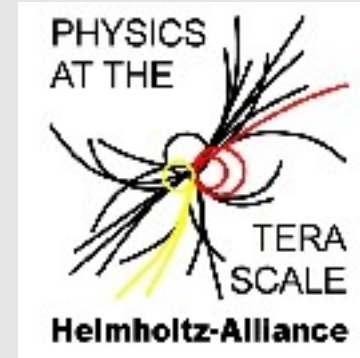




# BAT

## A Bayesian Analysis Toolkit



A. Caldwell<sup>1</sup>, D. Kollár<sup>1</sup>, **K. Kröninger<sup>2</sup>**

<sup>1</sup>*Max-Planck Institut für Physik, München*

<sup>2</sup>*II. Physikalisches Institut, Universität Göttingen*

- Aims of data analyses:
  - Compare data and models
  - Judge validity of models
  - Estimate parameters

$$p(\lambda|\vec{x}) = \frac{p(\vec{x}|\lambda)p_0(\lambda)}{\int p(\vec{x}|\lambda)p_0(\lambda)d\lambda}$$

## BAT: Software package tailored for Bayesian analyses

- Provide flexible environment to phrase arbitrary problems
- C++ based framework (flexible, modular)
- Provide set of numerical tools
- Interfaces to ROOT, Minuit, CUBA + user defined

## User defined

### Definition of models

- Parameters
- Conditional probabilities
- Prior probabilities

### Data set definition

- Read in from file  
(.txt, .root, histogram,  
user defined)
- Create on-the-fly

## Common tasks

### Analysis

- Normalize (Integration)
- Optimize (Minimization)
- Marginalize (Sampling)
- Goodness-of-fit
- Model comparison
- Error propagation
- Write output file
- Graphical output

- **Integration:**

- Simple Monte Carlo algorithms (sampled mean, importance sampling)
- Interface to CUBA (VEGAS, ...)

- **Marginalization:**

- MCMC (Metropolis)

- **Error propagation:**

- Calculate any function of the parameters during a run

- **Optimization:**

- Monte Carlo (hit&miss)
- Metropolis (quasi-simulated annealing)
- Interface to Minuit

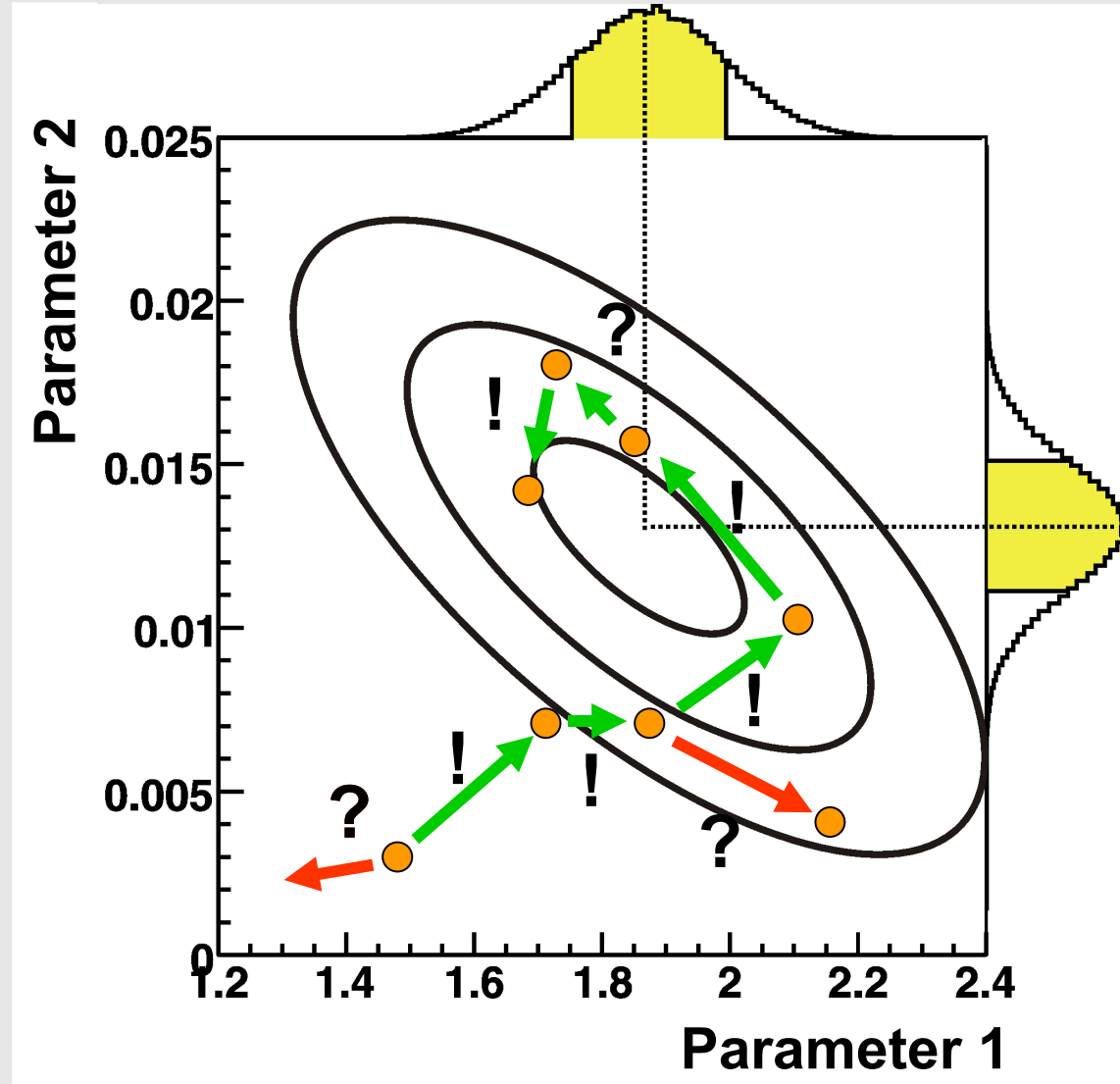
- **Goodness-of-fit:**

- Ensemble tests and p-value

**Key: Markov Chain  
Monte Carlo (MCMC)**

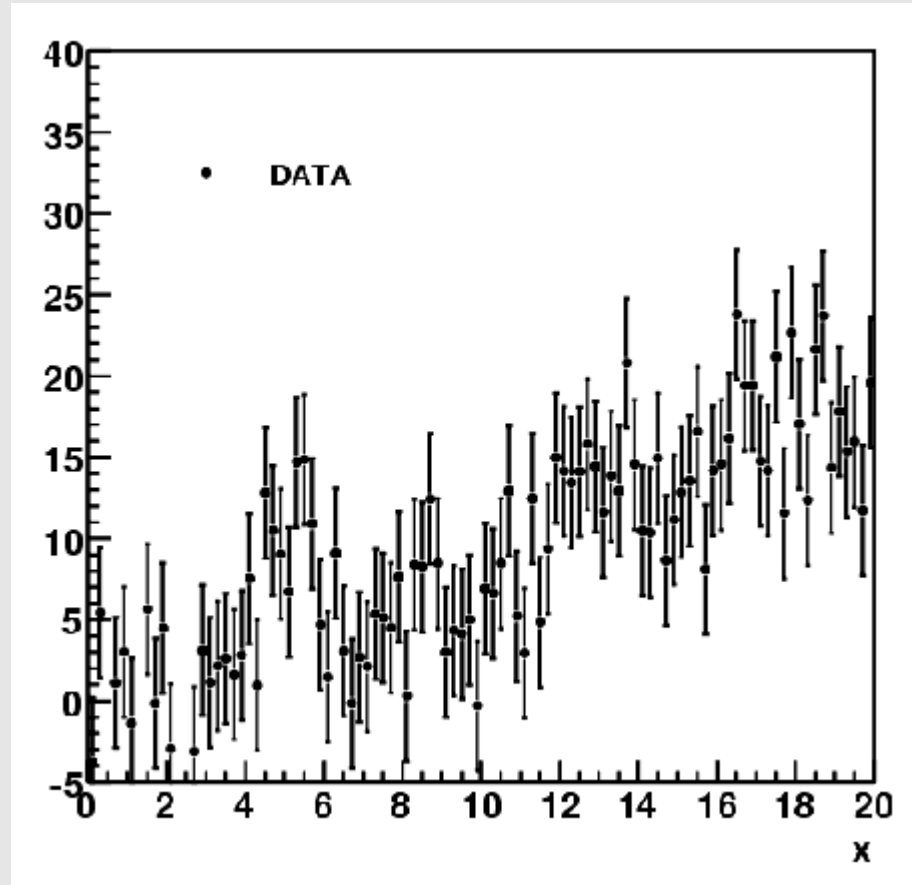
## Markov Chains:

- Sample parameter space by migrating to regions of larger probability
- Converge towards underlying distribution
- Marginalize pdf while walking
- Calculate any function of parameters while walking (error propagation)
- Find global maximum



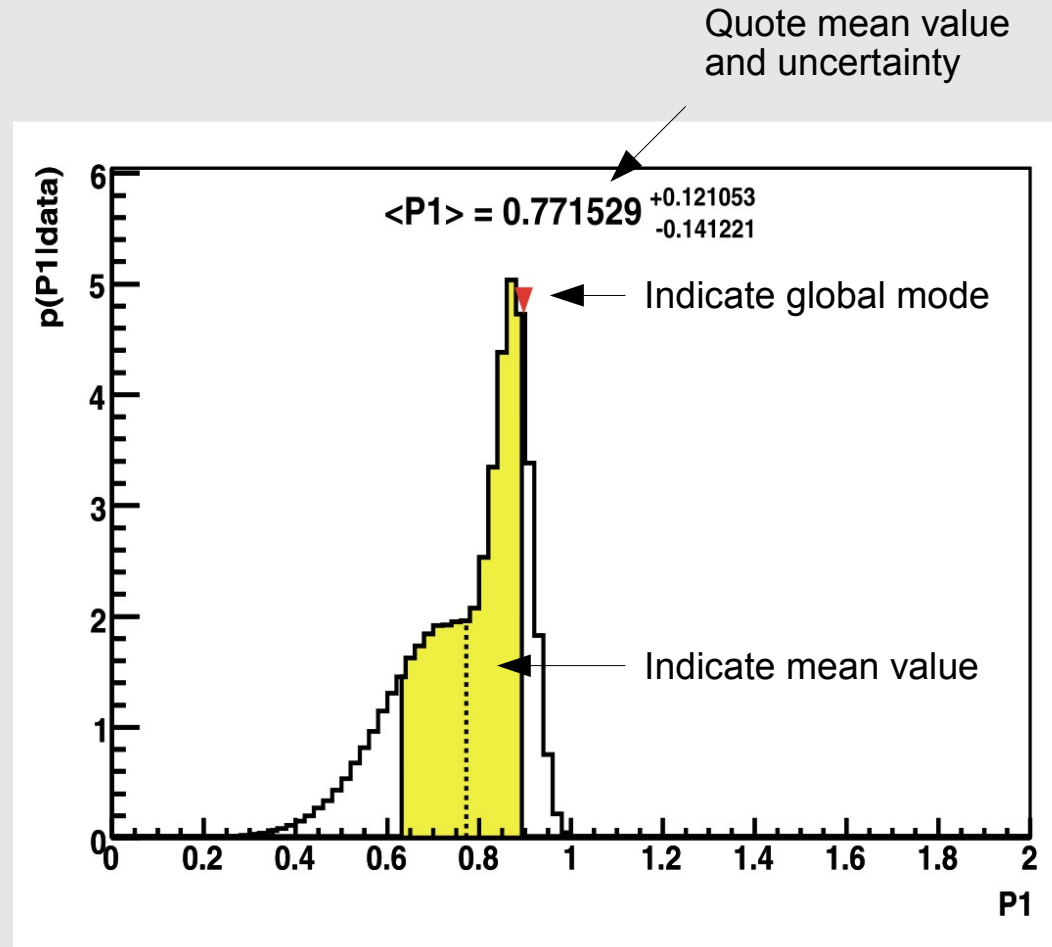
## Example for fitting:

- Generate spectrum with 2<sup>nd</sup> order polynomial plus Gaussian at  $x = 5$
- Test different hypotheses:
  - 2<sup>nd</sup> order polynomial
  - Gaussian plus constant
  - Gaussian plus line
  - Gaussian plus 2<sup>nd</sup> order polynomial
- Flat prior probabilities
- Peak position in  $[2, 18]$ , width  $< 4$



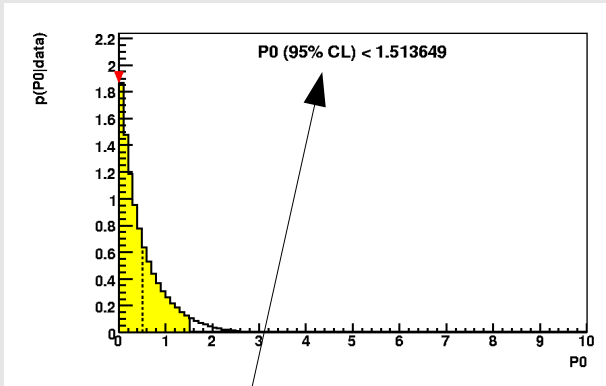
## Marginalized distributions:

- Project sampled distribution onto one (or two) parameters. Integrate over all others.
- Mode of marginalized distribution is in general not equal global mode. User's decision.
- Full information in Markov chain.
- Default output:
  - Global mode
  - Mean value
  - Central and smallest 68% interval

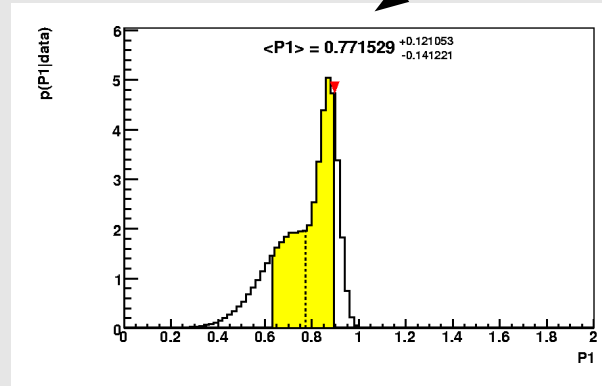


## Marginalized distributions (1-d)

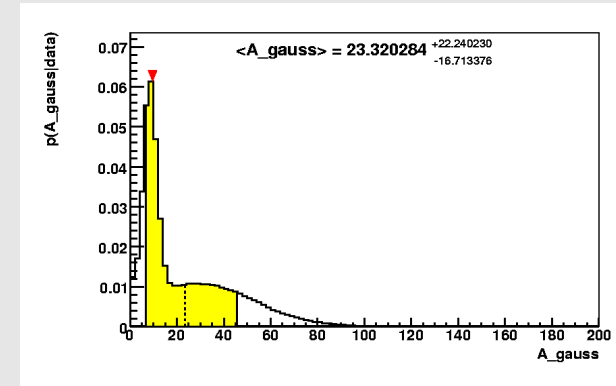
By default, calculate mean, mode and uncertainties



Offset

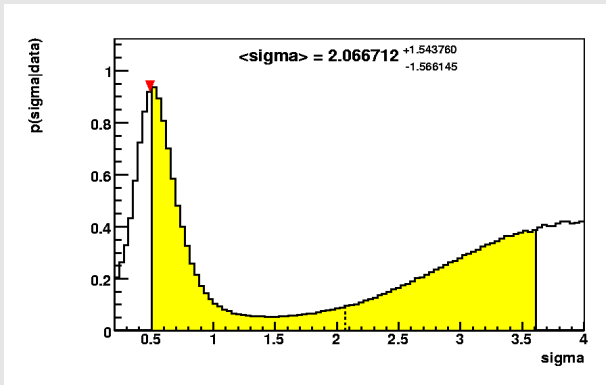


Slope

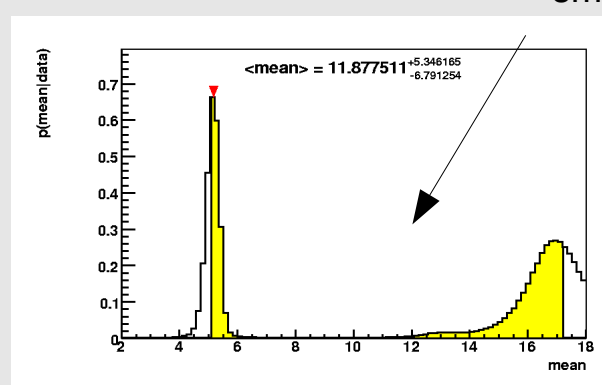


Central interval might not be appropriate, chose smallest interval

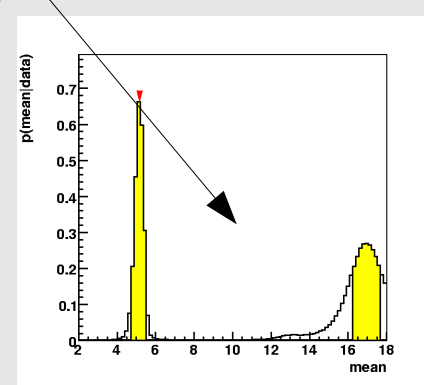
Peak strength



Peak width



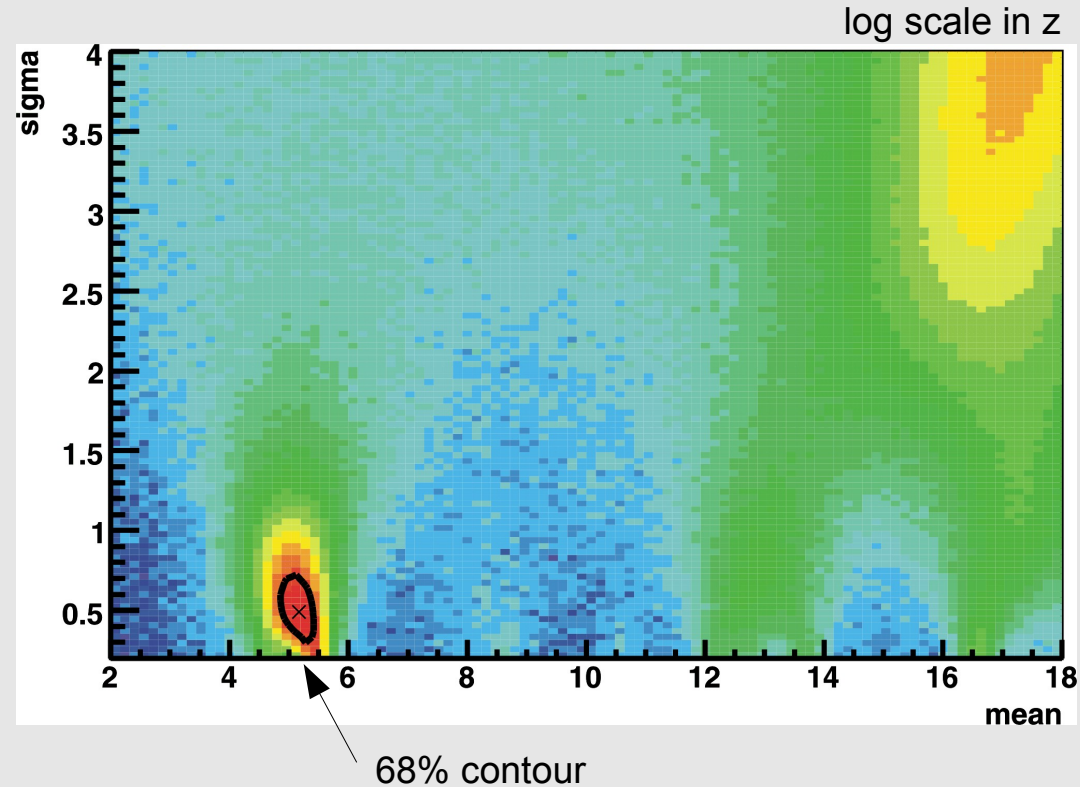
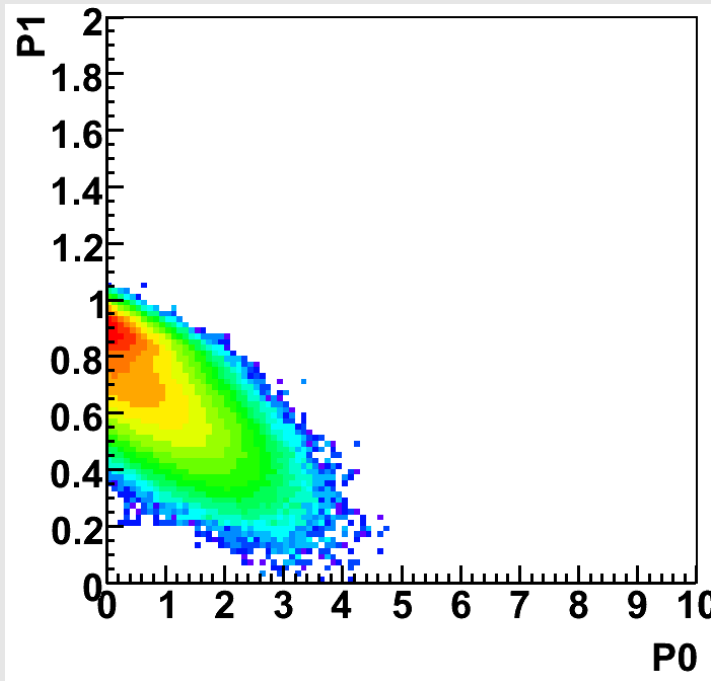
Peak position



Peak position

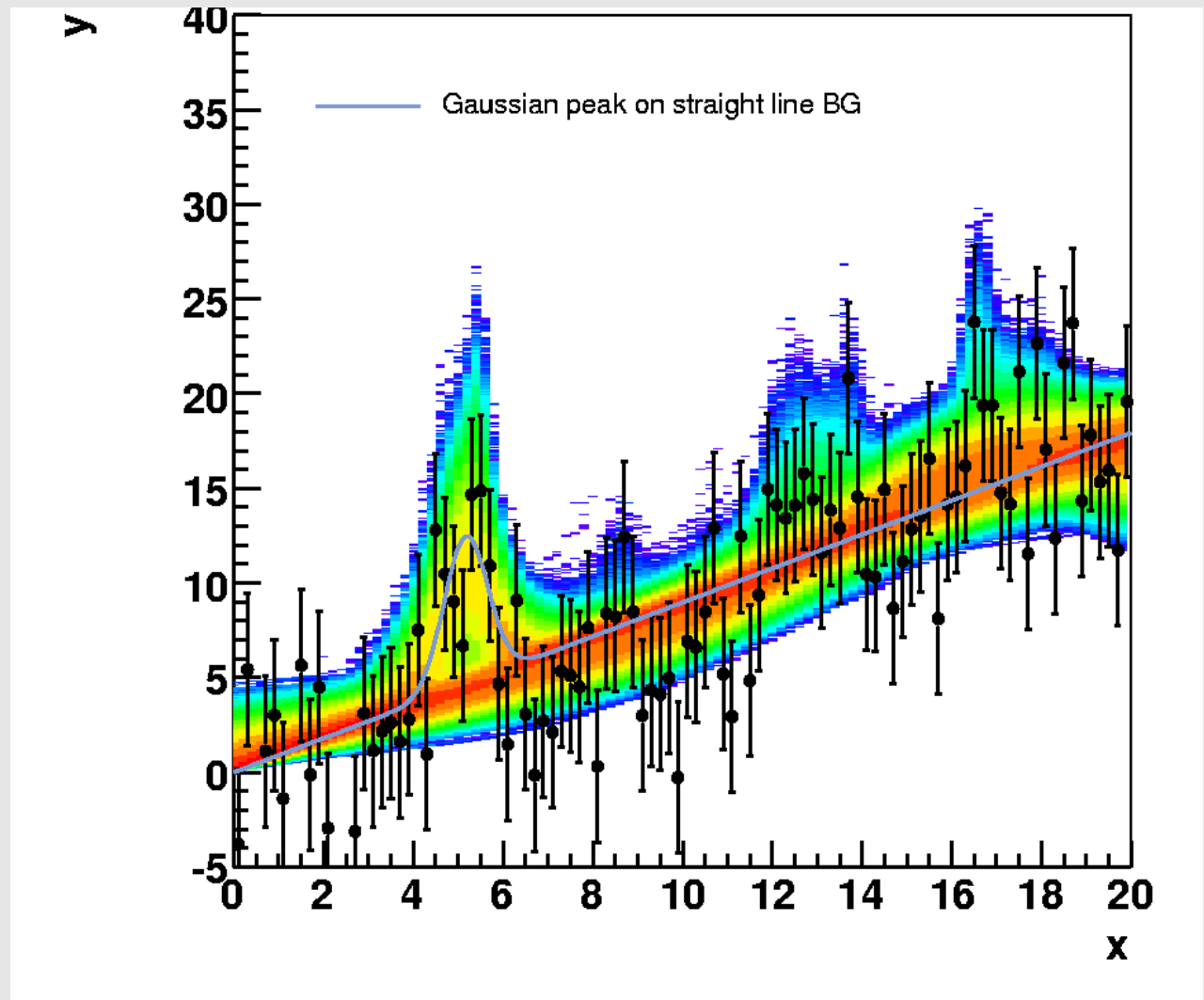


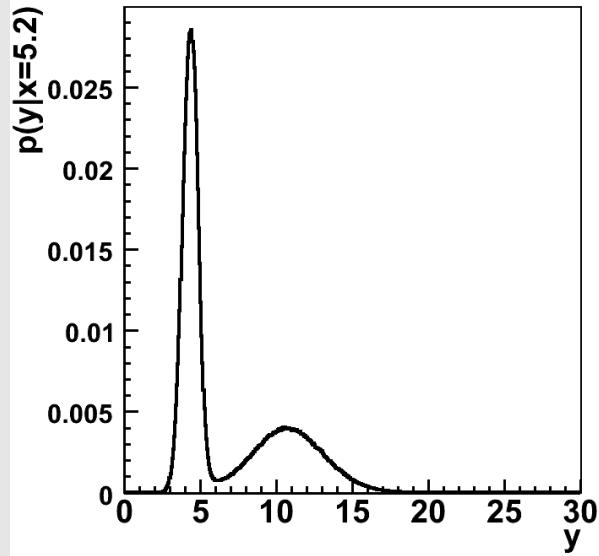
## Marginalized distributions (2-d)



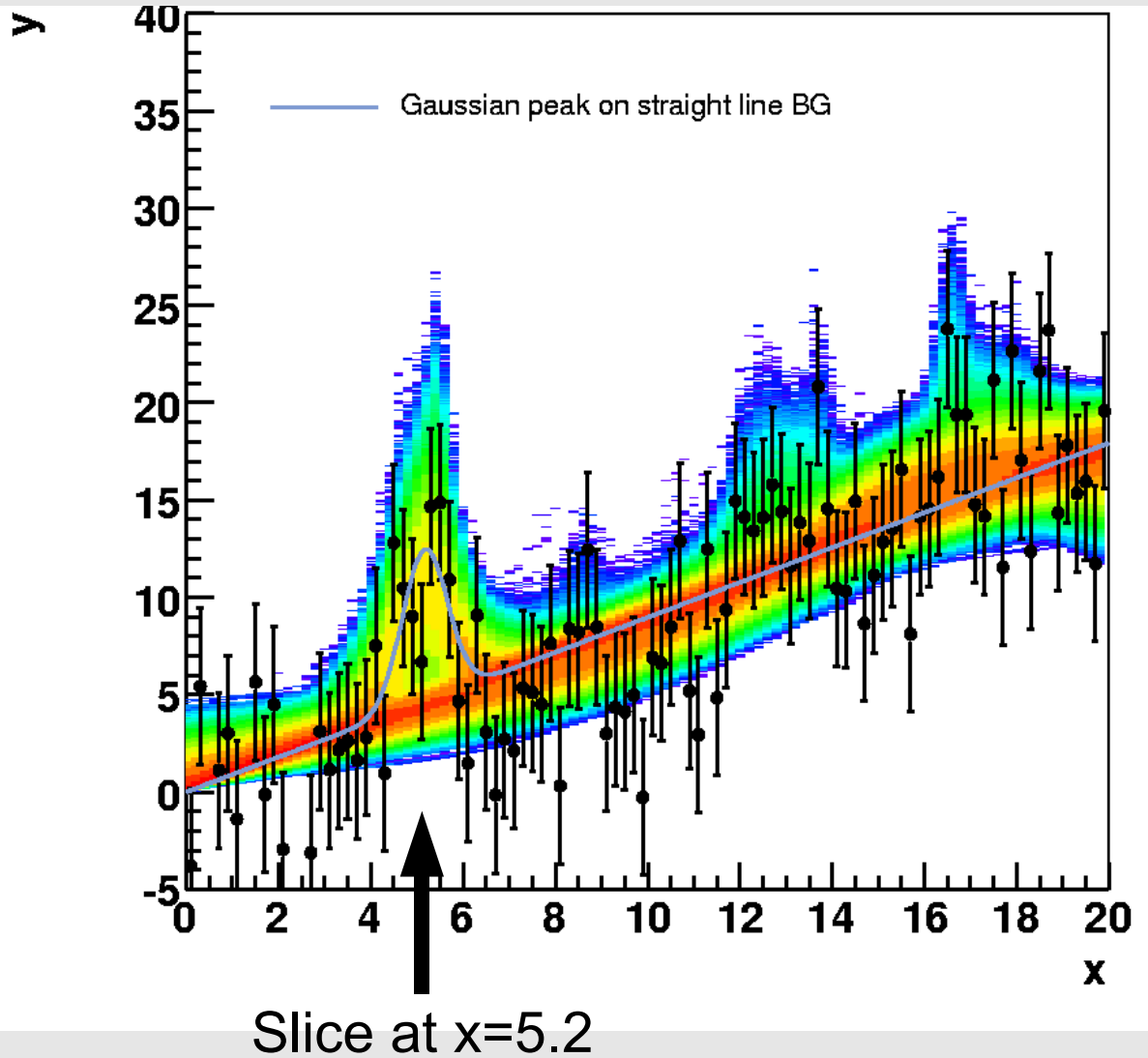
- All distributions (1-d and 2-d) are stored during single run
- Markov Chain saved to ROOT file

Error band contains 68% of estimated y-values given an x-value.



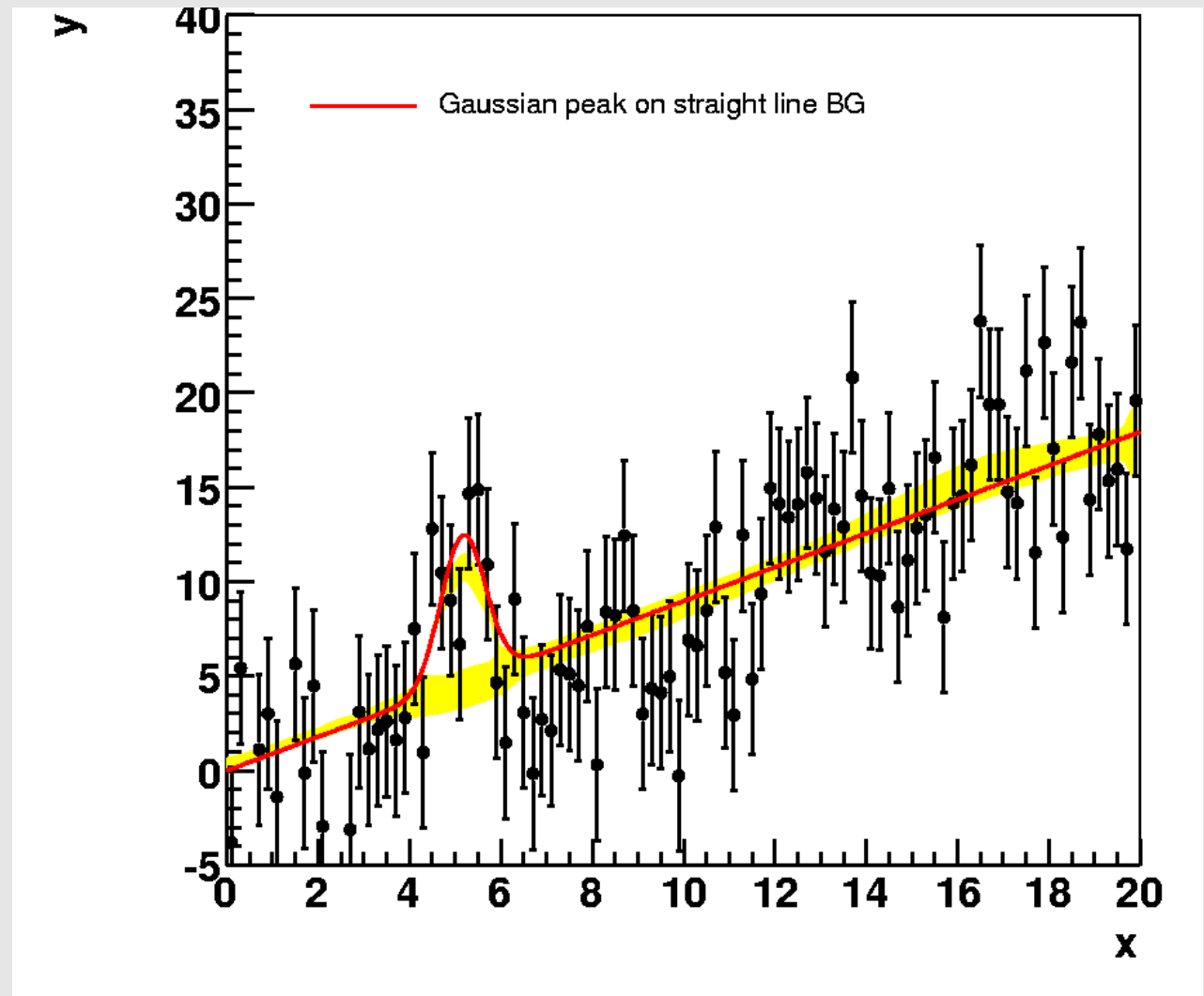


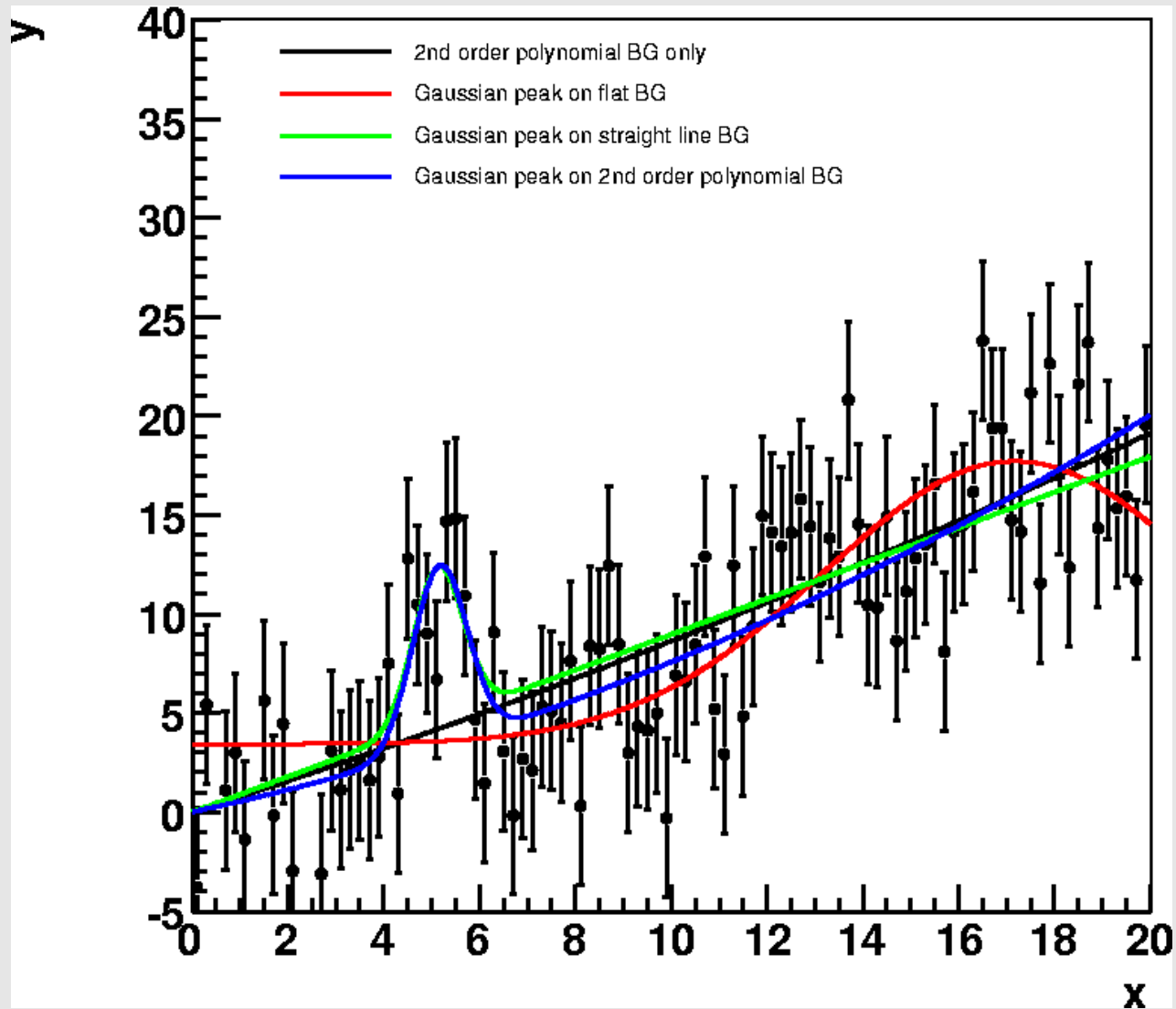
Probability density for true  $y$  at  $x=5.2$



Slice at  $x=5.2$

Error band contains 68% of estimated y-values given an x-value.





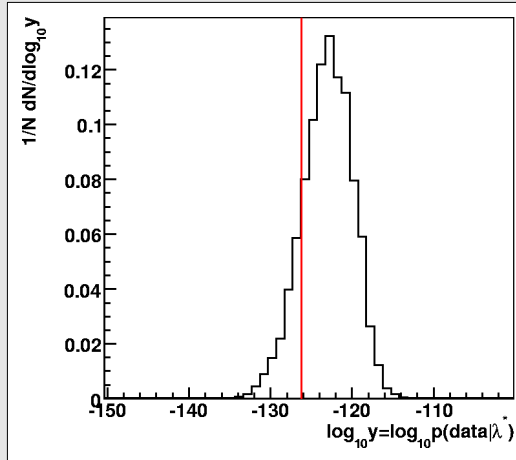
## How to judge if the model describes the data (at all)?

### Strategy:

- Find best parameters from original data
- Generate data sets using these parameters (ensemble tests)
- Calculate and histogram the likelihood
- Compare likelihood distribution with likelihood of original data
- Calculate the p-value:
  - probability likelihood equal to or less than for original data
  - Returns value between 0 and 1
  - Large p-value means good agreement

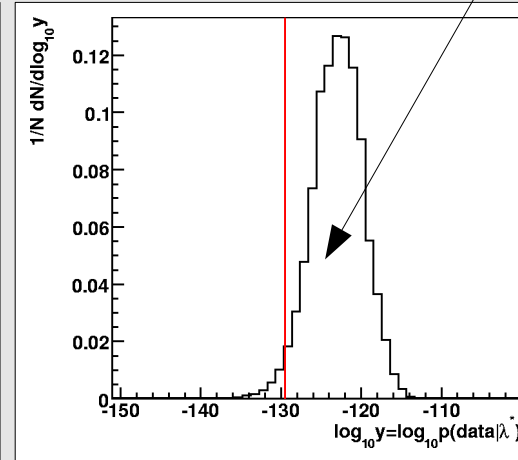
For a good model, the probability to find a better matching between data and theory (larger likelihood) should be small.

Polynomial  
p-value = 0.232



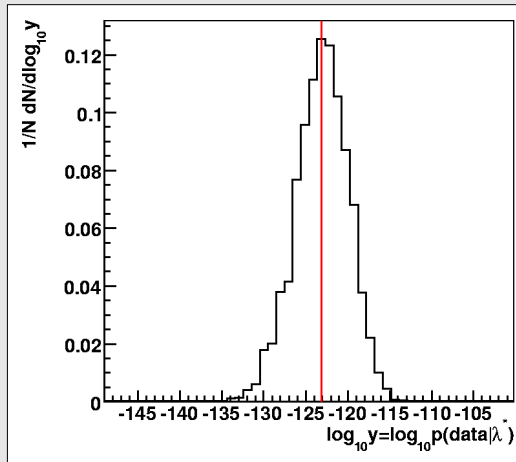
Model OK

Gauss + const.  
P-value = 0.041



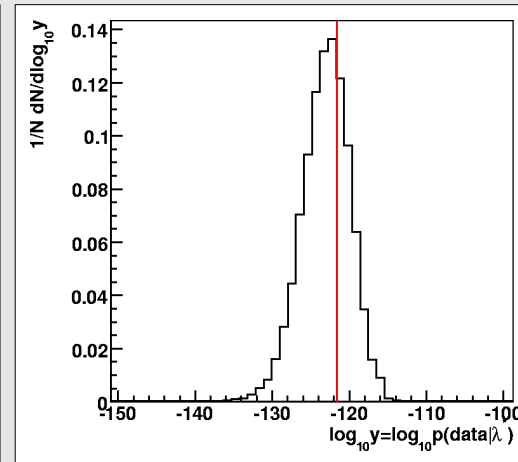
Model not appropriate

Gauss + line  
p-value = 0.540



Model well suited to describe data

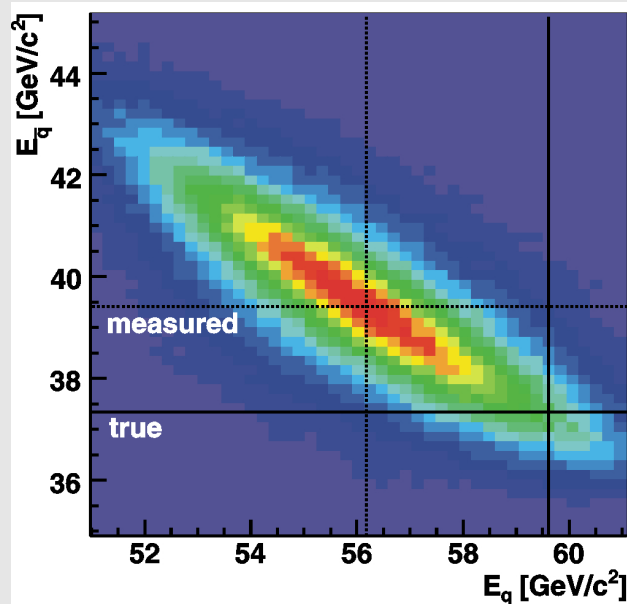
Gauss + polyn.  
P-value = 0.778



Model well suited to describe data

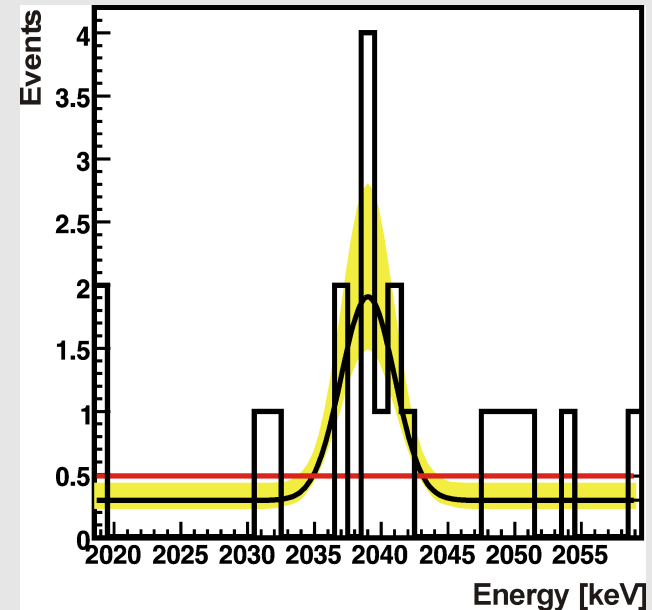
Your choice! Data not „strong“ enough for clear distinction

## ATLAS (HEP)



Correlation between 2 jets combining to a W

## GERDA ( $0\nu\beta\beta$ )



Energy spectrum in the region of interest. The model contains signal and flat background



- Toolkit tailored for Bayesian statistics
- “Fitting” just one example, package more general
- Strong numerical tools and interfaces implemented
- A variety of possible applications in physics
- Publication in preparation (1-2 weeks)
- **First release expected in July 2008 ([www.mppmu.mpg.de/bat](http://www.mppmu.mpg.de/bat))**
- **ROOTified version being worked on**
- **Developers and beta-testers welcome!**