Measurements and data processing

Ivan Prochazka

Consultations on request TN314

Czech Technical University, Prague

Course Goals

- high precision / accuracy (1)
- correct interpretation of results (2)
- marginal effect identification (3)
- low signal extraction from the noise background / data mining (4)

Course Concept

- “open concept”
  - questions / comments related to the subject welcome
  - language is no limitation
- based on local tradition and experience:
  - photon counting,
  - high precision & accuracy laser ranging,
  - Lidar,
  - precise timing etc.
- Measurement, data processing and laboratory demo
- contributions from students to the course appreciated (see next)

Requirements

- 3 tests within the semester, announced in advance (~ 10 questions / test, language is no limitation)
- minimum 50 % of correct answers in each test
- one spare term for the three tests
- !! WARNING just one single spare term / test !!
- final note will be an average of the three test results (improvement possible by active contribution ..)

Course Structure / Schedule

1. Definition of terms
(measurements, observations, errors characterization, precision, accuracy, bias)

2. Types of measurements and related error sources
(direct, indirect, substitution, event counting, ...)

3. Normal errors distribution
(histogram, r.m.s., r.s.s., averaging, ...)

4. Normal errors distribution consequences
(examples, demo, test#1)

5. Data fitting and smoothing I.
(interpolation, fitting, least square algorithm, mini-max methods, weighting methods)

6. Data fitting and smoothing II
(parameters estimate, fitting strategy, solution stability)

Course Structure / Schedule II

1. Data fitting and smoothing III
   (polynomial fitting, "best fitting" polynomial, splines, demo)

2. Data editing
   (normal data distribution, k * sigma, relation to data fitting, deviations from normal distribution, tight editing criteria, test #2)

3. Signal mining
   (noise properties, correlation, lock-in measurements)

4. Signal mining methods
   (Correlation estimator, Fourier transform application)

5. Signal mining methods – examples
   (Time correlated photon counting, laser ranging, relation to data editing and data fitting)

6. Review, test #3

References

- 1. Horák, Z.: Praktická fyzika. SNTL, Praha
- 3. Water measurement manual, [online] [cit. 2005-Jan-02], < http://www.usbr.gov/pmts/hydraulics_lab/pubs/wmm/chap03_02.html >
- - Chapter 3.2 - Measurement accuracy - Definitions of Terms Related to Accuracy
- 4. Wikipedia – The Free Encyklopedia, Accuracy and precision, [online] [cit. 2005-Jan-02], < http://en.wikipedia.org/wiki/Accuracy >
- 5. Wikipedia – The Free Encyklopedia, Interpolation, [online] [cit. 2005-Jan-02], < http://en.wikipedia.org/wiki/Interpolation >
- 6. Wikipedia – The Free Encyklopedia, Curve fitting, [online] [cit. 2005-Jan-02], < http://en.wikipedia.org/wiki/Curve_fitting >
- 7. Wikipedia – The Free Encyklopedia, Moving Average, [online] [cit. 2005-Jan-02], < http://en.wikipedia.org/wiki/Moving_average >
- 8. Moore A., Statistical Data Mining Tutorials, [online] [cit. 2005-Jan-02], < http://www.autonlab.org/tutorials/ >
- 9. BERKA, K.: Měření, pojmy, teorie, problémy. Academia, Praha, 1977
- 10. Broz, J. a kol.: Základy fyzikálního měření. SPN, Praha
- 11. Solomon R.C. Douglas and David M. Harrison, Dept. of Physics, Univ of Toronto - Least Squares Fitting of Data from the Physical Sciences & Engineering, [online] [cit. 2009-Feb-010], < http://www.upscale.utoronto.ca/PVB/Harrison/MSW2004/MSW2004_Talk.html >
- 12. Data Mining: What is Data Mining? [online] [cit. 2009-Feb-010], < http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm >
- 13. Photon Counting using Photomultiplier tubes, [online] [cit. 2009-Feb-010], < http://sales.hamamatsu.com/assets/applications/ETD/PhotonCounting_TPHO9001E04.pdf >
- 14.University of Michigan – Error Analysis Tutorials, [online] [cit. 2009-Feb-10], < http://instructor.physics.lsa.umich.edu/ip-labs/tutorials/errors/vocab.html >
- 15.Data Fitting Manual, [online] [cit. 2009-Feb-10], < http://bima.astro.umd.edu/wip/manual/node11.html >
- 16. Wikipedia – The Free Encyklopedia, Accuracy and precision, [online] [cit. 2009-Feb-10], < http://en.wikipedia.org/wiki/Accuracy >
- 17. Matějka K. a kol., Vybrané analytické metody pro životní prostředí, 1998, Vydavatelství ČVUT - Chapter: Statistika a chyby měření (pp. 57-63)

Measurements 1

- Units SI
- fundamental (kg, m, s, A, mol, candela, K)
- derived (m/s, …)
- standards SI , national, local,..

Measurements 2

- type of measurement

direct x indirect
absolute x relative
substitute
compensation …
(examples)

- Event counting
(examples)

Measurement errors

- Raw errors
- measurement errors
  - systematic
  - random errors

Precision and accuracy

- !!! WARNING - language dependent !!!
  přesnost cz
  genauigkeit ge
  točnosť ru

- PRECISION
  Relative, internal, consistency, data spread

- ACCURACY
  “absolute”, related to standards

RANDOM ERRORS - Precision

- measurement errors caused by random influences
- various influences randomly combined
- random behaviour = > statistical treatment
- increasing the number of measurements, the random error influence can be decreased

SYTEMATIC ERRORS - Accuracy

- A measure of the closeness of a measurement /its average/ to the true value.
- Includes a combination of random error (precision) and systematic error (bias) components.
- It is recommended to use the terms "precision" and "bias", rather than "accuracy," to convey the information usually associated with accuracy.

- *definition according to* USC Information Sciences Institute, Marina del Rey, CA

SYTEMATIC ERRORS – Accuracy 2

- errors of references, scales, …
- measurement linearity
- external effects dependency
- in general – very difficult to estimate !!
- increasing the number of measurements, the systematic error influence cannot be decreased

RANDOM and SYTEMATIC ERRORS
How to estimate them ?

- It is recommended to use the terms "precision" and "bias", rather than "accuracy”,
- precision may be estimated by statistical data treatment,
- bias may be determined as a result of individual contributors,
- To estimate the bias, <u>all the individual contributors must be identified and determined.</u>

Type of measurements versus errors
comments

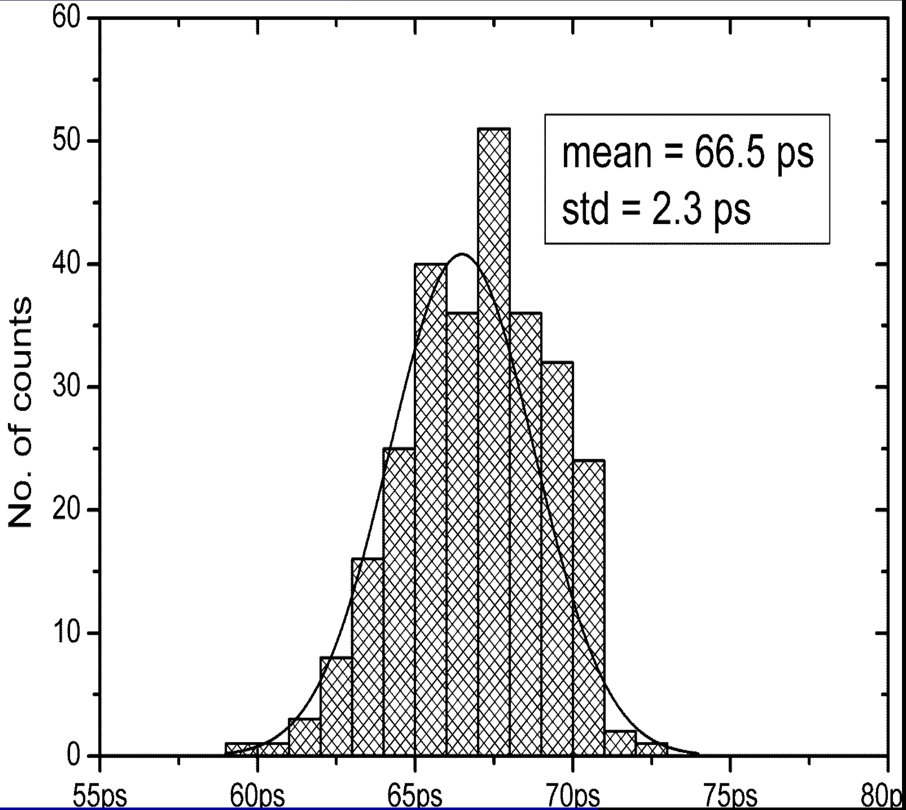
- comparative, compensation measurements are reducing the systematic errors,

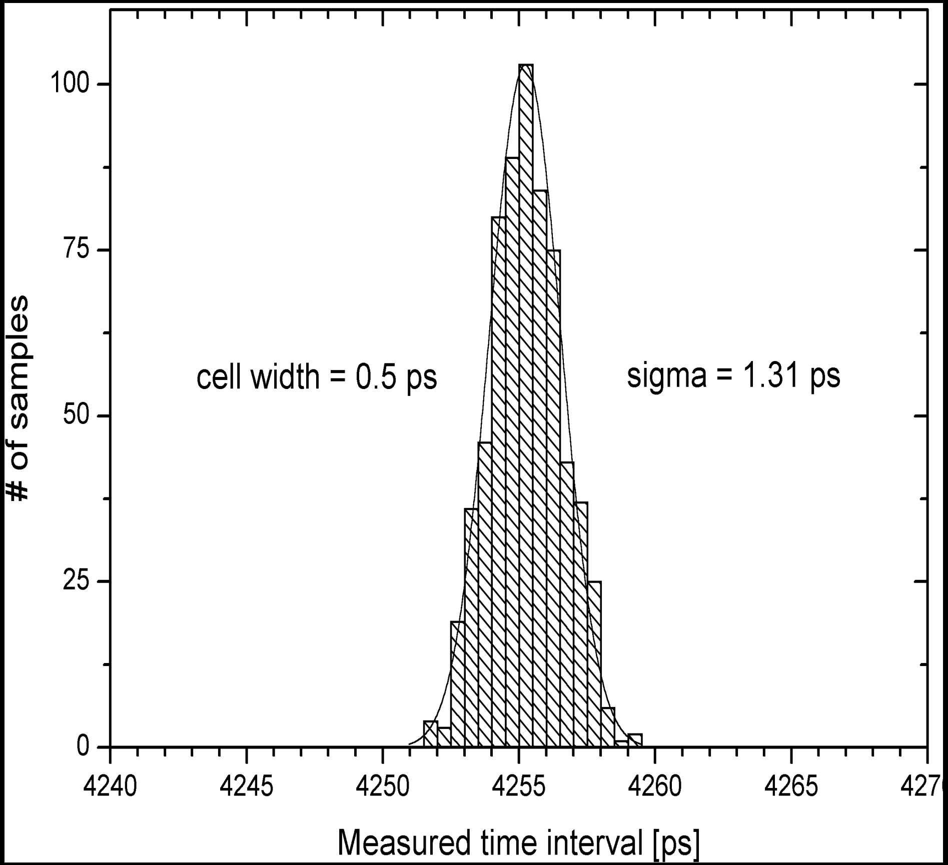
- more direct measurement is reducing both the error types,

- event counting (“clean measurement”) is drastically reducing the systematic errors,
  - the random errors can be predicted and effectively reduced
  - biases may be reduced by quantum level counting

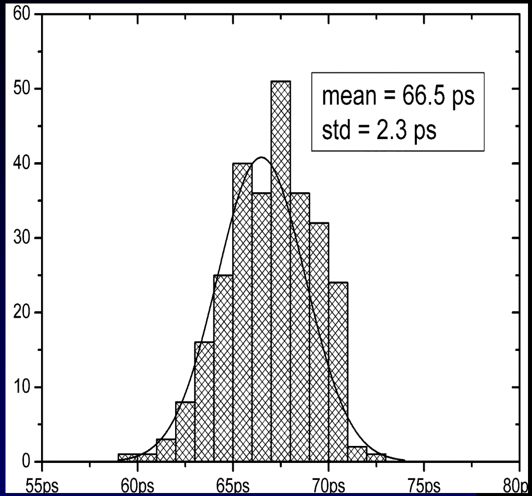
Random errors distribution – measured values

Histogram – statistical graph showing the frequency of occurrence, probability or Number of events





Random errors distribution – Gauss formula



3 KEY PRESUMPTIONS

1. Large number of errors (‘elementary’)
2. Equal size of all these errors
3. Random signs of errors

= > normal / Gauss distribution of errors

$$p(x) \approx e^{-\frac{(x-x_0)^2}{2\sigma^2}}$$

where p(x) … is a probability, that we will measure the value x

x_0 …. is a real value

σ ….. parameter – standard deviation
is a measure of precision

Random errors distribution – Gauss 2

$$p(x) \approx e^{-\frac{(x-x_0)^2}{2\sigma^2}}$$

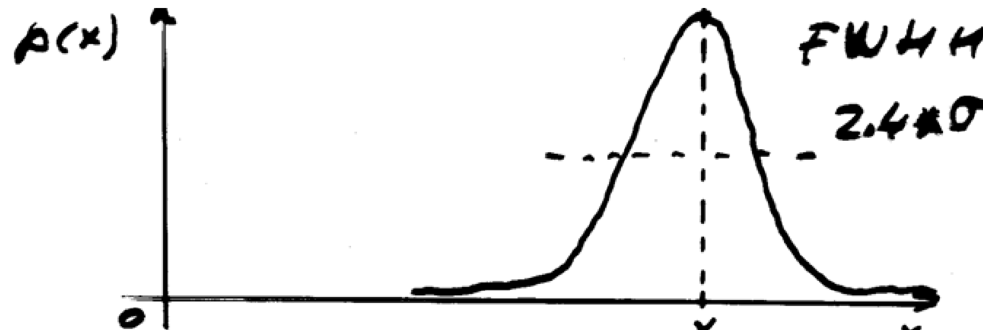
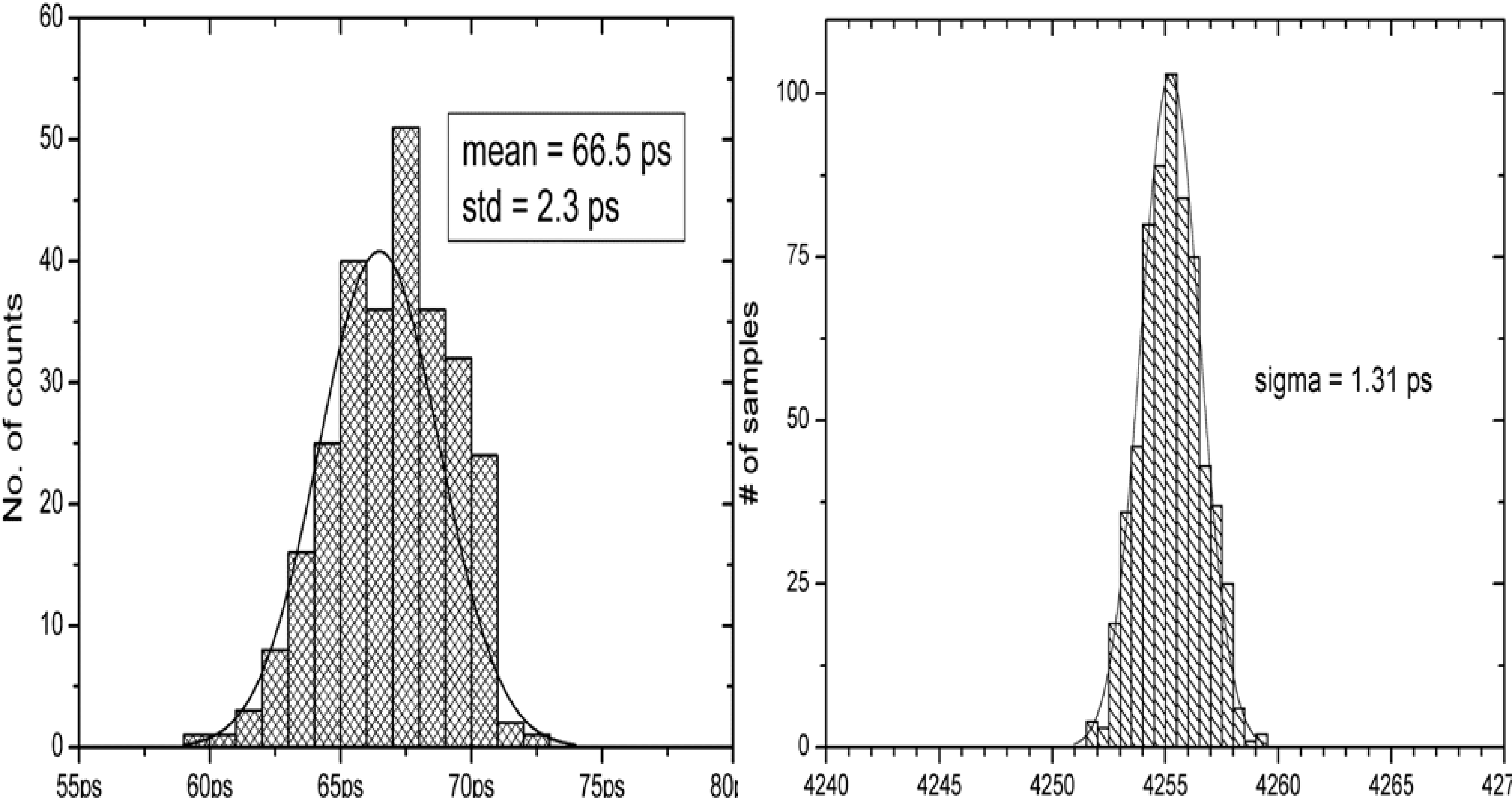


figure Gaussian distribution

- PROPERTIES
- Full Width Half Maximum … FWHM ~ 2.4 * σ is a measure of precision
- symetrical x_0
- approaches fast zero for (ABS($x-x_0$)) -> σ

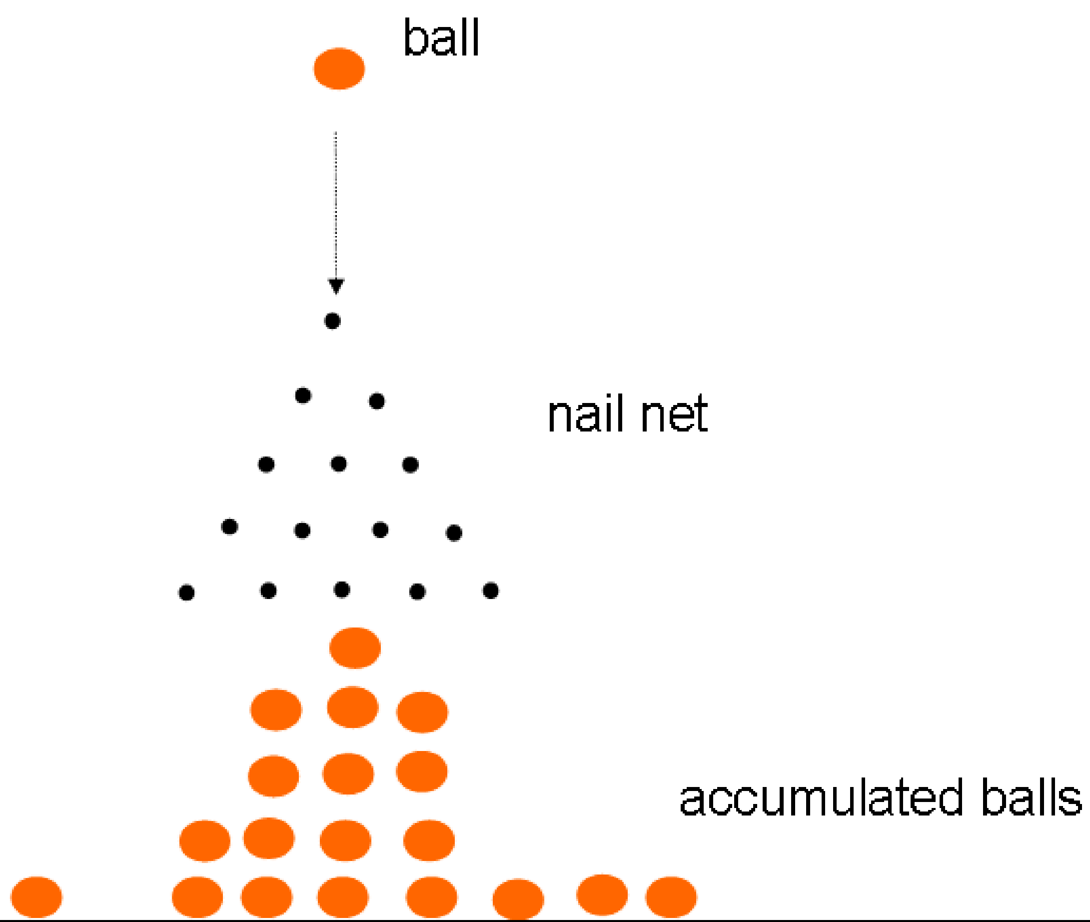
Random errors distribution – Gauss 3

Random errors distribution – measured values



Random errors distribution – DEMO

large number and equal size of elementary errors, random sign of errors



Consequences of normal distribution - 1

- the most probable value x_0 is an arithmetic average

$$x_0 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

1. the precision of the mean s is increasing with

$$s \approx \frac{1}{\sqrt{n}}$$

where

x_i	are the measured values
x_0	is a mean value
n	is a total No. of measurements

Consequences of normal distribution - 2

Example

Repeating the measurement 100 times, the random error of the resulting mean value will be 10 times lower.

The standard deviation σ may be estimated from the Root Mean Squares of the individual deviations,

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_0)^2}{n}}$$

where

x_i are the measured values

x_0 is a mean value

n is a total No. of measurements

Consequences of normal distribution - 3

By definition of probability :

$$\int_{-\infty}^{\infty} p(x) = 1$$

Assuming the p(x) for the normal distribution, one can evaluate, that

for a mean value = o

$$\int_{-3\sigma}^{3\sigma} p(x) > 0.99$$

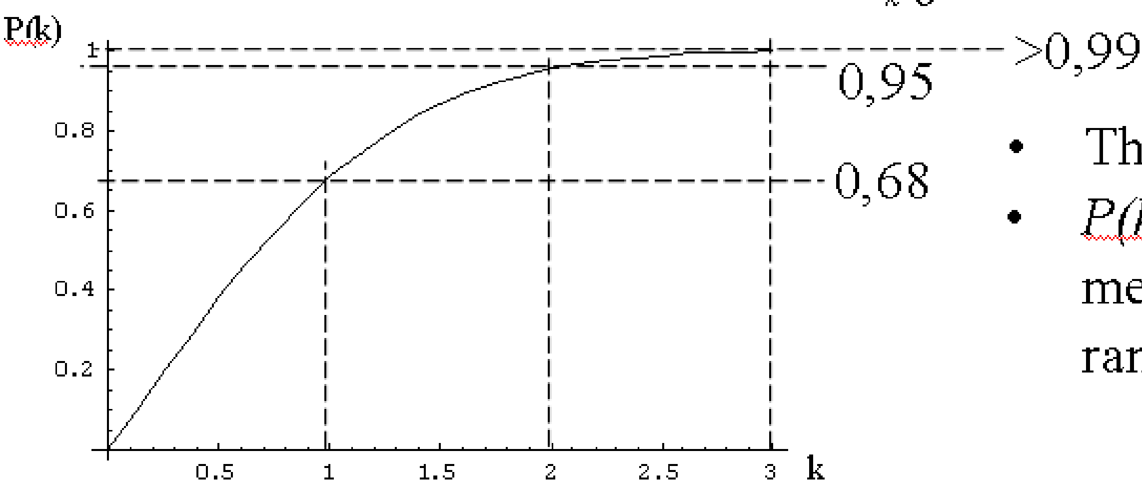
It means, that almost all the measured values (>99%) are within the limits +/- 3 σ.

Consequences :

1. the criterion +/- 3 σ may be used to separate the measurements from the raw errors / noise

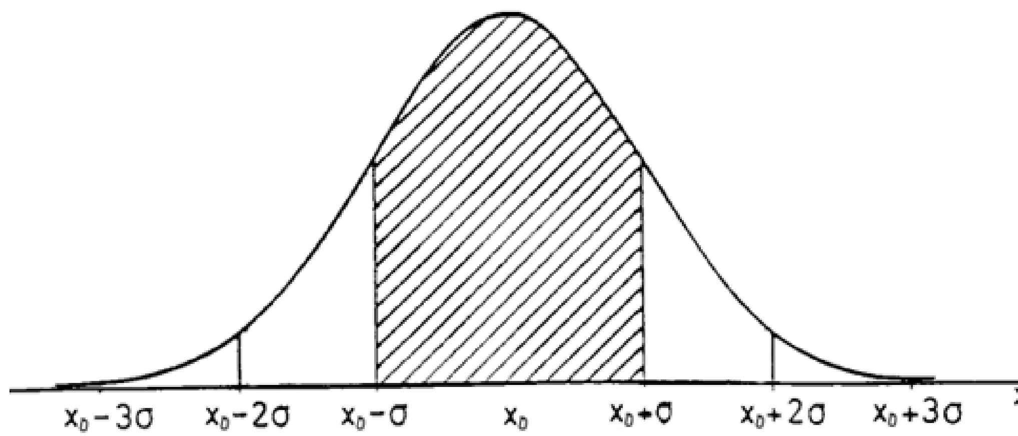
Consequences of normal distribution - 4

- Let's define probability $P(k) = \int_{-k\cdot\sigma}^{k\cdot\sigma} p(x)\,dx$



- The graph of *P(k)* for $k \in \langle 0, 3 \rangle$
- *P(k)* means probability, that measured value will be in the range $x_0 \pm k\sigma$

- The graph of *p(x)* – Gauss distribution of errors
- x_0 is mean value
- σ is precision (rms)



RANDOM ERRORS Example
Car manufacturing production – precision / accuracy

- Question — how precise / accurate (?) must be each component to guarantee that only < 1 / 1000 car will be not acceptable due to parts miss-match ?

- Problem — high precision / accuracy = > high manufacturing costs
  low precision / accuracy = > high repairs costs

- Solution — probability of off-tolerance component must be ~ 1 $*10^{-6}$

  = > probability of good comp. p(x) >= 0.999 999
  = > solve for integration limits k * sigma

  = > precision / accuracy of manufacturing must be about 6 times better than a limit, for which the parts fit

Consequences of normal distribution #5 Random Errors Averaging limits

- The precision of the mean value is increasing with SQR(N)
- BUT - How long ? What is the limit ?
- Answer - as long as the entire experiment is stable / reproducible

EXAMPLE Ocean level increase (~ 1 mm / year ??)

- Let's consider ocean waves ~ 1 m peak-peak, 10 seconds
- To get 1 mm precision, we have to average 1 million level readings, this would take 10 millions of seconds => > 100 days
- This will not work, ocean tides (6 hr, 12 hr, month,….), wind, ocean currents etc… would limit the final precision

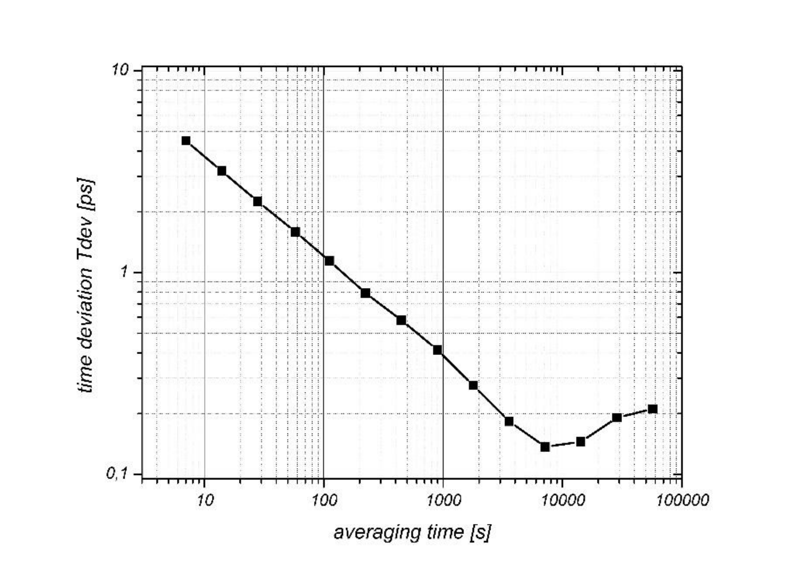
- In addition – the ACCURACY issue !
  Continental drift ~ 10 mm / year
  Invariant coordinates ?

Consequences of normal distribution #6
Random Errors Averaging limits

Allan variation - definition

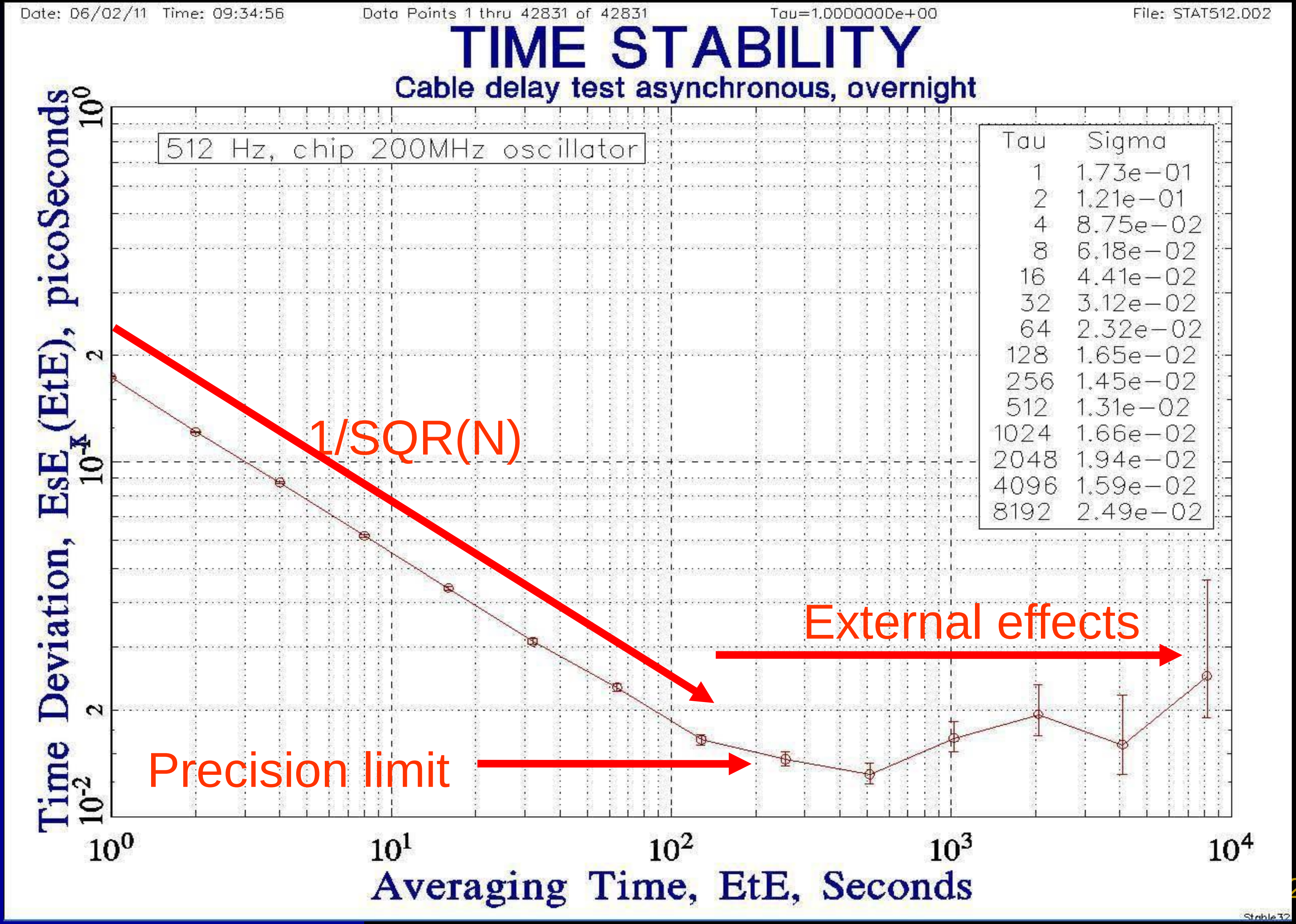
$$\sigma_y = \frac{1}{2(M-1)} \sum_{i=1}^{M-1} [y_{i+1} - y_i]^2$$

- *where y_i is i-th measurement, M is number of measured data*

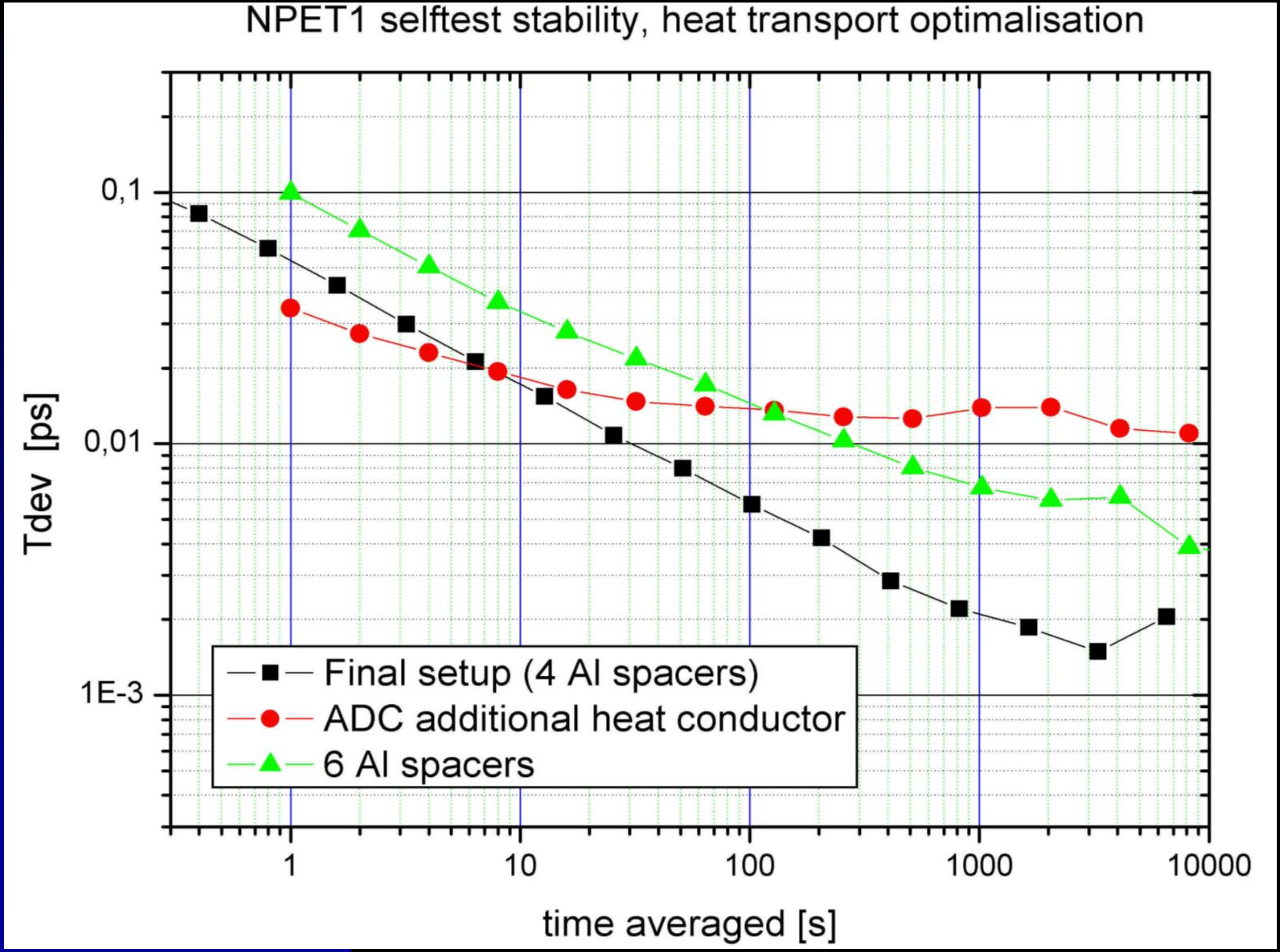


- log / log scale graph
- 1 / SQR(N) displayed as a line limitations clearly visible
- time and frequency measurements

Consequences of normal distribution #7
Allan variance example – time interval measurements



Consequences of normal distribution #7a
Allan variance example – time interval measurements



Consequences of normal distribution # 8
Precision of event counting

- Precision σ of the result of event counting may be estimated as

  $$\sigma = SQRT(n)$$

  where n is a count No.

- Consequence – accumulating more counts, higher precision of the result is obtained

- The counts outside the range

  $n +/- 3\sigma$

  indicate a new effect and vice versa

Consequences of normal distribution # 9
Precision of event counting - examples

- Referendum pools
  statistical sample, ~ 1800 respondents
  only 2 possibilities YES / NO , both ~ equal probability
  σ = SQRT(900) = 30 … => σ = 3.3%

- Consequence – the confidence of a pool with 1800 respondents is ~ 3% (one sigma).
  To predict as "almost sure (>99%)"
  the difference must be >= 10%

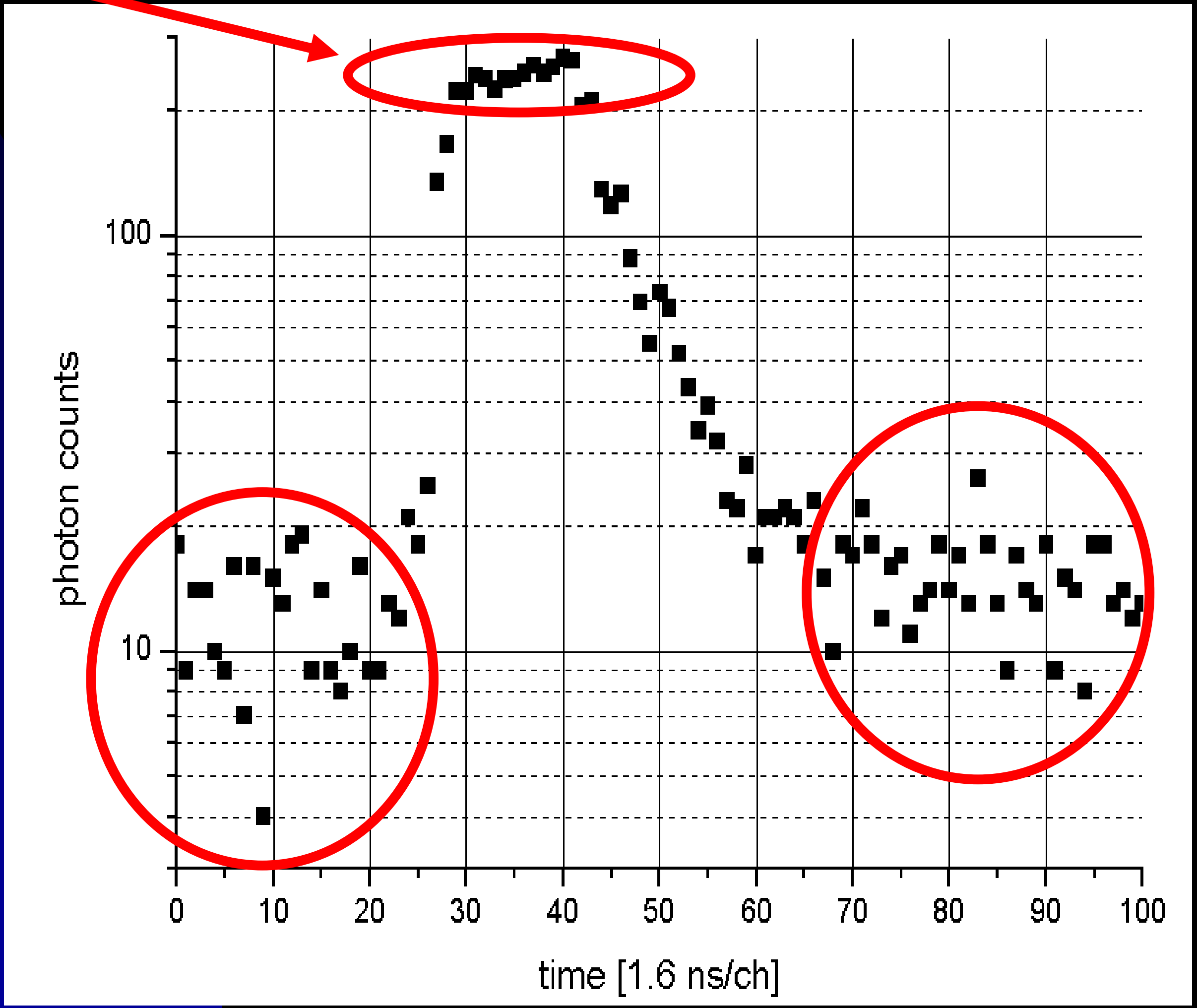
- Example – UK Wales "independence" referendum
  totally ~ 1.2 million voters
  results was 49.8 versus 50.2 %
  was it predictable ?

Consequences of normal distribution # 10
Precision of event counting - examples

Mean = 225

σ = 15 (6%)

Histogram of event counting



Mean = 11

σ = 3.3 (30%)

3 * σ = 10

Range (1,21)

Mean = 16

σ = 4 (25%)

3 * σ = 12

Range (4,28)

Consequences of normal distribution # 11
Precision of event counting - examples

The following vector is the result of the event counting measurements with random errors.
Identify numbers outside the range of 95% correctly identificated data.

51	29	50	50	56
52	42	51	49	48
65	48	56	52	47
49	49	70	36	50
49	51	51	46	49
50	48	46	51	52

1)X_0, mean estimation:

$$x_0 = \frac{\sum x_i}{n} \qquad x_0 = \frac{1493}{30} \cong 49{,}8$$

2)σ estimation:
event counting ⇒ $\sigma \approx \sqrt{x_0} \qquad \sigma = 7{,}05$

3)range estimation: 95% ≈ 2•σ = 14,1 range is from x_0-2•σ to x_0+2•σ
range: 35,7 - 63,9

4)data identification: numbers to throw away: **65, 29, 70**

Precision of a combined measurement

- Presumption
  - the result is a function of several independently measured quantities, each having normal distribution

$$y = y(x_1, x_2, \ldots, x_n)$$

- Standard deviation of the final distribution is then given by

$$\sigma(y) = \sqrt{\sum_{j=1}^{n} \left[\left(\frac{\partial y}{\partial x_j} \right)^2_{\langle x_j \rangle} \cdot \sigma^2(x_j) \right]}$$

where $\sigma(x_j)$ are standard deviations (RMS) of the particular quantities

Combined measurement 2 – Examples

1. Sum of difference (i.e. height of a chimney):

$$y = x_1 \pm x_2 \Longrightarrow \sigma(y) = \sqrt{\sigma^2(x_1) + \sigma^2(x_2)}$$

2. Product of ratio (i.e. surface of a rectangle):

$$y = x_1 \cdot x_2 \text{ or } y = \frac{x_1}{x_2} \Longrightarrow \sigma(y) = \langle y \rangle \cdot \sqrt{\left(\frac{\sigma(x_1)}{x_1}\right)^2 + \left(\frac{\sigma(x_2)}{x_2}\right)^2}$$

(the angle brackets ⟨ ⟩ symbolize an average value)

3. Multiplication by constant *C*

$$y = x \cdot C \Longrightarrow \sigma(y) = |C| \cdot \sigma(x)$$

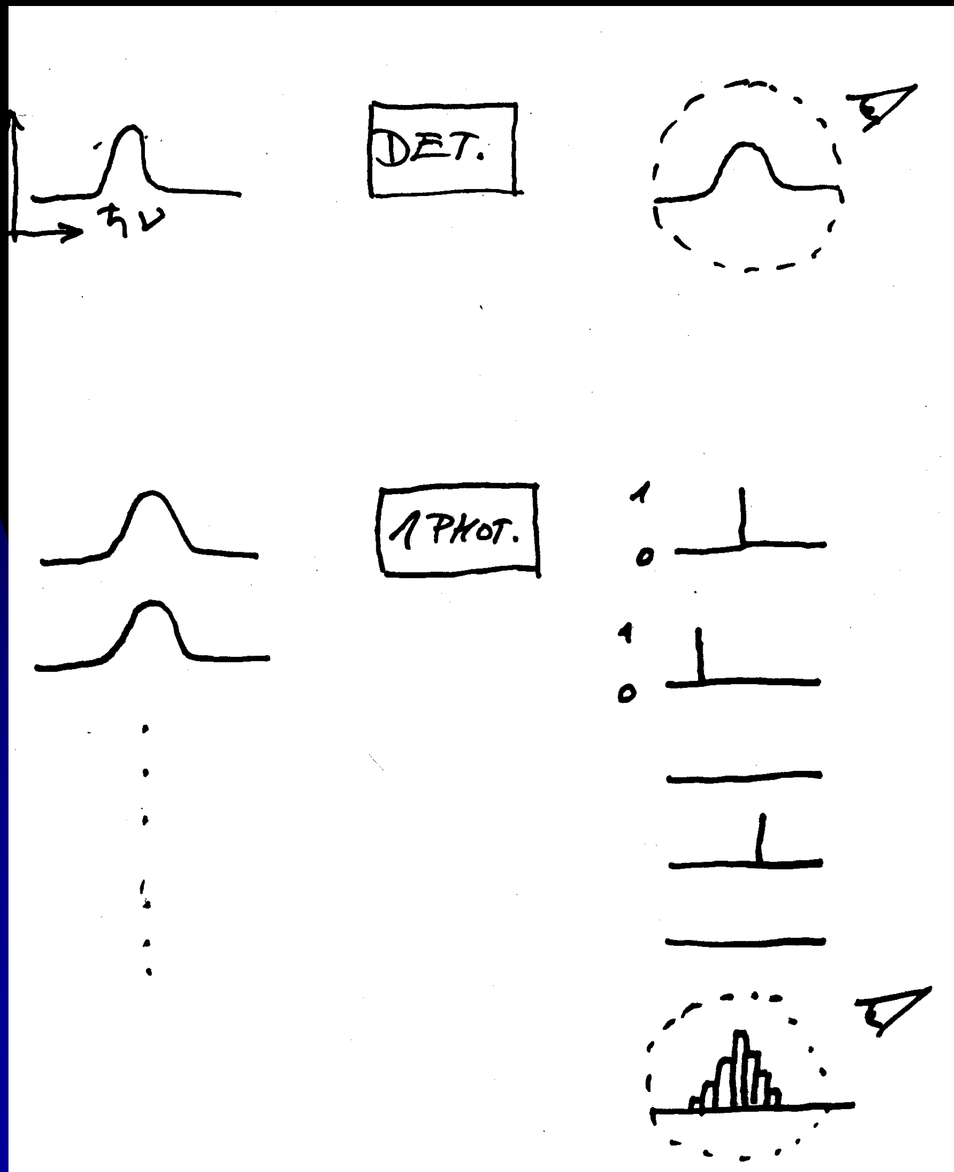
4. Variable to the power of *k*

$$y = x^k \Longrightarrow \sigma(y) = |k| \cdot \langle y \rangle \cdot \frac{\sigma(x)}{|x|}$$

Photon counting # 1

Intensity

time



“strong signal”

“single photon”

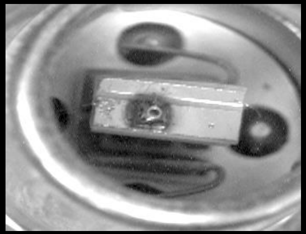
Photon counting detectors

- GENERAL LIMITATION - the dark count rate increases with area

- VACUUM / PHOTOCATHODE based
  apertures 3 mm up to ~ 1 meter
- dark count rate may be reduced by cooling

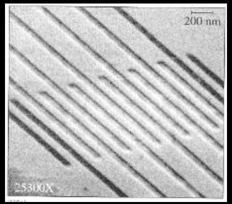
Hamamatsu 10 inch tube

- SEMICONDUCTING detectors
  apertures 5 µm up to 500 µm room / TE temp.
  up to 5 mm 77 K
- However, cooling impairs some detector parameters



Si SPAD 100um, TE cooled

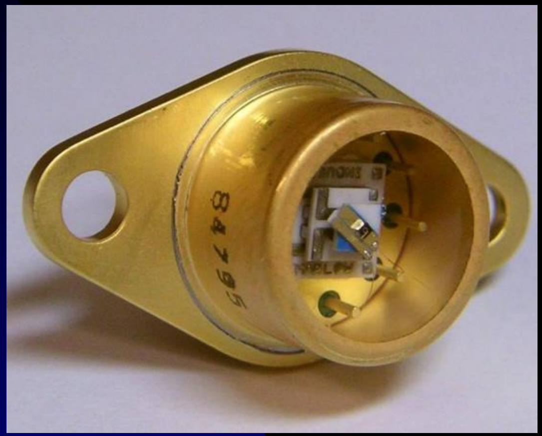
- SUPERCONDUCTING detectors
  apertures max. 10 µm



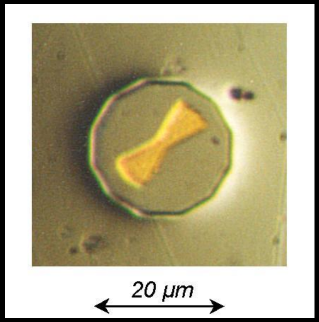
Superconducting detector 2 x 2 µm

Single Photon Detectors

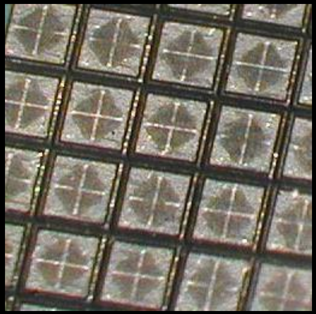
Several examples of SPAD structures made by CTU



Si ,200 um,TE3 cooled

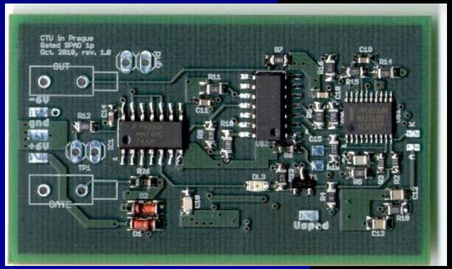


GaAs messa



GaAsP, 350 um

Active quenching and gating circuit



60 mm

Complete detector packages

130 mm

WHY Single photon detection ?

Not just „ .. higher sensitivity .. „

- quantum nature of light = > two states detected 0 / 1
- NO analog signal processing, inherently digital
  = > minimizing systematic errors
- Extremely low signals (<< 0.001 photon/pulse) detectable
  High dynamical range without degrading timing performance
- Measurement by-products optical signal intensity & shape precision

$$\sigma \approx \frac{1}{\sqrt{N}}$$

- picosecond resolution, sub-ps stability
- ps accuracy achievable
- Space qualified devices available

Photon counting data processing #1

1. **converting the probability of registration a photoelectron to a signal strength,**
2. **extracting the echo signal of interest from the backgound noise,** (10^{20})
3. **estimating the probability, that the extracted "signal" is a real, useful signal and not a result of a statistical nature of the noise.**

probability of detection p intensity I

$$p \approx I$$

Photon counting data processing # 2

$$p \approx I$$

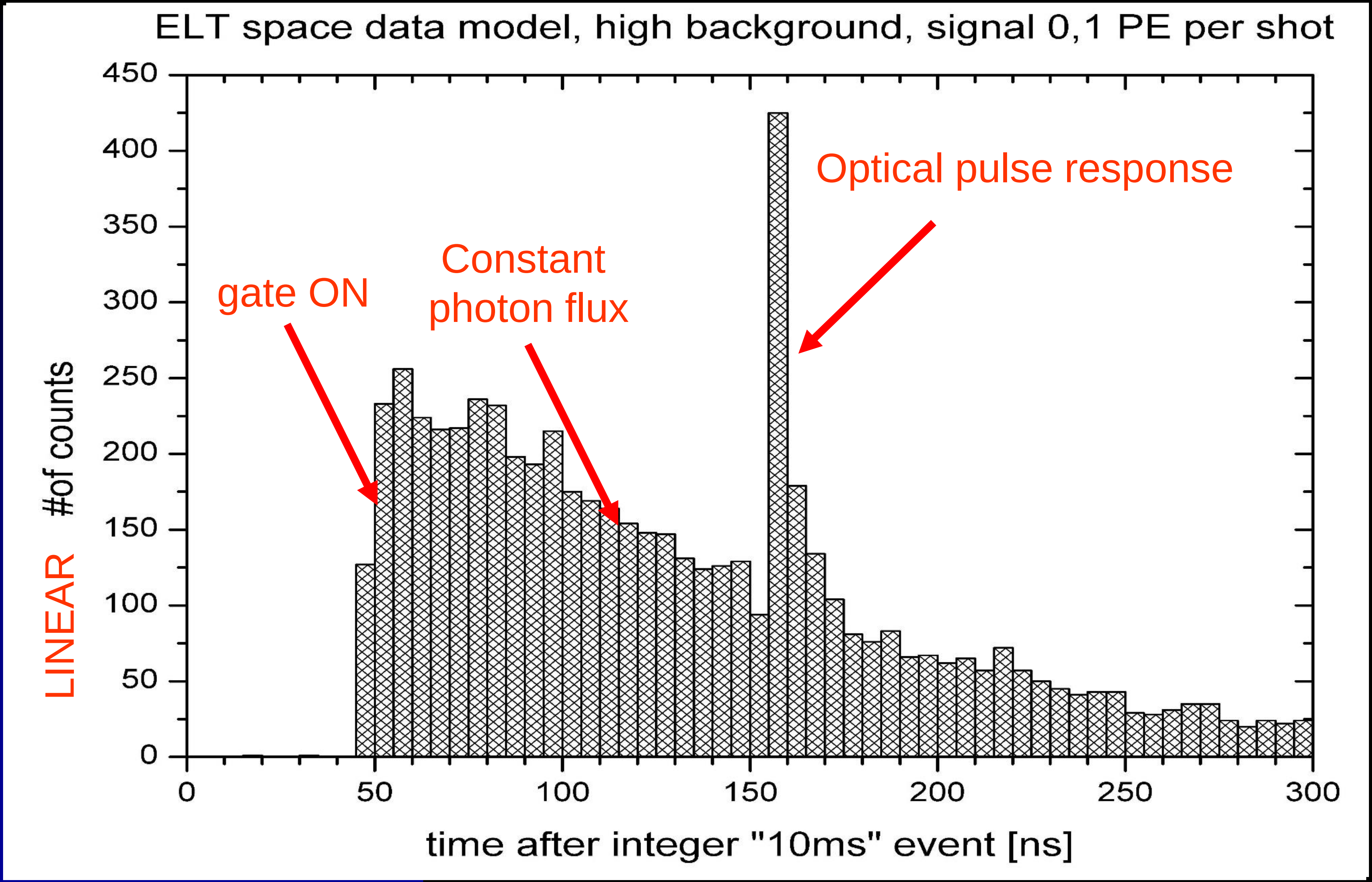
photoelectrons registered	**Ni**
number of all measurements, when the detection might occur	**Na**
total number of measurements	**N_{tot}**

$$p_i = \frac{N_i}{N_a}$$

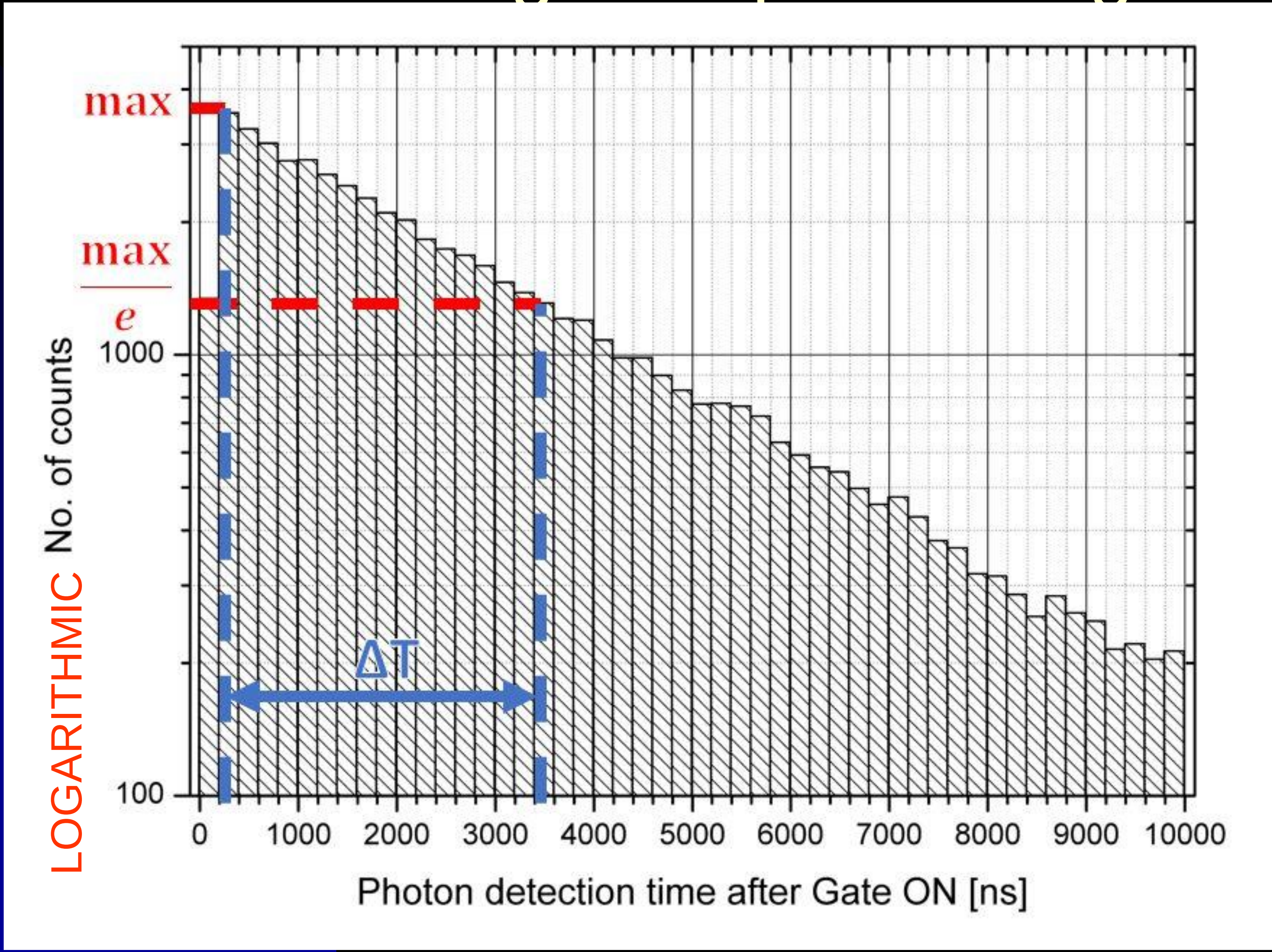
$$N_a = N_{tot} - \sum_{k=1}^{i-1} N_k$$

$$I_i \approx \frac{N_i}{N_{tot} - \sum_{k=1}^{i-1} N_k}$$

Photon counting data processing # 3



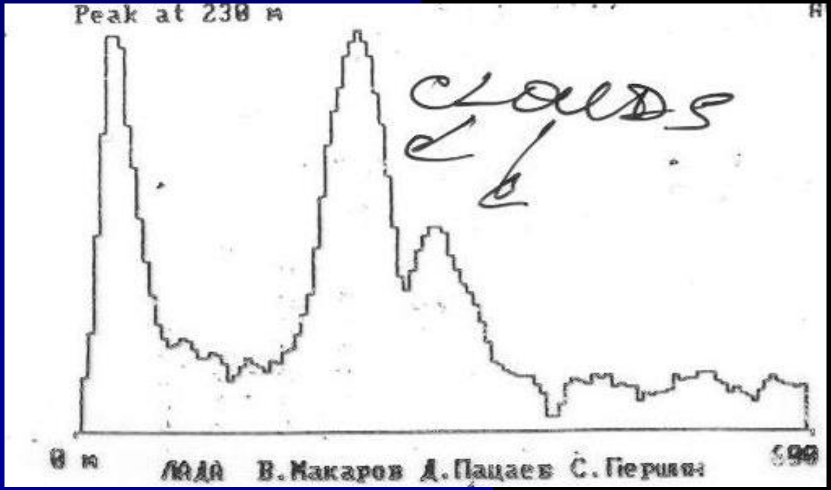
Photon counting data processing # 4



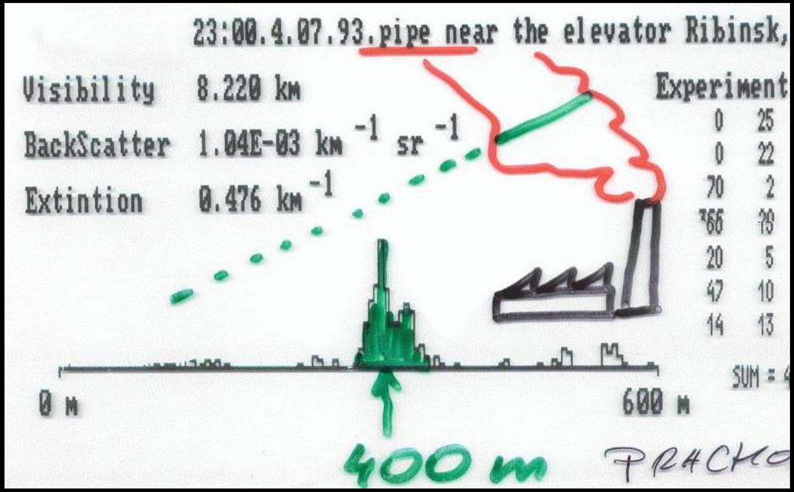
Application # 6 **Photon counting LIDAR**

- **Photon counting LIDAR**
- for metrology and ecology
- laser diode transmitter
- photon counting detector Si

Cloud height monitoring
air traffic control



Air pollution propagation monitoring
ecology



Photon counting LIDAR data processing # 1

LASER pulses: 1024000 SPAD pulses: 236319(N), 236222(N+S), Gate 4mks

daylight 3mm aperture 1msample tree+atmosphere

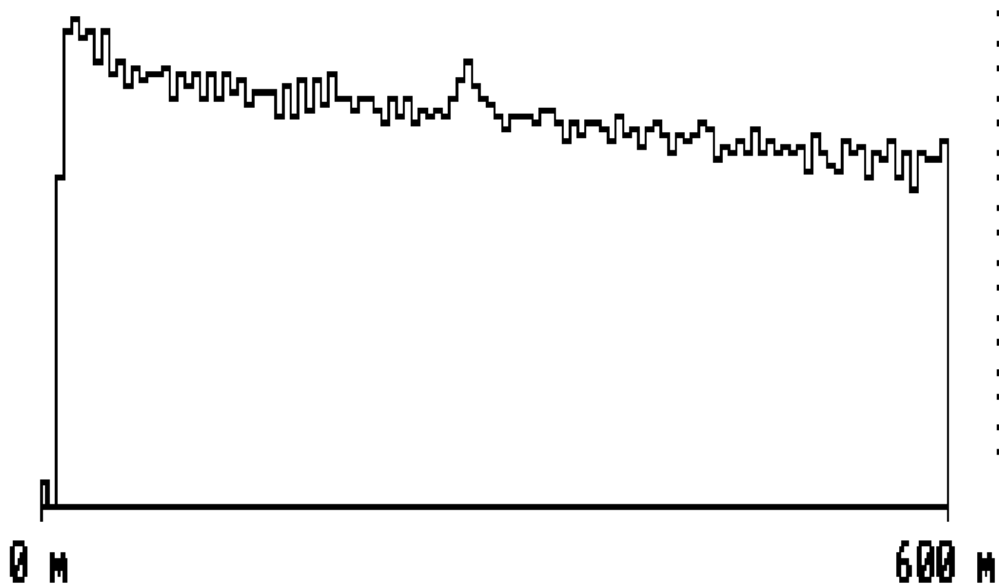
Visibility	BackScattering	Extintion
0.846 KM	1.29E-01 KM^{-1} SR^{-1}	4.623 KM^{-1}

CURSOR	CHANNEL	DISTANCE	COUNTER
LEFT	0	0 M	140
PEAK	56	280 M	2279

TOTAL COUNTER SUM = 236319

EXPERIMENT SIGNAL+NOISE

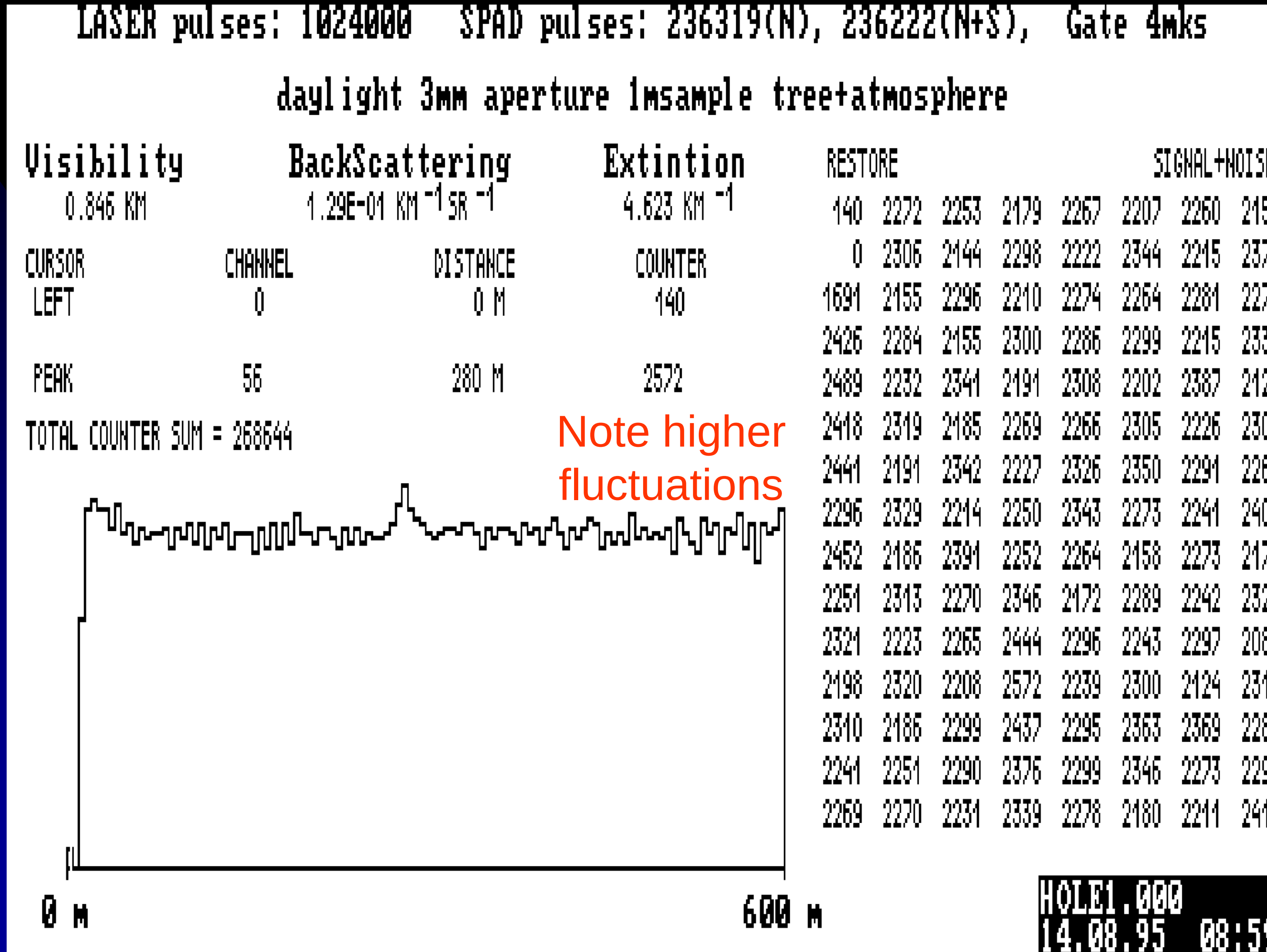
140	2204	2115	1979	1990	1874	1856	1709
0	2232	2008	2082	1946	1986	1815	1884
1689	2081	2146	1998	1988	1914	1865	1796
2419	2201	2010	2075	1994	1939	1807	1844
2476	2146	2178	1972	2008	1853	1943	1677
2399	2225	2029	2038	1967	1936	1808	1813
2416	2098	2170	1996	2015	1969	1857	1778
2268	2225	2047	2012	2025	1900	1812	1878
2416	2084	2205	2010	1952	1800	1834	1700
2213	2200	2089	2089	1869	1905	1805	1809
2277	2110	2080	2171	1971	1863	1845	1618
2151	2197	2023	2279	1918	1906	1703	1798
2256	2065	2102	2154	1962	1954	1895	1766
2184	2122	2089	2095	1961	1935	1814	1770
2206	2135	2030	2058	1939	1794	1761	1862



HOLE1.000
14.08.95 08:59

Photon count No vers. range

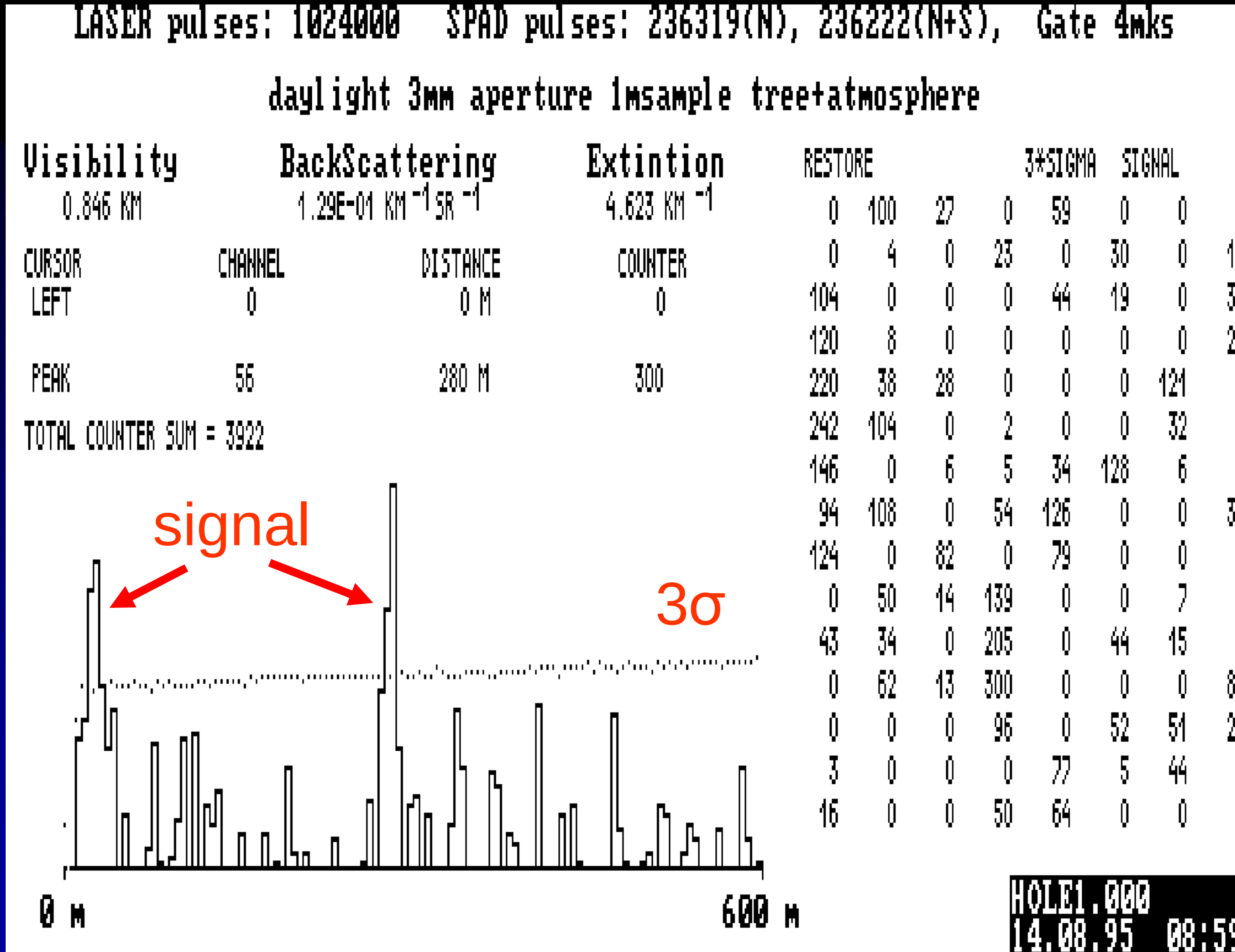
Photon counting LIDAR data processing # 2



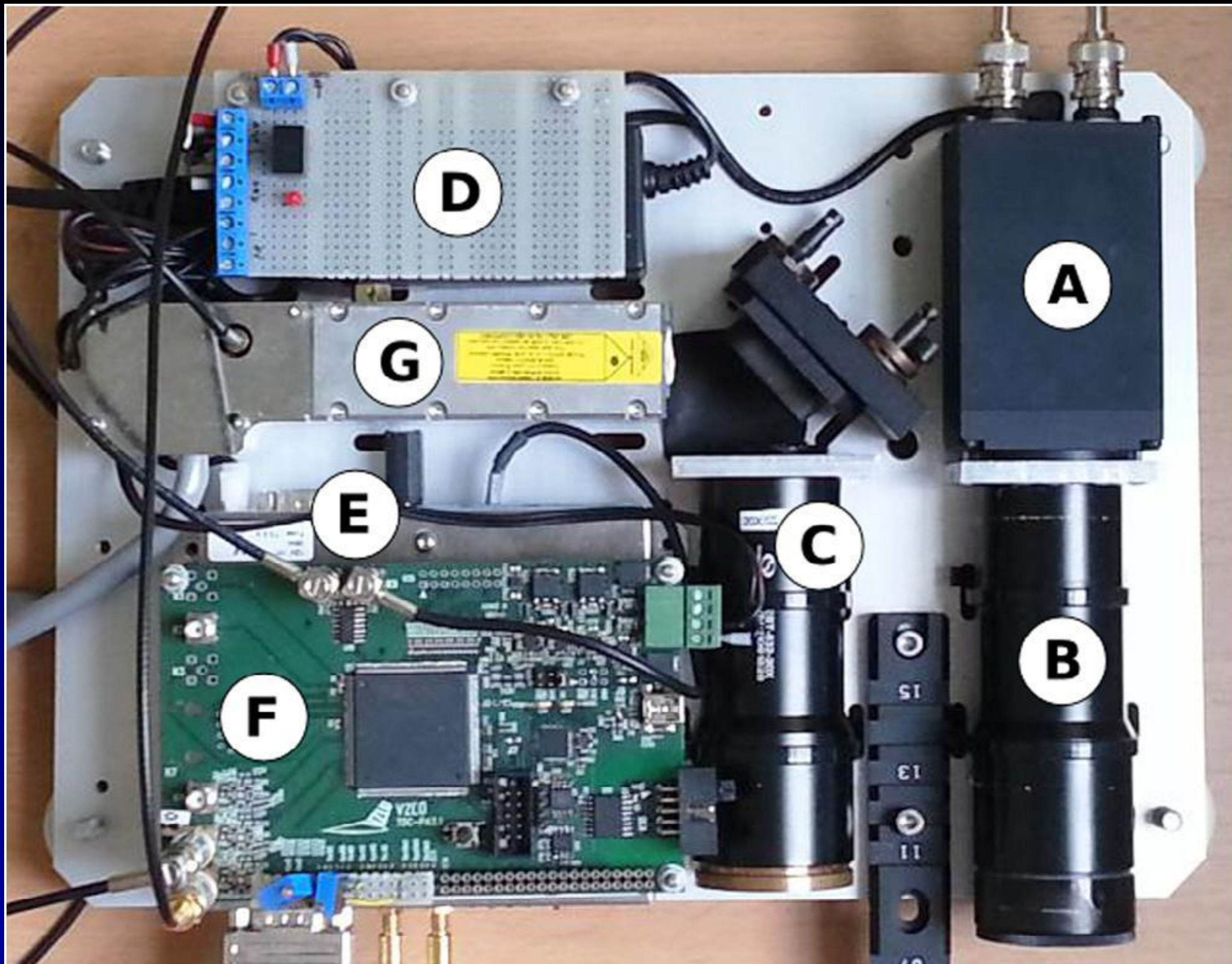
Intensity (probability) versus range

Photon counting LIDAR data processing # 3

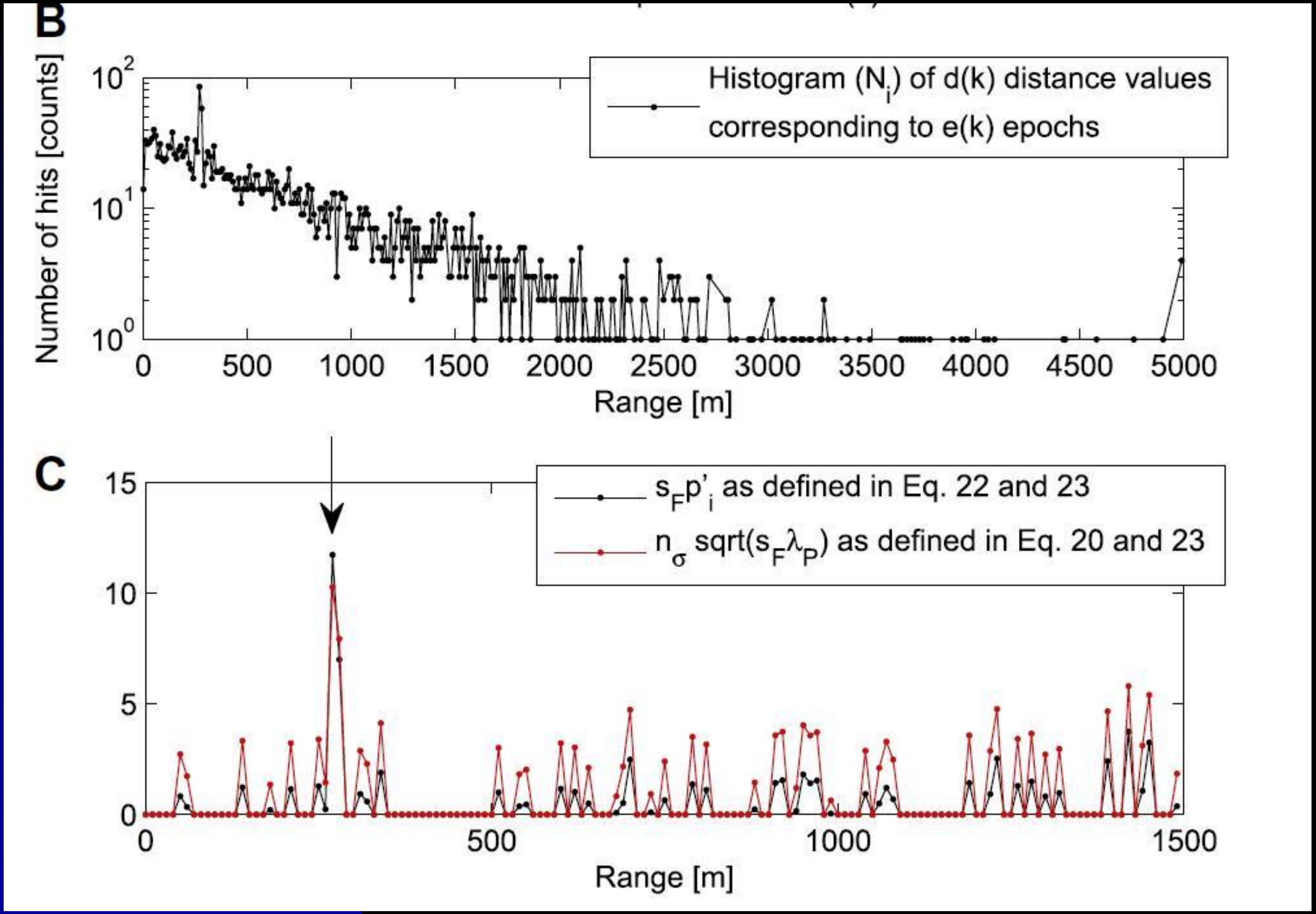
Intensity (probability) versus range



Photon Counting Lidar for planetary missions
Demo unit, CTU in Prague



Photon Counting Lidar for planetary missions
Demo unit, CTU in Prague



Data fitting and smoothing

- APPLICATION
- Repeated measurements of slowly varying effects
- (optionally) investigation of their dependence on unknown parameters

- GOALS
- Data smoothing : random errors reduction / precision increase / precision estimate
- Indirect measurement : determination of unknown parameters on the basis of a single variable measurements

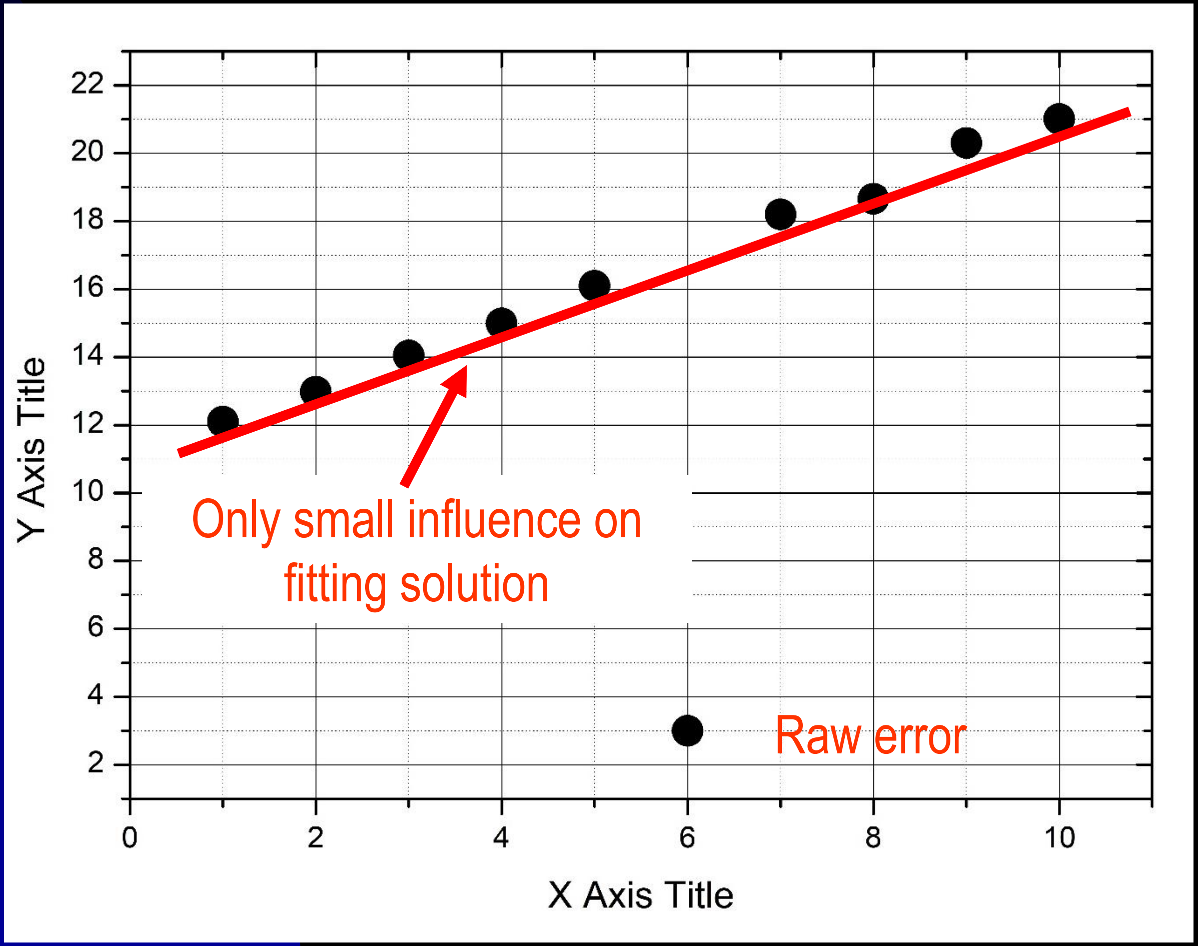
Data fitting and smoothing #2

- “Best fit”
- least square fit (> 90% of cases)
  minimum of sum of squares
- mini-max fit
  minimum of maximal deviation
  Chebychev polynom solution
- and many other
  weighted average...

Data fitting and smoothing #2a

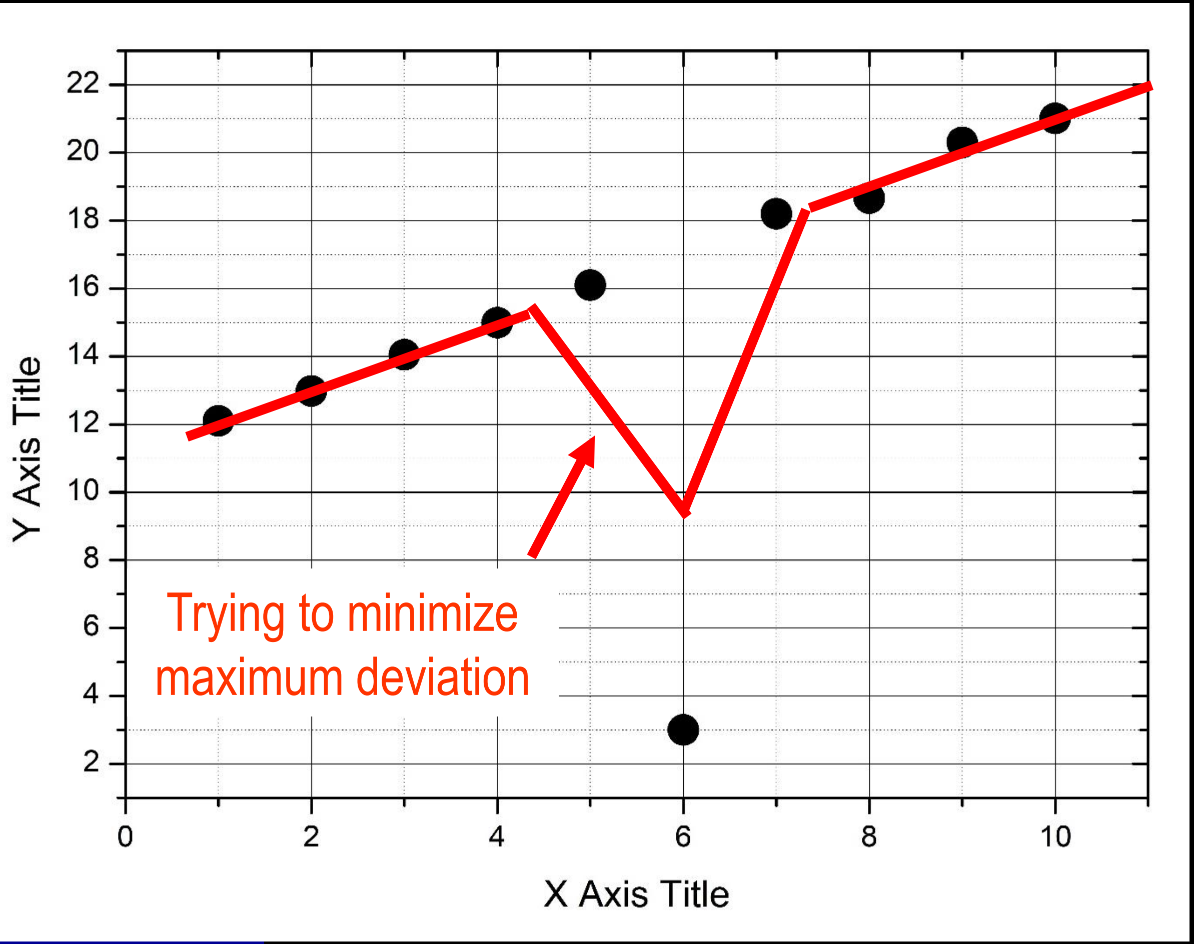
least square fit (> 90% of cases)

minimum of sum of squares of (o-c) residuals



Data fitting and smoothing #2b

mini-max fit : minimum of maximal deviation
Chebychev polynom solution



Data fitting and smoothing # 3

- TYPE of SOLUTION
- 1. known type of dependence
  F(a,b,c…, t)
  where F( ) is a known function
  a,b,c… are known with a limited precision
- Example
  motion equation, heat transport, electric circuit, …
- 2. un-known type of dependence

Data fitting and smoothing # 4

- SOLUTION STABILITY
- Well x ill defined parameters (correlated)
- parameter selection
- consequent increase of number of parameters

- STABILITY ROUGH ESTIMATE
- create two (interleaved) sub-sets of data compare the solutions

Data fitting and smoothing # 5

- MARGINAL EFFECTS IDENTIFICATION
- If the residuals after fitting with a function F indicate significant dependence, it indicates the presence of an effect, which is not described by the function F.
- Example
  F … dependence of a height of a snow man as a function of temperature and sunshine.
  … It is not predicting the heights increase :-)

Least squares fit – Normal Equations

Least square fit definition

$$\Sigma_i [F(a_1 + \Delta a_1, a_2 + \Delta a_2, ..., a_n + \Delta a_n, t) - M_i]^2 \text{ -> minimum}$$

$$(A)(B) = (C)$$

(A) ... square matrix of the n x n dimension

(B) ... vector of desired elements corrections

(C) ... n dimension vector

$$A_{jk} = {}_i\Sigma_1^N \ (\delta F/\delta a_j)_i \ (\delta F/\delta a_k)_i$$

$$C_j = {}_i\Sigma_1^N \ [M_i - F(a_1, ..., t)_i] \ (\delta F/\delta a_j)_i$$

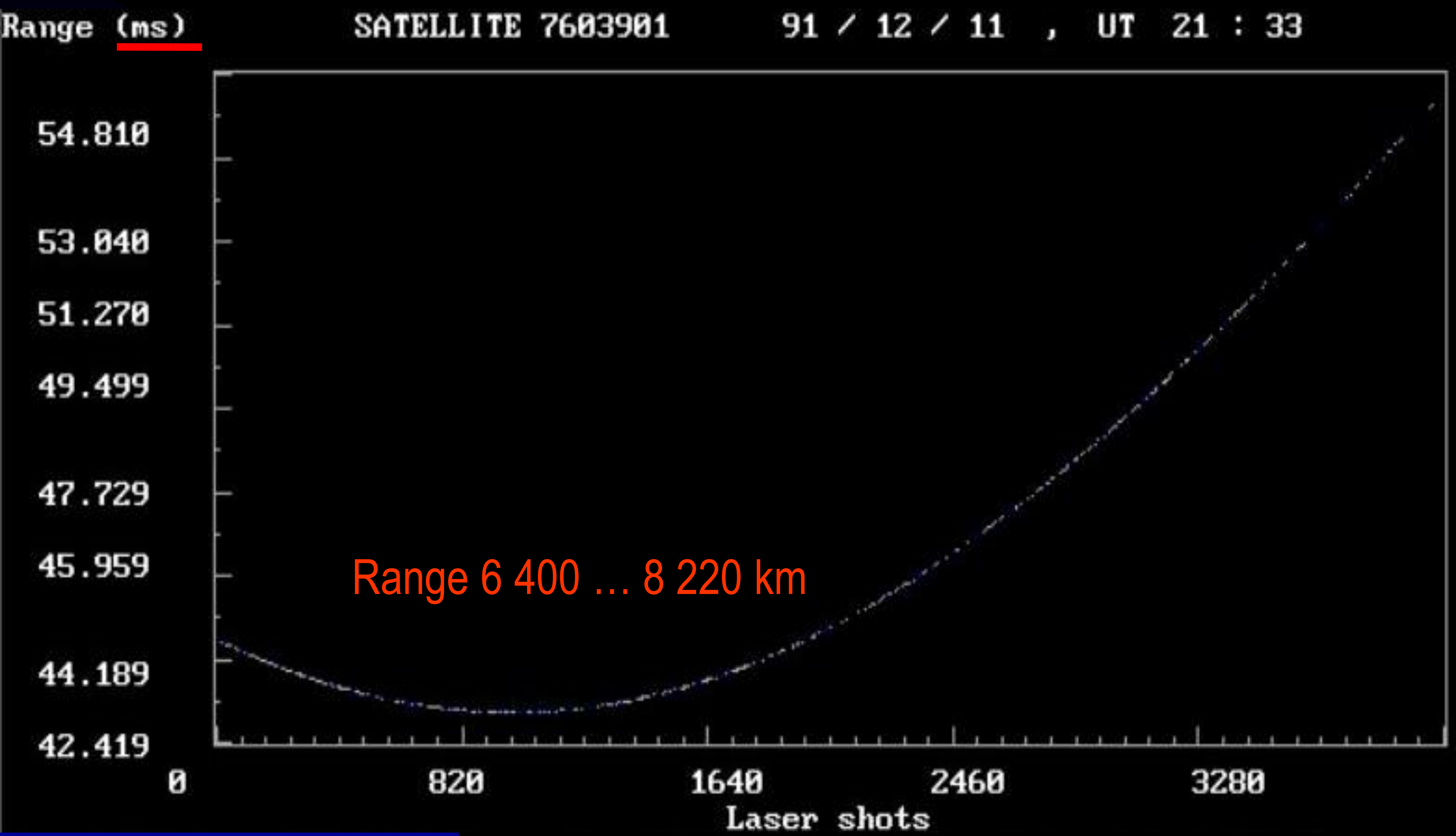
$$(\delta F/\delta a_j)_i = [F(a_1, a_2, ..., a_j + d_j, ..., a_n, t)_i - F(a_1, ..., a_n, t)_i] / d_j$$

Results are correction of parameters

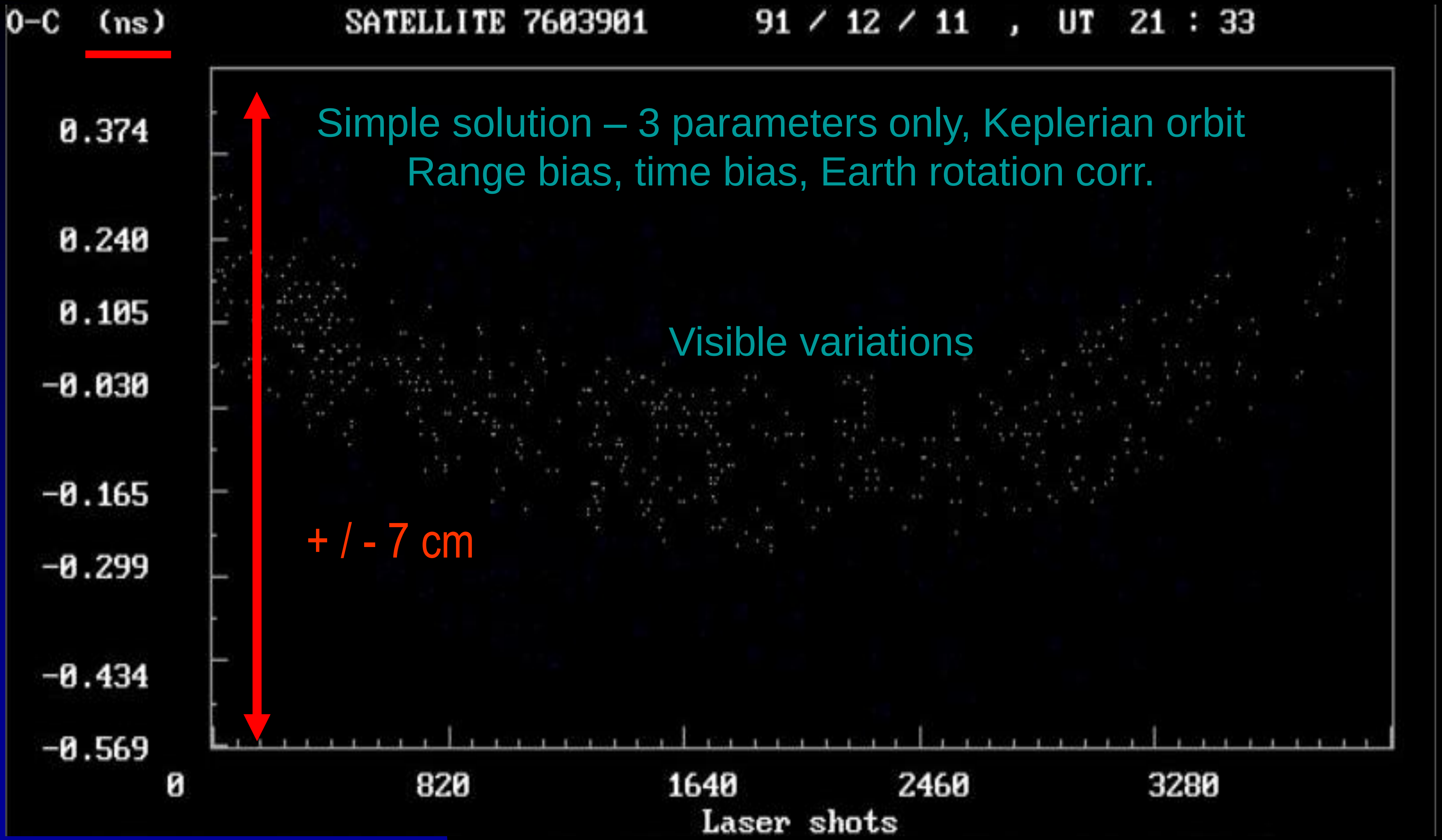
Root Mean Square – data scatter

- Where
- F_i is the fitting function value in the i-th point
  - x_i is the i-th data point
  - n is the total number of data points
  - k is the number of (solved for) parameter

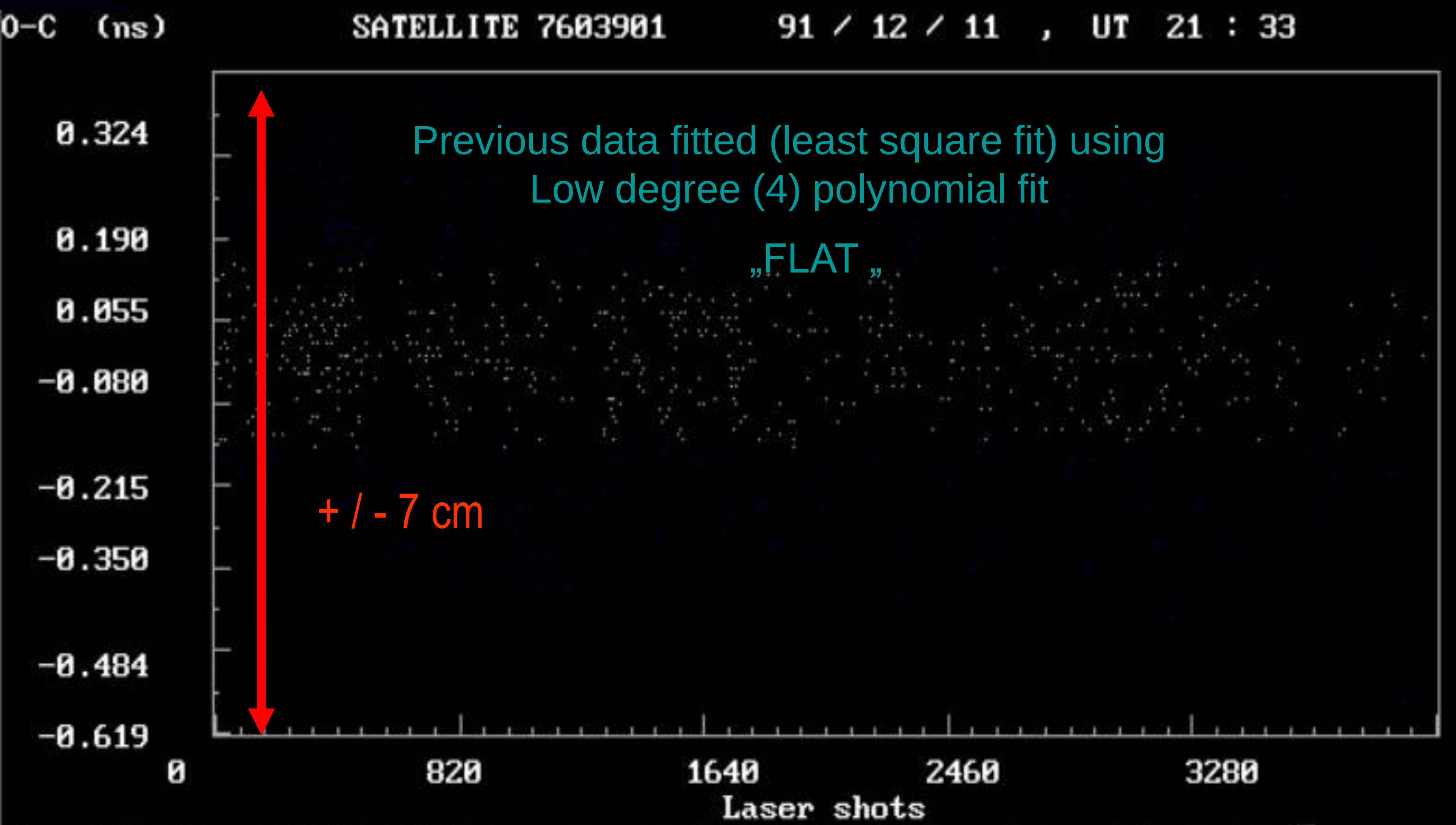
Data fitting and smoothing
SLR – raw data



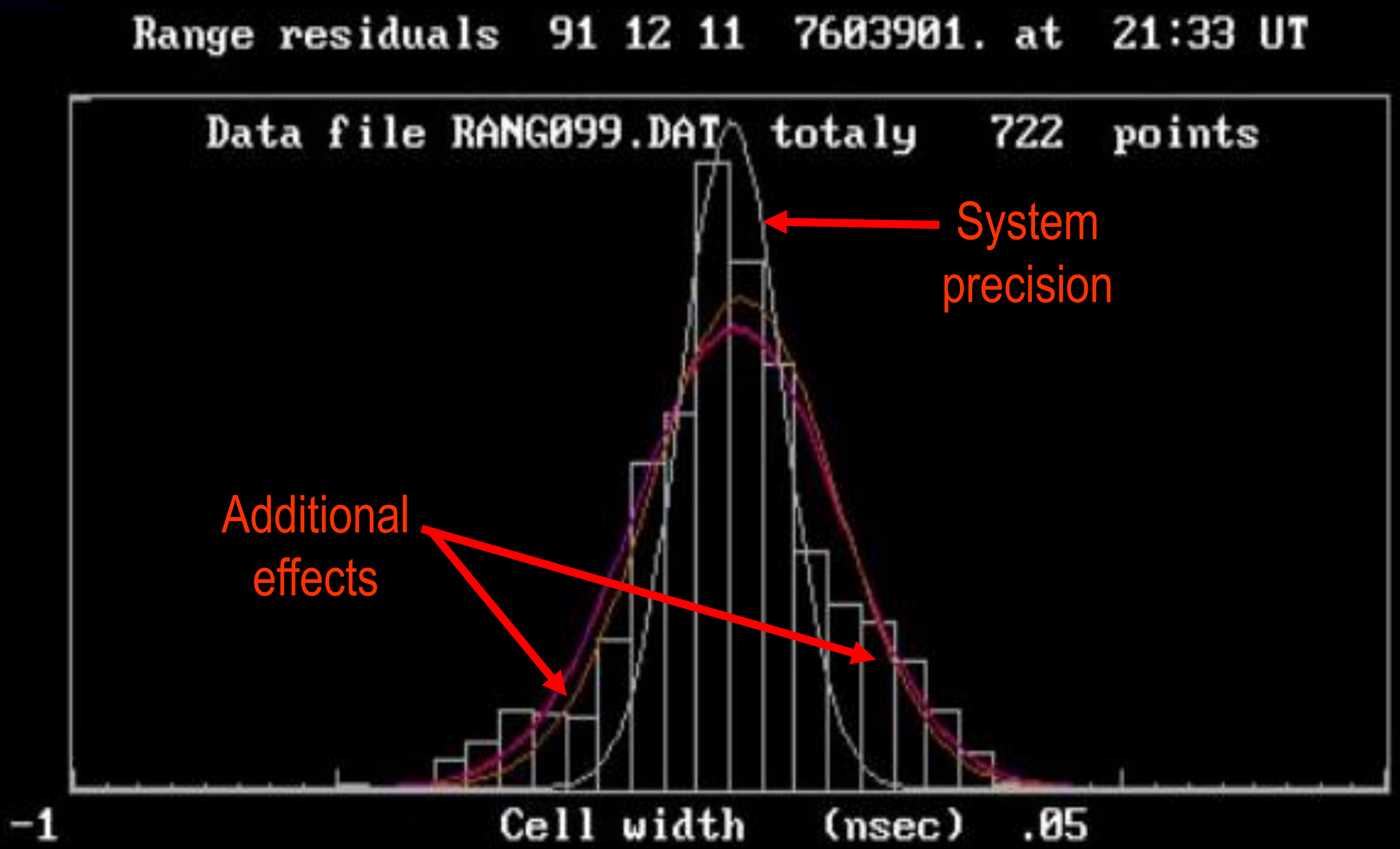
SLR data – Least square fit improved orbit



SLR data - Least square fit orbit + polynom

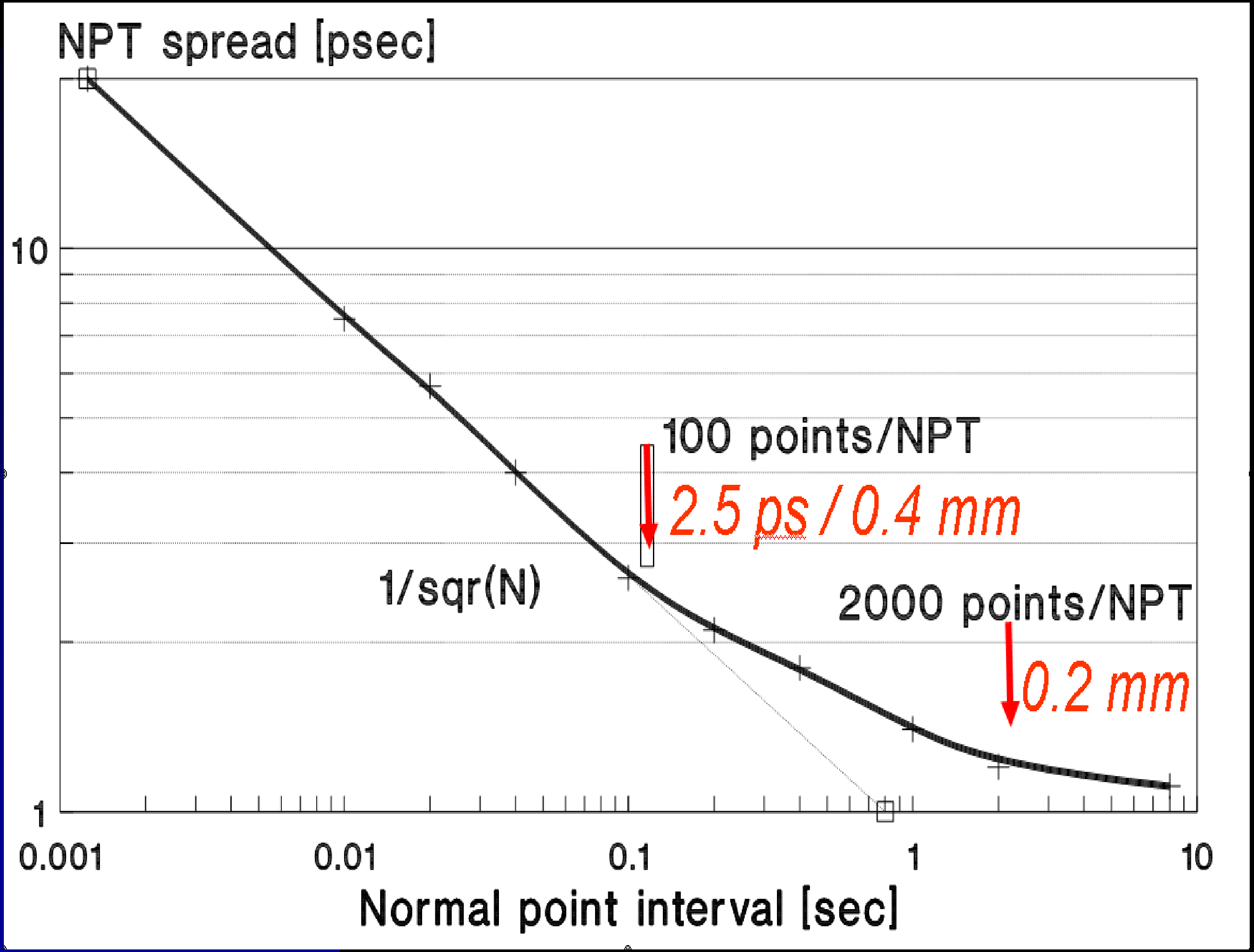


Histogram of final data fit



Data fitting and smoothing

SLR – fitted data averaging



Empiric rules for the best fitting polynom

- General
  The polynom degree should be as low as it fits the data "good"
  (It fits the data with the lowest possible RMS …)

- Strict limitation
  $M < 10$ unless special procedures are applied

- Number of points
  $M << N$ and / or $M^2 < N$
  M is the degree of the polynom and N is the number of points

- wide gaps in the data series:
  A is the width of gasp, B is the width of all range of data
  If A : B is high then $M \leq B / A$

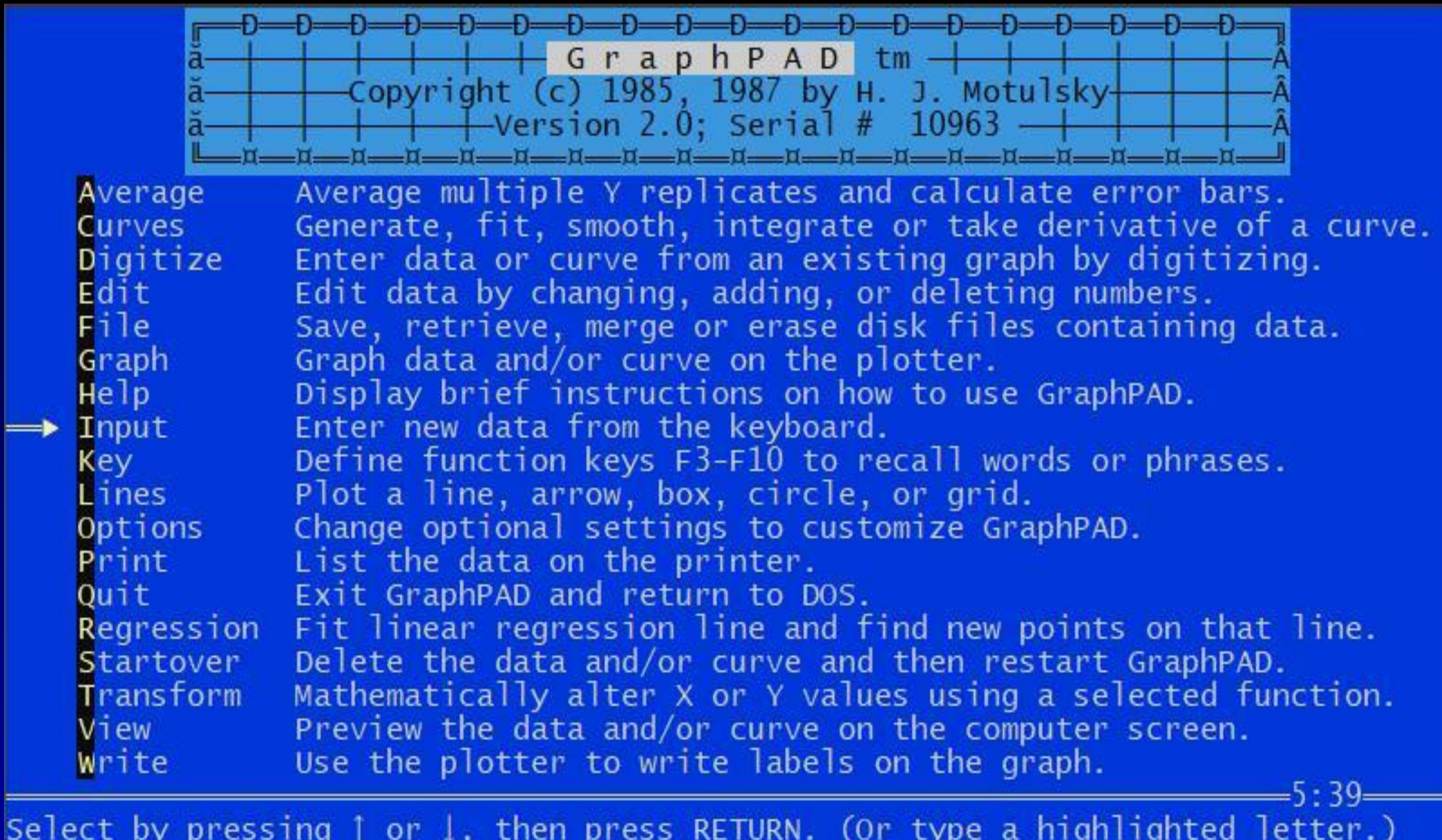
- Serial Correlation Coefficient $SCC \leq 0.5$

Data fitting and smoothing
Example # 1

- Experimental data - telescope pointing error in elevation „y“ [arc min]

- Measured for various azimuth „x“ 0 – 358 degrees
- 1. What is a good type of fitting function ?
- 2. Find the parameters of this fitting function

Data fitting and smoothing
Example # 1



Data fitting and smoothing
Example # 1

```
   Equation Menu:   (Press ESC for Main Menu.)
  A: Exponential decay
         Y = A*exp(-B*X) +C*exp(-D*X) + E
  B: Exponential association
         Y = A*[1-exp(-B*X)] + C*[1-exp(-D*X)] +E
  C: Exponential growth
         Y = A*exp(B*X) + C*exp(D*X)
  D: Rectangular hyperbola (binding isotherm)
         Y = A*X/(B + X)
  E: Double rectangular hyperbola
         Y = A*X/(B + X)  + C*X/(D + X) + E*X
  F: Sigmoid curve (log scale)
         A=bottom, B=top, C=log(EC50), D=`Hill' Slope
  G: Competition curve (log scale)
         A=bottom, B=top, C=log(EC50)
  H: Competition curve, 2 components (log scale)
         A=bottom, B=top, C=% site 1, D&E=log(EC50s)
  I: Sine wave
         Y= A + B * sin(C*X + D)
  J: Polynomial (1)
         Y= A*X^B + C*X^D + E
  K: Polynomial (2)   [This polynomial equation creates a `generic' curve.]
         Y = A +B*X + C*X^2 + D*X^3 + E*X^4
  L: Mixed exponential
         Y=A*exp(-B*X) + C*[1 - exp(-D*X)] + E*X
```

Data fitting and smoothing
Example # 1

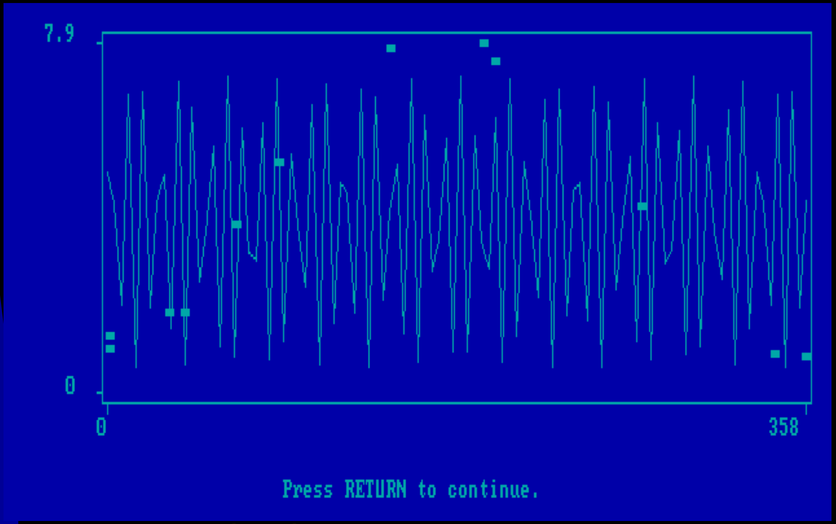
```
Sine wave
                   Y= A + B * sin(C*X + D)
-Enter estimates; the values will be changed later.



  Enter A: [ 3.856051]  3.856051
  Enter B: [-3.275097] -3.275097
  Enter C: [ 1.00096]  1.00096
  Enter D: [-0.348034] -0.348034

Enter the range of X values over which the curve is to be plotted.:
  (The X values of the data range from  1  to  358  .)
  Minimum X: [ 0 ]  0
  Maximum X: [ 358 ]  358

Do you want to view the curve generated by your estimates (Yes/No)? [No]
```



Example # 1

- Where was the mistake?
- The data seemed to be periodical, but the fit output is total nonsense
- We forgot to input information of the period we expect !
- USE EVERY SINGLE BIT OF INFORMATION YOU HAVE
- Let's try once more including this information...
(period coefficient estimate is ~ 0.017 deg/rad)

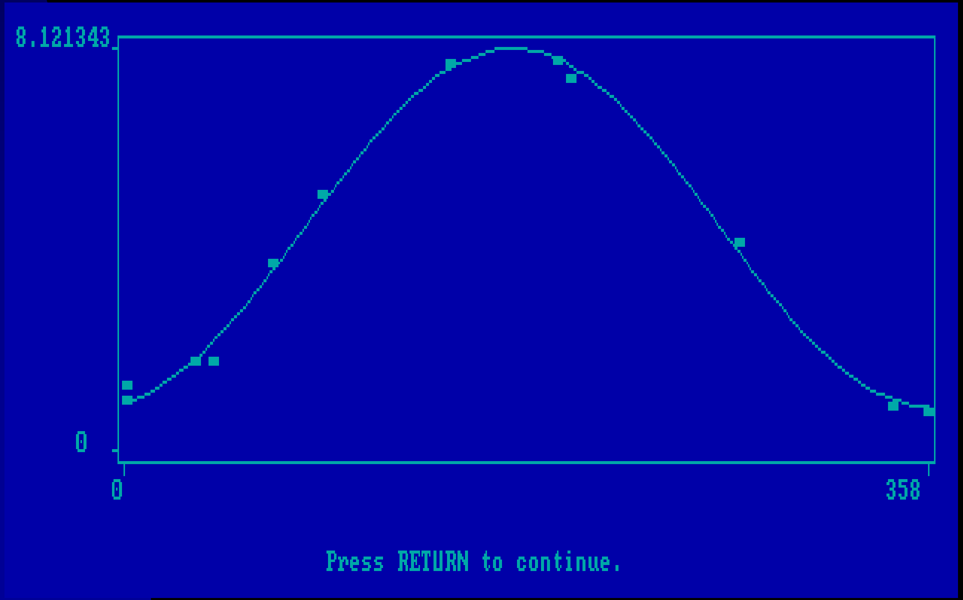
Data fitting and smoothing
Example # 1

Sine wave
Y= A + B * sin(C*X + D)

Final Results. Sum of Squares= 0.511746 (df= 9)
Goodness-of-fit assessed using actual distances; r squared=0.994.

Parameter	Value	*Std. Error	%Error
A	4.51084	.0745794	1.7%
B	-3.610954	.0934589	2.6%
C	.017	(Constant)	
D	1.777424	.0327	1.8%

* The std. error values are estimates. Don't use for calculating statistics.



OK

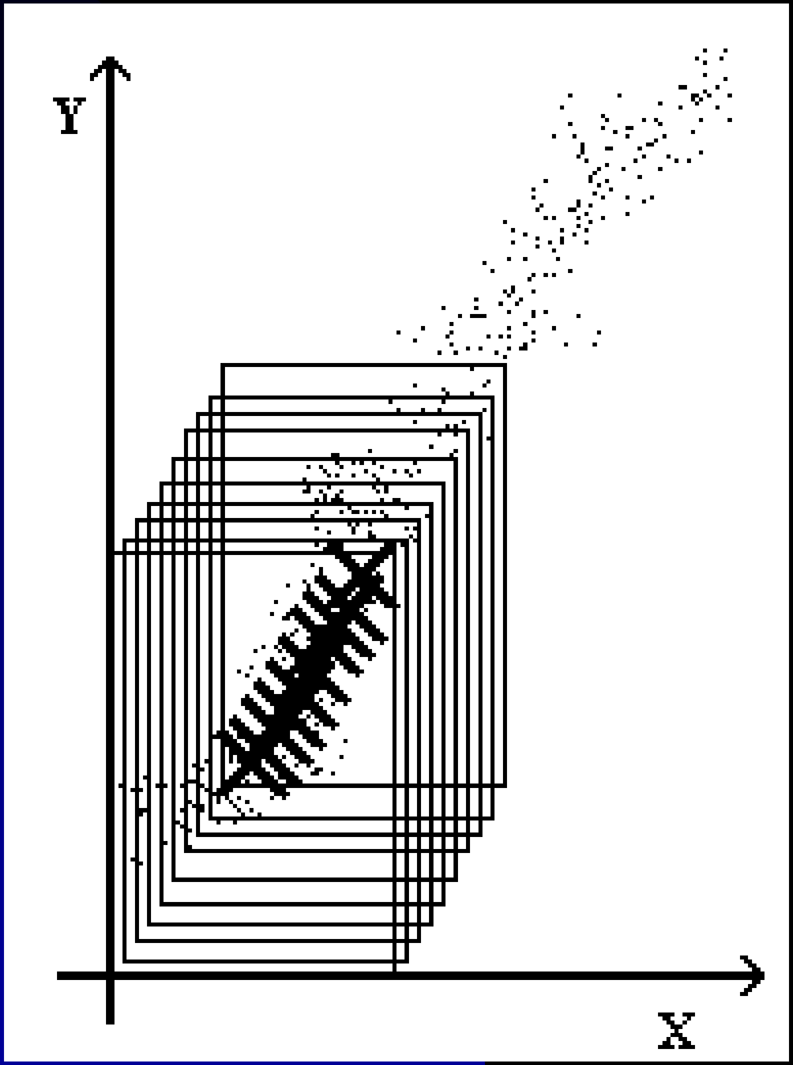
Data fitting and smoothing
Example # 1

- SERIOUS CONCLUSIONs
- USE EVERY SINGLE BIT OF INFORMATION YOU HAVE
- The initial parameter estimate is critical for correct solution
- with only one exception – which type of fitting function ?
- Ordinary polynom – the polynom parameters are direct solution of normal equations

Data fitting and smoothing
Moving average

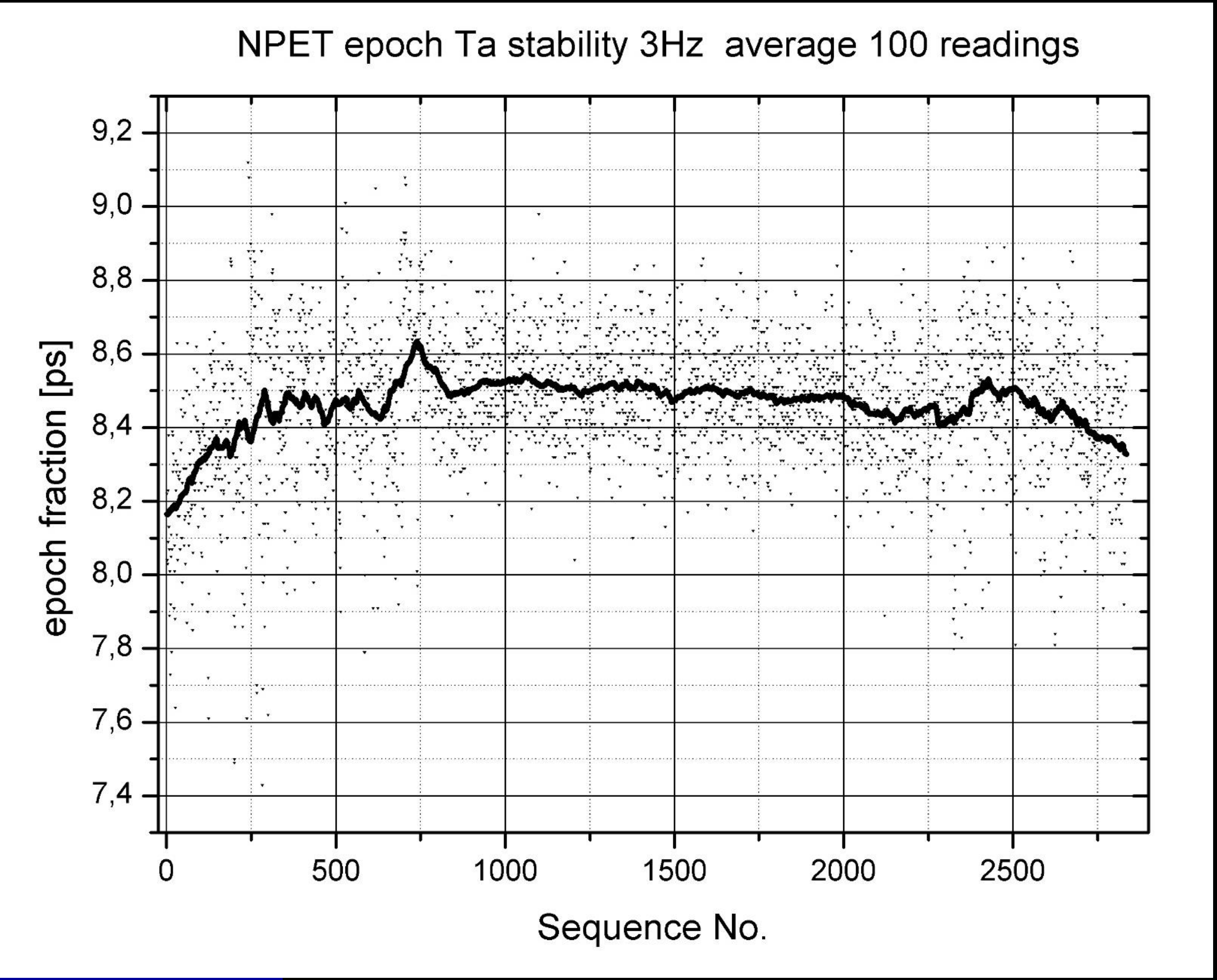
- simple method to smooth / fit a series of equidistant data
- moving average in the i-th interval = mean of the values in the interval <i-k, i+k>, where k is an positive integer
- spread inside the window is 1/SQR(n) smaller than original one
- various definitions of moving average value on both the ends of the interval

Data fitting and smoothing
Moving average #2

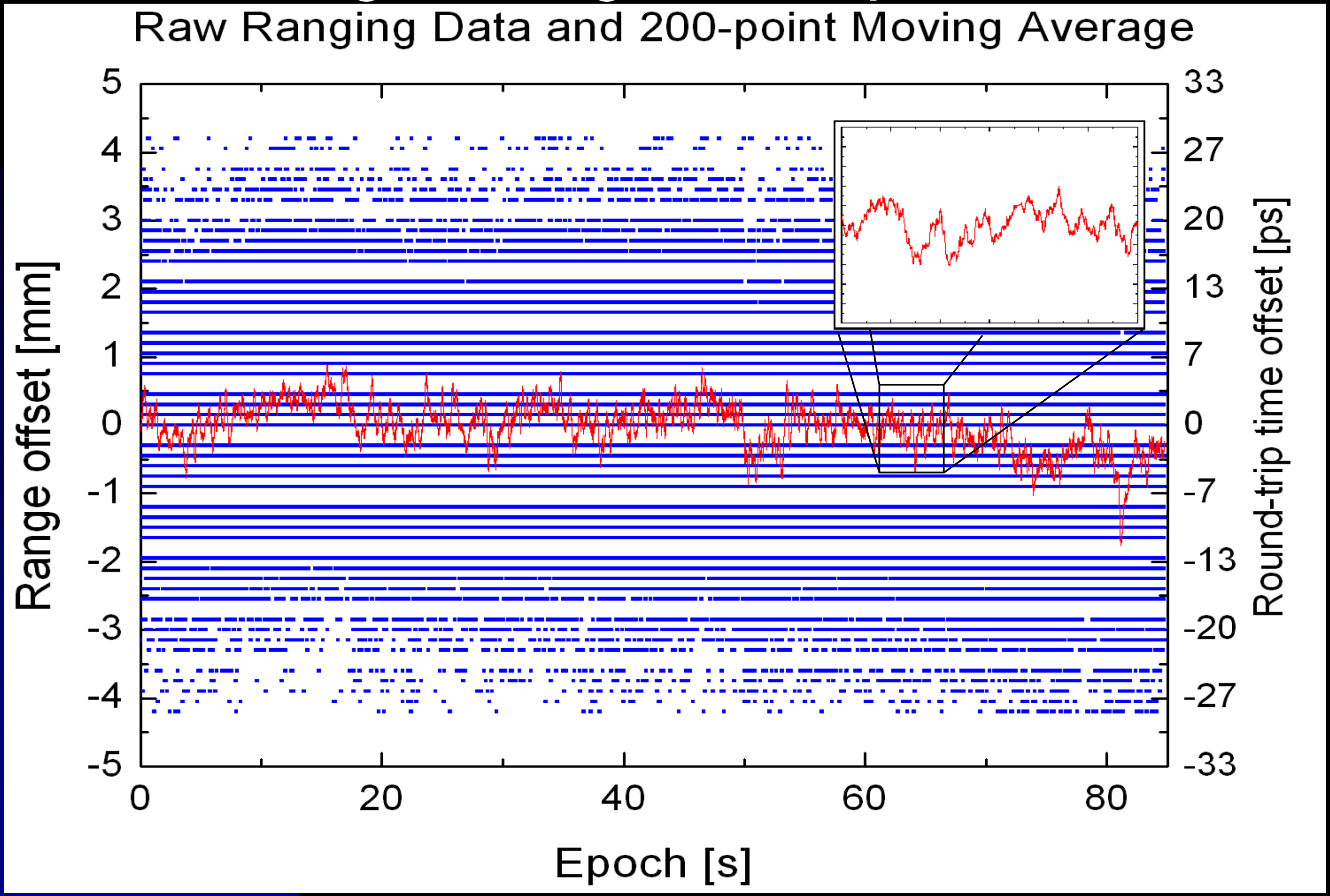


- windows moving by one point
- data from the beginning and the end are uncertain...
- <u>spread</u> inside the window is 1/SQR(n)
  smaller than original one
- the result is smoothed curve sequence of points,
- <u>number</u> of points is (almost) equal to original one

Data fitting and smoothing
Moving average example # 1

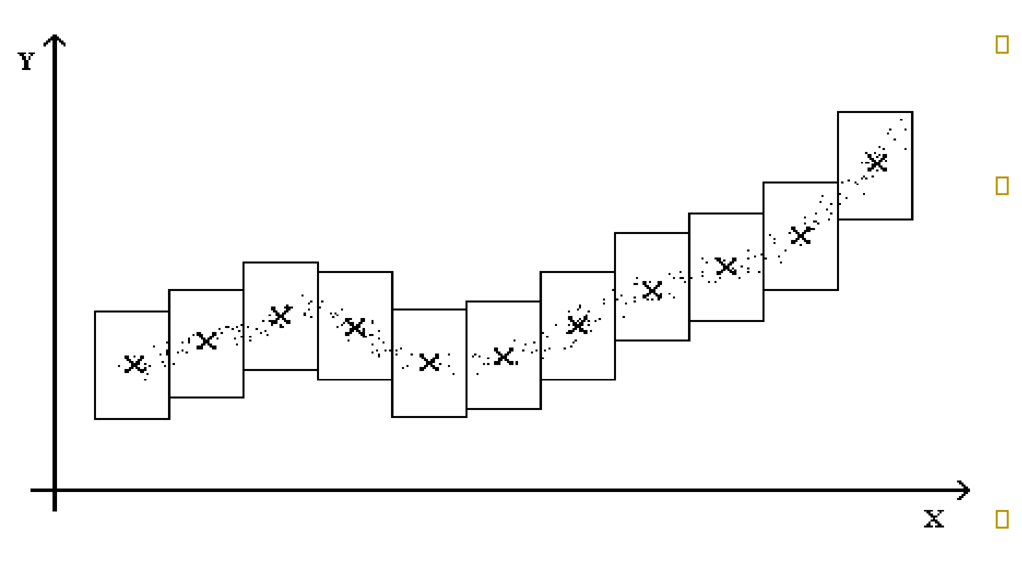


Moving average example # 2



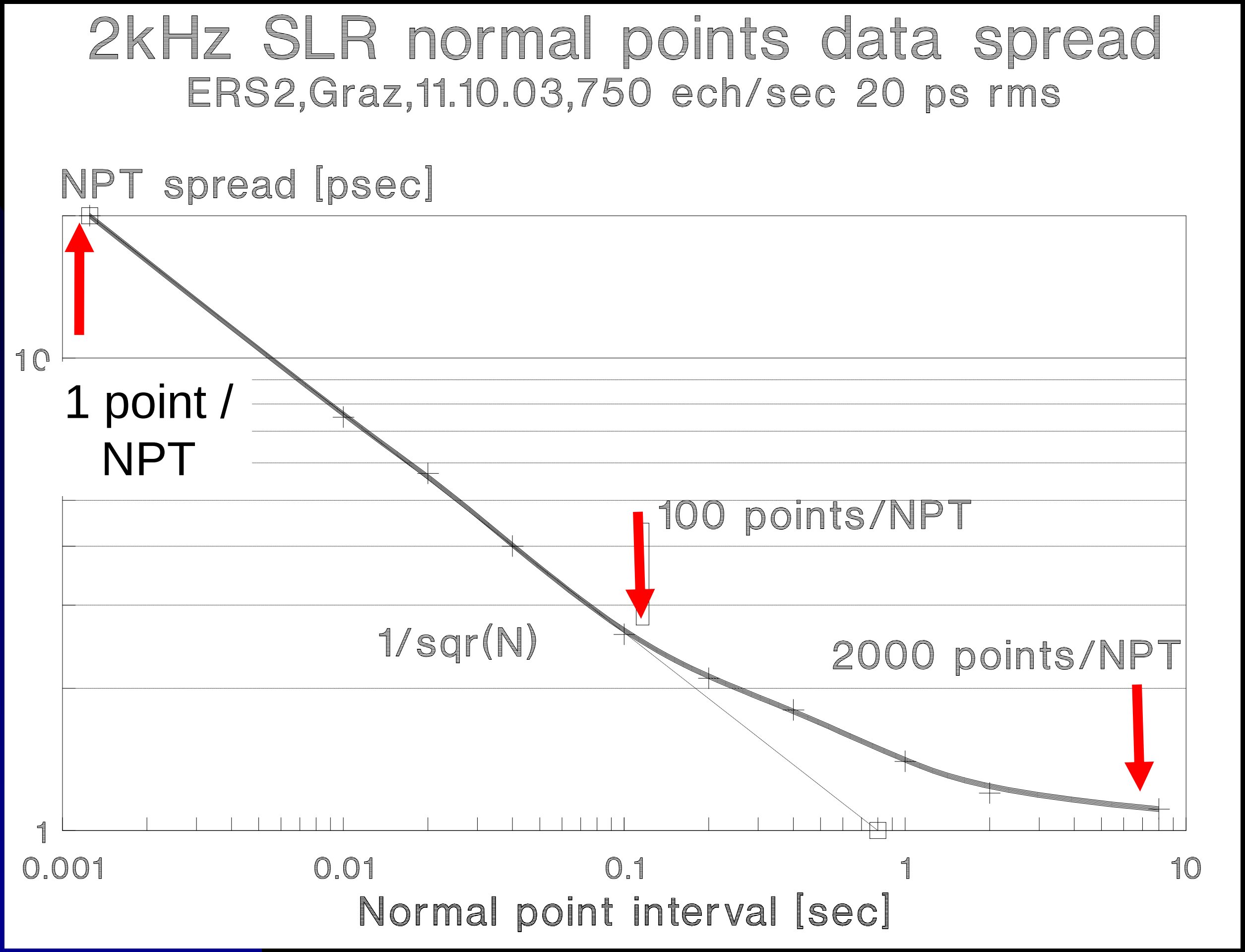
- Moving average data spread (RMS) is much bigger than in normal distribution = >
- New physical effect was discovered, (L.Kral et al, 2005)

Data fitting and smoothing
Normal points



- normal point is an arithmetic average of the data in a window
- windows are not overlapping
- spread of normal points is 1 / SQR(n) lower than the original one where n is number of points in the window
- Both ends are well defined
- Number of Normal points is substantially lower than original data points

Normal points example # 1

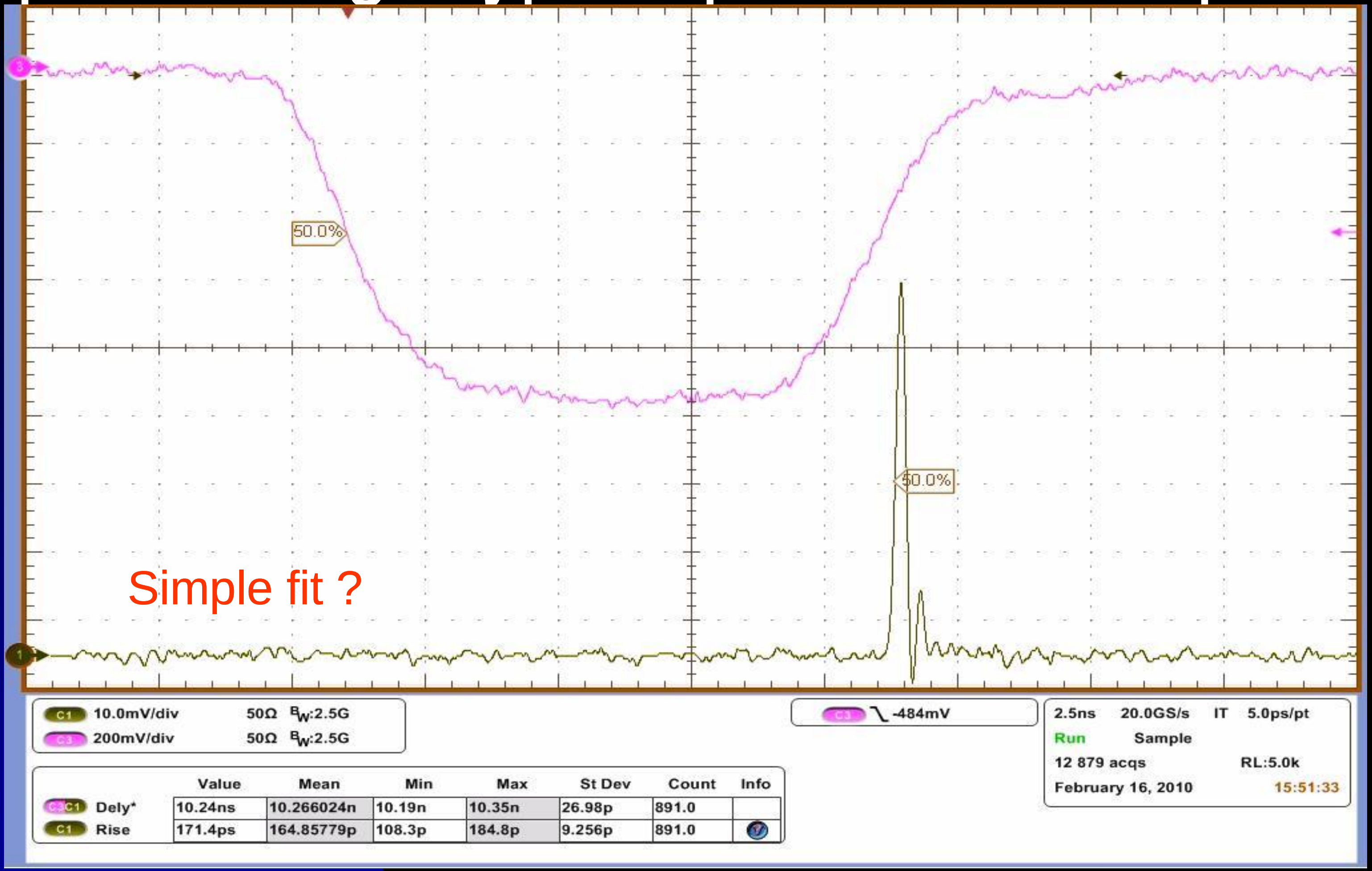


- deviation from ideal > 100 echoes / NPT 2.5 psec
- saturation : > 2000 echos / NPT 1.0 psec

Data fitting and smoothing
Splines

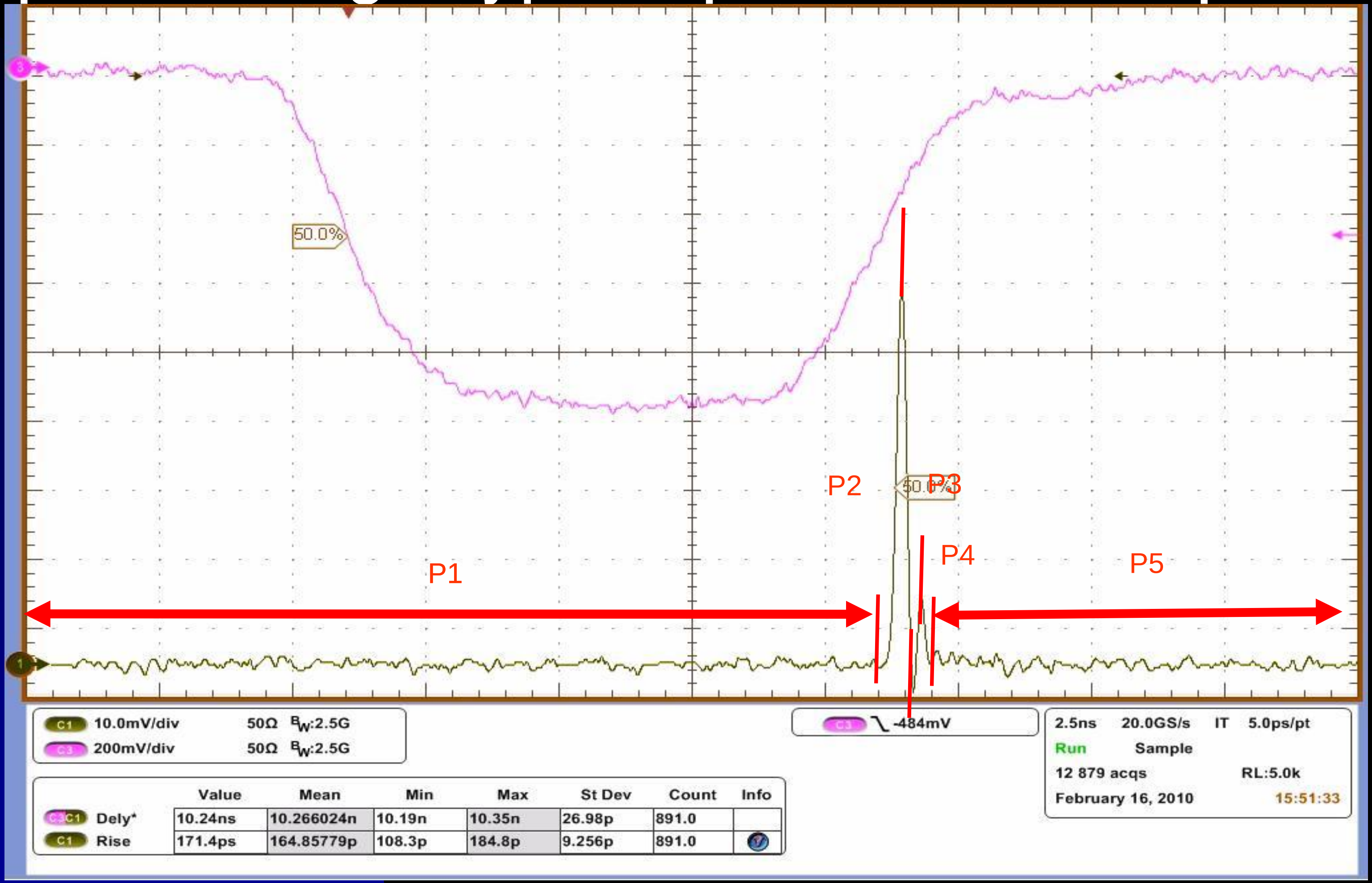
- data fitting by the series of low degree polynomials
- in the node /point of change from one polynomial to the other one / the value and the first derivative of both the polynomials must be equal
- most often used scheme - the sequence of 3rd degree polynomials
- used to fit data, which can not be fitted by classical polynomials / for example : pulse shapes,…/

Spline fitting - typical problem example #1



No single polynom will fit correctly the lower trace

Spline fitting - typical problem example #1



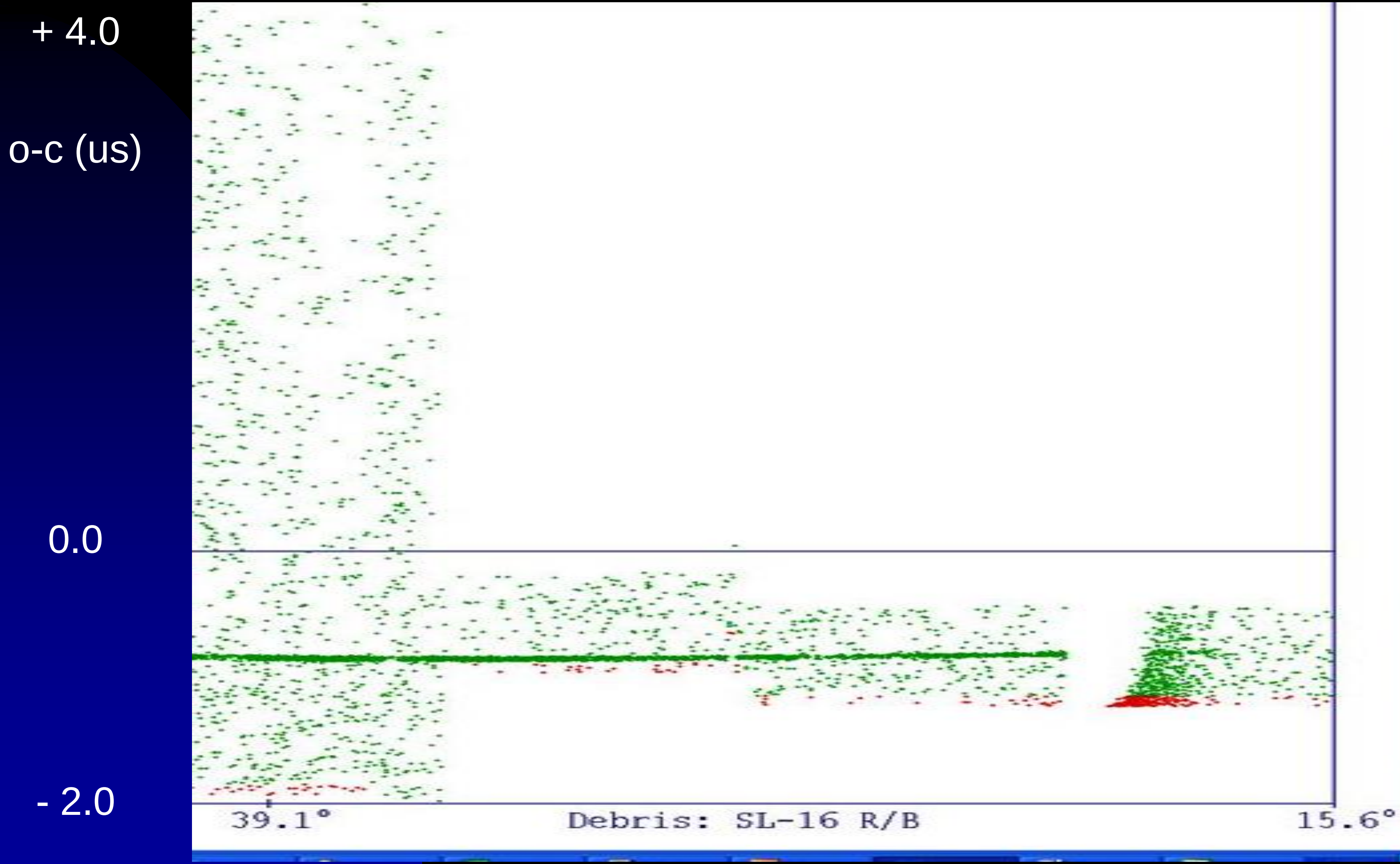
in the node - point of change from one polynomial to the other one – the value and the first derivative of both the polynomials must be equal

Data editing

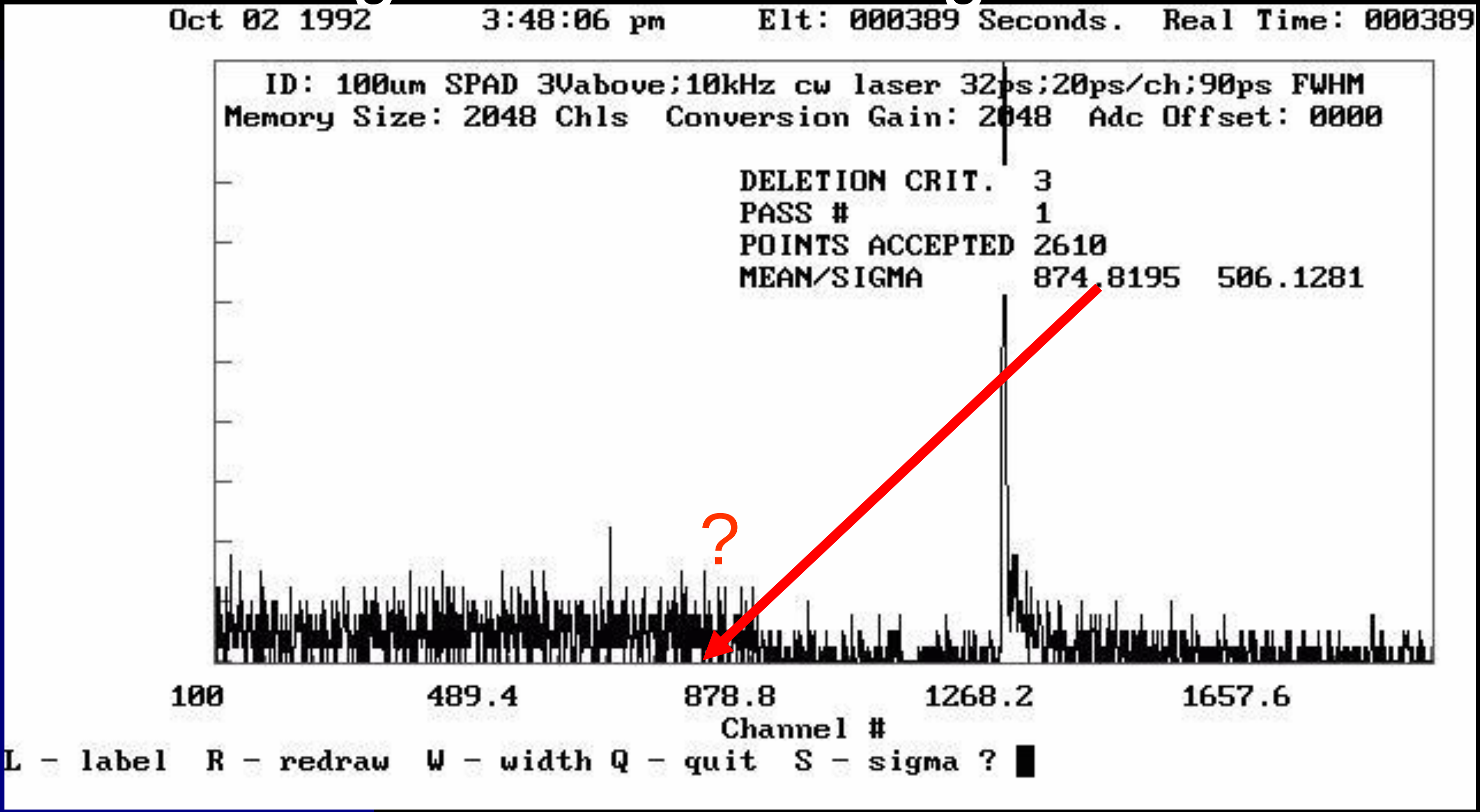
- normal distribution and deviations from it
- relation to data fitting
- probability of deviations > 3 * sigma and bigger
- proper selection of the editing criteria
  k * sigma … for k = 2.0 … 3.0
- applicable for S / N > ~ 0.3
- non-symetrical distribution
- normal distribution + DC offset
  = > convergence problem
  may be solved by tight editing criteria

Too high No of raw errors – simple “3*sigma” editing does not work

Space debris tracking, G.Kirchner, Graz August 2013



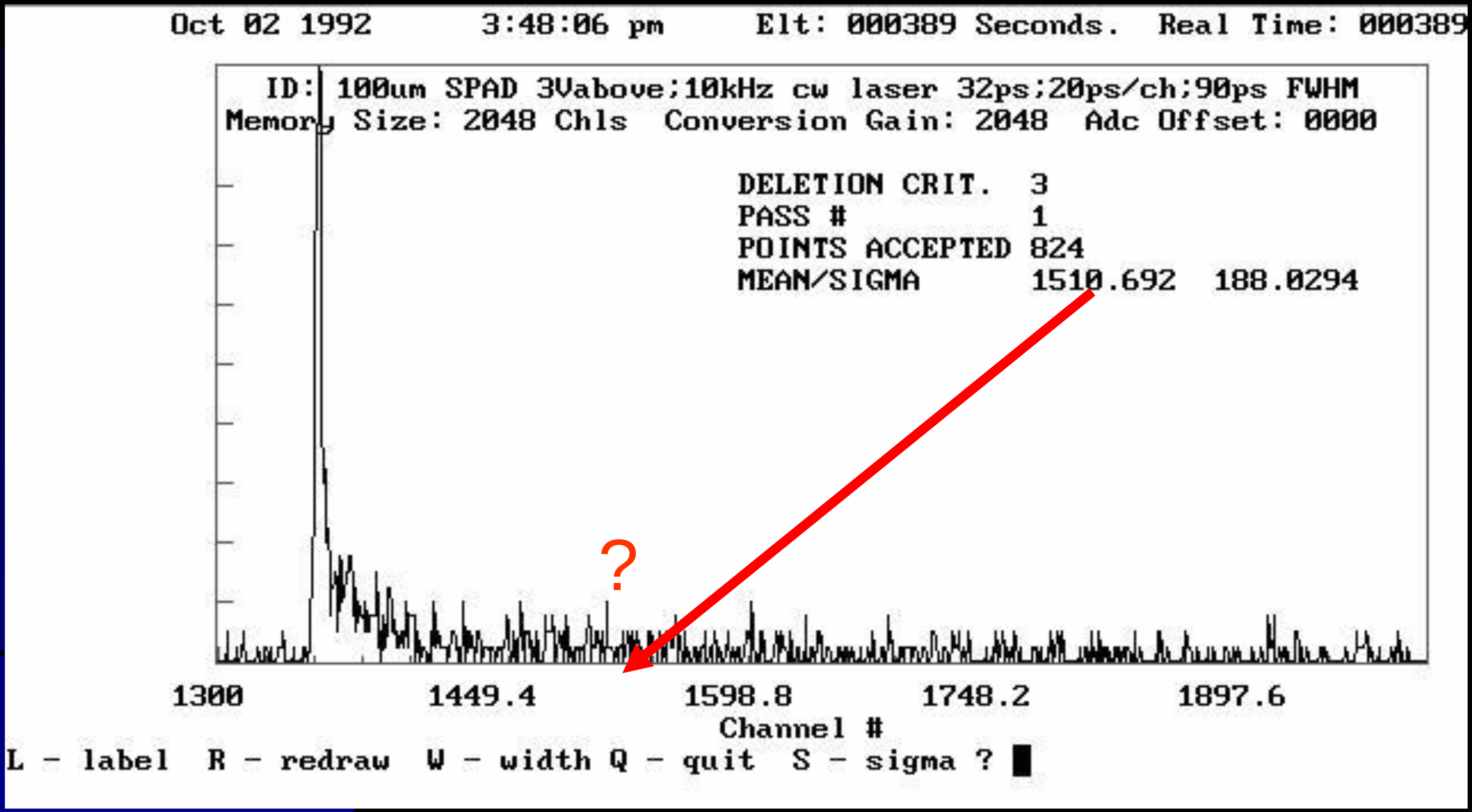
Data fitting and smoothing TCPC demo 1



- In a large amount of noise we have to locate desired correct value exactly (select narrow “data window” and tight editing criteria)
- Standard editing procedure “3*sigma “ does not make any sense, see graph..

Data editing

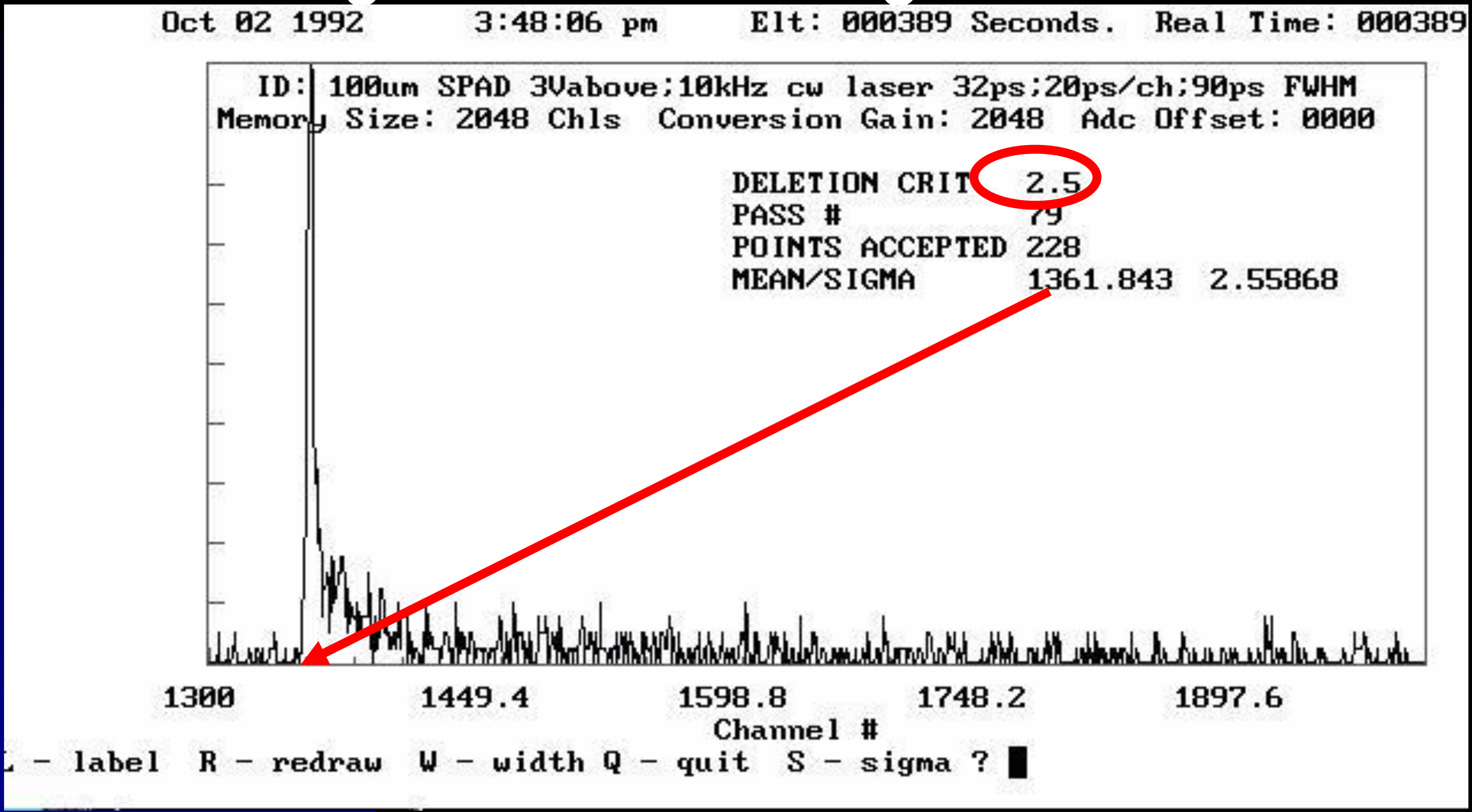
Data fitting and smoothing TCPC demo 2



- Even if we choose the right range of the data, the result still doesn't have to make sense
- After setting the proper value of SIGMA...

Data editing

Data fitting and smoothing TCPC demo 3



- ... we get the proper mean value, at least (correct data window and 2.5*sigma)

Data mining GOALS

- (1) Identification of useful signal within a “noise”
- (2) estimation of probability of correct signal identification
- < = > Eliminating the raw errors
  in a case, when number of raw errors is much larger than a number of useful signal
- In this chapter the term “noise” has a meaning of raw error
- In a previous example we have demonstrated that simple criteria like k * sigma will not work for very noisy data sets

Data mining # 2

- GENERAL RULE
- The signal is correlated
- noise is random

- STRATEGY
- The key problem – identification of effects, with which the signal is correlated

- EXAPLES

impulse	effects	epoch
periodic	effects	period
other	effects	time
		known effect
		etc..

EXAMPLEs of data mining / correlation

- direct TV broadcasting
  - direction
  - frequency
  - polarization
  - modulation (timing)

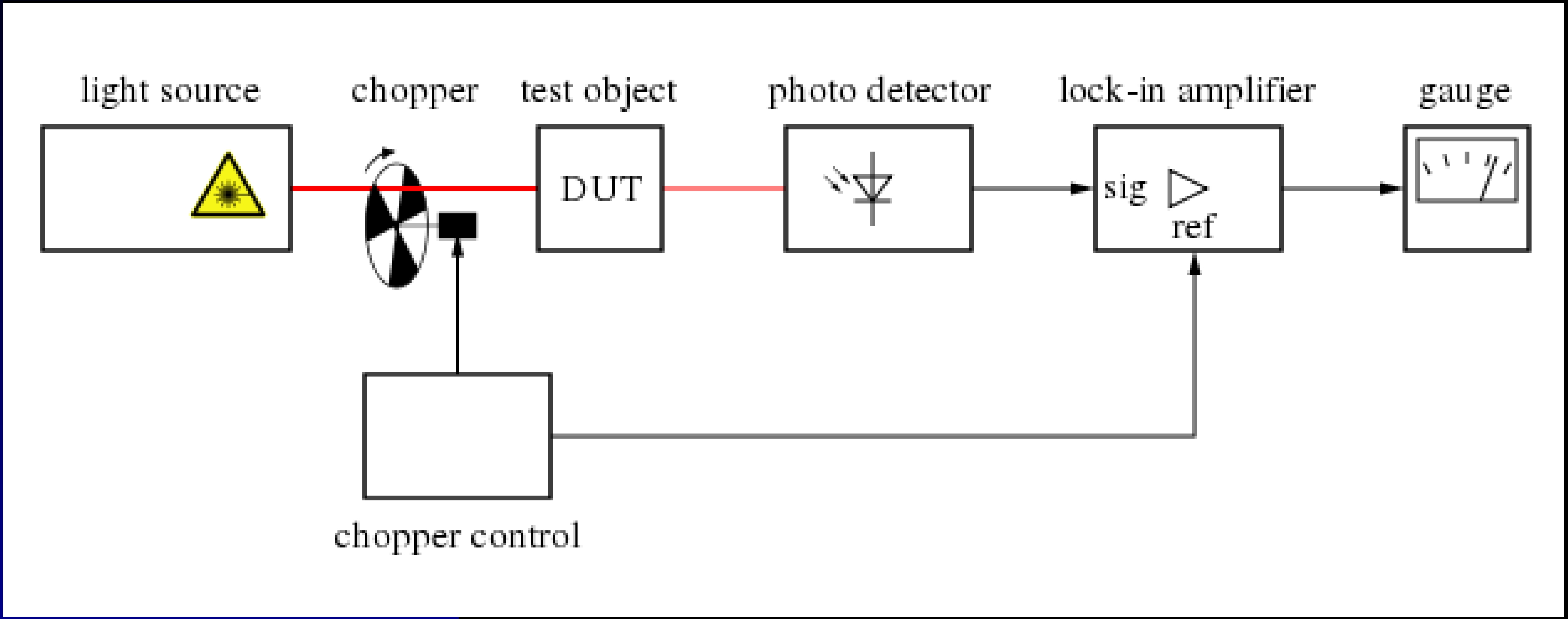
- Satellite Laser Ranging
  - direction
  - wavelength
  - epoch

Lock-in measurements

- used in experiments, in which there is a low degree of correlation
- additional “modulation” is applied to the experiment
- the signal ix extracted from the S + N on the basis of its correlation to the (known) external effect
- “lock-in amplifier” for low voltage / current measurements

Lock-in measurements #2

- Weak optical signal detection



Data mining
"Correlation Estimator"

- Enables to identify the known pattern in the noisy background
- Used in experiments, in which we can compare the original (for example transmitted) signal with the noisy (received) signal
- The problem is solved on the principle of maximizing the (auto)-correlation function
- The (fast) Fourier transformation approach (effective especially in 2D solutions, image processing,..)
- application in
  - radio-location
  - precise / impulse / timing
  - image processing (robotics)
  - etc.

“Correlation Estimator” # 2

For continuous functions *f* and *g*, the cross-correlation is defined as:

$$(f \star g)(\tau) \stackrel{\mathrm{def}}{=} \int_{-\infty}^{\infty} f^*(t)\ g(t+\tau)\, dt,$$

where f^* denotes the complex conjugate of *f* and τ is the time lag.

Similarly, for discrete functions, the cross-correlation is defined as:

$$(f \star g)[n] \stackrel{\mathrm{def}}{=} \sum_{m=-\infty}^{\infty} f^*[m]\ g[m+n].$$

Wikipedia

Data mining

"Correlation Estimator" # 3

- The cross-correlation of functions *f*(*t*) and *g*(*t*) is equivalent to the convolution of *f**(−*t*) and *g*(*t*). I.e.:

$$f \star g = f^*(-t) * g.$$

- If *f* is Hermitian, then $f \star g = f * g.$
- $(f \star g) \star (f \star g) = (f \star f) \star (g \star g)$
- Analogous to the convolution theorem, the cross-correlation satisfies:

$$\mathcal{F}\{f \star g\} = (\mathcal{F}\{f\})^* \cdot \mathcal{F}\{g\},$$

where $\mathcal{F}$ denotes the Fourier transform, and an asterisk again indicates the complex conjugate. Coupled with fast Fourier transform algorithms, this property is often exploited for the efficient numerical computation of cross-correlations. (see circular cross-correlation)

Wikipedia