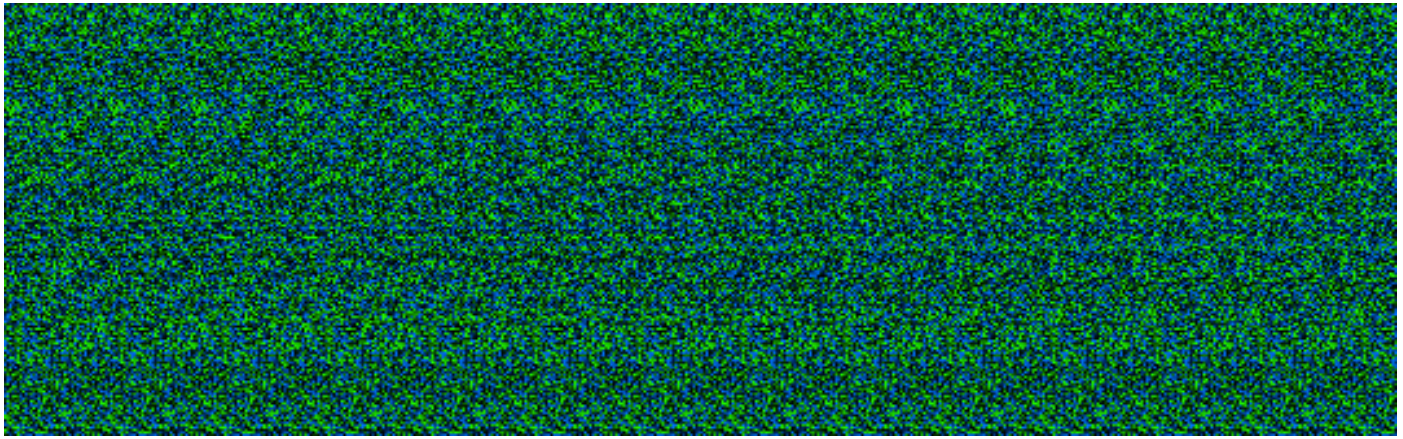
Estimating Fake Lepton Backgrounds in Early Data with kNN



G. Bauer, J. Bendavid, G. Gomez Ceballos,
P. Everaerts, *K. Hahn,* P. Harris, M. Klute,
C. Loizides, S. Nahn, C. Paus, M. Rudolph,
K. Sumorok, K. Sung, S. Tkaczyk, S. Xie

Berkeley Workshop on Early LHC Data
05/07/09

Outline

- Motivations
- Introduction to kNN
  - Overview of the method
  - Specifics for fake lepton estimation
- Results of preliminary studies
- Conclusions

- Please Note
  - Focusing on fake electrons in this talk
  - Results not endorsed by a certain experiment at this point ...
    - The review process to start soon ...

Motivation

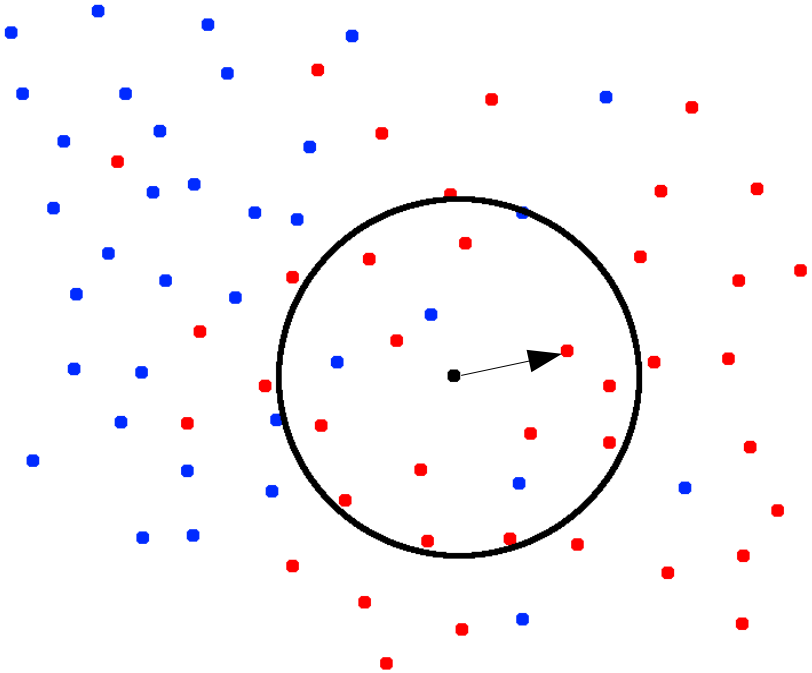
- Fakes/QCD a major background to EWK lepton channels
  - Difficult to model fakes with MC, instead estimate BG from data
  - Several historical techniques for this ...
    - eg : “ABDC”, fakeableObject, etc.
  - Can one formulate a complimentary approach?
- Fake rejection not precise in early running
  - Emphasis will be on high efficiency rather than high purity
  - Cuts tighten later as understanding of data matures
  - Can residual differences in discriminating variables be exploited?
- Established techniques involve extrapolation
  - Measure rates in QCD rich regions, assume same in signal region
  - Some methods use MC for fake distribution shapes
  - Can we estimate fakes using only data, with quantities likely to be well understood at startup?

Motivation (2)

- Classification : another possible approach to the problem
  - Utilize differences in lepton candidate features to discriminate between fake/real leptons
- Note: not discussing classification for rejection here
  - Instead proposing it as means for background estimation
- Design Choices :
  - Non-parametric : not all ID distributions easy/possible to model
    - no maximum-likelihood
  - Supervised : make use of auxiliary datasets (ie: Z's, QCD)
    - no self-clustering
  - As simple as possible : life complicated enough with early data
    - no neural network

k-NearestNeighbor (kNN)

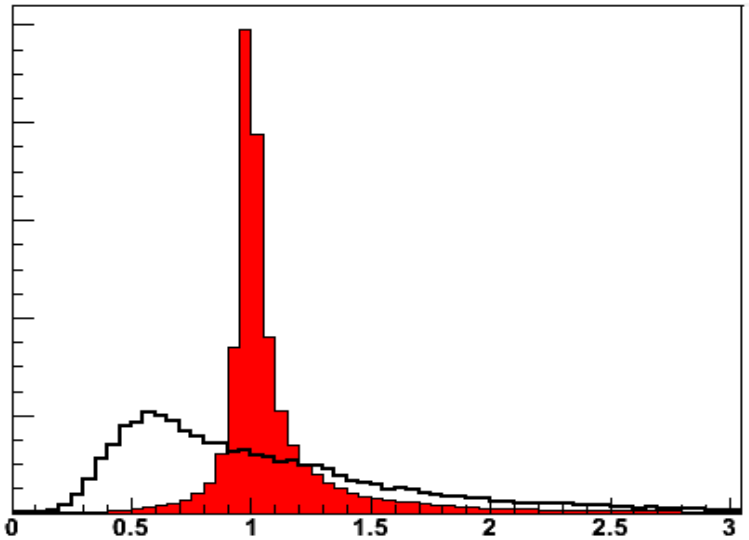
- A supervised non-parametric classification algorithm
  - One of the oldest/simplest machine-learning techniques
    - Uses in computer vision, bioinformatics, handwriting recognition
    - Active research on kNN variations and optimization
  - Robust consistency results ...
    - Misclassification error approaches ideal (Bayes Minimum Error) classifier for large statistics



- The concept (w.r.t. fakes)
  1) Construct "training" samples of real and fake ID'ed leptons
  2) Compare each ID'ed object in signal sample with prototypes
     - Calculate 'distance' to prototype neighbors using ID variables
  3) Classify object from majority class of k nearest neighbors
     - Calculate posterior probabilities

Distinguishability & Metric

- Method requires features that can be used to distinguish classes
  - Find those with wide differences in **fake**/real distributions after cuts
  - Rejecting those with obviously low discriminating power, left with :
    - Et, Isolation, shower shape, d0, E/p
- Determine 'distance' to prototypes via variables, 'metric' weights



Example : E/P for real/fake electrons

$$d^2(x_{i,\,test},\ x_{i,\,train}) = \sum_i w_i (x_{i,\,test} - x_{i,\,train})^2/\Delta x_i^2$$

  - Standardize variables such that scale effects don't prioritize variables
    - Divide by range (Δx) over which variables can/are observed to vary
  - Start with simplest metric: Euclidean, all weights (w_i) = 1

Toy Example

- Simple example to illustrate the technique
  - 2 classes (fake, real) , 2 ID variables (E/P, H/E), k=5
- Define a metric to determine 'distance'

$$d = \sqrt{(eop_{test}-eop_{train})^2/\Delta eop^2 + (hoe_{test} - hoe_{train})^2/\Delta hoe^2}$$

- Consider observation: E/P=0.88, H/E=0.012, Δeop=1, Δhoe=0.15
  - Select k=5 nearest neighbors from real/fake training samples

Class	EOP	HOE	d
real	٠.٧٩	٠.٠١٣	٠.٠٩٠
fake	٠.٧٧	٠.٠١٣	٠.١١٠
fake	٠.٦٦	٠.٠١٥	٠.٢٢٠
real	٠.٦٤	٠.٠٢٠	٠.٢٤٦
fake	٠.٥٥	٠.٠٠٨	٠.٣٣١
real	١.٢٥	٠.٠٥٦	٠.٤٧٢
fake	٠.٥	٠.٠٩٢	٠.٦٥٥
...	...	...	...

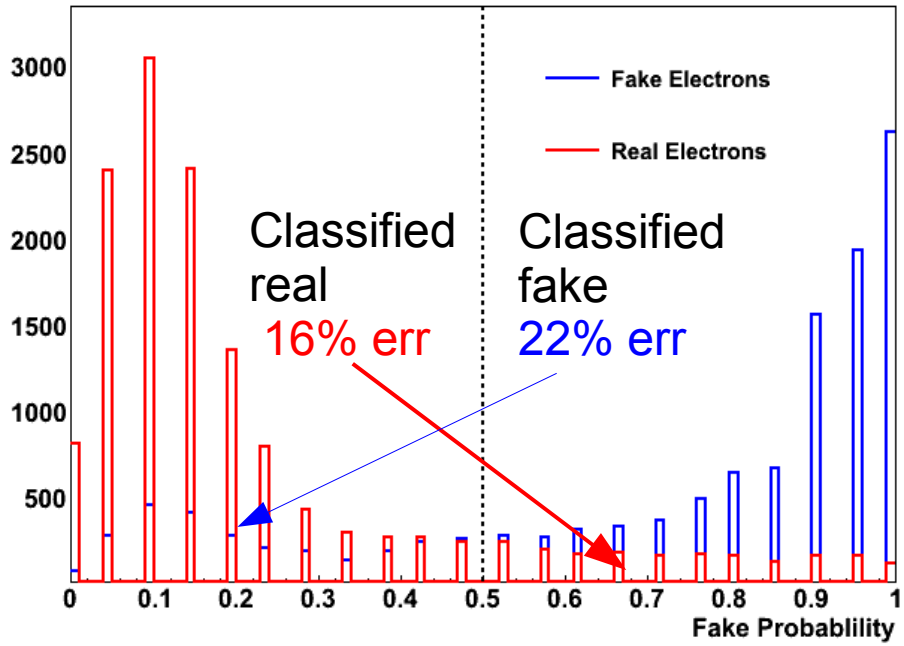
- Determine posterior fake probability for observation
  - $P(fake|x) = k_{fake}/k$
  - Here P(fake|x) = 60%
- Classify by majority vote
  - P(fake|x) > P(real|x) here so we classify observation as fake

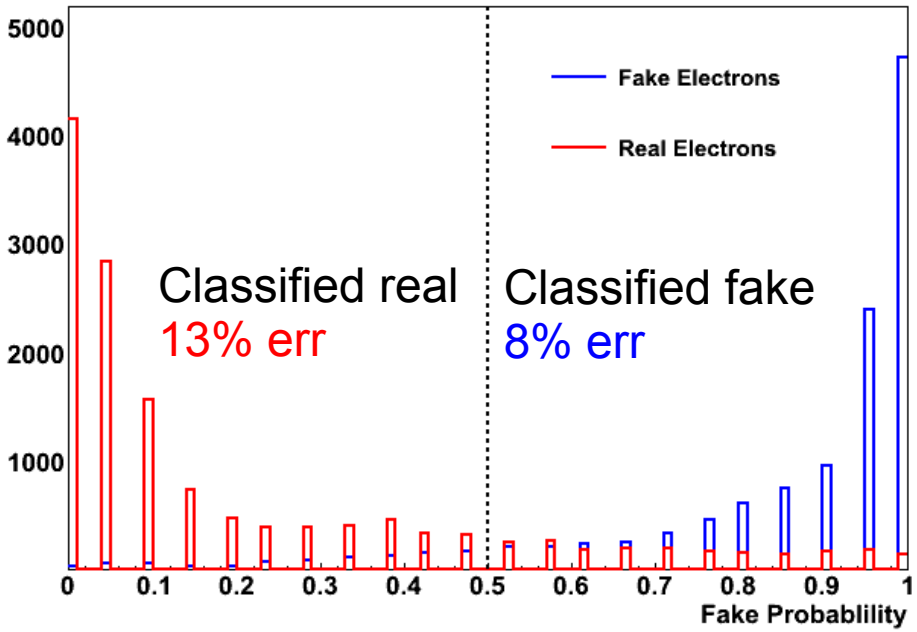
kNN Preliminaries

- Data samples
  - Using Pythia W/eν, QCD (EM enriched + Heavy Flavor)
    - In data : Et-scaled Z electrons and dijets with W/Z rejection
  - Split into separate samples: real/fake training, rest used as “data”
  - Apply appropriate triggers, electron ID cuts to everything
  - Training sample sizes each = 1K
    - Want equal sizes for electrons/fakes for unbiased fake probability
    - Want training samples large as possible, but increases CPU time
- Choice of “k” = 31
  - Impacts fake probability, trade-offs here ...
    - k too small, sensitive to noise. k too large, no longer local
    - Rule of thumb: $\sqrt{N_{training}}$ a good starting point
  - k should be odd to break ties

Metric Learning

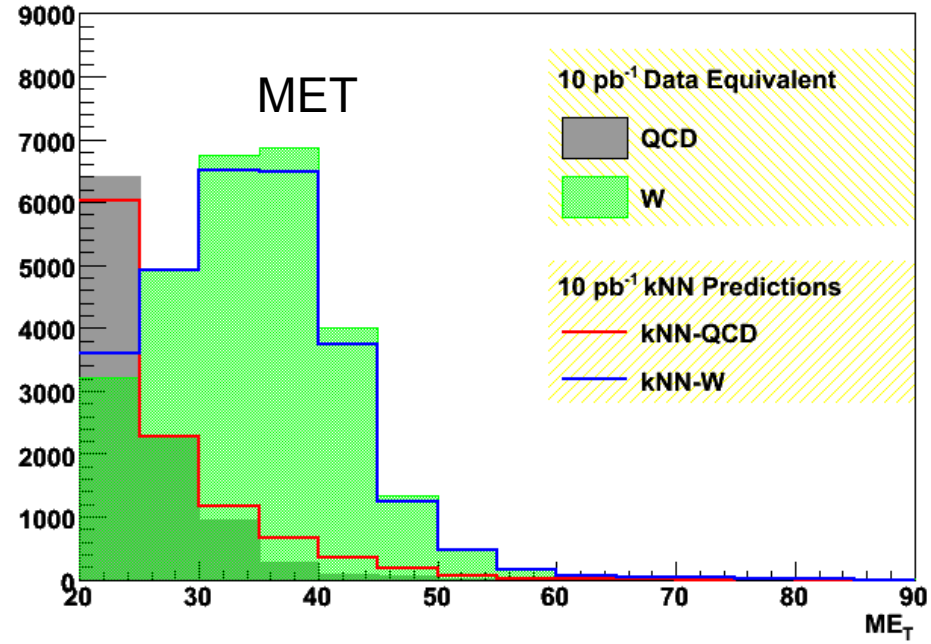
- Judge performance from posterior class probabilities
  - Calculate in training data
  - P(fake) = k_{fake} / k
  - Errors = % lost by misclassification
- Euclidean performance OK
  - But doesn't utilize relative importance of features
- Methods to learn "best" metric from training data
  - "Large Margin Nearest Neighbor" (LMNN) provided best results
    - Minimizes/Maximizes same-class/different-class distances
    - New weights determined

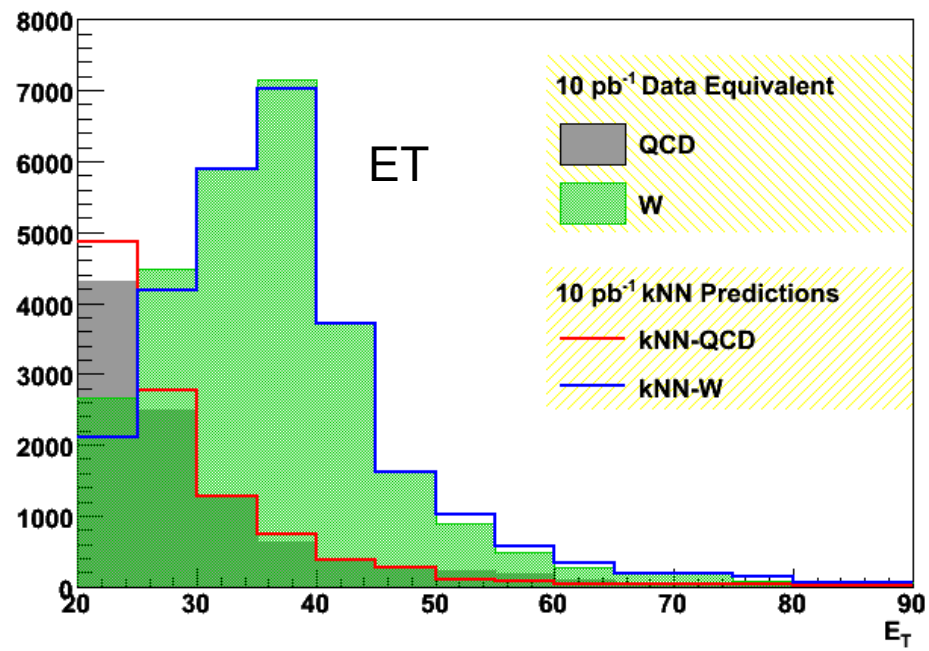




Results

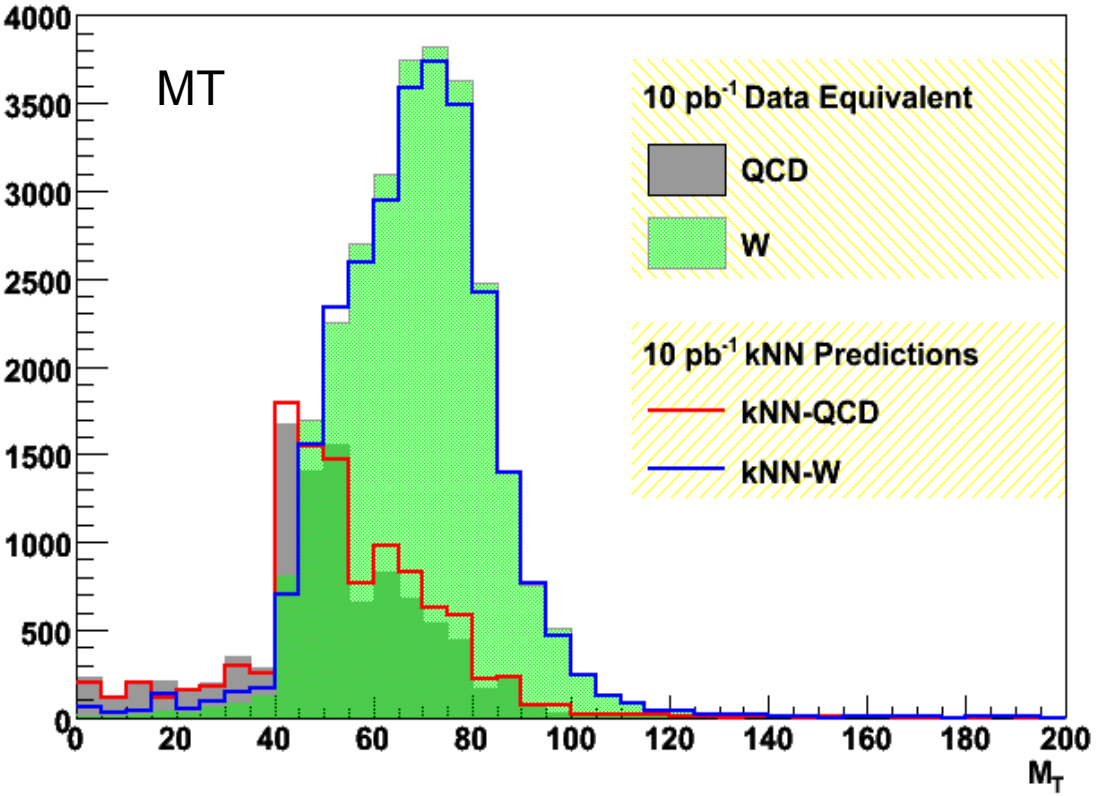
- Use training samples to classify 10pb-1 “data” samples
  - Gray = L.O. QCD fakes
  - Green = L.O. W→eν
- Predictions look good ...
  - Red = classified fakes
  - Blue = classified real
- Error w.r.t. Truth ≈ 2%
  - Some over/under fake predictions at low/high Et
  - Some under/over fake predictions at low/high MEt





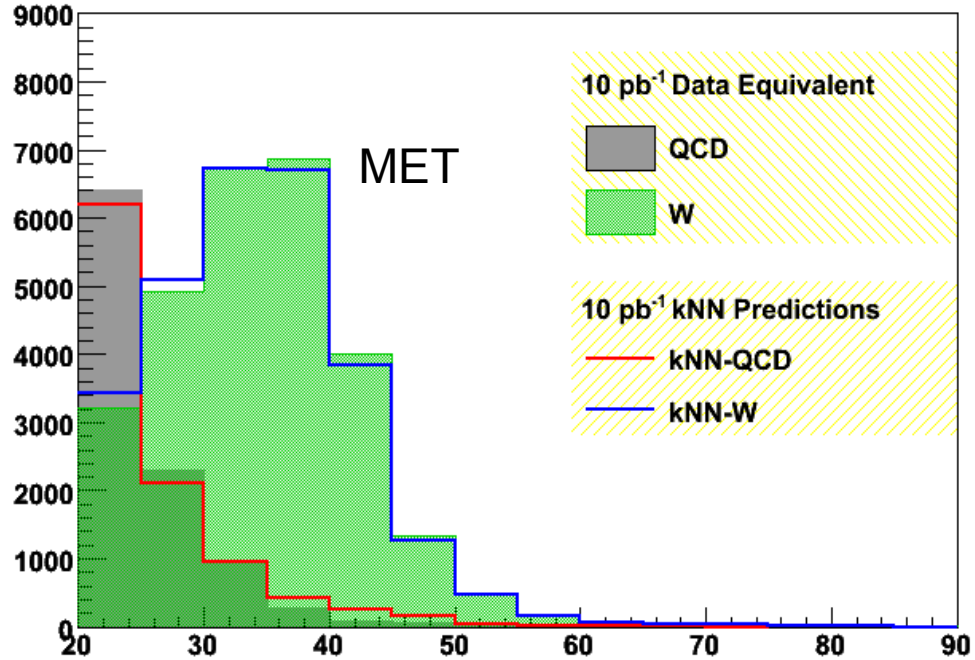
A Closer Look : Skew

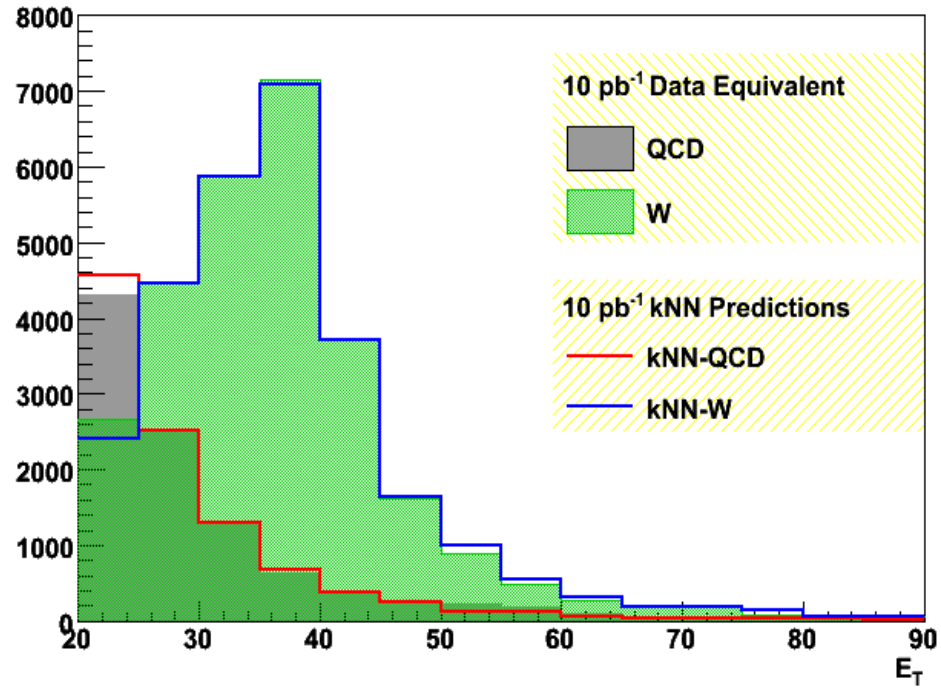
- mT shows fake over-prediction above 40 GeV
- Result of *skew* in the dataset
  - Many more W's than fakes in this region in data
  - Error rate low but misclassified W's sizable w.r.t to true fakes

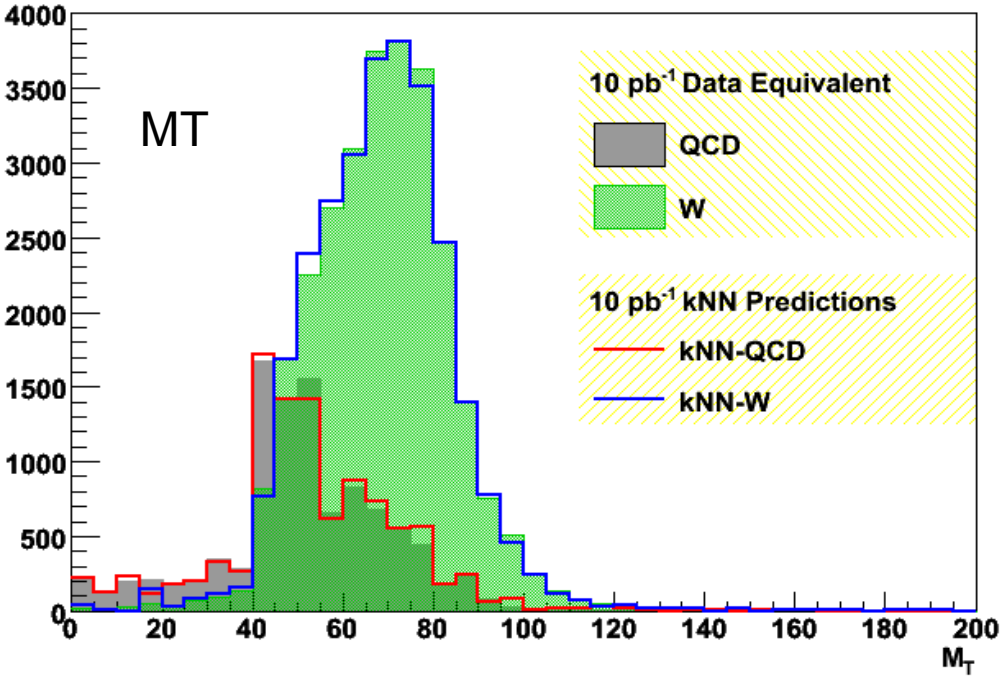


- Skewness is a common problem in classification ...
  - Try to address by tuning metric for smaller real electron error
    - Introduces new problems, larger fake error
  - Correct using classification efficiency measured in training

Corrected Predictions



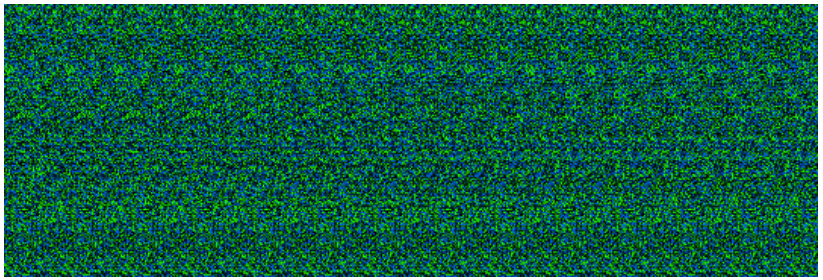




- W-bias in predictions much reduced
  - Still small discrepancies at low/high ET and MET
- Signal error w.r.t. Truth ≈ 0.5%

Conclusions

- kNN is a simple yet effective method for fake classification
  - A *completely* data-driven means of background estimation
  - Makes use of quantities that should be well measured at startup
  - Provides independent fake estimates w.r.t other methods on the market
    - Eventually average with other results for smaller overall error?
- Preliminary results are very encouraging
  - High accuracy achieved, proper error analysis on-going
  - Still room for improvement
- Technique allows for accurate resolution of fakes in data
  - As always, will need to “stare” hard at the data first!

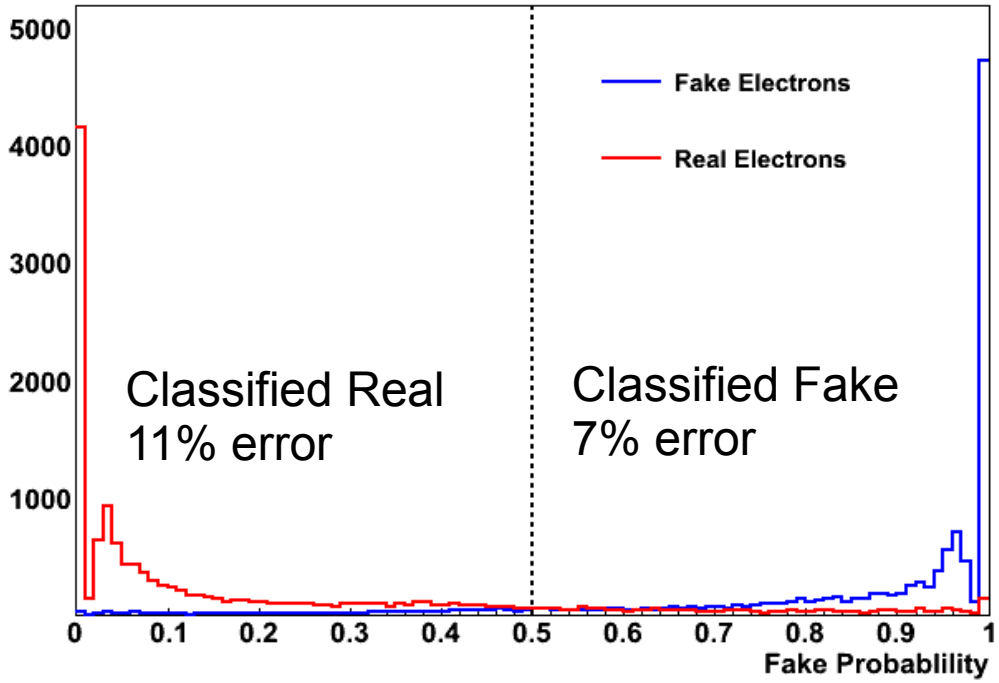


(http://www.eyetricks.com/stereograms/onlinetools/stereocreator.htm)

Backups

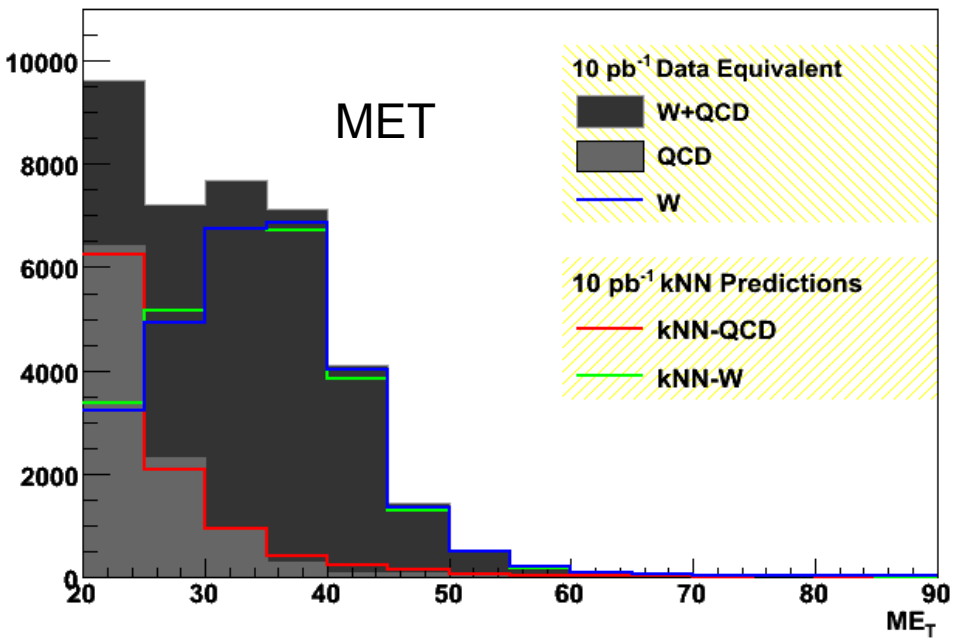
Distance Weighting

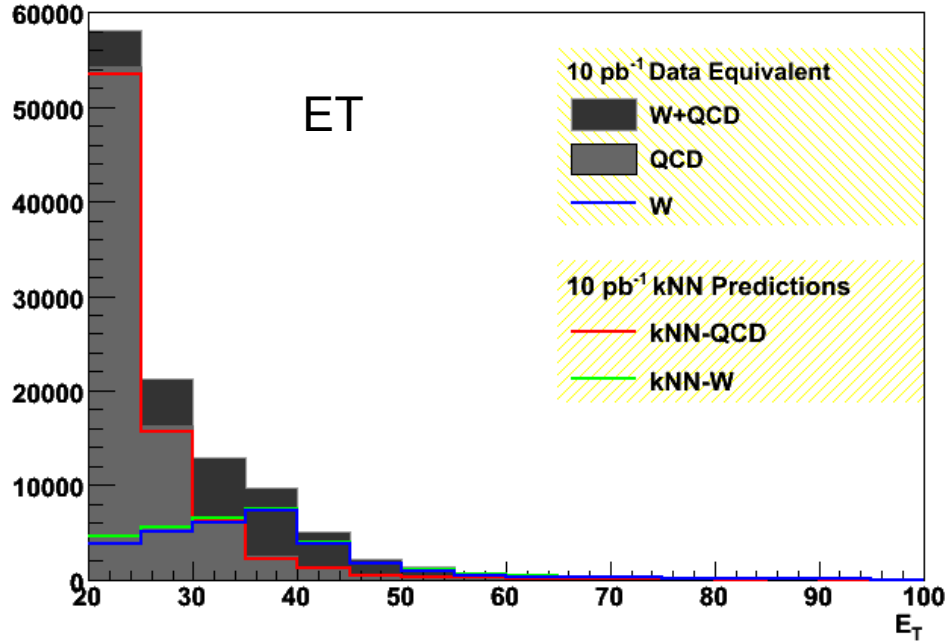
- Simple majority vote ignores some information
  - Distances to individual k neighbors may differ widely

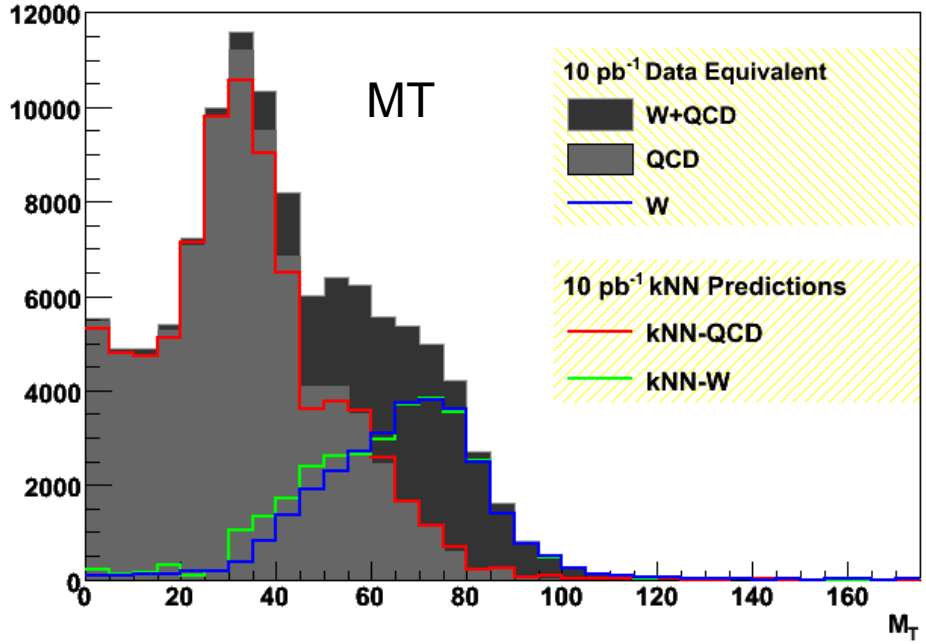


- Weight contribution by inverse distance to prototype
  - Close neighbors now count more than those far away
  - Smears out discrete probabilities from simple majority
  - Only slight improvement in classification accuracy

Predictions : Without MET Cut







- MET < 20 GeV a useful control region
- ET, MET, MT results show larger fake under-prediction at low values
  - Why don't the corrections address this? Investigating ...