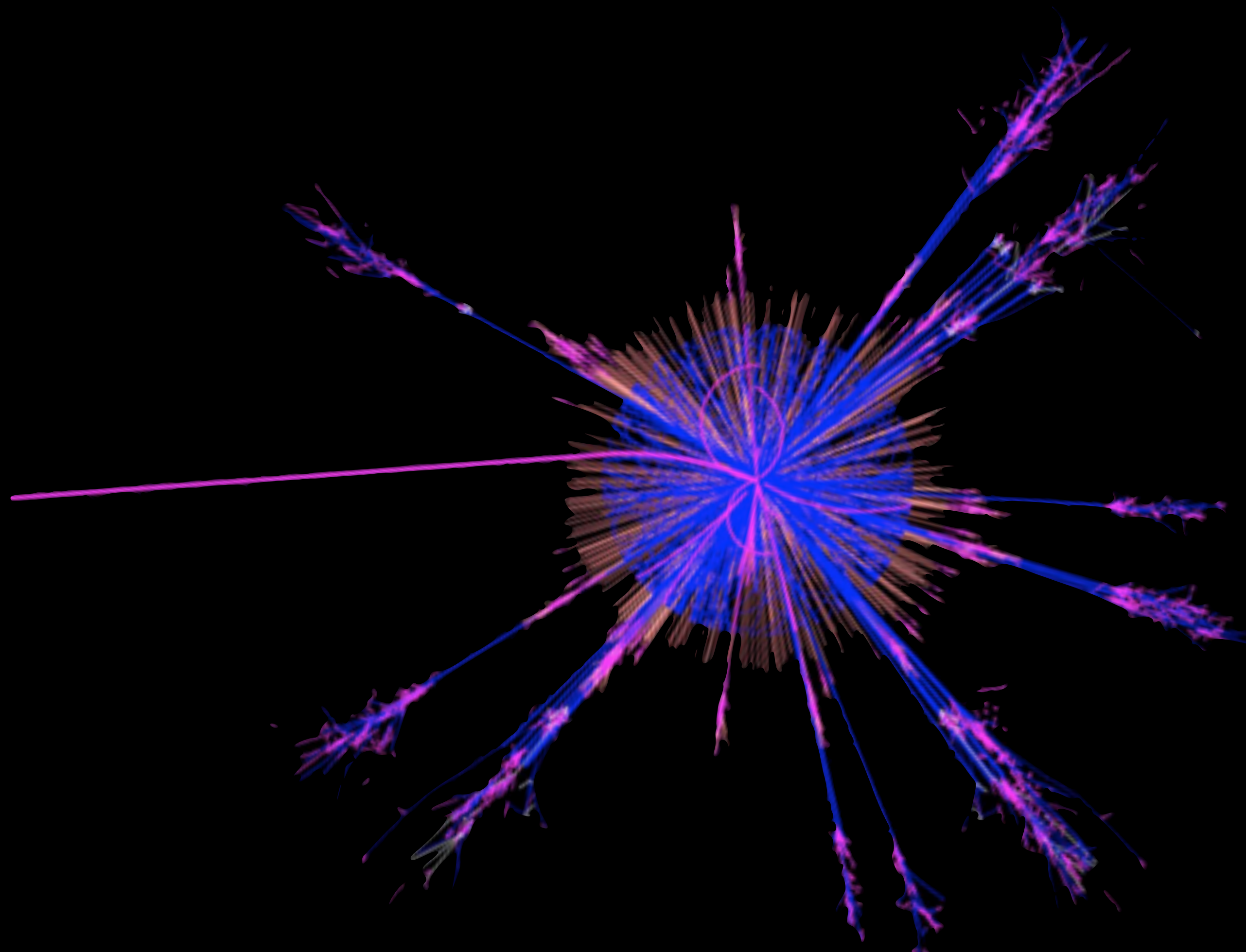
NYU Center for Data Science

Center for Cosmology and Particle Physics

LECTURES ON

STATISTICS



@KyleCranmer
New York University
Department of Physics
Center for Data Science

Introduction

Statistics plays a vital role in science, it is the way that we:

- quantify our knowledge and uncertainty
- communicate results of experiments

Big questions:

- how do we make discoveries, measure or exclude theoretical parameters, ...
- how do we get the most out of our data
- how do we incorporate uncertainties
- how do we make decisions

Statistics is a very big field, and it is not possible to cover everything in 3 hours. In these talks I will try to:

- **explain** some fundamental ideas & prove a few things
- **enrich** what you already know
- **expose** you to some new ideas

I will try to go slowly, because if you are not following the logic, then it is not very interesting.

- Please feel free to ask questions and interrupt at any time

LECTURE NOTES

Practical Statistics for the LHC

Kyle Cranmer
Center for Cosmology and Particle Physics, Physics Department, New York University, USA

Abstract
This document is a pedagogical introduction to statistics for particle physics. Emphasis is placed on the terminology, concepts, and methods being used at the Large Hadron Collider. The document addresses both the statistical tests applied to a model of the data and the modeling itself . I expect to release updated versions of this document in the future.

Links:
On Authorea
arxiv:1503.07622

Contents

Further Reading

By physicists, for physicists

G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.

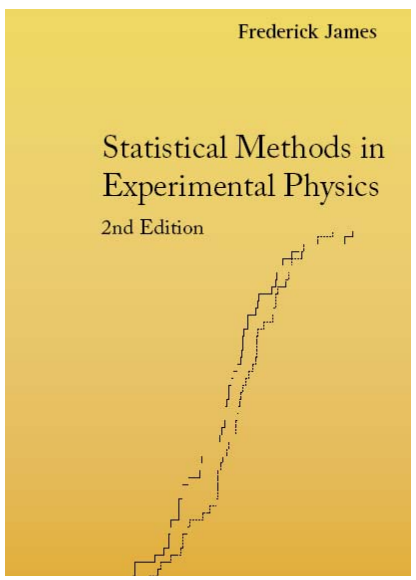
R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;

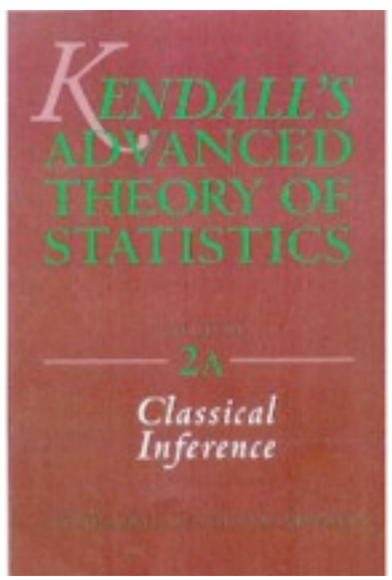
F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006;

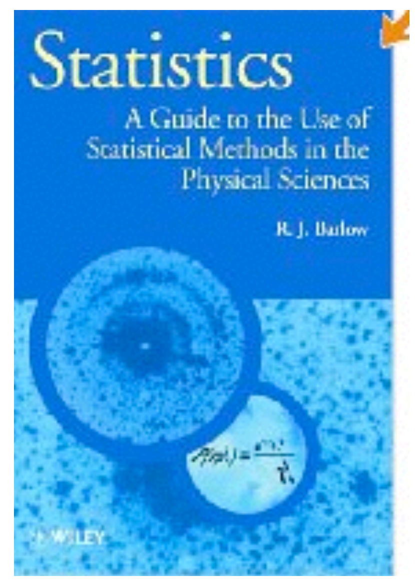
- W.T. Eadie et al., North-Holland, 1971 (1st ed., hard to find);

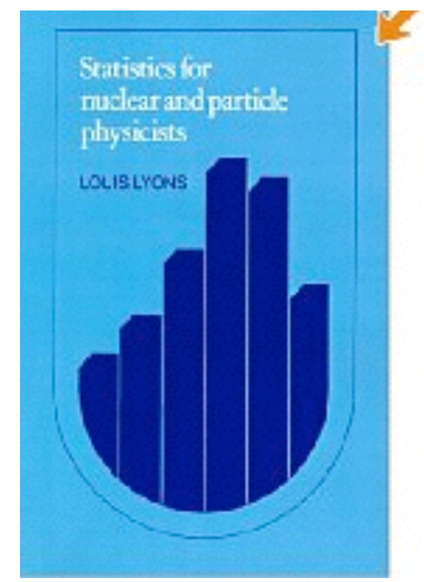
S.Brandt, Statistical and Computational Methods in Data Analysis, Springer, New York, 1998.

L.Lyons, Statistics for Nuclear and Particle Physics, CUP, 1986.









My favorite statistics book by a statistician:

Stuart, Ord, Arnold. "Kendall's Advanced Theory of Statistics" Vol. 2A *Classical Inference & the Linear Model.*

Other Lectures

Fred James's lectures

http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic_Training&id=AT00000799

http://www.desy.de/~acatrain/

Glen Cowan's lectures

http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Louis Lyons

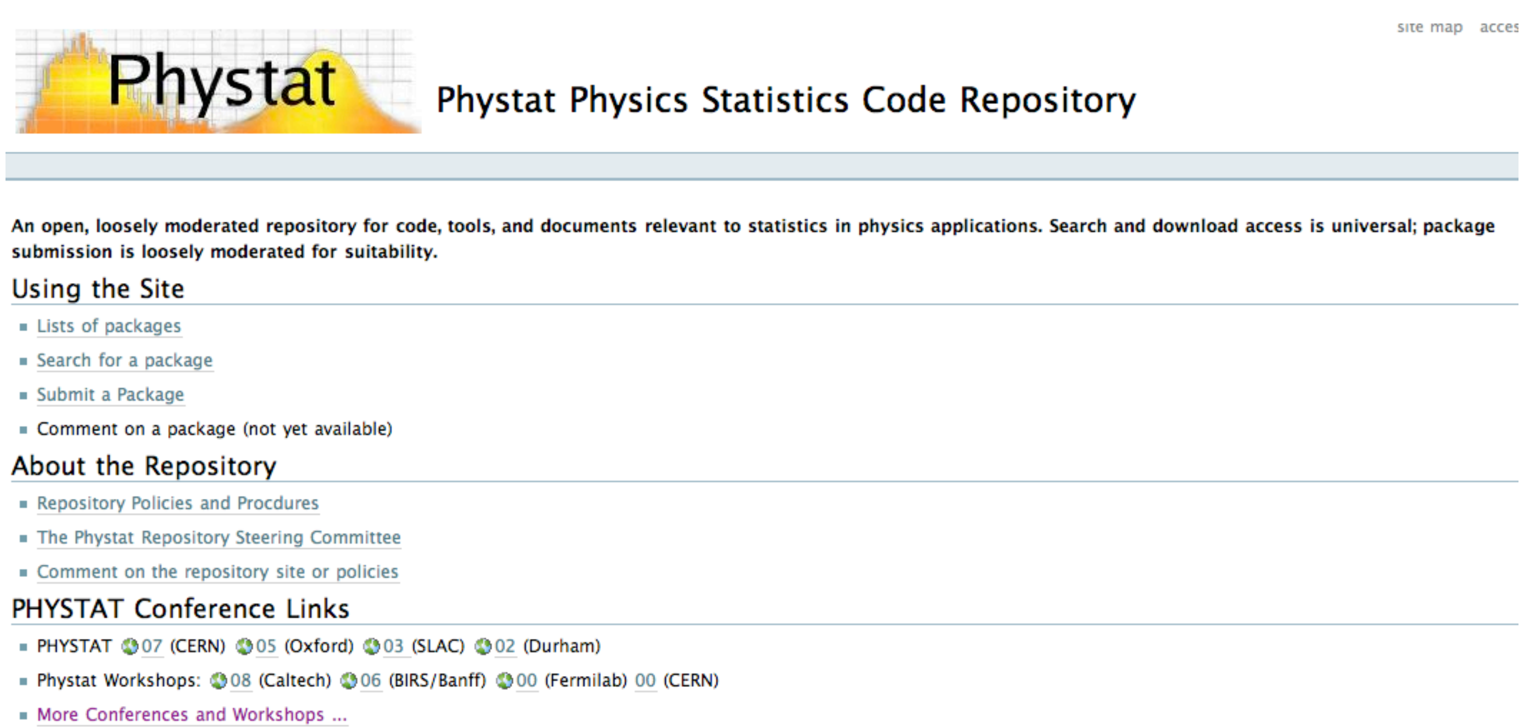
http://indico.cern.ch/conferenceDisplay.py?confId=a063350

Bob Cousins gave a CMS lecture, may give it more publicly

Gary Feldman "Journeys of an Accidental Statistician"

http://www.hepl.harvard.edu/~feldman/Journeys.pdf

The PhyStat conference series at PhyStat.org:



Outline

Lecture 1: Preliminaries

- Probability Density Function vs. Likelihood
- Point estimates (measurements) and maximum likelihood estimators

Part 2: Building a probability model

- Examples of different “narratives”
- A generic template for high energy physics

Lecture 2: Hypothesis testing

- The Neyman-Pearson lemma and the likelihood ratio
- Composite models and the profile likelihood ratio
- Review of ingredients for a hypothesis test

Lecture 3: Limits & Confidence Intervals

- The meaning of confidence intervals as inverted hypothesis tests
- LHC-style CLs
- Asymptotic properties of likelihood ratios
- Bayesian approach

LECTURE 1

Terms

The next lectures will rely on a clear understanding of these terms:

- Random variables / “observables” x
- Probability mass and probility density function (pdf) $p(x)$
- Parametrized Family of pdfs / “model” $p(x|\alpha)$
- Parameter α
- Likelihood $L(\alpha)$
- Estimate (of a parameter) $\hat{\alpha}(x)$

Random variable / observable

"Observables" are quantities that we observe or measure directly

- They are random variables under repeated observation

Discrete observables:

- number of particles seen in a detector in some time interval
- particle type (electron, muon, ...) or charge (+,-,0)

Continuous observables:

- energy or momentum measured in a detector
- invariant mass formed from multiple particles

When dealing with discrete random variables, define a **Probability Mass Function** as probability for i^th possibility

$$P(x_i) = p_i$$

Defined as limit of long term frequency

- probability of rolling a 3 := $\text{limit}_{\#\text{trials}\to\infty}$ (# rolls with 3 / # trials)
  - you don't need an infinite sample for definition to be useful

And it is normalized

$$\sum_i P(x_i) = 1$$

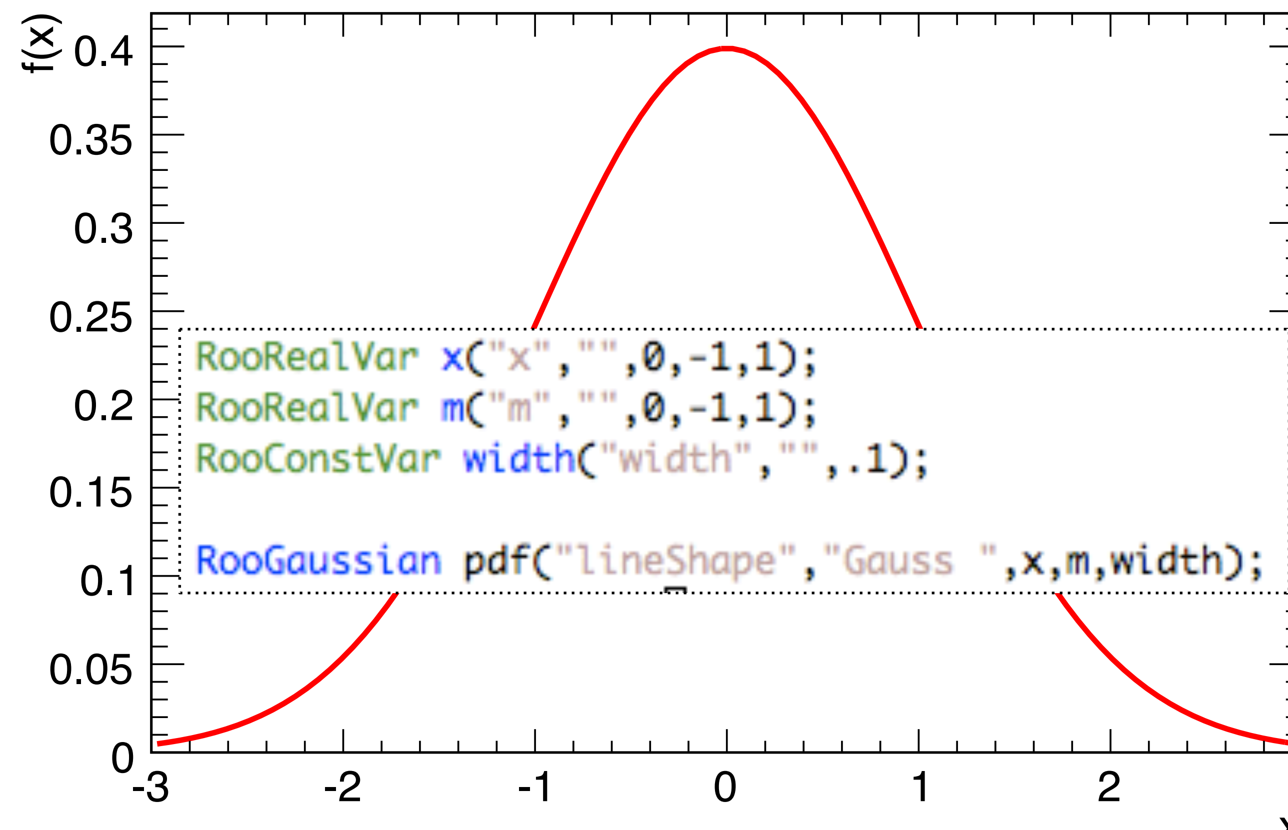
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function**

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

PDFs are always normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



Parametrized families / models

Often we are interested in a parametried family of pdfs

- We will write these as: $f(x|\alpha)$ said "f of x given α"
  - where α are the parameters of the "model" (written in greek characters)

A discrete example:

- The Poisson distribution is a probability mass function for n, the number of events one observes, when one expects μ events

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

A continuous example

- The Gaussian distribution is a probability density function for a continuous variable x characterized by a mean μ and standard deviation σ

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Likelihood Function

Consider the Poisson distribution describes a discrete event count n for a real-valued mean μ.

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

The **likelihood** of μ given n is the same equation evaluated as a function of μ

- Now it's a continuous function
- But it is not a pdf!

$$L(\mu) = Pois(n|\mu)$$

Common to plot the -ln L (or -2 ln L)

- helps avoid thinking of it as a PDF
- connection to χ^2 distribution

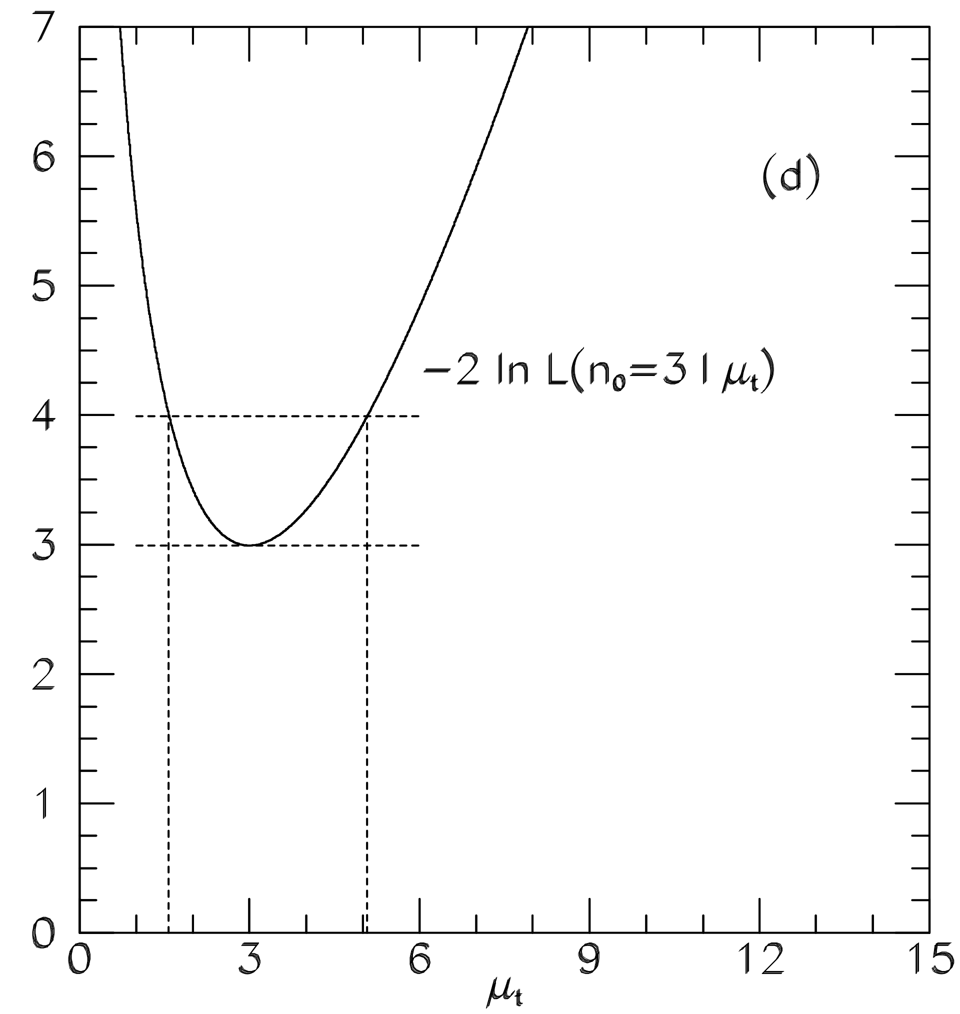


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

Repeated observations

In particle physics we are usually able to perform repeated observations of x that are **independent & identically distributed**

- These repeated observations are written $\{x_i\}$
- and the likelihood in that case is

$$L(\alpha) = \prod_i f(x_i|\alpha)$$

- and the log-likelihood is

$$\log L(\alpha) = \sum_i \log f(x_i|\alpha)$$

MEASUREMENT / ESTIMATORS

Estimators

Given some model $f(x|\alpha)$ and a set of observations $\{x_i\}$ often one wants to estimate the true value of α (assuming the model is true).

An **estimator** is function of the data written $\hat{\alpha}(x_1, \ldots x_n)$

- Since the data are random, so is the resulting estimate
- often it is just written $\hat{\alpha}$, where the x-dependence is implicit
- one can compute expectation of the estimator

$$E[\hat{\alpha}(x)|\alpha] = \int \hat{\alpha}(x) f(x|\alpha) dx$$

Properties of estimators:

- **bias** $E[\hat{\alpha}(x)|\alpha] - \alpha$ (unbiased means bias=0)
- **variance** $E[(\hat{\alpha}(x) - \bar{\alpha})^2|\alpha] = \int (\hat{\alpha}(x) - \bar{\alpha})^2 f(x|\alpha) dx$
- **asymptotic bias** limit of bias with infinite observations

MAXIMUM LIKELIHOOD ESTIMATORS

There are many different possible estimators, but the most well-known and well-studied is the maximum likelihood estimator (MLE)

$$\hat{\alpha}(x) = \text{argmax}_\alpha L(\alpha) = \text{argmax}_\alpha f(x|\alpha)$$

This is just the value of α that maximizes the likelihood

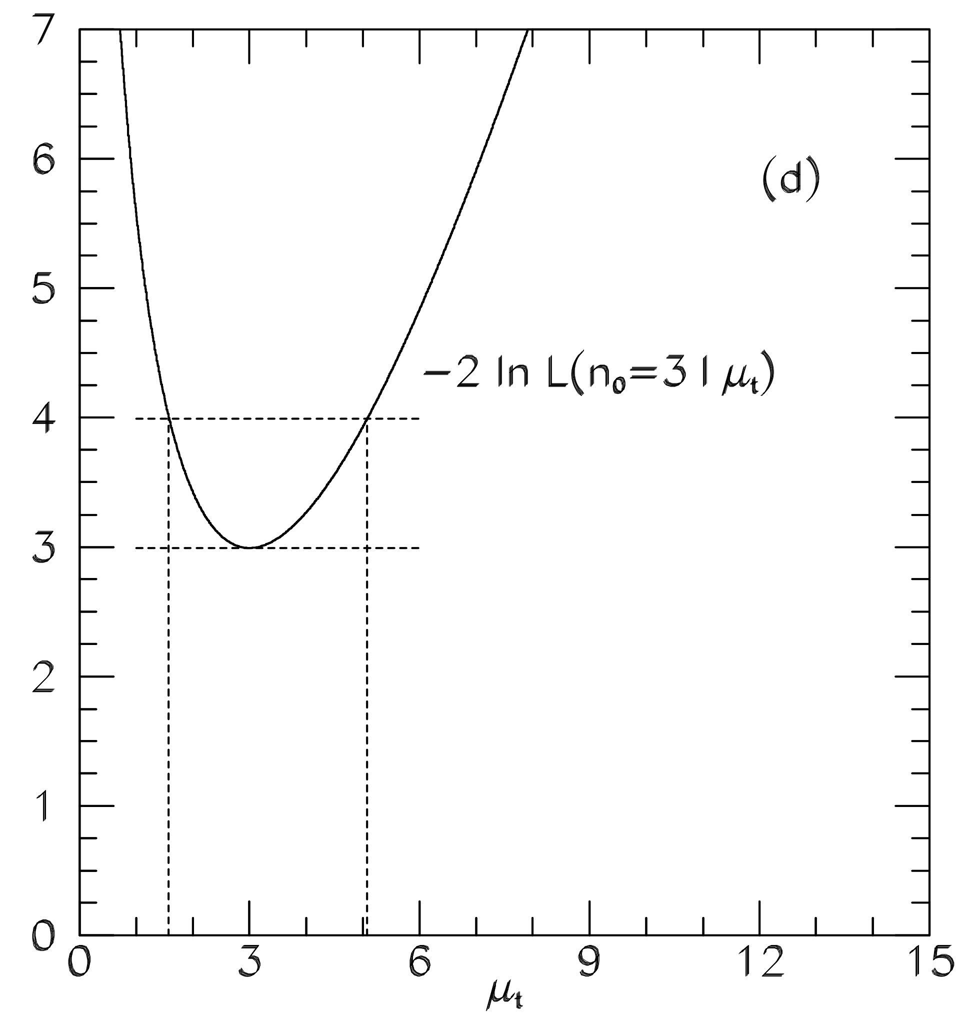


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

Example: the Poisson distribution

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

Maximizing $L(\mu)$ is the same as minimizing -ln $L(\mu)$

$$-\frac{d}{d\mu}\ln L(\mu)\big|_{\hat{\mu}} = 0 = \frac{d}{d\mu}\left(\mu - n\ln\mu + \underbrace{\ln n!}_{\text{const}}\right) = 1 - \frac{n}{\mu}$$

$$\Rightarrow \hat{\mu} = n$$

In this case, the MLE is unbiased b/c E[n]=μ

A SECOND EXAMPLE

Consider a set of observations $\{x_i\}$ and we want to estimate the mean of a Gaussian with known σ

which gives

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$-\frac{d}{d\mu} \ln L(\mu)\Big|_{\hat{\mu}} = 0 = \frac{d}{d\mu}\left(\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} + \underbrace{\ln\sqrt{2\pi}\sigma}_{\text{const}}\right) = \sum_i \frac{(x_i-\mu)}{\sigma^2}$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_i x_i$$ (an unbiased estimator) .

However, the MLE $\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i-\mu)^2$ is biased

It can be shown that $\hat{\sigma}^2 = \frac{1}{N-1}\sum_i (x_i-\mu)^2$ is unbiased

Thus, the MLE is **asymptotially unbiased .**

Note: if $\hat{\sigma}^2$ is an unbiased estimate of σ^2, then $\sqrt{\{\hat{\sigma}^2\}}$ is a biased estimate of σ.

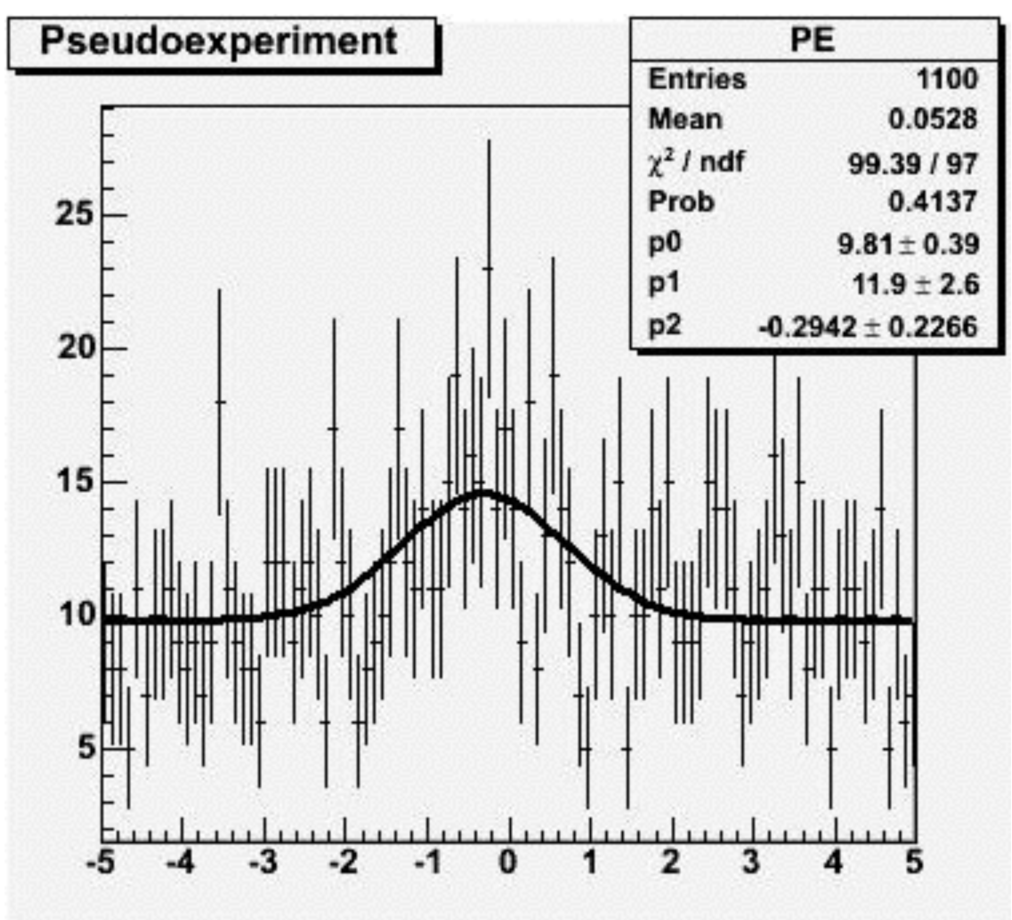
"GREEDY BUMP BIAS"

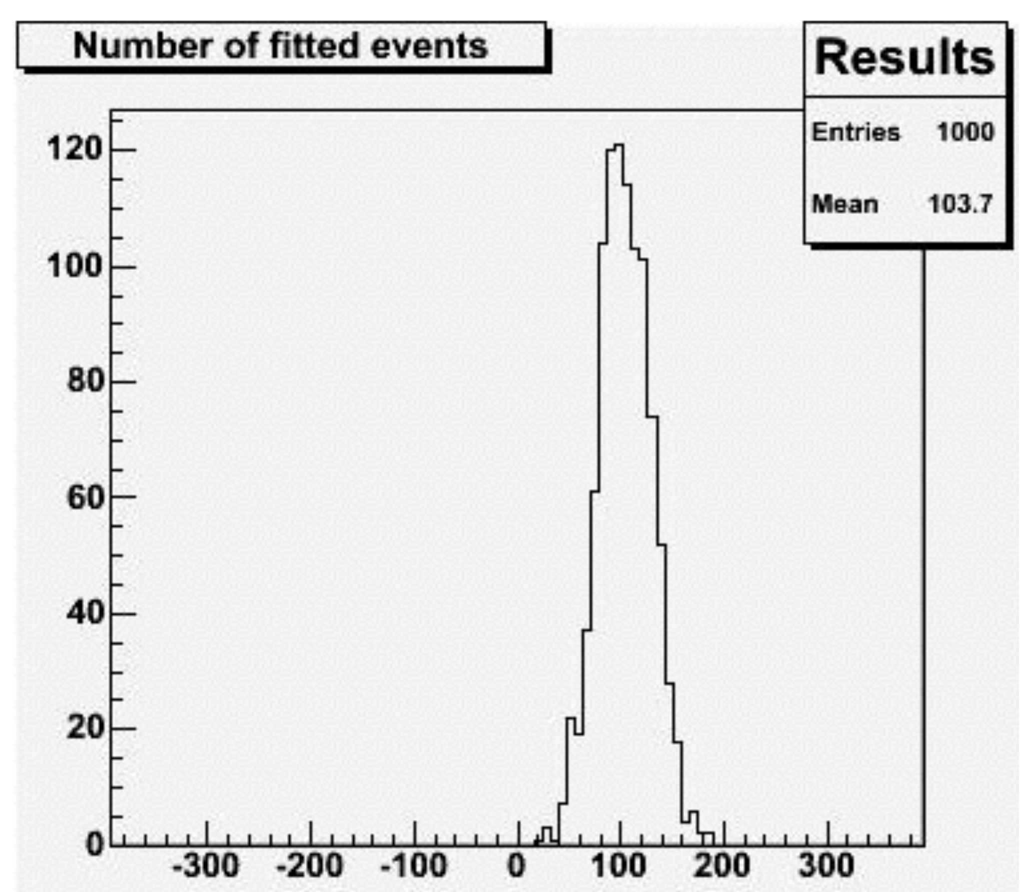
MLE of cross-section when the mass is also allowed to float is also biased.

- fit preferentially adjusts mass to find upward fluctuations
- For fixed s/b, the bias goes away asymptotically.

See "Greedy bump bias" by Tommaso Dorigo

http://www.science20.com/quantum_diaries_survivor/bump_hunting_ii_greedy_bump_bias





COVARIANCE AND CORRELATION

Define covariance cov[x,y] (also use matrix notation V_{xy}) as

$$\text{cov}[x,y] = E[xy] - \mu_x\mu_y = E[(x-\mu_x)(y-\mu_y)]$$

Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\text{cov}[x,y]}{\sigma_x\sigma_y}$$

If x, y, independent, i.e., $f(x,y) = f_x(x)f_y(y)$, then

$$E[xy] = \int\int xy\, f(x,y)\, dxdy = \mu_x\mu_y$$

$\rightarrow$ $\text{cov}[x,y] = 0$ x and y, 'uncorrelated'

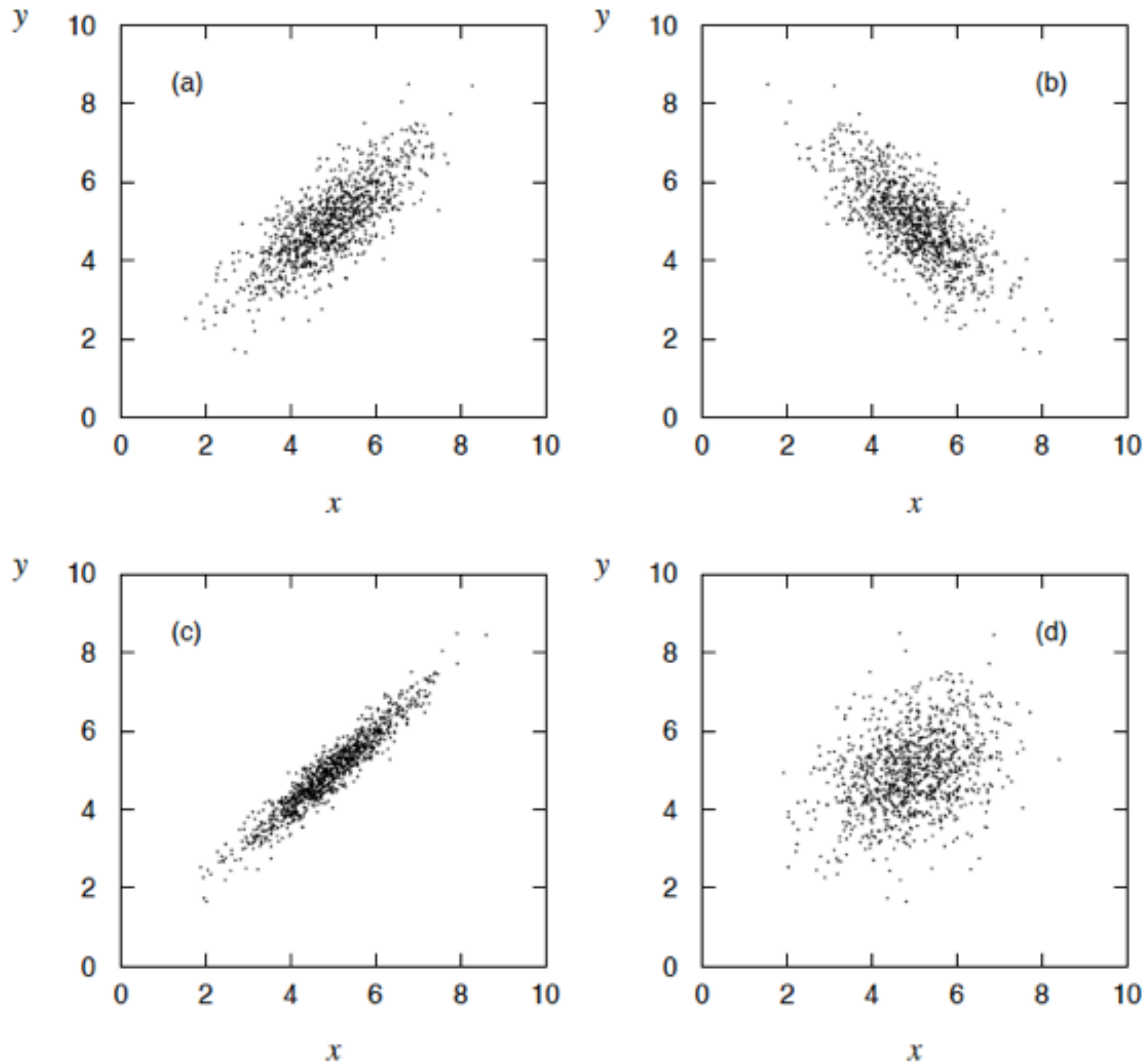
N.B. converse not always true.

[G. Cowan]

CORRELATION (CONT.)

$\rho = 0.75$

$\rho = -0.75$

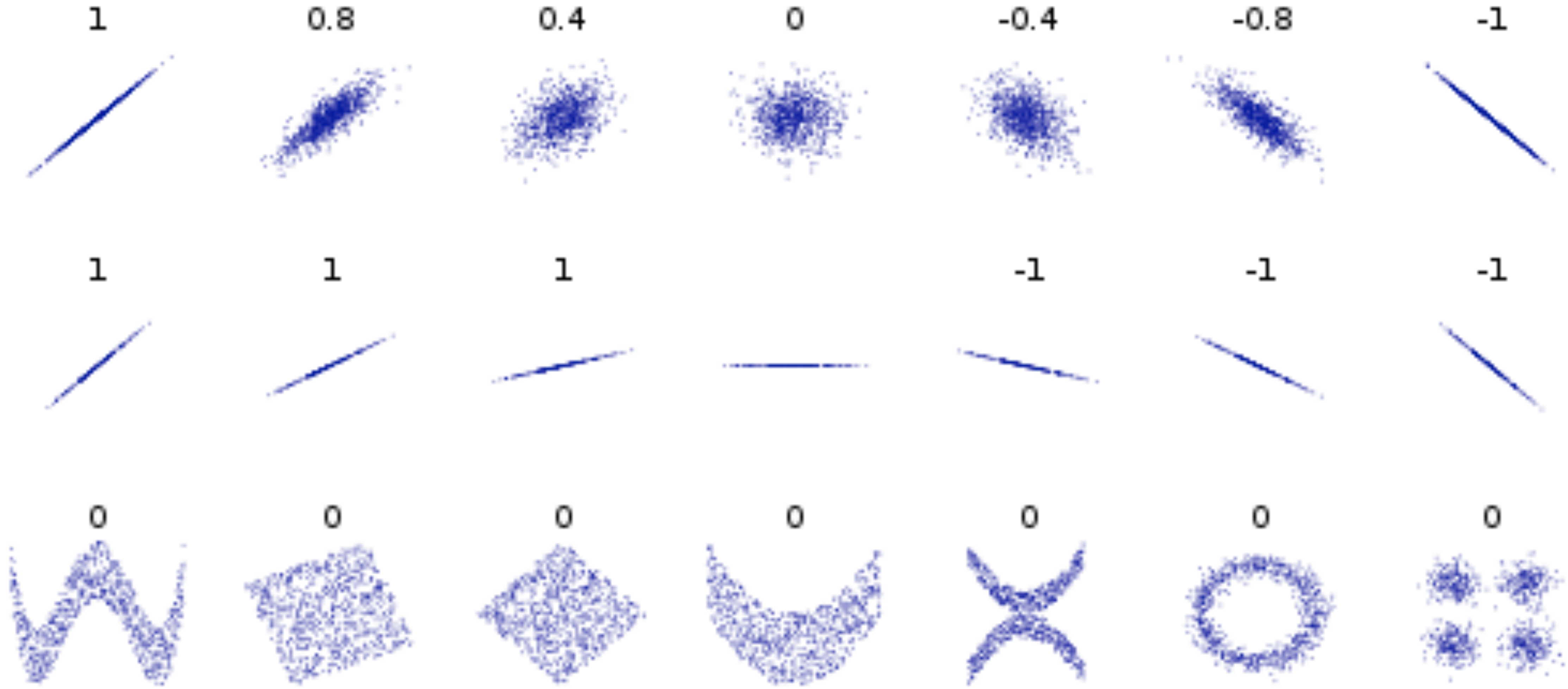


$\rho = 0.95$

$\rho = 0.25$

[G. Cowan]

Correlation (cont.)



http://en.wikipedia.org/wiki/Correlation_and_dependence

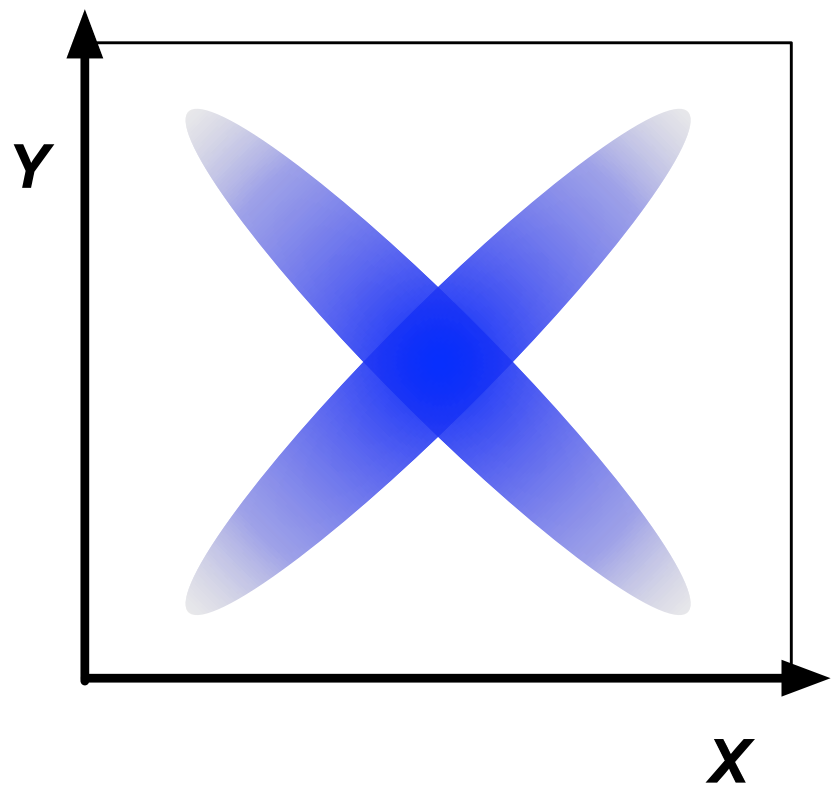
Mutual Information

Mutual Information is a more general notion of ‘correlation’

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right),$$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

- it is symmetric: $I(X;Y) = I(Y;X)$
- if and only if X,Y totally independent: $I(X;Y)=0$
- possible for X,Y to be uncorrelated, but not independent



Mutual Information doesn’t seem to be used much within HEP, but it seems quite useful

CRAMÉR-RAO BOUND

The minimum variance bound on an estimator is given by the Cramér-Rao inequality:

- simple univariate case:

$$\mathrm{var}(\hat{\theta}) = E[(\theta - \hat{\theta})^2]$$

- For an unbiased estimator the Cramér-Rao bound states

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- where $I(\theta)$ is the Fisher information

$$(\mathcal{I}(\theta))_{i,j} = \mathrm{E}\left[\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\frac{\partial}{\partial\theta_j}\ln f(X;\theta)\middle|\theta\right].$$

- General form for multiple parameters:

$$\mathrm{cov}[\hat{\theta}|\theta]_{ij} \geq I_{ij}^{-1}(\theta)$$

Maximum Likelihood Estimators *asymptotically* reach this bound

BAYES THEOREM

BAYES' THEOREM

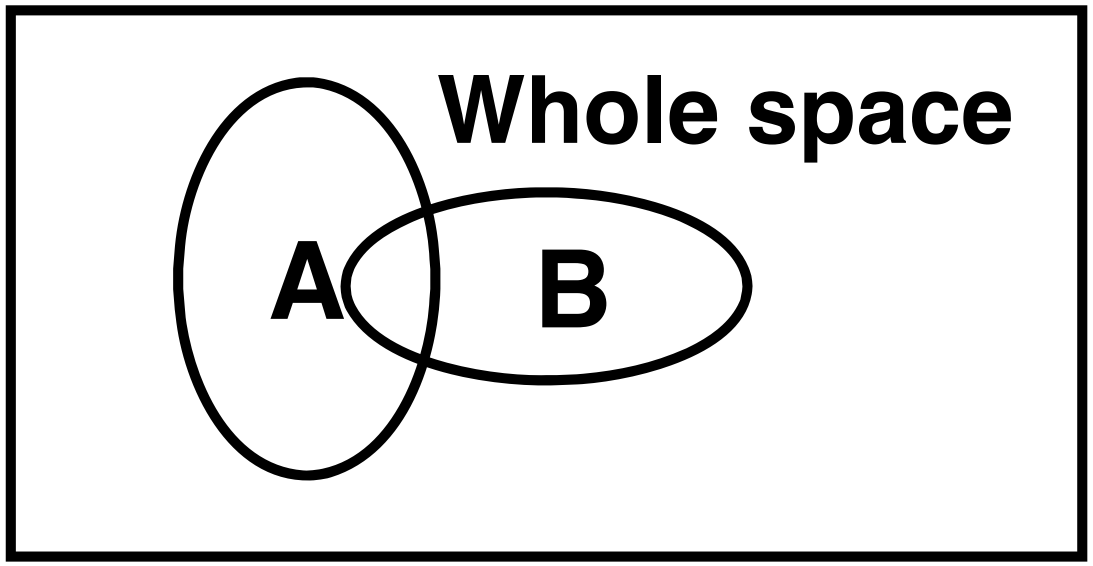
Bayes' theorem relates the conditional and marginal probabilities of events A & B

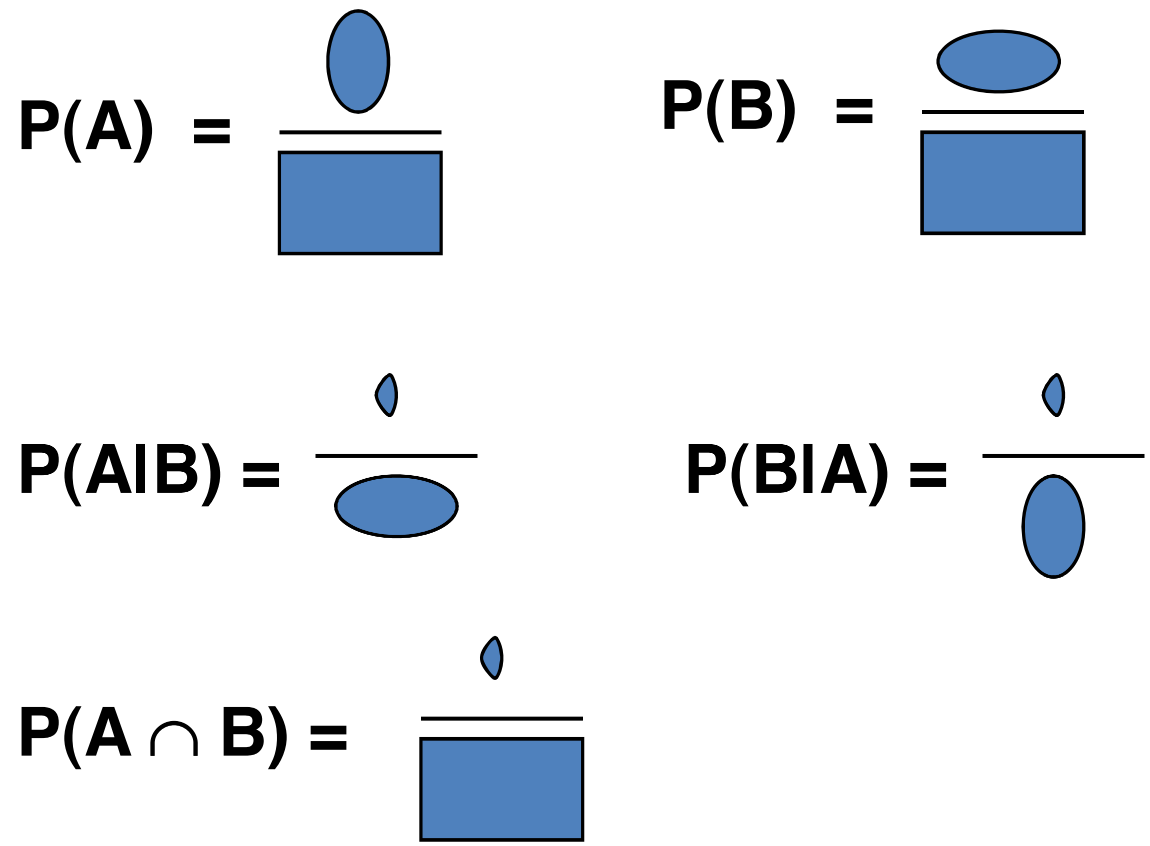
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- **P(*A*)** is the prior probability. It is "prior" in the sense that it does not take into account any information about *B*.
- **P(*A* | *B*)** is the conditional probability of *A*, given *B*. It is also called the posterior probability because it is derived from or depends upon the specified value of *B*.
- **P(*B* | *A*)** is the conditional probability of *B* given *A*.
- **P(*B*)** is the prior or marginal probability of *B*, and acts as a normalizing constant.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\mathcal{N}} \propto L(\theta)\pi(\theta)$$

P, Conditional P, and Derivation of Bayes' Theorem in Pictures





Don't forget about "Whole space" Ω I will drop it from the notation typically, but occasionally it is important.

Bob Cousins, CMS, 2008

$\Rightarrow$ **P(BIA) = P(AIB) × P(B) / P(A)**

Louis's Example

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

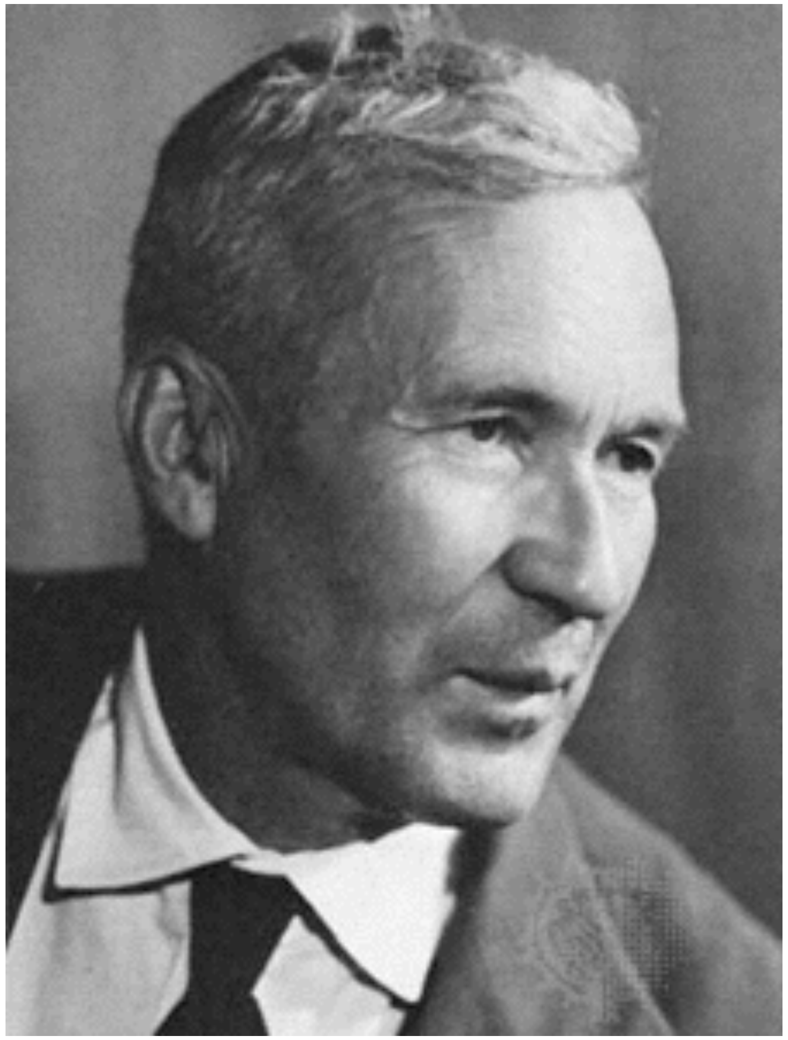
but

P (female ; pregnant) >>>3%

AXIOMS OF PROBABILITY

These Axioms are a mathematical starting point for probability and statistics

1. probability for every element, E, is non-negative $P(E) \geq 0 \quad \forall E \subseteq \mathcal{F} = 2^{\Omega}$
2. probability for the entire space of possibilities is 1 $P(\Omega) = 1.$
3. if elements E_i are disjoint, probability is additive $P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$



Kolmogorov axioms (1933)

Consequences:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\Omega \setminus E) = 1 - P(E)$$

Different definitions of Probability

Frequentist

- defined as limit of long term frequency
- probability of rolling a 3 := limit of (# rolls with 3 / # trials)
  - you don't need an infinite sample for definition to be useful
  - sometimes ensemble doesn't exist
    - eg. P(Higgs mass = 125 GeV), P(it will snow tomorrow)
- Intuitive if you are familiar with Monte Carlo methods
- compatible with orthodox interpretation of probability in Quantum Mechanics. Probability to measure spin projected on x-axis if spin of beam is polarized along +z

$$|\langle \rightarrow | \uparrow \rangle|^2 = \frac{1}{2}$$

Subjective Bayesian

- Probability is a degree of belief (personal, subjective)
  - can be made quantitative based on betting odds
  - most people's subjective probabilities are not ***coherent*** and do not obey laws of probability

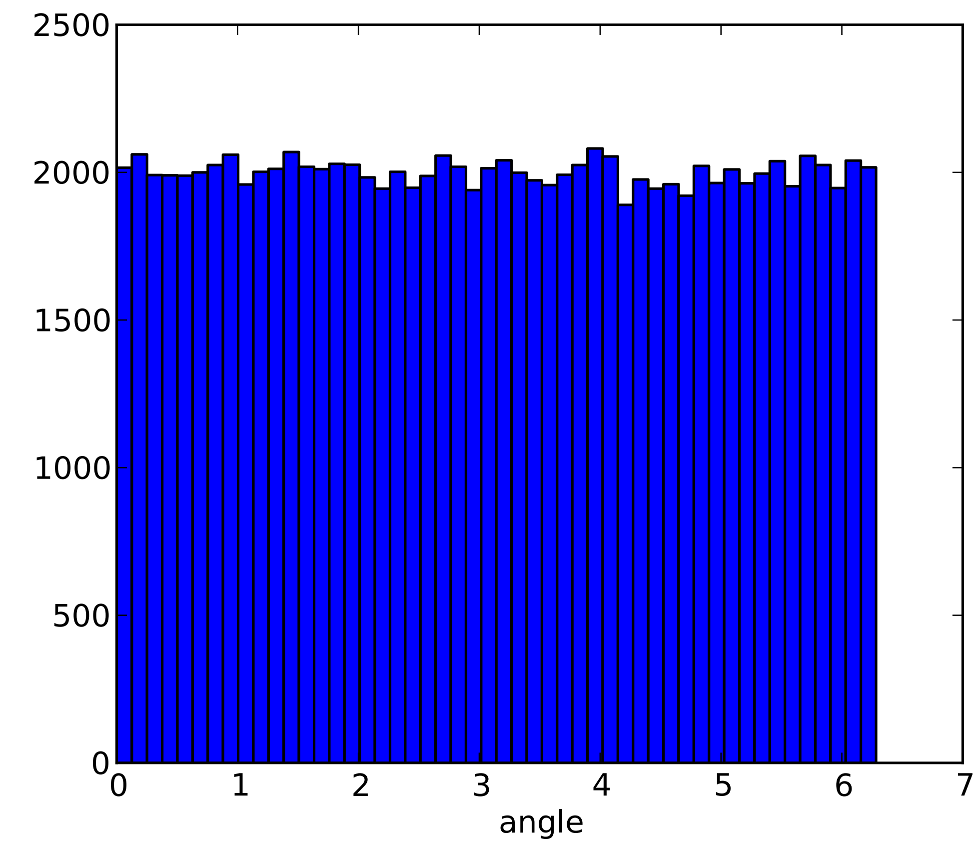
http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/#3.1

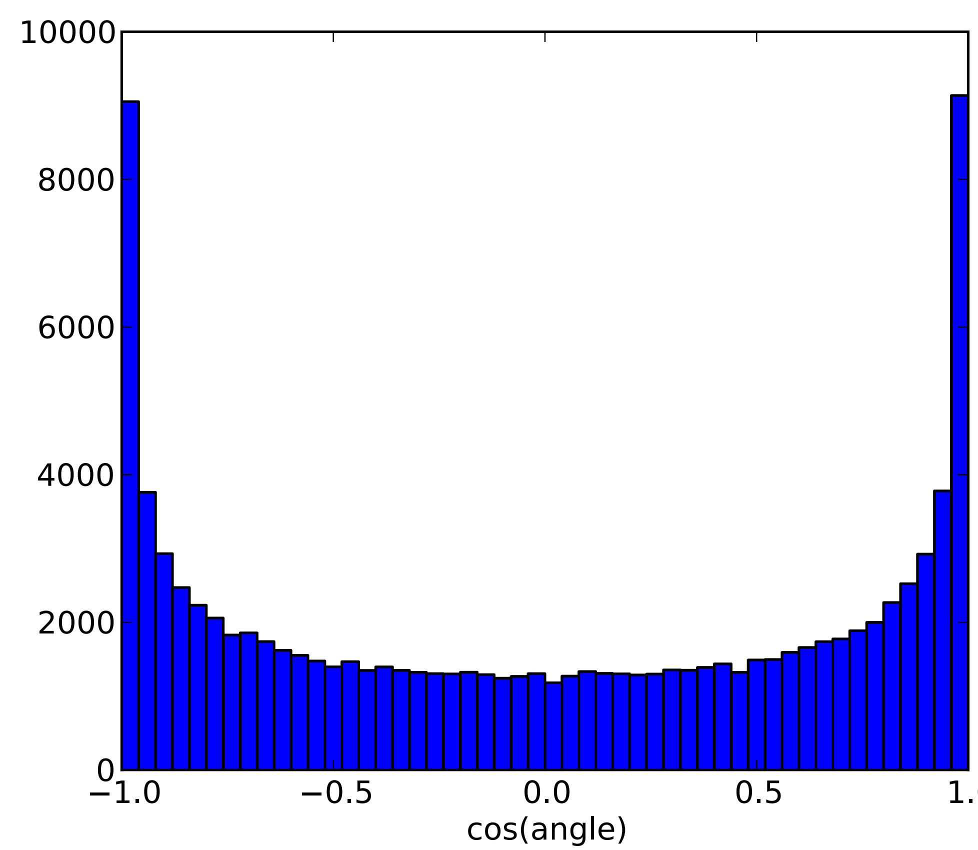
TRANSFORMATION PROPERTIES:
PDF VS. LIKELIHOOD

CHANGE OF VARIABLES

What happens with x⟶ cos(x)

```python
import numpy as np
import matplotlib.pyplot as plt

N_MC=100000  # number of Monte Carlo Experiments
nBins = 50 # number of bins for Histograms

data_x, data_y = [],[]  #lists that will hold x and y

# do experiments
for i in range(N_MC):
    # generate observation for x
    x = np.random.uniform(0,2*np.pi)

    y = np.cos(x)
    data_x.append(x)
    data_y.append(y)

#setup figures
fig = plt.figure(figsize=(13,5))
fig_x = fig.add_subplot(1,2,1)
fig_y = fig.add_subplot(1,2,2)

fig_x.hist(data_x,nBins)
fig_x.set_xlabel('angle')

fig_y.hist(data_y,nBins)
fig_y.set_xlabel('cos(angle)')

plt.show()
```





Change of variables

If $f(x)$ is the pdf for x and $y(x)$ is a change of variables, then the pdf $g(y)$ must satisfy

$$P(x_a < x < x_b) \equiv \int_{x_a}^{x_b} f(x)dx = \int_{y(x_a)}^{y(x_b)} g(y)dy \equiv P(y(x_a) < y < y(x_b))$$

We can rewrite the integral on the right

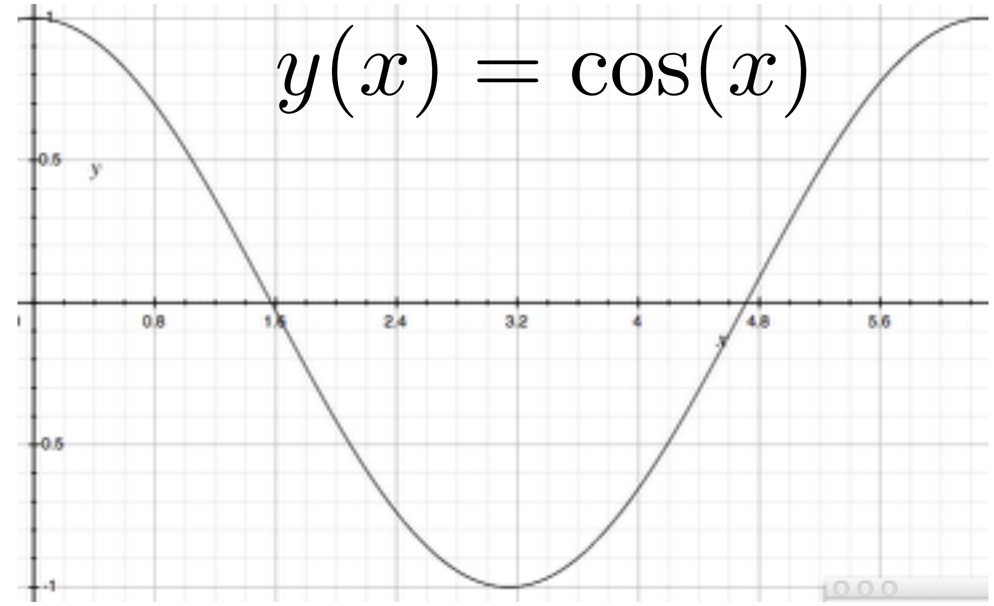
$$\int_{y(x_a)}^{y(x_b)} g(y)dy = \int_{x_a}^{x_b} g(y(x)) \left| \frac{dy}{dx} \right| dx$$

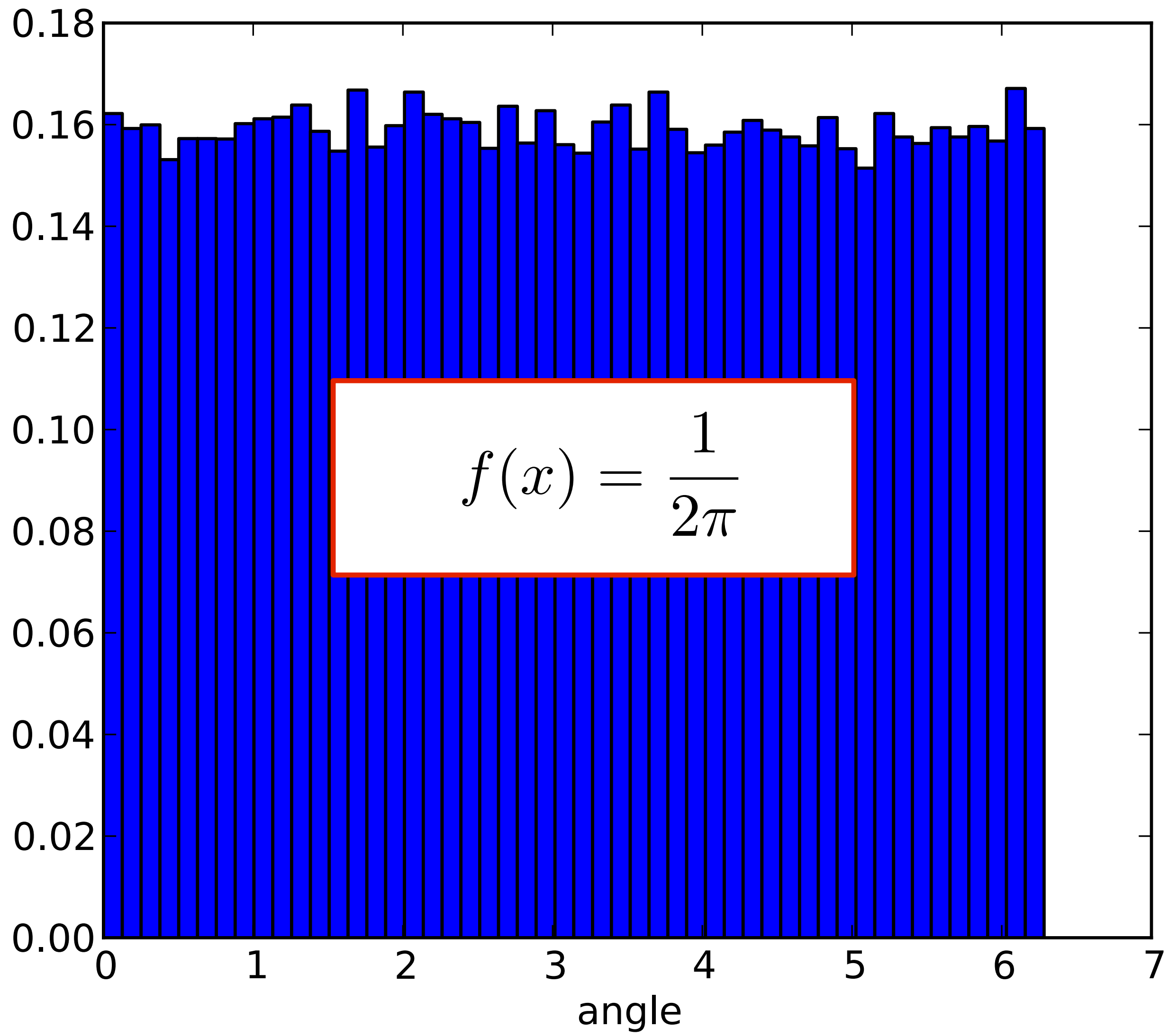
therefore, the two pdfs are related by a Jacobian factor

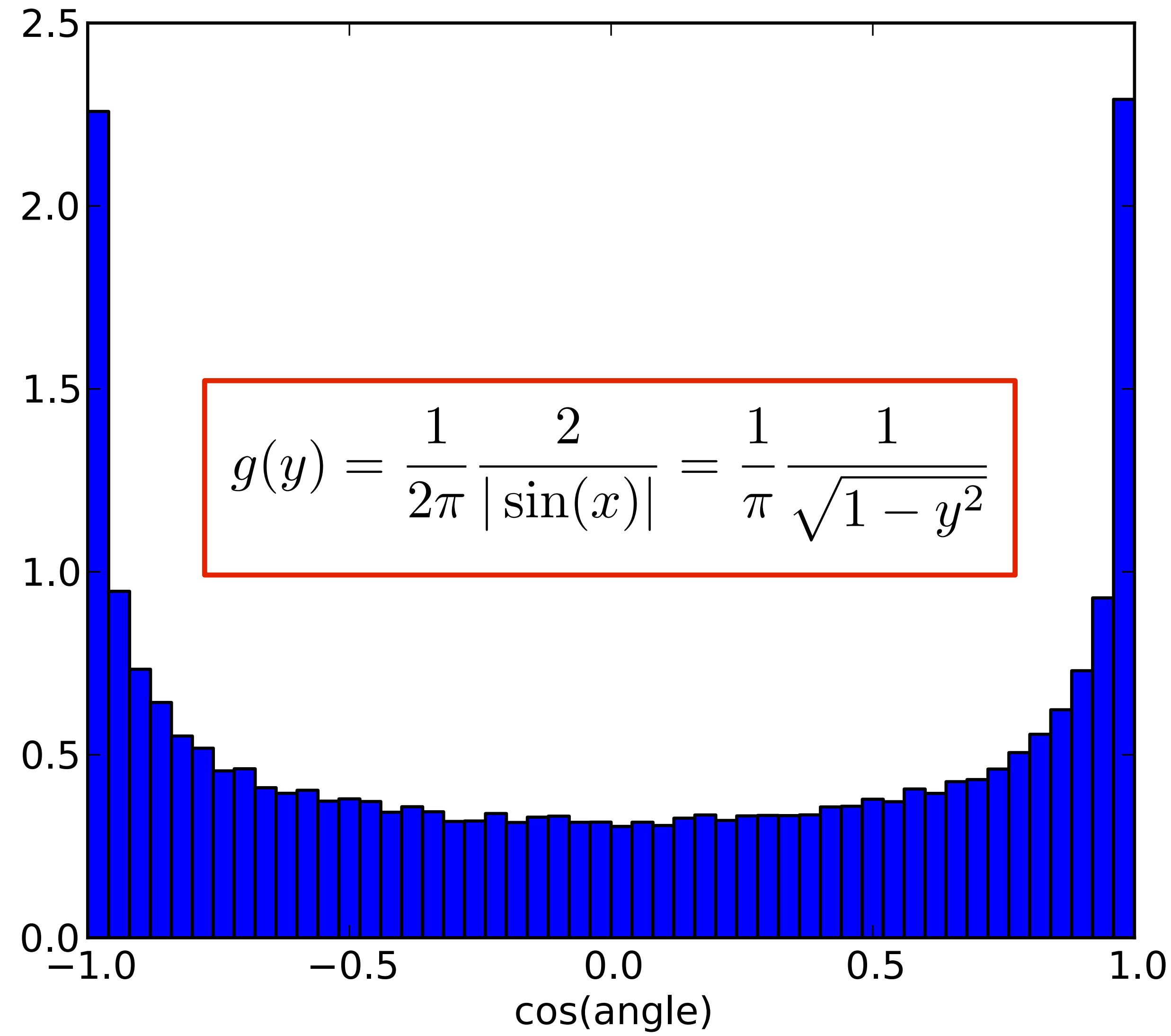
$$f(x) = g(y) \left| \frac{dy}{dx} \right|$$

AN EXAMPLE

$$f(x) = g(y)\left|\frac{dy}{dx}\right|$$







Change of variable x, change of parameter θ

- For pdf $p(x|\theta)$ and change of variable from x to y(x):

  $$p(y(x)|\theta) = p(x|\theta) \,/\, |dy/dx|.$$

  Jacobian modifies probability *density*, guaranties that

  $$P(\, y(x_1) < y < y(x_2)\,) \;=\; P(x_1 < x < x_2\,),$$ i.e., that

  Probabilities are invariant under change of variable x.
  - Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).
  - Likelihood *ratio* is invariant under change of variable x. (Jacobian in denominator cancels that in numerator).
- For likelihood $\mathcal{L}(\theta)$ and reparametrization from θ to u(θ):

  $$\mathcal{L}(\theta) \;=\; \mathcal{L}(u(\theta)) \quad (!).$$
  - Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization of parameter θ (reinforcing fact that $\mathcal{L}$ is *not* a pdf in θ).

Bob Cousins, CMS, 2008

Probability Integral Transform

Consider a specific change of variables related to the cumulative for some arbitrary *f(x)*

$$y(x) = \int_{-\infty}^{x} f(x')dx'$$

Using our general change of variables formula:

$$f(x) = g(y) \left| \frac{dy}{dx} \right|$$

We find for this case the Jacobian factor is

$$\left| \frac{dy}{dx} \right| = f(x)$$

Thus $g(y) = 1$

Probability Integral Transform

"…seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years"
– Egon Pearson (1938)

Given continuous $x \in (a,b)$, and its pdf $p(x)$, let

$$y(x) = \int_a^x p(x')\, dx' .$$

Then $y \in (0,1)$ and $p(y) = 1$ (uniform) for all y. (!)

So there always exists a metric in which the pdf is uniform.

***Many* issues become more clear (or trivial) after this transformation*. (If x is discrete, some complications.)**

The specification of a Bayesian prior pdf $p(\mu)$ for parameter μ is equivalent to the choice of the metric $f(\mu)$ in which the pdf is uniform. This is a *deep* issue, not always recognized as such by users of flat prior pdf's in HEP!

***And the inverse transformation provides for efficient M.C. generation of p(x) starting from RAN().**

Modeling:
The Scientific Narrative

Building a Model of the Data

Before one can discuss statistical tests, one must have a "**model**" for the data.

- by "model", I mean the full structure of P(data | parameters)
  - holding parameters fixed gives a PDF for data
  - provides ability to generate pseudo-data (via Monte Carlo)
  - holding data fixed gives a **likelihood function** for parameters
    - note, likelihood function is not as general as the full model because it doesn't allow you to generate pseudo-data

Both Bayesian and Frequentist methods start with the model

- it's the objective part that everyone can agree on
- it's the place where our physics knowledge, understanding, and intuiting comes in
- building a better model is the best way to improve your statistical procedure

The Scientific Narrative

The model can be seen as a quantitative summary of the analysis

- If you were asked to justify your modeling, you would tell a **story** about why you know what you know
  - based on previous results and studies performed along the way
- the quality of the result is largely tied to how convincing this story is and how tightly it is connected to model

I will describe a few “narrative styles”

- The “Monte Carlo Simulation” narrative
- The “Data Driven” narrative
- The “Effective Modeling” narrative

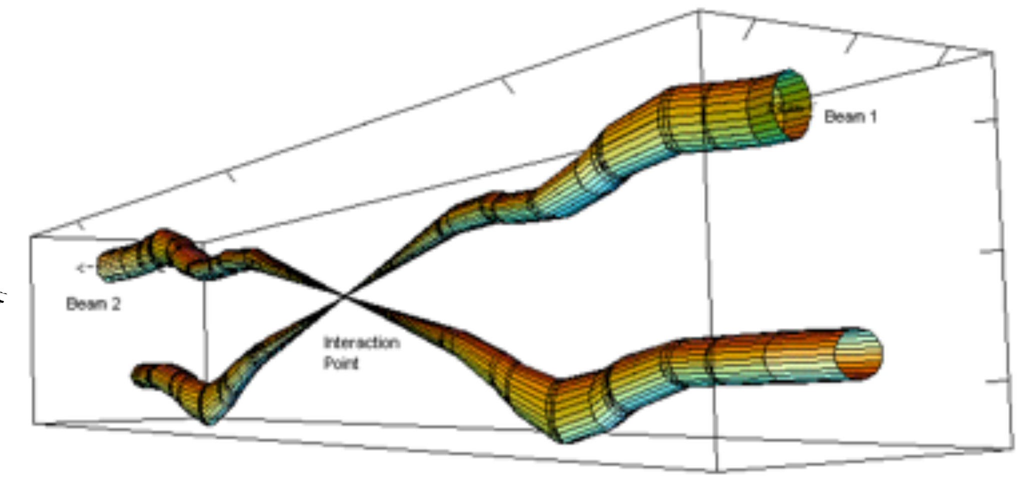
Real-life analyses often use a mixture of these

1) The language of the Standard Model is Quantum Field Theory
Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$



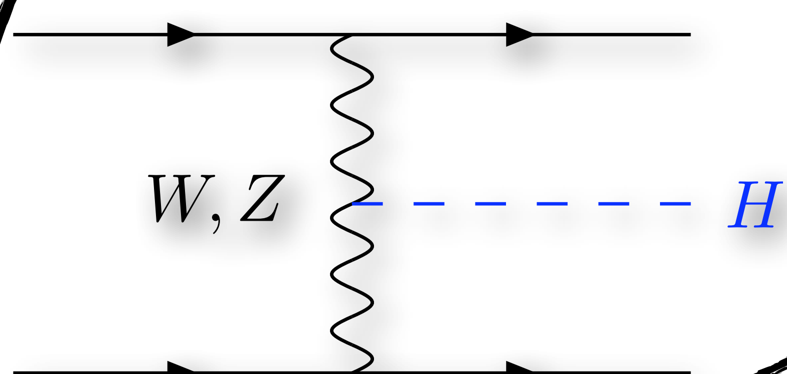
Relative beam sizes around IP1 (Atlas) in collision

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

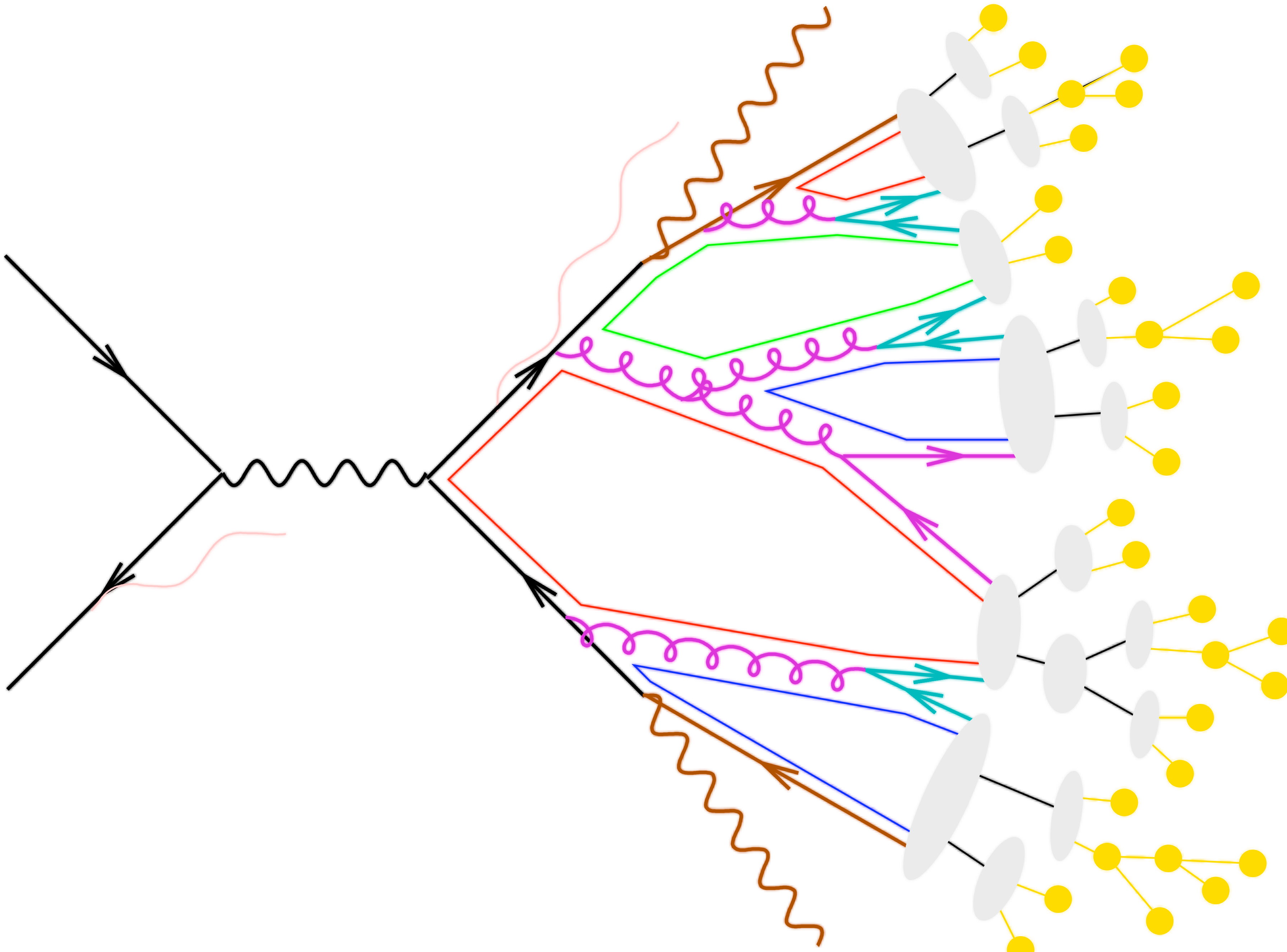
$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



THE SIMULATION NARRATIVE

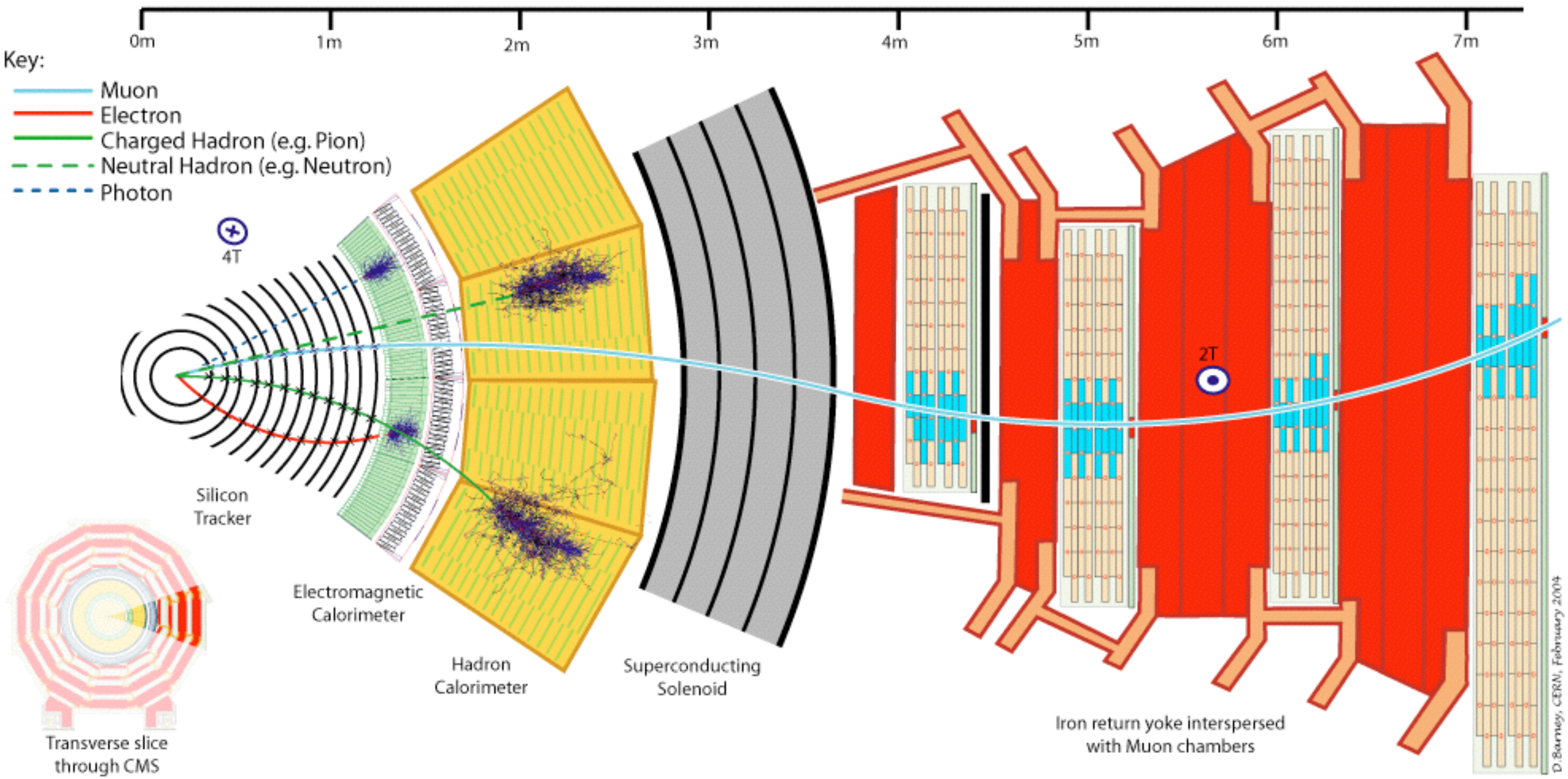
2) a) Perturbation theory used to systematically approximate the theory.
b) splitting functions, Sudokov form factors, and hadronization models
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**



- hard scattering
- (QED) initial/final state radiation
- partonic decays, e.g. $t \to bW$
- parton shower evolution
- nonperturbative gluon splitting
- colour singlets
- colourless clusters
- cluster fission
- cluster $\to$ hadrons
- hadronic decays

THE SIMULATION NARRATIVE

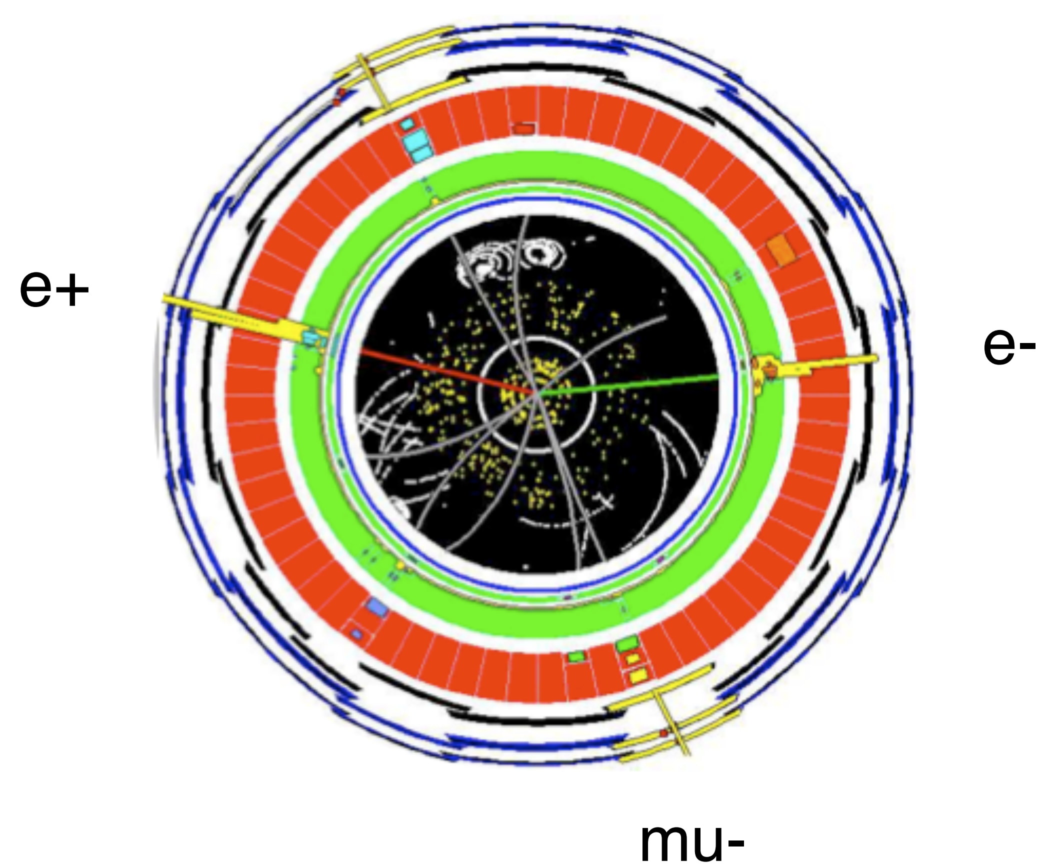
3) Next, the interaction of outgoing particles with the detector is simulated.
Detailed simulations of particle interactions with matter.
Accept/reject style Monte Carlo integration of very complicated function
P(detector readout | initial particles)

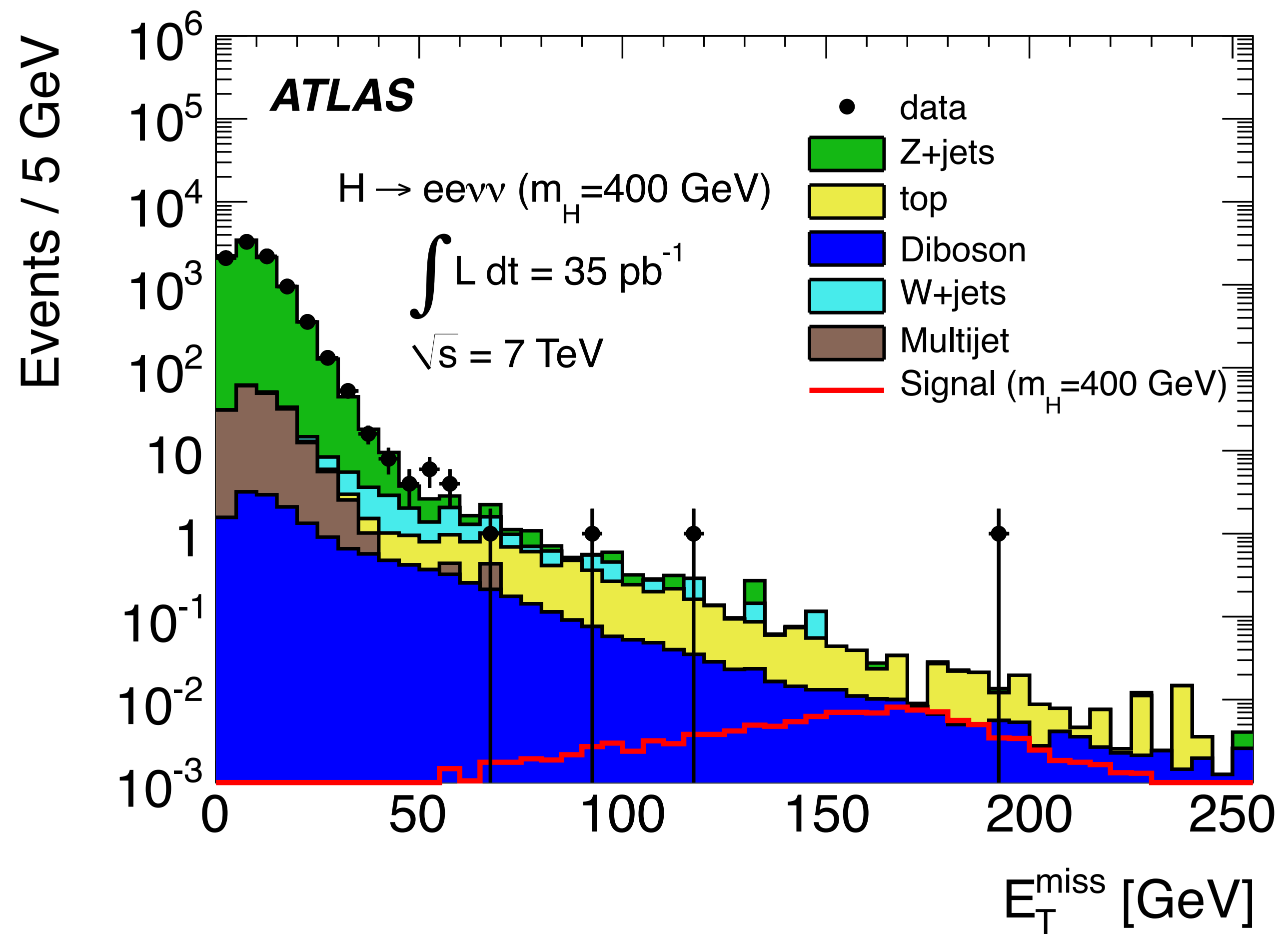


THE SIMULATION NARRATIVE

4) From the simulated response of the detector, we run reconstruction algorithms on the simulated data as if it were from real data. This allows us to look at distribution of any observable that we can measure in data.

P(observable | detector readout)



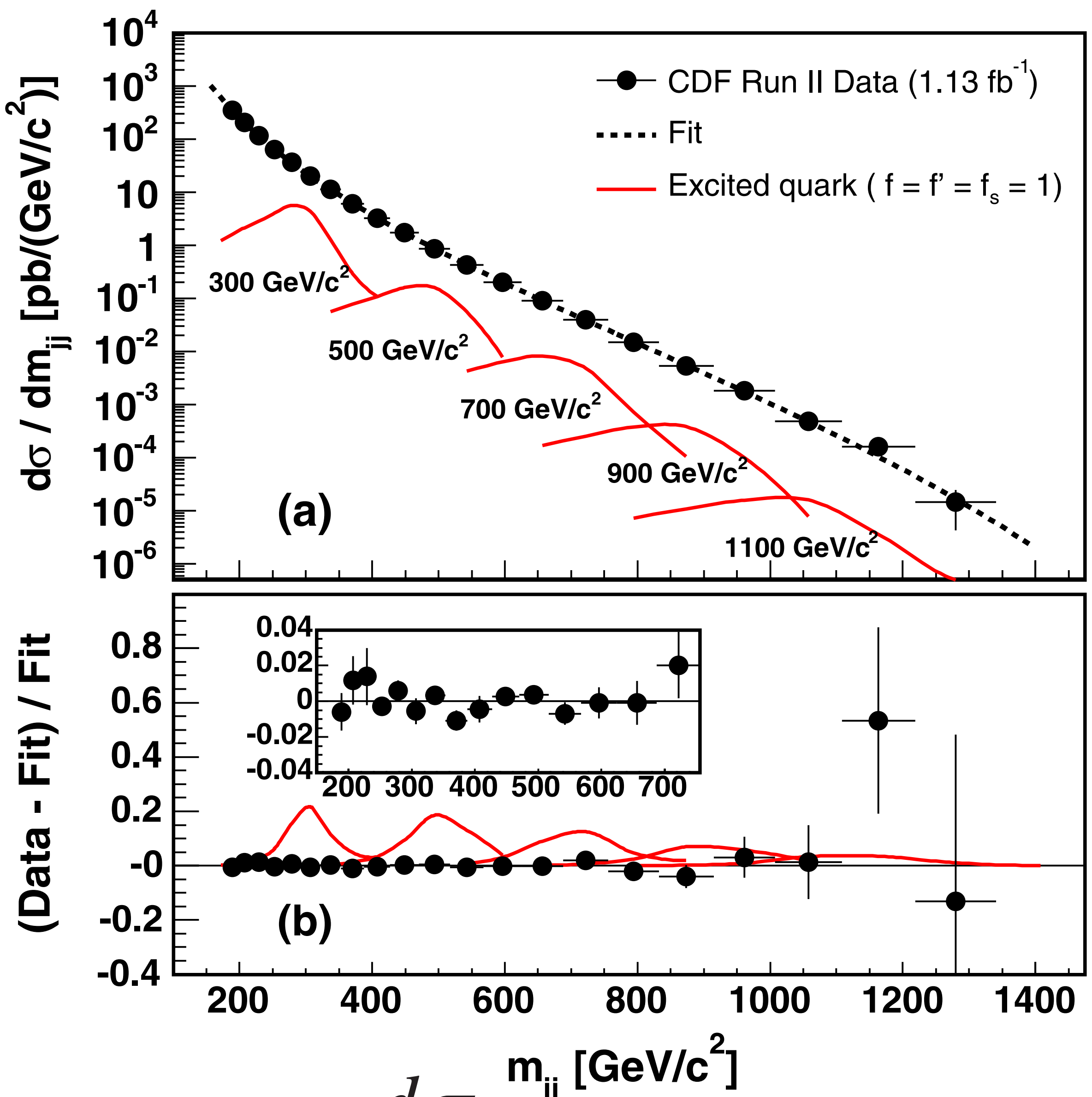


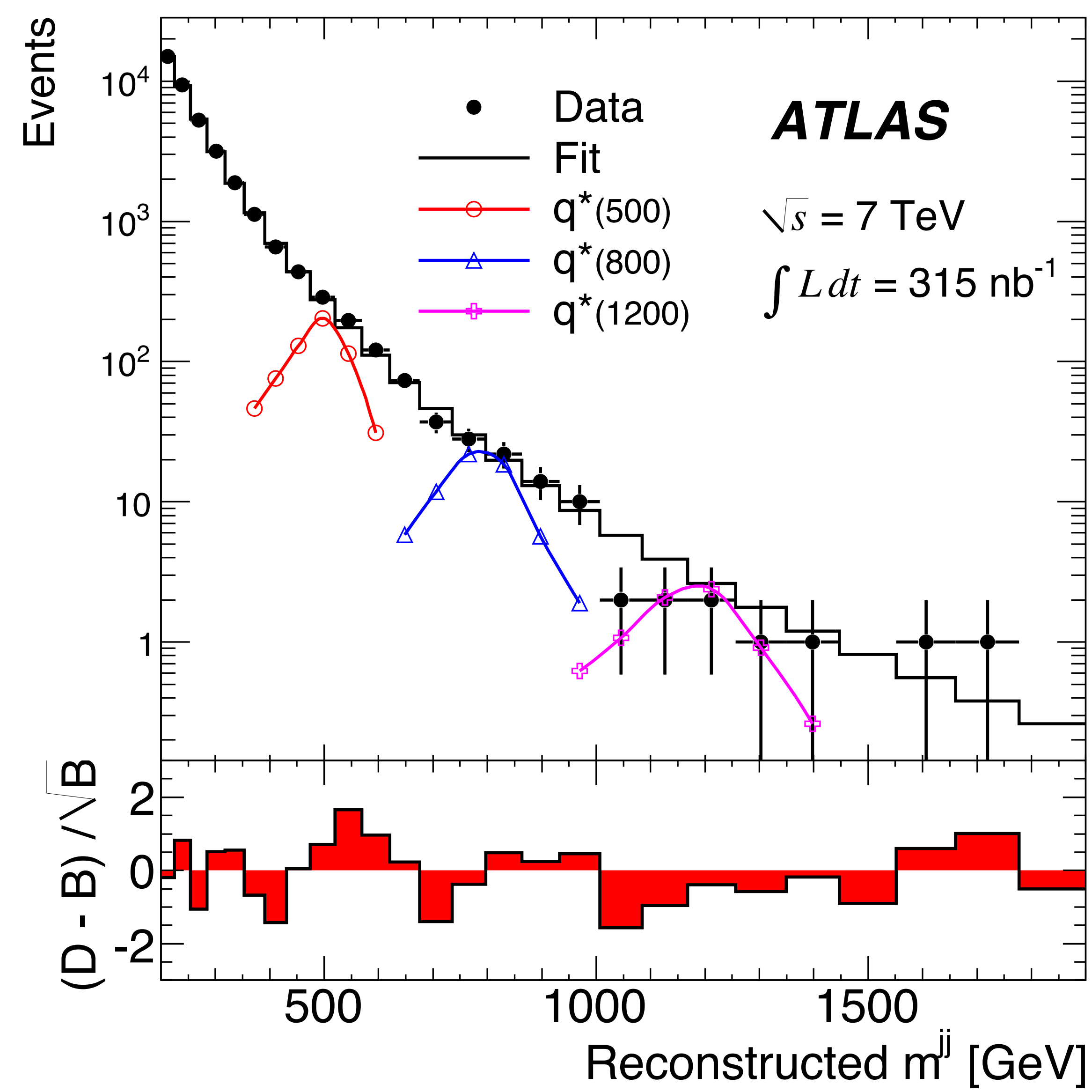
The Effective Model Narrative

In contrast, one can describe a distribution with some parametric function

- “we fit background to a polynomial”, exponential, ...
- While this is convenient and the fit may be good, the narrative is weak

PHYSICAL REVIEW D **79,** 112002 (2009)



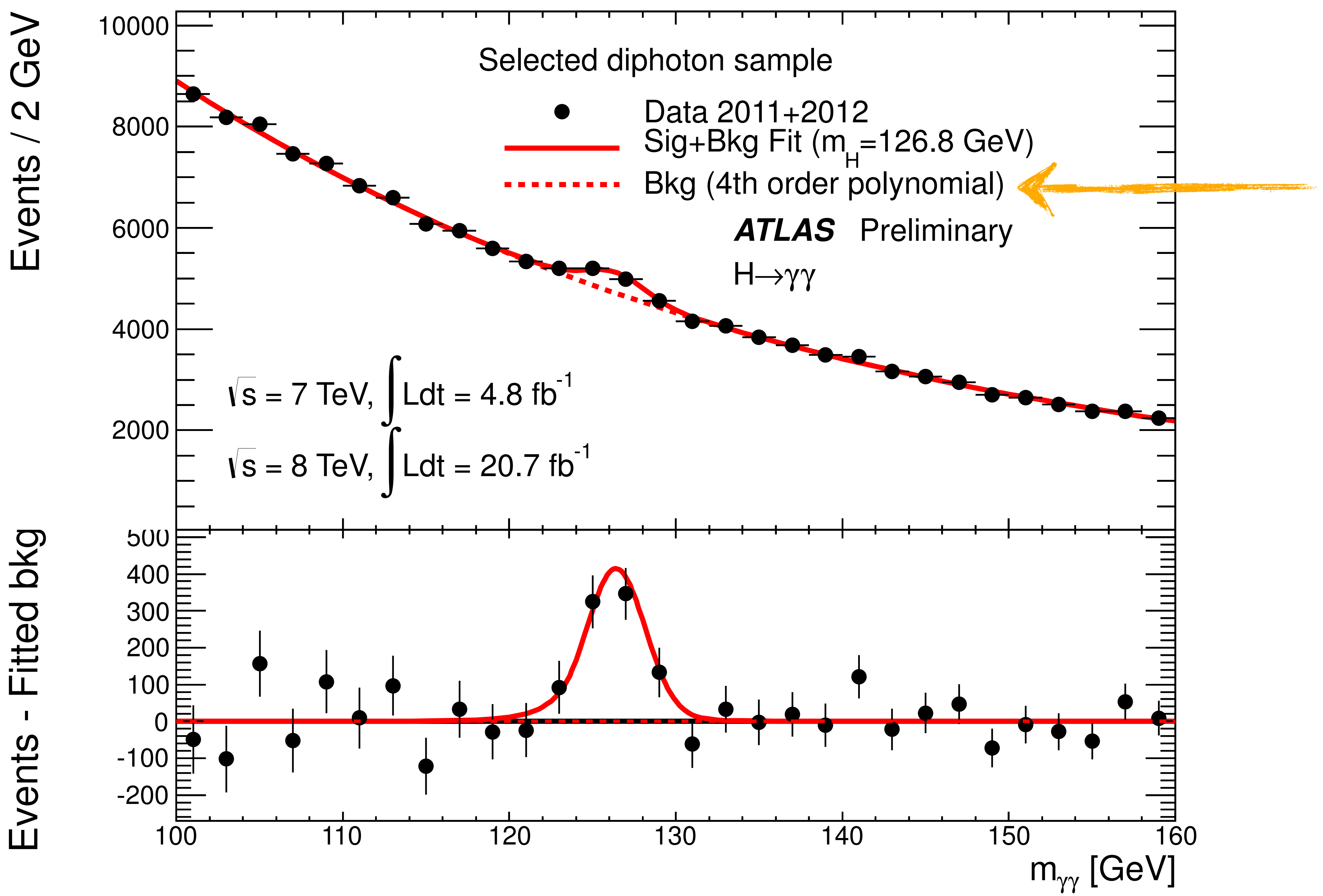


$$\frac{d\sigma}{dm_{jj}} = p_0(1-x)^{p_1}/x^{p_2+p_3\cdot\ln(x)}, \qquad x = m_{jj}/\sqrt{s},$$

The Effective Model Narrative

In contrast, one can describe a distribution with some parametric function

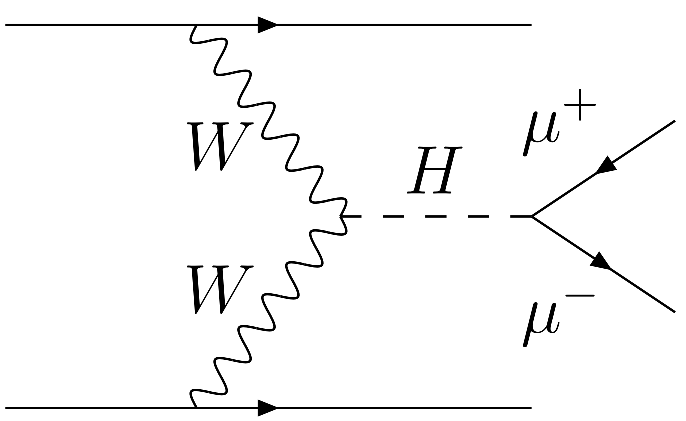
- “we fit background to a polynomial”, exponential, ...
- while this is convenient and the fit may be good, the narrative is weak
- often effective, parametric model is “validated” with simulation

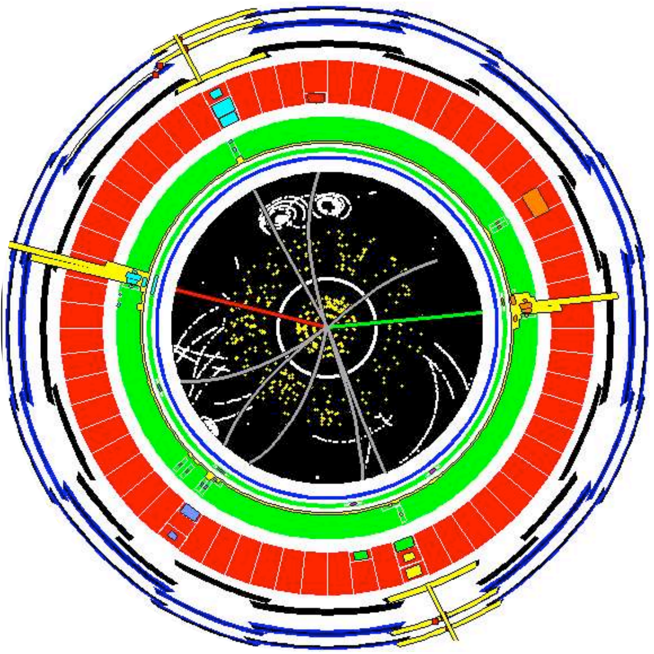


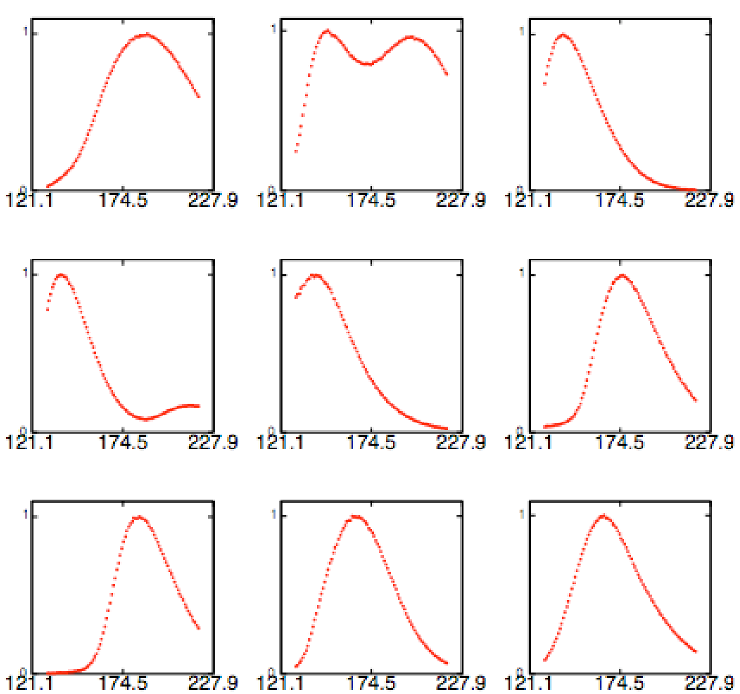
THE PARAMETRIZED RESPONSE NARRATIVE

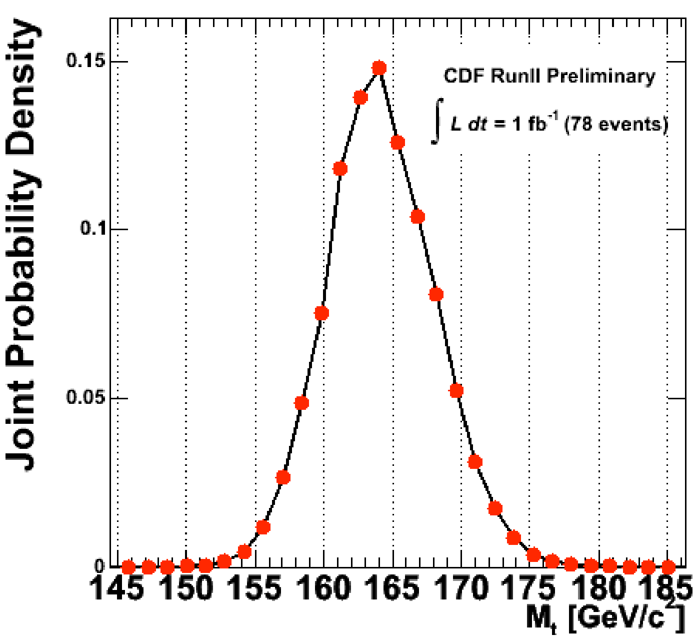
The Matrix-Element technique (aka MELA) is conceptually similar to the simulation narrative, but the detector response is parametrized.

- one still does integration over the unobserved “true” 4-momentum, but does not need to do much larger integration over interactions inside detector

$L(x|H_0) =$ 

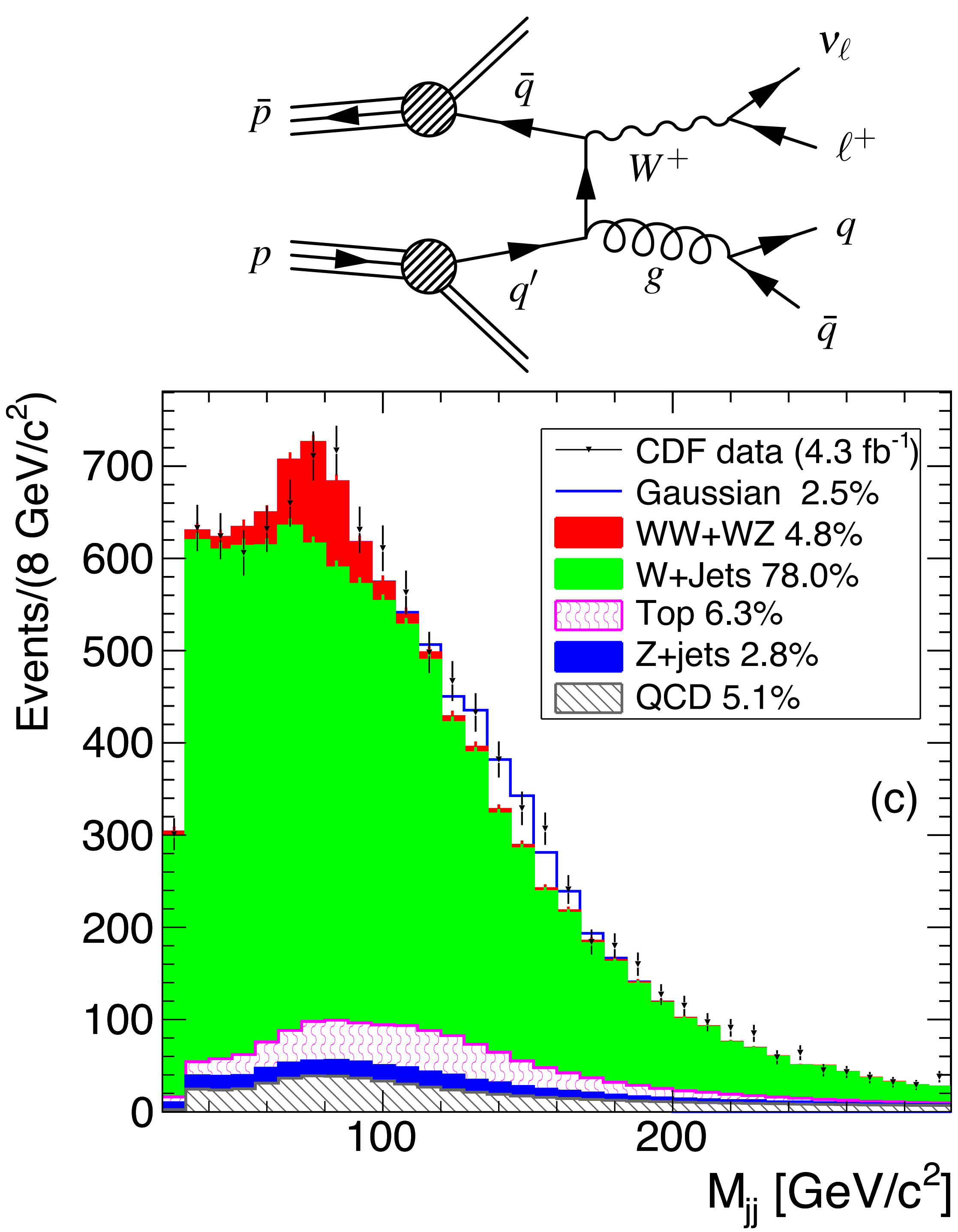
$\oplus$ 

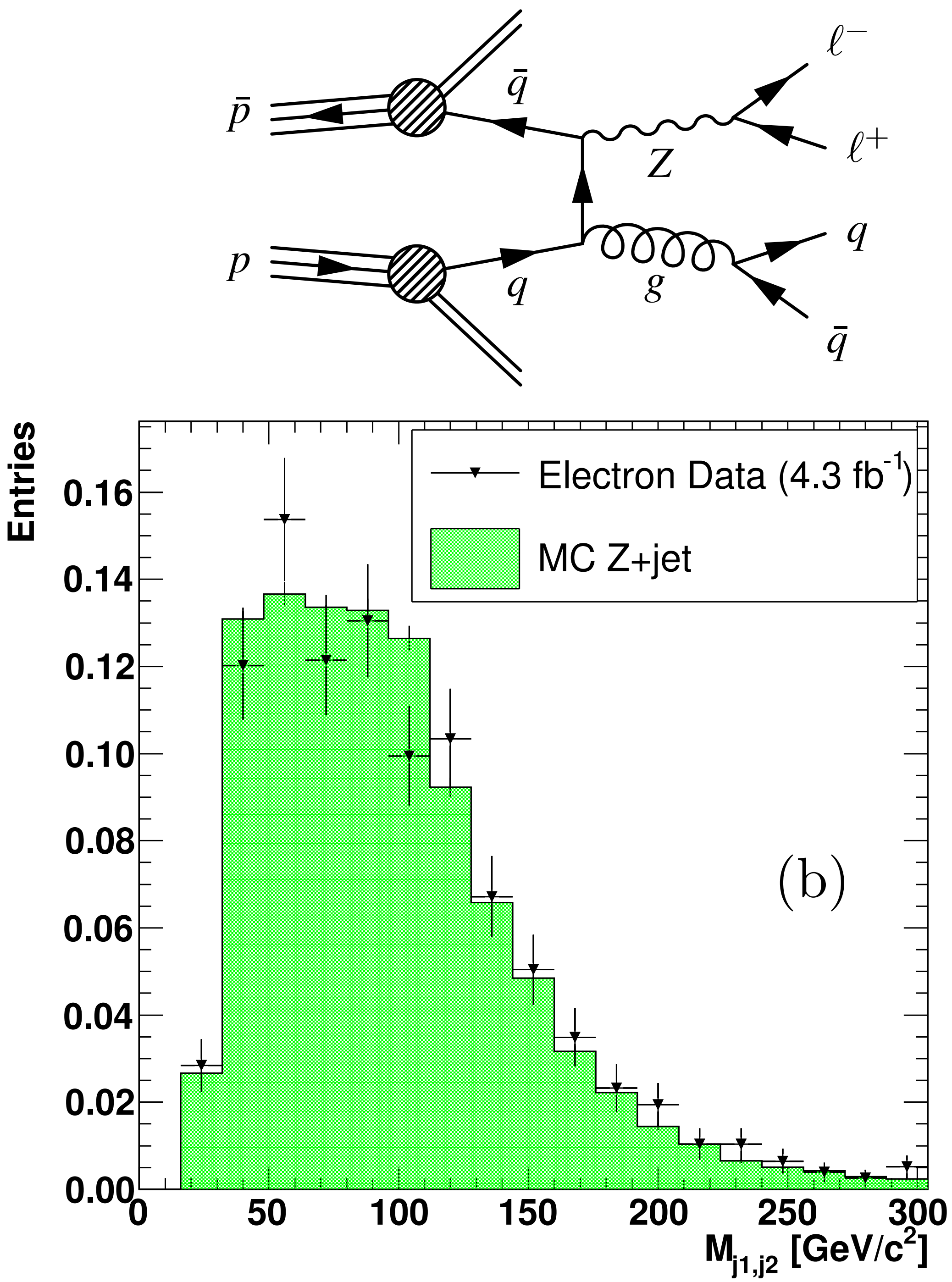




CHOICE: DATA DRIVEN VS. SIMULATION

In the case of the CDF bump, the Z+jets control sample provides a data-driven estimate, but limited statistics. Using the simulation narrative over the data-driven is a **choice**. If you trust that narrative, it's a good choice.

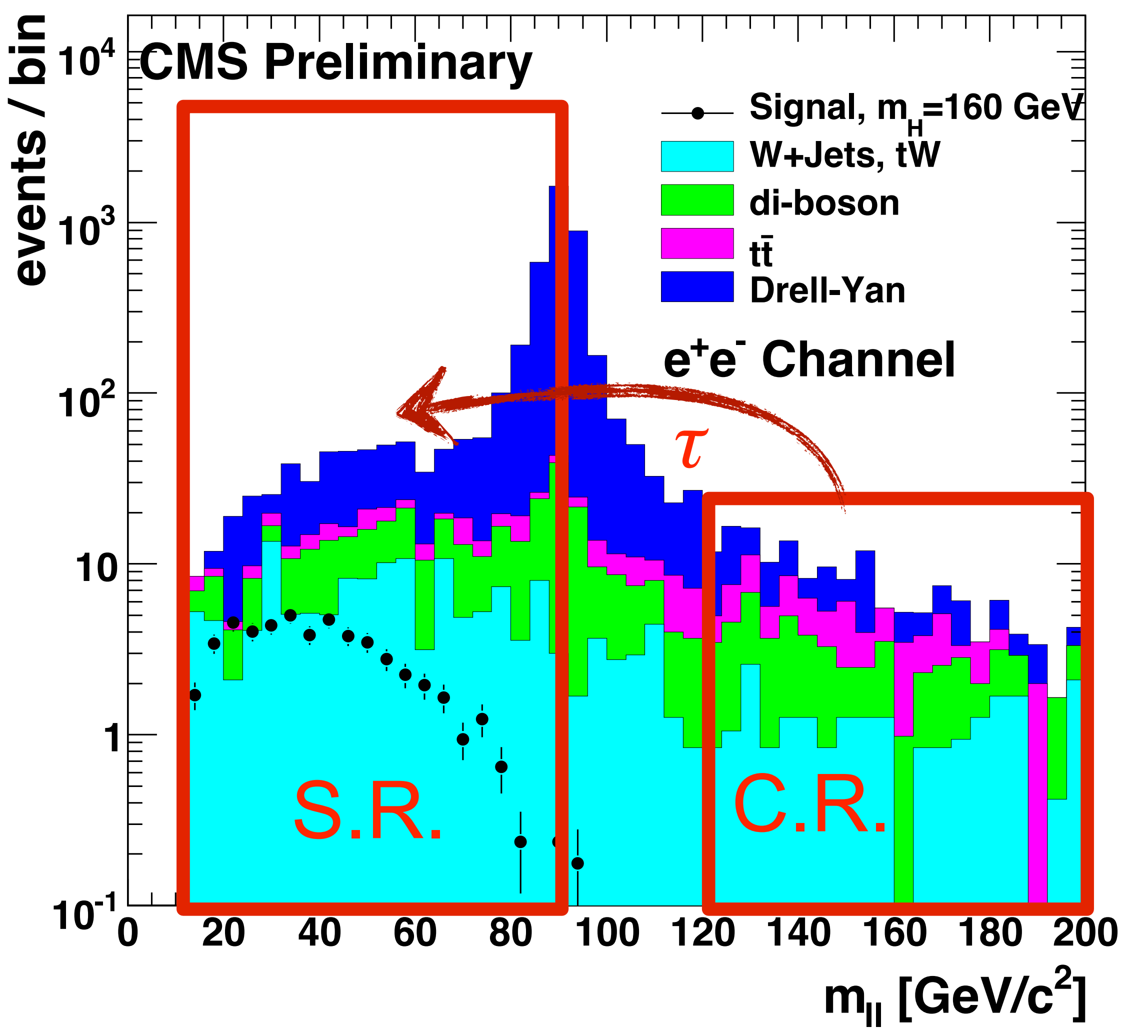




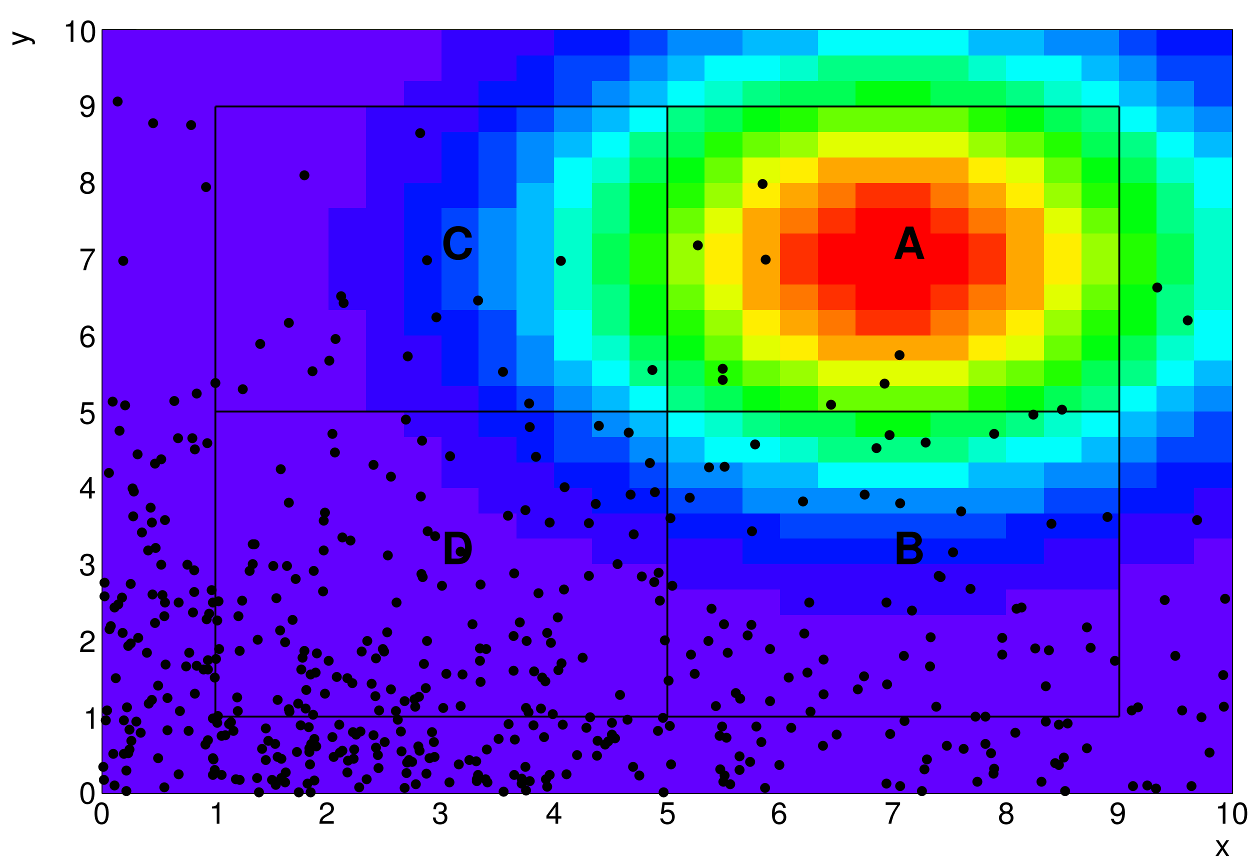
The data-driven narrative

Regions in the data with negligible signal expected used as control samples

- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties



"ABCD" method



What do we mean by uncertainty?

Let's consider a simplified problem that has been studied quite a bit to gain some insight into our more realistic and difficult problems

- number counting with background uncertainty
  - in our main measurement we observe n_{on} with s+b expected

$$\text{Pois}(n_{\text{on}}|s+b)$$

- and the background has some uncertainty
  - but what is "background uncertainty"? Where did it come from?
  - maybe we would say background is known to 10% or that it has some pdf π(b)
    - then we often do a **smearing** of the background:

$$P(n_{\text{on}}|s) = \int db\, \text{Pois}(n_{\text{on}}|s+b)\, \pi(b),$$

  - Where does π(b) come from?
    - did you realize that this is a Bayesian procedure that depends on some prior assumption about what b is?

The "ON/OFF" problem

Now let's say that the background was estimated from some control region or sideband measurement.

- We can treat these two measurements simultaneously:
  - main measurement: observe n_{on} with $s+b$ expected
  - sideband measurement: observe n_{off} with τb expected

$$\underbrace{P(n_{\rm on}, n_{\rm off}|s,b)}_{\text{joint model}} = \underbrace{\text{Pois}(n_{\rm on}|s+b)}_{\text{main measurement}}\;\underbrace{\text{Pois}(n_{\rm off}|\tau b)}_{\text{sideband}}$$

  - In this approach "background uncertainty" is a statistical error
  - justification and accounting of background uncertainty is much more clear

How does this relate to the smearing approach?

$$P(n_{\rm on}|s) = \int db\,\text{Pois}(n_{\rm on}|s+b)\,\pi(b),$$

- while $\pi(b)$ is based on data, it still depends on some original prior $\eta(b)$

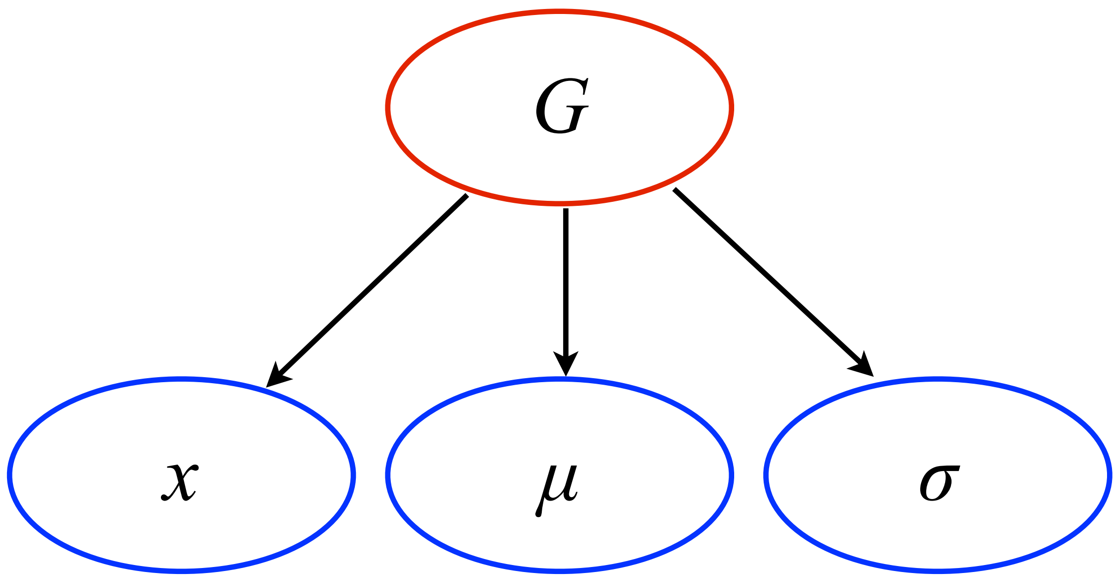
$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

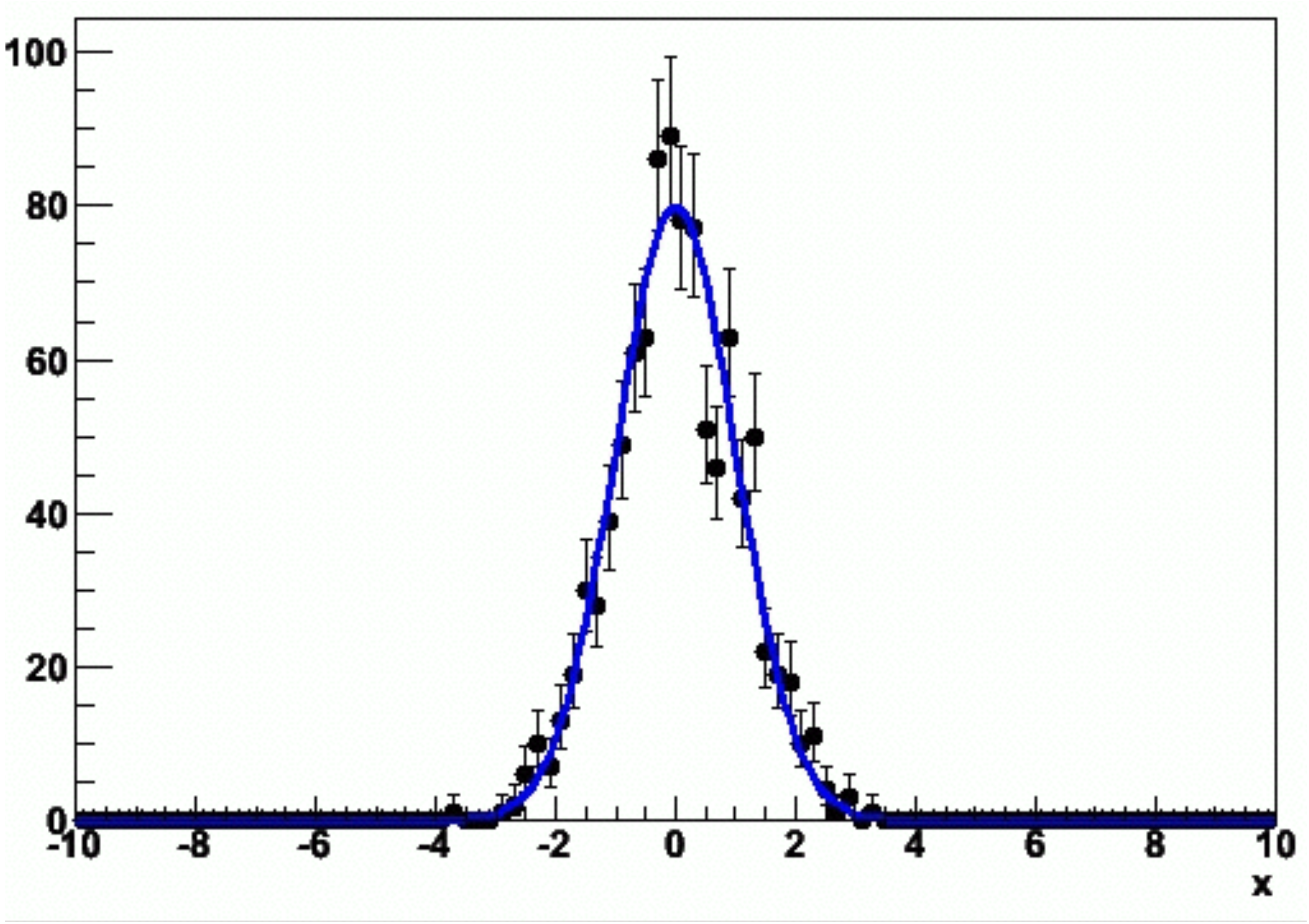
A GENERAL PURPOSE STATISTICAL MODEL

VISUALIZING PROBABILITY MODELS

I will represent PDFs graphically as below (directed acyclic graph)

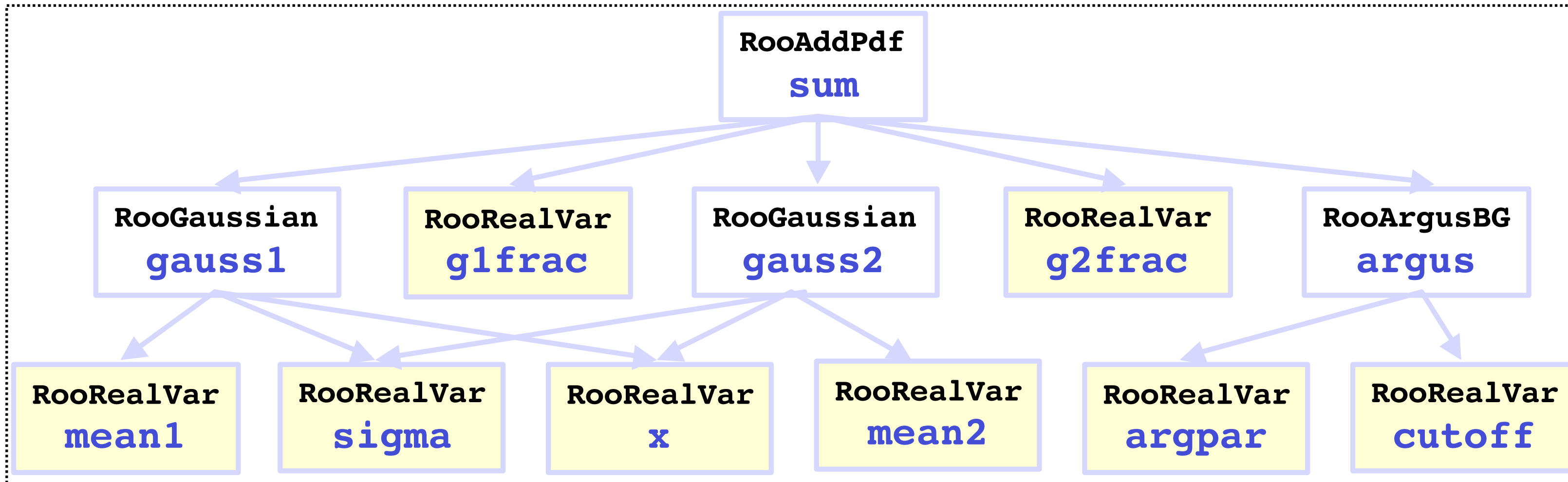
- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- every node is a real-valued function of the nodes below



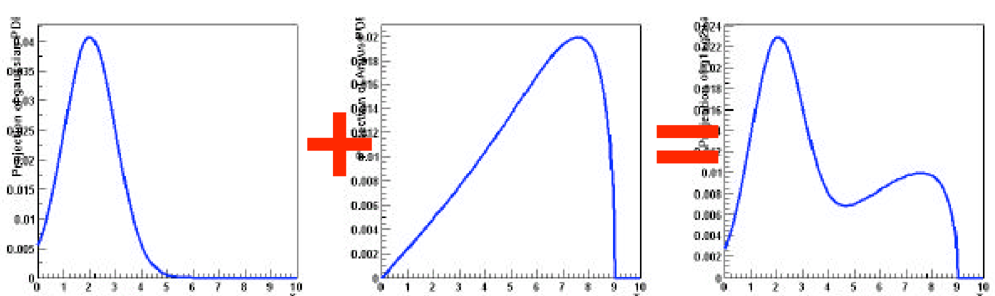


RooFit: A data modeling toolkit

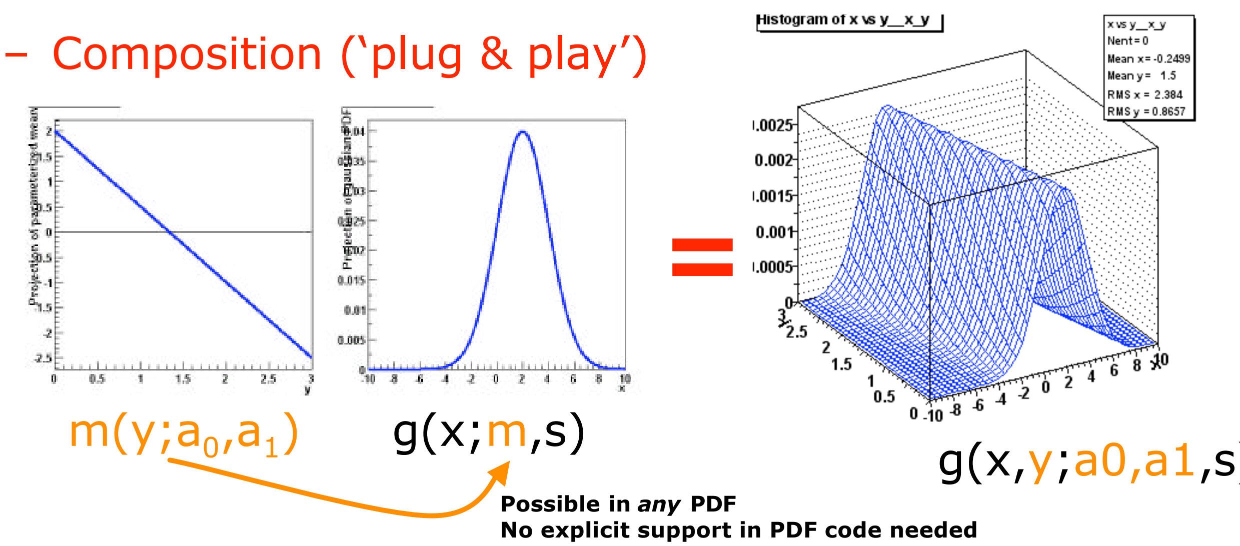
RooFit is a major tool developed at BaBar for data modeling.
RooStats provides higher-level statistical tools based on these PDFs.



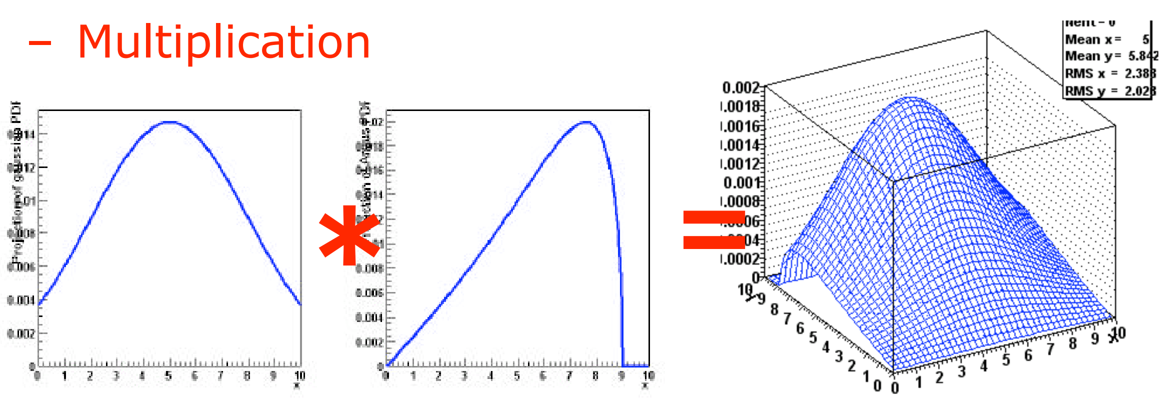
– Addition



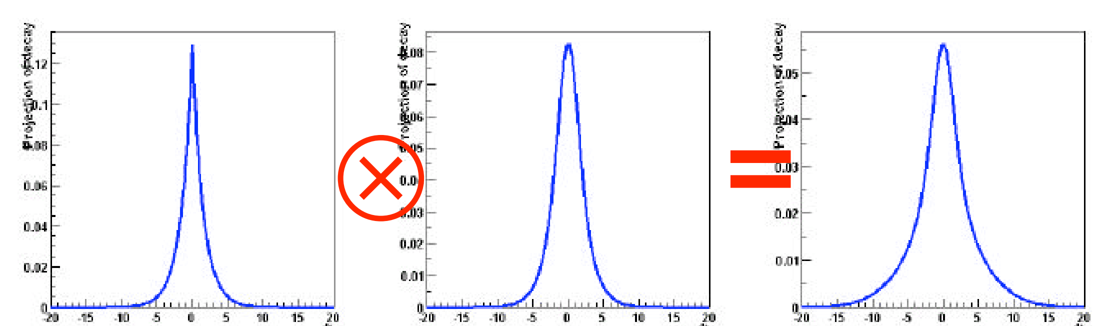
– Composition ('plug & play')



– Multiplication



– Convolution



Wouter Verkerke,

Wouter Verkerke, UCSB

Marked Poisson Process

Channel: a subset of the data defined by some selection requirements.

- eg. all events with 4 electrons with energy > 10 GeV
- n: number of events observed in the channel
- ν: number of events expected in the channel

Discriminating variable: a property of those events that can be measured and which helps discriminate the signal from background

- eg. the invariant mass of two particles
- $f(x)$: the p.d.f. of the discriminating variable x

$$\mathcal{D} = \{x_1, \ldots, x_n\}$$

Marked Poisson Process / Extended Likelihood:

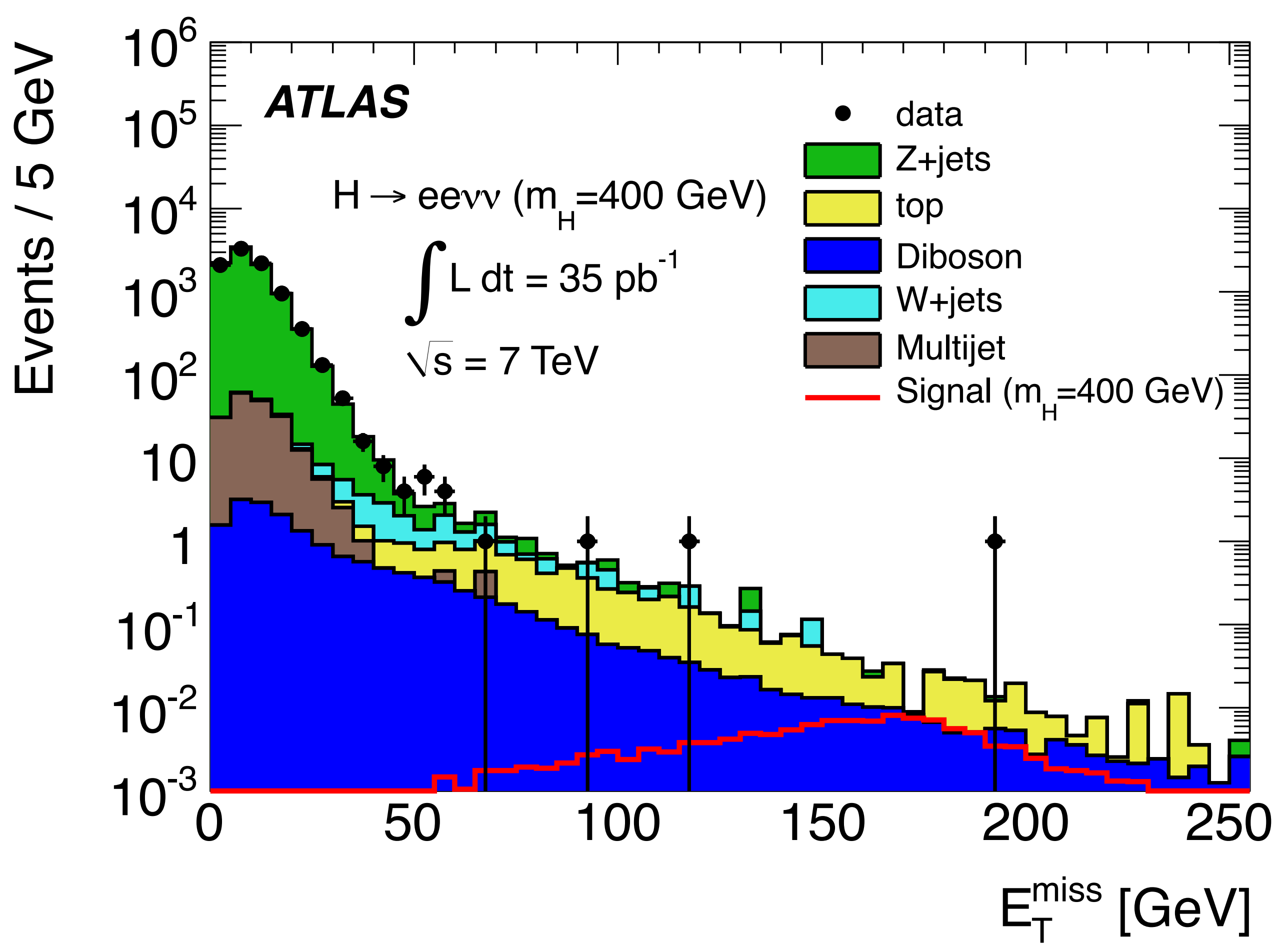
$$\mathbf{f}(\mathcal{D}|\nu) = \text{Pois}(n|\nu) \prod_{e=1}^{n} f(x_e)$$

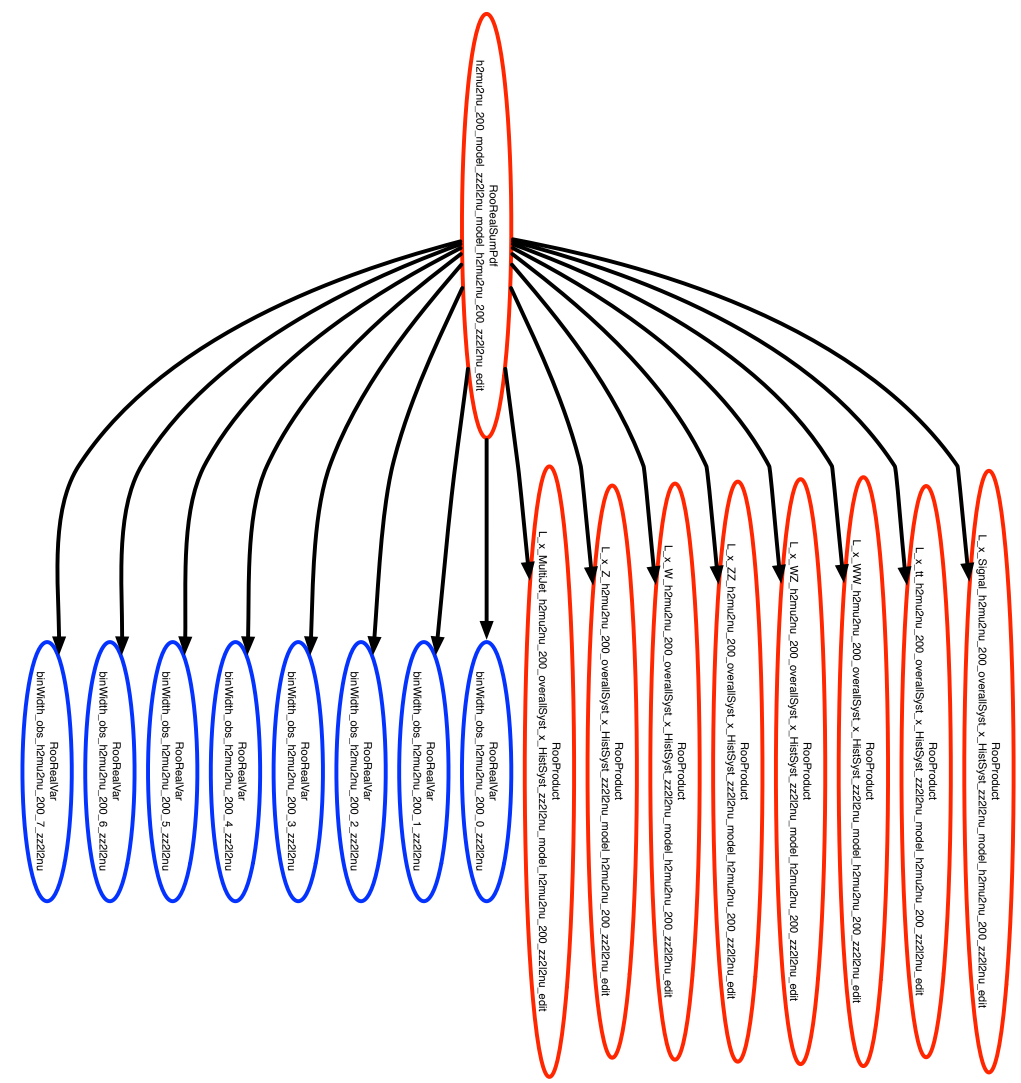
MIXTURE MODEL

Sample: a sample of simulated events corresponding to particular type interaction that populates the channel.

- statisticians call this a mixture model

$$f(x) = \frac{1}{\nu_{\rm tot}} \sum_{s\in{\rm samples}} \nu_s f_s(x) \;, \qquad \nu_{\rm tot} = \sum_{s\in{\rm samples}} \nu_s$$





Parametrizing the model $\boldsymbol{\alpha} = (\mu, \boldsymbol{\theta})$

Parameters of interest (*μ*): parameters of the theory that modify the rates and shapes of the distributions, eg.

- the mass of a hypothesized particle
- the "signal strength" μ=0 no signal, μ=1 predicted signal rate

Nuisance parameters (*θ* or *α_p*): associated to uncertainty in:

- response of the detector (calibration)
- phenomenological model of interaction in non-perturbative regime

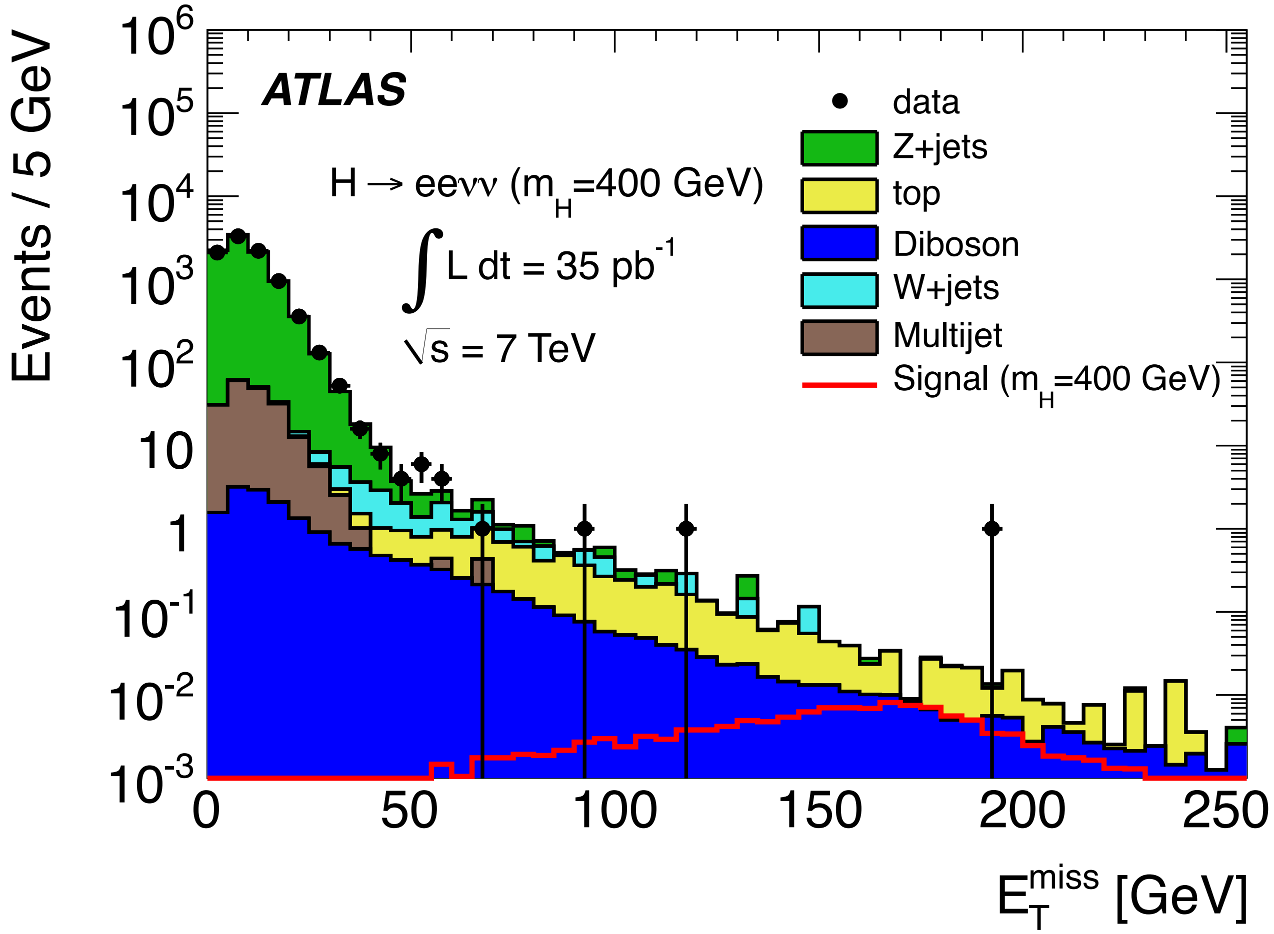
Lead to a parametrized model: $\nu \to \nu(\boldsymbol{\alpha}), f(x) \to f(x|\boldsymbol{\alpha})$

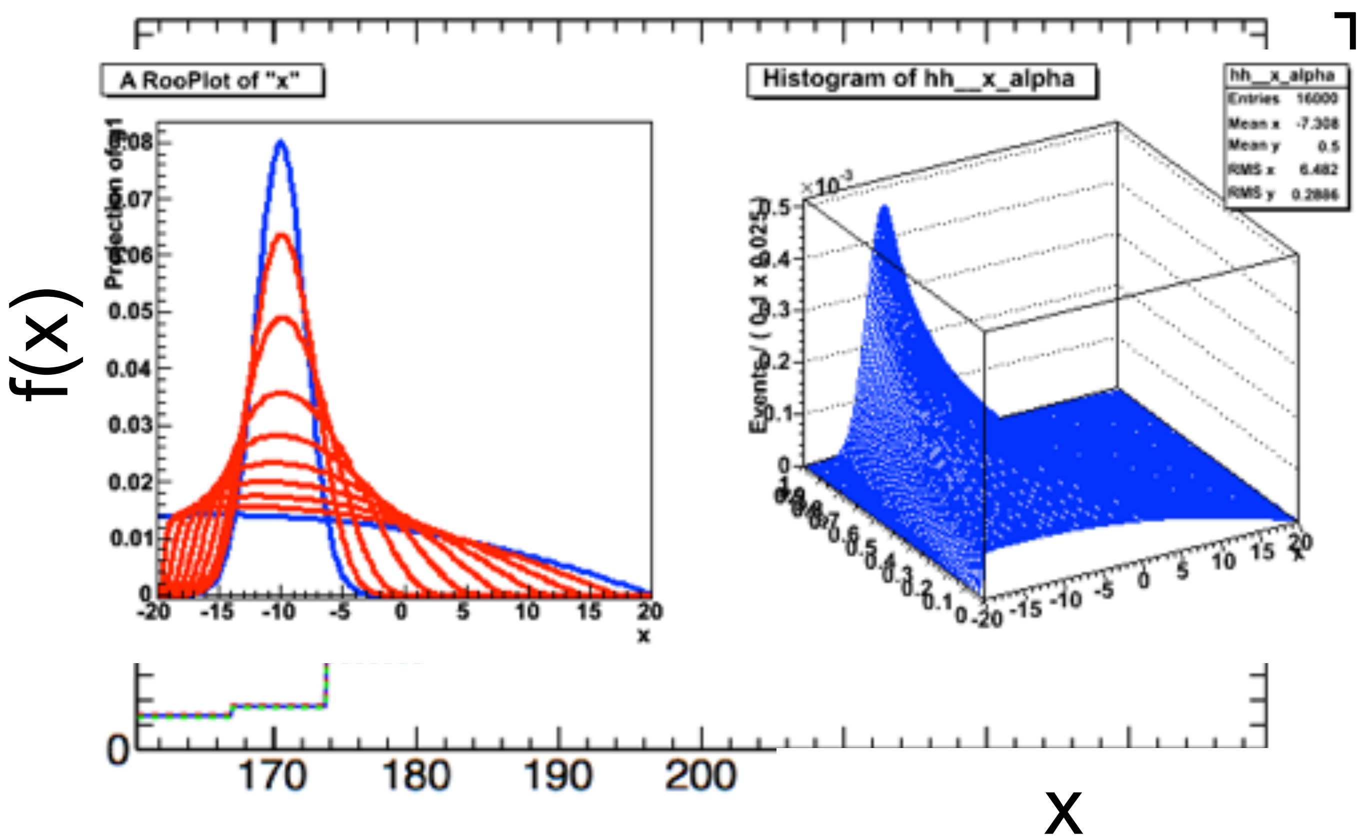
$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \text{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

INCORPORATING SYSTEMATIC EFFECTS

Tabulate effect of individual variations of sources of systematic uncertainty

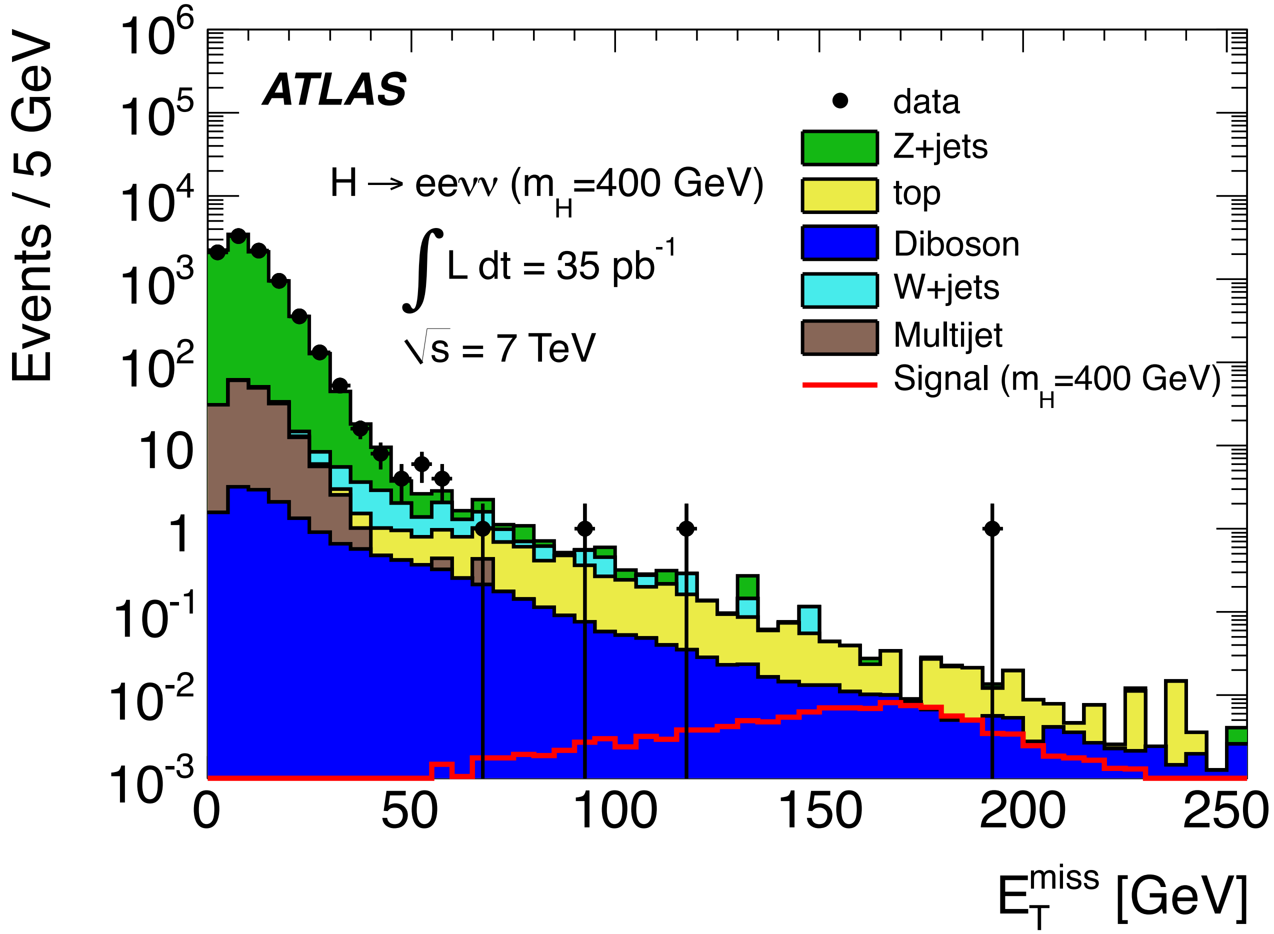
- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

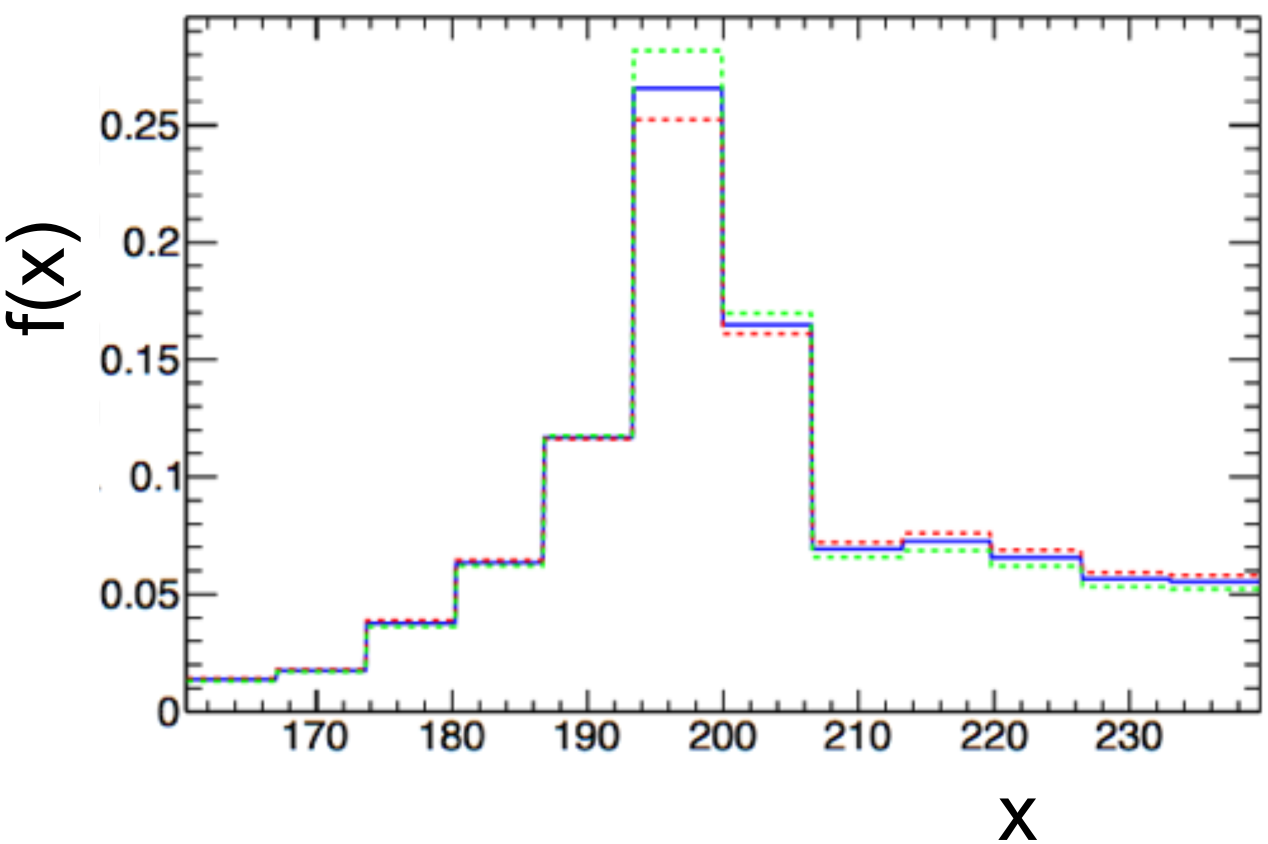


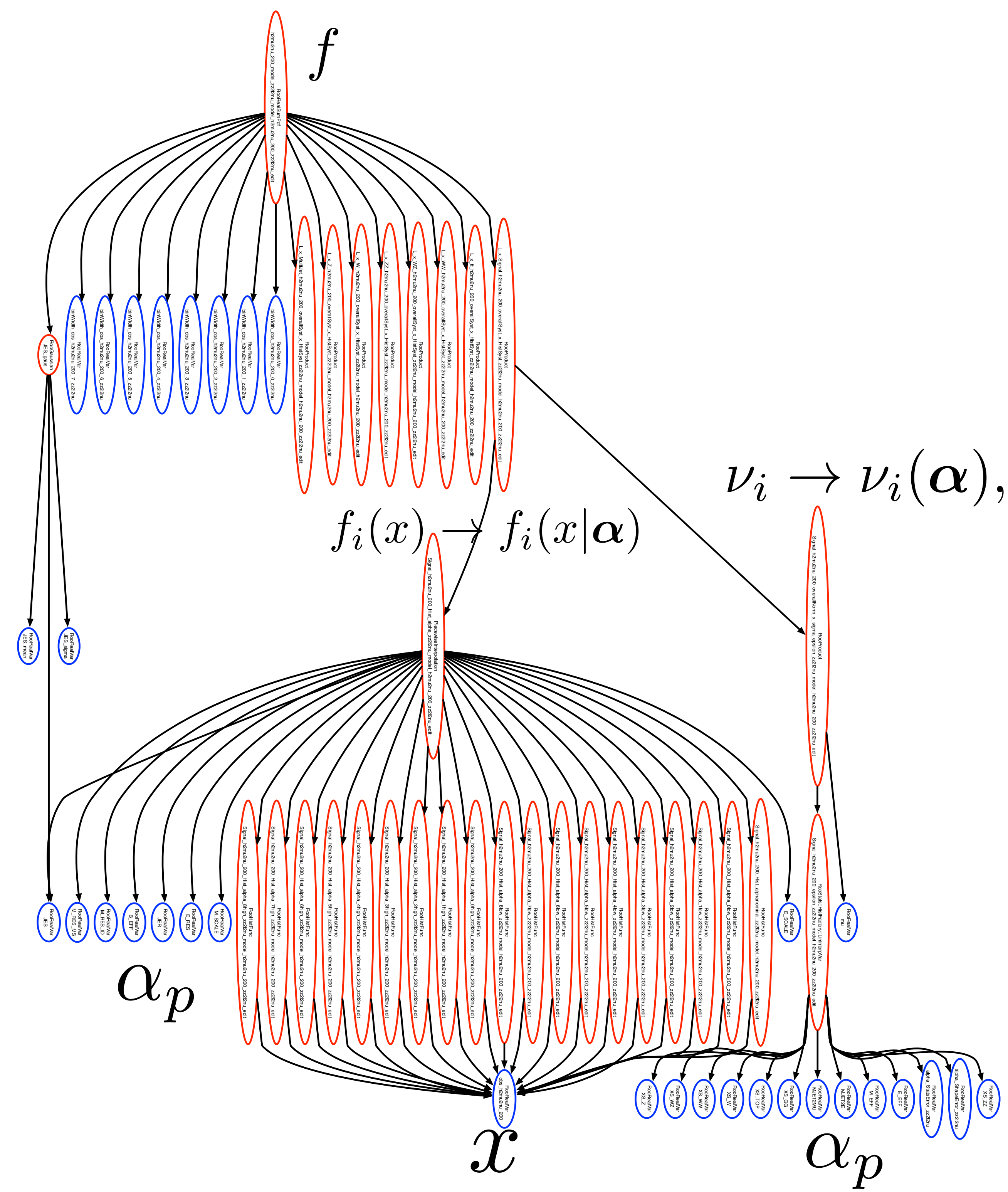


$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \text{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

VISUALIZING THE MODEL FOR ONE CHANNEL

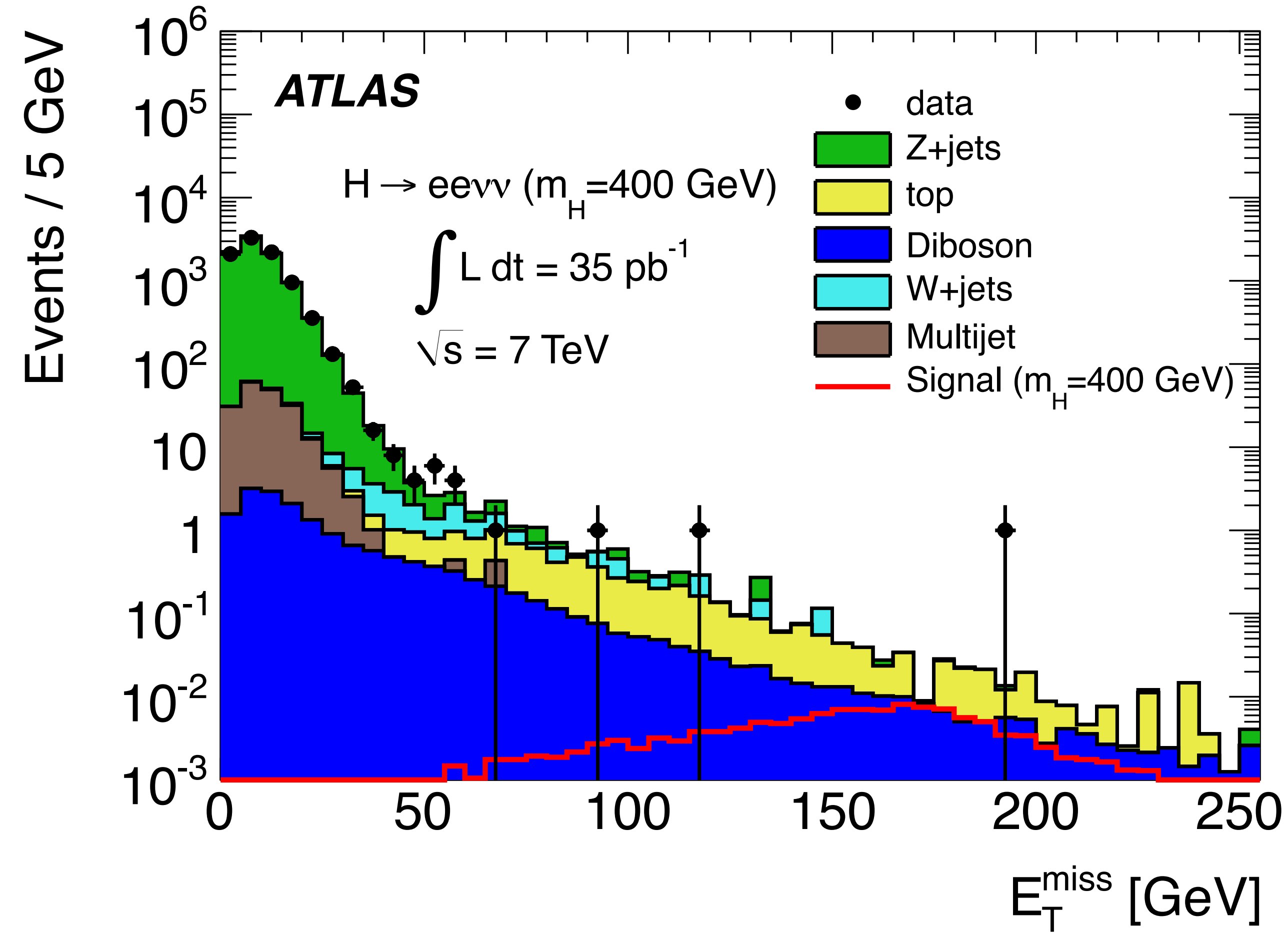


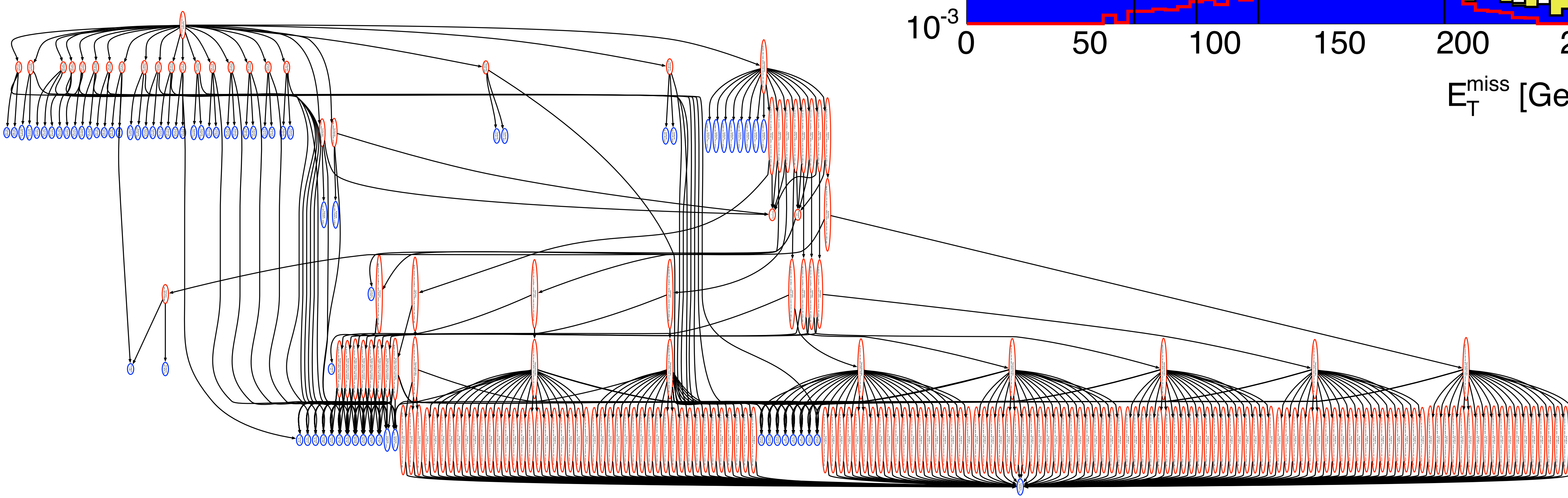




VISUALIZING THE MODEL FOR ONE CHANNEL

After parametrizing each component of the mixture model, the pdf for a single channel might look like this





Simultaneous Multi-Channel Model

Simultaneous Multi-Channel Model: Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

$$\mathbf{f}_{\rm sim}(\mathcal{D}_{\rm sim}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right]$$

where $\mathcal{D}_{\rm sim} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\rm max}}\}$

Control Regions: Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

- attempt to describe systematics in a statistical language
- Prototypical Example: “on/off” problem with unknown ν_b

$$\mathbf{f}(n, m|\mu, \nu_b) = \underbrace{\text{Pois}(n|\mu + \nu_b)}_{\text{signal region}} \cdot \underbrace{\text{Pois}(m|\tau\nu_b)}_{\text{control region}}$$

Constraint terms

Often detailed statistical model for auxiliary measurements that measure certain nuisance parameters are not available.

- one typically has MLE for α_p, denoted a_p and standard error

Constraint Terms: are idealized pdfs for the MLE.

$$f_p(a_p|\alpha_p) \quad \text{for} \quad p \in \mathbb{S}$$

- common choices are Gaussian, Poisson, and log-normal
- New: careful to write constraint term a frequentist way
- Previously: $\pi(\alpha_p|a_p) = f_p(a_p|\alpha_p)\eta(\alpha_p)$ with uniform η

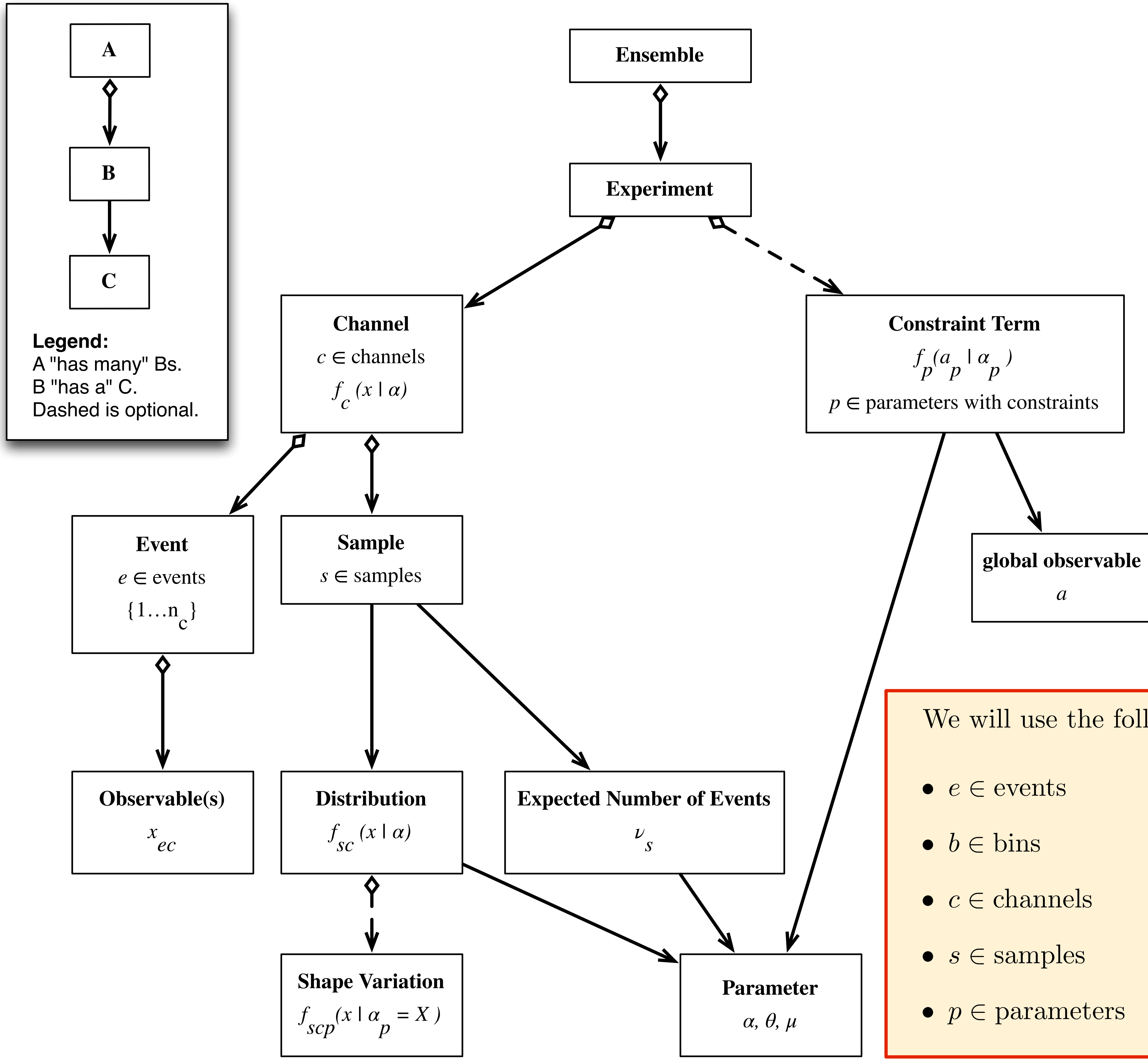
Simultaneous Multi-Channel Model with constraints:

$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p)$$

where

$$\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\max}}\}, \quad \mathcal{G} = \{a_p\} \quad \text{for} \quad p \in \mathbb{S}$$

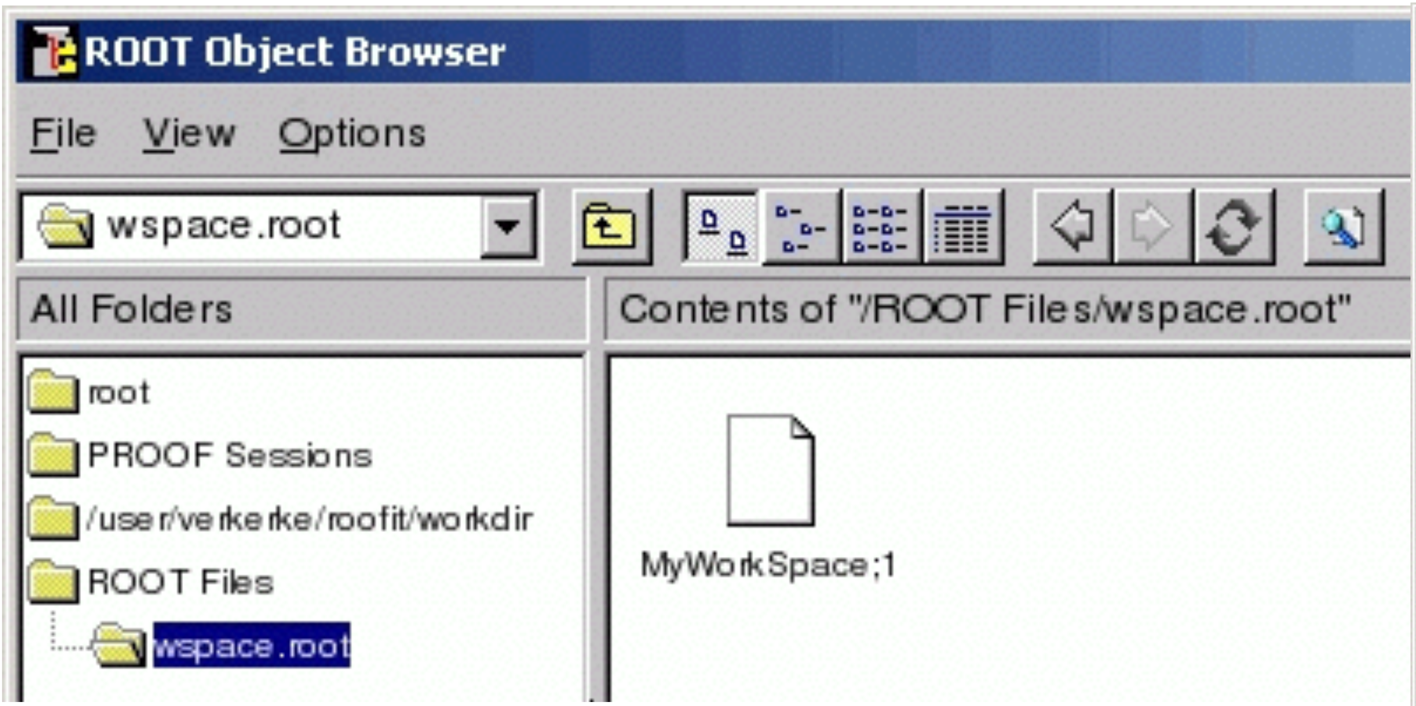
CONCEPTUAL BUILDING BLOCKS

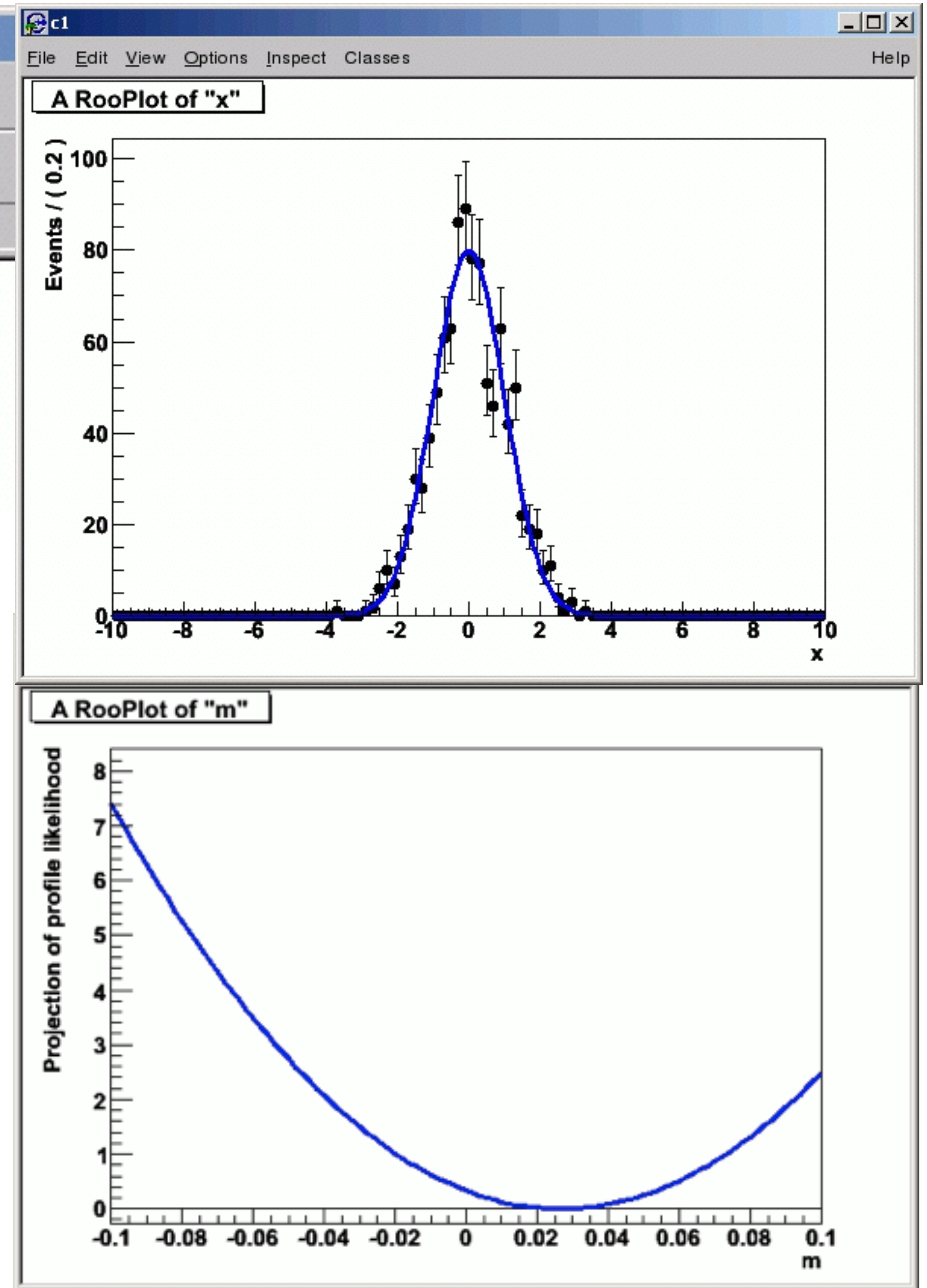


We will use the following mnemonic index conventions:

- $e \in$ events
- $b \in$ bins
- $c \in$ channels
- $s \in$ samples
- $p \in$ parameters

EXAMPLE OF DIGITAL PUBLISHING





RooFit's Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Need this for combinations, as p-value is not sufficient information for a proper combination.

HistFactory

32 page documentation of HistFactory tool + manual

- currently a “living document”

Information | Discussion (0) | Files | Linkbacks

http://cds.cern.ch/record/1456844

Preprint

Report number	CERN-OPEN-2012-016
Title	**HistFactory: A tool for creating statistical models for use with RooFit and RooStats**
Author(s)	Cranmer, Kyle (New York U.) ; Lewis, George (New York U.) ; Moneta, Lorenzo (CERN) ; Shibata, Akira (New York U.) ; Verkerke, Wouter (NIKHEF, Amsterdam)
Collaboration	ROOT Collaboration
Abstract	The HistFactory is a tool to build parametrized probability density functions (pdfs) in the RooFit/RooStats framework based based on simple ROOT histograms organized in an XML file. The pdf has a restricted form, but it is sufficiently flexible to describe many analyses based on template histograms. The tool takes a modular approach to build complex pdfs from more primative conceptual building blocks. The resulting PDF is stored in a RooWorkspace which can be saved to and read from a ROOT file. This document describes the defaults and interface in HistFactory 5.32.

Combined ATLAS Higgs Search

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

- Models for individual channels come from about 11 sub-groups performing dedicated searches for specific Higgs decay modes
- In addition low-level performance groups provide tools for evaluating systematic effects and corresponding constraint terms

Higgs Decay	Subsequent Decay	Additional Sub-Channels	m_H Range	L [fb^{-1}]
$H \to \gamma\gamma$	–	9 sub-channels ($p_{T_t} \otimes \eta_\gamma \otimes$ conversion)	110-150	4.9
$H \to ZZ$	$\ell\ell\ell'\ell'$	$\{4e, 2e2\mu, 2\mu 2e, 4\mu\}$	110-600	4.8
	$\ell\ell\nu\nu$	$\{ee, \mu\mu\} \otimes$ {low pile-up, high pile-up}	200-280-600	4.7
	$\ell\ell qq$	{b-tagged, untagged}	200-300-600	4.7
$H \to WW$	$\ell\nu\ell\nu$	$\{ee, e\mu, \mu\mu\} \otimes$ {0-jet, 1-jet, VBF}	110-300-600	4.7
	$\ell\nu q\overline{q'}$	$\{e, \mu\} \otimes$ {0-jet, 1-jet}	300-600	4.7
$H \to \tau^+\tau^-$	$\ell\ell 4\nu$	$\{e\mu\} \otimes$ {0-jet} $\oplus$ {1-jet, VBF, VH}	110-150	4.7
	$\ell\tau_{\text{had}}3\nu$	$\{e, \mu\} \otimes$ {0-jet} $\otimes$ $\{E_T^{\text{miss}} \gtrless 20$ GeV$\}$ $\oplus$ $\{e, \mu\} \otimes$ {1-jet, VBF}	110-150	4.7
	$\tau_{\text{had}}\tau_{\text{had}}2\nu$	{1-jet}	110-150	4.7
$VH \to b\bar{b}$	$Z \to \nu\bar{\nu}$	$E_T^{\text{miss}} \in \{120-160, 160-200, \geq 200$ GeV$\}$	110-130	4.6
	$W \to \ell\nu$	$p_T^W \in \{< 50, 50-100, 100-200, \geq 200$ GeV$\}$	110-130	4.7
	$Z \to \ell\ell$	$p_T^Z \in \{< 50, 50-100, 100-200, \geq 200$ GeV$\}$	110-130	4.7

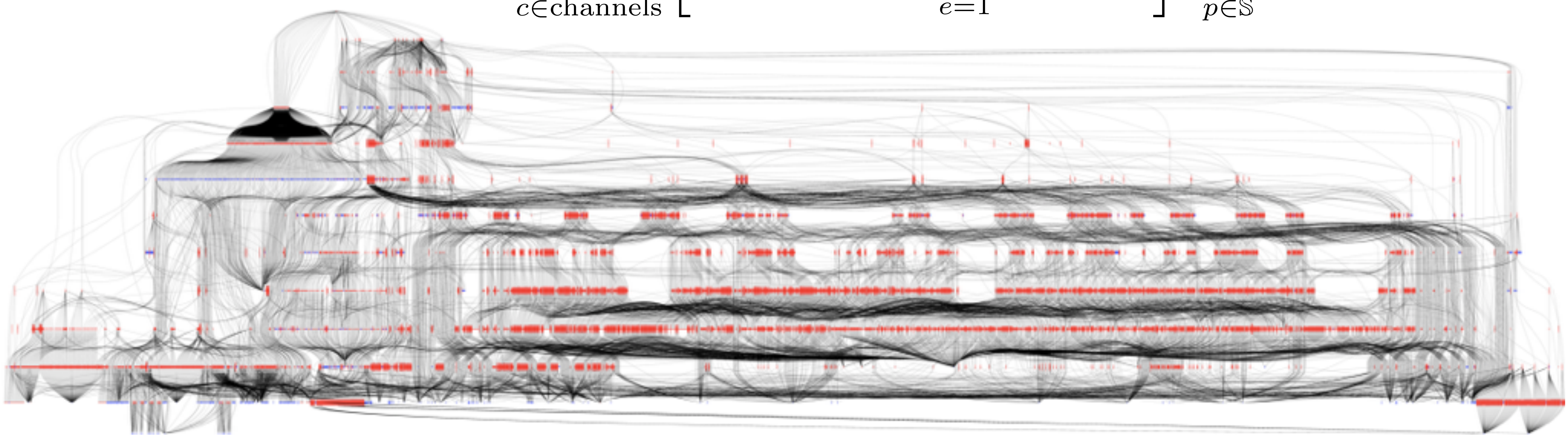
VISUALIZING THE COMBINED MODEL

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

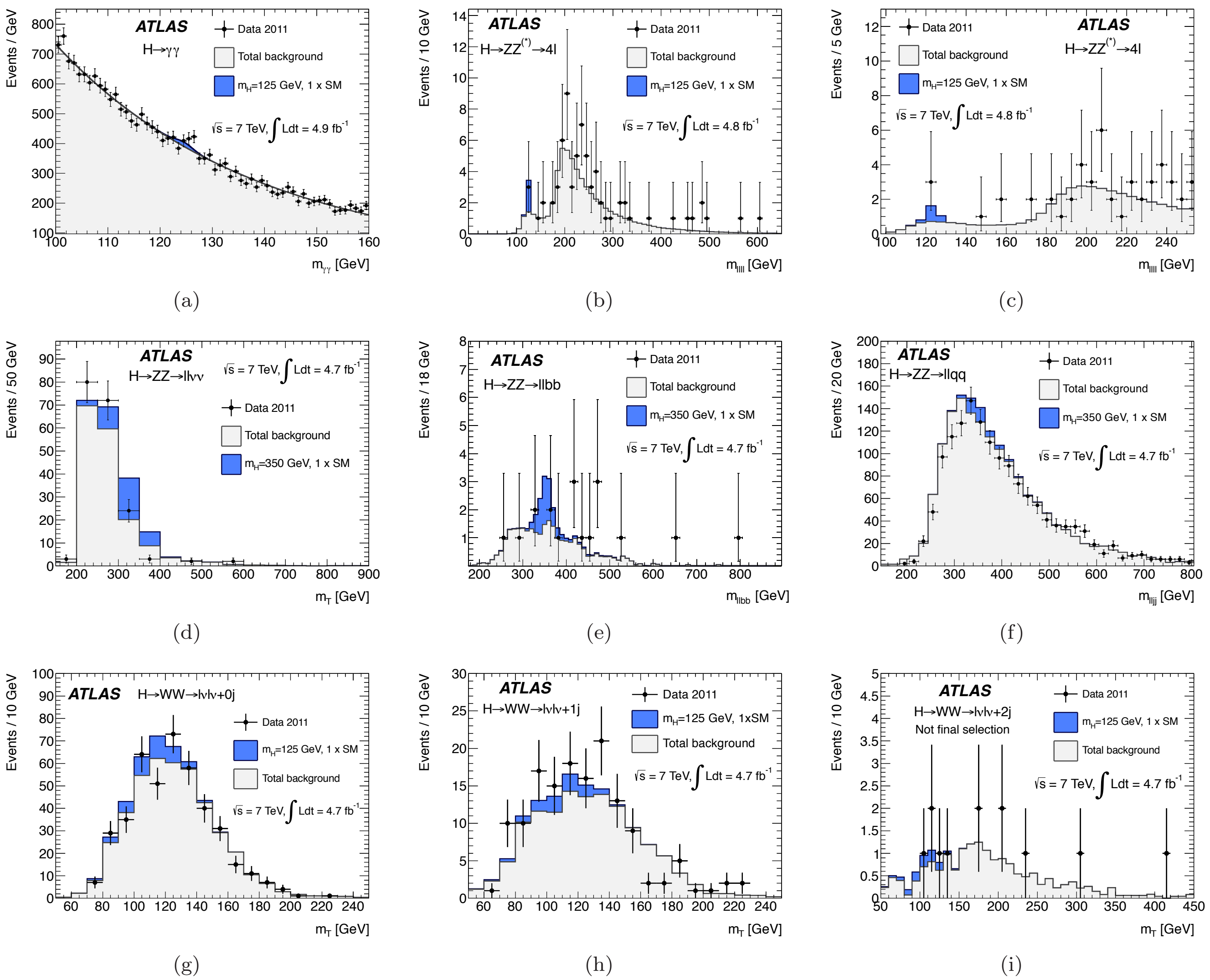
RooFit / RooStats: is the modeling language (C++) which provides technologies for collaborative modeling

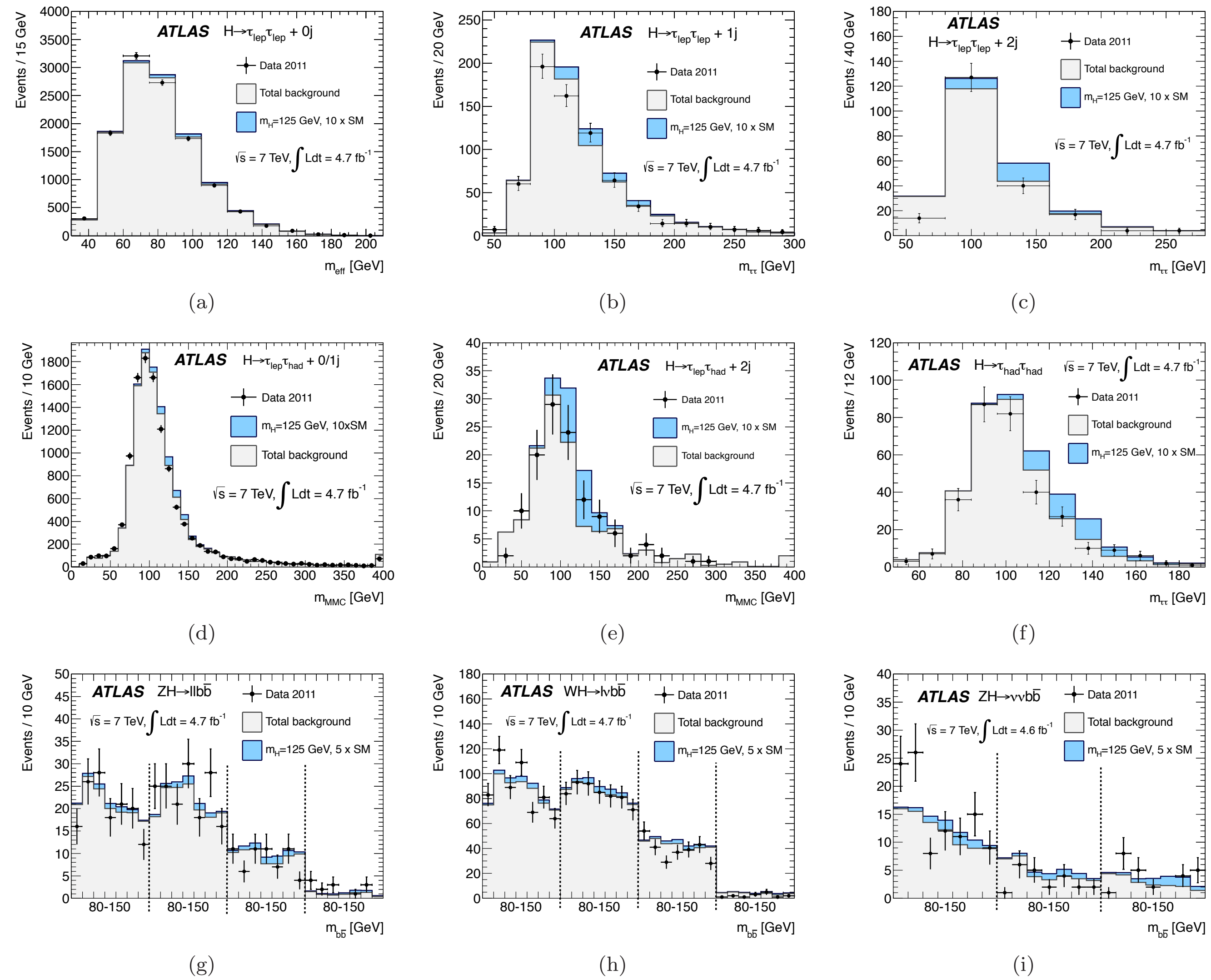
- provides technology to publish likelihood functions digitally
- and more, it's the full model so we can also generate pseudo-data

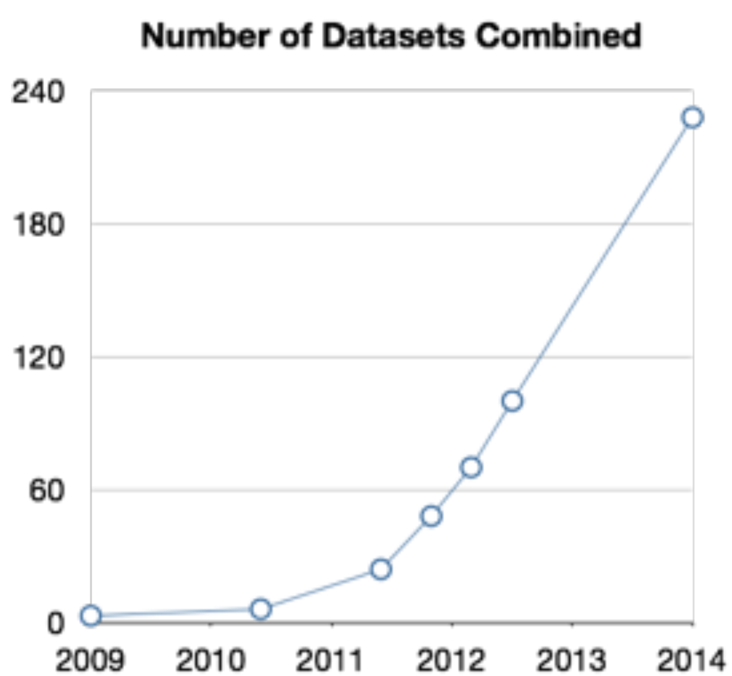
$$\mathbf{f}_{\rm tot}(\mathcal{D}_{\rm sim}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c\in{\rm channels}} \left[{\rm Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p\in\mathbb{S}} f_p(a_p|\alpha_p)$$

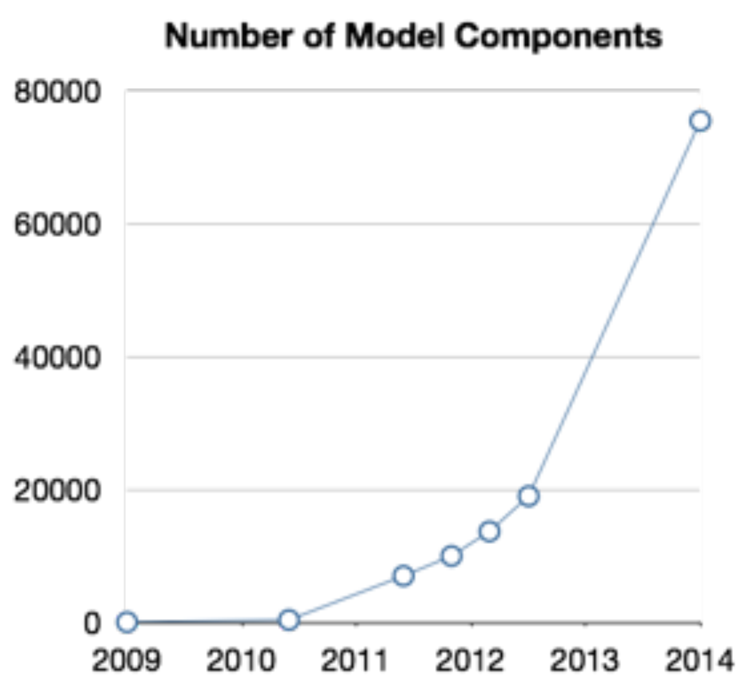


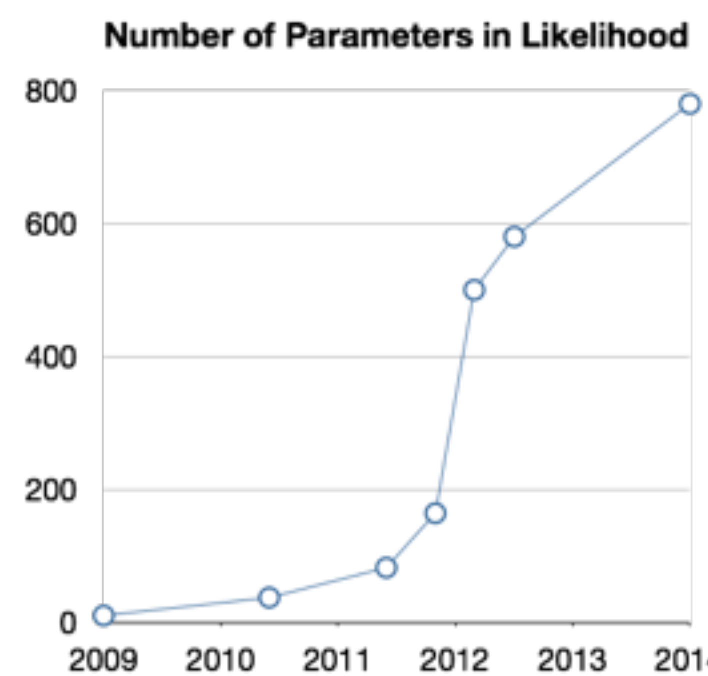
EVOLUTION OF MODEL COMPLEXITY











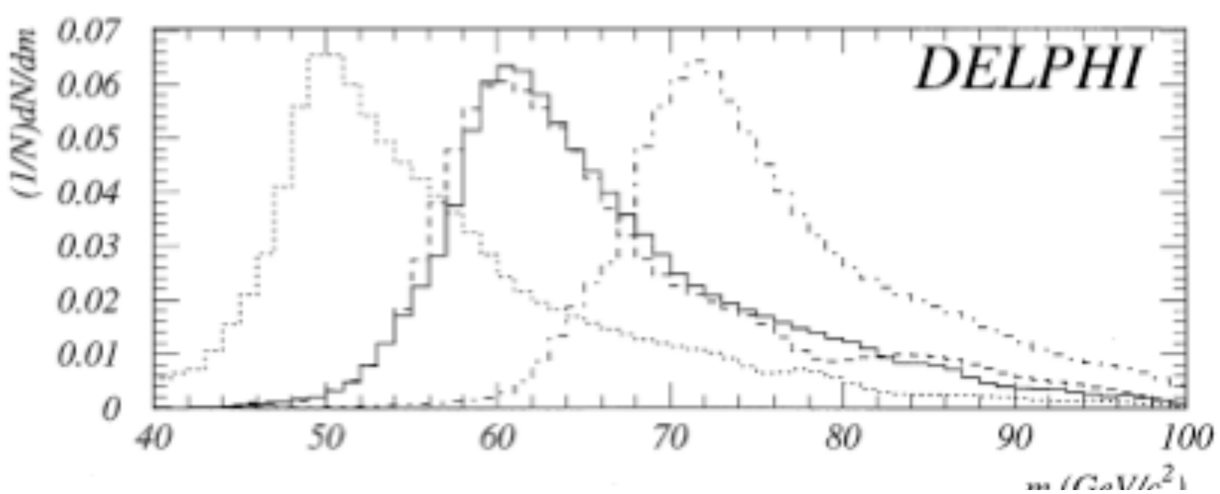
EXTRAS

Histogram Interpolation

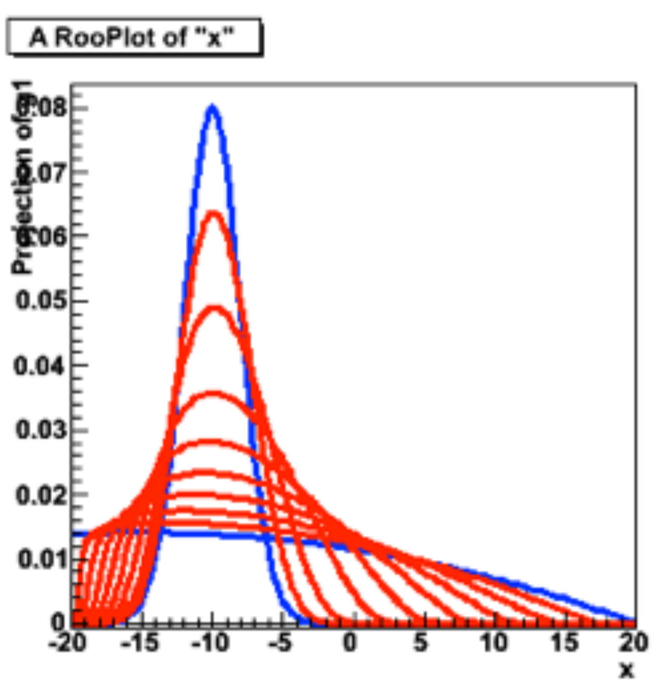
Several interpolation algorithms exist: eg. Alex Read's "horizontal" histogram interpolation algorithm (RooIntegralMorph in RooFit)

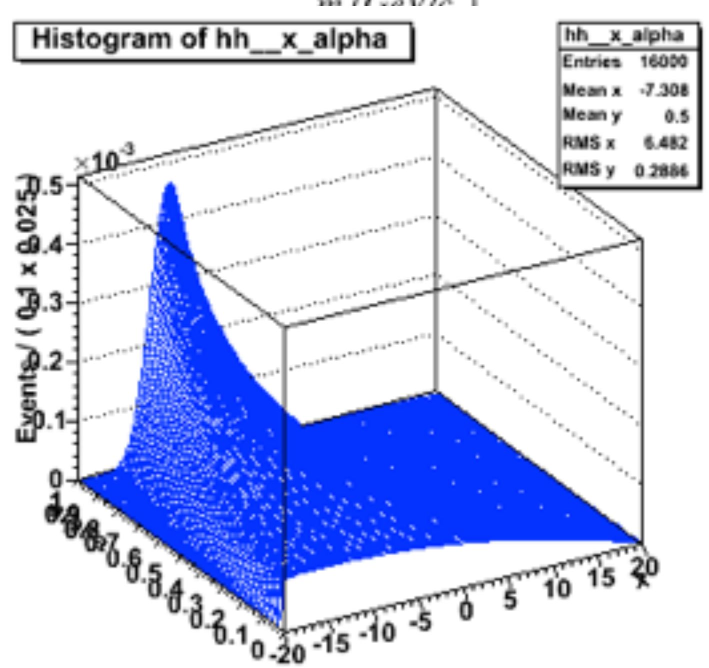
- take several PDFs, construct interpolated PDF with additional **nuisance parameter** α

A.L. Read / Nuclear Instruments and Methods in Physics Research A 425 (1999) 357–360



Simple "vertical" interpolation bin-by-bin.





Alternative "horizontal" interpolation algorithm by Max Baak called "RooMomentMorph" in RooFit (faster and numerically more stable)

COMMON CONSTRAINTS TERMS

Many uncertainties have no clear statistical description or it is impractical to provide

Traditionally, we use Gaussians, but for large uncertainties it is clearly a bad choice

- quickly falling tail, bad behavior near physical boundary, optimistic p-values, ...

For systematics constrained from control samples and dominated by statistical uncertainty, a Gamma distribution is a more natural choice [PDF is Poisson for the control sample]

- longer tail, good behavior near boundary, natural choice if auxiliary is based on counting

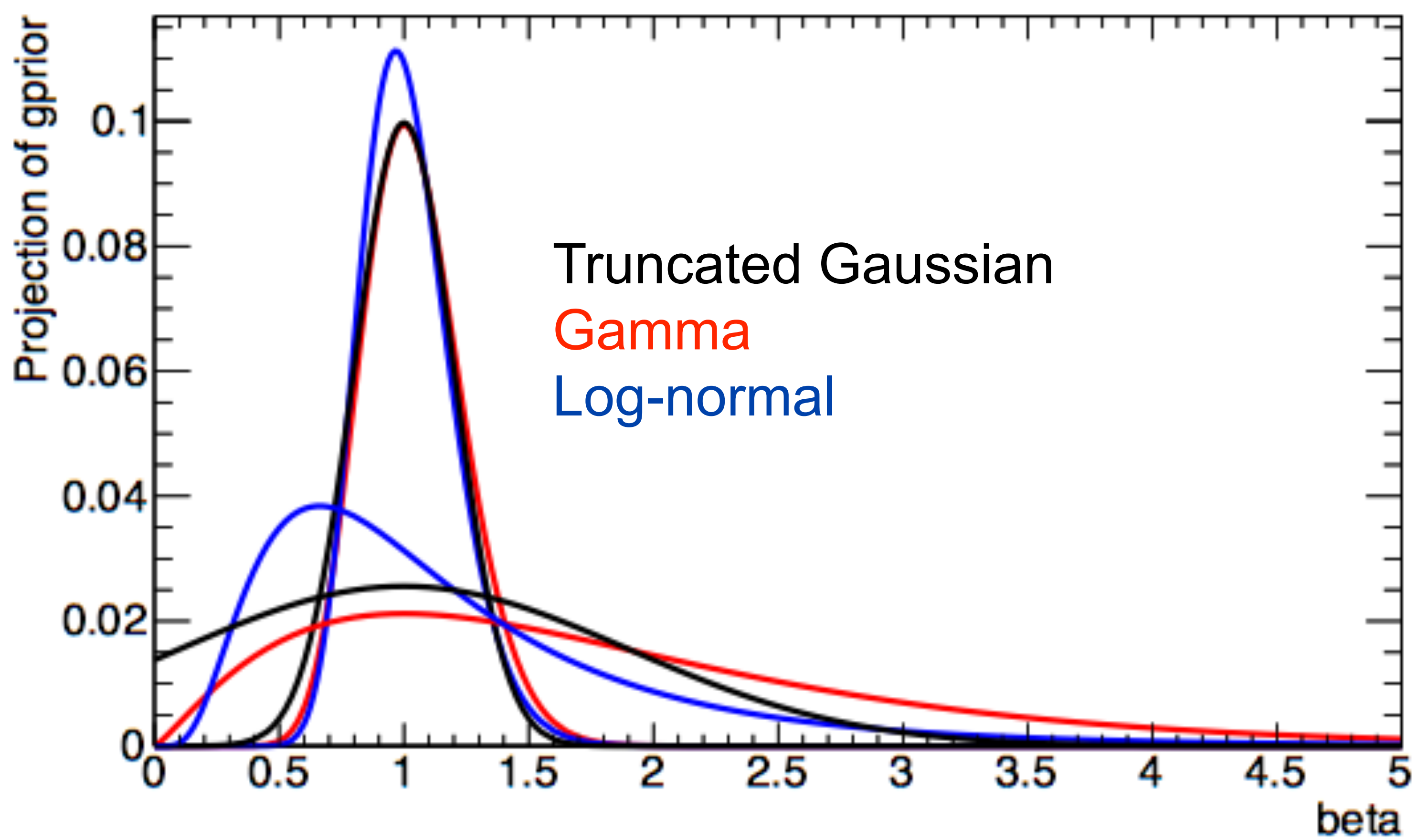
For "factor of 2" notions of uncertainty log-normal is a good choice

- can have a very long tail for large uncertainties

None of them are as good as an actual model for the auxiliary measurement, if available

To consistently switch between frequentist, Bayesian, and hybrid procedures, need to be clear about prior vs. likelihood function

PDF(y| β)	Prior(β)	Posterior(β|y)
Gaussian	uniform	Gaussian
Poisson	uniform	Gamma
Log-normal	1/β	Log-Normal

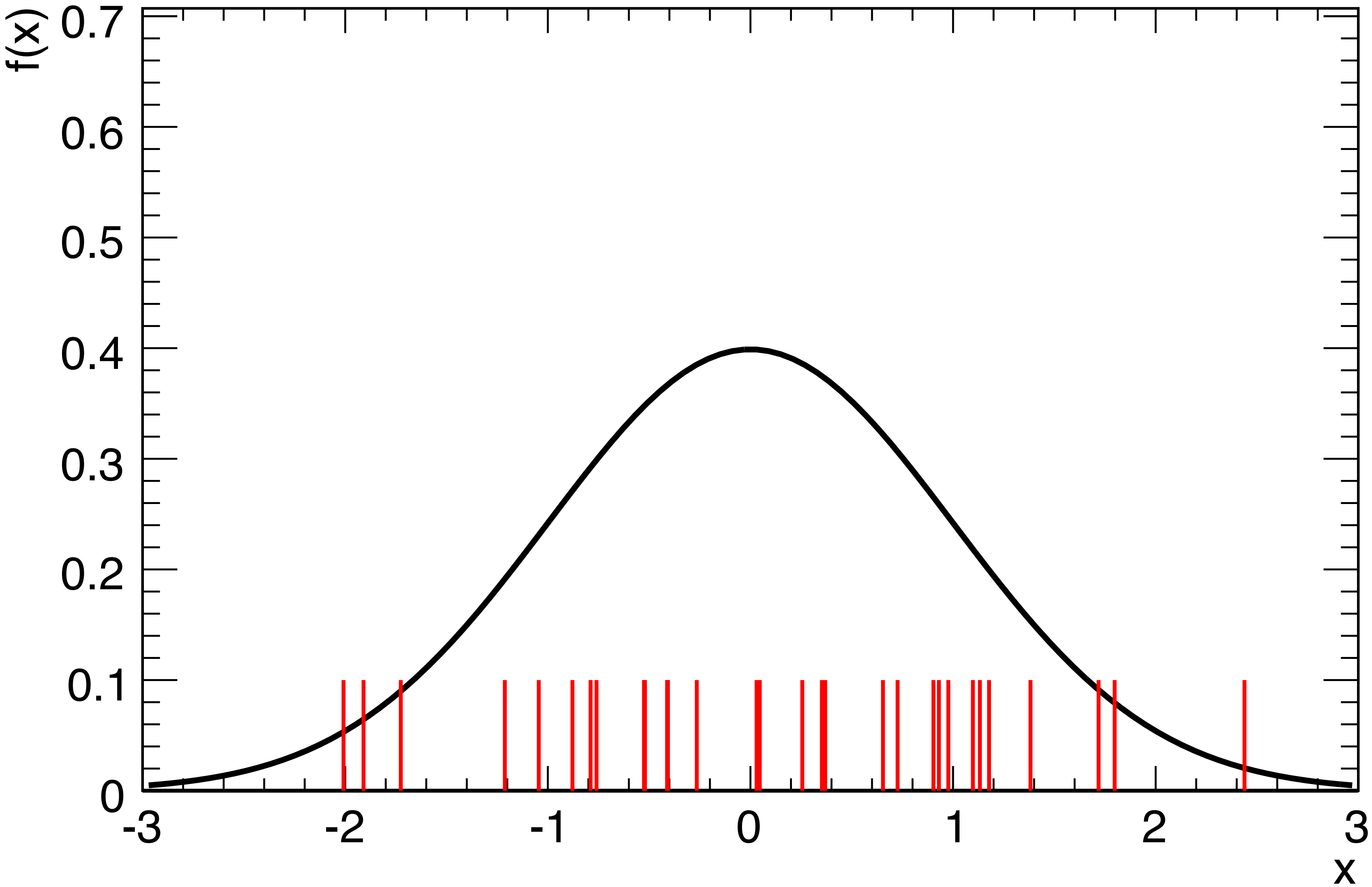


Parametric vs. Non-Parametric PDFs

No parametric form, need to construct **non-parametric** PDFs

From Monte Carlo samples, one has empirical PDF

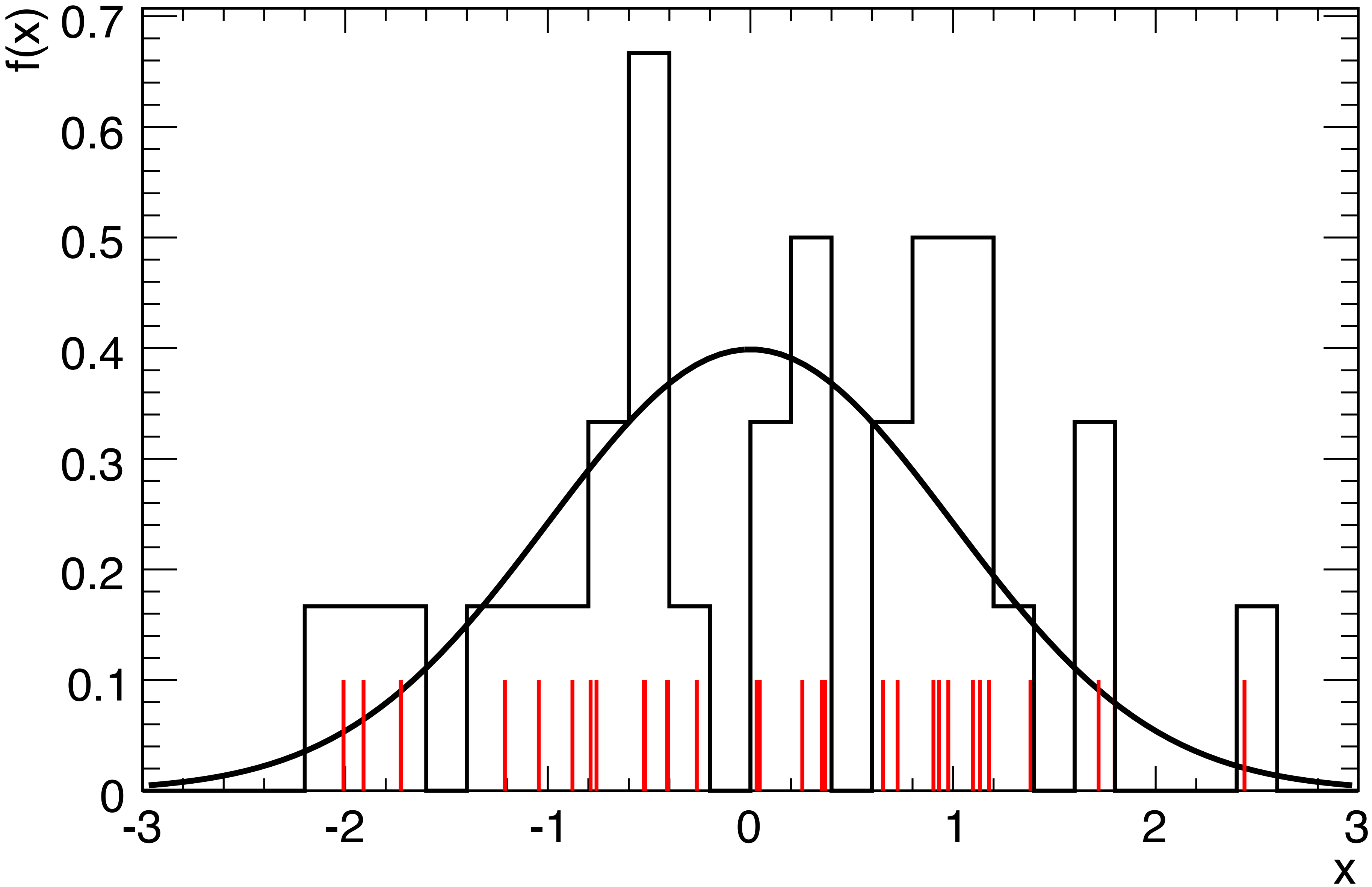
$$f_{emp} = \frac{1}{N} \sum_{i}^{N} \delta(x - x_i)$$



Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

$$f_{hist}^{w,s}(x) = \frac{1}{N} \sum_i h_i^{w,s}$$

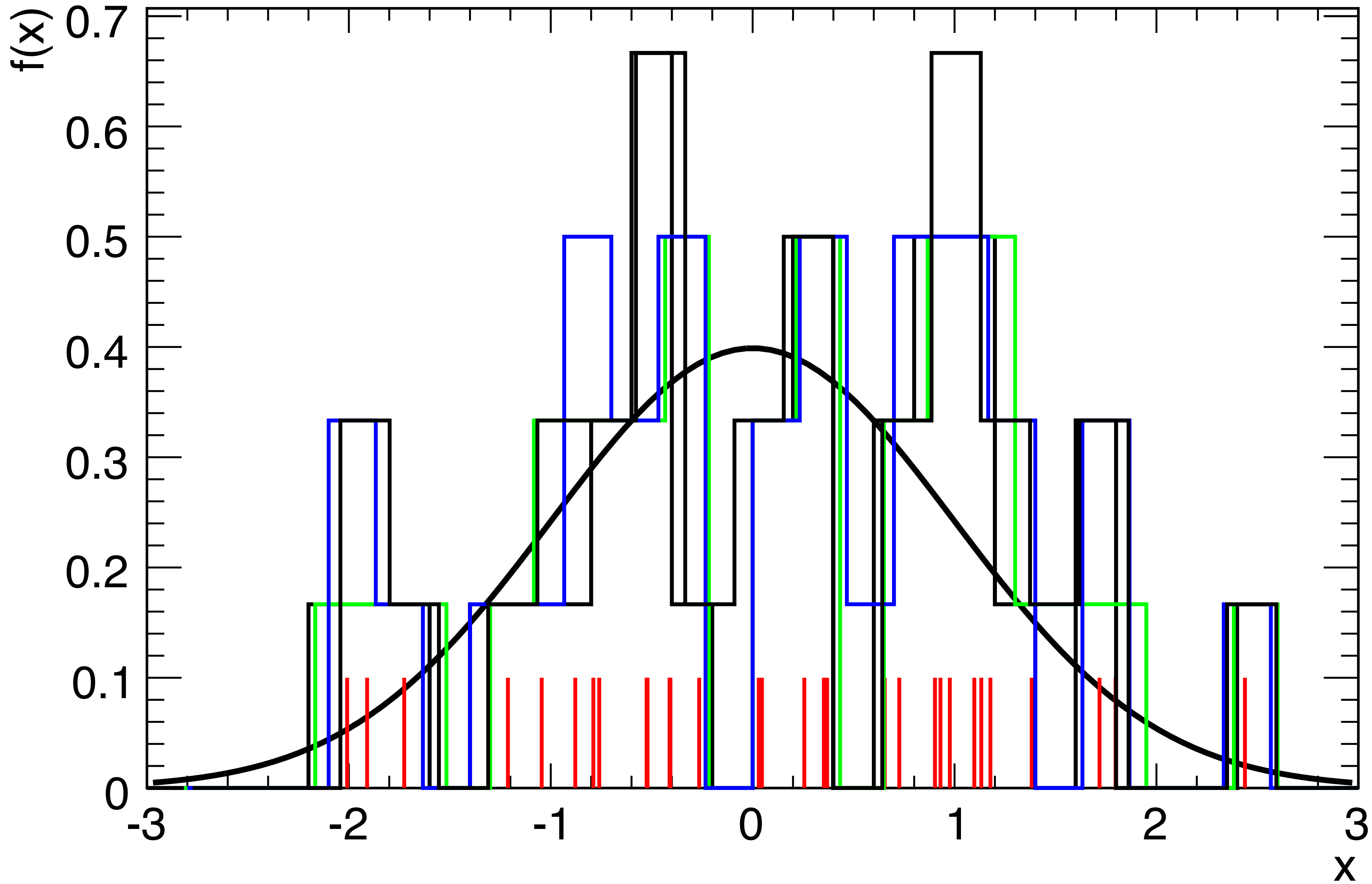


Parametric vs. Non-Parametric PDFs

Classic example of a non-parametric PDF is the histogram

but they depend on bin width and starting position

$$f_{hist}^{w,s}(x) = \frac{1}{N} \sum_i h_i^{w,s}$$

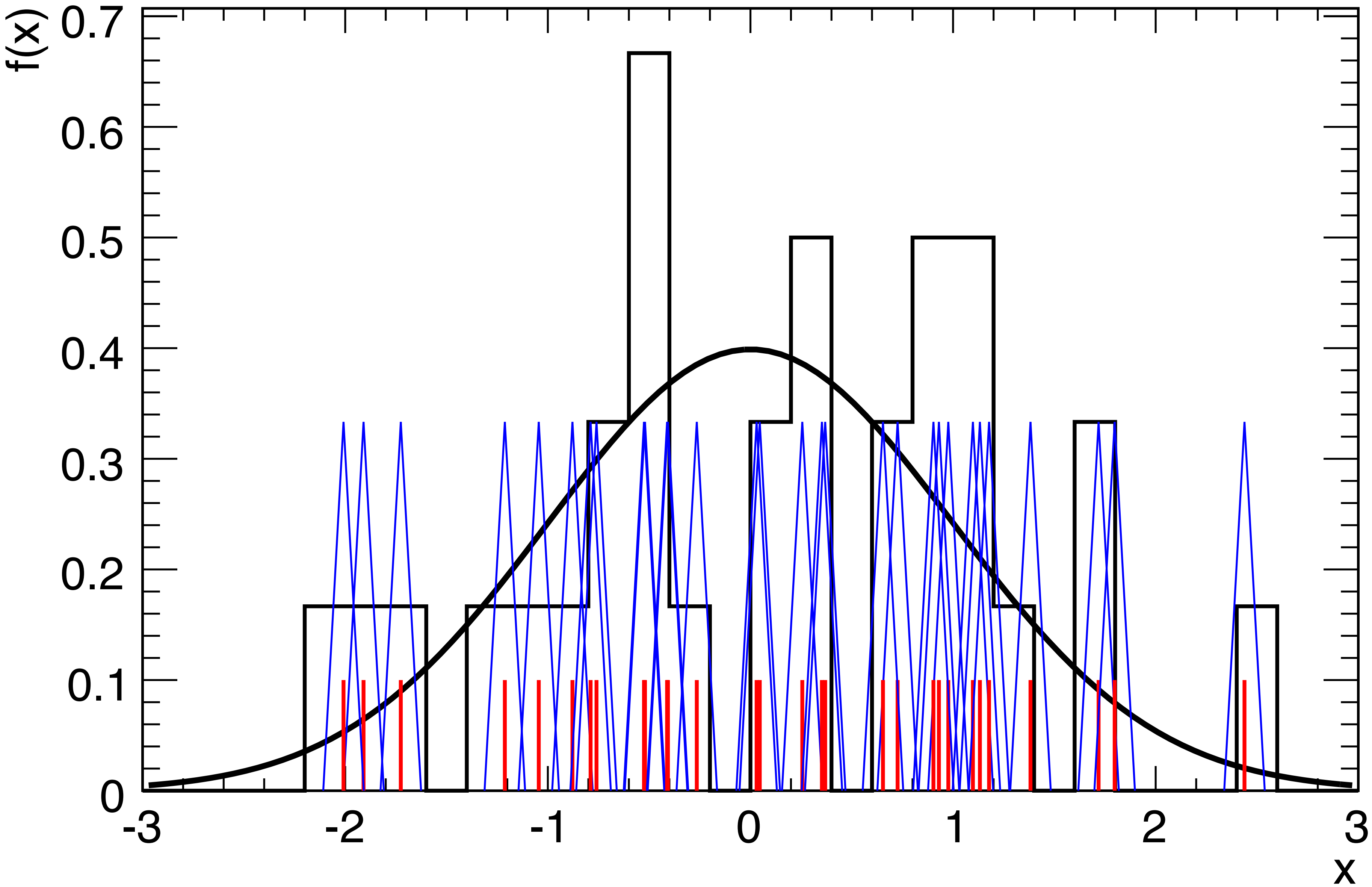


Parametric vs. Non-Parametric PDFs

Classic example of a non-parametric PDF is the histogram

"Average Shifted Histogram" minimizes effect of binning

$$f^w_{ASH}(x) = \frac{1}{N}\sum_i^N K^w(x - x_i)$$

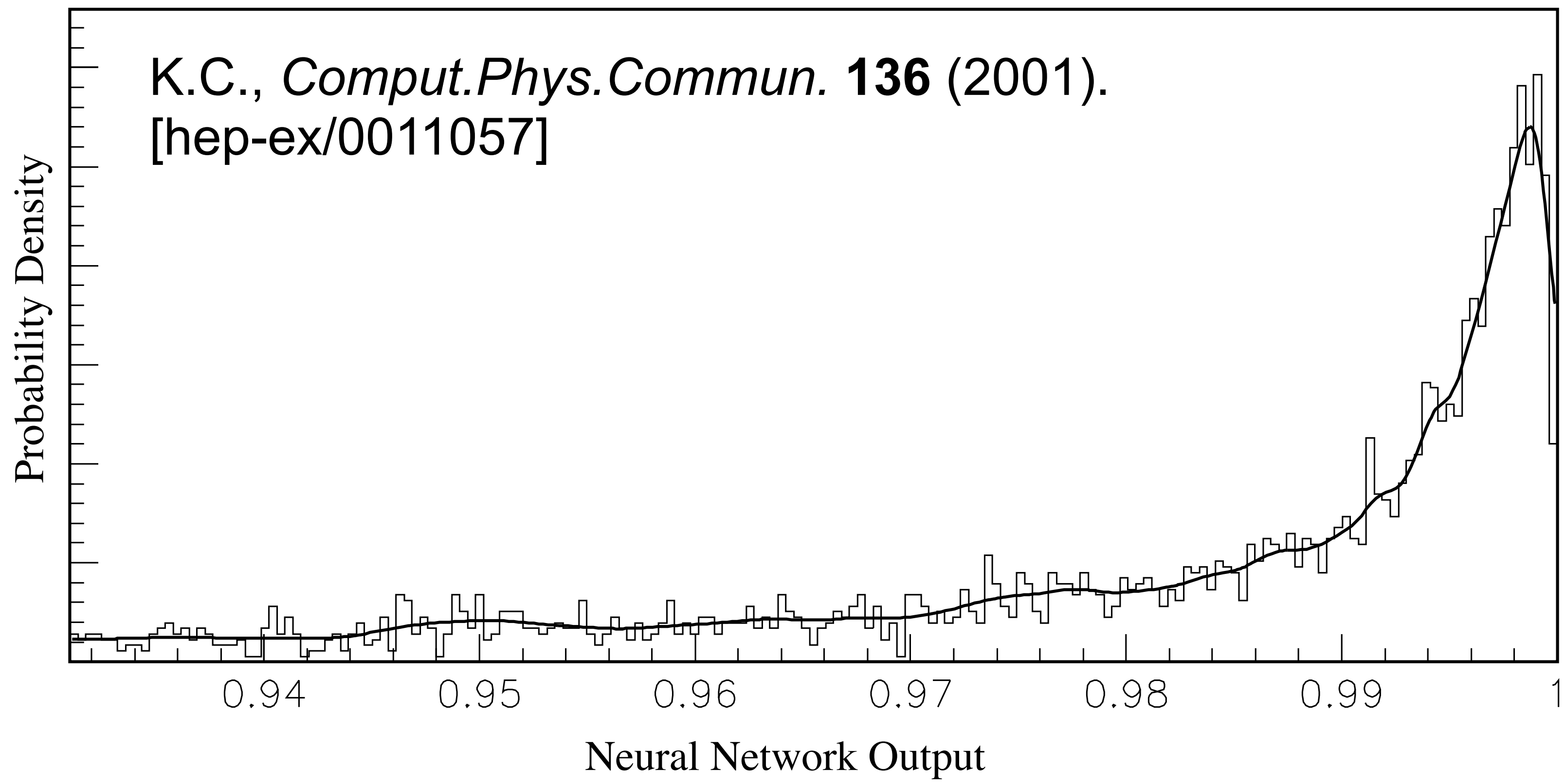


Kernel Density Estimation

Kernel estimation is the generalization of Average Shifted Histograms

$$\hat{f}_1(x) = \sum_i^n \frac{1}{nh(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right)$$

$$h(x_i) = \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{\hat{f}_0(x_i)}} n^{-1/5}$$



"the data is the model"

Adaptive Kernel estimation puts wider kernels in regions of low probability

Used at LEP for describing pdfs from Monte Carlo (KEYS)

MULTIVARIATE, NON-PARAMETRIC PDFS

Kernel Estimation has a nice generalizations to higher dimensions

- practical limit is about 5-d due to curse of dimensionality

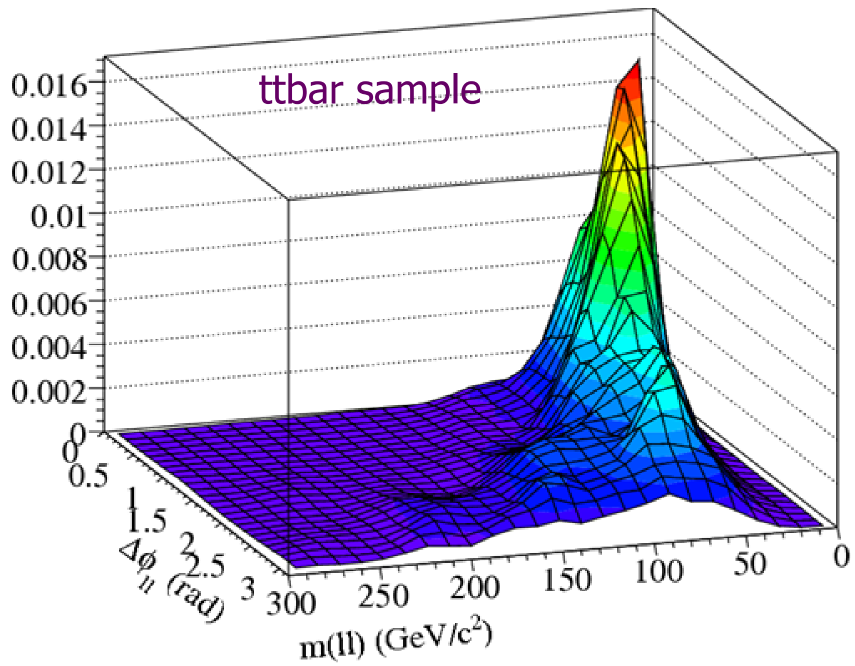
Max Baak has coded N-dim KEYS pdf described in Comput.Phys.Commun. **136** (2001) in RooFit.

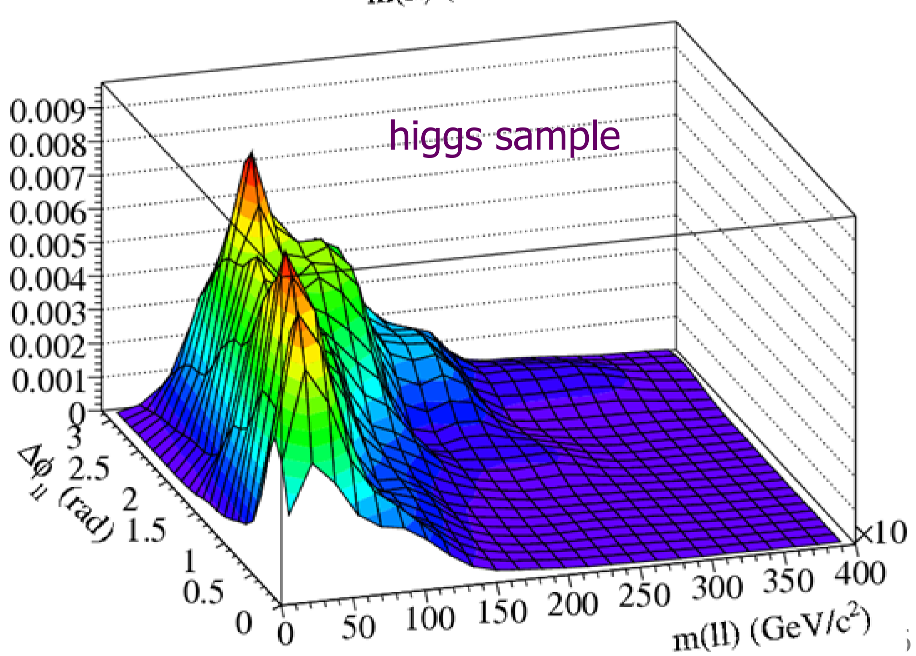
These pdfs have been used as the basis for a multivariate discrimination technique called "PDE"

$$D(\vec{x}) = \frac{f_s(\vec{x})}{f_s(\vec{x}) + f_b(\vec{x})}$$

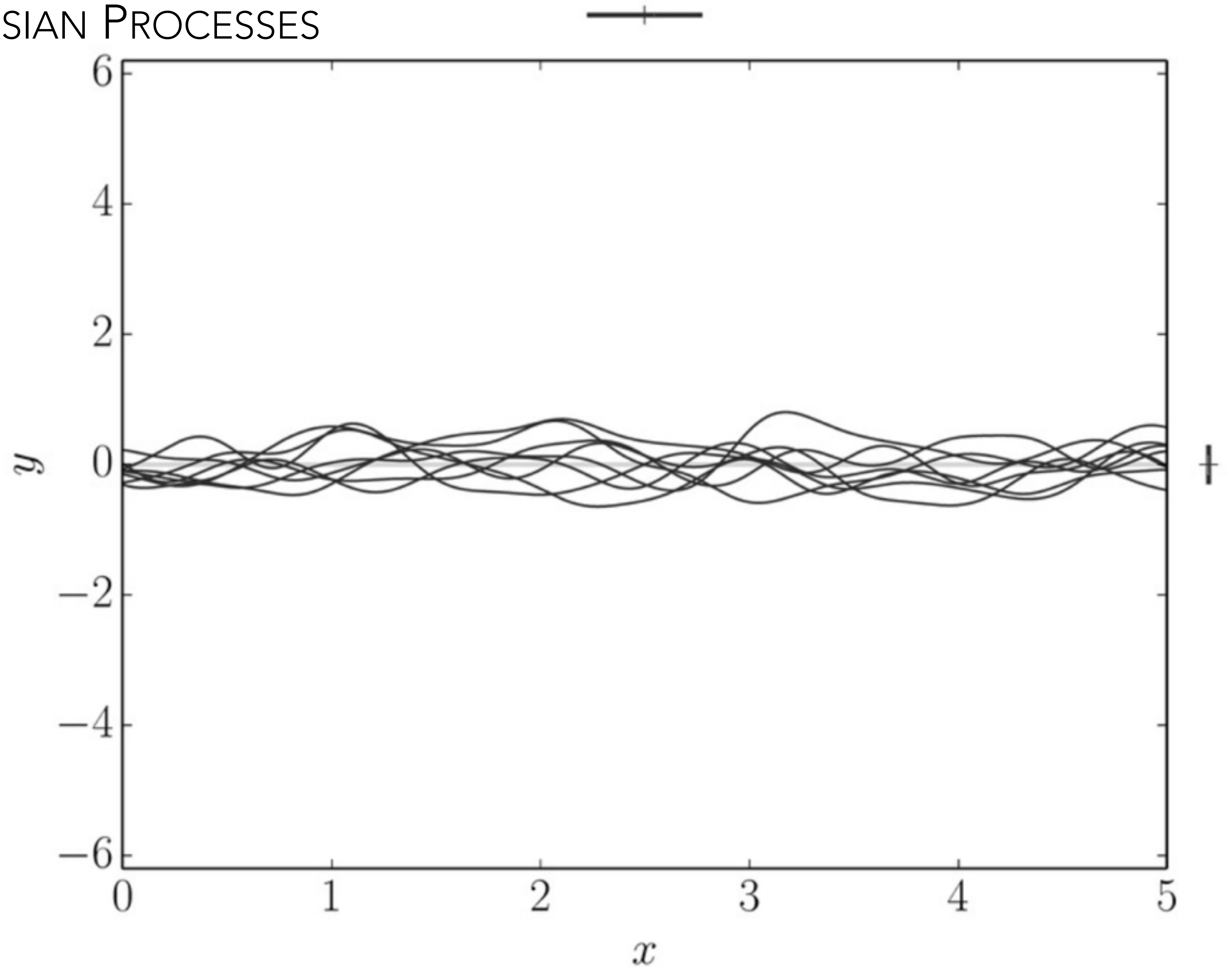
Correlations

- 2-d projection of pdf from previous slide.
- RooNDKeys pdf automatically models (fine) correlations between observables ...



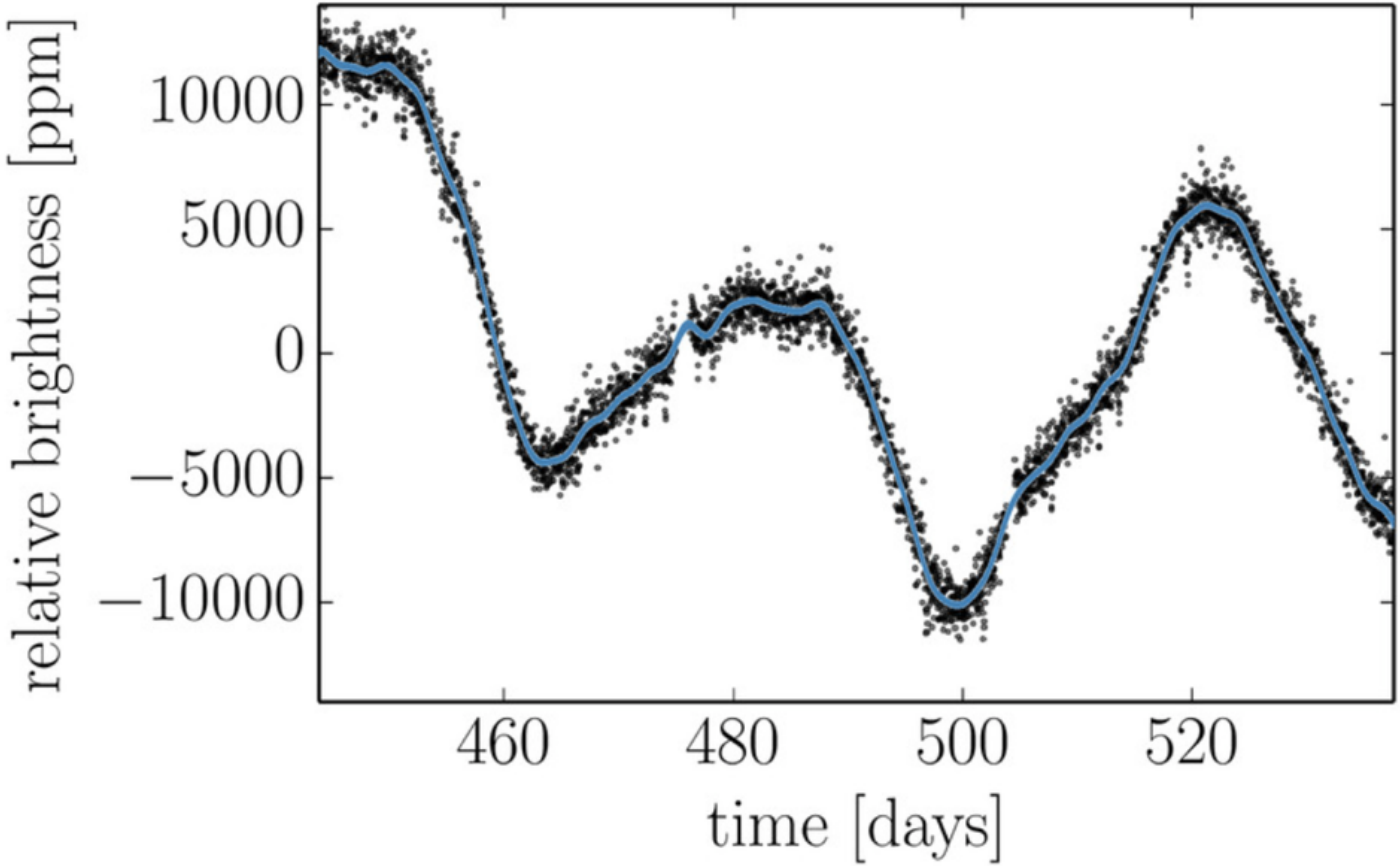


Gaussian Processes



https://speakerdeck.com/dfm/pydata-time-series-analysis-gps-and-exoplanets

AN EXOPLANET EXAMPLE



$$\log p(\boldsymbol{y}\,|\,\boldsymbol{x},\,\boldsymbol{\sigma},\,\boldsymbol{\theta},\,\boldsymbol{\alpha}) = -\,\frac{1}{2}\,[\boldsymbol{y} - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]^{\mathrm{T}}\,K_{\boldsymbol{\alpha}}(\boldsymbol{x},\,\boldsymbol{\sigma})^{-1}\,[\boldsymbol{y} - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})] \\ -\,\frac{1}{2}\,\log\det K_{\boldsymbol{\alpha}}(\boldsymbol{x},\,\boldsymbol{\sigma}) - \frac{N}{2}\,\log 2\,\pi$$

where

$$[K_{\boldsymbol{\alpha}}(\boldsymbol{x},\,\boldsymbol{\sigma})]_{ij} = {\sigma_i}^2\,\delta_{ij} + k_{\boldsymbol{\alpha}}(x_i,\,x_j)$$