Boosted Decision Trees, an Alternative to Artificial Neural Networks

B. Roe, University of Michigan

Collaborators on this work

- H. Yang, J. Zhu, University of Michigan
- Y. Liu, I. Stancu, University of Alabama
- G. McGregor, Los Alamos National Lab.
- “and the good people of Mini-BooNE”

What I will talk about

I will review Artificial Neural Networks (ANN), introduce the new technique of boosted decision trees and then, using the miniBooNE experiment as a test bed compare the techniques for distinguishing signal from background

Outline

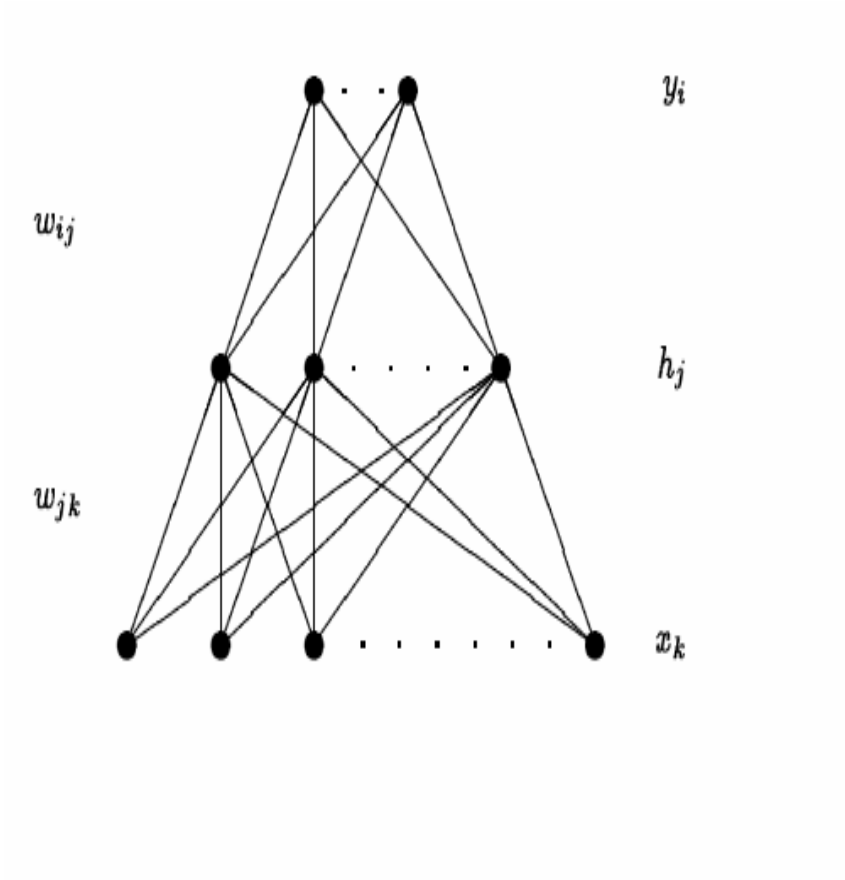
- What is ANN?
- What is Boosting?
- What is MiniBooNE?
- Comparisons of ANN and Boosting for the MiniBooNE experiment

Artificial Neural Networks

- Use to classify events, for example into “signal” and “noise/background”.
- Suppose you have a set of “feature variables”, obtained from the kinematic variables of the event

Neural Network Structure

- Combine the features in a non-linear way to a “hidden layer” and then to a “final layer”
- Use a training set to find the best w_{ik} to distinguish signal and background



A one hidden layer feed-forward neural network architecture.

Feedforward Neural Network--I

$$F_i(\vec{x}) = g\left[\frac{1}{T}\sum_j \omega_{ij} g\left(\frac{1}{T}\sum_k \omega_{jk} x_k + \theta_j\right) + \theta_i\right].$$

This corresponds to a network where the x_k, the input layer, are combined with weights ω_{jk} and offsets θ_j to give a hidden layer $h_j = g\left(\frac{1}{T}\sum_k \omega_{jk} x_k + \theta_j\right)$, and the h_j then combined in a similar manner to give an output layer y_i. Sometimes there are several hidden layers, defined in the obvious way by iterating the procedure given in the previous equation. The hidden layer enables non-linear modeling of the input data.

Feedforward Neural Network--II

$$F_i(\overrightarrow{x}) = g\left[\frac{1}{T}\sum_j \omega_{ij} g\left(\frac{1}{T}\sum_k \omega_{jk} x_k + \theta_j\right) + \theta_i\right].$$

T is a system parameter which scales the size of F_i.

The weights ω_{ij} and ω_{jk} are the parameters to be fitted to the data distributions and $g(x)$ is the non-linear neuron activation function, typically of the "sigmoid" form,

$$g(x) = \frac{1}{2}[1 + \tanh(x)] = (1 + e^{-2x})^{-1}.$$

Determining the weights

- Suppose want signal events to give output =1 and background events to give output=0
- Mean square error given N_p training events with desired outputs o_i either 0 or 1, and ANN results t_i.

$$E = \frac{1}{2N_p} \sum_{p=1}^{N_p} \sum_i (o_i^{(p)} - t_i^{(p)})^2$$

Back Propagation to Determine Weights

$$w_{t+1} = w_t + \Delta w_t,$$

$$where$$

$$\Delta w_t = -\eta \frac{\partial E}{\partial w}$$

$$+\alpha \Delta w_{t-1}, "momentum_term_to_stabalize"$$

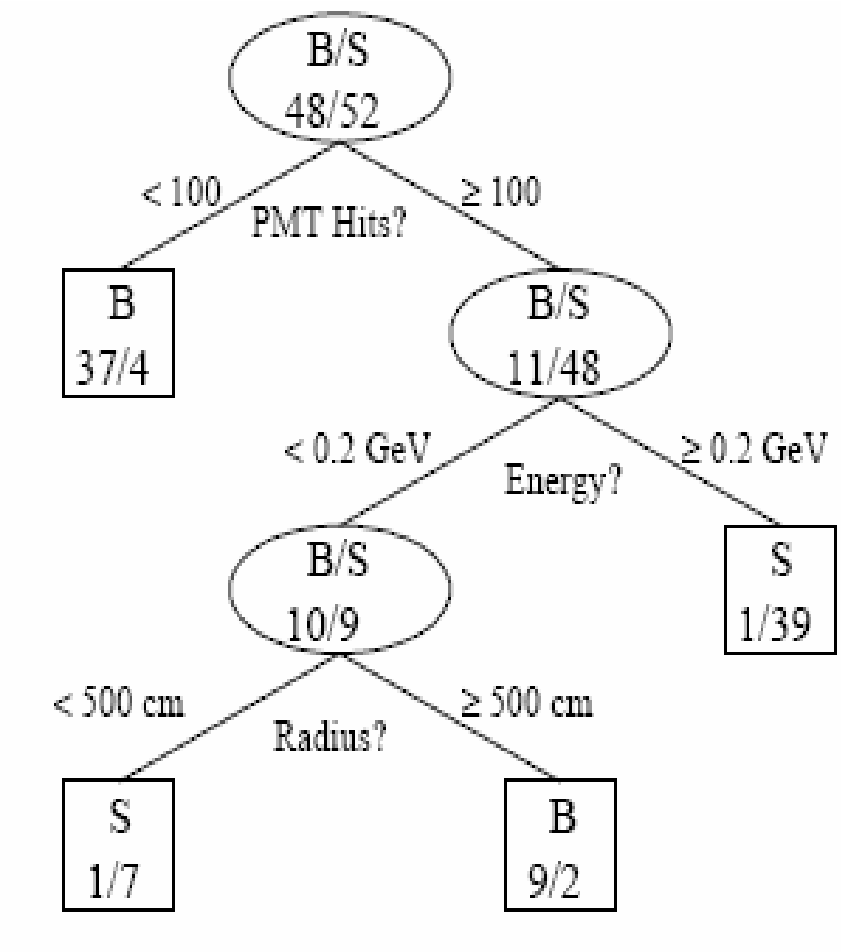
$$+\sigma, "noise_term_to_avoid_local_\min ima"$$

Boosted Decision Trees

- What is a decision tree?
- What is “boosting the decision trees”?
- Two algorithms for boosting.

Decision Tree

- Go thru all PID variables and find the best to split events
- For each of the two subsets repeat the process
- Continuing a tree is built. Ending nodes are called leaves



Select signal and background leaves

- Assume an equal weight of signal and background training events
- If more than ½ of the weight in a leaf corresponds to signal events, it is a signal leaf; otherwise it is a background leaf
- Signal events on a background leaf or background events on a signal leaf are misclassified

Criterion for “best” split

- Purity:

$$P = \frac{\sum_s W_s}{\sum_s W_s + \sum_b W_b}$$

- Gini: Note Gini is 0 for all signal or all background

$$Gini = (\sum_{i=1}^{n} W_i) P(1-P)$$

- Criterion is to minimize Gini_left + Gini_right

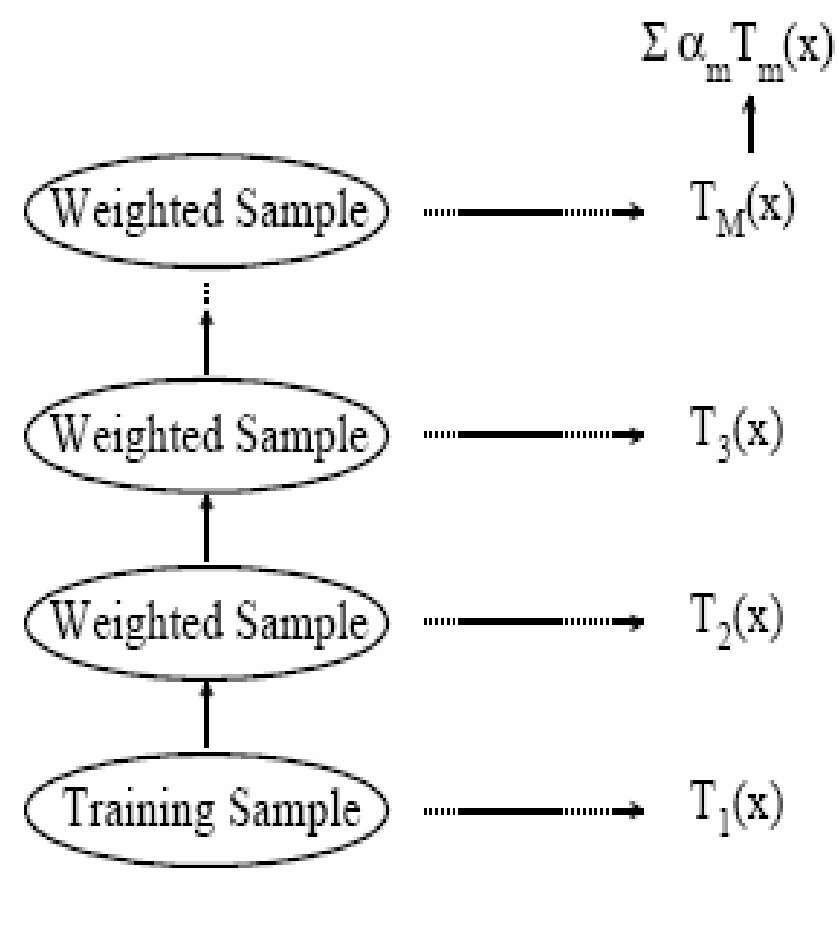
Criterion for next branch to split

- Pick the branch to maximize the change in Gini,

$$Criterion = Gini_{father} - Gini_{left-daughter} - Gini_{right-daughter}$$

Boosting the Decision Tree

- Give the training events misclassified under this procedure a higher weight and build a new tree
- Continuing, build perhaps 1000 trees and average the results (1 if signal leaf, -1 if bkrd leaf)



Training and Testing Samples

- An ANN or boosted decision tree set are trained with a sample of events—the training sample.
- However, it would be biased to use this sample to evaluate how good the classifier is. It is optimized for this individual set.
- A new set, the testing sample is used to evaluate the performance of the classifier after tuning. All results quoted here are for the testing sample.

How determine change of weights

- Two commonly used algorithms for boosting the decision trees are:

  AdaBoost

  ε-boost (or “shrinkage”)

Definitions

- X_i = set of particle ID variables for event i
- Y_i = 1 if the ith event is signal,-1 if bkrd
- W_i= weight of ith event
- $T_m(x_i)$= 1 if ith event lands on signal leaf, and -1 if ith event lands on bkrd leaf
- $I(y_i \neq T_m(x_i))$ =1 if misclassify, 0 if classify correctly

AdaBoost

- Define

$$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq T_m(x_i))}{\sum_{i=1}^{N} w_i}$$

$$\alpha_m = \beta \ln((1 - err_m)/err_m)$$

$$\beta = 1 usual, 0.5 our_use$$

- Change weight for misclassified events

$$w_i \rightarrow w_i e^{\alpha_m I(y_i \neq T_m(x_i))}$$

Example

- Suppose the weighted error rate is 10%, i.e., err=0.1
- Then alpha = (1/2)ln((1-.1)/.1)= 1.1
- Weight of a misclassified event is multiplied by exp(1.1)~3

Scoring events with AdaBoost

- Renormalize weights

$$w_i \rightarrow w_i \Big/ \sum_{i=1}^{N} w_i$$

- Score by summing over trees

$$T(x) = \sum_{m=1}^{N_{tree}} \alpha_m T_m(x)$$

Epsilon Boost (shrinkage)

- After tree *m*, change weight of misclassified events, typical ε ~0.01

$$w_i \rightarrow w_i e^{2\varepsilon I(y_i \neq T_m(x_i))}$$

- Renormalize weights

$$w_i \rightarrow w_i / \sum_{i=1}^{N} w_i$$

- Score by summing over trees

$$T(x) = \sum_{m=1}^{N_{tree}} \varepsilon T_m(x)$$

Comparison of methods

- Epsilon boost changes weights a little at a time
- AdaBoost can be shown to try to optimize each change of weights. Lets look a little further at that

AdaBoost Optimization

$f_i(x)$ =classifier, with values $= +a_i$ or $-a_i$, with a_i a positive constant

$F(x) = \sum_{i=1}^{N} f_i(x)$. (sum over trees.)

Can show that AdaBoost minimizes the expectation value, $E(e^{-yF(x)})$ by a series of Newton-like updates. Furthermore, the minimum value of $E(e^{-yF(x)})$ is $F(x) = \frac{1}{2}\ln\left(\frac{P(y=1|x)}{P(y=-1|x)}\right)$., which is $\frac{1}{2}$ the log-odds of the probabilty that $Y = 1$, given x. This minimization is closely related to maximizing the negative binomial log likelihood (cross-entropy). They can both be shown to have the same minimizer. Further, with y^* 1 or 0, and $p(x)$ = probability that $y^* = 1$ given x, then with $F(x) = \frac{1}{2}\ln\frac{p(x)}{1-p(x)}$ it can be shown that $e^{-yF(x)} = \frac{|y^*-p(x)|}{\sqrt{p(x)(1-p(x))}} = \chi$-statistic.

AdaBoost Fitting is Monotone

χ is a gentler alternative to a χ^2 minimizer. For finite samples, cross-entropy and ADA-boost give slightly different loss functions for differing values of yf, but both can be shown to be monotone and smooth. The criteria will continually drive estimates toward purer solutions (in terms of probability estimates). AdaBoost is quite resistant to overfitting. (ϵ-boosting, can sometimes reach a broad maximum and slowly overfit when plotted vs number of trees.)

If minimize $E(y - F(x))^2$, then it is NOT monotone. Classifications that are "too correct" are penalized as much as misclassification errors.

References

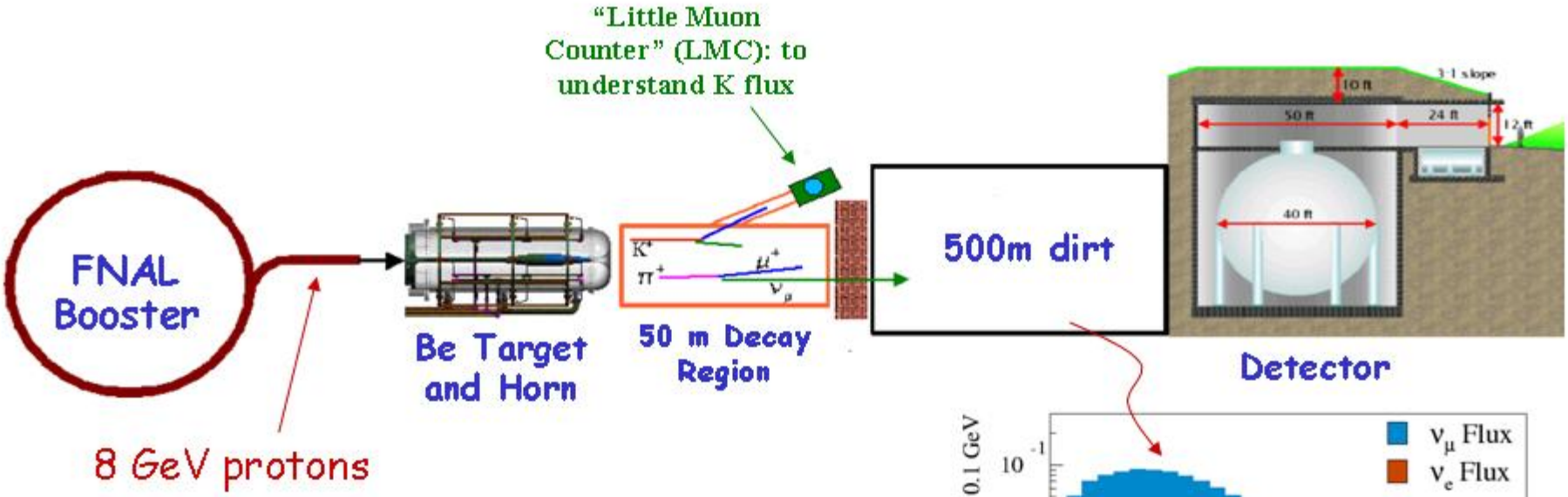
- R.E. Schapire ``The strength of weak learnability.'' *Machine Learning* **5** (2), 197-227 (1990). First suggested the boosting approach for 3 trees taking a majority vote
- Y. Freund, ``Boosting a weak learning algorithm by majority'', *Information and Computation* **121** (2), 256-285 (1995) Introduced using many trees
- Y. Freund and R.E. Schapire, ``Experiments with an new boosting algorithm, *Machine Learning: Proceedings of the Thirteenth International Conference,* Morgan Kauffman, SanFrancisco, pp.148-156 (1996). Introduced AdaBoost
- J. Friedman, T. Hastie, and R. Tibshirani, ``Additive logistic regression: a statistical view of boosting'', *Annals of Statistics* **28** (2), 337-407 (2000). Showed that AdaBoost could be looked at as successive approximations to a maximum likelihood solution.
- T. Hastie, R. Tibshirani, and J. Friedman, ``The Elements of Statistical Learning'' Springer (2001). Good reference for decision trees and boosting.

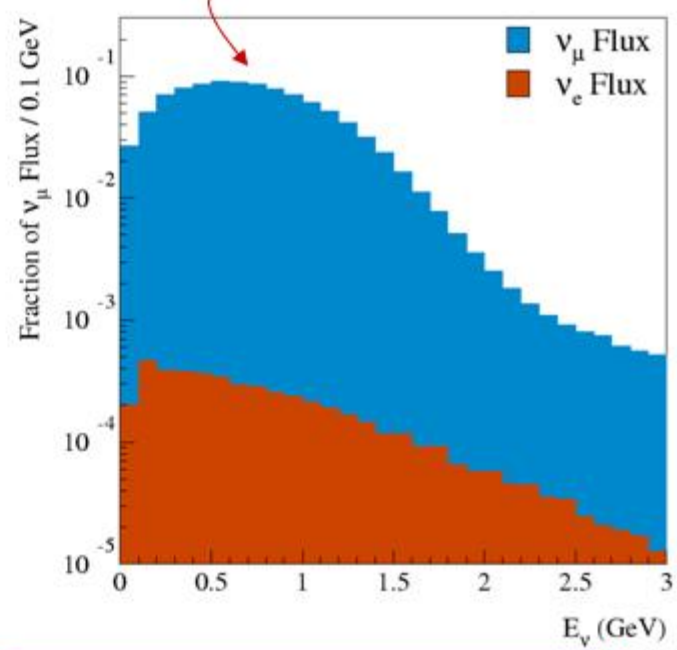
The MiniBooNE Experiment

The MiniBooNE Collaboration

Y.Liu, I.Stancu
University of Alabama
S.Koutsoliotas
Bucknell University
E.Hawker, R.A.Johnson, J.L.Raaf
University of Cincinnati
T.Hart, R.H.Nelson, E.D.Zimmerman
University of Colorado
A.A.Aguilar-Arevalo, L.Bugel, J.M.Conrad, J.Link, J.Monroe, D.Schmitz, M.H.Shaevitz, M.Sorel, G.P.Zeller
Columbia University
D.Smith
Embry Riddle Aeronautical University
L.Bartoszek, C.Bhat, S.J.Brice, B.C.Brown, D.A.Finley, R.Ford, F.G.Garcia, P.Kasper, T.Kobilarcik, I.Kourbanis, A.Malensek, W.Marsh, P.Martin, F.Mills, C.Moore, E.Prebys, A.D.Russell, P.Spentzouris, R.Stefanski, T.Williams
Fermi National Accelerator Laboratory
D.Cox, A.Green, T.Katori, H.Meyer, R.Tayloe
Indiana University
G.T.Garvey, C.Green, W.C.Louis, G.McGregor, S.McKenney, G.B.Mills, H.Ray, V.Sandberg, B.Sapp, R.Schirato, R.Van de Water, N.L.Walbridge, D.H.White
Los Alamos National Laboratory
R.Imlay, W.Metcalf, S.Ouedraogo, M.Sung, M.O.Wascko
Louisiana State University
J.Cao, Y.Liu, B.P.Roe, H.J.Yang
University of Michigan
A.O.Bazarko, P.D.Meyers, R.B.Patterson, F.C.Shoemaker, H.A.Tanaka
Princeton University
P.Nienaber
St. Mary's University of Minnesota
B.T.Fleming
Yale University

Neutrino Beam

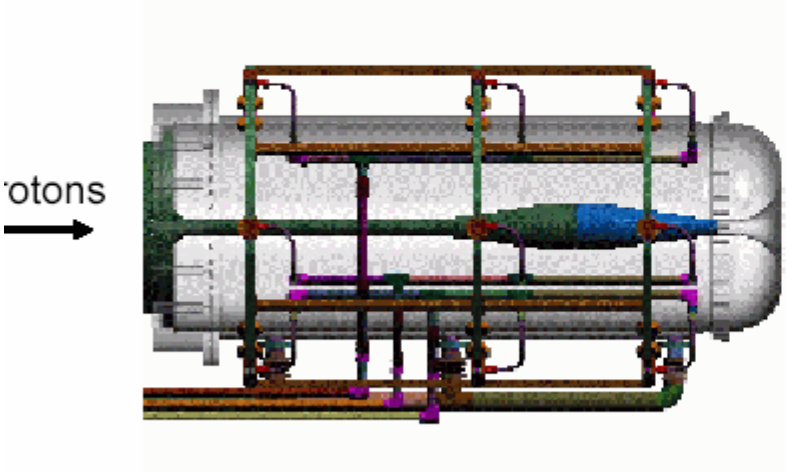




- Proton flux ~ 6E16 p/hr (goal 9E16 p/hr)
  - ~ 1 detected neutrino/minute
  - L/E ~ 1

Target and magnetic horn

ncreases neutrino intensity by 7x



70 kA in 140 μsec pulses @ 5 Hz

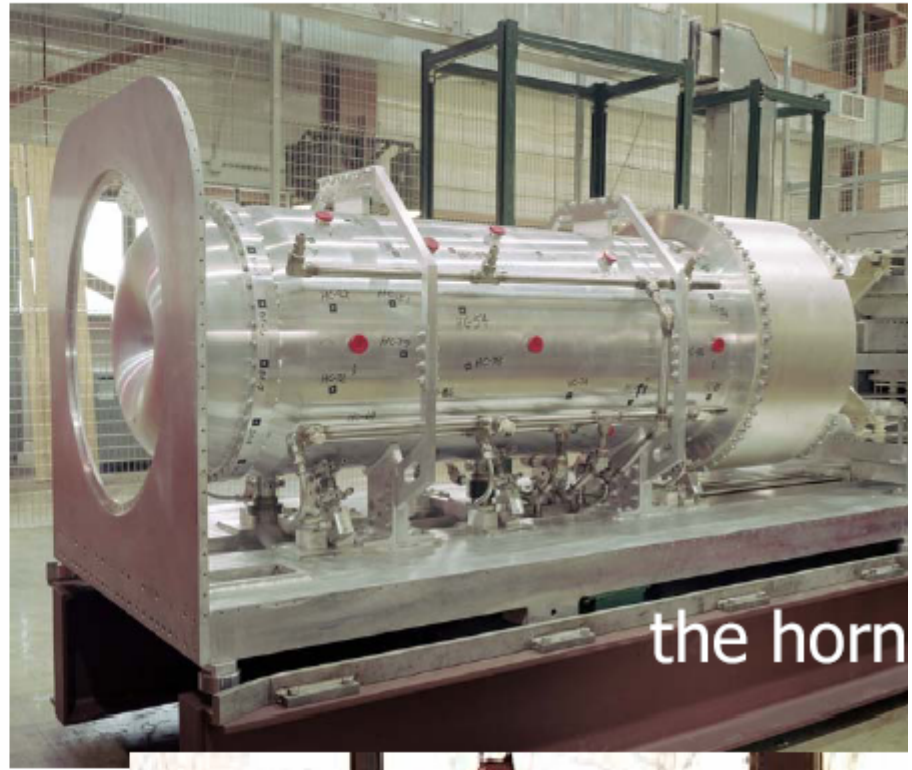
Currently positive particles are being focused, selecting neutrinos $\pi^+ \rightarrow \mu^+ \nu_\mu$

the horn current can be reversed to select antineutrinos $\pi^- \rightarrow \mu^- \bar{\nu}_\mu$

Prior to run, tested to 10M pulses

Ran 97 million pulses before ailing

World's longest-lived horn

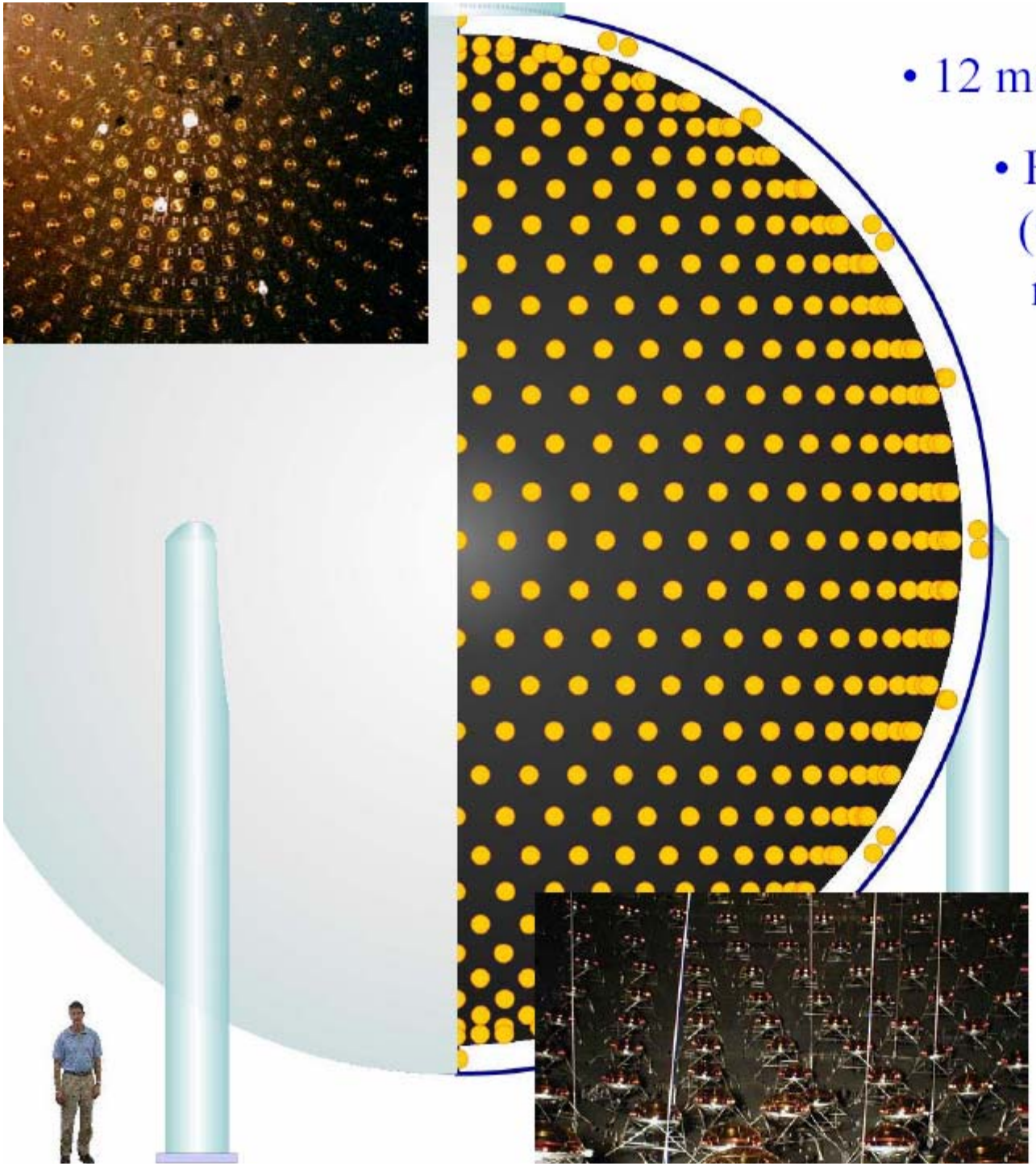


Horn now being replaced

- It acquired eight times more pulses than any previous horn anywhere.

Horn and Target Region

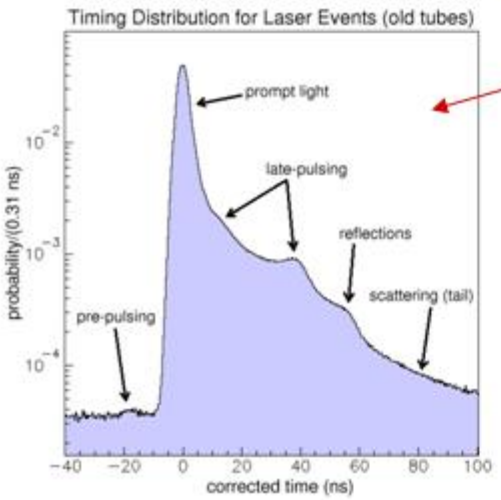
- Primary beam position monitor: air multiwire
- Target: 71 cm beryllium metal (1.7 λ_0), resides inside horn
- Horn:
  - Inner conductor thickness: 3 mm
  - Outer conductor thickness: 25 mm
  - Peak current: 170 kA
  - Pulse width: 140 μs
  - Voltage: ~4 kV



- 12 meter diameter sphere
  - Filled with 950,000 liters (900 tons) of very pure mineral oil
    - Light tight inner region with 1280 photomultiplier tubes
    - Outer veto region wit 241 PMTs.
    - Oscillation Search Method:

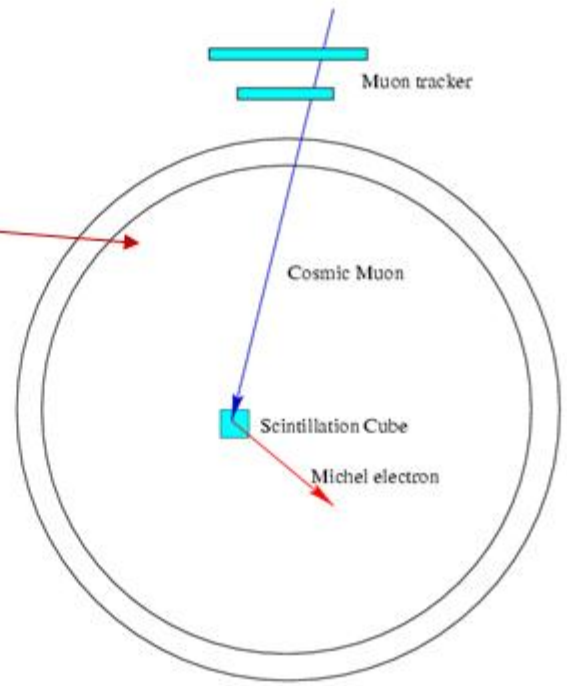
      Look for ν_e events in a pure ν_μ beam

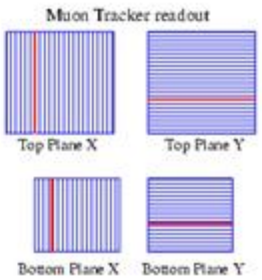
Characterizing the Detector

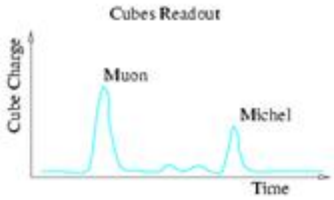


- Laser Calibration
  - Laser pulses illuminate one of 4 flasks which scatter light isotropically
  - Used to understand PMT response

- Cosmic Muons
  - Muon Tracker used in conjunction with "cubes" to trigger on a particular endpoint (energy)
  - Vital in understanding energy scale







aration of ν_μ from ν_e events

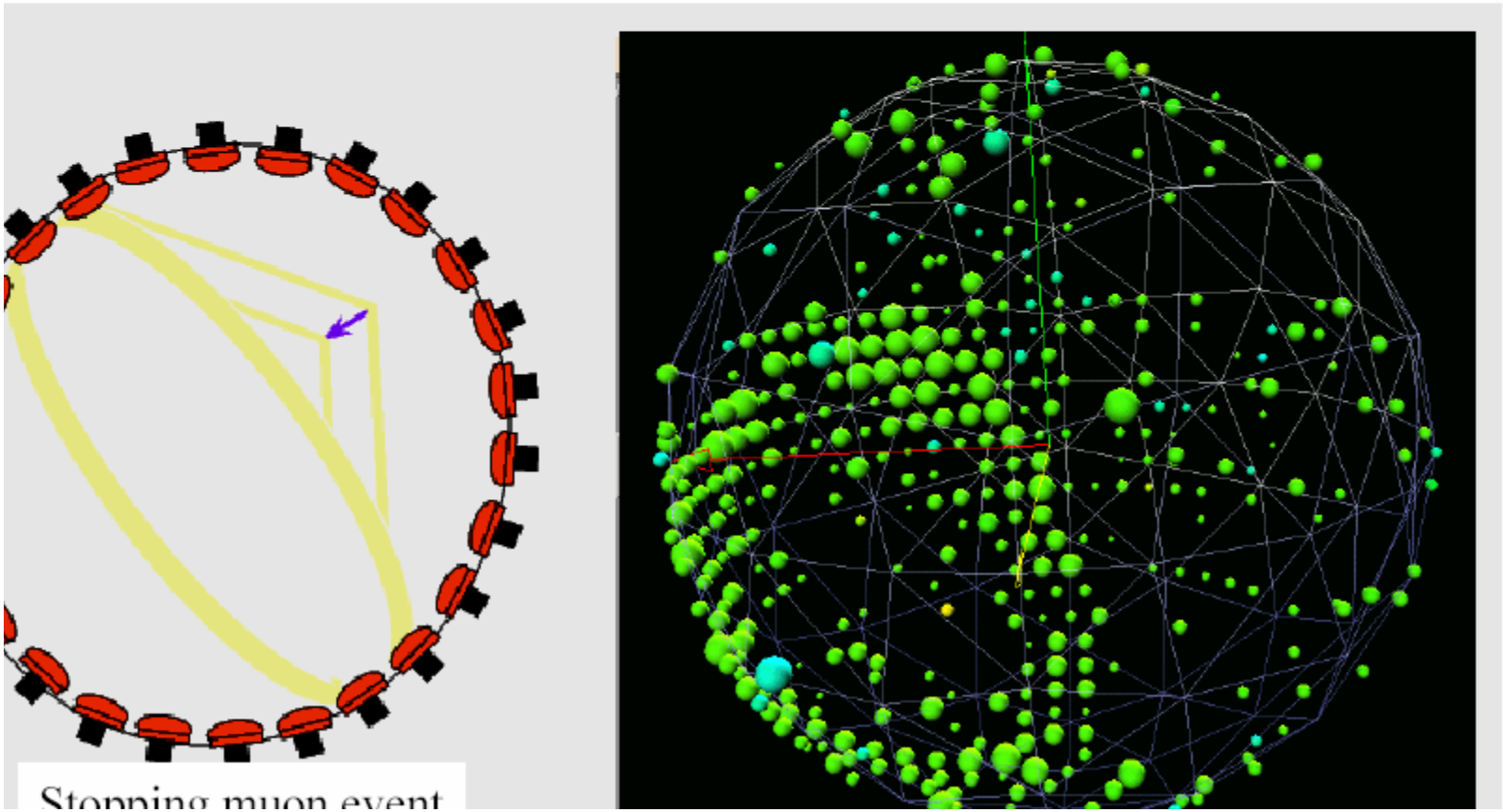
Exiting ν_μ events fire the veto

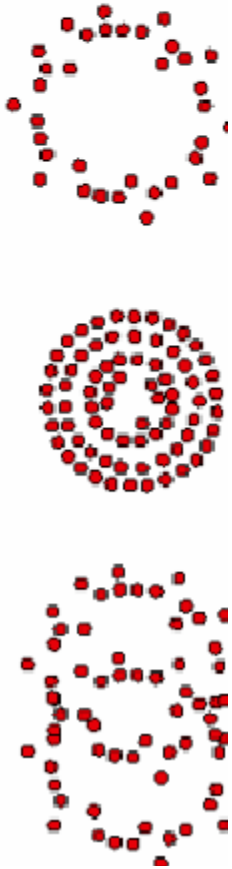
Stopping ν_μ events have a Michel electron after a few μsec

Also, scintillation light with longer time constant $\Rightarrow$ enhanced for slow pions and p

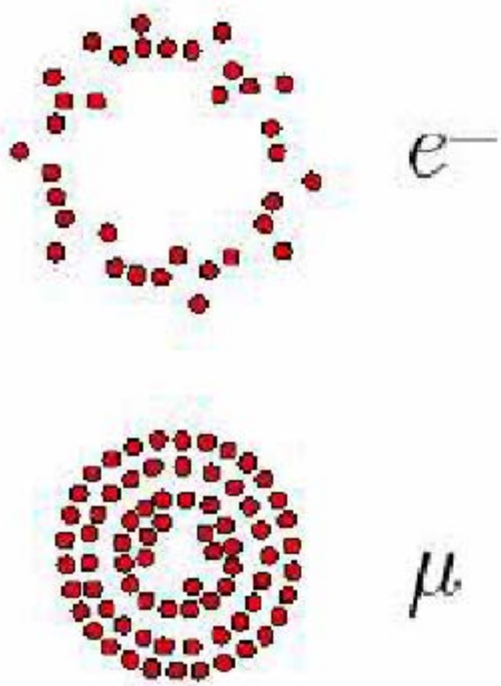
Čerenkov rings from outgoing particles

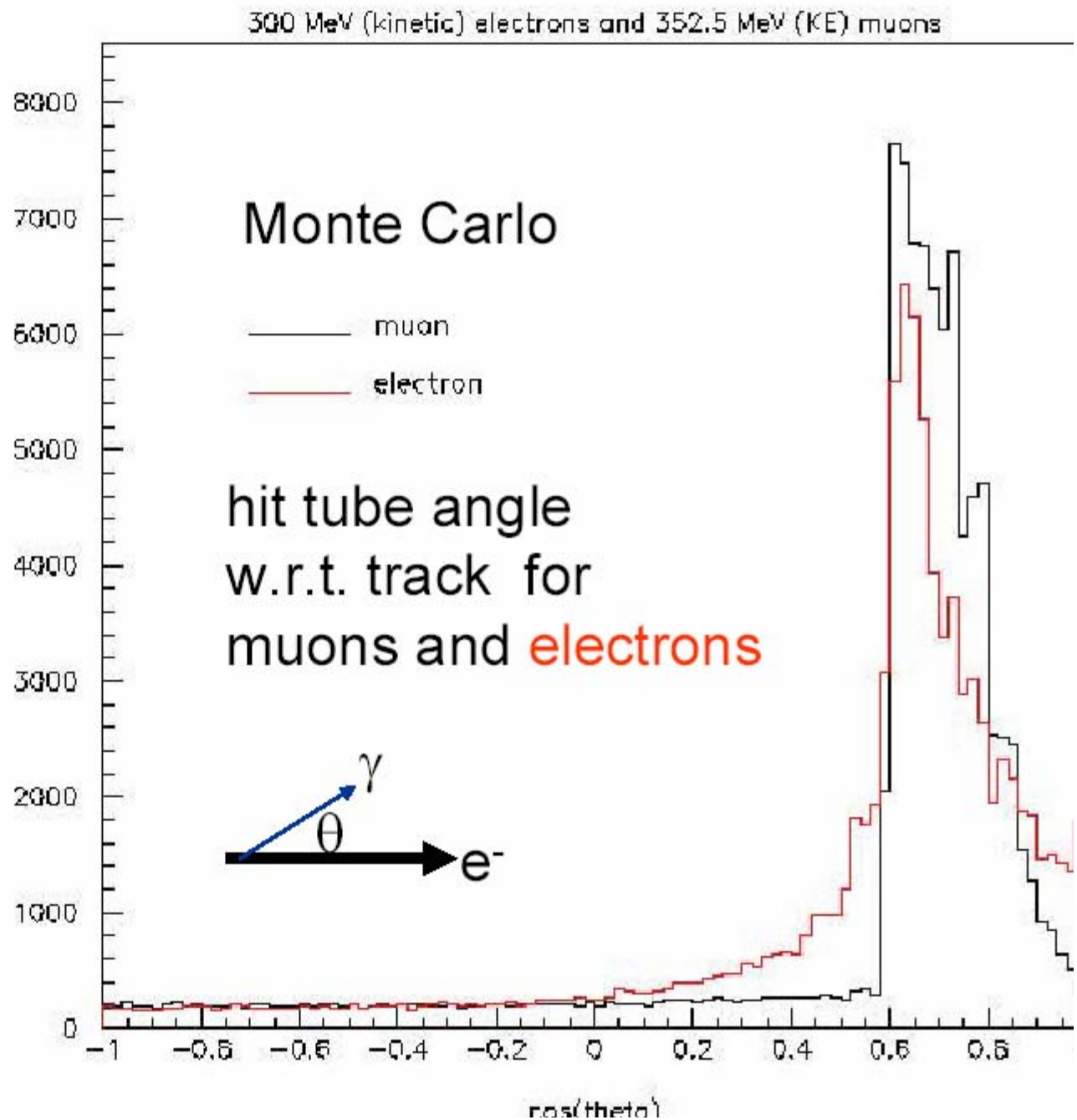
- Shows up as a ring of hits in the phototubes mounted inside the MiniBooNE sphere
- Pattern of phototube hits tells the particle type





electrons are fuzzy
muons are sharp

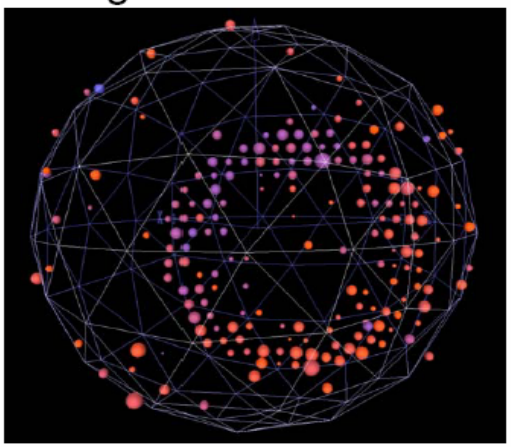




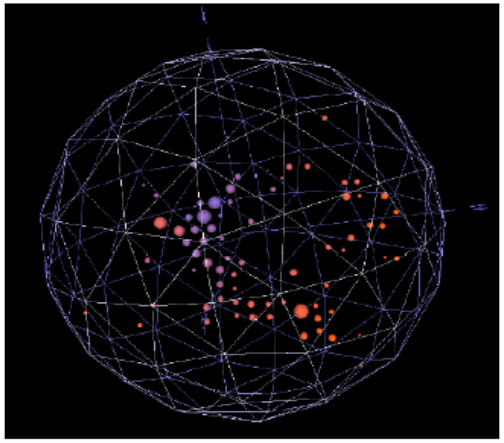
Examples of data events

Pattern of hit tubes (with charge and time information)
allows reconstruction of track location and direction
and separation of different event types.

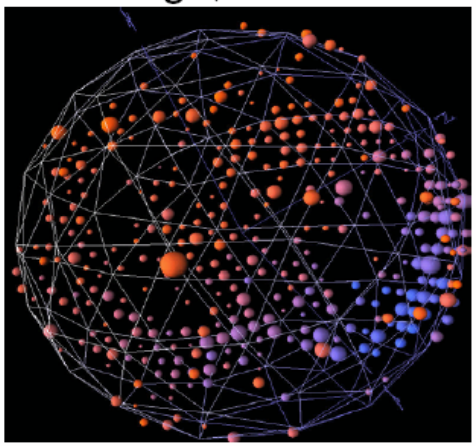
e.g. candidate events: size = charge, color = time



muon
from ν_μ interaction



Michel electron
from stopped μ decay
after ν_μ interaction



$\pi^0 \rightarrow$ two photons
from ν_μ interaction

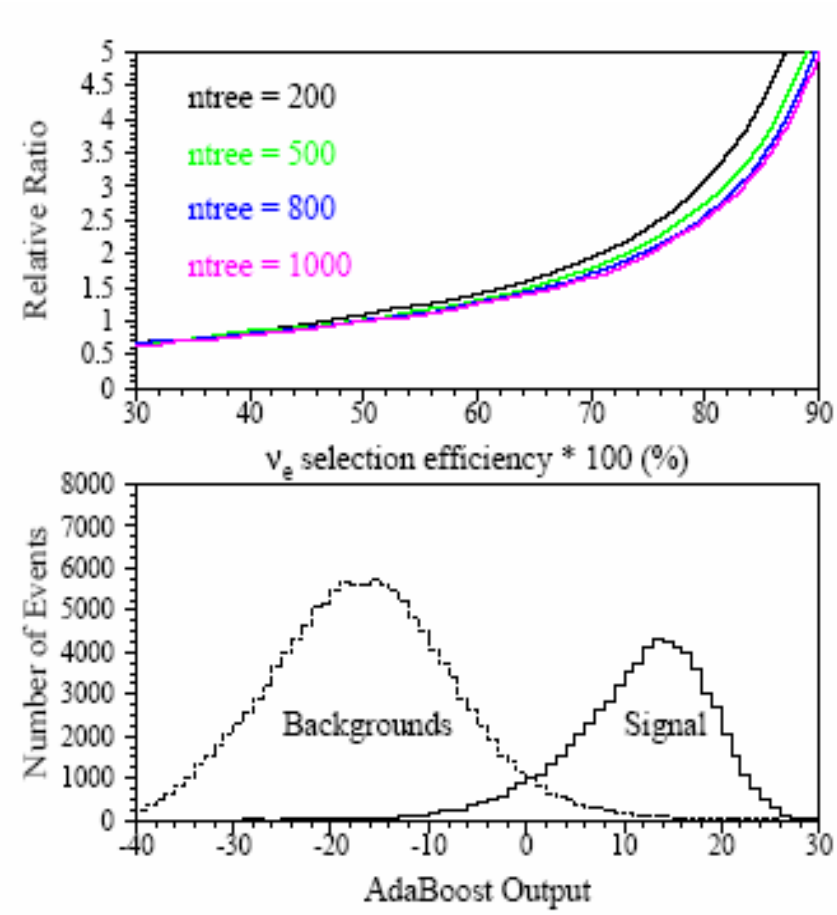
Comparison of Neural Nets and Boosting

Numerical Results

- There are 2 reconstruction-particle id packages used in MiniBooNE, rfitter and sfitter
- The best results for ANN and Boosting used different numbers of variables, 21 or 22 being best for ANN and 50-52 for boosting
- Results quoted are ratios of background kept by ANN to background kept for boosting, for a given fraction of signal events kept
- Only relative results are shown

Boosting Results versus Ntree

- Top: N_bkrd divided by N_bkrd for 1000 trees and 50% nue selection efficiency vs nue efficiency for CCQE events.
- Bottom: AdaBoost output for background and signal events for 1000 trees

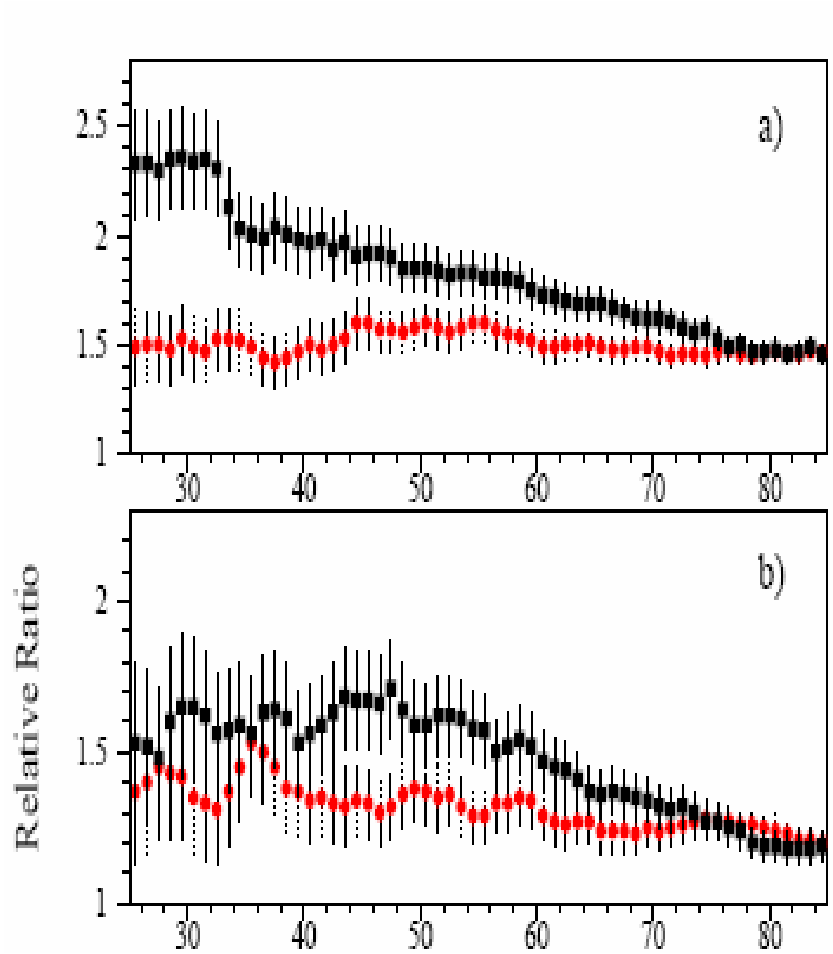


Numerical Results from rfitter

- Train against all kinds of backgrounds—21 ANN variables and 52 boosting variables: for 40-60% of signal kept, the ratio of ANN to boosting background varied from 1.5 to 1.8
- Train against nc pi0 background—22 ANN variables and 52 boosting variables: for 40-60% of signal kept, the ratio of ANN to boosting background varied from 1.3 to 1.6

Comparison of Boosting and ANN

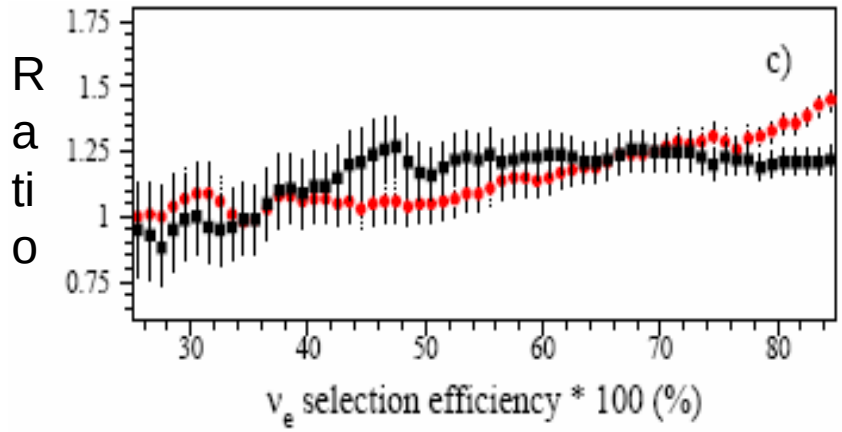
- A. Bkrd are cocktail events. Red is 21 and black is 52 training var.
- B. Bkrd are pi0 events. Red is 22 and black is 52 training variables
- Relative ratio is ANN bkrd kept/Boosting bkrd kept



Percent nue CCQE kept

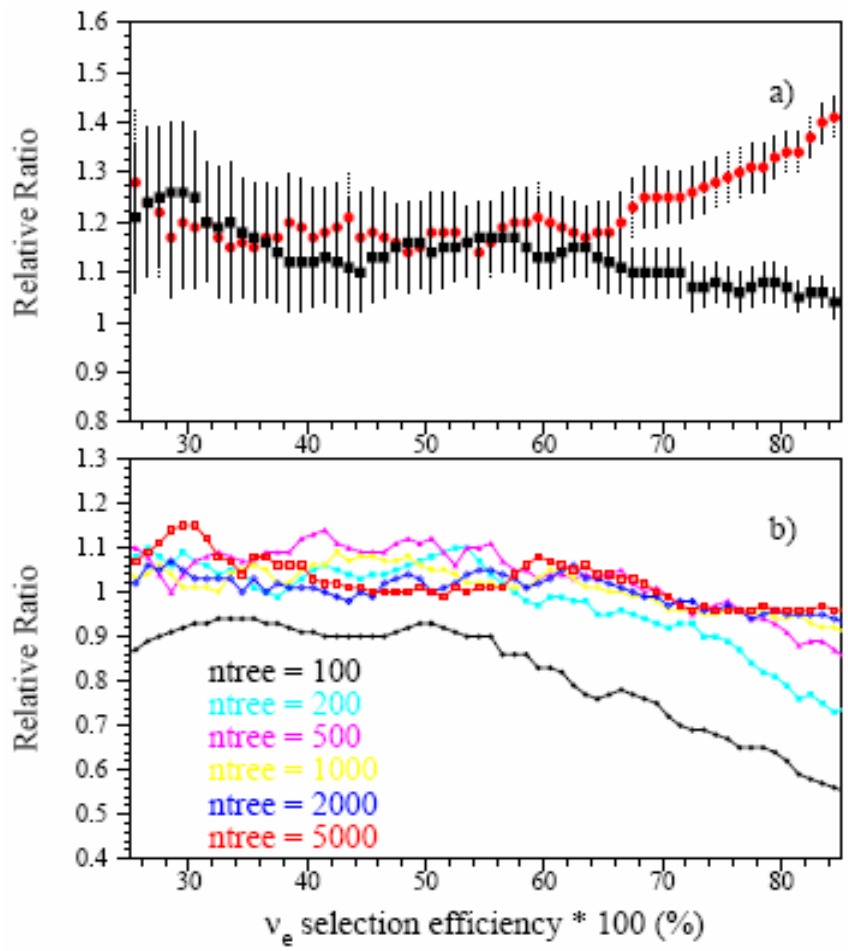
Comparison of 21 (or 22) vs 52 variables for Boosting

- Vertical axis is the ratio of bkrd kept for 21(22) var./that kept for 52 var.
- Red is if training sample is cocktail and black is if training sample is pi0
- Error bars are MC statistical errors only



AdaBoost vs Epsilon Boost and differing tree sizes

- A. Bkrd for 8 leaves/ bkrd for 45 leaves. Red is AdaBoost, Black is Epsilon Boost
- B. Bkrd for AdaBoost/ bkrd for Epsilon Boost Nleaves = 45.



Numerical Results from sfitter

- Extensive attempt to find best variables for ANN and for boosting starting from about 3000 candidates
- Train against pi0 and related backgrounds—22 ANN variables and 50 boosting variables: for the region near 50% of signal kept, the ratio of ANN to boosting background was about 1.2

Number of parameters to fit

- In ANN if 22 input variables and a hidden layer of 22, 2X(22 X (22+1)= 1012. ANN updates weights every few events.
- In boosting, if 1000 trees of 45 leaves, then 1000 X45 X2 (which var and cut point) = 90,000.

For ANN

- **For ANN one needs to set temperature, hidden layer size, learning rate… There are lots of parameters to tune**
- For ANN, if one

  a. Multiplies a variable by a constant, var17→2.var 17

  b. Switches two variables

  var 17←→var 18

  c. Puts a variable in twice

  The result is very likely to change

For boosting

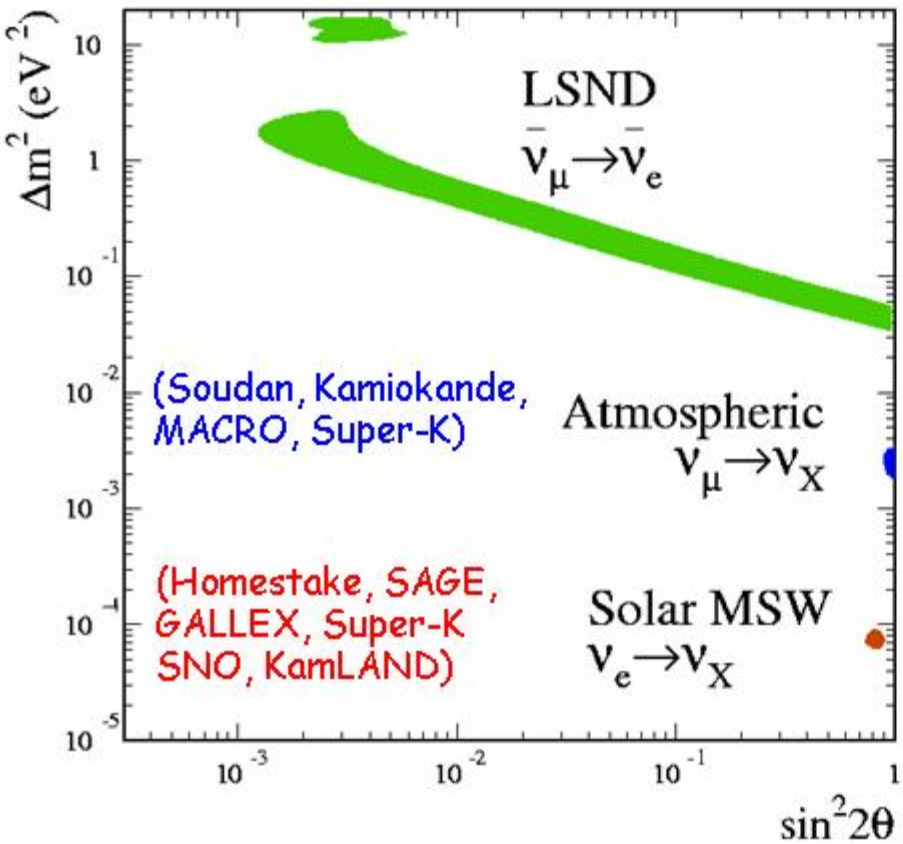
- Boosting can handle more variables than ANN; it will use what it needs.
- Duplication or switching of variables will not affect boosting results.
- Suppose we make a change of variables y=f(x), such that if x_2>x_1, then y_2>y_1. The boosting results are unchanged. They depend only on the ordering of the events
- There is considerably less tuning for boosting than for ANN.

Conclusions

- **For MiniBooNE boosting is better than ANN by a factor of 1.2—1.8**
- **AdaBoost and Epsilon Boost give comparable results within the region of interest (40%--60% nue kept)**
- **Use of a larger number of leaves (45) gives 10--20% better performance than use of a small number (8).**
- **It is expected that boosting techniques will have wide applications in physics.**
- **Preprint Physics/0408124; submitted to Phys. Rev.**

Motivation for the Experiment - LSND

State of Oscillation Results

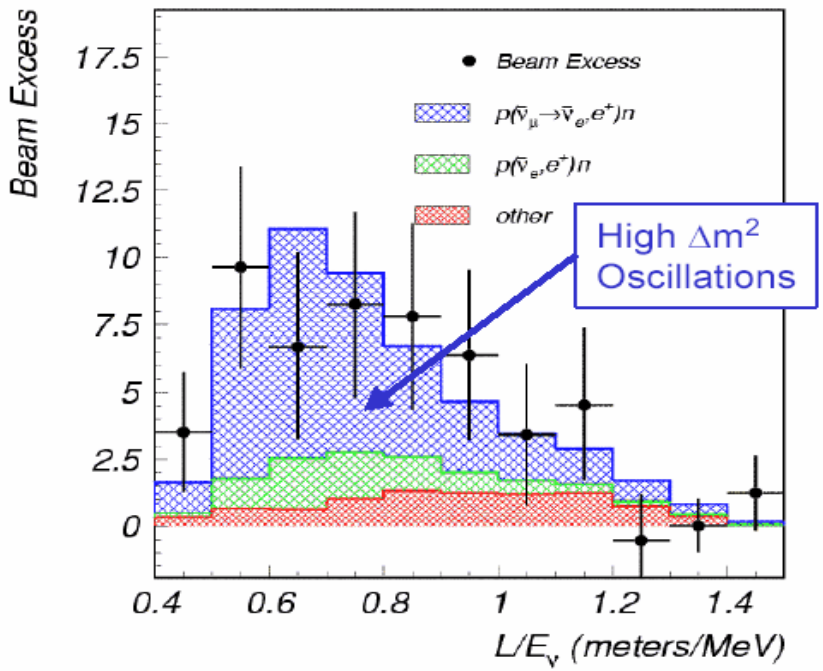


- Simplest model has three neutrino mass eigenstates, but…
- Data indicates 3 mass differences
  - $\Delta m^2_{atm} \sim 2\text{-}3\times10^{-3}$ eV^2
  - $\Delta m^2_{sol} \sim 7\times10^{-5}$ eV^2
  - $\Delta m^2_{LSND} \sim .1\text{-}10$ eV^2

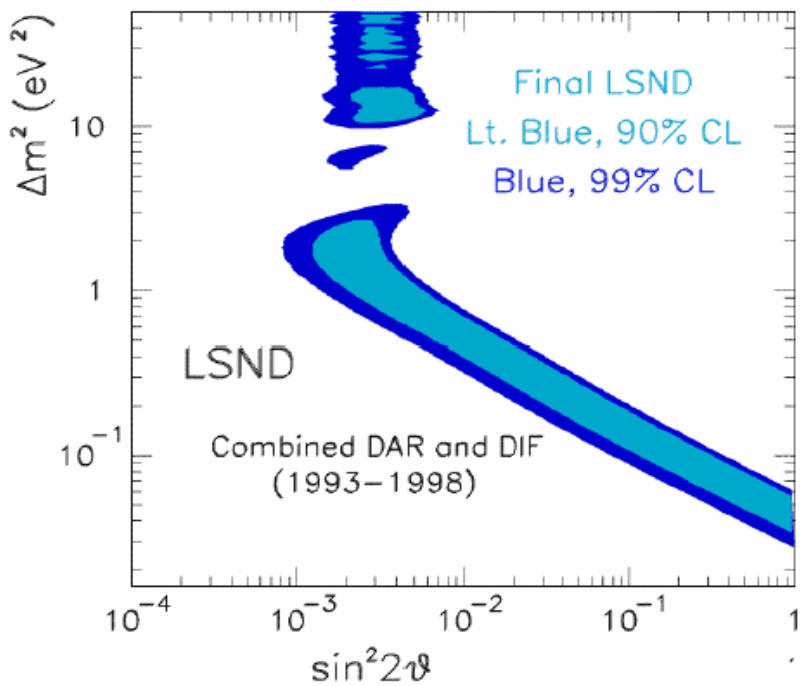
LSND final result

LSND sees excess above backgrounds

– Excess: 87.9 ± 22.4 ± 6.0 evts.

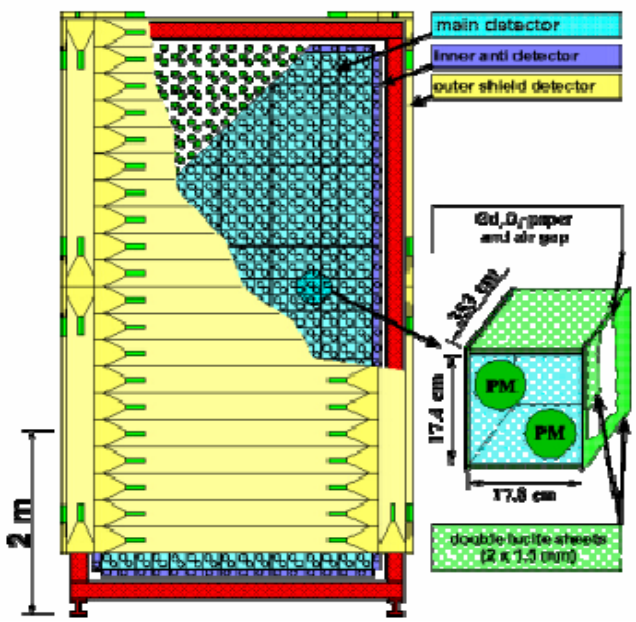


- Corresponding osc. probability: (0.264 ± 0.067 ± 0.045)%
- 3.3 σ evidence for oscillation.

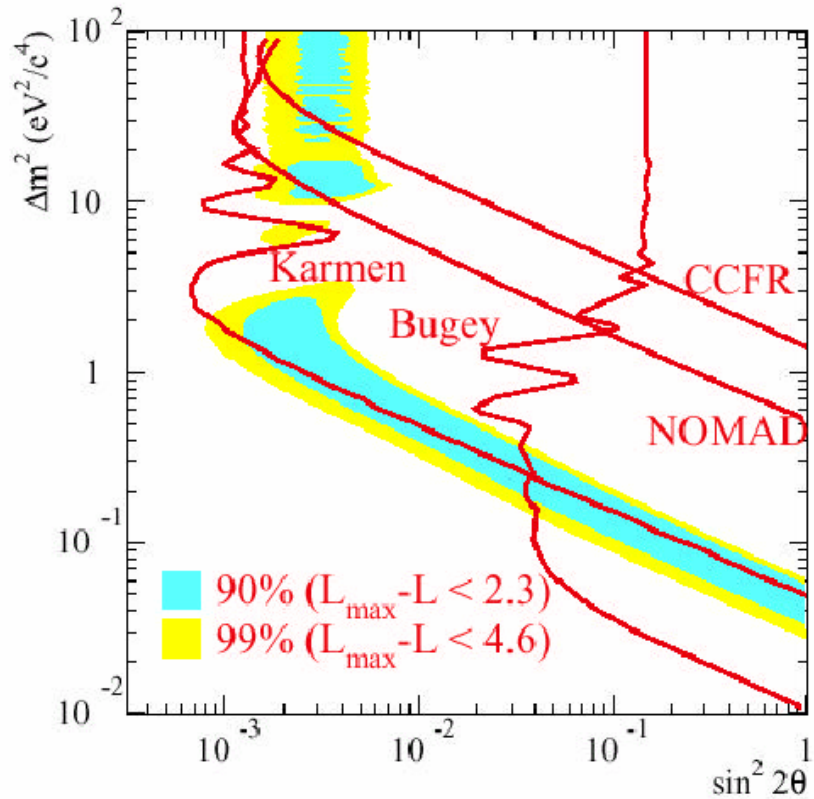


KARMEN II (1997-2001)

- Pulsed 800 MeV pot (ISIS)
  - DAR beam (90° to target)
  - 17.6 m baseline
- 56 tons of liquid scintillator
  - 512 modules
  - Gd-doped (8 MeV γ)
- $\times$10 less statistics than LSND (less intensity & size)



- 11 events observed
- 12.3 ± 0.6 events expected

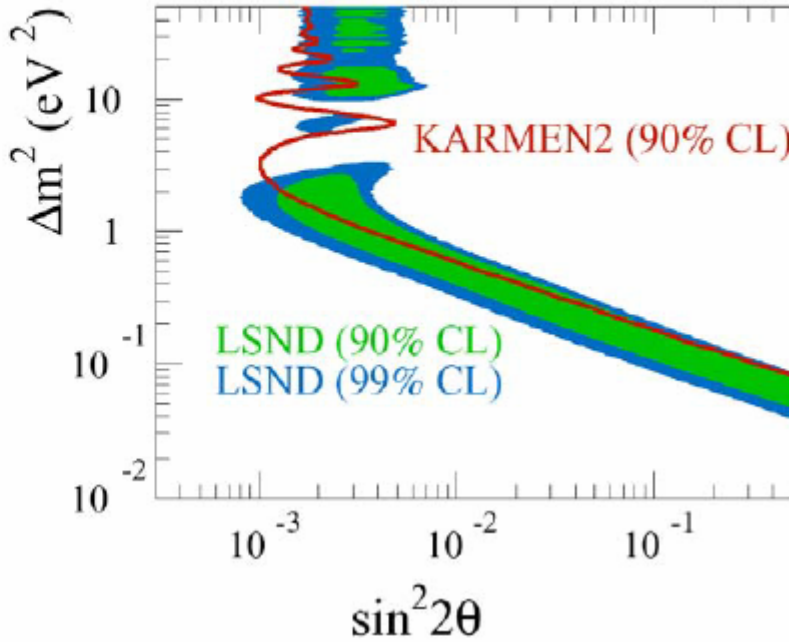


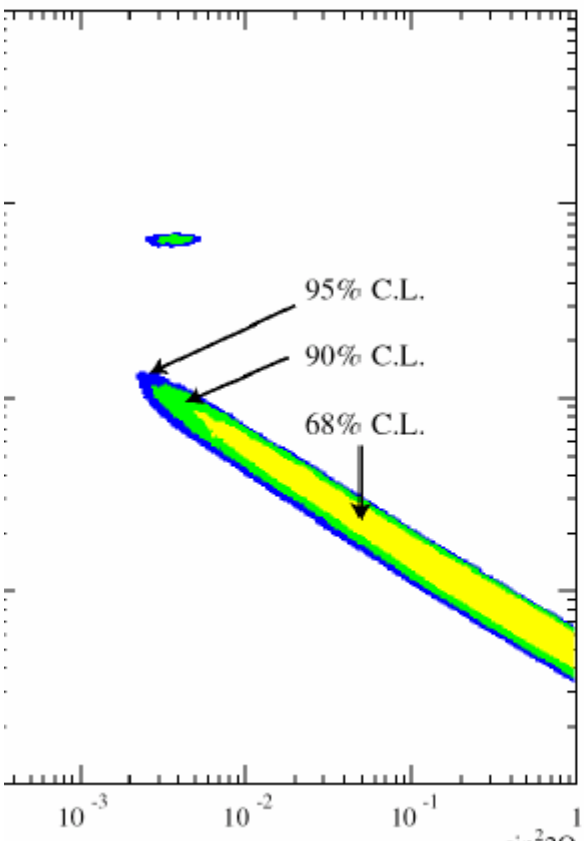
KARMEN Experiment

milar beam and detector to LSND

Closer distance and less target mass
⇒ x10 less sensitive than LSND

int analysis with LSND gives restricted
gion (Church et al. hep-ex/0203023)





- KARMEN also limits $\mu^+ \rightarrow e^+ \bar{\nu}_e \nu$ branching r
  BR < 0.9 x 10^{-3} (90% CL)
- LSND signal would require:
  1.9×10^{-3} < BR < 4.0 x 10^{-3} (90% CL)

⇒ $\mu^+ \rightarrow e^+ \bar{\nu}_e \nu$ **unlikely to explain LSND sig**

Interpretations of the LSND Signal

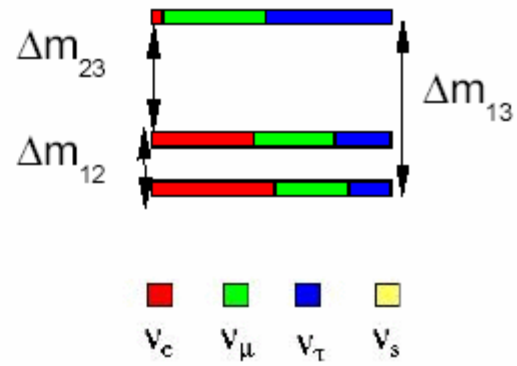
Not oscillations?

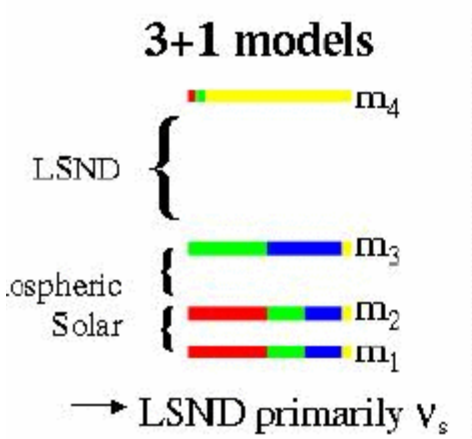
- Anomalous muon decay? Ruled out by KARMEN2 at 90%CL (hep-ex/0302017)

$10^{-5}+10^{-3}\neq 1$ (Solar + Atmospheric ≠ LSND)

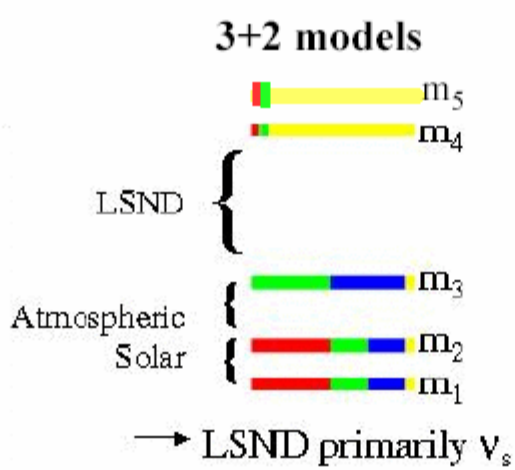
Sterile Neutrinos? (No weak coupling ⇐ invisible Z width)

CPT Violation?

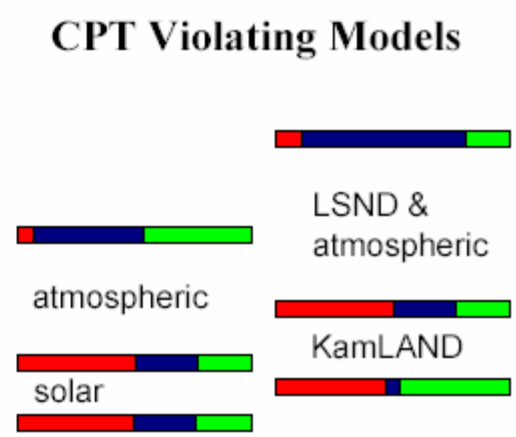




Increasingly disfavored



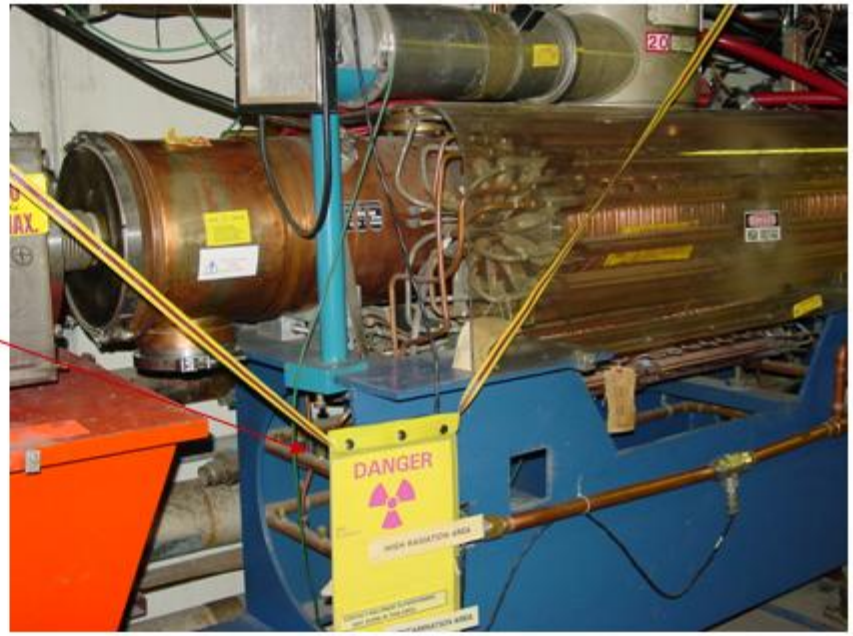
hep-ph/0305255



hep-ph/0108199, hep-ph/0212116

Delivering Protons

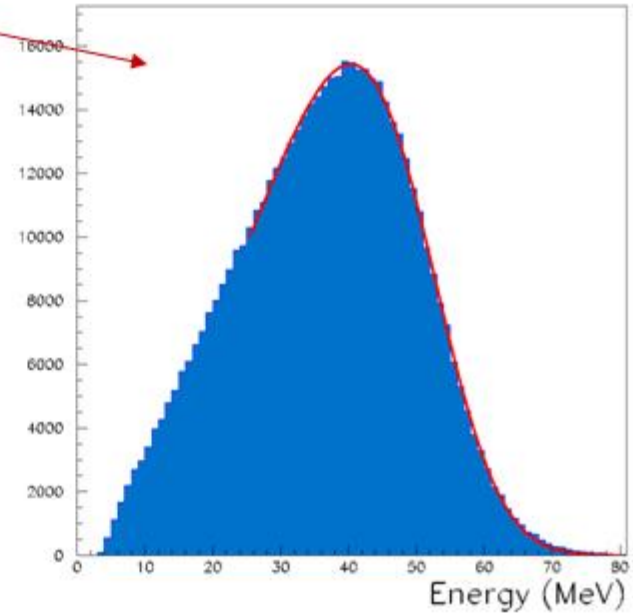
- Requirements of MiniBooNE greatly exceed the historical performance of the 30+ year old 8 GeV Booster, pushes…
  - Average repetition rate
  - Above ground radiation
  - Radiation damage and activation of accelerator components

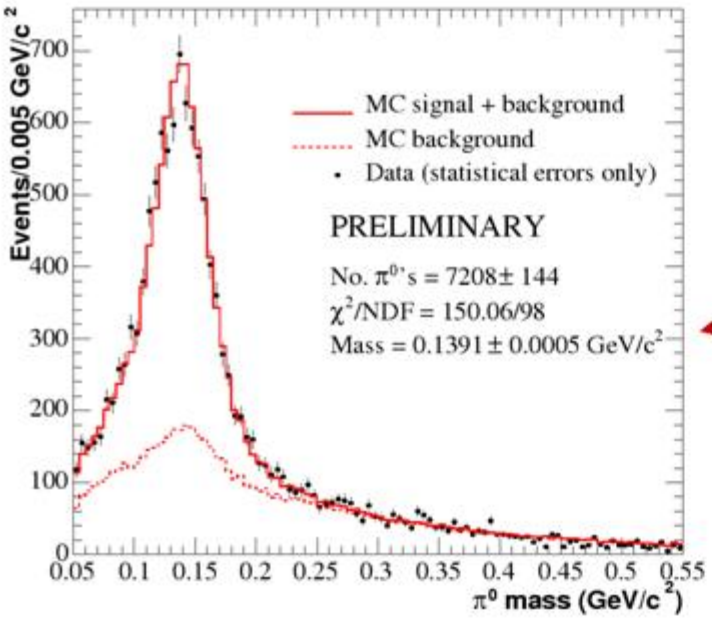


- Intense Program to improve the Booster
  - Shielding
  - Loss monitoring and analysis
  - Lattice improvements (result of Beam Physics involvement)
  - Collimation system

The Detector (cont'd)

- Elecrons from muon decay (Michel electrons)
  - Vital for understanding signal events.

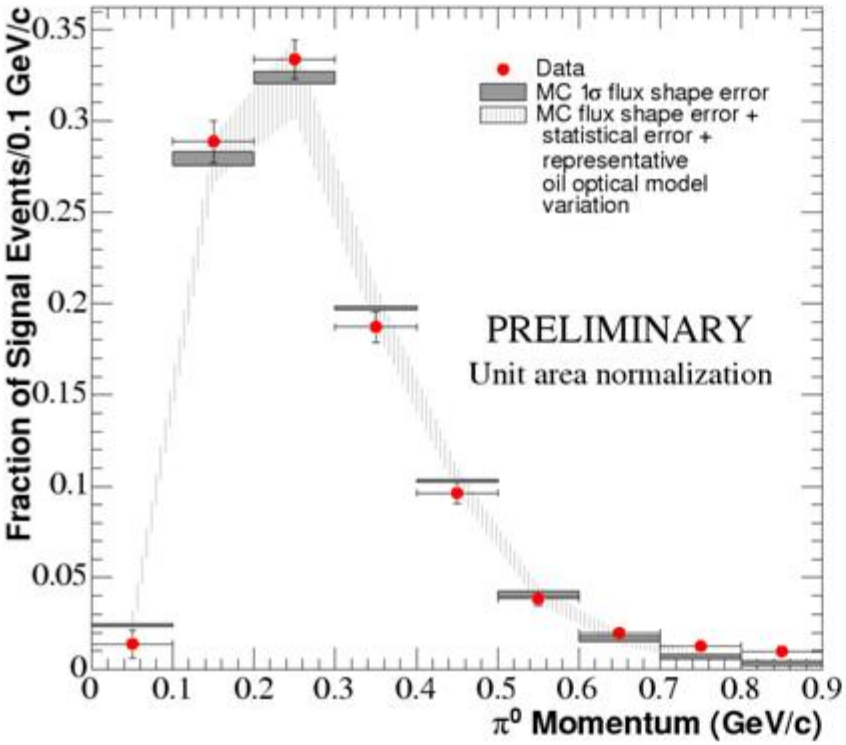




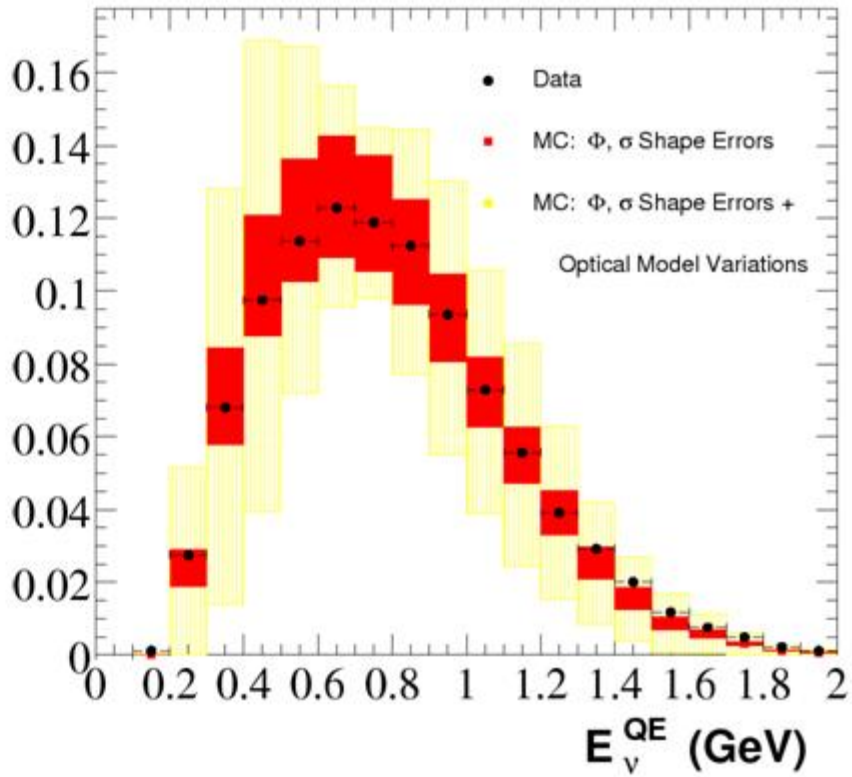
- π^0 Events
  - Help to understand higher energy ν_e
  - Help fix energy scale

Understanding the Data: Examples

π^0

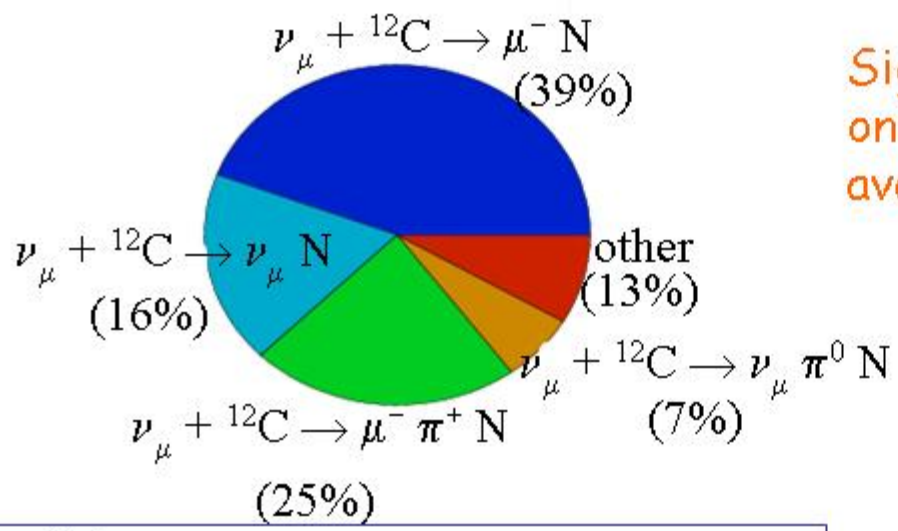


E_ν



Event Rates

ν_μ Reactions

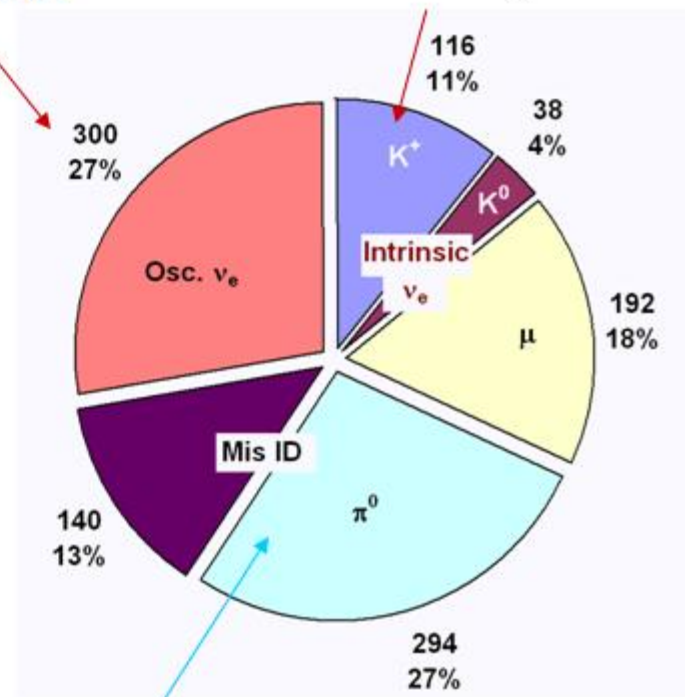


- MC:
  - Parametrized pBe interaction (our fit of BNL E910)
  - GEANT4 model of target, horn, and beamline
  - NUANCE MC of neutrino interaction
  - Full hit-based MC of detector, including oil, PMT's etc.

Contributions to ν_e Signal

Hope LMC+HARP will help with K background

Signal: Based on LSND average

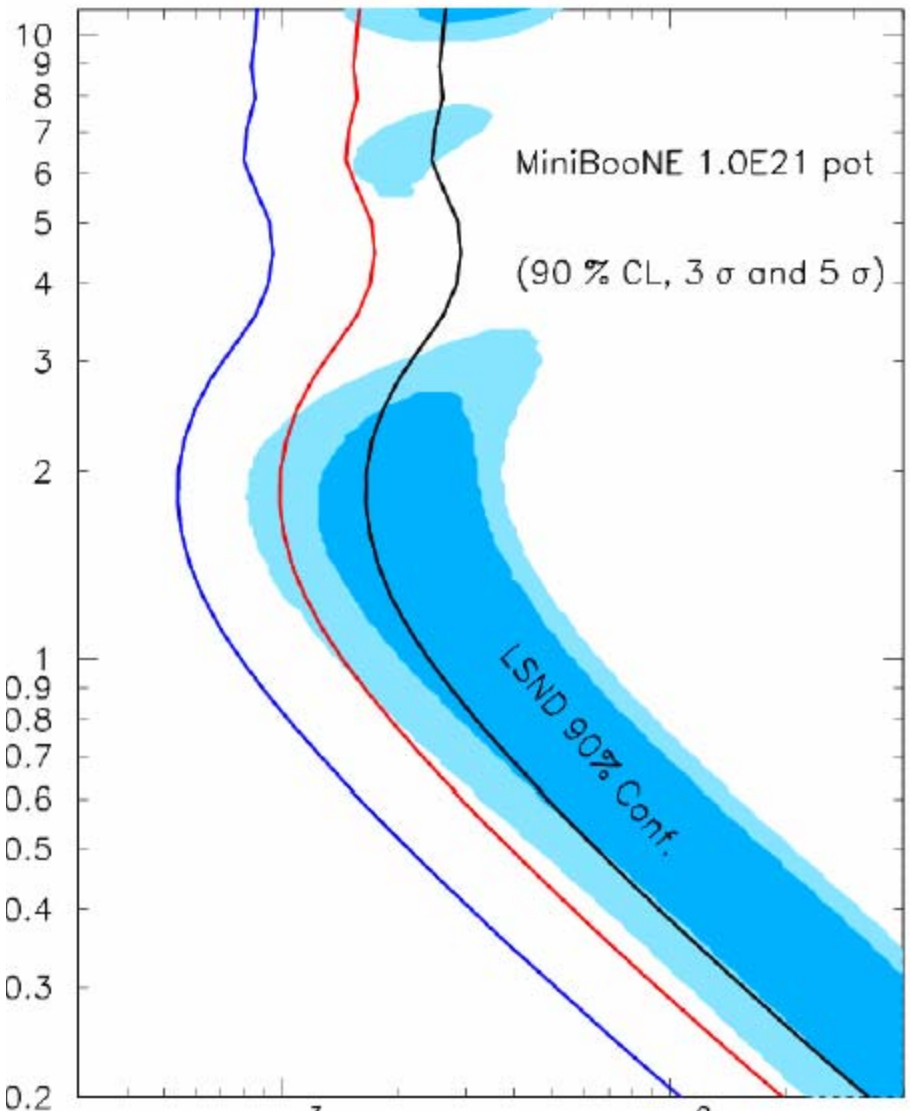


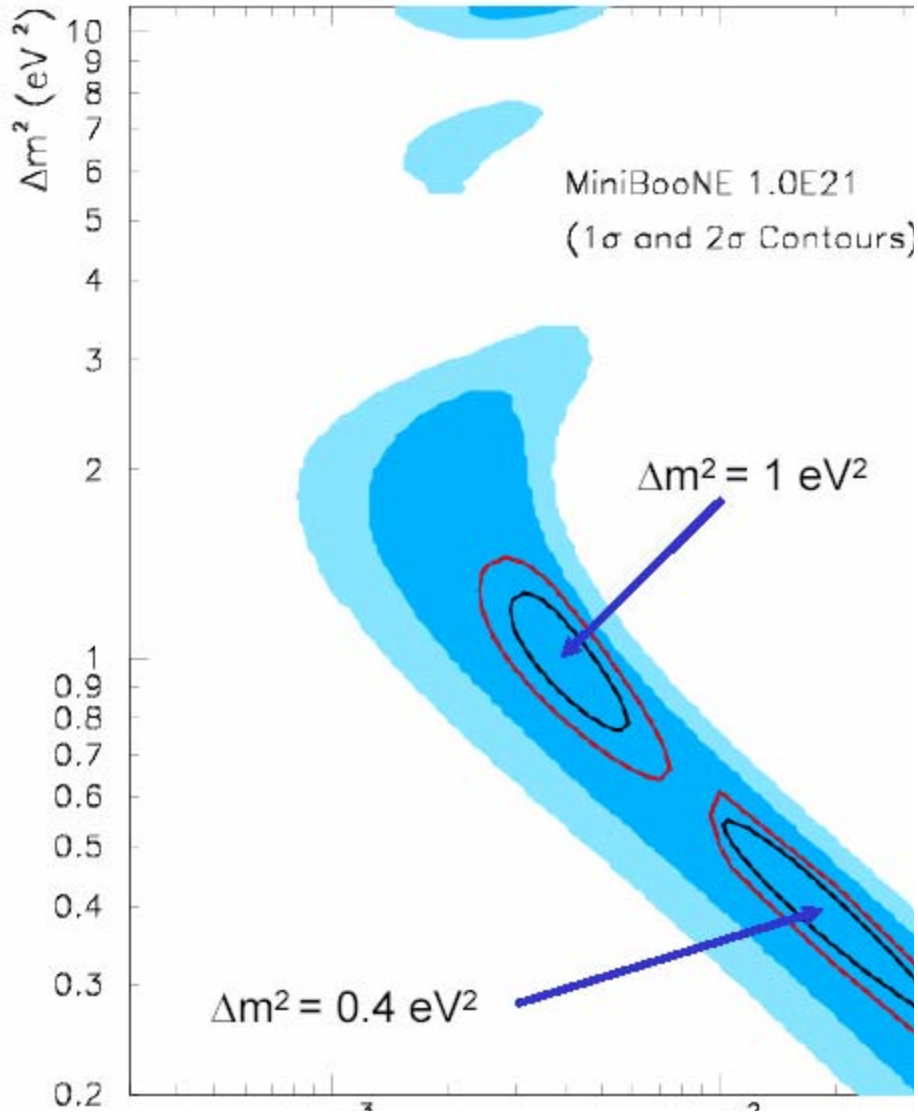
Can be constrained with data

There are None So Blind…

- This is a difficult analysis, and there are many opportunities for unintentional bias
- Therefore, we consider a blind analysis essential
- General philosophy: guilty until proven innocent
- Events go "into the box" unless they are specifically tagged as being non-signal events, e.g
  - Muons
    - Single μ-like ring
    - Topological cuts
  - π^0
    - No Michel electron
    - Clear two-ring fit, both with E>40 MeV
- Will only look at remaining data when
  - We have enough protons, AND
  - We are confident that we model the beam and detector well.

- Oscillation sensitivity and measurement capability
  - Data sample corresponding to 1×10^{21} pot
  - Systematic errors on the backgrounds average ~5%





Summary and Outlook

- MiniBooNE has collected 3.5E20 protons
- The experiment has made impressive progress in understanding both the detector and the data.
- Fermilab is about to go into a 13 week shutdown, during which…
  - The MiniBooNE horn will be replaced.
  - Improvements will be made to the Booster, which should allow it to achieve the MiniBooNE intensity goals
- NuMI will start in early 2005, BUT MiniBooNE should be able to continue taking data, albeit at a reduced intensity
  - *We could not have said this a few months ago.*
  - 5E20 proton by early 2005
  - 1E21 somewhere between mid-2006 and mid-2007
- Will not release ν_e appearance result before 5E20, but other physics along the way, e.g.
  - NC π^0 cross-section
  - ν_μ disappearance result

MiniBooNE consists of about 70 scientists from 12 institutions.

Y. Liu, I. Stancu *Alabama*
S. Koutsoliotas *Bucknell*
E. Hawker, R.A. Johnson, J.L. Raaf *Cincinn*
T. Hart, R.H. Nelson, E.D. Zimmerman *Colora*
A. Aguilar-Arevalo, L.Bugel, L. Coney, J.M. C
J. Formaggio, J. Link, J. Monroe, K. McConn
D. Schmitz, M.H. Shaevitz, M. Sorel, L. Wan
G.P. Zeller *Columbia*
D. Smith *Embry Riddle*
L.Bartoszek, C. Bhat, S J. Brice, B.C. Brown
D.A. Finley, B.T. Fleming, R. Ford, F.G.Garc
P. Kasper, T. Kobilarcik, I. Kourbanis,
A. Malensek, W. Marsh, P. Martin, F. Mills,
C. Moore, P. Nienaber, E. Prebys,
A.D. Russell, P. Spentzouris, R. Stefanski,
T. Williams *Fermilab*
D. C. Cox, A. Green, H.-O. Meyer, R. Tayloe *Indiana*
G.T. Garvey, C. Green, W.C. Louis, G.McGr
S.McKenney, G.B. Mills, V. Sandberg,
B. Sapp, R. Schirato, R. Van de Water,
D.H. White *Los Alamos*
R. Imlay, W. Metcalf, M. Sung, M.O. Wascko *Louisiana State*
J. Cao, Y. Liu, B.P. Roe, H. Yang *Michigan*