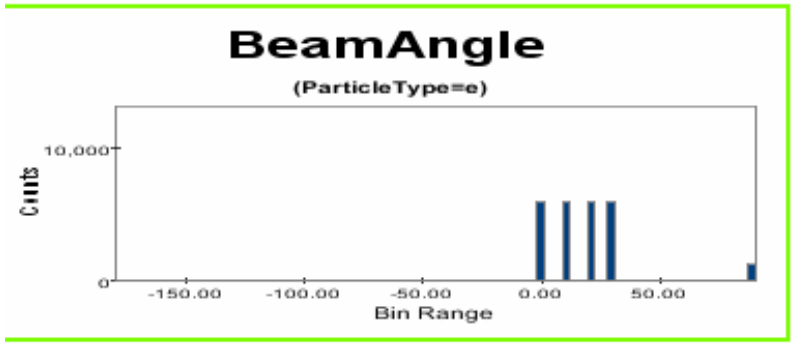
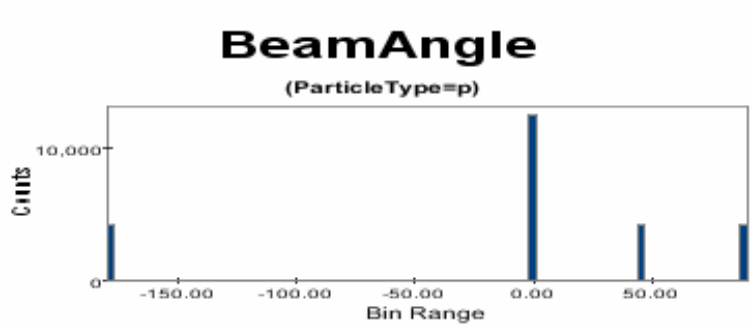
Electron-proton separation

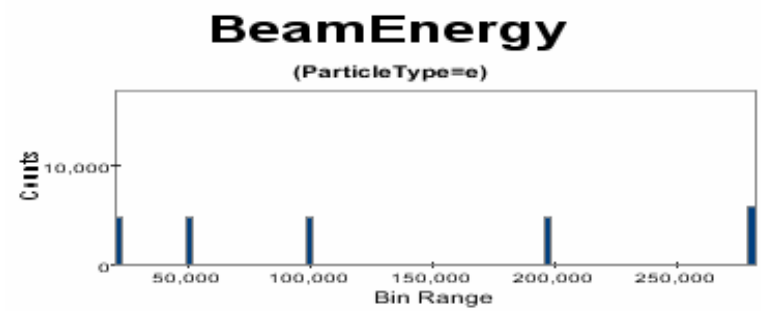
or our first contact with Insightful Miner

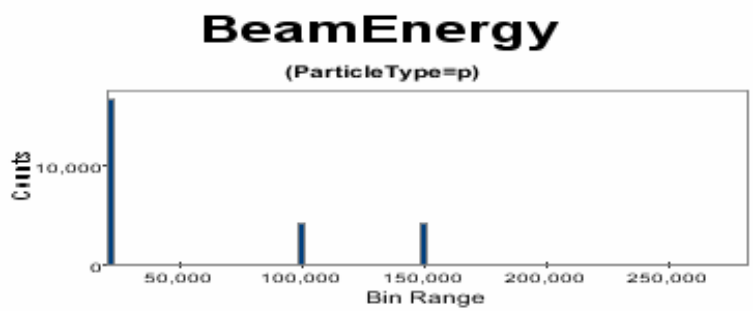
Goals and data sets

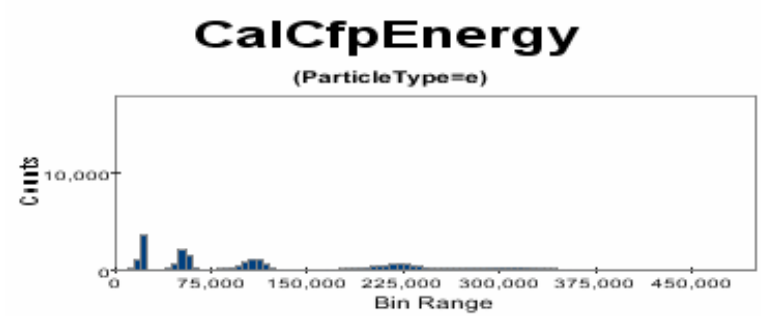
- Long-term
  - check cuts for high energy electron identification using BT data
  - develop a classification tree for selecting electrons
- Short term
  - Practice with Insightful Miner
    - test as many components as possible, we have a 30 days demo version and plan to purchase license next year with fresh funds
- Data set
  - Combined most e and p runs from Golden runs list with minimal cuts (BtSysTest, see appendix)
  - Had to convert them into txt files as IM does not like ROOT
    - now have a python script to select variables from a tuple and convert into tab separated file

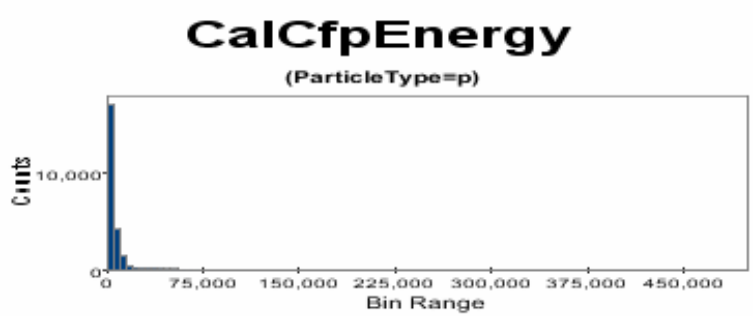

BeamAngle
(ParticleType=e)
Counts
10,000
0
-150.00
-100.00
-50.00
0.00
50.00
Bin Range

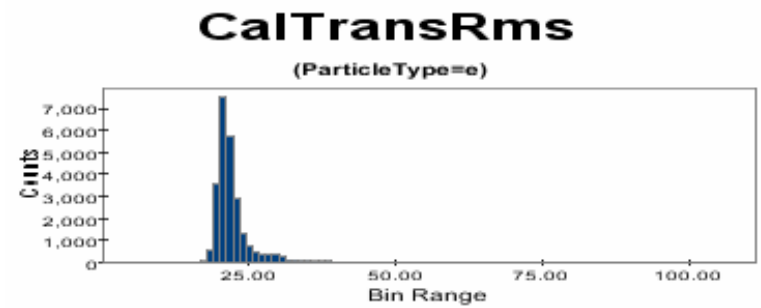

BeamAngle
(ParticleType=p)
Counts
10,000
0
-150.00
-100.00
-50.00
0.00
50.00
Bin Range

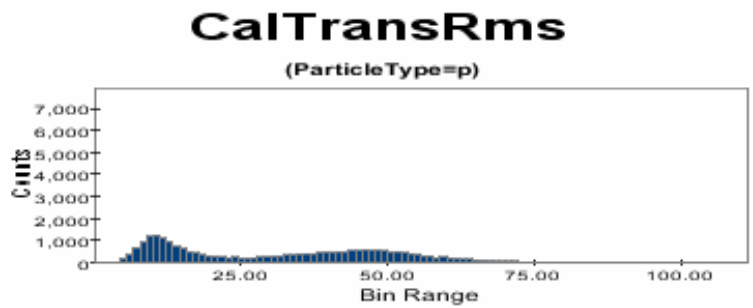

BeamEnergy
(ParticleType=e)
Counts
10,000
0
50,000
100,000
150,000
200,000
250,000
Bin Range


BeamEnergy
(ParticleType=p)
Counts
10,000
0
50,000
100,000
150,000
200,000
250,000
Bin Range

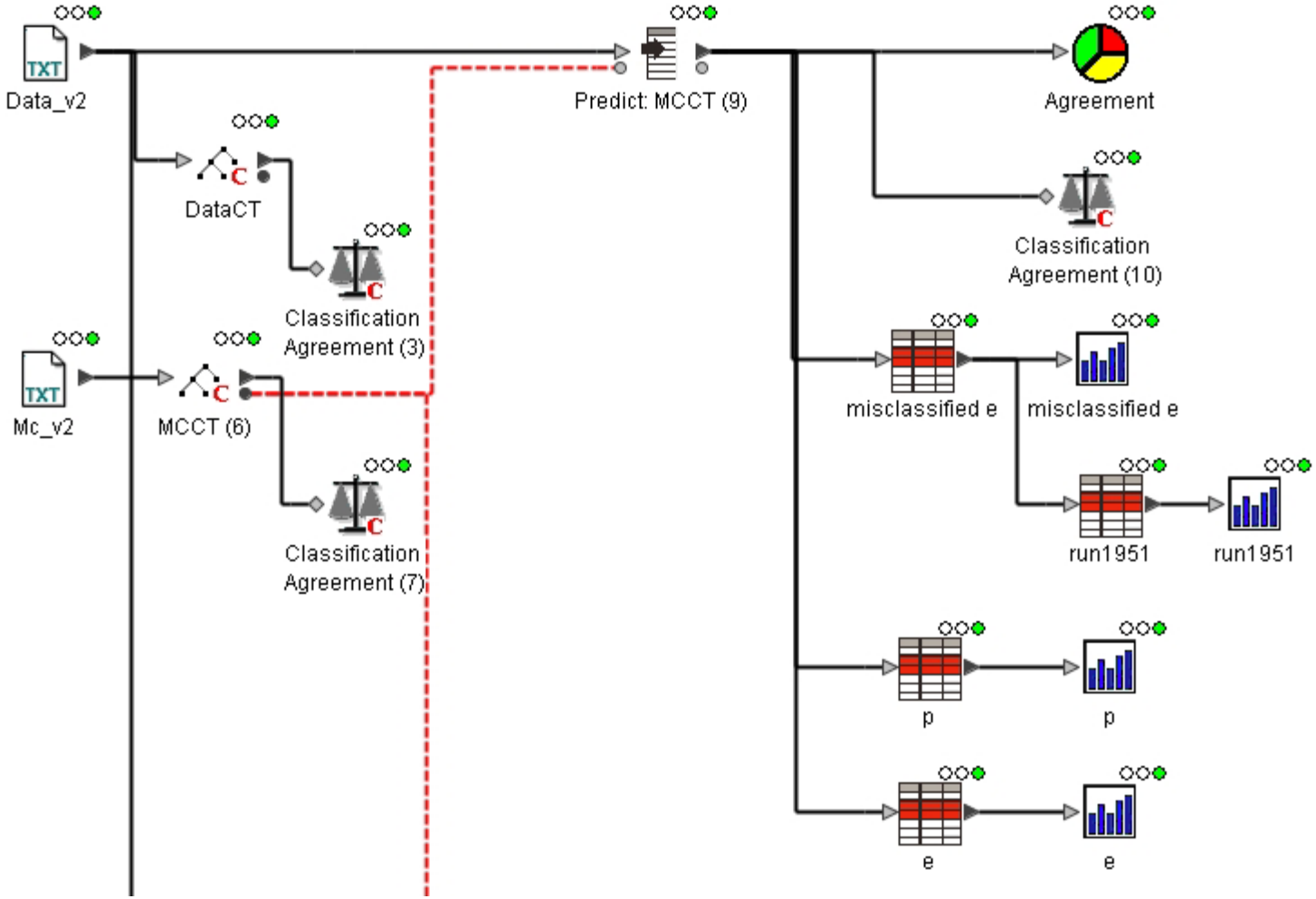

CalCfpEnergy
(ParticleType=e)
Counts
10,000
0
0
75,000
150,000
225,000
300,000
375,000
450,000
Bin Range


CalCfpEnergy
(ParticleType=p)
Counts
10,000
0
0
75,000
150,000
225,000
300,000
375,000
450,000
Bin Range

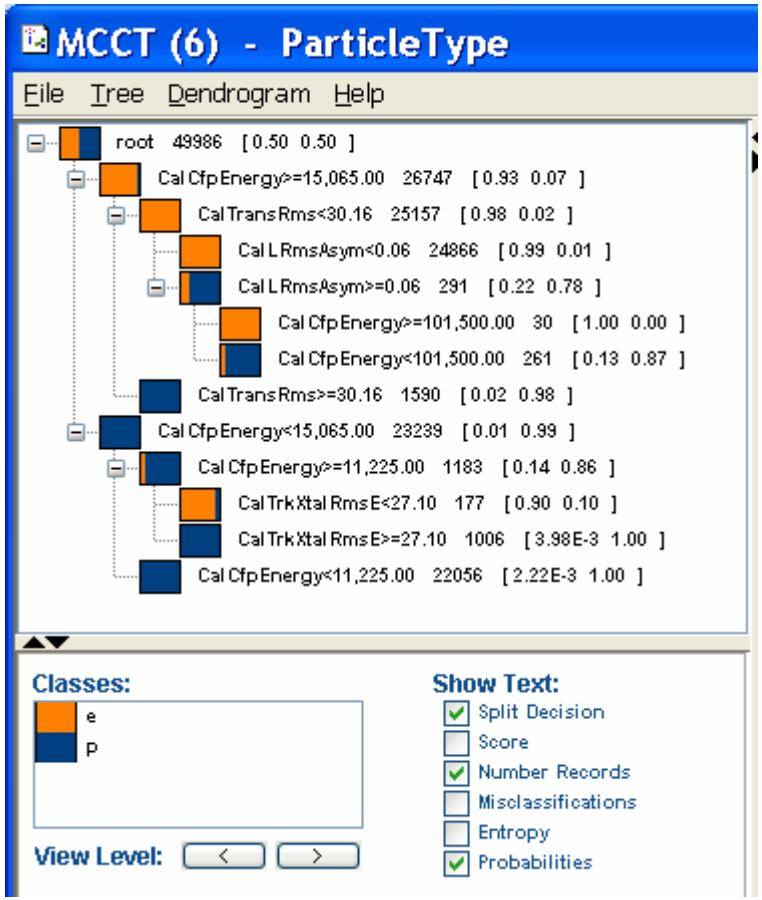

CalTransRms
(ParticleType=e)
Counts
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
25.00
50.00
75.00
100.00
Bin Range

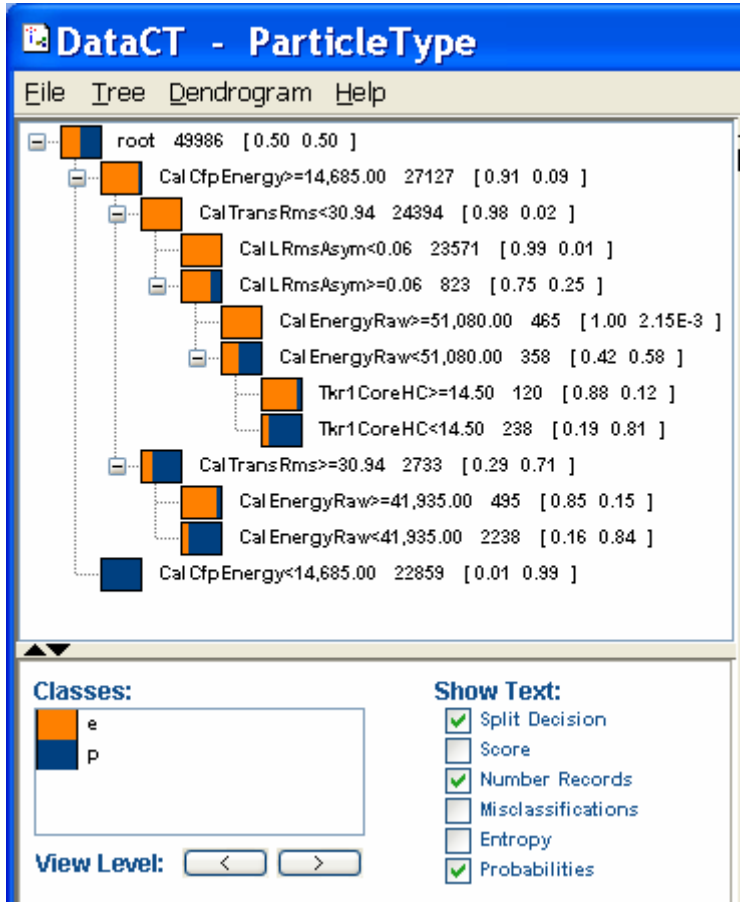

CalTransRms
(ParticleType=p)
Counts
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
25.00
50.00
75.00
100.00
Bin Range

Analysis – step 1



Classification Trees





- CT are slightly different
  - they are different data sets
- Split points reflect known discrepancies

Some results – CT agreement

1. trained on MC

		Predicted		Totals
		e	p	
Observed	e	24870	120	24990
	p	203	24793	24996
Totals		25073	24913	49986

	Observed		Overall
	e	p	
% Agree	99.5%	99.2%	99.4%

2. trained on Data

		Predicted		Totals
		e	p	
Observed	e	24405	585	24990
	p	246	24750	24996
Totals		24651	25335	49986

	Observed		Overall
	e	p	
% Agree	97.7%	99.0%	98.3%

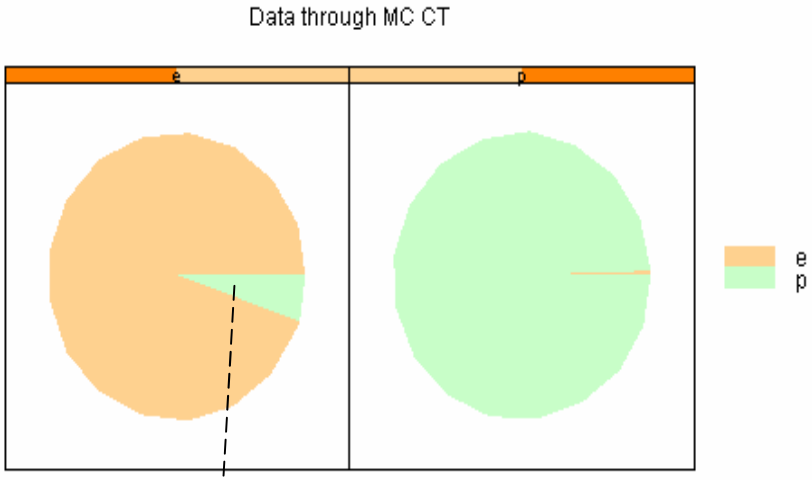
3. trained on MC applied to Data

		Predicted		Totals
		e	p	
Observed	e	23696	1294	24990
	p	152	24844	24996
Totals		23848	26138	49986

	Observed		Overall
	e	p	
% Agree	94.8%	99.4%	97.1%

- Classification intrinsically harder on data sample (1. vs 2.)
  - Discrepancies in variables?
  - Residual contaminations? Note e/p asymmetry in agreement
- p prediction higher (0.4%) wrt data CT (3. vs 1.)
  - due to CalTransRms shifted to higher values for p and e in data wrt MC?
- p contamination in e sample increase factor 10 (3. vs 1.)
  - let's check real p contamination in data sample

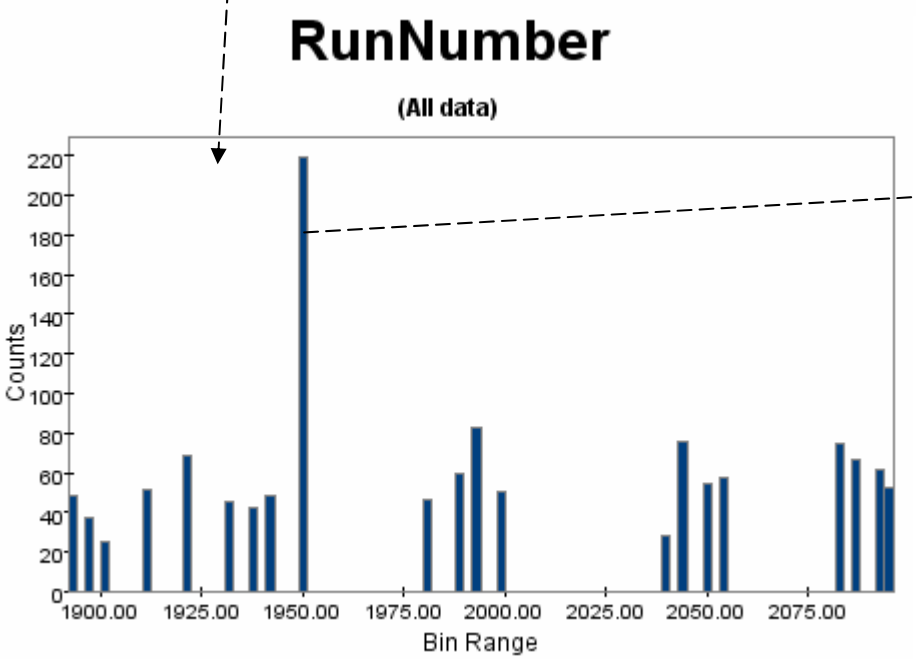
Closer look at misclassified e

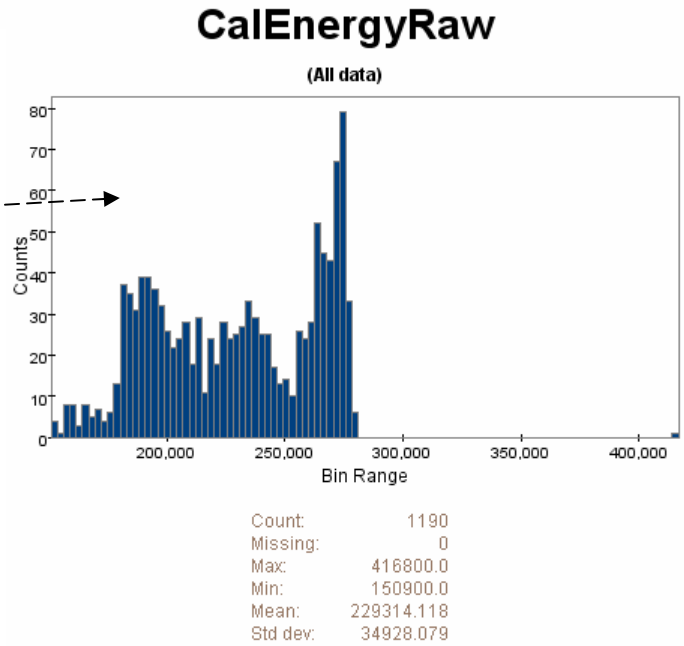


A good fraction of these come from run 1951

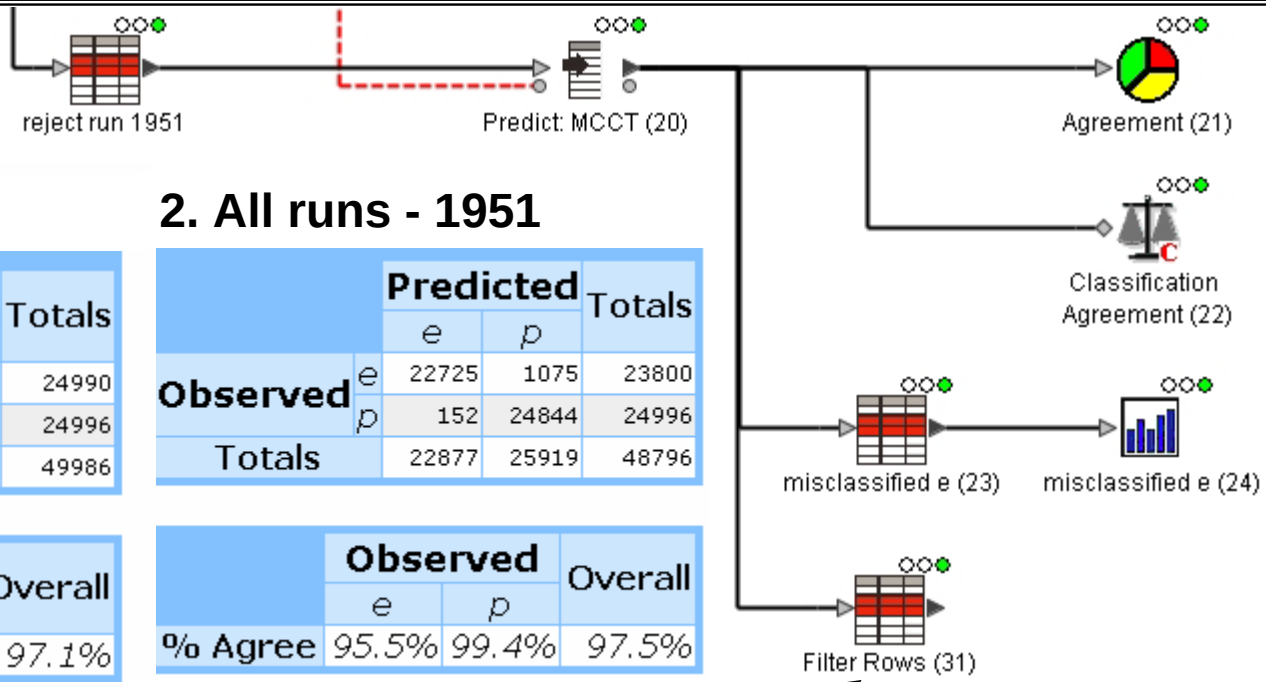
Gaps Special

Run	Particle	Energy	Impact point	BeamXYdir
BT-1796-v6r0925p2-GLAST	Electrons	99.GeV	(190.50, 13.70)	(-0.53, 0.33)
BT-1834-HEAD1.131	Electrons	99.GeV	(578.00, -1.00)	(-0.49, 0.33)
BT-1846	Electrons	99.GeV	(-201.00, 0.00)	(-0.53, 0.33)
BT-1951	Electrons	282.GeV	(749, -7, -95)	(90.0, 0.20)





Analysis - step 2 remove run 1951



1. All runs

		Predicted		Totals
		e	p	
Observed	e	23696	1294	24990
	p	152	24844	24996
Totals		23848	26138	49986

	Observed		Overall
	e	p	
% Agree	94.8%	99.4%	97.1%

2. All runs - 1951

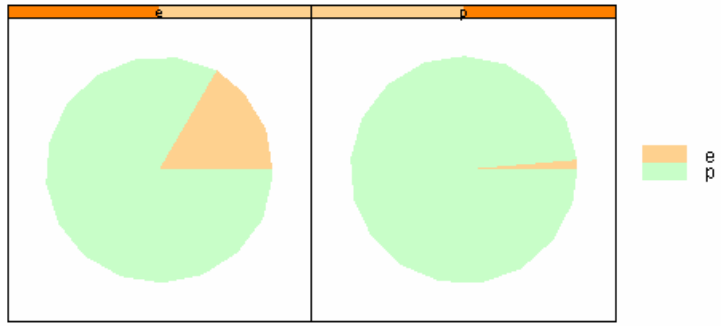
		Predicted		Totals
		e	p	
Observed	e	22725	1075	23800
	p	152	24844	24996
Totals		22877	25919	48796

	Observed		Overall
	e	p	
% Agree	95.5%	99.4%	97.5%

(CalCfpEnergy > 15065 & CalTransRms>30.16 & CalTransRms <30.94) | (CalCfpEnergy>14685 & CalCfpEnergy< 15065)

Select events classified differently in first two split due to variable shifts

- electrons are mainly misclassified!
- but they are not the single source of differences (176 events)



Conclusions

- Comparison with Alex cuts?
- Data must be clean to assess prediction power of a CT
- We deliberately went the wrong path building CT first and reverse-engineered data cleaning
- Current discrepancies in MC are not negligible and do produce misclassifications
  - Should we optimize the CT AFTER the main cut on CalCfpEnergy which eliminates 90% of protons?
- learning IM and we like it
- we can read Bill's IM worksheet and are now going through pass5 analysis
  - plan to check that directly against BT data

Appendix - Data set and cuts

- Data-v6r0922p4, MC GLAST-v7r1117p1
- CalTransRms>0 && CalTransRms>0 && Tkr1ToTTrAve>0 && CalCfpEnergy<500000
- Must implement cut on GemDeltaEventTime>10000
- Proton runs: 2237, 2251: , 2252, 2253, 2363, 1755
- Electron runs:
  - 2082: 'CalEnergyRaw> 1000 && CalCfpEnergy> 1000 '
  - 2087: 'CalEnergyRaw> 1000 && CalCfpEnergy> 1000 '
  - 2092: 'CalEnergyRaw> 1000 && CalCfpEnergy> 1000 '
  - 2096: 'CalEnergyRaw> 1000 && CalCfpEnergy> 1000 '
  - 2039: 'CalEnergyRaw> 2000 && CalCfpEnergy> 20000
  - 2044: 'CalEnergyRaw> 2000 && CalCfpEnergy> 2000 '
  - 2050: 'CalEnergyRaw> 2000 && CalCfpEnergy> 2000 '
  - 2054: 'CalEnergyRaw> 2000 && CalCfpEnergy> 2000 '
  - 1981: 'CalEnergyRaw> 5000 && CalCfpEnergy> 5000 '
  - 1988: 'CalEnergyRaw> 5000 && CalCfpEnergy> 5000 '
  - 1911: 'CalEnergyRaw> 5000 && CalCfpEnergy> 5000 ‘
  - 1993: 'CalEnergyRaw> 5000 && CalCfpEnergy> 5000 ',
  - 1999: 'CalEnergyRaw> 5000 && CalCfpEnergy> 5000 ',
  - 1892: 'CalEnergyRaw> 5000 && CalCfpEnergy> 10000 ',
  - 1898: 'CalEnergyRaw> 5000 && CalCfpEnergy> 10000 ',
  - 1902: 'CalEnergyRaw> 10000 && CalCfpEnergy> 50000 ',
  - 1922: 'CalEnergyRaw> 10000 && CalCfpEnergy> 50000 ',
  - 1932: 'CalEnergyRaw> 10000 && CalCfpEnergy> 60000 ',
  - 1938: 'CalEnergyRaw> 10000 && CalCfpEnergy> 50000 ',
  - 1942: 'CalEnergyRaw> 10000 && CalCfpEnergy> 50000 ',
  - 1951: 'CalEnergyRaw> 150000 && CalCfpEnergy> 100000'