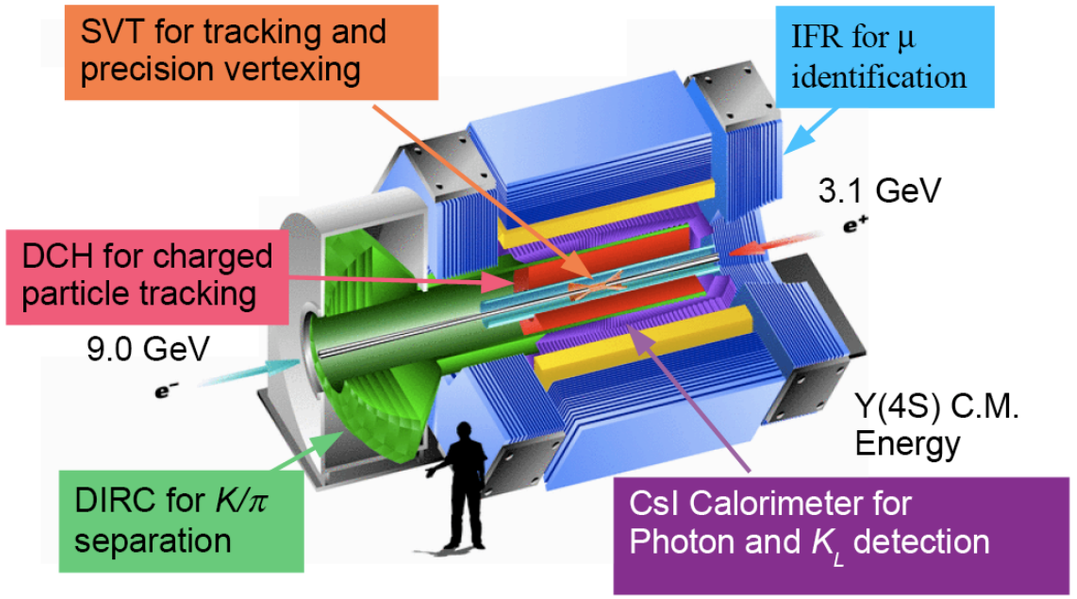
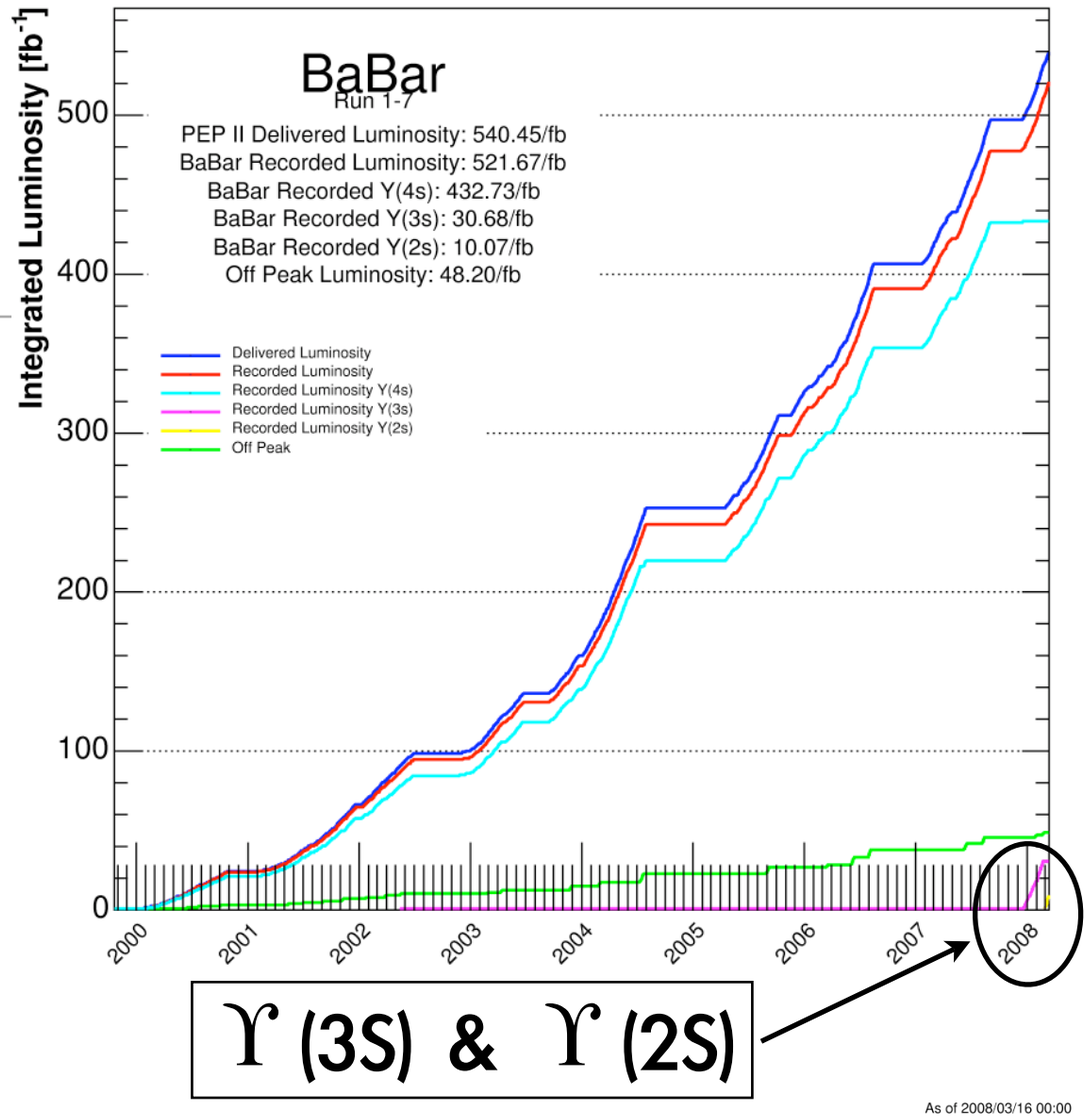
# Dalla presa dati alla pubblicazione @ *BaBar*

Alfio Lazzaro
Milano Atlas Meeting
18 Marzo 2008

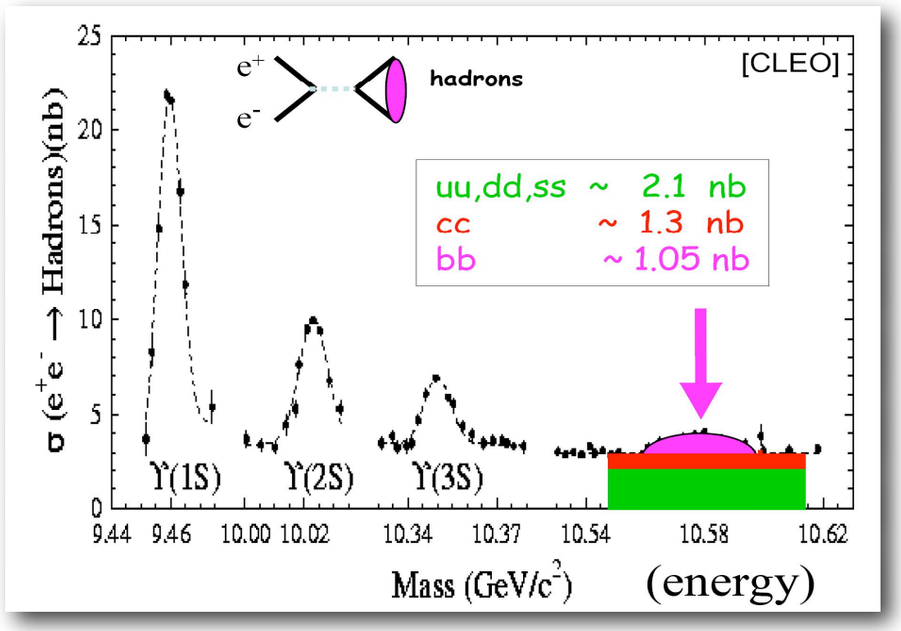
# BaBar @ SLAC

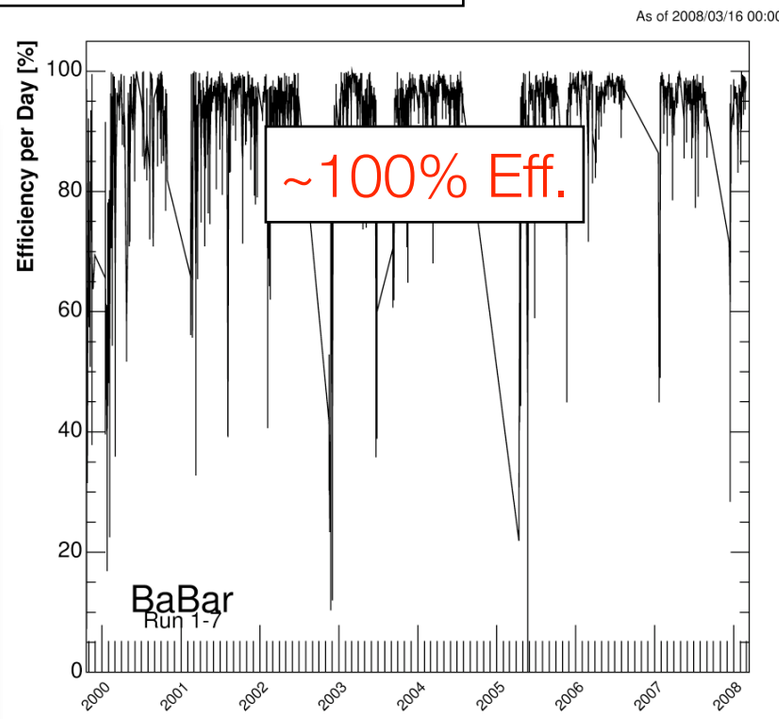
- Operativo dal 1999
- Energia nel CM: ~10.58 GeV



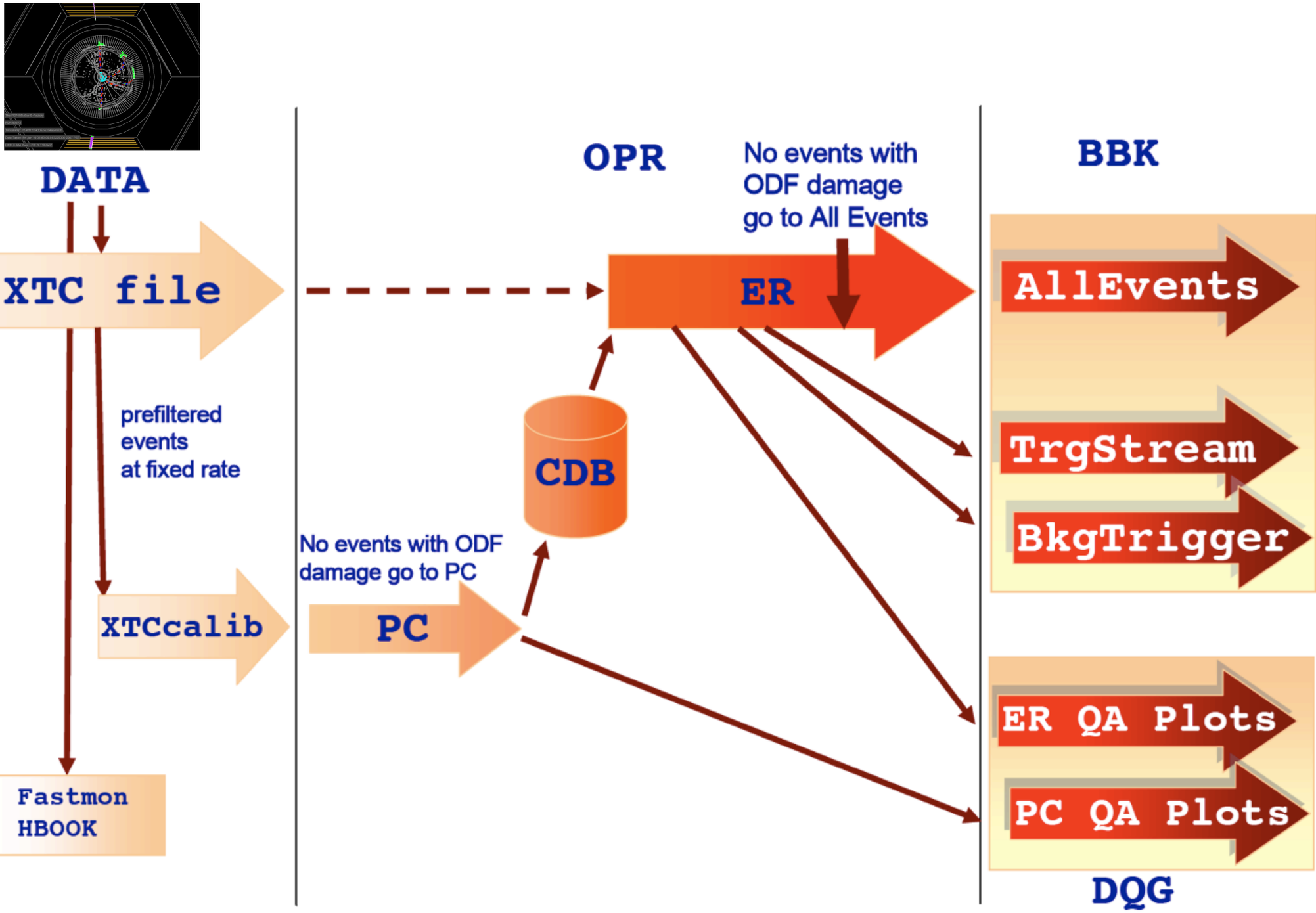




- Ambiente molto pulito:

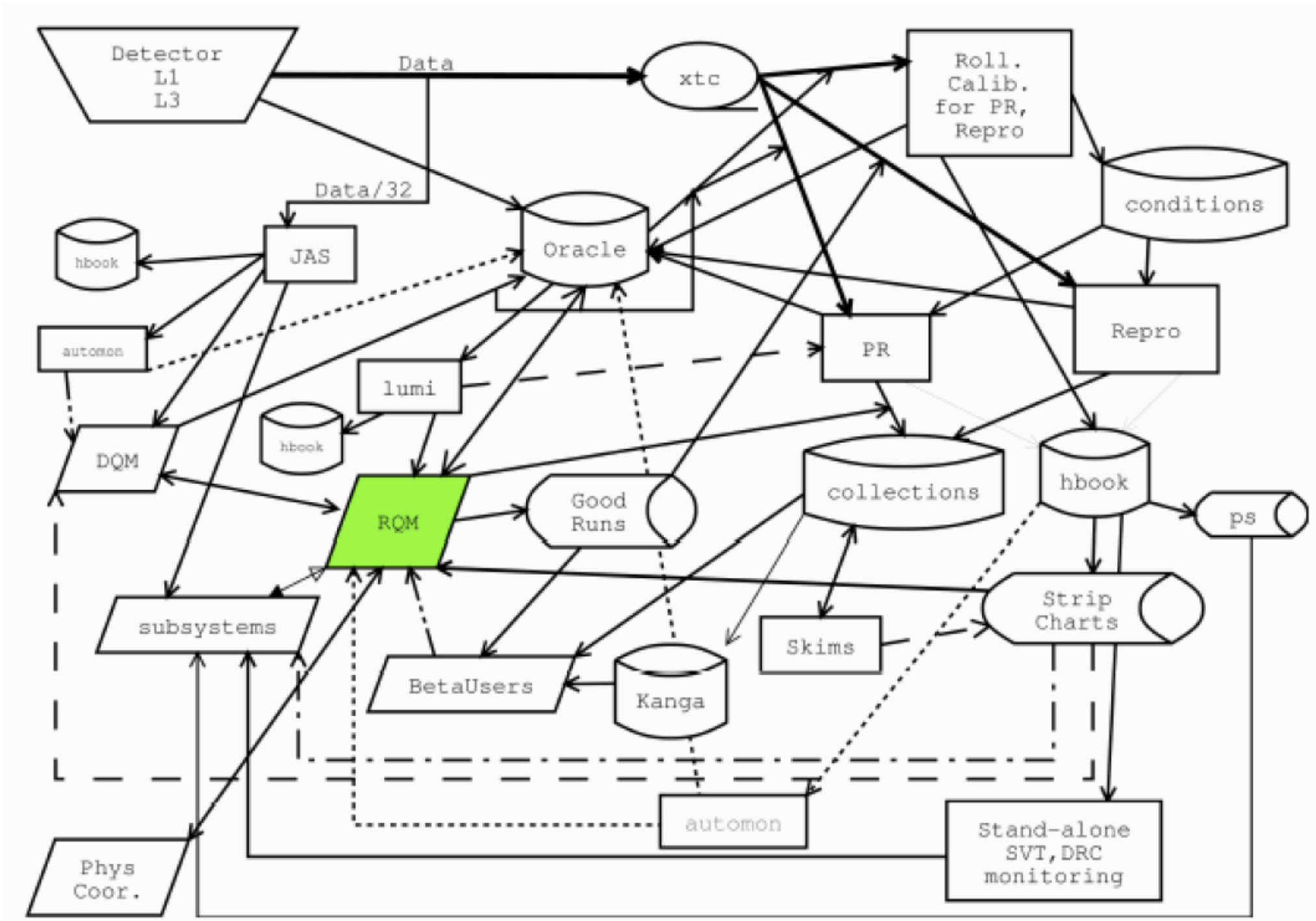


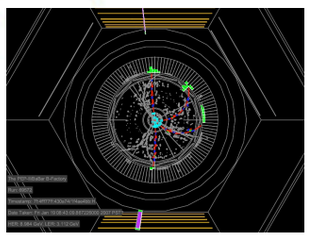
# I dati: dall'acquisizione al riprocessamento (mondo ideale)



# I dati: dall'acquisizione al riprocessamento (odissea reale)



# I flags



After 2 hours, if the run is Colliding beams and the XTCcalib has >700 events

- ✓ Good
- Flawed
- ✘ Unusable
- Detector Study

Only upon request for special purposes and usually only ER is done in order to preserve CDB integrity

OPR

QA files

DQG

- ✓ Good
- Flawed
- ✘ Unusable

Only collections with a good PC processing and a good ER processing done after the PC are checked

The DQG decides internally the criteria to define the global flag of each processing and, when necessary evaluates each single case.

They are also available in the DB while waiting for the DQG evaluation

BBK

- ✓ Good
- ✘ Unusable

Datasets

Data to Users

Only Good

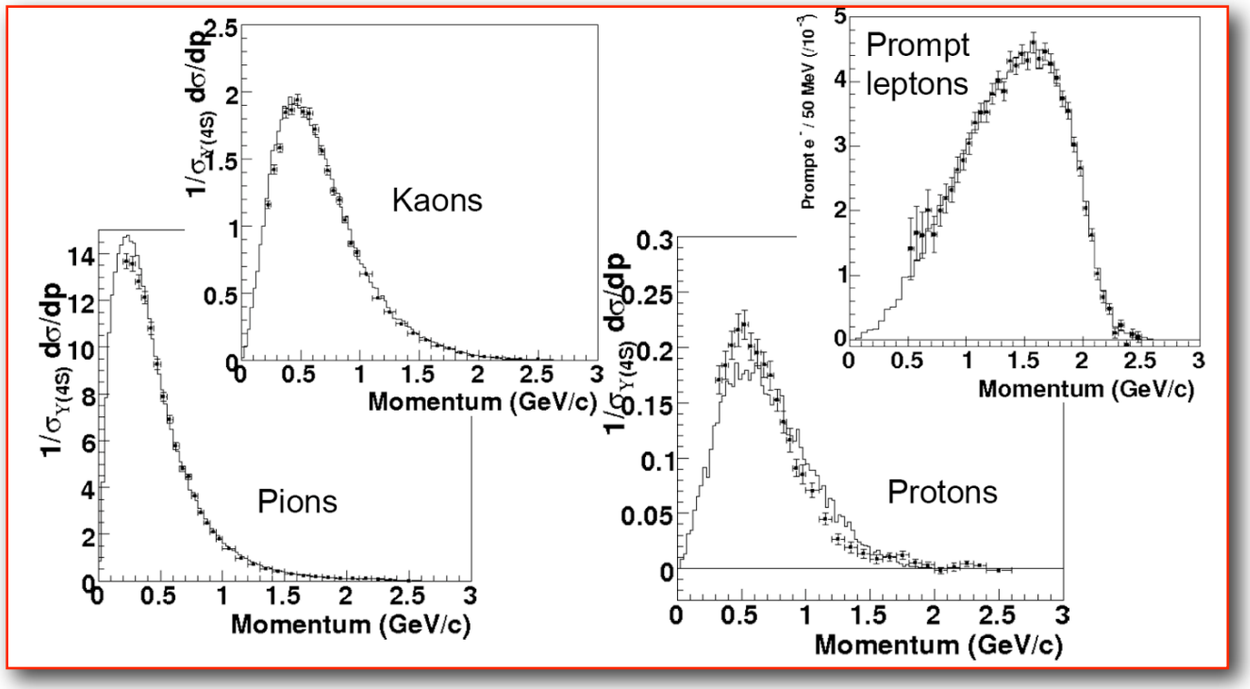
Skimming

A good collection must be flagged GOOD, be Colliding beams and Physics type

# La simulazione

- Diversi generatori per i vari processi integrati nel BaBar-Software
- Basata su GEANT4.8
- I MC vengono prodotti inserendo le condizioni del detector per i vari periodi di presa dati (mensilmente)
  - Produzione di MC generici
  - Produzione di MC esclusivi
- Validazione basata su confronto con control-sample di dati
  - Attualmente l'accordo è a livello del 95%-99% in molti processi
- Successivamente gli eventi MC seguono la stessa validazione usata nei dati

| | |
|---|---|
| EvtGen | B physics |
| KK2F | ISR + 2 fermion |
| Tauola | tau physics |
| JETSET | AKA Pythia, inclusive production of B and udsc |
| AfkQed | ISR + leptons or hadrons |

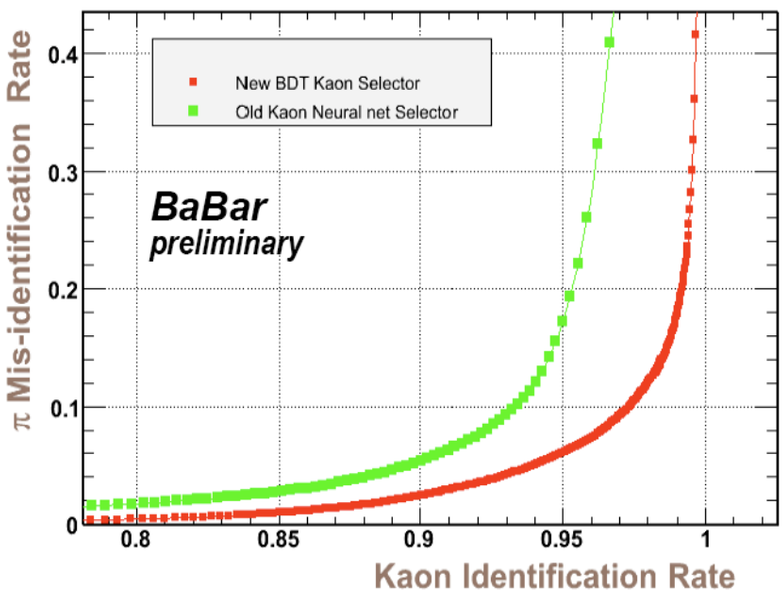


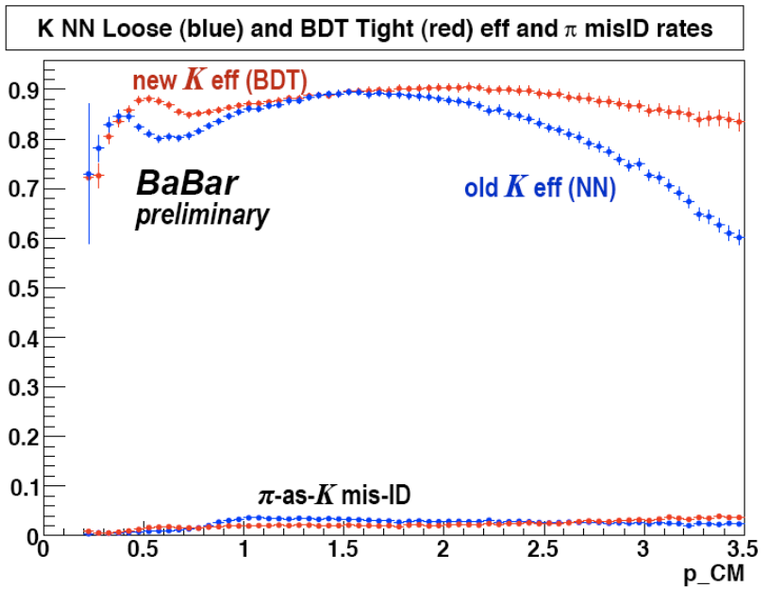
# La release di ricostruzione

- La release contiene tutto il codice dell'esperimento
  - Vari pacchetti in CVS (~1000)
    - Ogni pacchetto ha almeno un coordinatore
  - Si distinguono i pacchetti per l'online, offline, analisi fisica
    - Un super-coordinatore per ognuno dei tre gruppi
- Le release vengono aggiornate continuamente, introducendo nuovi tag dei pacchetti
  - fast build delle release giornalmente
  - full build circa una volta a settimana
- Test di validazione per scegliere la release (non necessariamente la stessa) per
  1. acquisizione
  2. ricostruzione dei dati
  3. simulazione
  4. riprocessamento dati
  5. skimming
  6. analisi fisica

# Riprocessamento dei dati

- Necessario quando la release incorpora significanti miglioramenti
  - Ultimo riprocessamento effettuato nel 2007, migliorando le efficienza di ricostruzione delle tracce
    - Aumento di efficienze nei segnali ~10%, diminuzione dei fondi ~10%
    - Impiegati circa 8 mesi in vari siti (SLAC, RAL, Padova)
- Attualmente è in corso il riprocessamento finale (The Final Game), atteso il completamento per Dicembre 2008
  - Migliorato tutto il codice per il PID, usando Boosted Decision Tree
  - Attualmente è in corso la validazione

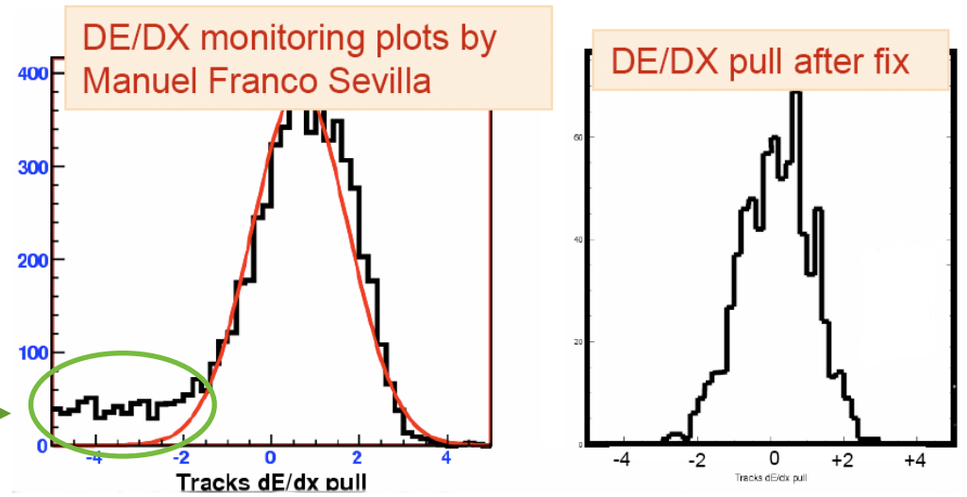


# Data-quality

- I dati riprocessati vengono accuratamente validati, confrontando con dati ricostruiti con precedenti release
  - Analizzate varie variabili legate ad ogni sub-detector
    - Un esperto per ogni sub-detector settimanalmente fornisce feedback sui vari dati riprocessati

Problema trovato nel PID (2007) ⟶



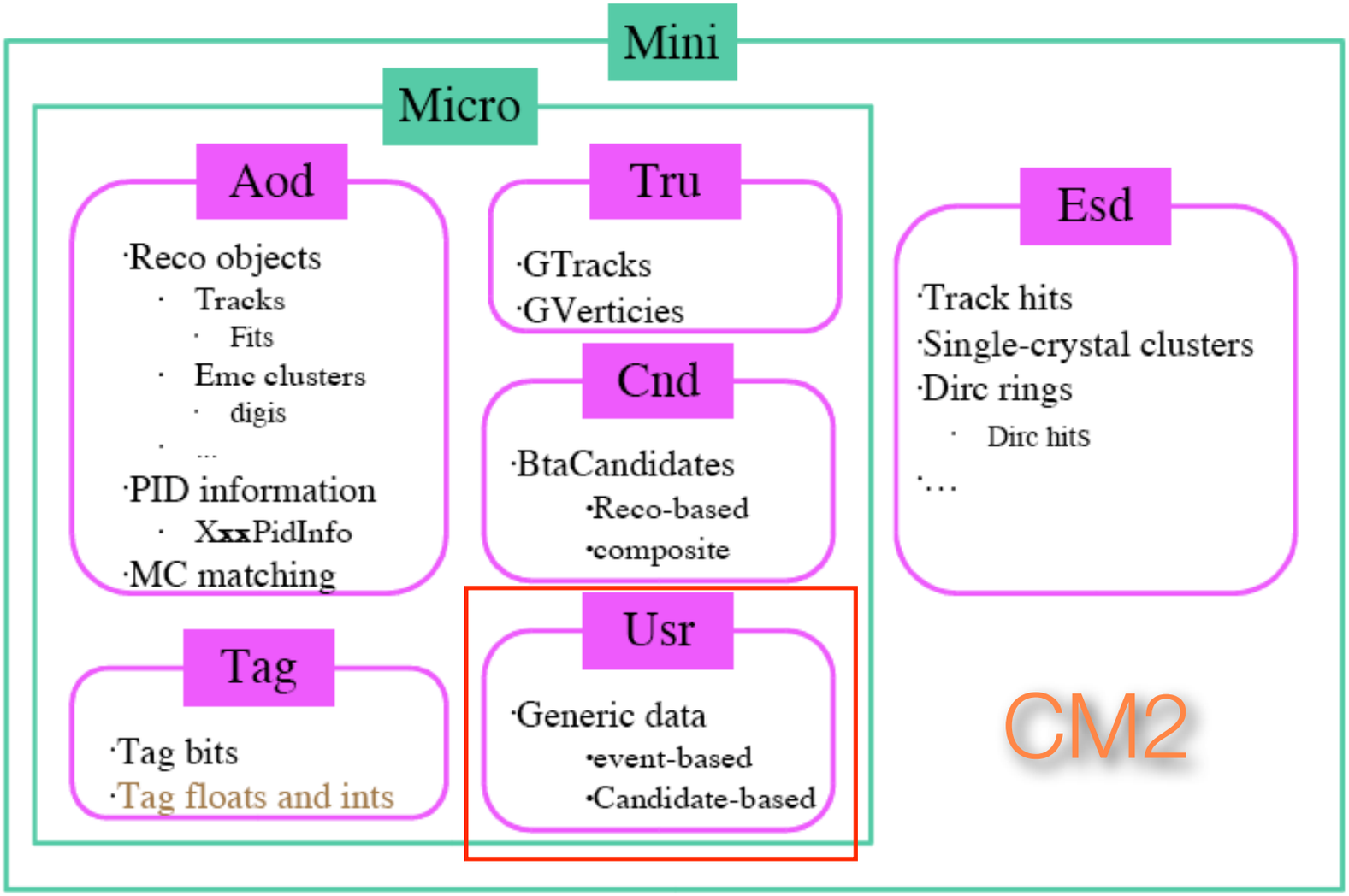

- Analizzate alcune variabili di fisica: numero di fotoni, carichi, quantità legate alle risonanze,..., ma non analisi complete
- Analisi campione complete di particolari processi ($B \to D^*D^*$, $B \to K^*\gamma$), vengono eseguite su un sample ridotto di dati (validazione 10 $fb^{-1}$)
- Solo per nuovi riprocessamenti si effettua una validazione su un più ampio dataset (validazione 60 $fb^{-1}$) per un gruppo maggiore di analisi

# Skims

- I dati che superano la validazione vengono raccolti in tag per l'analisi fisica successiva
- Normalmente ogni analisi necessita di un set limitato di dati, ~3% del campione totale
  - vengono costruiti dei filtri per una preselezione degli eventi
    - Esempio analisi $B^0 \rightarrow \eta'$ K: si chiede che nell'evento siano presenti almeno una $\eta'$ e una K per formare un *B* con un certo taglio sull'impulso nel CM
  - ad ogni evento vengono assegnate le uscite dei tagbit che corrispondo all'uscita dei filtri
    - I dati con un certo set di tagbit vengono raccolti in stream (copie fisiche o copie virtuali dei dati)
    - Le stream vengono rese disponibili nei vari siti per l'analisi fisica: SLAC, RAL, IN2PR, CNAF, GRIDKA, ...

# Organizzazione delle informazione nei dati

- Standard file ROOT, con tutte le informazioni (Mini), tranne i Trigger Filter
- Gli analisti possono inserire le loro informazioni di fisica direttamente durante lo skimming dei dati (Usr Data)
  - Rapida ricostruzione per l'analista




- È possibile un riprocessamento dei dati direttamente da mini (mini-to-mini conversion) se non e' richiesto di cambiare i Trigger Filter
- Diverse centinaia di skim prodotte

# Physics Analysis

- La release per la ricostruzione ed analisi degli eventi usata dagli analisti
- Vengono verificati i pacchetti per
  - Tracking and Neutral reconstruction
  - PID
  - Particle Composition
  - Vertexing
  - *B*-Tagging
  - Fitting tools
- I dati vengono selezionati con tagli preliminari
- Le variabili di interesse per l'analisi (impulsi, posizioni, masse, energie, PID, ...) vengono salvate in root-tree
- La ricostruzione si spezza in diversi jobs che runnano parallelamente nella farm, ognuno producendo un root-tree di uscita

# Root-tree esempio

- BRecMode
- BVertexChi2
- BVertexCovZ
- BVertexNDof
- BVertexProb
- BVertexStatus
- BVertexZ
- BcategoryTag04
- Bcharge
- BtagTag04
- DE
- DE_Err
- KsCtau2
- KsCtauErr2
- KsDeltaVtxIP2
- KsDistanceXY2
- KsEnergy2
- KsFlightLength2
- KsFlightLengthXY2
- KsKinkAngle2
- KsMass2
- KsMomentum2
- KsMomentumCMS2
- KsPionChi212
- KsPionChi222
- KsPionNDchHits12
- KsPionNDchHits22
- KsPionNSvtHits12
- KsPionNSvtHits22
- KsPionSvtPattern12
- KsPionSvtPattern22
- KsPionnDof12
- KsPionnDof22
- KsPionprobChi212
- KsPionprobChi222
- KsPionstartFoundRange12
- KsPionstartFoundRange22
- KsPiontrackLength12
- KsPiontrackLength22
- KsPx2
- KsPy2
- KsPz2
- KsVertexProb2
- Leg0
- Leg2
- R2all
- bEnergy
- bMass
- bMass_Err
- bPx
- bPy
- bPz
- combMassReco10
- combMassReco20
- combMassReco21
- cosBBeam
- cosBTBeam
- cosBTROE
- energyCMS
- energyHER
- energyLER
- etaCharge1
- etaCosHel1
- etaEnergy1
- etaEnergyCMS1
- etaMass1
- etaMassReco1
- etaMomentum1
- etaMomentumCMS1
- etaPi0Veto1
- etaPx1
- etaPy1
- etaPz1
- evtNum
- gamDistGams3
- gamDistTrks3
- gamEnergy11
- gamEnergy21
- gamEnergy3
- gamEnergyB3
- gamEnergyCMS3
- gamLat3
- gamNBumps3
- gamNXtal3
- gamPhi3
- gamPi0EtaVeto3
- gamPxCMS3
- gamPyCMS3
- gamPzCMS3
- gamR2prime3
- gamSecondMoment3
- gamStatus3
- gamTheta3
- lowerID
- mES
- mES_Err
- mHat
- mHat_Err
- mMiss
- mMiss_Err
- mcBRecID
- mcBRecMode
- mcBRecZ
- mcBTagID
- mcBTagMode
- mcBTagZ
- mcIsCharmed
- mcMatchInfoDiff
- mcMatchInfoNoMatch
- mcMatchMode
- mcNBDauGamma0
- mcNBDauGamma1
- mcNBDauK0
- mcNBDauK1
- mcNBDauKL0
- mcNBDauKL1
- mcNBDauMu0
- mcNBDauMu1
- mcNBDauOther0
- mcNBDauOther1

# I dati a Milano: Selezione degli eventi

- I root-tree prodotti vengono copiati a Milano
- Concateniamo i vari file in catene (classe TChain)
- Per la selezione finale degli eventi usiamo un selettore (classe TSelector)
  - Abbiamo esteso le funzionalità della classe TSelector
    - Riutilizzare il codice più possibile
    - Tutto (o quasi) è configurabile senza modificare il codice, tramite opzioni
  - Ogni analisi ha il suo proprio selettore con specifici tagli
- L'ottimizzazione dei tagli, il calcolo delle efficienze, la stima del fondo sono tutte implementate all'interno del selettore
- L'uscita del selettore è interfacciata con un foglio di calcolo (XML) per calcoli successivi e consente di scrivere direttamente tabelle LaTex (Pietro B.)
- Gli eventi che superano i tagli del selettore vengono salvate in un root-tree per l'input al fit Maximum Likelihood per l'estrazione dei risultati finali

# Selettore esempio

```
if (cos(gamTheta3)<-0.74 || cos(gamTheta3)>0.93) return kFALSE;
coreSelectorObj().CountCandidates("gamTheta");

if (gamStatus3!=0) return kFALSE;
coreSelectorObj().CountCandidates("gamStatus");

if (gamNXtal3<5) return kFALSE;
coreSelectorObj().CountCandidates("gamNumXtal");

if(gamDistGams3<25) return kFALSE;
coreSelectorObj().CountCandidates("Isolat. from gammas");

if(gamDistTrks3<25) return kFALSE;
coreSelectorObj().CountCandidates("Isolat. from tracks");

if (vetoCandMassReco_myPi0LooseMassVeto3>0.120 &&
    vetoCandMassReco_myPi0LooseMassVeto3<0.150)
  return kFALSE;
coreSelectorObj().CountCandidates("vetoPi0forGamma(default)");
```

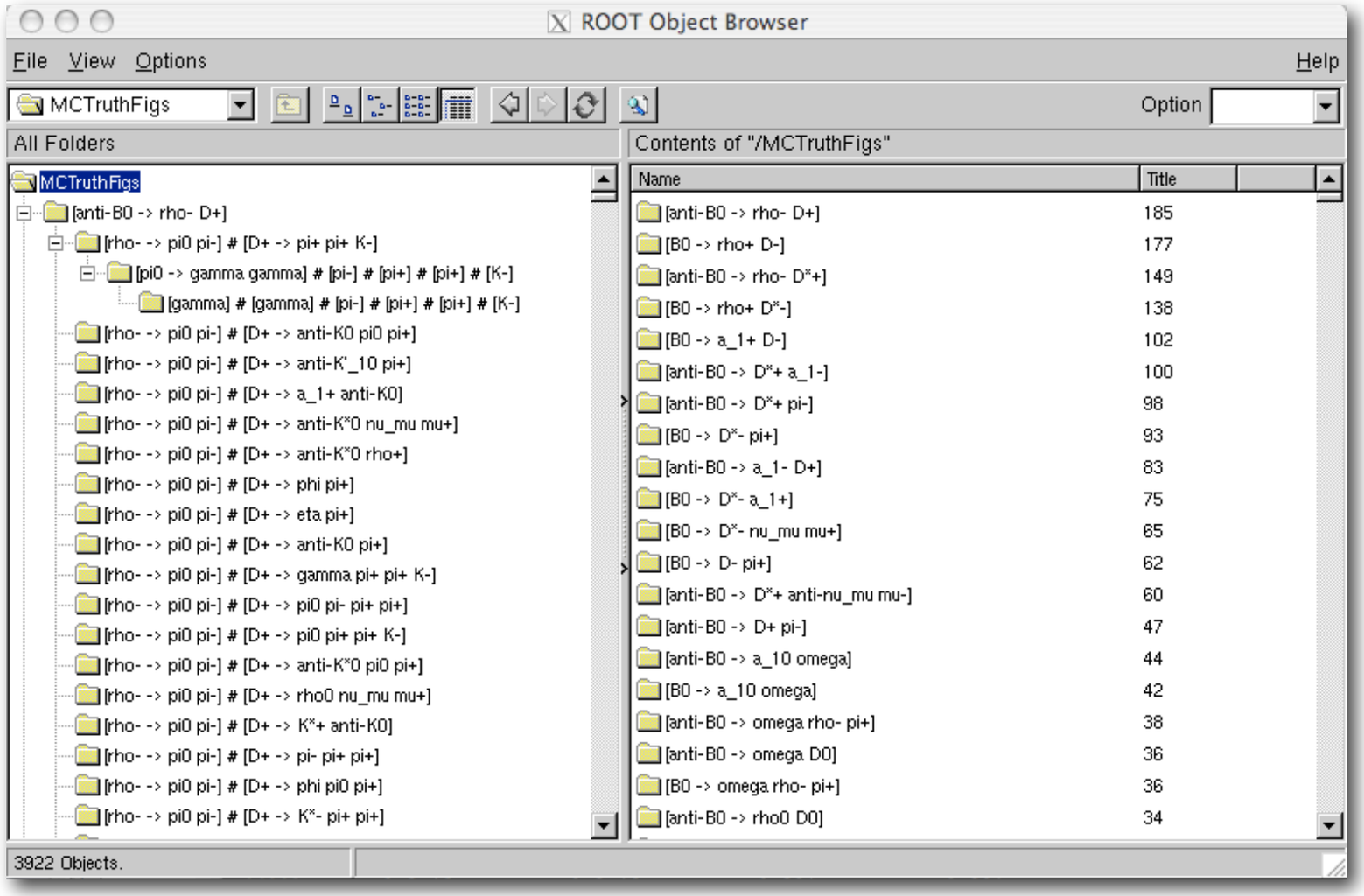
```
pcbabar6:analysis lazzaro$ root EtaGGKsGamma_5532_AllEventsSkim-R22d.root

  ********************************
  *                              *
  *     ROOT Version 5.18/00     *
  *   Compiled by Alfio Lazzaro  *
  *                              *
  ********************************

...using style 'Plain'
Loading libraries...
RooFit v2.31 -- Developed by Wouter Verkerke and David Kirkby
                Copyright (C) 2000-2008 NIKHEF, University of California & Stanford University
                All rights reserved, please read http://roofit.sourceforge.net/license.txt

Instance New CoreSelector object.
Use global pointer gCoreSelector.

Inizialize MiMacros::DrawVars macros.

Done!

    For approved plots use: gROOT->SetStyle("BABAR");
  To add a BABAR label use: BABARLabel();
  To add a better-scaling BABAR label use: BABARSmartLabel();
    type "BABARSmartLabel(-2);" for options

root [0]
Attaching file EtaGGKsGamma_5532_AllEventsSkim-R22d.root as _file0...
root [1] ntp1->Process("selector_B0ToEtaGGKsGamma.C+","w")
```

Opzione

```
>>>>>>>>>>>>>>>>>>>>>>>>>>>

Running Option: Unblind FitData Table

************ EtaGGKsGamma_5532_AllEventsSkim-R22d ************


Cut                          & Abs-Cands (%) & Rel-Cands (%) & Abs-Evts (%)  & Rel-Evts (%)
reconstruction               & 383903        & 383903        & 203132        & 203132
preliminary                  & 86.0          & 86.0          & 95.1          & 95.1
gamEnergyCMS                 & 81.0          & 94.3          & 94.0          & 98.8
secondMoment                 & 80.0          & 98.7          & 92.9          & 98.9
gamTheta                     & 78.4          & 98.0          & 91.1          & 98.1
gamStatus                    & 78.4          & 100.0         & 91.1          & 100.0
gamNumXtal                   & 78.4          & 100.0         & 91.1          & 100.0
Isolat. from gammas          & 73.8          & 94.1          & 86.3          & 94.7
Isolat. from tracks          & 72.3          & 98.0          & 84.7          & 98.1
vetoPi0forGamma(GammaConv)   & 72.3          & 100.0         & 84.7          & 100.0
vetoPi0forGamma(default)     & 69.0          & 95.5          & 81.4          & 96.1
vetoEtaforGamma(default)     & 65.5          & 94.9          & 78.1          & 96.0
eta mass                     & 57.0          & 87.1          & 74.8          & 95.7
pi0Veto Eta                  & 52.6          & 92.1          & 70.4          & 94.2
K0S cuts                     & 43.5          & 82.8          & 61.0          & 86.7
\etagg Decay Angle           & 37.1          & 85.2          & 56.5          & 92.6
pEtaCMS                      & 24.3          & 65.5          & 40.3          & 71.3
vetoEtaggEne                 & 23.7          & 97.6          & 39.4          & 97.7
vetoPi0Ene                   & 23.0          & 96.8          & 38.2          & 97.0
B Primary Photon Energy      & 22.9          & 99.7          & 38.1          & 99.8
Final                        & 87887         &               & 77345         &

<<<<<<<<<<<<<<<<<<<<<<<<<<<
```

# Stima del fondo

- Ricostruzione degli eventi MC di fondo generici
- Salviamo le informazioni relative alla MC generazione di tutti gli eventi ricostruiti
  - Usiamo questa informazione per capire che canali danno fondo dopo la selezione con il selettore
- Studi più approfonditi sono successivamente fatti con MC esclusi



# Soppressione del fondo

- Analisi Multivariate impiegando varie tecniche
  - Ottimizzazione dei tagli Multivariata
  - Discriminante di Fisher
  - Reti Neurali
  - Bagging & Boosted Decision Tree
- Pacchetti usati:
  - **TMVA -- Toolkit for Multivariate Data Analysis:** General framework for Multivariate data analysis

    http://tmva.sourceforge.net/
  - **SPR -- StatPatternRecognition:** General framework for Multivariate data analysis

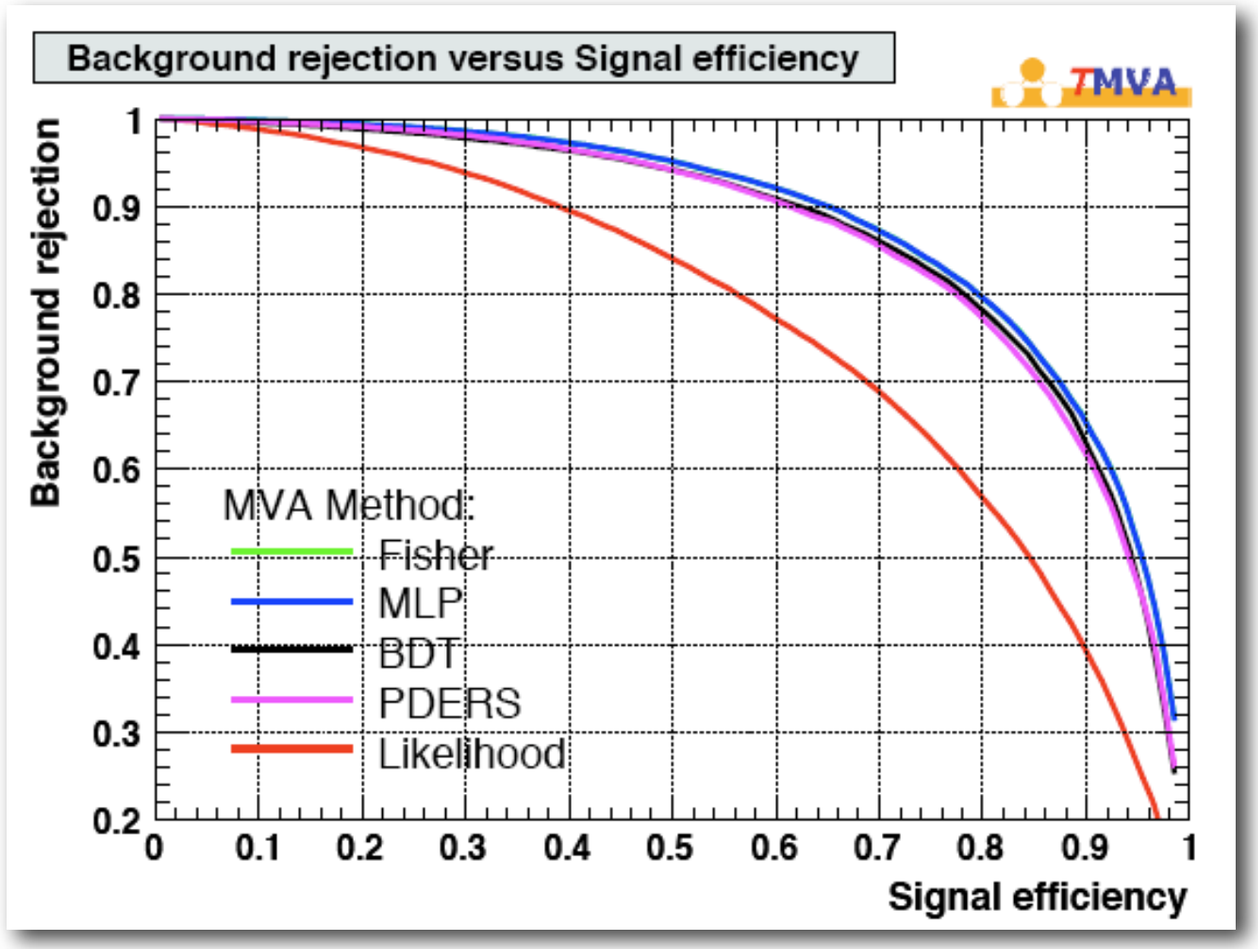
    http://sourceforge.net/projects/statpatrec
- Tecniche molto potenti per la separazione segnale/fondo, comunemente usate nel gruppo di Milano (e in BaBar in generale)

# Tecniche avanzate di soppressione segnale/fondo

- Recentemente ho dato un talk alla iCSC08 al CERN sull'argomento
  http://csc.web.cern.ch/CSC/2008/iCSC2008/iCSC2008.htm
- Riferimento bibliografico:

  **Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning - Data Mining, Interference, and Prediction", Springer Series in Statistics (2003)**

- I software sono facili da utilizzare, ben documentati, integrati in ROOT
  - Facile fare confronti tra le varie tecniche



# Fit Maximum Likelihood

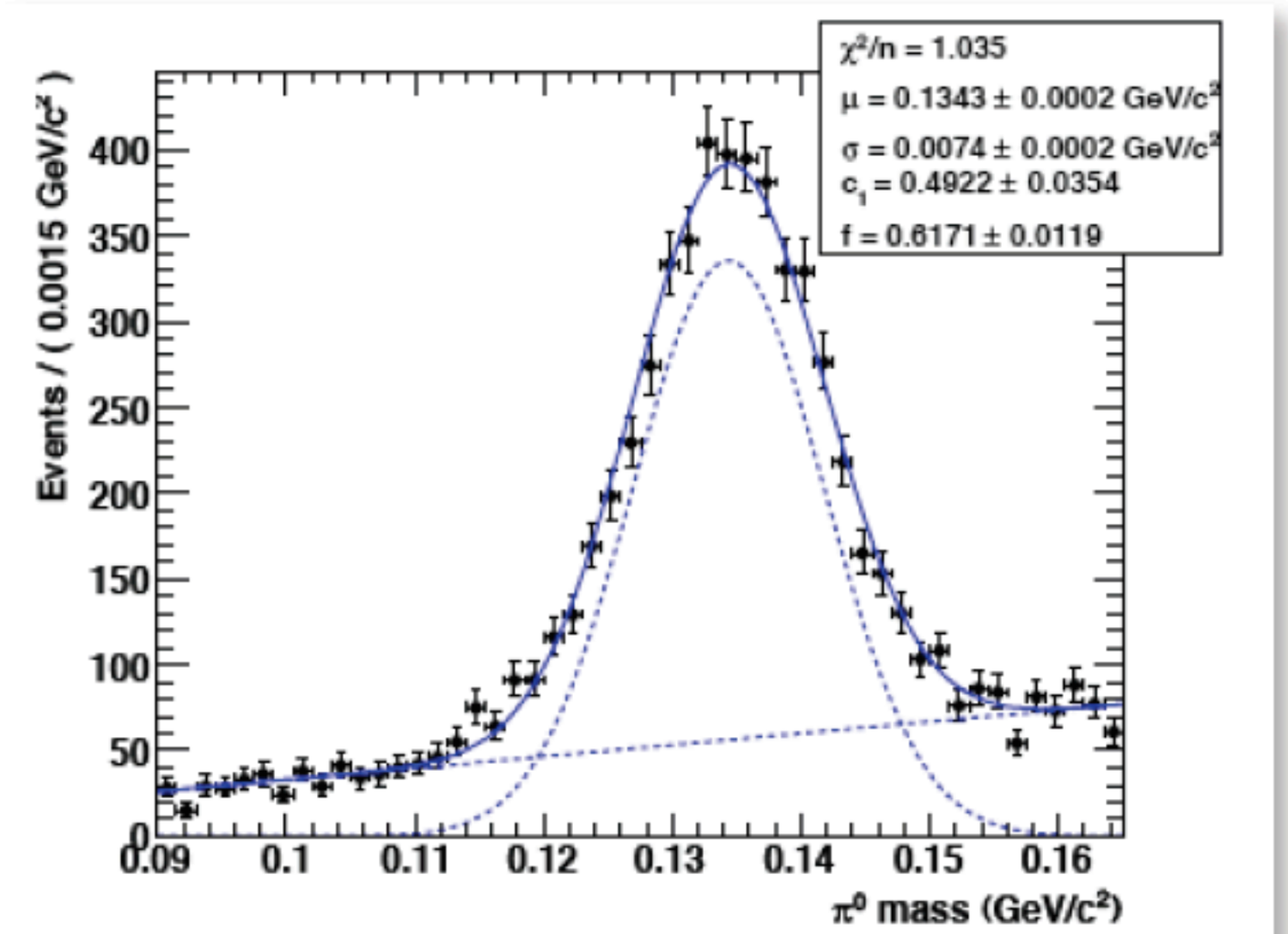
- Fit Multivariati Unbinned basato sul pacchetto RooFit
  - Integrato in ROOT
  - Esperienza con RooFit fin dagli inizi del suo sviluppo (entrambi gli autori, D. Kirkby e W. Verkerke, sono/erano membri di BaBar)
  - Praticamente usato in tutte le analisi di BaBar: grande esperienza accumulata nell'esperimento
  - Recente workshop tenuto a SLAC a Dicembre

http://www.slac.stanford.edu/BFROOT/www/doc/Workshops/2007/BaBar_RooFit/Agenda.html

- Comunque RooFit è pur sempre un insieme di classi C++ e spetta all'utente dover "lavorare" con il codice per scriversi il proprio programma
  - Poco versatile: se le analisi hanno aspetti in comune, è possibile pensare ad un sistema di scripting per il "riciclo" del codice
  - Basso livello: RooFit fornisce un set di classi base, non sufficienti per effettuare proiezioni, NLL scan, Upper Limit, esperimenti MC simulati, ...
- Serve qualcosa di più generale che permetta di configurare ROOT & RooFit

# MiFit

- Interfaccia ROOT & RooFit
- Interamente configurabile da semplici file di testo
  - Non serve alcuna conoscenza approfondita del codice
  - Costrutti superiori per la dichiarazione delle varie operazioni
- Generalizzato a qualsiasi tipo di analisi ML unbinned




- Attualmente il collo di bottiglia di molte analisi in BaBar è diventato il tempo di esecuzione dei fit: analisi Dalitz-plot a molti parametri liberi e molti eventi impiegano ore (giorni)
  - Diventa molto interessante usare tecniche di calcolo parallelo, attualmente mio oggetto di studio

# Coclusione

- L’esperienza di BaBar dice che è fondamentale avere un controllo attento dei dati e di tutto il codice di ricostruzione e di analisi fisica
- Particolarmente importante quando si ha un esperimento concorrente: Belle
- Tecniche avanzate consentono a BaBar di restare competitivo (stessa significatività) con Belle malgrado loro abbiano l’80% in più dei dati rispetto all’attuale campione di BaBar
  - particolarmente importante considerato che BaBar smetta la presa dati alla fine di questo mese, mentre Belle prosegue per almeno altri 2 anni

# Fine del talk