Summary of Session 2

Christian Kiesling
Max-Planck-Institut für Physik

- A total of 27 talks were presented
- Thanks to the speakers for delivering high quality presentations
- Apologies for not being able to mention all due to time constraints

Session 2: Data Analysis – Algorithms & Tools

Analysis Frames (common Session 1 & 2)	3 talks:	René Brun, Valeri Tioukov, Jens Zimmermann
Mathematical Methods	6 talks:	Florent de Dinechin, Vladislav Matouchek, Miroslav Morhac, Wolfgang Rolke, Jens Zimmermann, Victor Zlokazov
Analysis Methods	3 talks:	Mostafa Mjahed, Are Strandlie, Andre van Hameren
Simulation	3 talks:	Frank Gaede, Daniel Pomarede, Andreas Schaelicke
Tracking	4 talks:	Sergey Gorbunov, Takanori Kohno,Thomas Speer; Are Strandlie
Reconstruction	4 talks:	Paulo Vitor Magacho da Silva, Jose Seixas, Sergey Gorbunov
Neural Networks	4 talks:	Aatos Heikkinen, Ulrich Kerzel, Mostafa Mjahed, Andrea Piccione

René Brun:

Bitmap Indices for Fast End-User Physics Analysis in ROOT

- Bitmap indices are efficient data structures for accelerating multi-dimensional queries
- Supported by most commercial database management systems and data warehouses

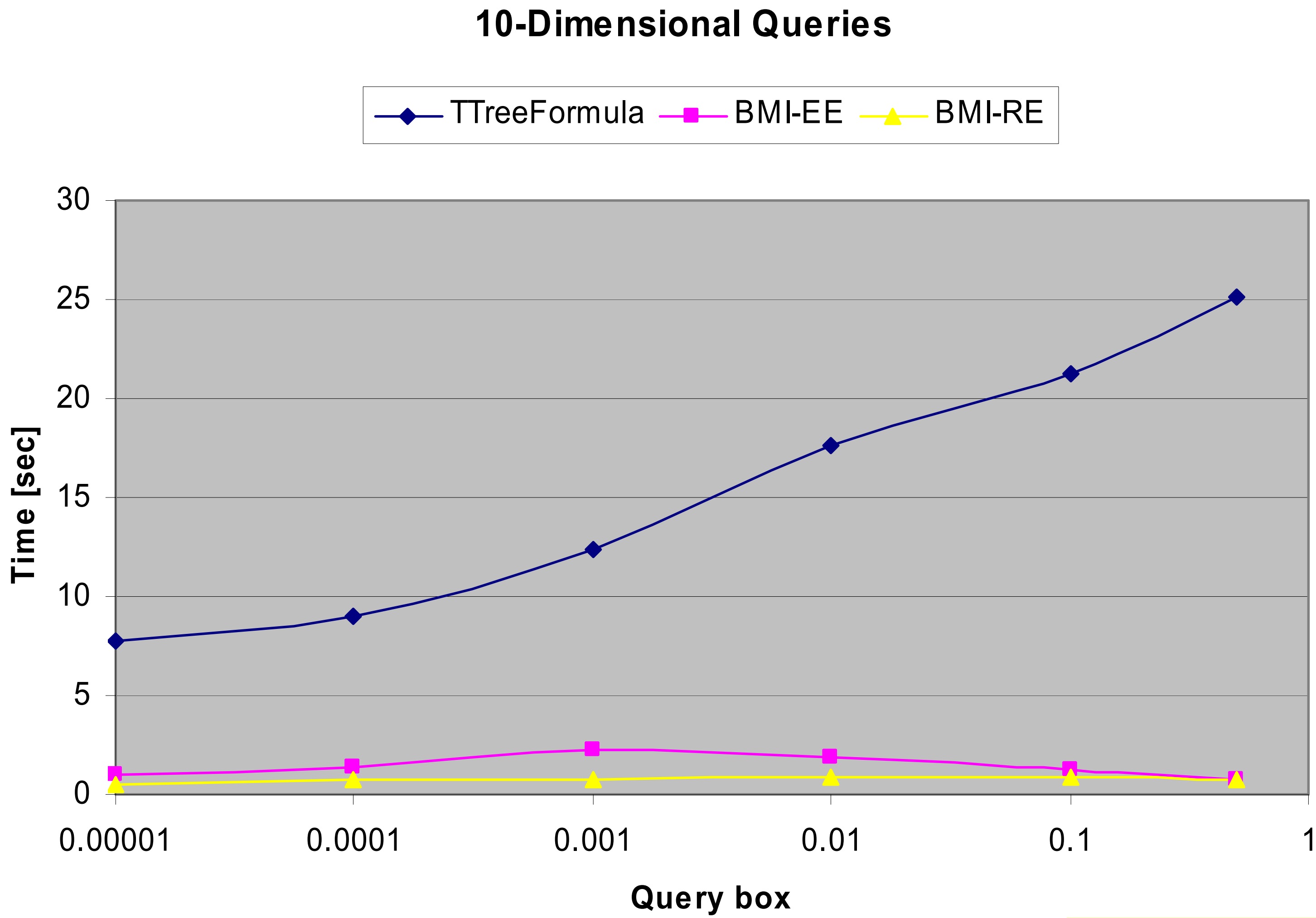
	$\pi_A(R)$	E^9	E^8	E^7	E^6	E^5	E^4	E^3	E^2	E^1	E^0	R^8	R^7	R^6	R^5	R^4	R^3	R^2	R^1	R^0
1	3	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0
2	2	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0
3	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0
4	2	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0
5	8	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
6	2	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0
7	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
9	7	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
10	5	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0
11	6	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
12	4	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0	0	0	0

(a) (b) (c)

list of attributes — equality encoding — range encoding

(here: a <= 3)

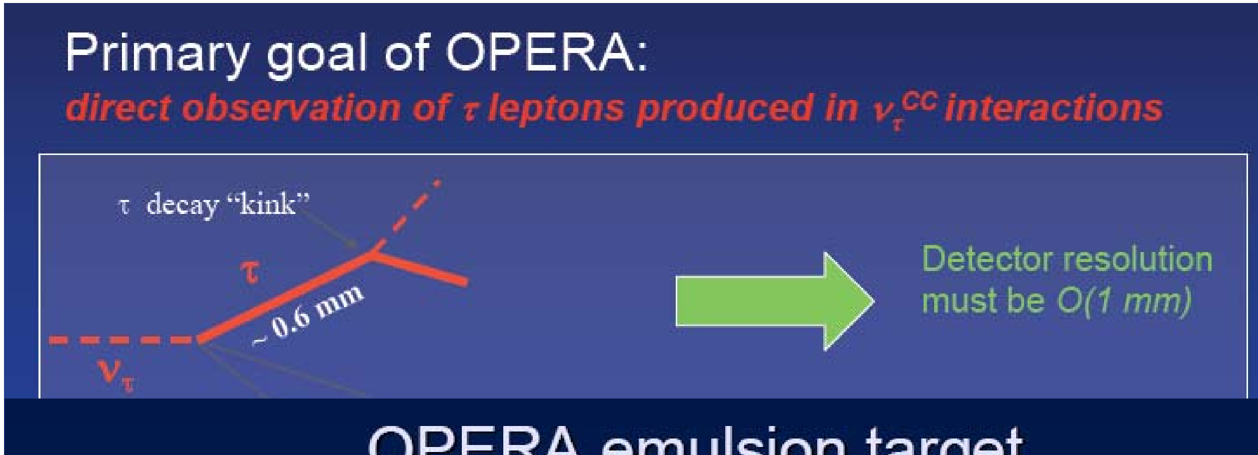
Query Performance - Approximate Answers (Error 0.1- 1%)

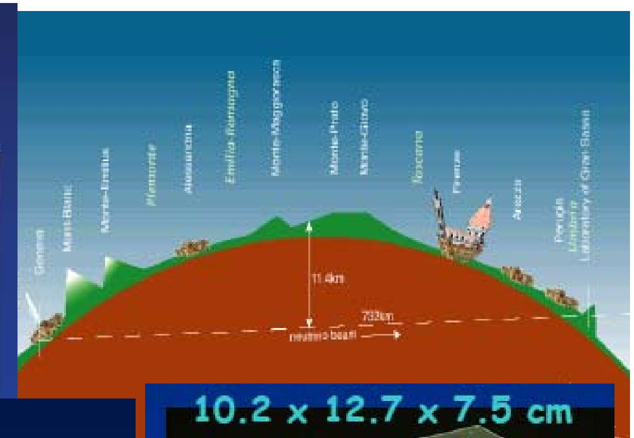


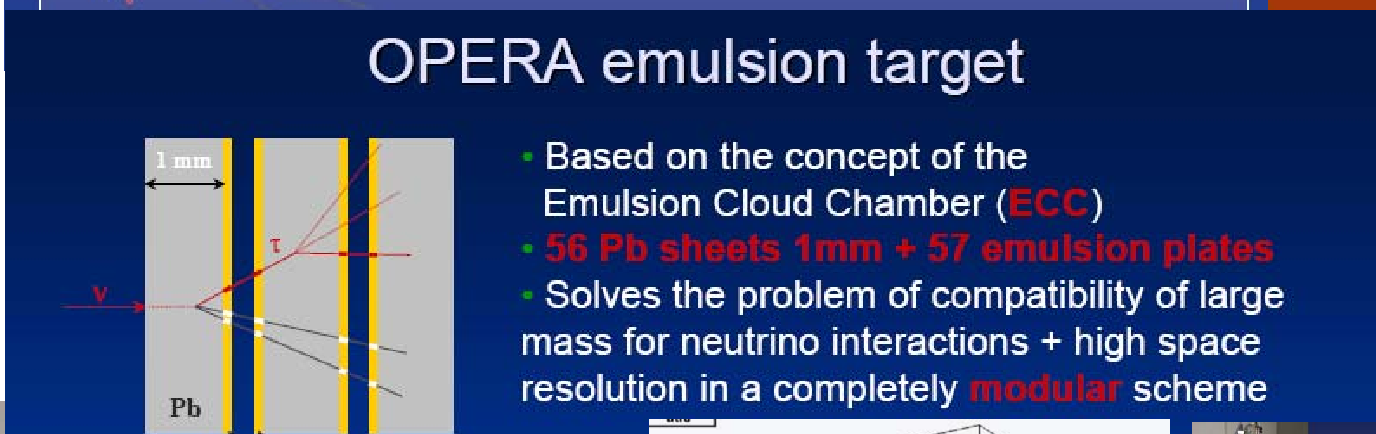
Performance improvement of bitmap indices over TTreeFormula up to a factor of 30.

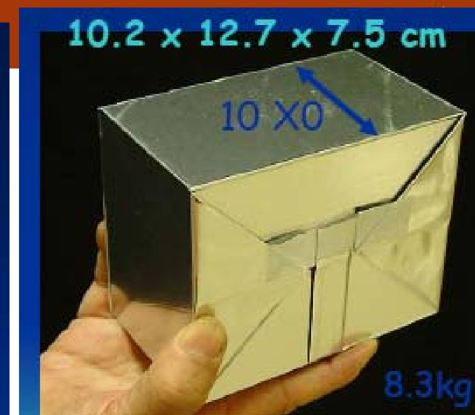
Valeri Tioukov:

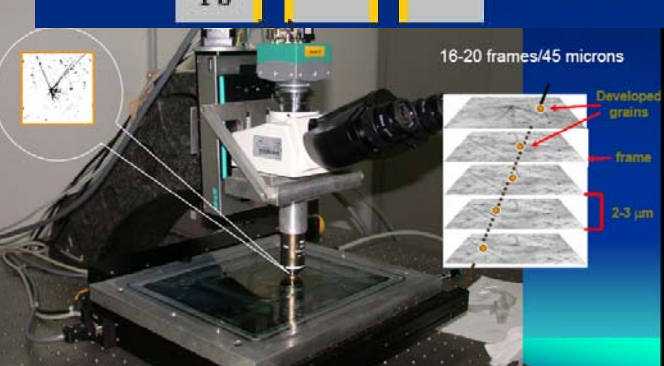
FEDRA – Framework for Emulsion Data Reconstruction and Analysis

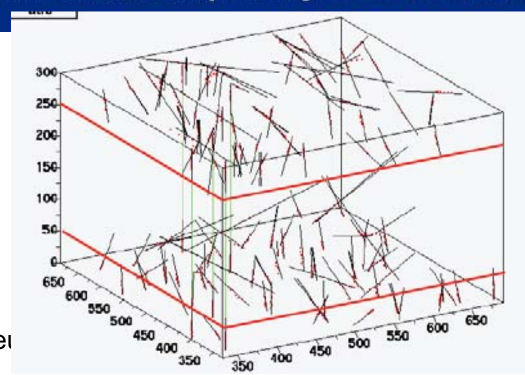


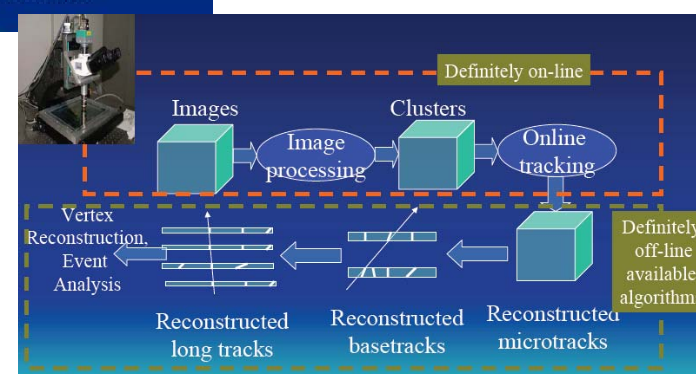








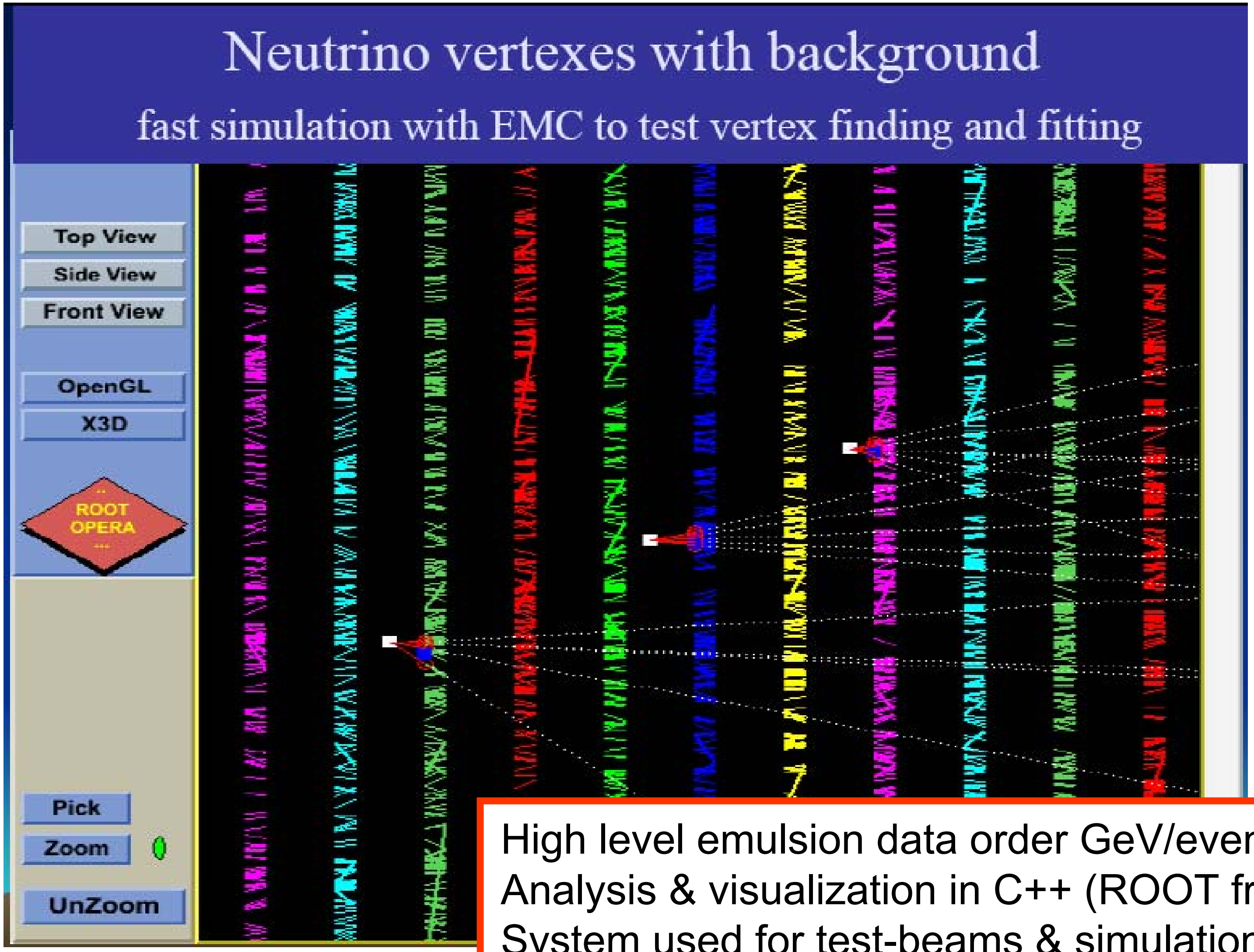




Basic FEDRA modules

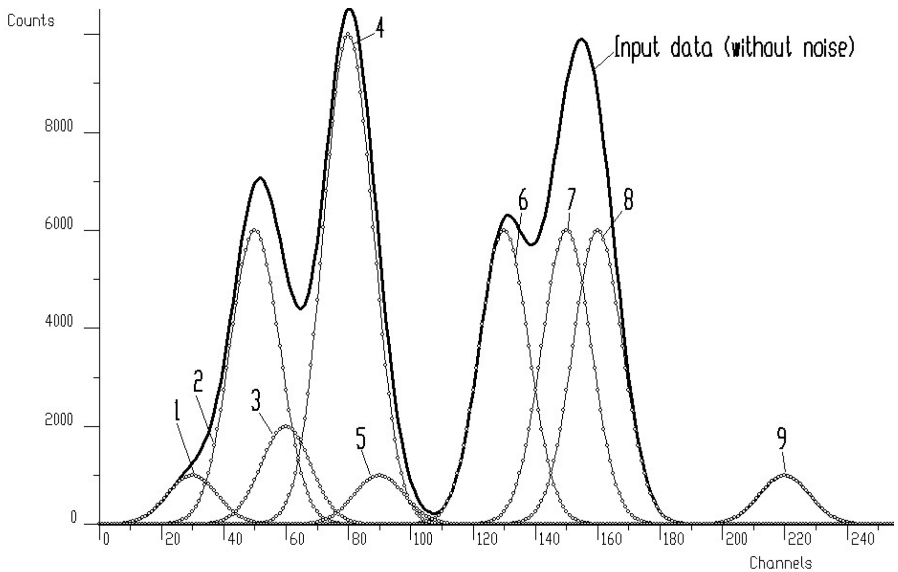
libraries of C++ classes based on ROOT structures

Valeri Tioukov



High level emulsion data order GeV/event
Analysis & visualization in C++ (ROOT framework)
System used for test-beams & simulation

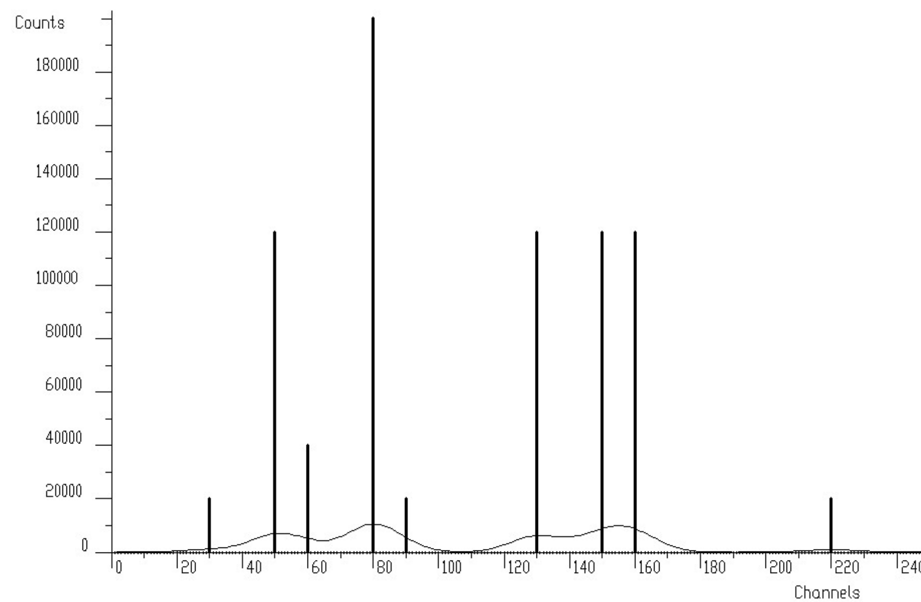
V. Matoušek and M. Morháč: Efficient Algorithms of Multidimensional γ Ray Sprectra Compression/Deconvolution



- Extraction of the correct physics from experimental spectra:
- limited resolution: overlapping signals deconvolution and decomposition
- remove smearing effects using known resolution function

Data compressed,
e.g., with Walsh transform

proposed new high-resolution peak searching algorithm based on Gold deconvolution for one- and two-dimensional spectra.



Wolfgang Rolke:
Setting Limits in the Presence of Nuisance Parameters

How do we set limits on a signal rate ?

Generally: x in signal region

$y \pm \Delta y$ in sidebands (background)

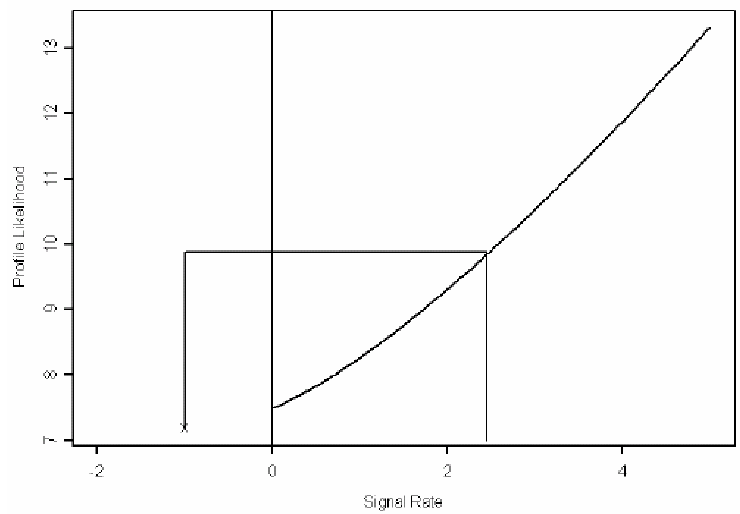
$\varepsilon \pm \Delta\varepsilon$ (signal efficiency)

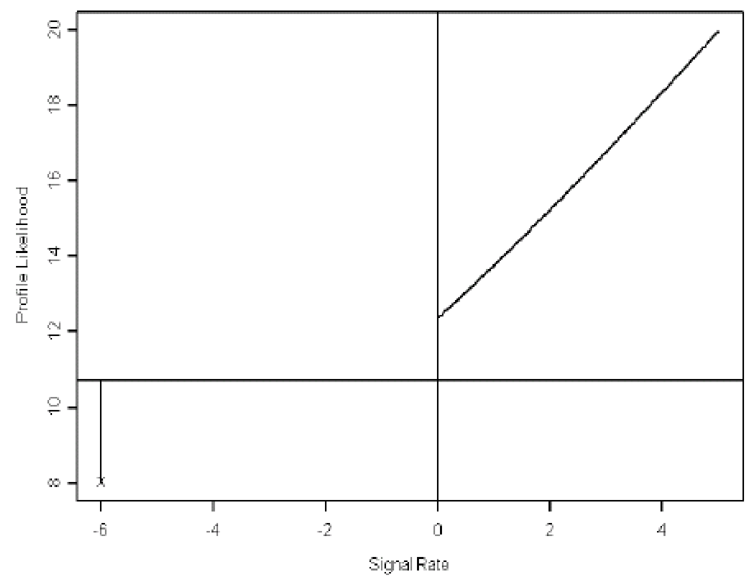
Problem already solved in MINUIT

BUT:
What if observed rate is below expected background ?

→ Profile Likelihood

Method implemented as ROOT class



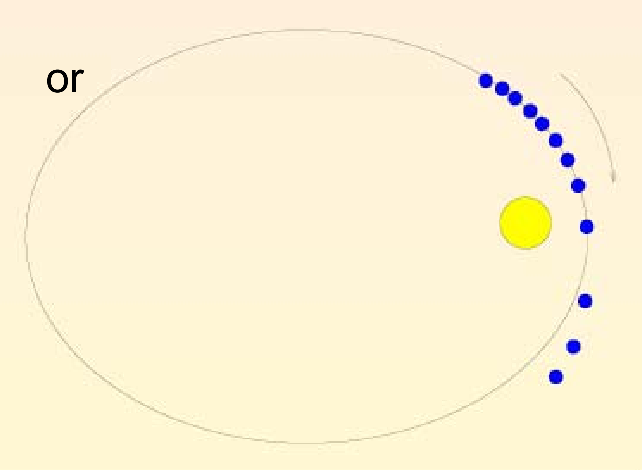


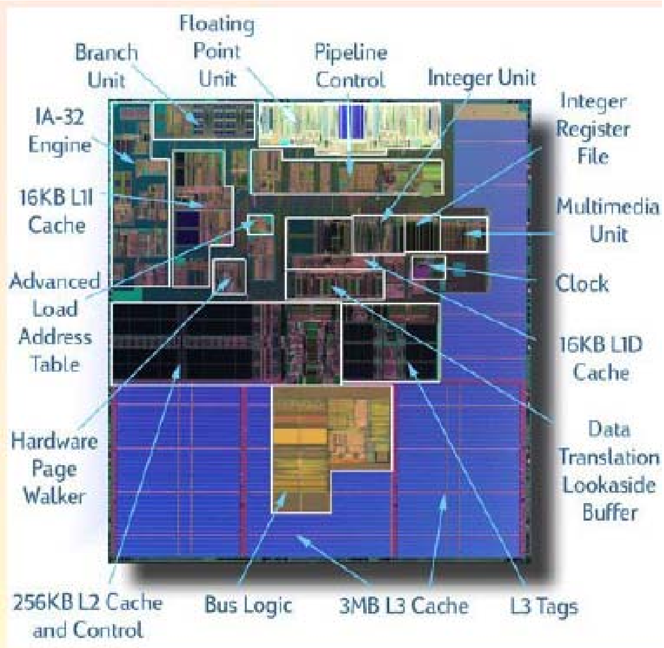
Florent de Dinechin:
High Precision numerical Accuracy in Physics Research

Double Precision not enough?

Add more than 10^{15} numbers ….

or





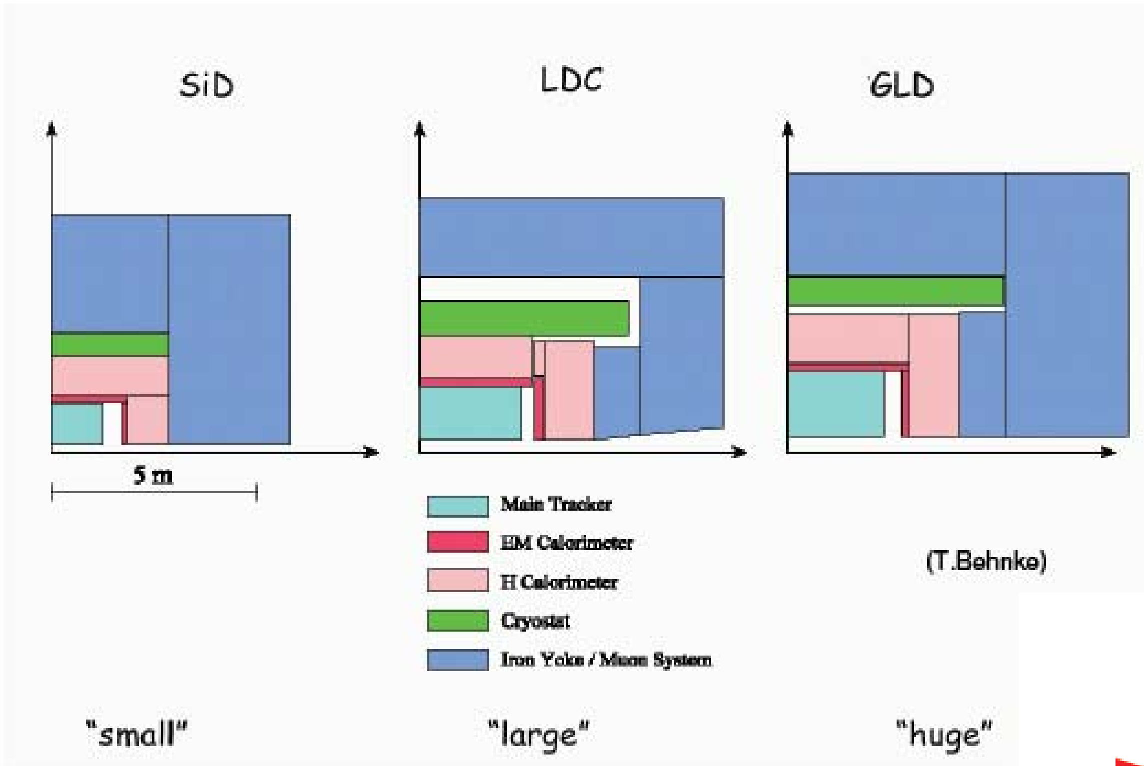
Itanium 2

Implement Quad Precision in Software ?

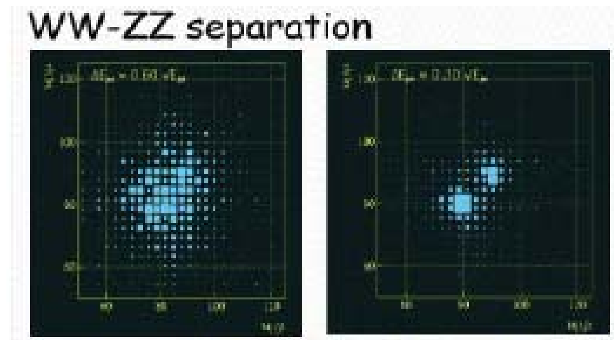
Yes! More flexible (still double OK for most calculations)

BUT: feasible also for very large codes ?

Frank Gaede: Simulation and Reconstruction Software for the ILC



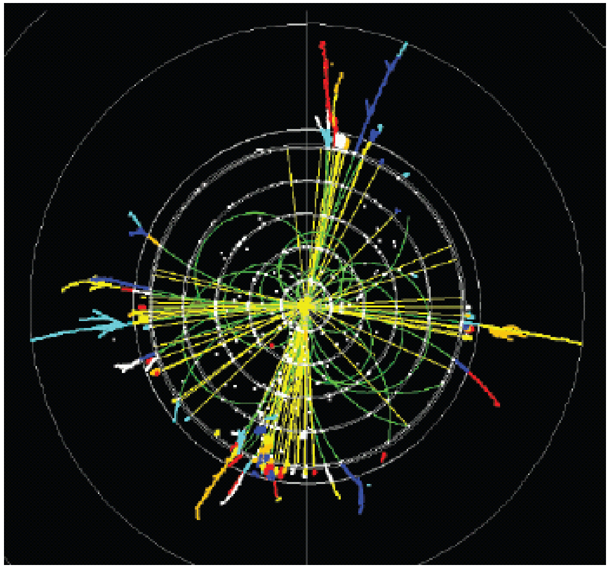
three interregional detector concept studies ongoing

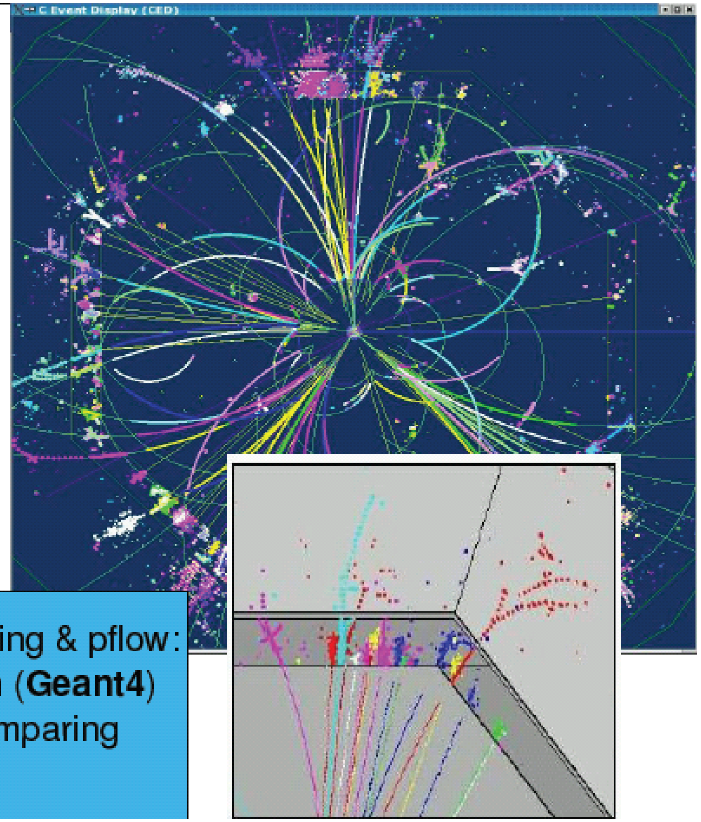


Need common **Simulation** and **Reconstruction** software to study detector concepts' performance !

Particle Flow

- reconstruct all single particles
- use tracker for charged particles
- use Ecal for photons
- use Hcal for neutral hadrons





Frank Gaede

Need software tools to improve clustering & pflow:
- detailed hadronic shower simulation (**Geant4**)
- framework for developing and comparing pflow algorithms

	Description
Simdet	fast Monte Carlo
SGV	fast Monte Carlo
Lelaps	fast Monte Carlo
Mokka	full simulation – Geant4
Brahms-Sim	Geant3 – full simulation
SLIC	full simulation – Geant4
LCDG4	full simulation – Geant4
Jupiter	full simulation – Geant4
Brahms-Reco	reconstruction framework (most complete)
Marlin	reconstruction and analysis application framework
hep.lcd	reconstruction framework
org.lcsim	reconstruction framework (under development)
Jupiter-Satelite	reconstruction and analysis
LCCD	Conditions Data Toolkit
LCIO	Persistency and datamodel
JAS3/WIRED	Analysis Tool / Event Display

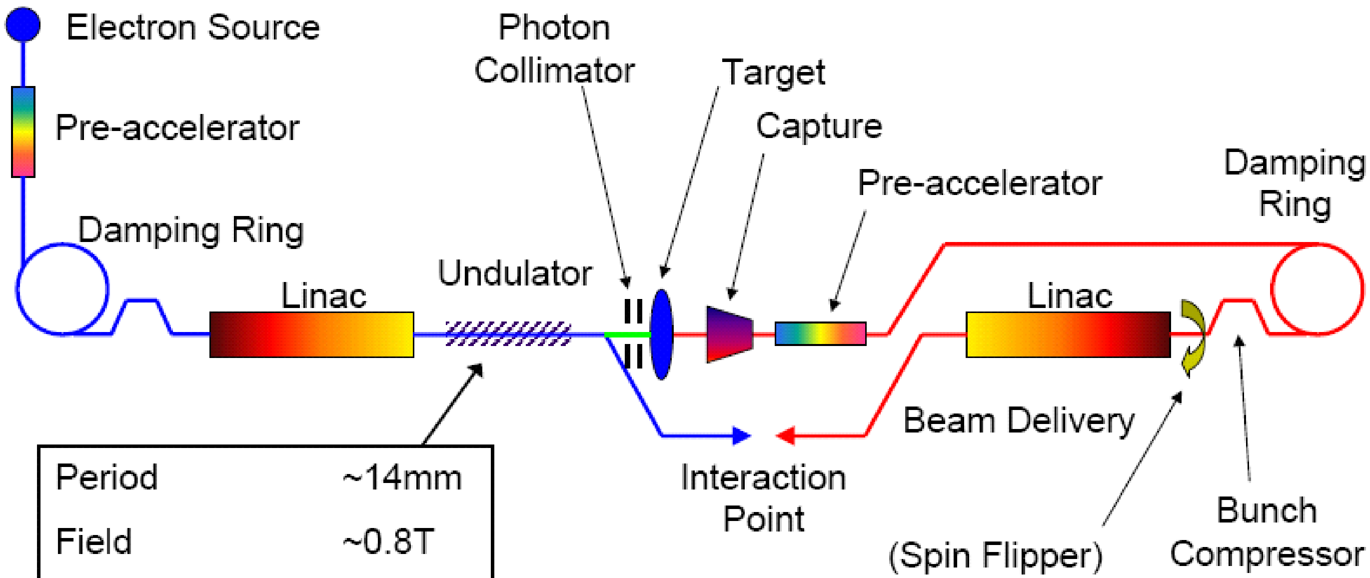
A Common Software Framework for the ILC ?

→ **M**odular **A**nalysis & **R**econstruction for the **L I N**ear Collider

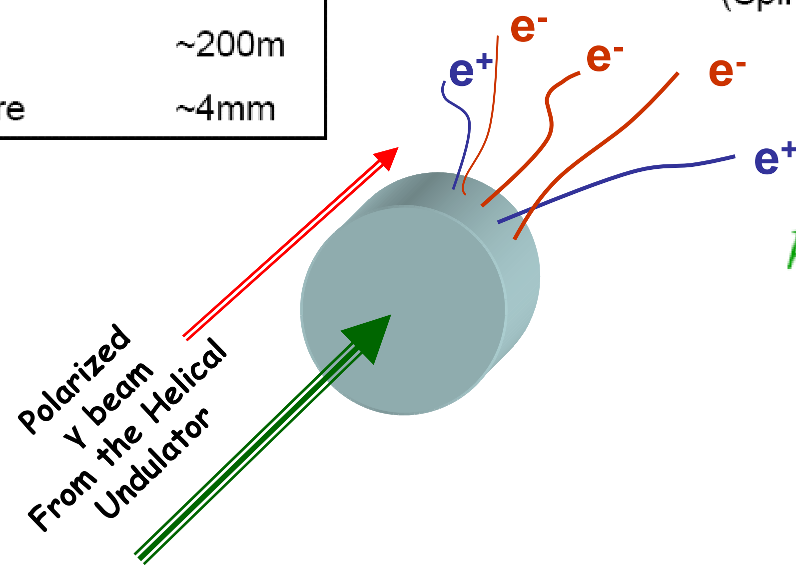
C++, Java

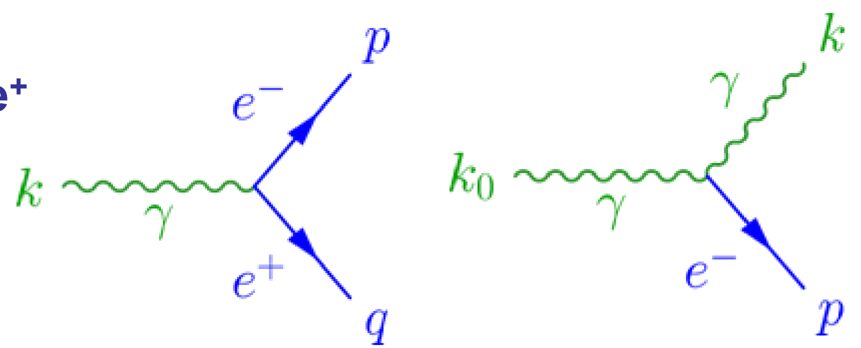
Holy grail: have an internationally used common software framework for the ILC that is used to support R&D and the detector concept studies !

Andreas Schälicke: Monte Carlo based studies of polarized positrons source for the International Linear Collider (ILC)



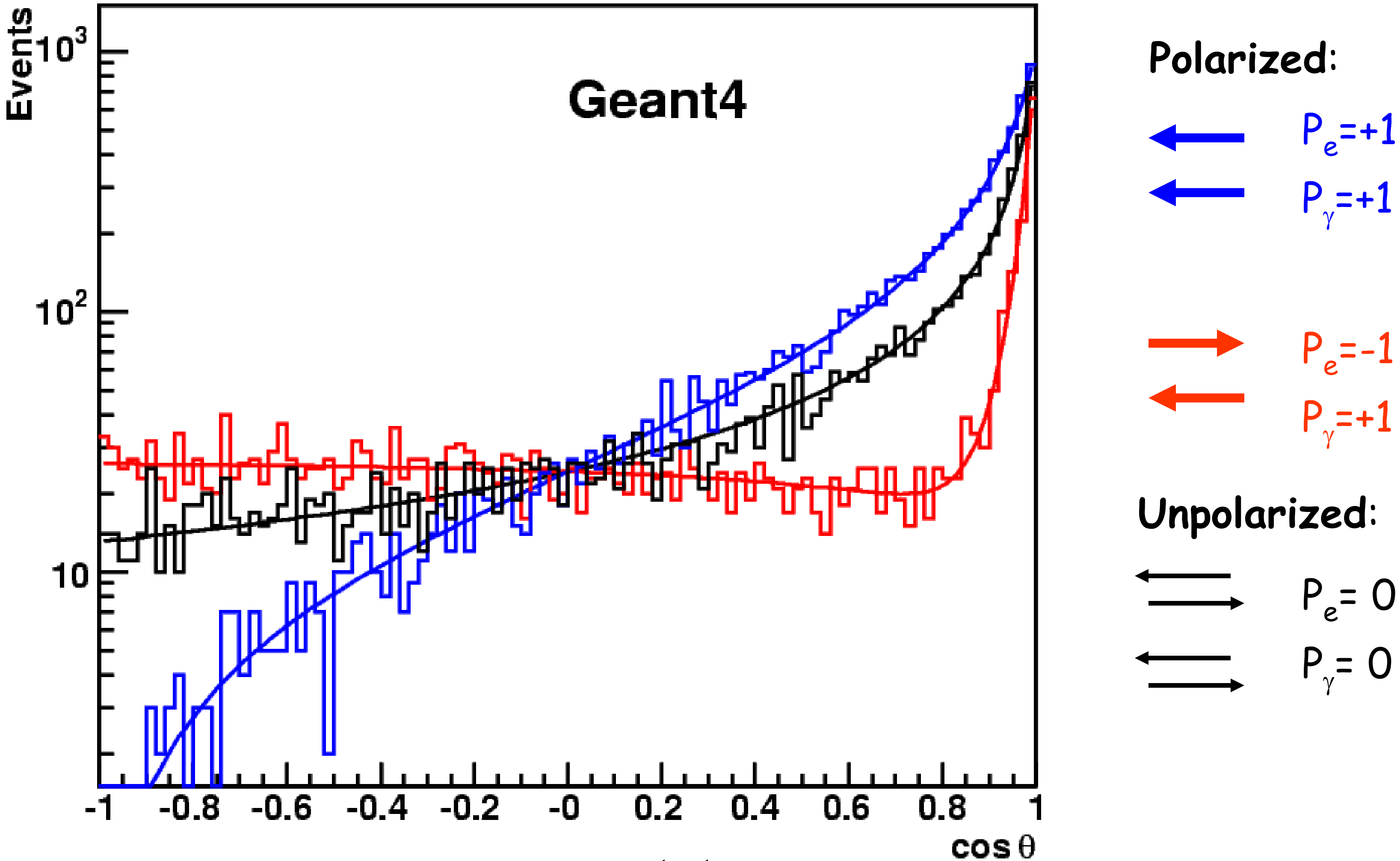
Period	~14mm
Field	~0.8T
Length	~200m
Aperture	~4mm





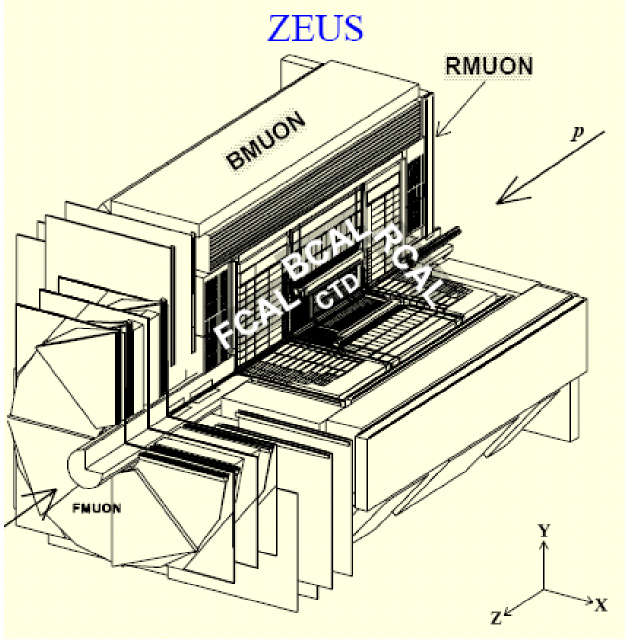
→ Stokes formalism

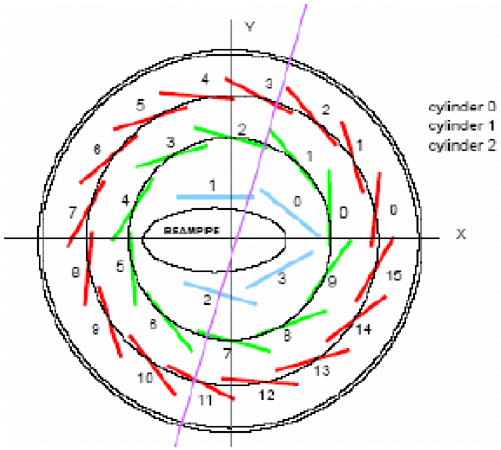
Some results from the new simulation (Compton Scattering): Andreas Schälicke



next steps:
compare with EGS4, experimental results (E166)

Takanori Kohno:
Alignment of the ZEUS MVD with Comic Rays



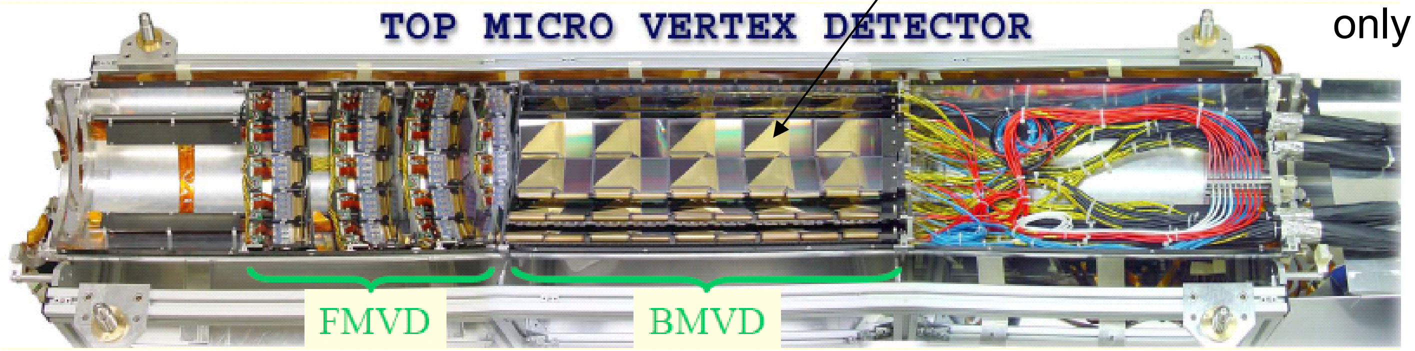


determine:

- relative position of ladders
- absolute position in ZEUS

Goal:
alignment precision < 10 μm

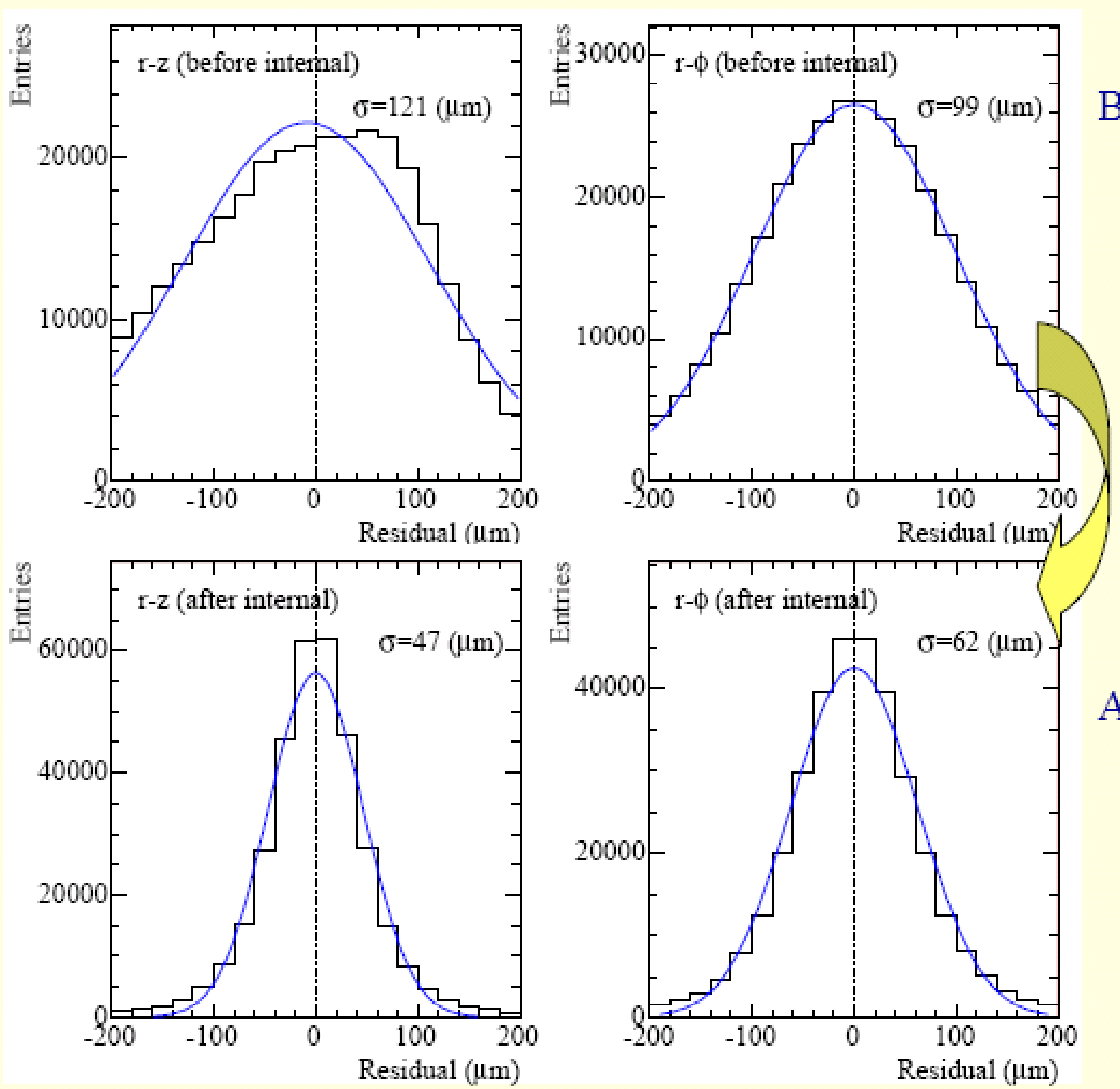
„ladders“



Cosmics:
only useful for
barrel

Effect of alignment procedure on residuals:
(local χ^2 minimation with linear approximation)

Takanori Kohno



Before internal alignment

- Residual centered
- Width became narrow

After internal alignment

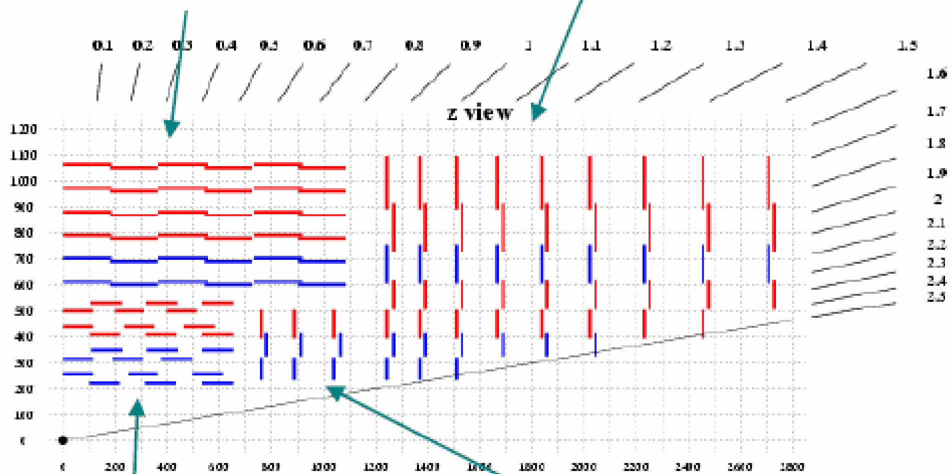
$$\sigma_{rz} = 47\ \mu m$$
$$\sigma_{r\phi} = 62\ \mu m$$

Residual offset < 10 μm

Thomas Speer: Track Reconstrucion in the CMS Detector

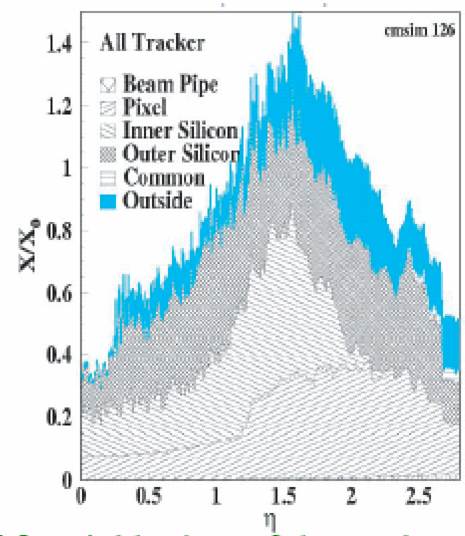
Outer Barrel (TOB): 6 layers
- Thick (500 μm) sensors
- Long Strips

Endcap (TEC): 9 disks pairs
- r<60cm: Thin sensors
- r>60cm: Thick sensors



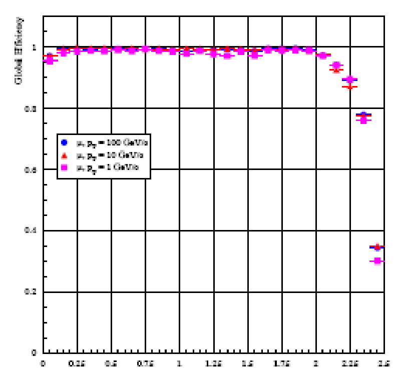
Inner Barrel (TIB): 4 layers
- Thin (320 μm)sensors
- Short Strips

Inner Disks (TID): 3 disks pairs
- Thin sensors

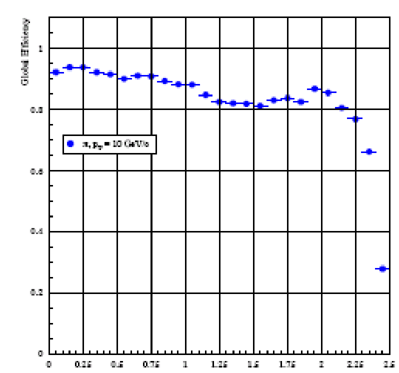


Material budget of the tracker

muons, p_T = 1, 10, 100 GeV/c



pions, p_T = 10 GeV/c



→ Combinatorial Kalman Filter

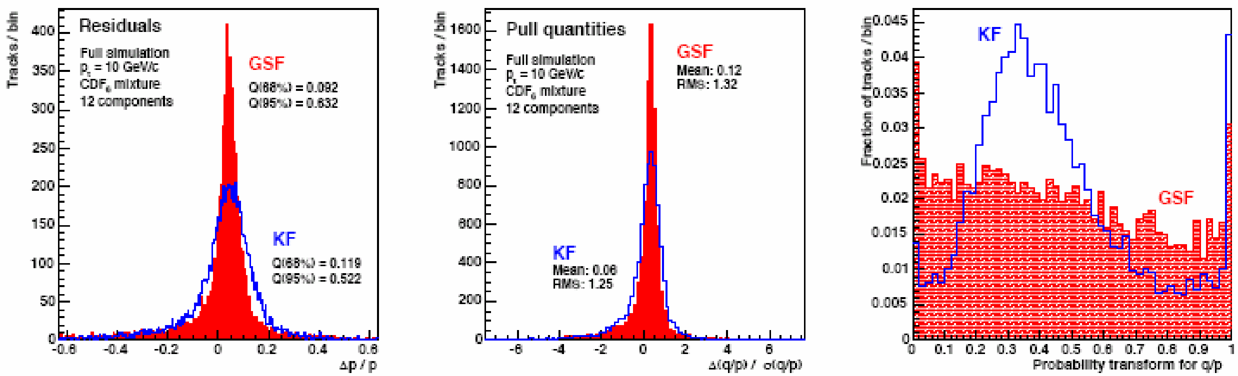
(equivalent to least squares fit)

efficient even for non-Gaussian errors

Thomas Speer

Alternative Filter Methods:

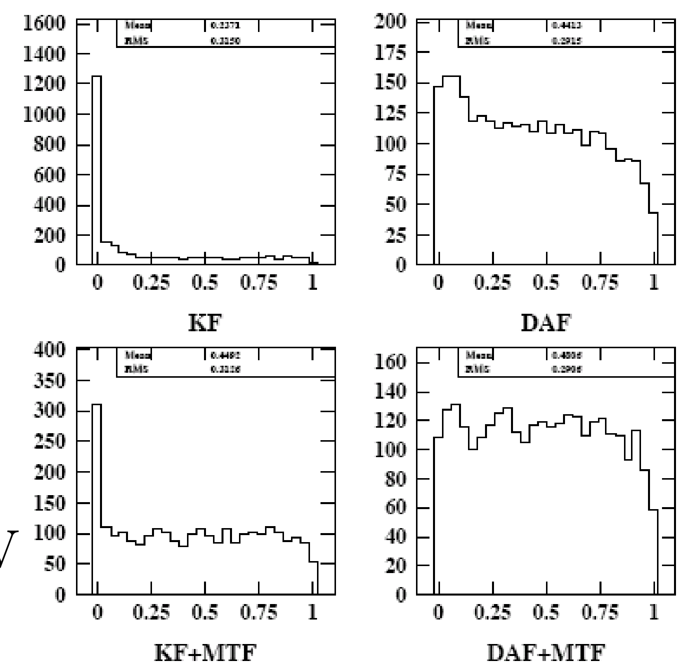
The Gaussian-sum Filter (non-linear generalization of KF)



non-Gaussian tails in hit residuals (multiple Coulomb scat.)

can improve core of distributions,
little improvement in tails (➔ first layer radiation)

χ^2 probability



Adaptive Filters:

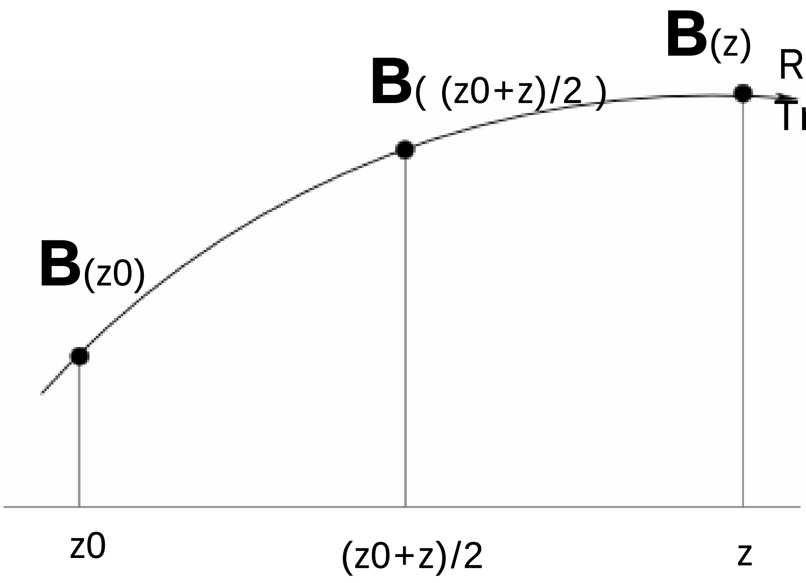
- Deterministic Annealing Filter (DAF)
- Multi-track fit (MTF)

➔ both need hit collection and track seeds

Bench mark : $H \to \tau^+\tau^- \quad M(H) = 500\ \mathrm{GeV}$

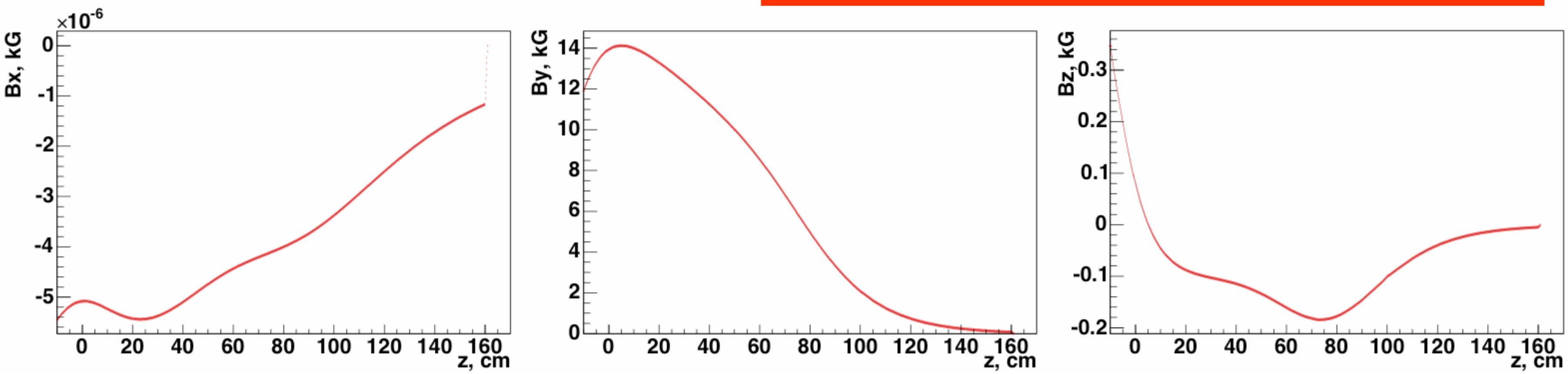
➔ Improvement in high track density

Sergey Gorbunov: Analytical Formula for Track Extrapolation in an Inhomogeneous Magnetic Field

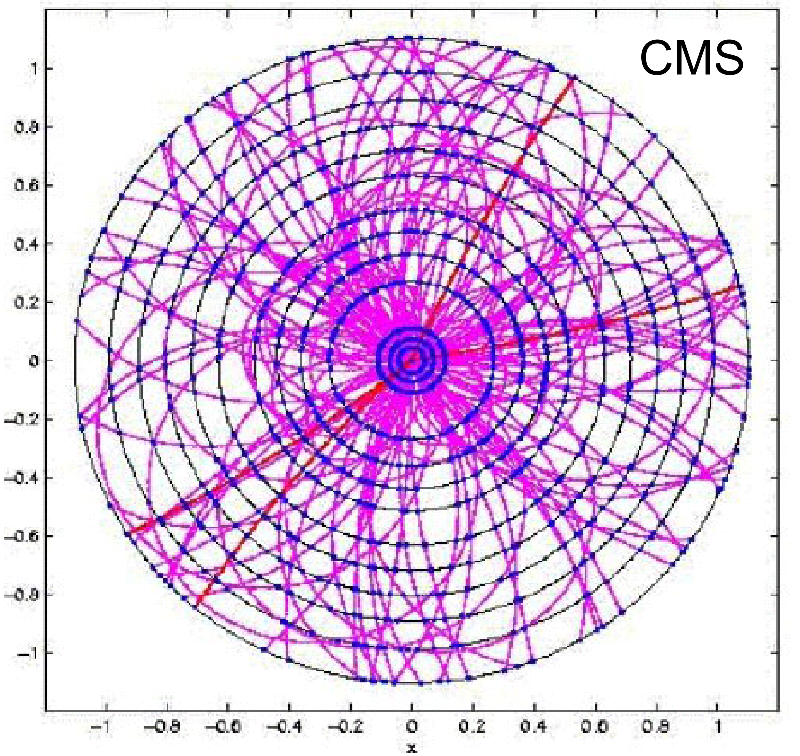


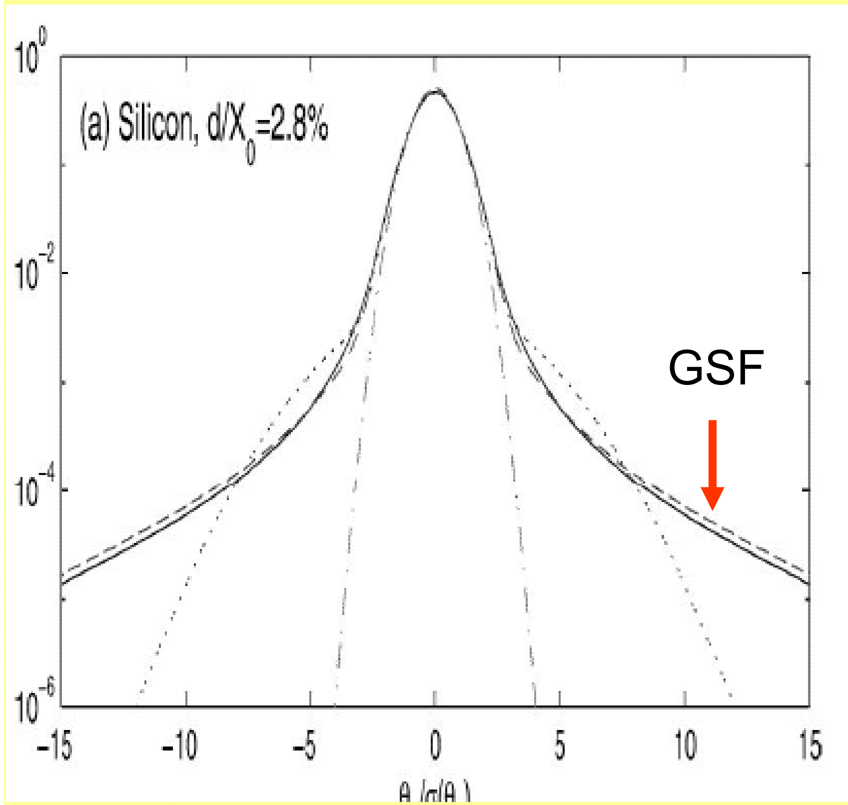
$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = F_l = \kappa \cdot q \cdot \mathbf{v} \times \mathbf{B}$$

3[rd] order analytical formula
(same accuracy as 4[th] order Runge-Kutta)



Are Strandlie: Modeling non-Gaussian Tails of Multiple Coulomb Scattering in Track Fitting with GS-Filter





Treatment of non-Gaussian tails implemented in simple detector geometry

Standard method: Kalman Filter

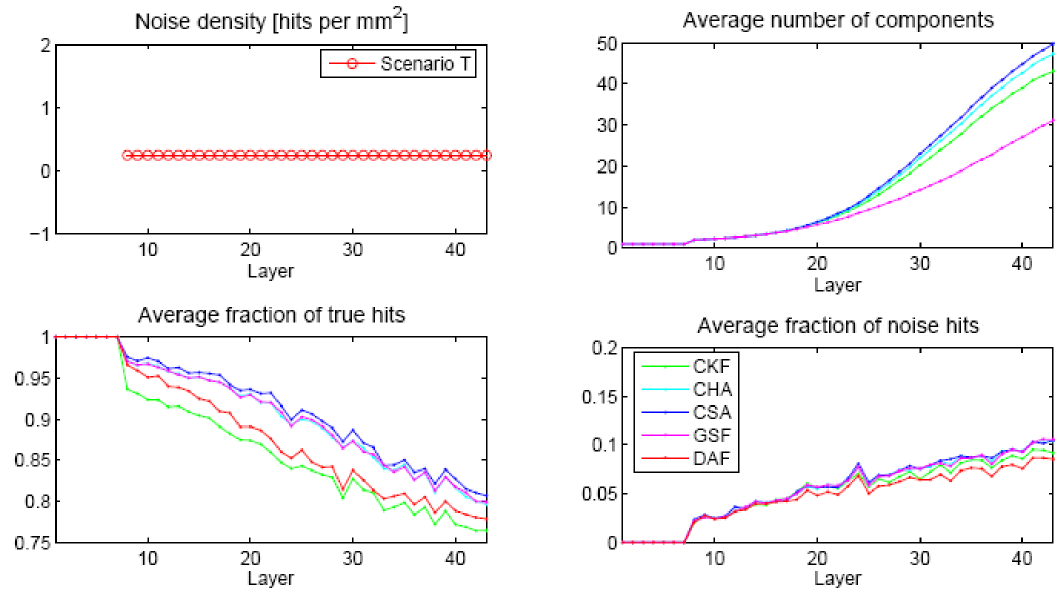
Gaussian Sum Filter slightly more precise than KF

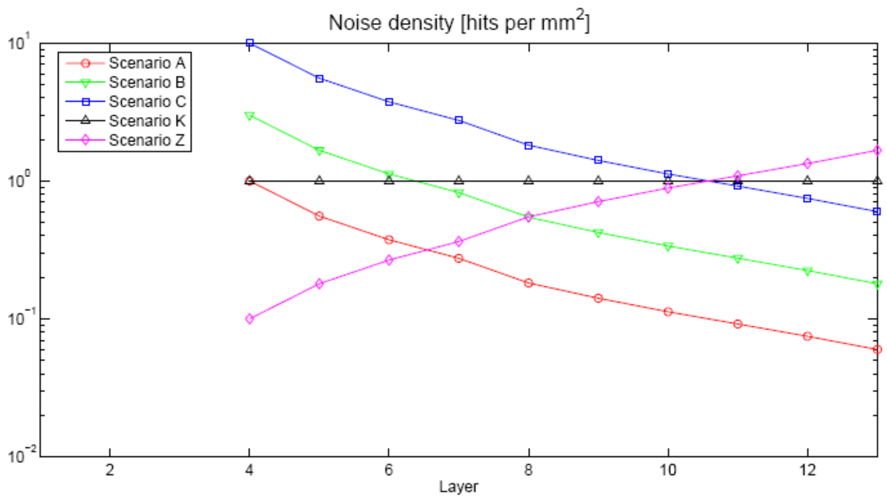
Are Strandlie: Adaptive Filters for Track Finding

Filters studied:

- Combinatorial Kalman Filter („standard filter“ with seeds)
- Competition with hard/soft assignment
- Gaussian Sum Filter
- Deterministic Annealing Filter

Example: ATLAS tracker

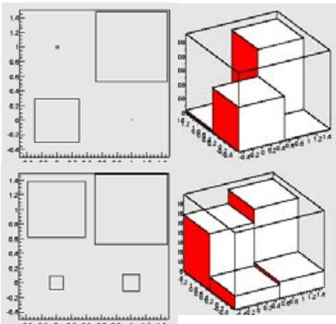




Method	CKF	CHA	CSA	GSF	DAF
Avg. true hits	37.4	39.1	39.4	39.1	38.0
Avg. noise hits	2.23	2.36	2.39	2.35	2.04
Avg. missg hits	3.40	1.56	1.25	1.56	2.97
Efficiency	0.97	0.99	1.00	0.99	1.00
Time per seed[1]	1.00	1.10	1.18	0.33	0.06
Fake rate	0.01	0.03	0.54	0.03	0.02
Time per seed[2]	0.14	0.16	0.64	0.08	0.06

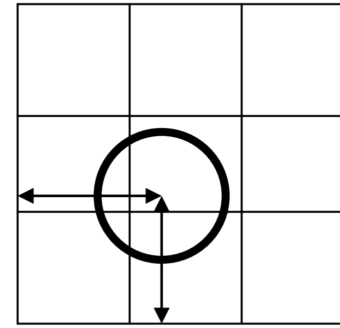
Scenario T

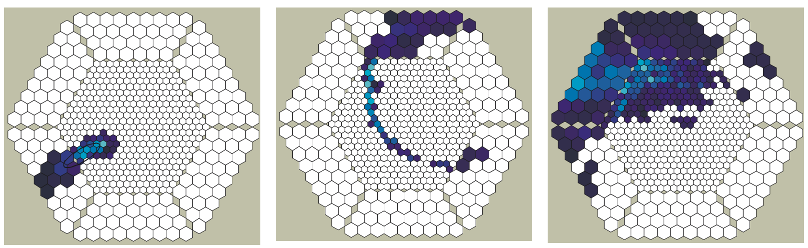
Jens Zimmermann: Statistical Learning Methods



Pileup vs.
Single photon

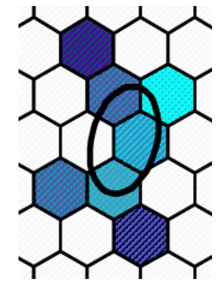
Reconstruction of the
incident position with
subpixel resolution



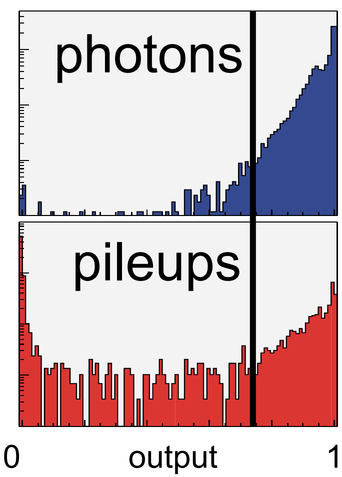


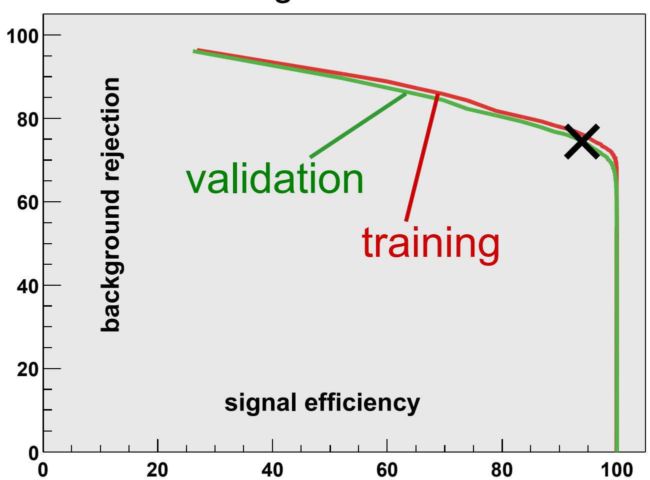
Gamma vs.
Myon vs.
Hadron event

Reconstruction
of the primary
photon energy

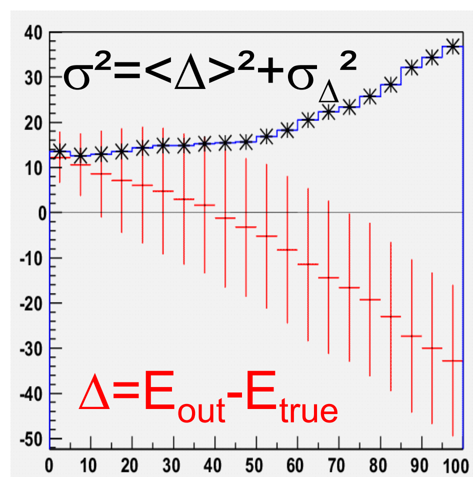


XEUS – Photon recognition

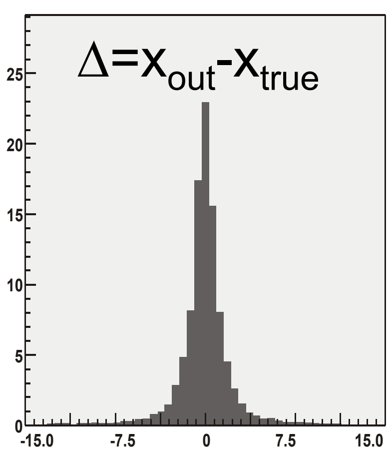




MAGIC – E[GeV]



XEUS – x[μm]



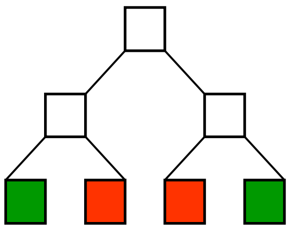
Classification, Regression

Decision Trees

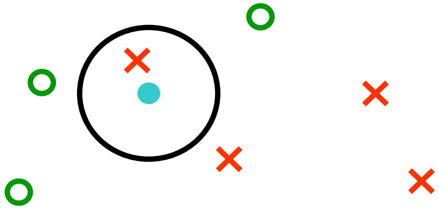
C4.5

CART

CAL5



Local Density Estimators



k-Nearest Neighbours

Naïve Bayes

“Maximum Likelihood”

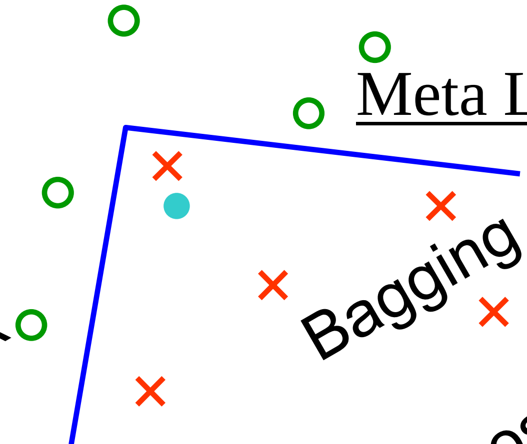
Linear Separation

Linear Discriminant Analysis

Support Vector Machine

Neural Network

Meta Learning Strategies



Bagging

Boosting

Random Subspace

implemented in unified system

Systematic Uncertainties: Comparison of the Performance of Statistical Learning Methods

Jens Zimmermann

NN: 96.5% vs. SVM: 95.7%

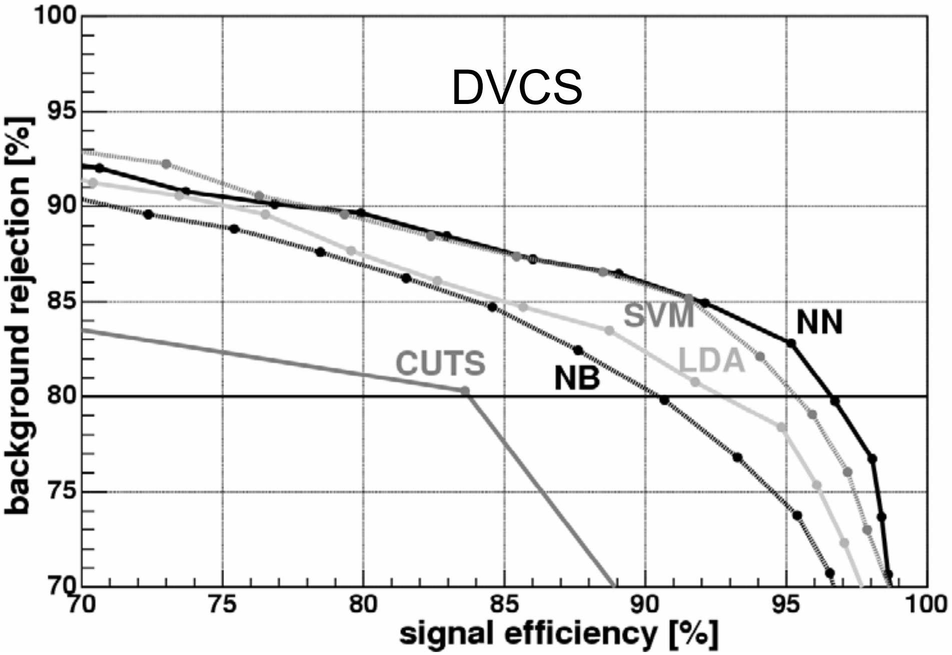
Statistically significant?

Build 95% confidence interval!

$$[\mu - z\sigma_\mu, \mu + z\sigma_\mu]$$

σ_μ is the variation over different parts of the test set

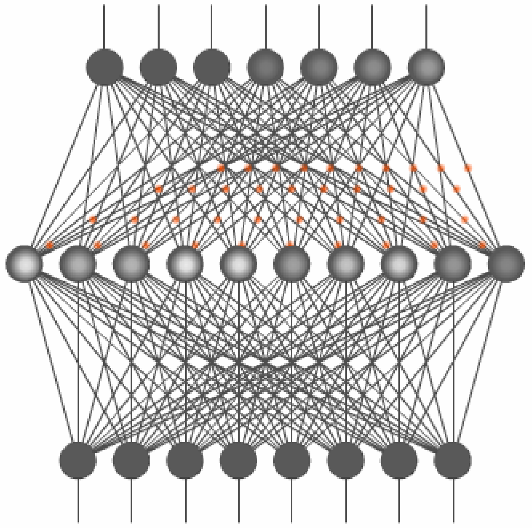
n	3	4	5	6	7	8	9	10	11	12	13	14	15
z	2.39	2.64	2.81	2.94	3.04	3.14	3.20	3.26	3.32	3.37	3.41	3.45	3.49



Neural Network	96.5%
$\Delta = 0.7\% \pm 0.9\%$	
Support Vector Machine	95.7%
$\Delta = 3.3\% \pm 1.5\%$ (!)	
Linear Discriminant Analysis	92.5%
$\Delta = 2.2\% \pm 2.1\%$ (!)	
Naive Bayes	90.5%
$\Delta = 6.8\% \pm 2.5\%$ (!)	
Cuts	83.6%

efficiencies for fixed rejection of 80%

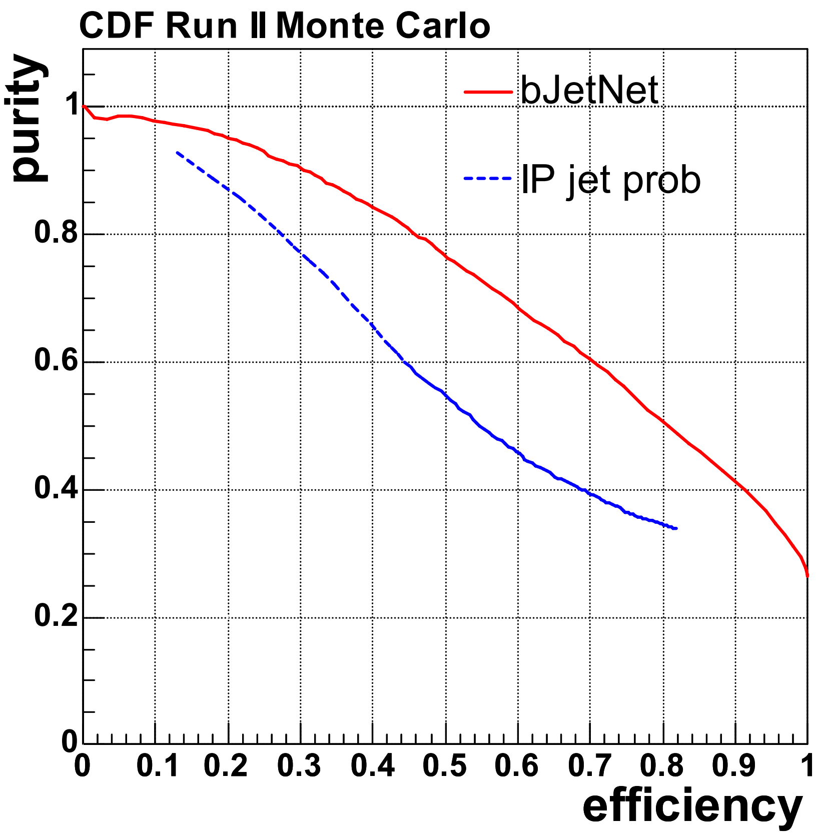
Ulrich Kerzel: The NeuroBayes Neural Network Package



Outputs („probabilities“)

Inputs
(preprocessing)

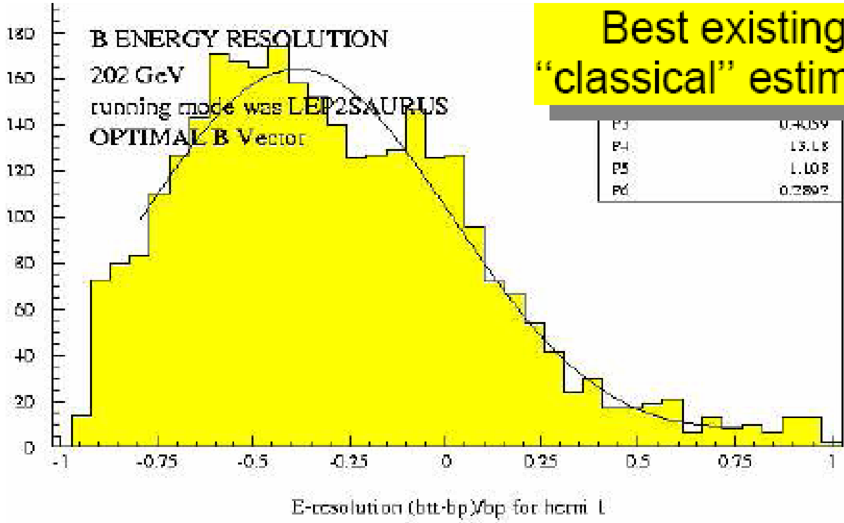
Input: „intelligent“ variables,
e.g. jet variables



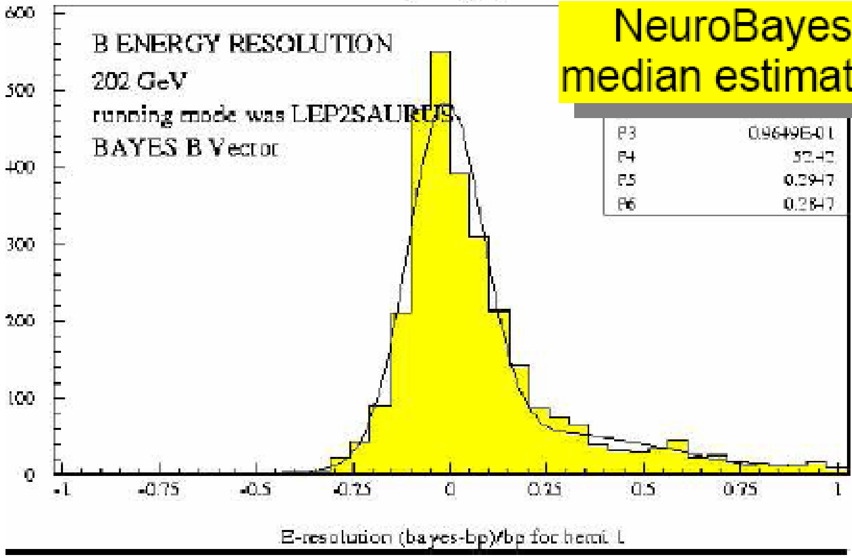
- Bayesian regularization
- Simple networks, small weights
- Remove non-significant nodes during training

Ulrich Kerzel

Recent real world application



Best existing "classical" estimator



NeuroBayes median estimator

Relative resolution of reconstructed B hadron energy in DELPHI at LEP II at 202 GeV energy

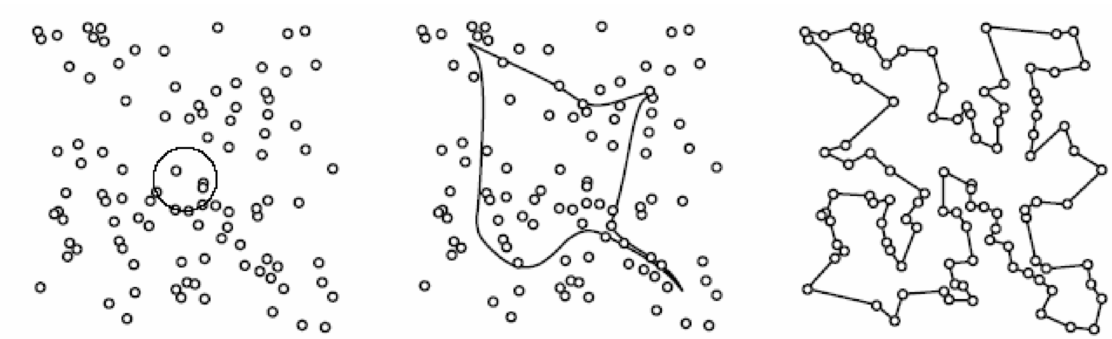
(completely inclusive)

core resolution 40% -> 10%

Spin-off: commercial applications
Company founded:
medicine, insurances, banks, …

Sergey Gorbunov:
Elastic Neural Net for Standalone Rich Ring Finding

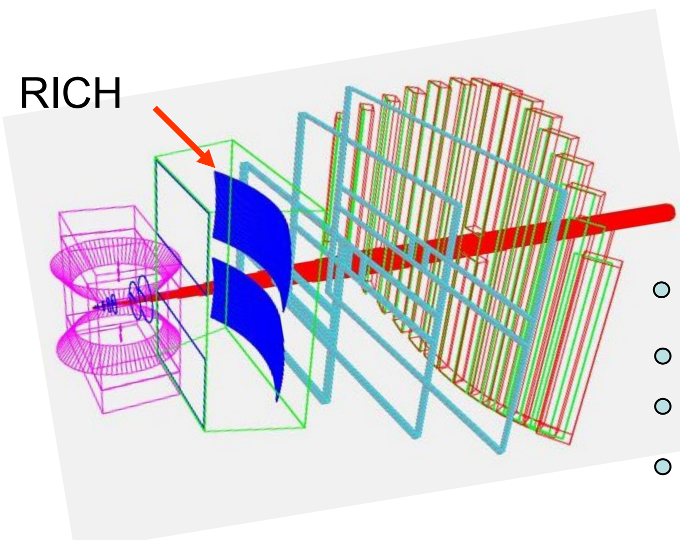
→ collective selforganization networks (unsupervised learning + neighbor relations)



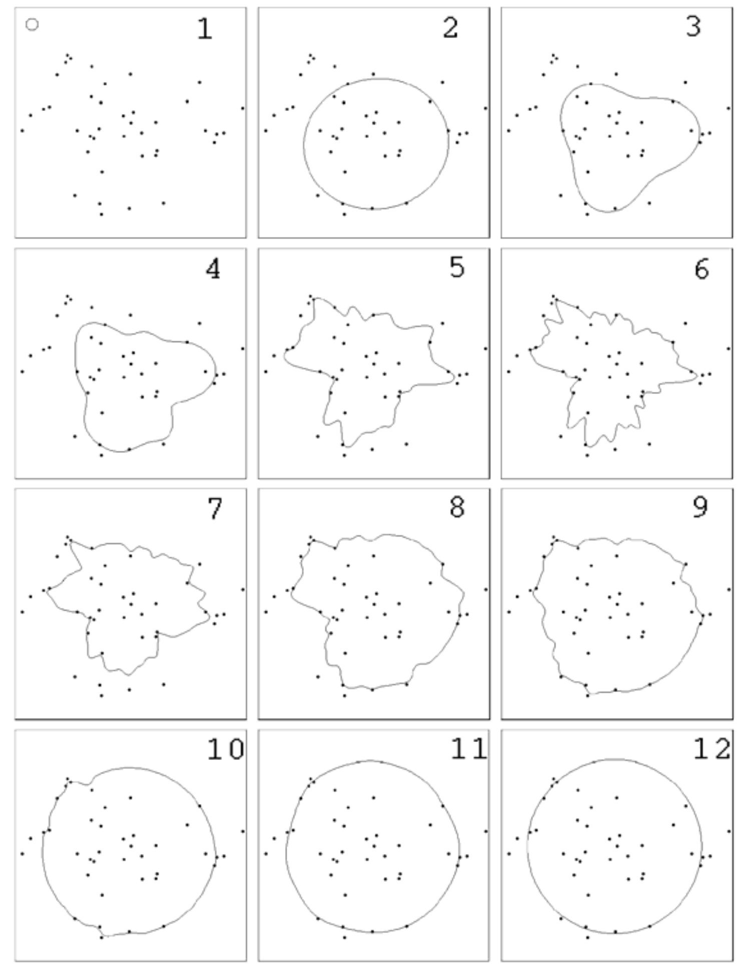
travelling salesman problem

CBM-experiment (GSI)

RICH

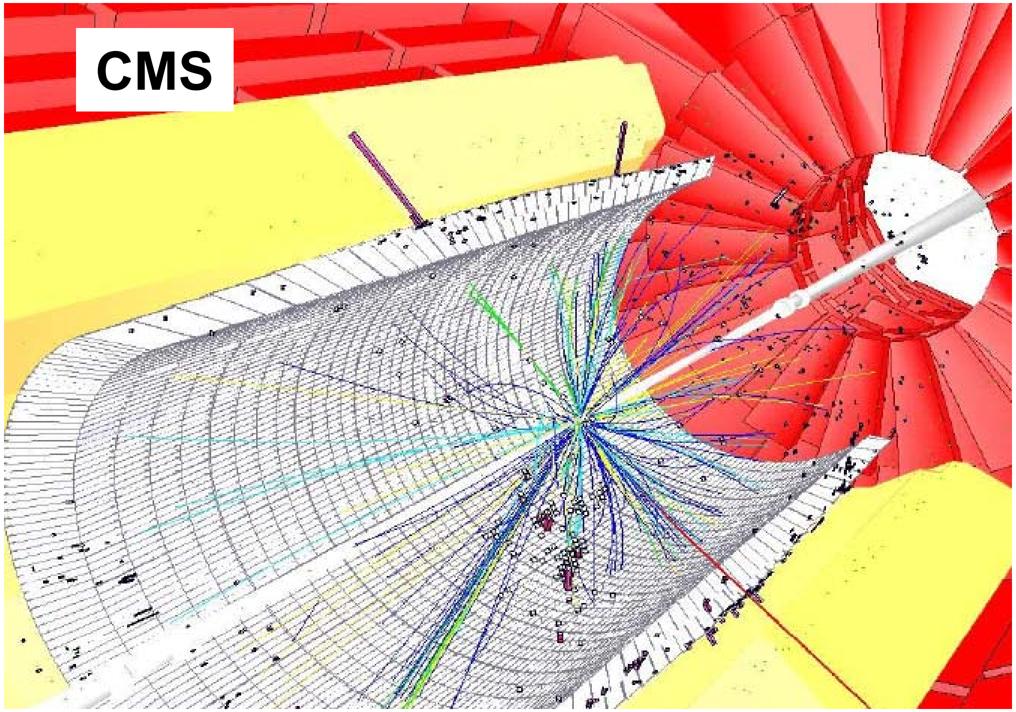


Few iterations necessary to find solution



- standalone
- simple
- efficient (92 %)
- Fast (1ms)

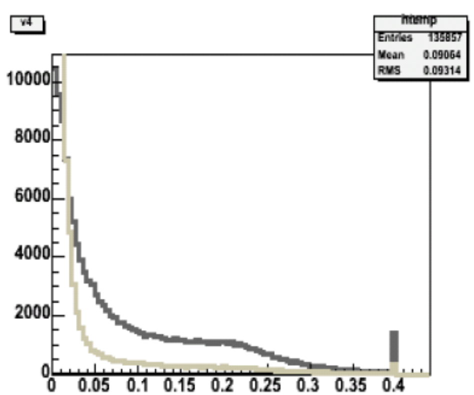
Aatos Heikkinen: Tagging b Jets Associated with Heavy Neurtral MSSM Higgs Bosons

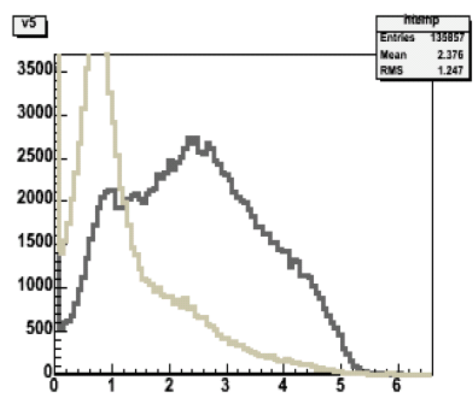


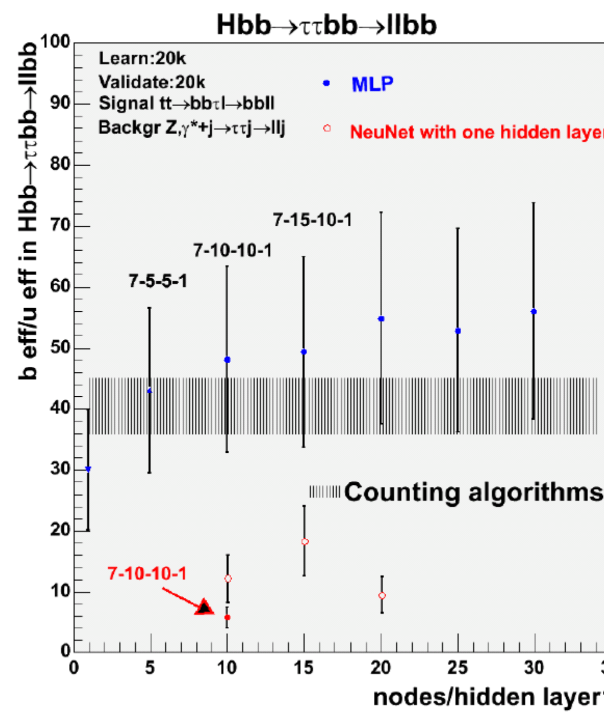
Standard method vs neural approach to solve pattern recognition problem:

find Higgs candidates in QCD background environment

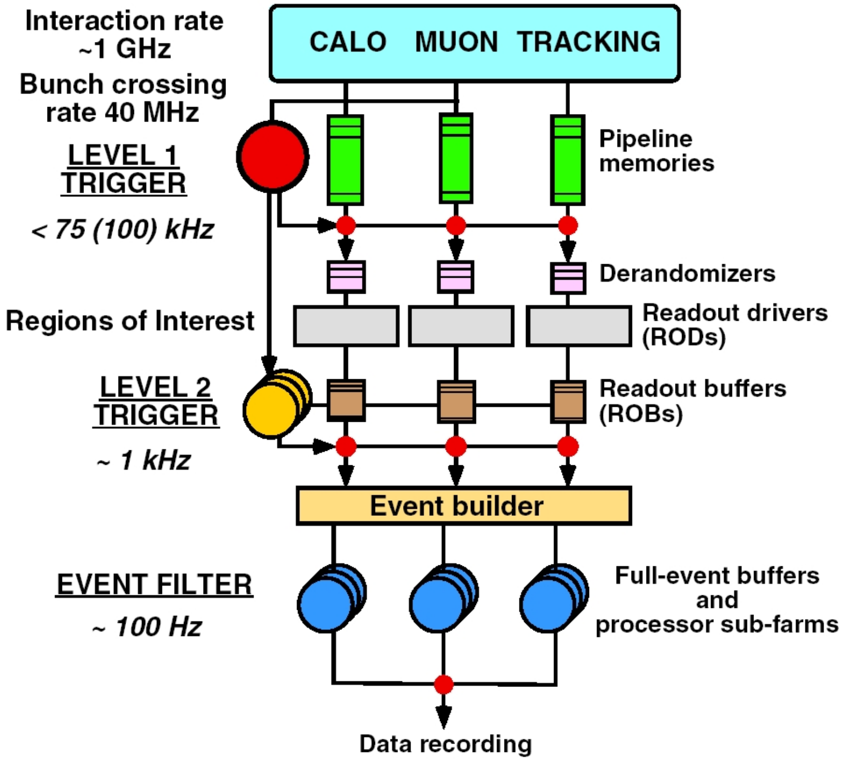
use available packages:
ROOT, NeuNet (MLP's)
SOM-PAK, LVQ-PAK

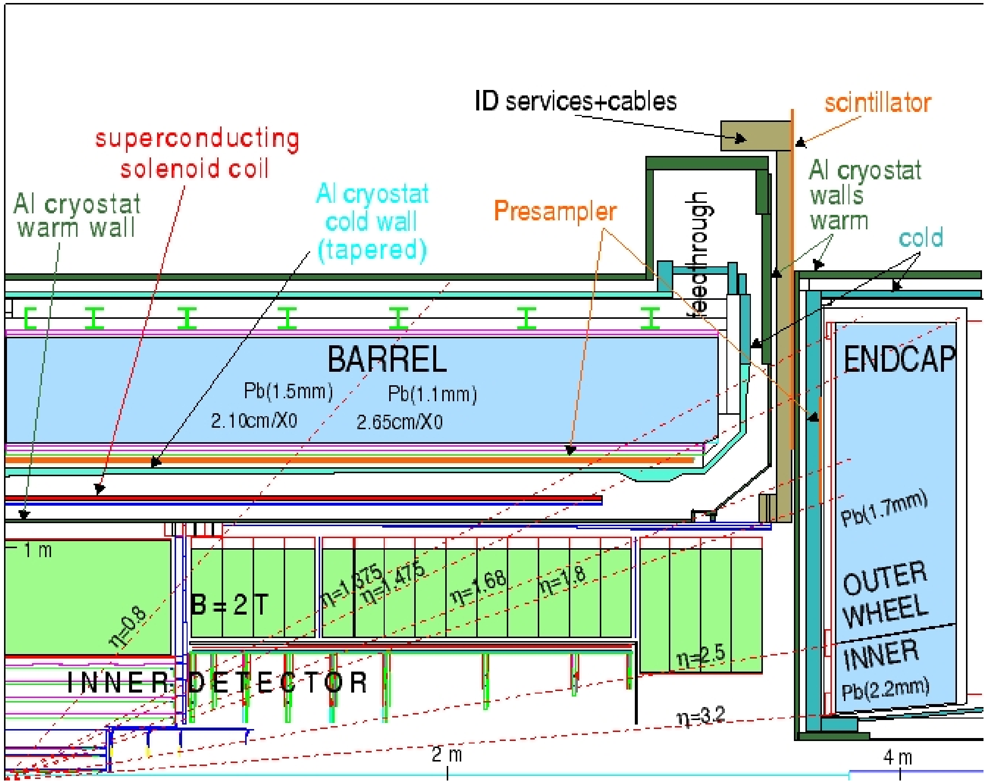






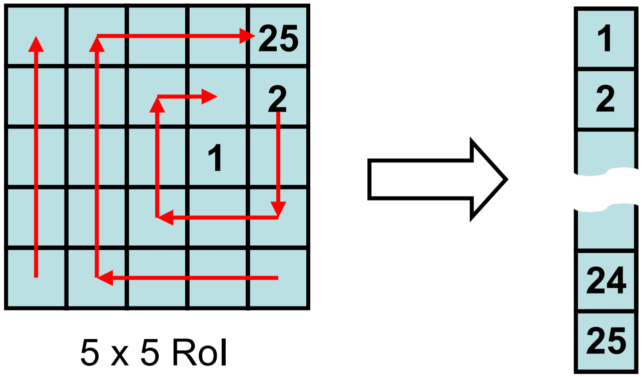
José M de Seixas: A Segmented Principal Component Analysis Applied to Calorimetry Information at ATLAS





Preprocessing of neural inputs:

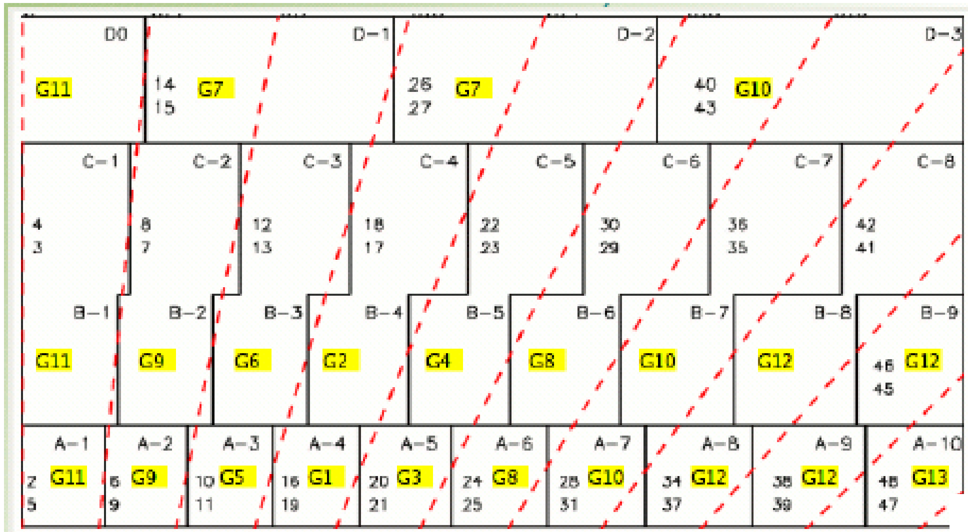
Principal component extraction



5 x 5 RoI

Had. Section

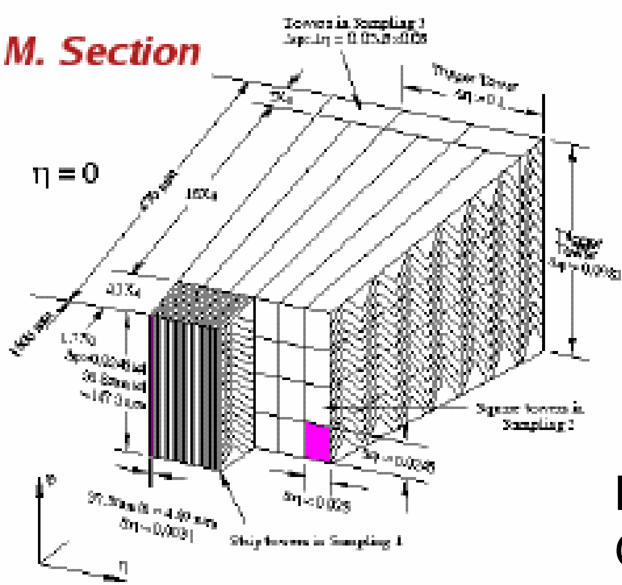
José M. Seixas



Neural Net for e/π separation:

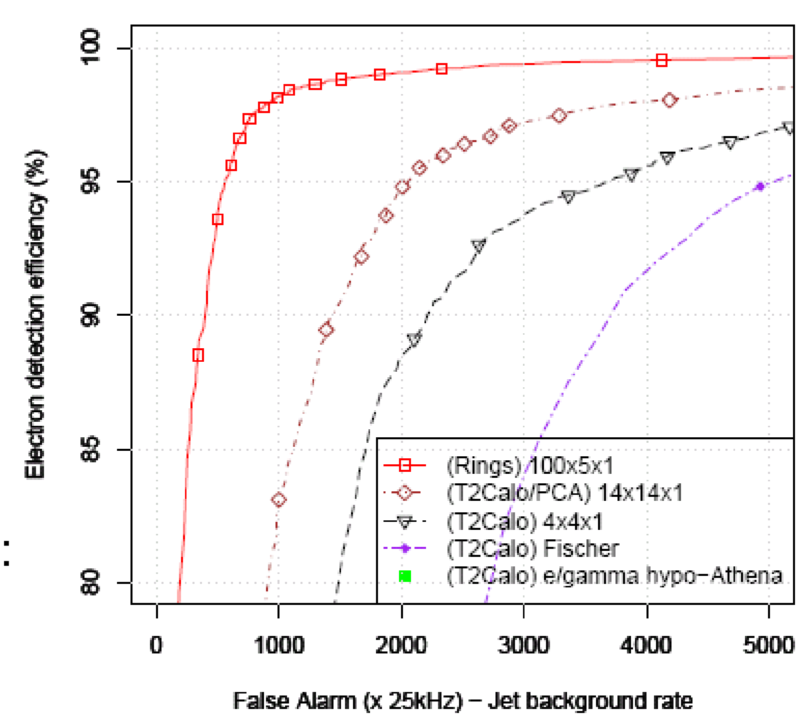
Nr. of inputs/nodes depending on detail of detector information

E.M. Section

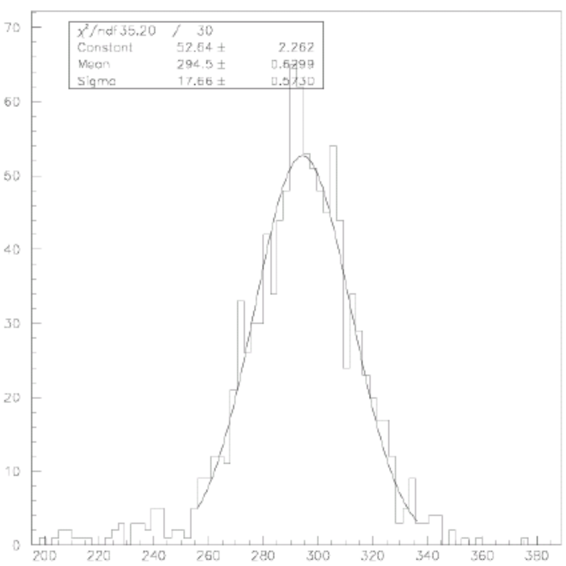


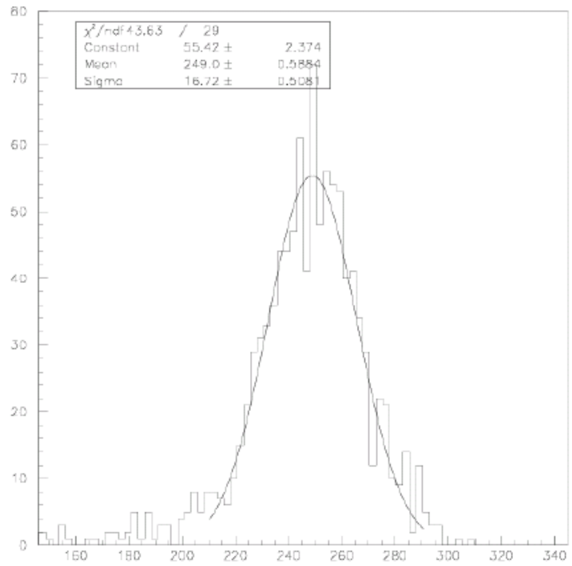
NN on DSP:
O(10 μs)

R.O.C. for e/jet discrimination



Paulo Vitor Magacho da Silva: Energy Reconstruction for a Hadronic Calorimeter Using Neural Networks





Energy Resolution

Network with
2 hidden layers

$$NNoutput\,,\quad \frac{\sigma}{\mu} = 6.0\,\%$$

$$Raw\ data\,,\quad \frac{\sigma}{\mu} = 6.7\,\%$$

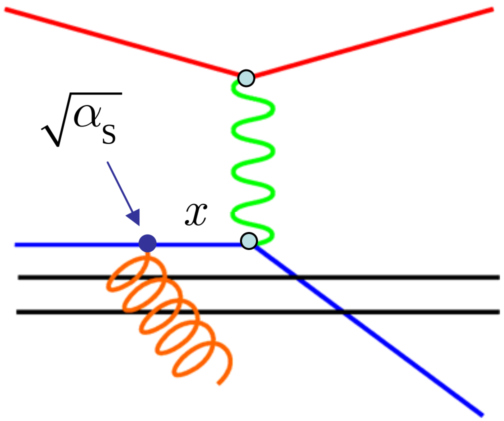
Network also improves linearity

	Non-linearity (%)
Raw	2.2
H1	1.3
Neural Net.	0.42

Andrea Piccione: Neural Networks Approach to Parton Distribution Fitting

- The cross section

$$\frac{d^2\sigma}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[[1+(1-y)^2]F_1 + \frac{1-y}{x}(F_2 - 2xF_1)\right]$$



- The structure function (QCD + parton model)

$$F_2(x,\ Q^2) = x\left[\sum_{q=1}^{n_f} e_q^2\, C^q \otimes q_q(x,\ Q^2) + 2n_f\, C^g \otimes g(x,\ Q^2)\right]$$

Traditional method: polynomial ansatz for parton distribution functions at some low Q_0

$$xq(x,Q_0) = Ax^B(1-x)^C\,[1 + D\sqrt{x} + Ex]$$

Neural approach: functional form „chosen“ by data

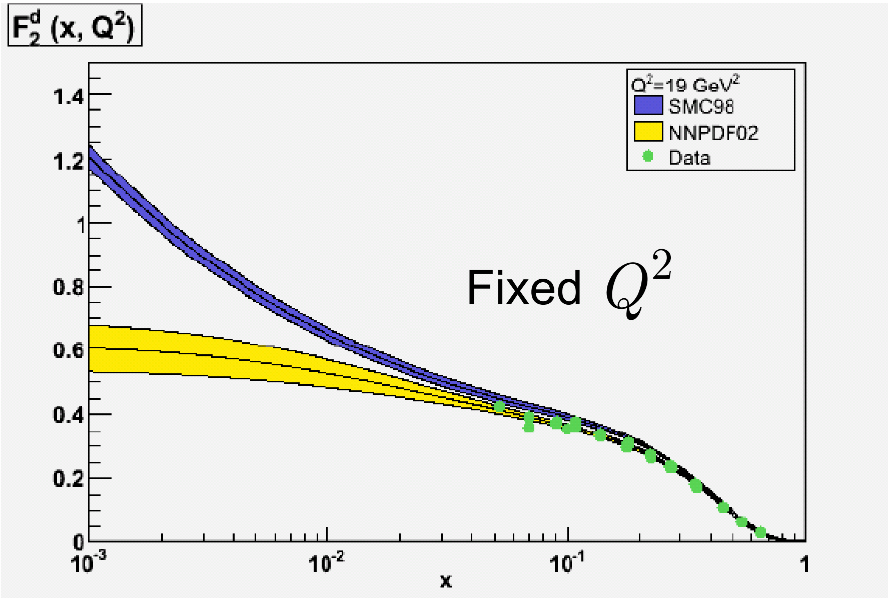
inputs: $x, Q^2, \ln x, \ln Q^2$

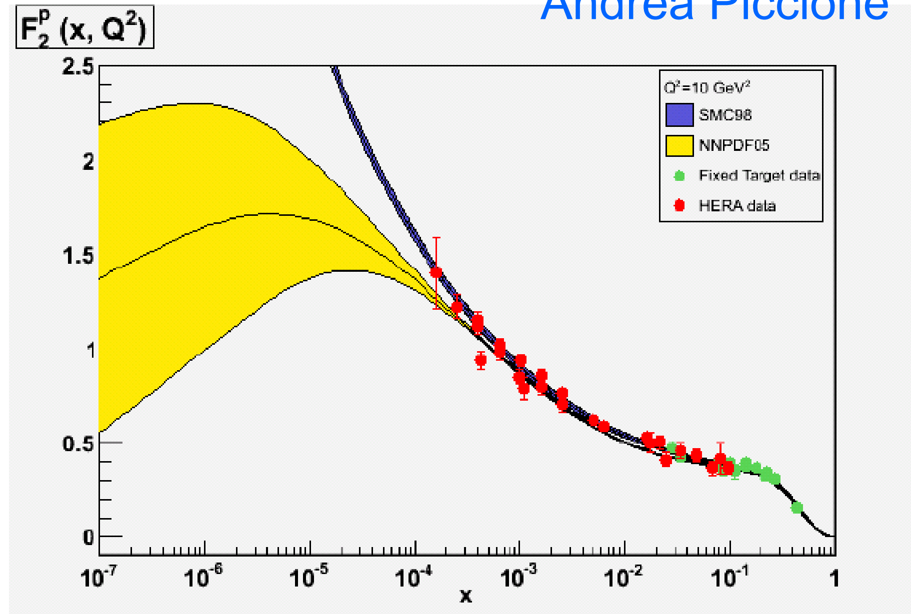
- Architecture: 4-5-3-1

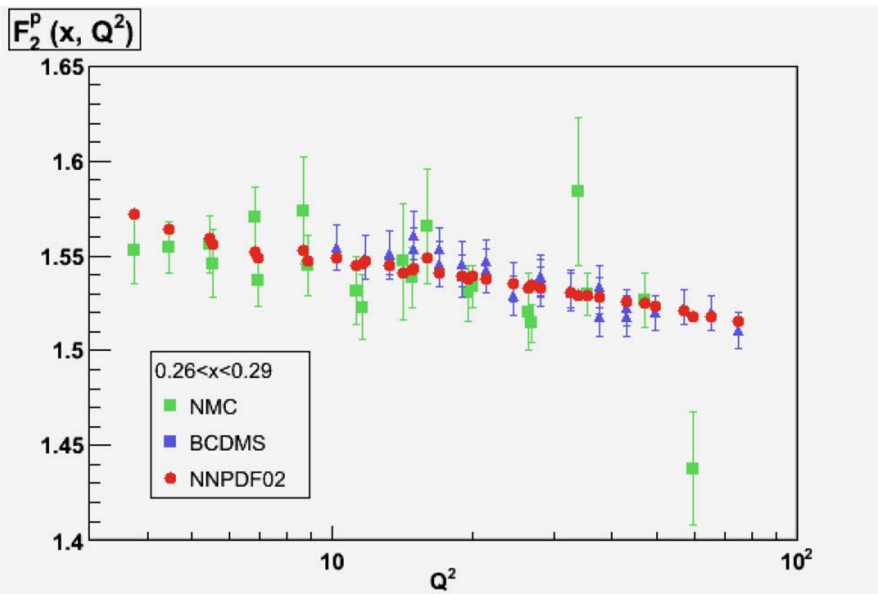
output: $F_2(x,Q^2)$

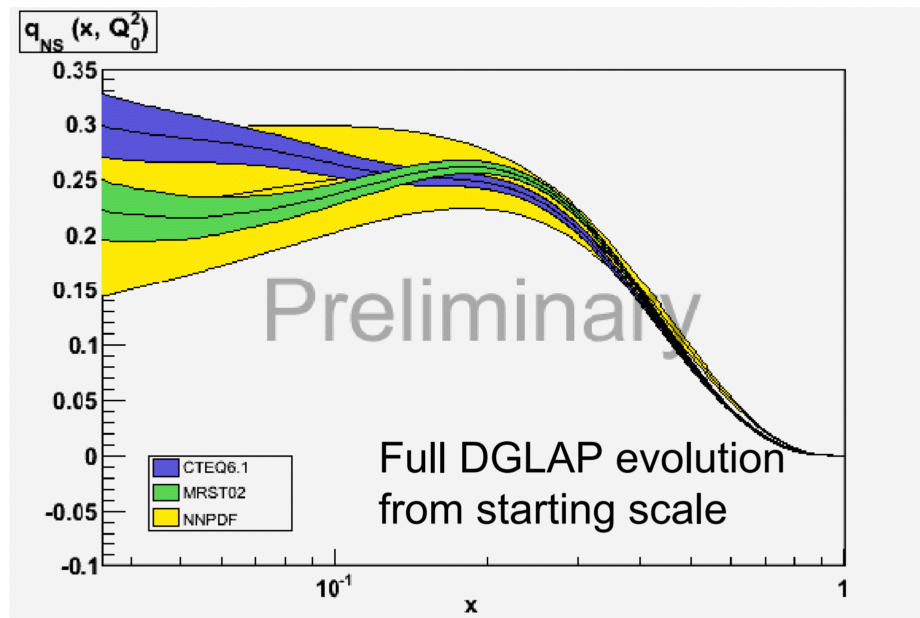
Train net with data set (replicas)

Andrea Piccione









Conclusions

- Session 2 was quite lively, good attendance
- Good workshop atmosphere
- In general high quality talks,
  showing originality in the spirit of the workshop
- Contributions of non-statistical and statistical (neural) algorithms
  in balance
- Personal opinion (maybe also shared by others):
  More time discussions („workshop“)
- Looking forward for the next workshop !