Second Machine Learning in High Energy Physics Summer School 2016

20-26 June 2016
Lund University

DATA DOPING

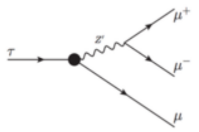
Solution for "Flavour of Physics" challenge

Dr. Vicens Gaitan
Grupo AIA

AGENDA

- “Flavour of Physics” Kaggle Challenge
- Why is so hard to “discover” the invariant mass?
- How to win the challenge: leasons learned
- Breaking the rules: Data Doping
- Machine Learning in HEP
- Conclusions

"FLAVOUR OF PHYSICS" KAGGLE CHALLENGE



Completed • $15,000 • 673 teams

Flavours of Physics: Finding $\tau \rightarrow \mu\mu\mu$

Mon 20 Jul 2015 – Mon 12 Oct 2015 (4 months ago)

Yandex Data Factory

SCHOOL OF DATA ANALYSIS

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Go Polar Bears ‡ *	1.000000	49	Mon, 12 Oct 2015 22:57:38
2	↑1	Alexander Gramolin ‡ *	0.999998	12	Mon, 12 Oct 2015 18:38:07
3	↓1	Josef Slavicek ‡ *	0.999897	25	Mon, 12 Oct 2015 21:49:53
4	—	Michal Wojcik	0.999225	35	Mon, 12 Oct 2015 23:57:46 (-3h)
5	—	rakhlin	0.998338	31	Mon, 12 Oct 2015 23:32:18 (-5.8h)
6	—	Archy ‡	0.997784	47	Mon, 12 Oct 2015 20:31:53 (-7.8h)
7	—	Faron	0.995918	66	Mon, 12 Oct 2015 18:15:46
8	—	Alejandro Mosquera	0.994946	28	Mon, 12 Oct 2015 15:23:51 (-19.7h)
9	—	Anton Laptiev	0.994894	61	Mon, 12 Oct 2015 23:56:37
10	—	Andrzej Prałat	0.993957	14	Mon, 12 Oct 2015 18:25:39 (-0.3h)
11	—	Ivanhoe	0.993692	35	Mon, 12 Oct 2015 23:17:39
12	—	George Solymosi	0.993646	95	Mon, 12 Oct 2015 23:58:45 (-0.6h)
13	—	PhysicsTau	0.993099	90	Mon, 12 Oct 2015 22:30:42
14	↑1	Grzegorz Sionkowski	0.992031	49	Mon, 12 Oct 2015 23:50:56 (-27.2h)
15	**↓1**	**Vicens Gaitan [0.989012 physically sound]**	**0.991860**	**85**	**Mon, 12 Oct 2015 20:56:04 (-5.9h)**
16	—	achm	0.991841	105	Mon, 12 Oct 2015 13:06:31 (-44.1h)
17	—	bgeol	0.991709	14	Tue, 06 Oct 2015 03:56:14 (-5.3d)

"FLAVOUR OF PHYSICS" KAGGLE CHALLENGE

- Training sample, $\tau \to \mu\mu\mu$
  - Signal - simulated
  - Background - real (taken from regions where signal cannot occur)
  - 40+ features
- Goal: Classify signal vs Background. Figure of merit: Weighted AUC
- Constrain1: Probability of signal canot be correlated with tau mass (CVM test)
- Control channel, $D \to \phi\pi$
  - well studied
  - has similar topology to $\tau \to \mu\mu\mu$
  - Available both MC and real data samples
- Constrain 2: Model will not discriminate real daa from MC for the control channel) (K-S test)

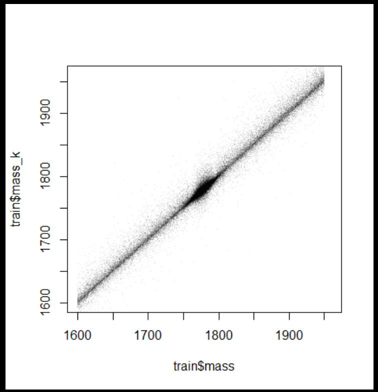
(More details in:

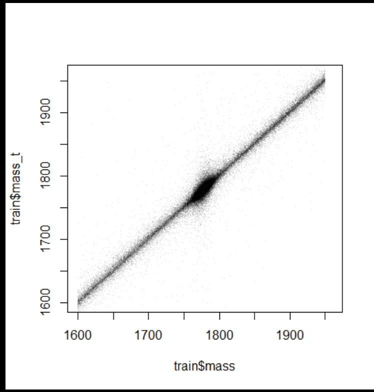
https://indico.cern.ch/event/433556/contributions/1930574/attachments/1230492/1803909/Ustyuzhanin__FoP_Summary.pdf)

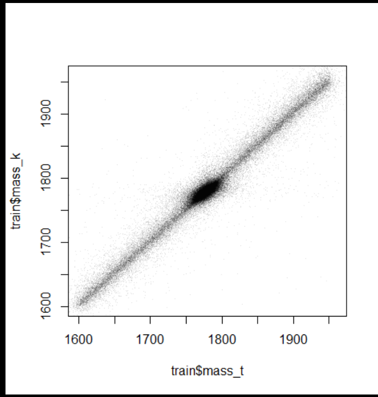
WHY IS SO HARD TO "DISCOVER" THE INVARIANT MASS?

- Fact 1: The tau invariant mass can be reconstructed with high accuracy from kinematic variables (p0,pt,eta) and/or lifetime & time of flight

```
ptau<-function(d){
  # muon mass in MeV/c^2
  mmu = 105.6583715
  # calculate tau energy
  tau_e = sqrt(d$p0_p ** 2 + mmu**2) + sqrt(d$p1_p ** 2 + mmu**2) +sqrt(d$p2_p ** 2 + mmu**2)
  # calculate pz of tau candidate
  tau_pz = d$p0_pt * sinh(d$p0_eta) + d$p1_pt * sinh(d$p1_eta)+d$p2_pt * sinh(d$p2_eta)
  # calculate momentum of tau candidate
  tau_p =sqrt(d$pt ** 2 + tau_pz ** 2)
  # calculate eta of tau candidate
  tau_eta = asinh(tau_pz / d$pt)
  # calculate mass of tau candidate
  tau_m2=tau_e ** 2 - tau_p ** 2
  tau_m2[tau_m2<0]=0
  tau_m_k=sqrt(tau_m2)
  #M = tau_p*LifeTime*c/FlightDistance
  #c Speed of light
  c= 299.792458
  tau_m_t=tau_p*d$LifeTime*c/d$FlightDistance
  return(list(tau_e,tau_pz,tau_p,tau_eta,tau_m))
}
```

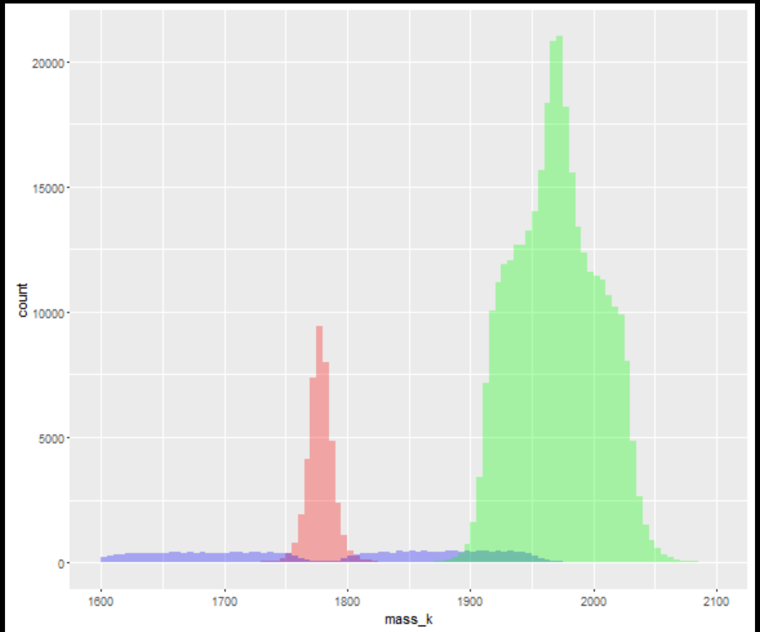




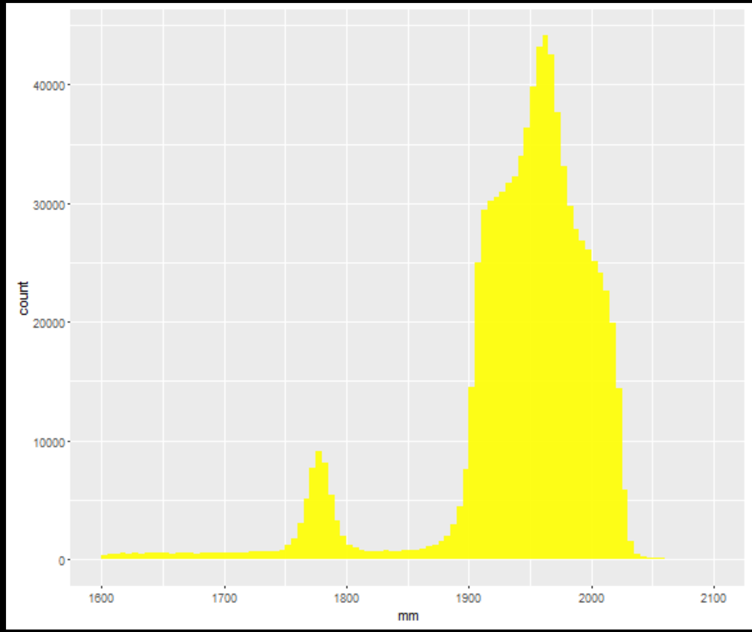


WHY IS SO HARD TO "DISCOVER" THE INVARIANT MASS?

- Fact 2: The reconstructed tau mass separates nicely signal and background because the background spectrum has a "hole" for decays coming from a true tau (Real data (background) in this window can contain "signal")



Train & Agreement

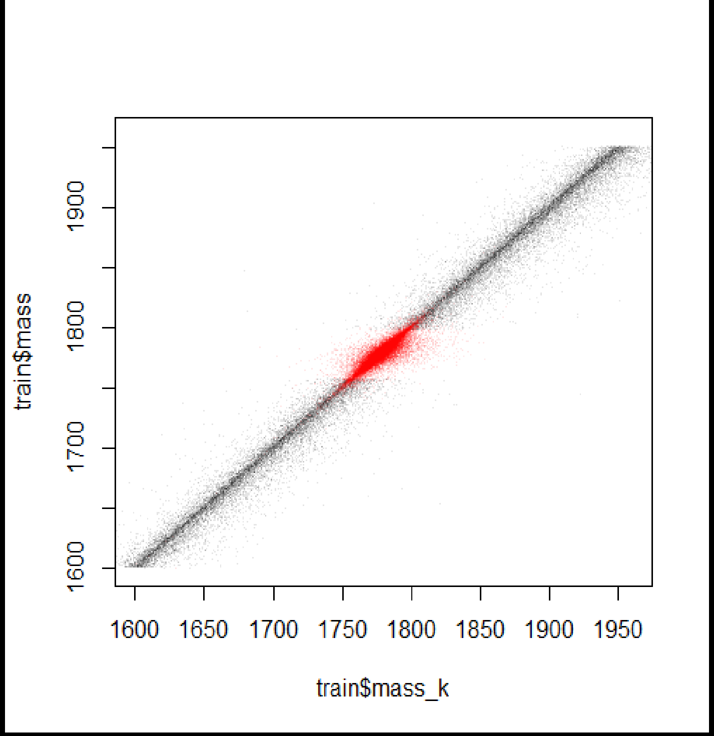


Test

WHY IS SO HARD TO "DISCOVER" THE INVARIANT MASS?

- Toy models (not taking into account agreement & correlation):
  - XGBoost
  - 3-fold CV
  - Par("max_depth"=5,"eta"=.1)

1. Using true mass : wAUC = 0.999999999999999(89) +/- 8.4e-16 (!!)
2. Using mass_k: wAUC = 0.997(51) +/- 0.00038
3. Using mass_t: wAUC = 0.996(81) +/- 0.00021
4. Using both wAUC = 0.999(83) +/- 0.00012



Reconstructed Mass is THE Golden Feature

WHY IS SO HARD TO "DISCOVER" THE INVARIANT MASS?

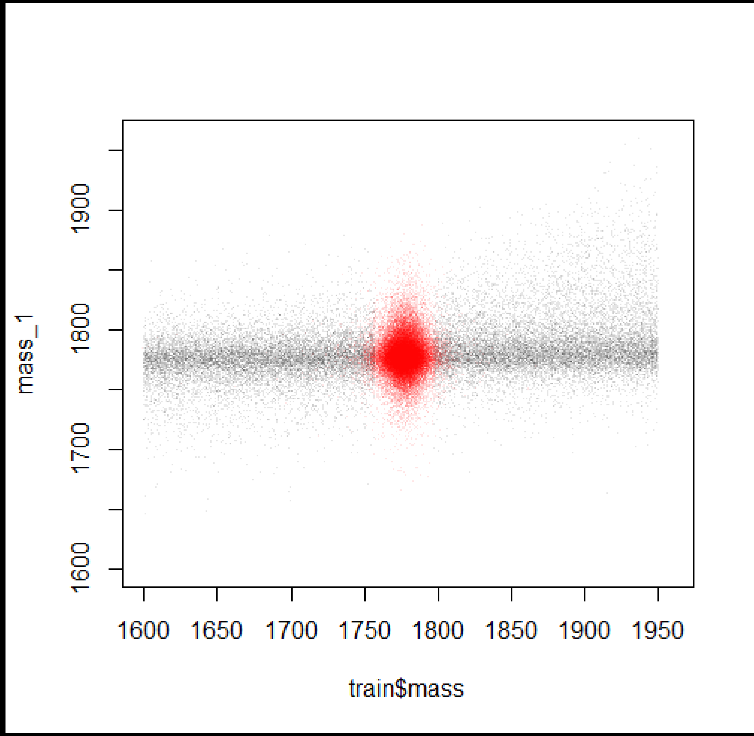
- BUT using as variables:

c("p0_p","p1_p","p2_p","p0_pt","p1_pt","p2_pt","p0_eta","p1_eta","p2_eta","pt","LifeTime","FlightDistance")

c("tau_e","tau_p","LifeTime","FlightDistance")

wAUC = 0.85(15) +/- 0.0032 ??????

And trying to fit the mass with XGBoost : we obtain:

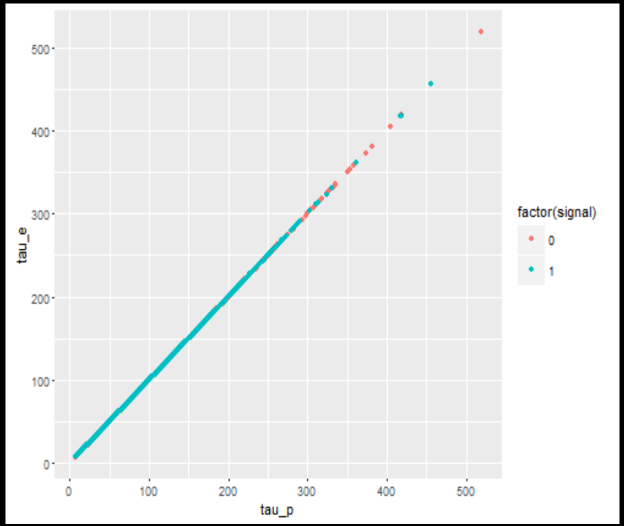


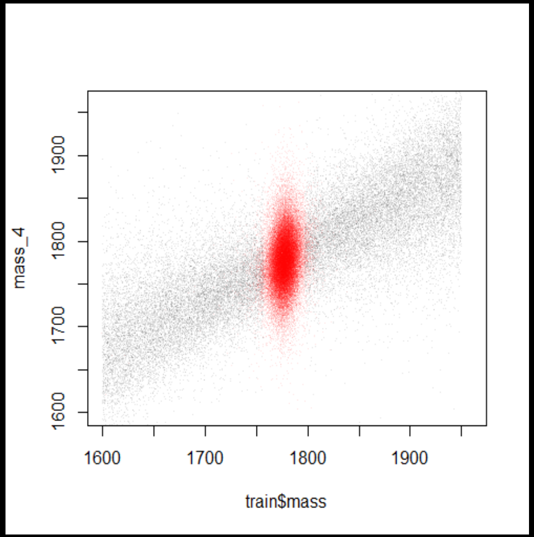
WHY IS SO HARD TO "DISCOVER" THE INVARIANT MASS?

The reason: Highly correlated variables: mass is an effect of 1 over 2500

- Solution: uncorrelate variables with PCA

- wAUC = 0.947(73) +/- 0.00099





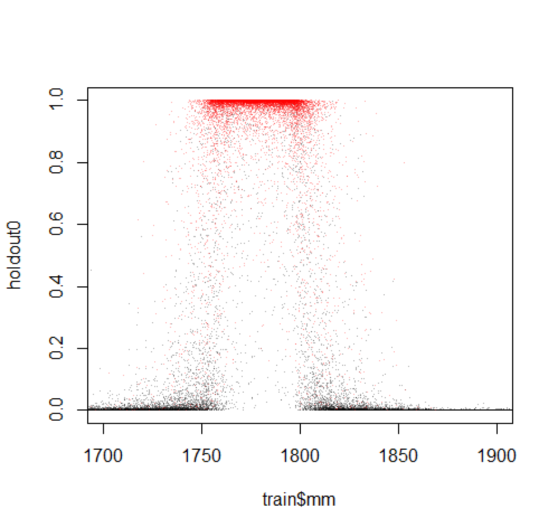
- Gradient Boosting Trees are not able build a representation of Invariant Mass
- Maybe Deep Learning can do it?

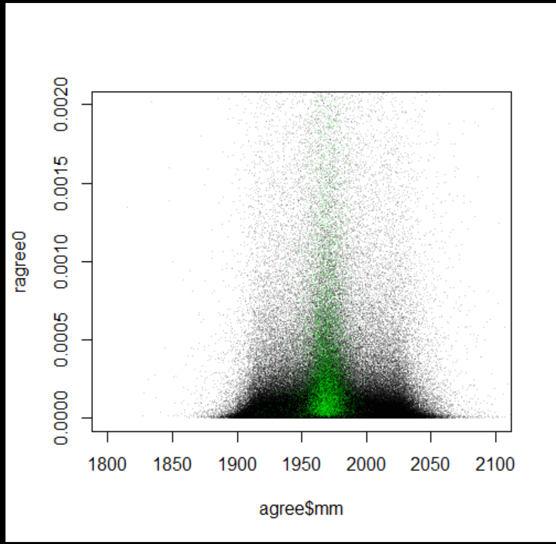
HOW TO WIN THE CHALLENGE: LEASONS LEARNED

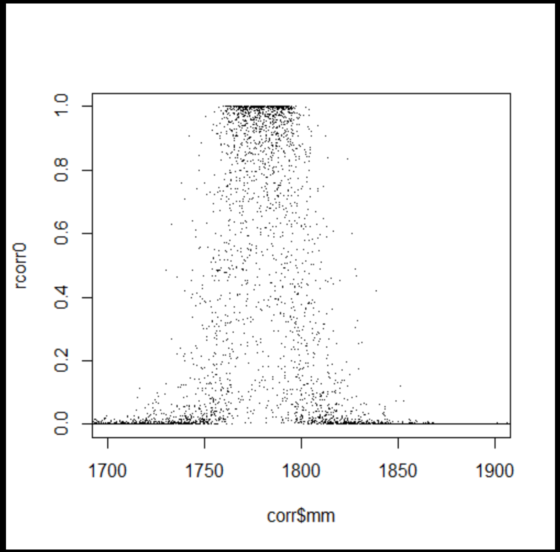
Recipe to win the challenge

1. Add the reconstructed tau mass (don't bother about mass correlation test)
2. Use all available variables (profit from bad simulated MC variables to separate signal from real background)

AUCw=0.9999920 CVM=0.0848 K-S=0.2226



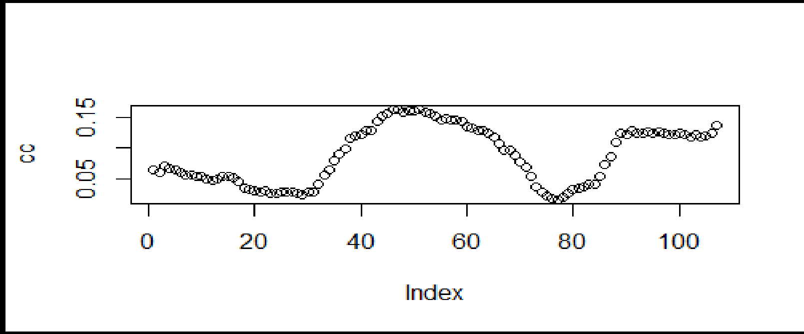




3. Hack the Correlation and Agreement Test ;)

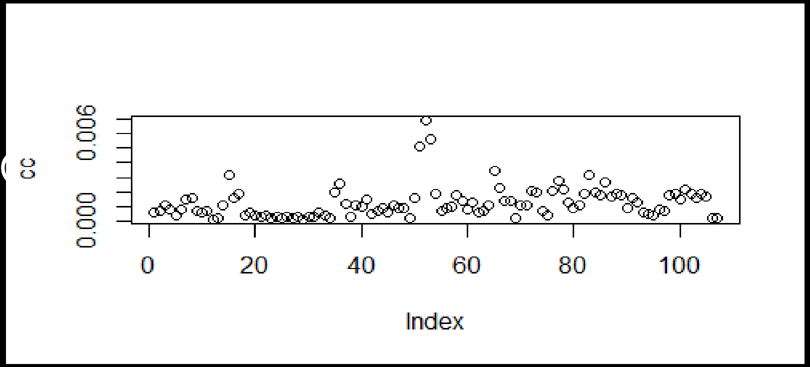
HOW TO WIN THE CHALLENGE: LEASONS LEARNED

Correlation Test : Correlation between classifier output p and mass over a rolling window:



CVM= 0.85

Define p' = .99 * p ^5000 + .01 * RND and calculate the CVM



- CVM= 0.0012 !!
- p' has very similar AUC as p
- p' for agreement sample ->0 because p' agreement <<1

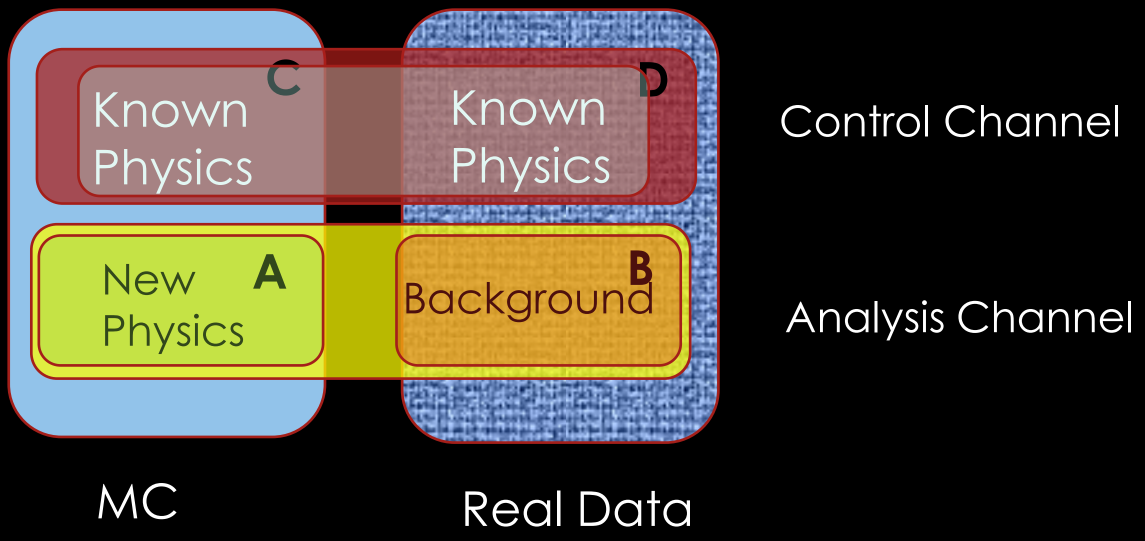
AUCw= 0.9999921   CVM=0.0014   K-S=0.0088

- Useless for physics : Just exploiting the background mass gap

BREAKING THE RULES: DATA DOPING

- Recipe to build a physically sound classifier:
  1. Not to use reconstructed mass, nor features allowing easy mass reconstruction
  2. Try to not use variable regions for which the Monte Carlo simulation doesn't agree with real data

In order to fullfill 2 we have to break the rules and take a look to the control channel



Goal: train a classifier able to separate A from B, but not C from D

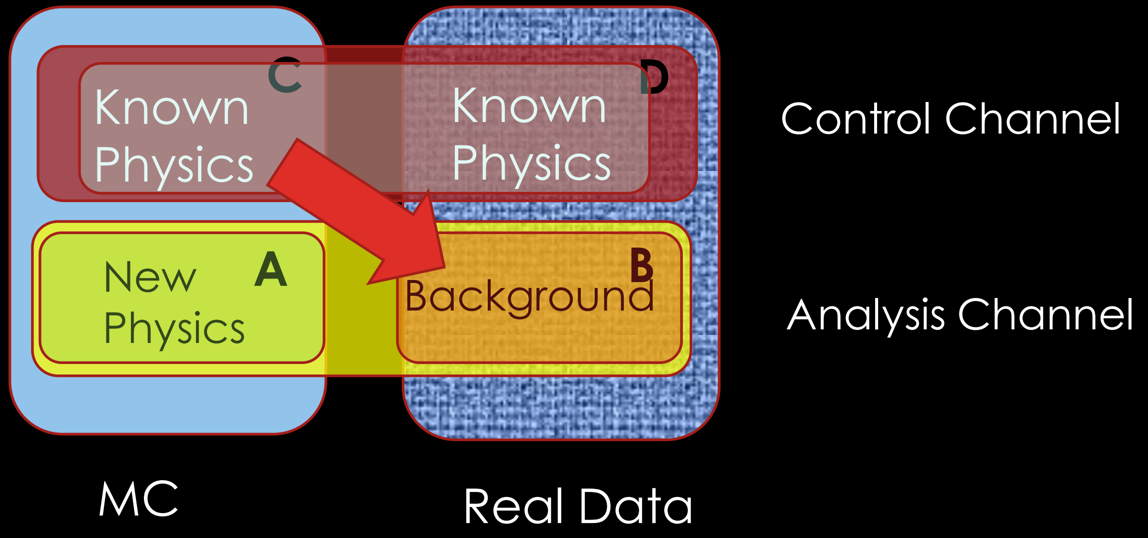
Max(wAUC(A,B)) with KS(C,D)<epsilon

Hypothesis: Control Channel & Analysis channel share the same MC "defects"

BREAKING THE RULES: DATA DOPING

- The idea is to "dope" (in the semiconductor meaning) the training set with a **small number of Monte Carlo events from the control channel , but labeled as background**.

  This disallow the classifier to pick features discriminating data and Monte Carlo.



There are two parameters that regularize the learning:

- The number of "doping" events
- the complexity of the classifier (for instance number of trees)

BREAKING THE RULES: DATA DOPING

Bck vs $\tau \rightarrow \mu\mu\mu$

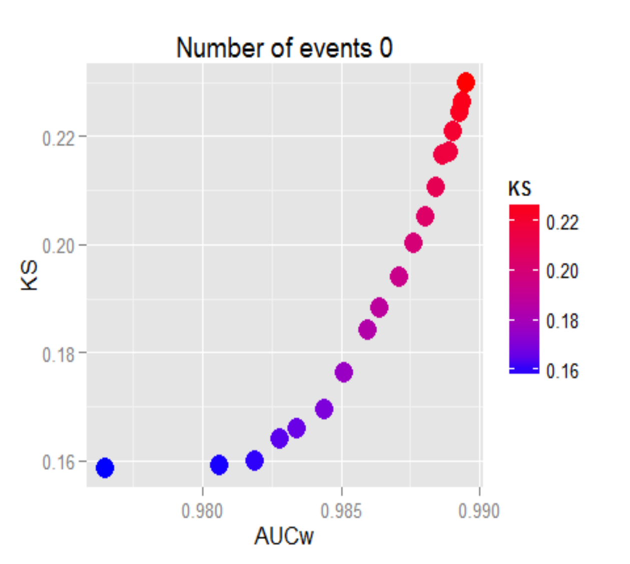
$D \rightarrow \phi\pi$ $D \rightarrow \phi\pi$

Data Doping

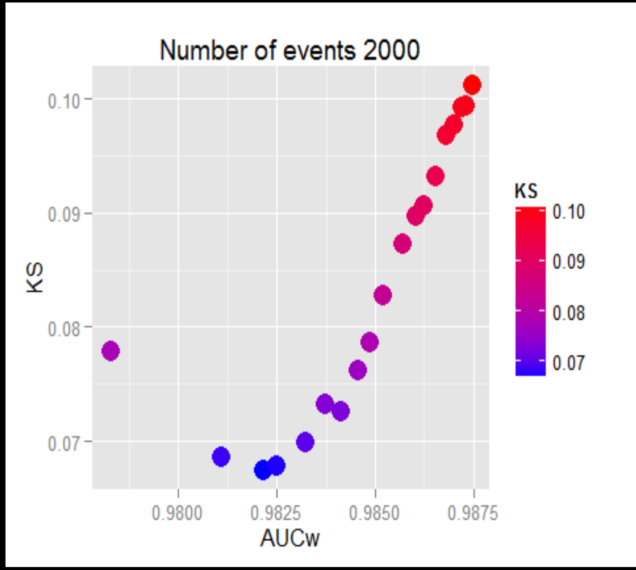
Bck + Eps * $D \rightarrow \phi\pi$ vs $\tau \rightarrow \mu\mu\mu$

BREAKING THE RULES: DATA DOPING

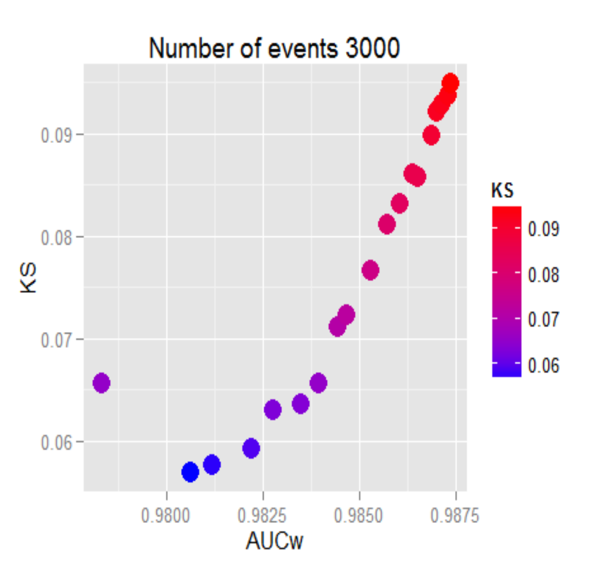
Grid search over Classifier complexity (n_ trees) and Number (weight) of doping events
Dammit! A new hyperparameter….



Free classifier

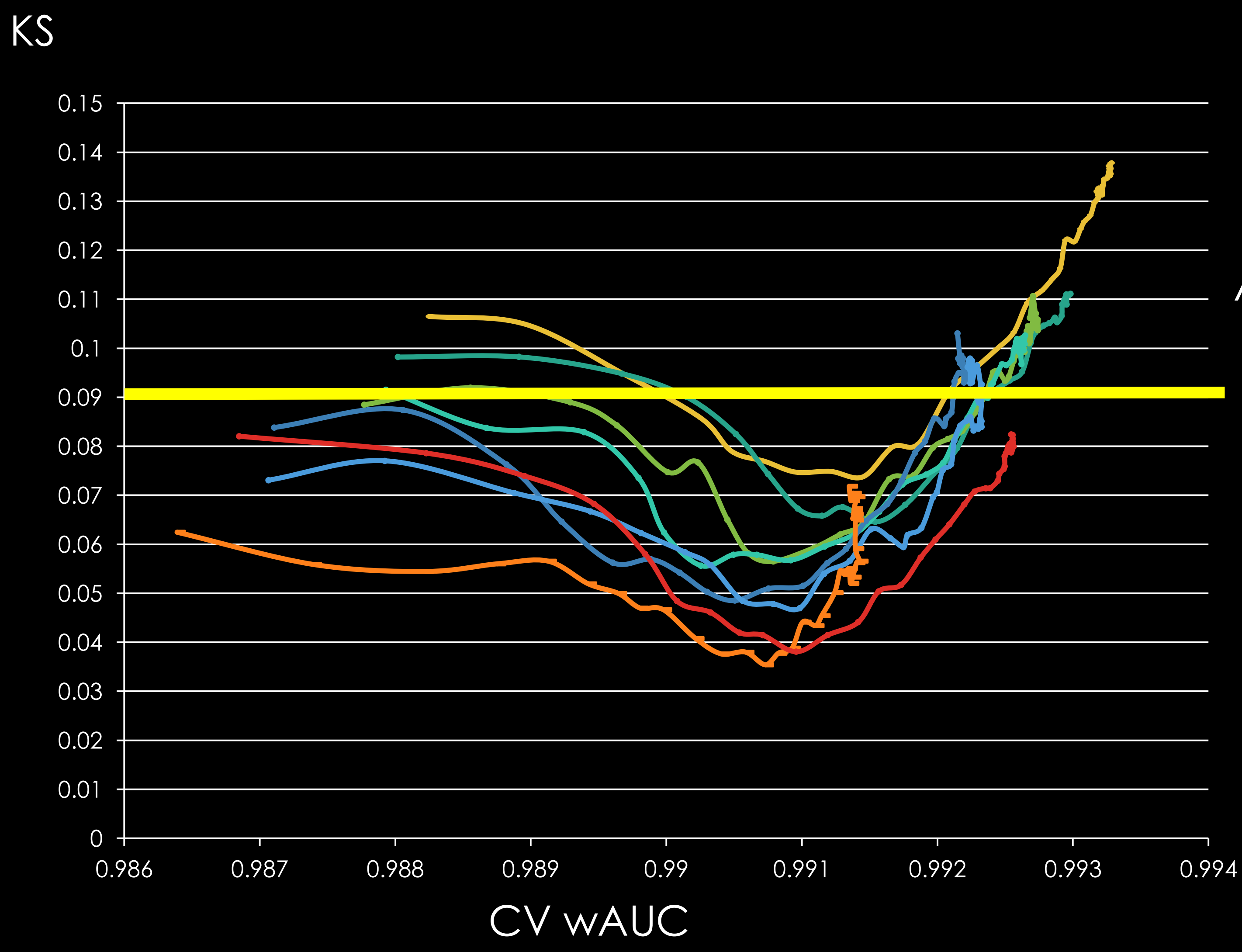


Doping events: 2000



Doping events: 3000

BREAKING THE RULES: DATA DOPING

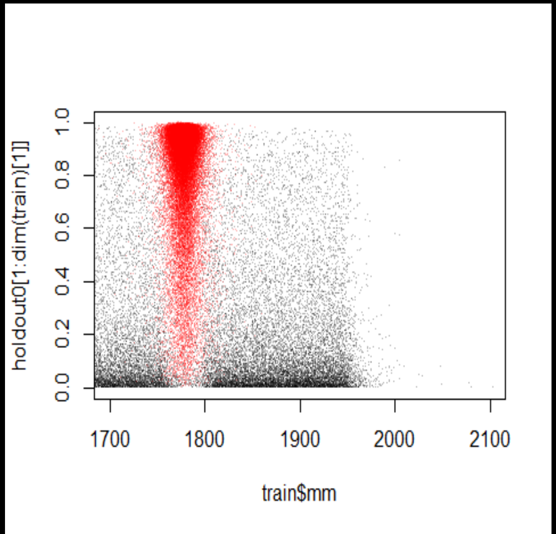


Optimal Doping Events: 3960

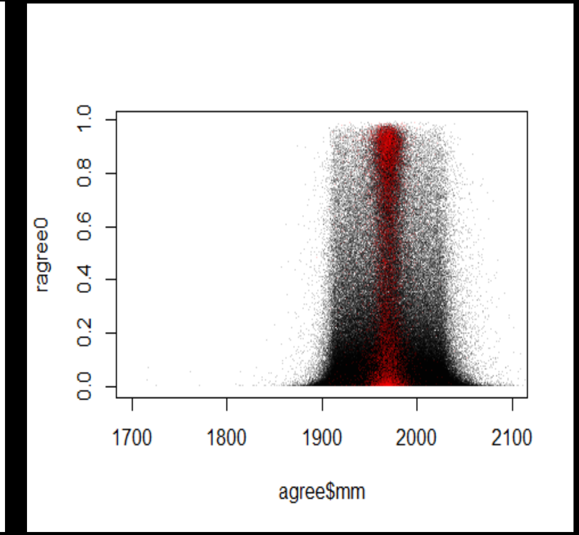
Private LB

AUCw= 0.9893 CVM=0.0011 K-S=0.087

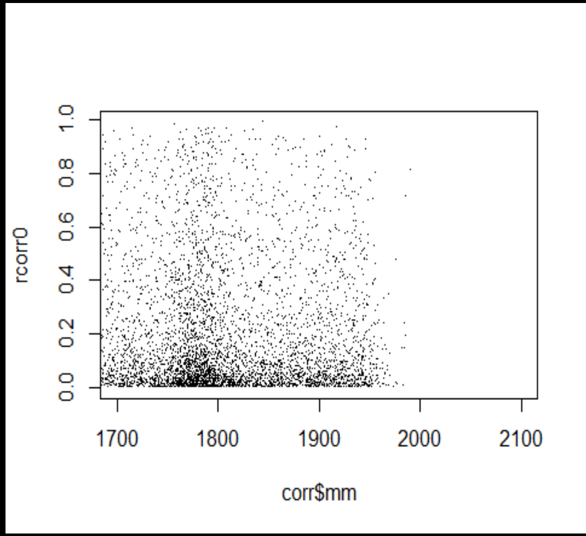
BREAKING THE RULES: DATA DOPING



Analysis Channel



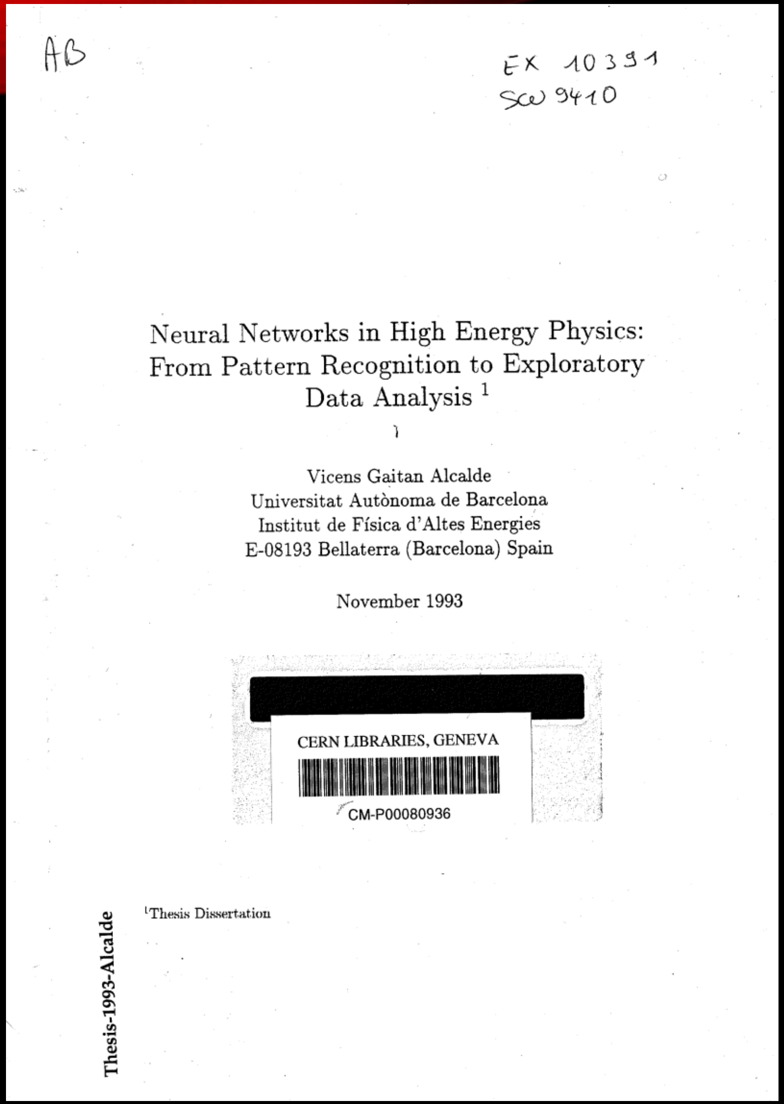
Control channel



Correlation Test

- Good discriminating power in analysis channel
- No separation for the control channel
- Classifier not correlated with mass

- Probably good for physics….



AB

EX 10391
Scw 9410

Neural Networks in High Energy Physics:
From Pattern Recognition to Exploratory
Data Analysis [1]

Vicens Gaitan Alcalde
Universitat Autònoma de Barcelona
Institut de Física d'Altes Energies
E-08193 Bellaterra (Barcelona) Spain

November 1993

[1] Thesis Dissertation

MACHINE LEARNING IN HEP

Today we have

- the right tools
- data availability
- the computer power

- But new physics is difficult to discover unless you know what are you looking for…

- A complementary approach can be to use unsupervised learning (only real data driven, we have lots of them)

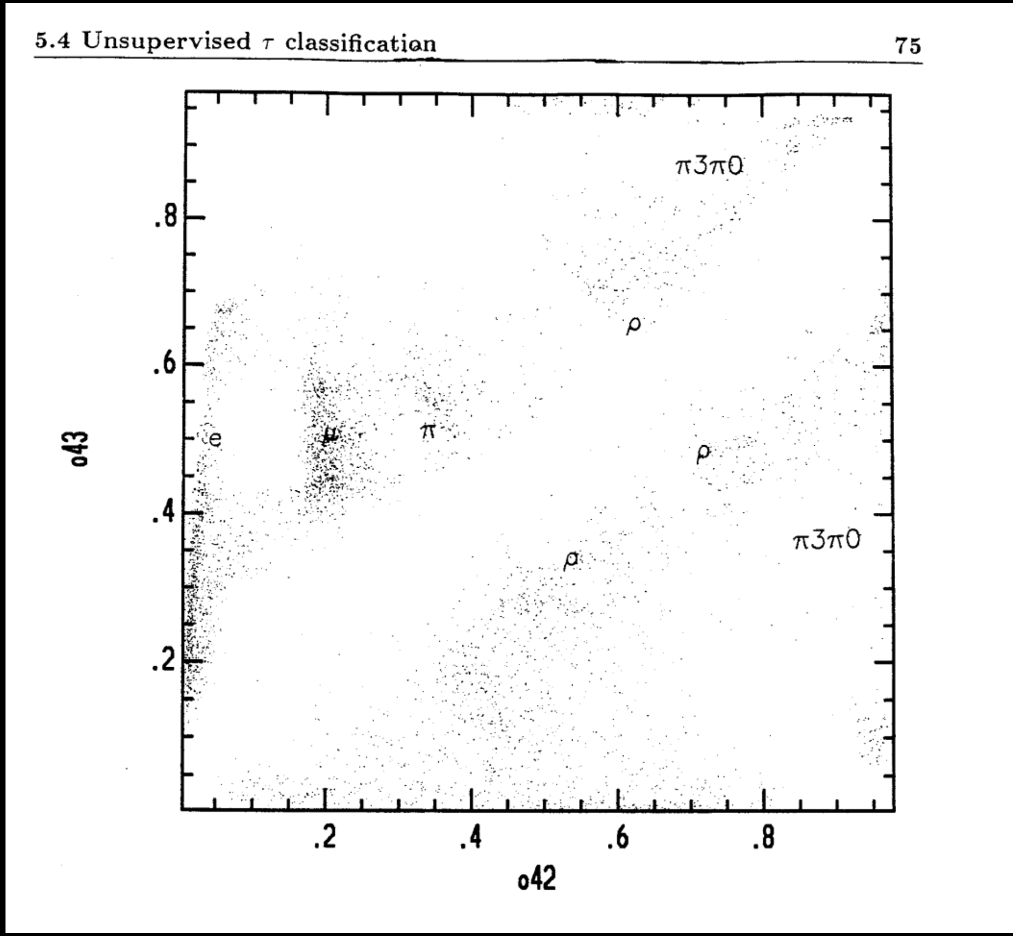
MACHINE LEARNING IN HEP

Example: exploring tau decay at LEP (ALEPH 1993)
(yes, e+ e- physics is cleaner…)

Feeding an autoencoder with “elaborated” detector data we are able to “discover” different decay modes looking at the compressed representation without a physics model (MC)

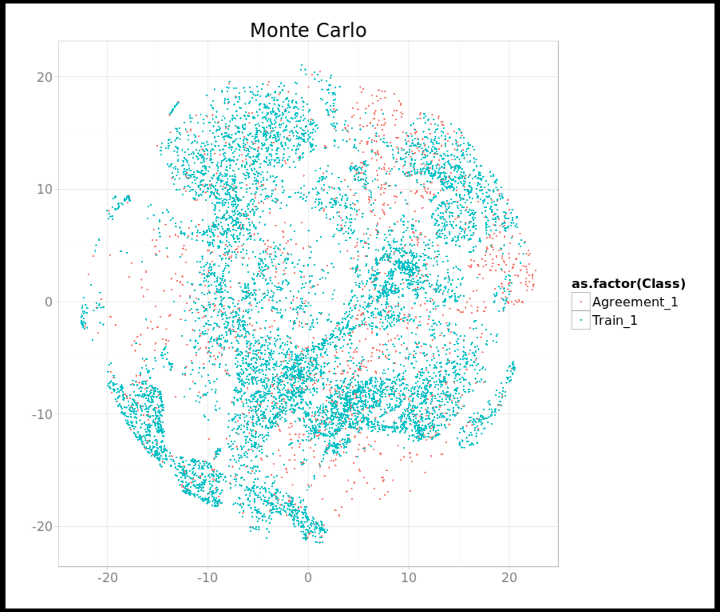
Today is possible to do “end to end” autoencoding from raw detector data

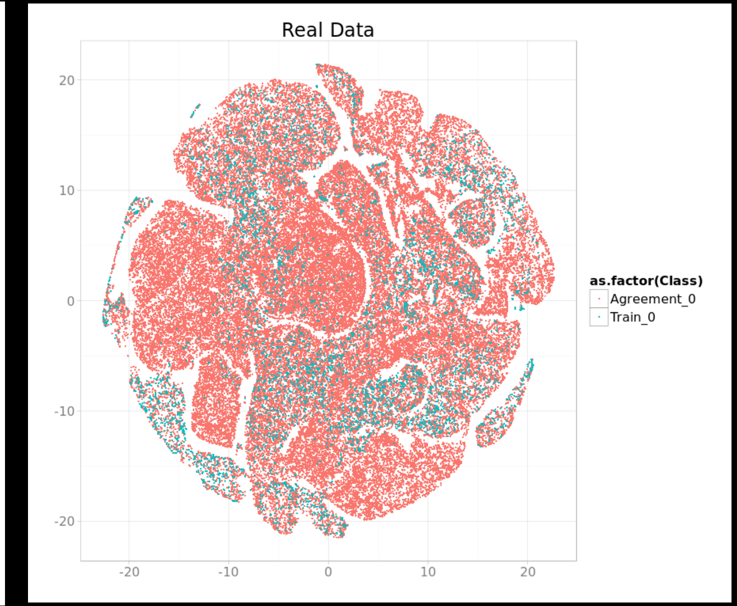
Input neuron	Variable Description	Kolmogorov C.L.
1	Number of charged tracks in hemisphere +	0.947
2	Number of charged tracks in hemisphere -	0.339
3	Number of neutral tracks in hemisphere +	0.010
4	Number of neutral tracks in hemisphere -	0.047
5	Total charged energy in hemisphere +	0.131
6	Total charged energy in hemisphere -	0.078
7	Total neutral energy in hemisphere +	0.874
8	Total neutral energy in hemisphere -	0.995
9	Number of identified μ in hemisphere +	1.000
10	Number of identified μ in hemisphere -	1.000
11	Number of identified electrons in hemisphere +	0.367
12	Number of identified electrons in hemisphere -	0.921
13	Number of identified γ in hemisphere +	0.258
14	Number of identified γ in hemisphere -	0.746
15	Planarity	0.489
16	Total momentum in hemisphere +	0.523
17	Total momentum in hemisphere -	0.534
18	Invariant mass	0.90621
-	Output neuron	0.457

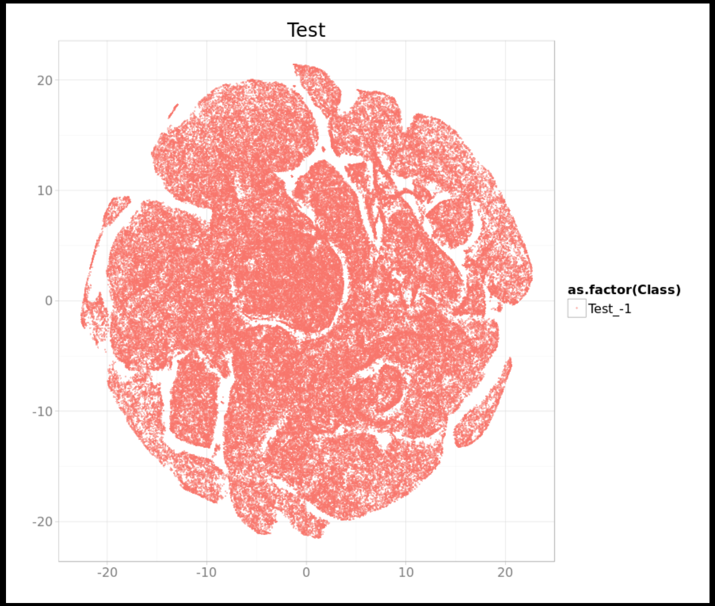


MACHINE LEARNING IN HEP

Example: t-sne with the challenge data: Look at the fine structure….







MC:
Control Channel
Signal

Real Data (all you see is real!)
Control Channel
Background

Test

CONCLUSIONS

- Machine learning algorithms alone can fail to discover tiny effects in the data (1777 MeV is only 1.e-4 of the energy at the center of mass)
  - Use your knowledge: Try to reduce your data using fundamental simmetries, like Lorentz invariance, detector geometry…
- Be aware of the test you are using to assure the classifier validity:
  - If a test can be "hacked" an enough powerfull machine learning algorithm will find the way
- If it is possible, try to use non supervised methods(without MC) to gain insight in your data