Theme : Machine learning for high energy physics (HEP)

Dissertation Title : A meta-algorithm for cascades : A high energy physics application

Vidhi Lalchand
Supervisor: Dr. Anita Faul
Physics Advisor: Dr. Chris Lester

May 18, 2016

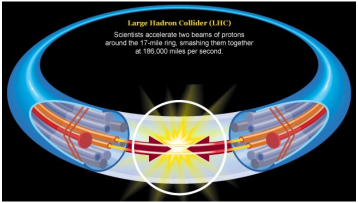
The future of HEP

'The future of high energy physics lies in large circular colliders colliding particles at really high energies' - Nima Arkani Hamed.

An important goal of the particle physics community worldwide is the search for new particles and physical processes that complement and go beyond the Standard Model of particle physics. Eg. Supersymmetry, the Higgs boson. As the LHC is set to hit higher energy frontiers the rate of data collection is set to increase by a factor of 10.

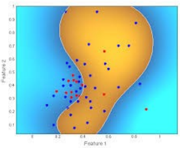
The process of discovery in HEP

Accelerator ring

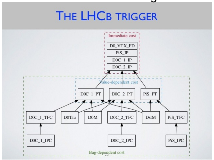


5σ → Discovery

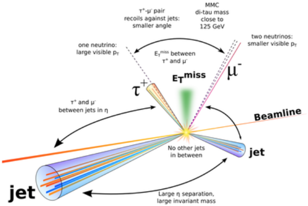
Multivariate Analysis



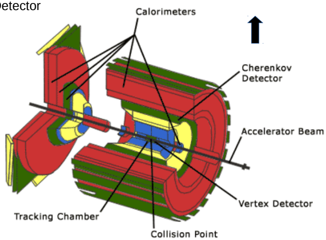
Preselection of interesting events



Collisions



Detector



Classification in HEP

There are 3 types of classification processes that take place at the LHC.

- Trigger classifiers - fast and sclalable, real time signal/background separation.

Classification in HEP

There are 3 types of classification processes that take place at the LHC.

- Trigger classifiers - fast and sclalable, real time signal/background separation.
- **Classification for discovery** - sophisticated models for signal/background separation which are slower to train and must learn to classify events from classes that mimic each other closely.

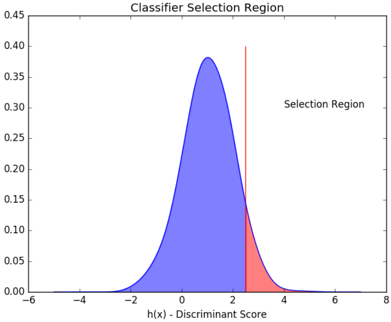
Classification in HEP

There are 3 types of classification processes that take place at the LHC.

- Trigger classifiers - fast and sclalable, real time signal/background separation.
- **Classification for discovery** - sophisticated models for signal/background separation which are slower to train and must learn to classify events from classes that mimic each other closely.
- Deep learning - Unsupervised learning, automatically inferring powerful latent features in the data.

Classification for discovery

Events generated in the collider are preprocessed and represented as high dimensional *feature vectors*. These can be represented as $x \in \mathbb{R}^d$. A classifier $h(x) : \mathbb{R}^d \to \mathbb{R}$ is trained to classify signal events (s) from background (b). It outputs a discriminant score taking small values in one class and large values in the other. By putting a threshold on the discriminant value, a selection region $\mathcal{H}$ is chosen.



Classification for discovery - the math

Let the selection region $\mathcal{H}$ be characterized by,

$$\mathcal{H} = \{\mathbf{x} : h(\mathbf{x}) > \theta\} \tag{1}$$

s: signal events in $\mathcal{H}$, or the true positives.
b: background events in $\mathcal{H}$, or the true negatives.

$$\text{AMS} = \frac{s}{\sqrt{b}}$$

This is the fundamental objective function the classifiers are tuned to maximise.

Classification for discovery - the math

The occurrence of background events follow a Poisson process. Over a given time period during which events are recorded, the expected number of *selected* background events is μ_b and its variance is also μ_b The normalized statistic,

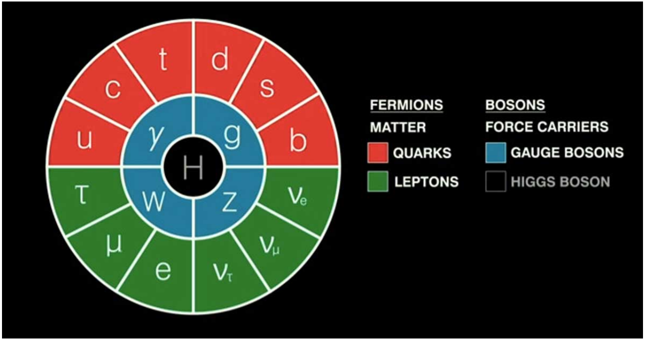
$$\hat{t} = (n - \mu_b)/\sqrt{\mu_b} \sim N(0, 1) \tag{2}$$

serves as a test statistic for detection of signal events. A fluctuation is considered sufficiently large to claim a discovery of the signal process if it exceeds 5σ, i.e. if $\hat{t} > 5$ ($\sigma = 1$ for the normalized test statistic).

$$(n - \mu_b)/\sqrt{\mu_b} = (s + b - b)/\sqrt{b} = s/\sqrt{b} \tag{3}$$

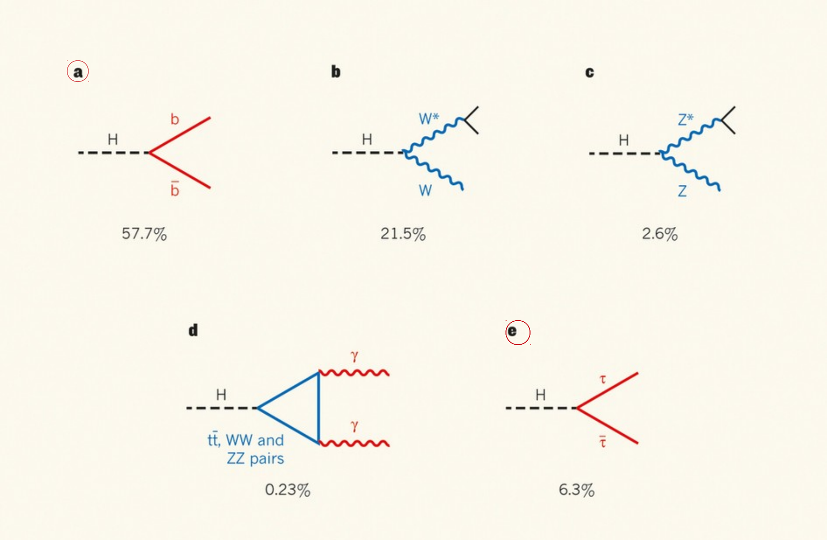
The Higgs particle (H)

In 2012, the ATLAS experiment claimed the discovery of a new particle, the Higgs boson. The existence of this particle provides support to the theory that that a field permeates the universe through which fundamental particles acquire mass, a theory which is cardinal for the completeness of the Standard Model of particle physics.



The Higgs Decay

A key property of a particle is how it decays into other particles. The Higgs has 5 experimentally accessible decay channels.



H $\rightarrow \tau\tau$

In the original discovery the Higgs was seen decaying into $\gamma\gamma$, WW and ZZ. The $H \rightarrow \tau\tau$ channel is particularly interesting as it hasnt been experimentally verified .i.e. the Higgs to tau-tau excess is not yet at 5σ. This decay channel is particularly hard to explore due to two reasons:

- The decay into two taus is not a unique channel, in fact the Z boson can also decay into two taus, further this happens a lot more frequently than the Higgs. The two decays produce events which have very similar signatures and this prevents a clean separation of the parent candidate.

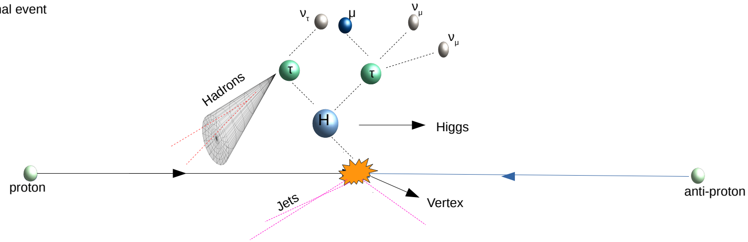
H $\rightarrow \tau\tau$

In the original discovery the Higgs was seen decaying into $\gamma\gamma$, WW and ZZ. The $H \rightarrow \tau\tau$ channel is particularly interesting as it hasnt been experimentally verified .i.e. the Higgs to tau-tau excess is not yet at 5σ. This decay channel is particularly hard to explore due to two reasons:

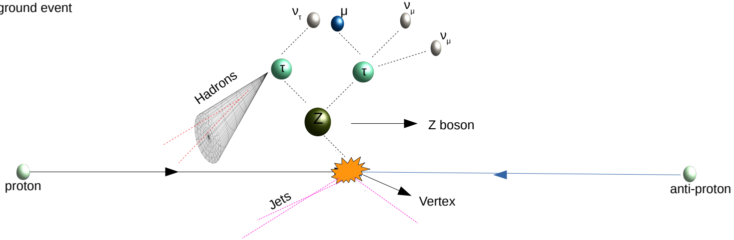
- The decay into two taus is not a unique channel, in fact the Z boson can also decay into two taus, further this happens a lot more frequently than the Higgs. The two decays produce events which have very similar signatures and this prevents a clean separation of the parent candidate.
- Taus are heavy and unstable, they decay instantaneously. Their dominant decay modes involve neutrinos and the presence of these undetectable particles in their decay make it difficult to reconstruct the mass of the Higgs on an event by event basis.

Decay signatures

Signal event



Background event



The formal problem

Inputs : $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), ...(\mathbf{x}_n, y_n, w_n)\}$
$\mathbf{x}_i \in \mathbb{R}^d$,
$y_i \in \{b, s\}$,
$w_i \in \mathbb{R}^+$
$\mathcal{S} = \{i : y_i = s\}$ and $\mathcal{B} = \{i : y_i = b\}$
$n_s = |\mathcal{S}|$ and $n_b = |\mathcal{B}|$

$$\sum_{i \in \mathcal{S}} w_i = N_s \quad \text{and} \quad \sum_{i \in \mathcal{B}} w_i = N_b \tag{4}$$

The constants have physical meaning, they are the expected total number of signal and background events during the time interval over which the data has been recorded (in the dataset used, it is the year 2012).

The formal statement

Based on the inputs provided we train a classifier, h, generating a selection region, $\mathcal{H} = \{\mathbf{x} : h(\mathbf{x}) = s\}$, $\mathbf{x} \in \mathbb{R}^d$.
$\hat{\mathcal{H}} = \{i : \mathbf{x}_i \in \mathcal{H}\}$ is the index set.

The quantities,

$$s = \sum_{i \in \mathcal{S} \cap \hat{\mathcal{H}}} w_i \quad \text{and} \quad b = \sum_{i \in \mathcal{B} \cap \hat{\mathcal{H}}} w_i \tag{5}$$

are the true positives and false positives which are used to calculate the AMS $= \dfrac{s}{\sqrt{b}}$.

The Data

- The ATLAS experiment have made publicly available a simulated (labelled) dataset that has been used by physicists in analysing the Higgs to tau-tau decay.
  1. Training Set: 250K events
  2. Test Set: 450K events
  3. Validation Set: 100k events

The Data

- The ATLAS experiment have made publicly available a simulated (labelled) dataset that has been used by physicists in analysing the Higgs to tau-tau decay.
    1. Training Set: 250K events
    2. Test Set: 450K events
    3. Validation Set: 100k events
- In nature, signal events occur much less frequently than background events. However, the dataset published has been enriched with signal events (30:70) to present a more balanced classification problem.

The Data

- The ATLAS experiment have made publicly available a simulated (labelled) dataset that has been used by physicists in analysing the Higgs to tau-tau decay.
  1. Training Set: 250K events
  2. Test Set: 450K events
  3. Validation Set: 100k events
- In nature, signal events occur much less frequently than background events. However, the dataset published has been enriched with signal events (30:70) to present a more balanced classification problem.
- To compensate for this bias, all events are weighted with importance weights reflecting their probability of occurrence.

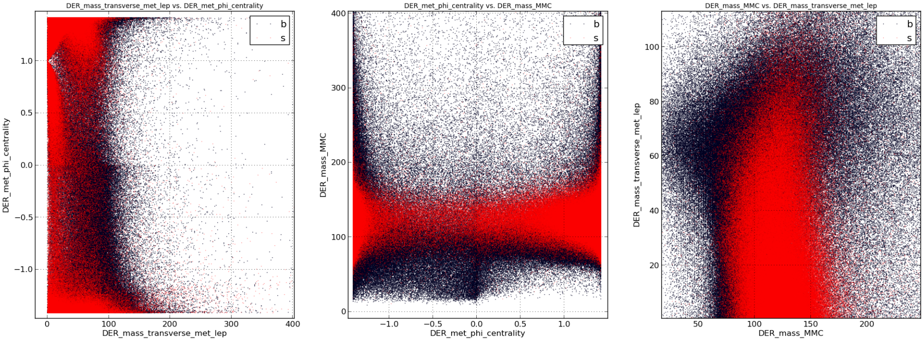
The machine learning perspective

Goal:
To propose a classifier h to identify a signal-rich region in the feature space where an excess of signal over background events are expected $\left(\frac{s}{\sqrt{b}}\right)$.

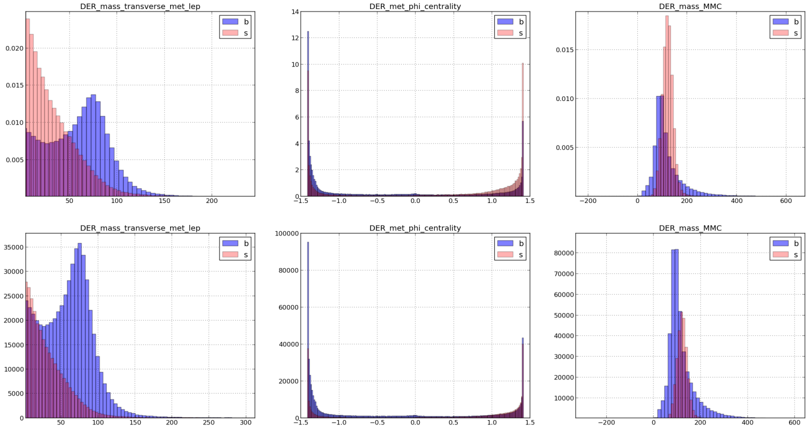
Challenges:

- The signal and background classes densely overlap.
- The discovery significance metric the AMS is unusual, noisy.

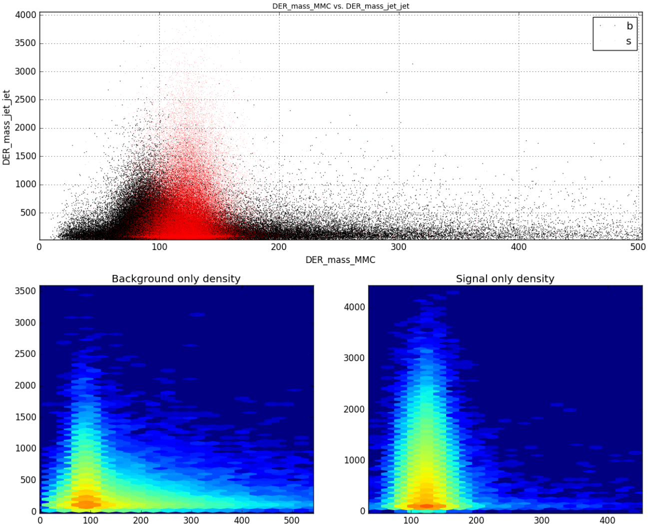
Class overlap



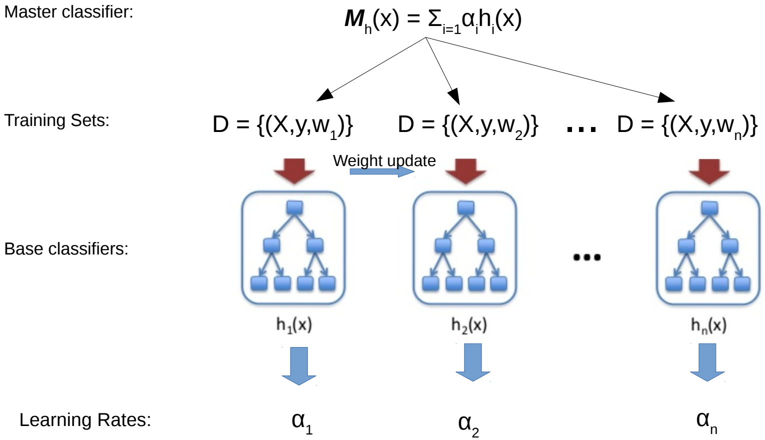
Class overlap



Density


DER_mass_MMC vs. DER_mass_jet_jet
b
s
DER_mass_jet_jet
DER_mass_MMC
0
100
200
300
400
500
500
1000
1500
2000
2500
3000
3500
4000
Background only density
Signal only density

Boosting Algorithm



Boosting Algorithm

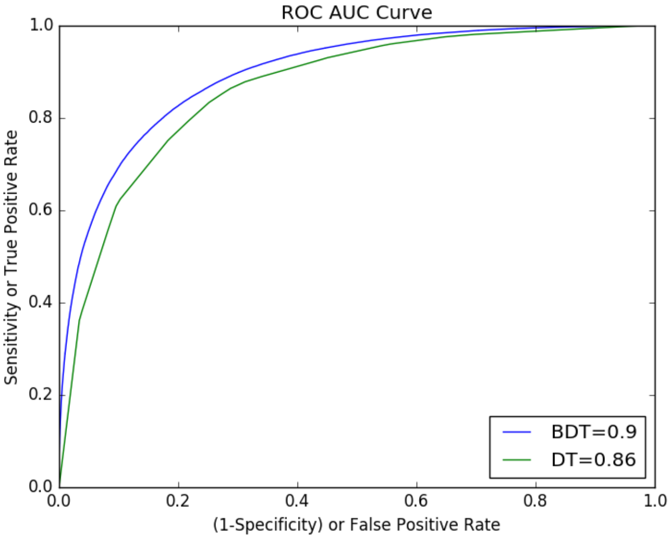
Algorithm 1 AdaBoost with M stages

Initialize the data weighting coefficients $\{w_i\} = 1/N \; \forall i = 1 \ldots N$
for all $m = 1 \ldots \mathrm{M}$ **do**
    Fit classifier $h_m(\mathbf{x})$ by minimizing the weighted error function $R(h_m) = \sum_{i=1}^{N} w_i \mathbf{I}(h_m(\mathbf{x}_i) \neq y_i)$
    Compute error rate ϵ_m as the fraction of misclassified samples.
    Compute classifier weight $\alpha_m = \frac{1}{2} \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$
    Update weights for stage $(m+1)$ by, $w_i^{(m+1)} = w_i^{(m)} e^{\alpha_m \mathbf{I}(h_m(\mathbf{x}_i) \neq y_i)}$
end for
$M_h(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m h_m(\mathbf{x})$
return

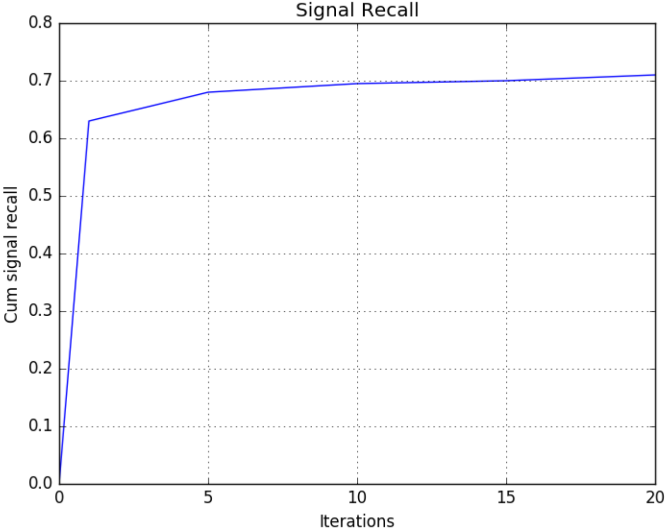
Baseline performance

Classifier	**Precision** [TP/(TP+FP)]	**Recall** [TP/P]	AMS
Weak Learner			
Background	0.79	0.79	
Signal	0.60	0.61	
Average	0.73	0.73	2.33σ
AdaBoost with weak learner			
Background	0.84	0.84	
Signal	0.69	0.68	
Average	0.79	0.79	2.87σ
Strong learner			
Background	0.84	0.90	
Signal	0.76	0.62	
Average	0.80	0.80	3.41σ
AdaBoost with strong Learner			
Background	0.86	0.89	
Signal	0.77	0.71	
Average	0.83	0.83	3.52σ

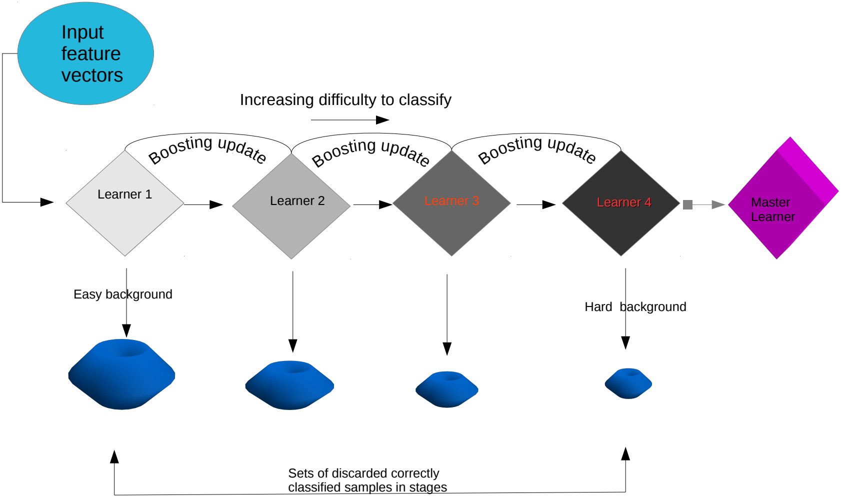
ROC curves


ROC AUC Curve
Sensitivity or True Positive Rate
(1-Specificity) or False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
BDT=0.9
DT=0.86

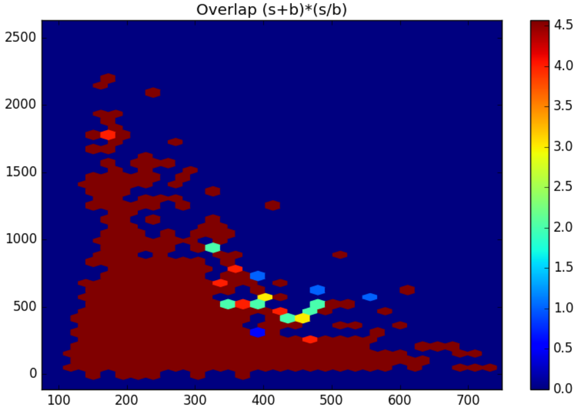
Signal Sensitivity



The *meta* approach



Fragmenting the feature space



Targets for Algorithm Design

- Adaptive selection of features.
- Diversity in component classifiers of the ensemble.
- Targetting overlapping regions during training.
- Focus on improving signal sensitivity $\Rightarrow$ maximixing AMS.

Open problems

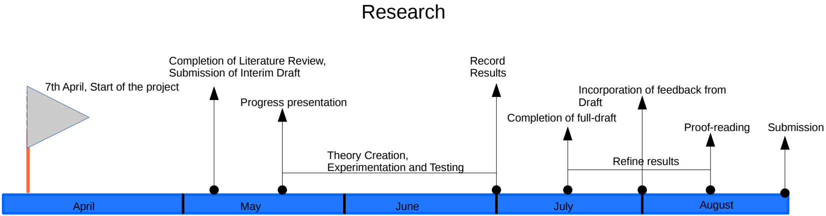
- Control for overfitting/overtraining (bias-variance trade-off).
- The choice of component classifiers at each stage of the classification.
- Demonstrate usage on datasets outside of a HEP context that demonstrate class overlap.

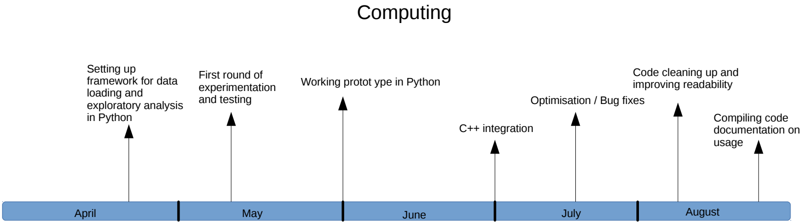
Application Domains

The main contribution of this work is to develop an algorithm that is generally applicable in domains where the data sets exhibit the problem of class-overlap in binary or multiple-classes.

- Medical Diagnosis
- Fraud detection
- Psychiatry (predicting criminality potential in individuals)
- Loan default predictions
- Geochemical distributions

Timeline


Research
Completion of Literature Review, Submission of Interim Draft
Record Results
Incorporation of feedback from Draft
7th April, Start of the project
Progress presentation
Completion of full-draft
Proof-reading
Submission
Theory Creation, Experimentation and Testing
Refine results
April
May
June
July
August


Computing
Setting up framework for data loading and exploratory analysis in Python
First round of experimentation and testing
Working protot ype in Python
Code cleaning up and improving readability
Optimisation / Bug fixes
C++ integration
Compiling code documentation on usage
April
May
June
July
August

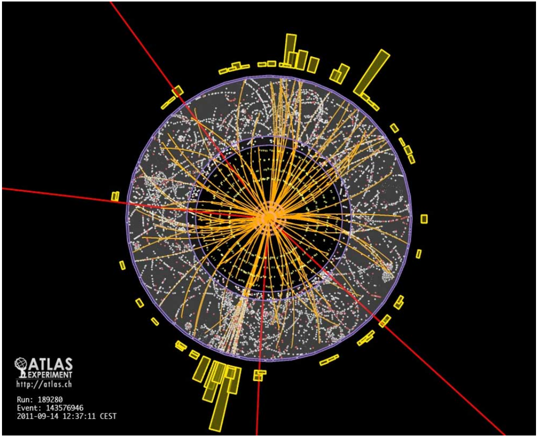
Critical references

- Journal of Machine learning Research, Workshop and Conference Proceedings 42:19-55, 2015
  The Higgs Boson Machine learning challenge
  http://jmlr.org/proceedings/papers/v42/cowa14.pdf
- ACM Transactions on Embedded Computing Systems, Vol. 9, No. 4, Article 1, April 2015
  A Review of Boosted Machine learning Techniques for Particle Identification at ALICE
  https://people.cs.uct.ac.za/ wngrya001/documentation/lit_review_particle_identify.pdf

TensorFlow Library

- Google's ***TensorFlow*** is an open source application grade machine learning library written in Python and C++ with full API support.
- It offers deep learning models on a flexible architecture and is accessible through a python API and a lesser documented C++ API.
- CERN has expressed interest in using TensorFlow technology for machine learning tasks in the realm of HEP.
- TensorFlow can run on multiple CPUs and GPUs with optional CUDA extensions.
- It was developed by Dr. Geoffrey Hinton and Dr. Jeff Dean.

- Thank you -


ATLAS
EXPERIMENT
http://atlas.ch
Run: 189280
Event: 143576946
2011-09-14 12:37:11 CEST