HIGGS BOSON CHALLENGE

FANTASTIC FOUR

- AKHILA IYENGAR
- LEONARDO CITRARO
- ARTHUR PERODOU
- BAUDOUIN ROULLIER

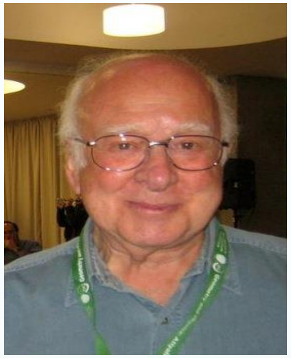
LIST OF CONTENTS

- Introduction
- Data Description
- Evaluation
- Support Vector Machines
- Neural Networks
- Deep Neural Networks
- Results

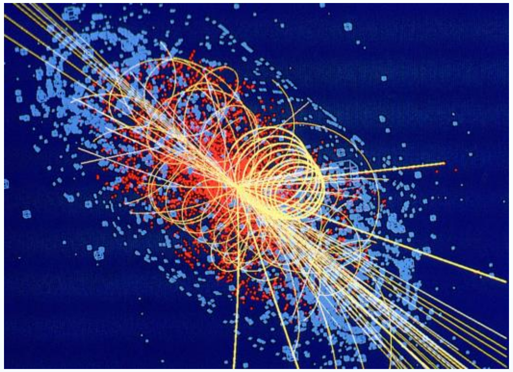
Introduction

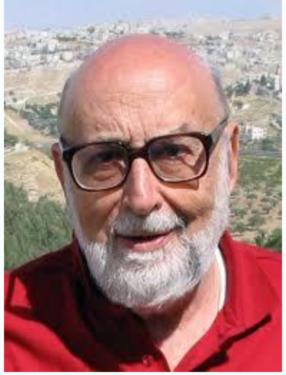
- Higgs boson is an elementary particle in the standard model of particle physics.

- Peter Higgs and Francois Englert discovered Higgs which was later confirmed by ATLAS and CMS at CERN



Peter Higgs

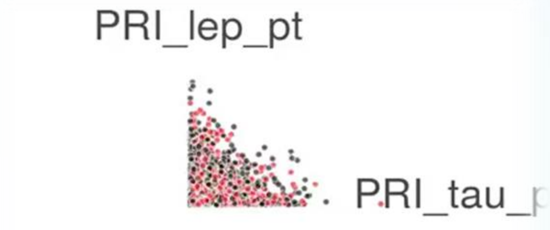




Francois Englert

- Small signal buried in huge background noise

- Goal is to classify the events into respective regions (Signal and Background)



Data Description

- Dataset comprises of simulated events of which **250,000 are training** set and **550,000 are test** set
- Training data has several attributes like event ID along with **30 features**, labels and weights while test data has event ID and 30 feature columns
- Data has signal (Higgs Boson) and background with associated **weights**

Binary classification problem

Training dataset
30 features
250’000 samples + weights

Test dataset
30 features
550’000 samples + weights

Evaluation

- Performance evaluation of predictive model is done by AMS (Approximate Median Significance).
- This makes events with higher weights more significant.
- Let s and b be the true and false positive rates respectively, the AMS is defined as:

$$AMS = \sqrt{2\left(\left(s + b + b_{reg}\right)\ln\left(1 + \frac{s}{b+b_{reg}}\right) - s\right)}$$

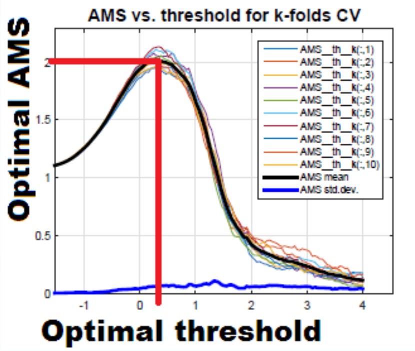
$$b_{reg} = 10$$

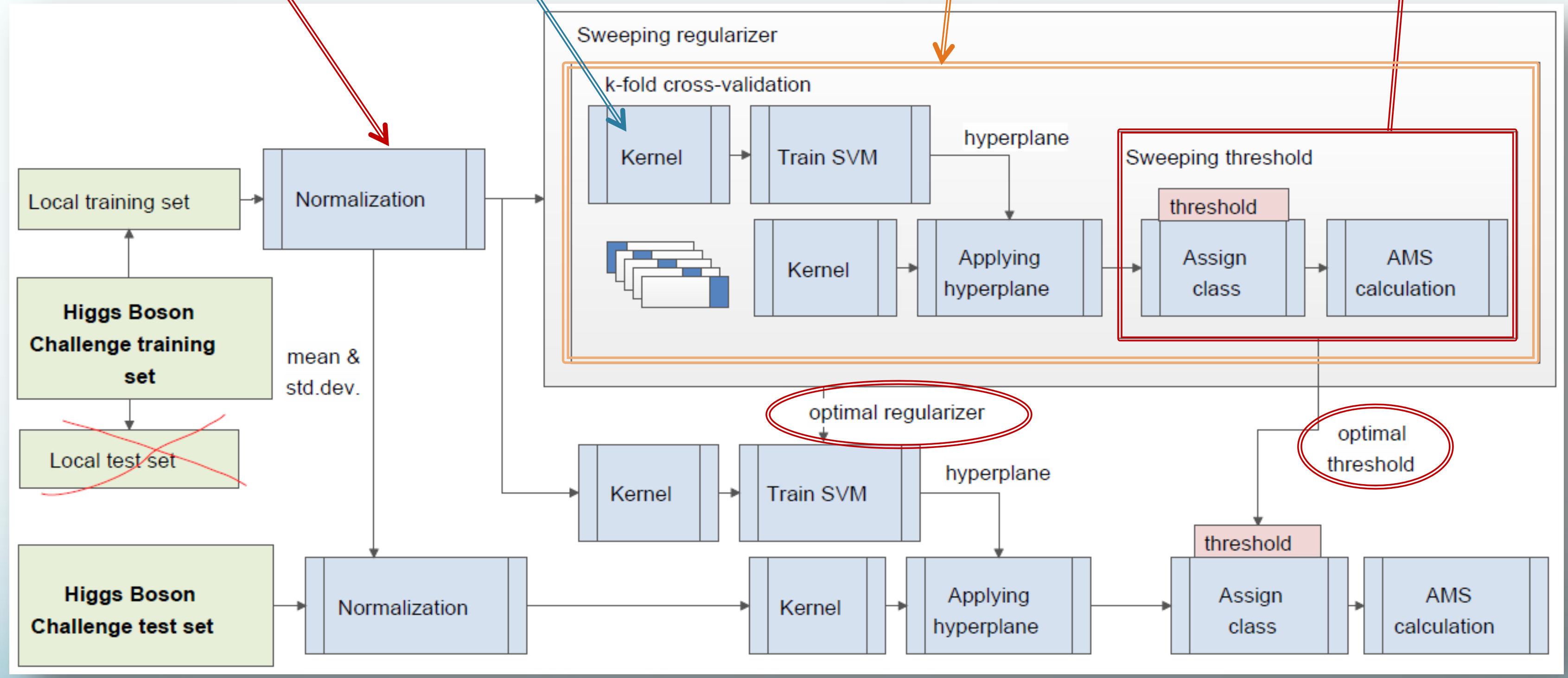
Support Vector Machine

Zero mean
unit std.

Homogeneous kernel map
(linear approximation of
non-linear mappings)
[VLFeat toolbox]

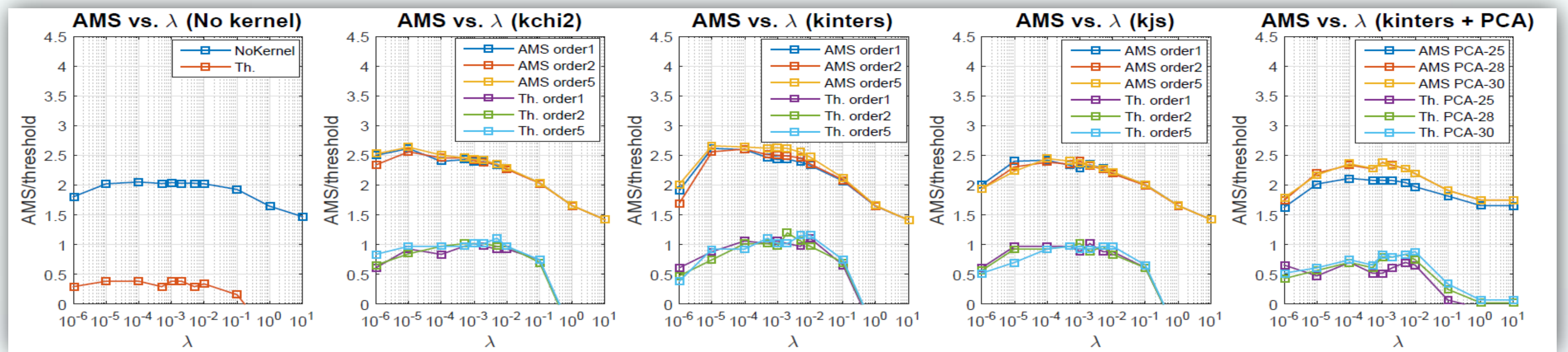
4-folds cross-validation





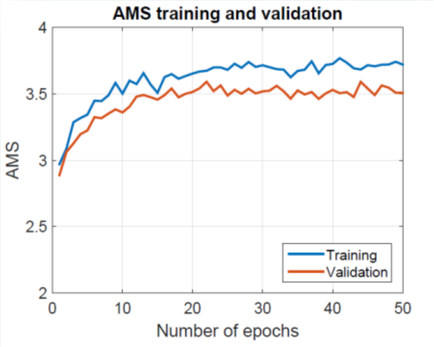
Results Support Vector Machine

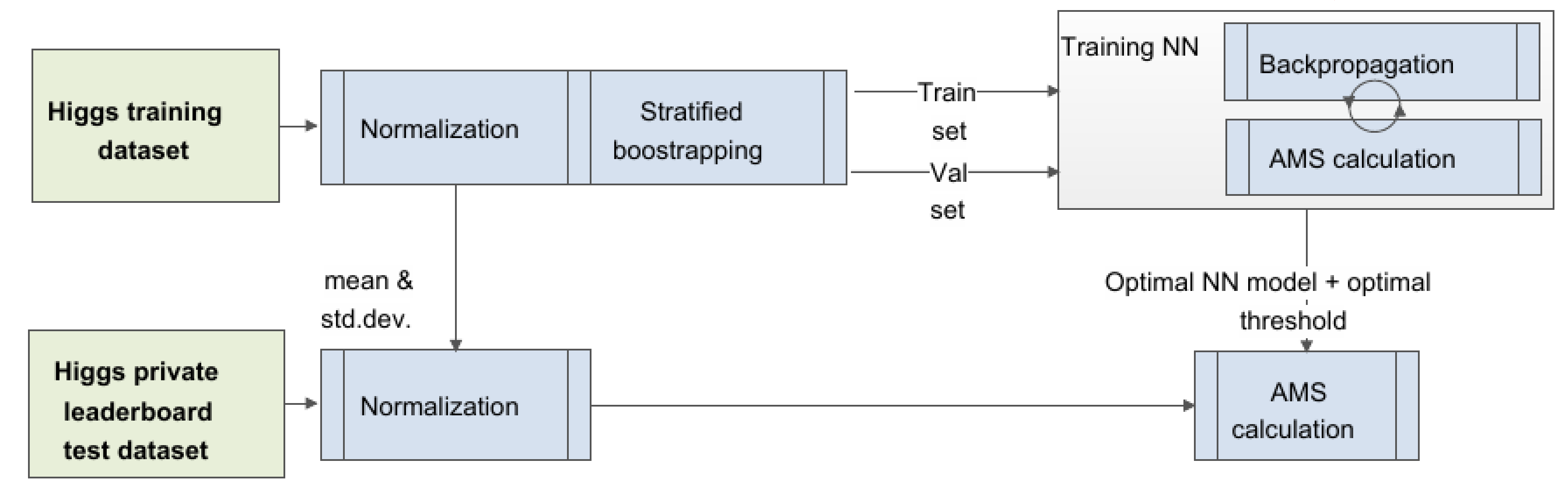
- Maximum AMS vs. Regularizer λ



- Maximum AMS scored in cross-validation: **2.66 (kernel intersect)**
- AMS scored using the test set: **2.44**
- Result far from the best AMS probably due to the linear kernel.
- For big data the SVM is slow and non-linear kernel require a huge amount of memory

Single Layer Neural Network





Higgs training dataset

Normalization

Stratified boostrapping

Train set

Val set

Training NN

Backpropagation

AMS calculation

mean & std.dev.

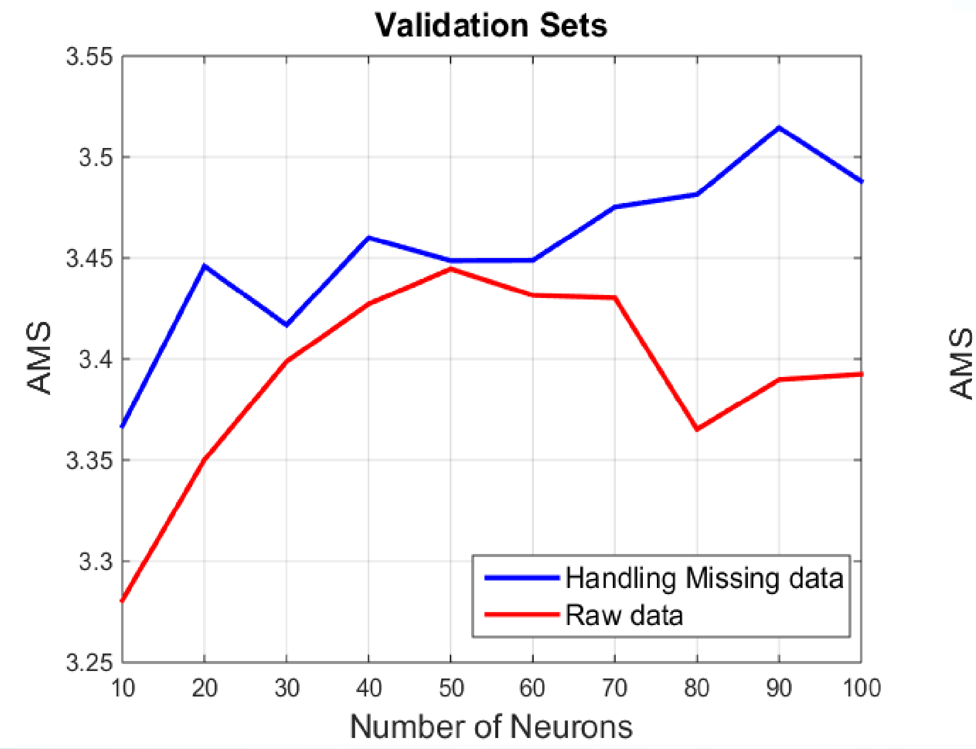
Optimal NN model + optimal threshold

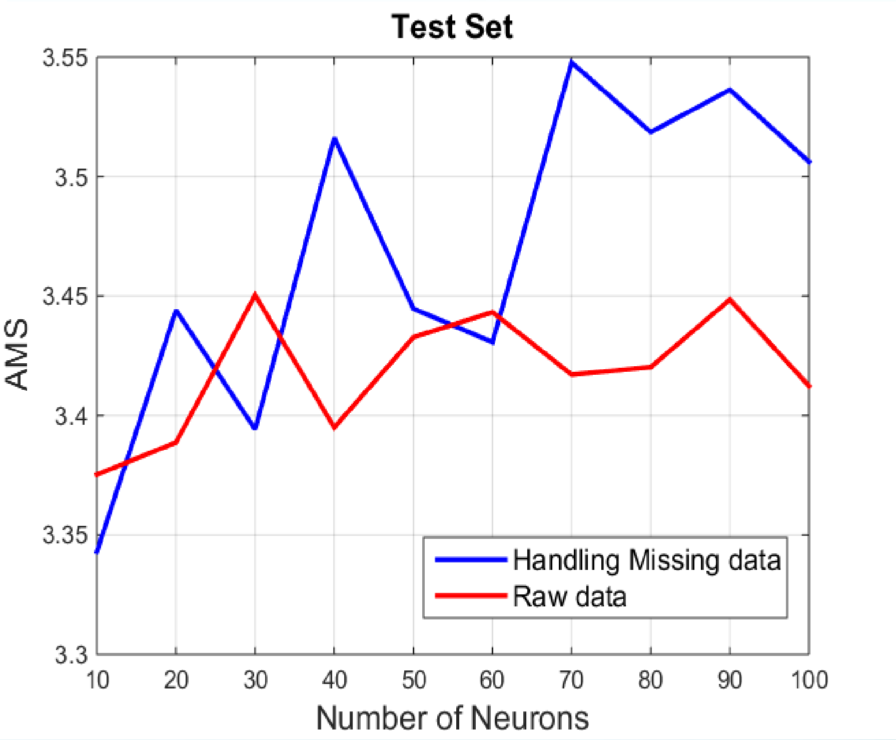
Higgs private leaderboard test dataset

Normalization

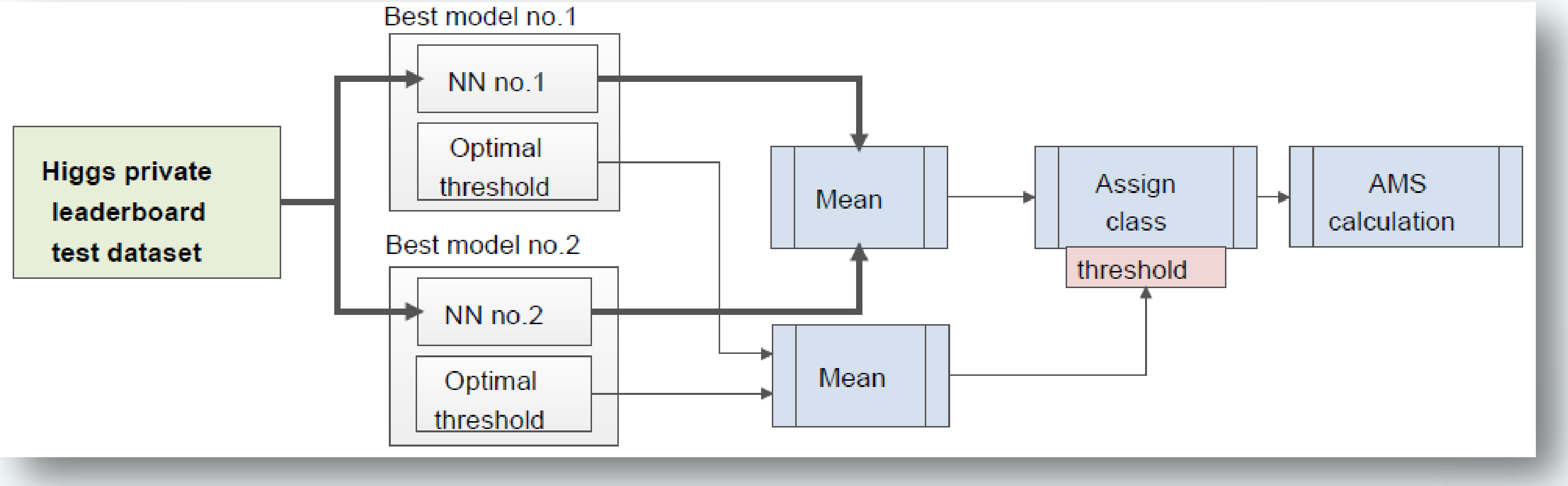
AMS calculation

Result Single Layer Neural Network + Missing Data





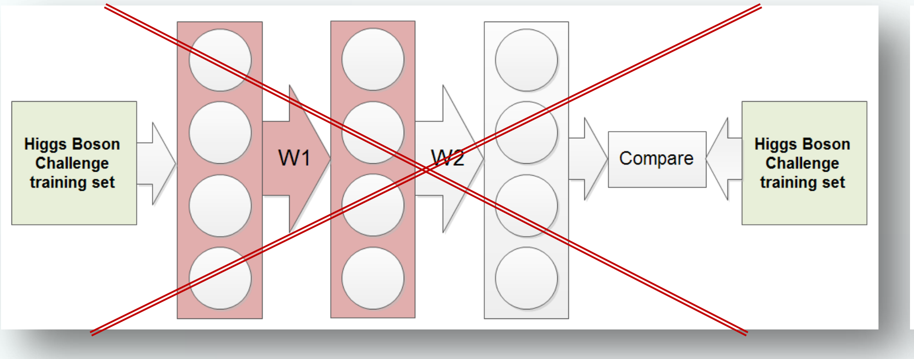
Averaging multiple Neural Networks



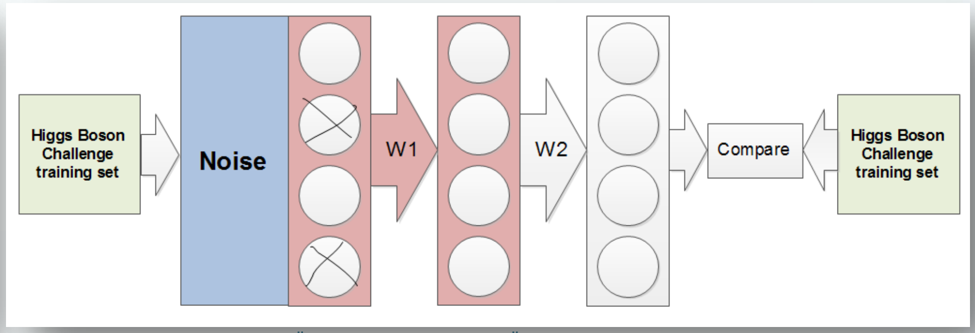
- Using k NN
- Averaging of their prediction
- Averaging of their optimal threshold

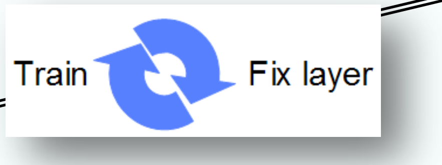
Deep Neural Network

Auto-encoder


Higgs Boson Challenge training set
W1
W2
Compare
Higgs Boson Challenge training set

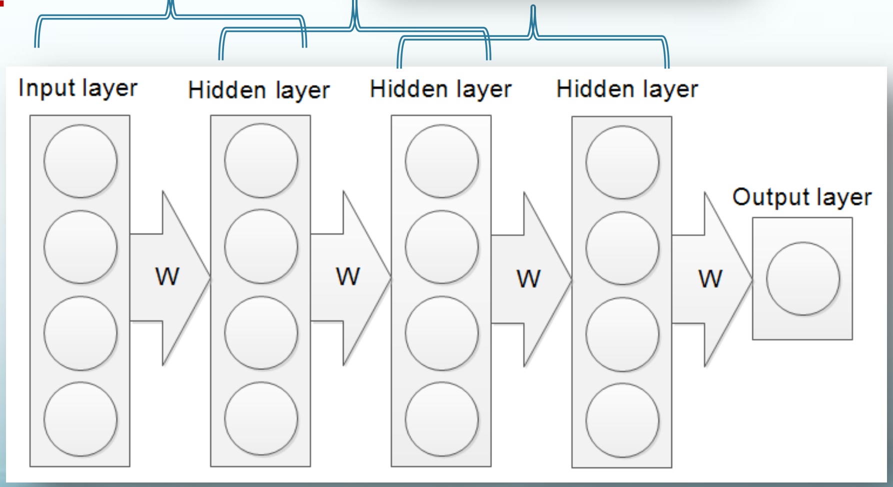
Denoising Auto-encoder


Higgs Boson Challenge training set
Noise
W1
W2
Compare
Higgs Boson Challenge training set



Thousand of ways to set W1 and W2 giving error=0

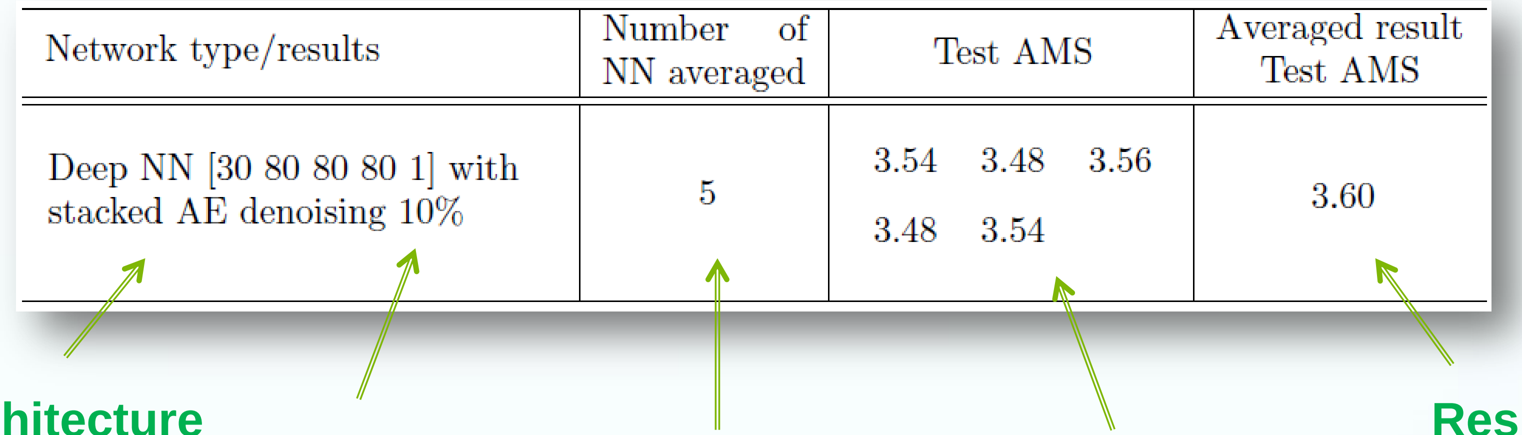
more meaningful mapping of the dataset



Greedy layer-wise unsupervised pre-trained process

← Deep network!

Multiple Averaged Deep Neural Network + Missing Data



Network type/results	Number of NN averaged	Test AMS	Averaged result Test AMS
Deep NN [30 80 80 80 1] with stacked AE denoising 10%	5	3.54 3.48 3.56 3.48 3.54	3.60

Architecture

Noise percentage Auto-Encoders

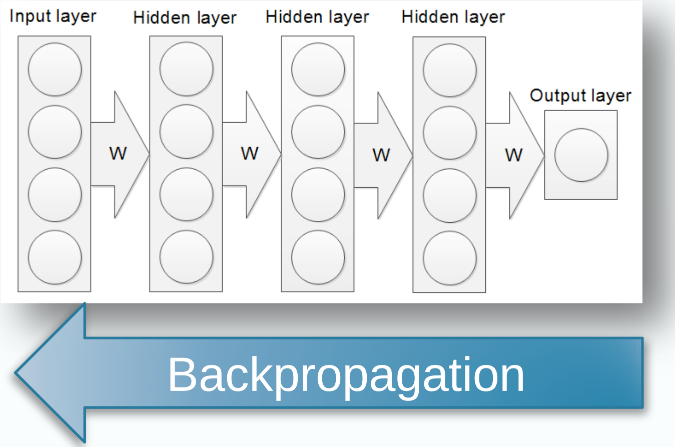
Number NN trained

Results of each DeepNN Mean:3.53

Result of averaging process

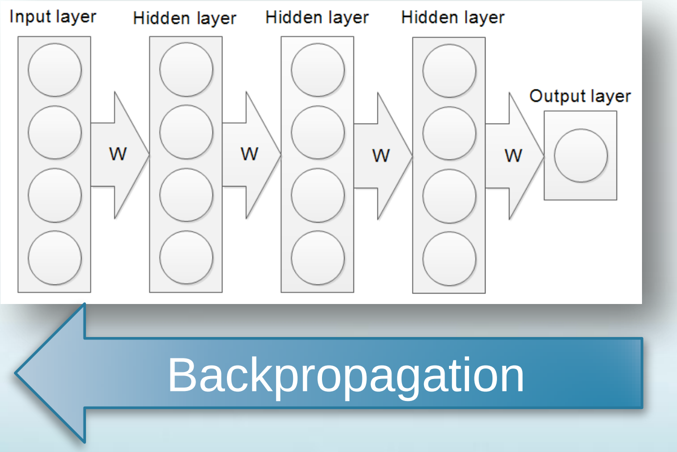
Error weights

Classic error function



$$E_n = \frac{1}{2}\sum_k (\widehat{y_{nk}} - t_{nk})^2$$

Weighted error function

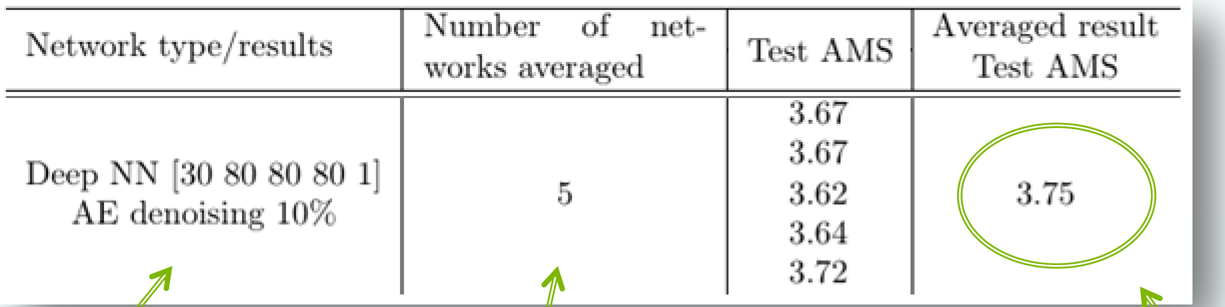


$$E_n = \frac{1}{2}\sum_k (\widehat{y_{nk}} - t_{nk})^2 \cdot \boldsymbol{w_n}$$

Set according to the weights of the training dataset

Forces the network to learn where matter

Multiple Averaged Deep Neural Network + Missing Data + Error weights



Network type/results	Number of networks averaged	Test AMS	Averaged result Test AMS
Deep NN [30 80 80 80 1] AE denoising 10%	5	3.67	3.75
		3.67	
		3.62	
		3.64	
		3.72	

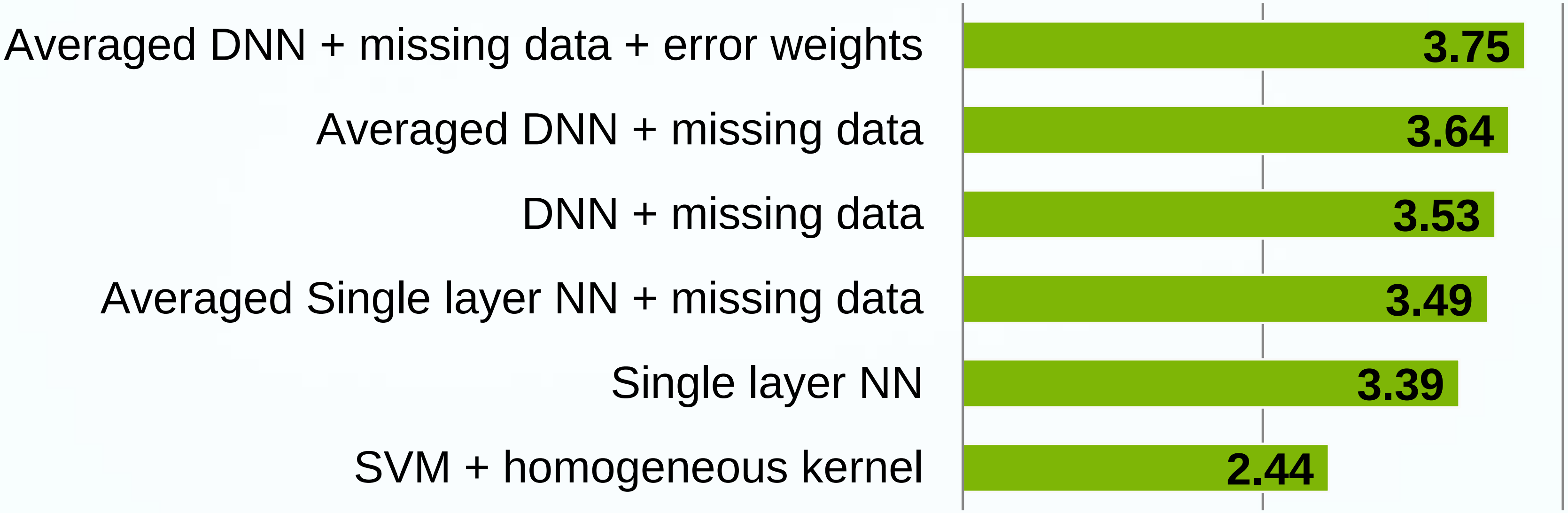
Same architecture as before

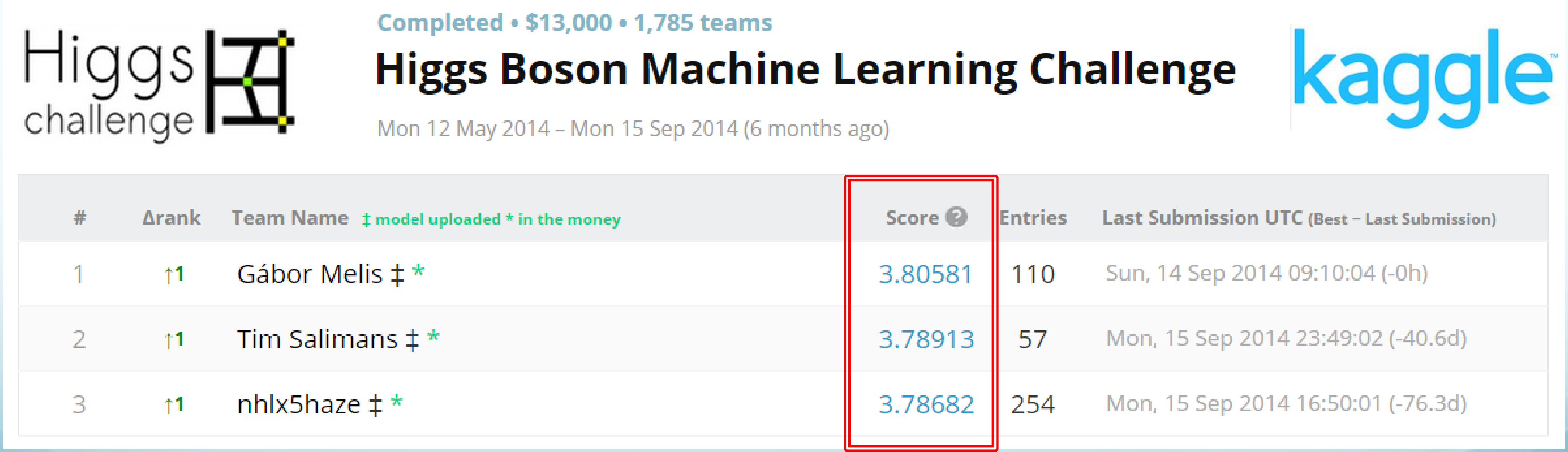
Number NN

Results of each DeepNN

Result of averaging process!!!!

Results





Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

Mon 12 May 2014 – Mon 15 Sep 2014 (6 months ago)

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Gábor Melis ‡ *	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↑1	Tim Salimans ‡ *	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	↑1	nhlx5haze ‡ *	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)

Our rank: 14th over ~1800 teams → **top 8% in 2 month**

Questions?