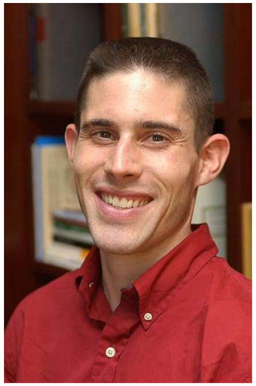
Vista as a tool for generator tuning

Stephen Mrenna
CD/FNAL

HERALHC

The MIT High-Pt group:



Bruce Knuteson

Conor Henderson

Ray Culbertson

Georgios Choudalakis

Dictionary.com Unabridged (v 1.1) - vis·ta /ˈvɪstə/

1. a view or prospect, esp. one seen through a long, narrow avenue or passage, as between rows of trees or houses.

2. such an avenue or passage, esp. when formally planned.

3. a far-reaching mental view: vistas of the future.

[Origin: 1650–60; < It: a view, n. use of fem. of visto (ptp. of vedere to see < L vidére)]

Dictionary.com Unabridged (v 1.1) - vis·ta /ˈvɪstə/

1. a view or prospect, esp. one seen through a long, narrow avenue or passage, as between rows of trees or houses.

2. such an avenue or passage, esp. when formally planned.

3. a far-reaching mental view: vistas of the future.

4. attempt to understand the entirety of high-pT collider data, with a focus on large scale discrepancies and shapes

[Origin: 1650–60; < It: a view, n. use of fem. of visto (ptp. of vedere to see < L vidére)]

How do we address the question:

Do we understand the entirety of high-pT (collider) data?

Implement the Standard Model

Look in many final states

Final State	Chi2	data	bkg			
1b3j1pmiss_sumPt400+ [73]	9.0	451	374.5	+-	18	(pyth_jj_200 =
2b1e+2j [-]	8.0	15	6.5	+-	1.9	(ttcp0z = 2.3
2j_sumPt0-400 [161]	6.0	69704	67013.6	+-	1171.2	(pyth_jj_018 =
2j2mu+1pmiss [-]	-5.0	2	12.2	+-	3	(mad_mu+mu-jj
1b2e+2j [-]	5.0	9	3.9	+-	1.5	(mrenna_e+e-jj
1j1ph1pmiss [5]	4.0	2591	2470.1	+-	37.7	(pyth_pj_045 =
2j1mu+1ph [-]	4.0	11	11.2	+-	2.2	(mrenna_mu+mu-
1e+1j1mu+ [-]	4.0	13	6.6	+-	2.1	(ztcp5i = 3.4
1e+2j1ph [-]	4.0	31	20.9	+-	2.7	(mad_aajj = 6.
3j2mu+ [-]	4.0	34	23.2	+-	2.7	(mrenna_mu+mu-
2b2j1pmiss_sumPt400+ [-]	-3.0	17	30.4	+-	4.2	(pyth_jj_200 =
1b2j_sumPt400+ [229]	3.0	4669	4518.6	+-	72.7	(pyth_jj_200 =
4j_sumPt0-400 [253]	-3.0	2611	2736.9	+-	42.3	(pyth_jj_040 =
2b1j1ph1pmiss [-]	3.0	6	2.7	+-	1.5	(pyth_jj_200 =
1b1j1mu+ [-]	3.0	67	53.8	+-	4.3	(pyth_jj_018 =
1j1ph [277]	3.0	31738	31149.8	+-	352.1	(pyth_pj_045 =
1e+1mu+ [-]	3.0	66	53.5	+-	3.2	(ztcp5i = 38.8
4j1mu+ [-]	3.0	73	61.3	+-	2.6	(pyth_jj_040 =
5j [269]	3.0	448	406	+-	14.5	(pyth_jj_040 =
1b5j [-]	3.0	8	8.9	+-	1.7	(pyth_jj_060 =
1b1j1pmiss_sumPt0-400 [-]	2.0	120	104	+-	7.2	(pyth_jj_040 =
2j1pmiss_sumPt0-400 [37]	2.0	2381	2281.2	+-	73.9	(pyth_jj_018 =

...

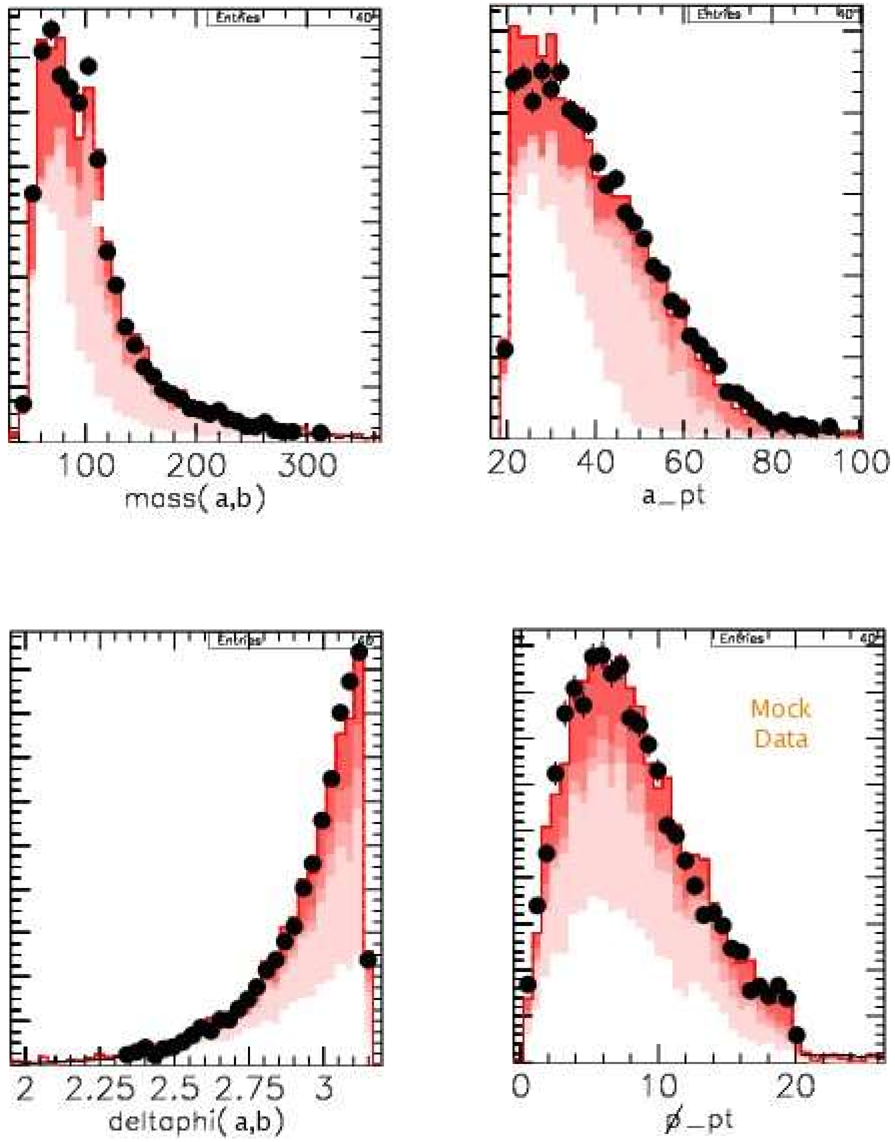
Have a measure of agreement

Final State	Chi2	data	bkg	
1b2e+2j [-]	5.0	9	3.9 +-	1.5

(mrenna_e+e-jjj = 1.9 , mad_e+e-jj = 0.5 , mrenna_e+e-jj = 0.4 , mad_e+e-b-b = 0.4 , ztopcz = 0.3 , pyth_jj_040 = 0.2 , mad_aajj = 0.1 , mrenna_e+vejjjj = 0.1 , hewk03 = 0.1 , wtop1z = 0.1)

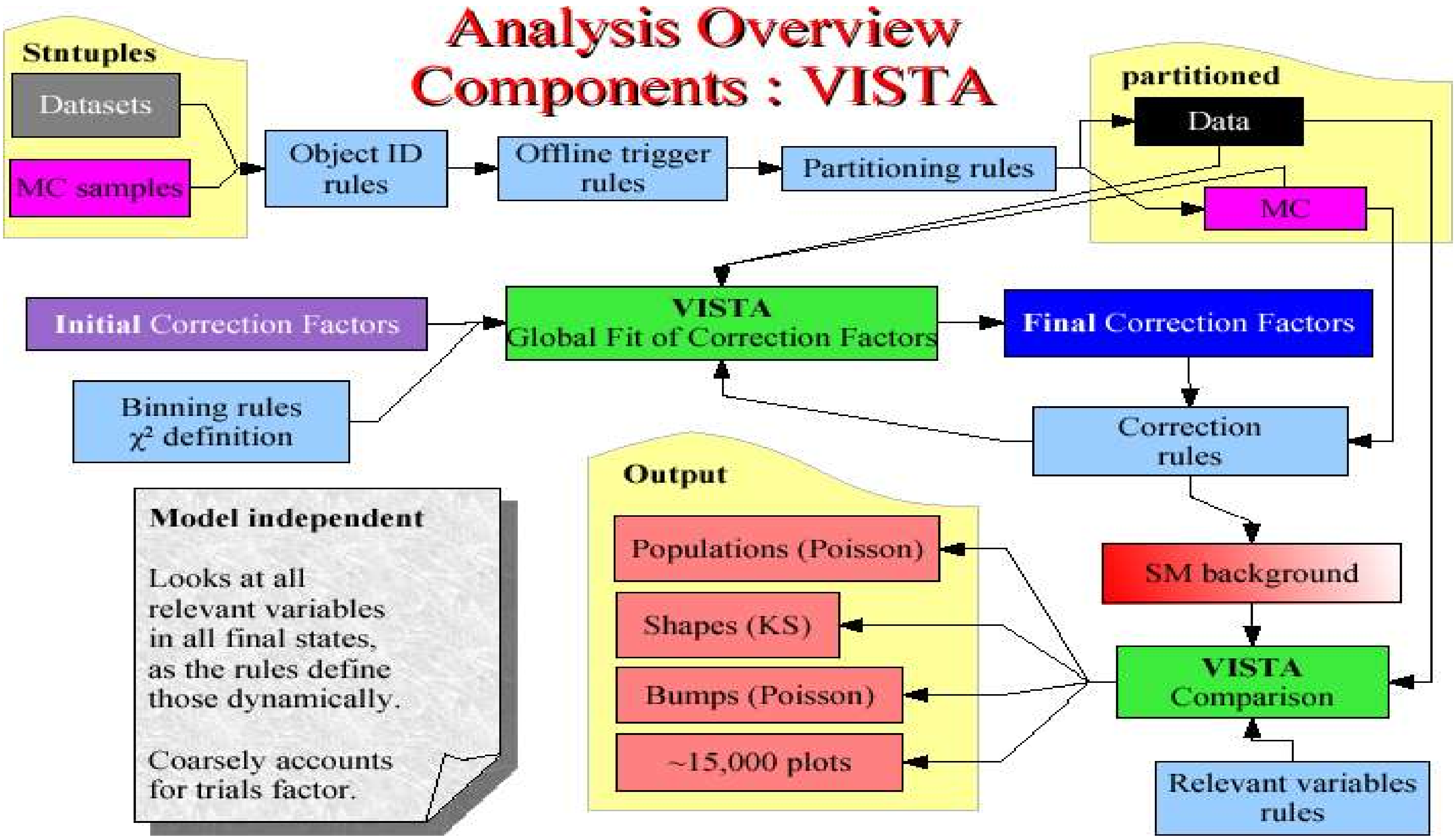
Must mix a cocktail

Check also consistency between shapes of object distributions in each final state

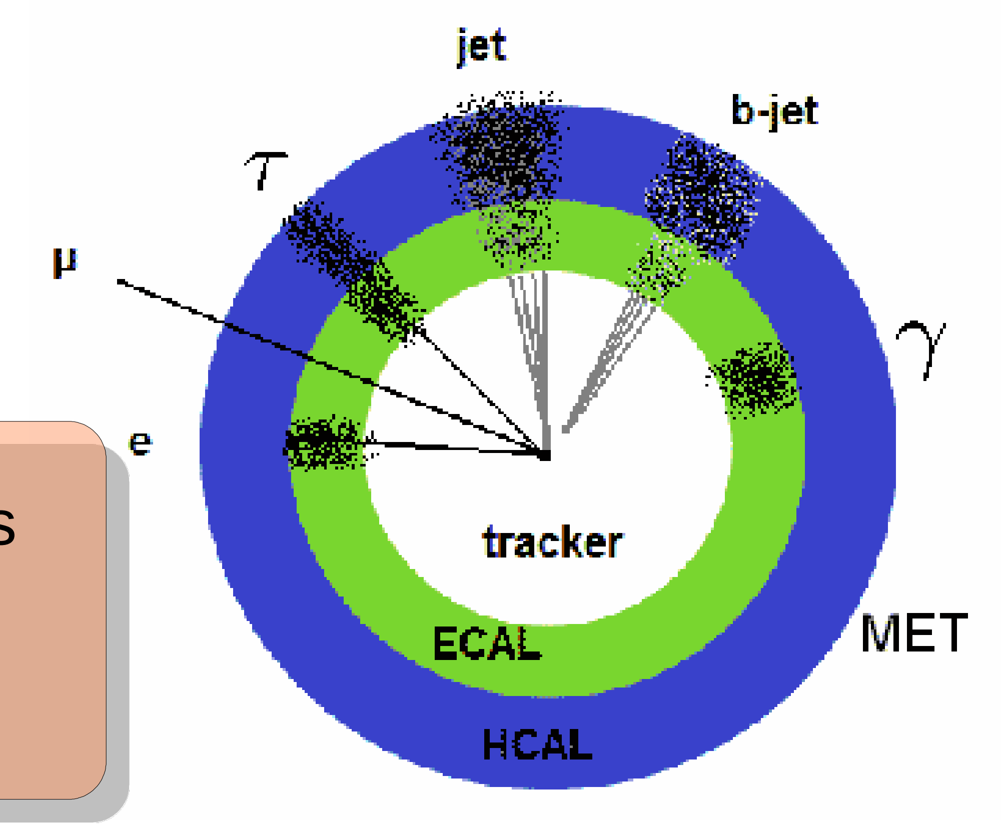


Vista Overview

1. Identify objects
2. Select events
3. Partition events
4. Implement the SM
5. Determine correction factors
6. Perform global comparison
7. Investigate discrepancies


Analysis Overview
Components : VISTA
Stntuples
Datasets
MC samples
Object ID rules
Offline trigger rules
Partitioning rules
partitioned
Data
MC
Initial Correction Factors
VISTA
Global Fit of Correction Factors
Final Correction Factors
Binning rules
χ^2 definition
Correction rules
Model independent
Looks at all relevant variables in all final states, as the rules define those dynamically.
Coarsely accounts for trials factor.
Output
Populations (Poisson)
Shapes (KS)
Bumps (Poisson)
~15,000 plots
SM background
VISTA
Comparison
Relevant variables rules

standard definition of objects.
objects are reasonably "tight".



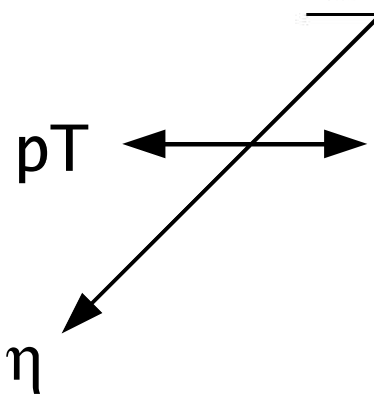
Same objects
in all
final states

Standard ID cuts
Objects identified if $p_T > 17$ GeV

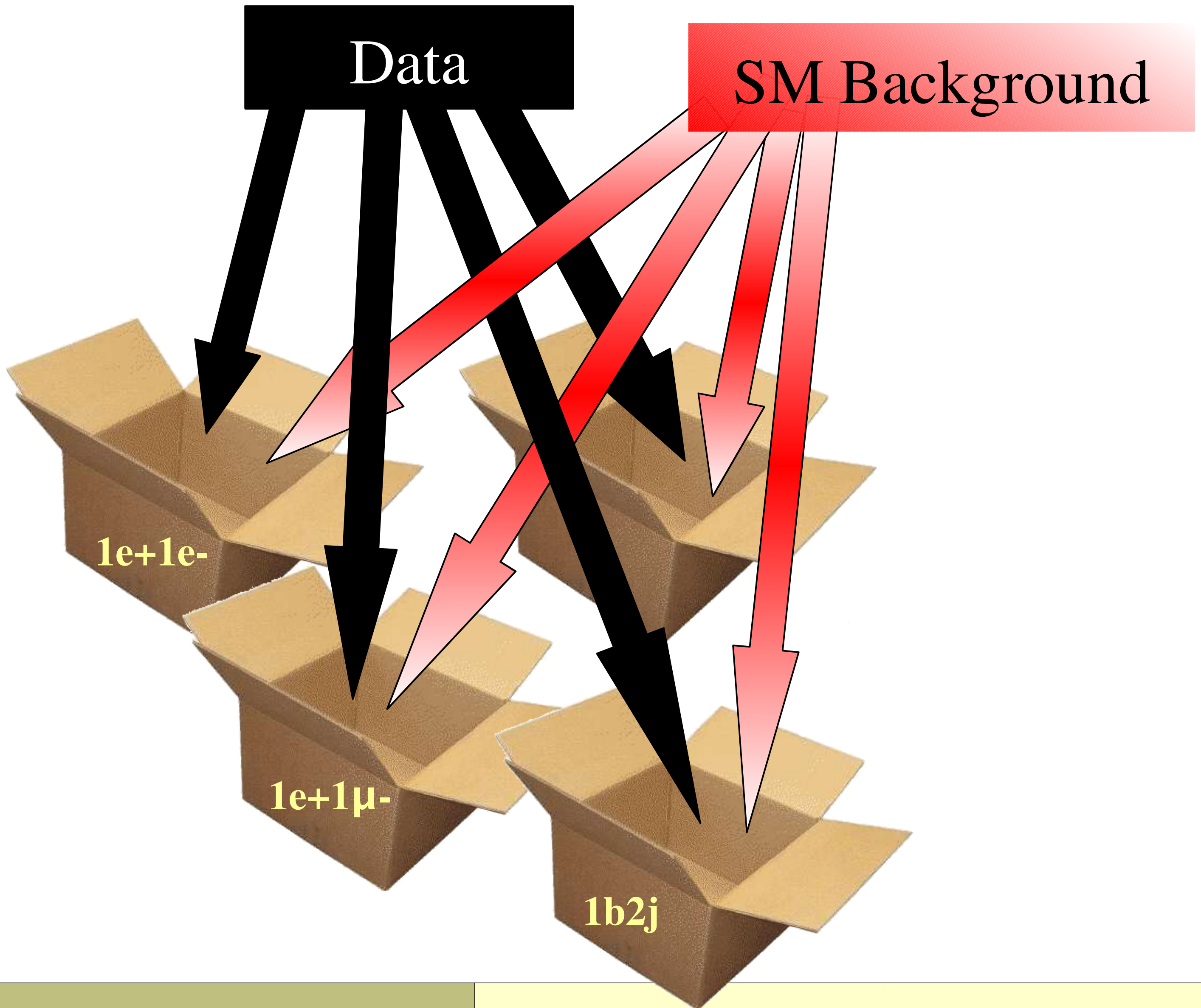
Cartoon of Vista mis-id matrix

Observed

Truth	e	μ	τ	γ	j	b
e	0.9		10^{-2}	10^{-3}	10^{-1}	10^{-3}
μ		0.9				
τ			10^{-1}		0.9	
γ				0.9	10^{-1}	
j	10^{-4}	10^{-6}	10^{-3}	10^{-4}	1	10^{-3}
b	10^{-3}	10^{-3}	10^{-3}	10^{-3}	0.6	0.4



Vista: partitioning into final states



~350 boxes

partition rules for Vista

- Group events exclusively, according to what final reconstructed objects they contain. Objects are : e, γ, μ, τ, b, j, pTmiss
- Any object with pT > 17 GeV counts.
- μ,τ, b : $|\eta| < 1.0$
- e, j, γ : $|\eta| < 2.5$

Monte Carlo samples used

Matched (MadEvent+Pythia) W/Z + jets

Pythia QCD dijets

Pythia QCD photon+jet(s)

Pythia QCD diphoton (+jets)

Herwig ttbar

Pythia VV

MadEvent + Pythia filling in holes (e.g. Z+2γ)

Using standard CDF samples wherever possible.

Entirely Modular: plug-in better
Monte Carlo where needed

Correction Factors

object-dependent, location-dependent, energy-dependent corrections applied to particle-level events to maximize the agreement between "theory" and data

1. consistent with external constraints

2. some theory-based (K-factors)

3. expt.-based (misid, trigger + prescale, energy scale, ...)

4. instead of defining a control region, one checks for consistency of the fit

5. <50 correction factors, but >10000 distributions

6. start with 0, and add those that are needed (debug detector and SM simultaneously)

Correction Factors

Code	Category	Explanation	
5001	luminosity	CDF integrated luminosity	
5102	k-factor	cosmic_ph	
5103	k-factor	cosmic_j	
5121	k-factor	1ph1j	photon+jet(s)
5122	k-factor	1ph2j	
5123	k-factor	1ph3j	
5124	k-factor	1ph4j+	
5130	k-factor	2ph0j	diphoton(+jets)
5131	k-factor	2ph1j	
5132	k-factor	2ph2j+	
5141	k-factor	W0j	W (+jets)
5142	k-factor	W1j	
5143	k-factor	W2j	
5144	k-factor	W3j+	
5151	k-factor	Z0j	Z (+jets)
5152	k-factor	Z1j	
5153	k-factor	Z2j+	
5161	k-factor	2j pt<150	dijet
5162	k-factor	2j 150<pt	
5164	k-factor	3j pt<150	multijet
5165	k-factor	3j 150<pt	
5167	k-factor	4j pt<150	
5168	k-factor	4j 150<pt	
5169	k-factor	5j+ low	

5211	misId	p(e->e)	central
5212	misId	p(e->e)	plug
5213	misId	p(mu->mu)	CMUP
5214	misId	p(mu->mu)	CMX
5216	misId	p(ph->ph)	central
5217	misId	p(ph->ph)	plug
5219	misId	p(b->b)	central
5245	misId	p(e->ph)	plug
5256	misId	p(q->e)	central
5257	misId	p(q->e)	plug
5261	misId	p(q->mu)	
5273	misId	p(j->b)	25<pt
5285	misId	p(q->tau)	15<pt<60
5286	misId	p(q->tau)	60<pt<200
5292	misId	p(q->ph)	central
5293	misId	p(q->ph)	plug
5401	trigger	p(e->trig)	central, pt>25
5402	trigger	p(e->trig)	plug, pt>25
5403	trigger	p(mu->trig)	CMUP, pt>25
5404	trigger	p(mu->trig)	CMX, pt>25

the global fit

$$\chi^2(\vec{s}) = \left(\sum_{k \in \mathrm{bins}} \chi_k^2(\vec{s}) \right) + \chi^2_{\mathrm{constraints}}(\vec{s})$$

$$\chi_k^2(\vec{s}) = \frac{(\mathrm{Data}[k] - \mathrm{SM}[k])^2}{\delta \mathrm{SM}[k]^2 + \sqrt{\mathrm{SM}[k]}^{\,2}}$$

SM = Integrated Luminosity ×
{σ_{LO} x k-factors} ×
{ID and misID probabilities} ×
{Trigger Efficiencies}

There are no systematic uncertainties on the procedure ?

Vista takes a different approach to the question of systematic uncertainties.

Finds values for the correction factors that allow the data to be described by the Standard Model.

The correction factors themselves are systematic shifts

The uncertainties on the values are the errors on these systematic shifts.

How can you possibly generate a complete Standard Model prediction?

At LEP 2, all 4 experiments had an internal web page containing a complete Standard Model cocktail, showing which Monte Carlo samples to mix together (and in what proportions) to obtain a complete Standard Model background estimate.

A similar cocktail was mixed at H1 in HERA Run I.

In this analysis, we have constructed such a cocktail for the first time at a hadron-hadron collider.

The size is intimately tied to the choice of offline trigger and identification pT thresholds.

Vista comparison

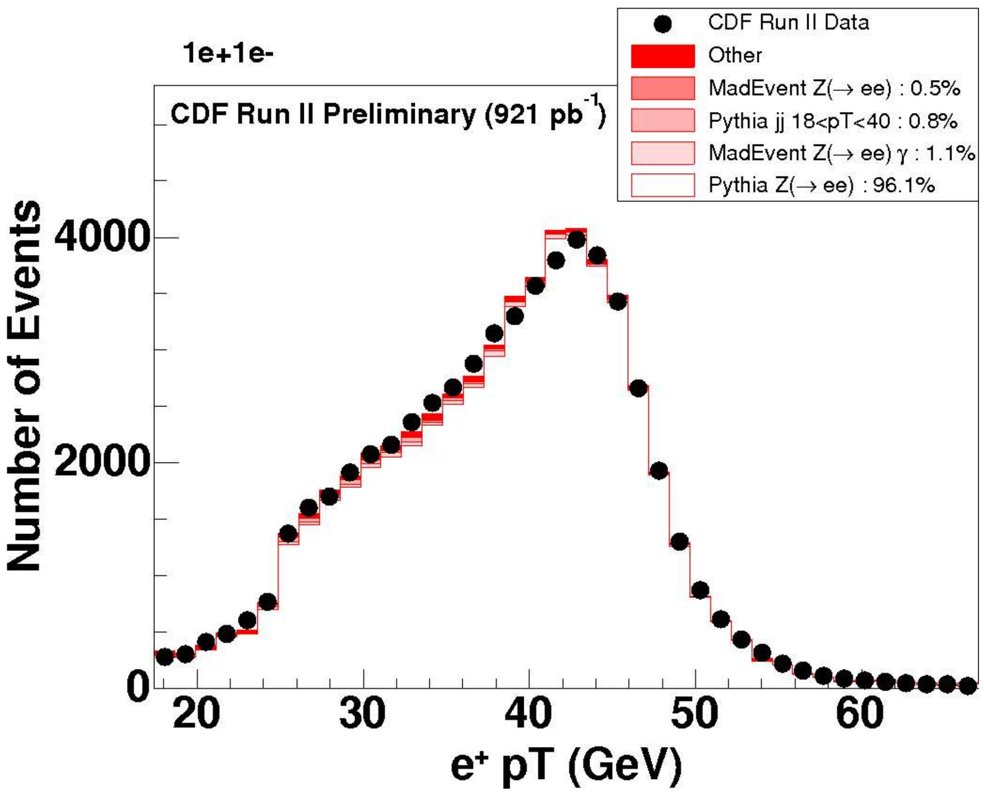
344 final state populations

Dissimilarities follow:

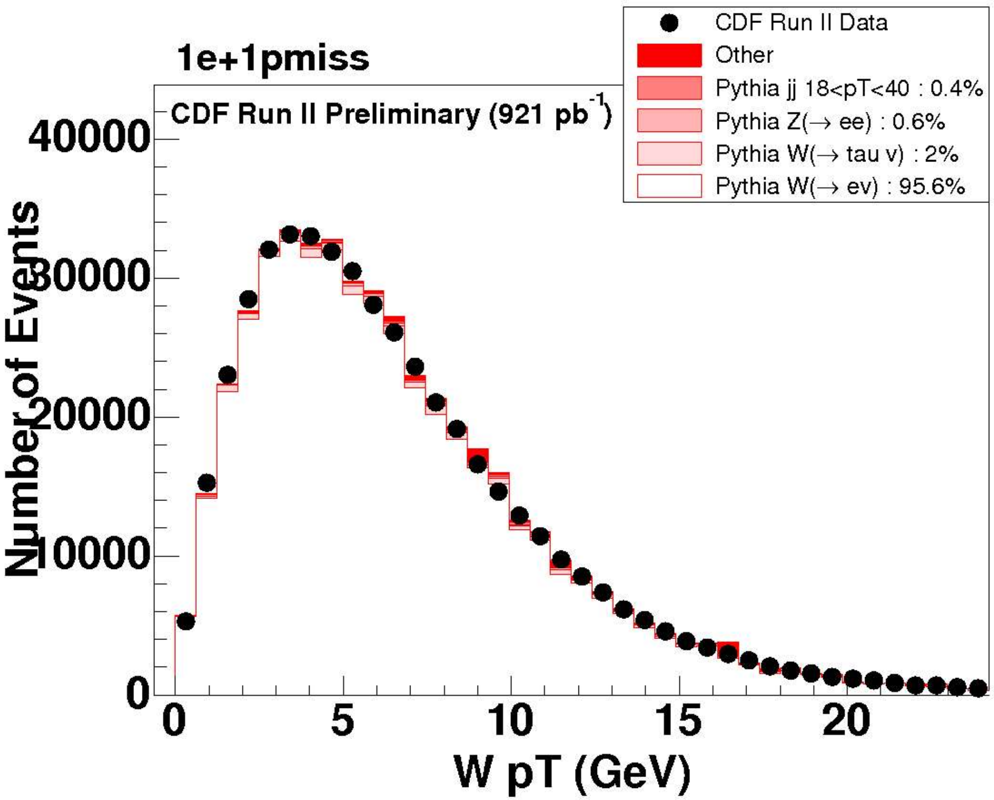
Final state	effect (σ)	data	bkg		stat. uncertainty
	-		113.7	+-	3.6
	-		1902.9	+-	50.8
	-		296.5	+-	5.6
	-		27	+-	4.6
	+		2015.4	+-	28.7
	-		37294.6	+-	524.3
	+		1751.6	+-	42
	+		695.3	+-	13.3
	-		967.5	+-	38.4
	+		11.6	+-	1.5
	+		551.2	+-	11.2
	+		27281.5	+-	405.2
	+		95	+-	4.7
	-		85.6	+-	8.2
	-		125	+-	13.6
	-		29.5	+-	4.6
	-		369.4	+-	21.1
	+		14.2	+-	1.8
	+		92437.3	+-	1354.5
	+		298.6	+-	7.7
			0.7	+-	1.2
			6.1	+-	2.5
			35.6	+-	4.9
			298.4	+-	14.7
			2.1	+-	1.6
			40898.8	+-	649.2
			8403.7	+-	144.7
			57.5	+-	11

Vista comparison

~15,000 plots of kinematic distributions

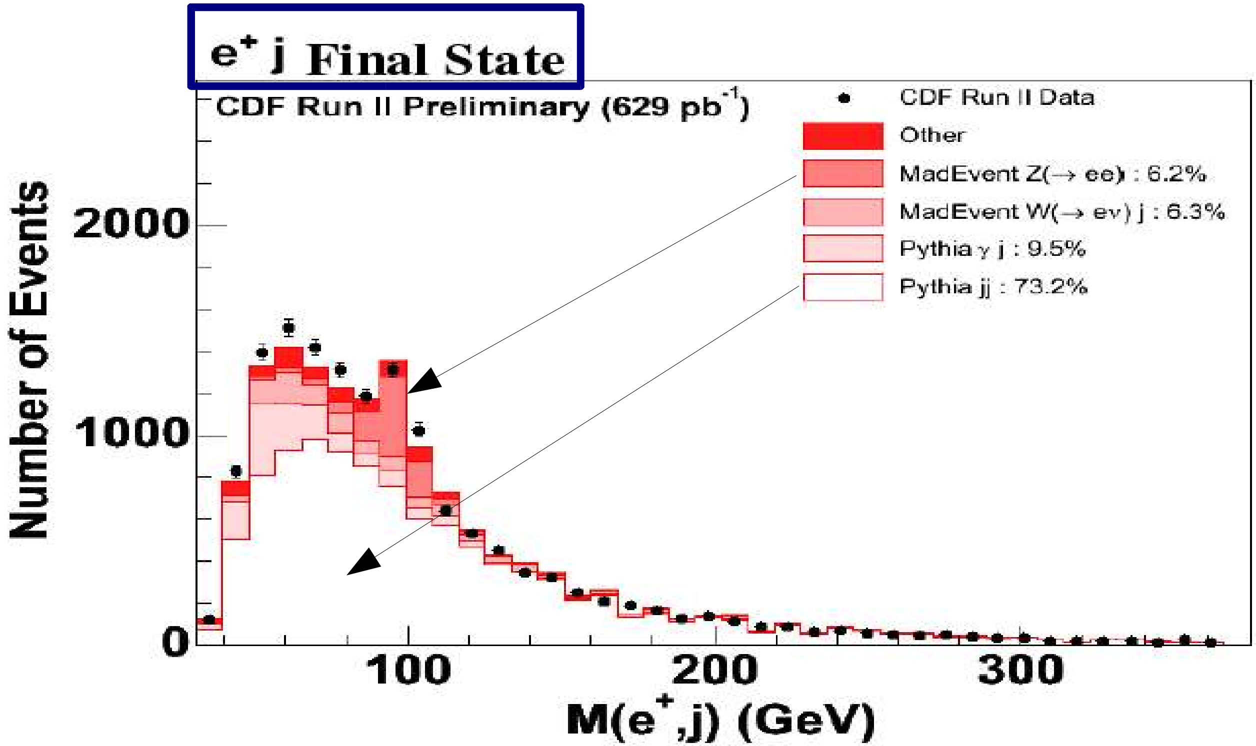


dominated by Z→ee



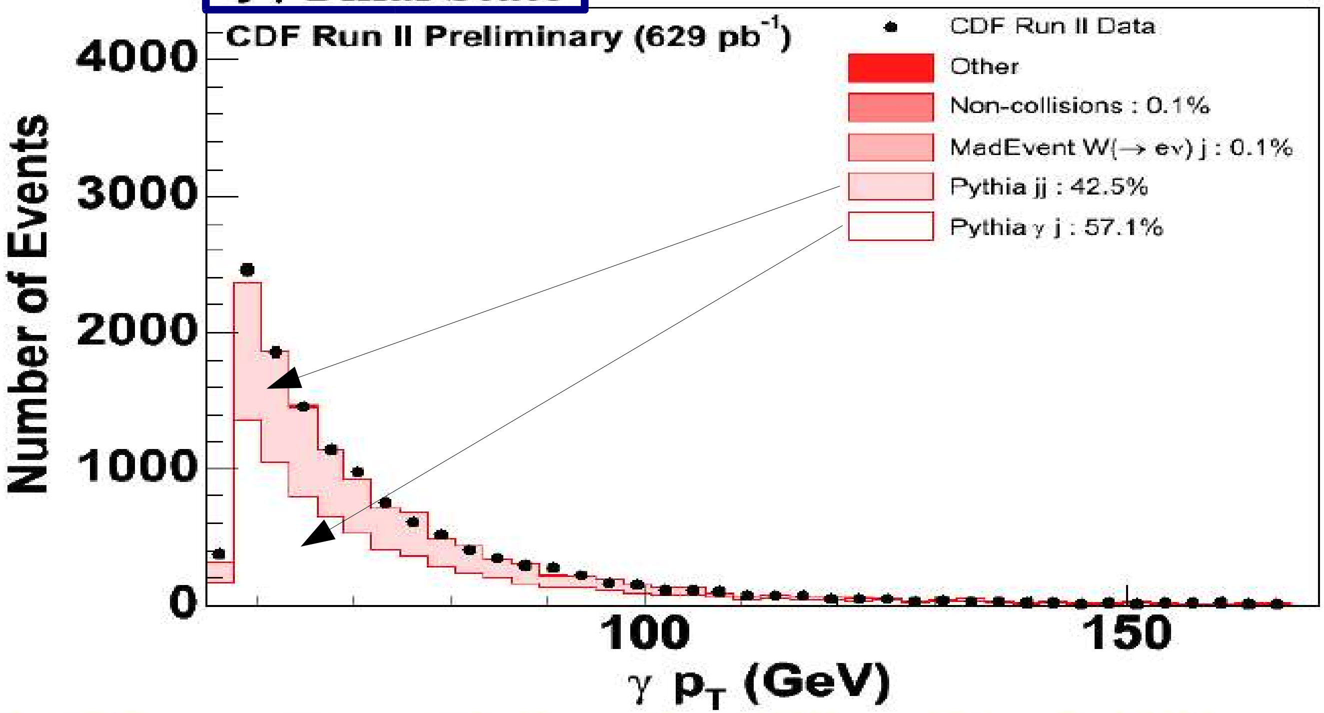
dominated by W→eν

Some final states are very powerful in determining fakes



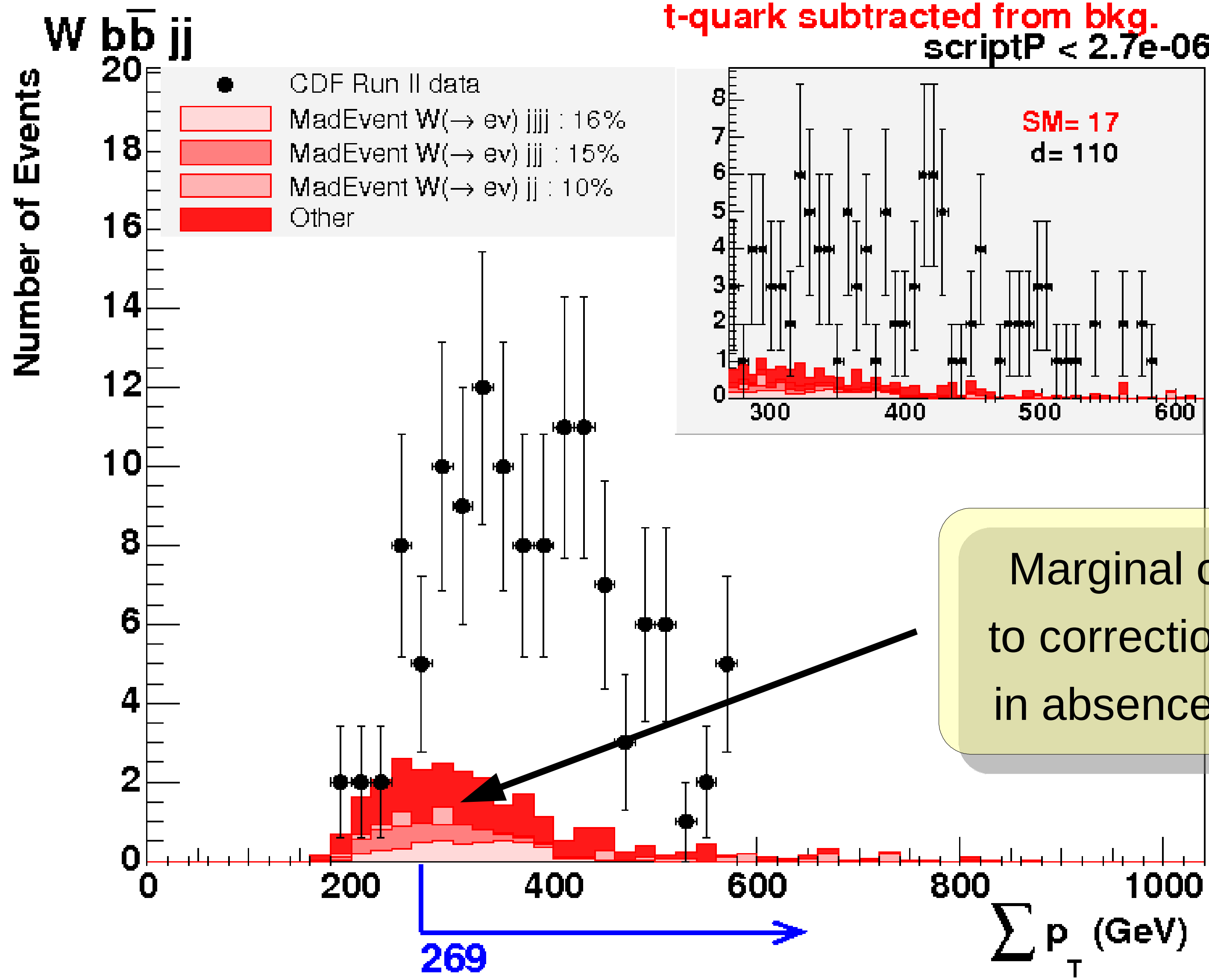
- ej final state dominated by jets reconstructed as electrons
- Also a peak at Z-mass, where electron is reconstructed as a jet

j γ Final State



- Real photon+jet production, plus di-jets with a jet faking a photon

Refitting after removing t-tbar



Marginal changes to correction factors in absence of t-tbar

Strengths

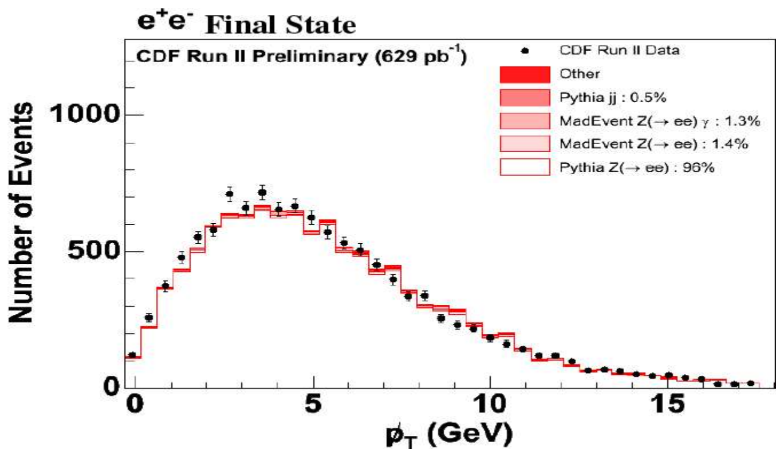
- Looks at many places
- Is not blind
- Fast
- Looks for both large and small effects
- Correction factors apply globally
- Uncertainties and correlations of correction factors

Limitations

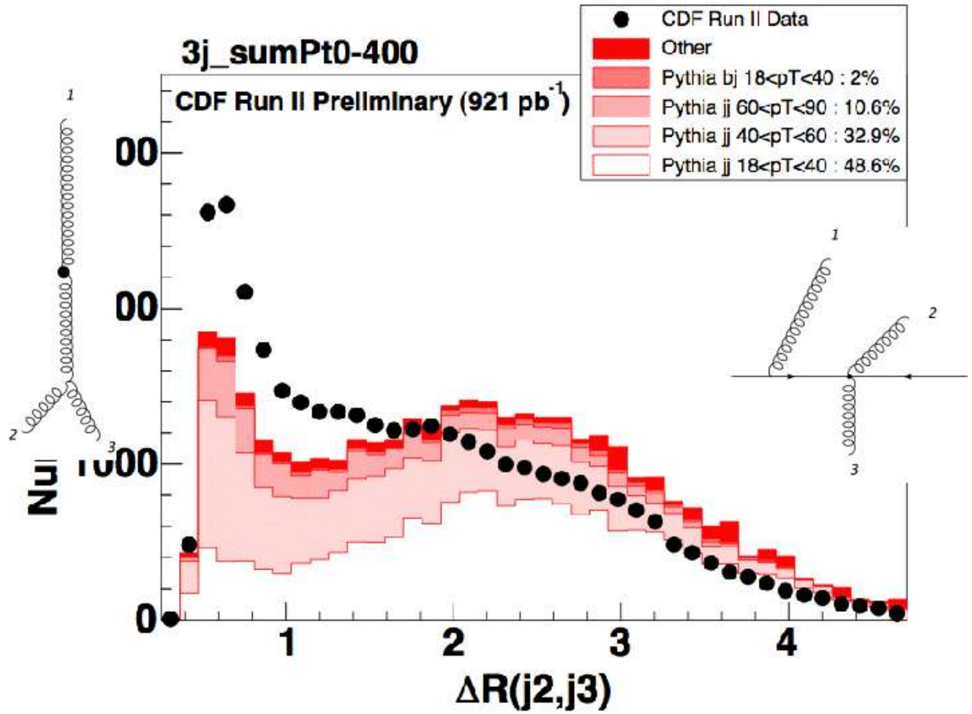
- Perfection in implementation is reached asymptotically
- Trials factor penalty
- Does not find C.L. regions.
- Stricter offline trigger reduces dataset
- Finite MC
- Some systematics are very hard to implement (CDFsim)

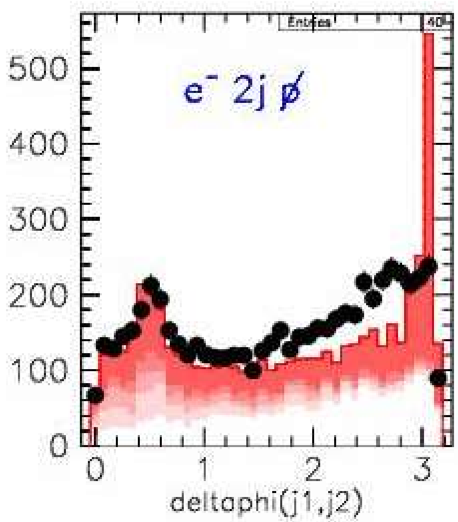
Successful "Debugging" of SM implementation

Large instrinsic kT



ISR too hard in Tune A





FSR for multiparton interactions

Questions raised by Vista spurred changes to RF Tunes

PYTHIA 6.2 Tunes

Use LO α_s with Λ = 192 MeV!

K-factor (Sjöstrand)

UE Parameters

ISR Parameter

Intrinsic KT

Parameter	Tune DW	Tune DWT	ATLAS	Tune QW	Tune QWT	Tune QK	Tune QKT
PDF	CTEQ5L	CTEQ5L	CTEQ5L	CTEQ6.1	CTEQ6.1	CTEQ6.1	CTEQ6.1
MSTP(2)	1	1	1	1	1	1	1
MSTP(33)	0	0	0	0	1	1	1
PARP(31)	1.0	1.0	1.0	1.0	1.0	1.8	1.8
MSTP(81)	1	1	1	1	1	1	1
MSTP(82)	4	4	4	4	4	4	4
PARP(82)	1.9 GeV	1.9409 GeV	1.8 GeV	1.1 GeV	1.1237 GeV	1.9 GeV	1.9409 GeV
PARP(83)	0.5	0.5	0.5	0.5	0.5	0.5	0.5
PARP(84)	0.4	0.4	0.5	0.4	0.4	0.4	0.4
PARP(85)	1.0	1.0	0.33	1.0	1.0	1.0	1.0
PARP(86)	1.0	1.0	0.66	1.0	1.0	1.0	1.0
PARP(89)	1.8 TeV	1.96 TeV	1.0 TeV	1.8 TeV	1.96 TeV	1.8 TeV	1.96 TeV
PARP(90)	0.25	0.16	0.16	0.25	0.16	0.25	0.16
PARP(62)	1.25	1.25	1.0	1.25	1.25	1.25	1.25
PARP(64)	0.2	0.2	1.0	0.2	0.2	0.2	0.2
PARP(67)	2.5	2.5	1.0	2.5	2.5	2.5	2.5
MSTP(91)	1	1	1	1	1	1	1
PARP(91)	2.1	2.1	1.0	2.1	2.1	2.1	2.1
PARP(93)	15.0	15.0	5.0	15.0	15.0	15.0	15.0

Issues to be considered

So far, the tuning procedure is not optimized.

Getting the bulk of the effect is often enough,
since correction factors can compensate

We have been lucky that the shape of MC
can be reproduced at the particle level

Our current stage of “debugging” is trying to
do “parameter scans” more intelligently

The Importance of Automation

Allows quick testing and feedback

Enables us to look at many channels at once

Want possibility to test hypotheses overnight or as soon as reasonably possible

Choice of objects, trigger thresholds, etc. reflect this.

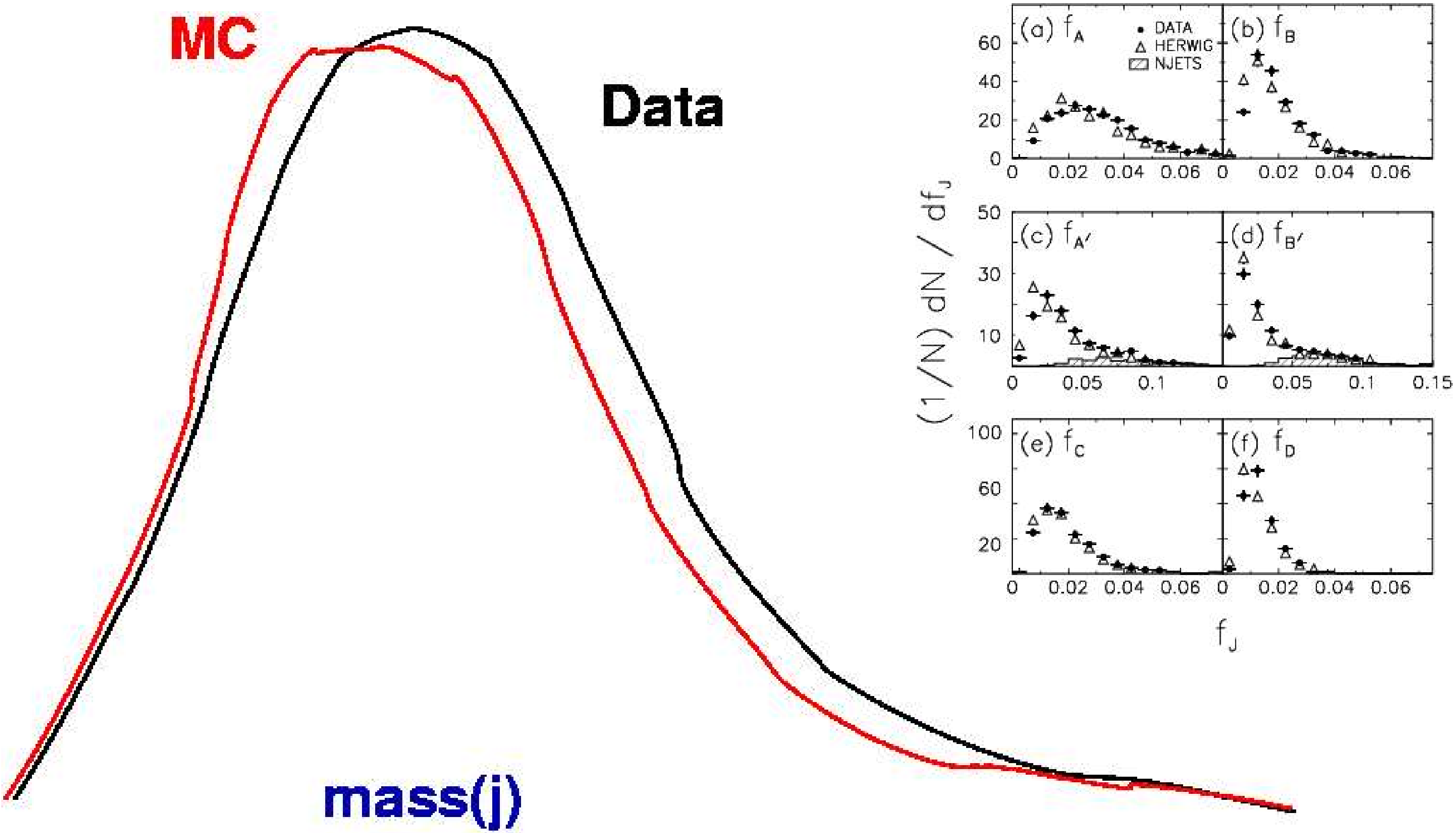
Aim at probing 1-2M "most interesting" high-pT events

Speed is an important consideration.

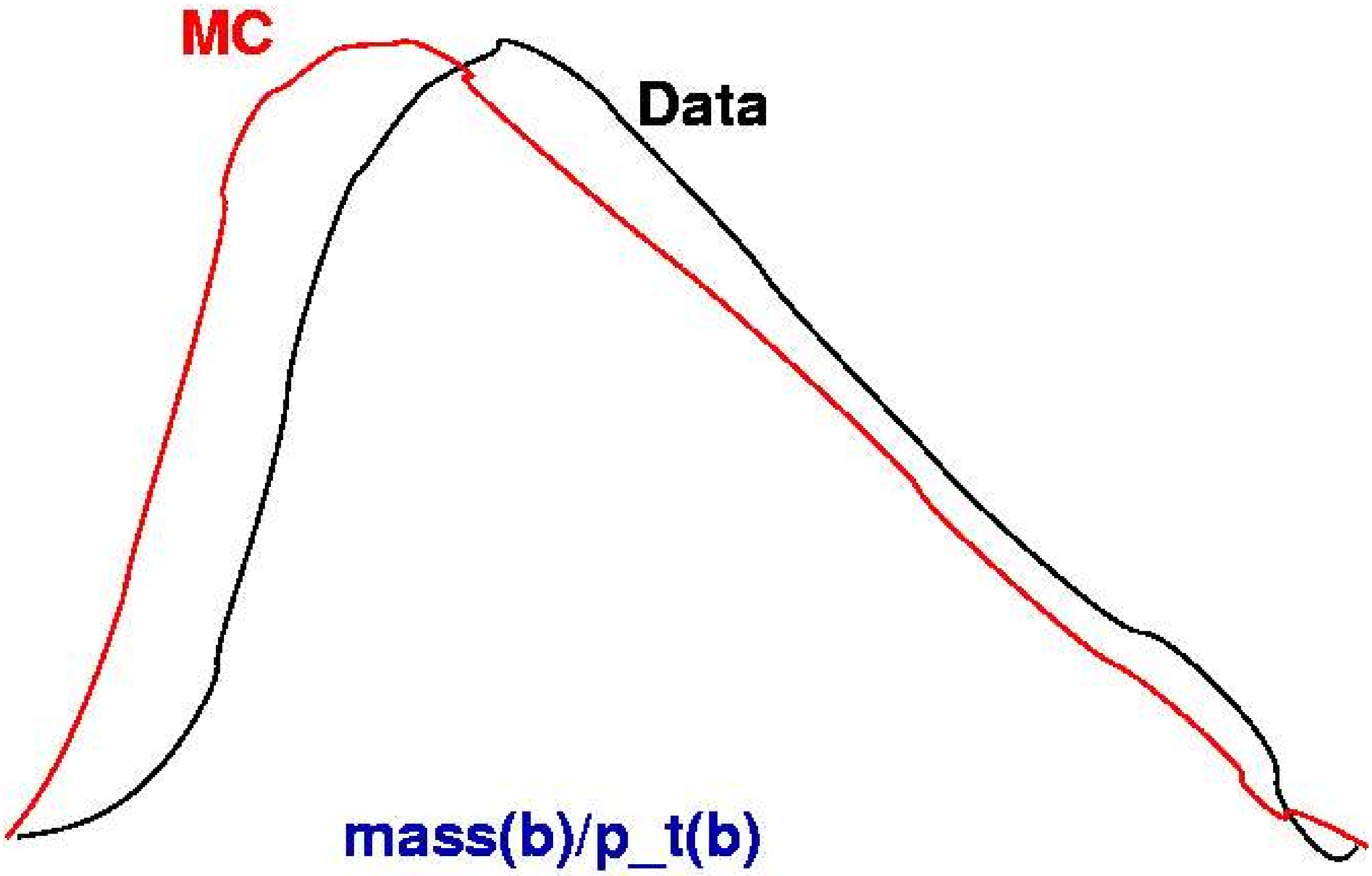
Want/need methods to quickly reweight "old" Monte Carlo and converge on a "right" answer

Other discrepant shapes

CDF Run1



Discrepant shapes



Parting Words

Vista is a methodology to answer the question: "Do we understand all the high-pT data?"

We have demonstrated that it is a useful part of our toolkit to understand the Standard Model

We got started "late" in Run2

We believe this should be done earlier as part of debugging the experiment

We should be prepared at the LHC start-up

- Getting ready for the 900 GeV run