Quaero

Making HEP Data Publicly Available

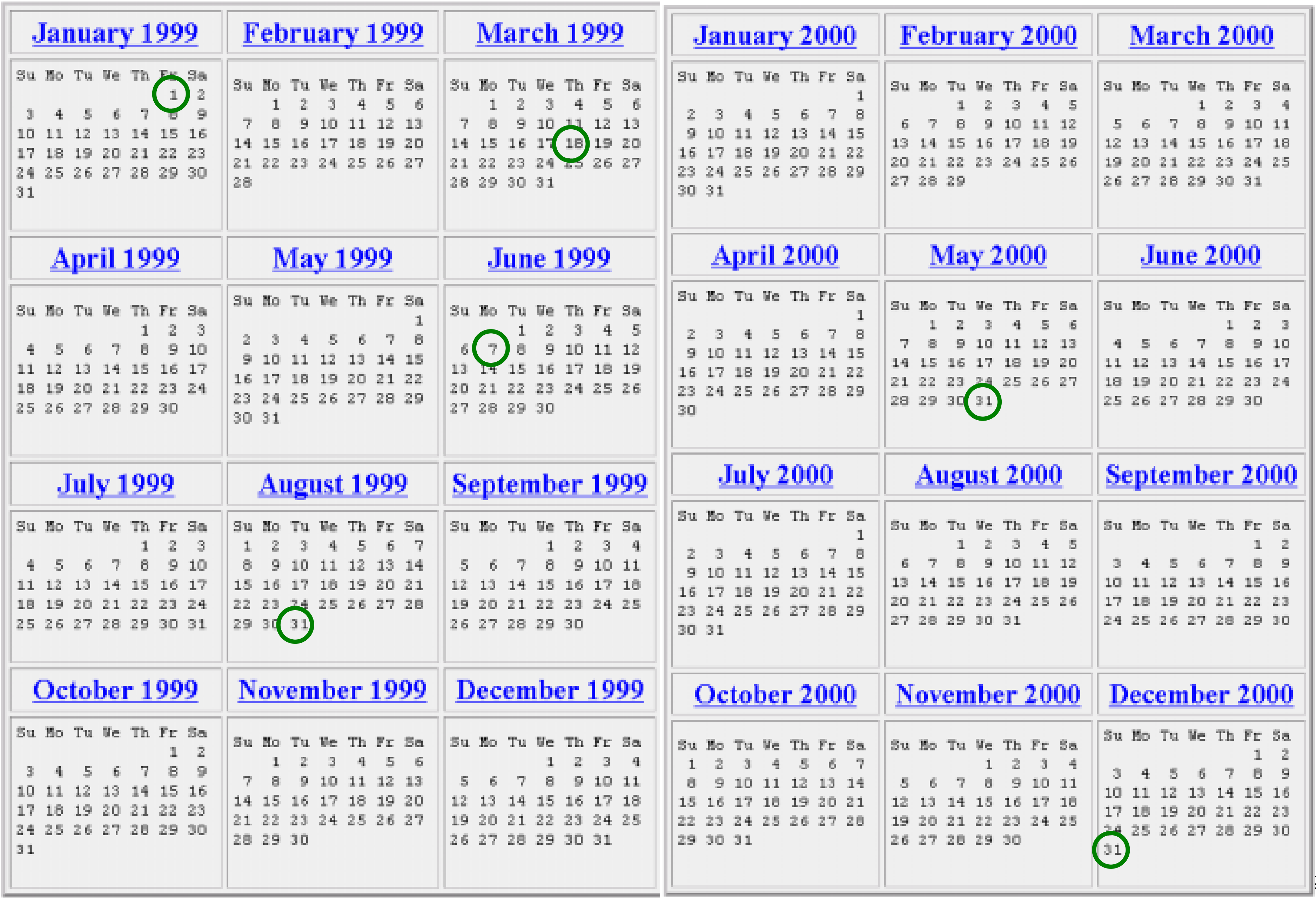
http://quaero.fnal.gov/

Motivation
Data
Algorithm
Examples

Bruce Knuteson
University of Chicago / CDF
UC Berkeley / LBNL / DØ

Motivation

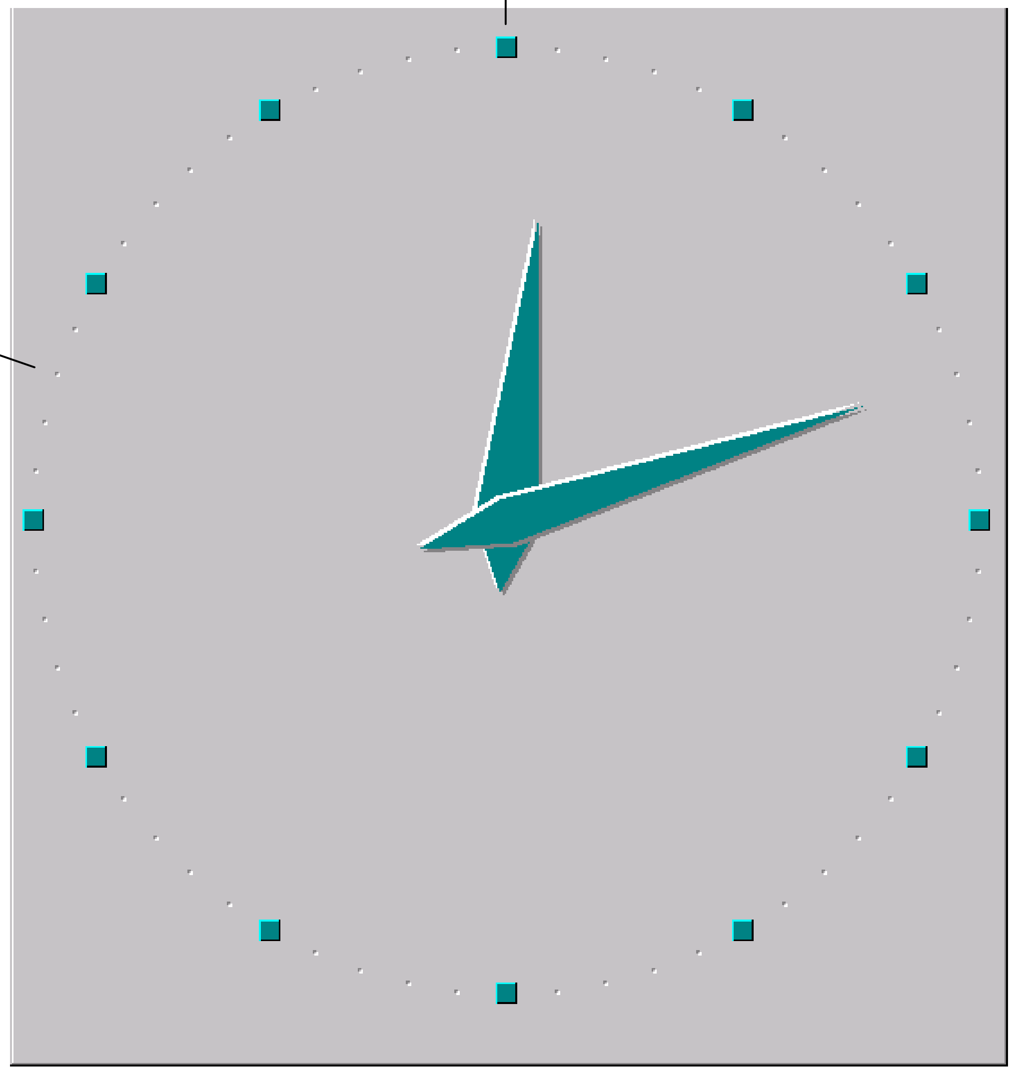
Time

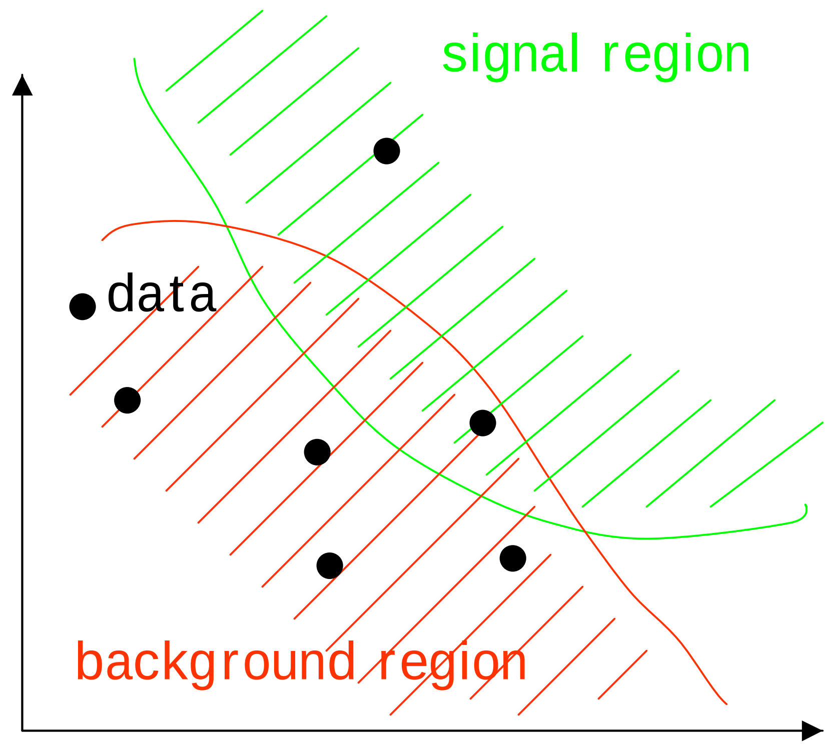


Time

Physicist has an idea

DØ data returns a verdict





How does one choose a set of cuts without bias?

WIRED NEWS

BUSINESS · POLITICS · WIRE SERVICE · CULTURE · TECHNOLOGY · TOP STORIES

LOOK FOR [ ] Wired News GO

Print this · E-mail it · Set E-mail Alerts

Secrets of the Atom Revealed

By Jeffrey Benner

2:00 a.m. July 27, 2001 PDT

You can find a lot of information on the Web, but you just couldn't find a decent picture of the subatomic universe online.

Until now.

Scientists at the Fermilab in Illinois, home to the world's most powerful atom smasher, announced Wednesday that data collected during the last big round of experiments into the depths of the atom is now available online.

Search for New Physics Using Quaero: A General Interface to DØ Event Data

DØ makes a subset of Tevatron Run I data publicly available

hep-ex/0106039

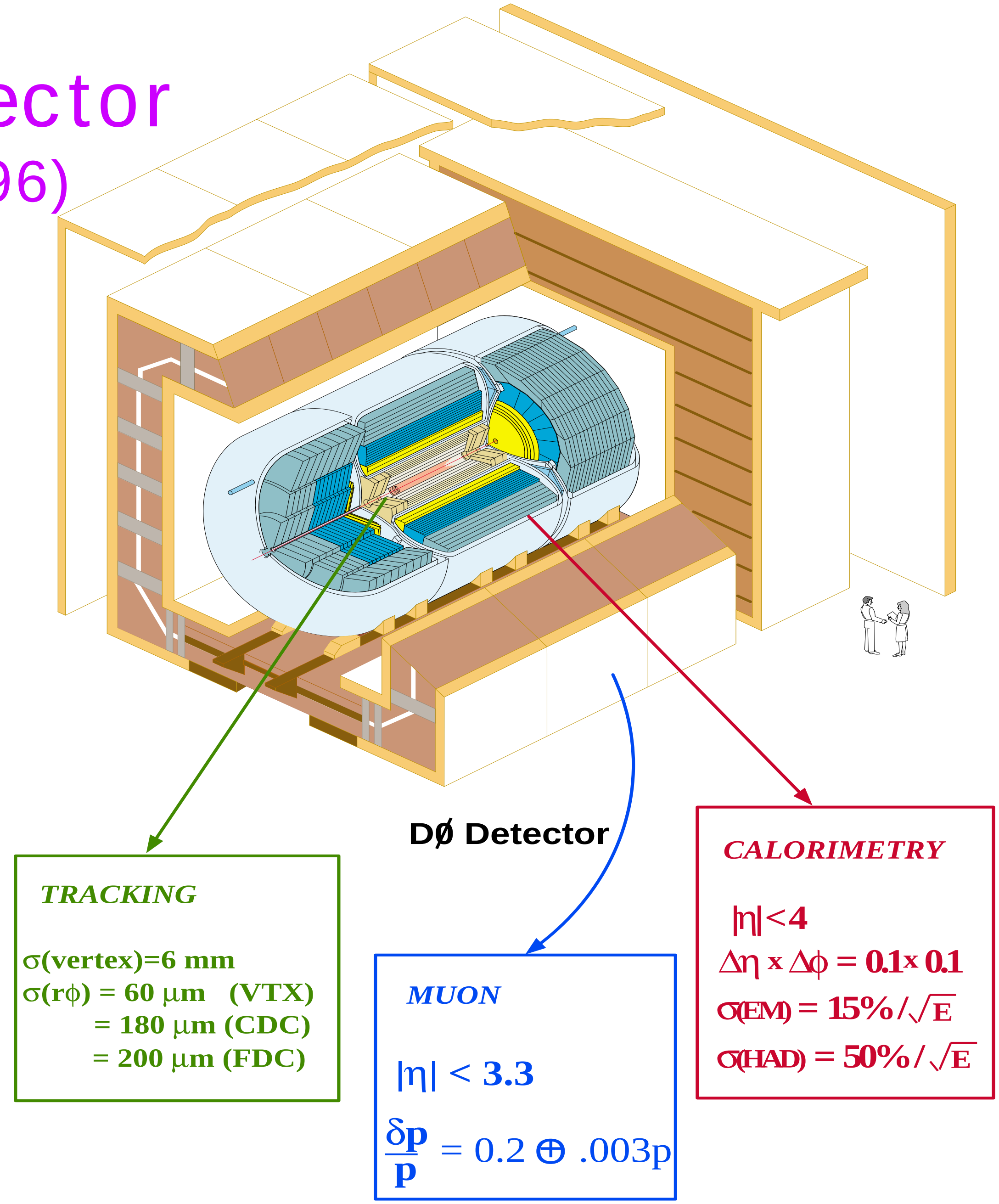
Search for New Physics Using QUAERO: A General Interface to DØ Event Data

(DØ Collaboration)


Motivation
Data
Algorithm
Examples

Run I detector
(1992-1996)

- Multipurpose detector
  - central tracking
  - muon spectrometer
  - U-LAr sampling calorimeter
- No central magnetic field
- Excellent electromagnetic and hadronic calorimeters



We have chosen 3 well-understood final states:

Final State	Selection criteria	ϵ_{ID}	$\int \mathcal{L}\, dt$
$e\mu$	$p_T^{e,\mu} > 15$ GeV	0.30	108 ± 5 pb^{-1}
	$\lvert\eta^{\mu}_{\mathrm{det}}\rvert < 1.7$		
$e\not{E}_T\, 2j$	$p_T^{e,j_{1,2}} > 20$ GeV	0.61	115 ± 6 pb^{-1}
	$\not{E}_T > 30$ GeV		
	$p_T^{e\not{E}_T} > 40$ GeV		
$ee\, 2j$	$p_T^{e_{1,2},j_{1,2}} > 20$ GeV	0.70	123 ± 7 pb^{-1}

ε_{ID} is the efficiency of identification requirements

additional jets are identified if $p_T^j > 15$ GeV and $|\eta| < 2.5$

Our background estimates come from a mishmash of sources:

	Standard model backgrounds				
Final State	multijets	W	Z	VV	$t\bar{t}$
$e\mu$	data	data	ISAJET	PYTHIA	HERWIG
$e\not{E}_T\ 2j$	data	VECBOS	–	PYTHIA	HERWIG
$ee\ 2j$	data	–	PYTHIA	PYTHIA	–

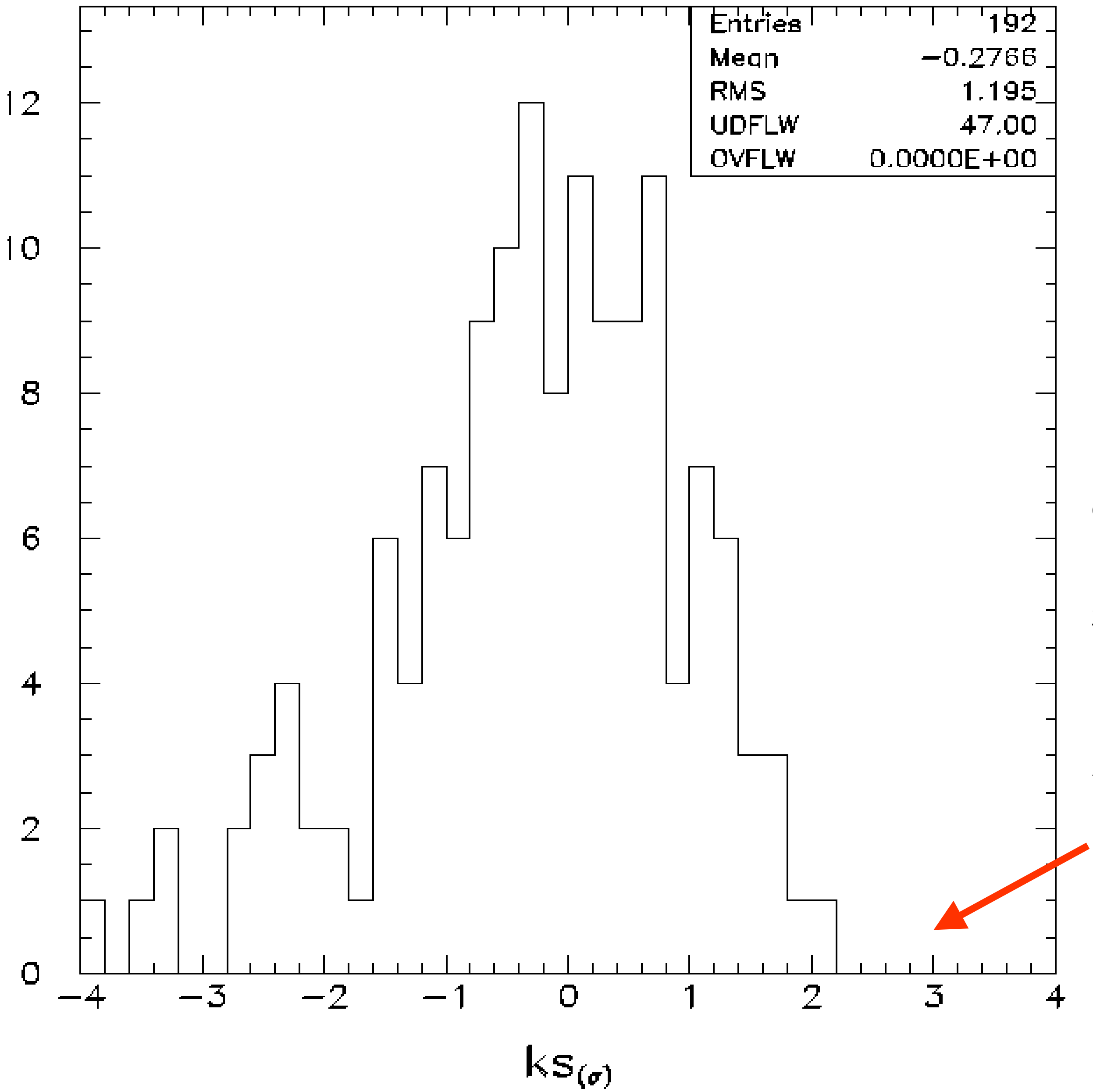
Systematic errors vary among the final states we consider, but roughly:

Systematic uncertainties

jet modeling	15%
trigger / lepton ID eff	10%
cross sections	10%
“faking” probabilities	10%
luminosity	5%

These uncertainties are incorporated into the cross section limit

Comparison of data and backgrounds



We analyzed 192 different distributions in these final states, including:

- transverse momenta
- invariant masses
- angular separation

and computed the Kolmogorov-Smirnov statistic for each.

Agreement is observed in all distributions

The data and background estimates have each been used in previous analyses:

- $e\mu$
  - Top quark cross section PRL 79 1203 (1997)
  - Scalar top PRL in collab. review
  - Sleuth PRD 62 92004 (2000)
- $e\not{E}_T$ 2j
  - Leptoquarks PRL 80 2051 (1998)
  - Sleuth PRL 86 3712 (2001) PRD 64 12004 (2001)
- ee 2j
  - Leptoquarks PRL 79 4321 (1997)
  - Sleuth PRL 86 3712 (2001) PRD 64 12004 (2001)

Data

These data and backgrounds are well understood

Motivation

Data

Algorithm

Examples

Quaero

A General Interface to DØ Data

PRL Manual
Examples

Signal

Final State: e mu (met) (nj) Smear? ☑

○ Pythia Input:

○ Signal File: Browse... xsec: pb

○ View ○ Search

Backgrounds: ☑ (nj) ☑ W(nj) ☑ Z(nj) ☑ VV(nj) ☑ tt(nj)

Variables

Constraints:

Variables:

v1

v2

v3

Requestor

Name:

Institution:

Email:

Brief description of model:

Submit

Help! - Bug history - DØ - Fermilab - Author

1) The signal Monte Carlo is processed

- (events are generated using Pythia, if requested)
- events are smeared with a fast detector simulation
- selection criteria are applied for desired final state
- particle identification efficiencies are considered

This gives

- total number of expected signal events in final state
- Monte Carlo signal events as they would look in the detector

The parametrized detector simulation is simple but sufficient:

- Partons are clustered and merged into jets
- Energies are smeared according to measured resolutions

e/γ: $\delta E/E = 15\%/\sqrt{E} \oplus 0.3\%$

jets: $\delta E/E = 80\%/\sqrt{E}$

μ: $\delta(1/p) = 0.18(p-2)/p^2 \oplus 0.003$

(Resolutions depend loosely on pseudorapidity)

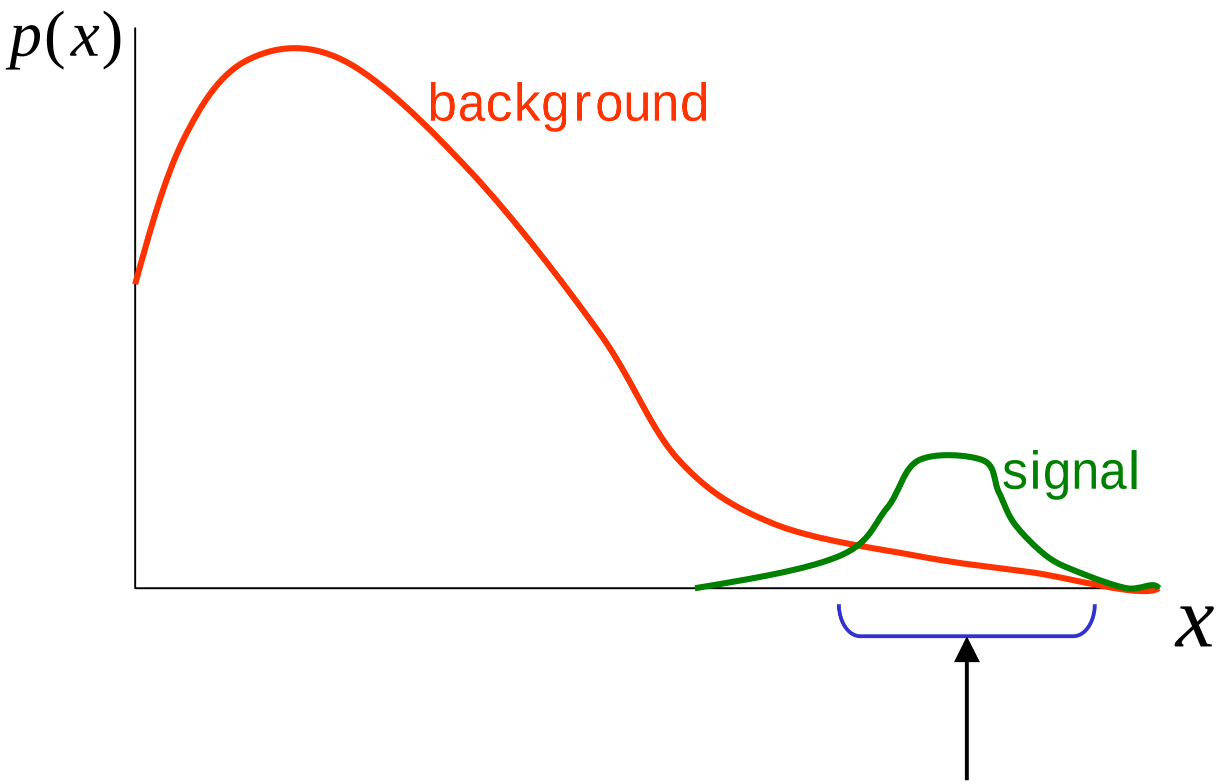
Why do we believe the fast detector simulation?

PRL 75 1028 (1995)	PRL 75 1034 (1995)	PRL 78 3634 (1997)
PRL 78 3640 (1997)	PRD 56 6742 (1997)	PRL 79 1441 (1997)
PRD 57 3817 (1998)	PRD 58 051101 (1998)	PRD 58 31102 (1998)
PRD 60 72002 (1999)	PRD 62 52005 (2000)	PRD 62 71701 (2000)
PRL 75 1456 (1995)	PRD 60 52003 (1999)	PRD 61 72001 (2000)
PRD 61 32004 (2000)	PRL 84 2792 (2000)	PRL 77 3309 (1996)
PRD 58 12002 (1998)	PRL 80 3008 (1998)	PRD 58 92003 (1998)
PRL 80 5498 (1998)	PRD 62 92006 (2000)	PRL 84 222 (2000)
PRL 81 524 (1998)	PRL 86 1156 (2001)	. . .

These publications cover a variety of analyses

- gauge boson couplings
- W/Z production
- searches for new physics

Note also the difference between signal and background:

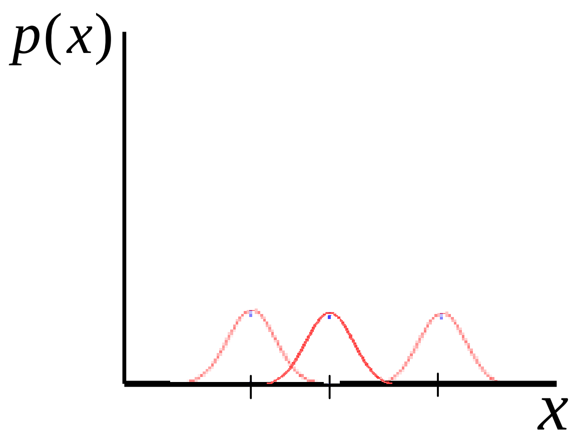


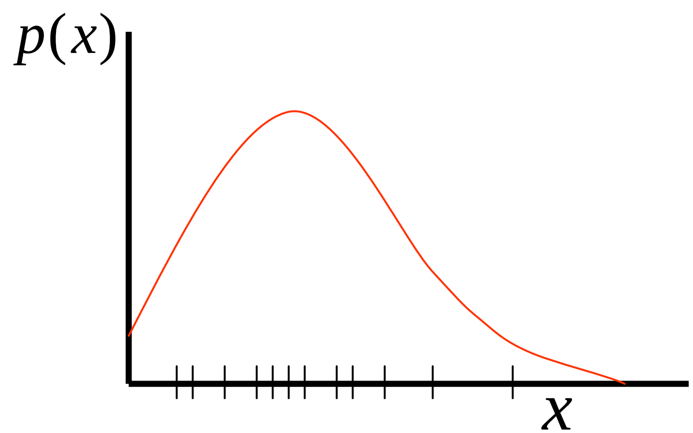
The region that we are interested in generally lies in the *tail* of the background, but in the *bulk* of the signal

(Correct modeling of the background tails is relatively more important than correct modeling of the signal tails)

2) An optimal region is chosen in the variables provided

a) Estimate signal and background densities using kernels





1) place “bumps of probability” around each Monte Carlo point

2) sum these bumps into a continuous distribution

$$p(x) = \sum_{i=1}^{N} \text{gauss}(x - x_i)$$

The multivariate generalization is immediate

b) Define a *discriminant*

$$D(x) = \frac{p(x|s)}{p(x|s) + p(x|b)}$$

and choose a cut on D(x) that minimizes

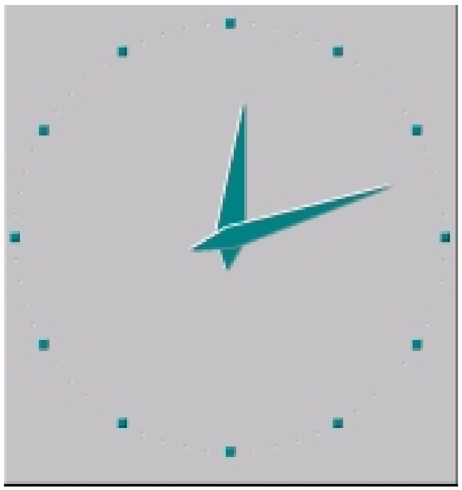
the 95% CL cross section limit you would expect to set assuming the data contains no signal.

We call 1/this quantity the "sensitivity"

Note that so far we have made no use of the data

3) Comparing number of observed events in the data to expected bkg, set 95% CL cross section limit on signal

4) Result is returned by email



Total elapsed time ≈ 1 hour

From: quaero@fnal.gov
Subject: Quaero Request #29

$W_R \rightarrow t\bar{b} \rightarrow e\not{E}_T\ 2j$

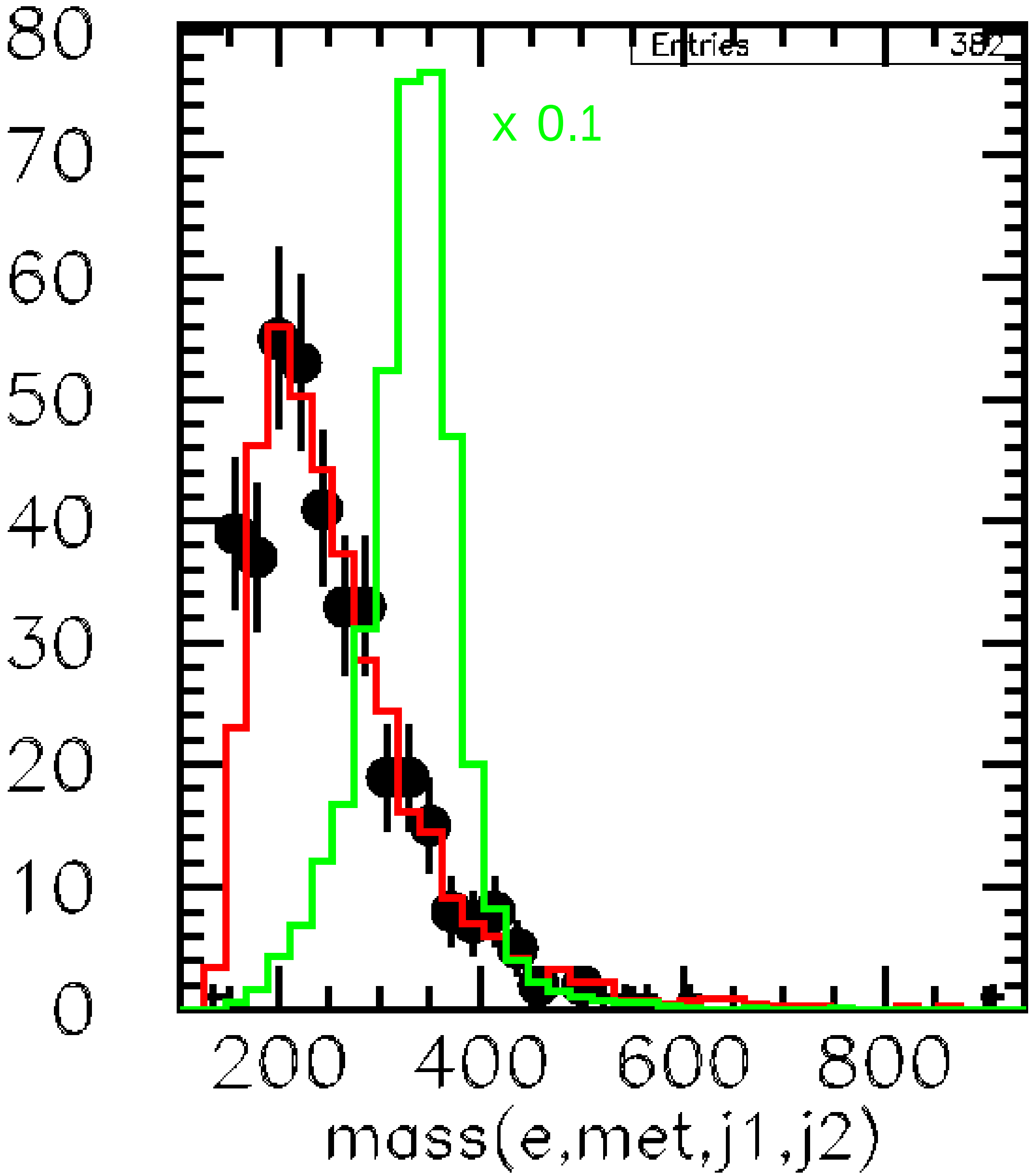
Result

Pythia cross section x branching ratio = 1.68 pb.

Upper limits on the cross section to this process at confidence levels of 50%, 90%, and 95% are found to be 0.8 pb, 1.8 pb, and 2.1 pb, respectively. Maximal sensitivity (0.73 pb^{-1}) is achieved in a region of variable space with 17.6 signal events expected, 32.7 +- 7.1 background events expected, and 36 events observed in the data.

Plots

Plots of the variables that you used are available for viewing at http://quaero.fnal.gov/quaero/requests/plots/29.ps. The red curve is the expected background; the green curve is your signal multiplied by a factor of 10; the black dots are D0 data.


80
70
60
50
40
30
20
10
0
Entries
382
x 0.1
200
400
600
800
mass(e,met,j1,j2)

There are a number of ways Quaero could be implemented

Don't
- Keep Quaero as an internal tool

Make data available with limited scope and internal review
- Restrict those who are allowed to use Quaero
- Review all Quaero results before releasing them

Make data available with general scope and more limited internal review

"Put the data out there"
- Make data available to all with no internal review

Many variations on these themes

We chose to make data available with general scope and limited internal review

Quaero Policy

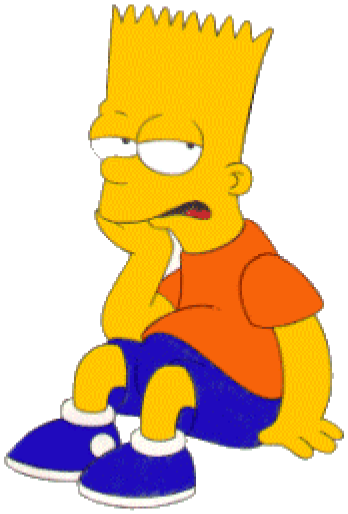
- Any "interesting" Quaero result will be reviewed by a DØ Quaero Review Board.
  - A Quaero result is "interesting" if an excess of data over background of more than 2.0 standard deviations is found.
  - If an interesting result is found, the requestor is notified that his request is under review, and the result of the request is sent to the review board.
  - If a fault is found the fault is rectified, the request is re-run, and the new result is sent to the requester (along with an explanation).
  - If the "interestingness" is not deemed to be due to any fault, the result is sent to the requester.
- In all cases the requester is free to publish the Quaero result in his or her own paper, so long as Quaero is referenced. The appropriate citation, including the Quaero request log number and request date, is included in the email with Quaero's result.

Motivation
Data
Algorithm
Examples

Examples

You're presumably thinking

"Surely this can't be as good as a real analysis . . ."

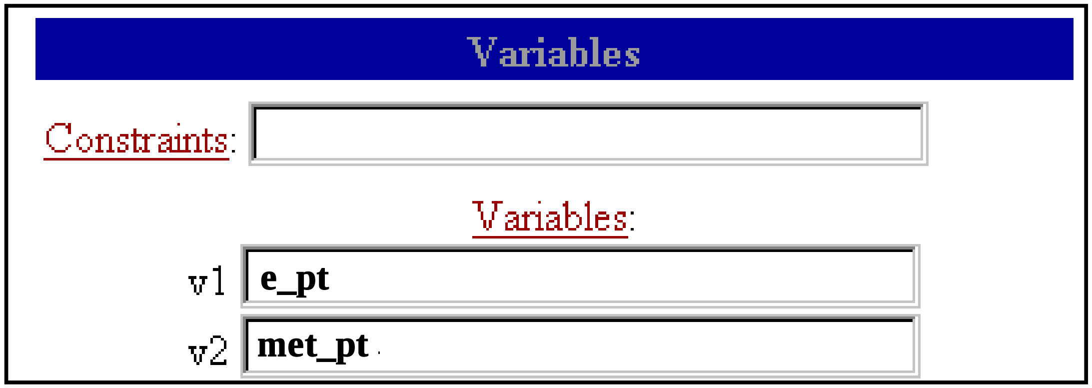


If so, think again!

Using Quaero, you can analyze the following in a day:

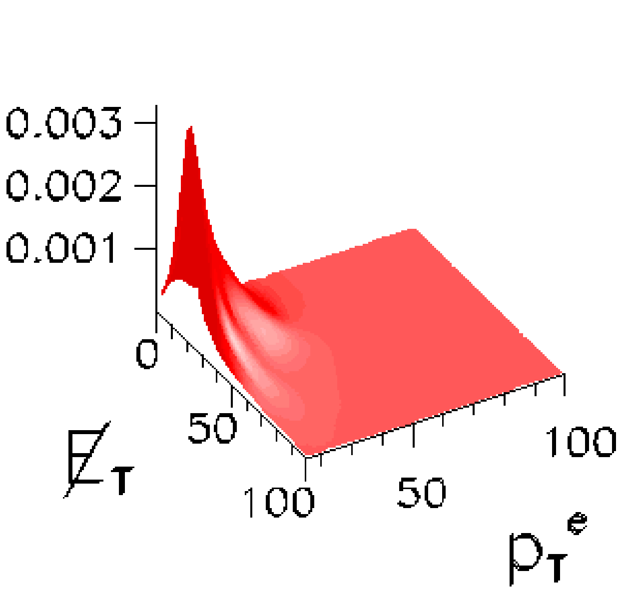
Examples #1

Standard Model $WW \rightarrow e\mu \not{E}_T$

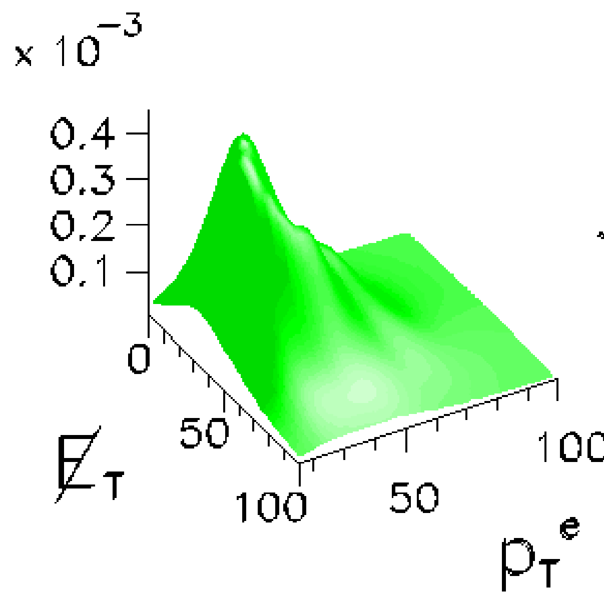


ε_{sig}	14%
$\hat{b}$	19.0 ± 4.0
N_{obs}	23
$\sigma^{95\%} \times \mathcal{B}$	1.1 pb

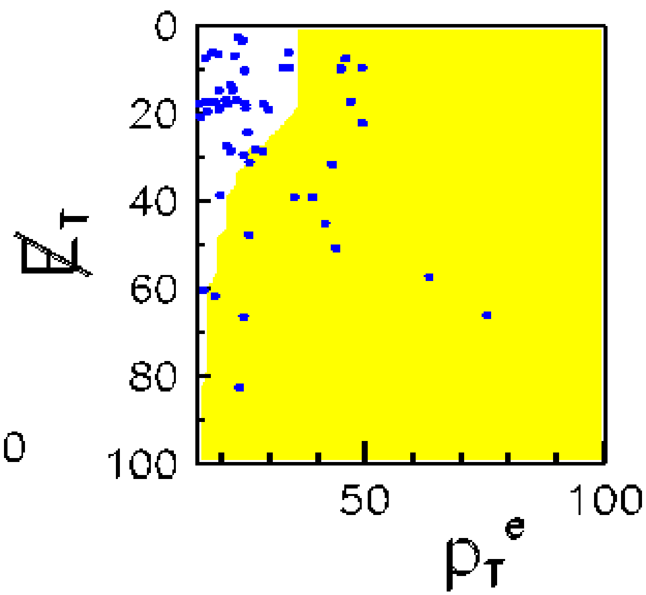
Background density



Signal density

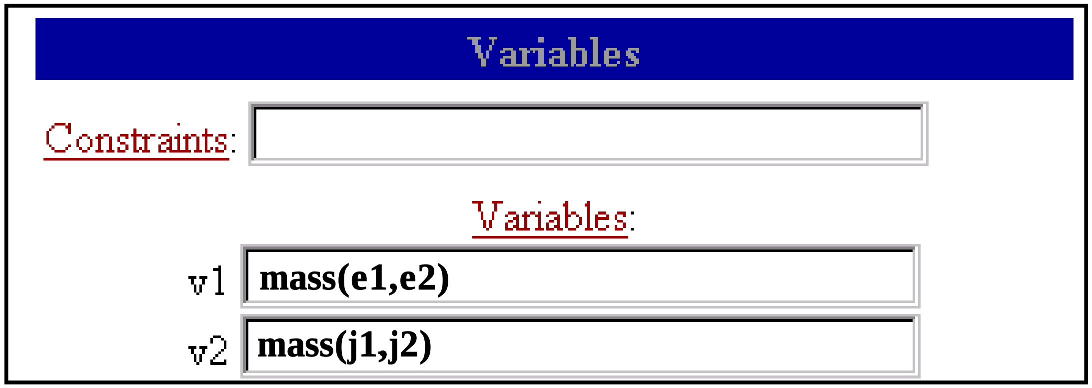


Selected region



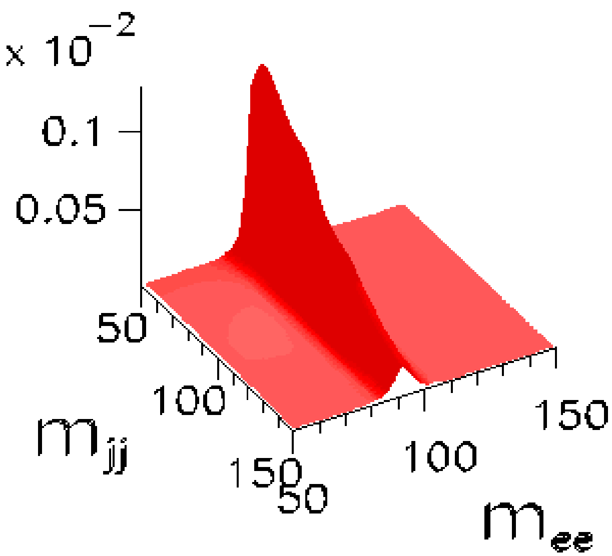
Examples #2

Standard Model ZZ→eejj

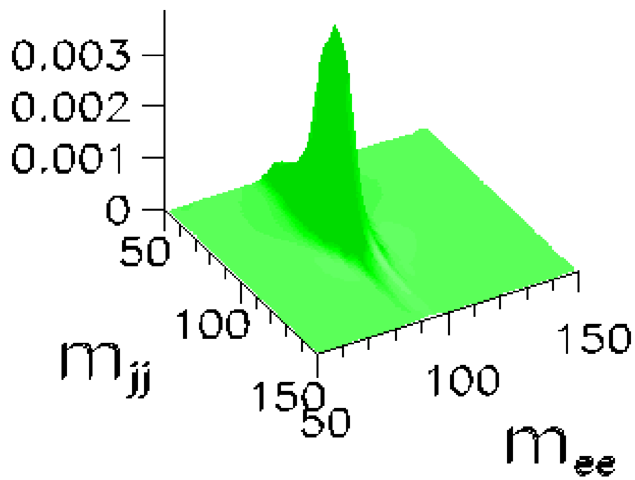


ε_{sig}	12%
$\hat{b}$	19.7 ± 4.1
N_{obs}	19
$\sigma^{95\%} \times \mathcal{B}$	0.8 pb

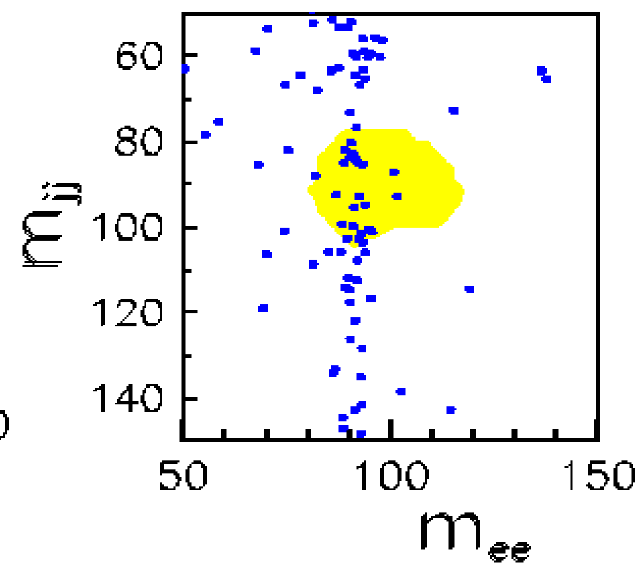
Background density



Signal density



Selected region



Standard Model WW $\rightarrow e\mu\not{E}_T$

Predicted $\sigma \times \mathcal{B}$ = 0.17 pb

Limit $\sigma^{95\%} \times \mathcal{B}$ = 1.1 pb

Standard Model ZZ $\rightarrow$ ee 2j

Predicted $\sigma \times \mathcal{B}$ = 0.05 pb

Limit $\sigma^{95\%} \times \mathcal{B}$ = 0.8 pb

These processes are currently at the edge of the Tevatron's sensitivity

CDF, "Observation of W^+W^- production . . . ." PRL 78 4536 (1997)

5 events observed on a background of 1.2 ± 0.3 events
in all dilepton final states (ee, eμ, μμ)

Examples #3

Standard Model $t\bar{t} \to e\mu \not{E}_T$ 2j

Variables

Constraints: j2_pt>15

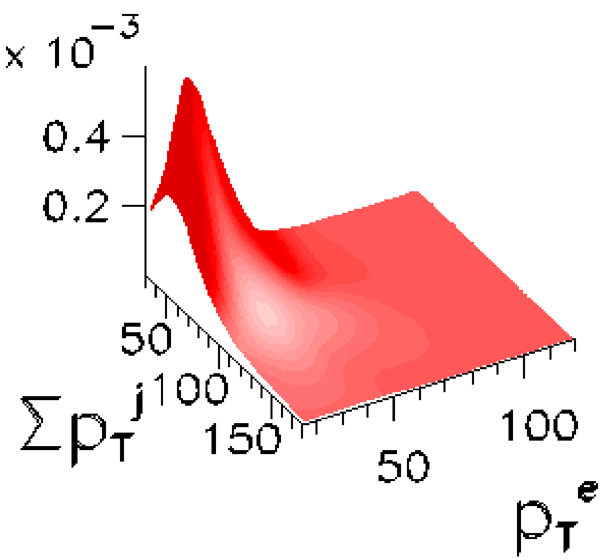
Variables:

v1 e_pt

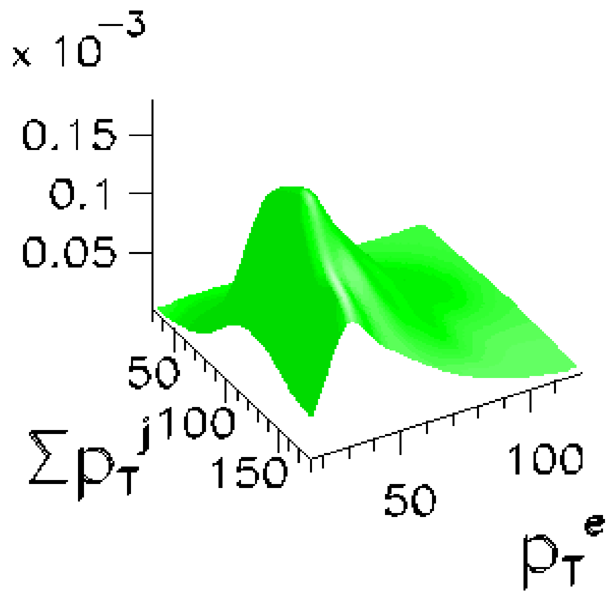
v2 j1_pt + j2_pt + j3_pt

ε_{sig}	14%
$\hat{b}$	0.6 ± 0.2
N_{obs}	2
$\sigma \times \mathcal{B}$	$0.14^{+0.15}_{-0.08}$ pb

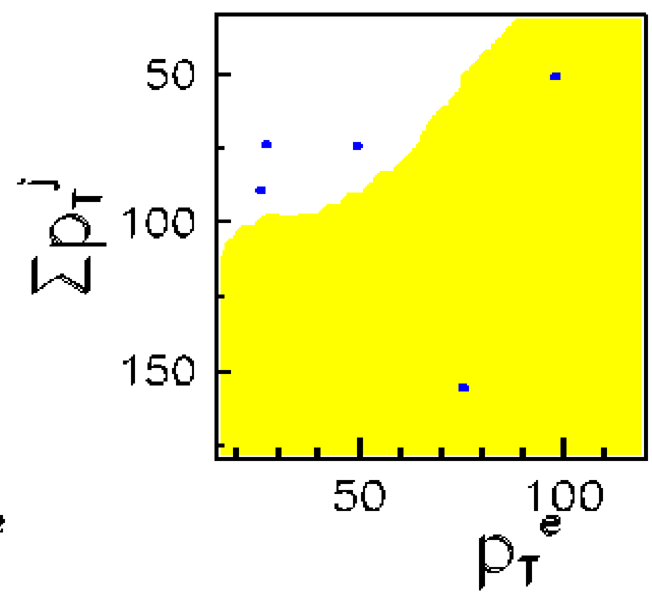
Background density



Signal density



Selected region



Examples #4

Standard Model $t\bar{t} \rightarrow e\not{E}_T\ 4j$

Variables

Constraints: j4_pt>15

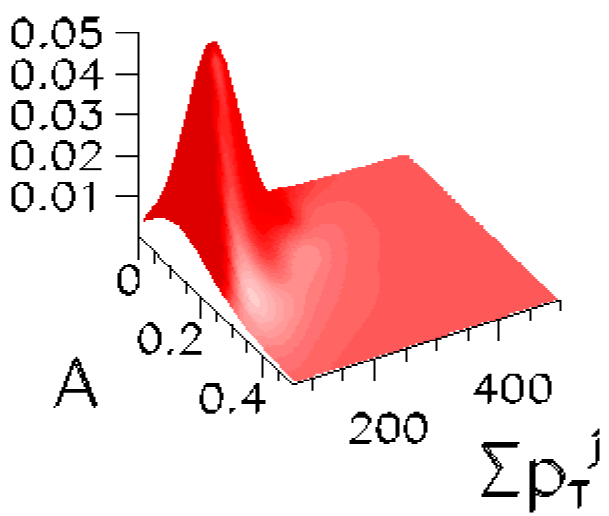
Variables:

v1 j1_pt + j2_pt + j3_pt + j4_pt + j5_pt

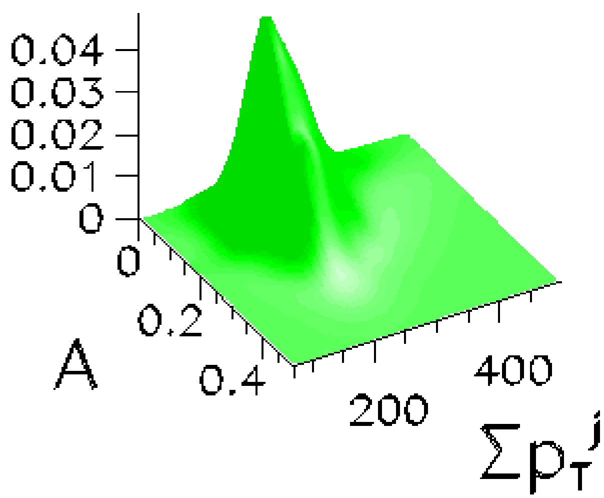
v2 aplanarity()

ε_{sig}	13%
$\hat{b}$	3.1 ± 0.9
N_{obs}	8
$\sigma \times \mathcal{B}$	$0.39^{+0.21}_{-0.19}$ pb

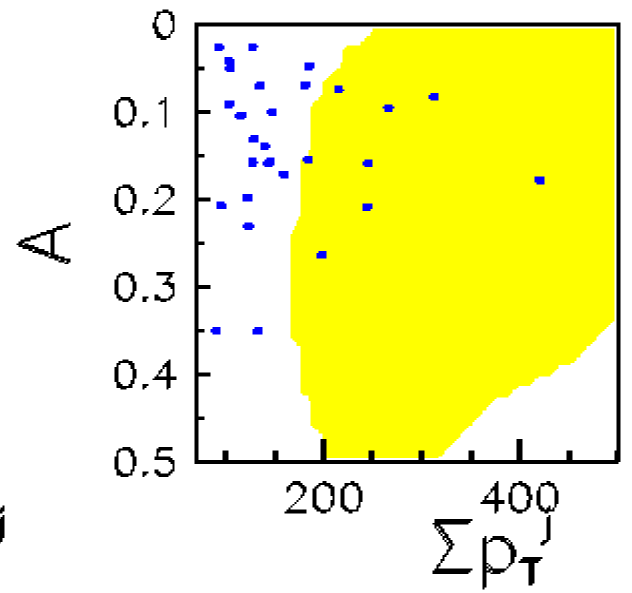
Background density



Signal density



Selected region



DØ PRL (1997, 125 pb^{-1})

in dileptons (ee, eμ, and μμ):

2.1σ (5 events with 1.4 expected)

in lepton+jets (e+jets and μ+jets):

2.6σ (19 events with 8.7 expected) w/o b-tag

3.6σ (11 events with 2.5 expected) w/ b-tag

We have used

- ½ of the dilepton sample (eμ) only
- ½ of the lepton+jets sample (e+jets) only
- no b-tagging

$\sigma \times \mathcal{B}$ is consistent with previous measurements

Examples #5

SM Higgs $\rightarrow$ WW $\rightarrow$ e$\not{E}_T$ 2j

Variables

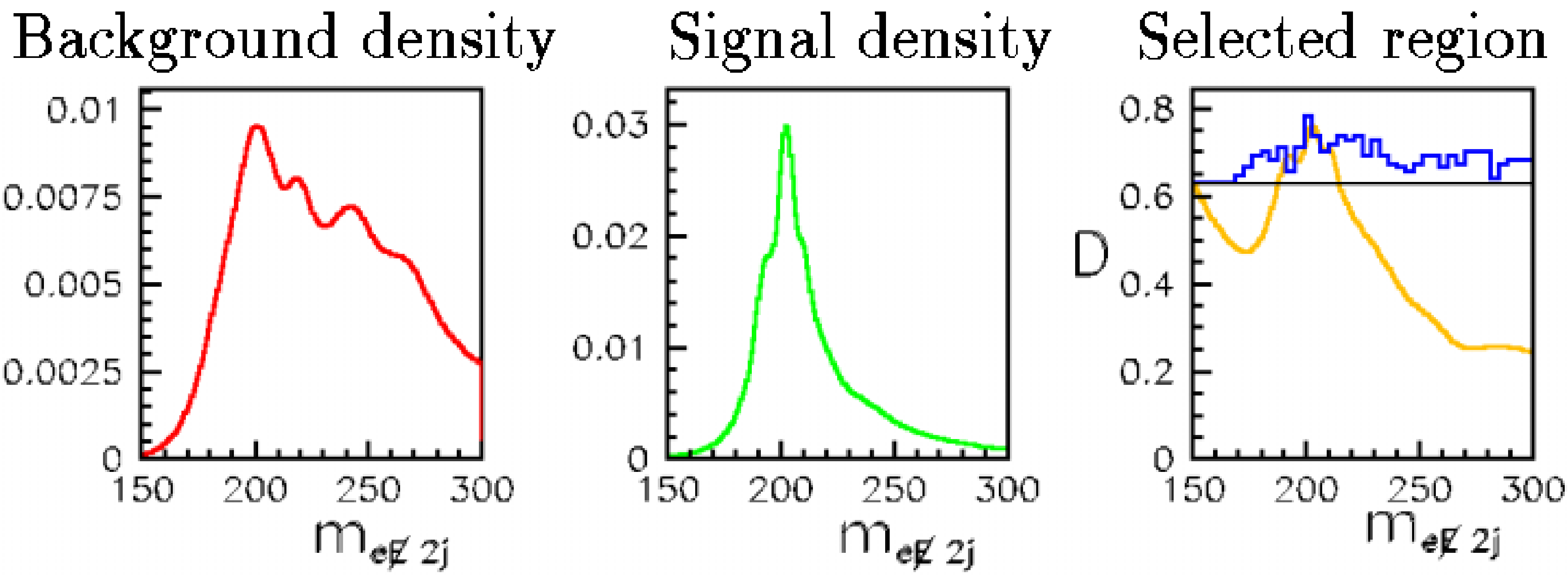
Constraints: **ChiSqdConstrainWW(e,met,j1,j2)<100**

Variables:

v1 **mass(e,met,j1,j2)**

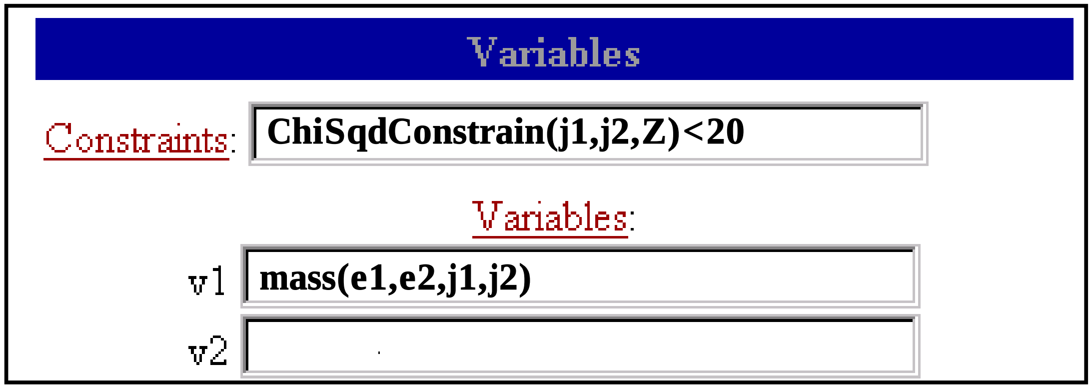
v2

ε_{sig}	7%
$\hat{b}$	66.0 ± 13.8
N_{obs}	69
$\sigma^{95\%} \times \mathcal{B}$	4.4 pb



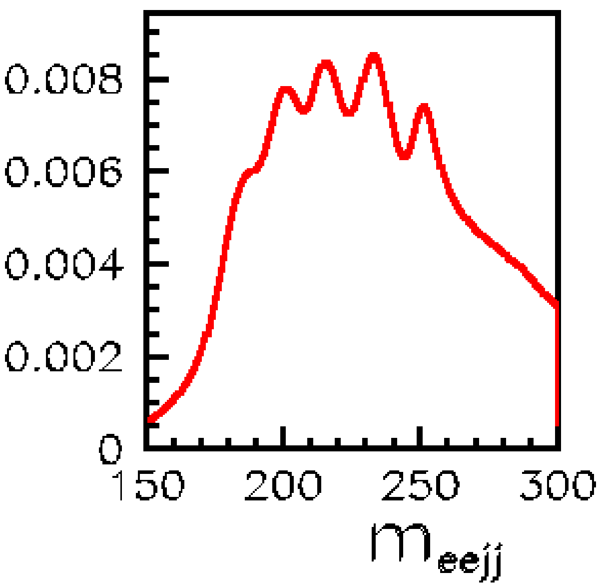
Examples #6

SM Higgs $\rightarrow$ ZZ $\rightarrow$ ee 2j

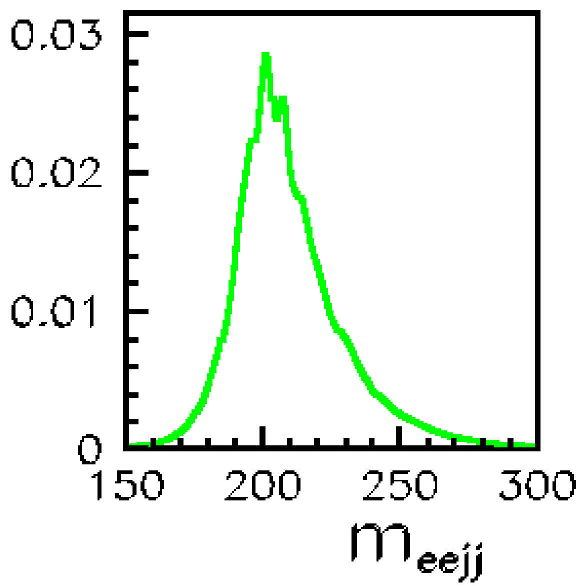


ε_{sig}	15%
$\hat{b}$	17.9 ± 3.7
N_{obs}	15
$\sigma^{95\%} \times \mathcal{B}$	0.6 pb

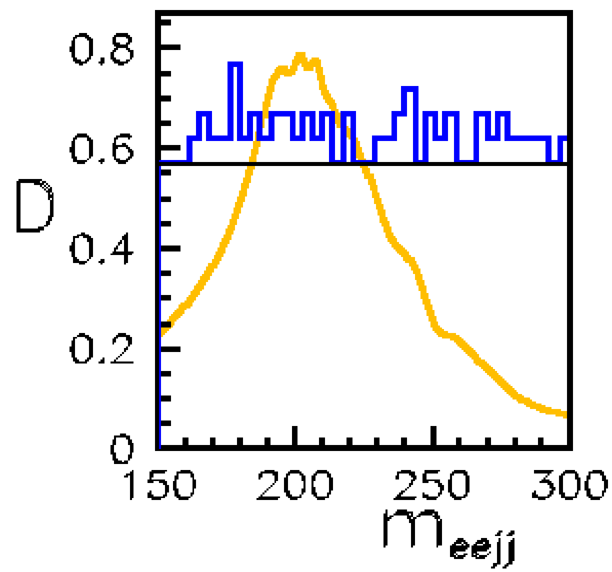
Background density



Signal density

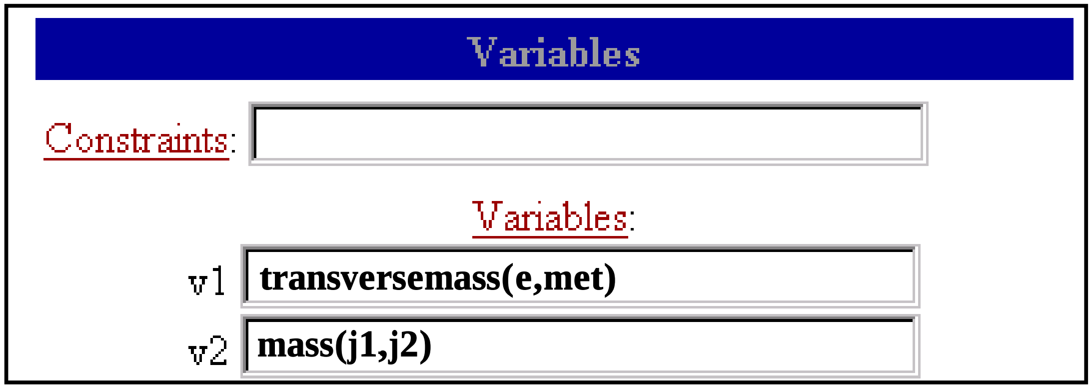


Selected region



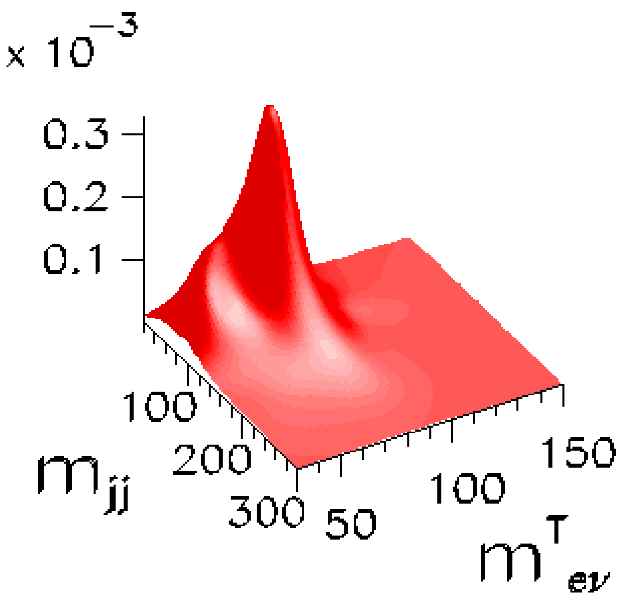
Examples #7

$W^* \to Wh \to e\not{E}_T\ 2j$

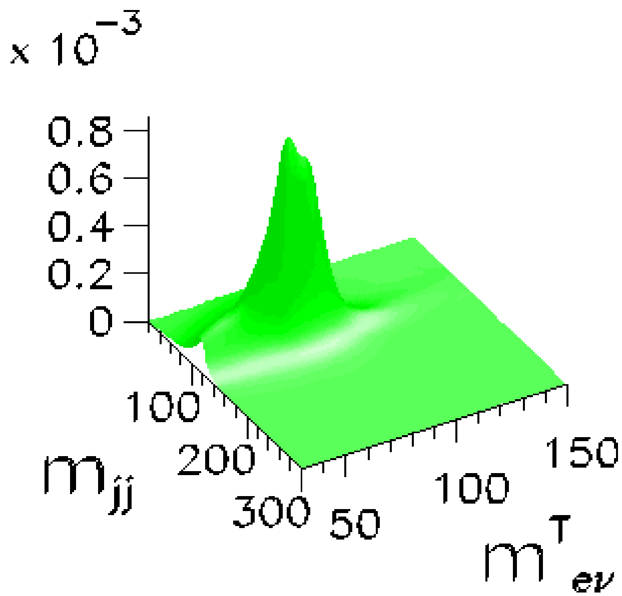


ε_{sig}	8%
$\hat{b}$	37.3 ± 8.2
N_{obs}	32
$\sigma^{95\%} \times \mathcal{B}$	2.0 pb

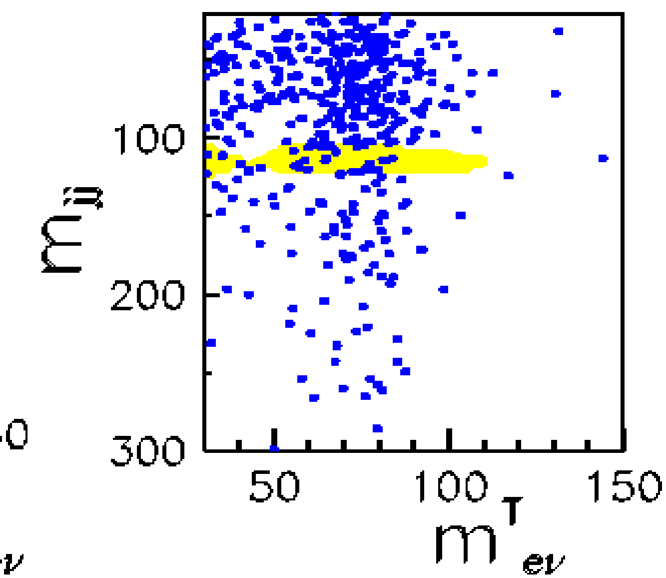
Background density



Signal density

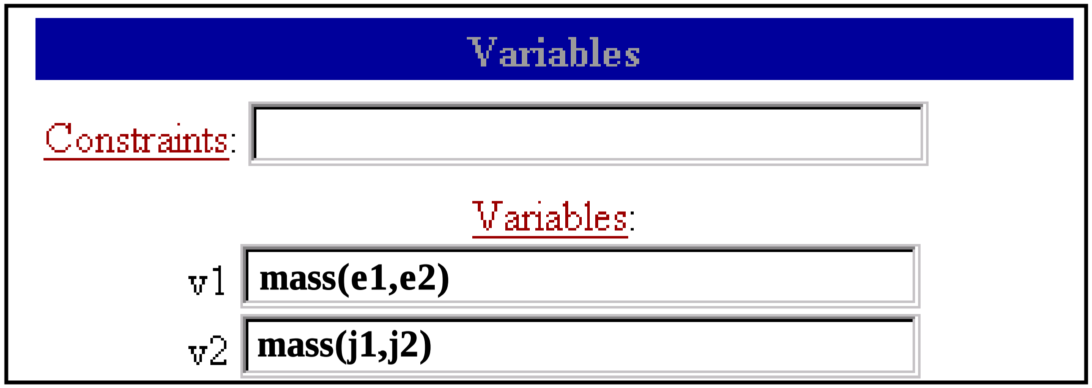


Selected region



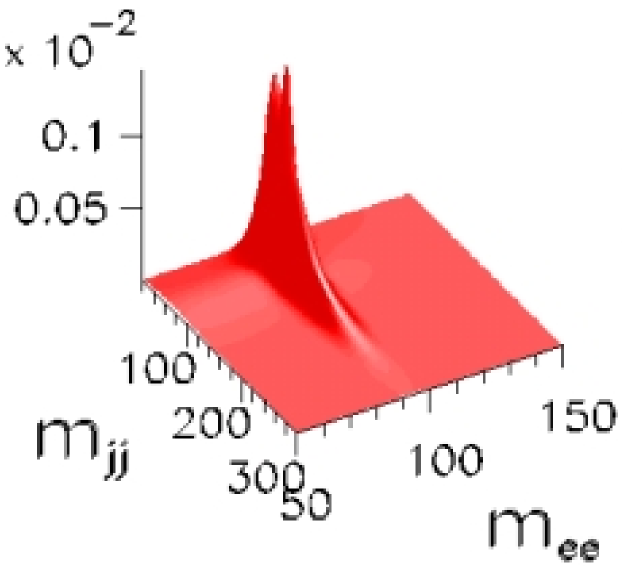
Examples #8

$Z^* \to Zh \to ee\ 2j$

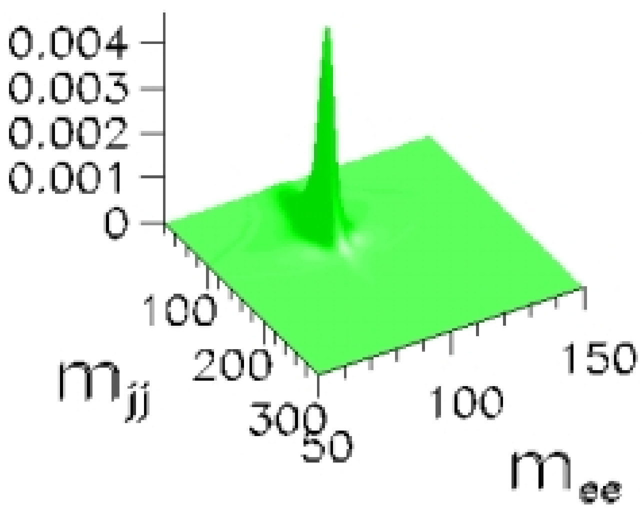


ε_{sig}	20%
$\hat{b}$	19.5 ± 4.1
N_{obs}	25
$\sigma^{95\%} \times \mathcal{B}$	0.8 pb

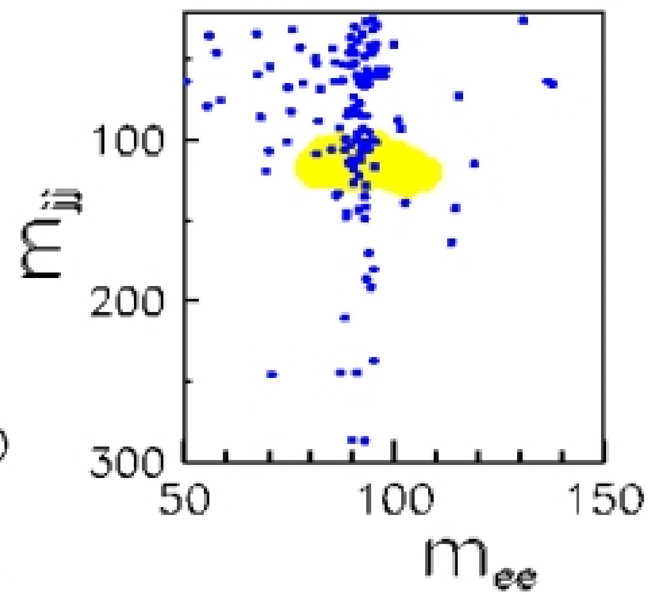
Background density



Signal density



Selected region



Heavy Higgs (m_h=200 GeV)

Standard Model $h \to WW \to e\not{E}_T\ 2j$

Predicted	$\sigma \times \mathcal{B}$	= 0.0047 pb
Limit	$\sigma^{95\%} \times \mathcal{B}$	= 4.4 pb

Standard Model $h \to ZZ \to ee\ 2j$

Predicted	$\sigma \times \mathcal{B}$	= 0.0005 pb
Limit	$\sigma^{95\%} \times \mathcal{B}$	= 0.6 pb

factor of 10^3 — no surprise here

In fact, our search has been more general:
$X \to WW \to e\not{E}_T\ 2j$
$X \to ZZ \to ee\ 2j$

Light Higgs (m_h=115 GeV)

Standard Model $W^* \to Wh \to e\not{E}_T\ 2j$

Predicted	$\sigma \times \mathcal{B}$	= 0.011 pb
Limit	$\sigma^{95\%} \times \mathcal{B}$	= 2.0 pb

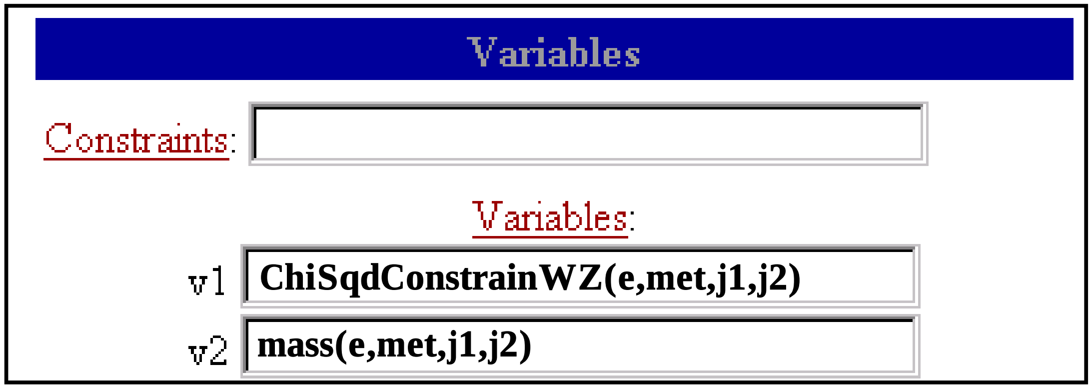
Standard Model $Z^* \to Zh \to ee\ 2j$

Predicted	$\sigma \times \mathcal{B}$	= 0.0023 pb
Limit	$\sigma^{95\%} \times \mathcal{B}$	= 0.8 pb

nothing expected, nothing seen

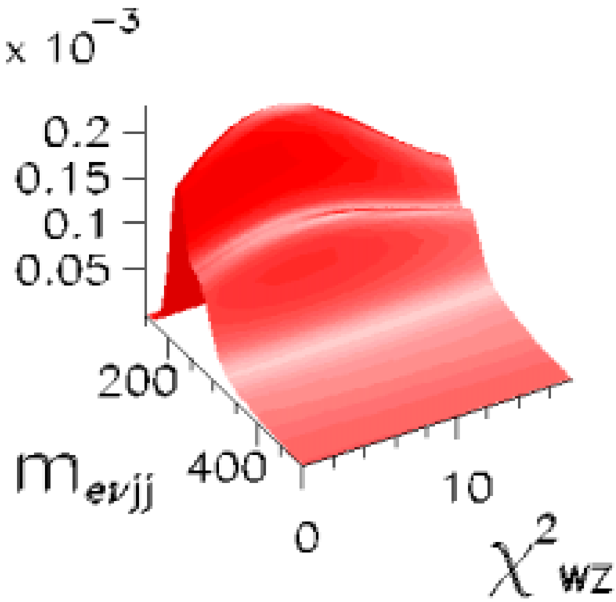
Examples

W' $\rightarrow$ WZ $\rightarrow$ e$\not{E}_T$ 2j

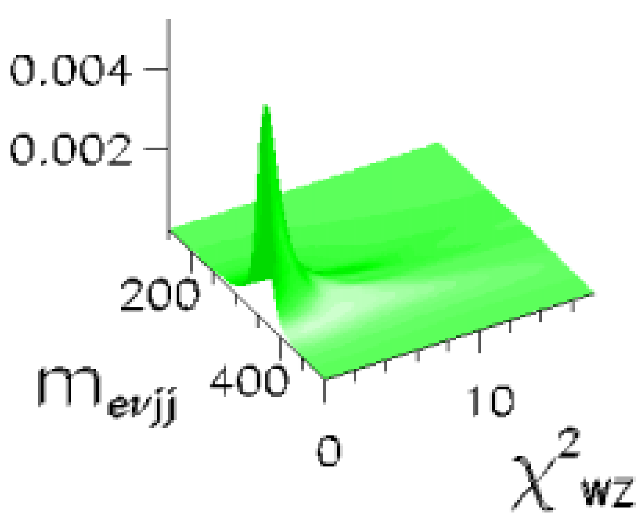


ε_{sig}	23%
$\hat{b}$	22.7 ± 5.2
N_{obs}	27
$\sigma^{95\%} \times \mathcal{B}$	0.7 pb

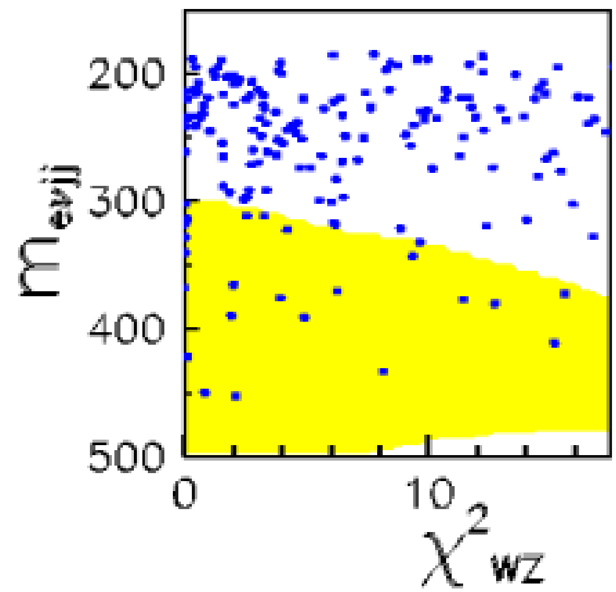
Background density



Signal density



Selected region



Examples #10

$Z' \rightarrow t\bar{t} \rightarrow e\not{E}_T\ 4j$

Variables

Constraints: **(j4_pt>15)&&(ChiSqdConstrain(e,met,W)<10)**

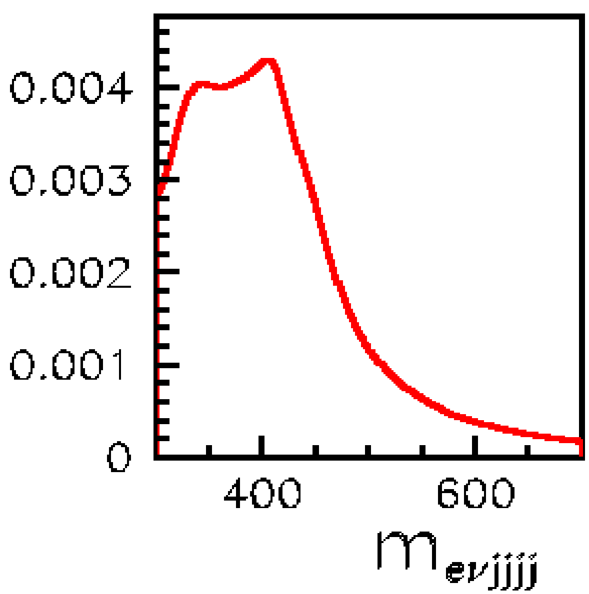
Variables:

v1 **mass(e,met,j1,j2,j3,j4)**

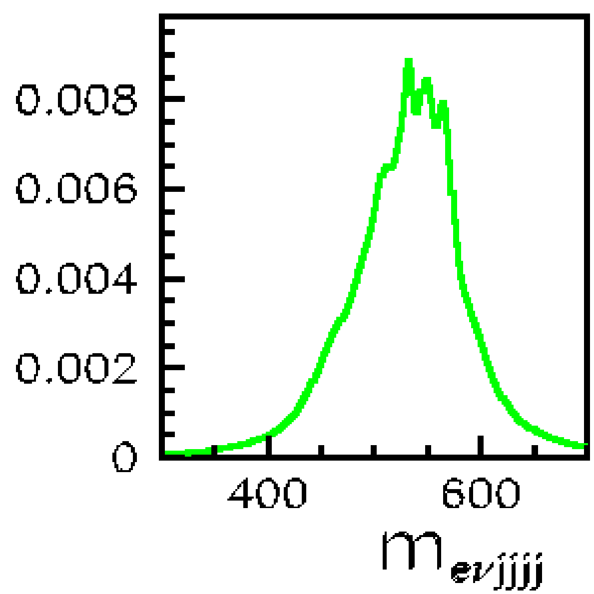
v2

ε_{sig}	14%
$\hat{b}$	3.8 ± 1.0
N_{obs}	2
$\sigma^{95\%} \times \mathcal{B}$	0.3 pb

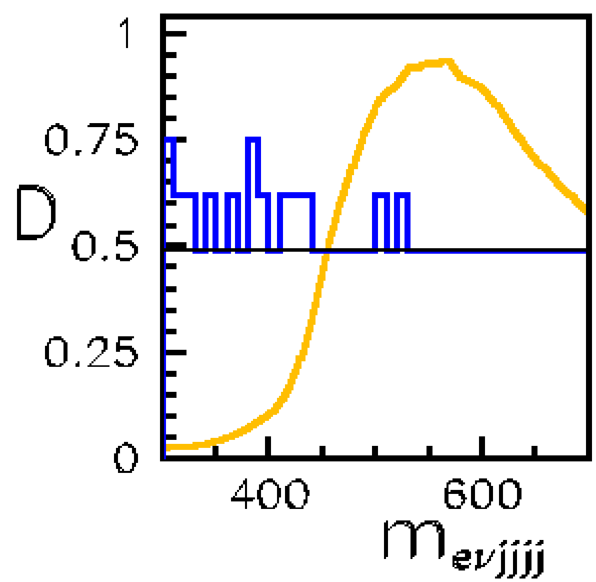
Background density



Signal density



Selected region



$W' \to WZ \to e\not{E}_T\ 2j$

"Extended gauge model" [Altarelli et al., Z. Phys. C 45 109 (1989)]

mass(W')	$\sigma^{95\%} \times \mathcal{B}$ (Quaero)	$\sigma^{95\%} \times \mathcal{B}$ (CDF)	$\sigma \times \mathcal{B}$ (theory)
200	3.4 pb	6.6 pb	0.65 pb
350	0.7 pb	2.0 pb	0.13 pb
500	0.2 pb	0.5 pb	0.02 pb

[CDF, hep-ex/0108004]

$Z' \to t\bar{t} \to e\not{E}_T\ 4j$

Standard Model (Z-like) couplings assumed

mass(Z')	$\sigma^{95\%} \times \mathcal{B}$ (Quaero)	$\sigma^{95\%} \times \mathcal{B}$ (CDF)
350	1.1 pb	-
450	0.9 pb	0.65 pb
550	0.3 pb	0.45 pb

[CDF, PRL 85 2062 (2000)]

Results are consistent with (and competitive with) previous measurements.

Leptoquarks $\rightarrow$ ee 2j

Variables

Constraints:

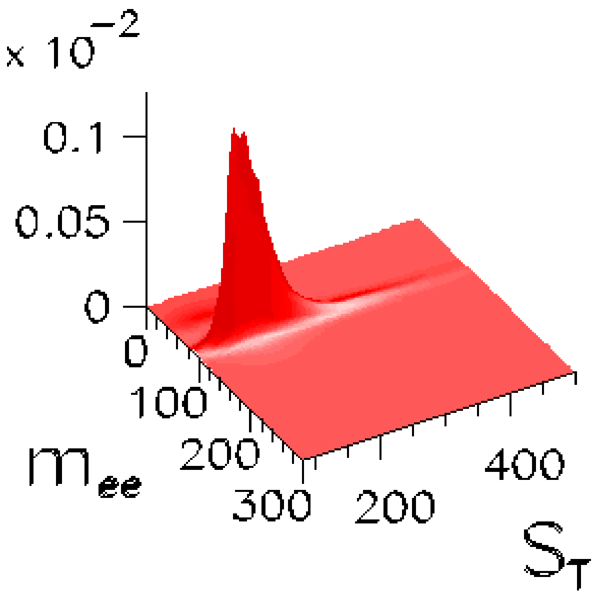
Variables:

v1 **e1_pt + e2_pt + j1_pt + j2_pt + j3_pt + j4_pt**

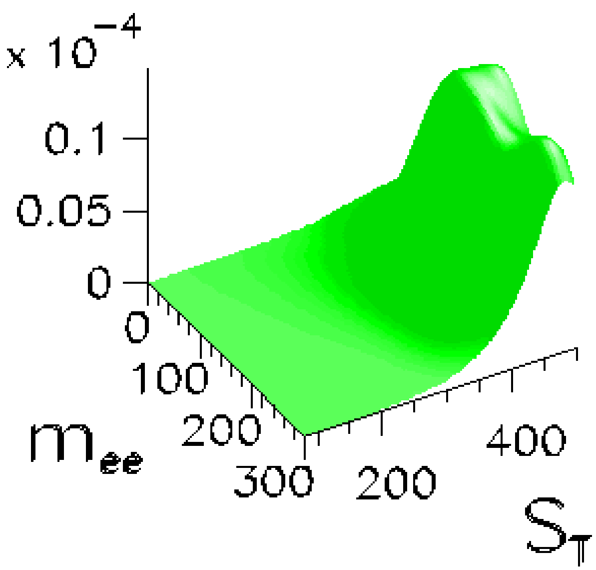
v2 **mass(e1,e2)**

ε_{sig}	33%
$\hat{b}$	0.3 ± 0.1
N_{obs}	0
$\sigma^{95\%} \times \mathcal{B}$	0.07 pb

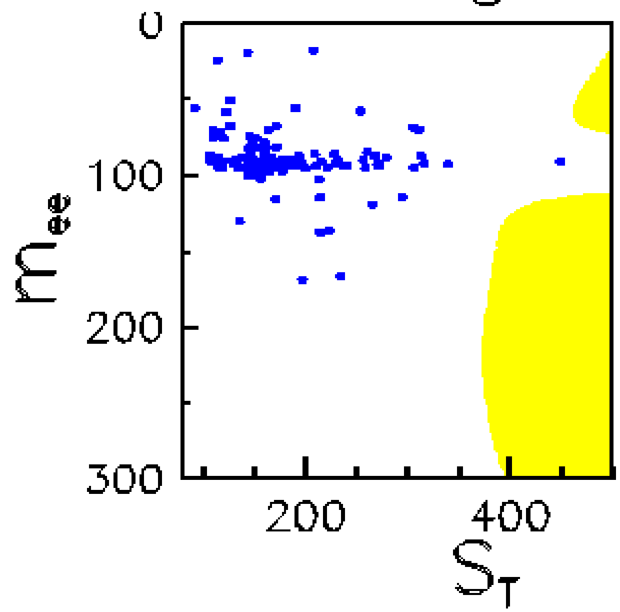
Background density



Signal density



Selected region



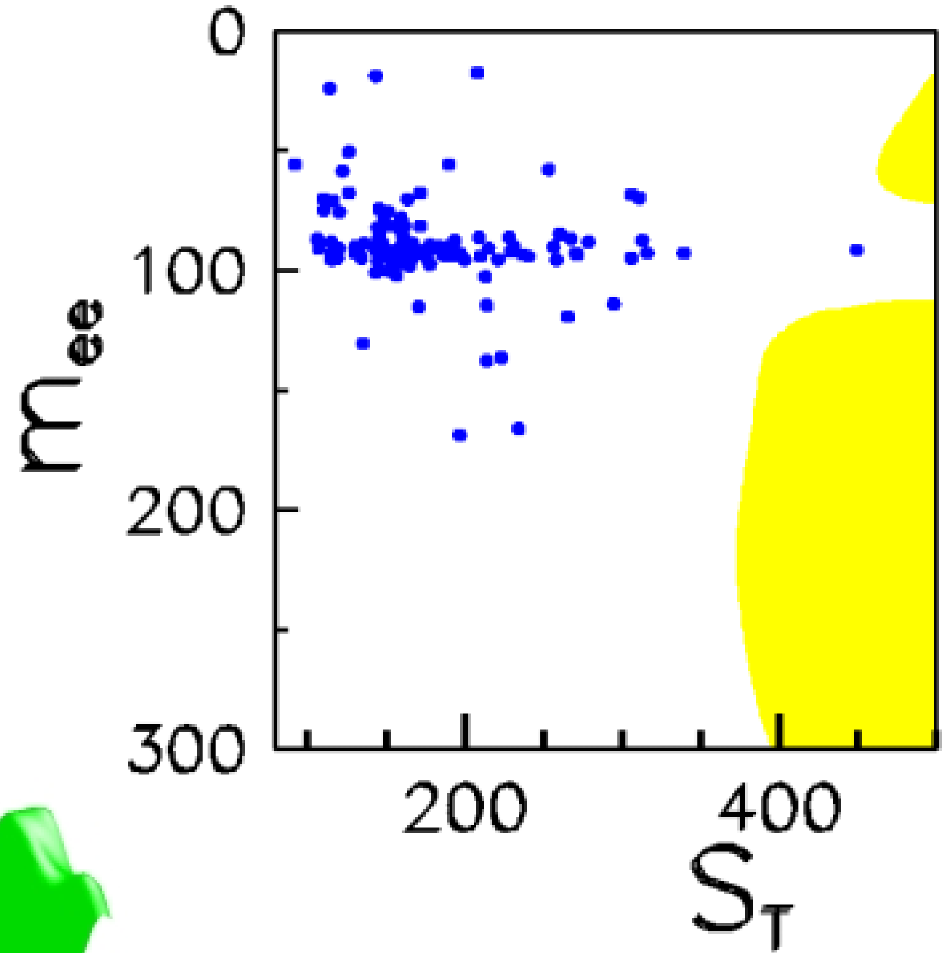
Leptoquarks → ee 2j

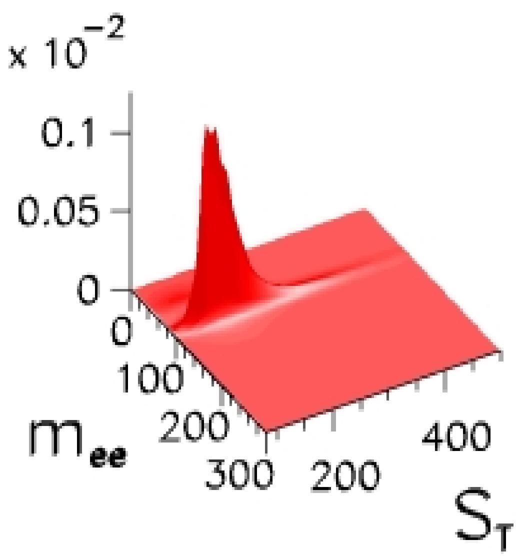
Quaero

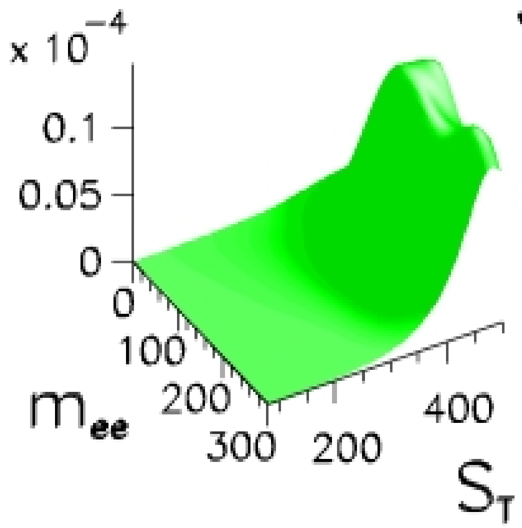
$$\hat{b} = 0.3 \pm 0.1$$

$$\varepsilon_{sig} = 0.33$$

$$\sigma^{95\%} \times \mathcal{B} = 0.07\ pb$$







Previous

0.4 ± 0.15

0.35

$0.07 \ pb$

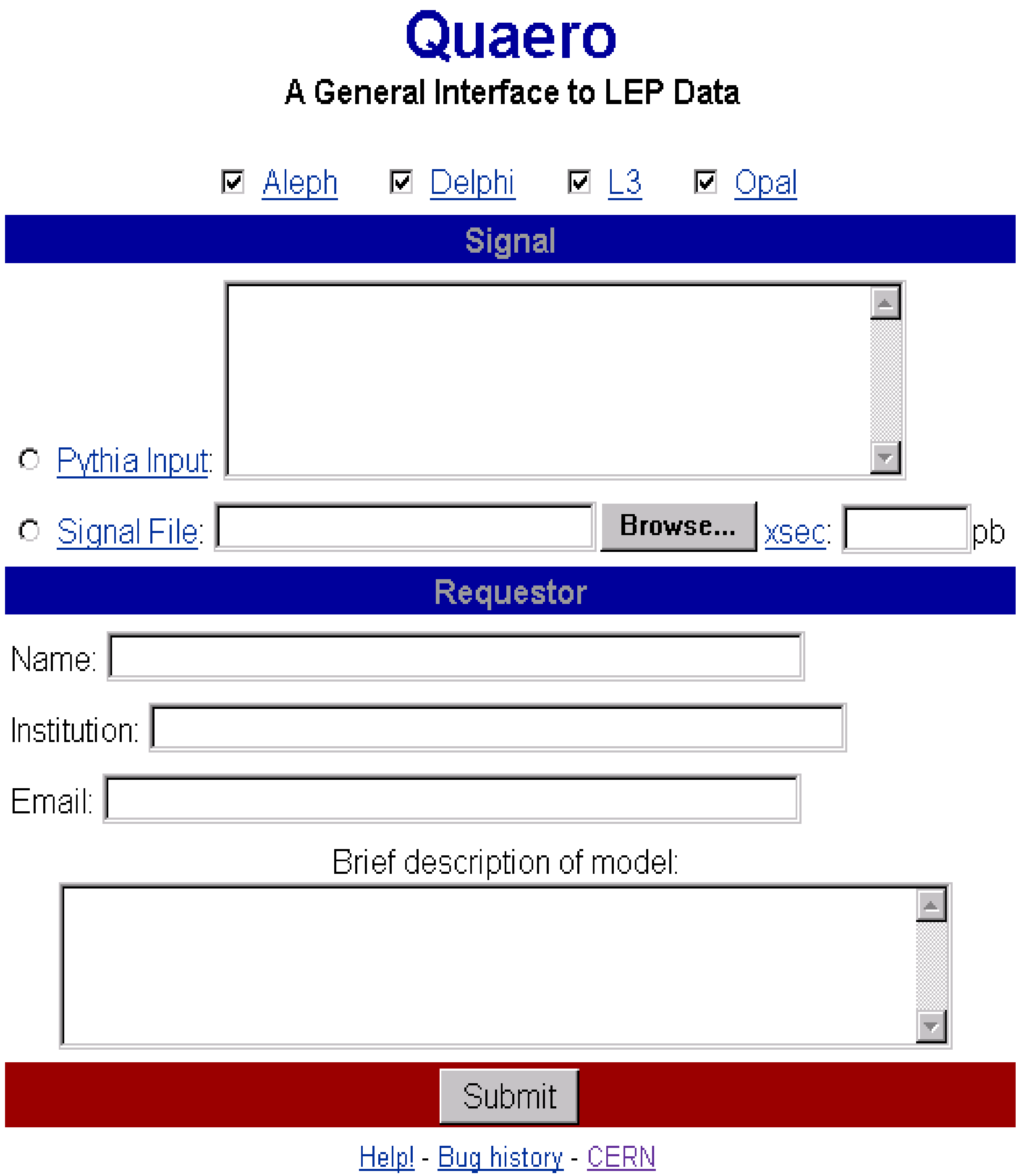
Nearly identical results

(to be overlayed)

Examples

Quaero performs as expected in 11 of 11 sample analyses


Motivation
Data
Algorithm
Examples
LEP


Quaero
A General Interface to LEP Data
Aleph
Delphi
L3
Opal
Signal
Pythia Input:
Signal File:
Browse...
xsec:
pb
Requestor
Name:
Institution:
Email:
Brief description of model:
Submit
Help! - Bug history - CERN

Current challenges:

1) Archiving data

   Can we archive ~~HEP~~ LEP data in a meaningful way?

2) Combining results from various final states

   What do the data *as a whole* say about a particular model?

3) Combining results from various experiments

   Can we do this consistently and correctly, with minimal headache?

4) Precision measurements

   Can Quaero do measurements, in addition to searches?

5) Systematic errors

   Can we allow an arbitrarily sophisticated treatment of systematics?

Any experiment wishing to use Quaero needs to provide 4 things:

- Data
  - Object 4-vectors
- Backgrounds
  - Object 4-vectors
- Systematic errors
  - Sources of error & effect on 4-vectors
- Detector simulation
  - (and reconstruction, if necessary)

Event format:

eventType
weight{err/mag,...} **sqrt(s)**{err/mag,...}
e+ E\{err/mag,...\} $\cos(\theta)$ ϕ
e- E\{err/mag,...\} $\cos(\theta)$ ϕ
b E\{err/mag,...\} $\cos(\theta)$ ϕ
b E\{err/mag,...\} $\cos(\theta)$ ϕ
uncl m\{err/mag,...\} E\{err/mag,...\} $\cos(\theta)$ ϕ **;**

Data event:

data	1	190.0			
e+	45.2	+0.11	0.21		
e-	47.3	-0.05	3.56		
b	46.0	-0.16	1.71		
b	48.2	-0.02	4.90		
uncl	0.44	3.3	+0.07	3.97	;

Background event:

ZZ 0.0041{1/0,12/0.0002,0201/0.0001} 190.0
e+ 45.2{0221/1.4} +0.11 0.21
e- 47.3{0221/1.5} -0.05 3.56
b{0211/j/0.001} 46.0{0222/3.6} -0.16 1.71
b{0211/j/0.05} 48.2{0222/3.7} -0.02 4.90
uncl 0.44{0222/0.07} 3.3{0222/0.32} +0.07 3.97 ;

Instead of returning cross section limits, Quaero will return a single number:

$$\mathcal{L}(\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})},$$

where $\mathcal{H}$ is the hypothesis being tested.

(Further manipulation of this quantity is then straightforward)

Will this new algorithm actually work in practice?

We'll see . . .

Conclusions

- **Quaero** is a method for making HEP data publicly available
  - Perform the analysis automatically
  - Results within the hour
- **Quaero** has been used to publish DØ Run I data
  - hep-ex/0106039, submitted to PRL
  - http://quaero.fnal.gov/