Data Analysis

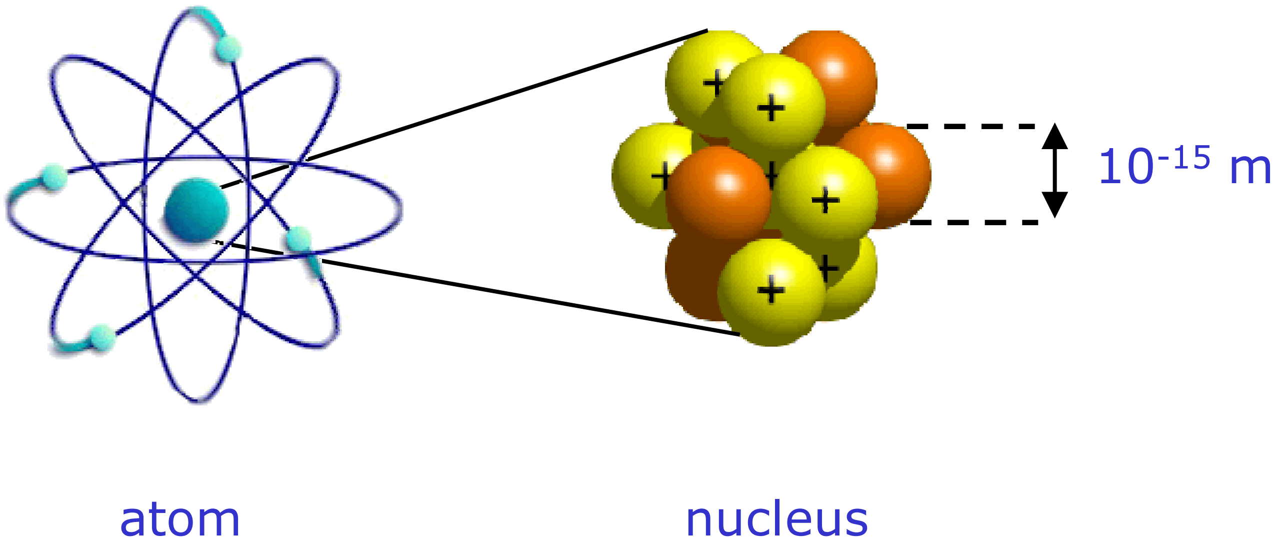
Wouter Verkerke
(NIKHEF)

HEP and data analysis

— General introduction

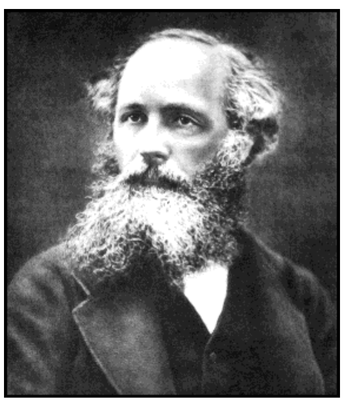
Wouter Verkerke, NIKHEF

Particle physics



Looking at the smallest constituents of matter → Building a consistent theory that describe matter and elementary forces

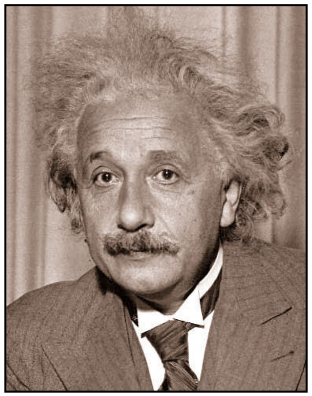
Newton



Maxwell

Theory of Relativity

Quantum Mechanics

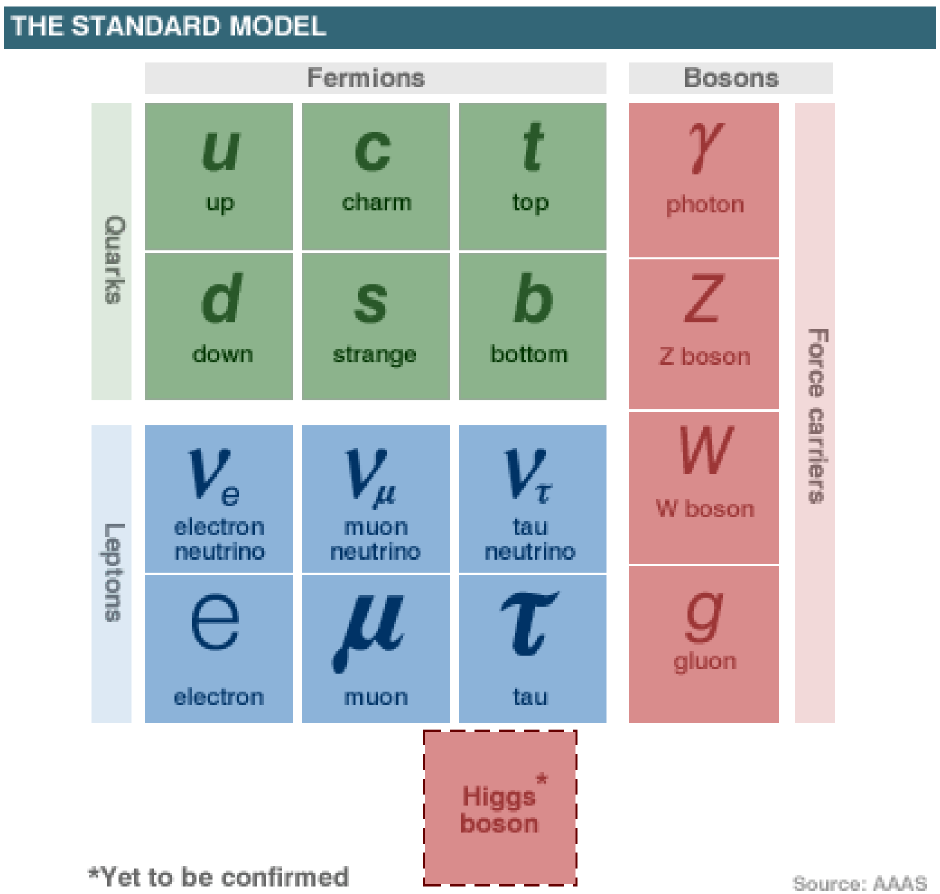


Einstein

Bohr

High Energy Physics

- Working model: 'the Standard Model' (a Quantum Field Theory)
  - Describes constituents of matter, 3 out of 4 fundamental forces

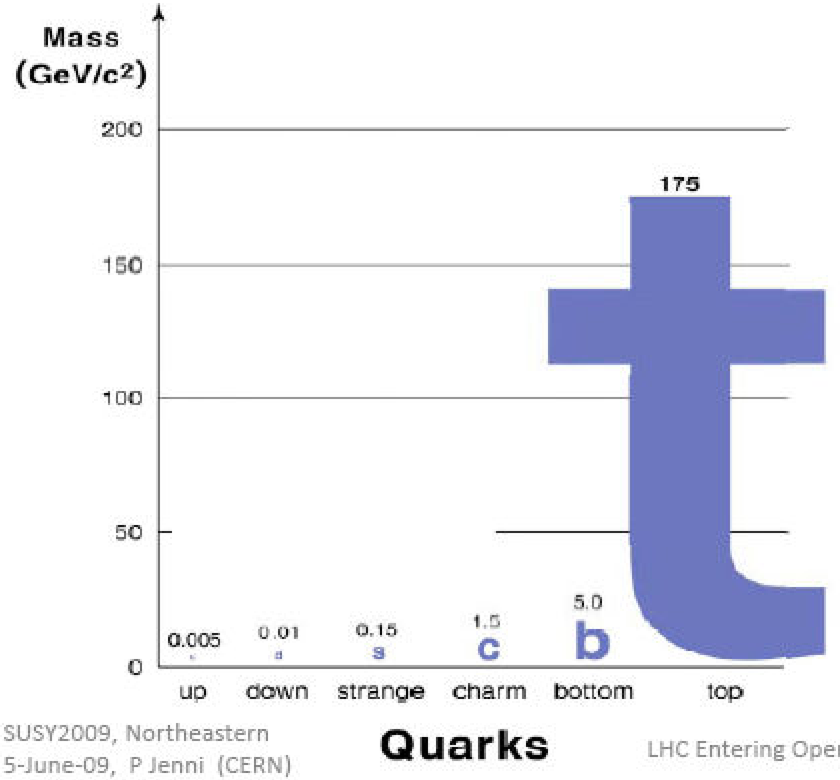


The standard model has many open issues

A most basic question is why particles (and matter) have masses (and so different masses)

The mass mystery could be solved with the 'Higgs mechanism' which predicts the existence of a new elementary particle, the 'Higgs' particle (theory 1964, P. Higgs, R. Brout and F. Englert)

Peter Higgs

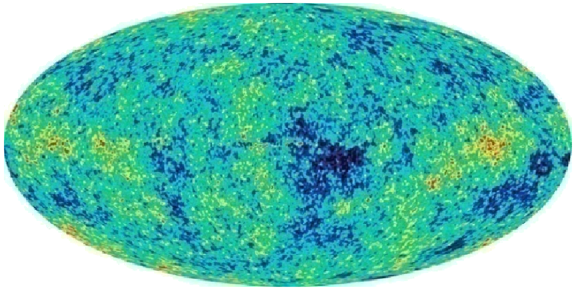


The Higgs (H) particle has been searched for since decades at accelerators, but not yet found…

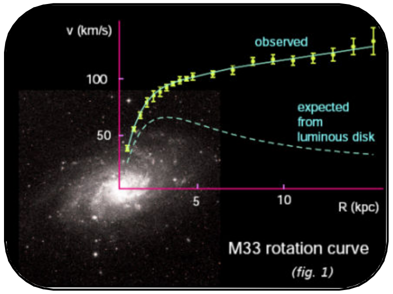
The LHC will have sufficient energy to produce it for sure, if it exists

Francois Englert

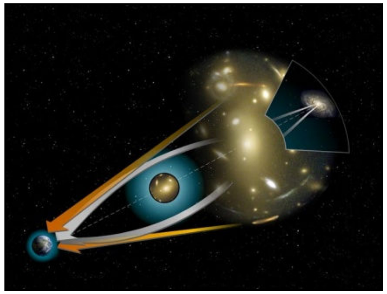
Link with Astrophysics



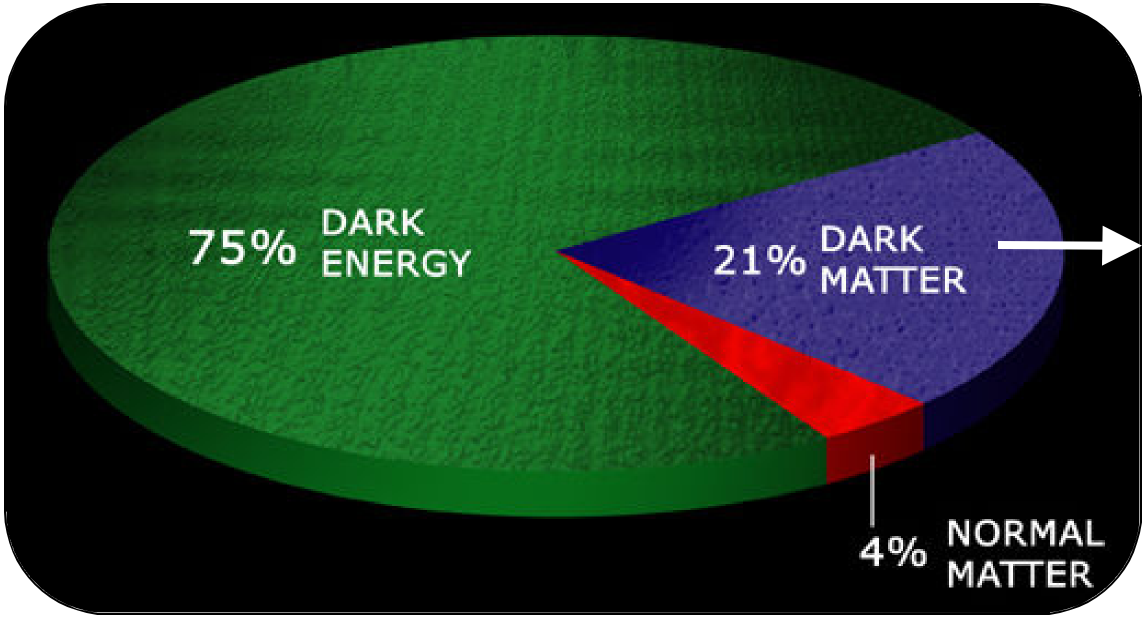
Temperature fluctuations
in Cosmic Microwave Background



Rotation Curves

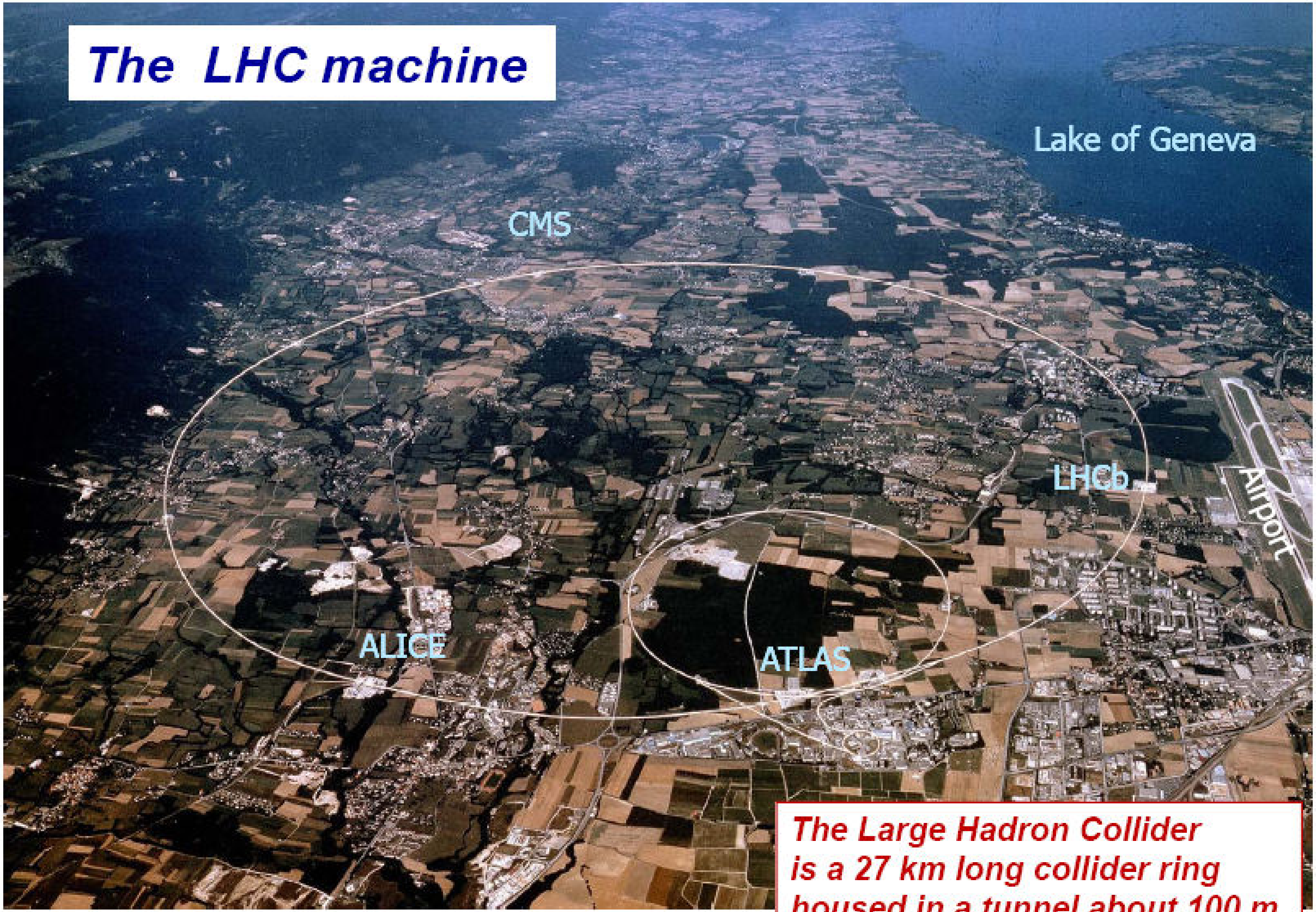


Gravitational
Lensing



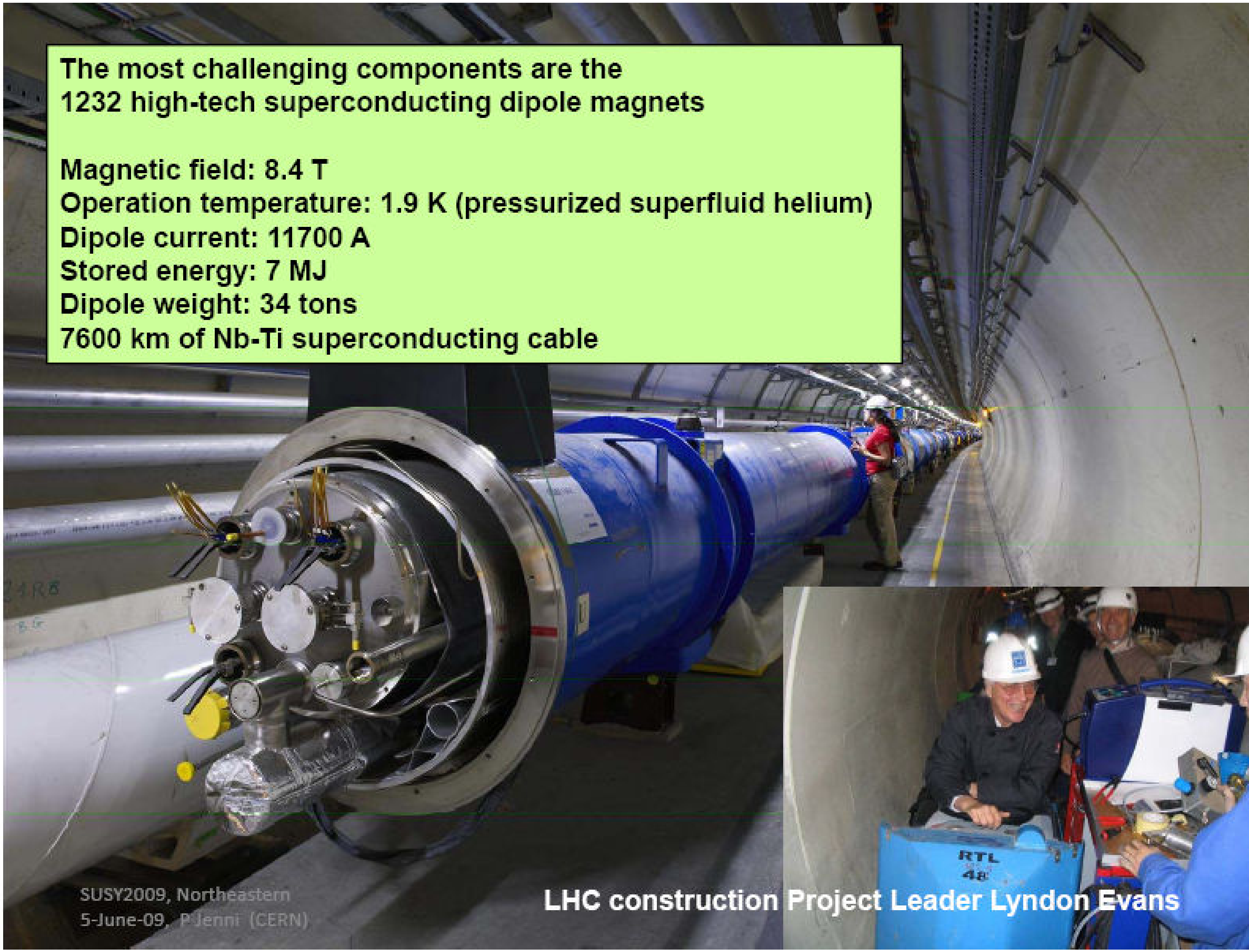
What is
dark matter?

Particle Physic today – Large Machines

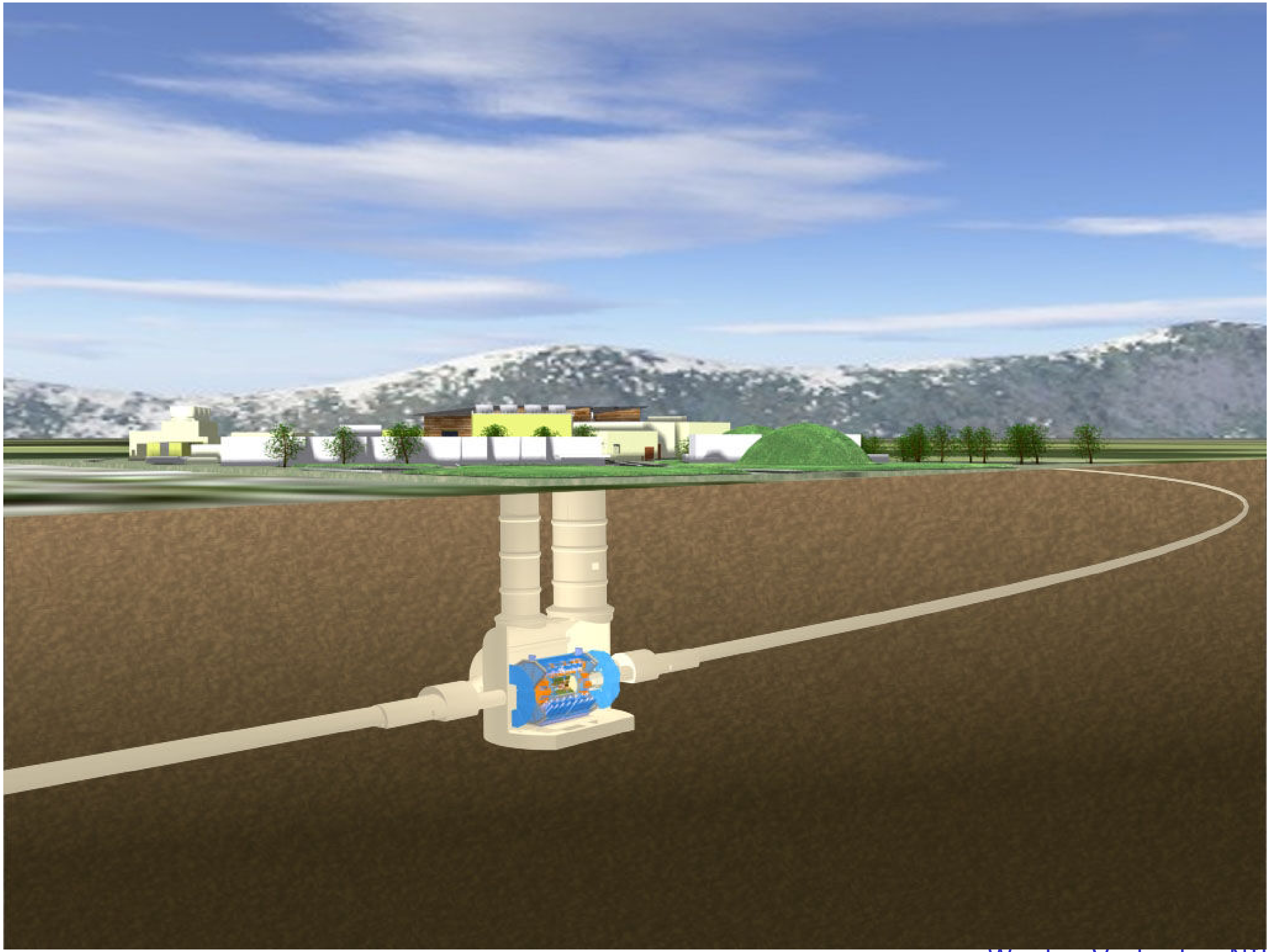


The Large Hadron Collider is a 27 km long collider ring housed in a tunnel about 100 m underground near Geneva

Detail of Large Hadron Collider

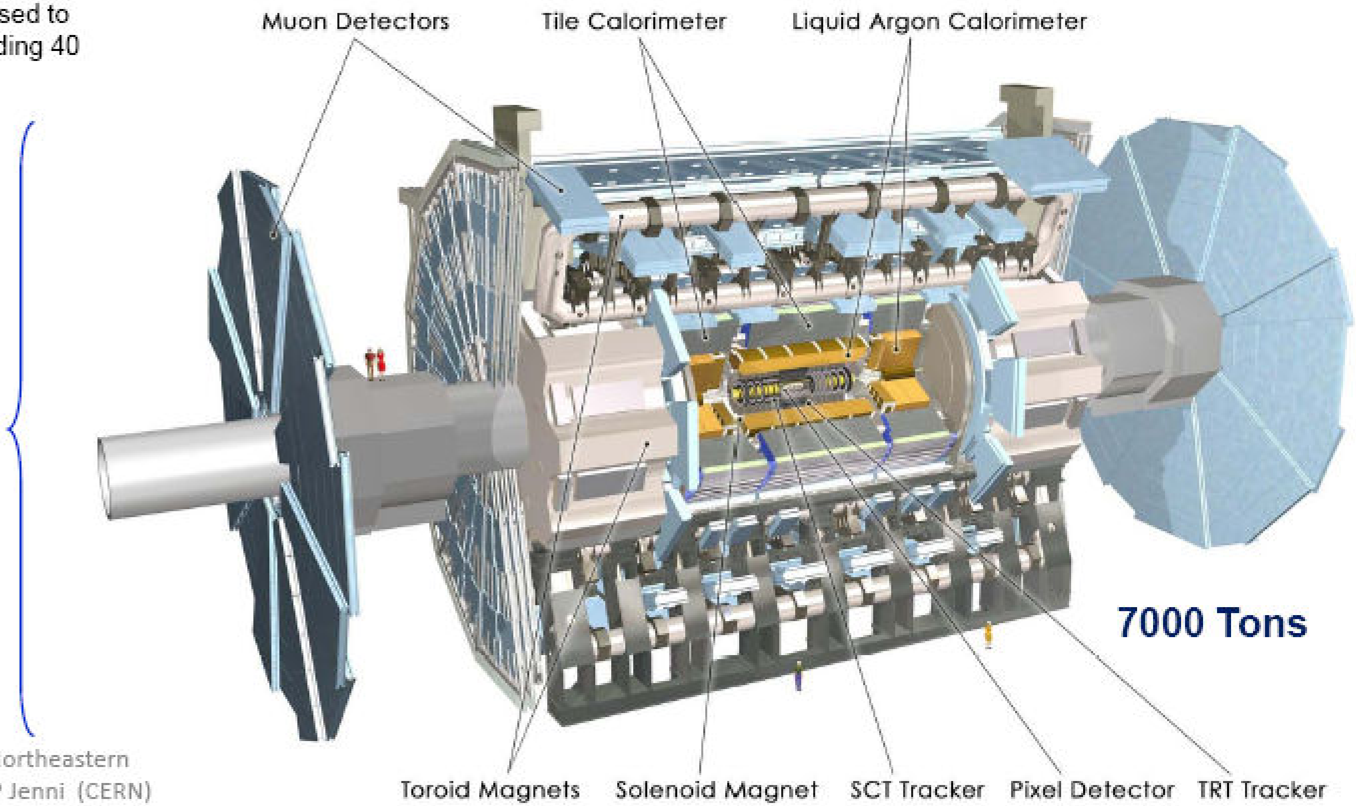


And large experiments underground

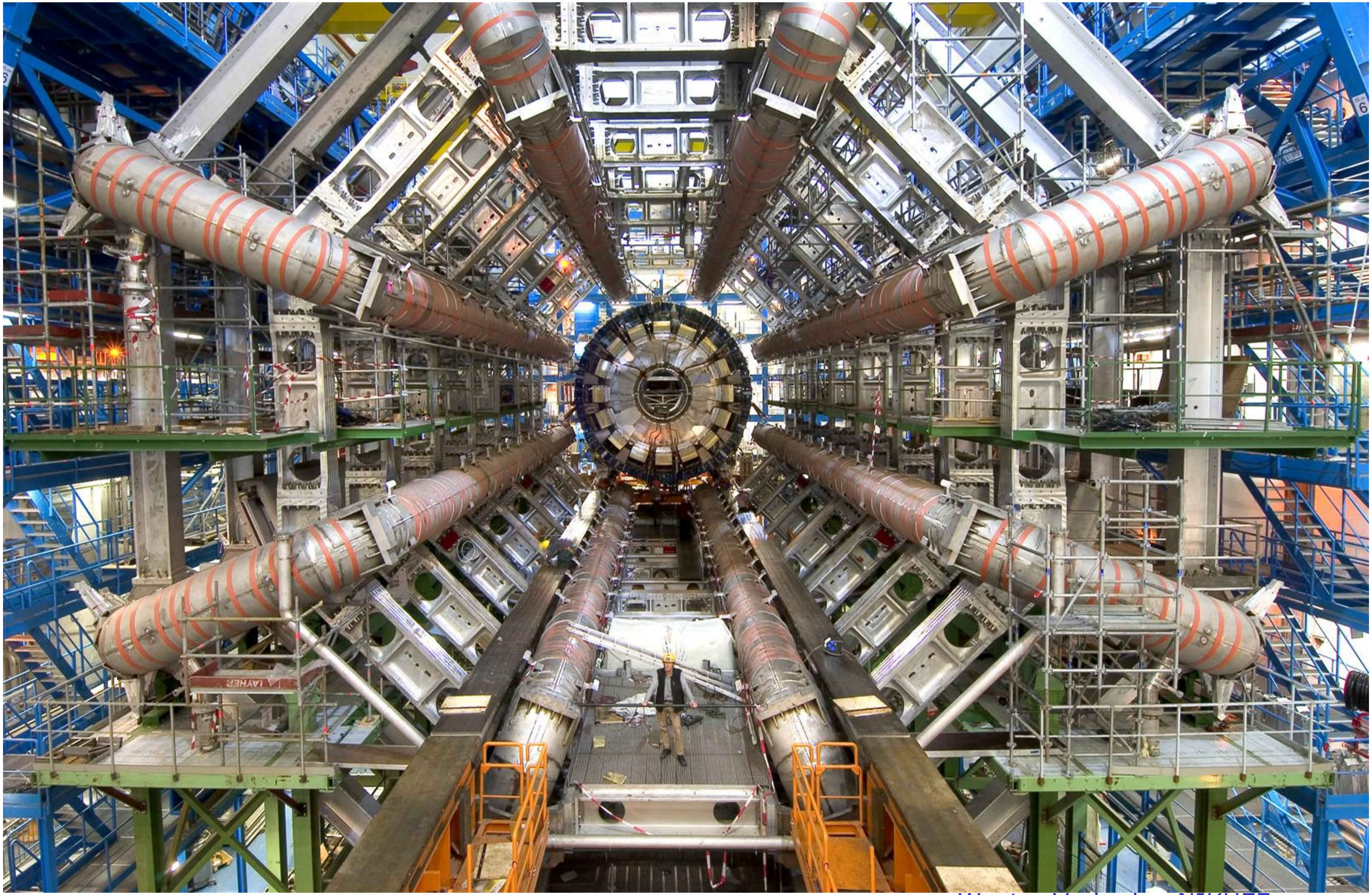


One of the 4 LHC experiments – ATLAS

ATLAS superimposed to
the 5 floors of building 40

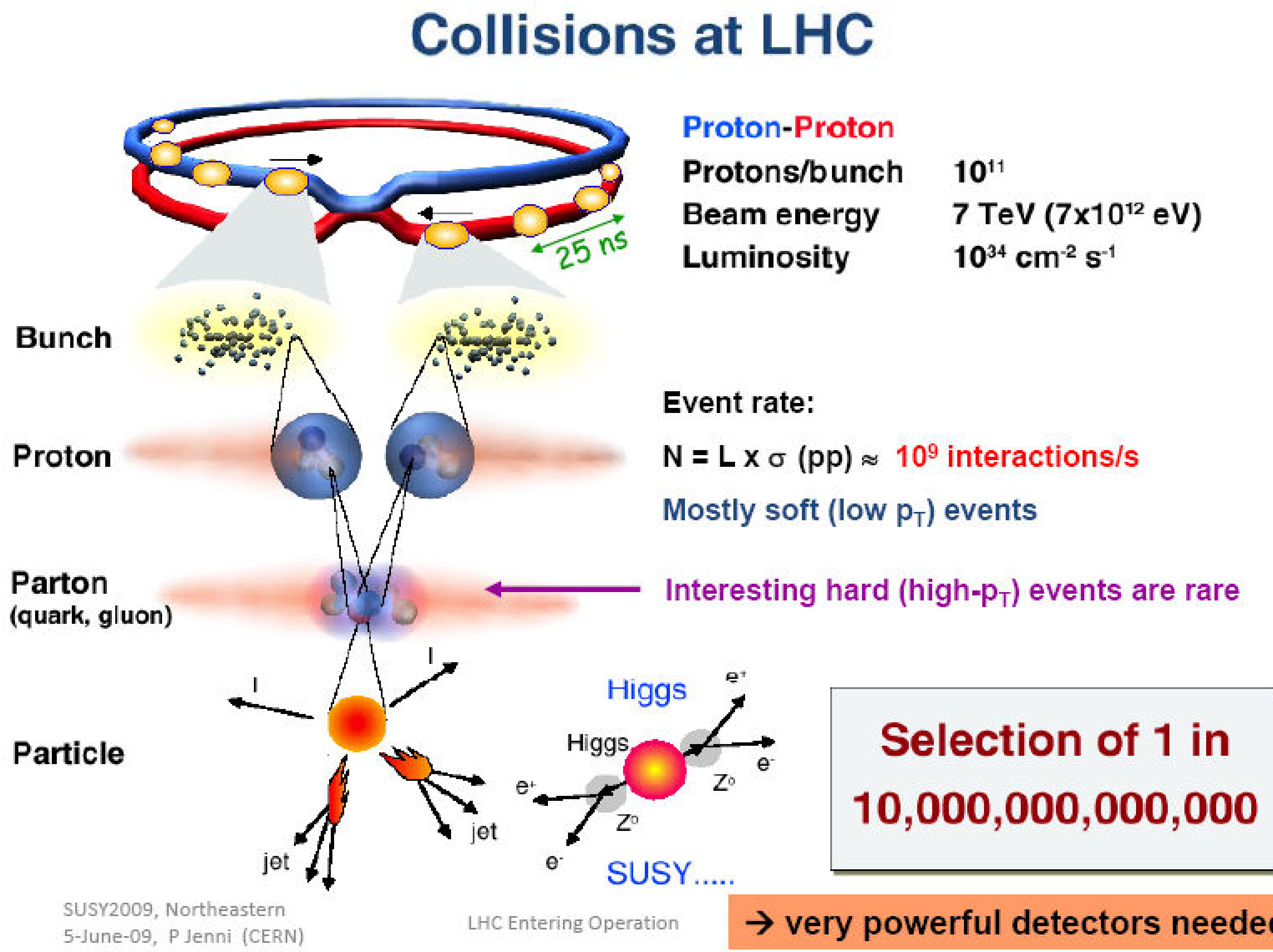


View of ATLAS during construction



Wouter Verkerke, NIKHEF

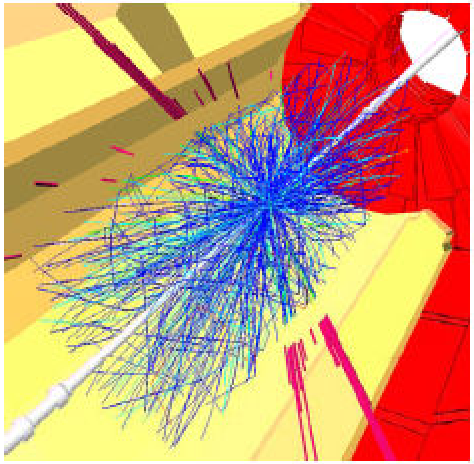
Collecting data at the LHC



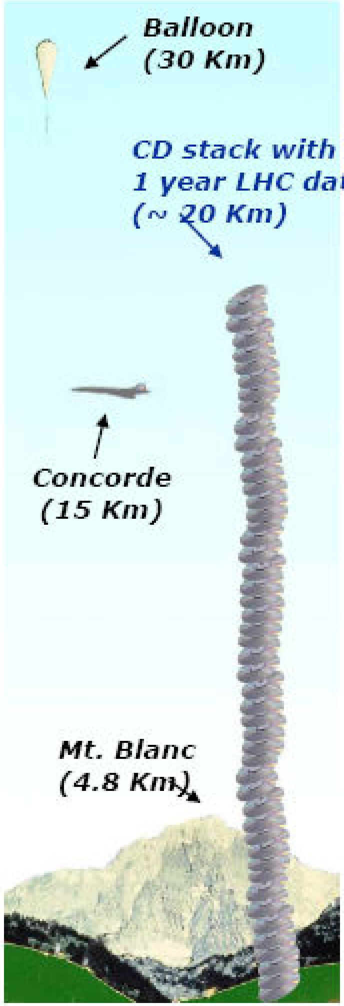
→ very powerful detectors needed

Data reduction, processing and storage are big issues

Worldwide LHC Computing Grid (wLCG)



WLCG is a worldwide collaborative effort on an unprecedented scale in terms of storage and CPU requirements, as well as the software project's size



GRID computing developed to solve problem of data storage and analysis

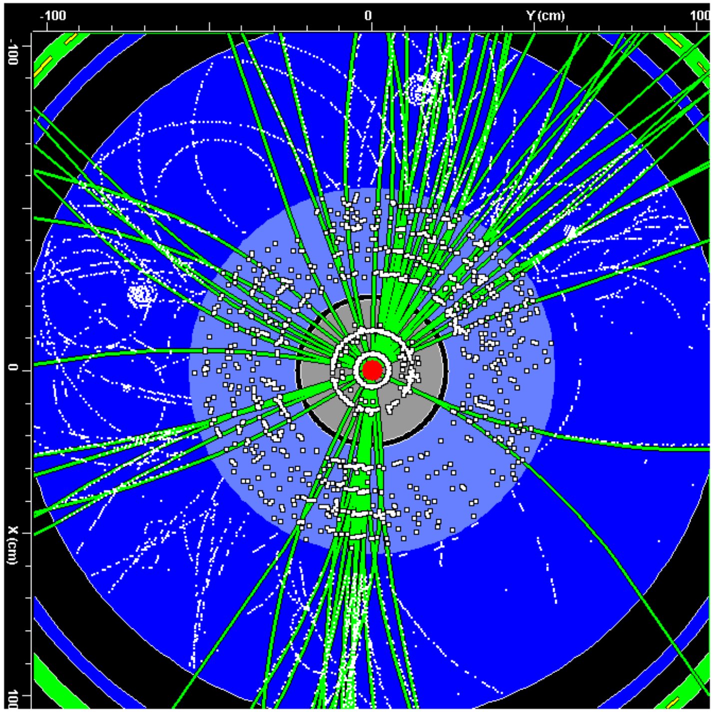
LHC data volume per year:
10-15 Petabytes

One CD has ~ 600 Megabytes
1 Petabyte = 10^9 MB = 10^{15} Byte

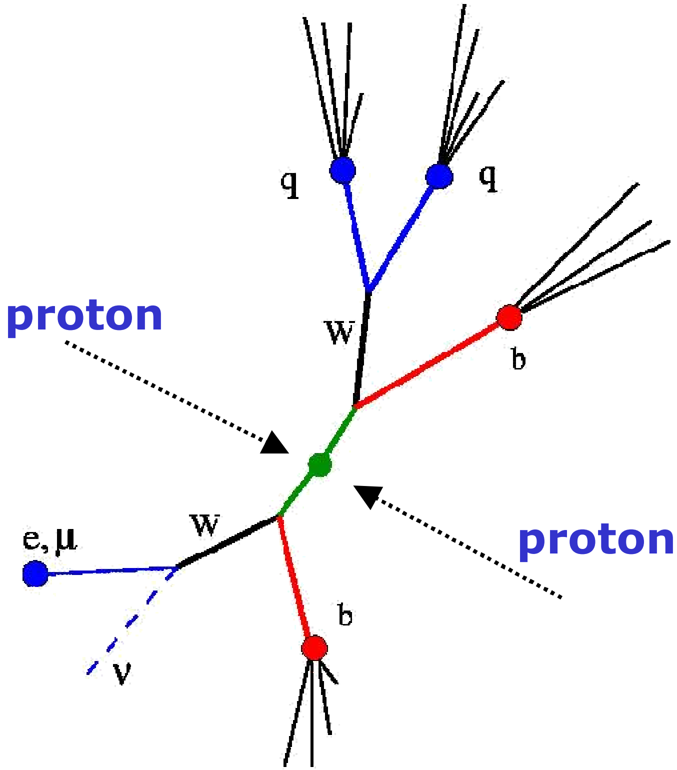
(Note: the WWW is from CERN...)

Analyzing the data – The goal

What we see in the detector

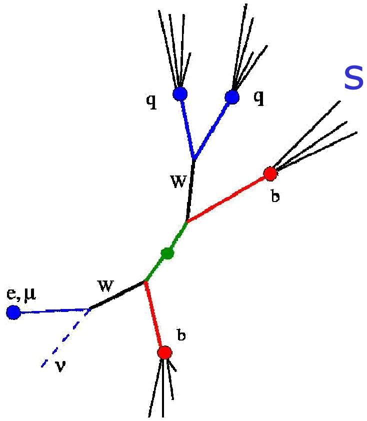


Fundamental physics picture



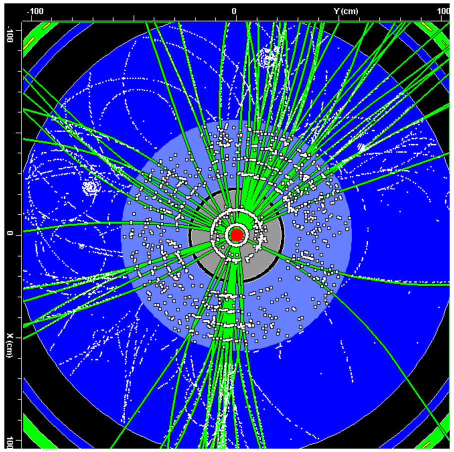
Extremely difficult
(and not possible on an event-by-event basis anyway due to QM)

Analyzing the data – in practice

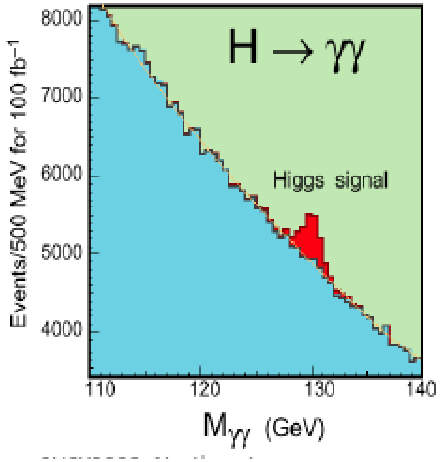


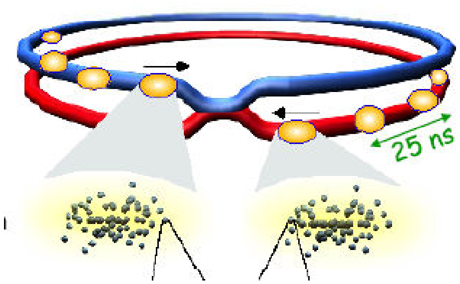
Physics simulation

detector simulation



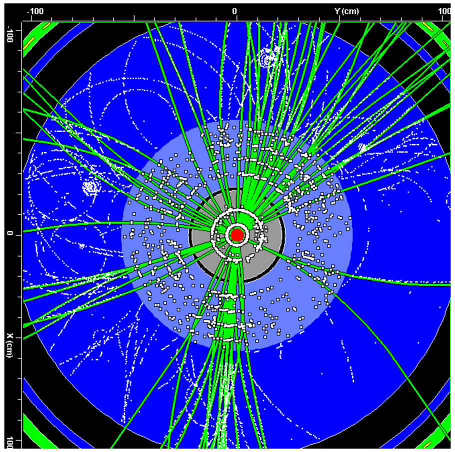
event reconstruction
data analysis





The LHC

ATLAS detector



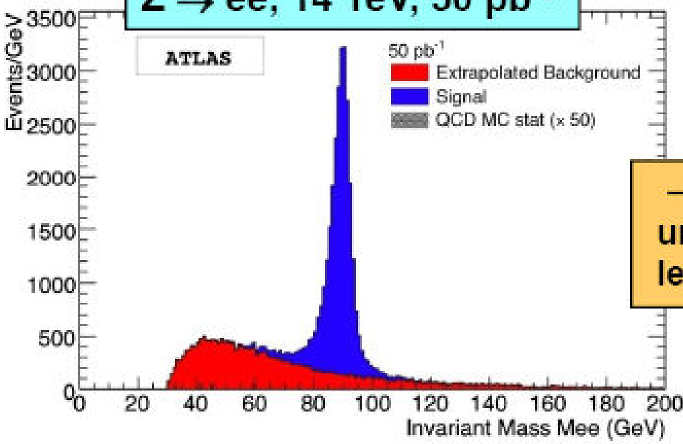
?

Wouter Verkerke, NIKHEF

'Easy stuff'

ATLAS and CMS early "signals": J/ψ, W, Z, top, the so-called "candles"

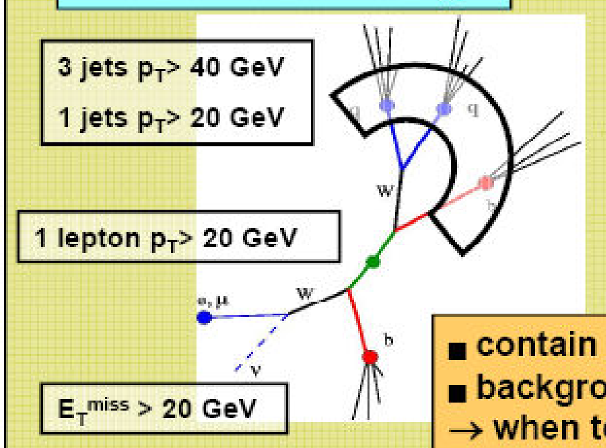
$Z \to ee$, 14 TeV, 50 pb^{-1}



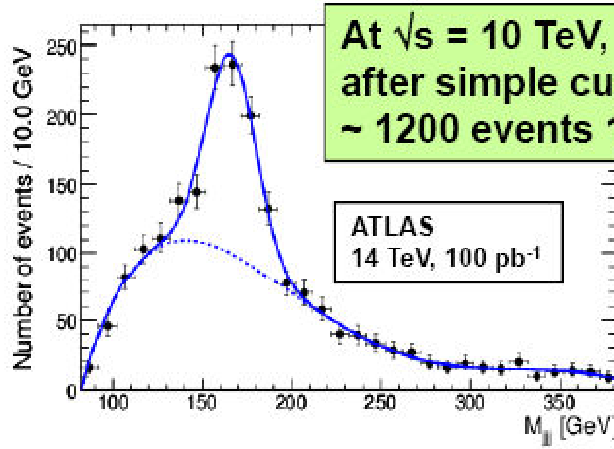
$\sqrt{s}$ = 10 TeV, after cuts:
~ 200 $Z \to ee$, $\mu\mu$ per day at L = 10^{31}
~ 40000 events 50 pb^{-1}

→ Muon Spectrometer alignment, EM calo uniformity, energy/momentum scale of full detector, lepton trigger and reconstruction efficiency, …

$tt \to bW\,bW \to bl\nu\,bjj$



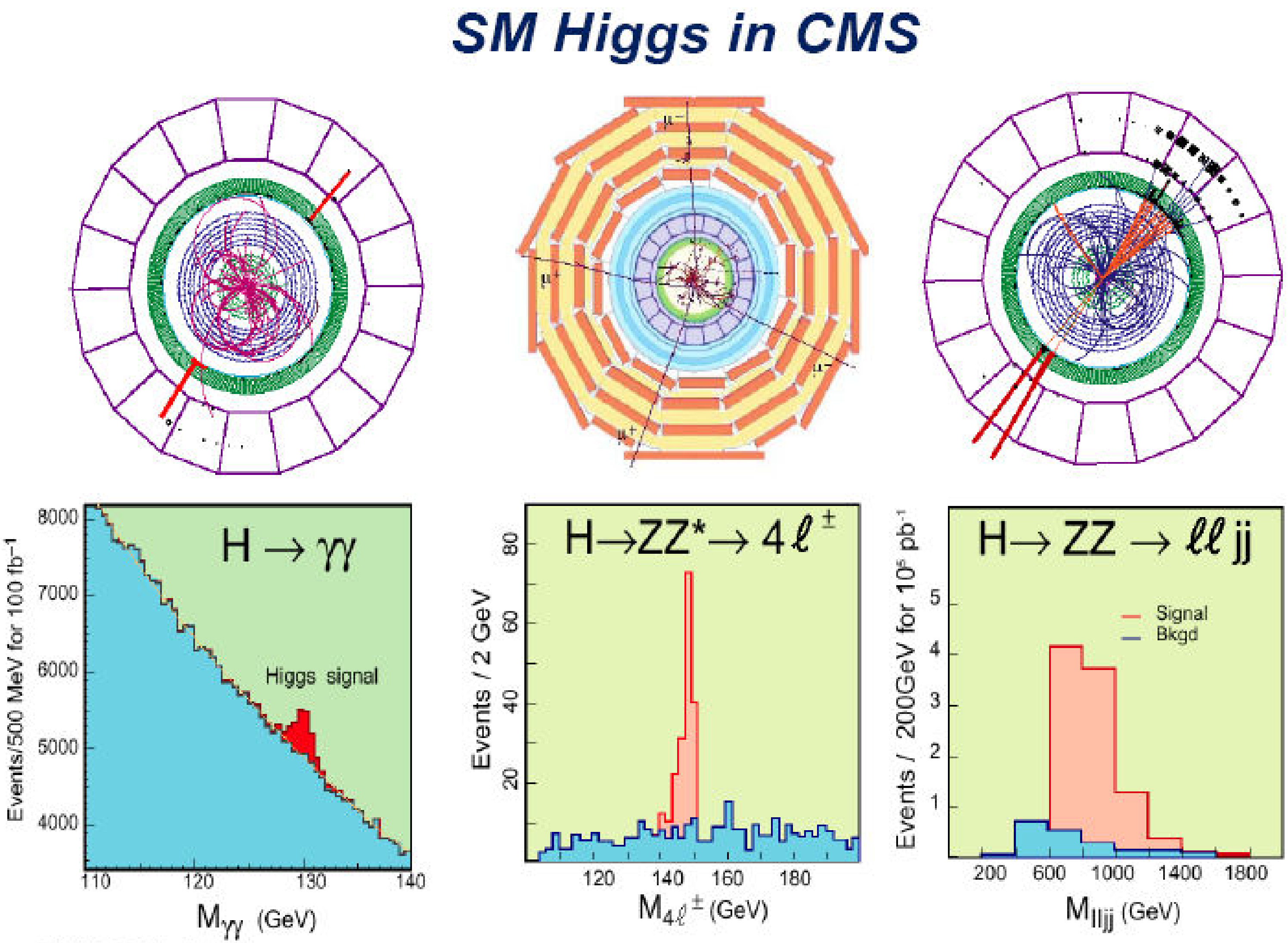
At $\sqrt{s}$ = 10 TeV, after simple cuts:
~ 1200 events 100 pb^{-1}



- **contain most physics objects: leptons, jets, E_T^{miss}, b-jets**
- **background to ~ all searches**

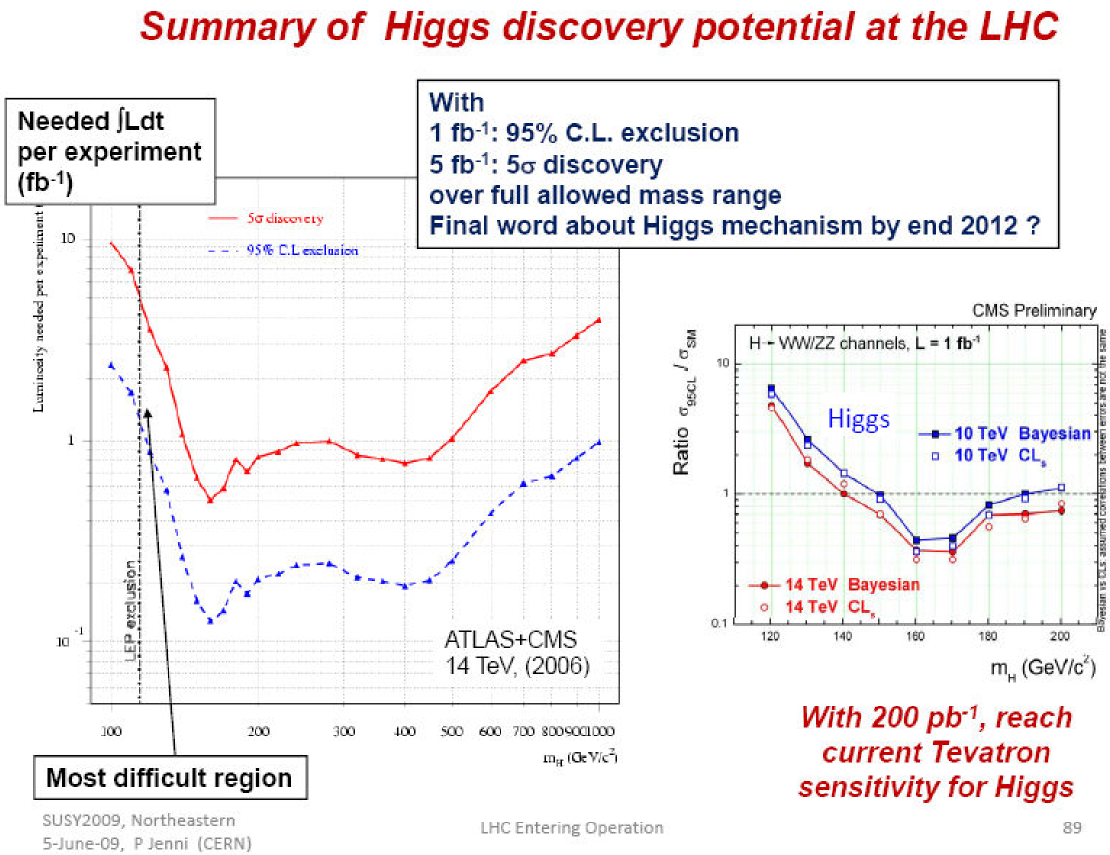
→ when top measured, experiment is ready for discovery phase

‘Difficult stuff’



What do we *expect* to see?

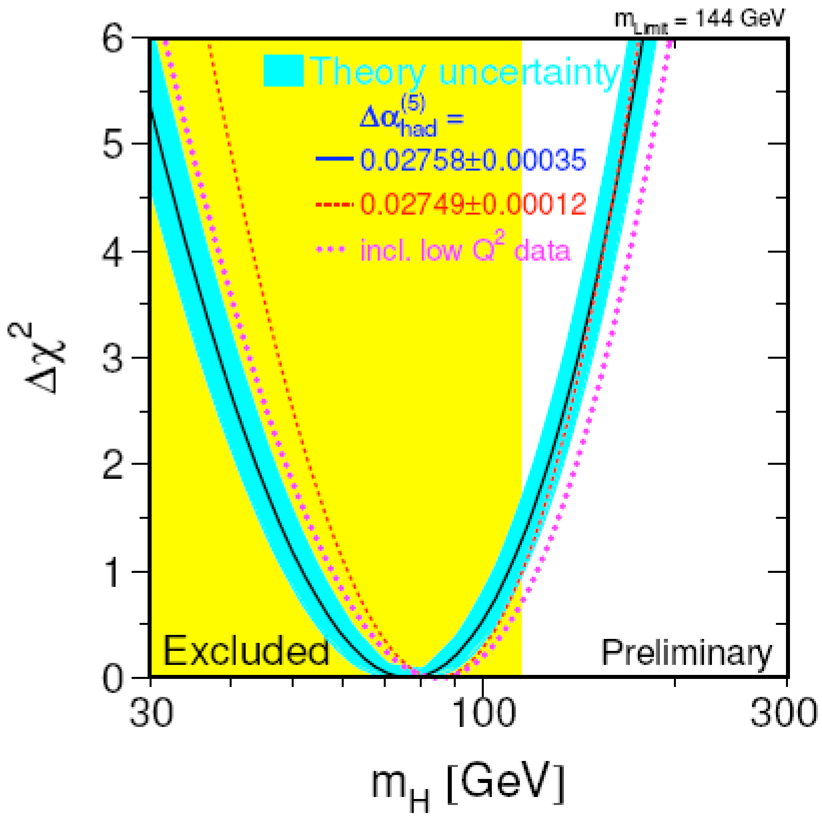
- Very active field of statistical data analysis
- Methods and details are important – for certain physics we only expect a handful of events after years of data taking

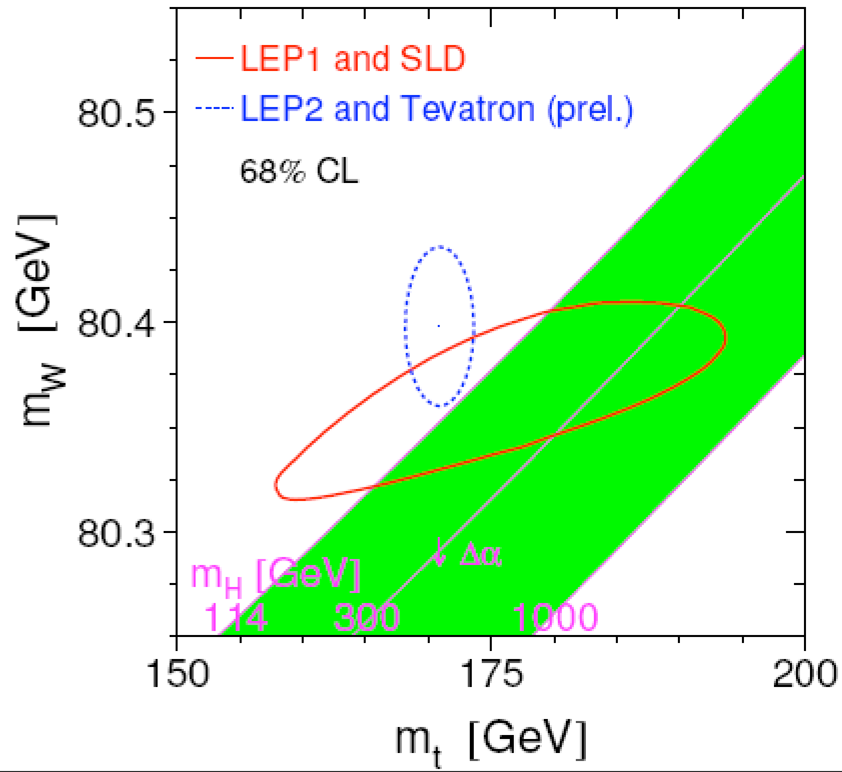


Examples of Likelihood Analysis

In these examples, a model that relates precision electroweak observables to parameters of the Standard Model was used

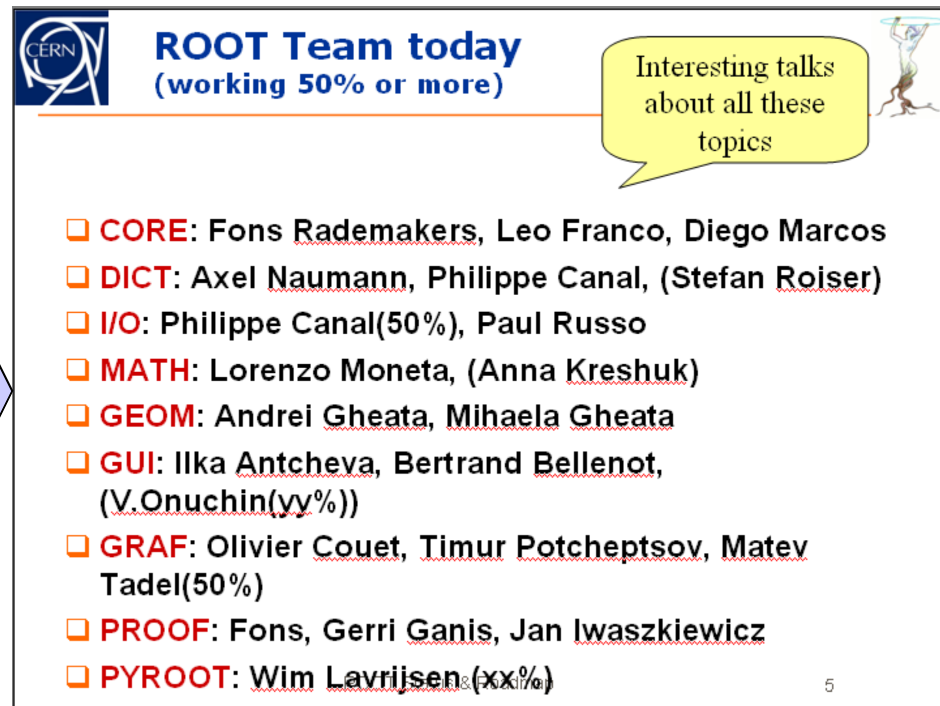
- the inference is based only on the likelihood function for data at hand
  - there is no prior, so it's not Bayesian. Not a Neyman Construction.
  - what is the meaning of this contour if it's not the Neyman Construction?





Tools for data analysis in HEP

- Nearly all HEP data analysis happens in a single platform
  - ROOT (1995-now)
  - And before that PAW (1985-1995)
- Large project with many developers, contributors, workshops



Choice of working environment R vs. ROOT

- ROOT has become *de facto* HEP standard analysis environment
  - Available and actively used for analyses in running experiments (Tevatron, B factories etc..)
  - ROOT is integrated LHC experimental software releases
  - Data format of LHC experiments is (indirectly) based on ROOT → Several experiments have/are working on summary data format directly usable in ROOT
  - Ability to handle *very* large amounts of data
- ROOT brings together a lot of the ingredients needed for (statistical) data analysis
  - C++ command line, publication quality graphics
  - Many standard mathematics, physics classes: Vectors, Matrices, Lorentz Vectors Physics constants…
- Line between 'ROOT' and 'external' software not very sharp
  - Lot of software developed elsewhere, distributed with ROOT (TMVA, RooFit)
  - Or thin interface layer provided to be able to work with external library (GSL, FFTW)
  - Still not quite as nice & automated as 'R' package concept

(Statistical) software repositories

- ROOT functions as moderated repository for statistical & data analysis tools
  - Examples TMVA, RooFit

- Several HEP repository initiatives, some contain statistical software
  - PhyStat.org (StatPatternRecognition, TMVA,LepStats4LHC)
  - HepForge (mostly physics MC generators),
  - FreeHep

- Excellent summary of non-HEP statistical repositories on Jim Linnemans statistical resources web page
  - From Phystat 2005
  - *http://www.pa.msu.edu/people/linnemann/stat_resources.html*

Roadmap for this course

- Basics of statistics
- Event classification
- Parameter estimation
- Confidence intervals, limits, significance
- Systematic uncertainties

Basic Statistics

- Mean, Variance, Standard Deviation
- Gaussian Standard Deviation
- Covariance, correlations
- Basic distributions – Binomial, Poisson, Gaussian
- Central Limit Theorem
- Error propagation

Wouter Verkerke, NIKHEF

Describing your data – the Average

- Given a set of *unbinned* data (measurements)

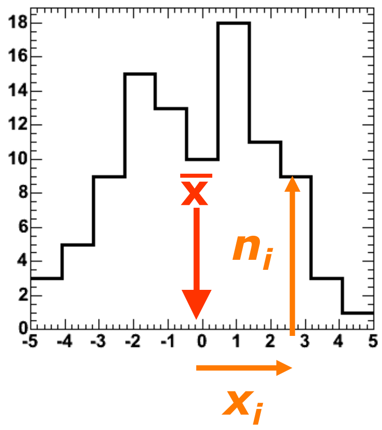
  $$\{ x_1, x_2, \ldots, x_N \}$$

  then the mean value of x is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- For *binned* data

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} n_i x_i$$



  - where n_i is bin count and x_i is bin center
  - Unbinned average more accurate due to rounding

Describing your data – Spread

- *Variance V(x)* of x expresses how much ***x*** is liable to vary from its mean value $\bar{x}$

$$
\begin{aligned}
V(x) &= \frac{1}{N}\sum_i (x_i - \bar{x})^2 \\
&= \frac{1}{N}\sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{N}\sum_i x_i^2 - \frac{1}{N}2\bar{x}\sum_i x_i + \frac{1}{N}\bar{x}^2\sum_i 1) \\
&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \overline{x^2} - \bar{x}^2
\end{aligned}
$$

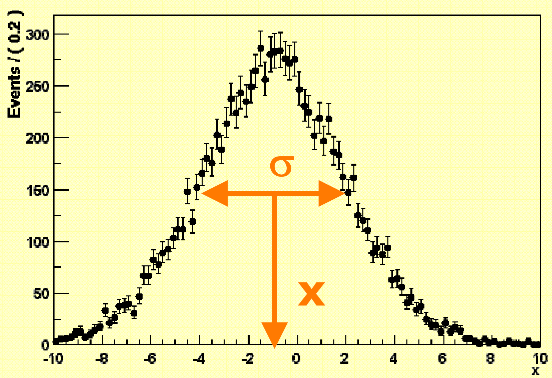
- ***Standard deviation*** $\sigma \equiv \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$

Different definitions of the Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_i (x^2 - \bar{x})^2}$$ is the S.D. of the ***data sample***

- Presumably our data was taken from a parent distributions which has mean μ and S.F. σ

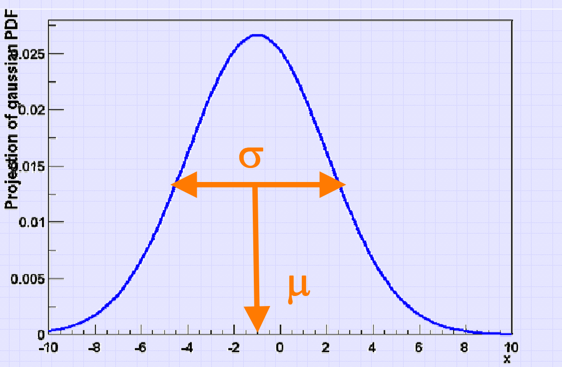
Data Sample



$\bar{x}$ – mean of our sample

σ – S.D. of our sample

Parent Distribution
(from which data sample was drawn)



μ – mean of our parent dist

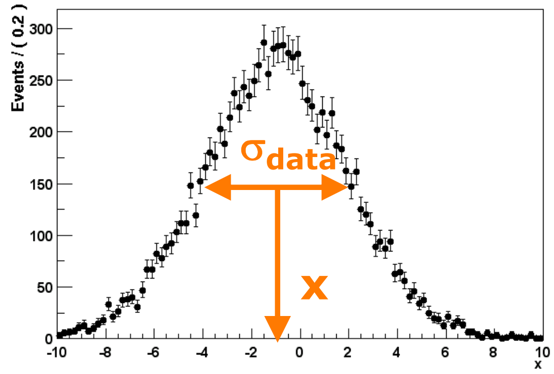
σ – S.D. of our parent dist

Beware Notational Confusion!

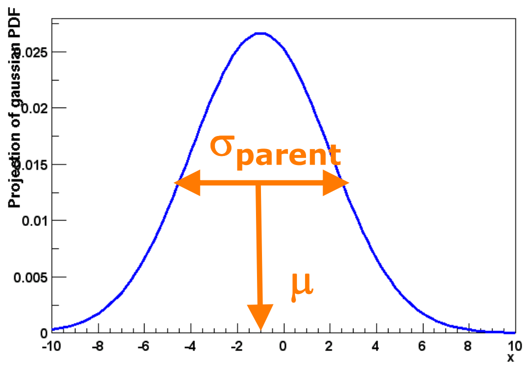
Different definitions of the Standard Deviation

- Which definition of σ you use, σ_{data} or σ_{parent}, is matter of preference, but be clear which one you mean!

Data Sample



Parent Distribution (from which data sample was drawn)



- In addition, you can get an unbiased estimate of σ_{parent} from a given data sample using

$$\hat{\sigma}_{\text{parent}} = \sqrt{\frac{1}{N-1}\sum_i (x^2 - \bar{x})^2} = \hat{\sigma}_{\text{data}}\sqrt{\frac{N}{N-1}}$$

$$\left(\sigma_{\text{data}} = \sqrt{\frac{1}{N}\sum_i (x^2 - \bar{x})^2}\right)$$

Wouter Verkerke, NIKHEF

More than one variable

- Given *2 variables x,y* and a dataset consisting of pairs of numbers

$$\{ (x_1,y_1),\ (x_2,y_2),\ \ldots\ (x_N,y_N) \}$$

- Definition of $\overline{x}$, $\overline{y}$, σ_x, σ_y as usual
- In addition, any *dependence between x,y* described by the ***co***variance

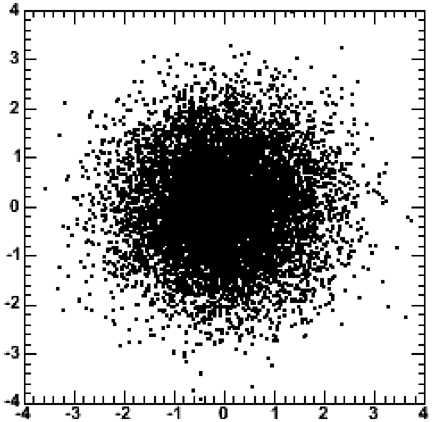
$$\begin{aligned}\mathrm{cov}(x,y) &= \frac{1}{N}\sum_i (x_i-\overline{x})(y_i-\overline{y}) \\ &= \overline{(x-\overline{x})(y-\overline{y})} \\ &= \overline{xy}-\overline{x}\,\overline{y}\end{aligned}$$

(has dimension D(x)D(y))

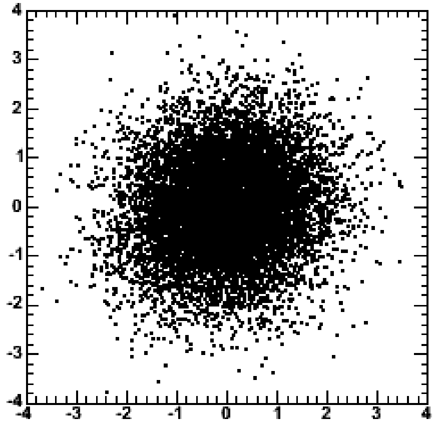
- The dimensionless ***correlation*** coefficient is defined as $\rho = \dfrac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \in [-1,+1]$

Visualization of correlation

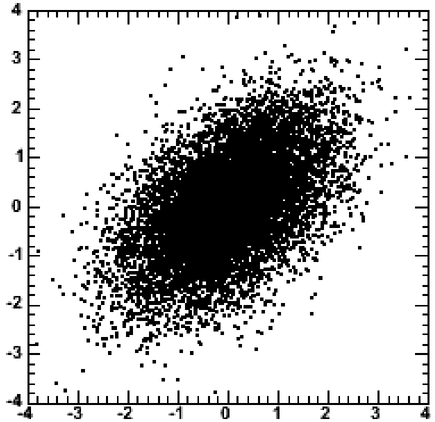
ρ = 0



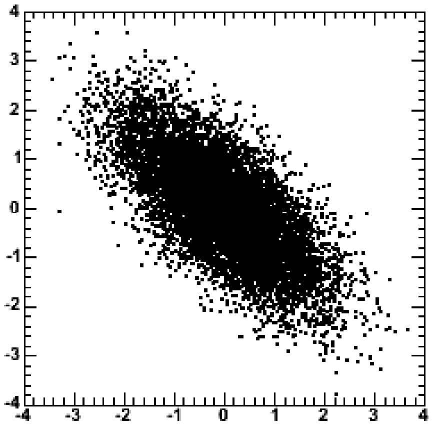
ρ = 0.1



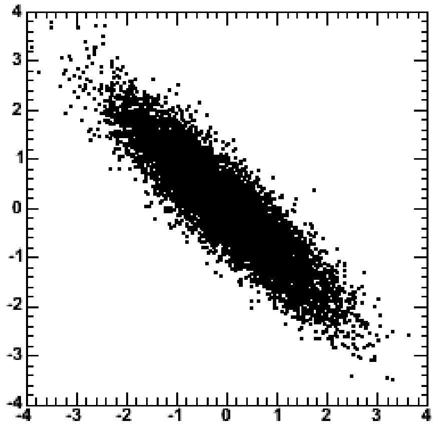
ρ = 0.5



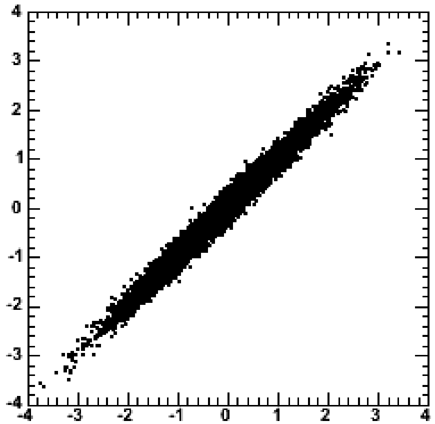
ρ = -0.7



ρ = -0.9



ρ = 0.99



Correlation & covariance in >2 variables

- Concept of covariance, correlation is easily extended to arbitrary number of variables

$$\mathrm{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)} x_{(j)}} - \bar{x}_{(i)} \bar{x}_{(j)}$$

- so that $V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)})$ takes the form of a *n x n symmetric matrix*

- This is called the ***covariance matrix***, or ***error matrix***
- Similarly the correlation matrix becomes

$$\rho_{ij} = \frac{\mathrm{cov}(x_{(i)}, x_{(j)})}{\sigma_{(i)} \sigma_{(j)}} \longrightarrow V_{ij} = \rho_{ij} \sigma_i \sigma_j$$

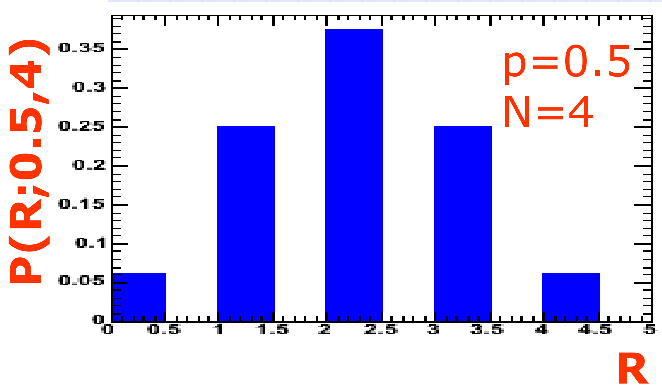
Basic Distributions – The binomial distribution

- Simple experiment – Drawing marbles from a bowl
  - Bowl with marbles, fraction **p** are black, others are white
  - Draw **N** marbles from bowl, *put marble back after each drawing*
  - Distribution of **R** black marbles in drawn sample:

Probability of a specific outcome e.g. 'BBBWBWW'

Number of equivalent permutations for that outcome

$$P(R;p,N) = p^R (1-p)^{N-R} \frac{N!}{R!(N-R)!}$$

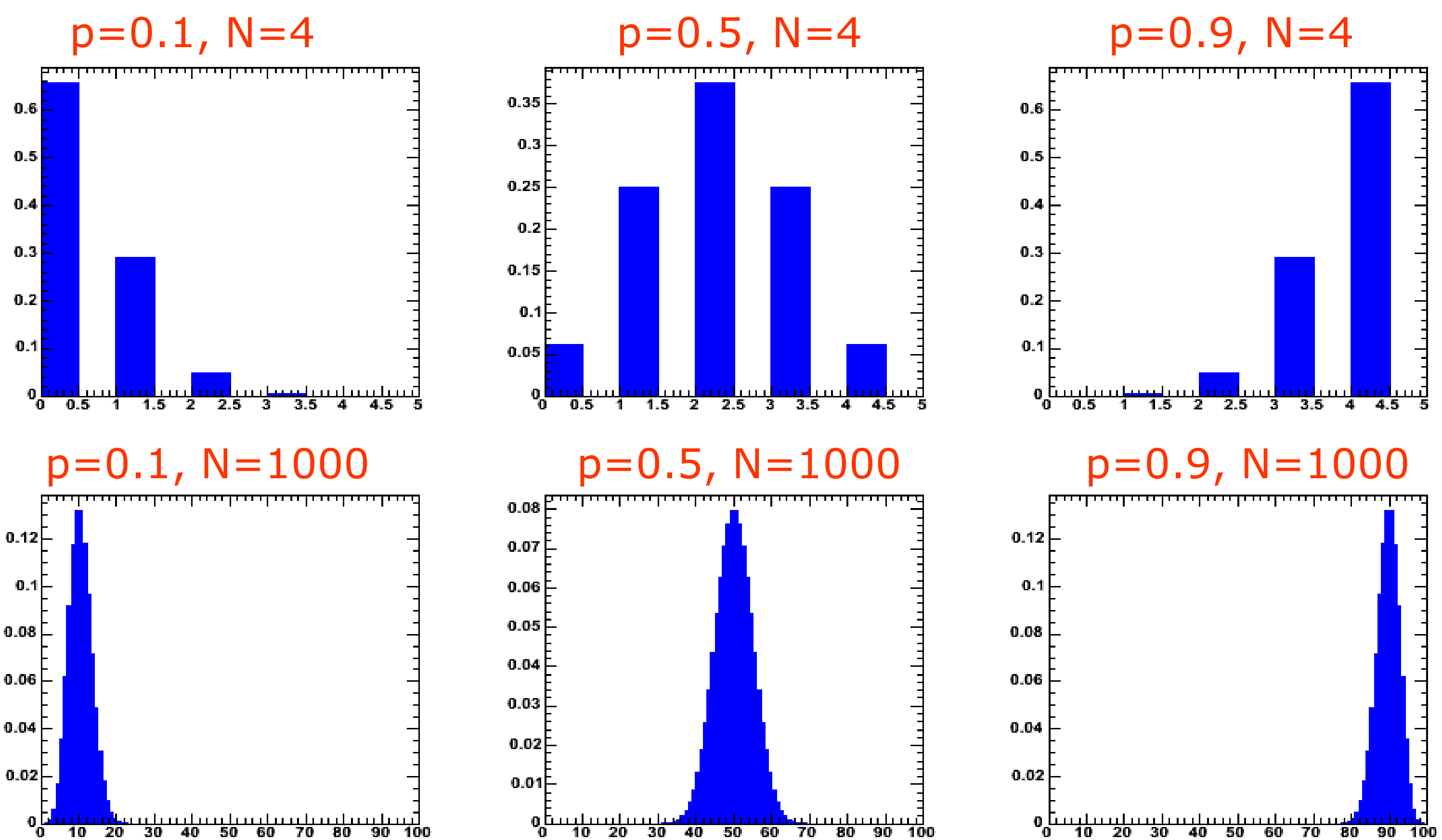


Binomial distribution

Properties of the binomial distribution

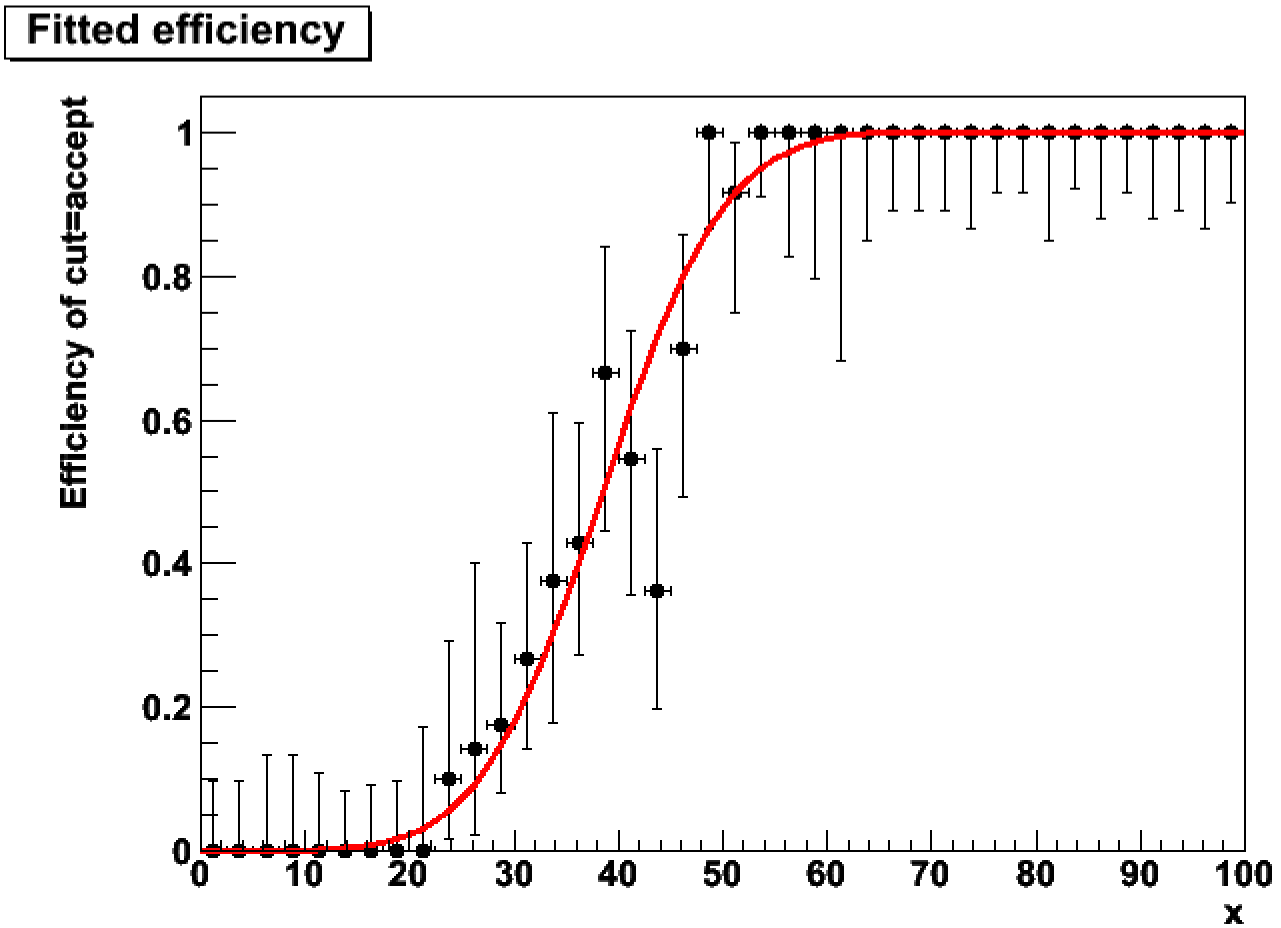
- Mean: $$\langle r \rangle = n \cdot p$$

- Variance: $$V(r) = np(1-p) \quad \Rightarrow \quad \sigma = \sqrt{np(1-p)}$$



HEP example – Efficiency measurement

- Example: trigger efficiency
  - Usually done on simulated data so that also untriggered events are available



Basic Distributions – the Poisson distribution

- Sometimes we don't know the equivalent of the number of drawings
  - **Example: Geiger counter**
  - Sharp events occurring in a (time) continuum

- What distribution to we expect in measurement over fixed amount of time?
  - Divide time interval λ in n finite chunks,
  - Take binomial formula with p=λ/n and let n→∞

$$P(r;\lambda/n,n)=\frac{\lambda^r}{n^r}(1-\frac{\lambda}{n})^{n-r}\frac{n!}{r!(n-r)!}$$

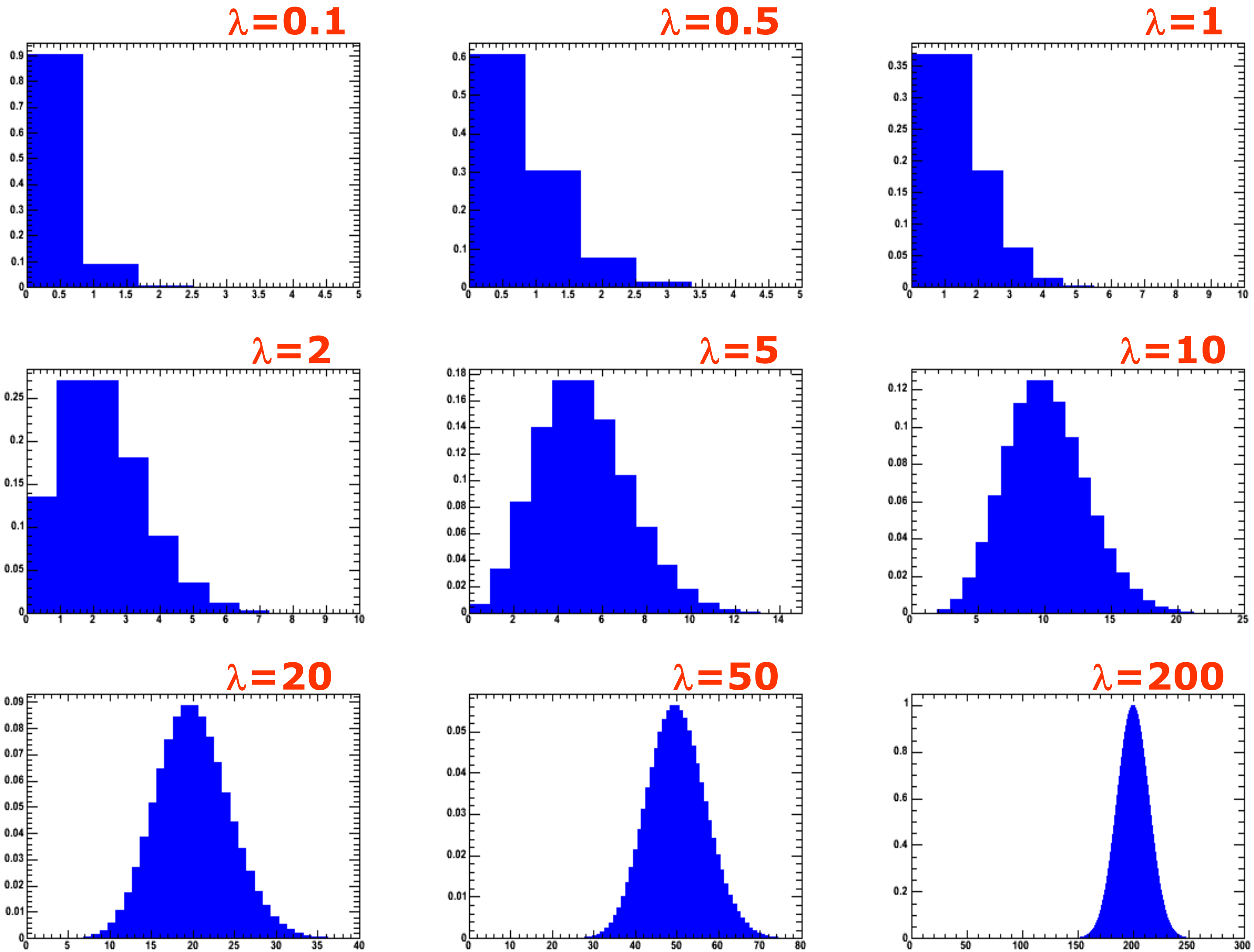
$$\lim_{n\to\infty}\frac{n!}{r!(n-r)!}=n^r,$$

$$\lim_{n\to\infty}(1-\frac{\lambda}{n})^{n-r}=e^{-\lambda}$$

$$P(r;\lambda)=\frac{e^{-\lambda}\lambda^r}{r!}$$

←Poisson distribution

Properties of the Poisson distribution



Wouter Verkerke, UCSB

More properties of the Poisson distribution $P(r;\lambda)=\frac{e^{-\lambda}\lambda^r}{r!}$

- Mean, variance: $\langle r \rangle = \lambda$

$$V(r)=\lambda \quad \Rightarrow \quad \sigma=\sqrt{\lambda}$$

- Convolution of 2 Poisson distributions is also a Poisson distribution with $\lambda_{ab}=\lambda_a+\lambda_b$

$$P(r)=\sum_{r_A=0}^{r} P(r_A;\lambda_A)P(r-r_A;\lambda_B)$$

$$=e^{-\lambda_A}e^{-\lambda_B}\sum\frac{\lambda_A^{r_A}\lambda_B^{r-r_A}}{r_A!(r-r_A)!}$$

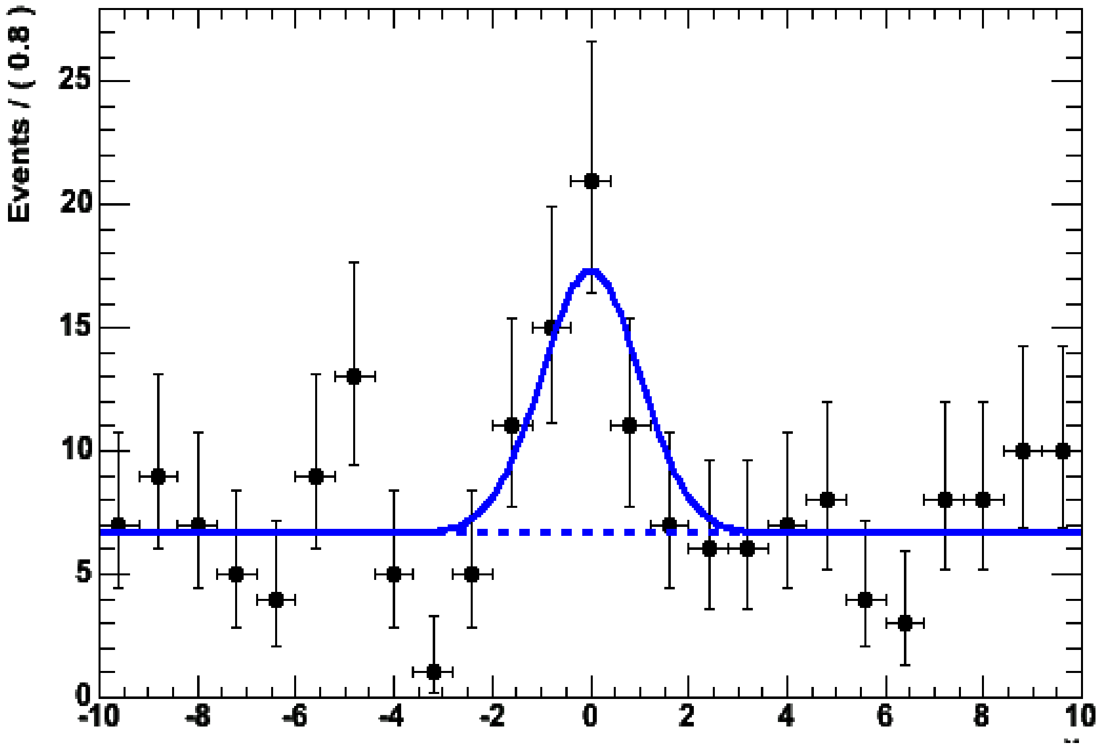
$$=e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\sum_{r_{A=0}}^{r}\frac{r!}{(r-r_A)!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^{r_A}\left(\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^{r-r_A}$$

$$=e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}+\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^r$$

$$=e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}$$

HEP example – counting experiment

- Any distribution plotted on data (particularly in case of low statistics



Basic Distributions – The Gaussian distribution

- Look at *Poisson distribution* in limit of *large N*

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Take log, substitute, r = λ + x,
and use $\ln(r!) \approx r\ln r - r + \ln\sqrt{2\pi r}$

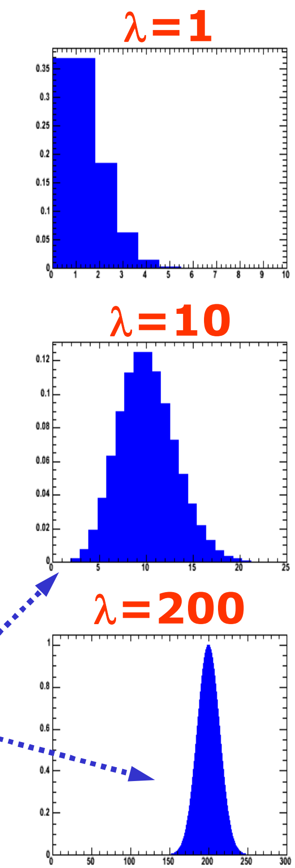
$$\begin{aligned}
\ln(P(r;\lambda)) &= -\lambda + r\ln\lambda - (r\ln r - r) - \ln\sqrt{2\pi r} \\
&= -\lambda + r\left[\ln\lambda - \ln(\lambda(1+\frac{x}{\lambda}))\right] + (\lambda + x) - \ln\sqrt{2\pi\lambda} \\
&\approx x - (\lambda - x)\left(\frac{x}{\lambda} + \frac{x^2}{2\lambda^2}\right) - \ln(2\pi\lambda) \\
&\approx \frac{-x^2}{2\lambda} - \ln(2\pi\lambda)
\end{aligned}$$

$\ln(1+z) \approx z - z^2/2$

Take exp

$$P(x) = \frac{e^{-x^2/2\lambda}}{\sqrt{2\pi\lambda}}$$

Familiar Gaussian distribution,
(approximation reasonable for N>10)



Properties of the Gaussian distribution

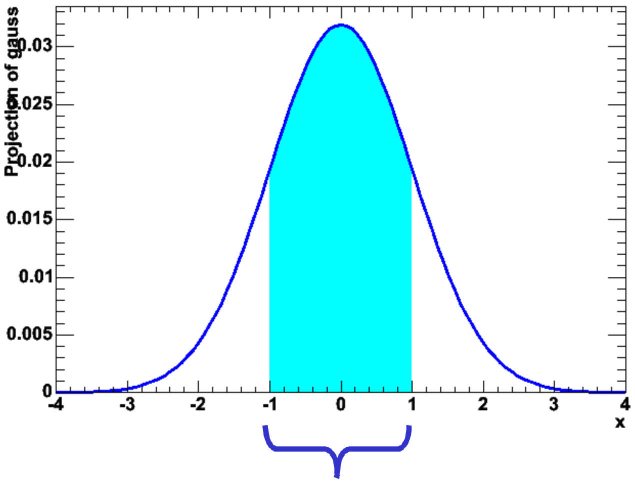
$$P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

- *Mean* and *Variance*

$$\langle x \rangle = \int_{-\infty}^{+\infty} x P(x;\mu,\sigma)dx = \mu$$

$$V(x) = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x;\mu,\sigma)dx = \sigma^2$$

$$\sigma = \sigma$$

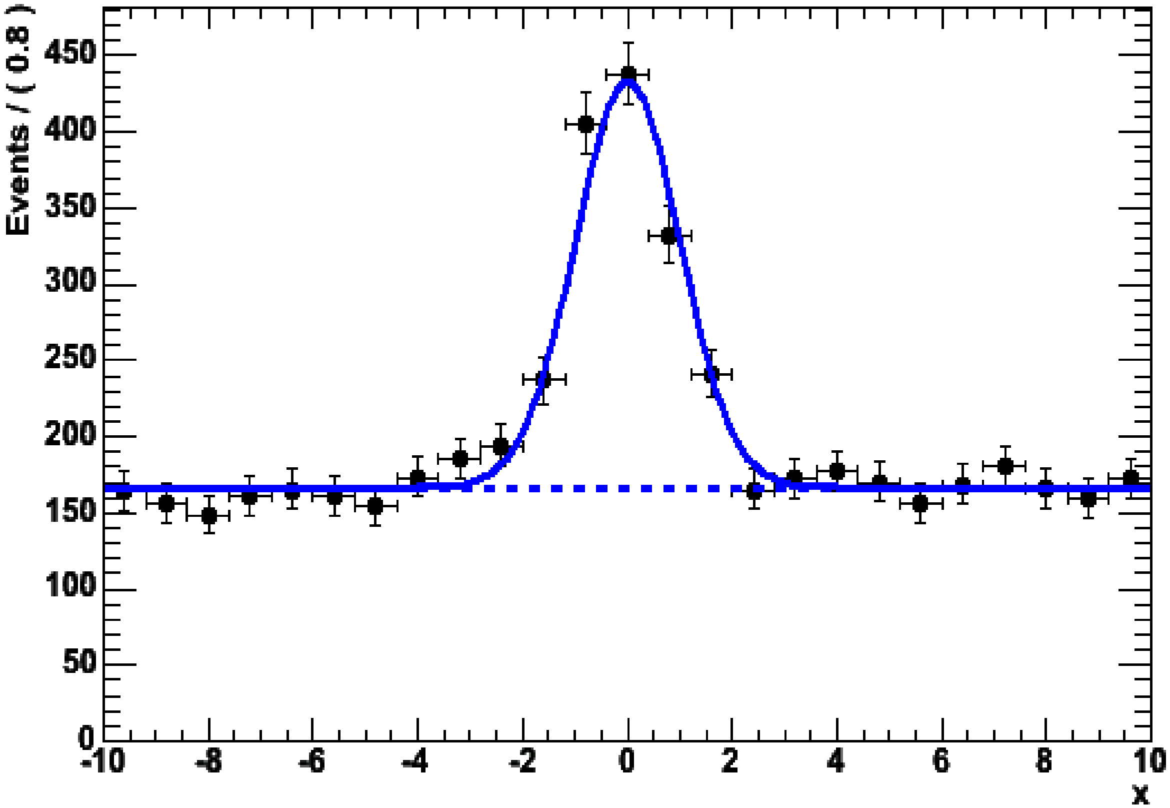


- Integrals of Gaussian

68.27% within 1σ	90% → 1.645σ
95.43% within 2σ	95% → 1.96σ
99.73% within 3σ	99% → 2.58σ
	99.9% → 3.29σ

HEP example – high statistics counting expt

- High statistics distributions from data



Errors

- Doing an experiment → making measurements

- Measurements not perfect → imperfection quantified in resolution or error

- Common language to quote errors
  - Gaussian standard deviation = sqrt(V(x))
  - 68% probability that true values is within quoted errors

    [NB: 68% interpretation relies strictly on Gaussian sampling distribution, which is not always the case, more on this later]

- Errors are usually Gaussian if they quantify a result that is based on many independent measurements

The Gaussian as 'Normal distribution'

- Why are errors usually Gaussian?

- The ***Central Limit Theorem*** says
  - If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

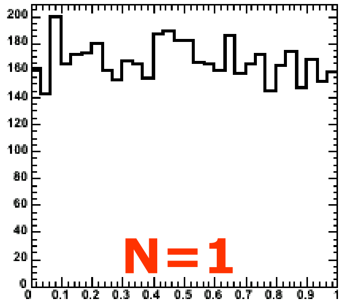
    (a) has expectation value $$\langle X \rangle = \sum_i \mu_i$$

    (b) has variance $$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$

    (c) becomes Gaussian as N → ∞

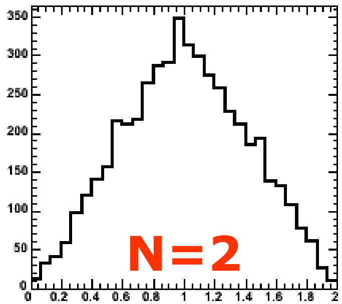
  - *Small print: tails converge very slowly in CLT, be careful in assuming Gaussian shape beyond 2σ*

Demonstration of Central Limit Theorem



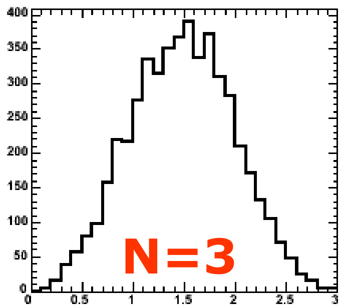
← 5000 numbers taken at random from a uniform distribution between [0,1].

- Mean = $^1/_2$, Variance = $^1/_{12}$

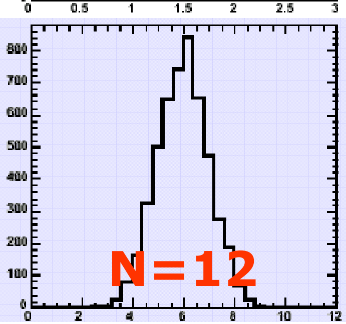


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1+x_2$.

- Triangular shape



← Same for 3 numbers, $X = x_1 + x_2 + x_3$



← Same for 12 numbers, overlaid curve is exact Gaussian distribution

Central Limit Theorem – repeated measurements

- Common case 1 : Repeated identical measurements
  i.e. $\mu_i = \mu$, $\sigma_i = \sigma$ for all i

C.L.T

$$\langle X \rangle = \sum_i \mu_i = N\mu \quad \Rightarrow \quad \langle \bar{x} \rangle = \frac{X}{N} = \mu$$

$$V(\bar{x}) = \sum_i V_i(\bar{x}) = \frac{1}{N^2} \sum_i V_i(X) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$

← **Famous sqrt(N) law**

Central Limit Theorem – repeated measurements

- Common case 2 : Repeated measurements with identical means but different errors (i.e weighted measurements, $\mu_i = \mu$)

$$\bar{x} = \frac{\sum x_i / \sigma_i^2}{\sum 1/\sigma_i^2}$$

Weighted average

$$V(\bar{x}) = \frac{1}{\sum 1/\sigma_i^2} \Rightarrow \sigma(\bar{x}) = \frac{1}{\sqrt{\sum 1/\sigma_i^2}}$$

'Sum-of-weights' formula for error on weighted measurements

Error propagation – one variable

- Suppose we have $f(x) = ax + b$

- How do you calculate V(f) from V(x)?

$$
\begin{aligned}
V(f) &= \langle f^2 \rangle - \langle f \rangle^2 \\
&= \langle (ax+b)^2 \rangle - \langle ax+b \rangle^2 \\
&= a^2\langle x^2 \rangle + 2ab\langle x \rangle + b^2 - a\langle x \rangle^2 - 2ab\langle x \rangle - b^2 \\
&= a^2\left(\langle x^2 \rangle - \langle x \rangle^2\right) \\
&= a^2 V(x)
\end{aligned}
$$

← **i.e. $\sigma_f = |a|\sigma_x$**

- More general: $V(f) = \left(\frac{df}{dx}\right)^2 V(x) \quad ; \quad \sigma_f = \left|\frac{df}{dx}\right| \sigma_x$
  - But only valid if *linear approximation is good in range of error*

Error Propagation – Summing 2 variables

- Consider $f = ax + by + c$

$$V(f) = a^2\left(\langle x^2\rangle - \langle x\rangle^2\right) + b^2\left(\langle y^2\rangle - \langle y\rangle^2\right) + 2ab\left(\langle xy\rangle - \langle x\rangle\langle y\rangle\right)$$

$$= a^2V(x) + b^2V(y) + 2ab\,\mathrm{cov}(x,y)$$

Familiar 'add errors in quadrature' **only valid in absence of correlations**, i.e. cov(x,y)=0

- More general

$$V(f) = \left(\frac{df}{dx}\right)^2 V(x) + \left(\frac{df}{dy}\right)^2 V(y) + 2\left(\frac{df}{dx}\right)\left(\frac{df}{dy}\right)\mathrm{cov}(x,y)$$

$$\sigma_f^2 = \left(\frac{df}{dx}\right)^2 \sigma_x^2 + \left(\frac{df}{dy}\right)^2 \sigma_y^2 + 2\left(\frac{df}{dx}\right)\left(\frac{df}{dy}\right)\rho\sigma_x\sigma_y$$

But only valid if *linear approximation is good in range of error*

The correlation coefficient ρ [-1,+1] is 0 if x,y uncorrelated

Error propagation – multiplying, dividing 2 variables

- Now consider $f = x \cdot y$

$$V(f) = y^2 V(x) + x^2 V(y)$$

(math omitted)

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

  – Result similar for $f = x / y$

- Other useful formulas

$$\frac{\sigma_{1/x}}{1/x} = \frac{\sigma_x}{x} \quad ; \quad \sigma_{\ln(x)} = \frac{\sigma_x}{x}$$

Relative error on x,1/x is the same

Error on log is just fractional error