First evidence for rare single top quark production at the DØ experiment

Brigitte Vachon
Canada Research Chair, McGill University, Montréal
for the DØ Collaboration

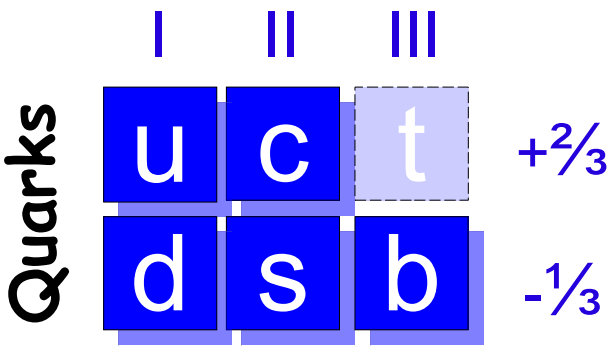
WNPPC, Banff
16 February, 2007

Outline

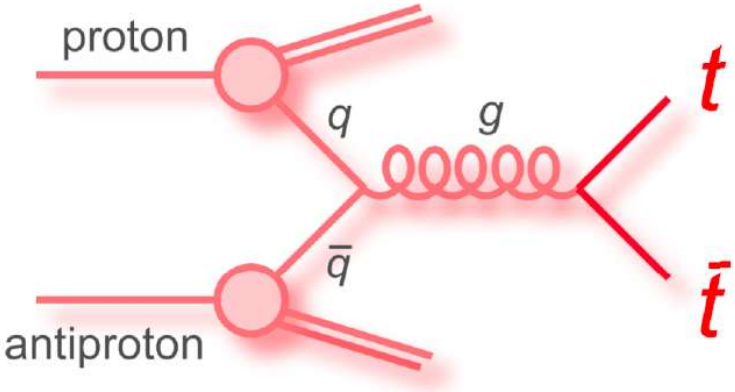
- Introduction
- DØ experiment at the Tevatron
- Event selection
- Event classification
- Cross-section measurement
- $|V_{tb}|$ measurement
- Summary

The Top Quark

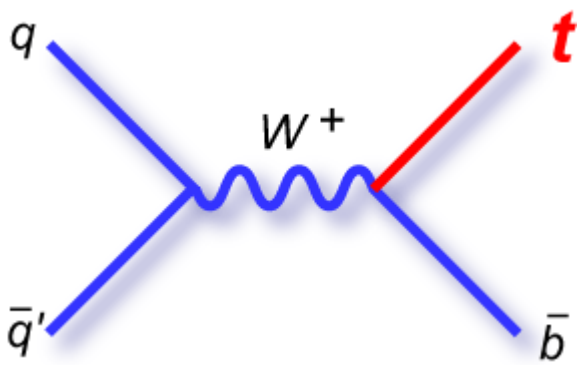
- Spin ½ fermion, charge +⅔
- Weak-isospin partner of the bottom quark
- Heaviest known fundamental particle
- M_{top} = 171.4 ± 2.1 GeV

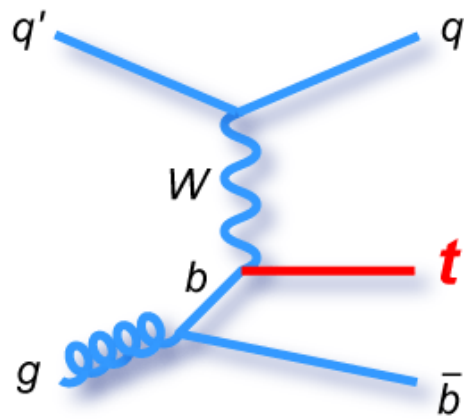


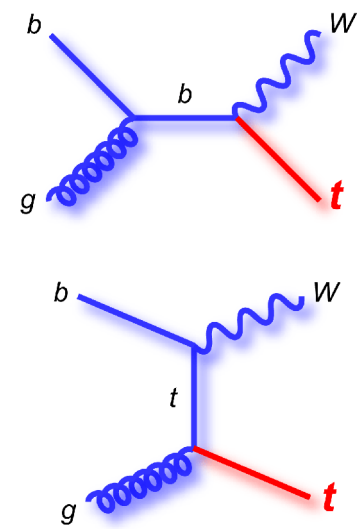
- Discovered in 1995 by CDF/DØ
- Produced in $t\bar{t}$ pairs at the Tevatron
- Cross-section = 6.8 ± 0.6 pb (NNLO)
- Measurements consistent with this value



Single Top Quark Production







s-channel: "*tb*"

$\sigma_{NLO} = 0.88 \pm 0.11$ pb

t-channel: "*tqb*"

$\sigma_{NLO} = 1.98 \pm 0.25$ pb

"*tW* production"

$\sigma_{NLO} = 0.21$ pb

Experimental Results (95% C.L.):

DØ (370 pb^{-1}): σ_{tb} **< 5.0 pb**

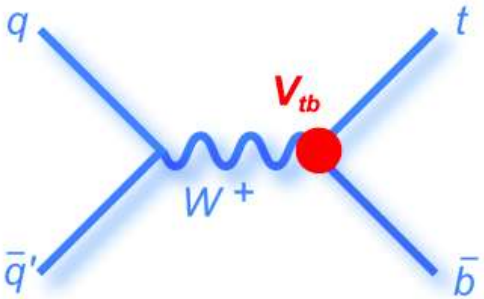
CDF (700 pb^{-1}): σ_{tb} **< 3.2 pb**

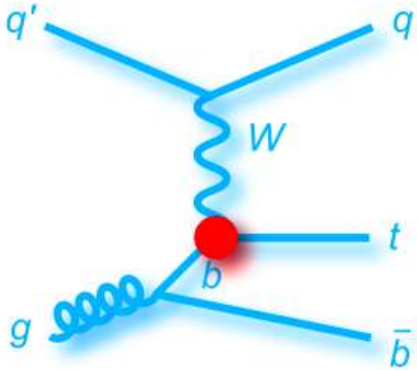
DØ (370 pb^{-1}): σ_{tqb} **< 4.4 pb**

CDF (700 pb^{-1}): σ_{tqb} **< 3.1 pb**

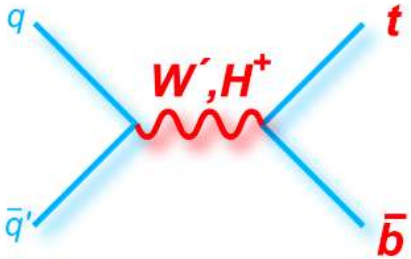
Motivation

- Direct test of Standard Model
- Study *Wtb* coupling
  - Measure $|V_{tb}|$ directly
  - Test unitarity of CKM matrix
  - Anomalous *Wtb* couplings

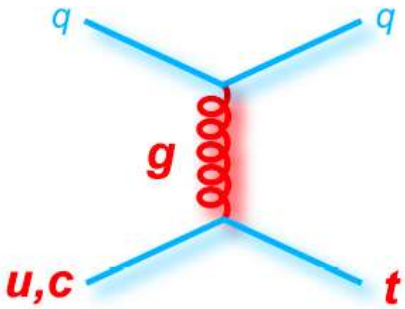




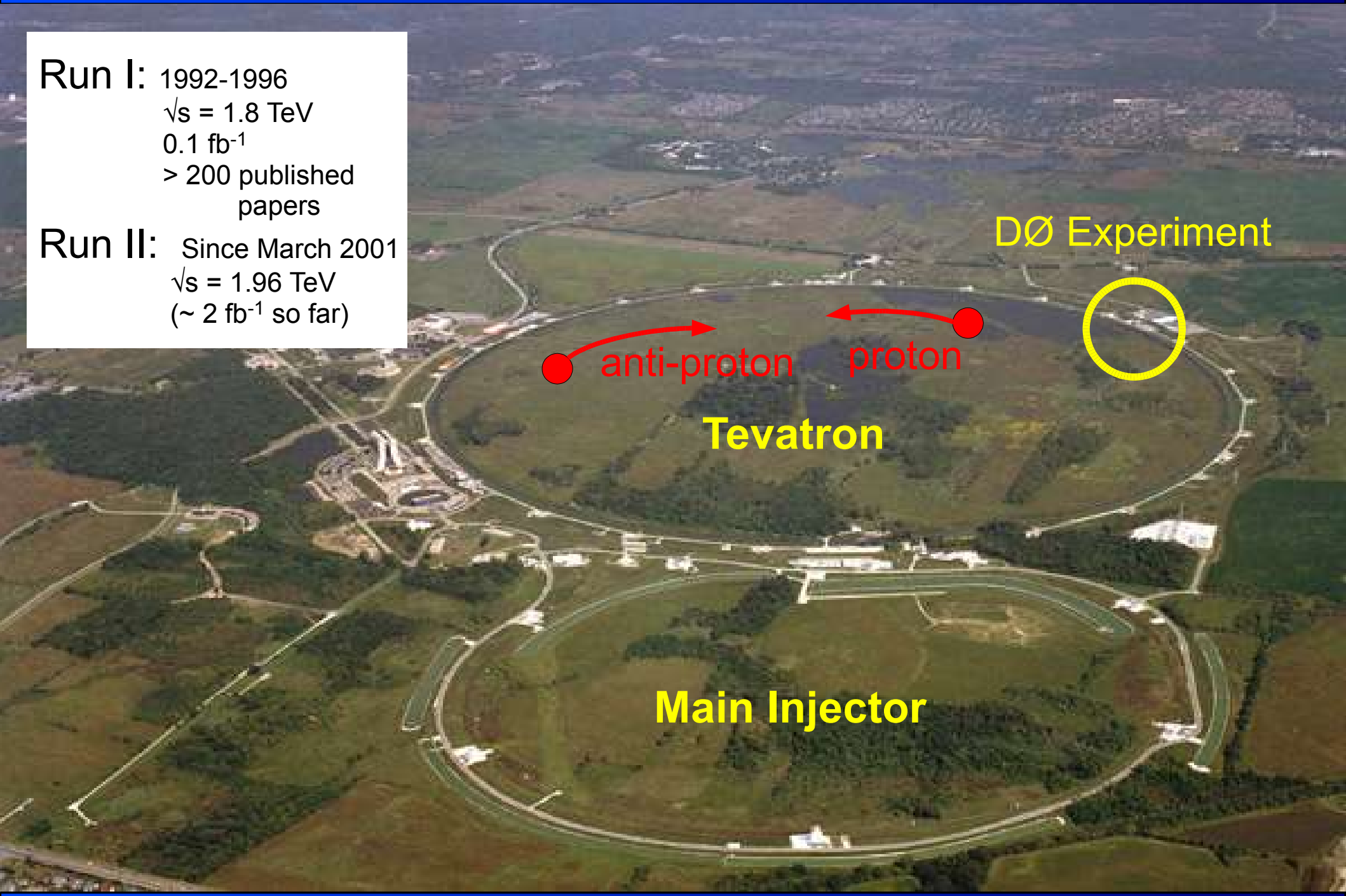
- Search for new physics
  - *s*/*t*-channel have different sensitivity to new physics
  - *s*-channel: Heavy resonances (W', H^+, KK state, technipion, etc.)
  - *t*-channel: FCNC (t-$Z/\gamma/g$-c/u couplings)



- Top polarization measurement
- Top quark partial decay width and lifetime measurement
- Analysis techniques development



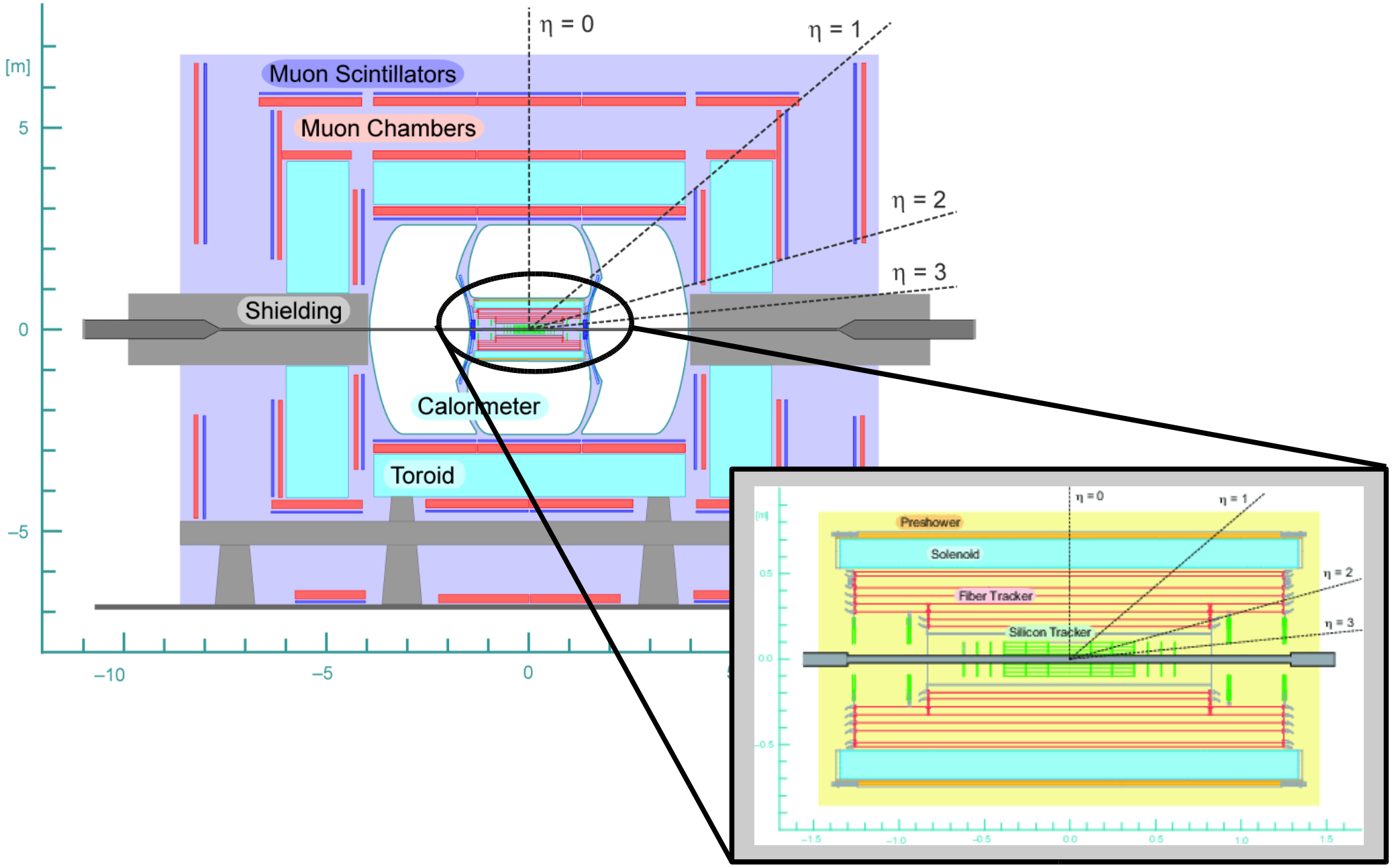
The Tevatron at Fermilab

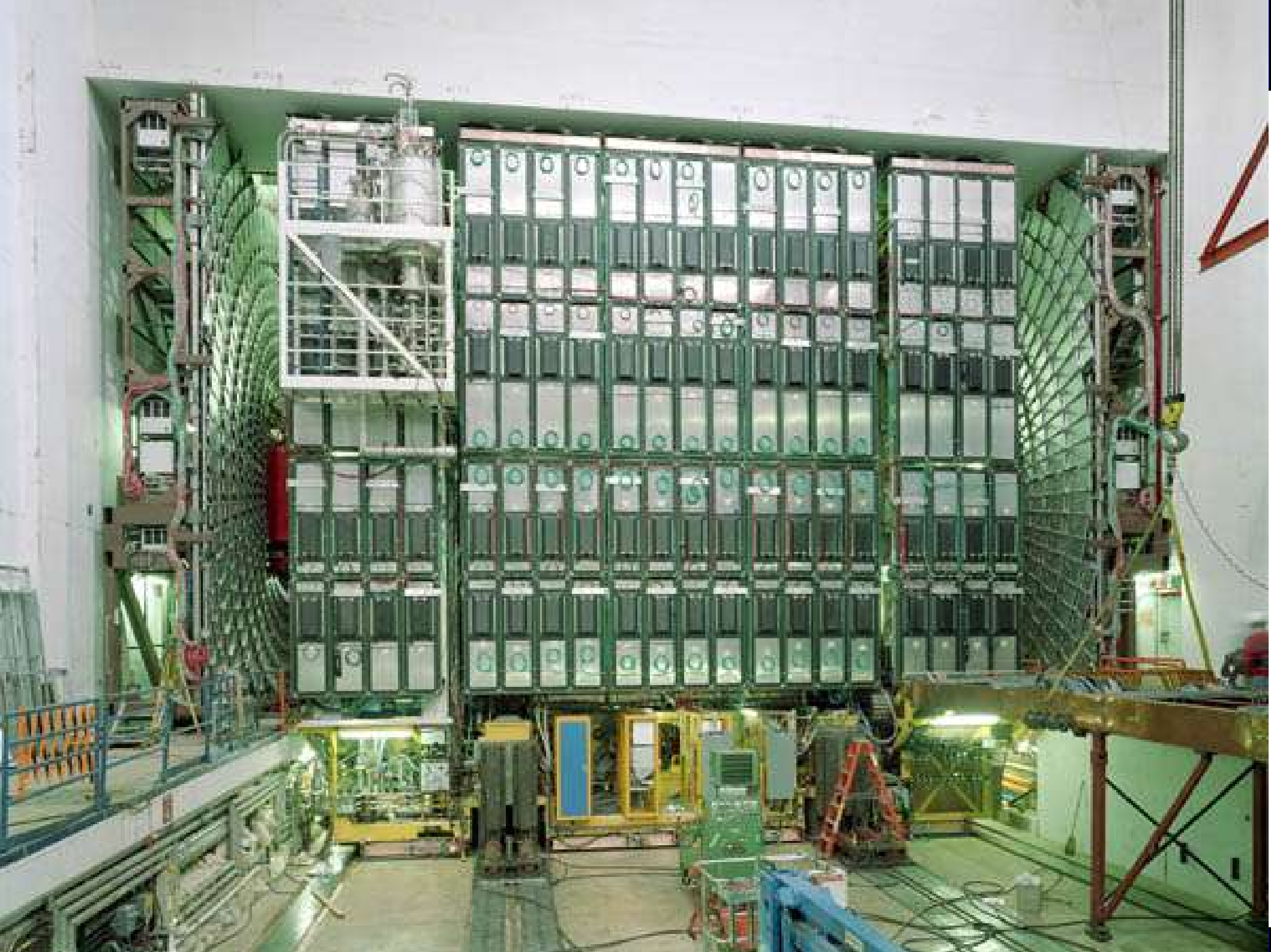


Run I: 1992-1996
$\sqrt{s}$ = 1.8 TeV
0.1 fb^{-1}
> 200 published papers

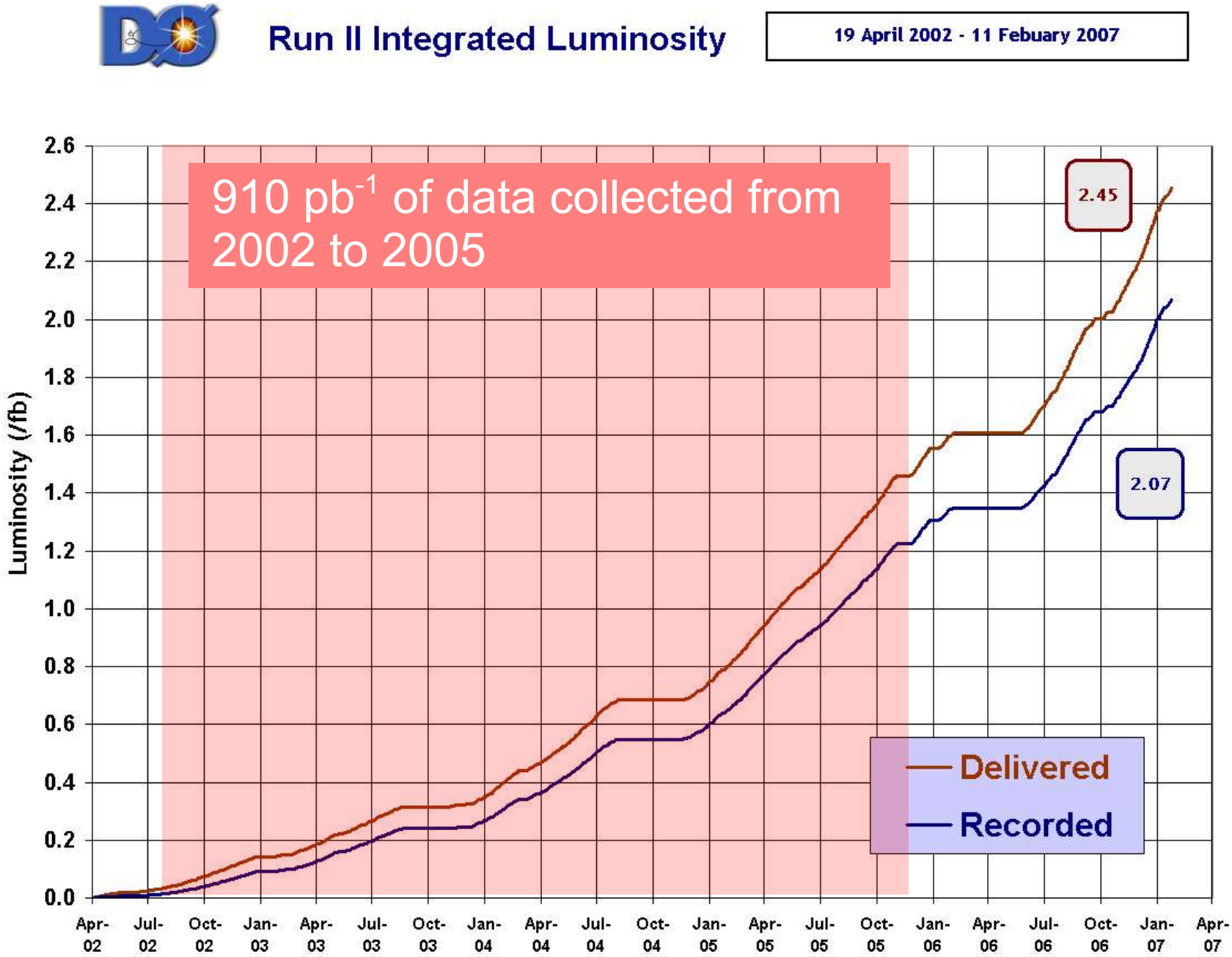
Run II: Since March 2001
$\sqrt{s}$ = 1.96 TeV
(~ 2 fb^{-1} so far)

The DØ Detector



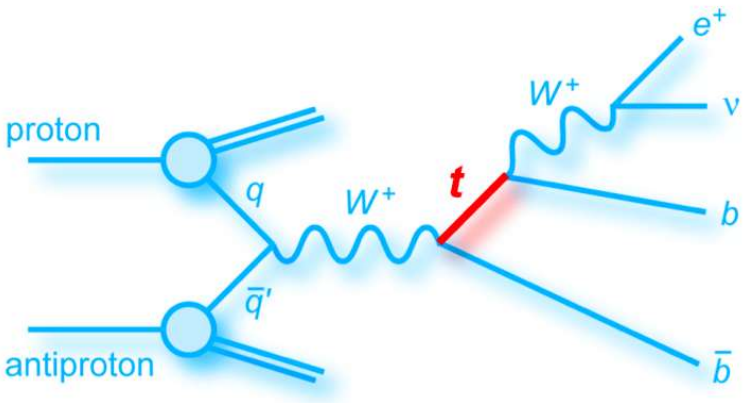


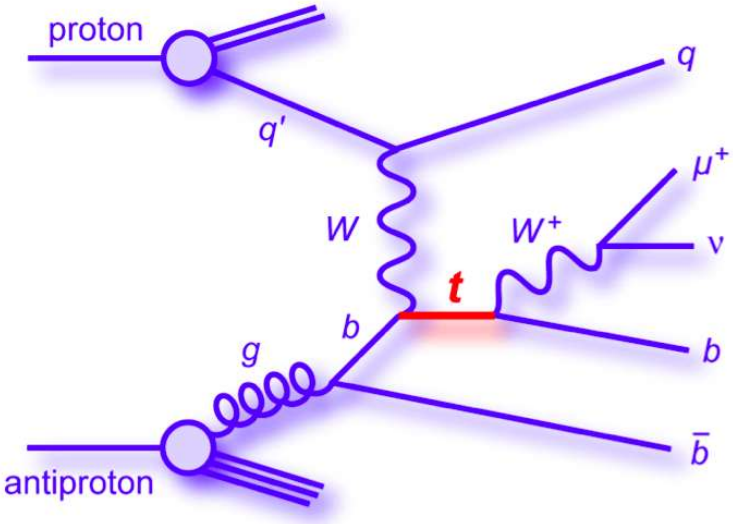
Dataset


Run II Integrated Luminosity
19 April 2002 - 11 Febuary 2007
910 pb^{-1} of data collected from 2002 to 2005
Luminosity (/fb)
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2.45
2.07
Delivered
Recorded
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07

Event Selection

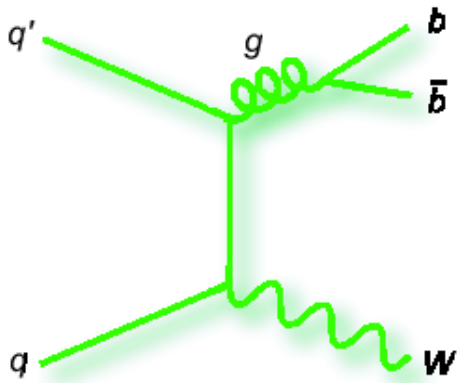
- One isolated electron or muon
  - Electron $p_T > 15$ GeV, $|\eta| < 1.1$
  - Muon $p_T > 18$ GeV, $|\eta| < 2.0$
- Missing transverse energy
  - $\not{E}_T > 15$ GeV
- At least two jets and one b-tagged jet
  - 2-4 jets with $p_T > 15$ GeV, $|\eta| < 3.4$
  - Leading jet $p_T > 25$ GeV, $|\eta| < 2.5$
  - Second Leading jet $p_T > 20$ GeV



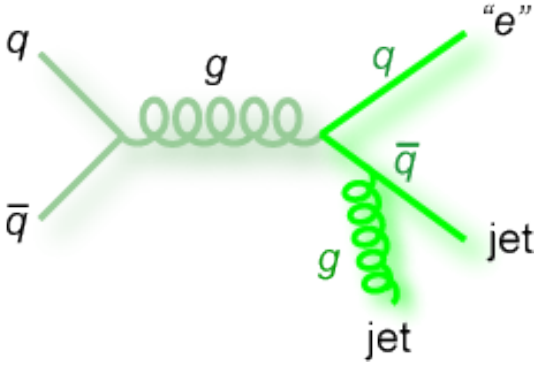


Signal and Background

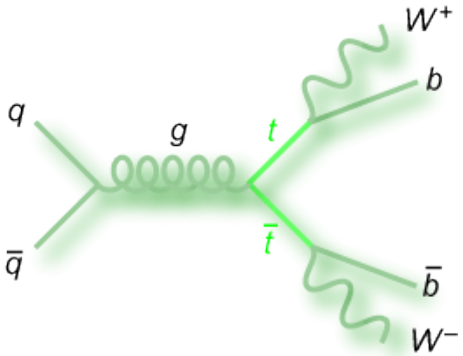
- *W/Z*+jets production ("real lepton")
  - W+jets background modeled using ALPGEN+PYTHIA (hadronization)
  - W+jets normalized to data
  - Wbb and Wcc fractions from data to take into account higher order effects
  - Z+jets and di-boson background very small and included in W+jets normalization



- Multi-jet events ("fake lepton")
  - Jet mis-identified as isolated electron/muon
  - Background modeled using data with non-isolated lepton and jets
  - Background normalized to data



- Top pair production
  - $t\bar{t}$ background modeled using ALPGEN+PYTHIA (hadronization)
  - Background normalized to NNLO cross-section



Event Yields

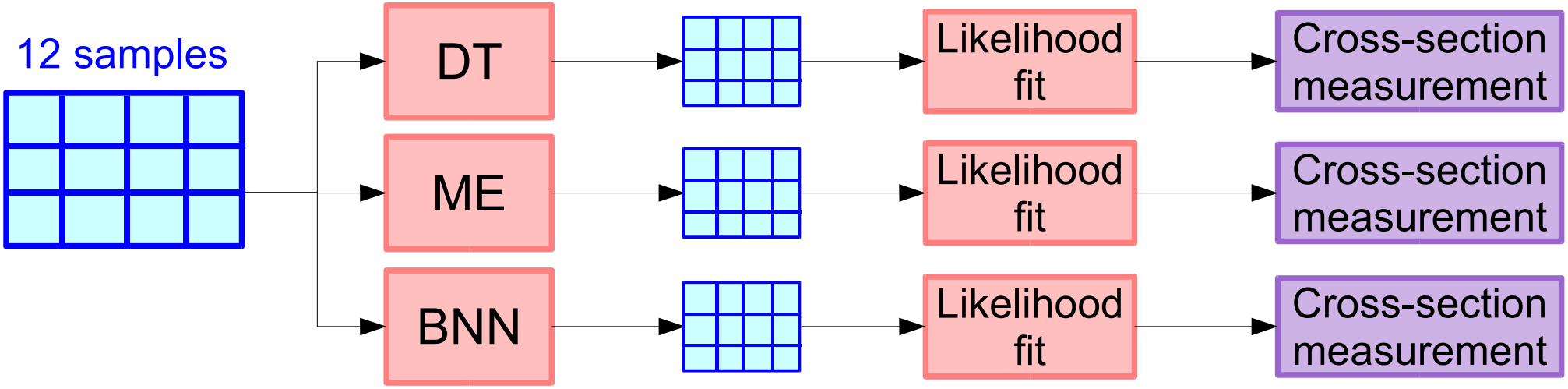
Source	Event Yields in 0.9 fb^{-1} Data Electron+muon, 1tag+2tags combined		
	2 jets	3 jets	4 jets
tb	16 ± 3	8 ± 2	2 ± 1
tqb	20 ± 4	12 ± 3	4 ± 1
$t\bar{t} \to ll$	39 ± 9	32 ± 7	11 ± 3
$t\bar{t} \to l$+jets	20 ± 5	103 ± 25	143 ± 33
$W+b\bar{b}$	261 ± 55	120 ± 24	35 ± 7
$W+c\bar{c}$	151 ± 31	85 ± 17	23 ± 5
$W+jj$	119 ± 25	43 ± 9	12 ± 2
Multijets	95 ± 19	77 ± 15	29 ± 6
Total background	686 ± 41	460 ± 39	253 ± 38
Data	697	455	246

- **Signal acceptances**: tb = (3.2 ± 0.4)%, tqb = (2.1 ± 0.3)%
- **Signal : background ratios**: tb+tqb combined are 1:10 to 1:50

Search Strategy

- Expected single top signal is **smaller** than total background uncertainties
  - Simple event counting method is not sensitive enough to presence of signal
  - Need to use multivariate discriminant:
    - **Boosted decision trees**
    - **Matrix elements**
    - **Bayesian neural networks**

- Divide analysis in **12** independent samples: (e,μ) x (2,3,4 jets) x (1,2 b-tags)
- Optimize discriminant for each independent search channel



12 Analysis Channels

- Data
- tb
- tqb
- $t\bar{t}$
- W + jets
- Multijets
- ±1σ uncertainty on background

W transverse mass

electrons

1 b-tag | 2 b-tags

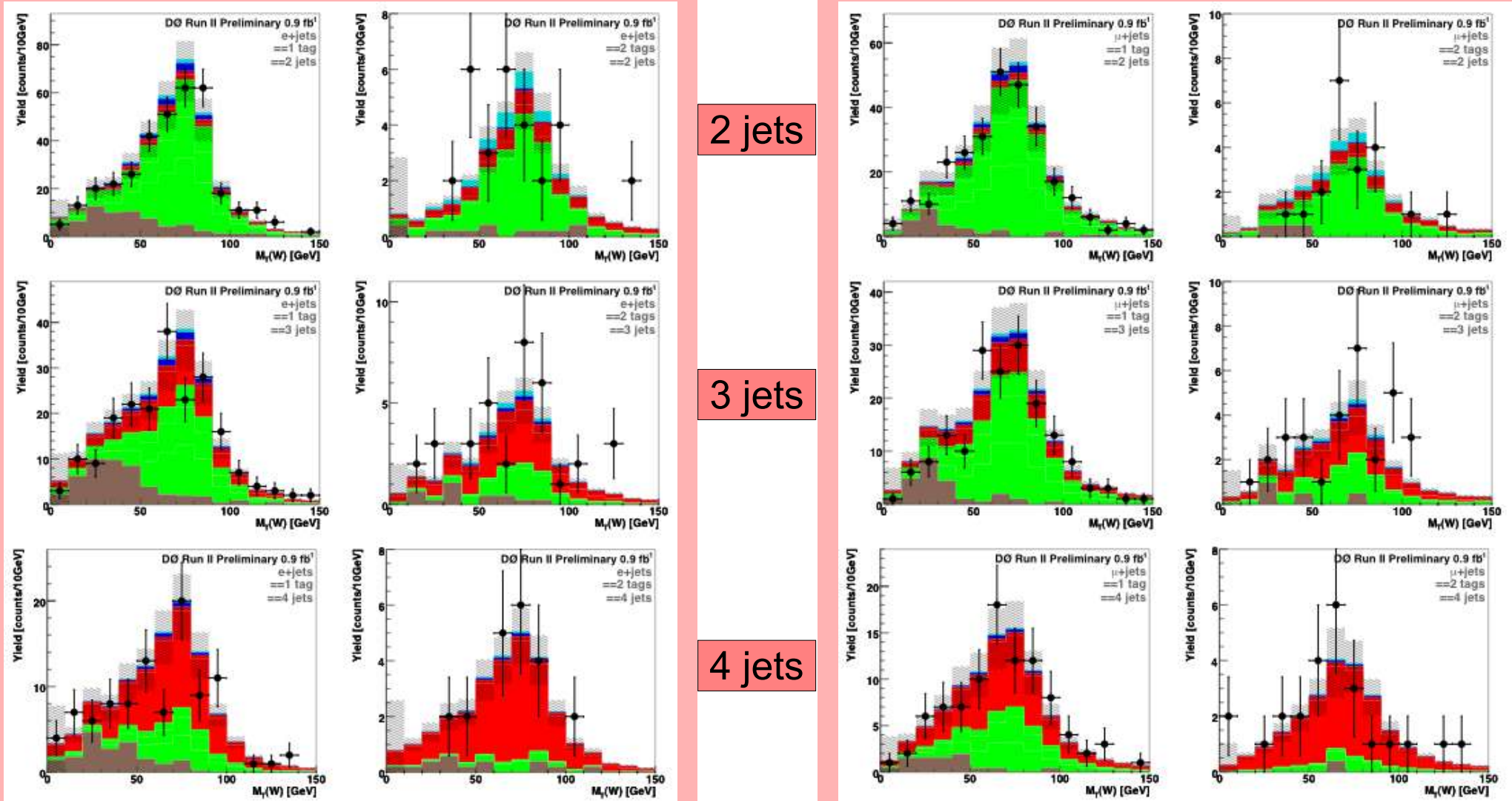
muons

1 b-tag | 2 b-tags

2 jets

3 jets

4 jets



Systematic Uncertainties

- Uncertainties assigned for each signal and background component in each analysis channels
- Correlations between channels and sources of systematics are taken into account.
- Two different types of uncertainties:
  - Normalization uncertainties
  - Uncertainties affecting shape of distributions (Jet energy scale, tag-rate function for b-tagging Monte Carlo events)

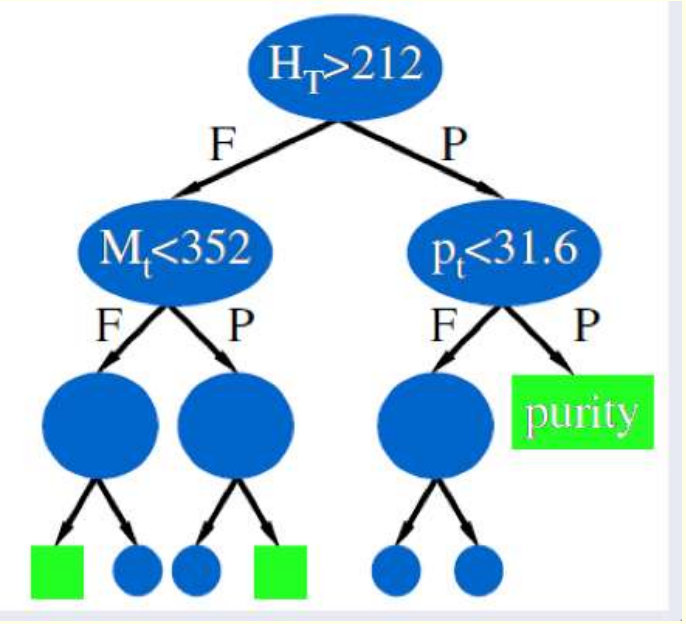
Source of Uncertainty	Size
Top pairs normalization	18%
W+jets & multijets normalization	18–28%
Integrated luminosity	6%
Trigger modeling	3–6%
Lepton ID corrections	2–7%
Jet modeling	2–7%
Other small components	Few %
Jet energy scale	1–20%
Tag rate functions	2–16%

Single top cross-section measurement uncertainties dominated by **statistical** uncertainties.

1 - Decision Trees

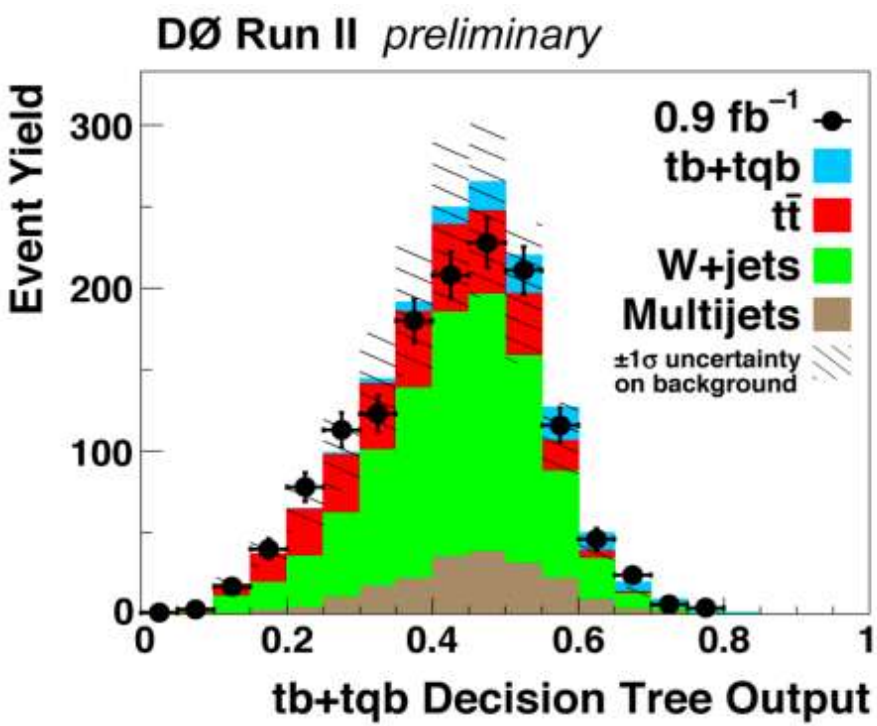
- **Machine-learning technique** widely used in social science
- Idea: use information from events that fail criteria in a simple cut-based analysis
- A **decision tree** consists in a chain of selection criteria that defines unique samples of events that are either ***background*-like** or ***signal*-like**

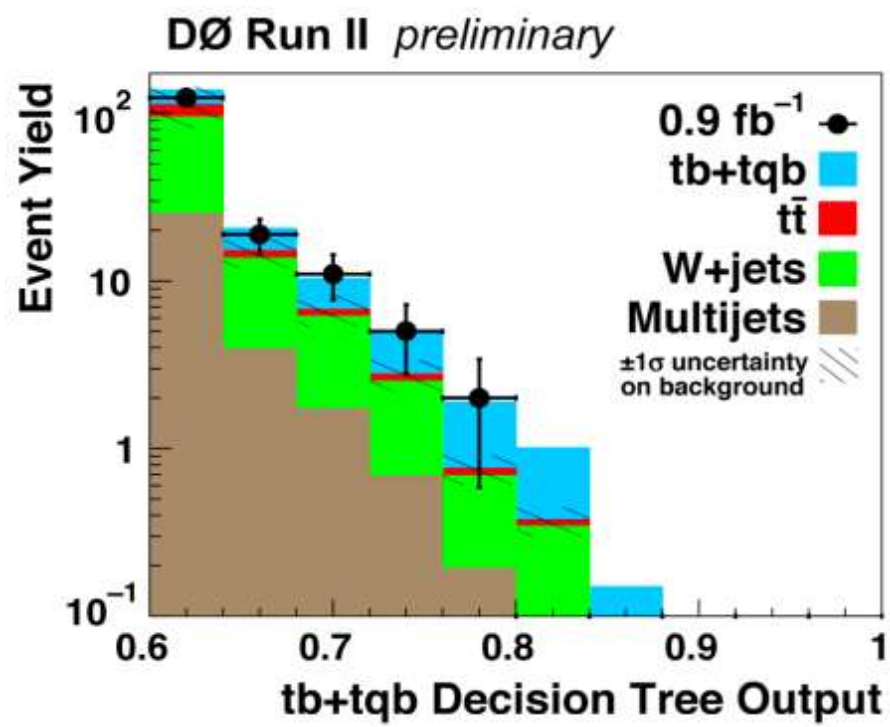
- For each "**node**" , pick the variable and cut value that provide the best separation between signal and background.
- Split the events based on whether they pass (**P**) or fail (**F**) the cut.
- Repeat recursively
- Terminate when improvement in signal to background separation stops or when too few events are left.
- For each terminal node ("**leaf**") calculate purity = $N_{signal}/(N_{signal}+N_{bkg})$
- **Decision Tree output** for each event = **leaf purity**



1 - Boosted Decision Trees

- "**Boosting**": Technique used to improve the performance of a weak classifier
- Construct a **weighted average** of the results of many **trees**
  - Improves performance
  - Dilutes the piecewise nature of decision trees
- In this analysis:
  - Trained 36 sets of decision trees: *(tb+tqb, tb, tqb) x (e,μ) x (2,3,4 jets) x (1,2 b-tags)*
  - Use 49 input variables





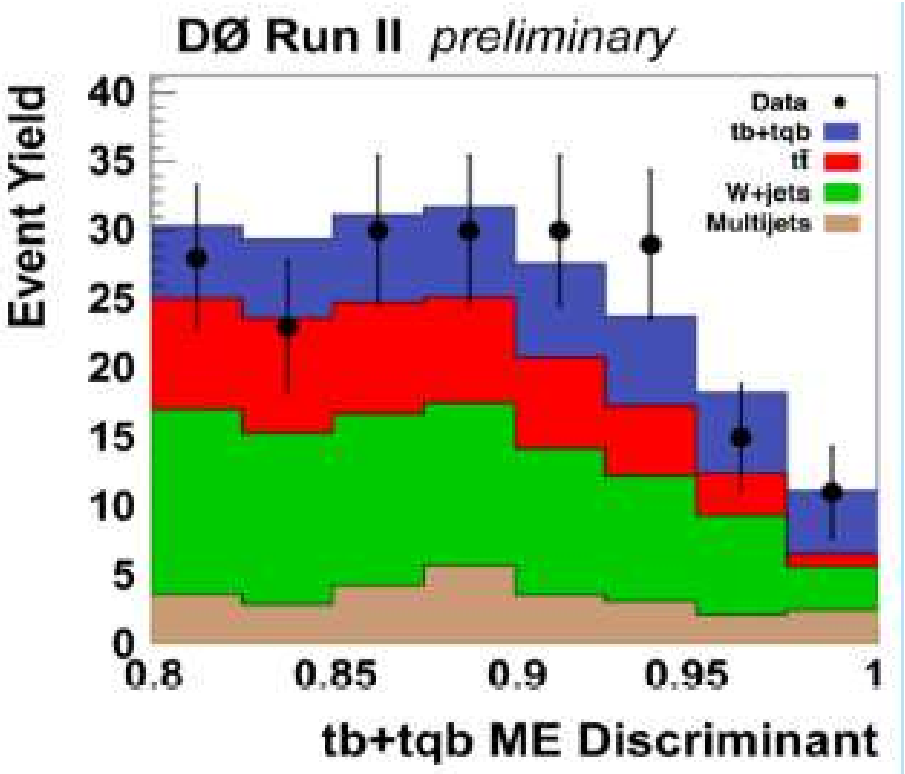
2 - Matrix Elements

- Method pioneered by DØ for top quark mass measurement
- Use 4-vectors of all reconstructed objects in an event
- Use **matrix elements** of main signal and background Feynman diagrams to compute an event probability density for signal and background hypothesis:

$$P_s(\vec{x}) = \frac{1}{\sigma_s} \, d\sigma_s(\vec{x})$$

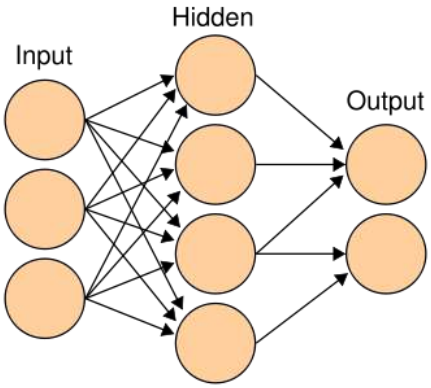
- Construct a **discriminant**:

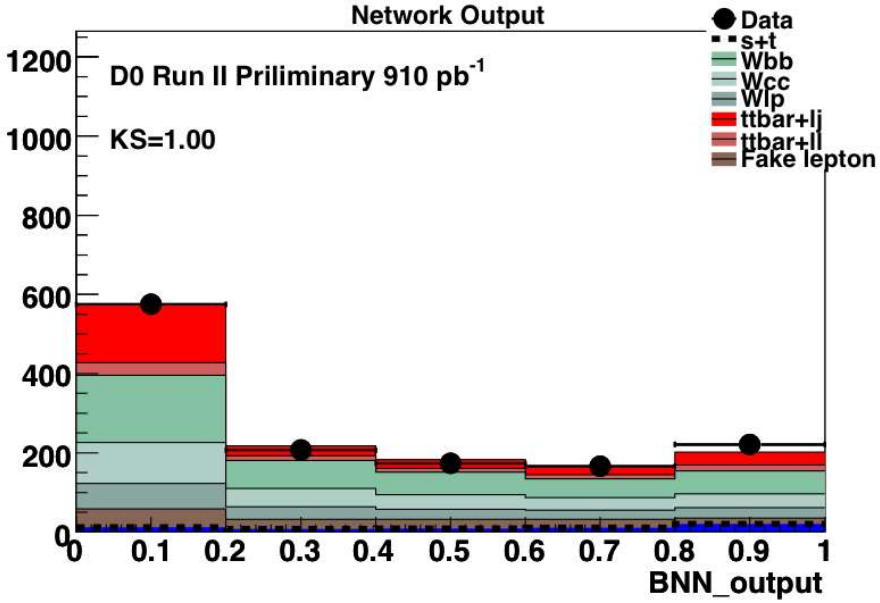
$$D_S(\vec{x}) \equiv P(S|\vec{x}) = \frac{P_S(\vec{x})}{P_S(\vec{x}) + P_B(\vec{x})}$$

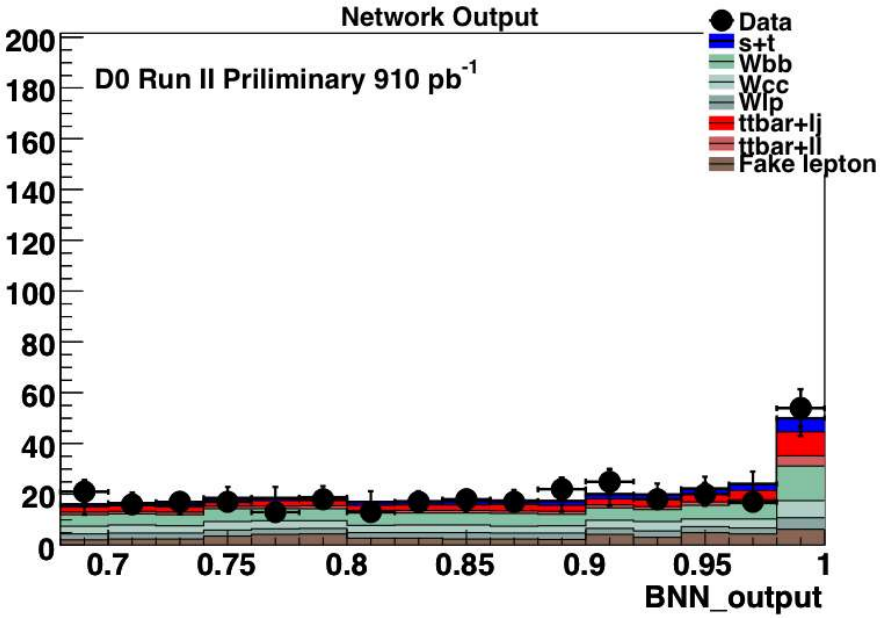


3 - Bayesian Neural Network

- Neural network is a non-linear statistical technique to data classification
- **Bayesian neural networks** consists in a weighted average of many networks, given a training sample
  - Less prone to over-training due to averaging
  - Network structure is less important, just use a large network
  - Computationally intensive







Cross-section measurement

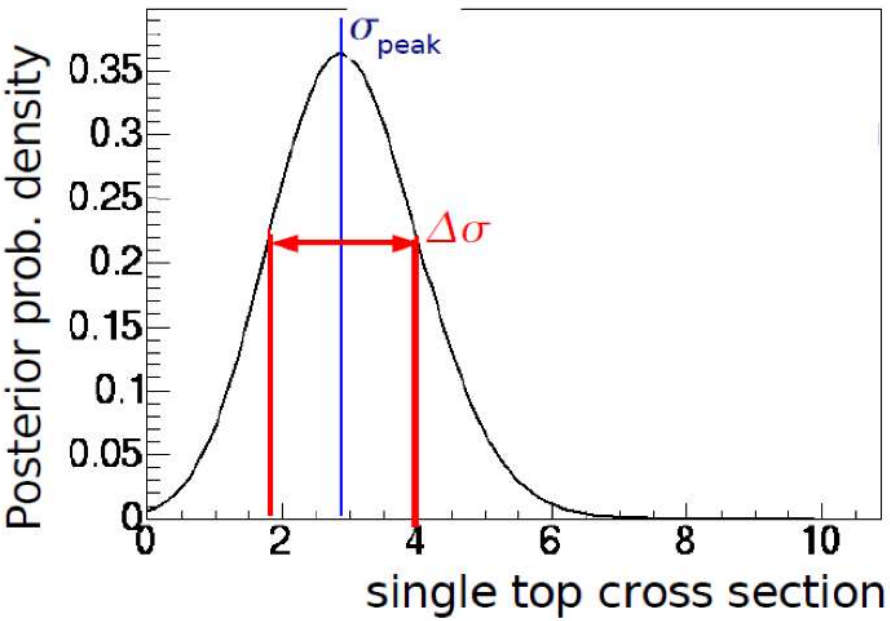
- Compute Bayesian posterior probability density

$$\text{Posterior Probability Density}(\sigma|\text{data}) \propto \int\int \text{Prob}(\text{data}|\sigma, a, b)\,\text{Prior}(\sigma)\,\text{Prior}(a, b)\,da\,db$$

- Use shape of output discriminants (DT, ME, BNN)
- Combined all individual search channels

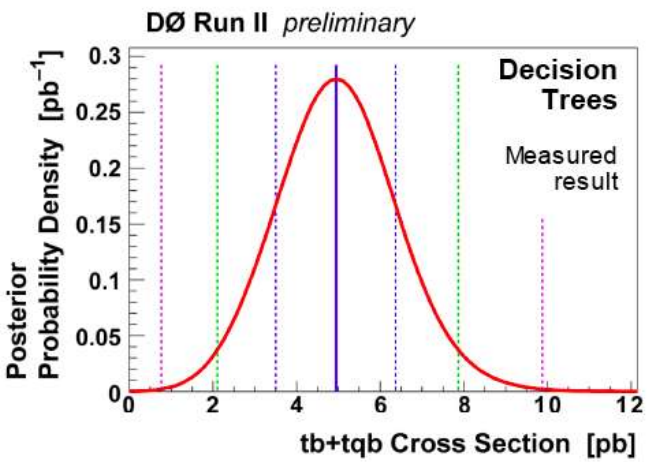
$$\text{Prob}(\text{data}|\sigma, a, b) = \sum_{i=1}^{Nbins} \text{Prob}(\text{data}_i|\sigma_i, a_i, b_i)$$

- Flat prior in signal cross-section (σ > 0)
- Shape and normalization systematics treated as nuisance parameters
- Bin-by-bin correlations between uncertainties and search channels taken into account
- Cross-section measurement obtained from peak position and uncertainty obtained from width



Results

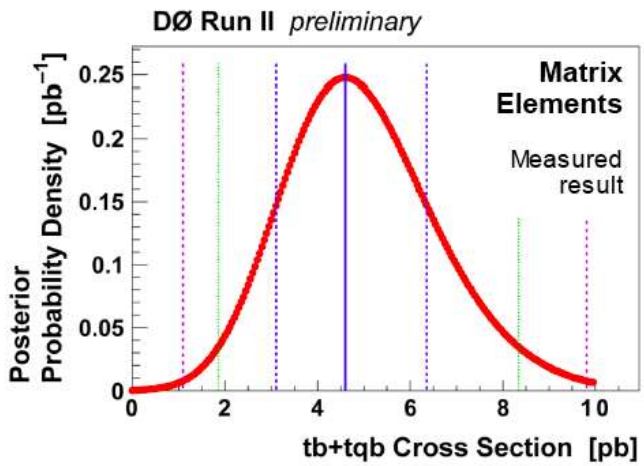
Decision Trees



$\sigma(tb+tqb) = 4.9 \pm 1.4$ pb

Measured p-value = 0.035 %
Measured significance = 3.4 σ
Compatibility with SM = 11%

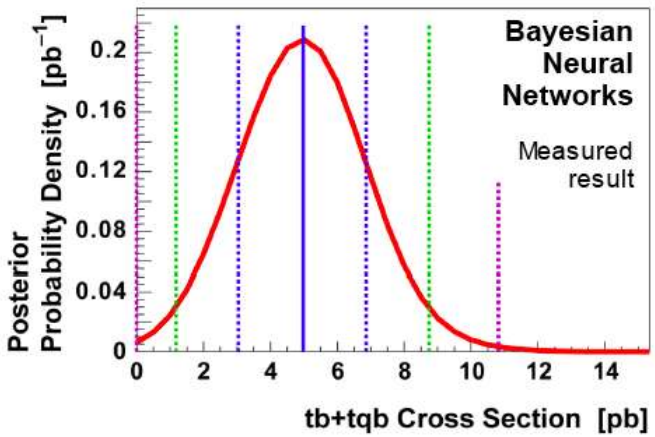
Matrix Elements



$\sigma(tb+tqb) = 4.6\ ^{+1.8}_{-1.5}$ pb

Measured p-value = 0.21 %
Measured significance = 2.9 σ
Compatibility with SM = 21%

Bayesian Neural Networks



$\sigma(tb+tqb) = 5.0 \pm 1.9$ pb

Measured p-value = 0.89 %
Measured significance = 2.4 σ
Compatibility with SM = 18%

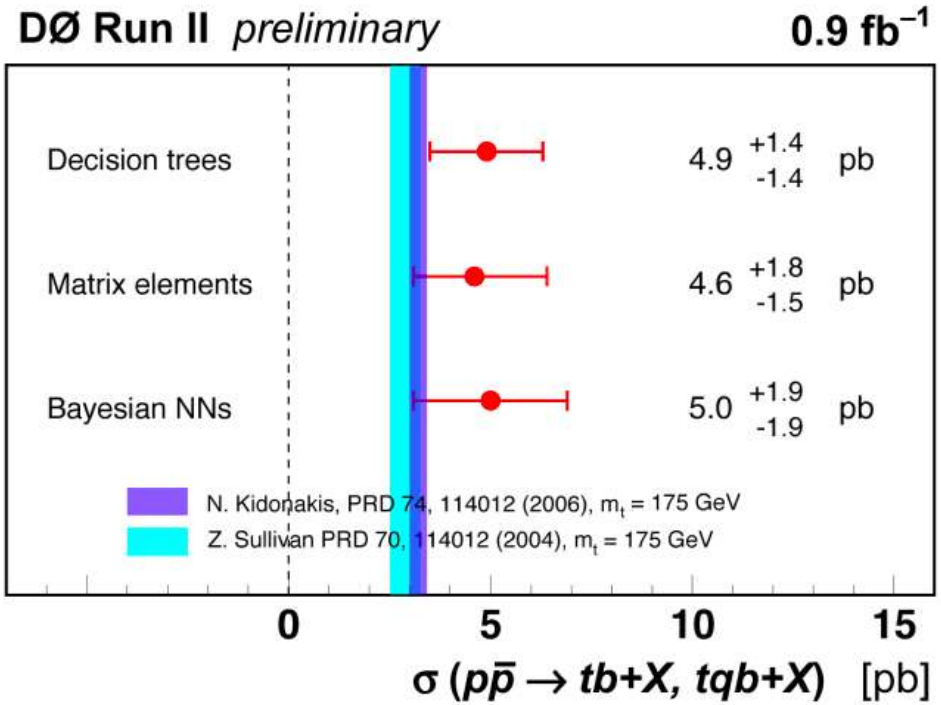
p-value: probability to measure a cross-section value at least as large as the one observed assuming background-only hypothesis.

significance: p-value expressed in terms of number of sigma

compatibility with SM: probability to measure a cross-section value at least as large as the one observed assuming SM cross-section

Results

Results from three different methods are consistent with each other



Currently working on combination of the three measurements

Correlation between measured cross sections			
	DT	ME	BNN
DT	100 %	39 %	57 %
ME		100 %	29 %
BNN			100 %

Measuring $|V_{tb}|$

In the SM:

- $V_{td}^2 + V_{ts}^2 + V_{tb}^2 = 1$
- Constraints on V_{td} and V_{ts} give $V_{tb} = 0.999100^{+0.000034}_{-0.000004}$

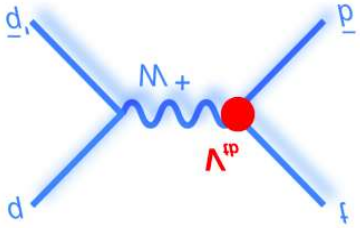
Physics beyond the SM:

- $V_{td}^2 + V_{ts}^2 + V_{tb}^2 < 1$
- No constraints on V_{tb}

- Use the measurement on the single top cross-section to obtain first direct measurement of $|V_{tb}|$
- General form of *Wtb* vertex

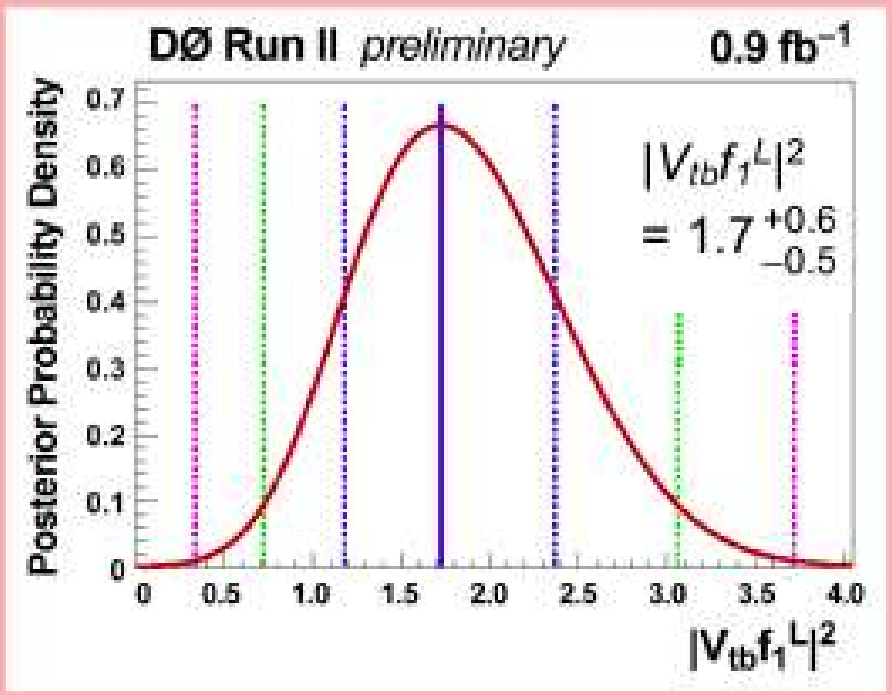
$$\Gamma^{\mu}_{Wtb} = -\frac{g}{\sqrt{2}} V_{tb} \left\{ \gamma^{\mu} \left[f_1^L P_L + f_1^R P_R \right] - \frac{i\sigma^{\mu\nu}}{M_W} (p_t - p_b)_{\nu} \left[f_2^L P_L + f_2^R P_R \right] \right\}$$

- No need to assume only **three quark families** or **CKM unitarity**
- Assume
  - SM top quark decay: $|V_{td}|^2 + |V_{ts}|^2 \ll |V_{tb}|^2$
  - Pure V-A coupling: $f_1^R = 0$
  - CP conservation: $f_2^L = f_2^R = 0$

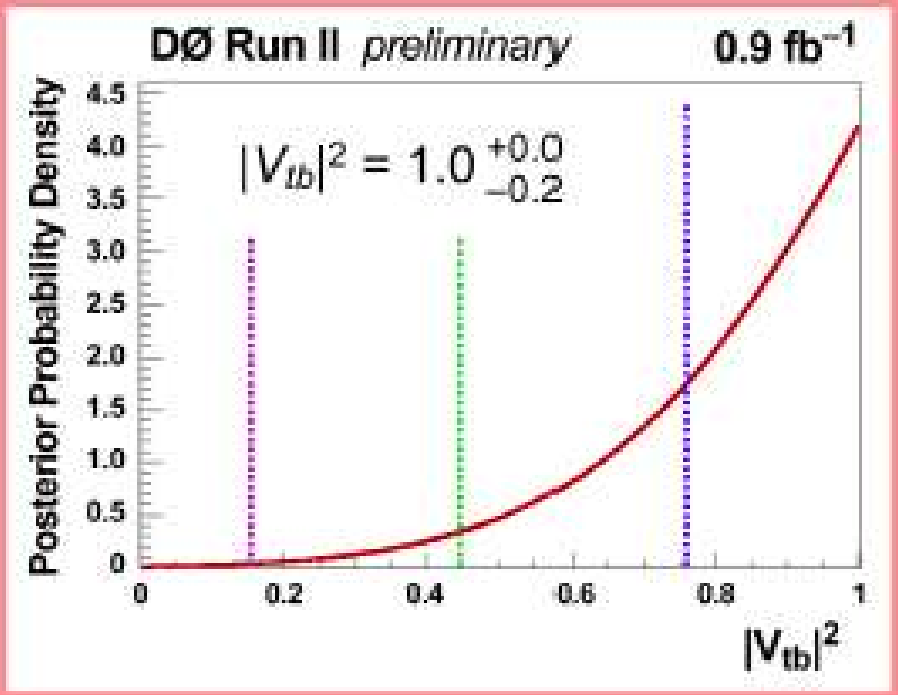


$|V_{tb}|$ measurement

Calculate a posterior probability density for $|V_{tb}|$



$|V_{tb}f_1^L| = 1.3 \pm 0.2$



$0.68 < |V_{tb}| \leq 1$ at 95% C.L.
(assuming $f_1^L = 1$)

Summary

- First Evidence for single top quark production at DØ

$$\sigma_{tb+tqb} = 4.9 \pm 1.4 \text{ pb} \quad (3.4\ \sigma \text{ significance})$$

- Direct measurement of CKM matrix element $|V_{tb}|$

$$0.68 < |V_{tb}| < 1.0 \text{ at } 95\% \text{ CL} \quad (f_1^L = 1)$$

- Challenging measurement – extracting a small signal from a large complex background
- Use three innovative event classification techniques:
  - Boosted decision trees
  - Matrix elements
  - Bayesian neural networks
- Results submitted to *Physical Review Letters*
- Open the door for studies of *Wtb* coupling and searches for new physics

Decision Tree - Boosting

- Technique to improve performance of a weak classifier
- Construct a linear combination of many filter functions
- Usually improves performance
- Averaging dilutes the piecewise nature of decision trees

Basic principle:

- Train a tree T_k
- Create a new tree: T_{k+1} = modify(T_k)
- Boosted result of event i: $T(i) = \sum_{n=1}^{N_{tree}} \alpha_k T_k(i)$

Adaptive boosting:

- Check which events are misclassified by T_k
- Increase weight of misclassified events
- Use modified training sample to build T_{k+1}

Decision Trees in single top analysis

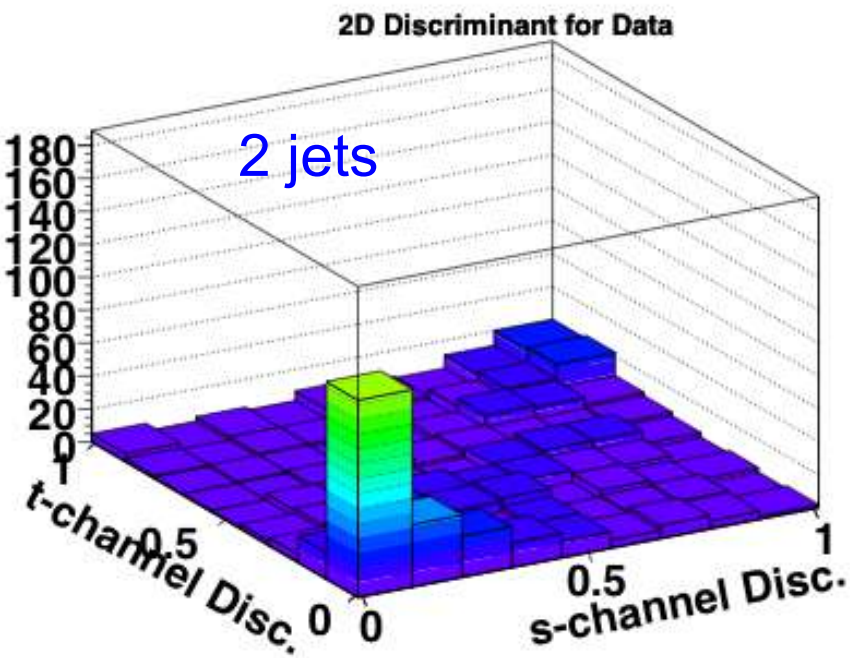
- Train 36 separate trees: (s,t,s+t) x (e,μ) x (2,3,4 jets) x (1,2 btags)
- For each signal train trees against sum of backgrounds ($t\bar{t}$, W+jets)
- Use 1/3 of MC for training, other 2/3 to calculate leaf purities

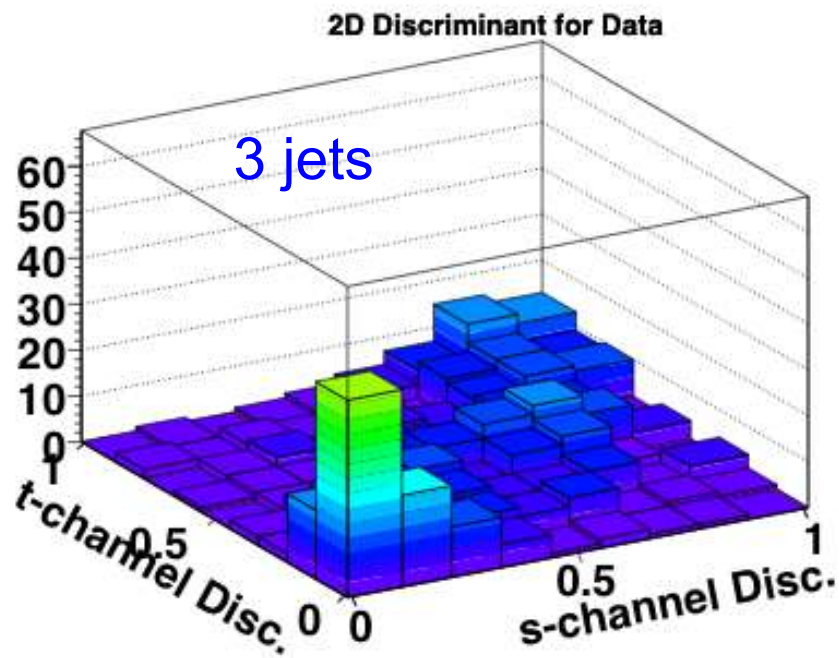
DT parameters used:

- Node splitting criterion: "Gini" factor (using "entropy" doesn't lead to better sensitivity)
- Minimum leaf size: 100 events (smaller value doesn't lead to statistically significant purity calculations, larger value degrades performance)
- Leaf class assignment: signal leaf if purity > 0.5
- Normalization: sum of signal and background event weights are each normalized to 1 (varying ratio of signal and background weights had little effect in sensitivity except in extreme cases where discrimination goes down)
- Using adaptive boosting "Adaboost": $\beta = 0.2$ (proportional to individual trees weight)
- Boosting cycle: 20 (i.e. Averaging 20 decision trees)

Matrix Element Method in single top

- Only use events with exactly 2 or 3 jets
- Use 4-momentum of all reconstructed objects: $\vec{x}=(\vec{p}_l, \vec{p}_{j_1}, \vec{p}_{j_2})$
- Take into account b-tagging information
- Matrix elements calculated using `Madgraph` leading-order matrix element generator
- Calculate discriminant for W+jets background only
- Use 2-D discriminant:





DT: Choosing input variables

- Detailed studies of Feynman diagrams for single top and background
- 49 discriminating variables were identified
  - Individual object kinematics
  - Angular correlation variables
  - Global event kinematics
- Adding variables does not degrade performance
- Reducing list of variables result in loss of sensitivity
- Same variables used for all search channels

Object Kinematics

p_T(jet1)
p_T(jet2)
p_T(jet3)
p_T(jet4)
p_T(best1)
p_T(notbest1)
p_T(notbest2)
p_T(tag1)
p_T(untag1)
p_T(untag2)

Angular Correlations

ΔR(jet1,jet2)
cos(best1,lepton)$_{\text{besttop}}$
cos(best1,notbest1)$_{\text{besttop}}$
cos(tag1,alljets)$_{\text{alljets}}$
cos(tag1,lepton)$_{\text{btaggedtop}}$
cos(jet1,alljets)$_{\text{alljets}}$
cos(jet1,lepton)$_{\text{btaggedtop}}$
cos(jet2,alljets)$_{\text{alljets}}$
cos(jet2,lepton)$_{\text{btaggedtop}}$
cos(lepton,Q(lepton)$\times z$)$_{\text{besttop}}$
cos(lepton,besttopframe)$_{\text{besttopCMframe}}$
cos(lepton,btaggedtopframe)$_{\text{btaggedtopCMframe}}$
cos(notbest,alljets)$_{\text{alljets}}$
cos(notbest,lepton)$_{\text{besttop}}$
cos(untag1,alljets)$_{\text{alljets}}$
cos(untag1,lepton)$_{\text{btaggedtop}}$

Event Kinematics

Aplanarity(alljets,W)
M(W,best1) ("best" top mass)
M(W,tag1) ("b-tagged" top mass)
H_T(alljets)
H_T(alljets−best1)
H_T(alljets−tag1)
H_T(alljets,W)
H_T(jet1,jet2)
H_T(jet1,jet2,W)
M(alljets)
M(alljets−best1)
M(alljets−tag1)
M(jet1,jet2)
M(jet1,jet2,W)
M_T(jet1,jet2)
M_T(W)
Missing E_T
p_T(alljets−best1)
p_T(alljets−tag1)
p_T(jet1,jet2)
Q(lepton)$\times\eta$(untag1)
$\sqrt{\hat{s}}$
Sphericity(alljets,W)

Matrix Element

Define discriminant:

$$D_S(\vec{x}) \equiv P(S|\vec{x}) = \frac{P_S(\vec{x})}{P_S(\vec{x}) + P_B(\vec{x})}$$

$P_S(\vec{x})$: Signal probability density function
$P_B(\vec{x})$: Bkg probability density function

Probability density functions are calculated numerically as the normalized differential cross-section for each processes:

$$P(\vec{x}) = \frac{1}{\sigma} \times \frac{\partial \sigma}{\partial \vec{x}}$$

where

$$d\sigma(\vec{x}) = \sum_{i,j} \int d\vec{y} \left[f_i(q_1, Q^2) \times f_j(q_2, Q^2) \times \frac{\partial \sigma_{i,j}(\vec{y})}{\partial \vec{y}} \times W(\vec{x}, \vec{y}) \times \Theta_{\text{parton}}(\vec{y}) \right]$$

parton PDF

hard scatter $\sim |M|^2$

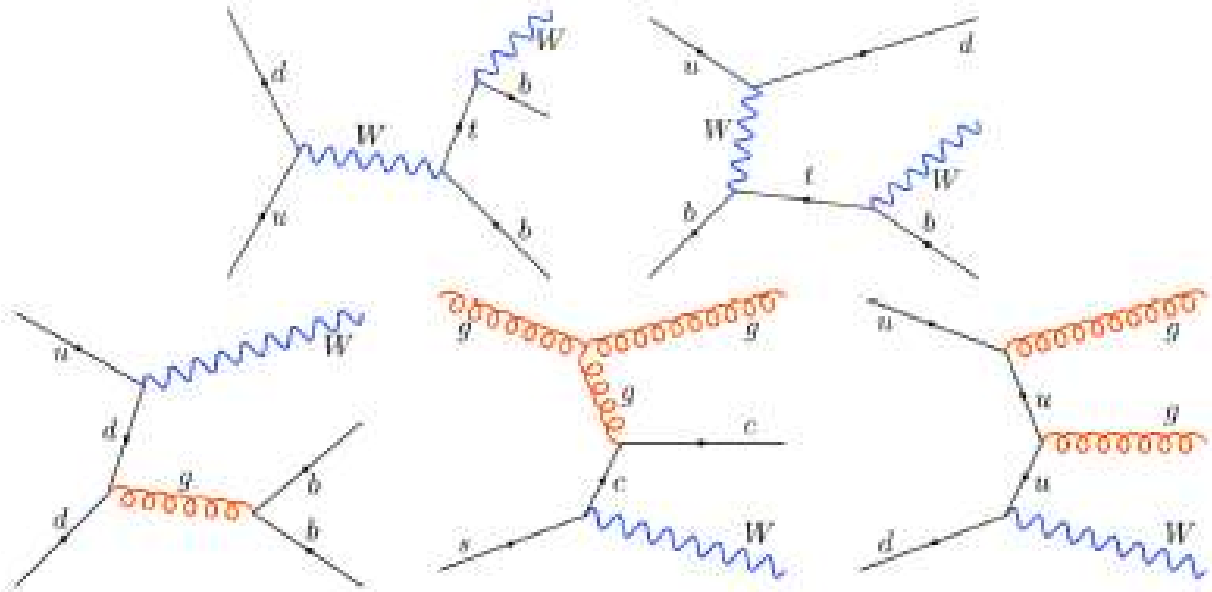
transfer function

parton-level cuts

2-jet channels

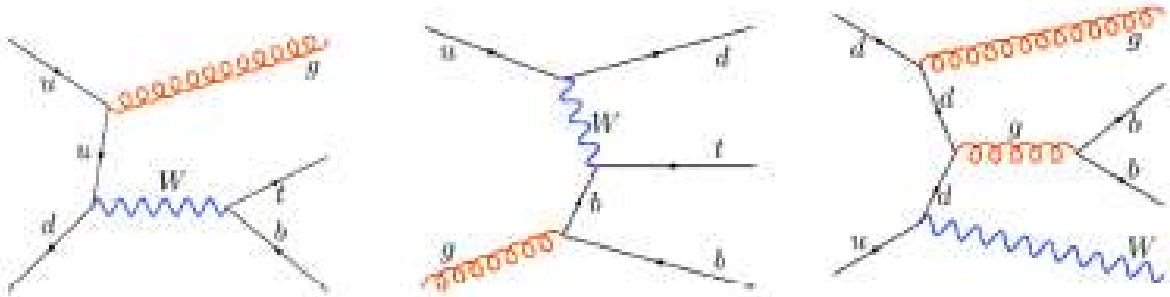
tb *tq*

Wbb *Wcg* *Wgg*



3-jet channels

tbg *tdb* *Wbbg*



Bayesian NN

Treat the training of a neural network as a problem of inference and solve using Bayes' theorem.

Interpret output of ONE neural network as a probability that the target t =1.

$$y(x,w)=p(t=1|x,w)$$

x: vector of input variables
w: vector of parameter values that define a neural network

Define posterior probability density over all possible weight sets (i.e. over all possible neural networks) as:

$$p(w|t,x)=\frac{p(t|x,w)\,p(w)}{p(t|x)}$$

$p(t|x,w)$: likelihood
$p(w)$: prior density
$p(t|x)$: probability of set of targets t given data x

Estimate of the probability that an event in data x' belongs to signal class t =1:

$$\boxed{\bar{y}(x'|t,x)=\int y(x',w)\,p(w|t,x)\,dw}$$

(weighted average of all possible neural networks)

In practice, use Monte Carlo Markov chain as an approximation:

$$\bar{y}(x'|t,x)\approx\frac{1}{N}\sum_{i=1}^{N} y(x',w_i)$$

Bayesian Neural Network

Advantages:

- In principle, provides best estimate of $p(1|x)$
- Less prone to over-training due to averaging
- Network structure less important, just use a large network

Disadvantages:

- Computationally intensive

Example of use in the search for single top:

- 24 input variables (currently limited by software package used)
- Networks with 40 hidden nodes
- Run 800 iterations of training. Each iteration consists in the average of the outputs of 20 training epochs.
- Final network is taken as the average of the last 100 iterations