Practical Statistics for Particle Physics

Daniel Whiteson, UC Irvine

HCPSS, 2014: Lecture 2

Outline

I. Mathematical preliminaries
II. Fitting
III. Data models
IV. Hypothesis testing
V. Tools and examples

Models

Full MC
Fast MC
Effective models
Data-driven models

Uncertainties

We have a recipe for

f(data|theory)

But is it right?

Uncertainties

We have a recipe for

f(data|theory)

But is it right?

Theory has lots
of nuisance parameters:
cross-sections, LO, NLO...
showering details
hadronization details
detector response

There is some point in NP
space which gives
the most accurate model
but we don't know
where it is!

Systematics

The Good

NP can be constrained in some control region. Uncertainty decreases with luminosity.

eg. Background cross-section

B-tagging efficiency

Jet energy scale

The Ugly

Underlying theoretical approach

eg PYTHIA vs HERWIG

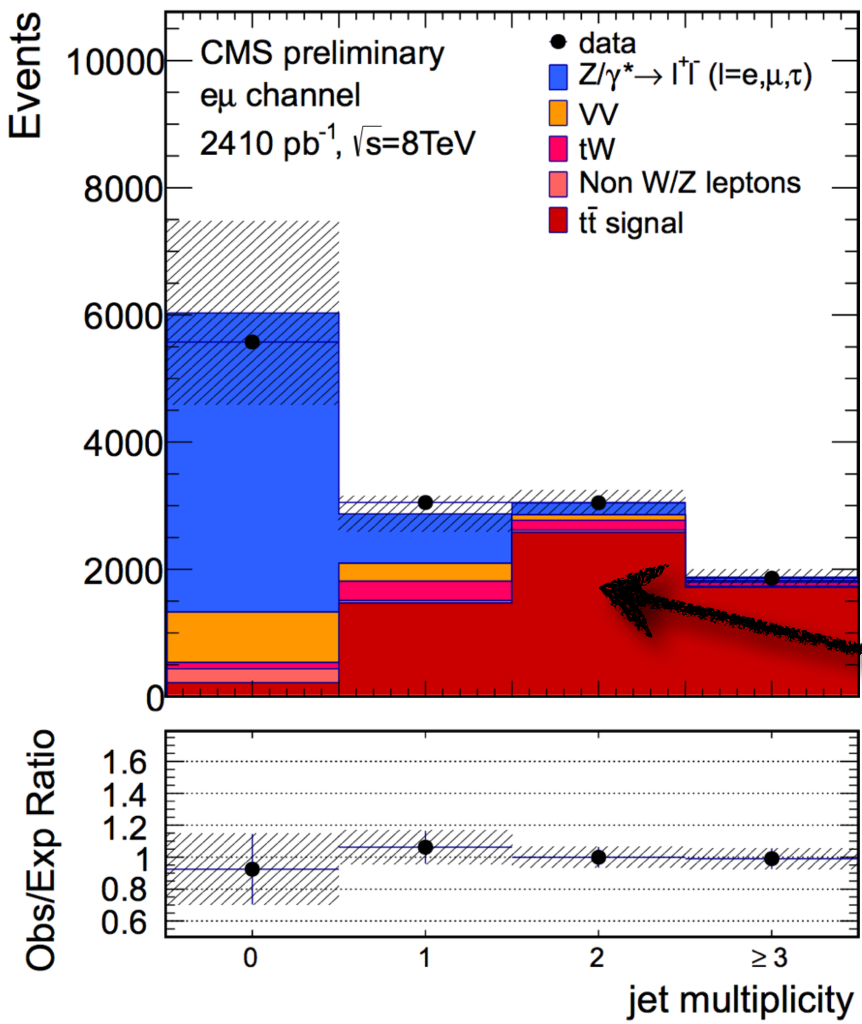
The Bad

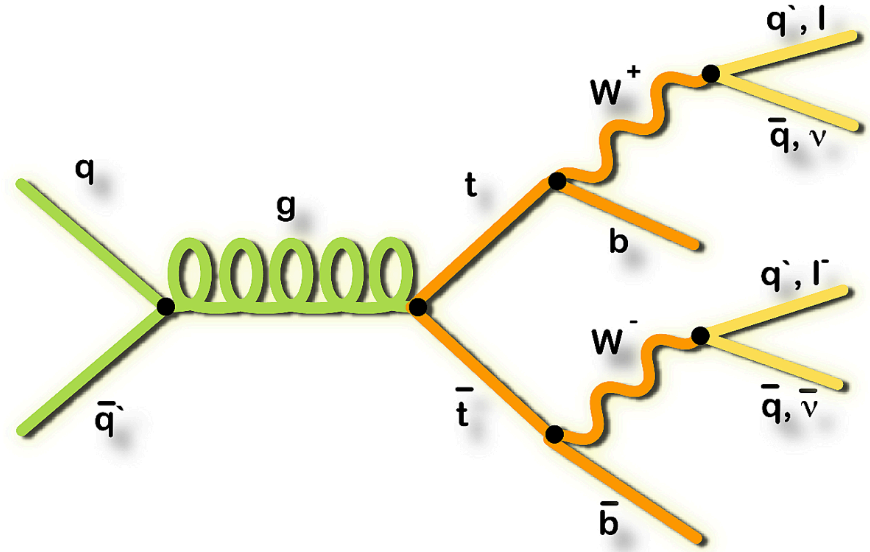
Parameters of underlying heuristic

eg PYTHIA tunes

(Pekka Servino)

Example

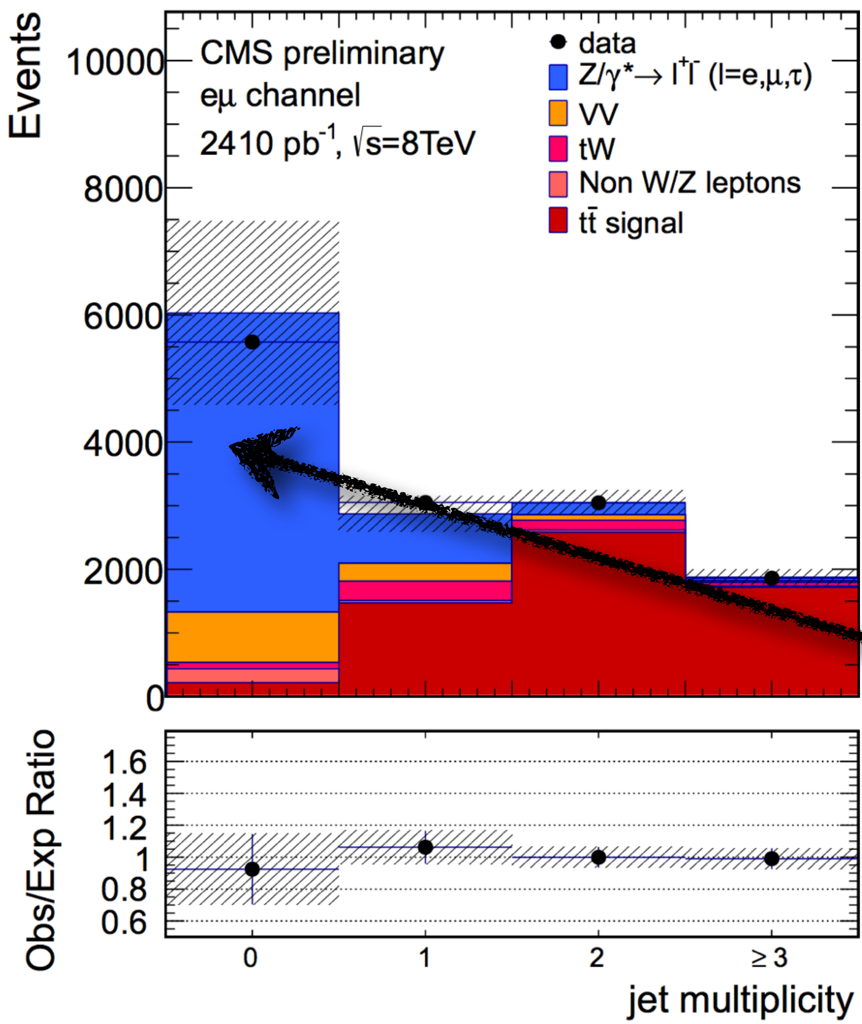


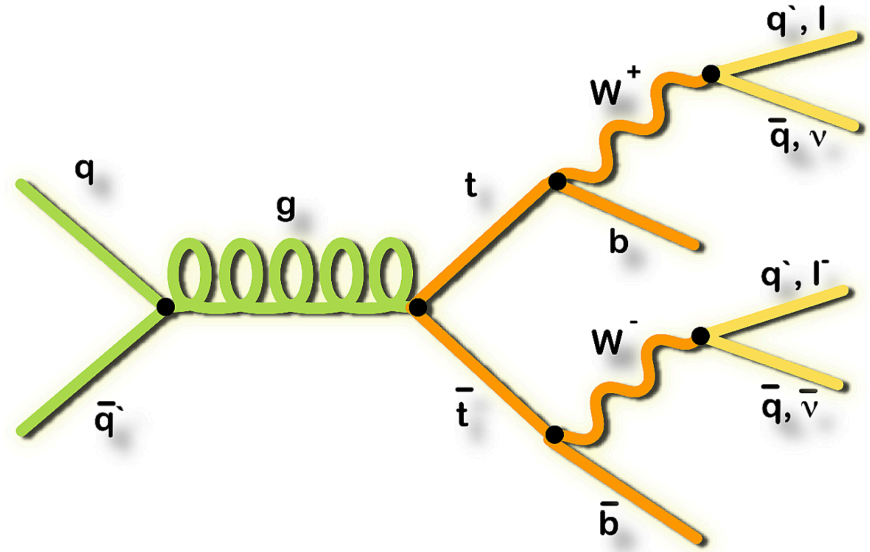


Signal region:

>= 2 jets

Example

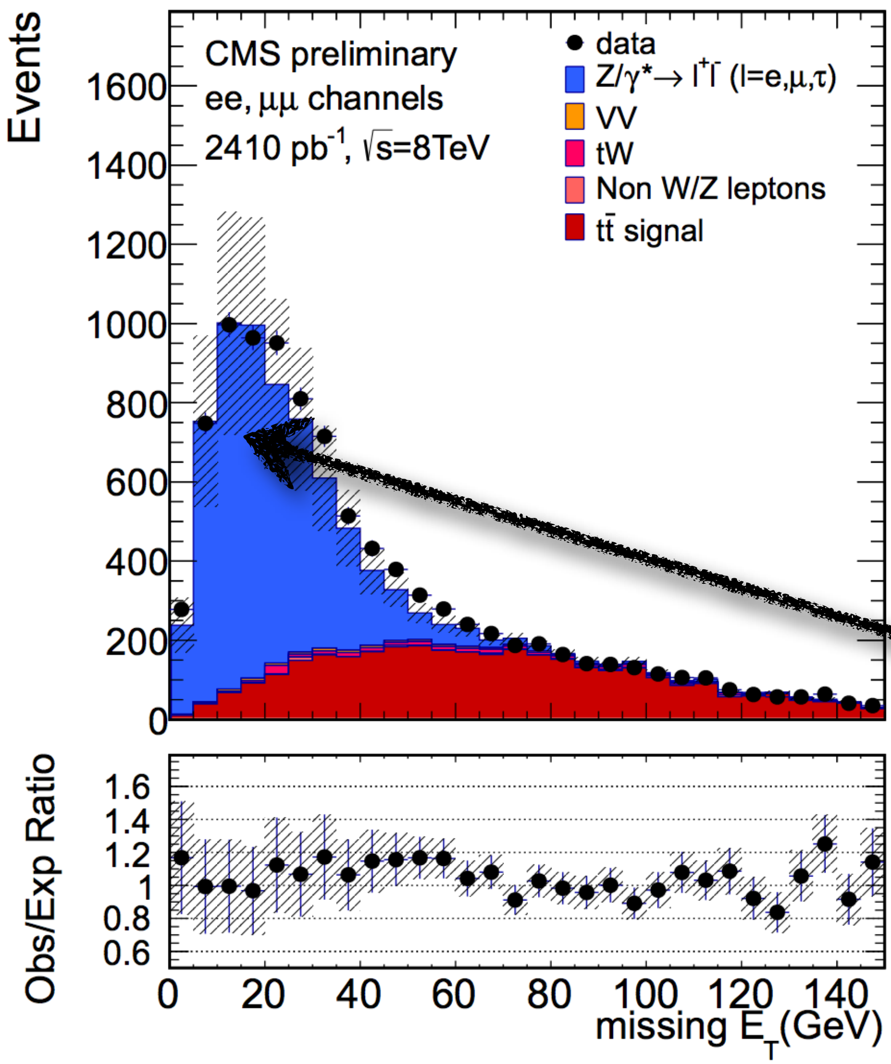


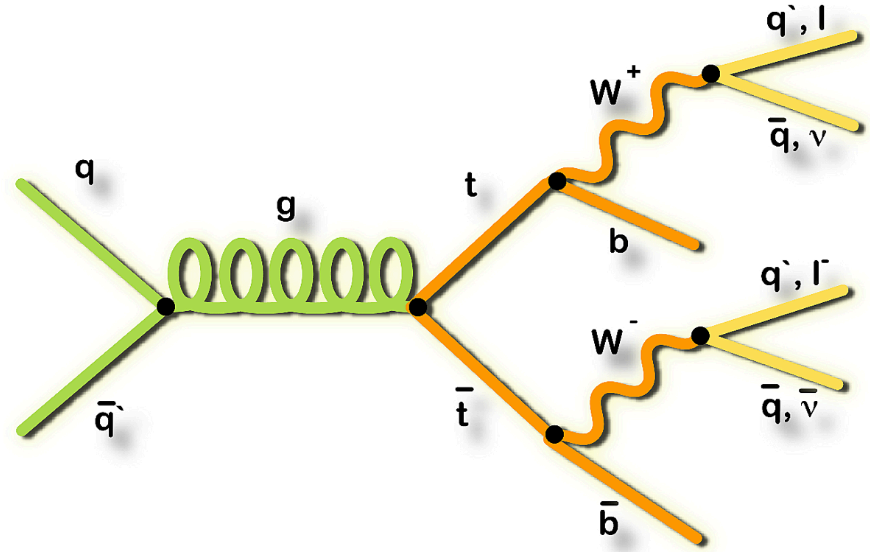


Z control region

== 0 jets

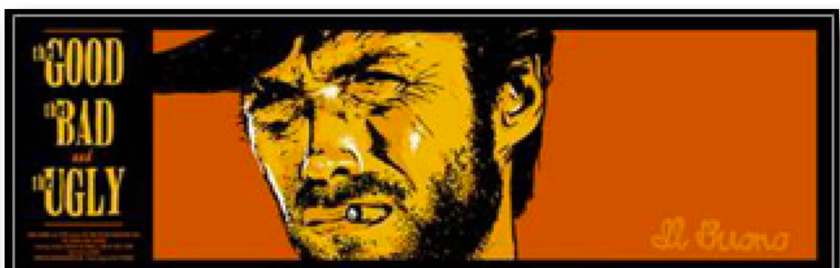
Example



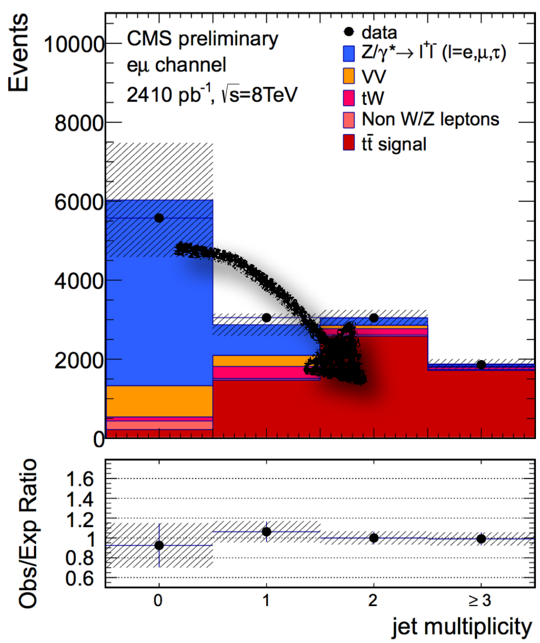


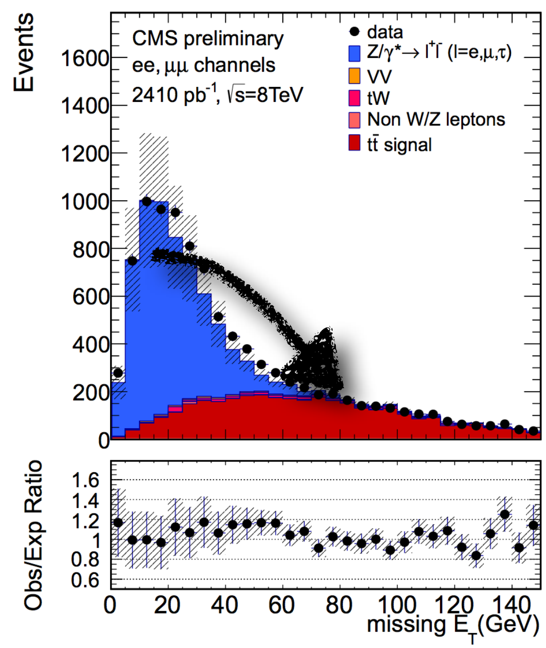
<u>Z control region</u>

Small Missing ET



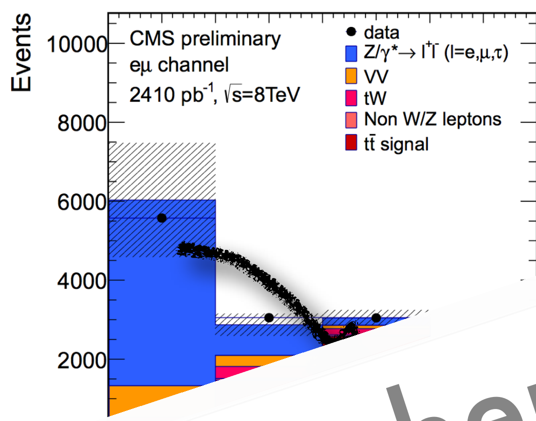
Extrapolation

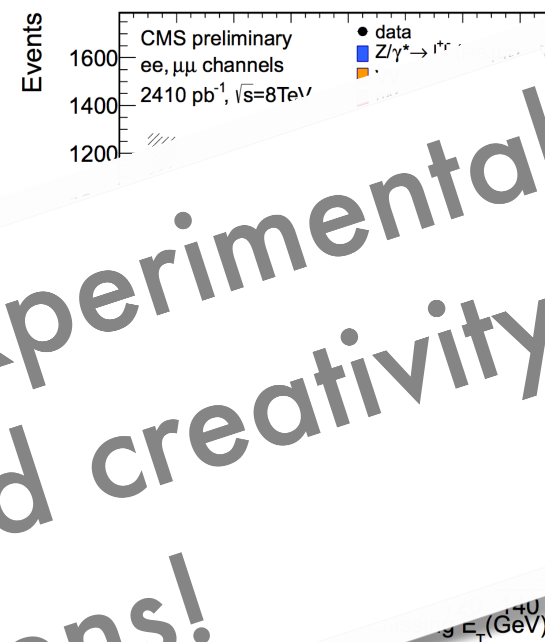




Measure background in one region, extrapolate knowledge to another. How well do we know rates of jet production, tails of MET?

Extrapolation

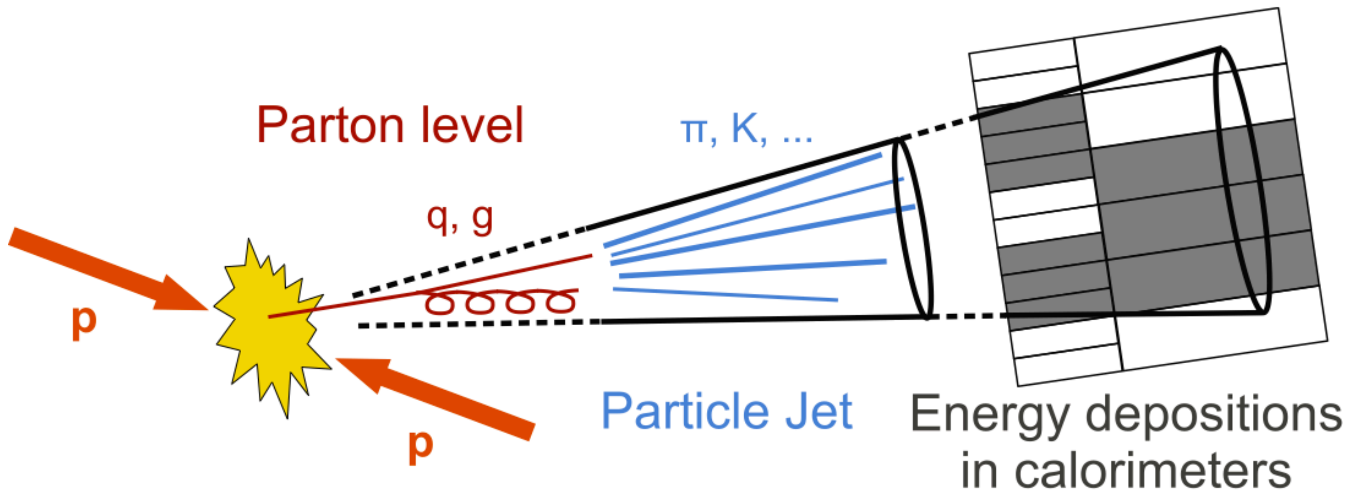


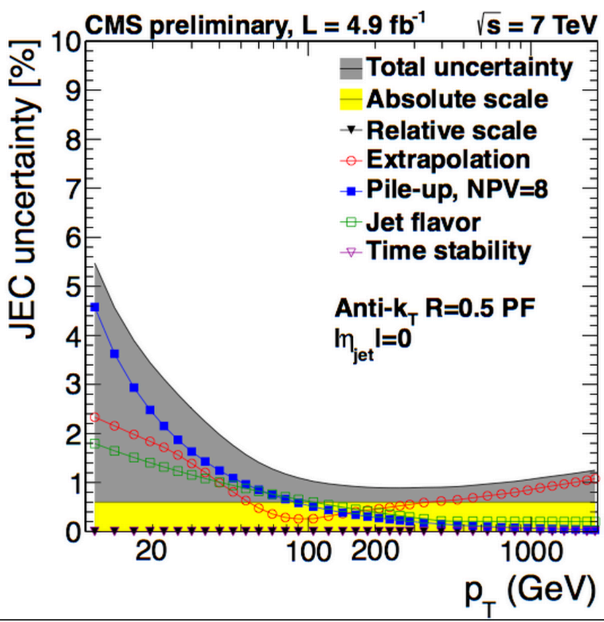


This is where experimental cleverness and creativity happens!

...und in one ...apolate knowledge ...another. How well do we know rates of jet production, tails of MET?

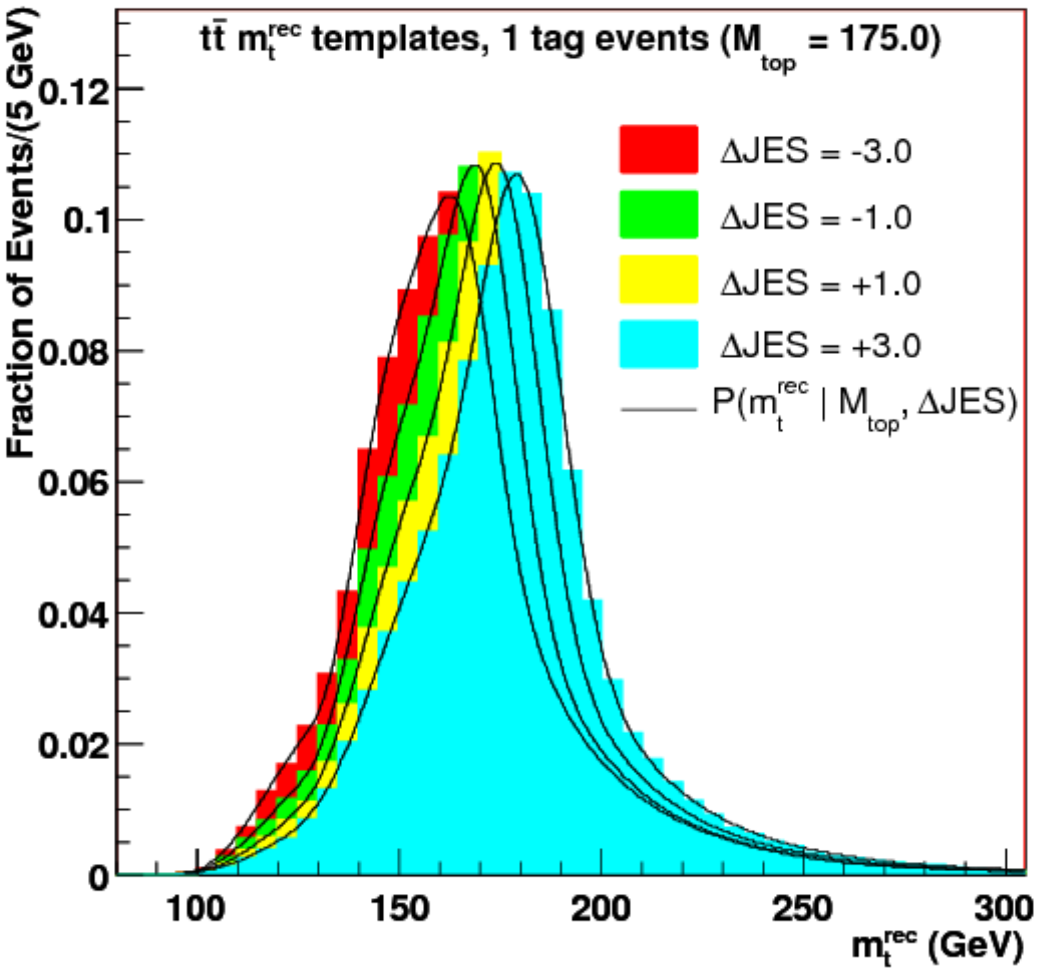
Jet energy scale





Many steps in jet production
Lots of opportunities for mistakes
Calibrate in jj, photon+jet
Extrapolate to your dataset

the shift method

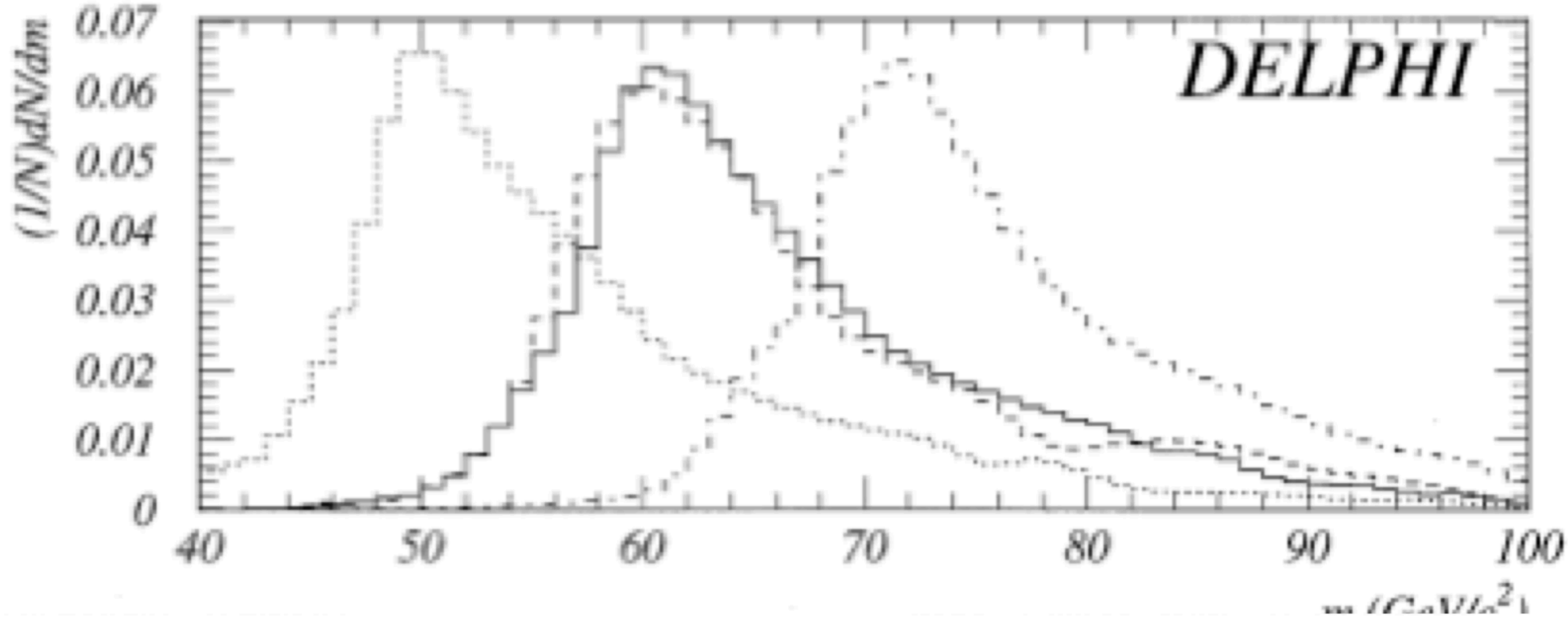


Generating samples at arbitrary values of NP can be expensive!

Often, just generate a few and interpolate.

Histogram interpolation

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

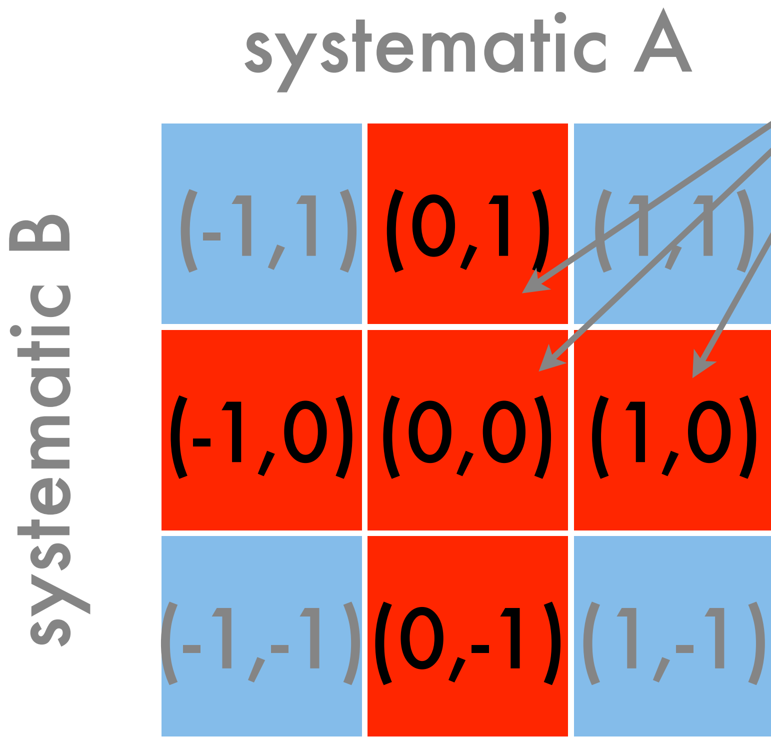


dimensions

systematic A

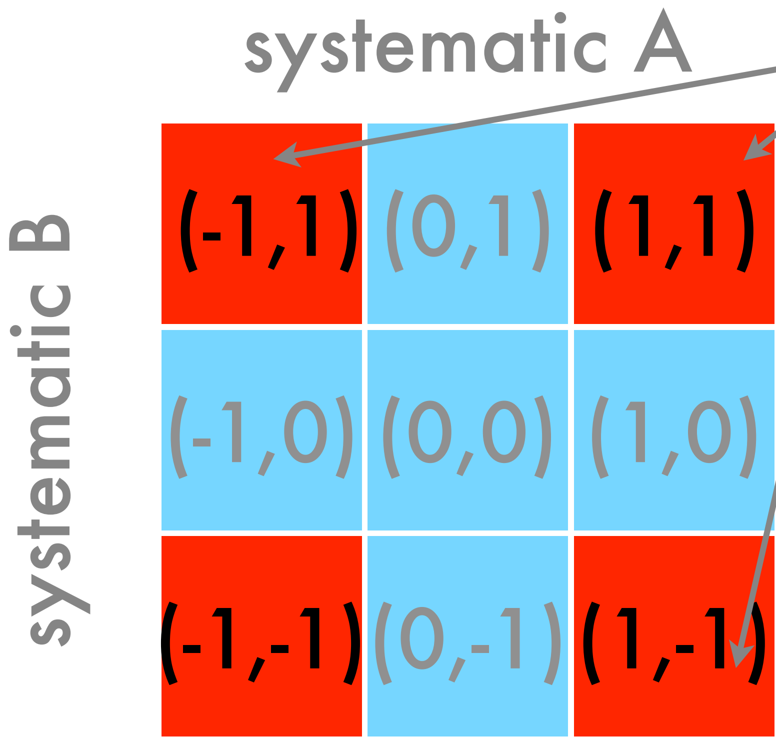
systematic B			
	(-1,1)	(0,1)	(1,1)
	(-1,0)	(0,0)	(1,0)
	(-1,-1)	(0,-1)	(1,-1)

dimensions



We often generate these off-axis points.

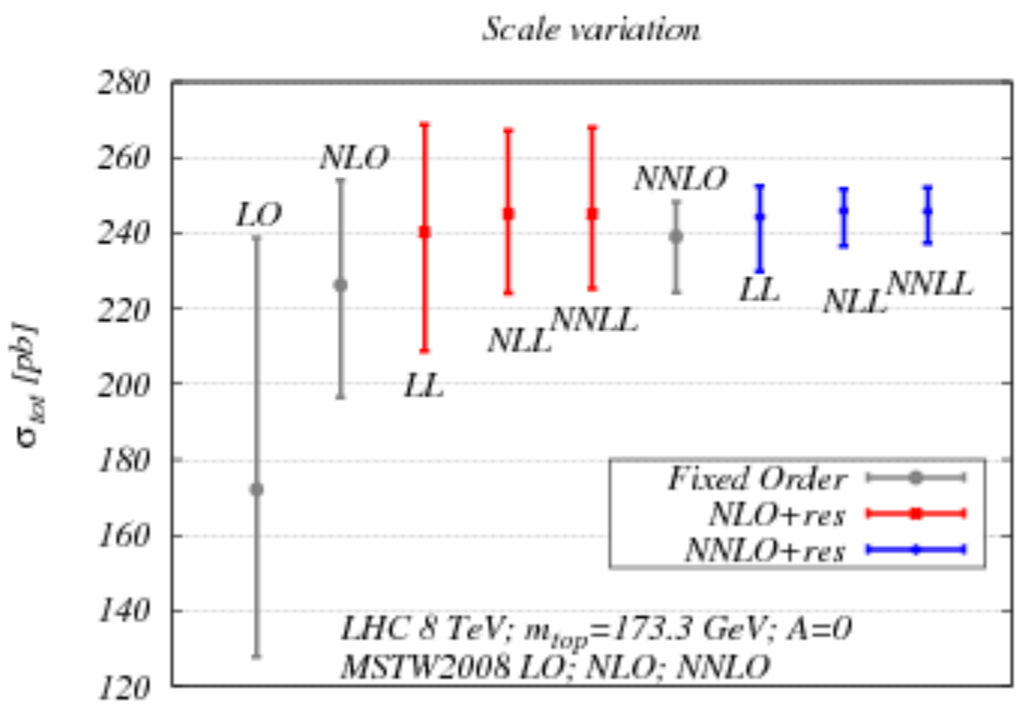
dimensions



We **almost never** generate these off-axis points.

We rely on linear interpolation

Uncertainties



Uncertainty:

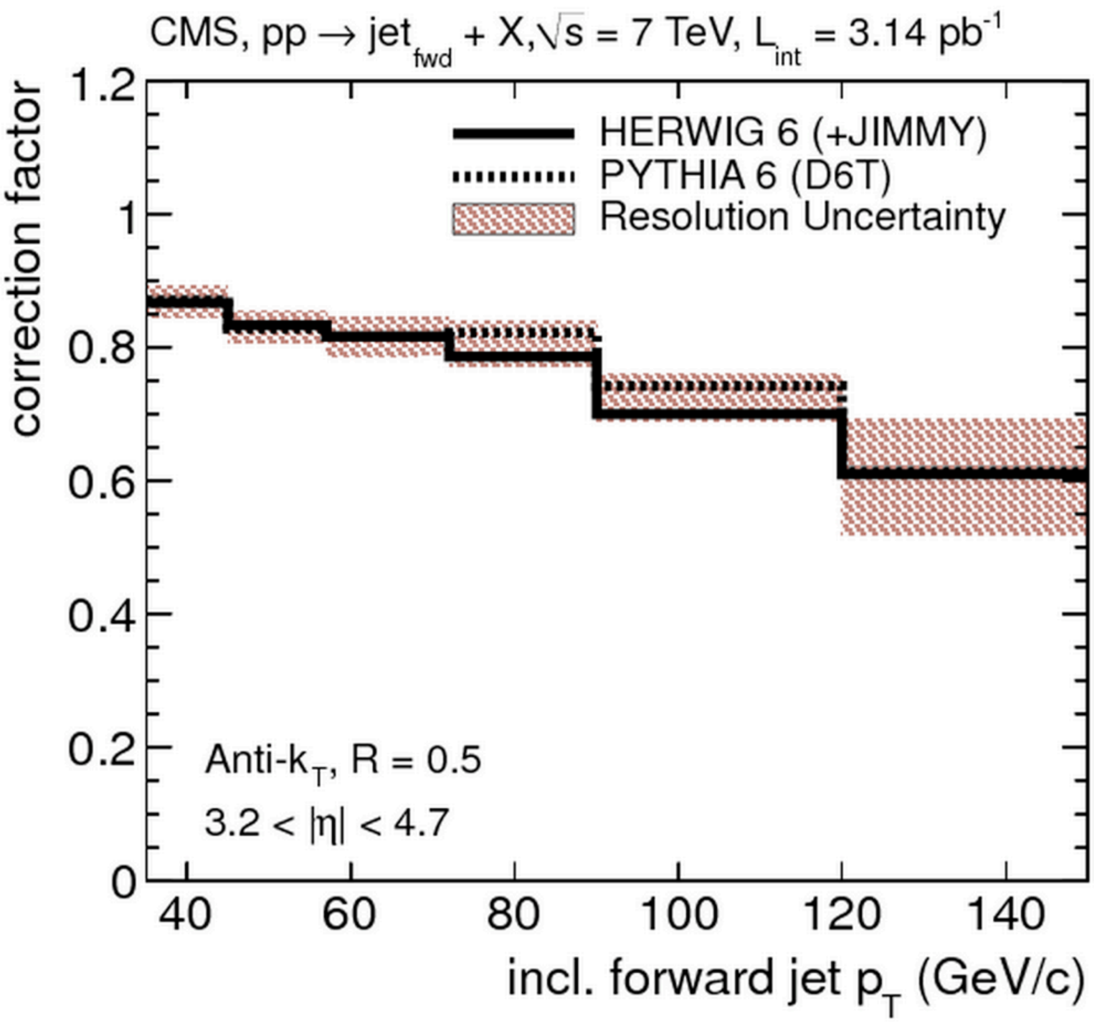
shift renormalization, factorization scales by 2, 1/2 measure change.

Why 2, 1/2?

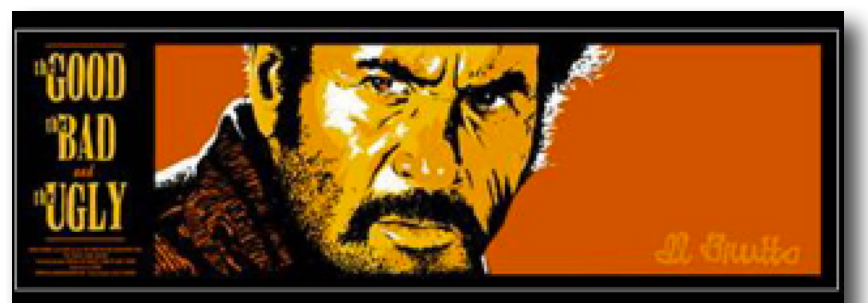
Just convention

Not 1 sigma!

Generators



CMS, pp → jet$_{fwd}$ + X, $\sqrt{s}$ = 7 TeV, L_{int} = 3.14 pb^{-1}
correction factor
HERWIG 6 (+JIMMY)
PYTHIA 6 (D6T)
Resolution Uncertainty
Anti-k_T, R = 0.5
3.2 < |η| < 4.7
0
0.2
0.4
0.6
0.8
1
1.2
40
60
80
100
120
140
incl. forward jet p_T (GeV/c)



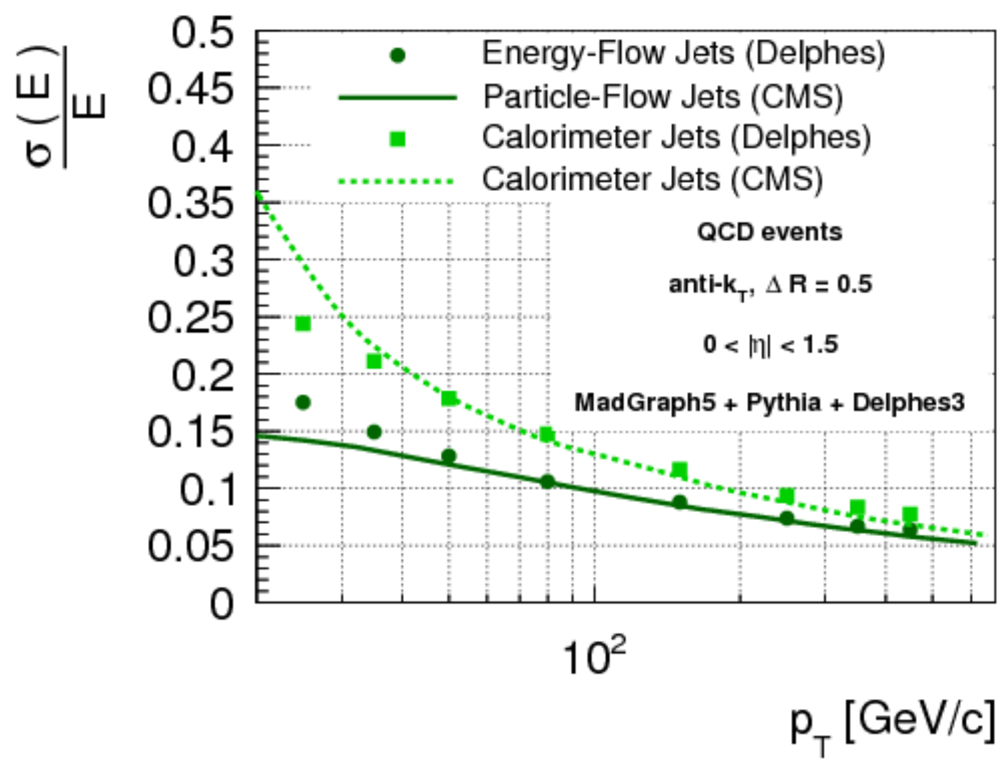
GOOD
BAD
UGLY

fast-MC model

fast MC model

DELPHES
fast simulation

Begin with generated events
but rather than simulating microphysics,
smear particles according to resolution.



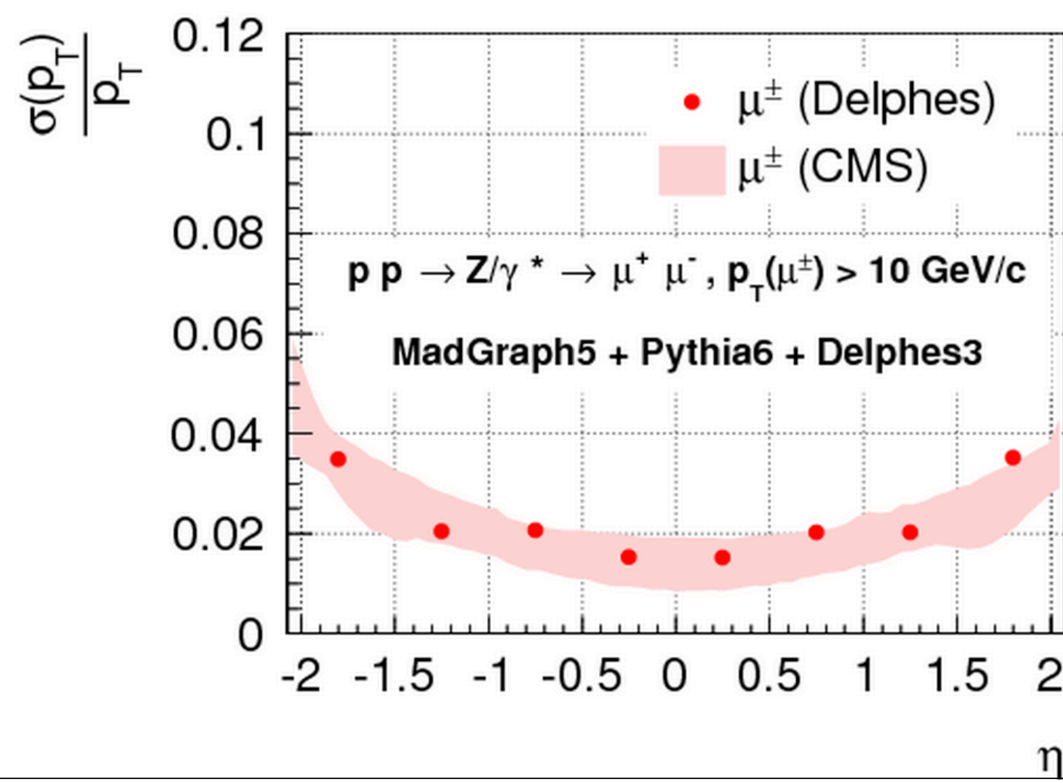
fast MC model

Less accurate, but same issues as full MC:

No analytic PDF

Uncertainties in simulation

DELPHES
fast simulation



The dream

f(data | final-state particles P)

x f(final state particles P | showered particles S)

x f(showered particles S | hard scatter products M)

x f(hard scatter products M | theory)

Sum over all possible intermediate P,S,M

ME approach

If we have a parametrized detector response,
can we parameterize

f(data | final-state particles P)

x f(final state particles P | showered particles S)

x f(showered particles S | hard scatter products M)

x f(hard scatter products M | theory)

ME approach

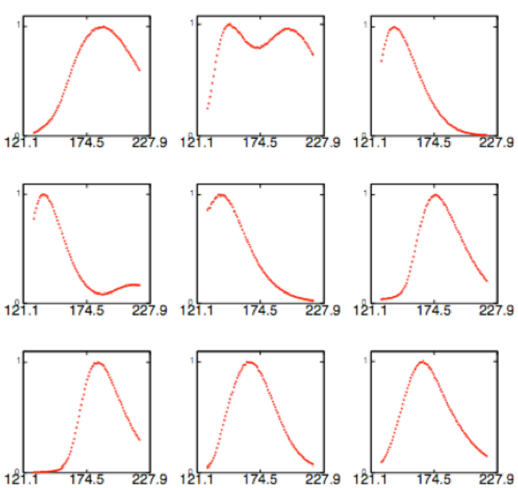
Yes we can!

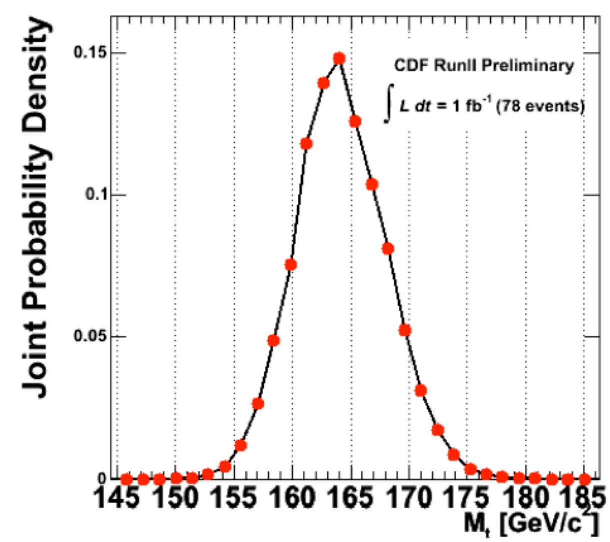
$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi \; |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

Phase-space Integral

Matrix Element

Transfer Functions





ME approach

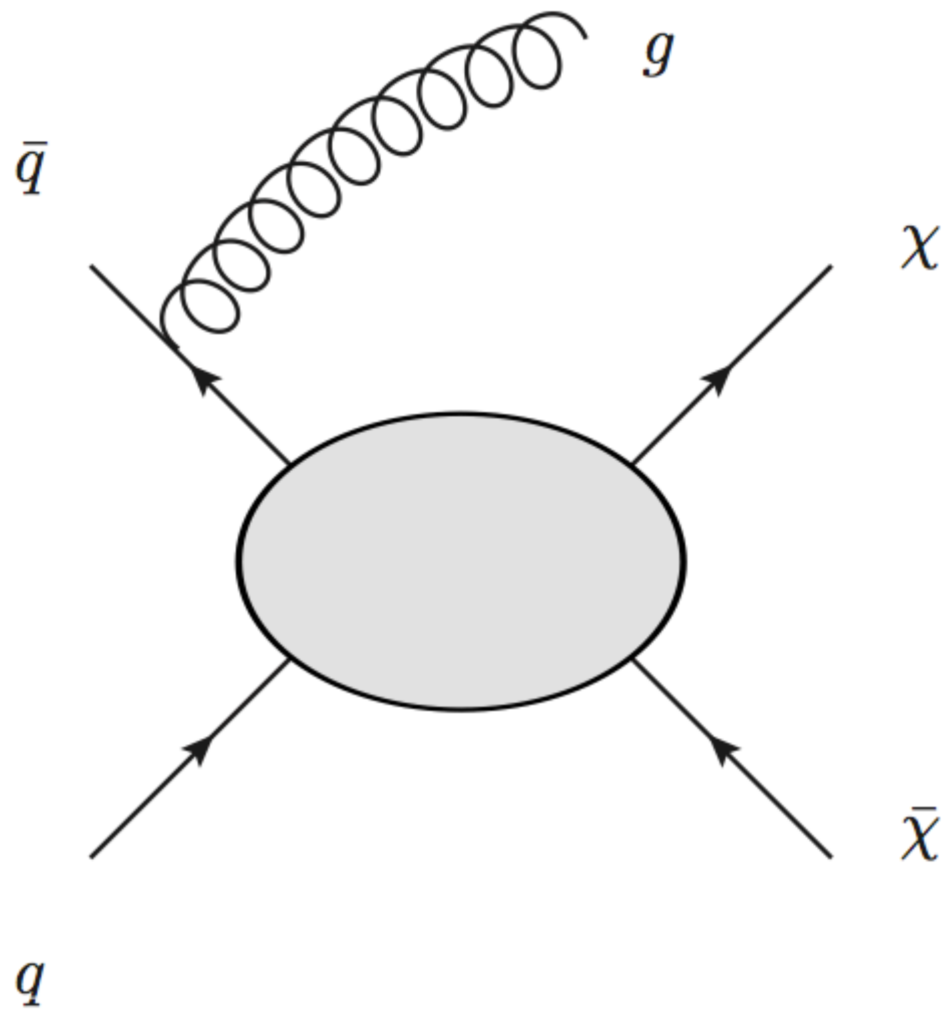
$$P(\mathbf{x}|M_t) = \frac{1}{N}\int \underbrace{d\Phi}_{\text{Phase-space Integral}} \; |\underbrace{\mathcal{M}_{t\bar{t}}}_{\text{Matrix Element}}(p;M_t)|^2 \prod_{jets} \underbrace{f(p_i,j_i)}_{\text{Transfer Functions}} f_{PDF}(q_1) f_{PDF}(q_2)$$

Transfer functions reflect a very complex process

By necessity, approximations, and therefore uncertainties.

Data-driven model

Example: dark matter



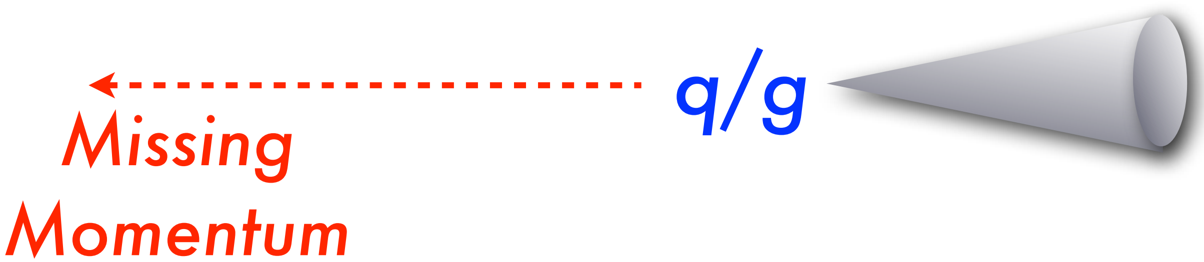
Final state:

Two WIMPs+jet

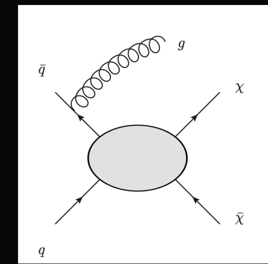
Detector signature

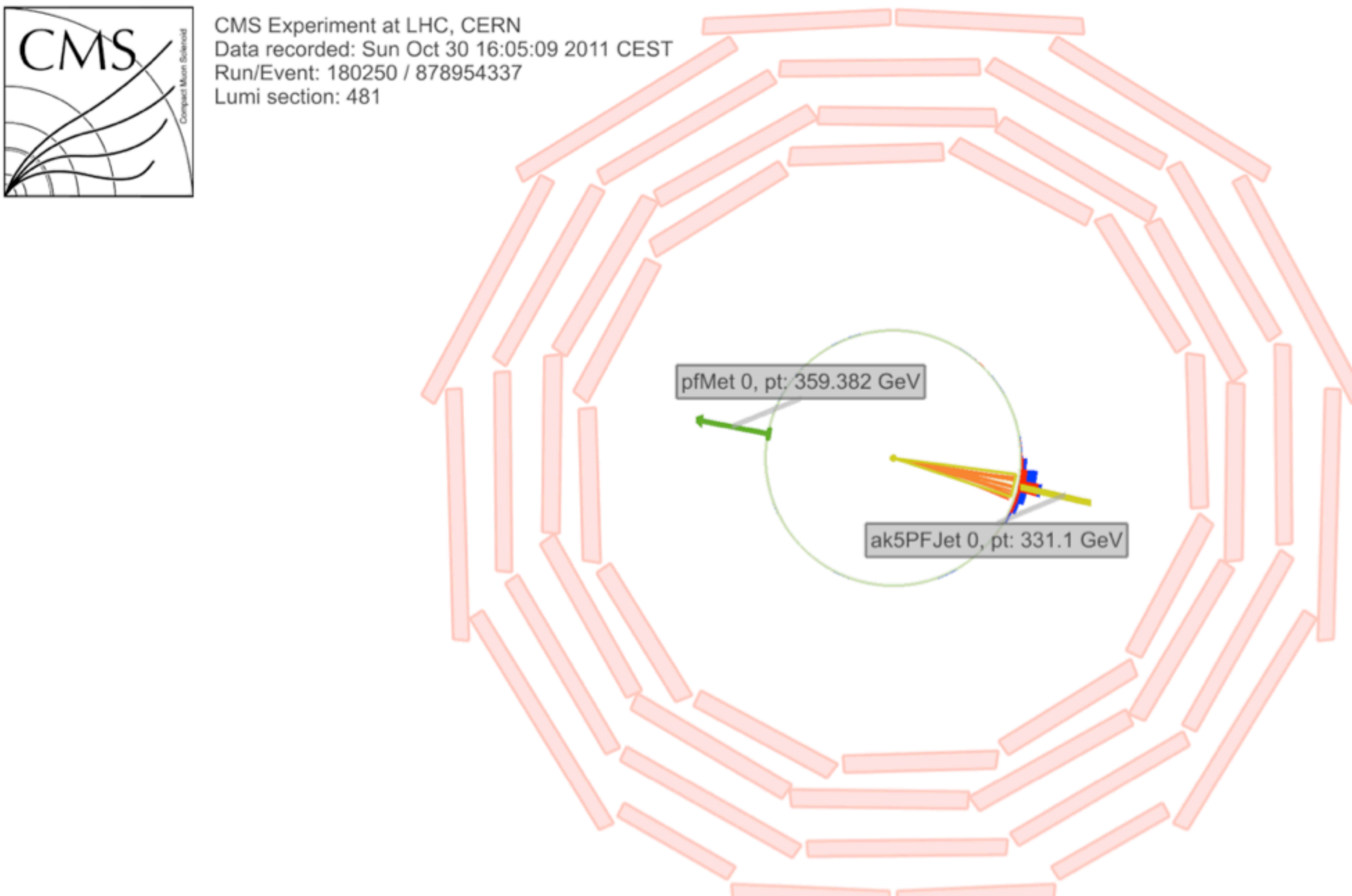
Jet + MET

Mono-jet

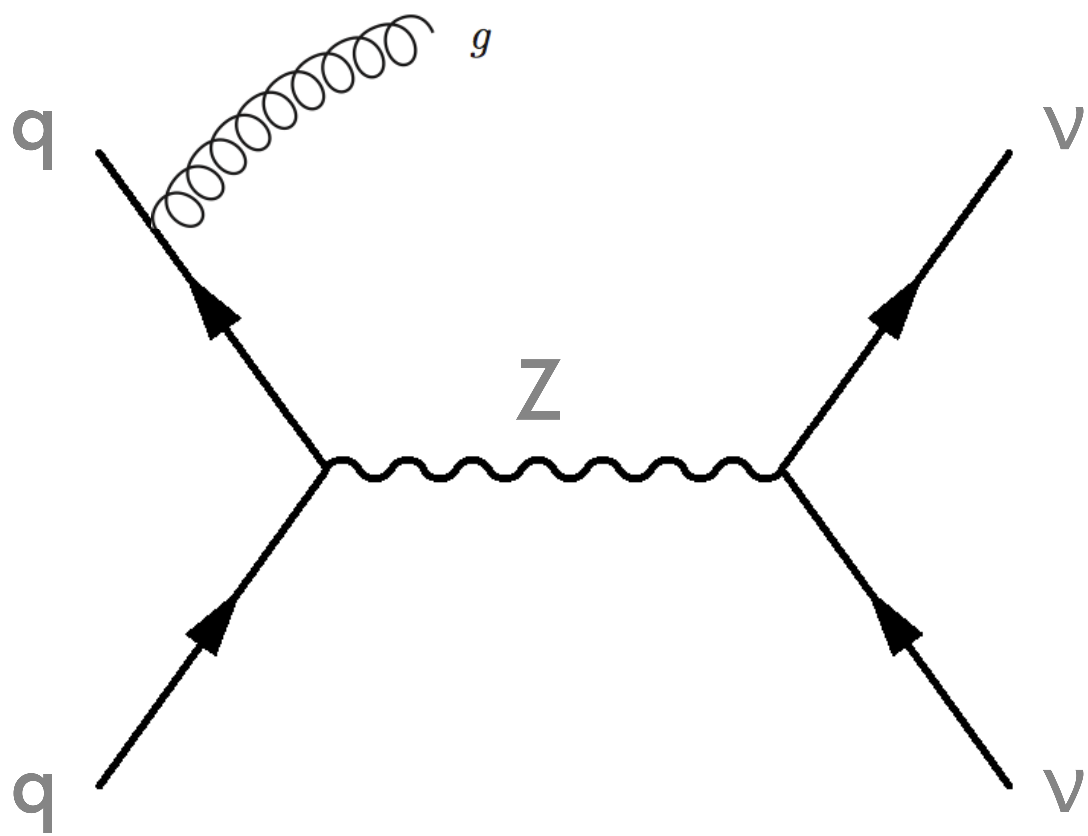


Event display





Backgrounds



Final state:

jet + MET

Process:

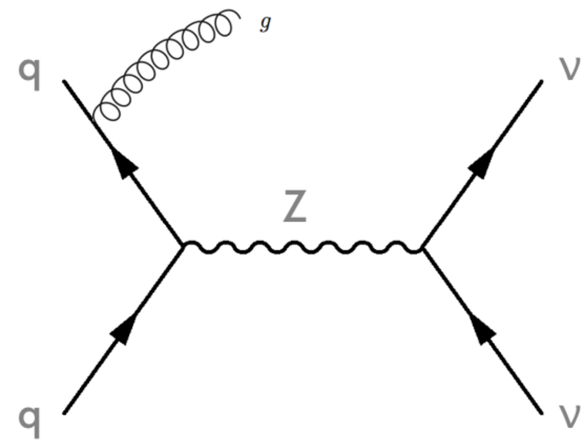
Z → ν ν, with jet

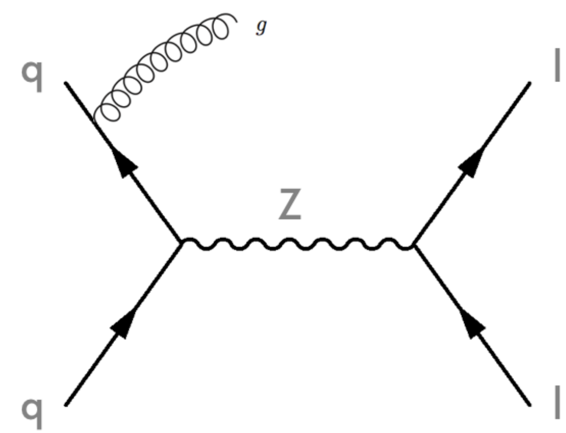
Backgrounds

How to estimate?

Idea: Z → ν ν from Z → ll

Approach:
(1) measure Z to ll + jet
(2) scale by known branching ratios





Details

N[Z(vv)] = N[Z(ll)] x BF[Z(vv)] / BF[Z (ll)]

Details

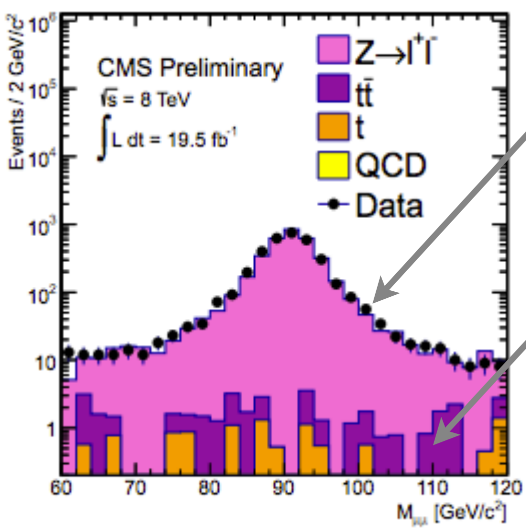
$$N[Z(vv)] = N[Z(ll)] \times BF[Z(vv)] / BF[Z(ll)]$$

$$N[Z(ll)] = N(ll) - N(bg) / \varepsilon$$

Details

N[Z(vv)] = N[Z(ll)] x BF[Z(vv)] / BF[Z (ll)]

N[Z(ll)] = N(ll) - N(bg) / ε

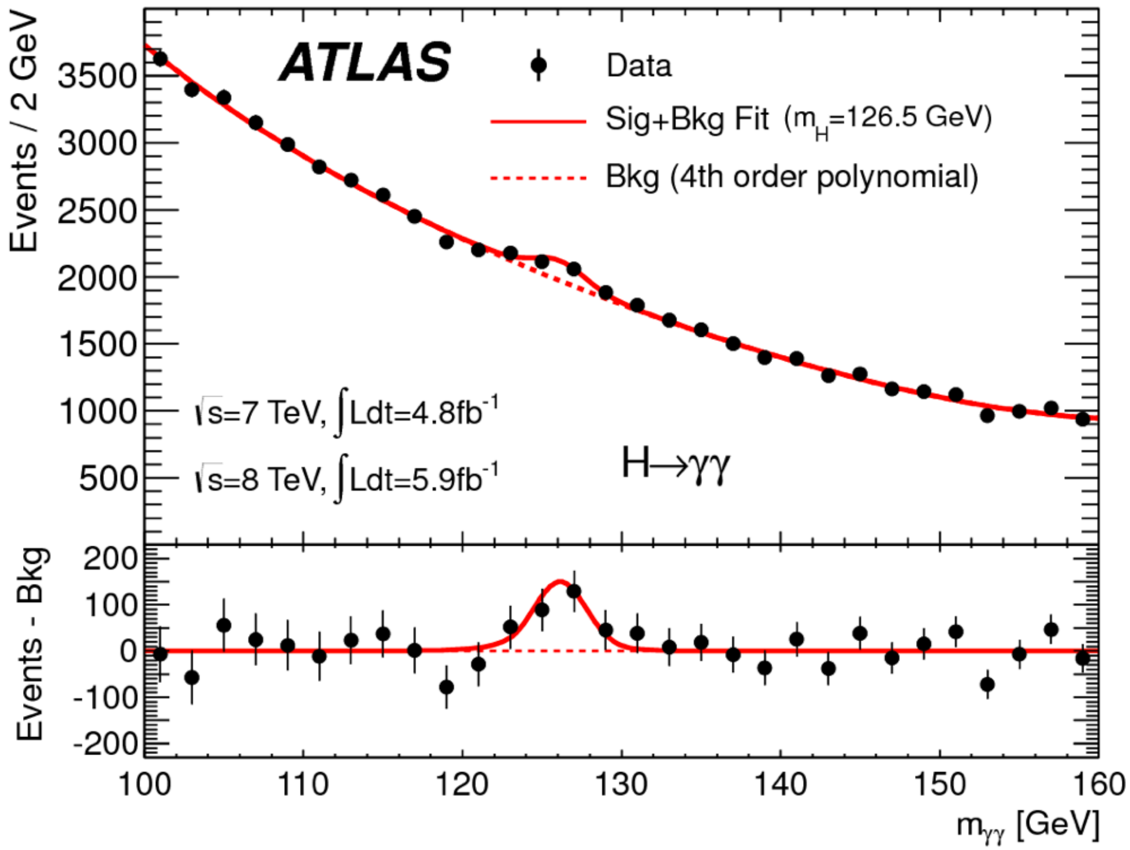


From simulation

CMS PAS EXO-12-048

Effective Model

Effective Model



High-level arguments

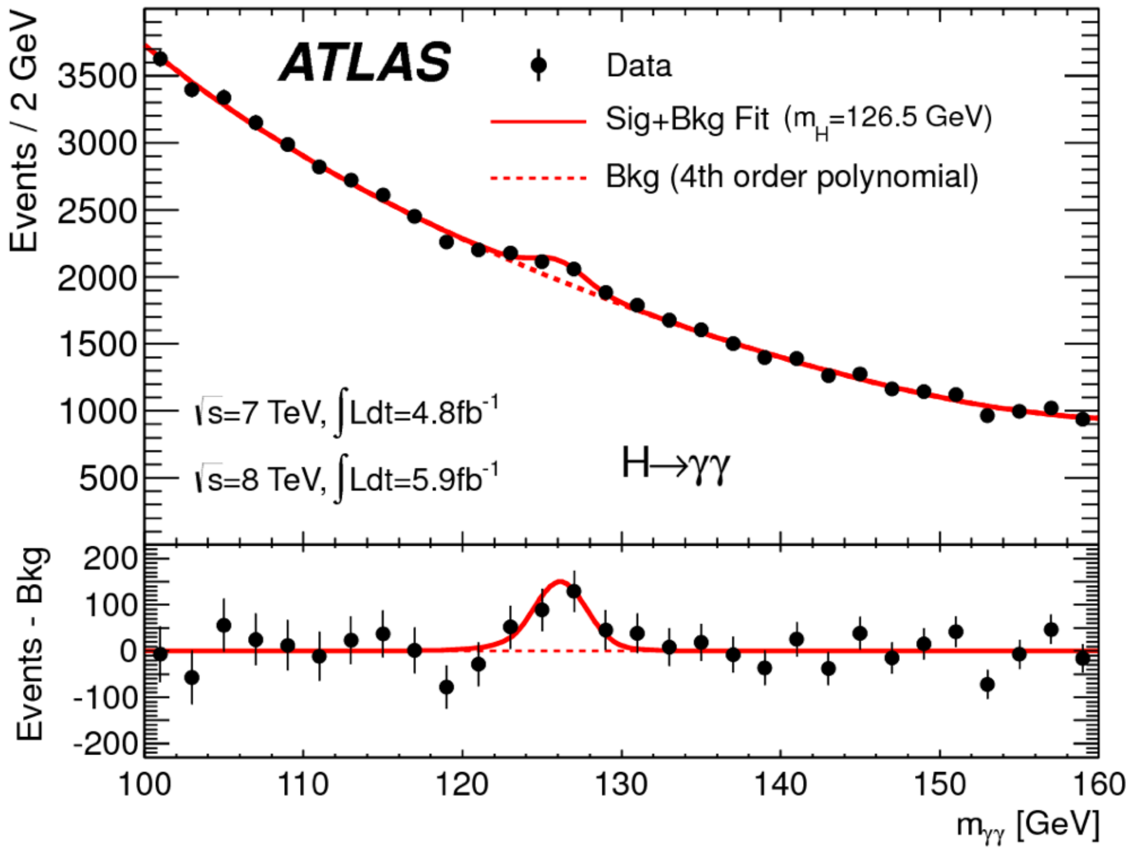
dependence of background

eg “smooth background”

shape of signal

eg “narrow resonance”

Effective Model



Uncertainties

Prediction under peak
depends on quality of side band fits

Background function
evaluate in control regions

Summary of models

<u>MC simulation</u>

- sample of events from on/off simulation
- estimate PDF from events

<u>Fast MC simulation</u>

- simpler generation model
- still estimate PDF from events

<u>Data-driven model</u>

- extrapolate from control regions

<u>Effective model</u>

- parametrized functional form

Summary of models

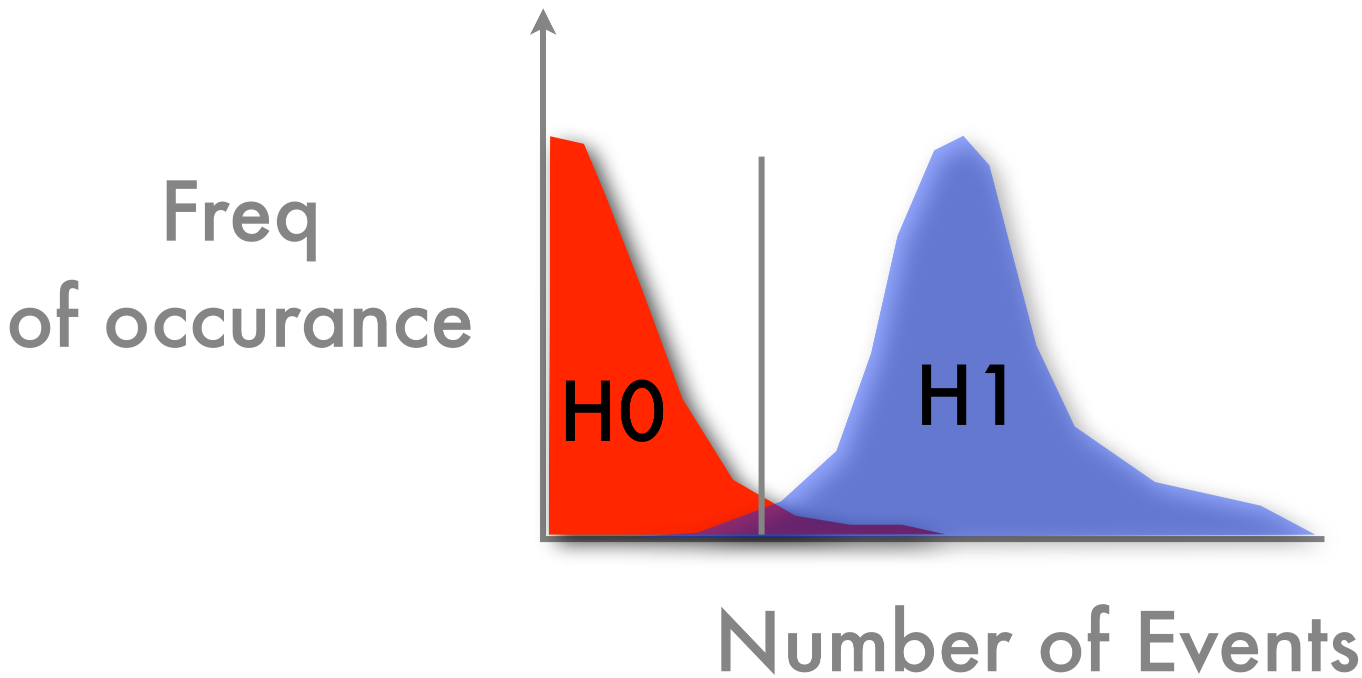
	Pros	Cons
MC Simulation	detailed descr of micro physics	very slow must reconstruct PDF
Fast MC	fast	approximate
Data-driven	Calculations by Nature	Extrapolations from CR have uncertainties
Effective model	fast, physical justification	approximate no details of underlying effects

Hypothesis Testing

Hypothesis Testing

	BSM Particle is real	BSM Particle is not real	
Claim Discovery	True Positive	False Positive Type I error	α
No Claim of Discovery	False Negative Type II error	True Negative	
	β, power=1-β		

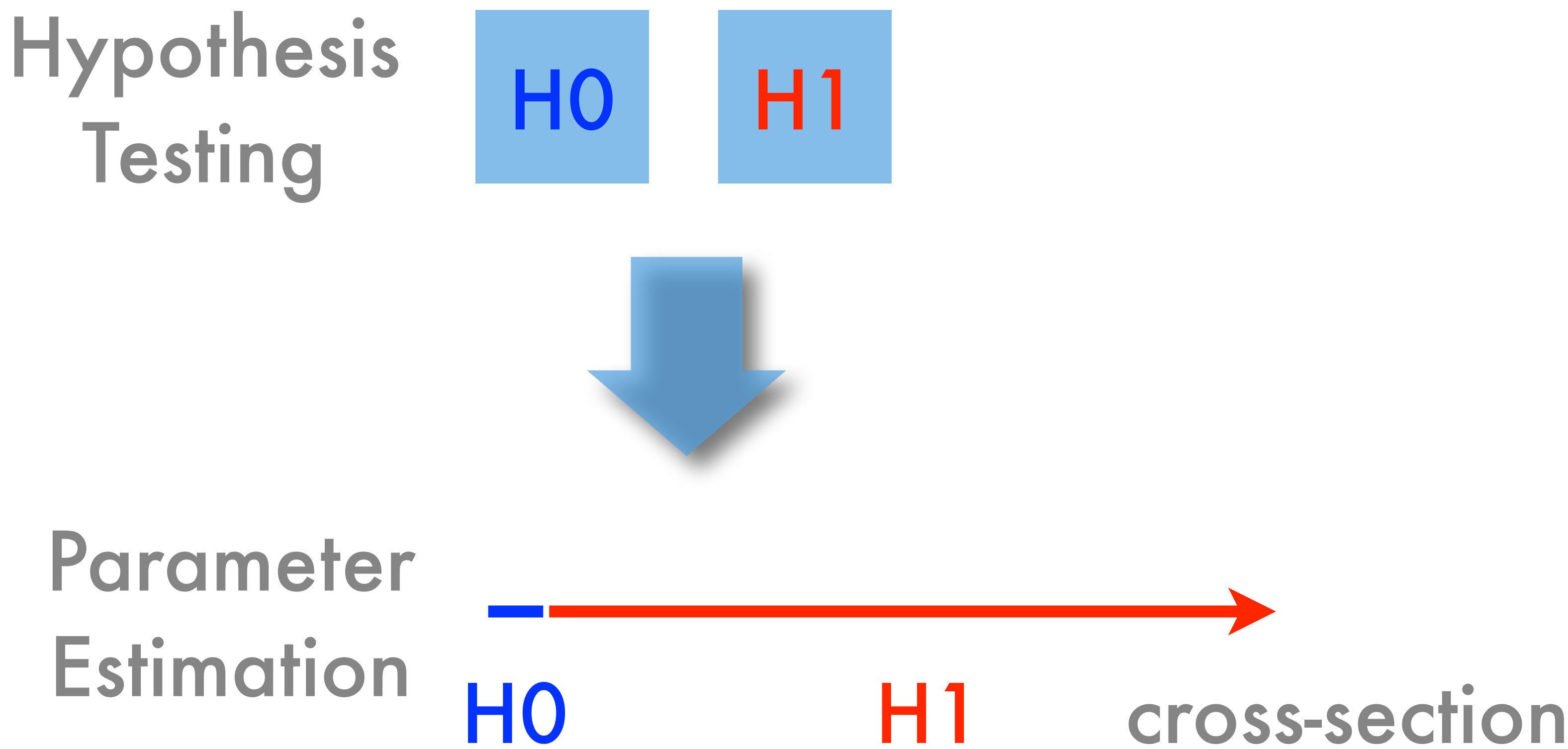
Example



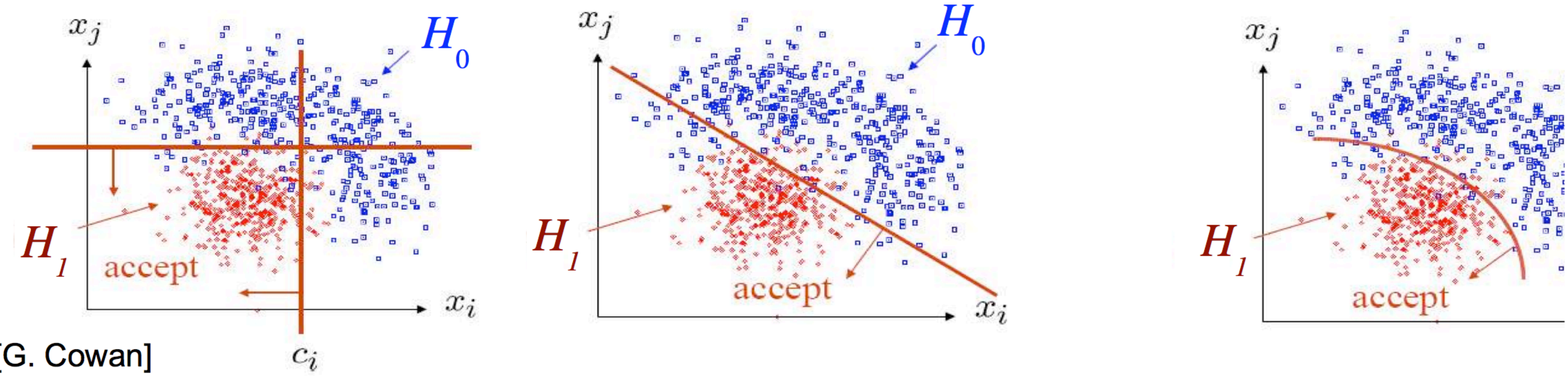
A threshold makes sense.
Choice of position balances
Type I/II errors

Typically:
fix α
minimize β

Generalize


Hypothesis
Testing
H0
H1
Parameter
Estimation
H0
H1
cross-section

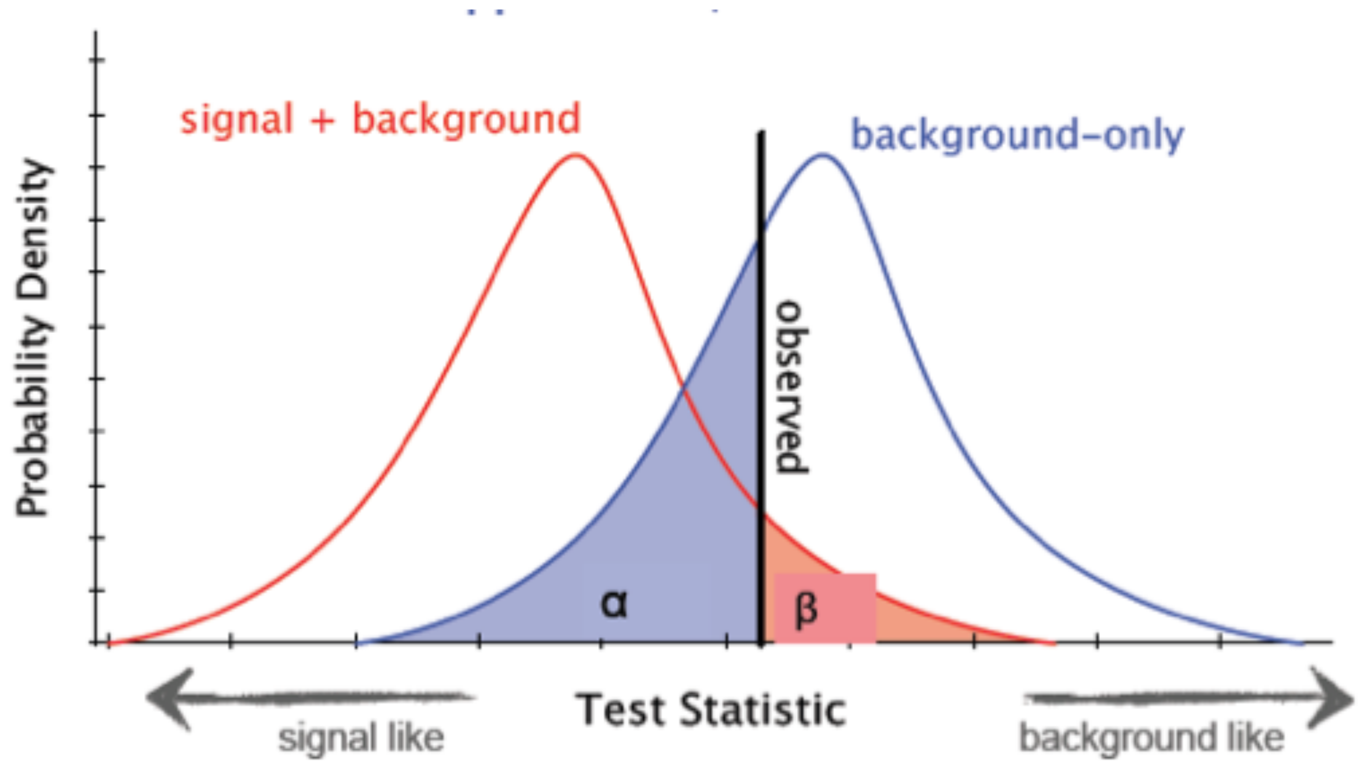
More complicated



[G. Cowan]

Test statistic

Reduce vector of observables to 1 number



How to build distribution of TS? (Usually MC)

How to choose TS?

(K. Cranmer)

Neyman-Pearson

Statement of the problem:

Given some prob that we wrongly reject the Null hypothesis

$$\alpha = P(x \notin W | H_0)$$

Find the region W (where we accept H_0) such that we minimize the prob

$$\beta = P(x \in W | H_1)$$

	BSM Particle is real	BSM Particle is not real
Claim Discovery	True Positive	False Positive Type I error α
No Claim of Discovery	False Negative Type II error β, power=1-β	True Negative

Neyman-Pearson

NP lemma says that the best test statistic is the likelihood ratio:

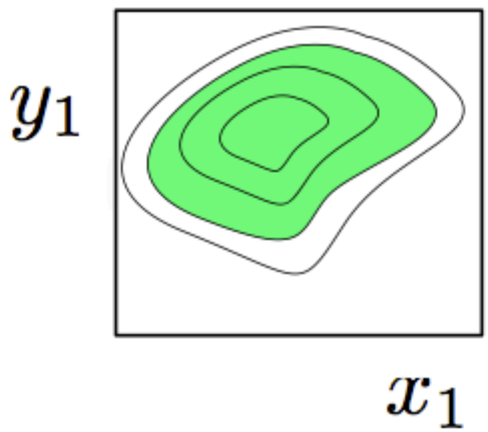
$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

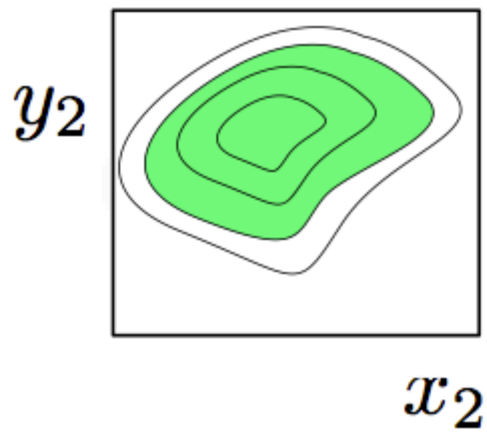
(Gives smallest β for fixed α)

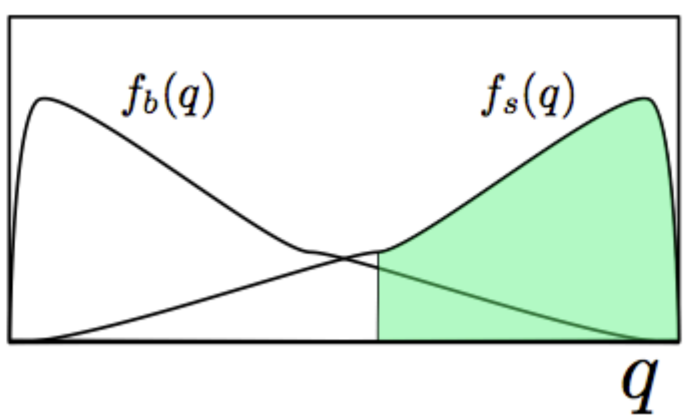
	BSM Particle is real	BSM Particle is not real	
Claim Discovery	True Positive	False Positive Type I error	α
No Claim of Discovery	False Negative Type II error	True Negative	
	β, power=1-β		

What does the TS do?

Finds a region in variable space



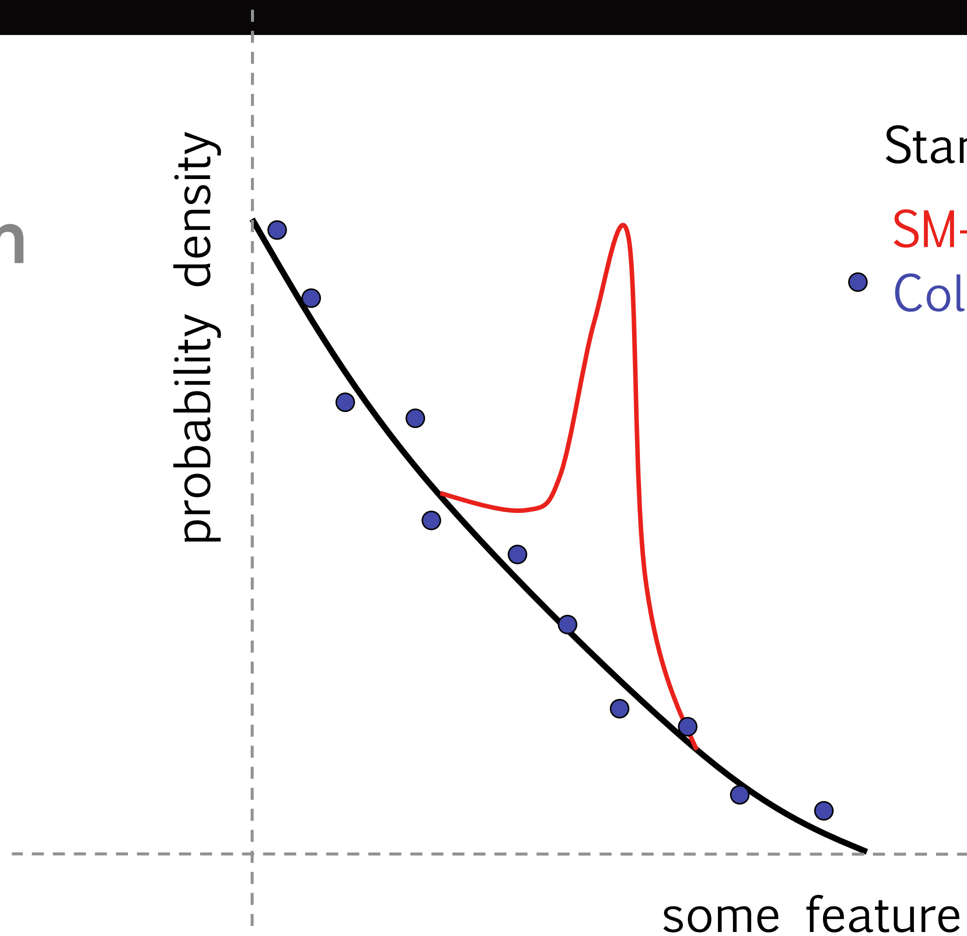




(K. Cranmer)

How to find NP

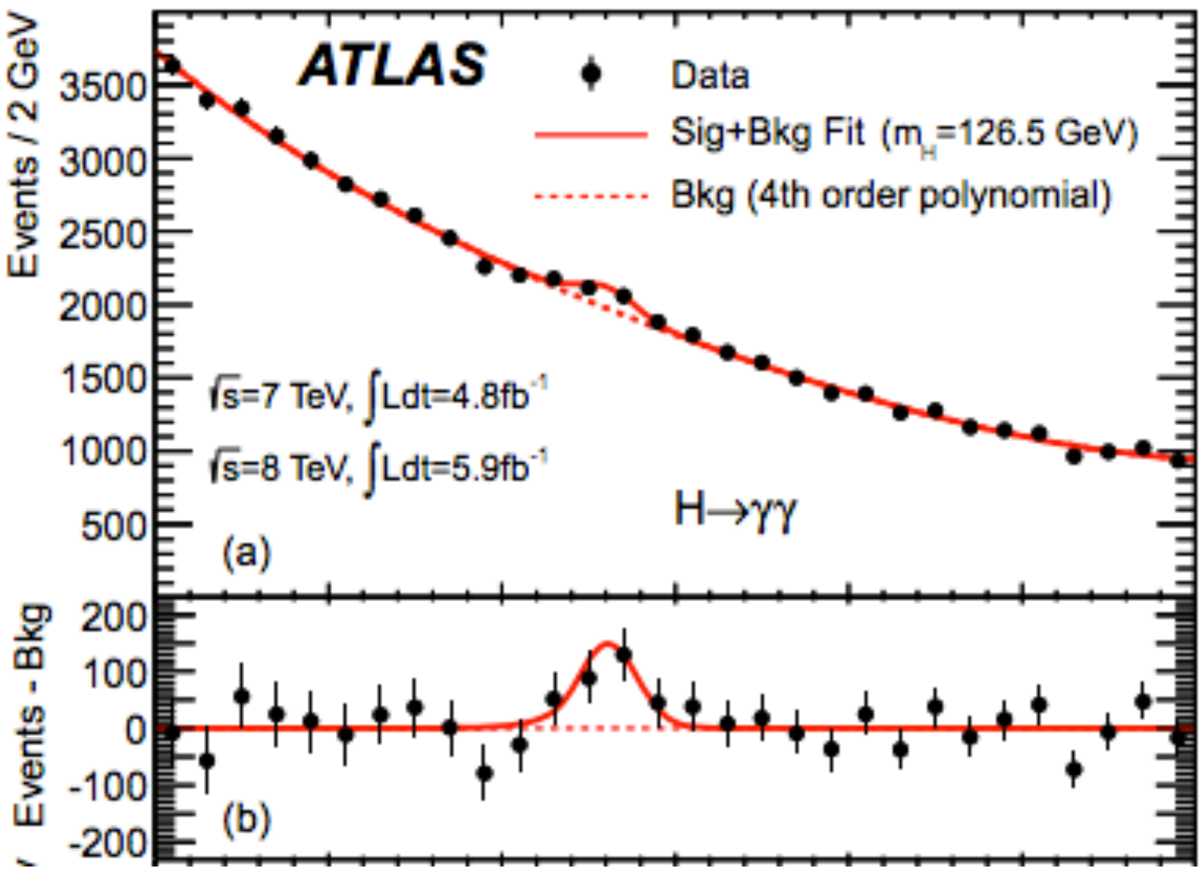
Isolate some feature in which two theories SM, SM+X can be best distinguished.



The data can tell us which hypothesis is preferred via a likelihood ratio:

$$\frac{L_{SM+X}}{L_{SM}} \qquad \frac{P(\text{data} \mid SM+X)}{P(\text{data} \mid SM)}$$

e.g.



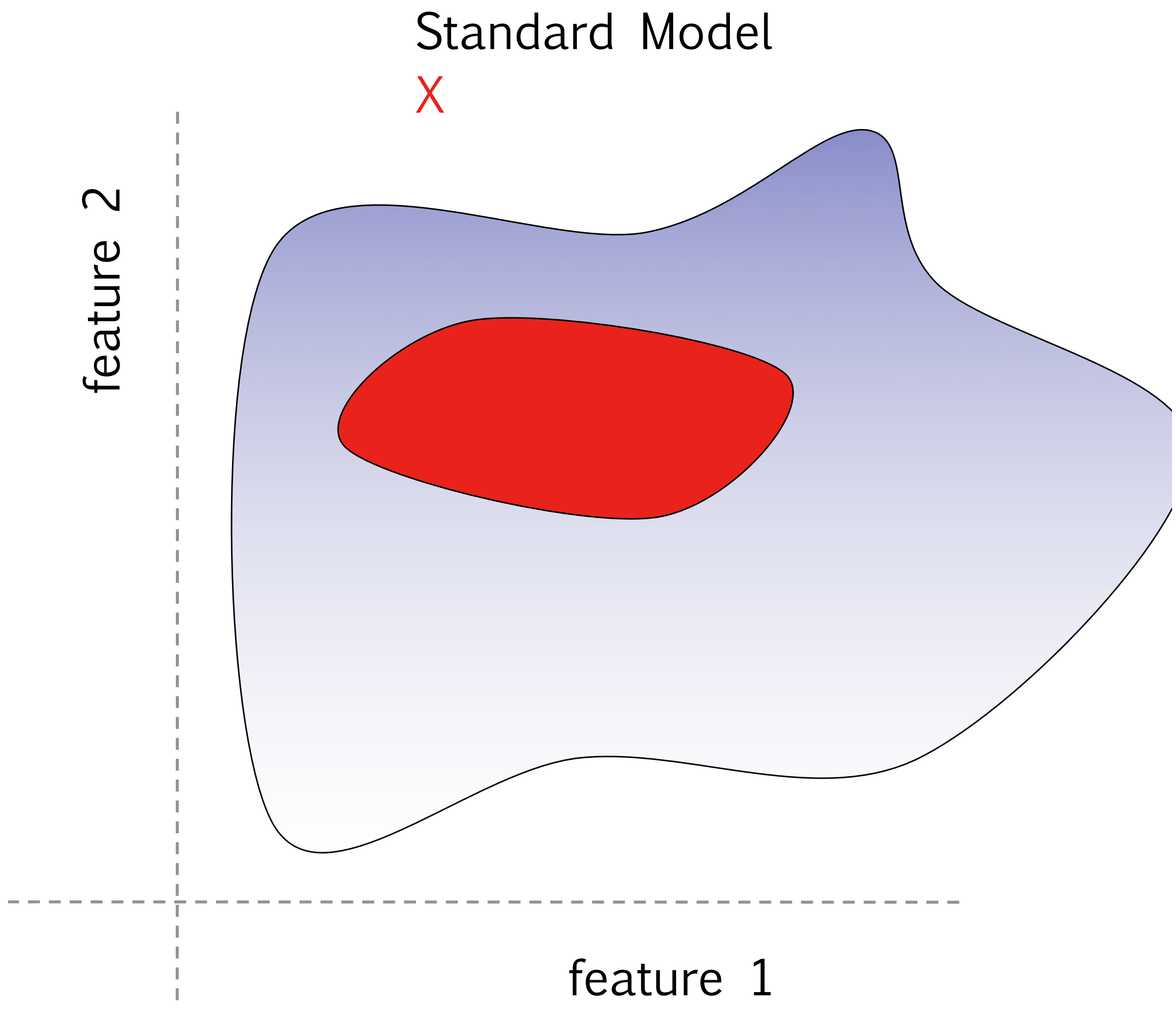
ATLAS
Data
Sig+Bkg Fit (m_H=126.5 GeV)
Bkg (4th order polynomial)
$\sqrt{s}$=7 TeV, $\int Ldt=4.8fb^{-1}$
$\sqrt{s}$=8 TeV, $\int Ldt=5.9fb^{-1}$
$H\rightarrow\gamma\gamma$
(a)
(b)
Events / 2 GeV
Events - Bkg
3500
3000
2500
2000
1500
1000
500
200
100
0
-100
-200

But...

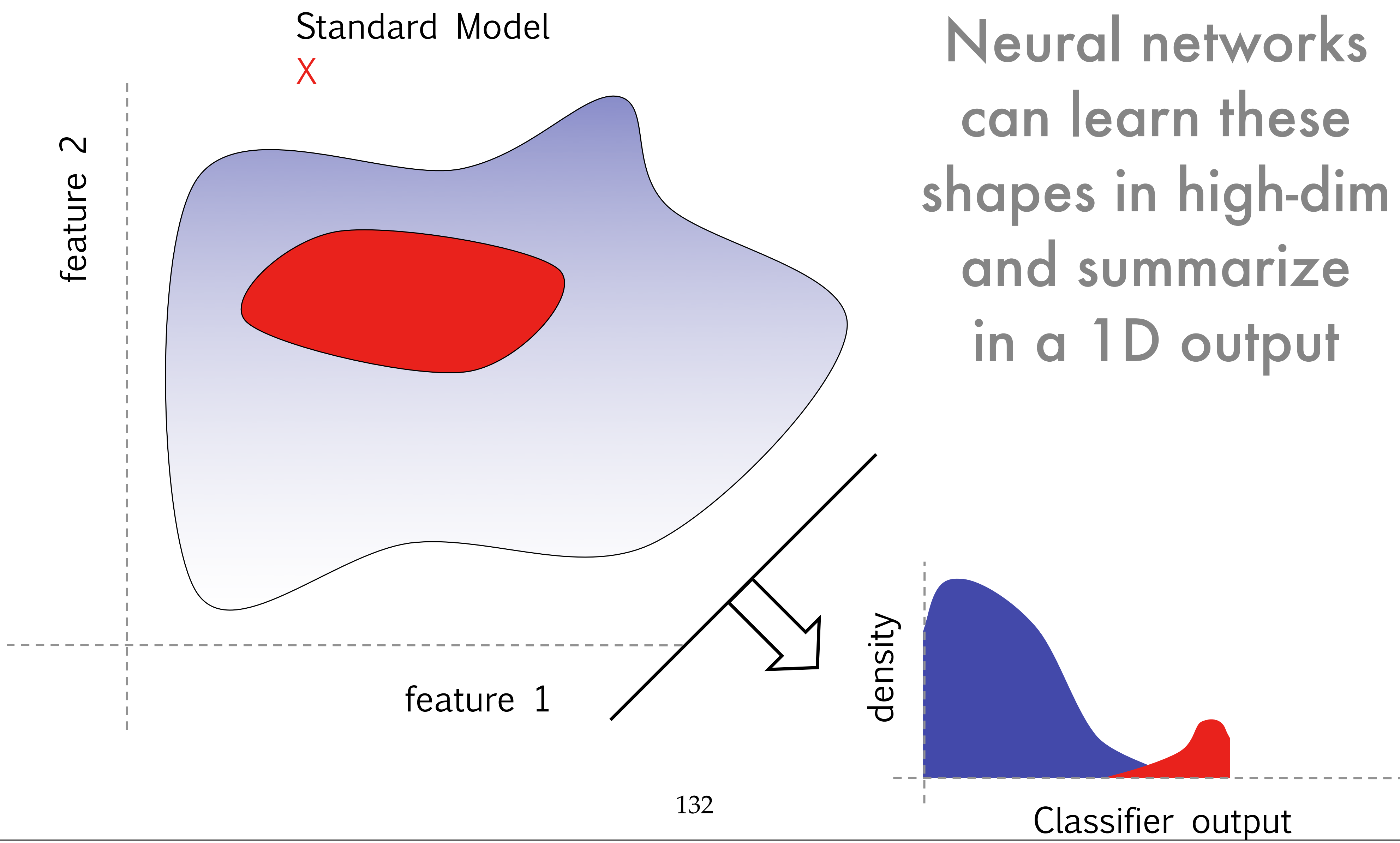
Reality is more complicated.

The full space can be very high dimensional.

Calculating likelihood in d-dimensional space requires ~100^d MC events.



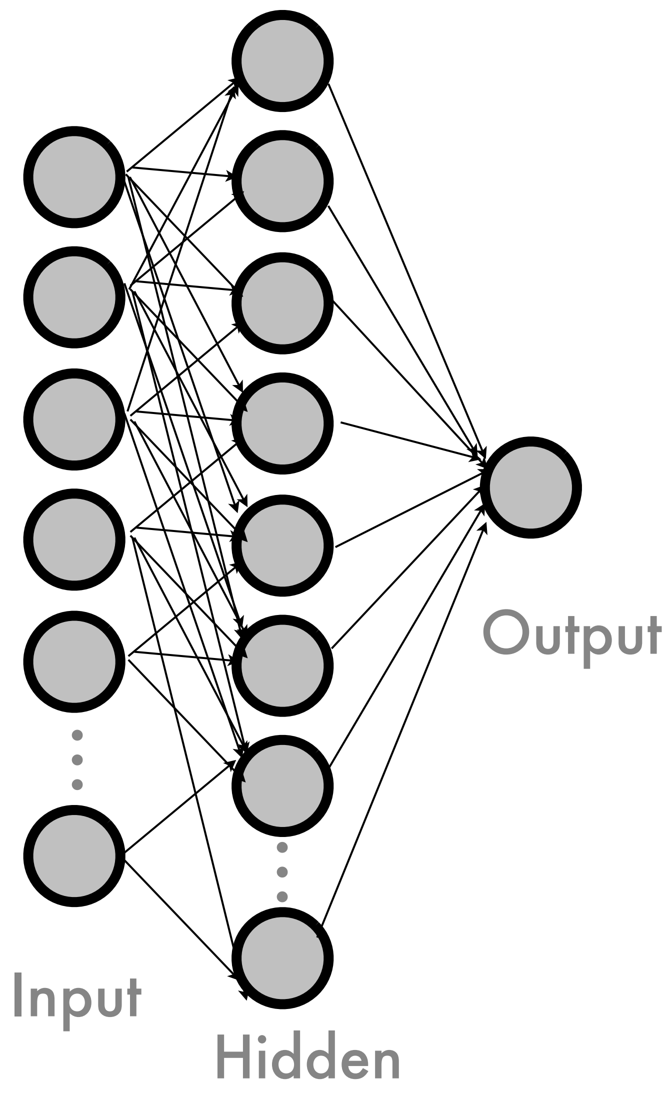
ML tools



Neural networks can learn these shapes in high-dim and summarize in a 1D output

Neural Networks

Essentially a functional fit with many parameters



Function

Each neuron's output
is a function of the
weighted sum of inputs.

Goal

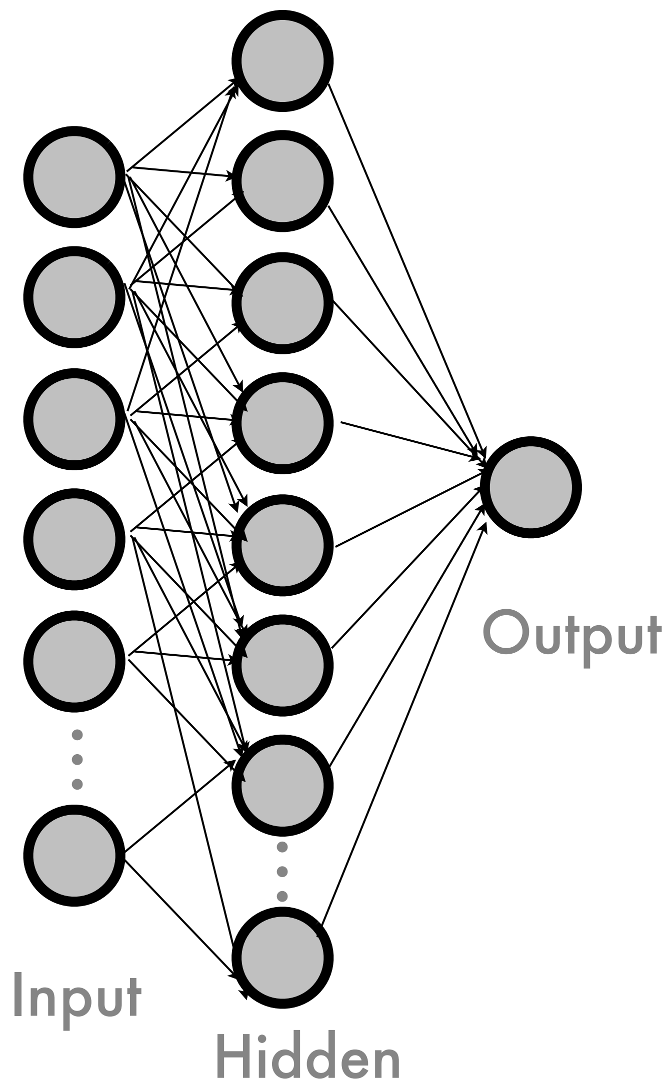
find set of weights
which give most useful function

Learning

give examples, back-propagate
error to adjust weights

Neural Networks

Essentially a functional fit with many parameters



Problem:
Networks with > 1 layer are very difficult to train.

Consequence:
Networks are not good at learning non-linear functions.
(like invariant masses!)

In short:
Can't just throw 4-vectors at NN.

Search for Input

ATLAS-CONF-2013-108

Can't just use 4v

Can't give it too many inputs

Painstaking search through input feature space.

Variable	VBF			Boosted		
	$\tau_{\rm lep}\tau_{\rm lep}$	$\tau_{\rm lep}\tau_{\rm had}$	$\tau_{\rm had}\tau_{\rm had}$	$\tau_{\rm lep}\tau_{\rm lep}$	$\tau_{\rm lep}\tau_{\rm had}$	$\tau_{\rm had}\tau_{\rm had}$
$m_{\tau\tau}^{\rm MMC}$	•	•	•	•	•	•
$\Delta R(\tau,\tau)$	•	•	•		•	•
$\Delta\eta(j_1,j_2)$	•	•	•			
m_{j_1,j_2}	•	•	•			
$\eta_{j_1}\times\eta_{j_2}$		•	•			
$p_{\rm T}^{\rm Total}$		•	•			
sum $p_{\rm T}$					•	•
$p_{\rm T}(\tau_1)/p_{\rm T}(\tau_2)$					•	•
$E_{\rm T}^{\rm miss}\phi$ centrality		•	•	•	•	•
$x_{\tau 1}$ and $x_{\tau 2}$						•
$m_{\tau\tau,j_1}$				•		
m_{ℓ_1,ℓ_2}				•		
$\Delta\phi_{\ell_1,\ell_2}$				•		
sphericity				•		
$p_{\rm T}^{\ell_1}$				•		
$p_{\rm T}^{j_1}$				•		
$E_{\rm T}^{\rm miss}/p_{\rm T}^{\ell_2}$				•		
$m_{\rm T}$		•			•	
$\min(\Delta\eta_{\ell_1\ell_2,{\rm jets}})$	•					
j_3 η centrality	•					
$\ell_1\times\ell_2$ η centrality	•					
ℓ η centrality		•				
$\tau_{1,2}$ η centrality			•			

Table 3: Discriminating variables used for each channel and category. The filled circles identify which variables are used in each decay mode. Note that variables such as $\Delta R(\tau,\tau)$ are defined either between the two leptons, between the lepton and $\tau_{\rm had}$, or between the two $\tau_{\rm had}$ candidates, depending on the decay mode.

Search for Input

ATLAS-CONF-2013-108

Can't just use 4v

Can't give it t… many inp…

Painstaking… through inp… feature space.

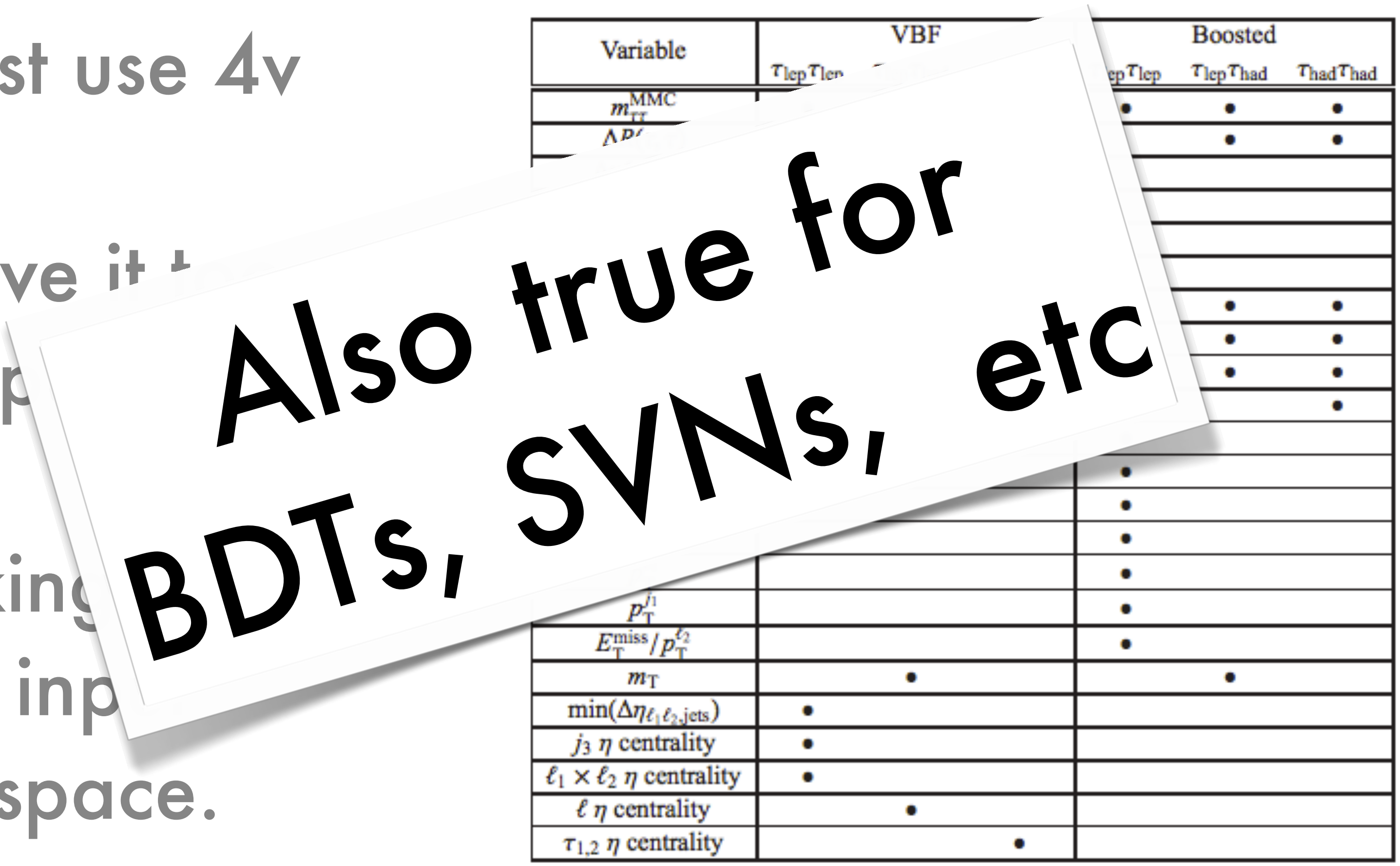
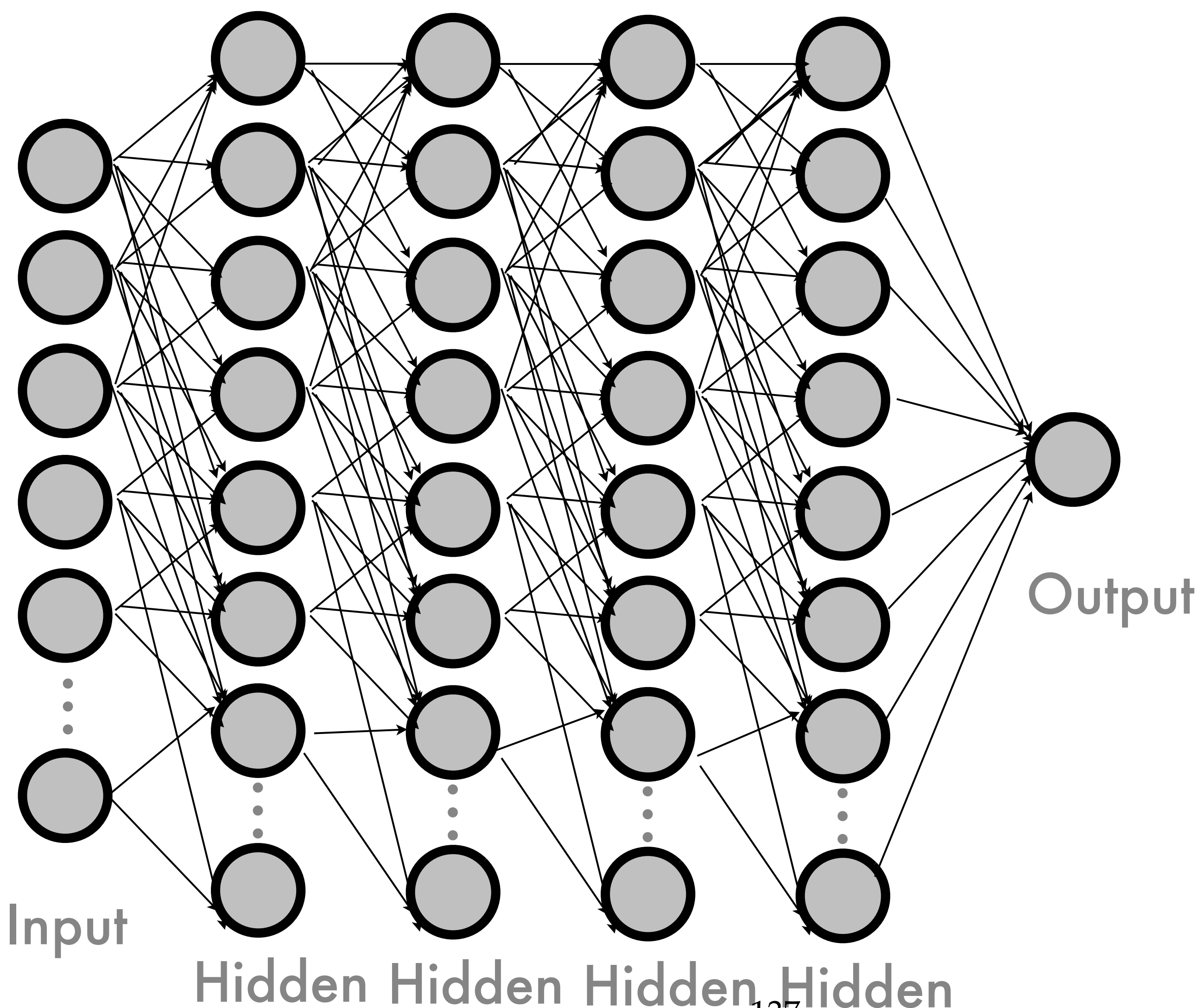


Table 3: Discriminating variables used for each channel and category. The filled circles identify which variables are used in each decay mode. Note that variables such as $\Delta R(\tau,\tau)$ are defined either between the two leptons, between the lepton and τ_{had}, or between the two τ_{had} candidates, depending on the decay mode.

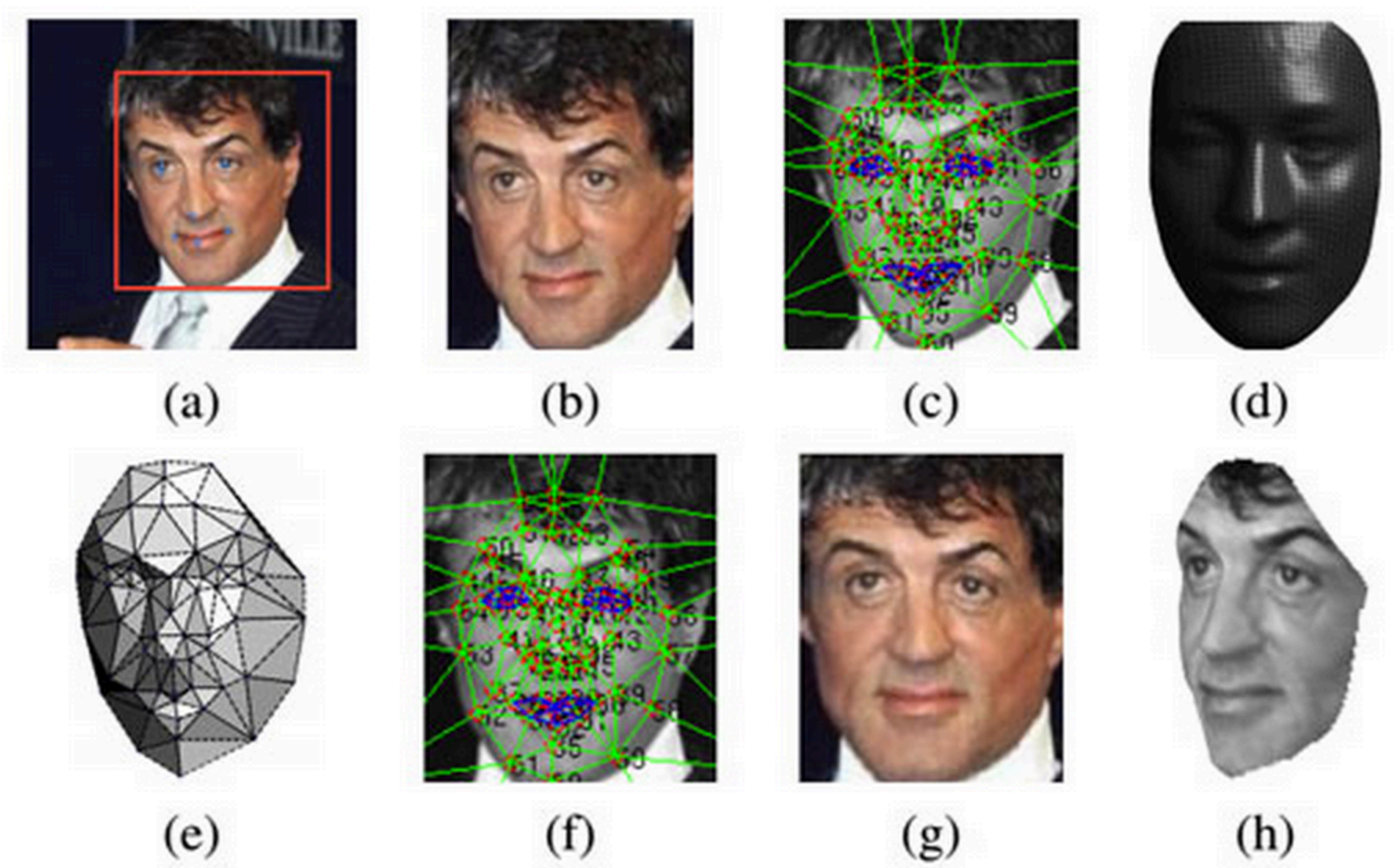
Deep networks



New tools let us train deep networks.

How well do they work?

Real world applications



Head turn: DeepFace uses a 3-D model to rotate faces, virtually, so that they face the camera. Image (a) shows the original image, and (g) shows the final, corrected version.

Paper

Deep Learning in High-Energy Physics: Improving the Search for Exotic Particles

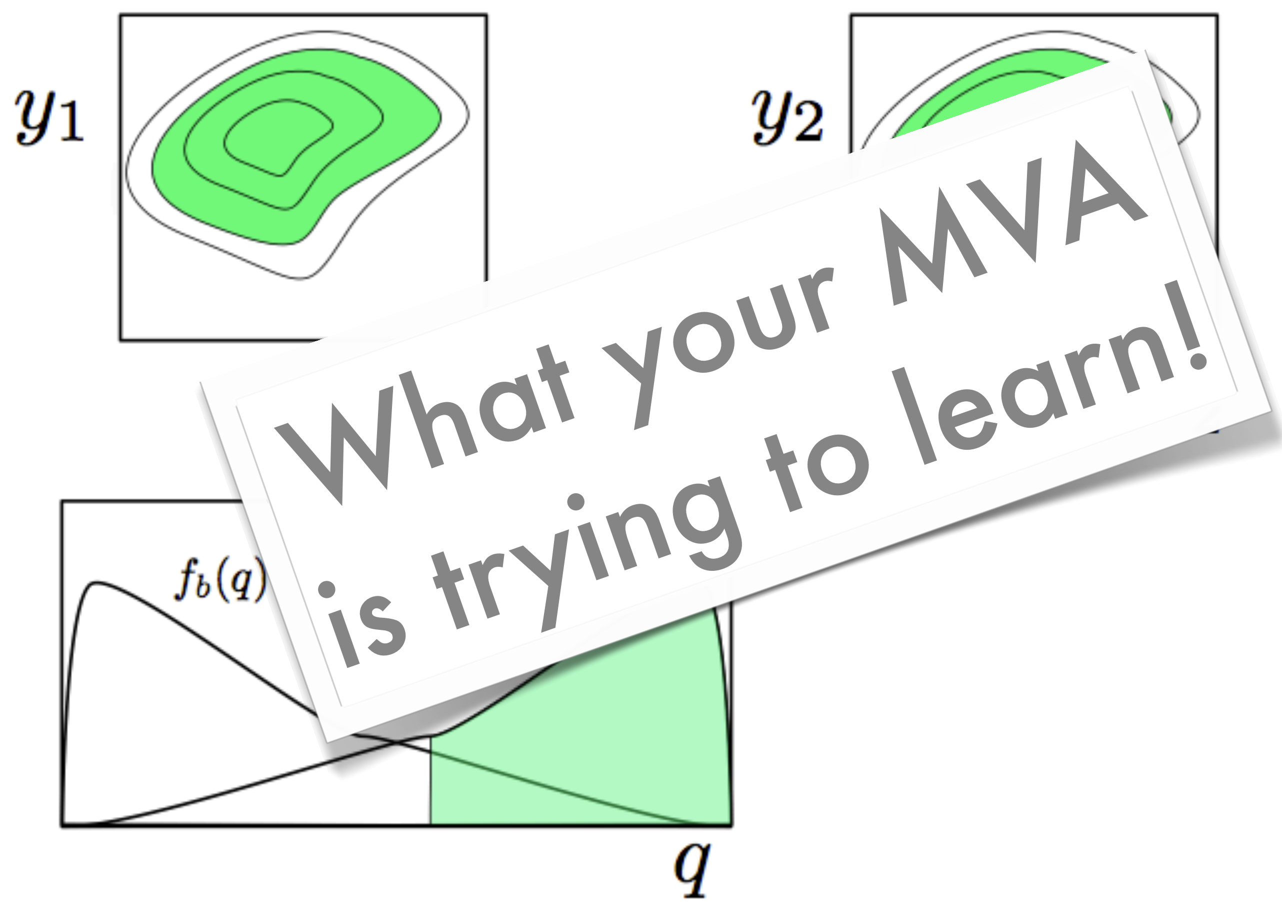
P. Baldi,[1] P. Sadowski,[1] and D. Whiteson[2]

[1] *Dept. of Computer Science, UC Irvine, Irvine, CA 92617*
[2] *Dept. of Physics and Astronomy, UC Irvine, Irvine, CA 92617*

arXiv: 1402.4735
Accepted in *Nature Comm.*

What does the TS do?

Finds a region in variable space



(K. Cranmer)

Test statistic

At LEP, this was used:

Define μ to be signal strength,
μ=0 is no signal
μ=1 is theory prediction

$$Q_{LEP} = L_{s+b}(\mu = 1)/L_b(\mu = 0)$$

Where the nuisance parameters
are fixed to their nominal values

Test statistic

Define μ to be signal strength,
μ=0 is no signal
μ=1 is theory prediction

At LEP, this was used:

$$Q_{LEP} = \frac{L(data|\mu = 1, b, \nu)}{L(data|\mu = 0, b, \nu)}$$

This also means the background estimate doesn't vary.

Tevatron

Still consider two points (0,1)
but now float the NPs at those points

$$Q_{TEV} = L_{s+b}(\mu = 1, \hat{\hat{\nu}})/L_b(\mu = 0, \hat{\hat{\nu}}')$$

Ratio of profiled likelihoods:
the model is adapted to the data
even in the signal region

LHC

Profile likelihood

$$\lambda(\mu=0) = \frac{L(data|\mu=0,\hat{\hat{b}}(\mu=0),\hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu},\hat{b},\hat{\nu})}$$

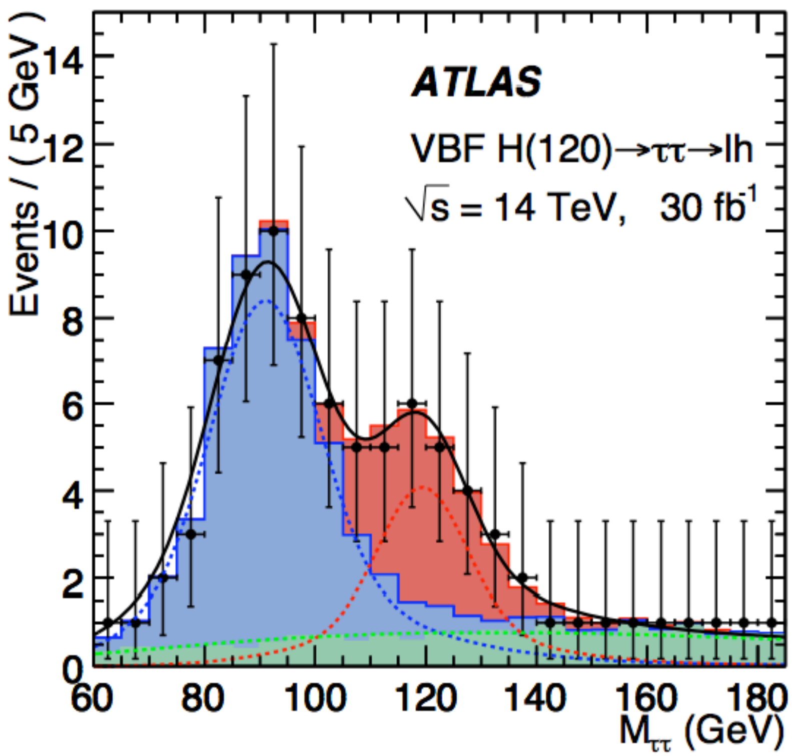
fit best value of NPs at μ=0

and at best fit value of μ

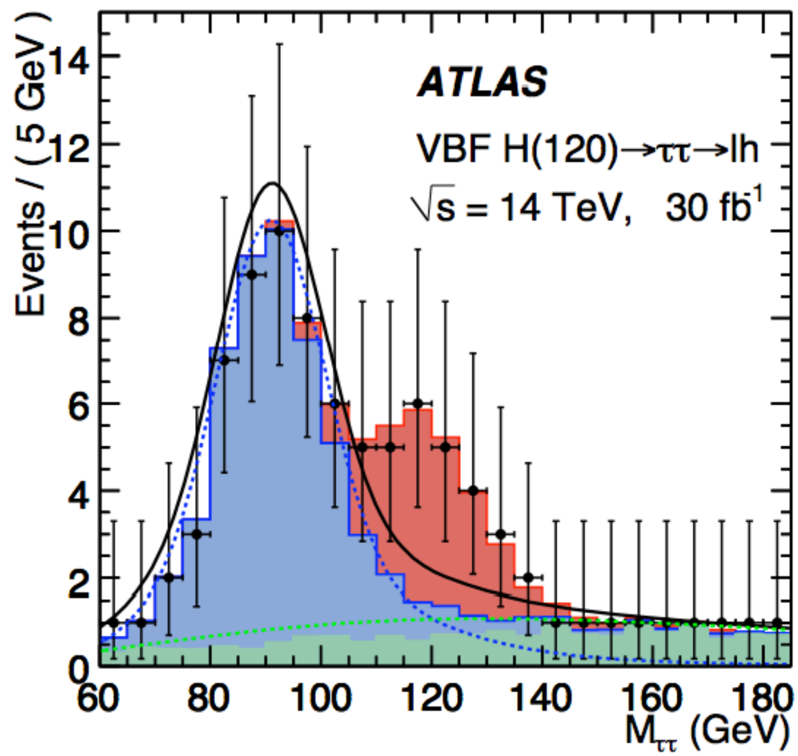
Two fits to data

$$\lambda(\mu=0) = \frac{L(data|\mu=0,\hat{\hat{b}}(\mu=0),\hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu},\hat{b},\hat{\nu})},$$

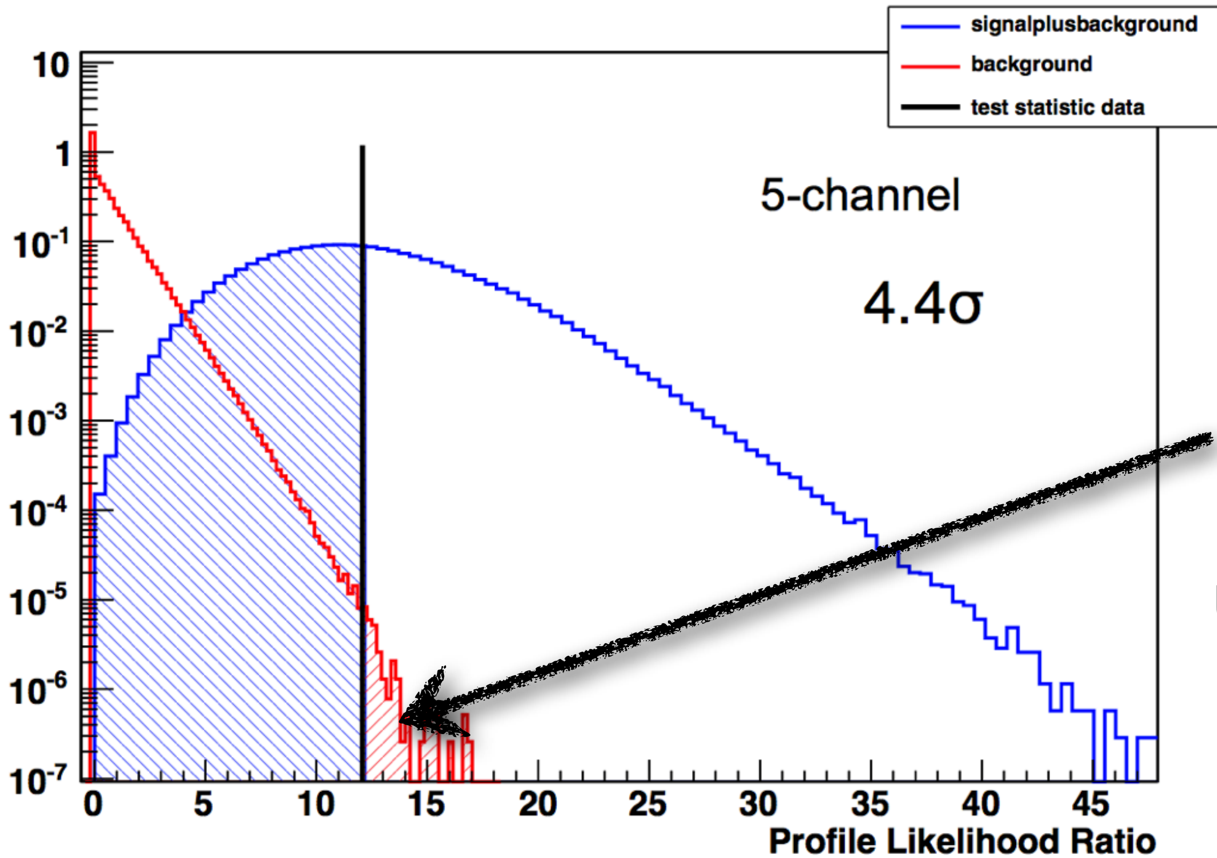
$L(data|\hat{\mu},\hat{b},\hat{\nu})$



$L(data|\mu=0,\hat{\hat{b}},\hat{\hat{\nu}})$



p values

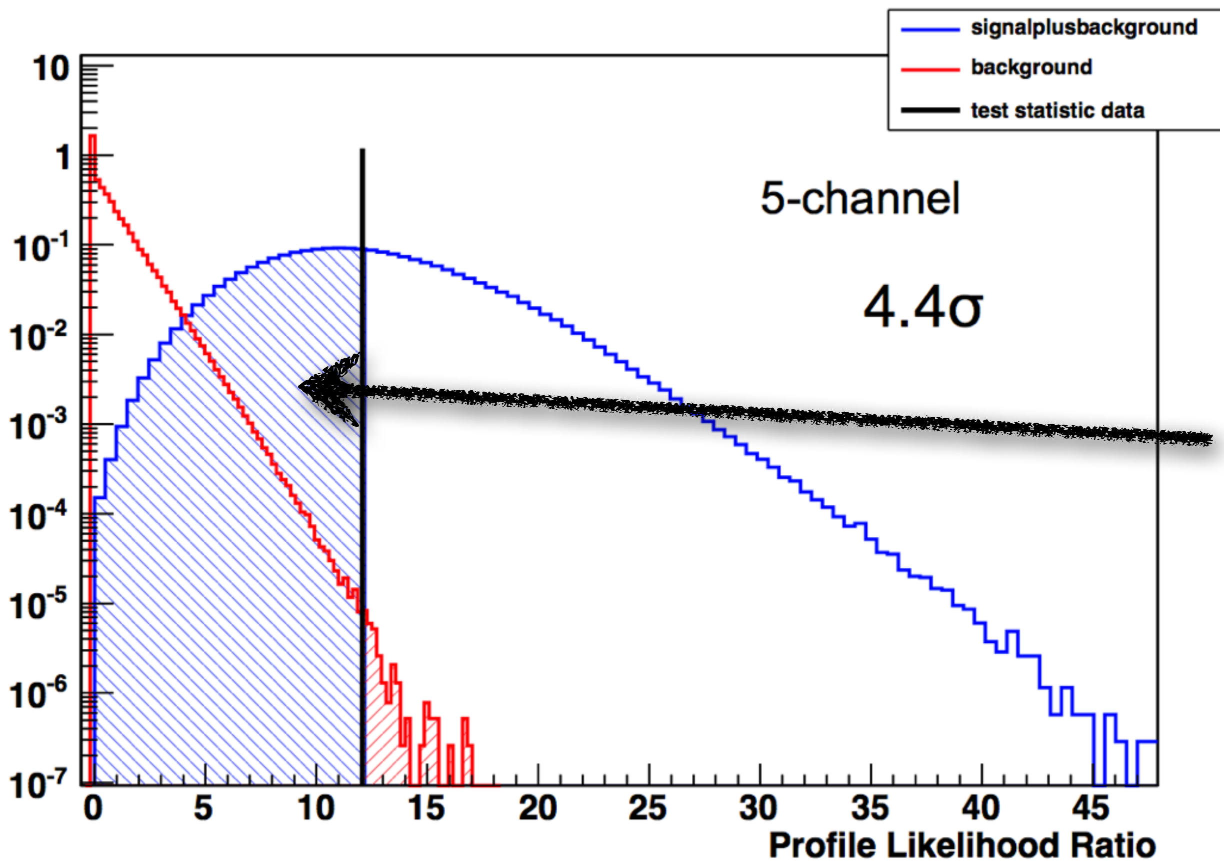


p_0 = probability to observe data or **more** signal-like under **background** hypothesis

$$p_0 = P\left(q_0 \geq q_o^{obs}\right)$$

(K. Cranmer)

p values



p_μ =
probability
to observe data
or less signal-like
under signal+b
hypothesis

(K. Cranmer)

Philosophy

Bayesian

&

Frequentist

Bayesian

Data: fixed

Parameter values: unknown

Probability: our lack of knowledge

PDFs over parameters: sensible

Frequentist

Data: one example from ens.
Parameter values: fixed (even if unknown)
Probability: rate of occurance
PDFs over parameters: not sensible

Bayesian Prob.

Bayes theorem:

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

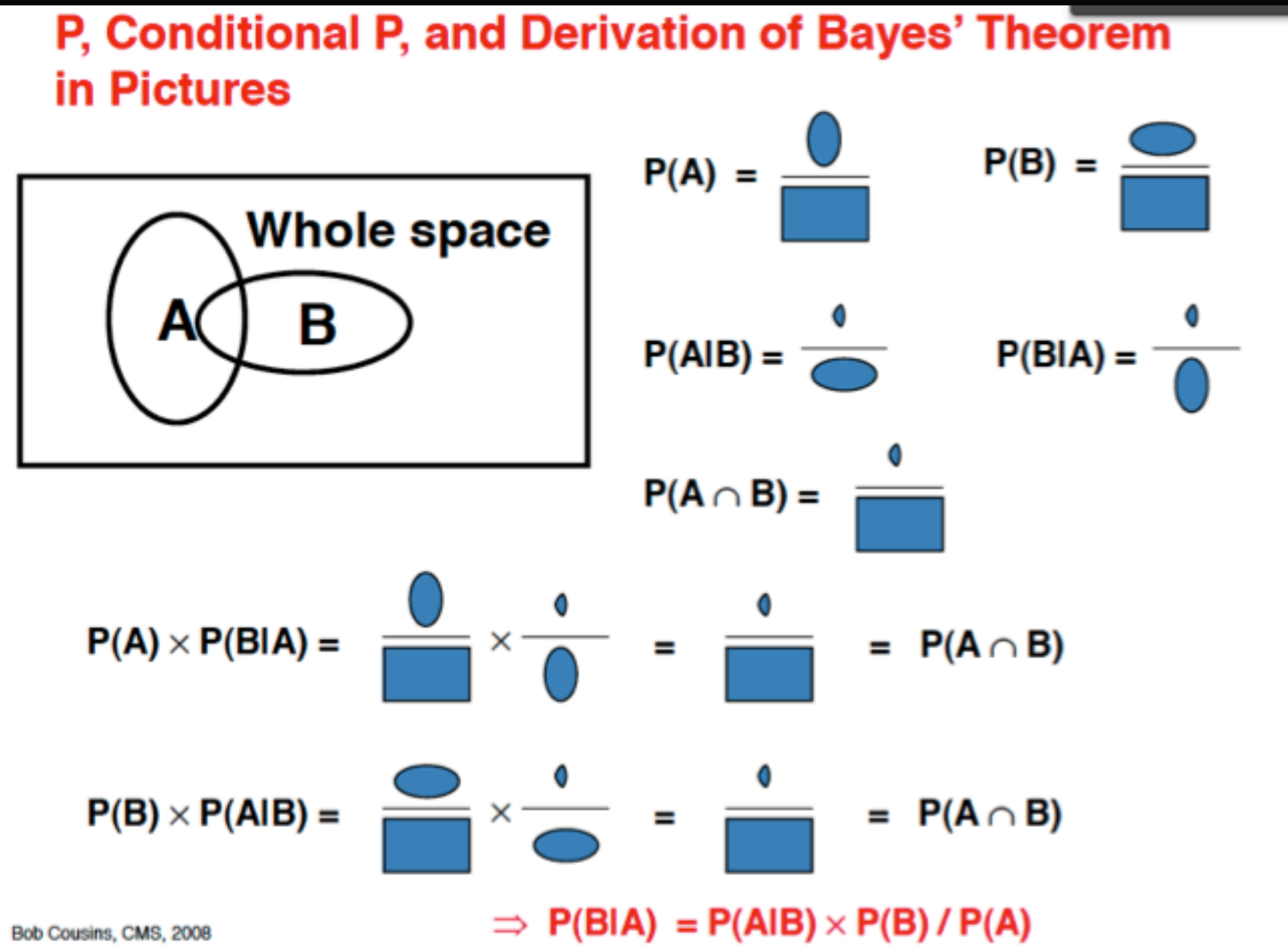
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

rearrange:

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

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

In Pictures


P, Conditional P, and Derivation of Bayes' Theorem in Pictures
Whole space
A
B
P(A) =
P(B) =
P(AIB) =
P(BIA) =
P(A ∩ B) =
P(A) × P(BIA) = × = = P(A ∩ B)
P(B) × P(AIB) = × = = P(A ∩ B)
⇒ P(BIA) = P(AIB) × P(B) / P(A)
Bob Cousins, CMS, 2008

Example 1

P(data|theory) != P(theory|data)

Theory = (male or female)

Data = (pregnant | not pregnant)

P(pregnant | female) ~ 3%

BUT

P(female | pregnant) >99%

Example 2

<u>Higgs search</u>

Expected bg = 0.1

Expected signal = 10

P(N| no Higgs) = 0.1

P(N| Higgs) = 10.1

What is P(Higgs|N=8)?

$$P(H|N=8) = \frac{P(N=8|H)P(H)}{P(N=8)}$$

Depends on P(H)!

(K Cranmer)

Parameter estimation

<u>Bayesian parameter estimation:</u>

Want to know the probability that some parameter θ is in some range $[\theta_0,\theta_1]$

- or -

Want to find a range $[\theta_0,\theta_1]$ that has probability of 0.95

Parameter estimation

<u>Bayesian parameter estimatio</u>

Want to know the probab

parameter θ is in som

- or -

Want t $[\theta_0, \theta_1]$ that has probability

of

<u>Remember:</u>

Probability reflects lack of knowledge, and prior information.

How?

The probability that the true value is inside an interal is:

$$1 - \alpha = \int_{\theta_{lo}}^{\theta_{hi}} p(\theta|x) d\theta$$

For lower or upper limits, choose zero or infinity as boundaries.
where we integrate out the nuisance parameters:

$$p(\theta|x) = \int d\nu p(\theta, \nu|x)$$

where

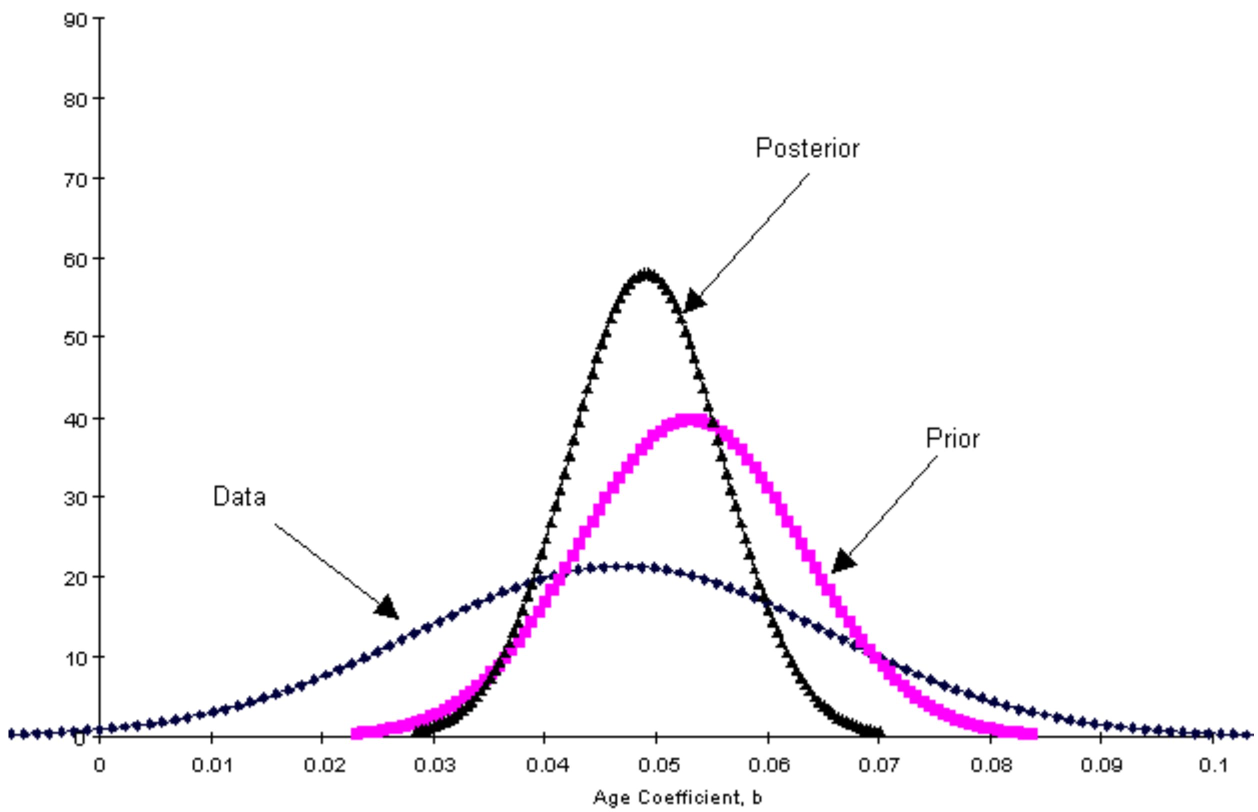
$$p(\theta, \nu|x) = \frac{p(x|\theta, \nu) p(\theta, \nu)}{p(x)}$$

These integrals can be very hard to do if the space is high dimensional.

Priors

Choice of prior p(θ)

- important but subjective choice

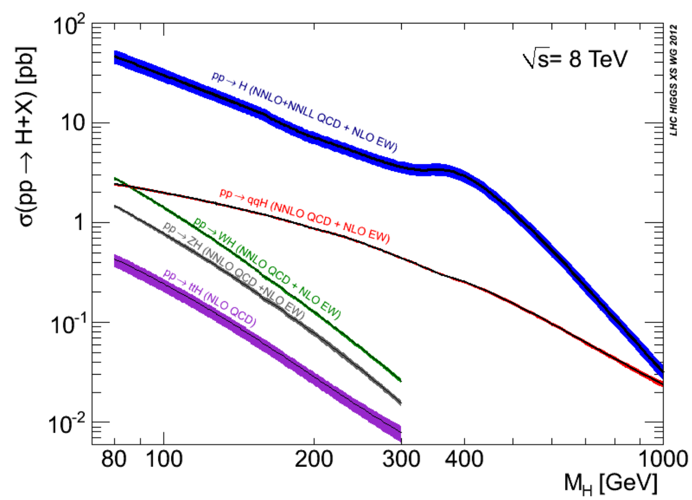


Priors

<u>Choice of prior p(θ)</u>

- Example: measuring Higgs cross-section
- Want to be unbiased: choose uniform prior?

$$\sigma=[0,\Lambda] \rightarrow \; P = k$$



- But σ and mass relationship makes this prior **not flat in mass**

- no uninformative prior across all transformations