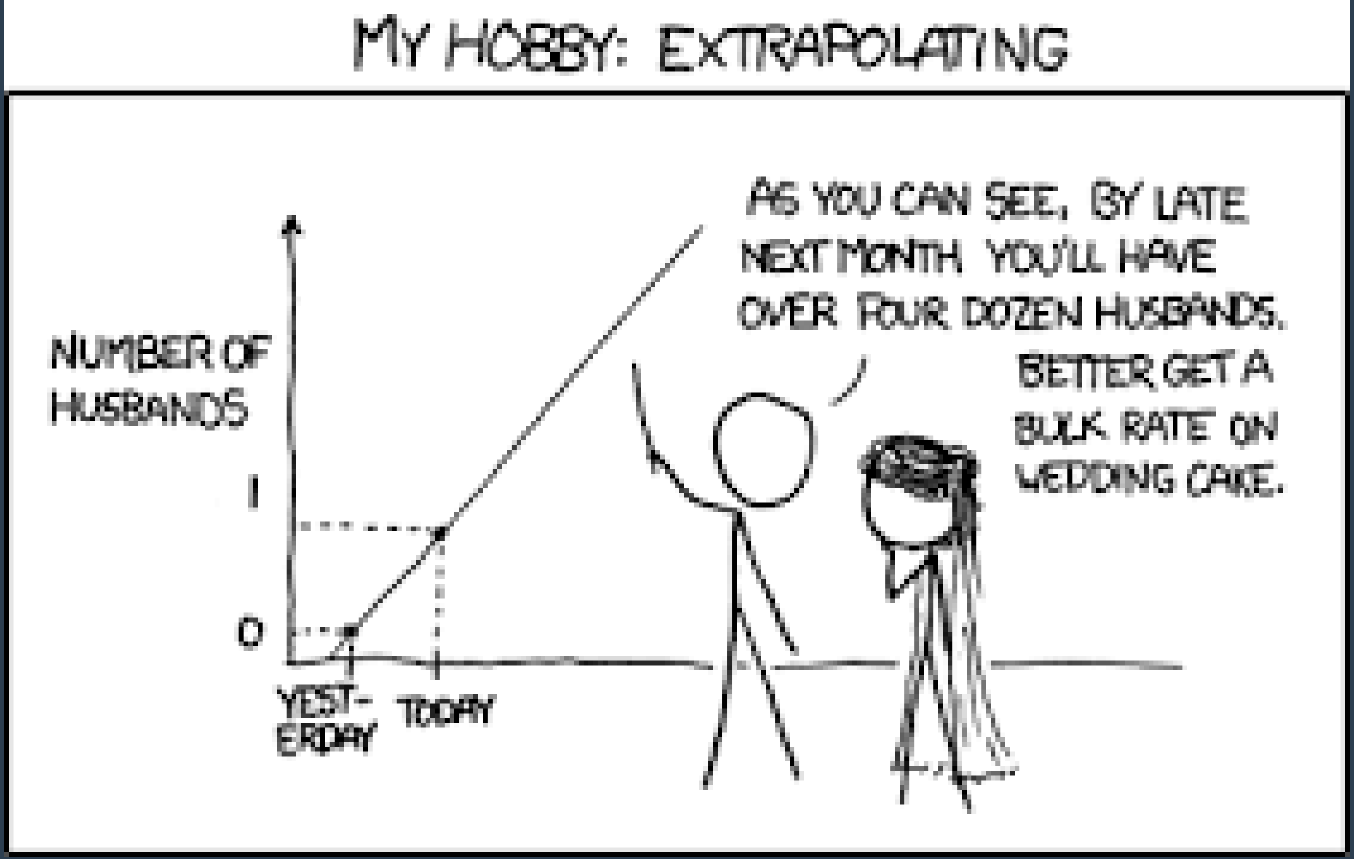


- **Personally I use R**
- **It is not tailored for particle physics**
  - Histogram fitting is somehow missing
- **It forces me to study and to think differently**
- **No built-in support for large data set**
  - I use SPARK clusters with parallel map/reduce

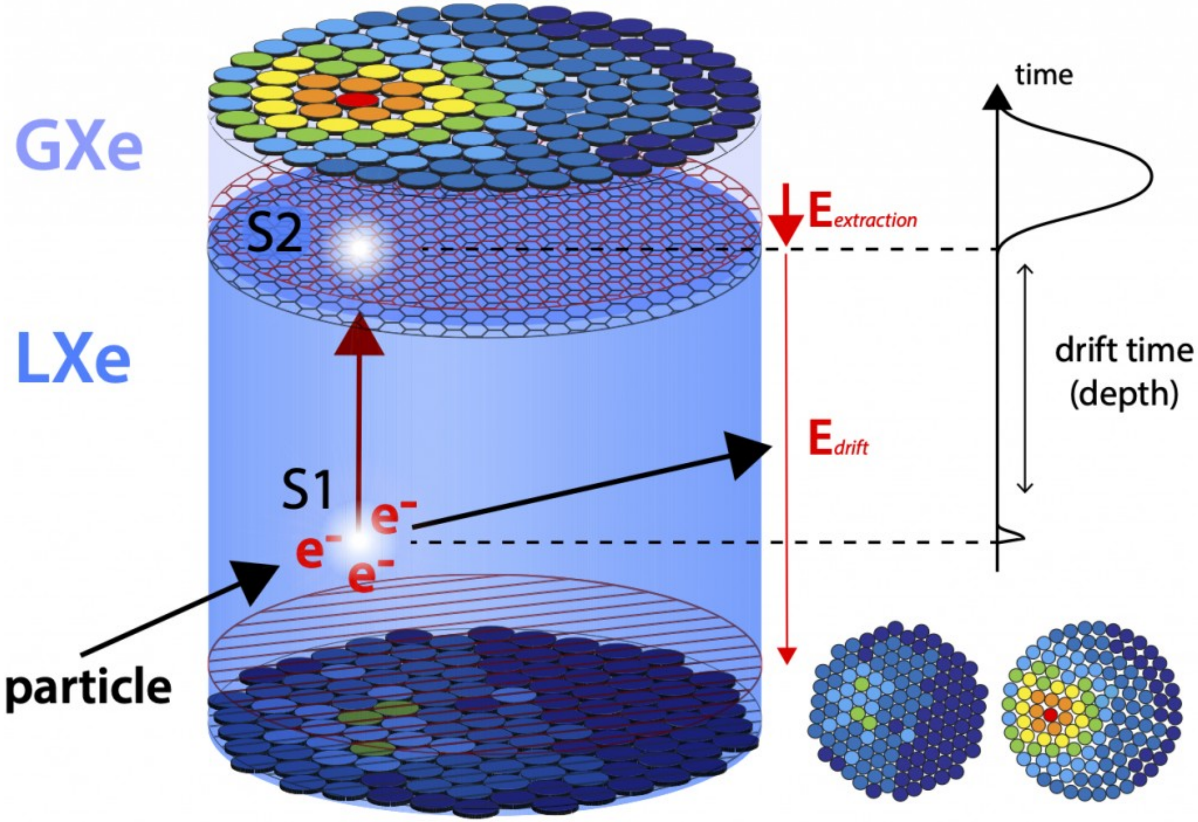
Conclusion

- **A rich environment that includes**
  - Theory
  - Experiments
  - Models
  - Data
  - Plus our idiomatic use of statistics
- **Large experiments are fueled by large international collaboration of physicists**

Thank you


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

XENON-nT



GXe
S2
LXe
S1
e⁻
e⁻
e⁻
particle
time
E extraction
E drift
drift time
(depth)