Data Processing

Kihyeon Cho

Syllabus

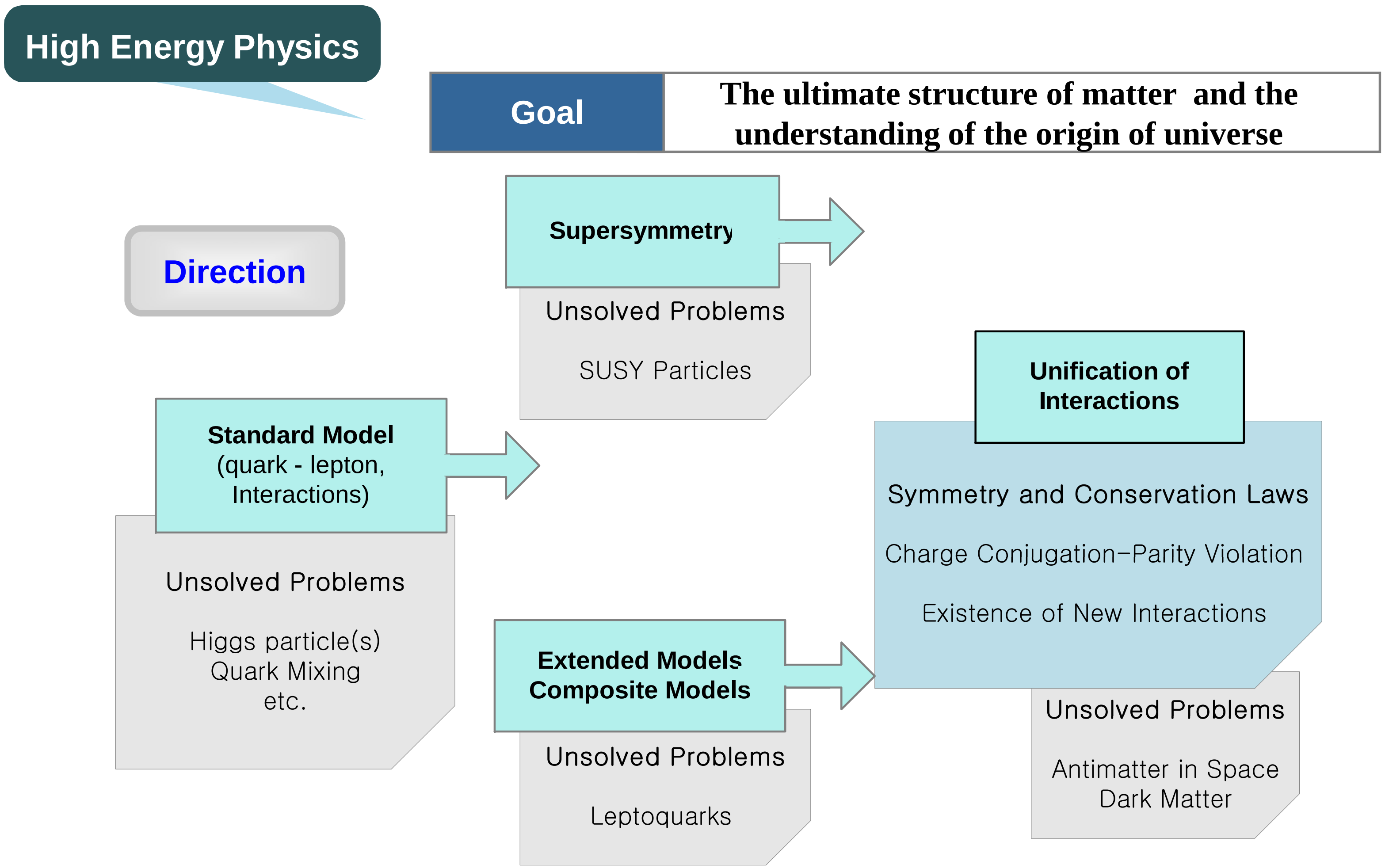
- Introduction (Chap. 1)
- Special Relativity (Chap. 2)
  - Special Relativity
  - Symmetry (Group)
- Quantum Mechanics (Chap. 3)
- Detector
- Data Processing
- Feynman diagram (Chap. 4)
- QED (Chap. 5)
- QCD (Chap. 6)
- Weak interaction (Chap. 7)

What cover In this Chapter?

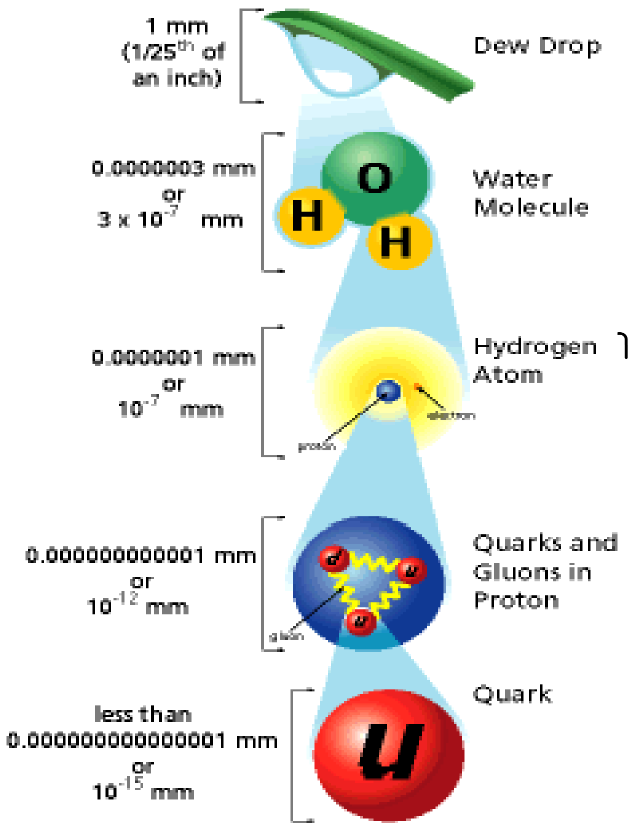
- **High Energy Physics**
- **Data Processing**
- **Fitting**
- **Conclusions**

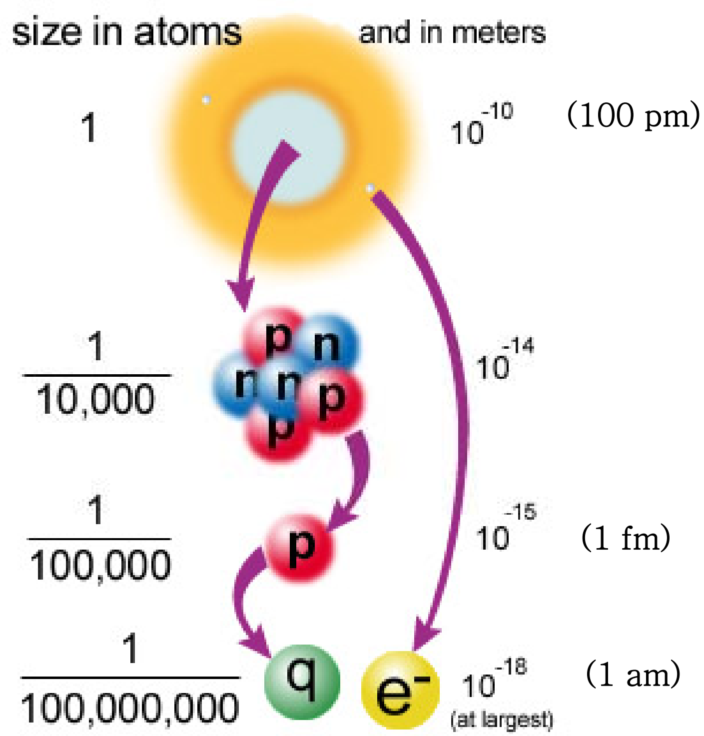
High Energy Physics

High Energy Physics



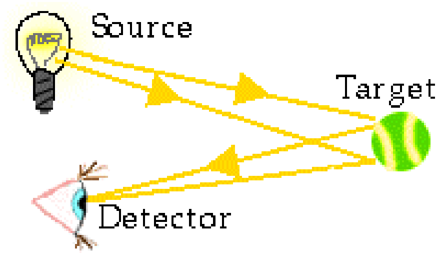
What is World Made of?

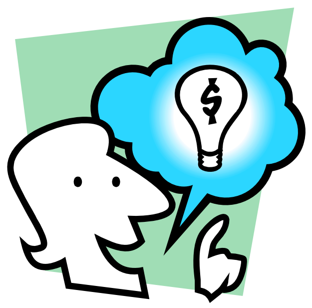




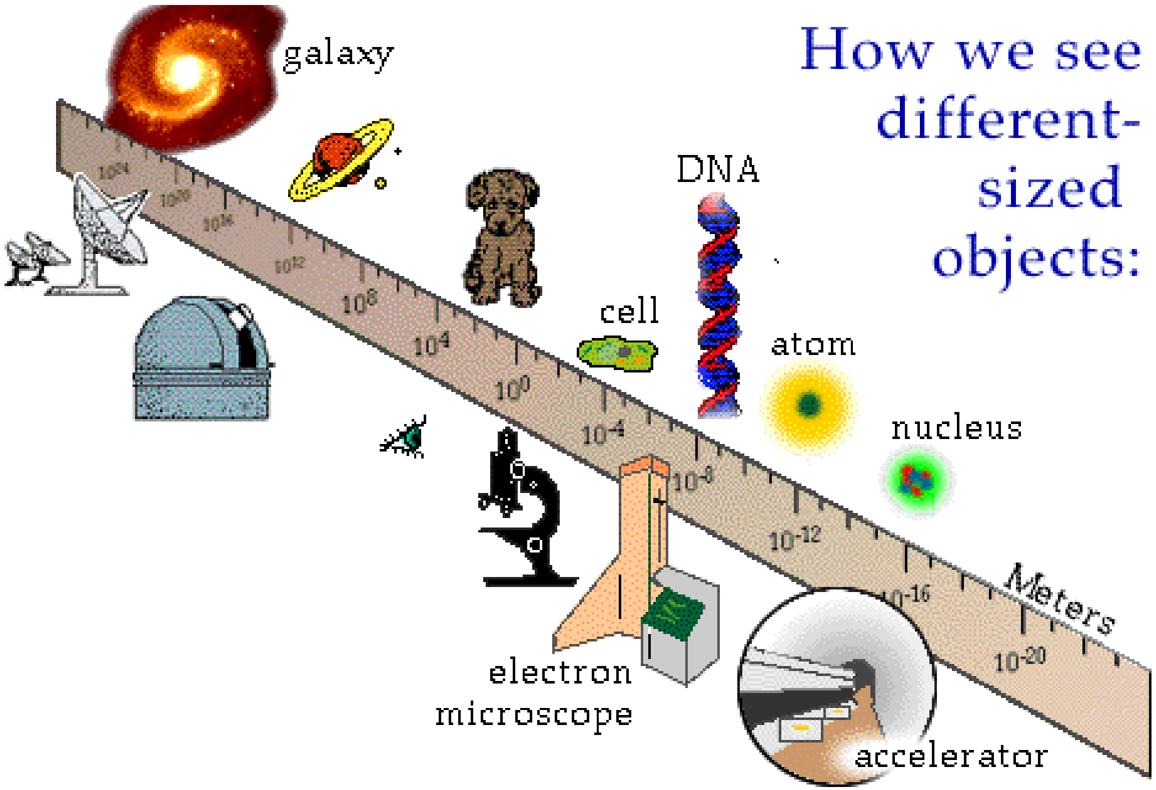
How to know any of this? (Testing Theory)

- Example
  - Light bulb (Source)
  - Tennis ball (target)
  - Eye (detector)





How to detect?

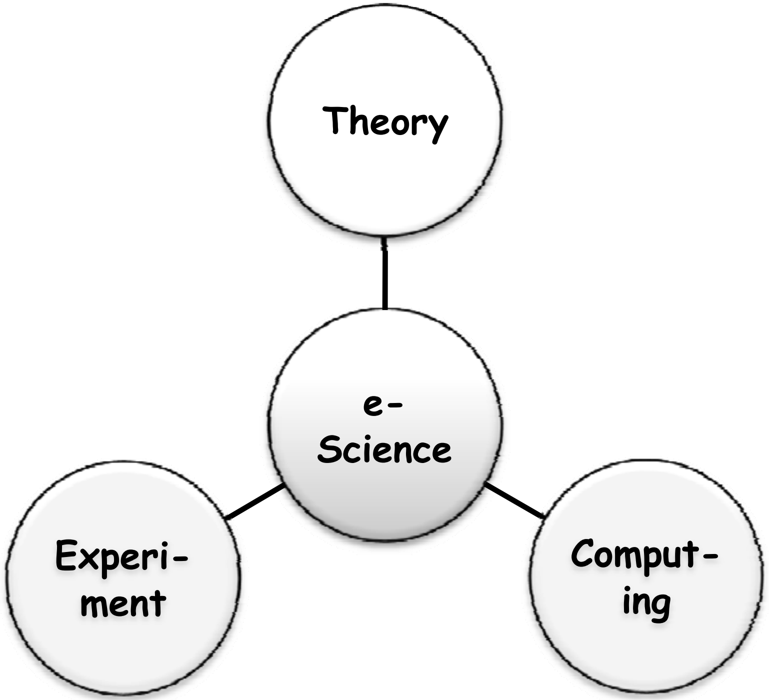


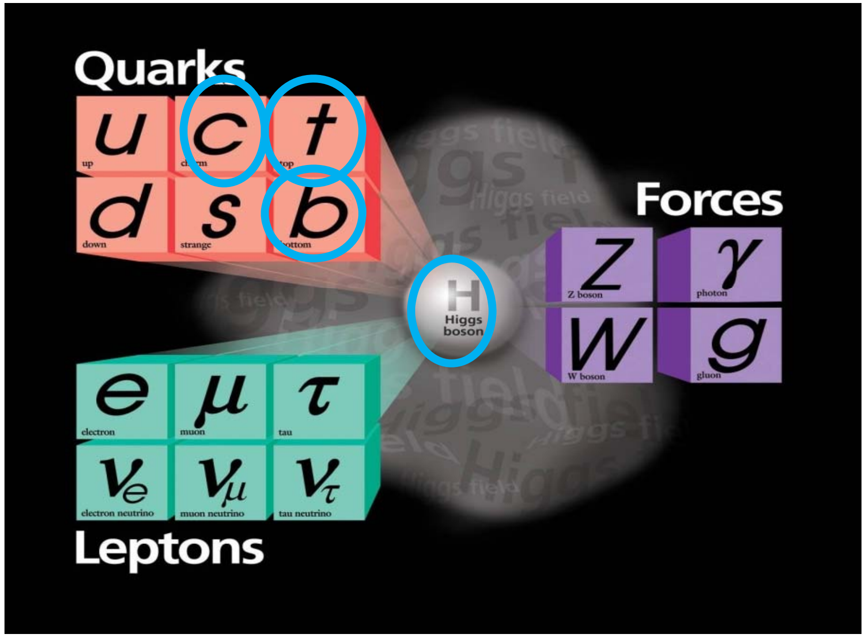
How do we experiment with tiny particles? (Accelerators)

- Accelerators solve two problems:
  - High energy gives small wavelength to detect small particles.
  - The high energy create the massive particles that the physicist want to study.

High Energy Physics Team

To probe the Standard Model and search for New Physics



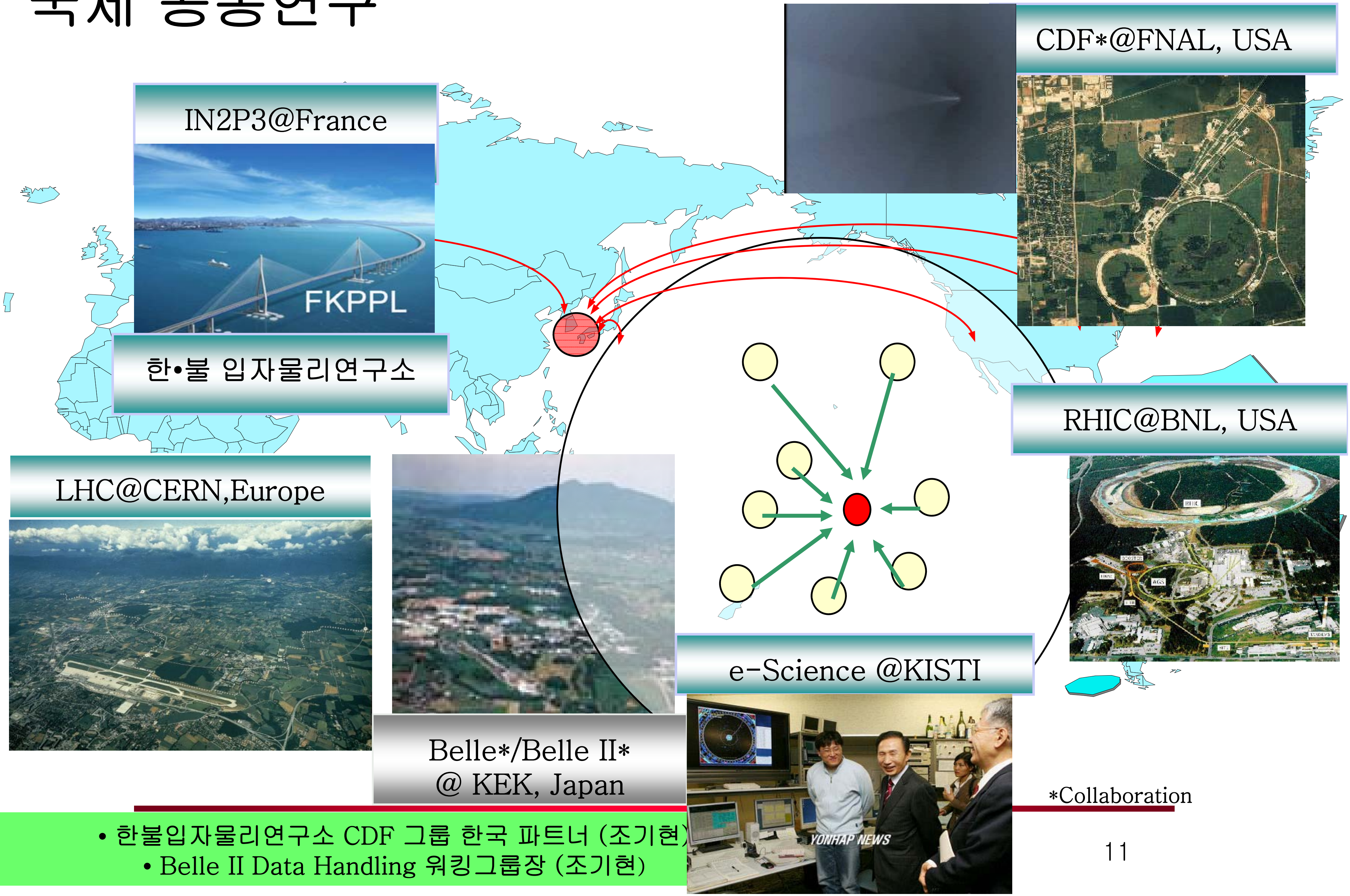


CDF
Belle / Belle II

cf. LHCb

K.Cho and H.W.Kim, JKPS (2009)

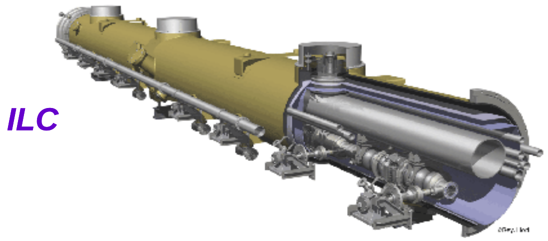
국제 공동연구



*Collaboration

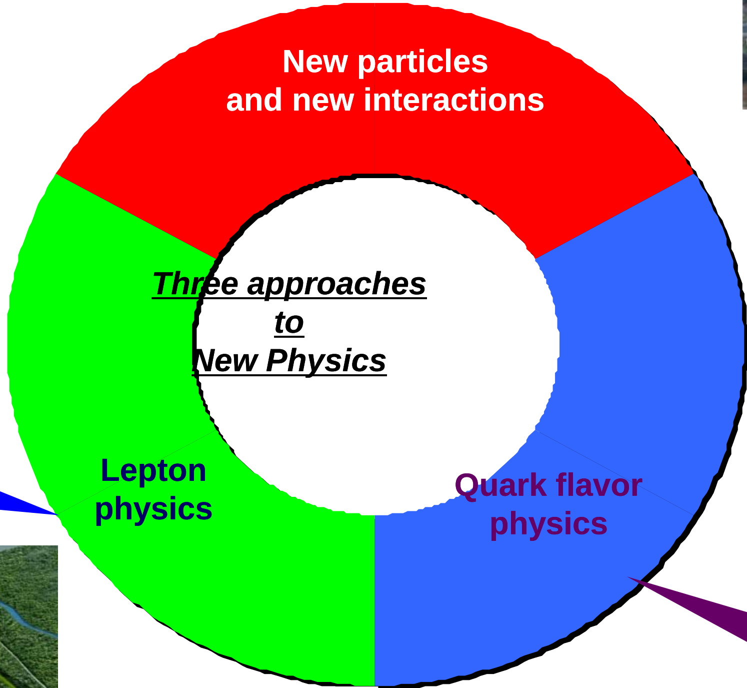
- 한불입자물리연구소 CDF 그룹 한국 파트너 (조기현)
- Belle II Data Handling 워킹그룹장 (조기현)

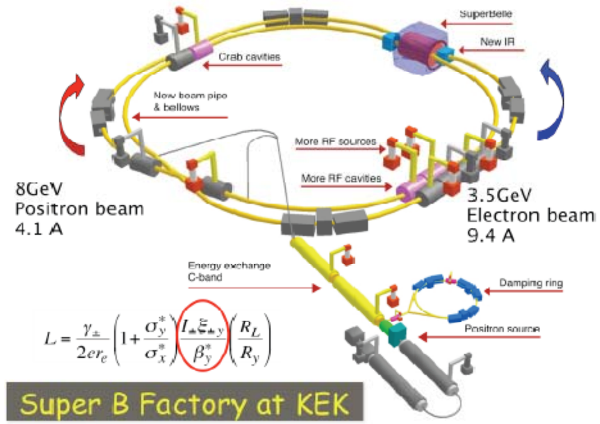
High Energy Physics in the LHC era



Energy frontier experiments
LHC, ILC, …

Higgs, SUSY, Dark matter,
New understanding of space-time…





ν exp., μ LFV, τ LFV,
g_μ-2, …

Super-B Factory,
K exp., *etc.*

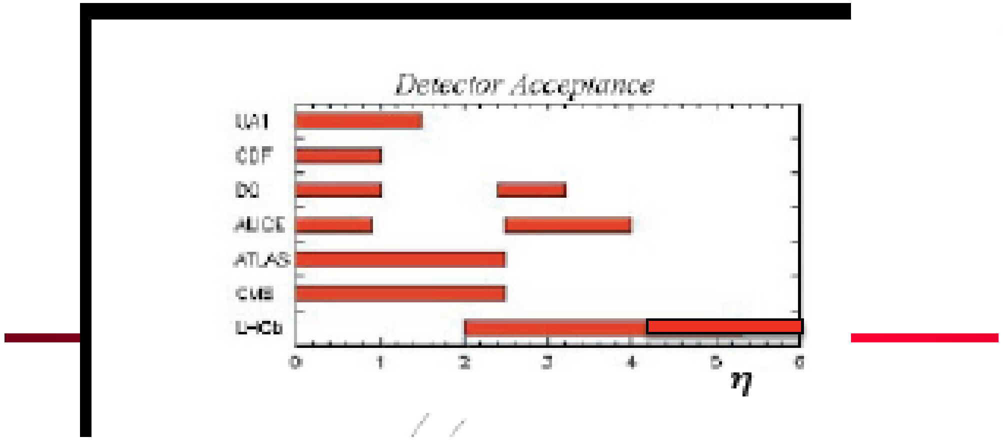
Neutrino mixing/masses,
Lepton number non-conservation…

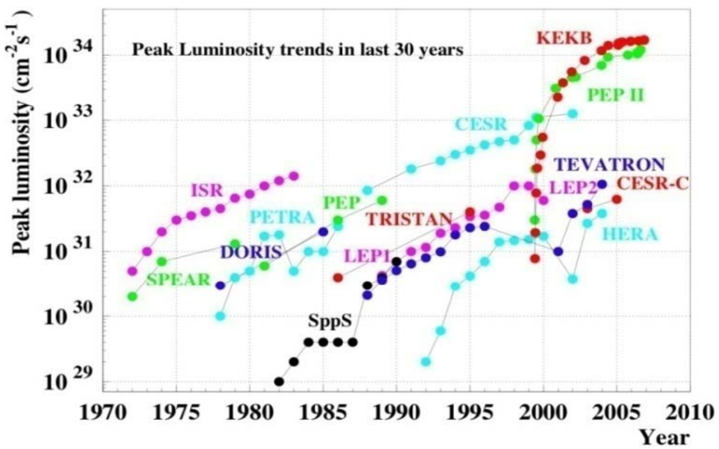
CP asymmetry, Baryogenesis,
Left-right symmetry, New sources

M. Yamakuchi, Belle II meeting (2008)

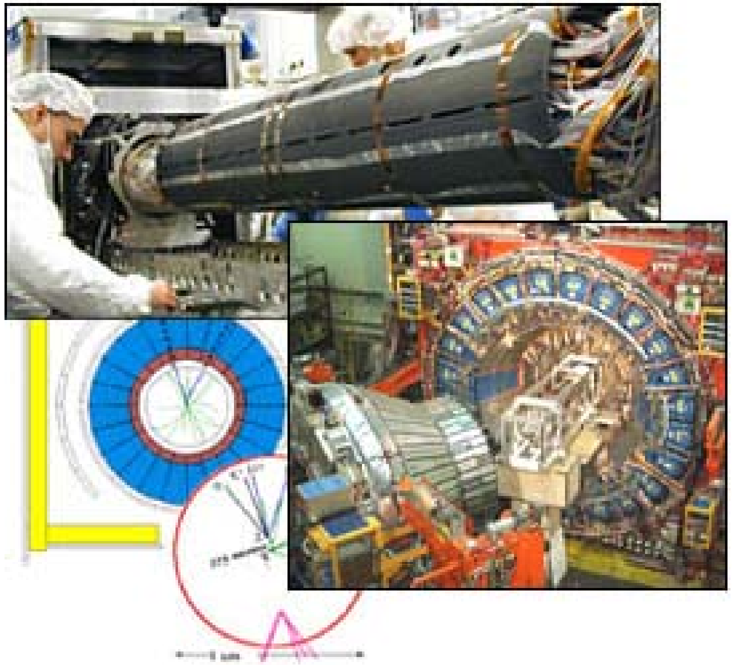
Heavy Flavor Physics Experiments

	Belle/Belle II	CDF	LHCb
Year	1998–2010 (Belle) 2014 – (Belle II)	2001–	2009–
Place	KEK, Japan	Fermilab, USA	CERN, Europe
Collaboration	13/47/~300(Belle II) (Nat./Ins./member)	15/63/620	15/54/730
σ	1 nb (10GeV)	150 μb (2TeV)	300~500 μb (7~14TeV)
Current Luminosity	1 ab^{-1}	8 fb^{-1}	180nb^{-1}





Data Processing



Why do we do experiments?

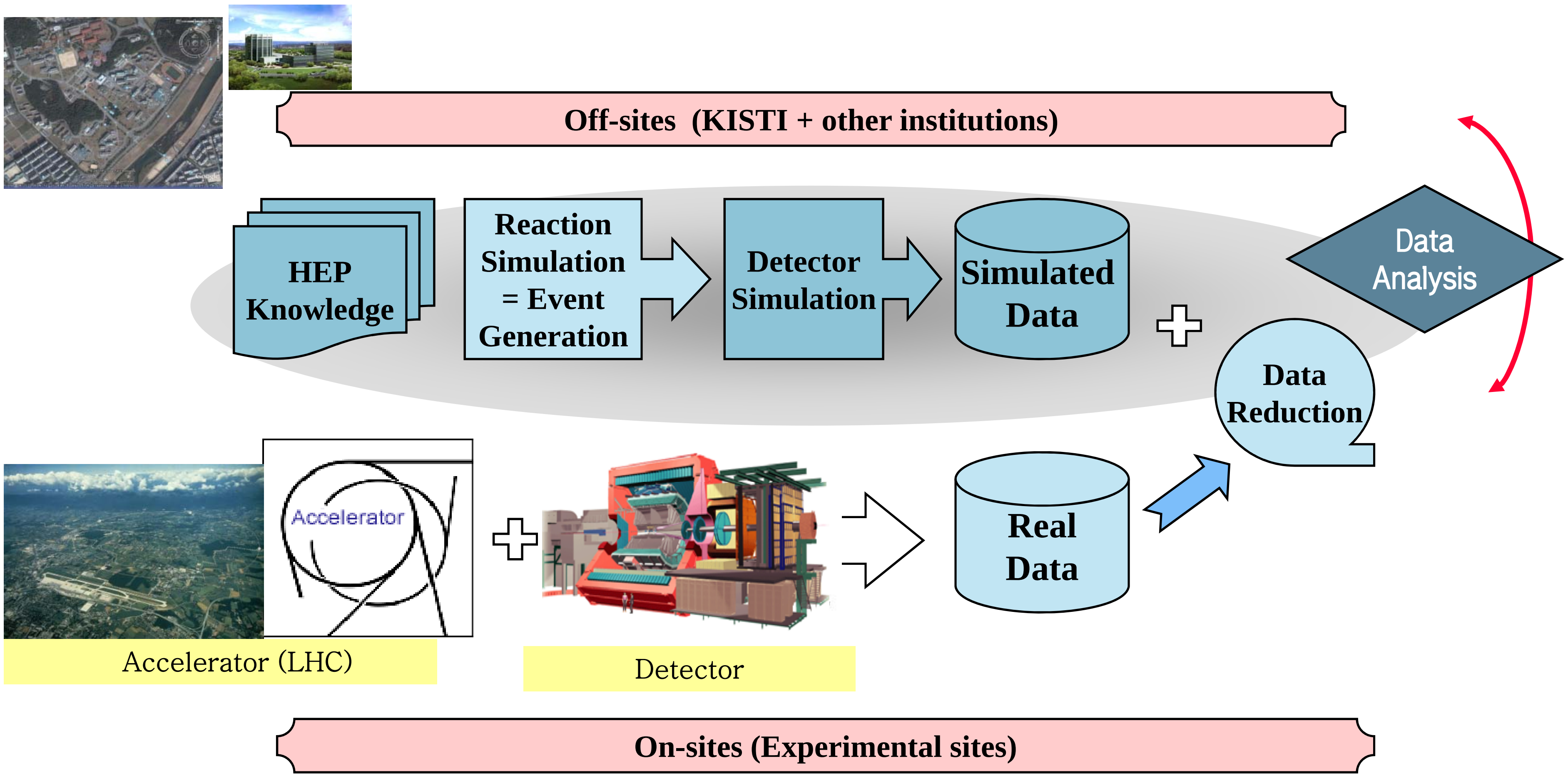
- Parameter determination
  - To set the numerical values of some physical quantities
  - Ex) To measure velocity of light

- Hypothesis testing
  - To test whether a particular theory is consistent with our data
  - Ex) To check whether velocity of light has suddenly increased by several percent since beginning of this year

Type of Data

- Real Data (on-site)
  - Raw Data : Detector Information
  - Reconstructed Data : Physics Information
  - Stream (Skim) Data : Selected interested physics
- Simulated Data (on-site or off-site)
  - Physics generation : pythia, QQ, bgenerator, CompHEP, ...
  - Detector Simulation : Fastsim, GEANT, ...

Typical Research Procedure



Error (σ)

- Error
  - Error : the difference between measurement and true value
  - True value
    - We don't know it
  - Statistical error
    - Error due to statistical fluctuation
  - Systematic error
    - More in nature of mistakes due to equipments and experimentalists

- Experimental value : Meas. ± stat. error ± sys. error

  Example) m(top) = 175.9 ± 4.8 ± 5.3 GeV/c^2 (CDF, 1998)

Why estimate errors?

- To know how accuracy of the measurement
- Example
  - The conventional speed of light $c=2.998 \times 10^8$ m/sec
  - When the new measurement $c=3.09 \times 10^8$ m/sec
  - Case 1. If the error is $\pm$ 0.15, then it is consistent.
    - Conventional physics is in good shape.
    - 3.09 ± 0.15 is consistent with 2.998×10^8 m/sec
  - Case 2 . If the error is ± 0.01, then it is not consistent.
    - 3.09 ± 0.01 is world shattering discovery.
  - Case 3. If the error is ± 2, then it is consistent.
    - However, the accuracy of 3.09 ± 2 is too low.
    - Useless measurement

⇒Whenever you determine a parameter, estimate the error or your experiment is useless.

Examples) Bad and Good

•L (Integrated Lum.) : 5169.26 nb-1 +/- 0.02

•Nsig : 4403

•A (Total acceptance for W->enu) : 0.22212

•Cross section * BR = Nsig / (A * L) = 4403 / (0.22212*5169.26)

= 3.83471 nb

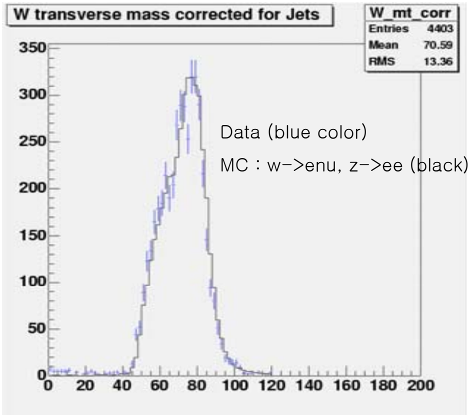
Cdfnote 6681,

Using 72.0 pb^{-1} of good Run II CDF data we measure

$$\sigma \cdot B(p\bar{p} \to W \to e\nu) = (2.753 \pm 0.015_{stat} \pm 0.084_{syst} \pm 0.165_{lum})\ \mathrm{nb}.$$

and

$$\sigma_Z \cdot Br(Z^0 \to e^+e^-) = 260.9 \pm 6.3(stat.) \pm 6.7(syst.) \pm 15.7(lum.)\ \mathrm{pb}$$



Data (blue color)

MC : w->enu, z->ee (black)

Bad

Background Estimations

• Integrated Lum. : 204.563 +/- 12.045 Pb-1

• $N_{\exp} = \int L\,dt \times \sigma \cdot Br \times (\varepsilon \cdot A)$

MC(pythia)	$\varepsilon \cdot A$	$\sigma \cdot Br\ (Pb)$	$N_{\exp}$
$W \to e\nu$	422976 / 2092000 = 0.202187	(cdf 6681) 2753.0 +/- 190	113864 +/- 7858
$W \to \tau\nu$	1178 / 458000 = 0.002572	(cdf 6447) 2620.0 +/- 270	1379 +/- 142
$Z/\gamma \to ee$	37241 / 2019500 = 0.018441	(cdf 6281) 261.5 +/- 31.9	987 +/- 120
$Z/\gamma \to \tau\tau$	1408 / 504000 = 0.002794	(cdf6281) 261.5 +/- 31.9	149 +/- 18
$t\bar{t}$	43649 / 732500 = 0.059589	(cdf6802) 4.7 +/- 2.4	57 +/- 29
QCD_{data}	-	-	3933 +/- 207 *(stat)*

Feb. 10. 2004 VEGY meeting 6

Good

How to reduce errors?

- Statistical error
  - Repeated measurement
  - N : the expected number of observation
  - σ = Sqrt(N) : the spread

- Systematic error
  - No exact formulae
  - Ideal case : All such effects should be absent.
  - Real world : An attempt to be made to reduce it.

How to solve systematic errors?

- Use constraint condition
  - Ex) Triangle
- Calibrations
- Energy and momentum conservation
  - E(after) – E(before) = 0
  - |P(after)| – |P(before)| = 0

How small of the systematic error?

- Systematic errors should be around statistical errors

The meaning of σ (error)

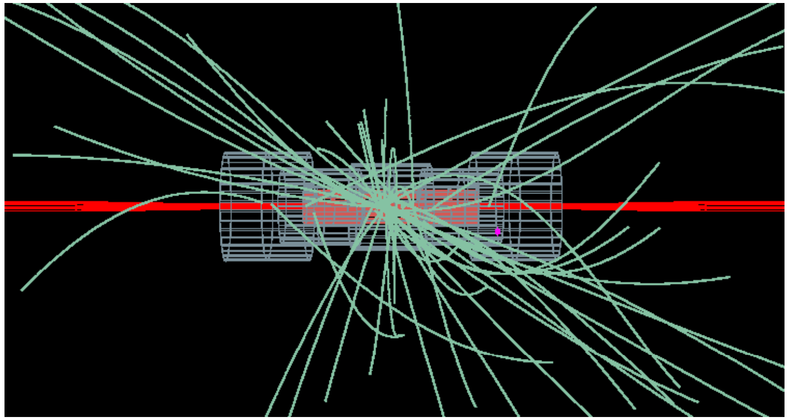
- Distributions x -> n(x)
  - Discrete
    - ex) # of times n(x) you met a girl at age x
  - Continuous :
    - ex) Hours sleep each night (x), # of people sleeping for time.

$\Rightarrow$ For an even larger number of observation and with small bin size, the histogram approach a continuous distribution.

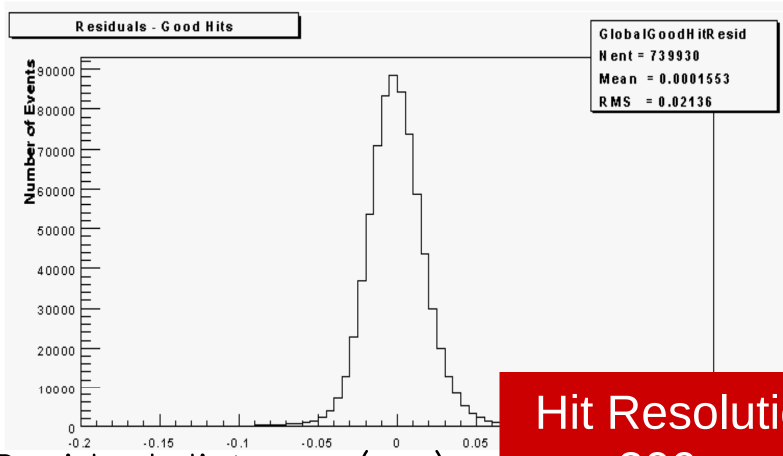
- Mean and Variance

- Gaussian distribution
  - In case of larger number of observation
  - It is important for error calculations

Tracking Performance

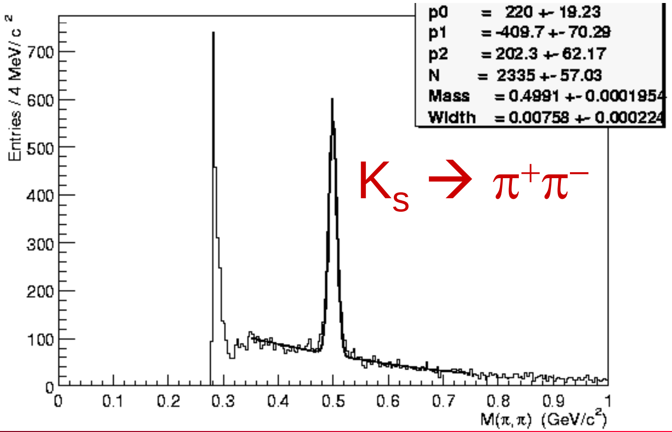


COT tracks

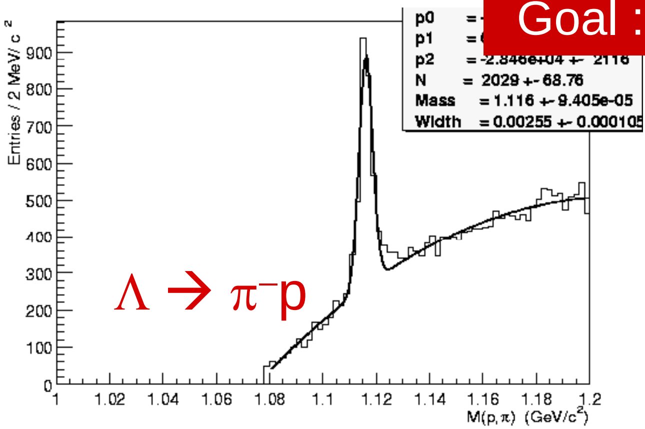


Residual distance (cm)

Hit Resolution
~200 μm
Goal : 180 μm



$K_S \rightarrow \pi^+\pi^-$



$\Lambda \rightarrow \pi^- p$

Mean and Variance

	True Value	Measurement
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s

In fact, we don't know the true value in the real world.

Mean

- Mean
  - N events has the value of (x1, x2, x3,... xN)

$$\bar{x} = \frac{\sum x_i}{N}$$

- Median – Observation or potential observation in a set that divides the set so that the same number of values, it is the middle value; for an even number it is the average of the middle tow
- Mode – Observation that occurs with the greatest frequency

  - When do not know true value

Variance

- Variance
  - When know true value

$$s^2 = \frac{\sum (x_i - \mu)^2}{N}$$

  - When do not know true value

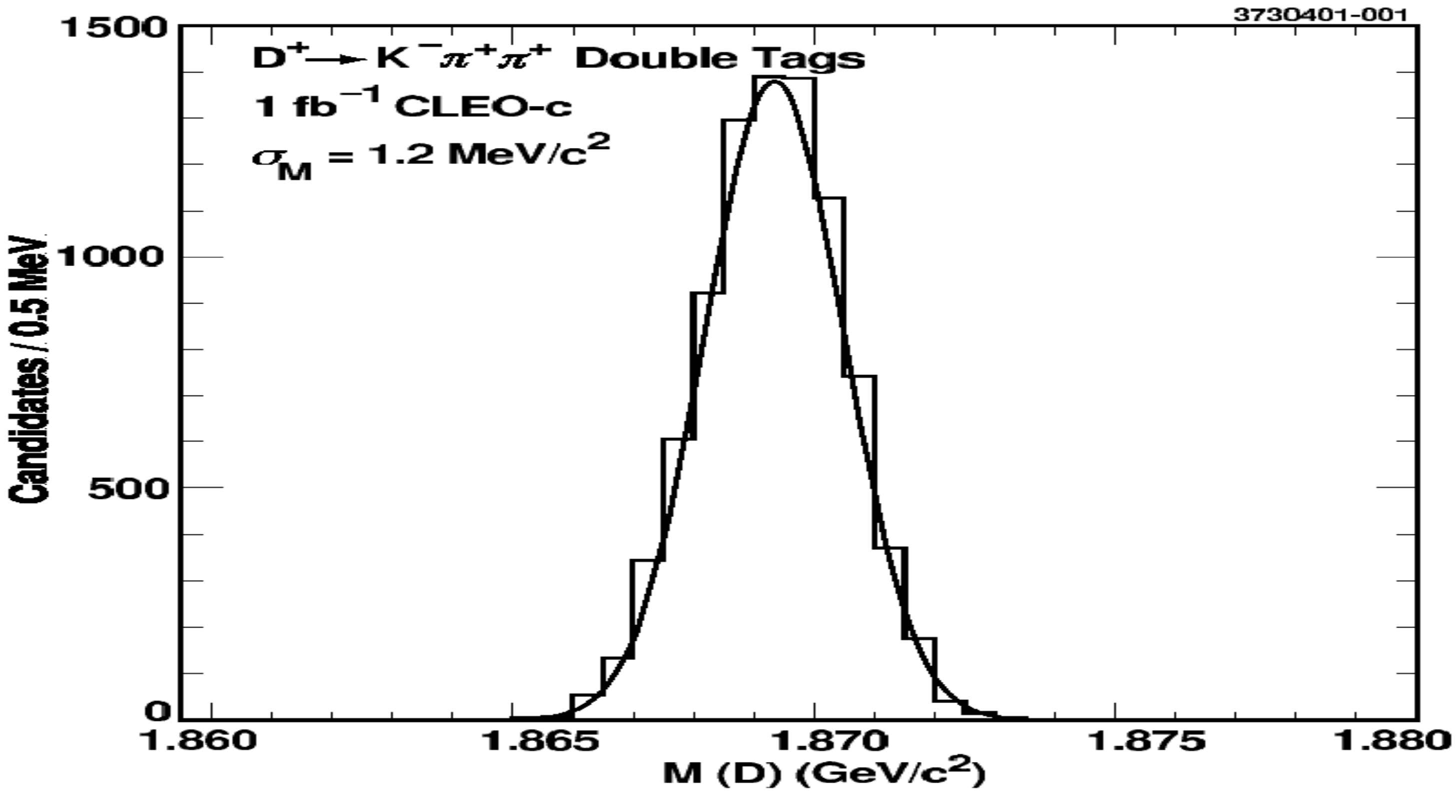
$$s^2 = \frac{\sum (x_i - \overline{x})^2}{N - 1}$$

Accuracy (δ)

- In order to know the accuracy of the measurement

$$\delta = \frac{s}{\sqrt{N}}$$

Gaussian Distribution



- In case of large size of data
- Gaussian distribution is the fundamental in error treatment.

Gaussian Distribution (cont'd)

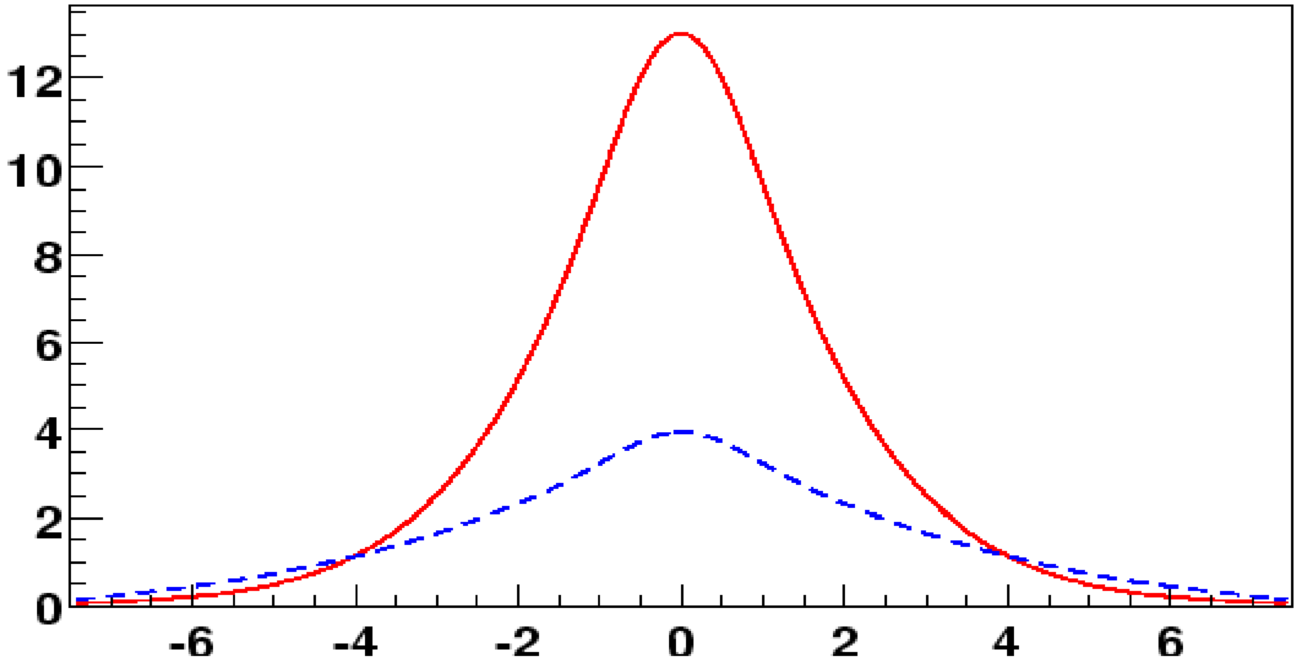
- The normalized function

$$y = \frac{1}{\sqrt{2\pi\sigma}} \exp\{-(x-\mu)^2 / 2\sigma^2\}$$

- Mean (μ)
- Width (σ)
- Width (σ) is smaller, distribution is narrower.
- Properties

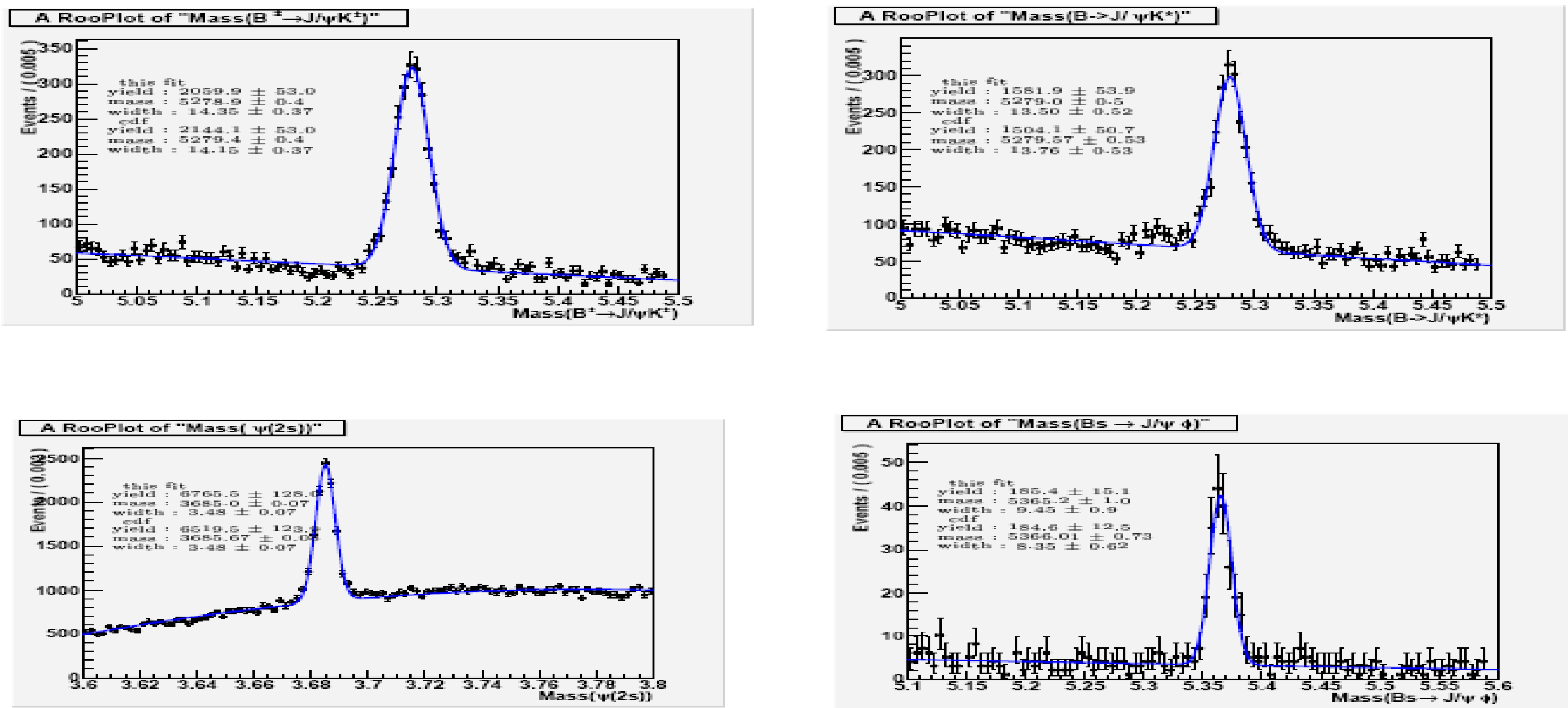
$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx = 0.68$$

Gaussian Distribution (cont'd)



- Mean (μ) is same as zero.
- However width (σ) is different.

Examples (Gaussian + BG)



D.J.Kong (2004.3.4)

CDF Secondary Vertex Trigger

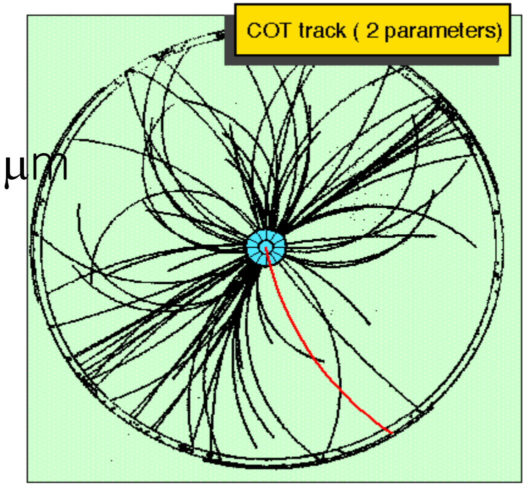
NEW for Run 2 -- level 2 impact parameter trigger

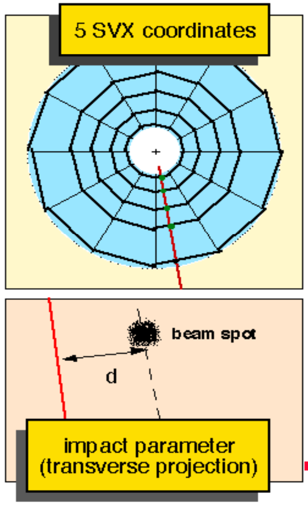
Provides access to hadronic B decays

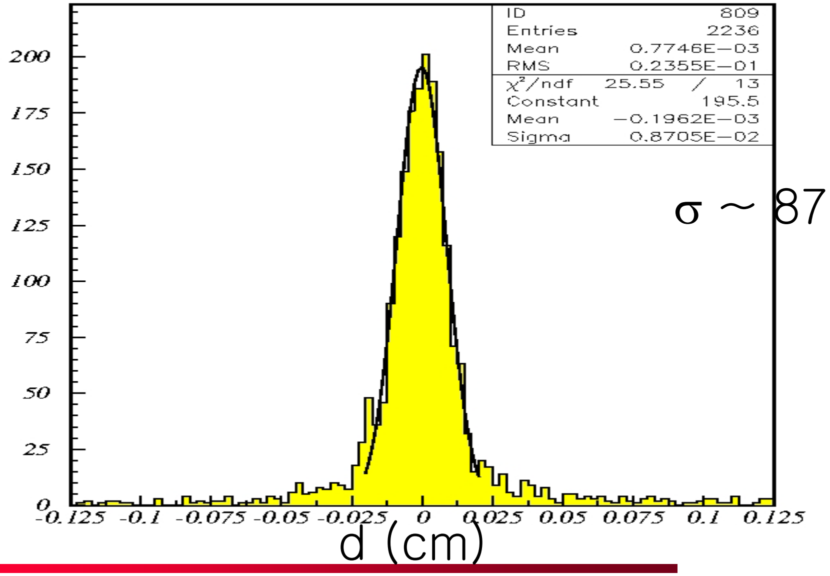
Data from commissioning run (no alignment or calibrations)

COT defines track at level 1

SVX measures impact parameter





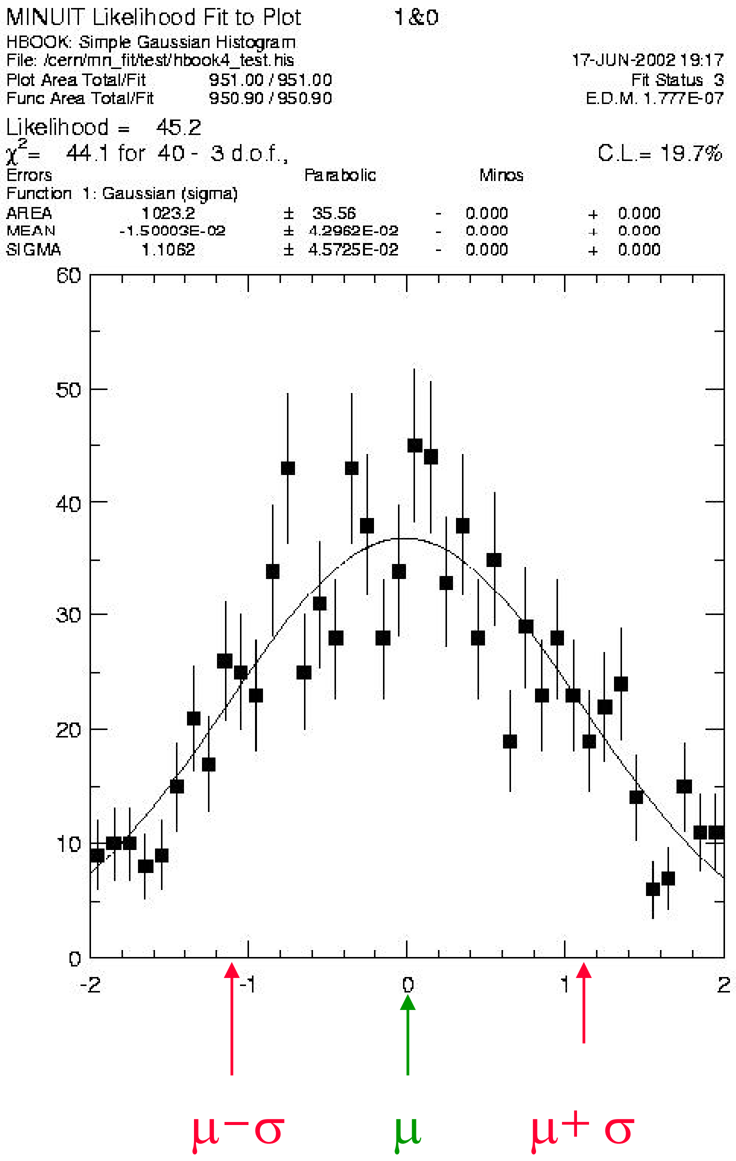


$\sigma \sim 87$ μm

Gaussian fitting

Using Mn_fit

$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx \Big/ \int_{-\infty}^{+\infty} f(x)dx = 0.68$$



Significant Figure

- The measured value has meaning by significant figures
- Significant Figure
  - It includes the first figure of uncertainty
  - All the figures between LSD (least significant digit) and MSD(Most significant digit)
    - LSD
      - If there is no point : The far right non-zero figure ex)23000
      - If there is point : The far right figure ex) 0.2300
    - MSD : The far left non-zero figure

Significant Figure (Example)

- 4 digit : 1234, 123400, 123.4, 1000.
- 4 digit : 10.10, 0.0001010,   100.0, 1.010×10^3
- 3 digit : 1010   cf) 1010. (Four digit of significant figure)

The calculation

- Add and Subtract
  - The last result is decided by the minimum point of calculations
  - Example)

```
    123
+      5.35
---------
    128.35
```

$128.3\cancel{5}$

```
     1.0001 ( 5 digit of SF)
+    0.0003 (1 digit of SF)
---------
     1.0004 (5 digit of SF)
```

Calculations (cont'd)

- Multiply and Divide
  - Same as the minimum digit of significant figure
  - Example)

    $16.3 \times 4.5 = 73.\cancel{35}$

    $\Rightarrow 73$

Propagation of Errors

- Suppose that *$(x_1, x_2, \ldots)$* is the variables, then variation of the function of *$F(x_1, x_2, \ldots)$* is as follows:
  - In case that there is no correlation between variables

$$\sigma_F^2 = \left(\frac{\partial F}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial F}{\partial x_2}\right)^2 \sigma_2^2 + \left(\frac{\partial F}{\partial x_3}\right)^2 \sigma_3^2 \ldots$$

Propagation of Errors (continued)

- Suppose that $(x_1, x_2, \ldots)$ is the variables, then variation of the function of $F(x_1, x_2, \ldots)$ is as follows:
  - In case that there is correlation between variables

$$\sigma_F^2 = \sum_{i,j} \left(\frac{\partial F}{\partial x_i}\right) \left(\frac{\partial F}{\partial x_j}\right) \sigma_i \sigma_j$$

$\Rightarrow$ Let us consider only non-correlation case.

Combining Errors

- Add or Subtract ($F=x_1+x_2$ or $F= x_1-x_2$)

$$\sigma_F = \sqrt{\sigma_1^2 + \sigma_2^2}$$

Example) $x_1 = 100. \pm 10.$

$+ \, x_2 = 400. \pm 20.$

$F = 500. \pm 22.$

Example) The error of the measurement

$$\sigma = \sqrt{\sigma_{stat}^2 + \sigma_{sys}^2}$$

Combining Errors (cont'd)

- *F=ax* (a is constant)

$$\sigma_F = a\sigma$$

Example) x =100. ± 10.

a = 5

\-\-\-\-\-\-\-\-\-\-\-\-\-

F = 500. ± 50.

Combining Errors (cont'd)

- Multiplication ($F = x_1 \cdot x_2$)

$$\sigma_F = x_1 x_2 \sqrt{(\sigma_1 / x_1)^2 + (\sigma_2 / x_2)^2}$$

Example) x_1 = 100. ± 10.

x_2 = 400. ± 20.

F = (400. ± 45.) X 10^2

Combining Errors (cont'd)

- Division ($F = x_1 / x_2$)

$$\sigma_F = (x_1 / x_2)\sqrt{(\sigma_1 / x_1)^2 + (\sigma_2 / x_2)^2}$$

Example) $x_1 = 100. \pm 10.$

$x_2 = 400. \pm 20.$

$F = 0.250 \pm 0.028$

Combining results Using weighting factor

- Cases
  - With different detection efficiencies (RunI, RunII)
  - With different parts of apparatus (SVX, COT)
  - With different experiment (CDF, D0)
  - With different decay mechanisms

    ex) Bs->Psi(2s) Phi

    1) Psi(2s) ->J/Psi mu+ mu-

    2) Psi(2s) -> mu+mu-

    ex) D0-> KsKs

    1) D*+ -> D0 pi+

    2) D*0 -> D0 pi0

Combining results Using weighting factor (cont'd)

- Average
  - There is N data whose values are $(x_1, x_2, \ldots x_k, \ldots x_N)$
  - Suppose that the error of X_k is σ_k

$$\bar{x} = \frac{\sum w_k x_k}{\sum_k w_k}$$

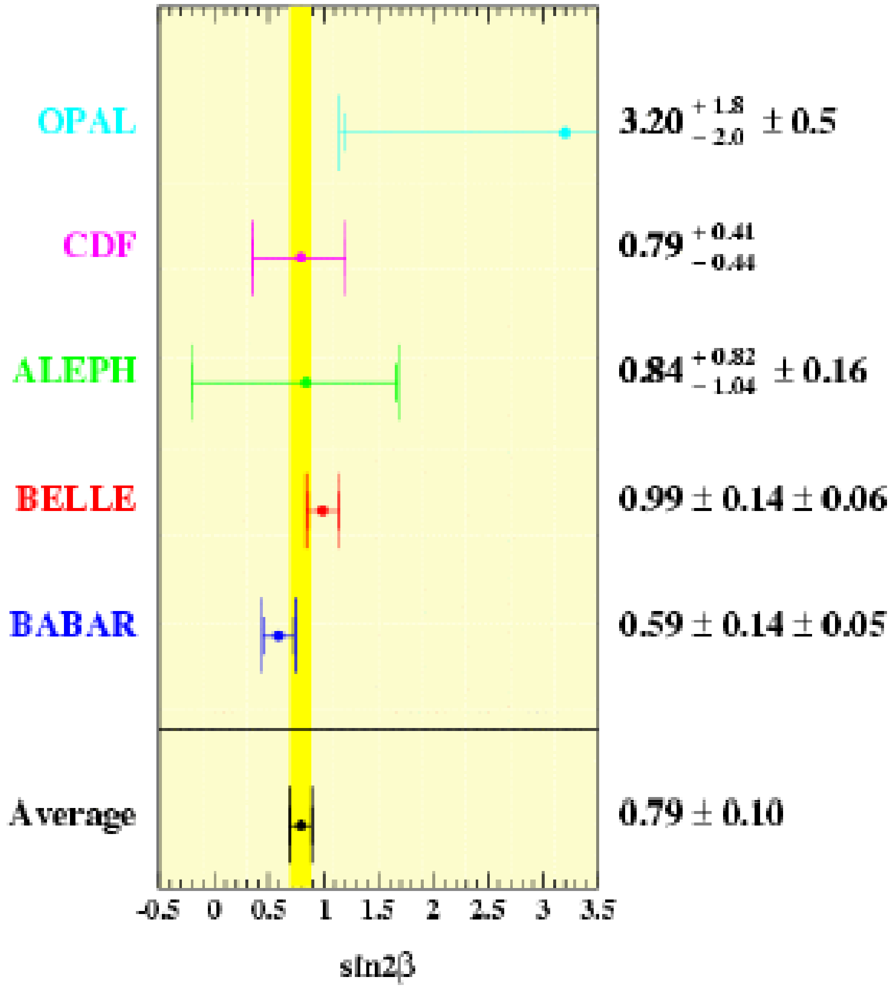
where weighting factor

$$w_k = 1/\sigma_k^{\ 2}$$

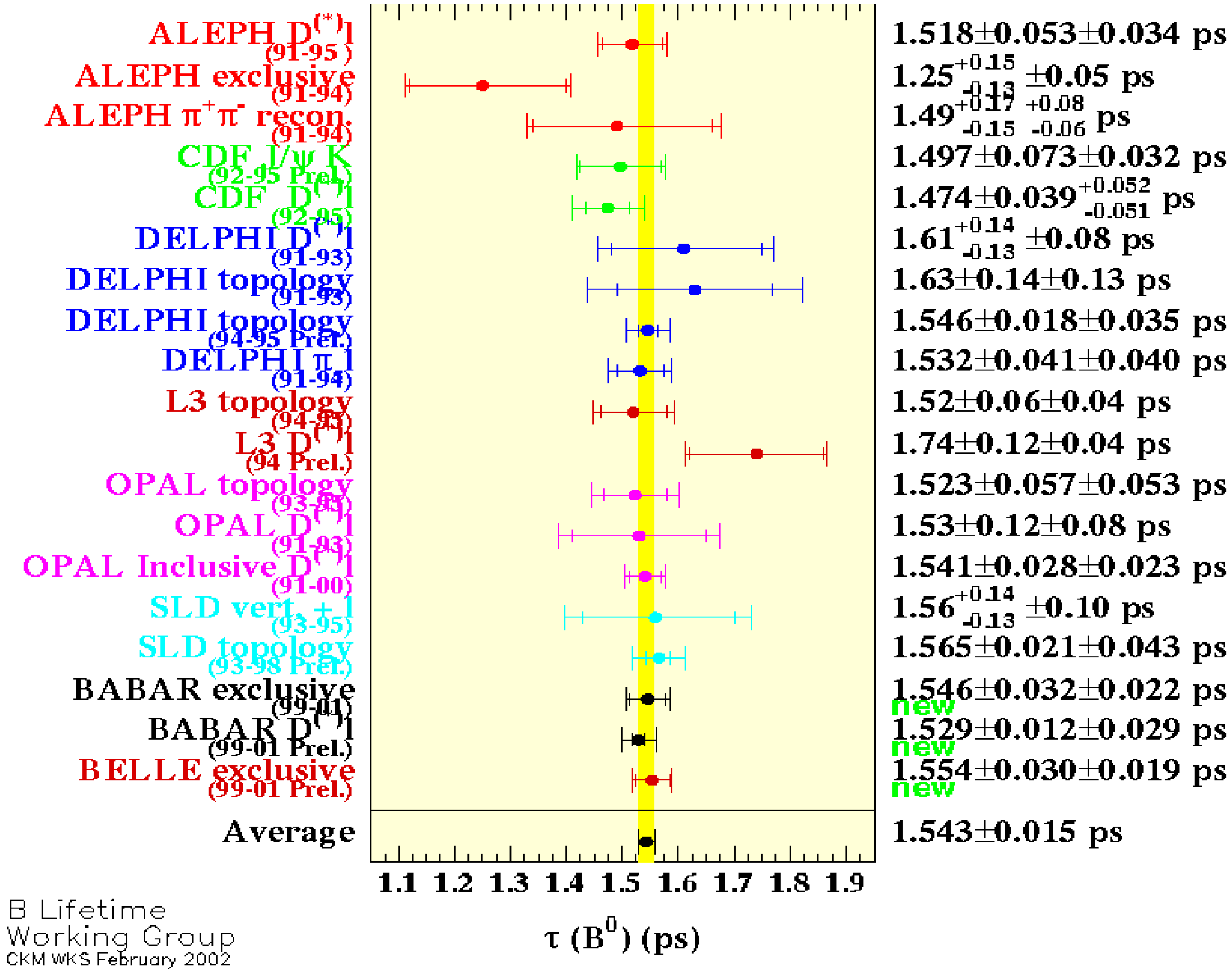
- Error :

$$\sigma^2 = 1/\sum w_k$$

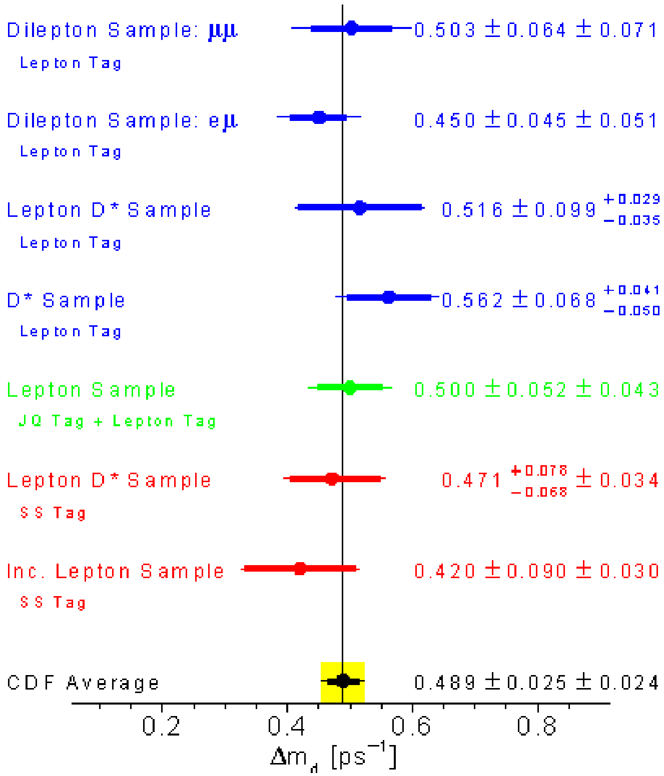
Ex) World Average of sin(2β)



Ex) B^0 lifetime summary



Ex) CDF B_d Mixing



Upper Limit

- Measurement ($B = B_m \pm \sigma$)

- Observation ($B_m > 5\sigma$)
  - Signal is greater than 5 sigma of error.
- Evidence ($3\sigma < B_m < 5\sigma$)
  - Signal is greater than 3 sigma of error, however less than 5 sigma.
- Upper Limit ($3\sigma > B_m$)
  - Signal is less than 3 sigma.

Upper Limit B_l (cont'd)

- Method I. General Case

Measurement $B = B_m \pm \sigma$

$B_l < B_m + 1.28\sigma$ (90% CL)

1.64σ (95% CL)

2.33σ (99% CL)

Measurement $B = B_m \pm \sigma$

Ex) $B_l = (3 \pm 5) \times 10^{-9}$

$B_l < (3+1.28 \times 5) \times 10^{-9}$ at 90% CL

or $B_l < 9.4 \times 10^{-9}$ at 90% CL

Upper Limit B_l (cont'd)

- Method 2. Negative B_m
  - Background Subtracted
  - Example)
    - $B_m = (-1 \pm 1) \times 10^{-9}$
    - $B_m = (\ \ 0 \pm 1) \times 10^{-9}$
  - Upper Limit at 90 % CL Level
    - g is Gaussian (Mean is B_m , width is σ)

$$\frac{\int_0^{B_l} g\,dB}{\int_0^{\infty} g\,dB} = 0.9$$

Compare Upper Limit (90% CL)

B_m	Method 1	Method 2
4	5.3	5.3
3	4.3	4.3
2	3.3	3.3
1	2.3	2.4
0.5	1.8	2.0
0	1.3	1.6
−0.5	0.8	1.4
−1	0.3	1.2
−2	−0.7	0.8
−3	−1.7	0.6
−4	−2.7	0.5

Assume
$\sigma=1$

Ex) CP Asymmetry in Charm
$(D^+ \to K^-K^+\pi^+)$

$$\eta(D) = \frac{N(D^+ \to K^-K^+\pi^+)}{N(D^+ \to K^-\pi^+\pi^+)}$$

$$\eta(D) = \frac{N(D^0 \to K^-K^+)}{N(D^0 \to K^-\pi^+)}$$

C.F.

$$A_{CP} = \frac{\eta(D) - \eta(\bar{D})}{\eta(D) + \eta(\bar{D})}$$

- Cabibbo Suppressed mode

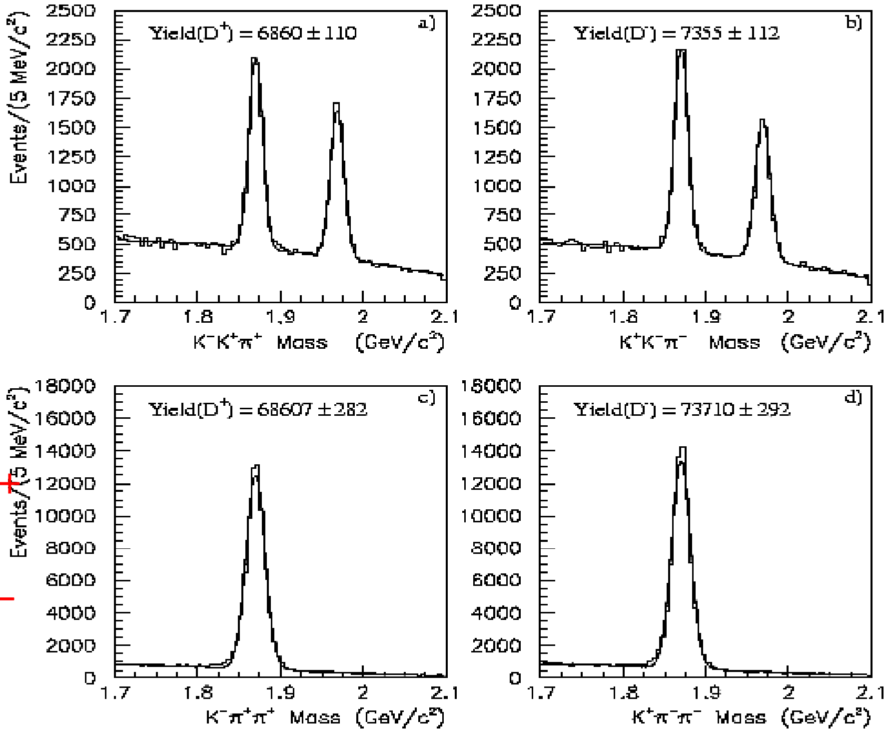
$D^+ \to K^-K^+\pi^+$

$D^- \to K^-K^+\pi^-$

- Cabbibo Favored mode

$D^+ \to K^-\pi^+\pi^+$

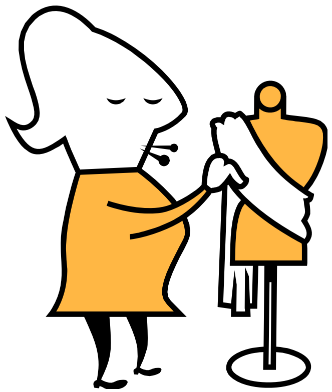
$D^- \to K^+\pi^-\pi^-$



A=0.006±0.011 ± 0.005

A < 0.025 at 95 %CL

Fitting



Fitting Methods

1. Moments
   - Simple, but inefficiency

2. Maximum likelihood Method
   - Can be used only if the theoretical distribution is known.
   - More general case

3. Least Square Method
   - In case of statistical error

Example) For a given N data of (x_i, y_i), let us fit using a linear equation of $y=ax+b$

1. Moment

- Method is to calculate the average
- Simplicity
- Example
  - A linear equation

$$y_i = ax_i$$

    - Parameter a is

$$a = (\sum_{i=1}^{n} \frac{y_i}{x_i}) / n$$

2. Maximum likelihood Method

- The likelihood *L*

$$L(\Gamma) = \prod_{i=1}^{n} y_i(\Gamma)$$

  - Where Γ is the parameter to find
  - y_i is the function given variable xi
- To find maximize *L*
- To maximize *l= log L*
- Normalization is essential.
- Ex) A linear equation

$$y_i = ax_i + b$$

$$L(a,b) = \prod_{i=1}^{n} y_i(a,b)$$

Maximum likelihood Method (cont'd)

- Can be used only if the theoretical distribution is known.
- The most powerful one for finding the values of unknown parameters
- No histogram needed (event by event)
- Efficient Method → Most case works
- We can transform one variable to another

Ex)

$$\lambda_0 = 1/\tau_0$$

3. Least Square Method

- Least Square Method for Simple Case
  - The first order of polynomials (linear equation y=ax+b)
  - For a given N data of (x_i, y_i), let us fit using a linear equation of y=ax+b
  - To find a and b which is the minimization of the sum of distance between data and equation . i.e. when we put Q as follows:

$$Q = \sum_i (a + bx_i - y_i)^2$$

  - Let us find a and b which satisfies the following equations

$$\frac{\partial Q}{\partial a} = 0 \quad \& \quad \frac{\partial Q}{\partial b} = 0$$

3. Least Square Method

- Least Square Method for Simple Case with errors
  - The first order of polynomials (linear equation y=ax+b)
  - For a given N data of (x_i, y_i, σ_i), let us fit using a linear equation of y=ax+b
  - To find a and b which is the minimization of the sum of distance between data and equation . i.e. when we put Q as follows:

$$Q = \sum_i [(a + bx_i - y_i)/\sigma_i]^2$$

  - Let us find a and b which satisfies the following equations

$$\frac{\partial Q}{\partial a} = 0 \quad \& \quad \frac{\partial Q}{\partial b} = 0$$

Least Square Method (Continued)

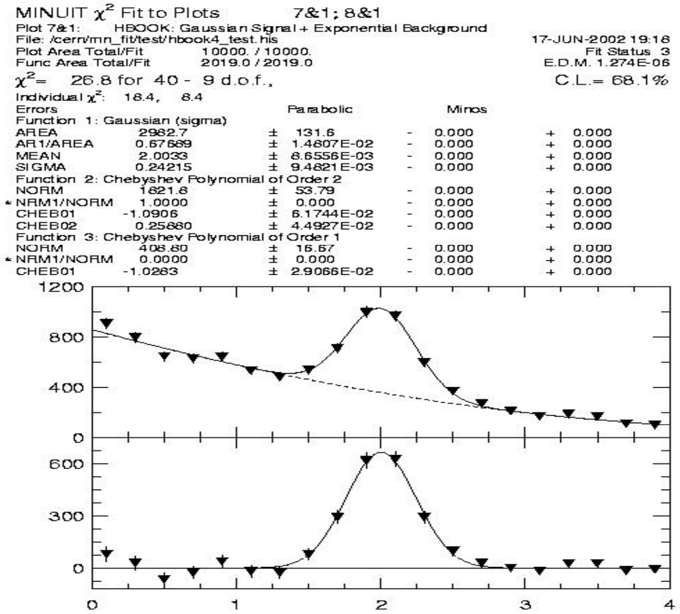
- Least Square Method for Linear Polynomials
  - m of unknown parameters (a_1, a_2, a_3,... a_m)
  - $F(x)=a_1 f_1(x)+a_2 f_2(x)+ \quad + a_m f_m(x)$
  - It is same as linear least square method
  - There will be m equations and solutions

- Least Square Method for Non-linear Equation
  - Let us expansion as a linear polynomial using Taylor series.

Least Square Method (Example)

- Mn_fit used
- Least Square Method

- Signal is gaussian.
- Background is Chebyshev polynomial.



근 사 이 론
(Least Squares)

양유철 ycyang@hep.knu.ac.kr

근사이론의 형태

근사이론(최소 제곱법)에는 두가지 형태의 문제와 관련이 있다.

첫번째 형태는 주어진 데이터에 함수를 맞추는 것으로서 그 데이터를 표현하는데에 사용될수 있는 어떤 부류의 함수들 중에서 데이터를 표현하는 데에 사용할 수 있는 가장 적절한 함수를 찾는 것과

예) Linear Least Squares 등

두번째 형태는 함수가 명시적으로 주어졌지만 다항식과 같은 단순한 형태의 함수 표현을 찾고자 하는것

예)

$$\sin \pi x = -4.12251 \ x^2 + 4.12251 \ x - 0.50465$$

절대 편차 이용 ?

1. $y_i = a_1 x_i + a_0$

오차 : $E(a_0, a_1) = \sum_{i=1}^{n} | y_i - (a_1 x_i + a_0) |$

1) $\frac{\partial}{\partial a_0} \sum_{i=1}^{n} | y_i - (a_1 x_i + a_0) | = 0$

2) $\frac{\partial}{\partial a_1} \sum_{i=1}^{n} | y_i - (a_1 x_i + a_0) | = 0$

=> 절대치 함수가 0에서 미분 불가능.
두방정식의 해를 반드시 구할 수 없음.
-> 최소 제곱법

선형 최소 제곱법

오차 : $$E(a_0, a_1) = \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2$$

1) $$\frac{\partial}{\partial a_0} \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2 = 2\sum_{i=1}^{n} (y_i - a_1 x_i - a_0)(-1) = 0$$

2) $$\frac{\partial}{\partial a_1} \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2 = 2\sum_{i=1}^{n} (y_i - a_1 x_i - a_0)(-x_i) = 0$$

정규 방정식(normal equation)

1) $$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

2) $$a_0 n + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$y = a_1 x - a_0$$

$$a_0 = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i y_i)(\sum_{i=1}^{n} x_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$a_1 = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

: Linear Least Squares

예제 1

x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

$$a_0 = \frac{(385)(81) - (55)(572.4)}{10(385) - (55)^2} = -0.360$$

$$a_1 = \frac{10(572.4) - (55)(81)}{10(385) - (55)^2} = 1.538$$

$$y = 1.538x - 0.360$$

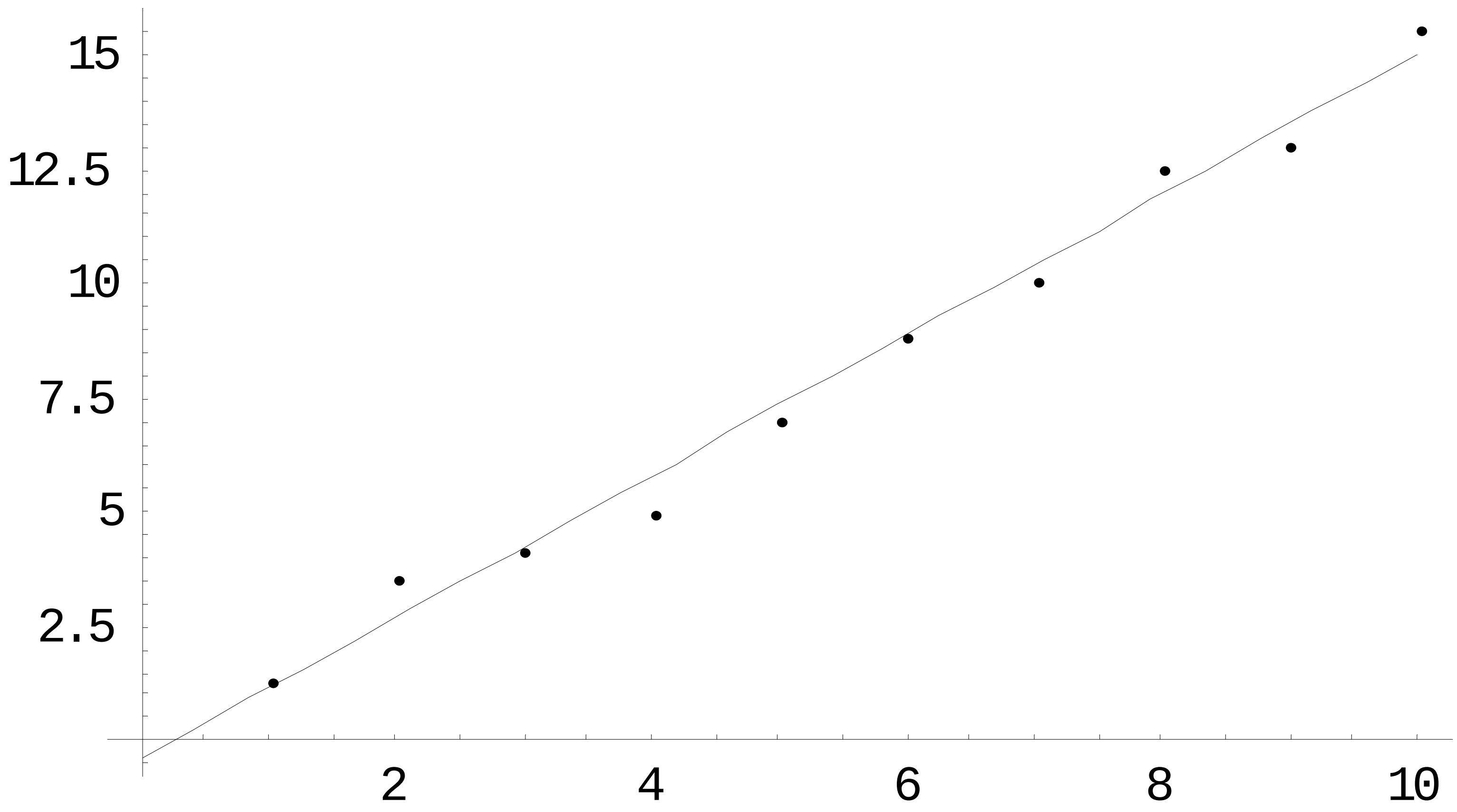


그림 – 예제 1

$$y_n(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_1 x + a_0$$

최소 제곱법을 이용하여 n차 대수 다항식을 구할수 있음.

최소 제곱 오차의 합을 최소

$$\frac{\partial}{\partial a_j} \sum_{i=1}^{m} [y_i - y_n(x_i)]^2 = 0$$ (j = 0,1,2,3, … ,n)

정규 방정식 : (n+1)개의 미지수 를 갖는 (n+1)개의 정규방정식

$$a_0 \sum_{i=1}^{m} x_i^0 + a_1 \sum_{i=1}^{m} x_i^1 + a_2 \sum_{i=1}^{m} x_i^2 + \cdots + a_n \sum_{i=1}^{m} x_i^n = \sum_{i=1}^{m} y_i x_i^0$$

$$a_0 \sum_{i=1}^{m} x_i^1 + a_1 \sum_{i=1}^{m} x_i^2 + a_2 \sum_{i=1}^{m} x_i^3 + \cdots + a_n \sum_{i=1}^{m} x_i^{n+1} = \sum_{i=1}^{m} y_i x_i^1$$

$$a_0 \sum_{i=1}^{m} x_i^n + a_1 \sum_{i=1}^{m} x_i^{n+1} + a_2 \sum_{i=1}^{m} x_i^{n+2} + \cdots + a_n \sum_{i=1}^{m} x_i^{2n} = \sum_{i=1}^{m} y_i x_i^n$$

예 제 2

I	1	2	3	4	5
X_i	0	0.25	0.50	0.75	1.00
Y_i	1.0000	1.2840	1.6787	2.1170	2.7183
$y(x_i)$	1.0051	1.2740	1.6482	2.1279	2.7130
$Y_i - y(x)$	−0.0051	0.0100	0.0005	−0.0109	0.0053

$$5a_0 + 2.5a_1 + 1.875a_2 = 8.7680$$
$$2.5a_0 + 1.875a_1 + 1.5625a_2 = 5.4514$$
$$1.875a_0 + 1.5625a_1 + 1.3828a_2 = 4.4015$$

$$y_2(x) = 0.84316x^2 + 0.86468x + 1.0051$$

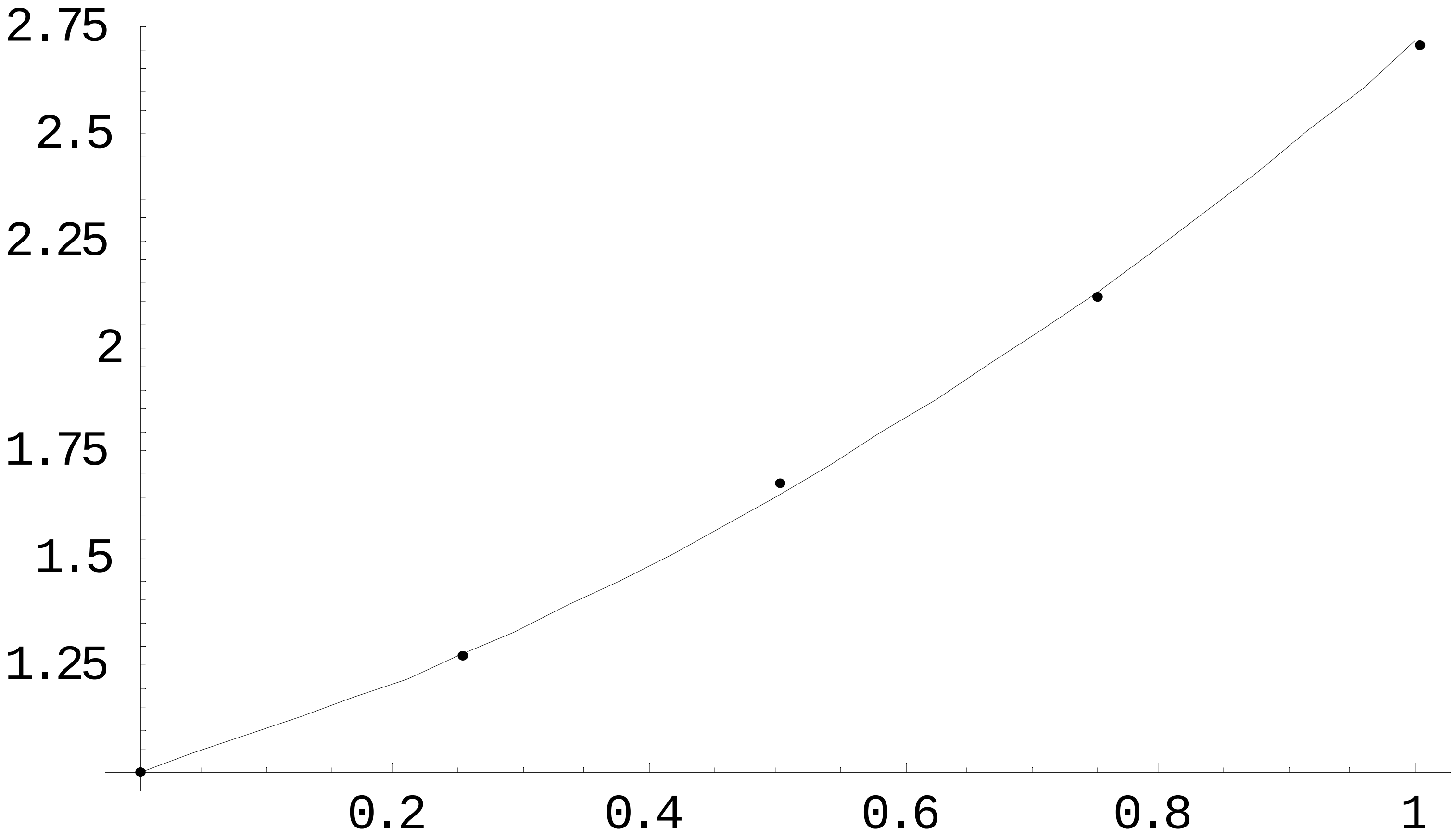


그림 - 예제 2

연속 최소 제곱법

$f \in C[a,b]$에서

오차 : $$E(a_0, a_1, \ldots, a_n) = \int_a^b (f(x) - P_n(x))^2 dx = \int_a^b (f(x) - \sum_{k=0}^{n} a_k x^k)^2 dx$$

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = \sum_{k=0}^{n} a_k x^k$$

오차의 최소화

$$\frac{\partial E}{\partial a_j} = -2 \int_a^b x^j f(x) dx + 2 \sum_{k=0}^{n} a_k \int_a^b x^{j+k} dx = 0$$

정규방정식

$$\sum_{k=0}^{n} a_k \int_a^b x^{j+k} dx = \int_a^b x^j f(x) dx, \quad (\mathrm{j} = 0,1,2, \cdots, \mathrm{n})$$

예 제 3

구간 [0.1] 상의 함수 $\sin(\pi x)$ 에 대한 2차 최소 자승 근사 다항식

$P_2(x) = a_2 x^2 + a_1 x + a_0$ 에 대한 정규 방정식은

$$a_0\int_0^1 1dx + a_1\int_0^1 xdx + a_2\int_0^1 x^2dx = \int_0^1 \sin(\pi x)dx \;\Rightarrow\; a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi}$$

$$a_0\int_0^1 xdx + a_1\int_0^1 x^2dx + a_2\int_0^1 x^3dx = \int_0^1 x\sin(\pi x)dx \;\Rightarrow\; \frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi}$$

$$a_0\int_0^1 x^2dx + a_1\int_0^1 x^3dx + a_2\int_0^1 x^4dx = \int_0^1 x^2\sin(\pi x)dx \;\Rightarrow\; \frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2-4}{\pi^3}$$

따라서 $a_0 = \dfrac{12\pi^2 - 120}{\pi^3} \approx -0.050465$, $\quad a_1 = -a_2 = \dfrac{720 - 60\pi^2}{\pi^3} \approx 4.12251$

$$f(x) = \sin \pi x \approx P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$$

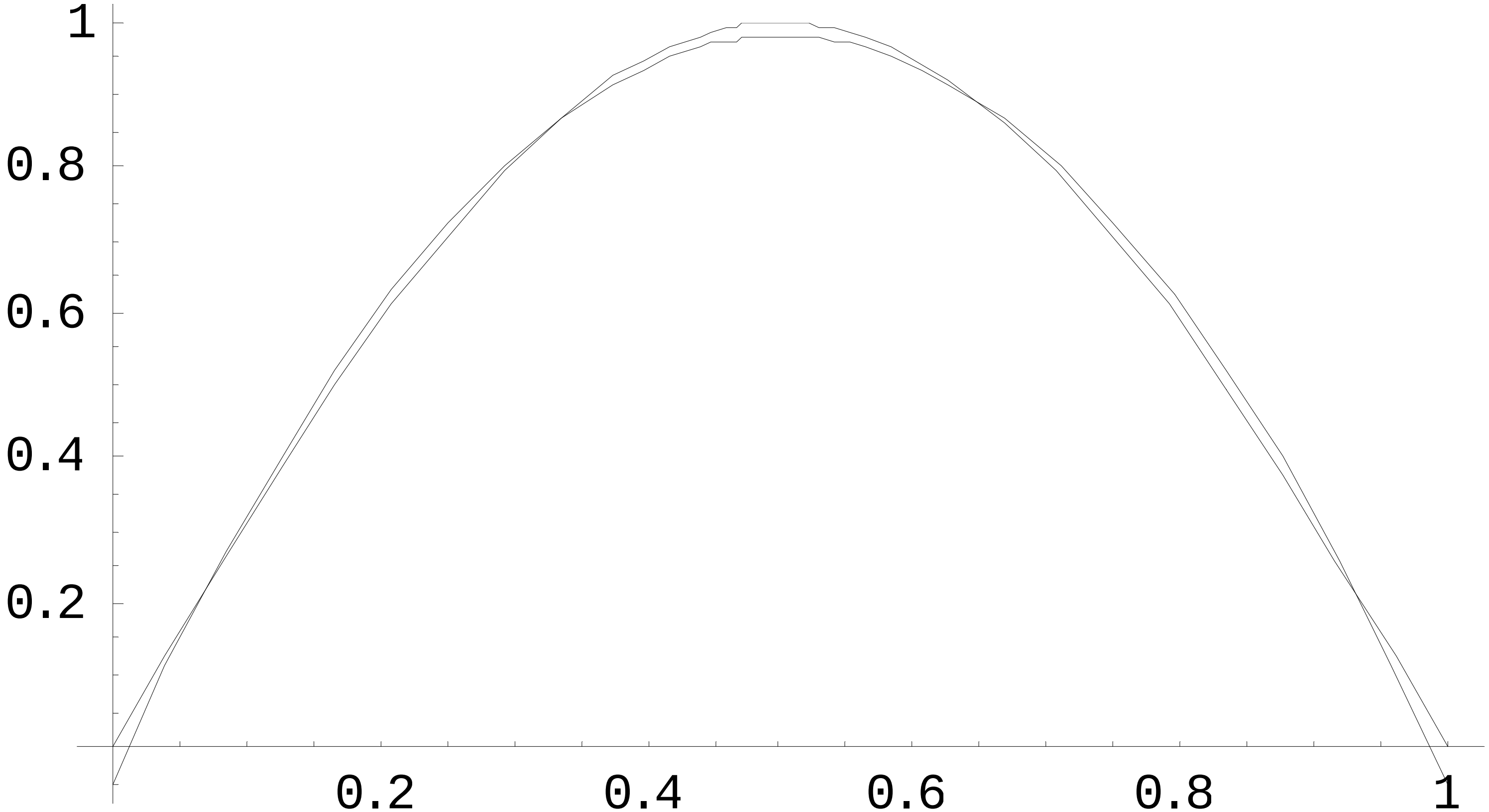


그림 – 예제 3

4. Maximum Likelihood vs. Least Square Method

	Maximum like.	Least Square
How easy	Normalization and maximization can be messy	Needs minimization
Efficiency	Usually most efficient	Sometime equivalent to max. like.
Input data	Individual events	Histograms
Estimate of goodness of fit	Very difficult	Easy
Zero event	Cover well	Troublesome

Maximum Likelihood = Least Square Method

- X-Y plane
- Errors in y-direction are Gaussian
- X-values are precisely determined

$\Rightarrow$The maximum likelihood and the least square methods are equivalent.

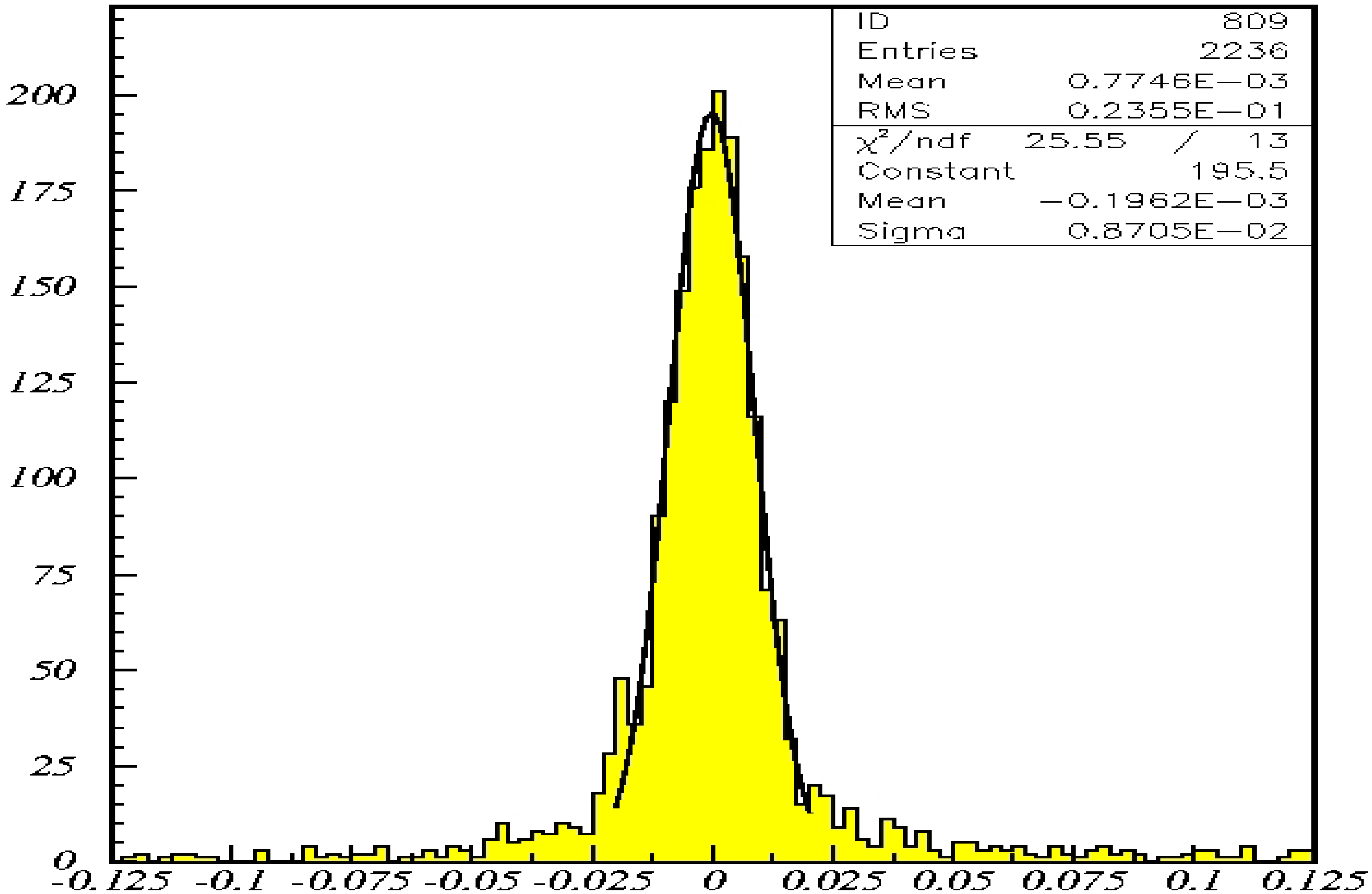
Example) Mass distributions

Fitting Package

- PAW
- Mn_fit
- Root
- ......

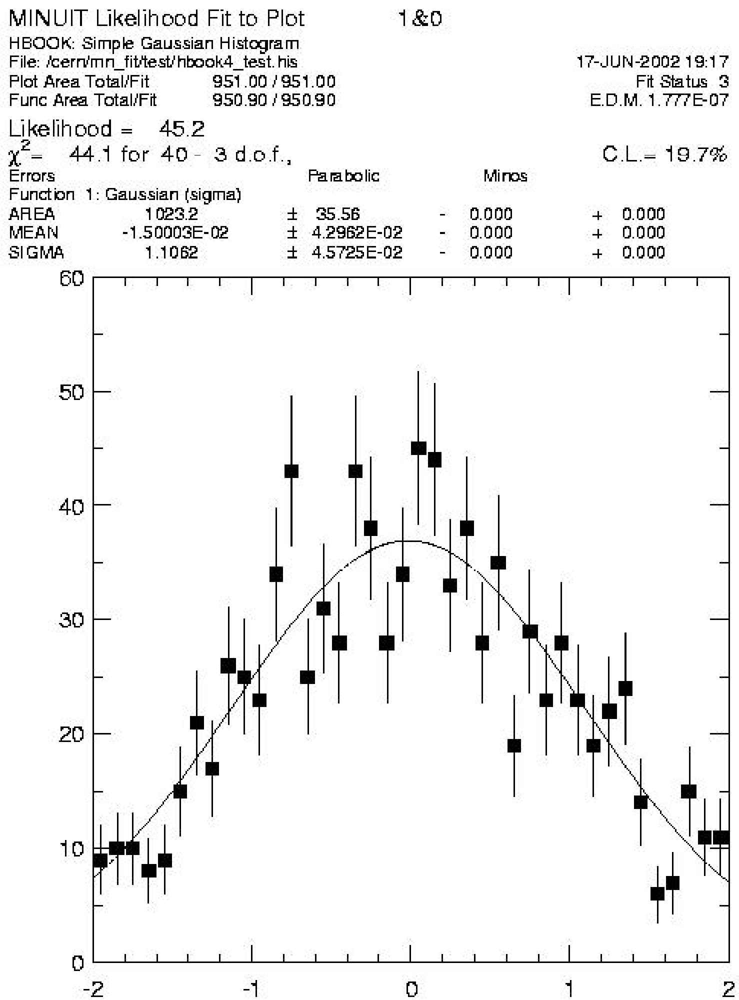
PAW

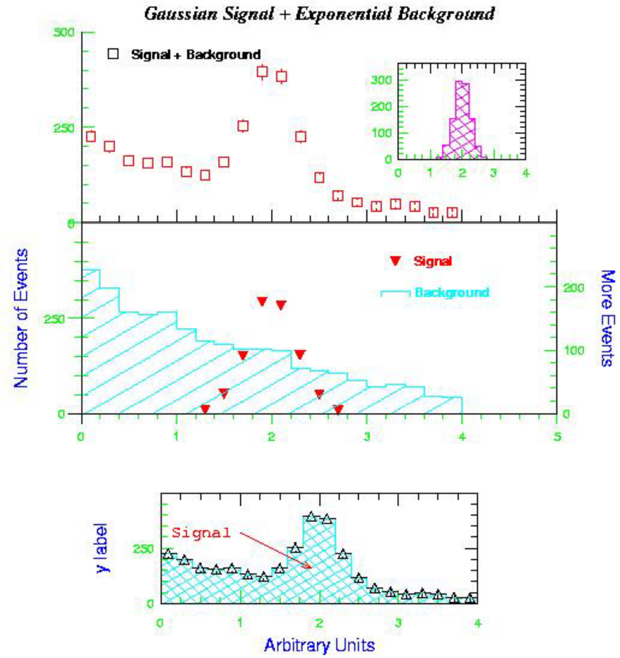
- Physics Analysis Workstation
- Inside of CERN library
- Ntuple – n dimensional variables
- Good to make histogram
- Include some fitting


ID 809
Entries 2236
Mean 0.7746E−03
RMS 0.2355E−01
χ²/ndf 25.55 / 13
Constant 195.5
Mean −0.1962E−03
Sigma 0.8705E−02
200
175
150
125
100
75
50
25
0
-0.125 -0.1 -0.075 -0.05 -0.025 0 0.025 0.05 0.075 0.1 0.125

Mn_fit

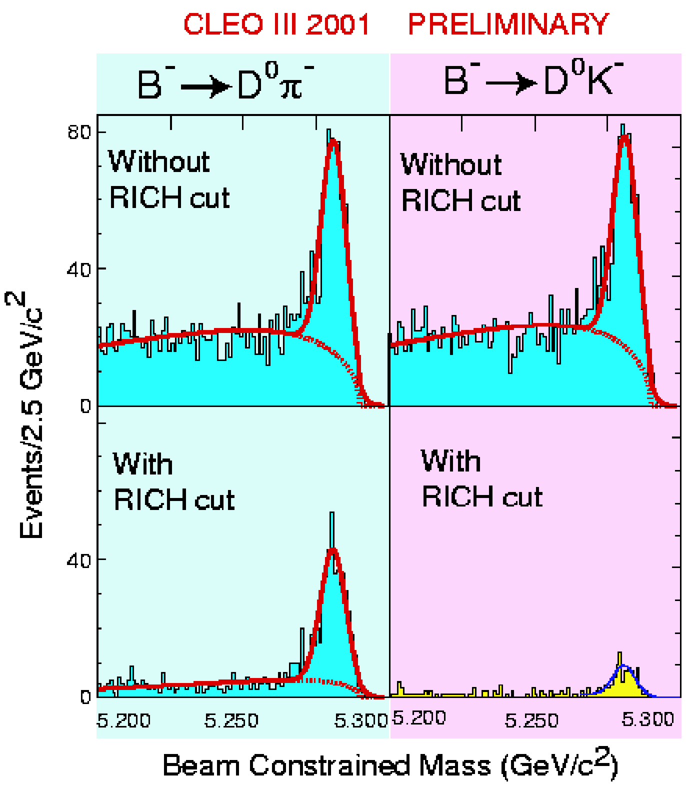
- Using fitting program in minuit at CERN library
- Powerful for fitting
- Easily check the results whether the fitting results are good or not.


MINUIT Likelihood Fit to Plot 1&0
HBOOK: Simple Gaussian Histogram
File: /cern/mn_fit/test/hbook4_test.his
17-JUN-2002 19:17
Plot Area Total/Fit 951.00 / 951.00
Fit Status 3
Func Area Total/Fit 950.90 / 950.90
E.D.M. 1.777E-07
Likelihood = 45.2
χ2= 44.1 for 40 - 3 d.o.f.,
C.L.= 19.7%
Errors Parabolic Minos
Function 1: Gaussian (sigma)
AREA 1023.2 ± 35.56 - 0.000 + 0.000
MEAN -1.50003E-02 ± 4.2962E-02 - 0.000 + 0.000
SIGMA 1.1062 ± 4.5725E-02 - 0.000 + 0.000


Gaussian Signal + Exponential Background
Signal + Background
Signal
Background
Number of Events
More Events
Signal
y label
Arbitrary Units

mn_fit (example)

- Signal is Gaussian
- Maximum likelihood is same as least square method



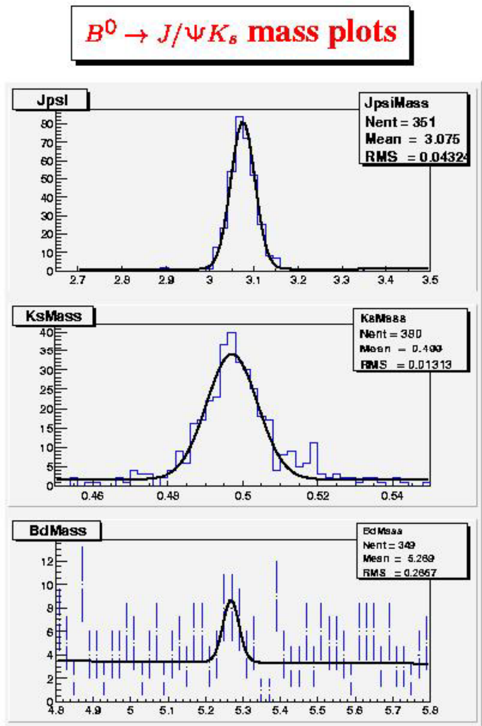
ROOT

- To Handle large data
- An object oriented HEP analysis Framework
- ROOT was created by Rene Brun and Fons Rademakers in CERN
- The ROOT system website is at http://root.cern.ch/

Differences from PAW

- Regular grammar (C++) on command line
- Single language (compiled and interpreted)
- Object Oriented (use your class in the interpreter)
- Advanced Interactive User Interface
- Well Documented code. HTML class descriptions for every class.
- Object I/O including Schema Evolution
- 3-d interfaces with OpenGL and X3D.

ROOT example



$N_B = 14 \pm 6$ events

- Data Processing is important for physicists

- Ref.
  - Louis Lyons, Statistics for nuclear and particle physicists (Cambridge Press)