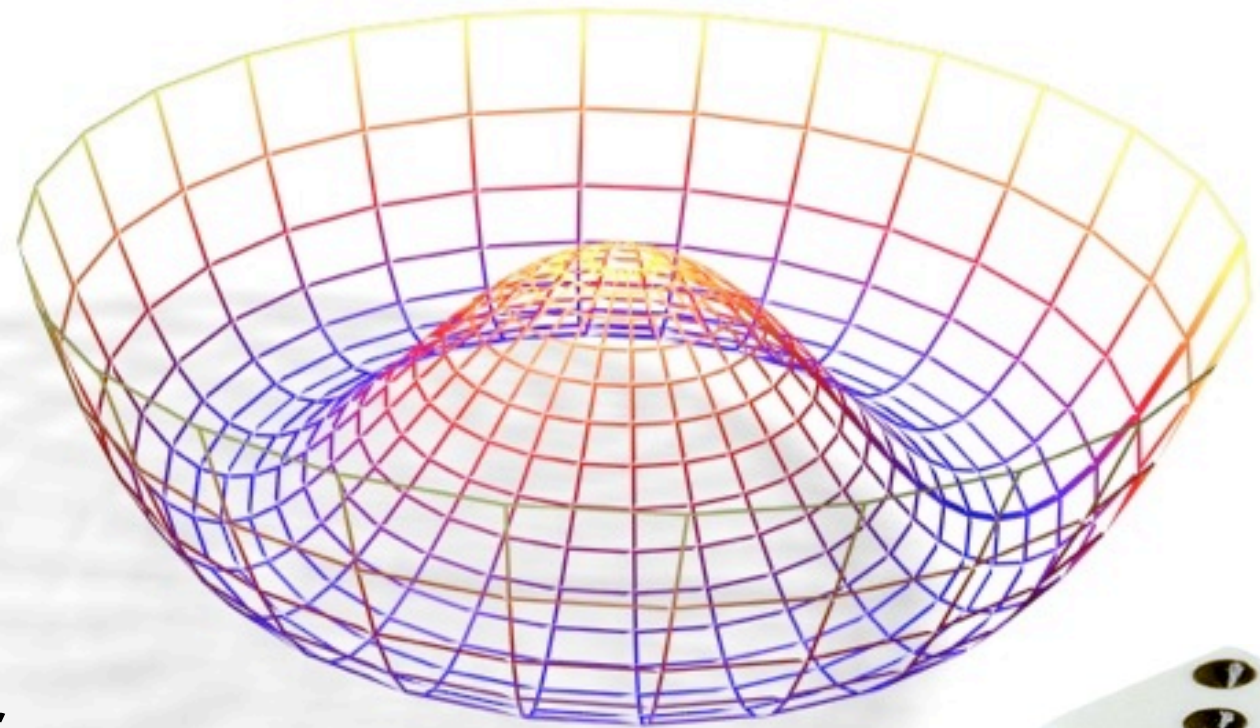




# *Practical Statistics for Particle Physics*



***Kyle Cranmer,***  
New York University

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

By physicists, for physicists

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

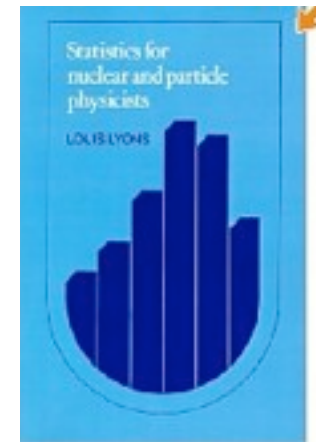
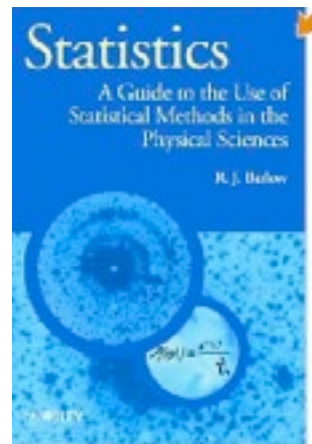
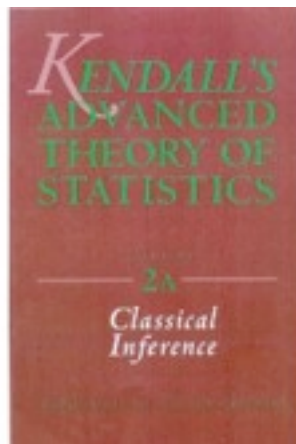
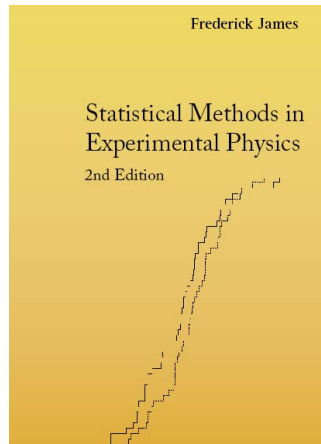
R.J.Barlow, *A Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley, 1989;

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*.

## Fred James's lectures

[http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic\\_Training&id=AT00000799](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](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](http://PhyStat.org):



site map access

### PhyStat Physics Statistics Code Repository

---

An open, loosely moderated repository for code, tools, and documents relevant to statistics in physics applications. Search and download access is universal; package submission is loosely moderated for suitability.

#### Using the Site

---

- [Lists of packages](#)
- [Search for a package](#)
- [Submit a Package](#)
- [Comment on a package \(not yet available\)](#)

#### About the Repository

---

- [Repository Policies and Procedures](#)
- [The PhyStat Repository Steering Committee](#)
- [Comment on the repository site or policies](#)

#### PHYSTAT Conference Links

---

- PHYSTAT [07](#) (CERN) [05](#) (Oxford) [03](#) (SLAC) [02](#) (Durham)
- [PhyStat Workshops](#): [08](#) (Caltech) [06](#) (BIRS/Banff) [00](#) (Fermilab) [00](#) (CERN)
- [More Conferences and Workshops ...](#)

## 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.

## Contents

1	Introduction . . . . .	3
2	Conceptual building blocks for modeling . . . . .	3
2.1	Probability densities and the likelihood function . . . . .	3
2.2	Auxiliary measurements . . . . .	5
2.3	Frequentist and Bayesian reasoning . . . . .	6
2.4	Consistent Bayesian and Frequentist modeling of constraint terms . . . . .	7
3	Physics questions formulated in statistical language . . . . .	8
3.1	Measurement as parameter estimation . . . . .	8
3.2	Discovery as hypothesis tests . . . . .	9
3.3	Excluded and allowed regions as confidence intervals . . . . .	11
4	Modeling and the Scientific Narrative . . . . .	14
4.1	Simulation Narrative . . . . .	15
4.2	Data-Driven Narrative . . . . .	25
4.3	Effective Model Narrative . . . . .	27
4.4	The Matrix Element Method . . . . .	27
4.5	Event-by-event resolution, conditional modeling, and Punzi factors . . . . .	28
5	Frequentist Statistical Procedures . . . . .	28
5.1	The test statistics and estimators of $\mu$ and $\theta$ . . . . .	29
5.2	The distribution of the test statistic and $p$ -values . . . . .	31
5.3	Expected sensitivity and bands . . . . .	32
5.4	Ensemble of pseudo-experiments generated with “Toy” Monte Carlo . . . . .	33
5.5	Asymptotic Formulas . . . . .	33
5.6	Importance Sampling . . . . .	36
5.7	Look-elsewhere effect, trials factor, Bonferoni . . . . .	37
5.8	One-sided intervals, CLs, power-constraints, and Negatively Biased Relevant Subsets . . . . .	37
6	Bayesian Procedures . . . . .	38
6.1	Hybrid Bayesian-Frequentist methods . . . . .	39
6.2	Markov Chain Monte Carlo and the Metropolis-Hastings Algorithm . . . . .	40
6.3	Jeffreys’s and Reference Prior . . . . .	40
6.4	Likelihood Principle . . . . .	41
7	Unfolding . . . . .	42
8	Conclusions . . . . .	42

## Lecture 1: Preliminaries

- Probability Density Function vs. Likelihood
- Point estimates 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

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

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



“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  $i^{\text{th}}$  possibility

$$P(x_i) = p_i$$



Defined as limit of long term frequency

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

And it is normalized

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

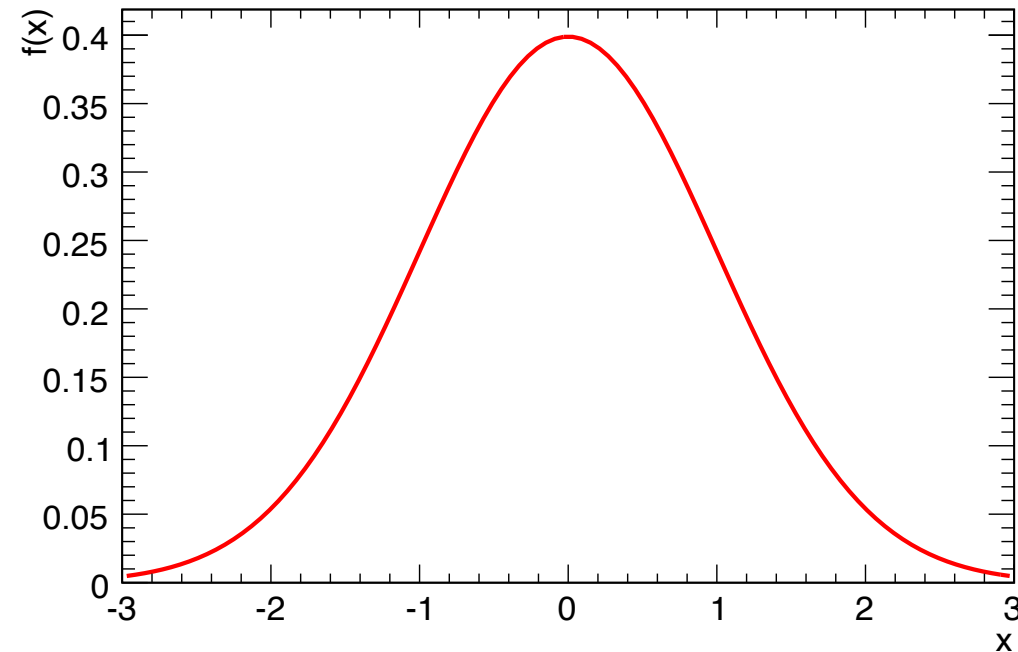
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function**

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

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

PDFs are always normalized

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



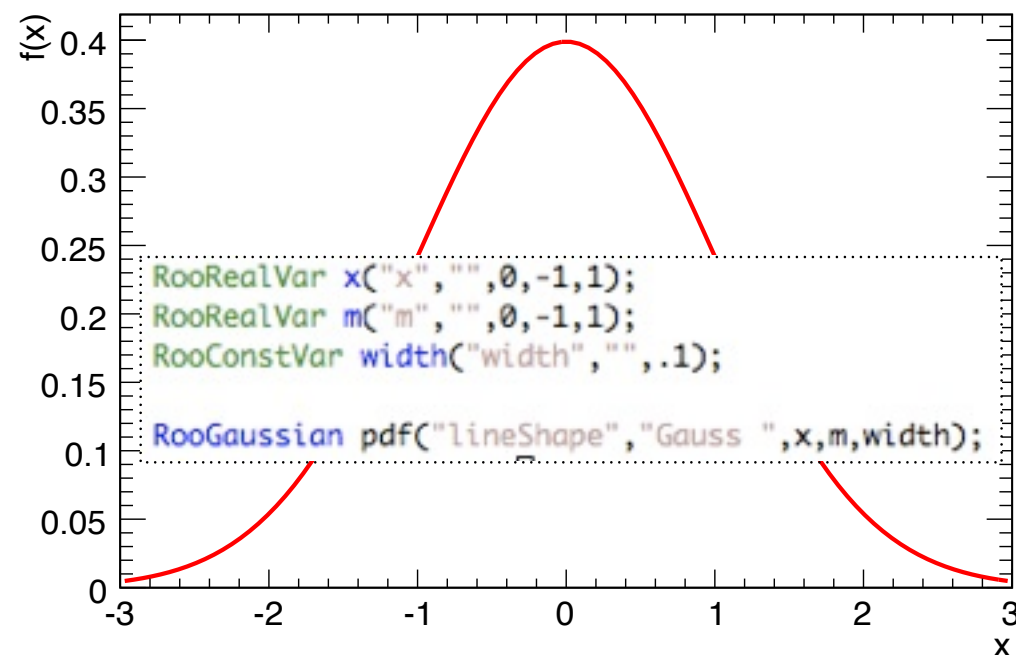
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function**

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

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

PDFs are always normalized

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



Often we are interested in a parametrized family of pdfs

- ▶ We will write these as:  $f(x|\alpha)$  said “ $f$  of  $x$  given  $\alpha$ ”
  - where  $\alpha$  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  $\mu$  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  $\mu$  and standard deviation  $\sigma$

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

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

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

The **likelihood** of  $\mu$  given  $n$  is the same equation evaluated as a function of  $\mu$

- ▶ 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  $\chi^2$  distribution

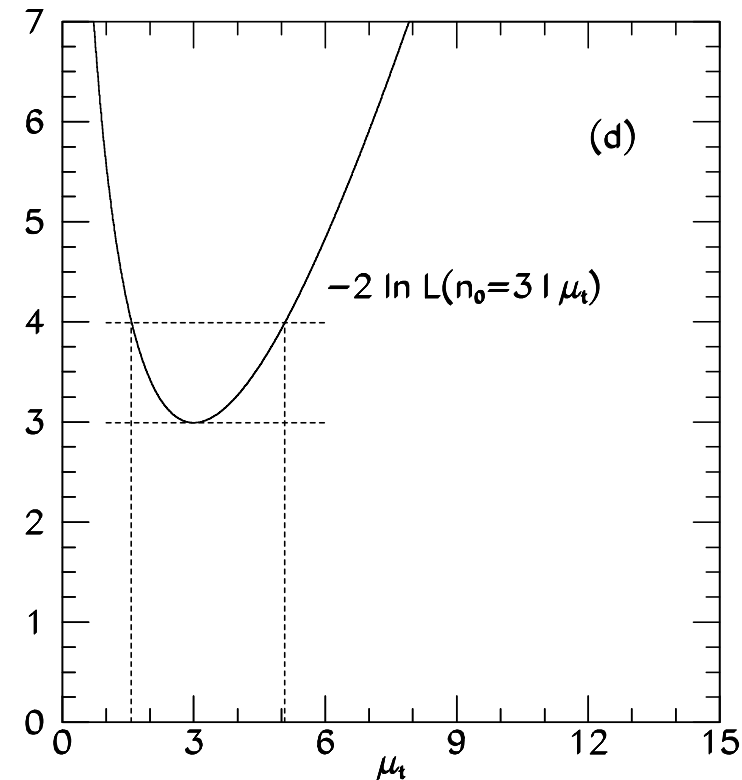


Figure from R. Cousins,  
Am. J. Phys. 63 398 (1995)

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

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

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

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

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



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

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

This is just the value of  $\alpha$  that maximizes the likelihood

## 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  $E[n]=\mu$

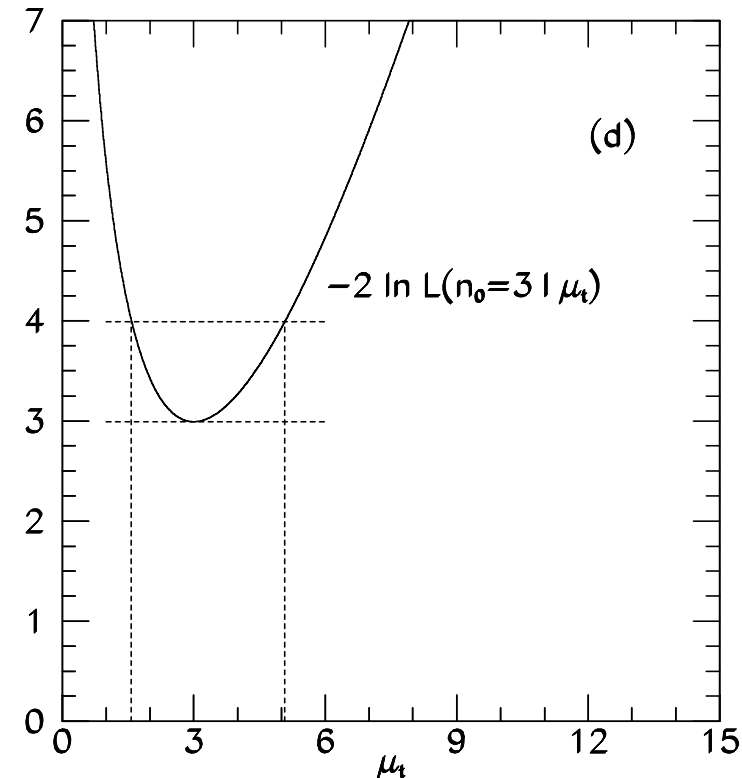


Figure from R. Cousins,  
Am. J. Phys. 63 398 (1995)

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

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

which gives

$$-\frac{d}{d\mu} \ln L(\mu) \Big|_{\hat{\mu}} = 0 = \frac{d}{d\mu} \left( \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} + \underbrace{\ln \sqrt{2\pi}\sigma}_{\text{const}} \right) = \sum_i \frac{(x_i - \mu)}{\sigma^2}$$
$$\Rightarrow \hat{\mu} = \frac{1}{N} \sum_i x_i \quad (\text{an unbiased estimator}).$$

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 asymptotically unbiased.

Define covariance  $\text{cov}[x,y]$  (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

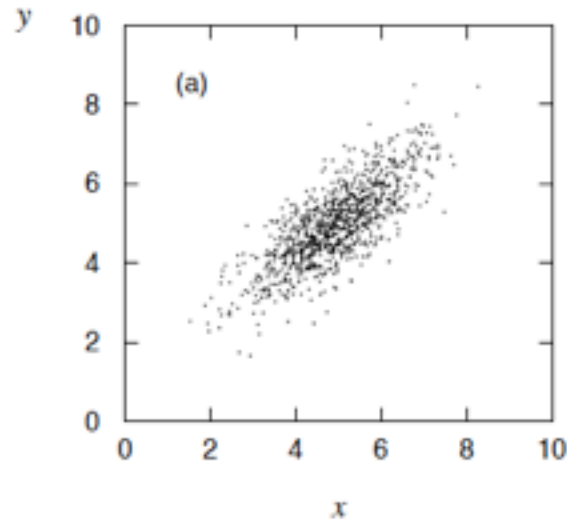
$$E[xy] = \int \int xy f(x, y) dx dy = \mu_x\mu_y$$

→  $\text{COV}[x, y] = 0$   $x$  and  $y$ , 'uncorrelated'

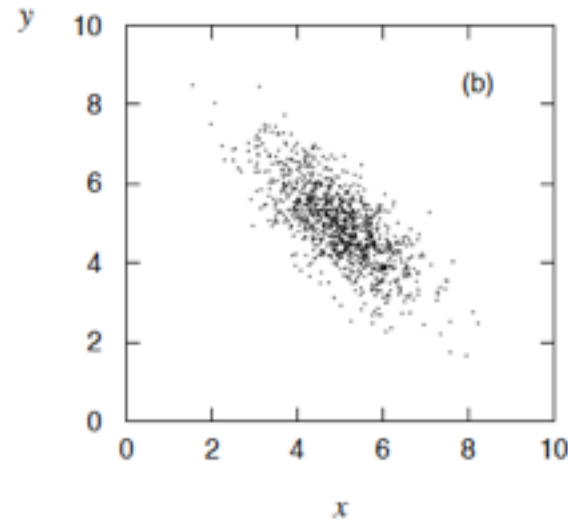
**N.B. converse not always true.**

# Correlation Coefficient examples

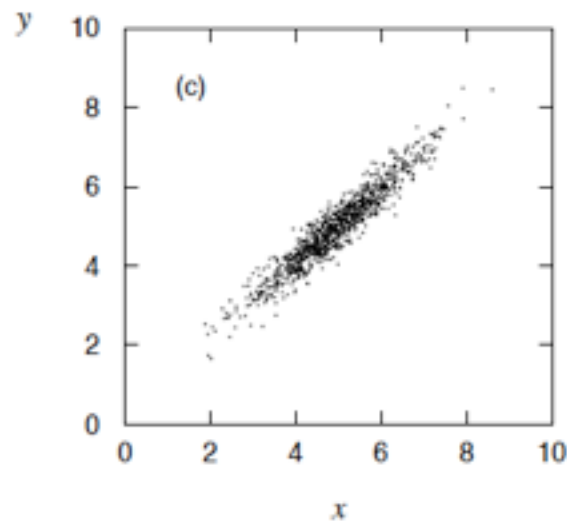
$$\rho = 0.75$$



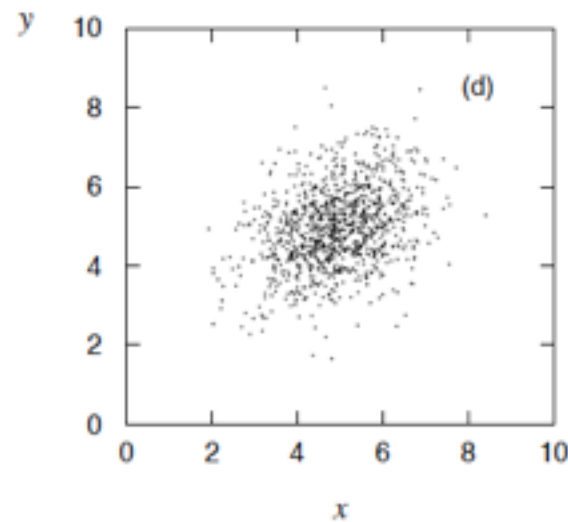
$$\rho = -0.75$$



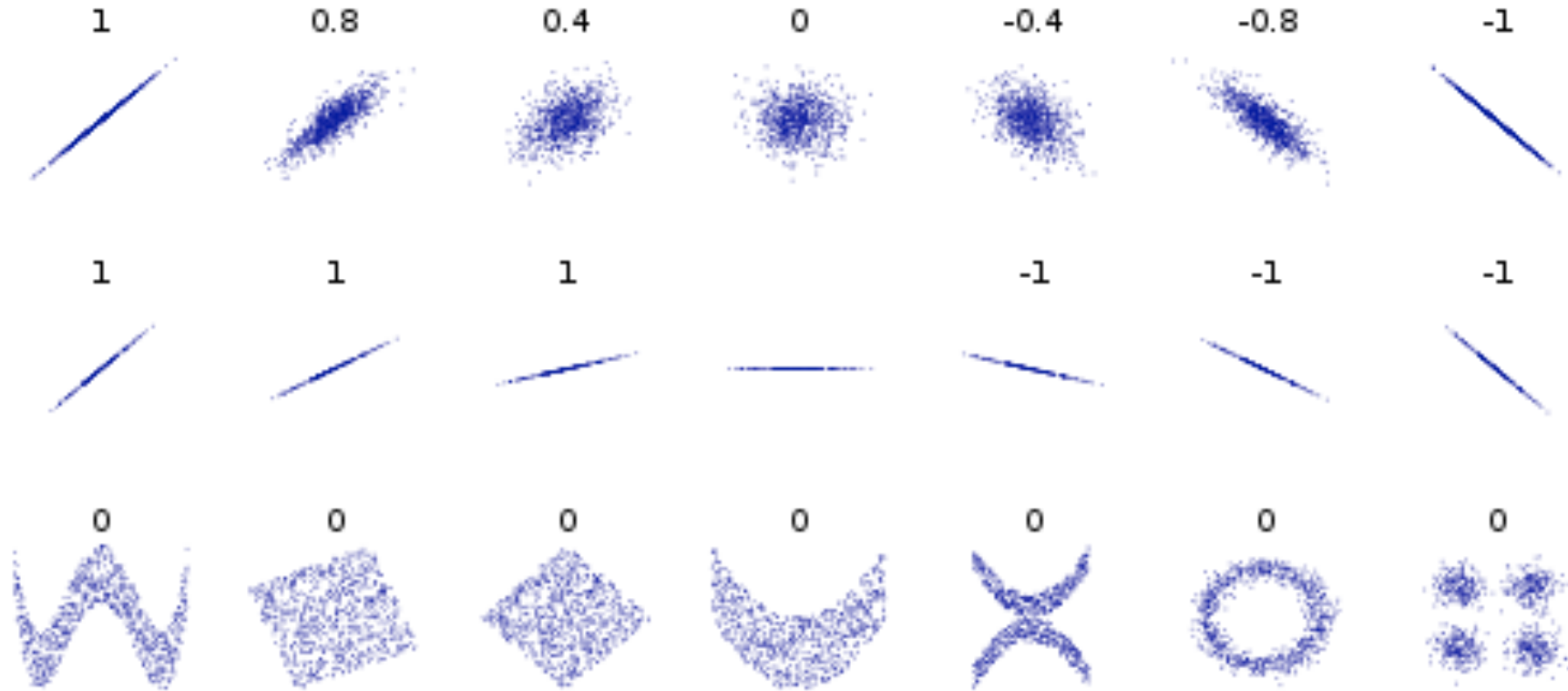
$$\rho = 0.95$$



$$\rho = 0.25$$



# Correlation Coefficient examples



[http://en.wikipedia.org/wiki/Correlation\\_and\\_dependence](http://en.wikipedia.org/wiki/Correlation_and_dependence)

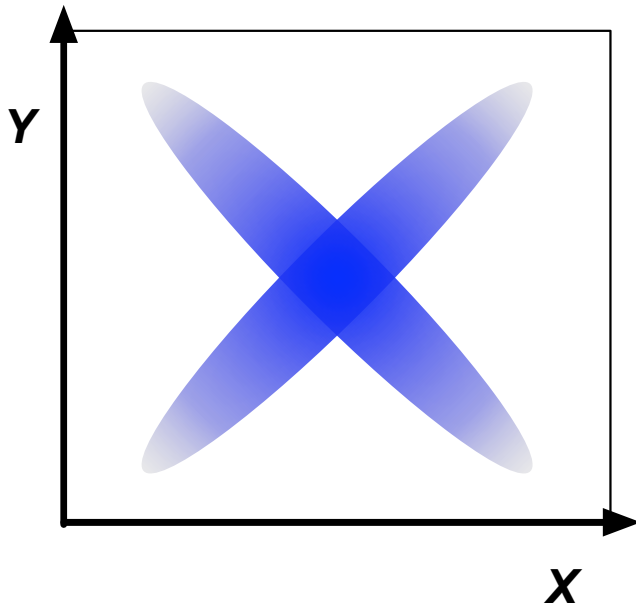
A more general notion of 'correlation' comes from  
**Information:**

**Mutual**

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

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

- ▶ 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

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

- ▶ simple univariate case:

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

- ▶ For an unbiased estimator the Cramér-Rao bound states

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

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

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

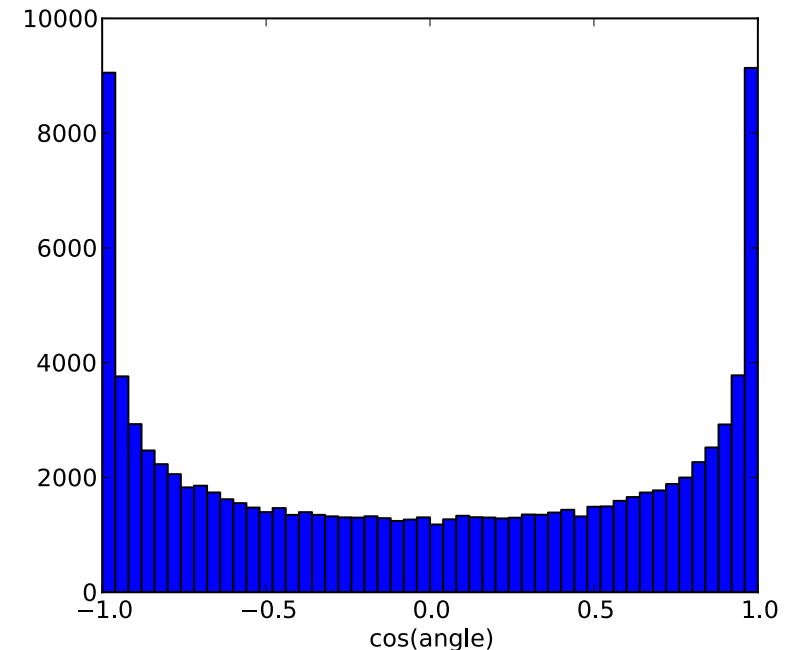
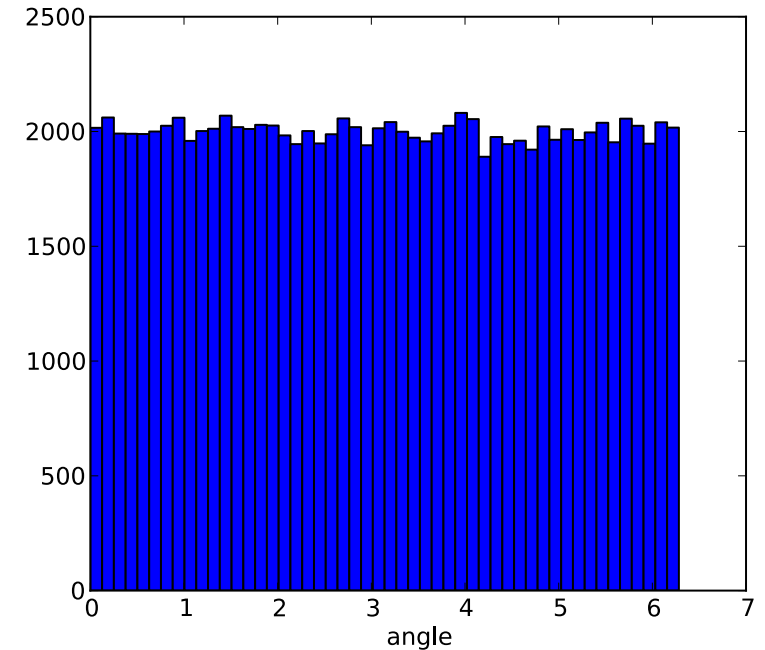
- ▶ General form for multiple parameters:

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

Maximum Likelihood Estimators *asymptotically* reach this bound

## What happens with $x \rightarrow \cos(x)$

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 N_MC=100000 # number of Monte Carlo Experiments
5 nBins = 50 # number of bins for Histograms
6
7 data_x, data_y = [], [] #lists that will hold x and y
8
9 # do experiments
10 for i in range(N_MC):
11     # generate observation for x
12     x = np.random.uniform(0,2*np.pi)
13
14     y = np.cos(x)
15     data_x.append(x)
16     data_y.append(y)
17
18 #setup figures
19 fig = plt.figure(figsize=(13,5))
20 fig_x = fig.add_subplot(1,2,1)
21 fig_y = fig.add_subplot(1,2,2)
22
23 fig_x.hist(data_x,nBins)
24 fig_x.set_xlabel('angle')
25
26 fig_y.hist(data_y,nBins)
27 fig_y.set_xlabel('cos(angle)')
28
29 plt.show()
```





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

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

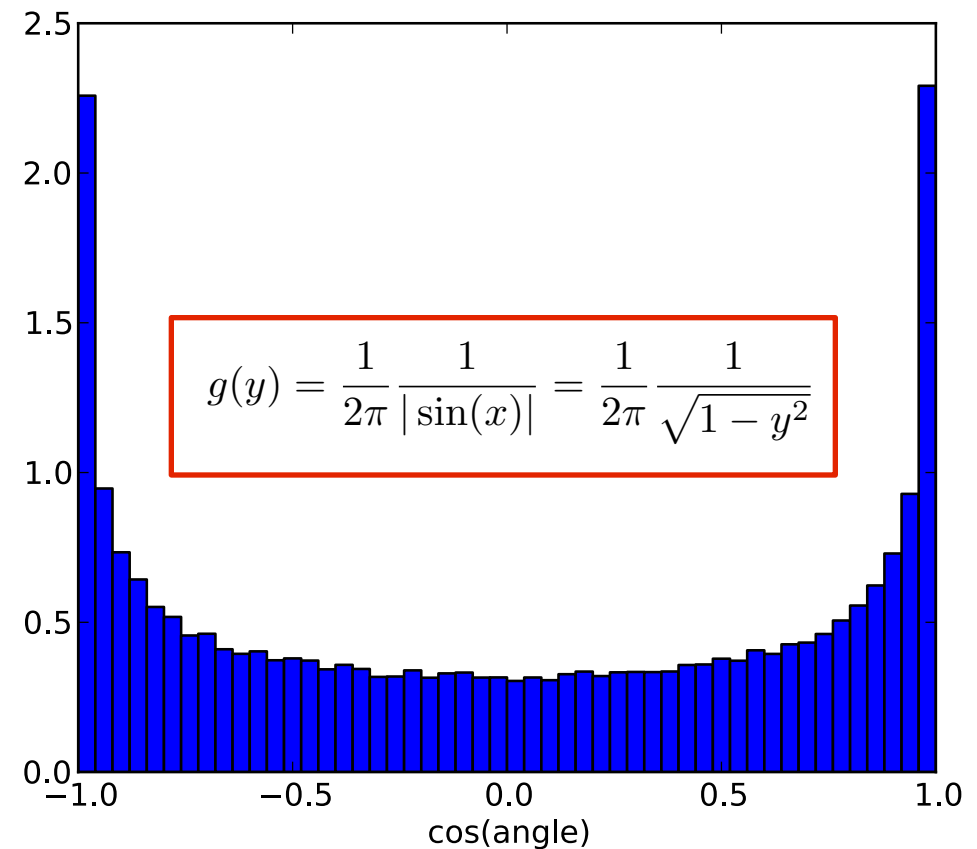
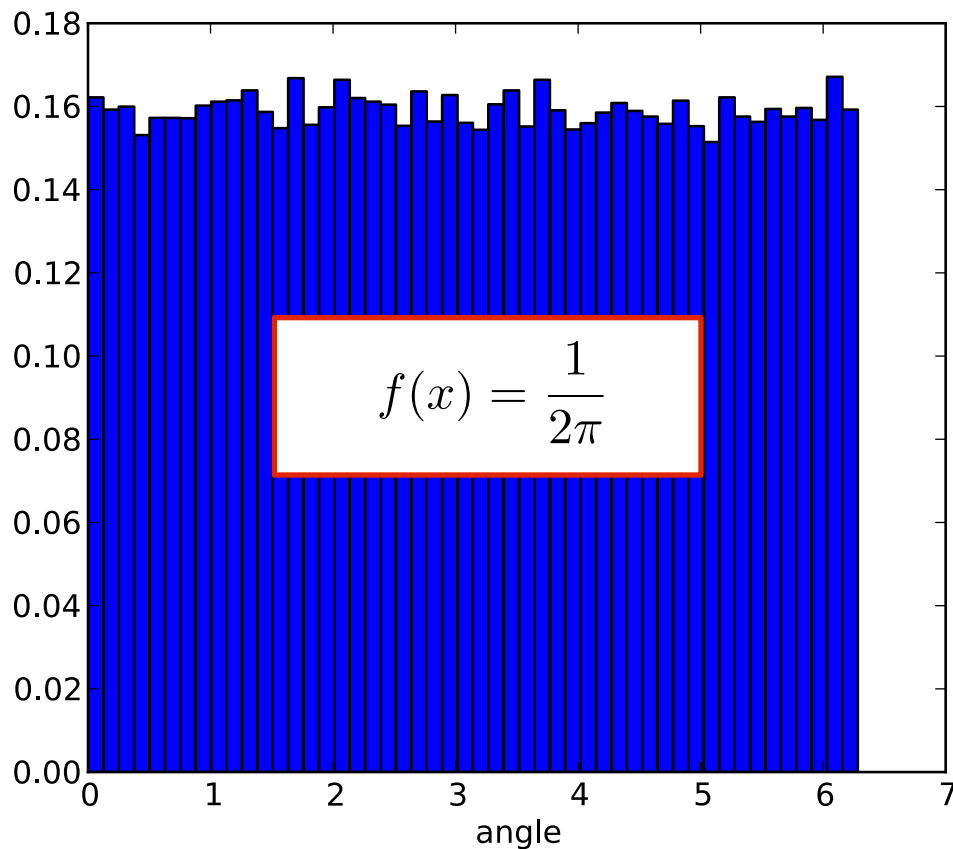
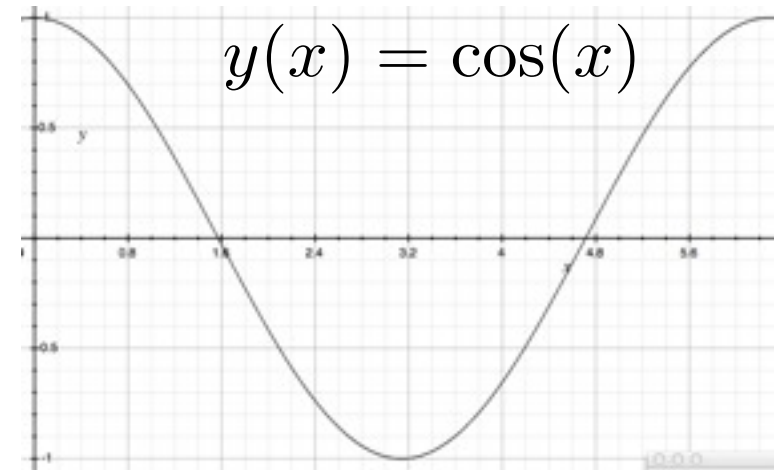
We can rewrite the integral on the right

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

therefore, the two pdfs are related by a Jacobian factor

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

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



## Change of variable $x$ , change of parameter $\theta$

- 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

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

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

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  $\mu$  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  $\text{RAN}()$ .

Bob Cousins, CMS, 2008

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

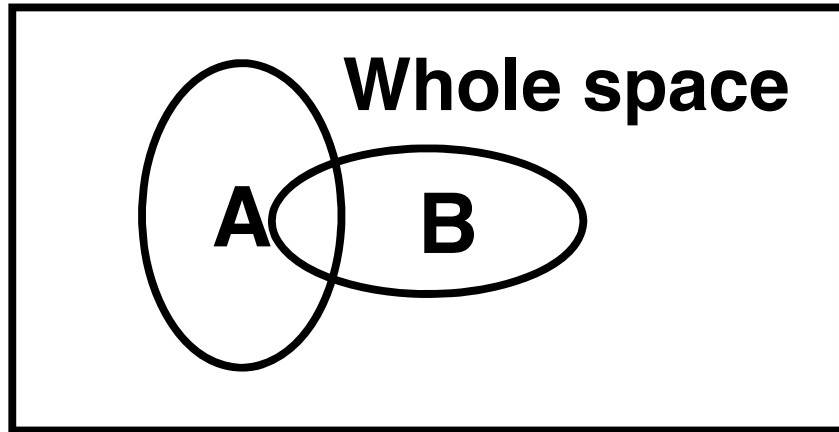
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- **$P(A)$**  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



$$P(A) = \frac{\text{Area of A}}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of B}}{\text{Area of Whole space}}$$

$$P(A|B) = \frac{\text{Area of A} \cap B}{\text{Area of B}}$$

$$P(B|A) = \frac{\text{Area of A} \cap B}{\text{Area of A}}$$

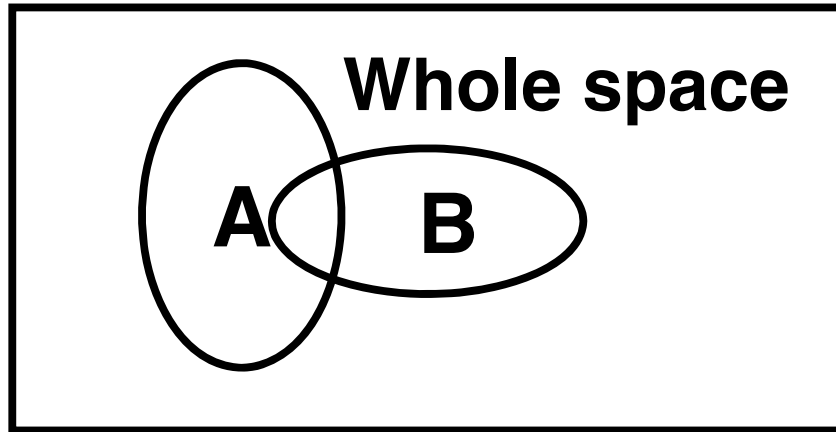
$$P(A \cap B) = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{Area of A}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of A}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{Area of B}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of B}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

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



$$P(A) = \frac{\text{Area of A}}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of B}}{\text{Area of Whole space}}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of B}}$$

$$P(B|A) = \frac{\text{Area of } A \cap B}{\text{Area of A}}$$

$$P(A \cap B) = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}}$$

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

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$



$$P(\text{Data};\text{Theory}) \neq P(\text{Theory};\text{Data})$$

Theory = male or female

Data = pregnant or not pregnant

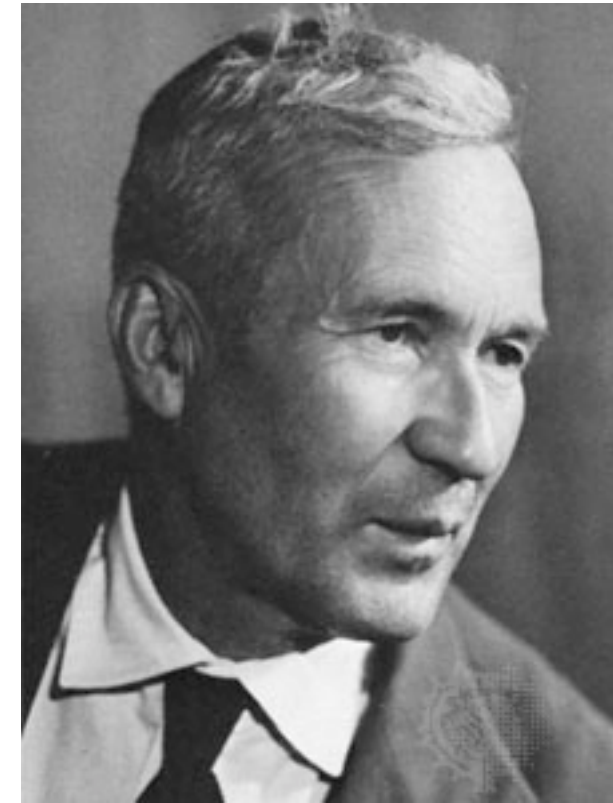
$$P(\text{pregnant ; female}) \sim 3\%$$

but

$$P(\text{female ; pregnant}) \gg \gg 3\%$$

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  $P(E_1 \cup E_2 \cup \dots) = \sum_i P(E_i).$



Kolmogorov  
axioms (1933)

Consequences:

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

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

## 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



## 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

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

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



# Part 2



# Modeling: The Scientific Narrative

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

- ▶ by “model”, I mean the full structure of  $P(\text{data} \mid \text{parameters})$ 
  - 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 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  $\Omega$  defines initial measure, sampled via Monte Carlo

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

$$P \rightarrow L\sigma$$

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

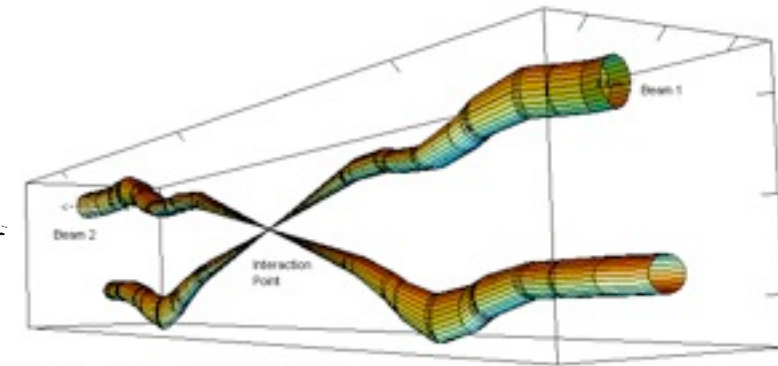


- 1) The language of the Standard Model is Quantum Field Theory  
Phase space  $\Omega$  defines initial measure, sampled via Monte Carlo

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

$$P \rightarrow L\sigma$$

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



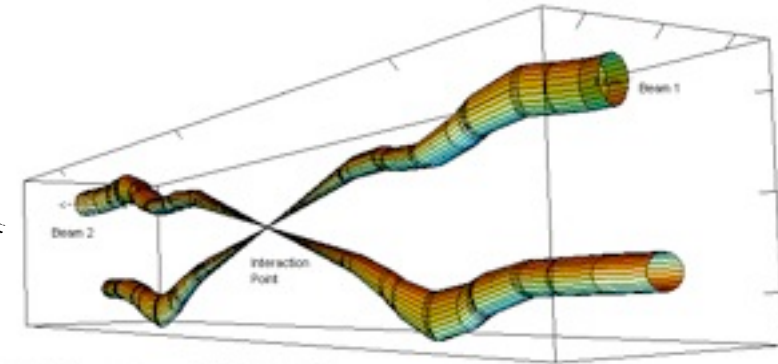
Relative beam sizes around IP1 (Atlas) in collision

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

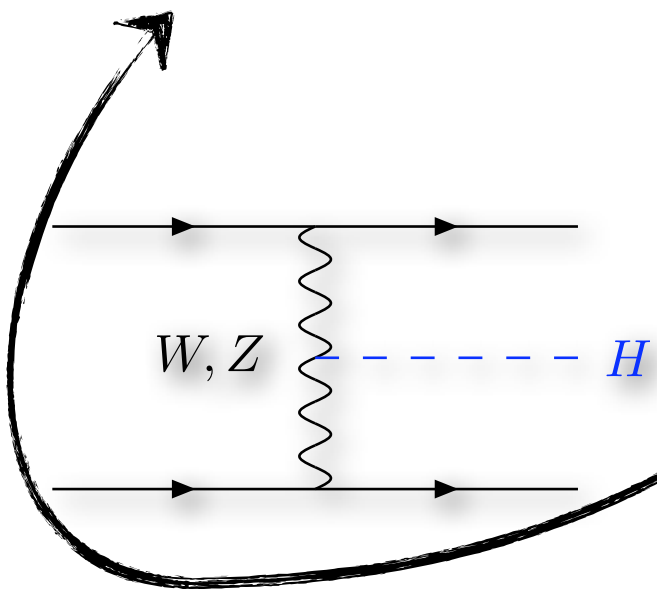
$$P = \frac{|\langle f|i \rangle|^2}{\langle f|f \rangle \langle i|i \rangle}$$

$$P \rightarrow L\sigma$$

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

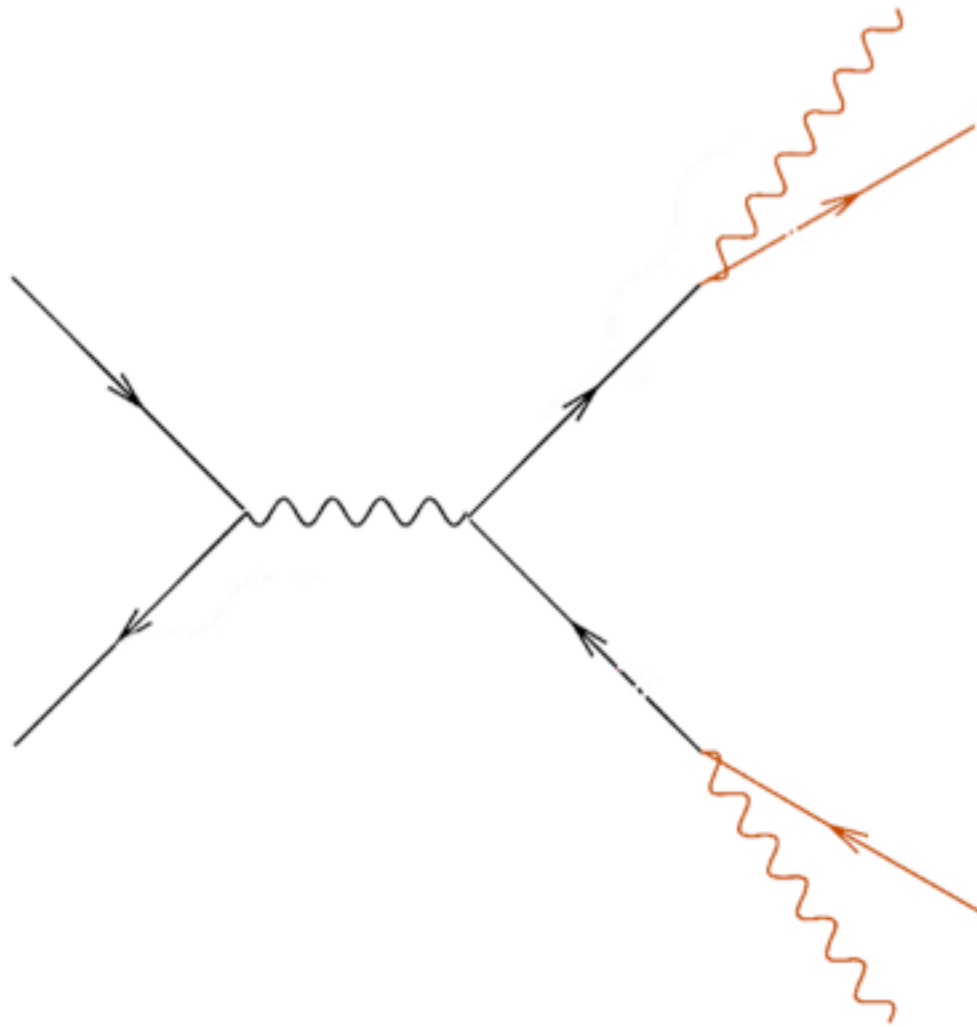


Relative beam sizes around IP1 (Atlas) in collision



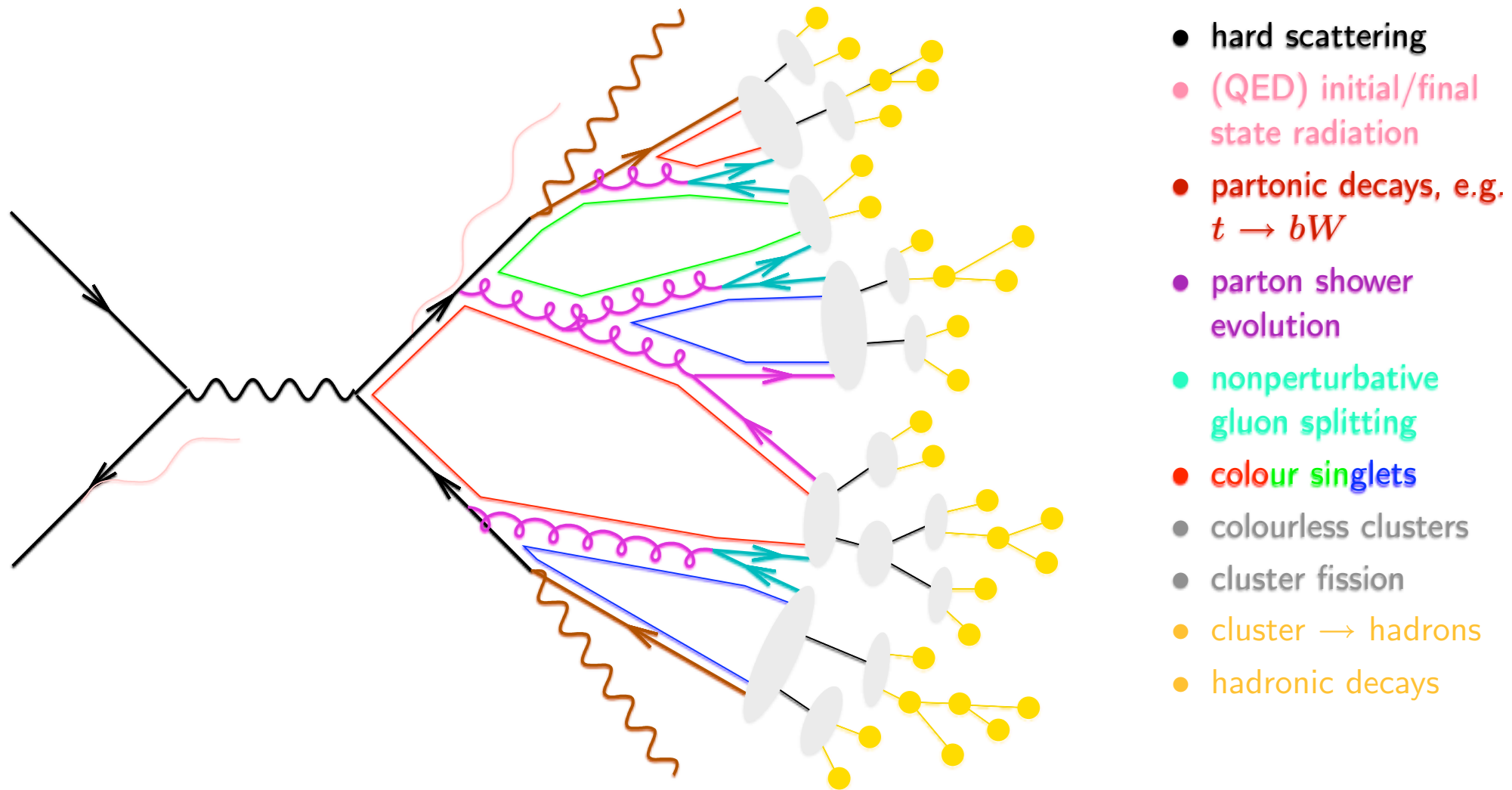
$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} + \underbrace{\bar{L}\gamma^\mu (i\partial_\mu - \frac{1}{2}g\boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2}g'Y B_\mu)L + \bar{R}\gamma^\mu (i\partial_\mu - \frac{1}{2}g'Y B_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}} + \underbrace{\frac{1}{2} |(i\partial_\mu - \frac{1}{2}g\boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2}g'Y B_\mu)\phi|^2 - V(\phi)}_{\text{W}^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} + \underbrace{g''(\bar{q}\gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L}\phi R + G_2 \bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

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

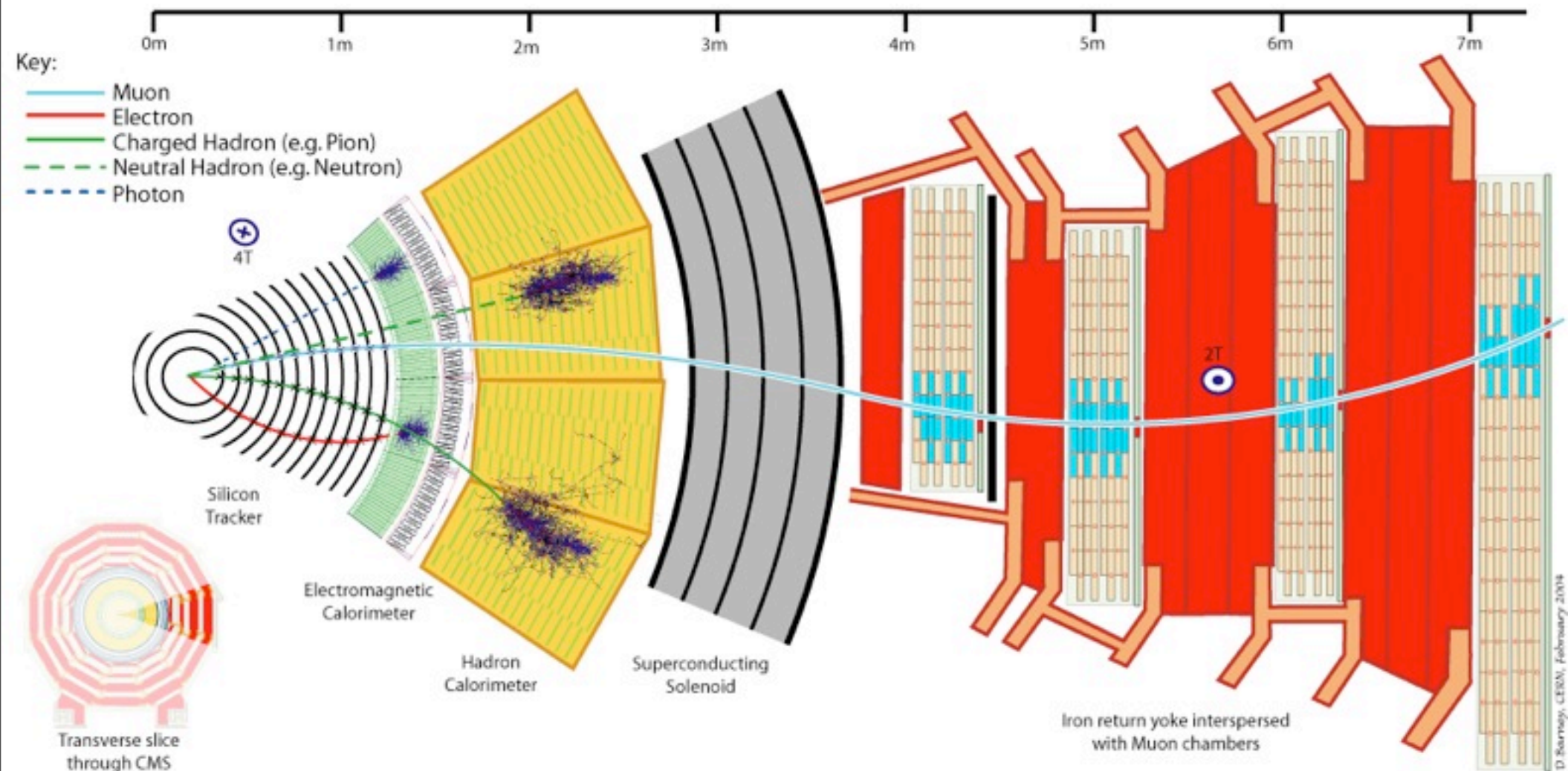


- hard scattering  
 $\sigma(\text{partons}) \sim \alpha_s^2$
- partonic decays, e.g.  
 $t \rightarrow bW$

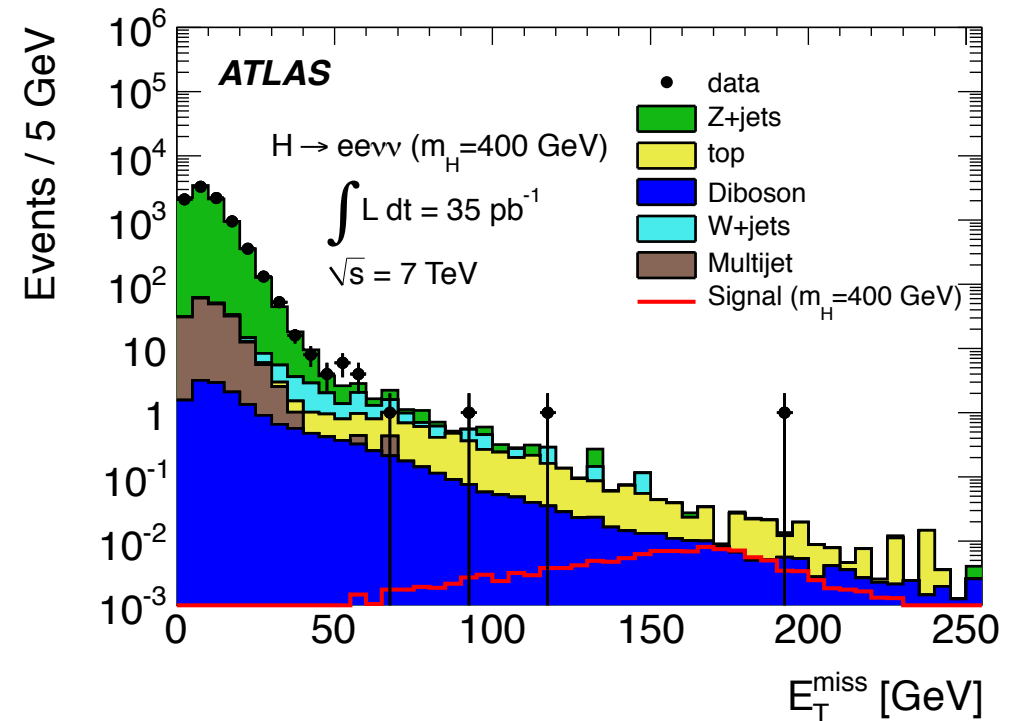
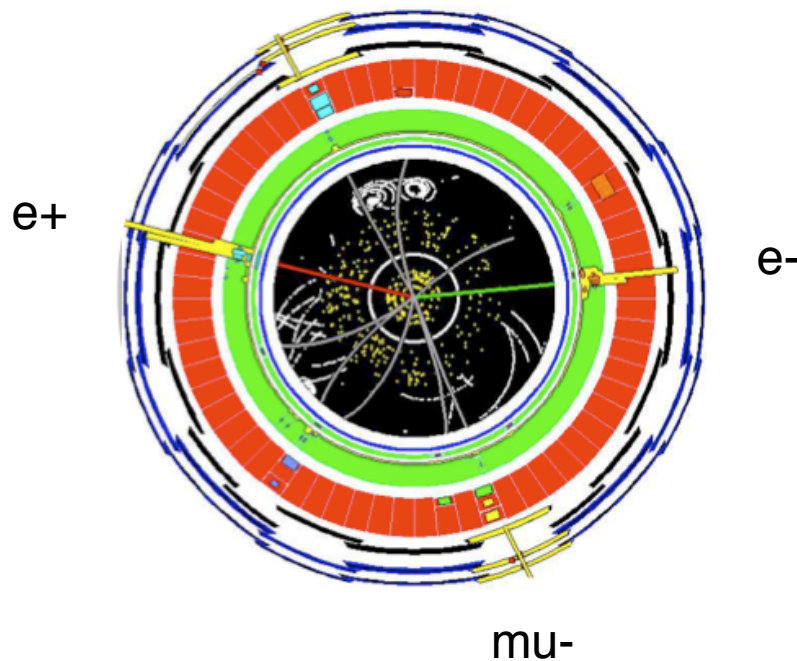
- 2) a) Perturbation theory used to systematically approximate the theory.  
b) splitting functions, Sudakov form factors, and hadronization models  
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**



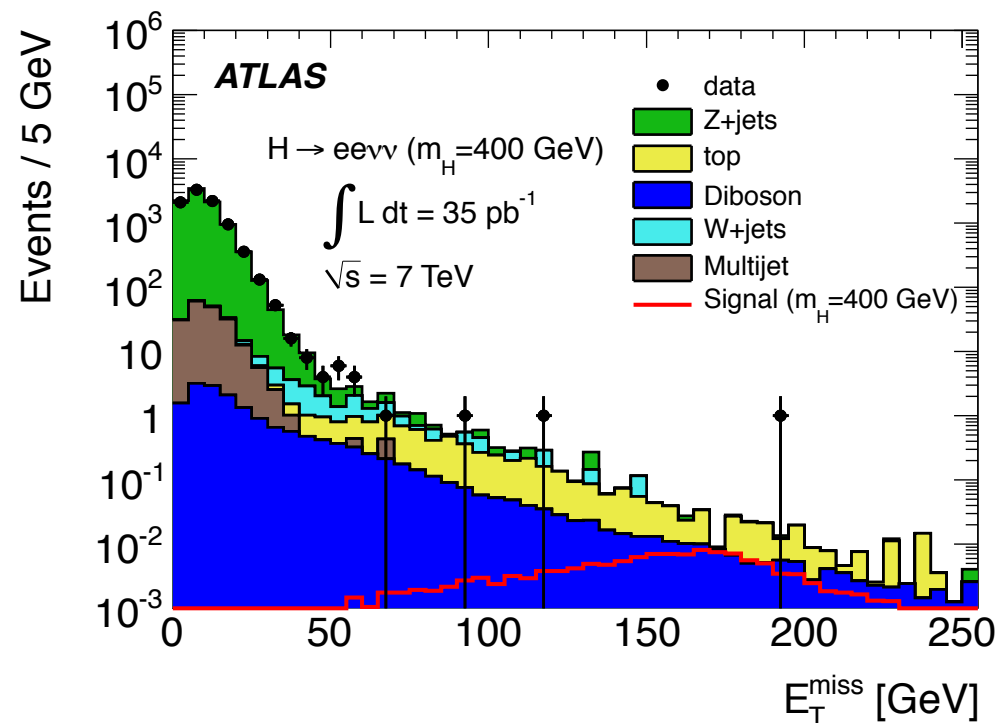
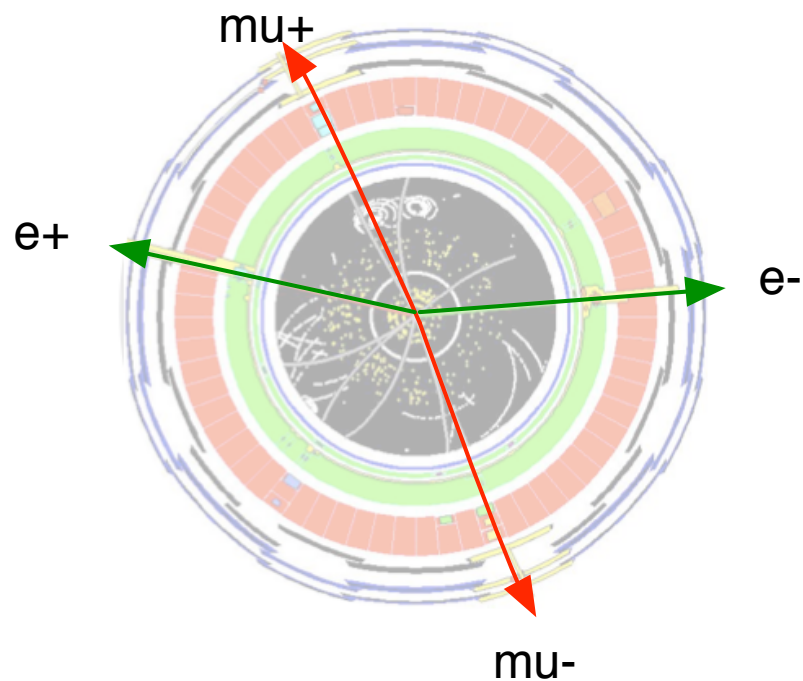
3) Next, the interaction of outgoing particles with the detector is simulated. Detailed simulations of particle interactions with matter. Accept/reject style Monte Carlo integration of very complicated function  $P(\text{detector readout} \mid \text{initial particles})$



- 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)**



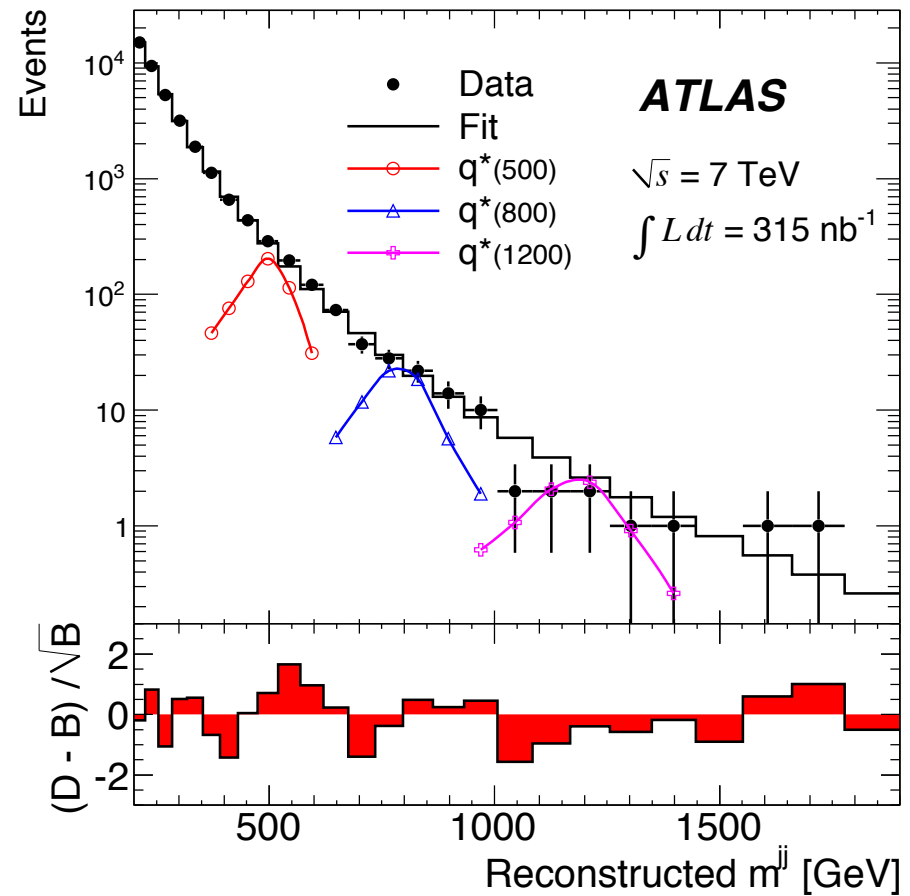
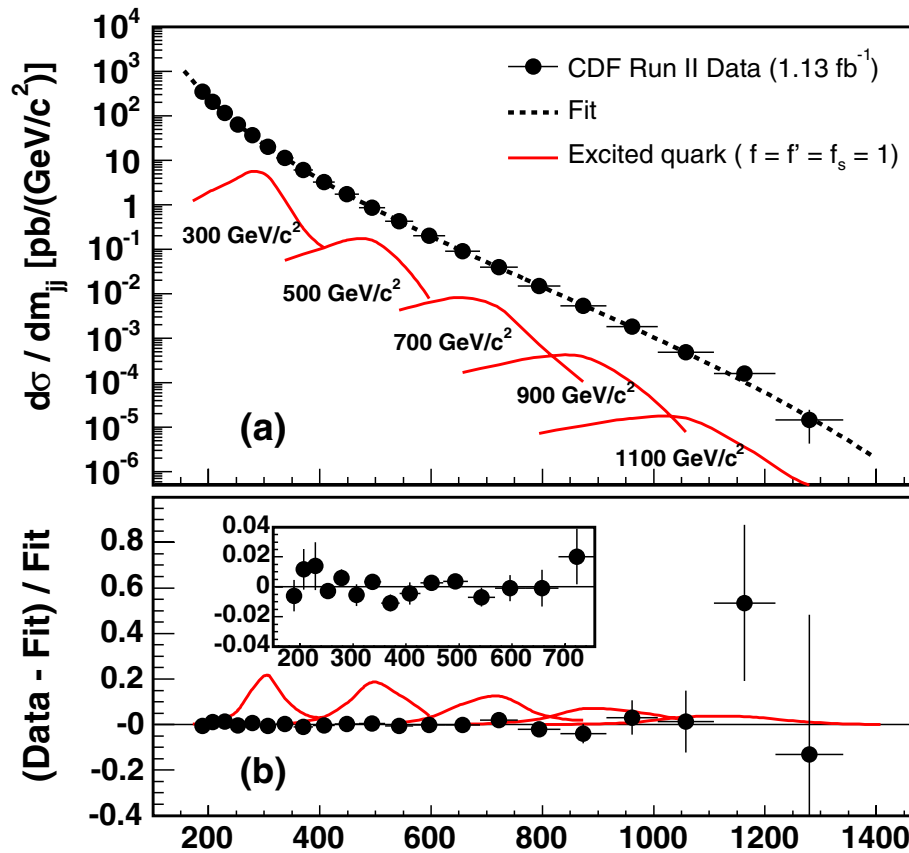
- 4) From the simulated response of the detector, we run reconstruction algorithms on the simulated data as if it were from real data. This allows us to look at distribution of any observable that we can measure in data.  
**P( observable | detector readout)**



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)

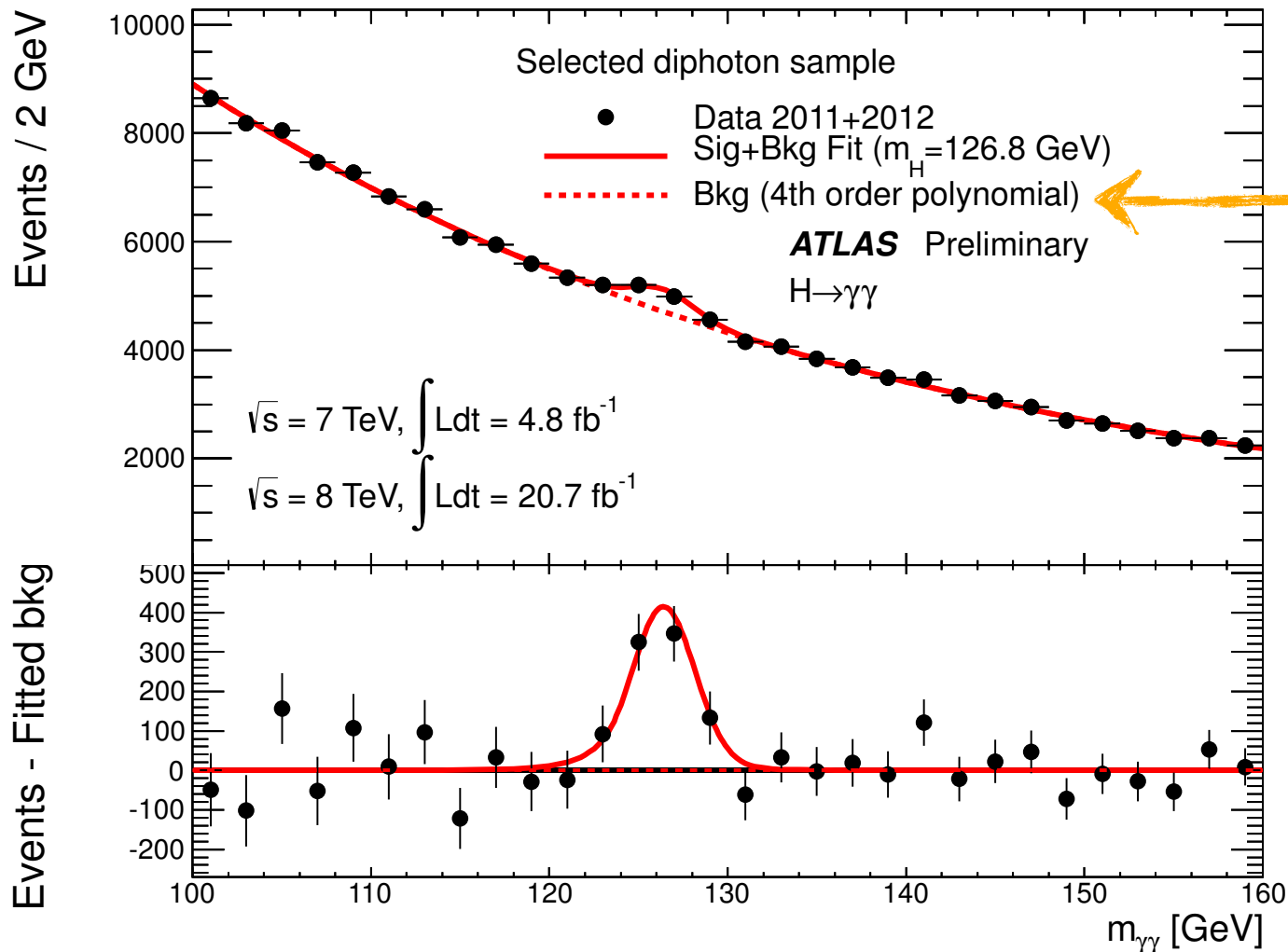


$$\frac{d\sigma}{dm_{jj}} = p_0(1 - x)^{p_1} / x^{p_2 + p_3 \cdot \ln(x)}, \quad x = m_{jj} / \sqrt{s},$$



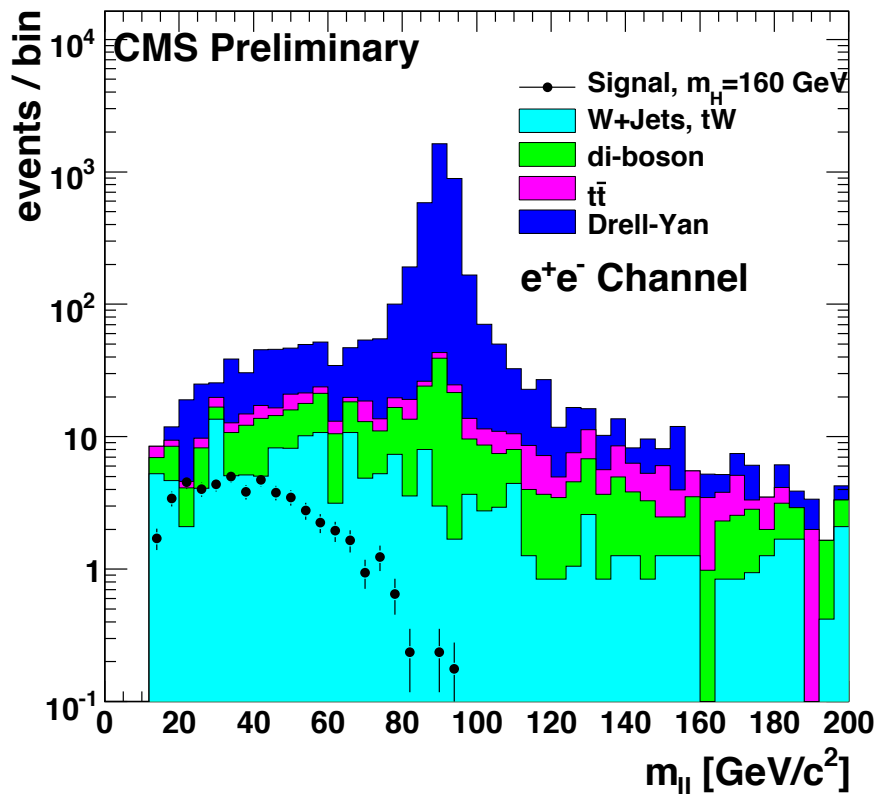
In contrast, one can describe a distribution with some parametric function

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

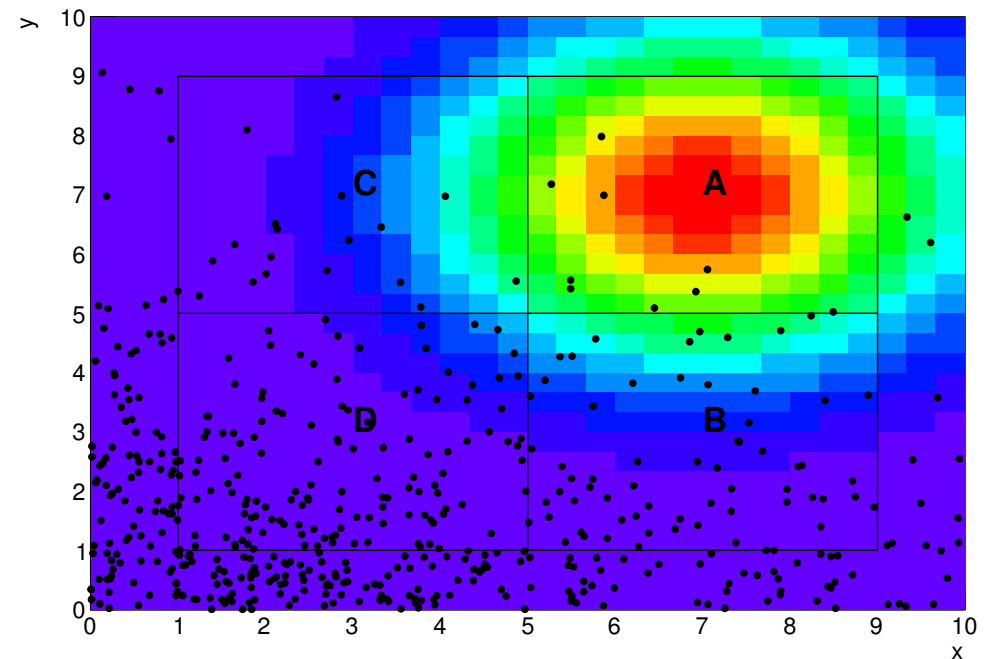


## Regions in the data with negligible signal expected used as control samples

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

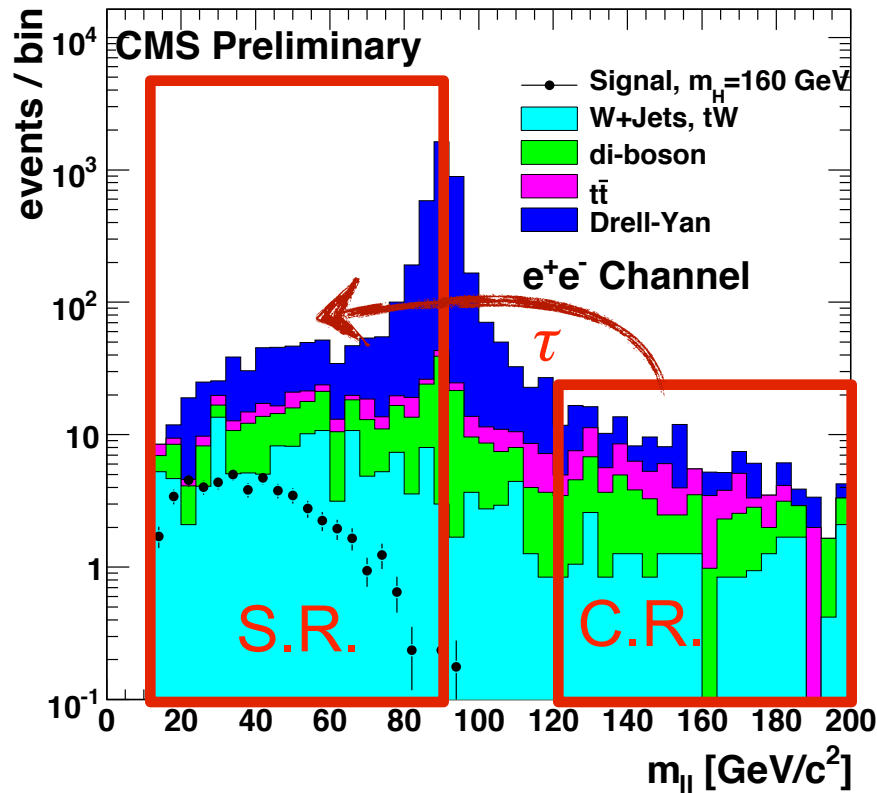


## “ABCD” method

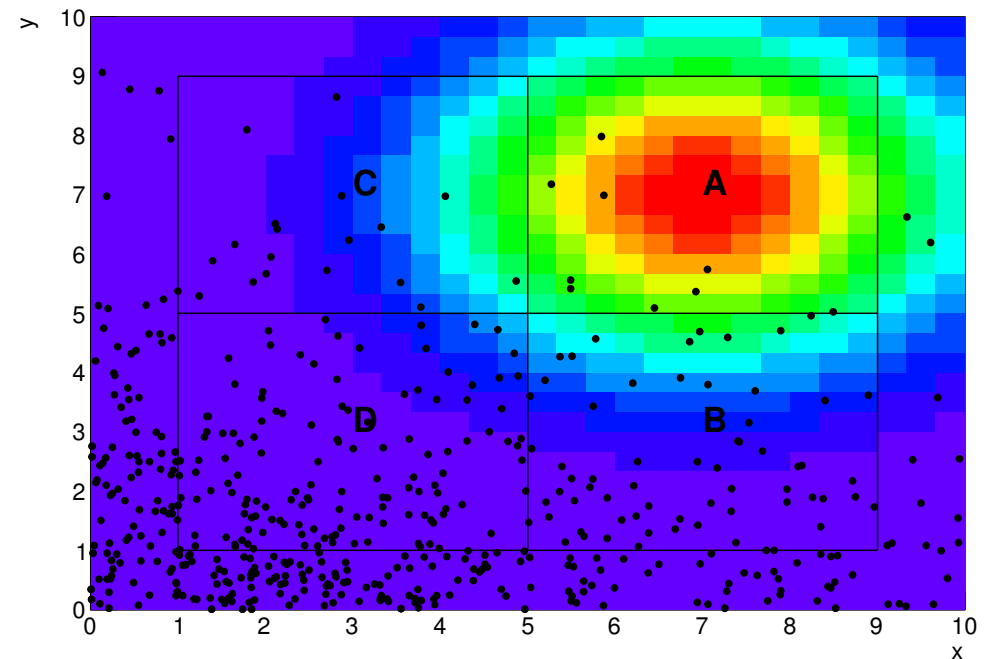


## Regions in the data with negligible signal expected used as control samples

- 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  $n_{\text{on}}$  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  $\pi(b)$ 
    - 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  $\pi(b)$  come from?
  - did you realize that this is a Bayesian procedure that depends on some prior assumption about what  $b$  is?

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  $n_{\text{on}}$  with  $s+b$  expected
- sideband measurement: observe  $n_{\text{off}}$  with  $\tau b$  expected

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

- 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?

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

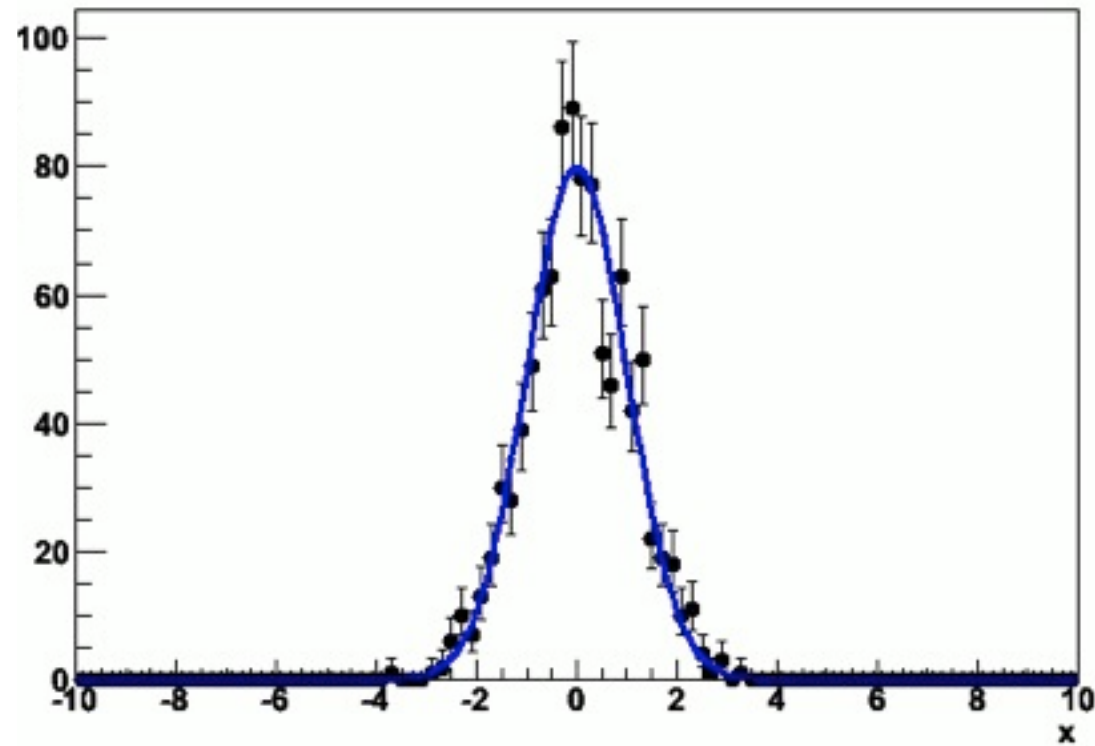
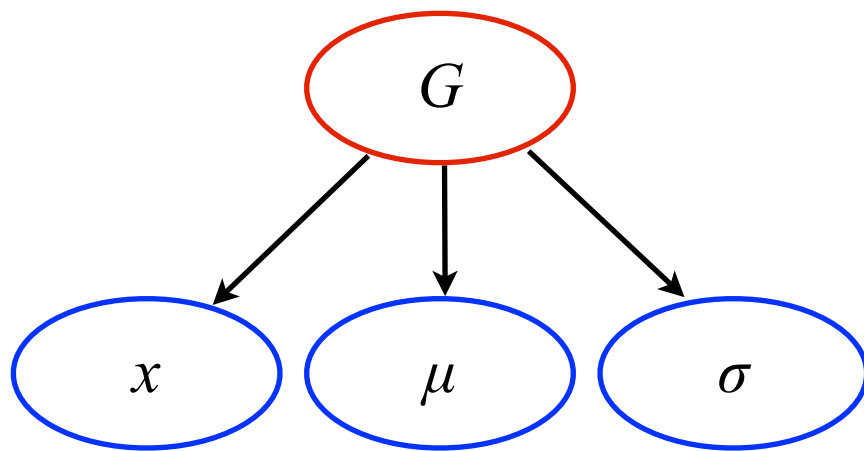
▶ while  $\pi(b)$  is based on data, it still depends on some original prior  $\eta(b)$

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

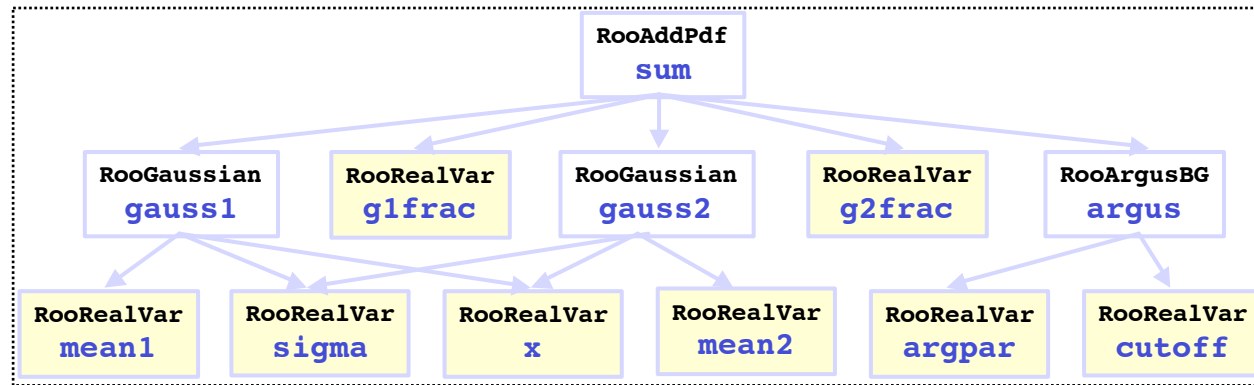
# A General Purpose Statistical Model

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

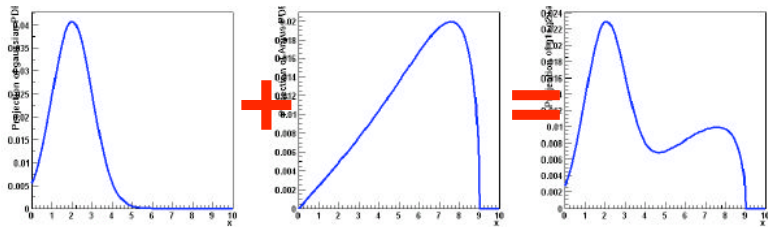
- ▶ eg. a Gaussian  $G(x|\mu, \sigma)$  is parametrized by  $(\mu, \sigma)$
- ▶ every node is a real-valued function of the nodes below



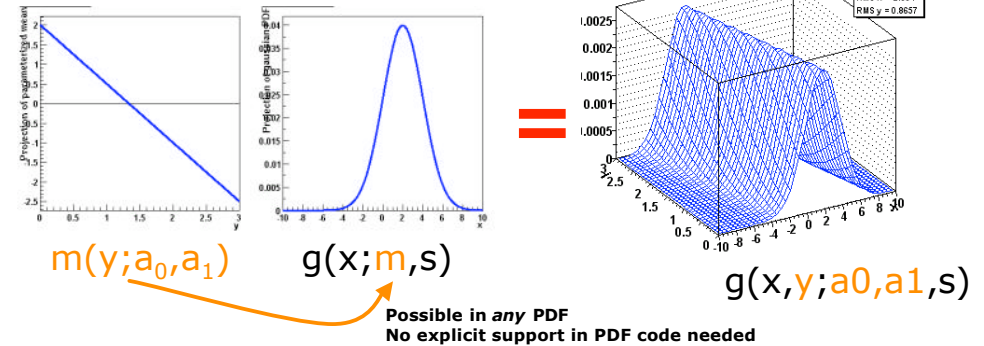
RooFit is a major tool developed at BaBar for data modeling.  
RooStats provides higher-level statistical tools based on these PDFs.



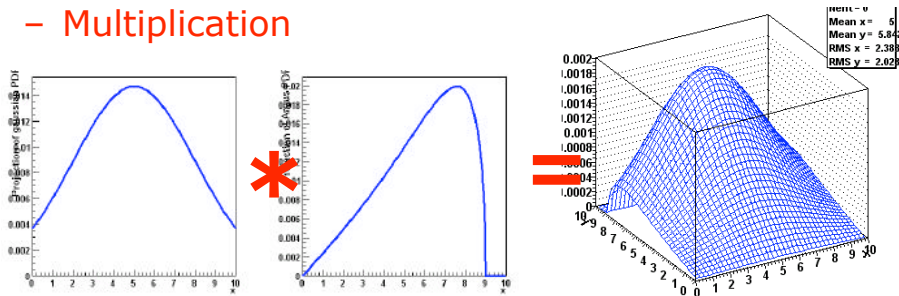
## - Addition



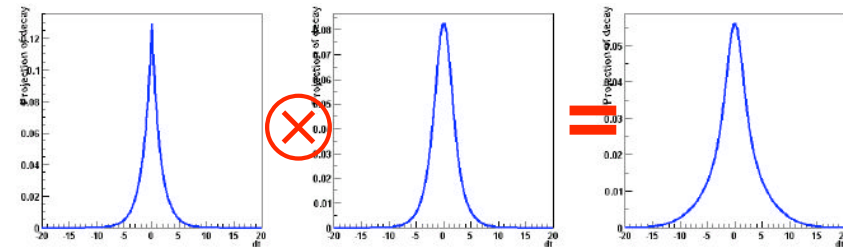
## - Composition ('plug & play')



## - Multiplication



## - Convolution



Wouter Verkerke,

Wouter Verkerke, UCSB



**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
- ▶  $\nu$ : 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  $x$

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

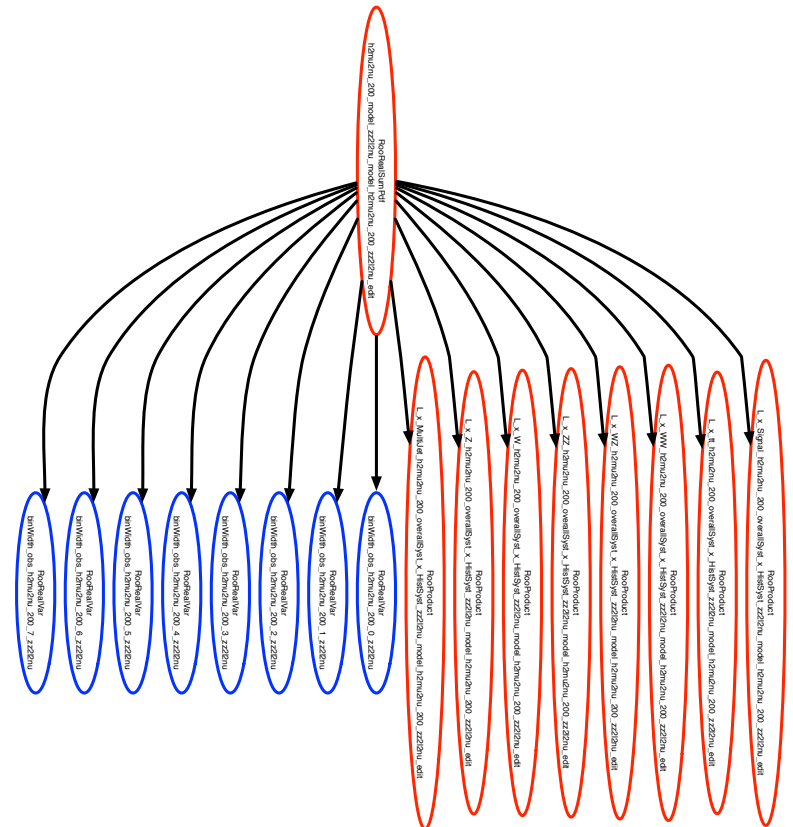
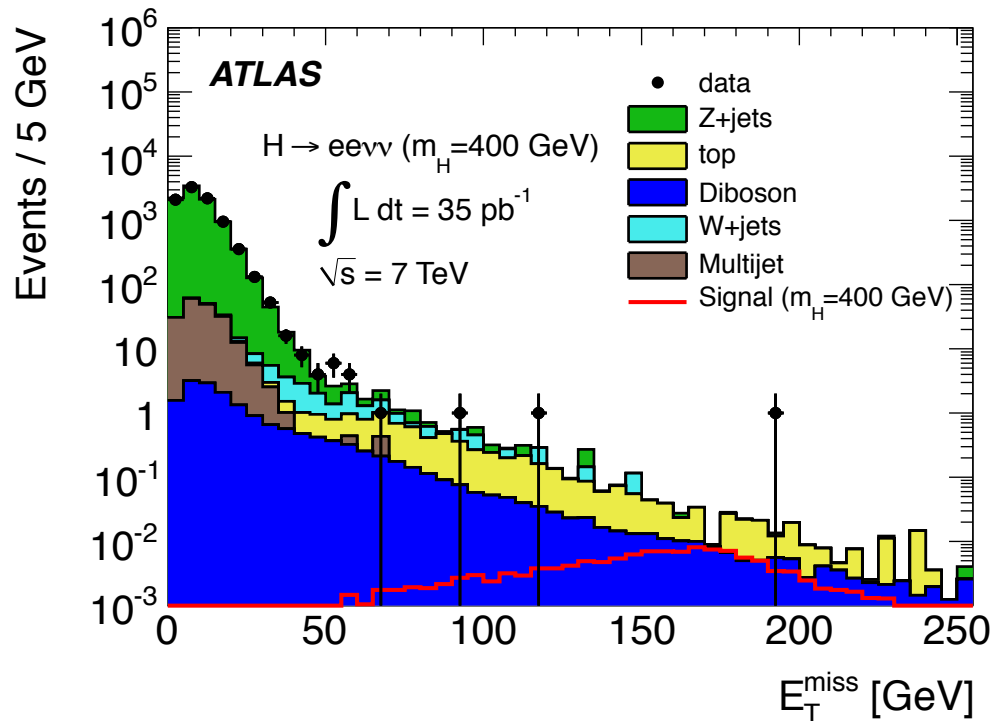
**Marked Poisson Process:**

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

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

▶ statisticians call this a mixture model

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



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

- ▶ the mass of a hypothesized particle
- ▶ the “signal strength”  $\mu=0$  no signal,  $\mu=1$  predicted signal rate

**Nuisance parameters ( $\theta$  or  $\alpha_p$ ):** associated to uncertainty in:

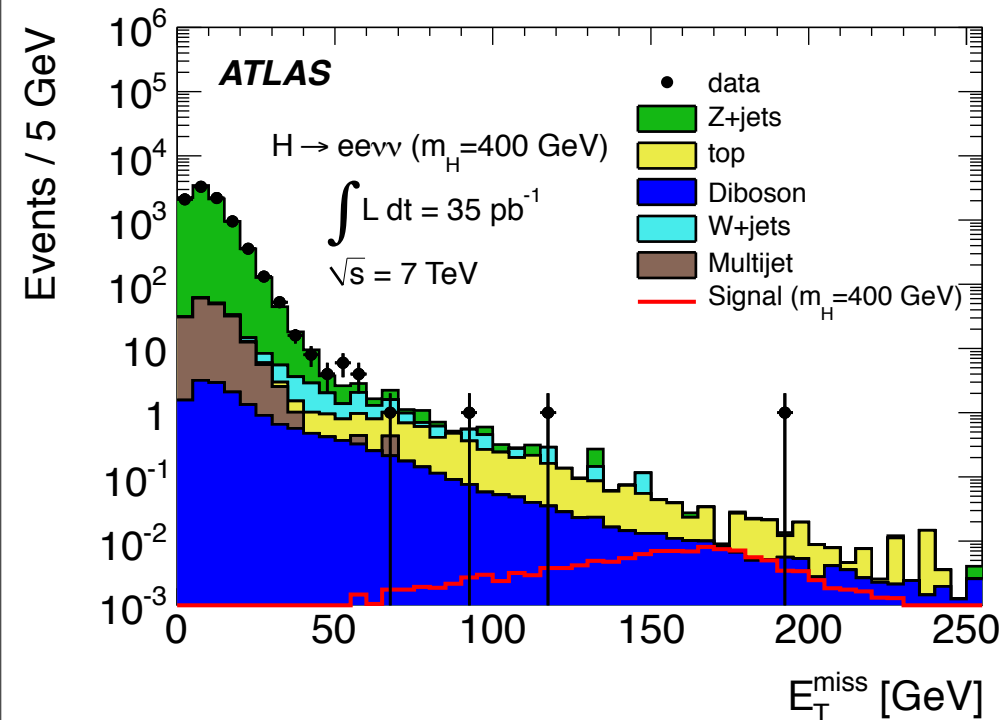
- ▶ response of the detector (calibration)
- ▶ phenomenological model of interaction in non-perturbative regime

**Lead to a parametrized model:**  $\nu \rightarrow \nu(\alpha), f(x) \rightarrow f(x|\alpha)$

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

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize  $p^{th}$  variation in terms of **nuisance parameter**  $\alpha_p$

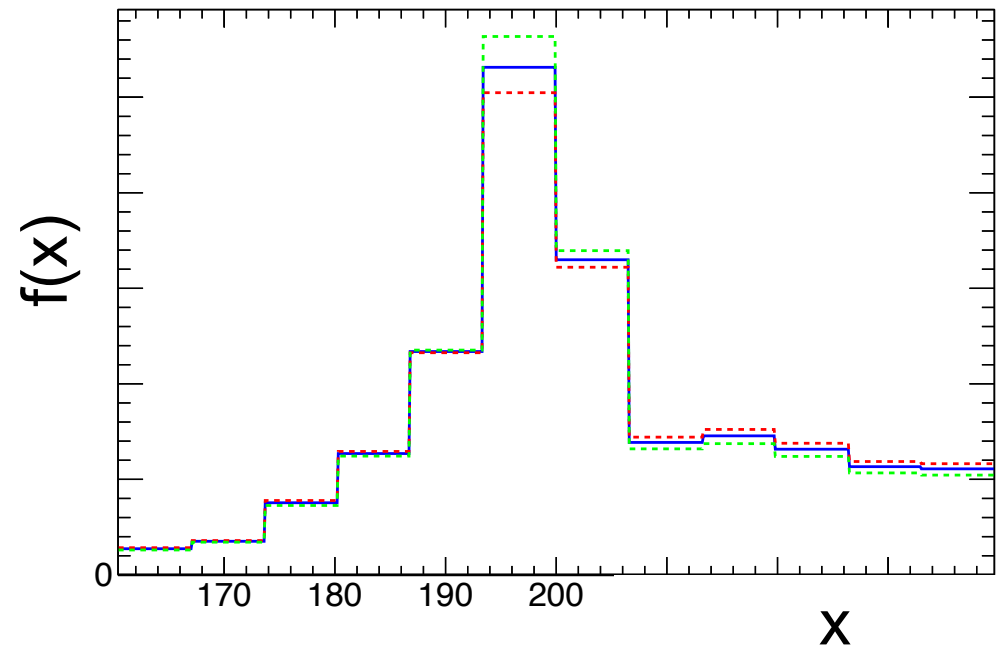
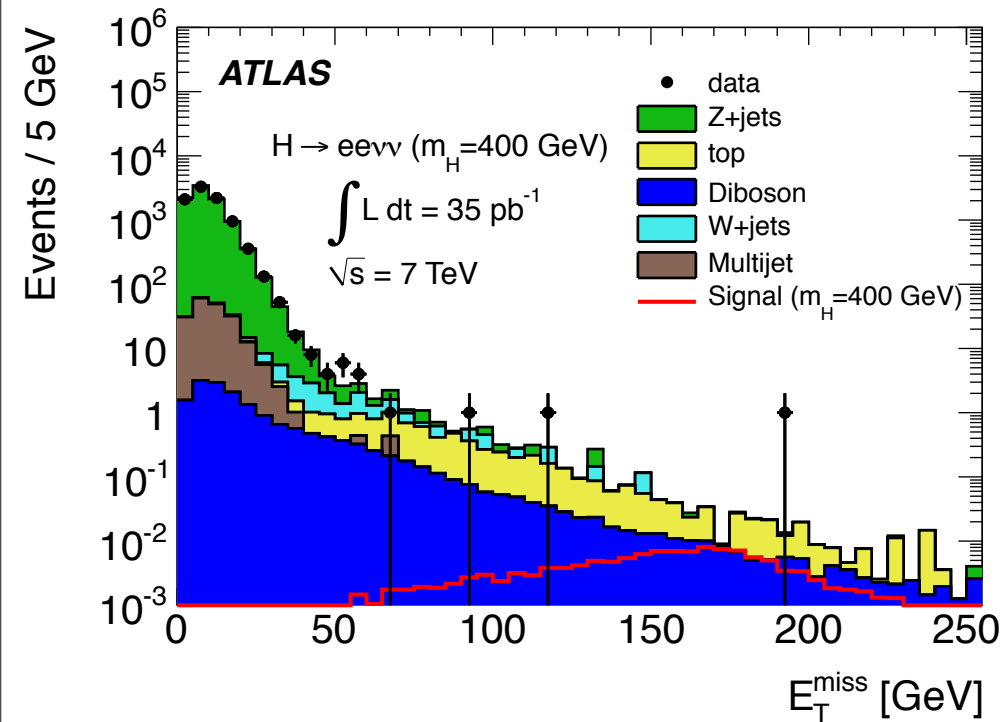


	Z+jets	top	Diboson	...
syst 1				
syst 2				
...				

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

Tabulate effect of individual variations of sources of systematic uncertainty

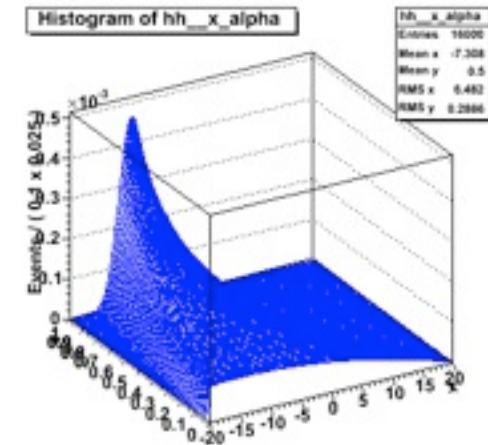
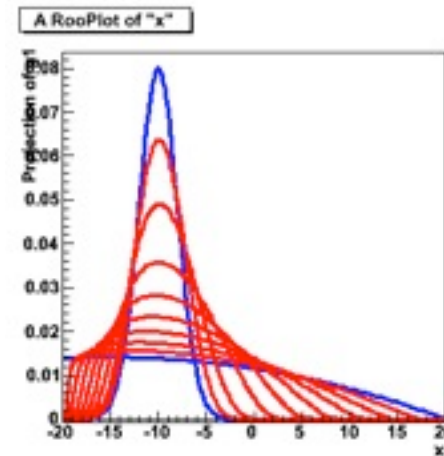
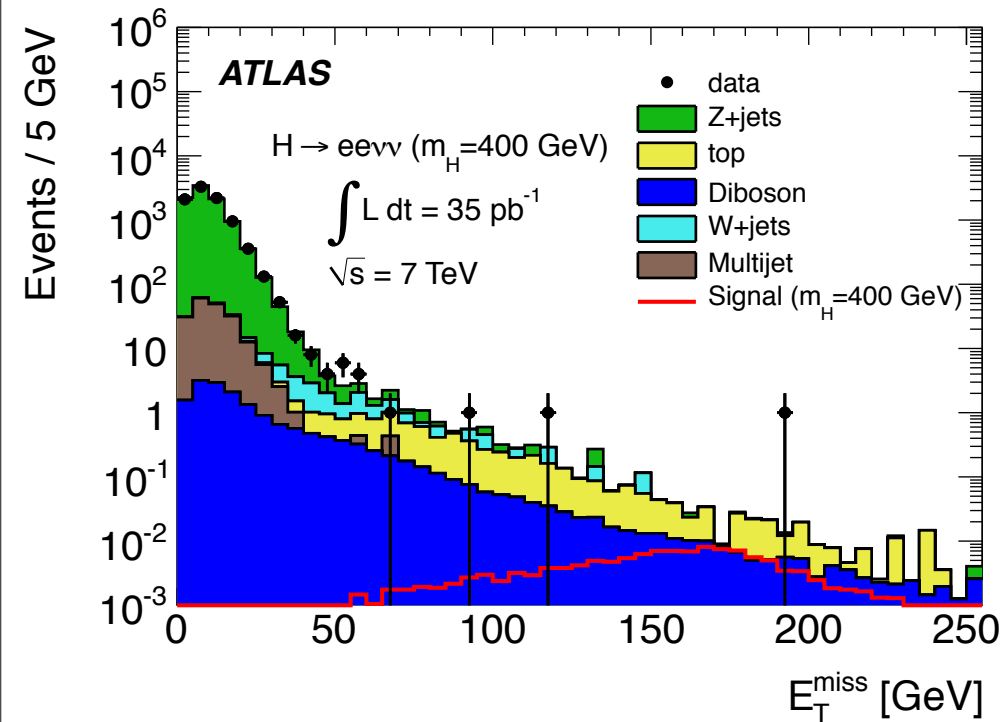
- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize  $p^{th}$  variation in terms of **nuisance parameter**  $\alpha_p$



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

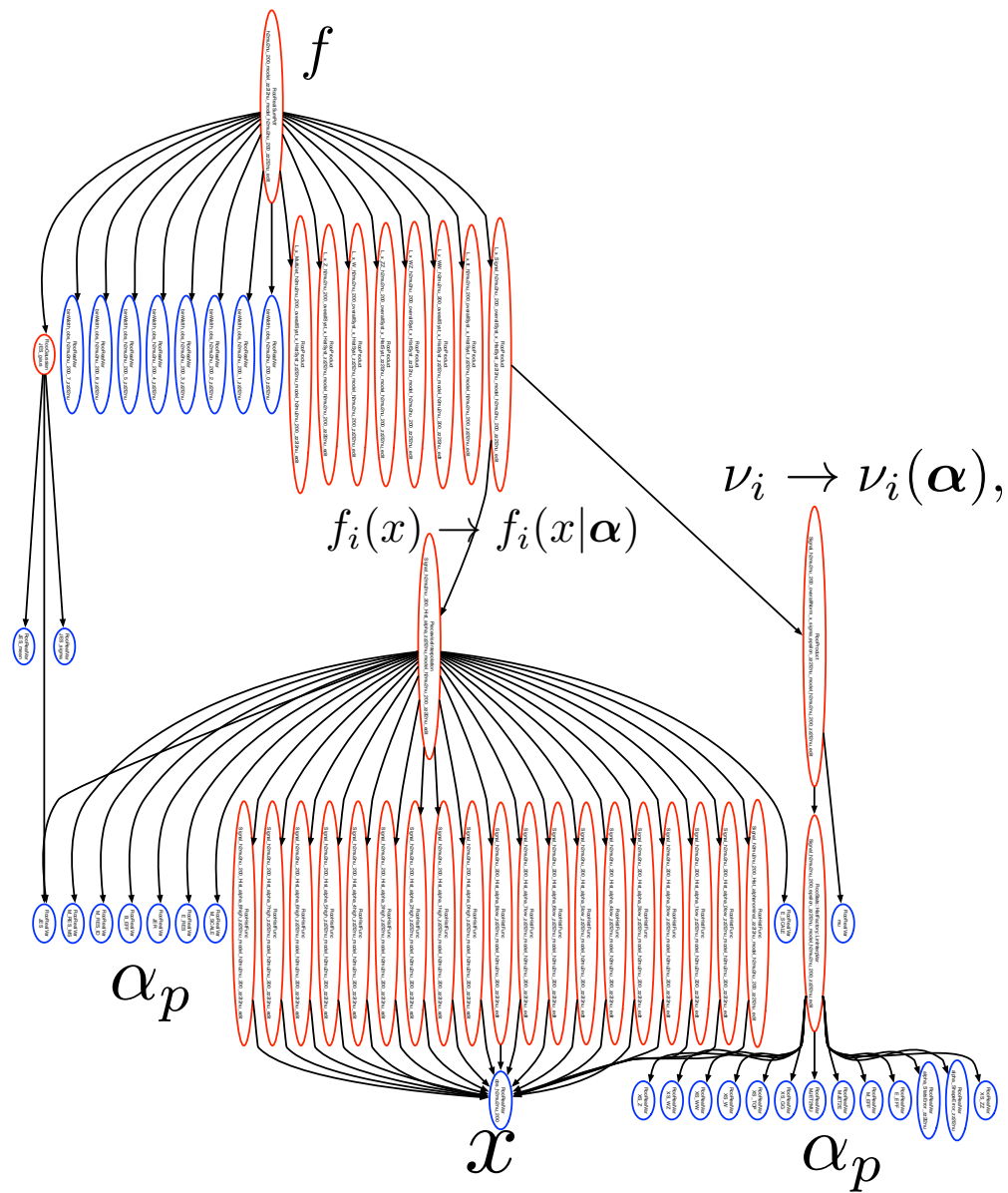
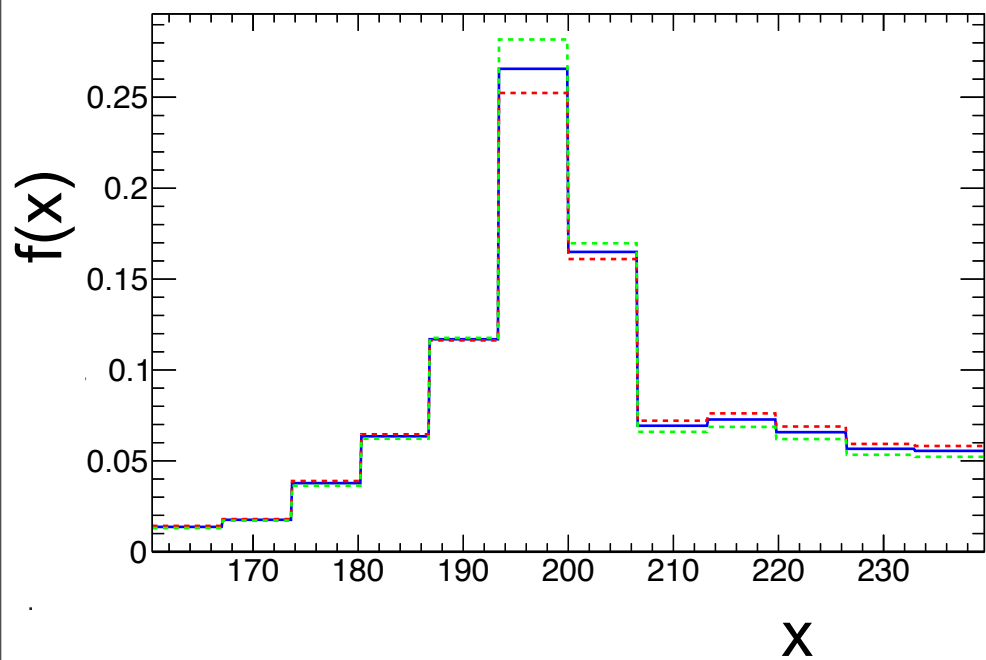
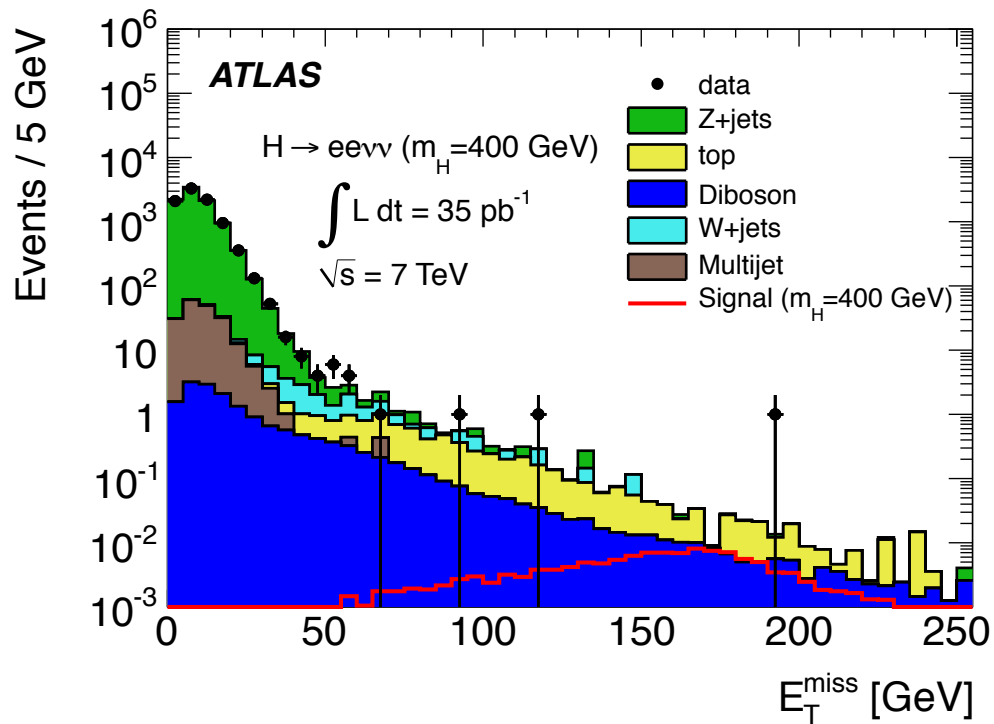
Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize  $p^{th}$  variation in terms of **nuisance parameter**  $\alpha_p$

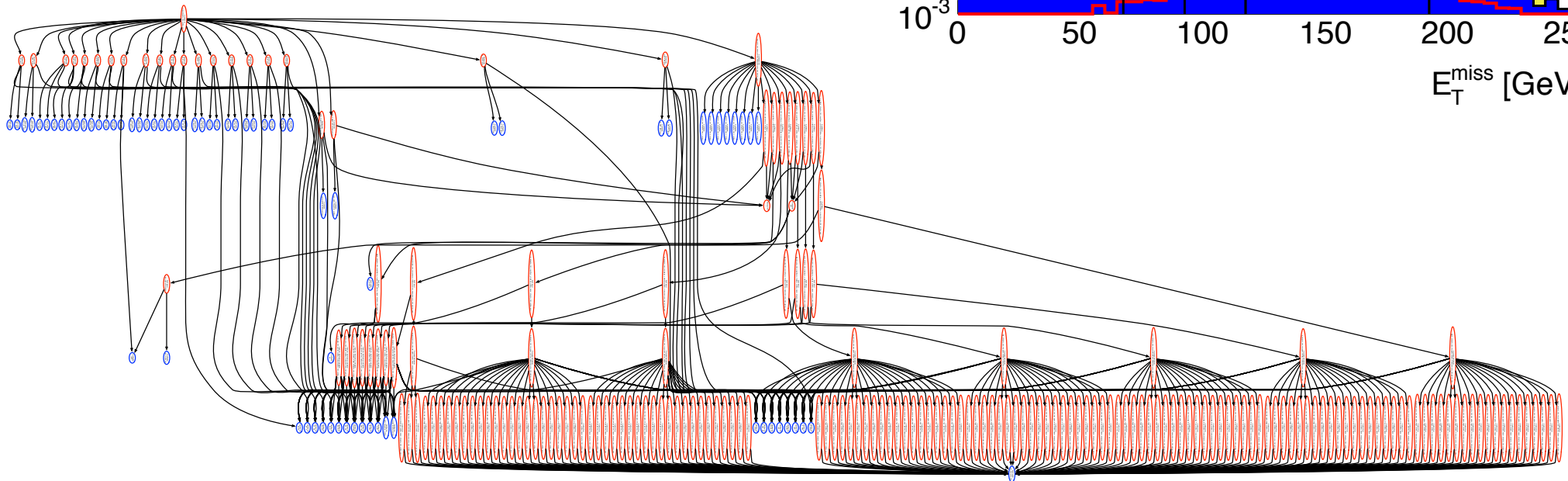
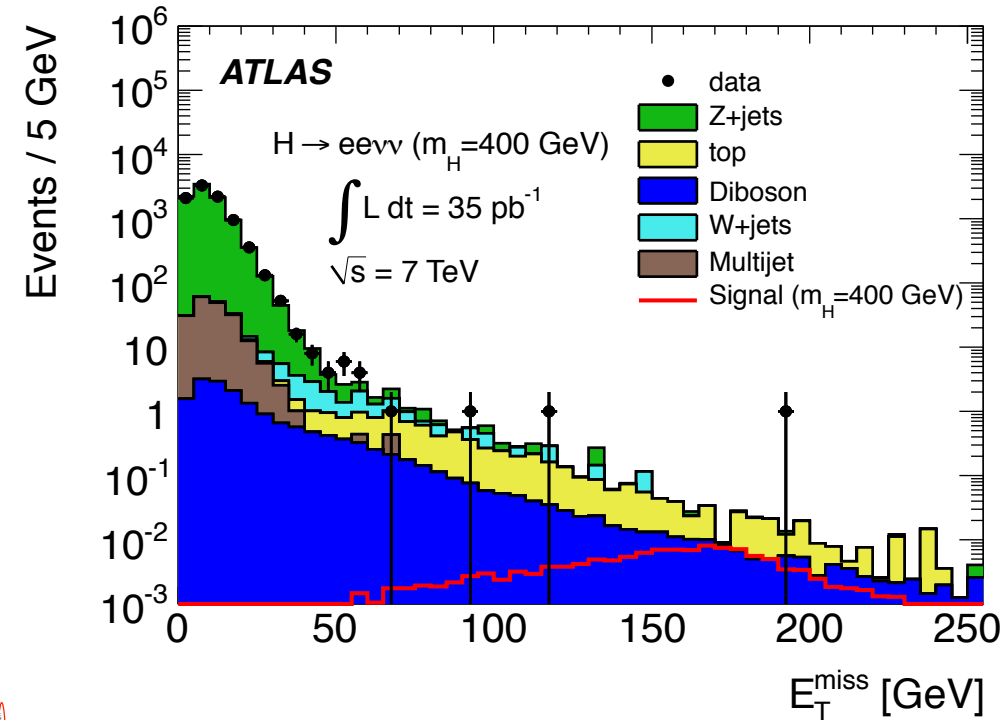


$$f(\mathcal{D}|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^n f(x_e|\alpha)$$

# 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:** 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.

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

where  $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \dots, \mathcal{D}_{c_{\text{max}}}\}$

**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  $\nu_b$

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

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

- ▶ one typically has MLE for  $\alpha_p$ , denoted  $a_p$  and standard error

**Constraint Terms:** are idealized pdfs for the MLE.

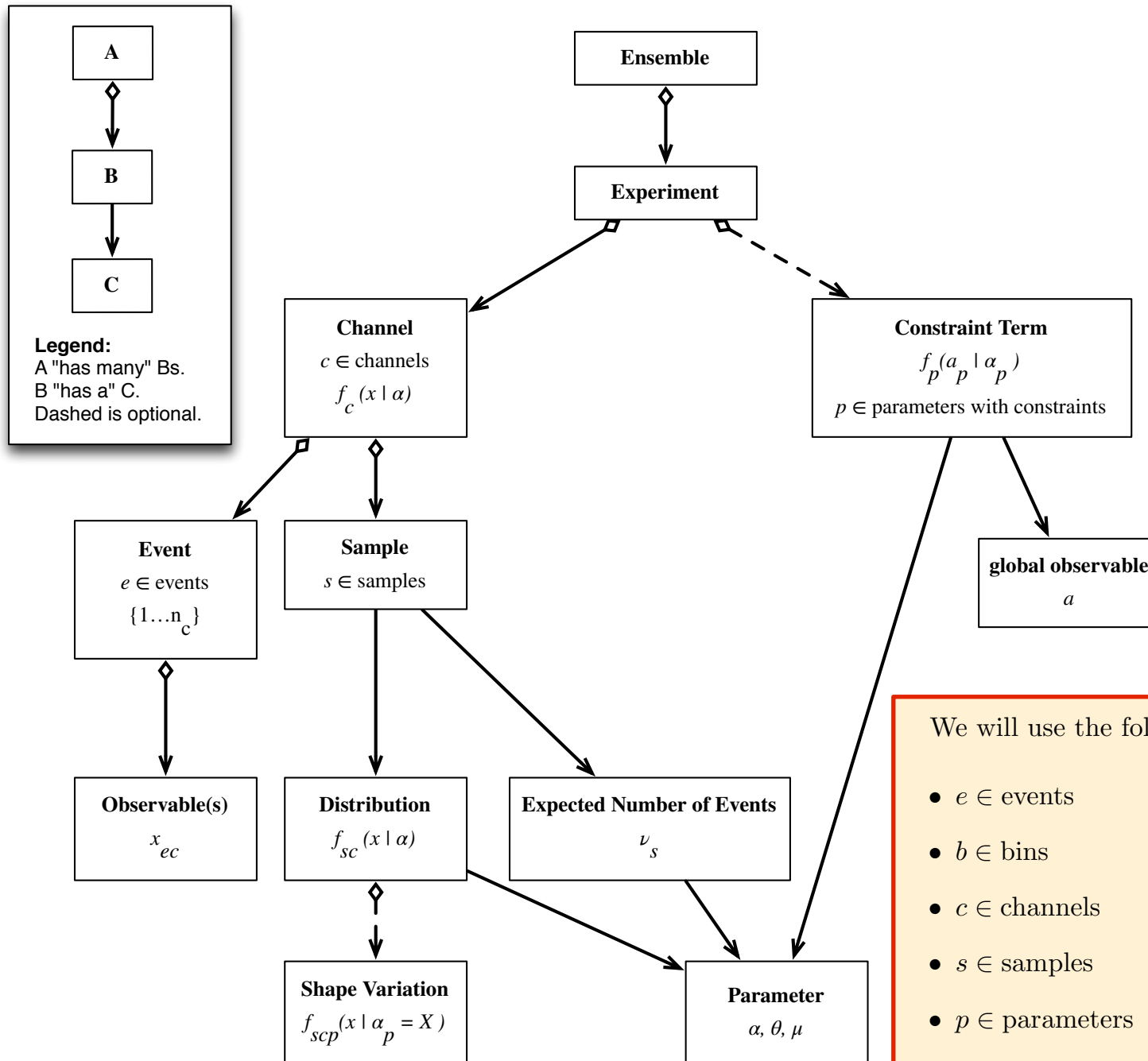
$$f_p(a_p|\alpha_p) \text{ for } p \in \mathbb{S}$$

- ▶ 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  $\eta$

**Simultaneous Multi-Channel Model with constraints:**

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

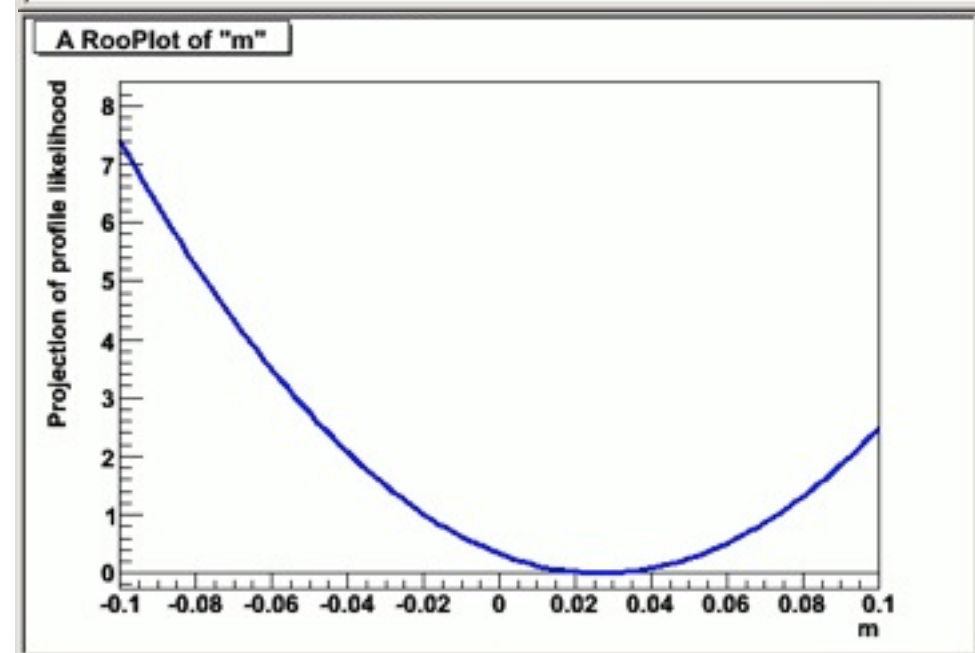
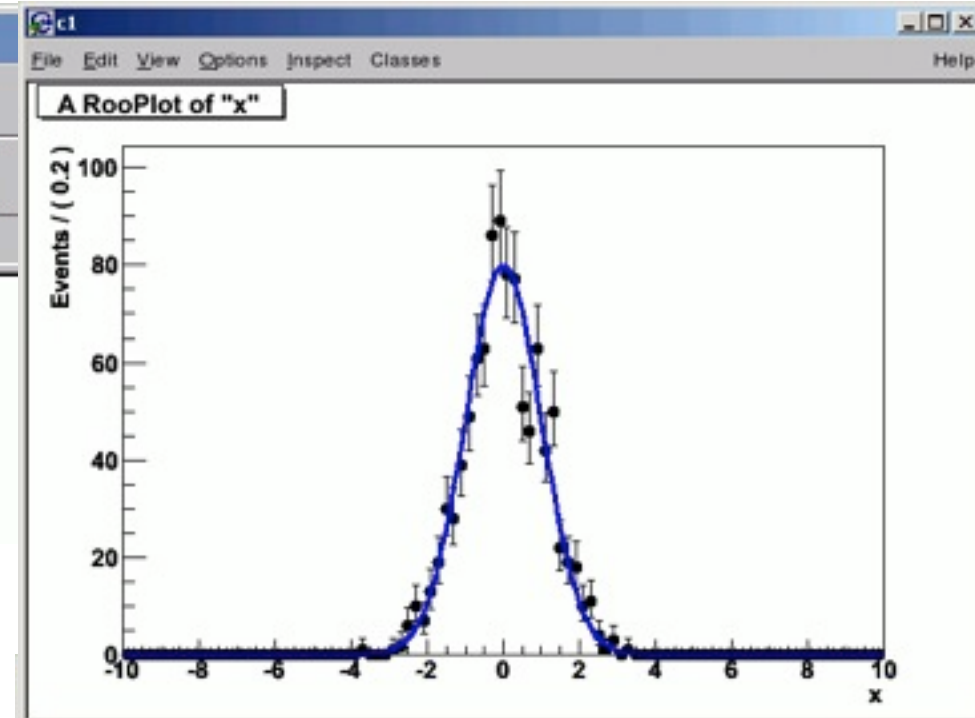
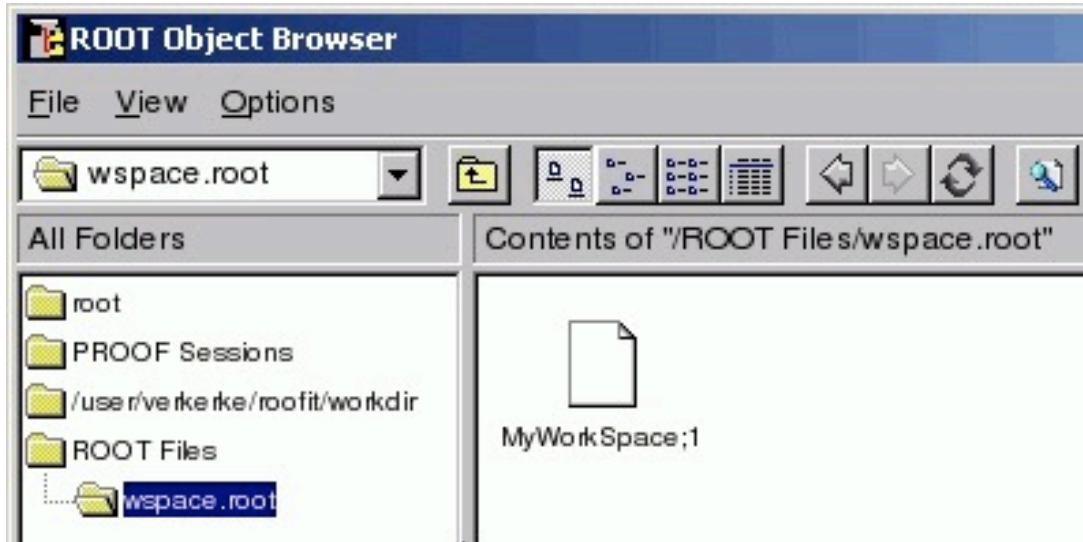
where  $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \dots, \mathcal{D}_{c_{\text{max}}}\}$ ,  $\mathcal{G} = \{a_p\}$  for  $p \in \mathbb{S}$



We will use the following mnemonic index conventions:

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


# 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.

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

Information	Discussion (0)	Files	Linkbacks
 <b>Preprint</b>			
Report number	CERN-OPEN-2012-016		
Title	<b>HistFactory: A tool for creating statistical models for use with RooFit and RooStats</b>		
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	<p>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 primitive 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.</p>		

32 page documentation of HistFactory tool + manual

▶ currently a “living document”

**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

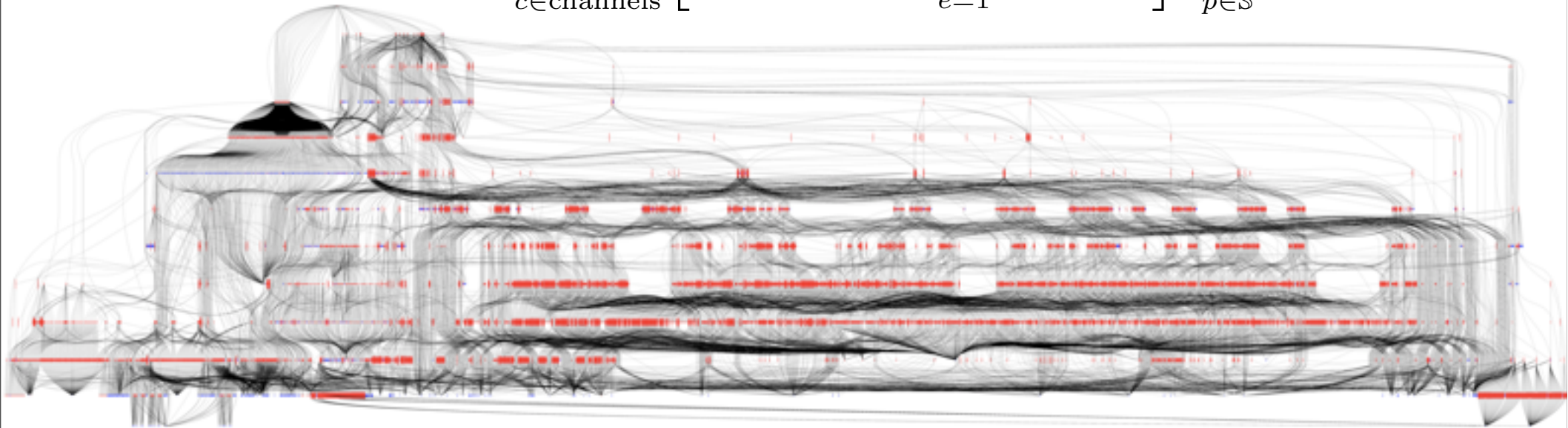
Higgs Decay	Subsequent Decay	Additional Sub-Channels	$m_H$ Range	L [fb <sup>-1</sup> ]
$H \rightarrow \gamma\gamma$	–	9 sub-channels ( $p_{T_i} \otimes \eta_\gamma \otimes$ conversion)	110-150	4.9
$H \rightarrow ZZ$	$lll'l'$	$\{4e, 2e2\mu, 2\mu2e, 4\mu\}$	110-600	4.8
	$ll\nu\nu$	$\{ee, \mu\mu\} \otimes \{\text{low pile-up, high pile-up}\}$	200-280-600	4.7
	$llqq$	$\{b\text{-tagged, untagged}\}$	200-300-600	4.7
$H \rightarrow WW$	$lvlv$	$\{ee, e\mu, \mu\mu\} \otimes \{0\text{-jet, 1-jet, VBF}\}$	110-300-600	4.7
	$lvqq'$	$\{e, \mu\} \otimes \{0\text{-jet, 1-jet}\}$	300-600	4.7
$H \rightarrow \tau^+\tau^-$	$ll4\nu$	$\{e\mu\} \otimes \{0\text{-jet}\} \oplus \{1\text{-jet, VBF, VH}\}$	110-150	4.7
	$l\tau_{\text{had}}3\nu$	$\{e, \mu\} \otimes \{0\text{-jet}\} \otimes \{E_T^{\text{miss}} \geq 20 \text{ GeV}\} \oplus \{e, \mu\} \otimes \{1\text{-jet, VBF}\}$	110-150	4.7
	$\tau_{\text{had}}\tau_{\text{had}}2\nu$	$\{1\text{-jet}\}$	110-150	4.7
$VH \rightarrow b\bar{b}$	$Z \rightarrow \nu\bar{\nu}$	$E_T^{\text{miss}} \in \{120 - 160, 160 - 200, \geq 200 \text{ GeV}\}$	110-130	4.6
	$W \rightarrow l\nu$	$p_T^W \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$	110-130	4.7
	$Z \rightarrow ll$	$p_T^Z \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$	110-130	4.7

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

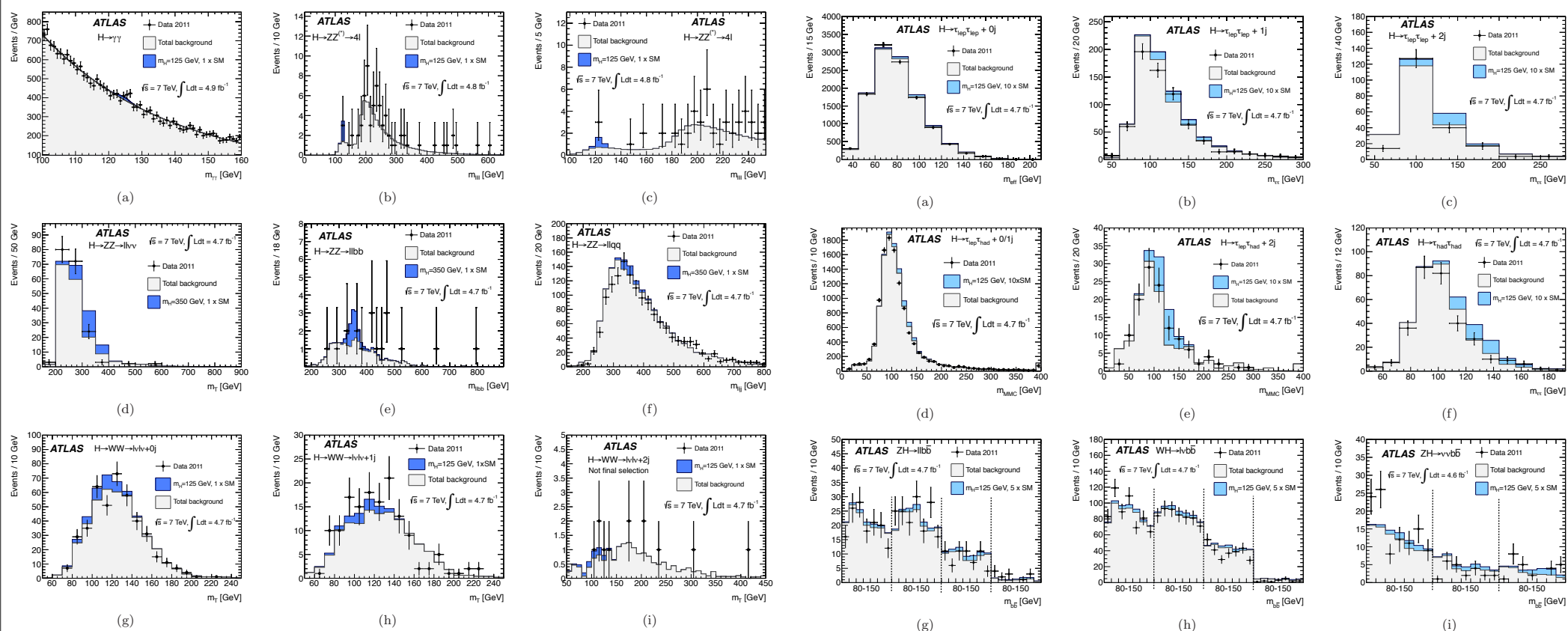
**RooFit / RooStats:** is the modeling language (C++) which provides technologies for collaborative modeling

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

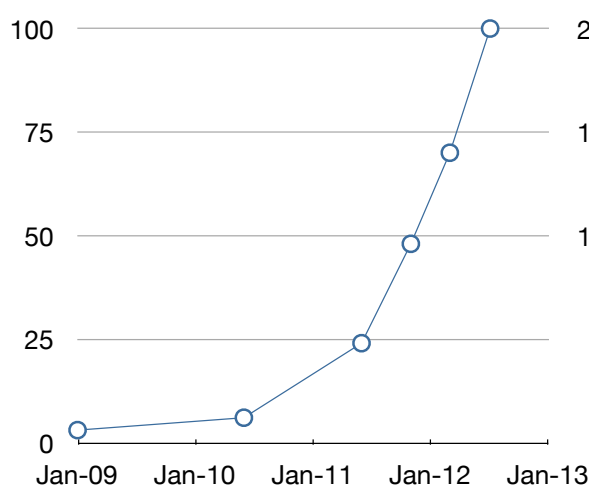
$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p | \alpha_p)$$



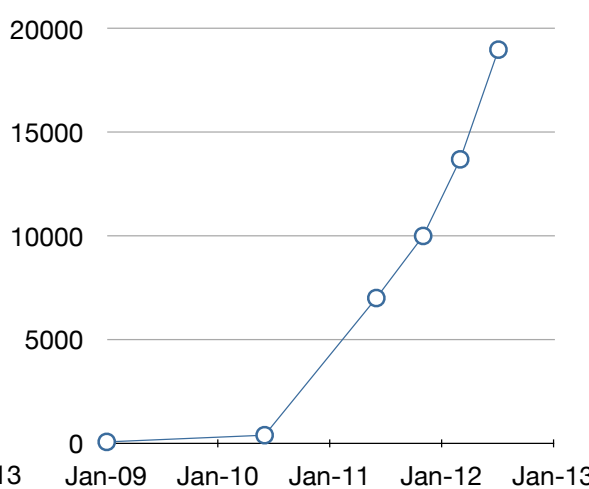
# Evolution of Model Complexity



**Number of Datasets Combined**



**Number of Model Components**



**Number of Parameters in Likelihood**

