DEEP LEARNING IN HIGH-ENERGY PHYSICS

Improving the Search for Exotic Particles

Peter Sadowski

February 9, 2015

University of California Irvine

DEEP LEARNING FOR PARTICLE COLLIDER DATA ANALYSIS

Problem:
Using simulated collision data, learn to distinguish signal from background.

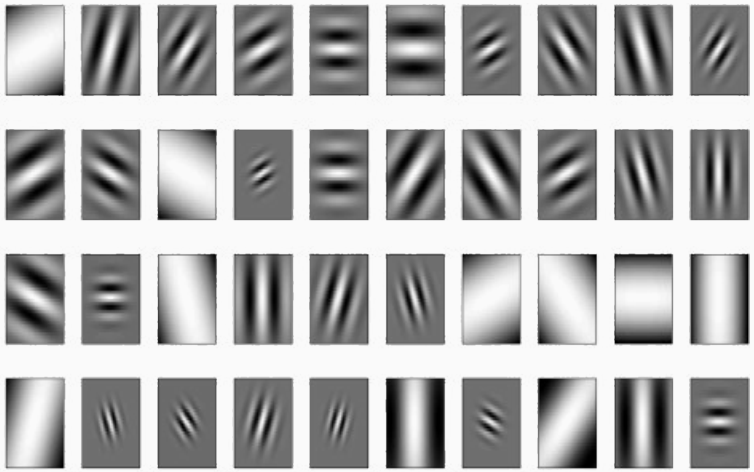
Current approach:
Shallow ML methods + feature engineering

Proposed approach:
Deep learning

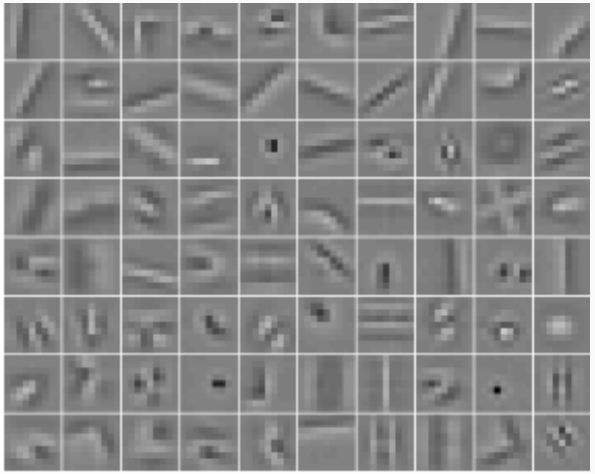
DEEP LEARNING FOR PARTICLE COLLIDER DATA ANALYSIS

Motivated by successes of deep learning in vision and speech.

- Huge progress on benchmark supervised learning tasks
- Replacement of **engineered** features with **learned** features



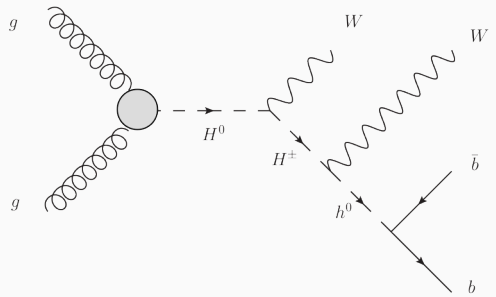
Engineered features



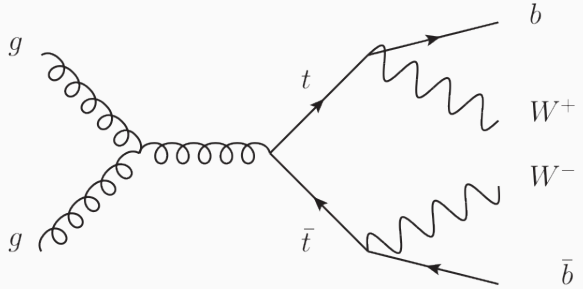
Learned features

DETECTING THE HIGGS BOSON

A two-class supervised learning problem:



Higgs-production



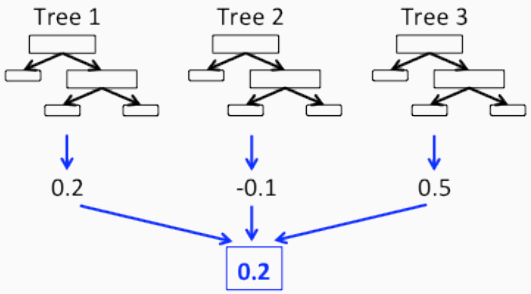
Primary background

Machine learning classifier:

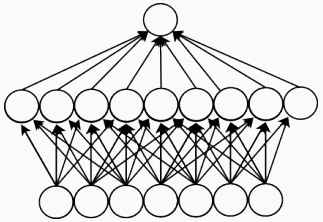
- 28 features
  - 21 low-level features
  - 7 high-level features derived by physicists
- 10M simulated collisions for training (50% each)
- 500k validation set
- 500k test set

DETECTING THE HIGGS BOSON

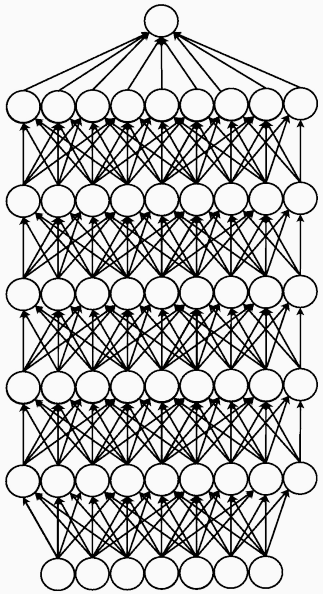
- Current approach: shallow models
  - Boosted decision trees* (BDT)
  - Shallow neural networks (NN)
- Our approach: deep neural networks (DNN)



BDT



NN



DNN

*Used for Higgs discovery in 2012

DETECTING THE HIGGS BOSON

Experimental comparison of classifiers:

- BDT and shallow NN as implemented in TMVA
- DNN trained on GPUs, using simple grid search to tune NN hyperparameters:
  - Learning rate
  - Weight-decay regularization
  - Number of hidden units
  - Number of hidden layers

DETECTING THE HIGGS BOSON

	Area Under ROC Curve for Test Set	
Technique	Low-level features	**All features**
BDT	0.73	**0.81**
NN	0.733 (0.007)	**0.816** (0.004)
DNN	0.880 (0.001)	**0.885** (0.002)

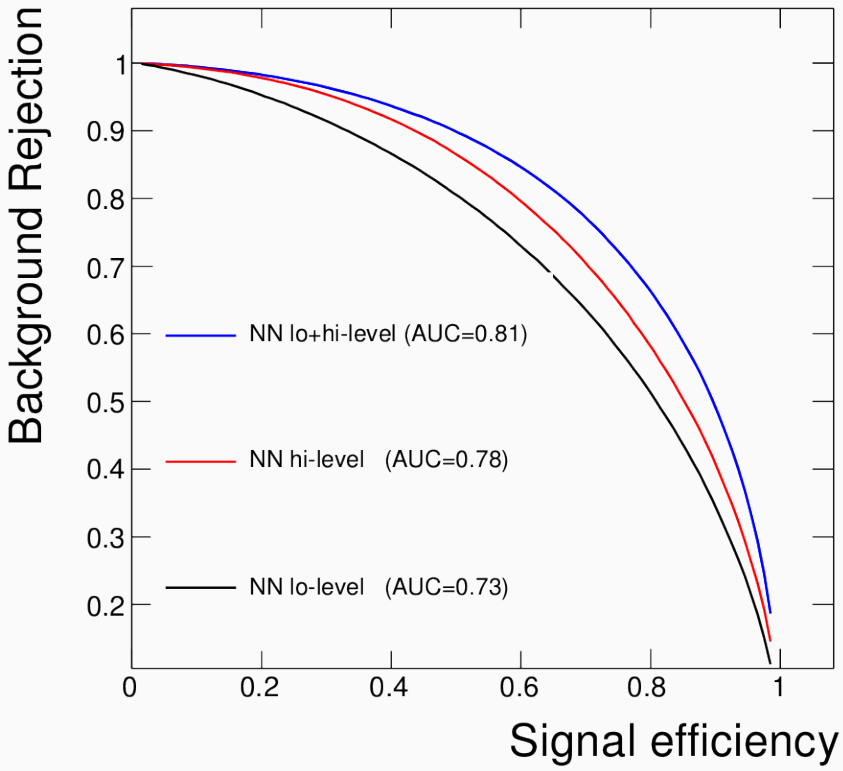
Deep learning improves AUC by 8% over shallow methods.

Deep learning does not require engineered features.

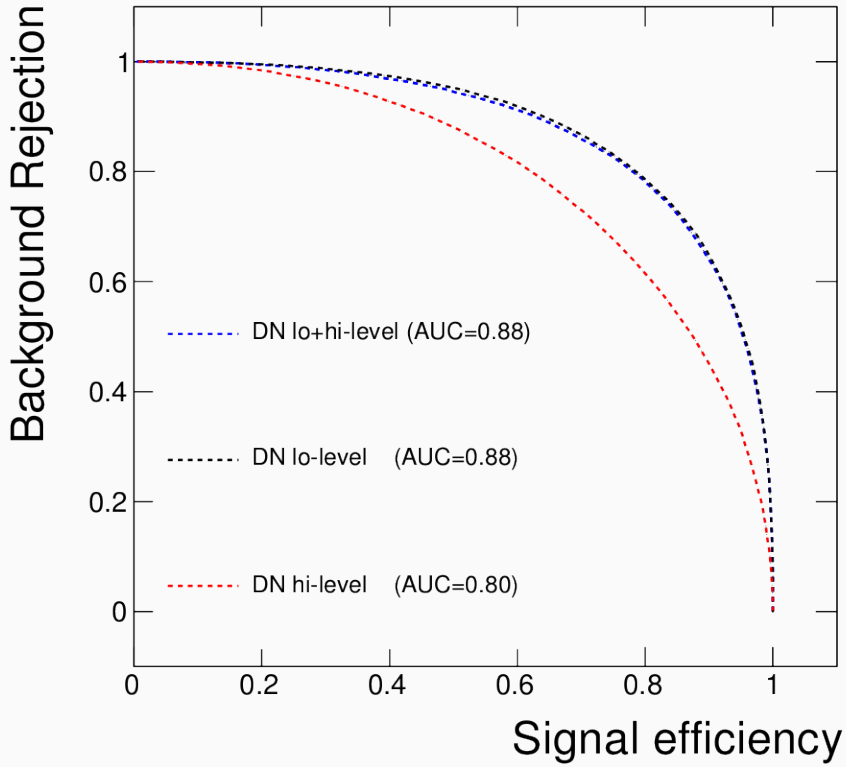
Baldi et al, Nature Communications 2014

DETECTING THE HIGGS BOSON

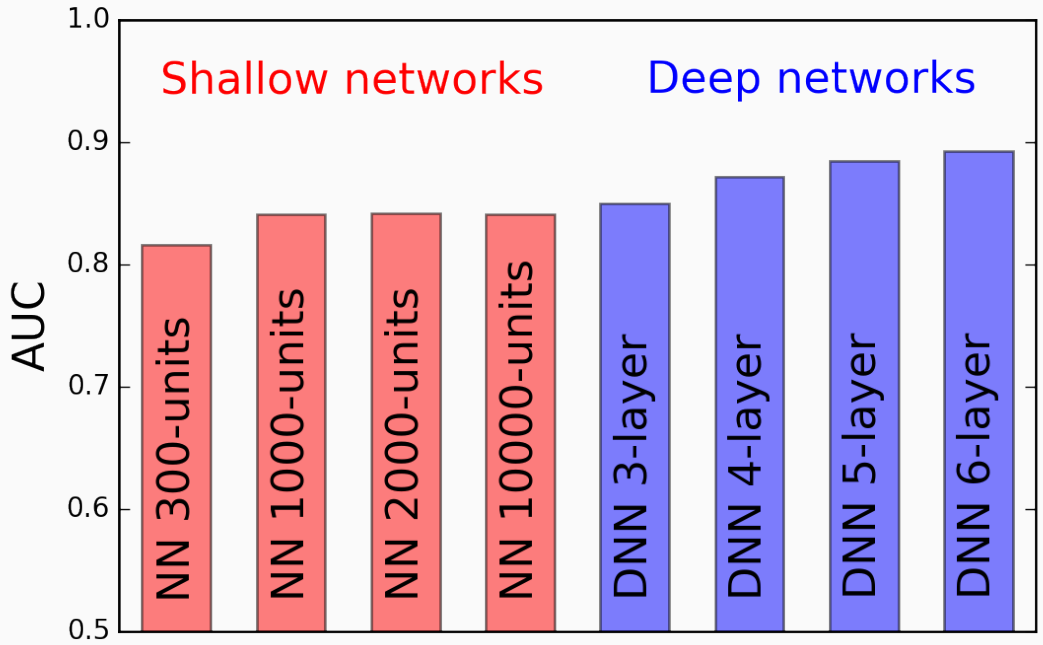
Shallow networks



Deep neworks



Analysis of wide shallow NNs vs. deep NNs

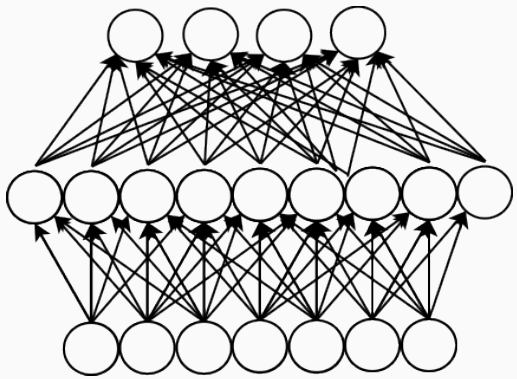


Depth helps more than width

DETECTING THE HIGGS BOSON

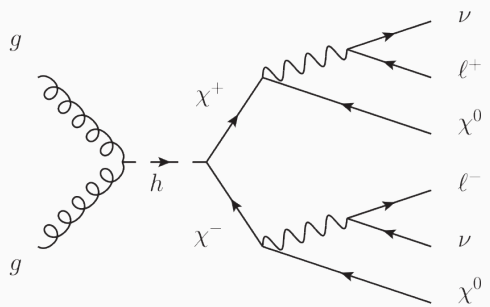
Mean Squared Error of networks trained to compute 7 high-level features from 21 low-level features.

Technique	Feature Regression MSE
Linear Regression	0.1468
NN	0.0885
DNN 3 layers	0.0821
DNN 4 layers	0.0818
DNN 5 layers	0.0815
DNN 6 layers	0.0812

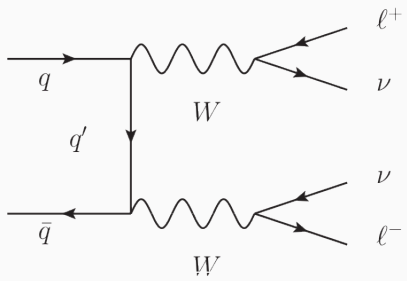


High-level features easier to learn with deep nets

Another experiment:



Process with hypothetical supersymmetric particles



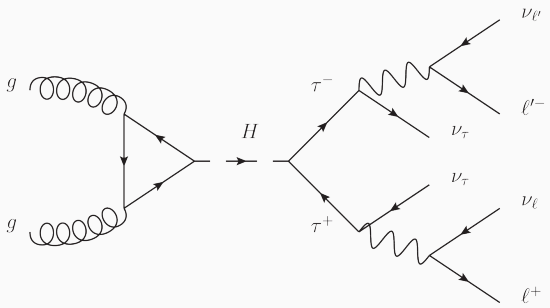
Background with W bosons

	AUC	
Technique	Low-level	**All features**
BDT	0.850 (0.003)	**0.863** (0.003)
NN	0.867 (0.002)	**0.875** ($<$ 0.001)
DNN	0.876 ($<$ 0.001)	**0.879** ($<$ 0.001)

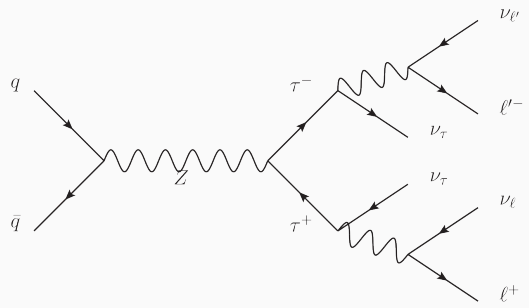
Deep learning again increases performance

HIGGS $\rightarrow \tau^+\tau^-$ DECAY

Another important problem:
Detect the hypothesized decay of the Higgs boson to fermions.



Higgs $\rightarrow \tau^+\tau^-$ decay



Background

Kaggle competition in Summer 2014

- 250k training examples
- Slightly different features

Dataset:

- 80 million examples (fast simulations)
- 10 low-level features
- 15 high-level features

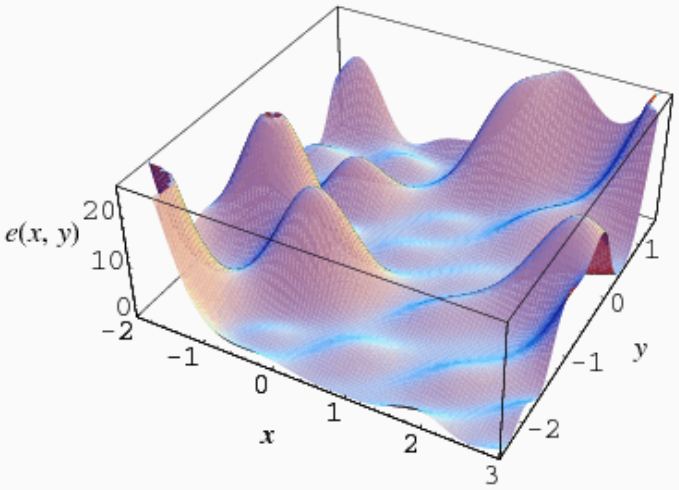
HIGGS $\rightarrow \tau^+\tau^-$ DECAY

DNN hyperparameters selected automatically with Bayesian optimization

DNN hyperparameter search space:

- Units per layer (100-500 units)
- Number of layers (2-8)
- Learning rate, LR decay, momentum parameters

Objective: Validation set error



HIGGS $\rightarrow \tau^+\tau^-$ DECAY

Bayesian Optimization results:

- Algorithm chose deepest possible network (8 layers)
- Only 274 hidden units per layer
- Overfitting observed with larger (wider) networks

Conclusion: Deep skinny networks help avoid overfitting

Constraints seem to force lower layers to learn generalizable features

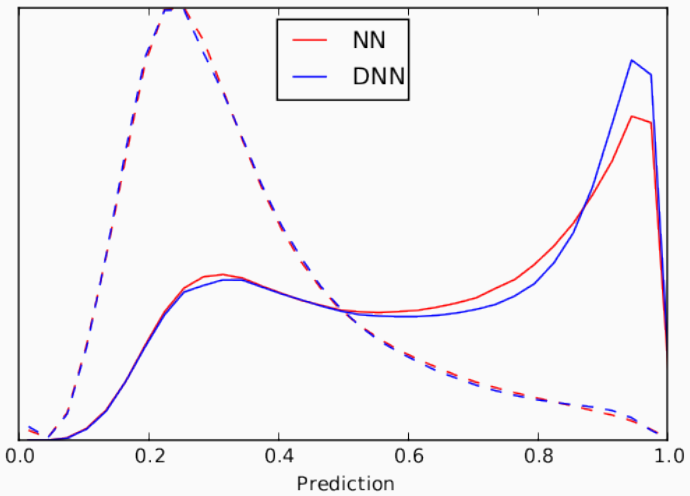
HIGGS $\to \tau^+\tau^-$ DECAY

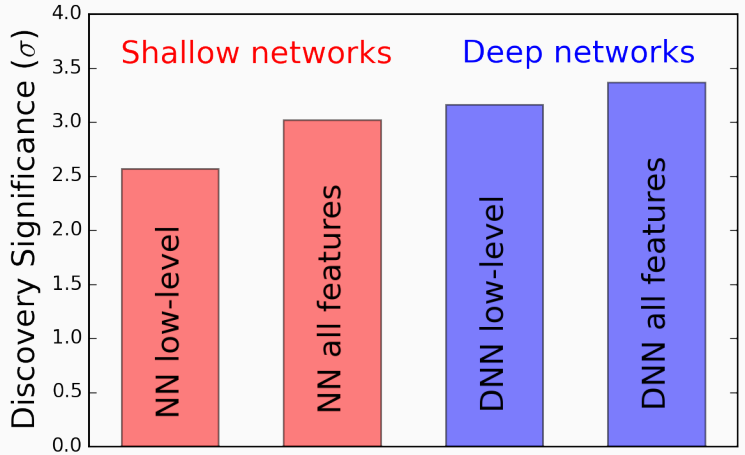
	AUC	
Technique	Low-level features	**All features**
NN	0.789 (0.0010)	**0.797** (0.0004)
DNN	0.798 (0.0001)	**0.802** (0.0001)

	Discovery significance	
Technique	Low-level features	**All features**
NN	2.57σ (0.006)	$\mathbf{3.02}\sigma$ (0.008)
DNN	3.16σ (0.003)	$\mathbf{3.37}\sigma$ (0.003)

20% reduction in experimental data needed for same significance

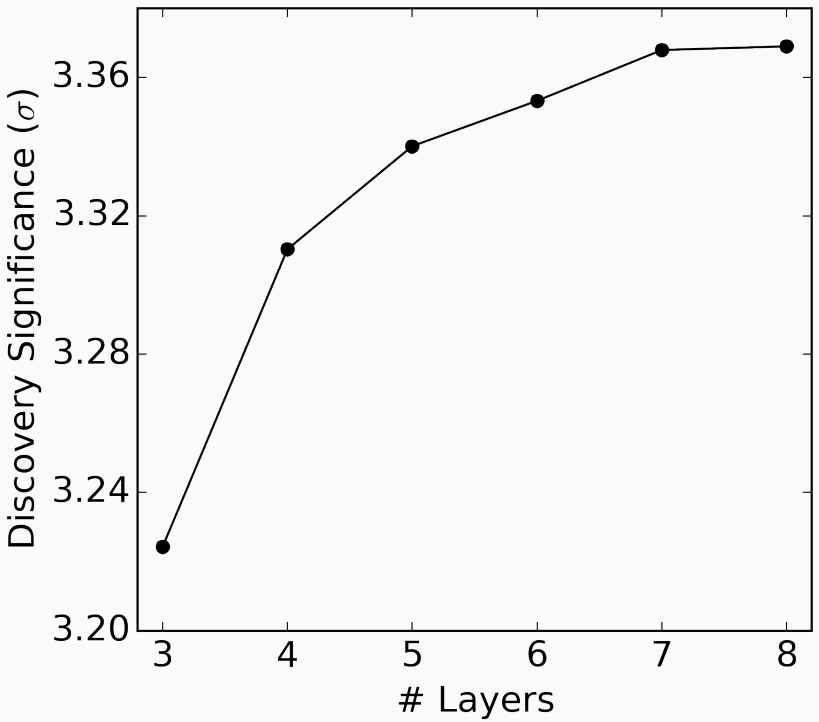
HIGGS $\rightarrow \tau^+\tau^-$ DECAY

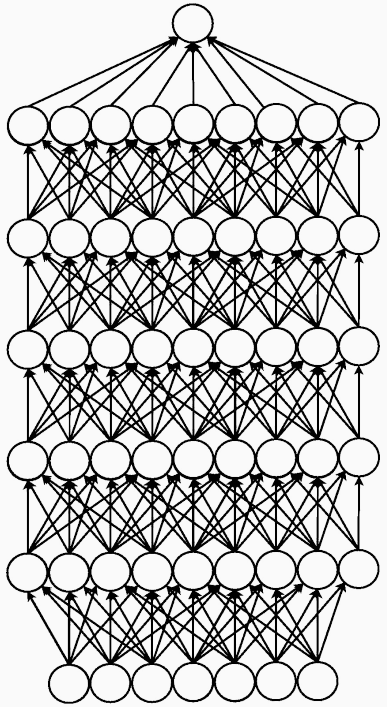




DNNs on low-level features beat NNs with engineered features

HIGGS $\rightarrow \tau^+\tau^-$ DECAY





Performance improves with up to 8 layers

CONCLUSIONS

1. Deep learning boosts the statistical power of HEP data analysis
    - More useful than feature engineering
    - Parallels success of deep learning in other domains
2. Neural networks have high capacity
    - Even shallow nets fit very complex functions
    - Overfitting to 80M examples
3. Deep tends to generalize better than shallow
    - Optimization algorithm chooses deep, skinny net
    - Seems to learn features that generalize

THANK YOU

Collaborators:

- Pierre Baldi, Dept. of Computer Science
- Daniel Whiteson, Dept. of Physics and Astronomy

Special thanks to:

- Mike Gelbart and other developers of Spearmint/Whetlab
- The developers of Pylearn2 and Theano

HIGGS $\rightarrow \tau^+\tau^-$ DECAY

Low-level features:

- 3D momenta, p, of each charged lepton;
- Momenta of particle 'jets' due to radiation of gluons or quarks;
- Imbalance of transverse momentum ($\not{E}_T$) in the final state.

High-level features include:

- Axial missing momentum, $\not{E}_T \cdot p_{\ell^+\ell^-}$;
- Angular distance between leptons, $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$;
- Missing mass, m_{MMC};