Multivariate Data Analysis in HEP. Successes, challenges and future outlook

Helge Voss (MPIK-Heidelberg)

ACAT 2014 Prague

"The presentation you are about to watch might look like (bad) parody. Its content is not fact checked. Its reporter is not a journalist. And his opinions might not be fully thought through."

freely adopted from "John Stuart"

Multivariate Data Analysis in HEP. Successes, challenges and future outlook

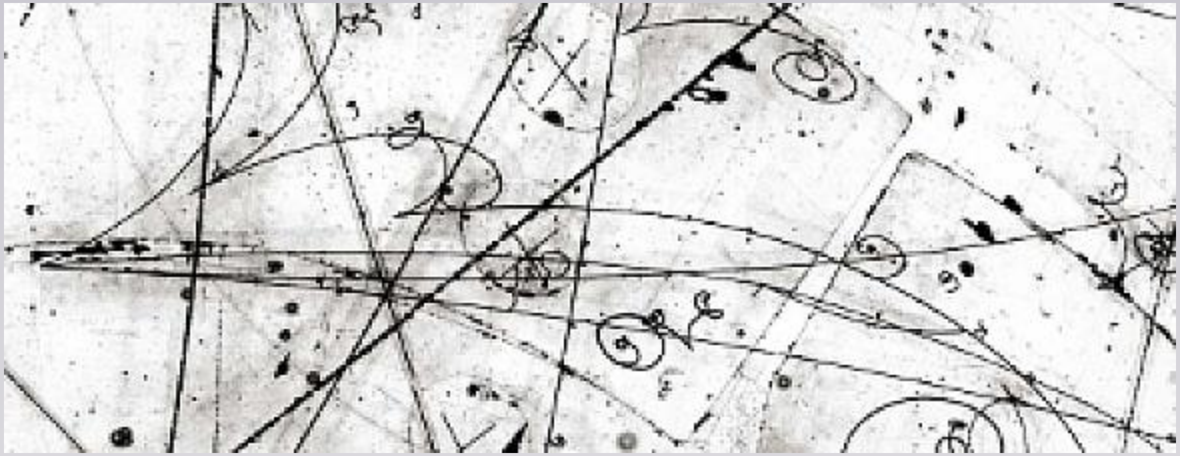
- A personal view of MVA history in HEP
- Some highlights

- Challenges

- What about the future ?

A Short History Of MVA

- **Already in the very beginning intelligent "Multivariate Pattern Recognition" was used to identify particles ….**

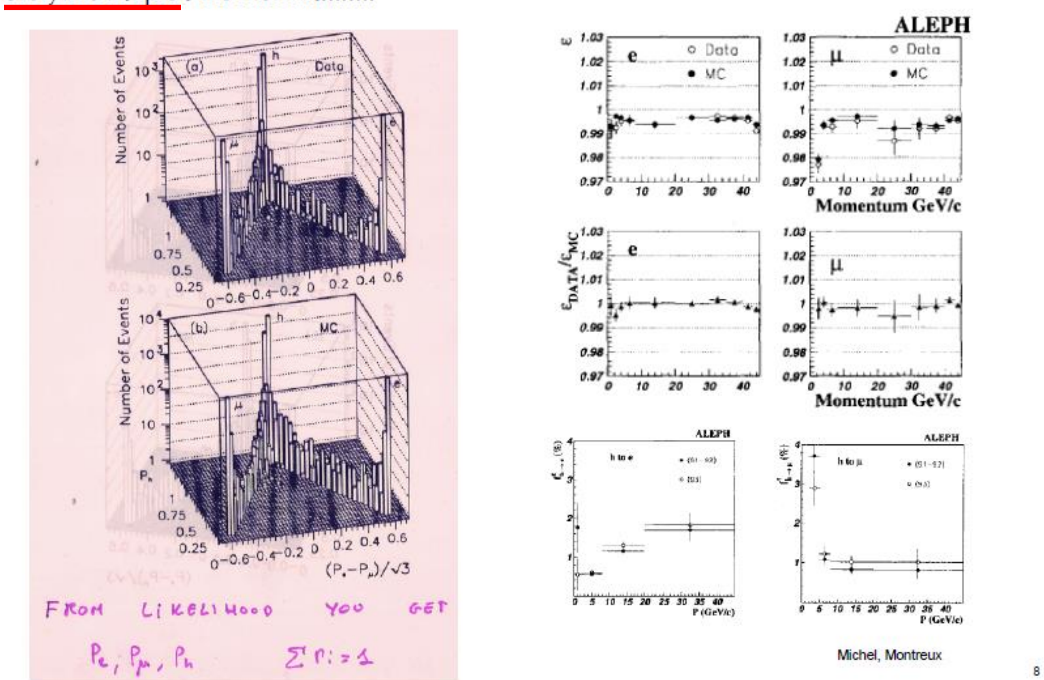


- **But later it became a bit 'out of fashion' with the advent of computers**
  - **… although I guess some Fisher-Discriminants (Linear Decision Boundary) were used her and there .. If I remember correctly my PhD supervisor mentioning such things being used back in MARKIII**

A Short History Of MVA

- **TAUPID ..** (my first encounter .. Again, that might absolutly not be the first or most important)
- **ALEPH (LEP) and later OPAL Tau-particle-identification with a "Likelihood" classifier (Naïve Bayesian)**

..... Particle identification was crucial for the understanding of τ decays in order to separate electrons, muons and hadrons. At the beginning, most people were using cuts, but a likelihood method TAUPID was soon developed by Zhiqing Zhang and Michel which proved so superior that everyone adopted the method........



Gigi Rolandi

A Short History Of MVA

- … and of course elsewhere..

November 1992

LU TP 92-23

Neural Networks in High Energy Physics

Carsten Peterson[1]

- .. Although people have always felt that the advance is somewhat slow…

High Energy Physics

The progress of exploiting ANN in high enregy physics has been somewhat slow. Partly this conservatism is due to the a misconception that ANN approaches contain an element of "black box magic" as compared to conventional approaches. I hope I have convinced the reader that this is not the case. Statistical interpretation of the answers makes the ANN approach as well-defined to use as the discriminant ones.

A Short History Of MVA

… and MVA usage in ‘particle searches’ was ‘taboo’

Until LEP2 Higgs search set on to break that, sometimes facing fierce resistance which were replied to like this:

“If you are out looking for a spouse and apply ‘hard cuts’, you’re also not going to get anywhere” ***(Wayne ? OPAL)***

NOTE: by the mid ’90s, ANNs were ‘already’ out of fashion in the machine learning community. Why didn’t we pick up on SVM or “Boosted Decision Trees (1996) ??

Sophisticated Multivariate Techniques Pay!

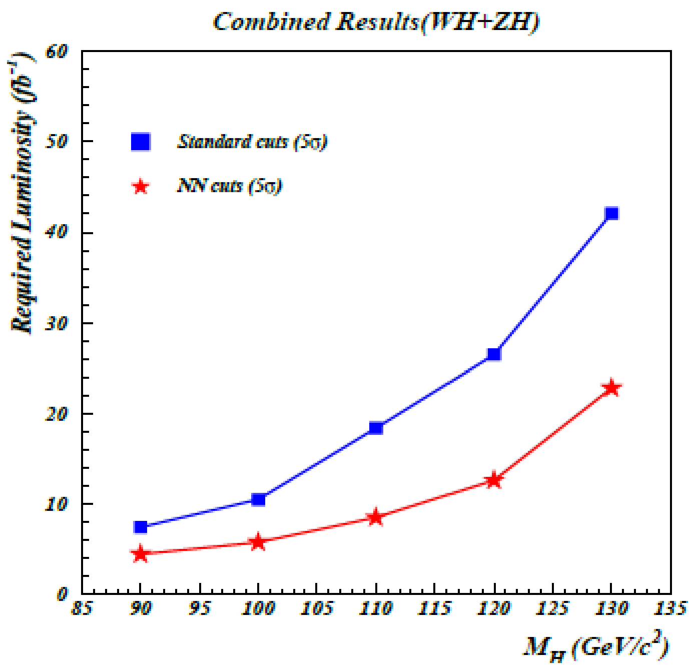
The searches for the Standard Model Higgs boson carried out by the four LEP experiments extended the sensitive range well beyond that anticipated at the beginning of the LEP programme [26]. This is due to the higher energy achieved and to more sophisticated detectors and analysis techniques

The LEP Working Group for Higgs Boson Searches / Physics Letters B 565 (2003) 61–75

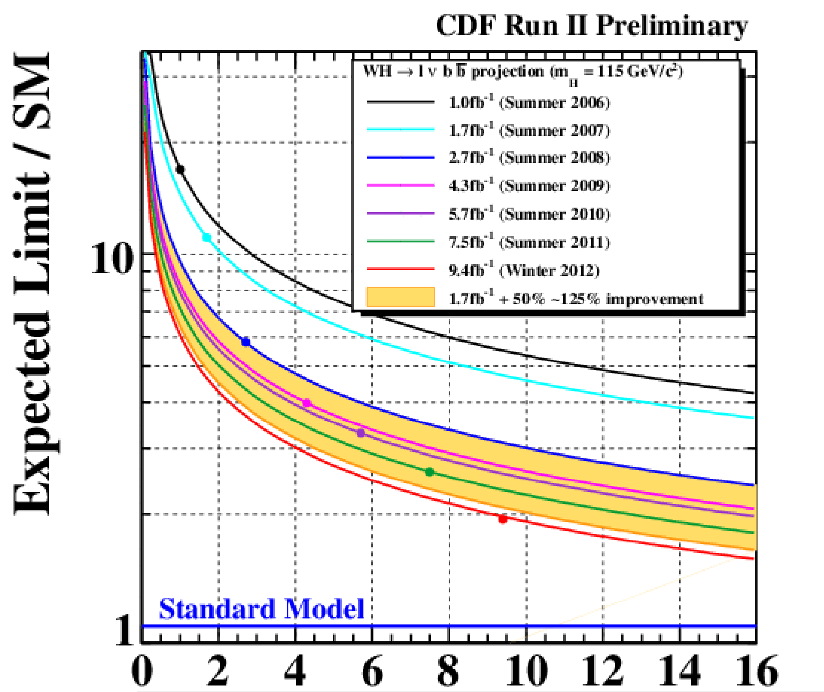
- **Well… sure, the 'more sophistication' is NOT ONLY multivariate techniques, but it sure plays its part of it**

Sophisticated Multivariate Techniques Pay!

- And obviously … the other side of the Atlantic was at least as active…



Run II Higgs study hep-ph/0010338 (Oct-2000)
P.C.Bhat, R.Gilmartin, H.Prosper, Phys.Rev.D.62 (2000) 074022



..and even ventured to Boosted Decision Trees!

Max-Planck-Institut für Kernphysik

Studies of Boosted Decision Trees for MiniBooNE Particle Identification

Hai-Jun Yang[a,c,1], Byron P. Roe[a], Ji Zhu[b]

[a] *Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[b] *Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

Abstract

Boosted decision trees are applied to particle identific
experiment operated at Fermi National Accelerator La
neutrino oscillations. Numerous attempts are made to tu
trees, to compare performance of various boosting algorit
variables for optimal performance.

→ **Made BDTs popular in HEP MiniBooNE, B.Roe et.a., NIM 543(2005)**

Decision Trees - 49 input variables

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)
- $\cos(\text{best1,lepton})_{\text{besttop}}$
- $\cos(\text{best1,notbest1})_{\text{besttop}}$
- $\cos(\text{tag1,alljets})_{\text{alljets}}$
- $\cos(\text{tag1,lepton})_{\text{btaggedtop}}$
- $\cos(\text{jet1,alljets})_{\text{alljets}}$
- $\cos(\text{jet1,lepton})_{\text{btaggedtop}}$
- $\cos(\text{jet2,alljets})_{\text{alljets}}$
- $\cos(\text{jet2,lepton})_{\text{btaggedtop}}$
- $\cos(\text{lepton}, Q(\text{lepton}) \times z)_{\text{besttop}}$
- $\cos(\text{lepton}_{\text{besttop}}, \text{besttop}_{\text{CMframe}})$
- $\cos(\text{lepton}_{\text{btaggedtop}}, \text{btaggedtop}_{\text{CMframe}})$
- $\cos(\text{notbest,alljets})_{\text{alljets}}$
- $\cos(\text{notbest,lepton})_{\text{besttop}}$
- $\cos(\text{untag1,alljets})_{\text{alljets}}$
- $\cos(\text{untag1,lepton})_{\text{btaggedtop}}$

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)
- $Q(\text{lepton}) \times \eta(\text{untag1})$
- $\sqrt{\hat{s}}$
- Sphericity(alljets,W)

- Adding variables does not degrade performance
- Tested shorter lists, lost some sensitivity
- Same list used for all channels

Yann Coadou (CERN) — Boosted decision trees: DØ single top evidence Workshop on Top Physics 20/10/07 15

D0 Single Top discovery

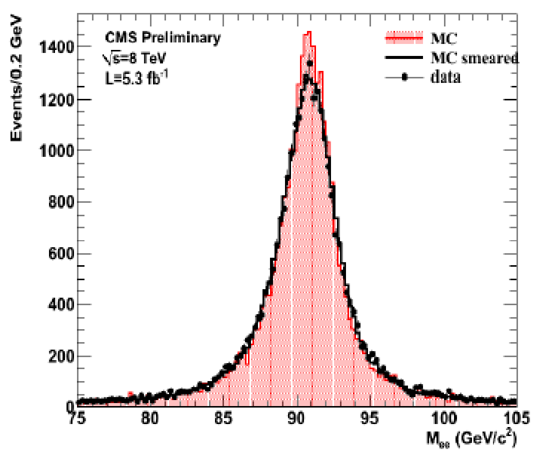
CMS Higgs Discovery
(such a nice example for MVA usage)

Max-Planck-Institut für Kernphysik

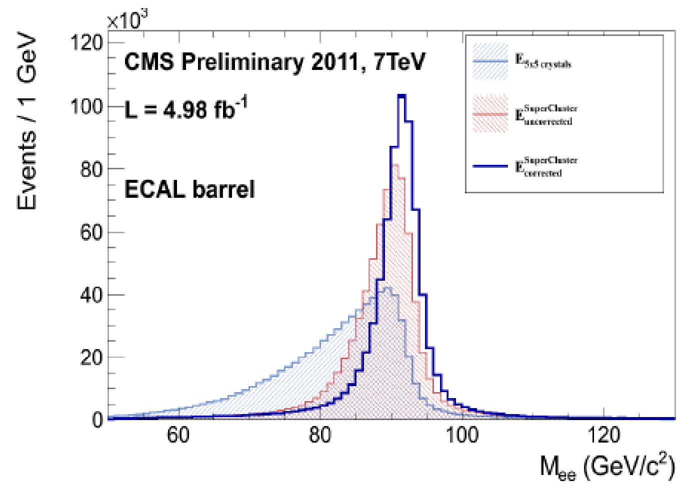
- MVA regression for energy calibration

Photon Energy Corrections, Scale and Resolution

- ECAL cluster energies corrected using a MC trained multivariate regression
  - Improves resolution and restores flat response of energy scale versus pileup
  - Inputs: Raw cluster energies and positions, lateral and longitudinal shower shape variables, local shower positions w.r.t. crystal geometry, pileup estimators
- Regression also used to provide a per photon energy resolution estimate
- **Energy Scale and resolution:** use $Z \to e^+e^-$



Non converted photons in the barrel |η|<1

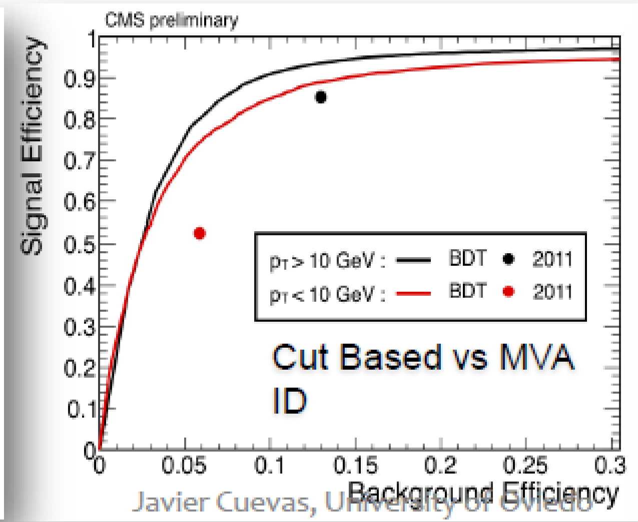


Effect of the regression on the $Z \to e^+e^-$ peak

PPC 2012 - KIAS - Nov 5th Javier Cuevas, University of Oviedo 14

CMS Higgs Discovery
(such a nice example for MVA usage)

- Multivariate electron identification in 2012
  - ECAL, tracker, ECAL-tracker-HCAL matching and impact parameter (IP) observables



$H \rightarrow \gamma\gamma$

- **Analysis selection (MultiVariate Analysis MVA)**
  - Vertex ID
    - Input variables: $\Sigma p_T^{2\,(tracks)}$, p_T balance wrt $\gamma\gamma$, conversions information
  - ID photons $p_{T1} > m_{\gamma\gamma} / 3$ $p_{T2} > m_{\gamma\gamma} / 4$
- **MVA Diphoton discriminant categories**
  - High score
    - signal-like events
    - good $m_{\gamma\gamma}$ resolution
  - Designed to be $m_{\gamma\gamma}$ independent
  - Trained on signal and background MC
  - Input variables:
    - Kinematic variables: $p_{T\gamma} / m_{\gamma\gamma}$, η_γ, $\cos(\varphi_1 - \varphi_2)$
    - Photon ID MVA output for each photon
    - Per-event mass resolutions for the correct and incorrect choice of vertex

PPC 2012 - KIAS - Nov 5th Javier Cuevas, University of Oviedo

LHCb $B_s \rightarrow \mu\mu$

CERN-PH-EP-2013-128
LHCb-PAPER-2013-046
July 18, 2013

Measurement of the $B^0_s \to \mu^+\mu^-$ branching fraction and search for $B^0 \to \mu^+\mu^-$ decays at the LHCb experiment

The analysis strategy is very similar to that employed in Ref. [12], with a different multivariate operator based on a boosted decision trees algorithm (BDT) [15, 16]. After trigger and loose selection requirements, $B^0_{(s)} \to \mu^+\mu^-$ candidates are classified according to dimuon invariant mass and BDT output.

MVA Techniques are Phantastic

- And can be successfully employed in numerous places
  - Almost everything we 'look at' or 'study' is depending on multiple variables ☺
- Detector signal reconstruction (i.e. cluster energy in calorimeter)
- Particle reconstruction/classification
- Event classification
- Automatic fault detection
  - At the machine?
  - At the detector – online histograms
  - …
- …

Issues to be be concerned about

- Which variables to choose
- Which classifier – modelling flexibility
- Test the generalisation properties of the fitted model
    - Issues with limited 'training/testing/validation samples sizes'

- And of course – the never ending story of - Systematic uncertatinties

MVA in the presence of systematic uncertainties

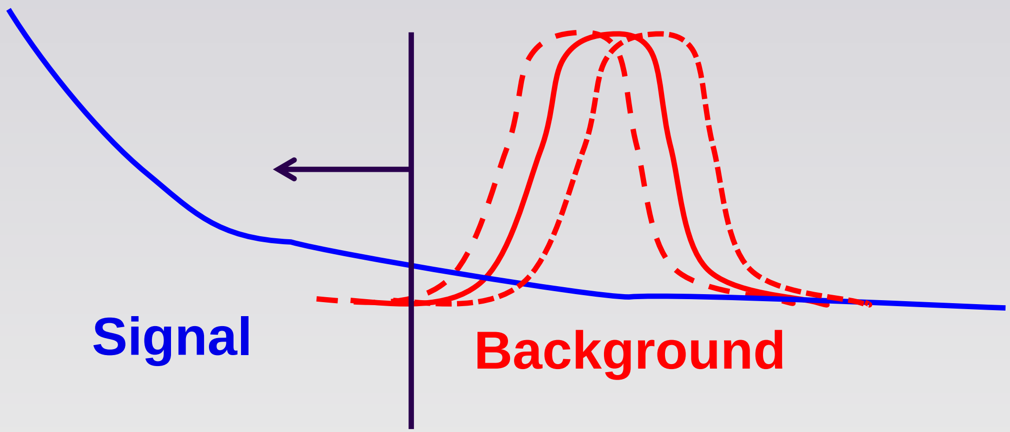
- minimize "systematic" uncertainties (robustness)

→"classical cuts" : do not cut near steep edges, or in regions of large sys. uncertainty

→ hard to translate to MVAs:

  → artificially degrade discriminative power (shifting/smearing) of systematically "uncertain" observables IN THE TRAINING

    → remove/smooth the 'edges' → MVA does not try to exploit them



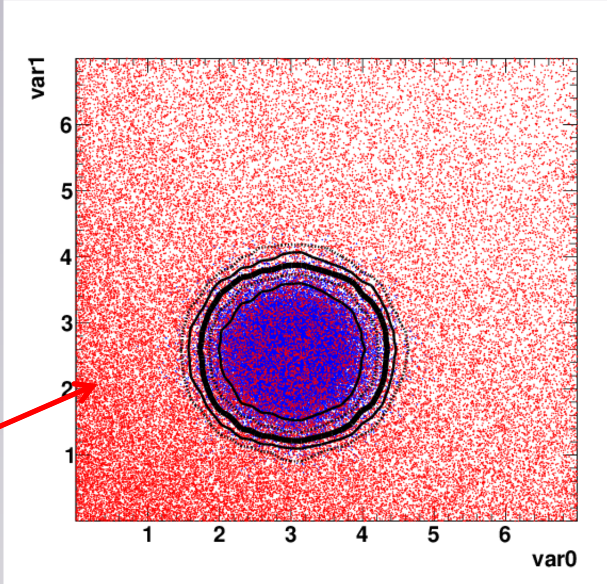
MVA in the presence of systematic uncertainties

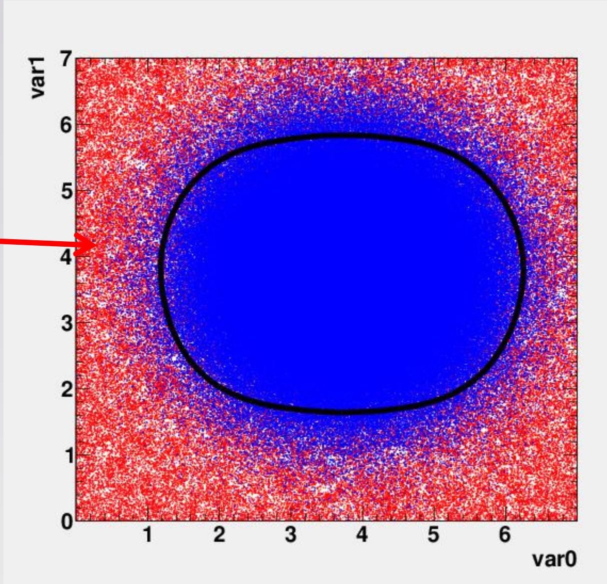
MVA-decision boundaries

- **Looser MVA-cut → wider boundaries in BOTH variables**

- **You actually want a boundary like THIS**
  - **Tight boundaries in var1**
  - **Loose boundaries in var0**

    - **YES it works !**

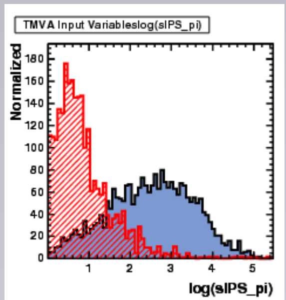


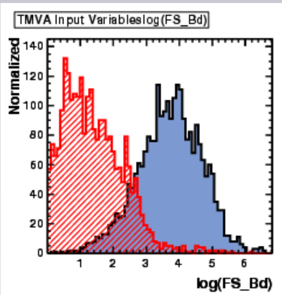


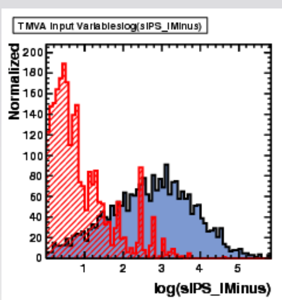
- **Sure, it is of course similar to 'mixing different Monte Carlos' as proposed earlier by others…**

What are MVAs ?

Multivariate Event Classification







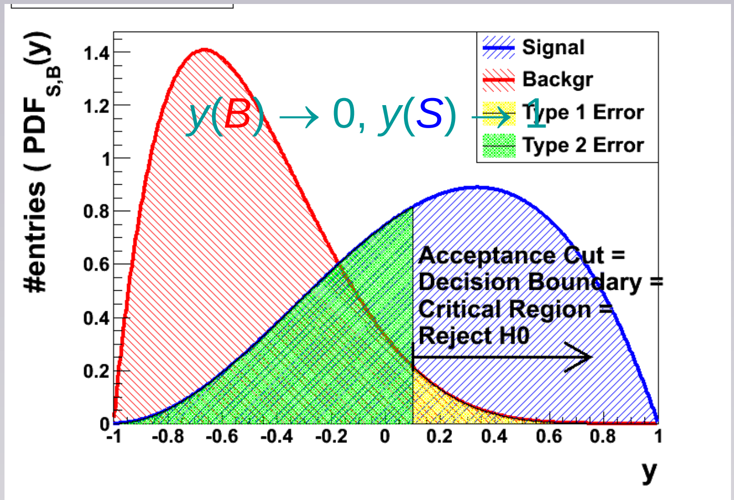
$\mathbb{P}^D$

"feature space"

- Each event, if Signal or Background, has "D" measured variables.
- Find a mapping from D-dimensional input/observable/"feature" space to one dimensional output → class labels

y(x): $R^n \rightarrow R$:

$\mathbb{P}^D \longrightarrow \mathbb{P}$



- y(x): "test statistic" in D-dimensional space of input variables
- distributions of y(x): $PDF_S(y)$ and $PDF_B(y)$
- used to set the selection cut!

→ efficiency and purity

y(x):
- > cut: signal
- = cut: decision boundary
- < cut: background

- y(x)=const: surface defining the decision boundary.
- overlap of $PDF_S(y)$ and $PDF_B(y)$ → separation power , purity

What is y(x) ??

- A neural network
- A decision tree
- A boosted decision tree forest
- A multidimensional/projective likelihood
- …

→ The same stuff out of which the 'dreams' of AI are born – or better.. which comes from AI or general 'pattern recognition' research

→ Stuff that powers nowadays 'BIG DATA business', search engines and social network studies for targeted advertisment … ☹

→ Stuff that is extremely fast evolving !!

Where do we go from here ?

- We like to think in HEP that we are "state of the art"
  - In MVA techniques, we've certainly always lagged behind
  - … and the gap is in my opinion growing rather than closing !

Deep Networks

Max-Planck-Institut für Kernphysik

Deep Learning in High-Energy Physics: Improving the Search for Exotic Particles

P. Baldi,[1] P. Sadowski,[1] and D. Whiteson[2]

[1]*Dept. of Computer Science, UC Irvine, Irvine, CA 92617*
[2]*Dept. of Physics and Astronomy, UC Irvine, Irvine, CA 92617*

19 Feb 2014

Collisions at high-energy particle colliders are a traditionally fruitful source of exotic particle discoveries. Finding these rare exotic particles requires solving difficult signal-versus-background classification problems, hence machine learning approaches are often used for this task. Standard approaches in the past have relied on 'shallow' machine learning models that have a limited capacity to learn complex non-linear functions of the inputs, and rely on a pain-staking search through manually constructed non-linear inputs. Progress on this problem has slowed, as a variety of techniques (neural networks, boosted decision trees, support vector machines) have shown equivalent performance. Recent advances in the field of deep learning, particularly with artificial neural networks, make it possible to learn more complex functions and better discriminate between signal and background classes. Using benchmark datasets, we show that deep learning methods need no manually constructed inputs and yet improve the AUC (Area Under the ROC Curve) classification metric by as much as 8% over the best current approaches. This is a large relative improvement and demonstrates that deep learning approaches can improve the power of collider searches for exotic particles.

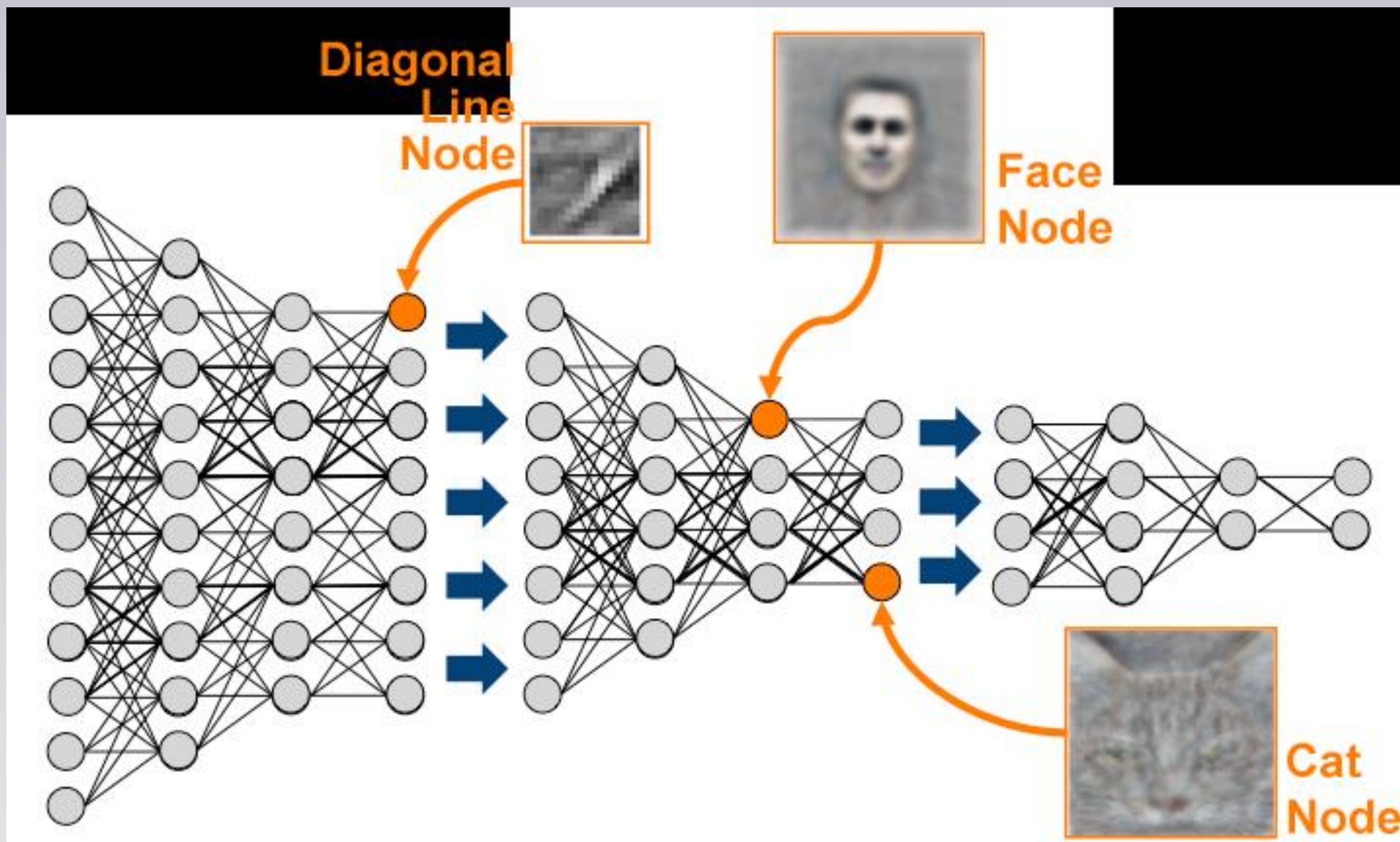
The field of *high energy physics* is devoted to the study of the elementary constituents of matter. By investigating the structure of matter and the laws that govern its

used in high-energy physics fail to capture all of the available information, even when boosted by manually-constructed physics-inspired features. This effectively re-

- **Yes… but look at the date ? 2014 !**
- **Deep networks became 'mainstream' after 2006 when "Google learned to find cats"**
  - **It has since revolutionised the fields of "speech and image recognition"**

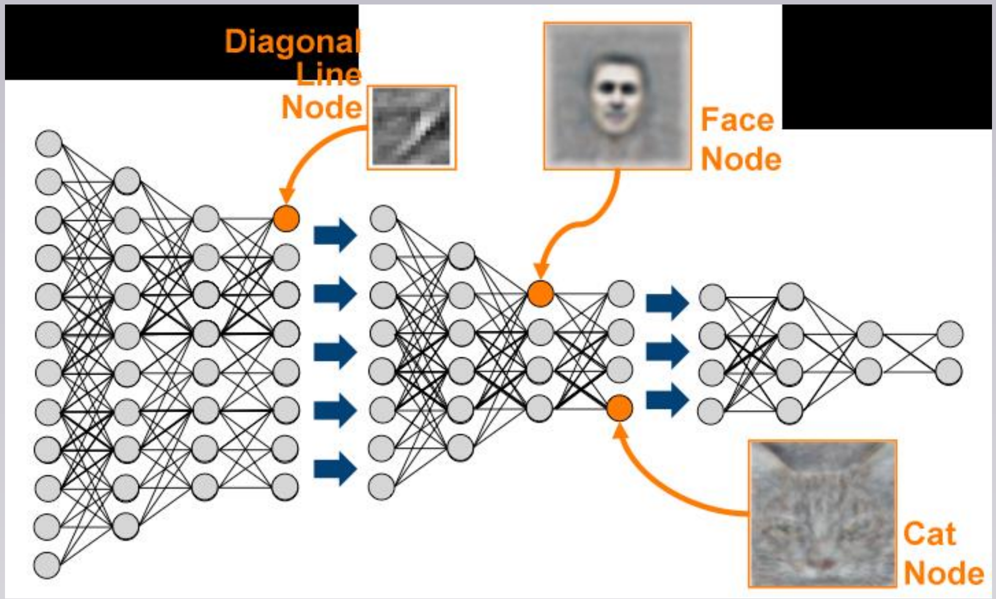
2006 : GOOGLE finds the Cat

Max-Planck-Institut
für Kernphysik



What is “DEEP LEARNING” ?

- A fairly ‘standard’ neural network architecture with MANY hidden layers
  - Conventional training (backpropagation) proves ineffective



- Pre-training individual layers as “auto-encoders” or “Restricted Boltzman Machines” (networks that are trained unsupervised and can ‘learn’ a probability density)

2006 ?

Max-Planck-Institut für Kernphysik

- 2006 .. When the world embraced “DEEP learning”
  - We celebrated MiniBooNE’s first venture into BDTs
  - TMVA is shipped as a ROOT “add on package”
    - And the majority of HEP physicists <u>started</u> to like MVAs
    - Easy accessibility of TMVA via ROOT and its ease of use are key ingredients to its success over ‘competitors’ like StatPatternRecognition or other non-HEP packages

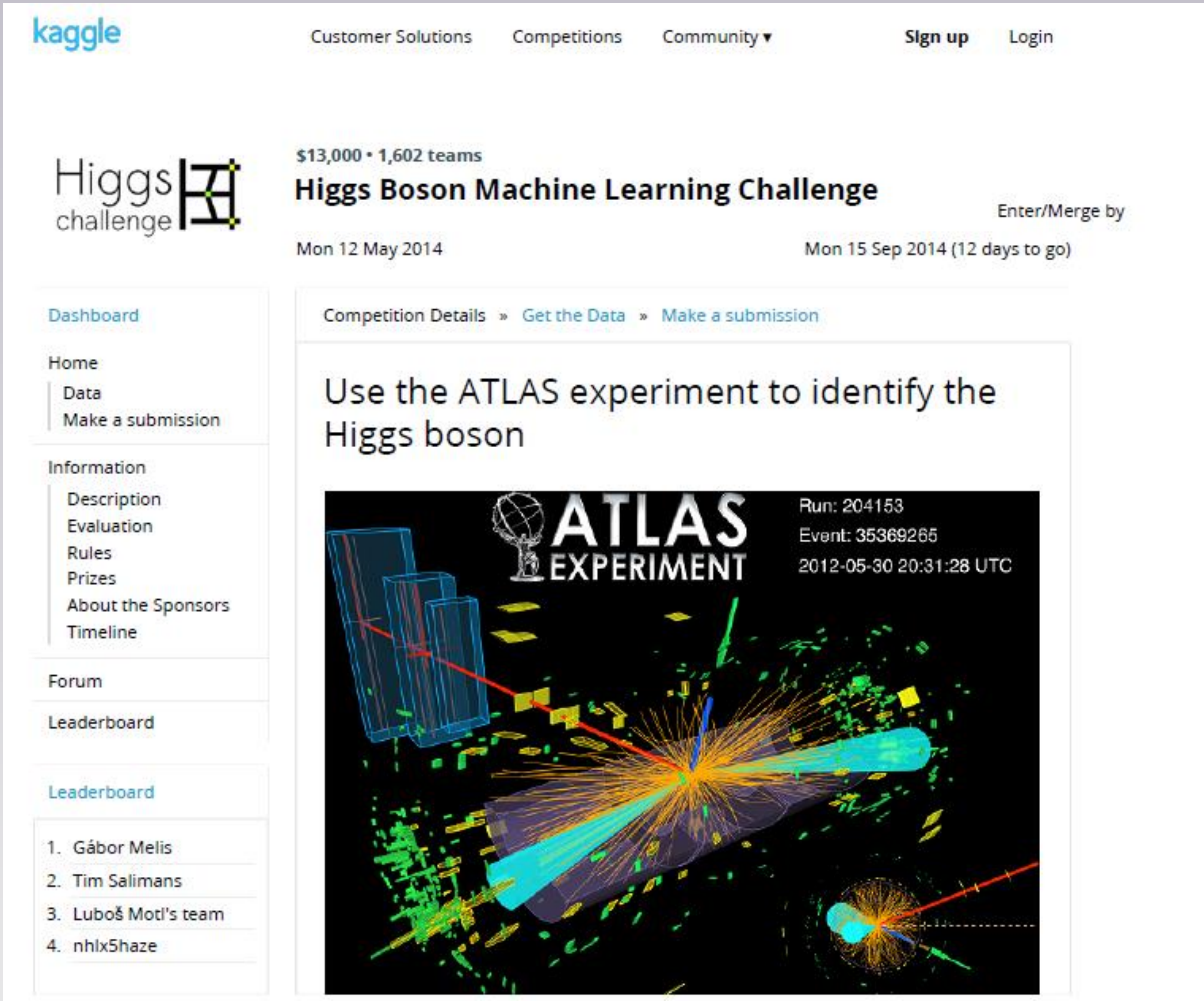
2006 ?

- Endless discussions about possible 'problems' with MVA (systematic etc) show:
- We (physicists) are smart and try to really understand what we are doing
- But it also shows:
  - Sophisticated techniques are of course more difficult to grasp than 'cuts'
  - To tap into the extremely valuable resource of pattern recognition techniques, the 'physicist way' of everyone re-inventing the wheel does not seem promising to me here.

Where do we go from Here

Max-Planck-Institut für Kernphysik

→ **But is THIS the answer ??**



Summary

- MVA's are great
- MVA's are widely used in HEP
- MVA's are even "widelier" used outside HEP
- MVA's are complicated to understand and to code !
- MVA's and work thereon still is not 'funded' buy HEP like "Detector/Accelerator development" for example is:

- note: before TMVA in ROOT, the majority of the HEP community only used/knew simple cuts which often perform much worse
  - significant improvement in physics reach (imagine how much a 20% better accelerator/detectors would cost?)
  - provide state of the art analysis tools for state of the art accelerator/detectors

→ And I think we should have a much larger concentrated effort to put HEP to 'state of the art' in pattern recognition, then this one paid TMVA position I was unsuccessfully asking for!

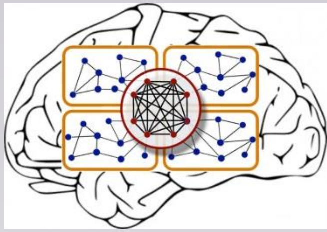
Backup

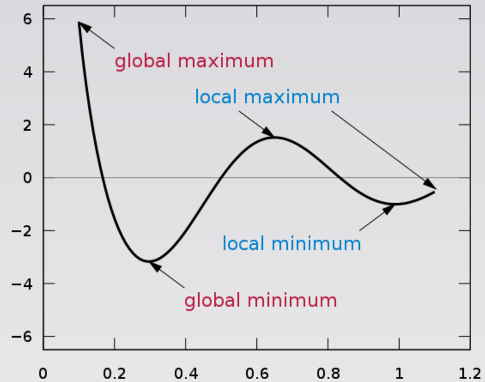
Neural Network "history"

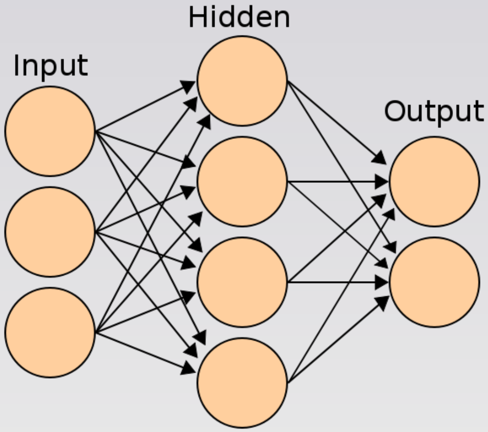
- **Developed to 'simulate' the working of the brain (McCulloch 1943)**
  - **somewhat but not terribly successful until:**
- **Backpropagation was invented (1974 – reinvented 1985)** (use 'chain rule' to calculate gradients of loss function ($\frac{\partial L(y(w_{ij}), y_{true})}{\partial w_{ij}}$) and adjust weights towards smaller L)
  - **but: "many layers" still didn't prove very helpful (**despite that fact that our brain has quite a few more than 2)

    → **vanishing gradient problem**

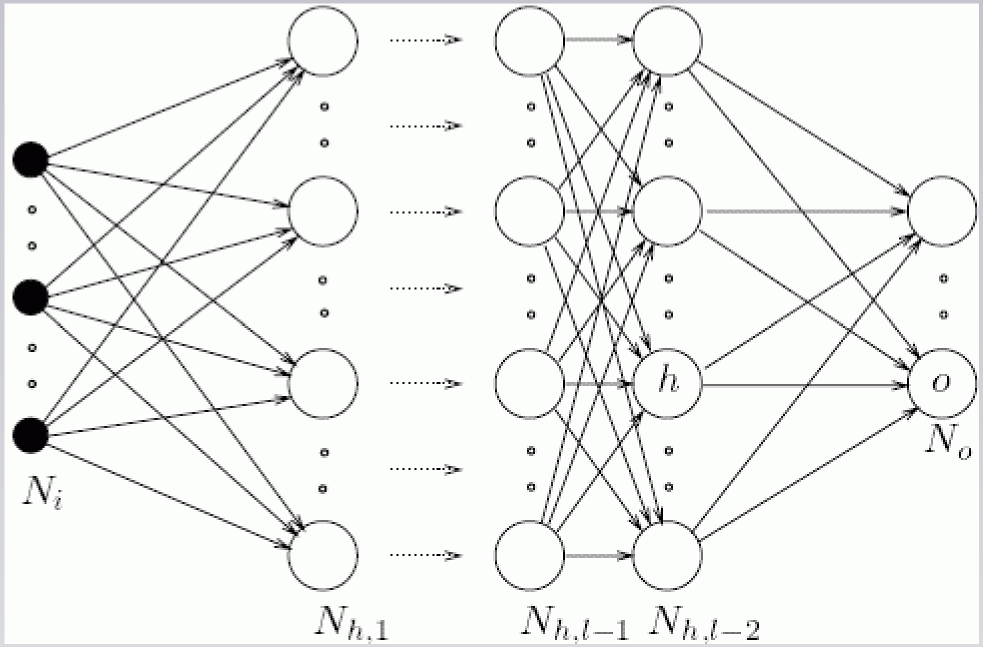
  → **and it finds 'local minima only'**







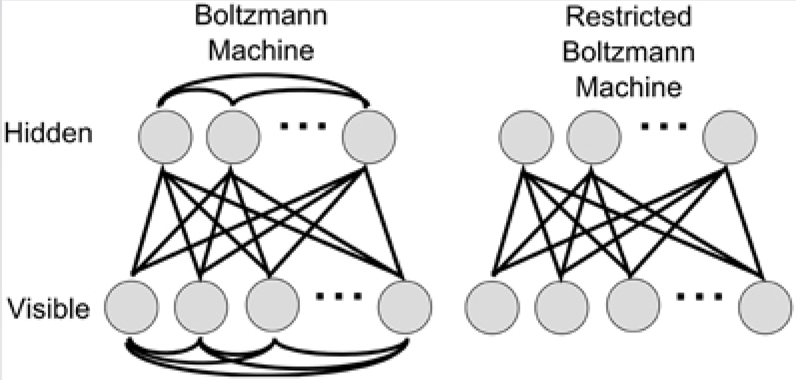
Deep Networks == Networks with many hidden layers



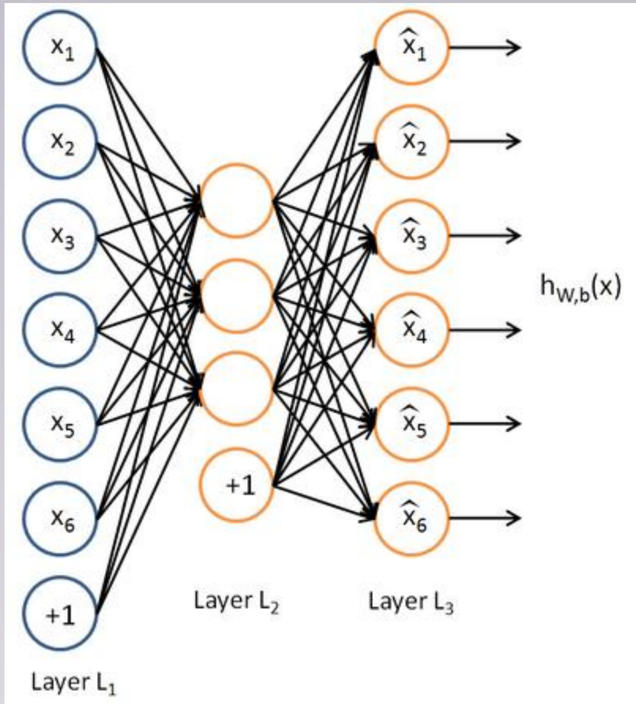
- **That's apparently "all" it means … although 'deep' somehow in statistical terms would mean apparently:**

Training deep networks

- The new trick is: pre-training + final backpropagation to “fine-tune”
  - → initialize the weights not ‘random’ but ‘sensibly” by
  - → ‘unsupervised training of’ each individual layer, one at the time, as an:
    - : auto-encoder (definite patterns)
    - : restricted-Boltzmann-machine (probabilistic patterns)



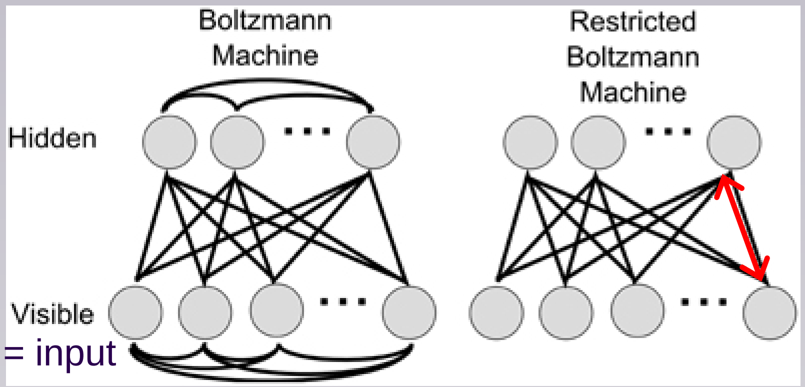
Auto-Encoder



- network that 'reproduces' its input
- hidden layer < input layer

→ hidden layer 'dimensionality reduction' needs to 'focus/learn' the important features that make up the input

Restricted Boltzmann Machine

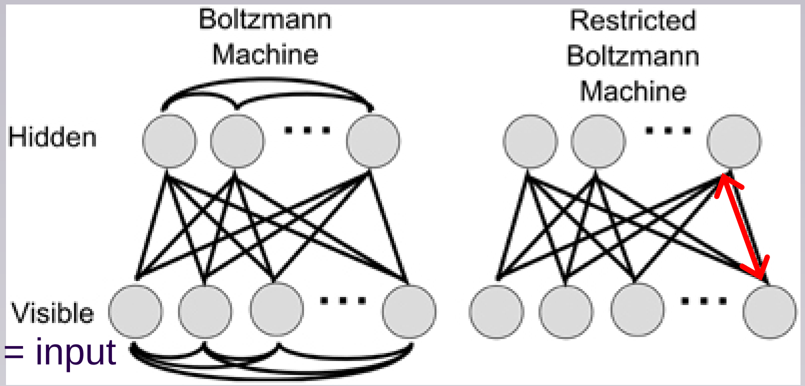


A network with 'symmetric' weights

i.e. not 'feed-forward'

- **if you 'train' it (i.e. determine the weights) it can 'learn' a probability distribution of the input (training events)**

Restricted Boltzmann Machine



A network with ‘symmetric’ weights

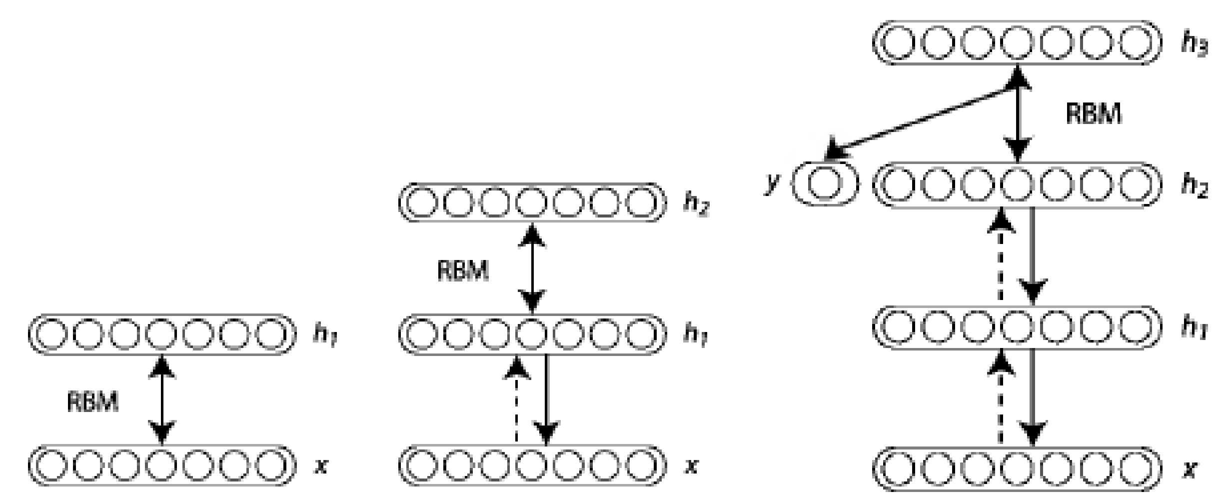
i.e. not ‘feed-forward’

- **if you ‘train’ it (i.e. determine the weights) it can ‘learn’ a probability distribution of the input (training events)**
- **… aeeh .. what the hell does THAT mean ??**
  - each network configuration (state) is ‘associated’ with an ‘energy’
  - the various states are populated according to the probability density given in the training data (given a particular energy I guess)

(hmm… given that I understood it ‘correctly’)

Deep Network training

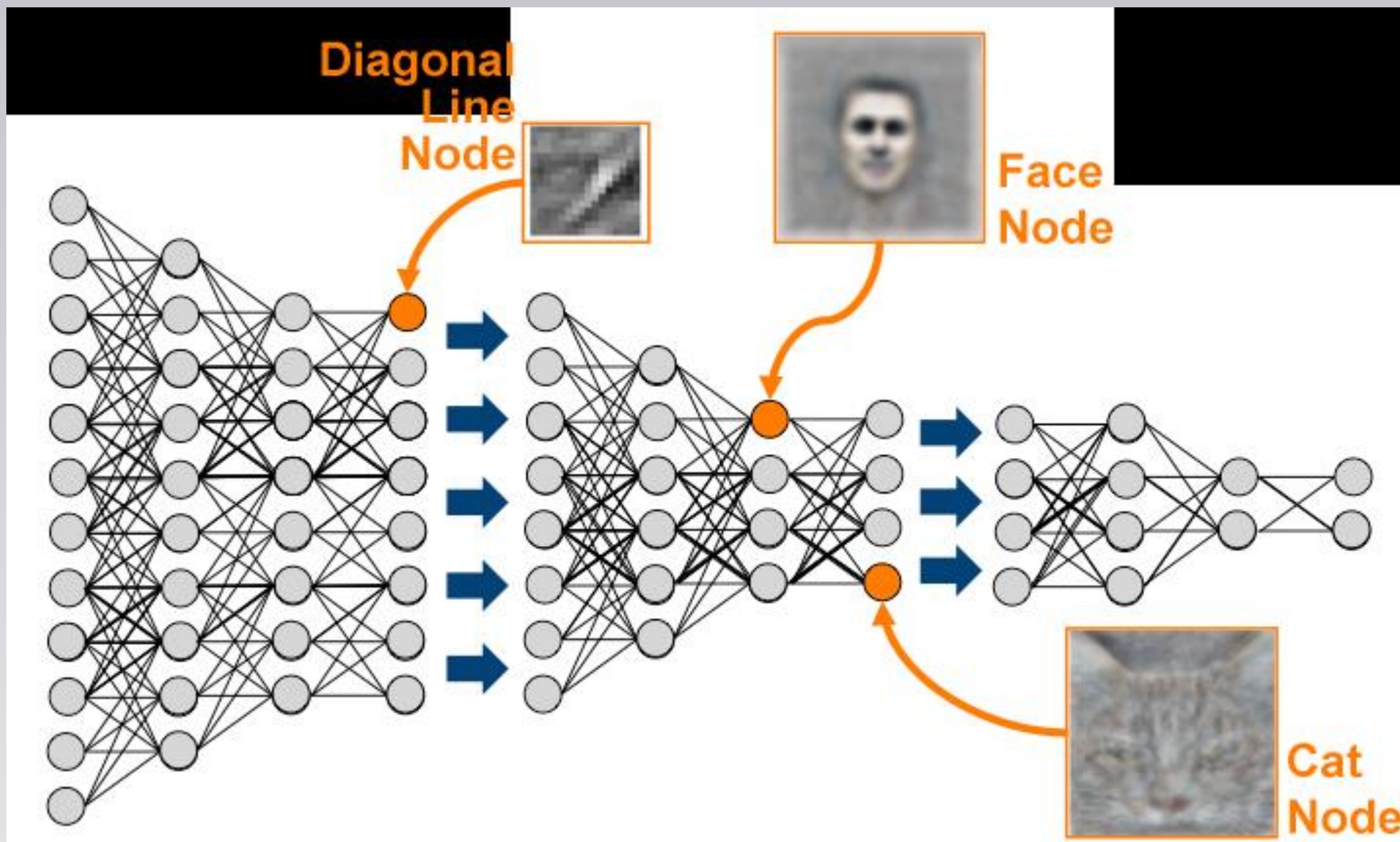
The "output" of the first layer (hidden layer of the first RBM trained) is used as input for training the second RBM etc..)



(a) Train RBM for $\mathbf{x}$ (b) Train RBM for $\mathbf{h}^1$ (c) Train RBM for $\mathbf{h}^2$ and y

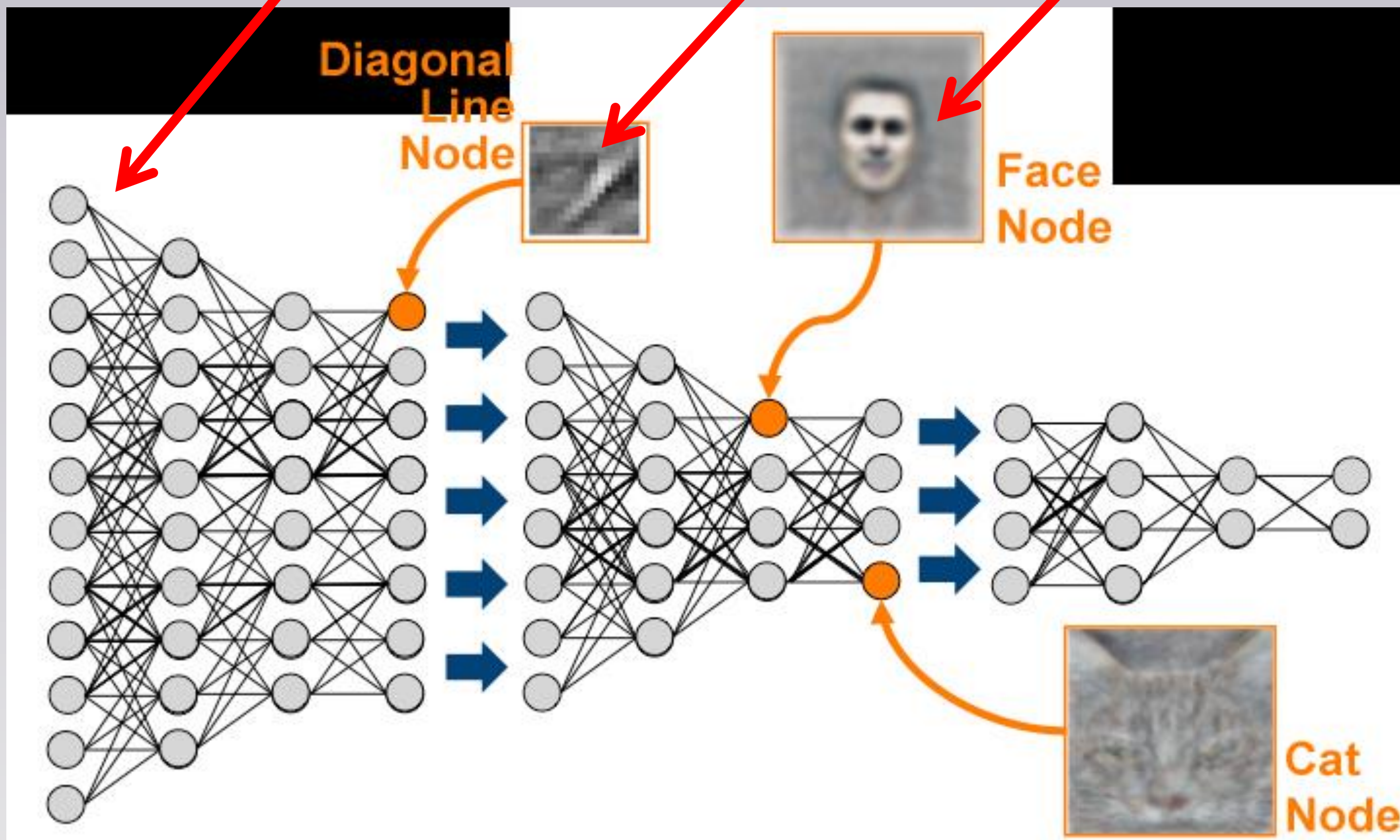
What does it do?

Max-Planck-Institut für Kernphysik



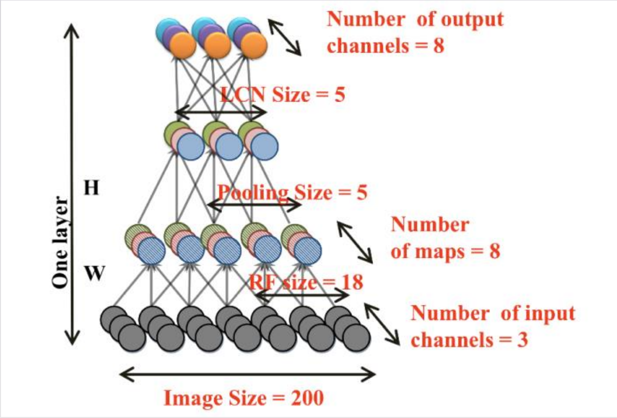
What does it do?

Could this be: 4-vectors invariant masses decays?



Something else we missed?

- **Plenty!** (somewhat embarrassingly but: as I said many times, our 'pattern recognition' tasks in HEP are very simple compared to "Artificial Intelligence" or even just "image recognition")
  - **Dropout ! (a new regularizer of weights against overfitting etc. → Bagging done in implicitly in ONE signle neural network)**
  - **'strange' activation functions → digital rather than analog output**
  - **what are 'convolutional networks' ?**
  - **what about more 'complicated' structures like, built out of many building blocks like this:**



http://www.deeplearning.net/tutorial/
http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/ReadingOnDeepNetworks