

# Machine learning for event selection in high energy physics

Shimon Whiteson<sup>a,\*</sup> Whiteson<sup>b</sup>

<sup>a</sup>*Informatics Institute, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, Netherlands*

<sup>b</sup>*Department of Physics & Astronomy, University of California, Irvine, 4129 Frederick Reines Hall, Irvine, CA 92697-4575*

---

## Abstract

The field of high energy physics aims to discover the underlying structure of matter by searching for and studying exotic particles, such as the top quark and Higgs boson, produced in collisions at modern accelerators. Since such accelerators are extraordinarily expensive, extracting maximal information from the resulting data is essential. However, most accelerator events do not produce particles of interest, so making effective measurements requires event selection, in which events producing particles of interest (signal) are separated from events producing other particles (background). This article studies the use of machine learning to aid event selection. First, we apply supervised learning methods, which have succeeded previously in similar tasks. However, they are suboptimal in this case because they assume the selector with the highest classification accuracy will yield the best final analysis; this is not true in practice, as such analyses are more sensitive to some backgrounds than others. Second, we present a new approach that uses stochastic optimization techniques to directly search for selectors that maximize either the precision of top quark mass measurements or sensitivity to the presence of the Higgs boson. Empirical results confirm that stochastically optimized selectors result in substantially better analyses. We also describe a case study in which the best selector is applied to real data from the Fermilab Tevatron accelerator, resulting in the most precise top quark mass measurement of this type to date. Hence, this new approach to event selection has already contributed to our knowledge of the top quark's mass and our understanding of the larger questions upon which it sheds light.

*Key words:* Machine learning, neural networks, evolutionary computation, high energy physics.

---

## 1. Introduction

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 interactions, this field strives to discover the fundamental properties of the physical universe. In experimental high energy physics, the goal is to test predictions made by current theories such as the *Standard Model* (1; 2; 3), which describes the behavior of three of the four fundamental forces.

The primary tools of experimental high energy physicists are modern accelerators, which collide protons and/or anti-protons to create exotic particles that occur only at extremely high energy densities. Such particles have not existed naturally since the first moments after the Big Bang, when the energy density of the universe was much higher. Observing these particles and measuring their properties may yield critical insights about the very nature of mass.

Two particles of particular interest are the *top quark* and the *Higgs boson*. The top quark, first observed in 1995 (4; 5), is nearly as massive as a gold nucleus, making it by far the most massive subatomic particle ever observed. The top quark is important because precise measurements of its mass can stringently test theories about the origins of particle mass (6; 8; 9; 10). Only the world's most powerful accelerator, the Fermilab Tevatron in Batavia, Illinois, has sufficient energy to produce top quarks. By contrast, the Higgs boson (11) has never been observed. In fact, it is the only remaining particle predicted by the Standard Model whose existence has not been experimentally verified (3). Since the Higgs boson is theorized to give mass to other particles through its interactions (12), it is central to current theories about particle mass. Hence, observing the Higgs boson is a paramount goal in high energy physics.

Producing and observing such particles requires extraordinary resources. The Tevatron accelerator and its particle detectors cost billions of dollars to construct and approximately a million dollars per day to operate. As a result, extracting maximal information from the resulting data is

---

\* Corresponding author: tel +31.020.525.8701; fax +31.020.525.7490  
*Email addresses:* s.a.whiteson@uva.nl (Shimon Whiteson),  
daniel@uci.edu (Whiteson).

essential. In this article, we study the use of machine learning methods to aid this process. In particular, we investigate their efficacy for *event selection* for top quark mass measurement and Higgs boson search.

In an accelerator event, protons and/or anti-protons are accelerated and annihilated. The resulting energy causes new particles to form, which can be observed via detectors that surround the point of collision. However, the vast majority of events do not produce particles of interest, such as the top quark or Higgs boson. For example, though the Tevatron produces approximately  $10^{10}$  events per hour, approximately one results in a top quark, on average. Therefore, good data analysis depends on effective event selection, in which events producing particles of interest (*signal*) are separated from those producing other particles (*background*). Event selection is difficult because several types of background can mimic the signal's characteristic signature. Hence, event selection in high energy physics is an exciting challenge for machine learning. In this article, we compare two different approaches to this problem.

The first approach is based on supervised learning methods, which are used to train classifiers that distinguish signal from background. Such methods have already proven successful in similar event selection problems by training neural networks (13; 14) or support vector machines (15) to classify events as signal or background. This supervised approach is most effective in the narrow class of problems in which the classification accuracy of the event selector is closely correlated with the quality of the resulting data analysis and systematic uncertainties are minimal. However, top quark mass measurement and Higgs boson search exemplify a broader class of problems where higher classification accuracy does not necessarily result in superior analysis performance. Instead, the top quark mass measurement is more sensitive to the presence of some background events than others, in ways that are difficult to predict *a priori*. The Higgs boson search is limited by systematic uncertainties on the background events. Therefore, selectors that maximize classification accuracy may perform worse than those that 1) increase the quantity of signal by tolerating harmless background, 2) reduce the quantity of signal to eliminate disruptive background, or 3) minimize the impact of systematic uncertainties.

To find such selectors, we introduce a second, novel approach that uses stochastic optimization techniques. Rather than maximizing classification accuracy, this approach directly optimizes selectors for their true purpose: maximizing either the precision of top quark mass measurements or sensitivity to the presence of the Higgs boson. Using NEAT (16), an evolutionary method for training neural networks, we optimize event selectors that operate either in conjunction with supervised classifiers or in lieu of them.

This article presents experiments that compare the performance of manually designed heuristic selectors to neural network selectors trained with backpropagation (17) or NEAT. In both top quark mass measurement and Higgs bo-

son search, the learning methods perform significantly better than the heuristic approach, confirming that machine learning can greatly benefit event selection in high energy physics. Furthermore, the NEAT selectors yield by far the best analyses, demonstrating the advantage of the stochastic optimization approach in an application area previously assumed the province of supervised methods.

Finally, this article describes a detailed case study in which the best performing selector is applied to real data gathered by the CDF II detector at the Tevatron. The result is a substantial reduction in uncertainty in the top quark mass measurement, yielding by far the most precise measurement of this type to date. Obtaining a similar reduction in uncertainty would otherwise require producing many more collisions at great expense. Hence, this new approach to event selection has already contributed substantially to our knowledge of the top quark's mass and our understanding of the larger questions upon which it sheds light.

The approaches we propose also offer potential benefits beyond the analysis of data gathered at the Tevatron. The future of high energy physics lies with the Large Hadron Collider (LHC), a new accelerator currently under construction. Once in operation, the LHC will produce collisions at much higher frequency and energy than the Tevatron. The resulting torrent of data will require highly effective event selection, which can potentially be aided by the methods presented in this article.

The remainder of this paper is organized as follows. Section 2 overviews the process of producing, detecting, selecting, and analyzing events in modern high energy accelerators. Sections 3 and 4 describe methods for performing event selection with the aid of supervised learning and stochastic optimization, respectively. Sections 5 and 6 compare the performance of these methods on the problems of top quark mass measurement and Higgs boson search, respectively. Section 7 describes the use of our optimized selector to produce a new top quark mass measurement with data from the Tevatron. Section 8 discusses the the implications of these results, Section 9 outlines our plans for future work, and Section 10 concludes.

## 2. Events in High Energy Physics

This section provides an overview of the process of producing, detecting, selecting, and analyzing events at modern high energy accelerators, with particular focus on measurement of the top quark's mass and the search for the Higgs boson.

### 2.1. Producing Events

To provide enough energy to produce massive exotic particles such as the top quark or the Higgs boson, one must accelerate and annihilate lighter particles and their anti-particles. Figure 1 shows the Fermilab Tevatron accel-

erator complex in Batavia, Illinois, which includes a series of smaller accelerators that seed the final 6.5-kilometer Tevatron ring. The Tevatron accelerates protons and anti-protons to a center-of-mass energy of 1.96 tera-electronvolts (TeV), the highest controlled energy events ever achieved.

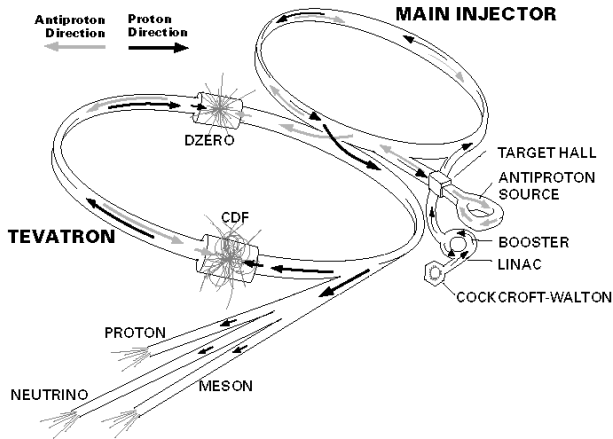


Fig. 1. Accelerator complex at Fermilab, showing the chain of lower energy accelerators used to prime the Tevatron, the world’s highest energy accelerator, which collides protons and anti-protons at two points (CDF II and DZERO) in the ring.

## 2.2. Detecting Events

Top quark and Higgs boson events are extremely rare. Though the Tevatron produces approximately  $10^{10}$  events per hour, only approximately 1 per hour yields a top quark and approximately 0.01 per hour produces a Higgs boson. Even when produced, detecting these particles is a great challenge. They cannot be directly observed since they are smaller than the wavelength of visible light. In addition, they decay into other particles almost immediately, in less than  $10^{-17}$  seconds.

Though the massive top quark and Higgs boson cannot be directly observed, the lighter stable particles to which they decay, called *decay products*, can be observed. Multiple layers of detectors (18) surround the point of collision for this purpose. As the decay products pass through these detectors, they interact with them in a way that allows their direction and energy to be observed. Figure 2 depicts CDF II, one of two detectors at the Tevatron.

## 2.3. Selecting Events

Since the vast majority of events do not produce particles of interest, culling signal from background is crucial to obtaining high quality analyses. This process occurs in two phases: 1) *pre-selection*, in which simple rules are applied to eliminate events that are trivially known to be background and 2) *final selection*, in which more subtle properties of the remaining events are examined to determine whether they contain particles of interest.

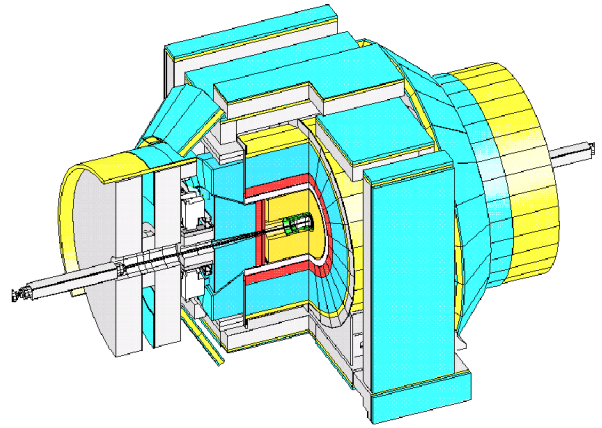


Fig. 2. The CDF II detector, approximately 10 meters high, is designed to study collisions of protons and anti-protons at the Fermilab Tevatron. The protons and anti-protons enter from the left and right and collide at the center, which is surrounded by concentric detector layers to measure the decay point (green), the direction and momentum (yellow) and the energy (red and cyan) of the event’s decay products.

Final selection can be performed using heuristics developed by physics experts (19) or, preferably, using machine learning, as will be detailed in Sections 3 and 4. The remainder of this section describes the pre-selection process for top quark mass measurement and Higgs search.

### 2.3.1. Pre-Selection for Top Quark Mass Measurement

In top quark mass measurement, the pre-selection phase discards all events that do not display the top quark’s characteristic signature. This signature emerges from the direction and energy of the top quark’s decay products, shown in Figure 3. The decay products consist of two leptons; two bottom quarks, which are seen in the detector as “jets” of lower energy particles; and an energy imbalance caused by missing neutrinos, which escape undetected. Pre-selection is safe because it has nearly 100% recall, i.e., signal events are almost never discarded.

Unfortunately, pre-selection does not have high precision because the characteristic signature is not specific to top quark decay (19). In fact, although pre-selection discards more than 99% of the events, 83% of those that remain consist of background that mimics the top quark’s signature. In particular, only five types of events may survive pre-selection. Their particular nature is not important for the process of final selection, but their relative frequencies are. These events are, in order of diminishing importance:

- (i) two gluons and a Z boson decaying to a pair of stable leptons (ee or mumu) (71% of the sample),
- (ii) The simultaneous measurement of three gluons and a W boson (6% of the sample),
- (iii) two gluons and a Z boson decaying to a pair of unstable (tau or tau-bar) leptons (3% of the sample),
- (iv) the simultaneous measurement of two gluons and two W bosons (2% of the sample), and
- (v) the simultaneous measurement of a Z boson and a W

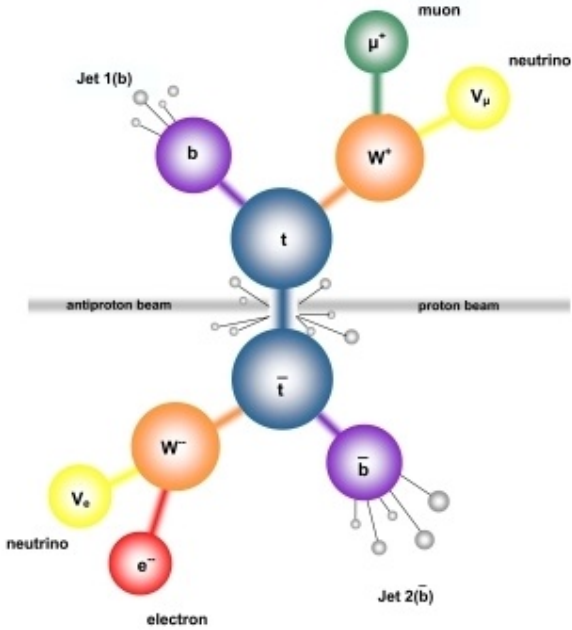


Fig. 3. Collision of proton and anti-proton resulting in the production of a pair of top quarks ( $t, \bar{t}$ ), which decay almost immediately to a pair of  $W$  particles ( $W^+, W^-$ ) and bottom quarks ( $b, \bar{b}$ ).  $W$  particles decay to leptons ( $\mu^+, e^-$ ) and their accompanying neutrinos ( $\nu_\mu, \nu_e$ ) while bottom quarks decay to jets of lower energy particles.

boson (1% of the sample).

While these backgrounds mimic the top quark's basic signature, they differ from top quark events in more subtle ways, e.g. the distribution of energy in the leptons or jets. Final selections, described in Sections 3 and 4, can further prune the data by exploiting these differences.

### 2.3.2. Pre-Selection for Higgs Boson Search

As in top quark mass measurement, pre-selection in Higgs boson search discards all events without the appropriate signature. In this case, the signature results from the Higgs boson's decay products, shown in Figure 4, which consist of two jets from bottom quarks and two jets from other quarks. As with the top quark, pre-selection for the Higgs boson has nearly perfect recall but poor precision, as this signature is not unique to the Higgs boson. Despite discarding more than 99% of the events, pre-selection leaves a pool of events still overwhelmingly dominated by background: only 1 in approximately 3500 events involves a Higgs boson. Though these backgrounds produce a similar signature, the energy of their jets will differ slightly, making it possible to suppress their influence during the final selection.

### 2.4. Analyzing Selected Events

The events that survive the final selection are used for data analysis, either to measure the mass of the top quark or to determine whether the Higgs boson was observed.

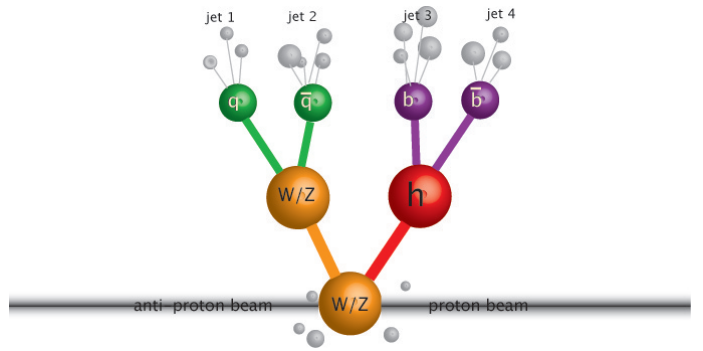


Fig. 4. Collision of proton and anti-proton resulting in a  $W$  or  $Z$  boson which radiates a Higgs boson  $h$ . The  $W$  or  $Z$  boson decays to two quarks ( $q, \bar{q}$ ), which are seen as jets, and the Higgs decays to two bottom quarks ( $b, \bar{b}$ ), also seen as jets.

In the former case, the top quark's mass can be measured by inferring the likely mass of the observed decay products in each event (21). In particular, probability density functions in top quark mass are obtained for every event. Combining these into a single joint probability density function and computing its expected value yields the final estimate of the top quark's mass (see Figure 5). Additionally, probabilities that the observed decay products are due to one of the major background processes are calculated and used to suppress the influence of background events in the mass measurement. Minimizing the uncertainty of this measurement is the primary goal of event selection for top quark mass measurement.

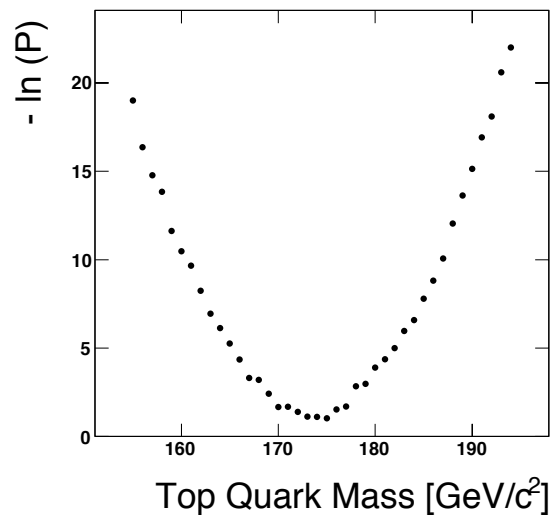


Fig. 5. The top quark's mass can be inferred from the observed decay products (21). An example measurement using simulated events shows a strong minimum in the negative log of the probability density near the simulated mass, 175  $\text{GeV}/c^2$ . The measured mass for this example is 173.6  $\text{GeV}/c^2$ .

In the latter case, the data is used to determine whether the presence of the Higgs boson can be confirmed. Since the Higgs boson occurs so rarely at the Tevatron, confirmation can occur only with powerful event selection and large

numbers of events. Confirmation is further hindered by a significant theoretical uncertainty in the nature of background events to the Higgs boson. However, even if there is not enough data to confirm the Higgs boson’s presence, the sensitivity of the selector can still be quantified. This is done using the *scale factor*, which measures how much more frequently than predicted by theory the Higgs boson must occur for it to be observed. The theoretical uncertainty in the background events yields a systematic uncertainty in the rate of Higgs boson events, and significantly increases the necessary scale factor. Minimizing the scale factor is the primary goal of event selection for Higgs boson search.

### 3. Supervised Learning for Event Selection

This section describes how final event selection can be performed with the aid of supervised learning methods which maximize classification accuracy. In a narrow class of problems, these are equivalent to a likelihood ratio test, which the Neyman-Pearson Lemma (7) suggests that are optimal. This approach is standard in the physics community and serves as a point of comparison for the optimization approach that will be described in Section 4.

Supervised methods can be used to train classifiers that separate signal from background. However, such classifiers cannot be trained on real accelerator data because the true labels for such events are unknown. Instead, they are trained with data drawn from simulators of the accelerator (22) and detector (23), which generate events and model the interaction of their decay products with the detector. These simulators are not “toys” but instead sophisticated models that capture our best current understanding of the underlying physical processes and have been extensively verified using data from previous accelerators (23).

Since the true mass of the top quark is not known, it is important that selectors for top quark mass measurement be robust across a range of likely masses. Hence, simulated events are generated using three likely mass values: 165, 175, and 185 giga-electron-volts per speed of light squared ( $\text{GeV}/c^2$ ). Each event is described using the following six features: 1) the mass of the system of two leptons, 2) the number of identified bottom quarks, 3) the imbalance of transverse momentum, indicating the presence of undetected neutrinos, 4) the total transverse energy of all decay products, 5) the minimum angle between a jet and the unbalanced transverse momentum, and 6) the minimum angle between a jet and a lepton. Figure 6 shows the distribution of values in the data set of simulated events, after pre-selection, for these six features.

The signal for the Higgs boson is more subtle than that of the top quark, and the backgrounds are both more numerous and less well understood theoretically. As a result, when searching for the Higgs boson, different features are used to describe each event. In particular, these features summarize the information contained in the decay products into three higher-level physics quantities: 1)  $C_W$ , which

indicates whether the decay products are consistent with a  $Wh$  event, 2)  $C_Z$ , which similarly indicates consistency with a  $Zh$  event, and 3)  $C_B$ , consistency with a background event. Figure 7 shows the distribution for simulated events after pre-selection for these three features.

Using these features to describe each simulated event, we can construct a training set by labeling each collision as signal or background. The remainder of this section describes two supervised learning approaches for training selectors with such data.

#### 3.1. Training Binary Classifiers

The simplest supervised learning approach to event selection involves training binary classifiers. This approach has proven successful on related event selection problems, using neural networks (13; 14) or support vector machines (15). In this article, we train feed-forward neural networks with backpropagation (17).

For top quark mass measurement, we use networks with six inputs, fourteen hidden nodes, and one output. These networks are fully connected, i.e., there is a link between every input and every hidden node and one between every hidden node and every output. In training, each event is labeled 1 if it is signal and 0 otherwise. In testing, an event is classified as signal if the network’s output is greater than a threshold  $t \in [0, 1]$ . Since we cannot quantify *a priori* the trade-off between precision and recall, we set  $t$  to the value that maximizes classification accuracy on the training set. To find this value, we sample the range  $[0, 1]$  at regular intervals of 0.025, computing the classification accuracy at each point.

For Higgs boson search, we use fully connected networks with three inputs, two layers of hidden nodes with four and three nodes respectively, and one output. In training, each signal event is labeled 1 and each background event 0. However, for testing, we do not apply a specific threshold to the output. Instead, the entire distribution of output values for all pre-selected events is used in measuring sensitivity to the presence of the Higgs boson. This process is described in more detail in Section 4.2.

#### 3.2. Training Multi-Class Classifiers

A potential disadvantage of the binary classification approach is that it gives all backgrounds the same label. Learning may be easier if the problem is cast as a multi-class classification task, where each type of background is treated as a separate class. To test this approach, we train a set of *one-against-all* (24) classifiers, each of which strives to distinguish a given class from all the others.

In the case of top quark mass measurement, there are six classes: one signal class and five background classes, as described in Section 2.3.1. We train six binary classifiers, each of which uses the same network topology described above. When training the  $k$ th classifier, each event is la-

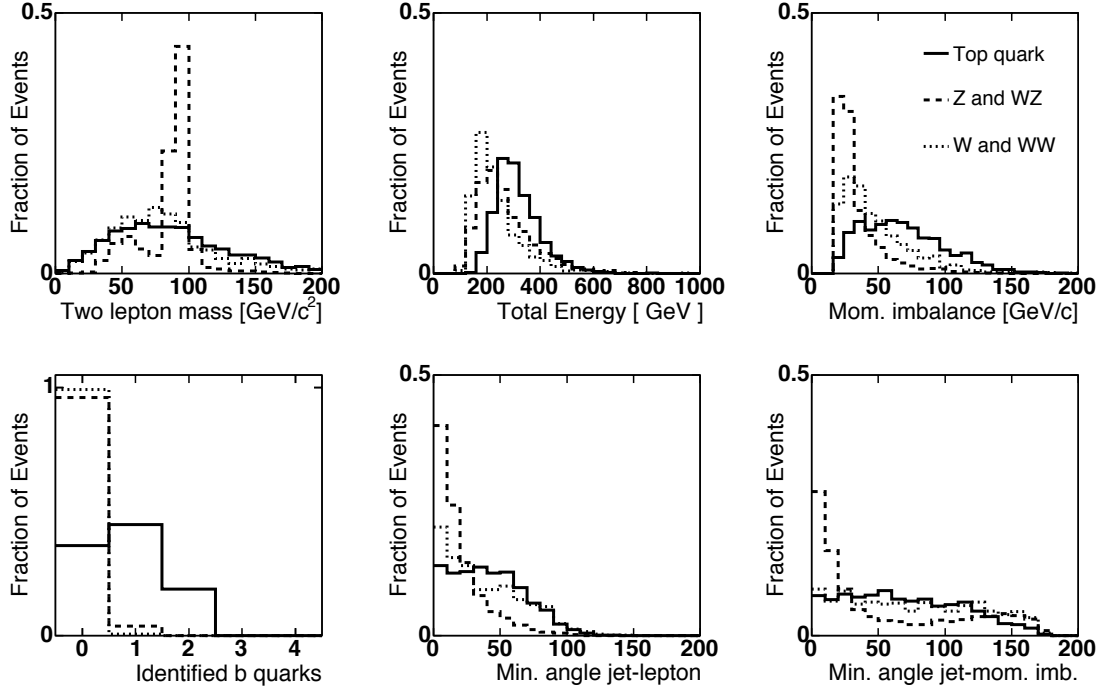


Fig. 6. The distribution in the training set, after pre-selection, of feature values for simulated events used to train selectors for top quark mass measurement. Clockwise from top left: the mass of the two leptons shows a strong peak for  $Z$  background; the total energy in the event is largest for the decays of the top quarks; the momentum imbalance is largest for top quark decay which produces undetectable neutrinos; the minimum angle between jets and leptons, or the minimum angle between jets and the momentum imbalance helps distinguish top quark decays from background; only top quark decays contain real bottom quarks, a powerful discriminator when identified.

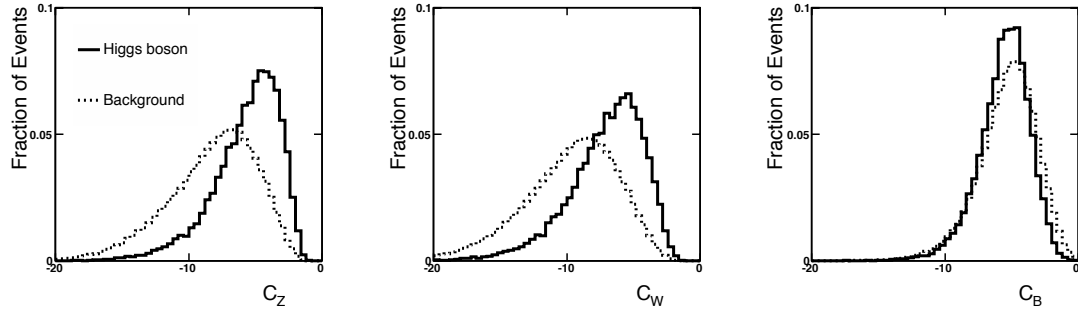


Fig. 7. The distribution in the training set, after pre-selection, of feature values for simulated events used to train selectors for Higgs boson search. From left,  $C_Z$ , which indicates whether the decay products are consistent with a  $Zh$  event,  $C_W$ , which similarly indicates consistency with a  $Wh$  event, and  $C_B$ , consistency with a background event.

beled 1 if it is in class  $k$  and 0 otherwise. Note that one of these classifiers, that which distinguishes signal from all five background classes, is identical to the binary classifier described in Section 3.1. In testing, an event's classification corresponds to the network with the highest output. Note that the multi-class approach is not applicable to Higgs boson search, which has a single dominant background.

#### 4. Optimization Methods for Event Selection

The supervised approach described above is most effective in the narrow class of problems in which the classification accuracy of the event selector is closely correlated

with the quality of the resulting data analysis. In this case, the Neyman-Pearson Lemma argues that such techniques are optimal.

However, top quark mass measurement and Higgs boson search exemplify a broader class of problems where higher classification accuracy does not necessarily result in better analysis. Instead, the analysis is more sensitive to the presence of some background events than others, in ways that are difficult to predict *a priori*. Therefore, selectors that maximize classification accuracy may perform worse than those that 1) increase the quantity of signal by tolerating harmless background, 2) reduce the quantity of signal to eliminate disruptive background, or 3) minimize the

impact of systematic uncertainties.

Since the costs of misclassification are not always the same, one way to address this challenge would be to treat event selection as a *cost-sensitive* supervised learning problem (25). In such problems, the cost of misclassification can vary depending on either the true class, the incorrectly predicted class, or both. As a result, the goal is not to minimize classification error but instead to minimize the total cost of misclassification. Specialized learning methods exist to tackle such problems. For example, the MetaCost algorithm (26) relabels training data such that a classifier trained to minimize classification error on the relabeled data will minimize the costs of misprediction. However, MetaCost and other methods like it are not applicable to the event selection problem because they require as input a cost matrix  $C$ , where  $C(i, j)$  is the cost of predicting class  $i$  when the true class is  $j$ . In event selection, this cost matrix is not known *a priori*. Furthermore, even if it were known, it is unlikely that cost-sensitive methods would substantially improve performance, since the disruption caused by any particular background event can depend on subtle features that are poorly correlated with that event's true type.

While cost-sensitive methods are not applicable to event selection, more general-purpose optimization methods are. Such methods search the space of possible selectors for one that maximizes a given fitness function. Previous studies (27) have shown that optimization techniques can compete with supervised methods in similar event selection problems by optimizing criteria related to classification accuracy. However, those studies used manually defined criteria that the designers hoped would yield effective selectors. In this article, we present a new approach that directly optimizes selectors for their true purpose: maximizing either the precision of the top quark mass measurement or sensitivity to the presence of the Higgs boson.

In the remainder of this section, we first define fitness functions for each event selection problem and then describe several optimization approaches that make use of these functions.

#### 4.1. Fitness Function for Top Quark Mass Measurement

The goal of event selection for top quark mass measurement is to obtain a measurement with maximal precision. Hence, the fitness of any given event selector is the precision of the mass measurement obtained using the subset of simulated events it selects. This choice of fitness function naturally accounts for the disruptive effects of the background events that a selector allows, and therefore gives optimization methods the possibility to realize significant performance gains.

All background events that survive pre-selection are roughly consistent with top quark events. Disruptive background events are those that are very consistent with top quark events of a specific top quark mass, and inconsistent with events at other masses. Though they contain no in-

formation about the true top quark mass, they sway the measured mass due to their apparent precision.

The most precise measurements are those with the smallest statistical uncertainty, which we measure by calculating the standard deviation of the mass estimates the selector produces on a series of 1000 independent trials at each of the three likely top quark masses. In each trial, we randomly select events, with replacement, from the pre-selected training set and feed them to the selector. The number of events in each trial was chosen to approximately equal the number of events accumulated by the CDF II detector. The events that survive selection are used to estimate the top quark mass, as described in Section 2.4. The standard deviation of these estimates reflects the statistical uncertainty of mass measurements produced by that selector.

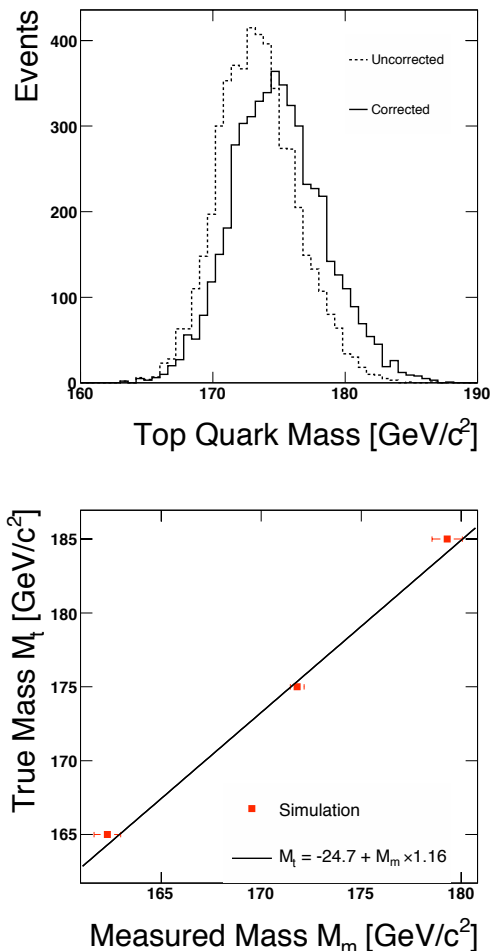


Fig. 8. At top, a histogram of top quark masses measured in 1000 independent trials where the true mass is 175 GeV/c<sup>2</sup>, both before and after bias correction. The fitness function for top quark mass measurement is negatively proportional to the width of the histogram after bias correction. At bottom, bias correction for top mass measurement using simulated events at three top quark masses (squares). A linear correction function (solid line) is fit to the measured masses.

Note that it is not necessary to optimize event selectors for accuracy because the mass measurement is calibrated for accuracy using simulated events, a process known as *bias*

correction (21). As shown in Figure 8, we can correct for any discrepancy between the measured mass  $M_m$  and the true mass  $M_t$  using simple linear regression, with the actual mass of the simulated events used as labels for training. Hence, background events that do not lower precision are harmless even if they introduce bias. The best selector is that which produces the most precise mass measurements, regardless of the resulting bias. The fitness function for the top quark mass measurement  $F_{M_t}$  is then the standard deviation of bias-corrected mass estimates:

$$F_{M_t} = \sqrt{\frac{1}{N} \sum_{i=1}^N (M_b^i - \bar{M}_b)^2} \quad (1)$$

where  $M_b^i$  is the  $i$ th mass measurement after bias correction and  $\bar{M}_b$  is the average of the  $N$  bias-corrected mass measurements.

#### 4.2. Fitness Function for Higgs Boson Search

The goal of event selection for Higgs boson search is to find a selector with maximal sensitivity to the presence of the Higgs boson. As mentioned in Section 2.4, this can be quantified by computing the selector’s scale factor, which measures how much the Higgs boson’s rate of production would need to increase for it to be observed or statistically excluded.

More specifically, the scale factor is the median 95% expected confidence level upper limit on the Higgs boson production rate, relative to the theoretically predicted rate. To compute confidence level limits, we conduct 4000 independent trials; each trial uses randomly selected simulated events, as with the fitness function for top quark mass measurement.

For each trial, we compare the distribution of the selector’s outputs for the simulated events to distributions we would expect for this selector given two hypotheses: 1) that the data contains only background events and 2) that the data contains both background and Higgs boson events. The greater the difference between the distributions for the two hypotheses, the more confident the resulting statistical limit will be (3). Increasing the scale factor, and therefore the rate of Higgs boson events, results in a larger difference between the two hypotheses and a more confident limit on the Higgs boson rate. We increase the scale factor of the Higgs boson production rate until the two distributions are different enough to allow us to set a 95% confidence level upper limit on the Higgs boson production rate, averaged over the 4000 trials. Figure 9 illustrates this process.

This scale factor is the final physics result of a Higgs search, and includes such effects as theoretical uncertainty in the simulation of the background events that can weaken the search. Choice of the scale factor as the fitness function allows optimization methods the opportunity to select events whose uncertainty will have minimal impact.

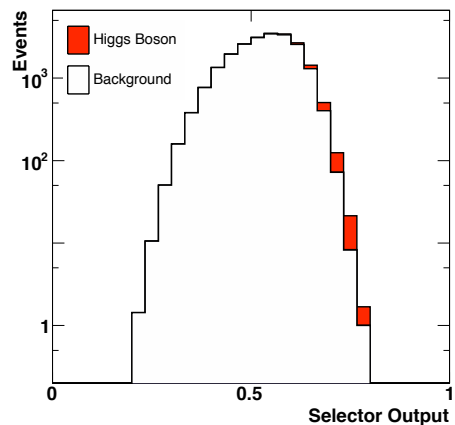
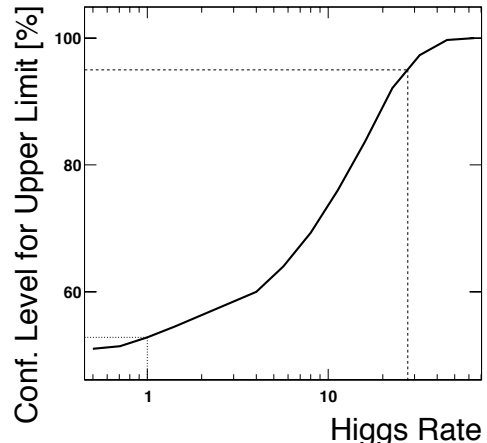


Fig. 9. At top, the confidence level of the statistical upper limit improves as the Higgs rate is increased; at the theoretically predicted rate (dotted line) the confidence is low, so the *scale factor* (dashed line) measures the relative increase in the Higgs rate necessary to achieve a 95% level. At bottom, example output of an event selector for Higgs boson search with simulated Higgs events (solid histogram) stacked on top of simulated background events (empty histogram). The Higgs production rate has been increased by the scale factor such that it is large enough to be excluded at 95% confidence level.

#### 4.3. Optimizing Binary Classifiers

The simplest way to use these fitness functions to improve event selection is in optimizing the threshold  $t$  of the binary classifier. As before, we sample the range  $[0, 1]$  at regular intervals of 0.025. However, at each point we compute the value of the fitness function, instead of the classification accuracy. Note that this approach is applicable only to top quark mass measurement, since the binary classifiers for Higgs boson search do not use a threshold.

#### 4.4. Optimizing Multi-Class Classifiers

Optimizing  $t$  could improve performance by effectively balancing the trade-off between precision and recall. However, it is still suboptimal because it treats all background



types equally. A selector that optimizes the output of the multi-class classifier (made up of multiple binary classifiers) could perform much better: by distinguishing between different background types, it could favor harmless events and discard disruptive ones. This approach is also applicable only to top quark mass measurement, since Higgs boson search has a single dominant background.

The one-against-all approach to multi-class classification does not have thresholds to tune. Nonetheless, its performance can be improved using stochastic optimization techniques. Instead of directly using the classifiers for selection, we use their classifications as input to a selector trained to maximize one of our fitness functions. This selector is also a neural network but its internal structure and weights are determined not by backpropagation but by a stochastic optimization method called NeuroEvolution of Augmenting Topologies (NEAT) (16). While many other optimization methods could be used in its place, we chose NEAT for event selection because of its previous empirical success on difficult optimization tasks (16; 28). Here we provide a brief overview of the NEAT method; a complete description is provided by Stanley and Miikkulainen (16).

Like other neuroevolutionary methods, NEAT uses evolutionary computation to train neural networks. In a typical neuroevolutionary system (29), the weights of a neural network are strung together to form an individual genome. A population of such genomes is then evolved by evaluating each one and selectively reproducing the fittest individuals through crossover and mutation. Most neuroevolutionary systems require the designer to manually determine the network’s topology (i.e. how many hidden nodes there are and how they are connected). By contrast, NEAT automatically evolves the topology to fit the given problem.

NEAT begins with a uniform population of simple networks with no hidden nodes and inputs connected directly to outputs. In addition to standard weight mutations, two special mutation operators incrementally introduce new structure to the population. Figure 10 depicts these operators, which add hidden nodes and links to the network. Only those structural mutations that improve performance tend to survive; in this way, NEAT searches through a minimal number of weight dimensions and finds the appropriate level of complexity for the problem.

#### 4.5. Optimizing Selectors Without Supervised Learning

A more radical departure from the traditional approach to training event selectors is to entirely eliminate the use of supervised methods. In this approach, the inputs to the NEAT selector are not the outputs of the one-against-all classifiers but instead the original features that served as inputs to those classifiers. As a result, training classifiers is no longer necessary. Instead, we treat event selection purely as an optimization problem and rely on NEAT to find a selector that maximizes the appropriate fitness function. Unlike the optimization approaches described above, this ap-

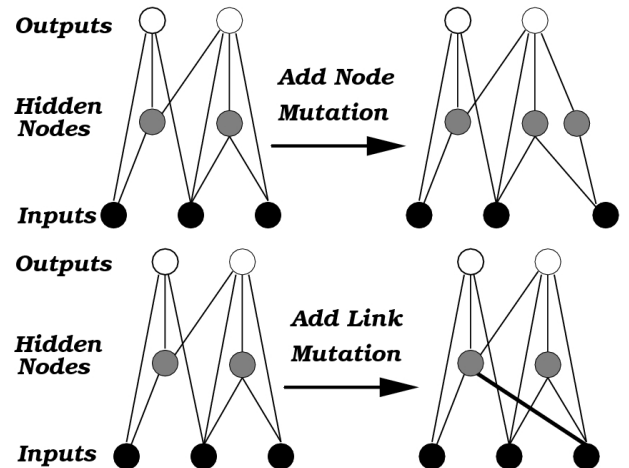


Fig. 10. Examples of NEAT’s structural mutation operators. At top, a new hidden node is added to the network by splitting an existing link in two. At bottom, a new link, shown with a thicker black line, is added to connect two existing nodes.

proach is applicable to both top quark mass measurement and Higgs boson search.

The networks have one input for each event feature (six for top quark mass measurement and three for Higgs boson search) and one output. These networks use the threshold  $t = 0.5$  in both training and testing. It is no longer necessary to tune  $t$  since NEAT evolves networks that are optimized for a fixed value of  $t$ .

Finding a good selection in this manner is challenging in part because of the size of the search space. The set of possible selections is the power set of the events. Hence, given  $n$  events, there are  $2^n$  possible selections. Nonetheless, directly searching for selectors that minimize mass measurement uncertainty yields much better performance than maximizing classification accuracy, as the results in the following sections confirm.

## 5. Comparative Results for Top Quark Mass Measurement

To assess the efficacy of the methods presented in Sections 3 and 4, we evaluated each one on the top quark mass measurement problem, averaging performance over ten independent runs, using 10,000 simulated events. These runs were conducted using ten-fold cross validation: in each run, 75% of the events are selected at random for training and the remaining 25% reserved for testing.

### 5.1. Supervised Learning Results

Figure 11 shows the classification accuracy on training data for networks trained with backpropagation on simulated pre-selected events, averaged over ten independent runs. As described in Section 3.2, each network is trained to identify one class of events. Binary classification uses only the network trained to identify top quark events, while multi-class classification uses all six networks. The fully-

connected, feed-forward networks have six inputs, fourteen hidden nodes, one output, and are trained with a learning rate of 0.001 and a momentum rate of 0.5. Accuracy during training is measured using a threshold  $t = 0.5$ .

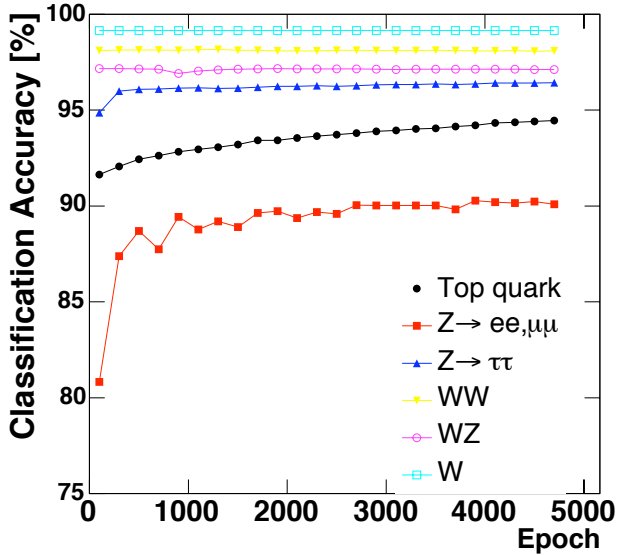


Fig. 11. Classification accuracy on training data for the six networks, each trained to recognize the signal (top quark) or one of the five backgrounds. Accuracy is averaged over all classes, weighted by the expected contributions. Networks trained to recognize rare backgrounds therefore have high accuracy. The signal network is used in the binary classification case; all six networks are used in the *one-against-all* multi-class approach.

On the data reserved for testing, the binary classifier had an average classification accuracy of  $93 \pm 1\%$ . The multi-class classifier identified the correct class with an accuracy of  $83 \pm 1\%$ . If the multi-class classifier is not penalized for labeling backgrounds with the wrong background class, its accuracy improves to  $91 \pm 1\%$ . Table 1 shows the precision and recall in training and testing for both the binary and multi-class classifiers. The similarity in performance between training and testing indicates that the classifiers are not overfit to the training data. Binary and multi-classifiers give mass measurements with an average uncertainty relative to the heuristic of  $0.91 \pm 0.04$  and  $0.90 \pm 0.05$ , respectively.

	Precision	Recall
Binary	0.84 (0.87)	0.82 (0.79)
Multi-Class	0.71 (0.73)	0.89 (0.88)
NEAT	0.15 (0.16)	0.97 (0.96)

Table 1  
Precision and recall for the binary classification, multi-class classification, and NEAT (with features) top-quark selectors. Performance in training is shown first, following by performance in testing in parentheses.

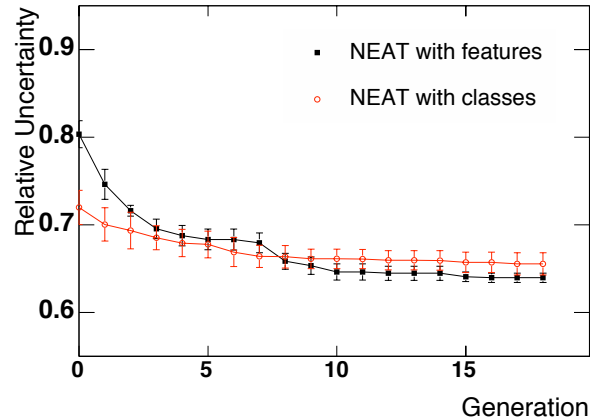


Fig. 12. Uncertainty in top quark mass relative to the heuristic in training data for the best network in each NEAT generation.

## 5.2. Optimization Results

If the threshold of the binary classifier is selected to minimize mass measurement uncertainty instead of maximizing classification accuracy, as described in Section 4.3, the resulting selectors allow for mass measurements with substantially lower average uncertainty,  $0.82 \pm 0.04$  relative to the heuristic.

Using NEAT to perform stochastic optimization yields even more precise measurements. Figure 12 shows mass uncertainty on the training set for the best network in each generation trained with NEAT. It compares the performance of NEAT optimizing multi-class classifiers to its performance optimizing directly on the features, without the help of supervised methods. The results are averaged over ten runs for each method. In testing, the average mass uncertainty of the final generation champions relative to the heuristic was  $0.66 \pm 0.02$  and  $0.64 \pm 0.02$  for the two approaches, respectively.

Table 1 also shows precision and recall in training and testing for the NEAT with features approach. The high recall and low precision suggests that the optimization approach improves the precision of the mass measurement by keeping more top quark events (yielding higher recall) but allowing more harmless background events (yielding lower precision).

Figure 13 shows the expected top quark mass uncertainty for each network evaluated during a typical NEAT run, as well as the fraction of selected events that are signal. Though NEAT finds networks with a signal fraction higher than 0.8, the networks with the lowest top quark mass uncertainty have a starkly lower signal fraction, less than 0.4. This result confirms our hypothesis that signal fraction is not well correlated with uncertainty in the resulting measurement, as the best performing selectors tolerate substantial harmless background.

Figure 14 summarizes the performance on testing data of all the machine learning methods we employed and com-

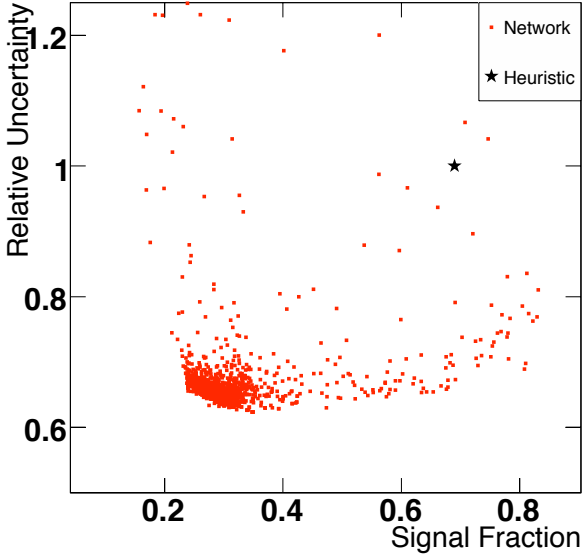


Fig. 13. Uncertainty in top quark mass relative to the heuristic, and signal fraction for each network evaluated during a typical NEAT run (red squares), and for the heuristic selection (black star).

compares it to the performance of a heuristic selector designed manually by physicists. Student's  $t$ -tests confirm with greater than 98% confidence the statistical significance of the differences between 1) the heuristic selector and each learning method, 2) each supervised method and each optimization method, and 3) the optimized binary classifier and each NEAT method.

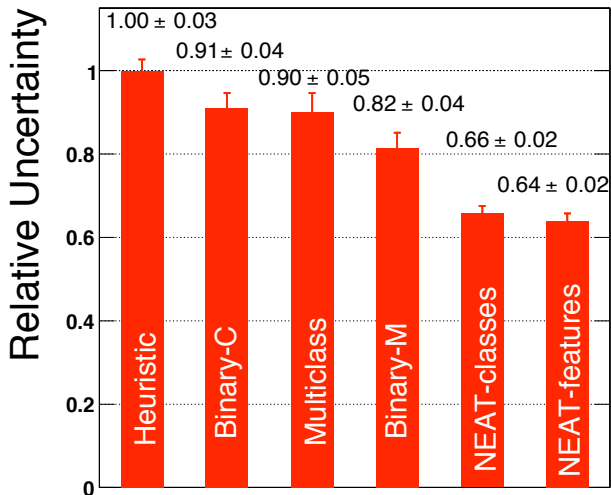


Fig. 14. Average mass measurement uncertainty on testing data for the heuristic selector, binary classifiers with  $t$  optimized for classification accuracy (Binary-C), or mass uncertainty (Binary-M), multi-class classifiers (Multiclass), NEAT with multi-class classifiers as inputs (NEAT-classes), and NEAT with the original features as inputs (NEAT-features).

## 6. Comparative Results for Higgs Boson Search

We also evaluated the relative performance of the supervised learning and optimization strategies on the problem of Higgs boson search. As before, each run was conducted using ten-fold cross validation with a 75%/25% split between training and testing.

### 6.1. Supervised Learning Results

Recall that the multi-class classification approach is not applicable to Higgs boson search since there is only a single dominant background. Thus, we tested only the binary classification approach. Figure 15 shows the classification accuracy on training data for networks trained with backpropagation on simulated pre-selected events, averaged over ten independent runs. We use a fully-connected, feed-forward neural network with three inputs, two layers of hidden nodes with four and three nodes respectively, and one output node. The network is trained using a learning rate of 0.001 and a momentum rate of 0.5.

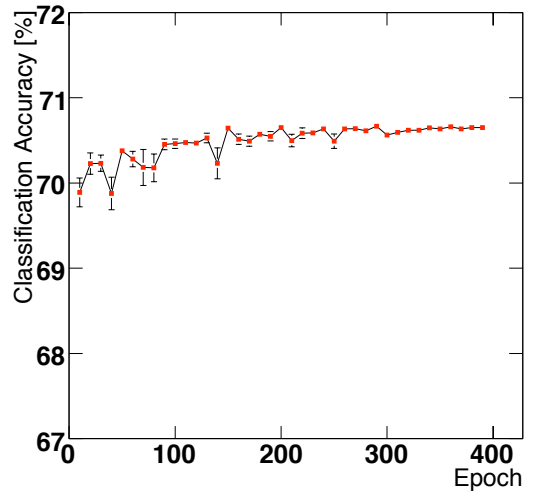


Fig. 15. Classification accuracy on training data averaged over ten runs for a network trained to distinguish the signal for a Higgs boson from the background process.

The classification accuracy does not improve dramatically during training due to the nature of the features, which themselves encapsulate a significant fraction of the classification information through their description of the consistency of each event with one of the two signal or the one background hypotheses. Table 2 shows precision and recall in training and testing. Before classification, the background dominates the pre-selected events; signal to background ratio is approximately 1:3500. Therefore, even relatively high classification accuracy and corresponding impressive reduction in background rate gives a fairly low precision, yielding a signal to background ratio of approx-

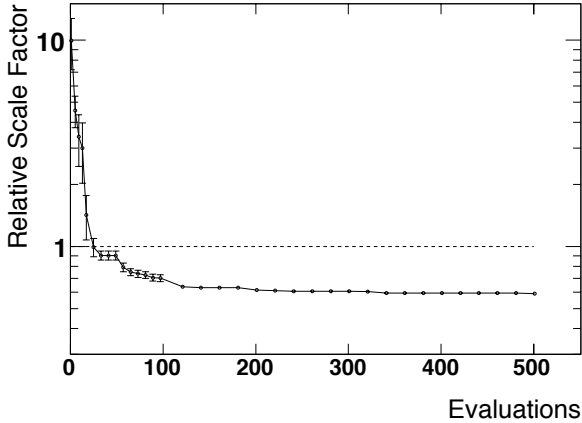


Fig. 16. Average scale factor on training data relative to the heuristic selector (dashed line) for the best network found so far by NEAT.

imately 1:5. The average resulting scale factor was  $0.92 \pm 0.04$ , relative to the heuristic.

	Precision	Recall
Binary	0.21 (0.22)	0.75 (0.76)

Table 2

Precision and recall for the binary Higgs boson classifier. Performance in training is shown first, following by performance in testing in parentheses.

## 6.2. Optimization Results

Since optimizing classifiers is not applicable in Higgs boson search, we tested only the pure optimization approach, where NEAT trains a selector that receives each event’s features directly. Figure 16 shows the scale factor on the training set for the best networks discovered by NEAT. For reasons of convenience, these experiments use a slightly different version of NEAT called rtNEAT (20). rtNEAT is a *steady-state* evolutionary method, which means it does not have explicit generations but instead one population that changes gradually over time. Thus, the results in this figure are plotted against total fitness evaluations rather than generations. As before, the results are averaged over ten runs for each method. In testing, the average scale factor of the final champions relative to the heuristic was  $0.67 \pm 0.03$ .

We do not present precision and recall results for this case because, when optimization methods are applied to Higgs boson search, such metrics are not meaningful and do not have a clear interpretation. The reason, as explained in Section 3.1, is that the output of the network is never interpreted as a class label and thus no threshold is applied to it; instead we merely seek the network that maps events to  $[0, 1]$  in a way that minimizes the scale factor.

Figure 17 summarizes the performance on testing data of both learning methods that were applied to event selection for Higgs boson search and compares it to the performance of a heuristic selector. Students t-tests confirm with

greater than 99% confidence the statistical significance of the differences between the NEAT method and both the heuristic and backpropagation methods.

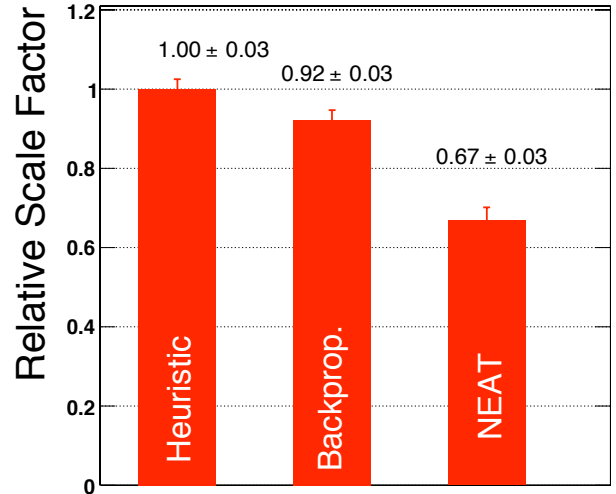


Fig. 17. Average scale factor on testing data for the heuristic selector, backpropagation and NEAT. All results are shown relative to the heuristic selector.

## 7. Application to Tevatron Events

The top quark was first observed in 1995 at the Fermilab Tevatron, which remains the sole accelerator with enough energy to produce it for direct study. The top quark’s mass is extraordinary on the scale of fundamental particles, nearly 40 times larger than the second most massive quark and comparable to the mass of a gold atom. Its enormous mass remains a puzzle from a theoretical standpoint, so physicists have focused great effort on probing the precious few top quarks collected in the five years since the CDF II detector became operational.

Consequently, precise measurements of the top quark’s mass are a chief scientific priority for the Fermilab program. Stringent efforts are made by the accelerator and detector teams to maximize the size of the accumulated data sample in order to reduce statistical errors, and improvement in techniques for extracting maximal precision from the collected data are the subject of much active research (31). Typically, the measurements are made and updated yearly, as the dataset grows. In this section we describe how our work has contributed to this effort. In particular, we detail the successful application of a NEAT selector to a re-analysis of the data collected by the CDF II detector.

Prior to this work, event selection for top quark mass measurement was done using the heuristic selector (19) described in this article. In analyzing the selected events, the measurement was dominated by statistical uncertainties due to the sparsity of observed top quarks. Our results demonstrating the potential of stochastic optimiza-

tion methods for event selection (33) convinced physicists that this approach could significantly increase the quantity of signal in the final selection and thus improve the precision of the resulting measurement.

The data sample used consists of events collected between March 2002 and March 2007. From more than five years of accumulated collisions, 642 events satisfy the pre-selection, of which we expect 131.6 on average to be signal. A neural network trained with NEAT in the manner described in Section 4.5 was used to produce a final selection containing 344 events, of which we expect 121.8 on average to be signal. Since the selector was trained on simulated data, it is critical that the real data is well described by the simulated data. Figure 18, which shows neural network output for the pre-selected data as well as for the simulated events, confirms that this is the case.

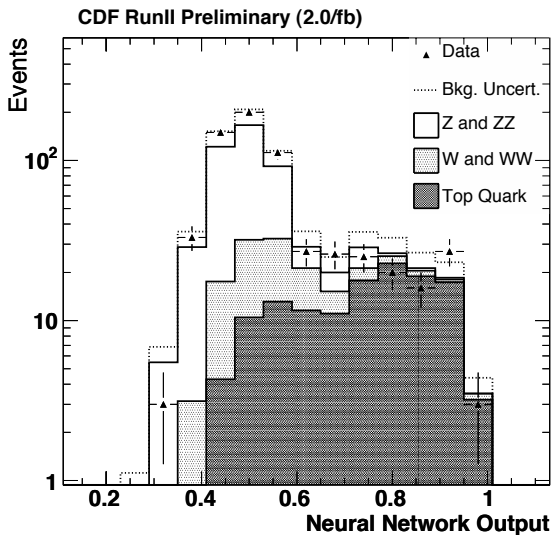


Fig. 18. The output of the final NEAT network evaluated on the collected data from the CDF II detector at the Tevatron (black triangles), compared to the output on simulated data (stacked solid histograms) (30).

Using this final selection, a new mass measurement was obtained (30):

$$M_{top} = 171.2 \pm 2.7(\text{statistical}) \pm 2.9(\text{systematic})$$

This is the world's most precise top quark mass measurement using this characteristic signature, by nearly a factor of two (32) (see Figure 19). It is the first measurement of this type not dominated by statistical uncertainties. Obtaining a similar reduction in uncertainty would otherwise require producing many more collisions at great expense.

Following this measurement, competing measurements of the top quark mass revisited the design of the heuristic selection to more closely follow the selection derived with our optimization. Thus, this new methodology has changed the way physicists devise heuristic selectors.

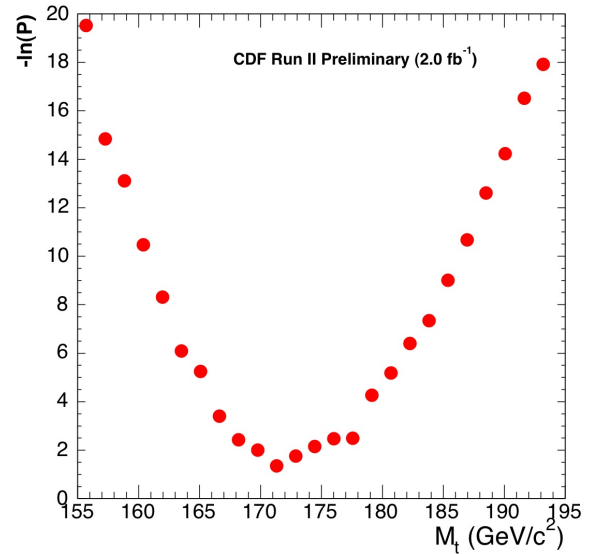


Fig. 19. Negative log of the probability density in top quark mass for the data collected from the CDF II detector at the Tevatron (30), showing a minimum around  $M_{top} = 171.2 \text{ GeV}/c^2$  and a width which corresponds to a statistical error of  $2.7 \text{ GeV}/c^2$ .

## 8. Discussion

The results presented in Sections 5 and 6 confirm the conclusion of earlier work (13; 14; 15) that machine learning methods can substantially outperform heuristic event selectors. However, previous results demonstrated only that learned selectors had higher classification accuracy, while these results directly verify that they can improve the quality of the resulting analysis. More importantly, these results confirm the advantage of treating event selection as an optimization rather than a supervised learning problem.

In top quark mass measurement, the chief difficulty lies in determining which backgrounds will be harmful to the precision of the resulting measurement. The heuristic selector (19) strives to eliminate as much background as possible, without regard to how disruptive any particular event may be. It explicitly removes events which have decay products consistent with the largest background,  $Z$  boson with gluons. The heuristic performs well in this regard, as the final selection has a signal fraction of approximately  $2/3$ . However, it also removes a lot of signal. By contrast, the selector evolved by NEAT allows approximately 25% more signal into the final sample than the heuristic. It is able to do so by tolerating background that is minimally disruptive to the measurement. As Figure 13 makes clear, the best performing selectors are those that tolerate substantial quantities of harmless background.

The lack of correlation between the signal fraction and the quality of the resulting analysis also explains why optimization methods outperform supervised methods. By simply maximizing classification accuracy, neither the binary nor multi-class classification approaches consider which background is worth tolerating. Even the most naïve optimization approach, which merely tunes the thresh-

old  $t$  of a binary classifier, outperforms the best purely supervised approach. Furthermore, the performance of NEAT when trained directly on the features suggests that supervised methods are not necessary for this task.

In Higgs boson search, event selection is made difficult by event-by-event differences in the theoretical uncertainties in the description of the physical process that causes background events. In this case, the heuristic selector makes the reasonable approximation that the three input features can be treated as probabilities that the event is Higgs boson signal ( $Wh$  or  $Zh$ ) or a background event. The heuristic forms a simple likelihood ratio of the signal and background features, which tends to give large values to signal events and smaller values to background events. This heuristic is suboptimal because it does not suppress events whose uncertainties weaken the final limit.

Supervised methods can marginally improve classification accuracy over these raw features, which results in a small but significant reduction in the resulting scale factor. However, as with top quark mass measurement, directly optimizing the selector yields by far the greatest performance.

In both tasks, the performance improvement is not only significant but also substantial. In top quark mass measurement, NEAT produces 29% lower uncertainty than the supervised approach. In Higgs boson search, NEAT produces a 37% lower scale factor than the supervised approach. Obtaining these improvements would otherwise require accumulating 66% or 88% more events, respectively, costing literally tens of millions of dollars and hundreds of person-years.

These savings are not hypothetical, since this new methodology for training event selectors has been applied, not only to simulated data, but to real data collected on the Fermilab Tevatron accelerator. As described in Section 7, using the best selector trained by NEAT results in the most precise top quark mass measurement of this type to date, by nearly a factor of two. Consequently, this novel approach to event selection has directly aided the progress of high energy physics, by contributing substantially to our knowledge of the top quark’s mass and our understanding of the larger questions upon which it sheds light.

## 9. Future Work

We believe that the intersection of machine learning and high energy physics is a highly promising but under-explored research area. Thus, we hope that the results presented in this article will mark only the beginning of a long and fruitful effort to bridge the gap between these two disciplines. Such research can be a boon to both fields, by providing machine learning researchers with a challenging, realistic proving ground for their methods and arming high energy physicists with the tools to make the most of their data. In this section, we outline some of our plans for future work.

### 9.1. Methods

Many other machine learning techniques besides those tested here could potentially aid event selection for top quark mass measurement and Higgs boson search. In the future, we aim to further broaden our search for the methods most suitable to these tasks. While many other traditional supervised methods, such as decision trees (34), Bayesian classifiers (35), nearest neighbor methods (36), or support vector machines (37), could be used in place of backpropagation, there is not much reason to believe they would significantly improve performance, as previous studies have shown little performance difference between such methods on related event selection problems (15).

However, it is entirely plausible that other optimization methods could fare better than NEAT. The full representation-learning power of NEAT may not be necessary on these tasks, in which case simpler methods such as hill climbing or simulated annealing (38) could do just as well. Other neuroevolutionary methods such as CoSyNE (39) or recently developed estimation of distribution algorithms (40) are also potential contenders.

Another possibility for improving performance lies in the area of *feature selection*, wherein a learning algorithm seeks the most useful subset of features describing each example. Since too few features doom the learner to sub-optimal performance and too many features can lead to intractable dimensionality, feature selection can be a critical issue. Since event selection methods in high energy physics have so far relied on a manually selected set of features, the potential exists for performance improvement via automatic feature selection. However, successfully applying it in this domain is not a trivial task. In fact, our preliminary experiments have shown that making additional features available to the learner does not improve performance, suggesting the current features selected by expert physicists are already quite good. Nonetheless, a careful study of feature selection in this application area is an important subject for future work. There are many feature selection methods for supervised learning (41), but only a few exist for optimization techniques. However, FS-NEAT (42), an extension to NEAT that automates the process of feature selection, could be particularly applicable.

Whereas feature selection aims to pick the right features from a given set, *feature extraction* or *feature construction* (43; 44) aims to build new, more useful features from those provided. This strategy could also yield significant performance dividends. However, it may prove just as challenging as feature selection. In fact, the optimization results presented in Section 5.2 (see Figure 12) show that NEAT does not perform better using features extracted by supervised learning rather than using the original features.

Finally, we plan to explore the use of *structured prediction* methods (45; 46) for event selection. These methods, which in recent years have proven effective in a wide range of tasks from natural language processing to computational

biology (47), represent a new hope for supervised methods in event selection. Unlike traditional supervised techniques, structured prediction methods can minimize arbitrary cost functions that consider dependencies between examples. Hence, such methods could directly maximize the precision of top quark mass measurements or sensitivity to the Higgs boson, while still fully exploiting the availability of labeled training data.

## 9.2. Applications

As described in Section 7, the methodology introduced in this article has already been applied to real data from the Tevatron to obtain a new top quark mass measurement. In the future, we hope to similarly employ the best selectors trained for Higgs boson search to aid final determinations of whether the Higgs boson was observed at the Tevatron.

However, the future of high energy physics lies not with the Tevatron but with the Large Hadron Collider (LHC), shown in Figure 20, which is currently under construction near Geneva, Switzerland (48). First collisions at the LHC will be in 2009 at a center-of-mass energy of 14 TeV, nearly an order of magnitude greater than the Tevatron. In addition, the rate of collisions will be two orders of magnitude higher. The resulting torrent of data will pose a plethora of new machine learning challenges. For example, event selection for top quark mass measurement at the LHC will be very different than at the Tevatron, since the LHC's higher energy levels are expected to produce top quark collisions much more frequently. In addition, collisions at the LHC will have enough energy to directly observe the Higgs boson, though doing so will require highly effective event selection.

More generally, the scale of high energy physics demands fast, intelligent, and automatic processing of enormous quantities of data. The potential for connections between machine learning and high energy physics stretch beyond the problem of event selection, to reconstruction and classification of the decay products inside each event. In the future, we hope to explore such topics as well.

## 10. Conclusion

This article describes the use of machine learning methods to aid the process of selecting events at high energy accelerators for the purpose of studying the fundamental nature of matter and its interactions. First, we apply supervised learning methods, which have succeeded previously in similar tasks. Second, we present a new approach that uses stochastic optimization techniques to directly search for selectors that maximize either the precision of top quark mass measurements or sensitivity to the presence of the Higgs boson. Empirical results confirm that stochastically optimized selectors result in substantially better analyses and therefore more powerful physical insight. This article also describes a case study in which the best selector is ap-

plied to real accelerator events from the Fermilab Tevatron, resulting in the most precise top quark mass measurement of this type to date and hence contributing significantly to the progress of high energy physics.

## Acknowledgments

Thanks to Peter Stone, Risto Miikkulainen, Ken Stanley, Razvan Bunescu, Misha Bilenko, and Gwenn Englebienne for their comments and suggestions on this research.

## References

- [1] S. Weinberg. A Model of Leptons *Physics Review Letters* 19:1264, 1967.
- [2] S.L. Glashow. Partial Symmetries of Weak Interactions *Nuclear Physics* 22:579, 1961.
- [3] W.-M. Yao et al. Review of Particle Physics *Journal of Physics, G* 33:1, 2006.
- [4] F. Abe *et al.* Observation of top quark production in  $p\bar{p}$  collisions with the collider detector at FermiLab. *Physical Review Letters*, 74:2626, 1995.
- [5] B. Abott *et al.* Observation of the top quark. *Physical Review Letters*, 74:2632, 1995.
- [6] M. Hashimoto, M. Tanabashi, and K. Yamawaki. Top mode standard model with extra dimensions. *Physical Review D*, 64:056003, 2001.
- [7] J. Neyman, E. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London*, A231: 289, 1933.
- [8] S. Heinemeyer. Impact of a precise top mass measurement. *Journal of High Energy Physics*, 0309:075, 2003.
- [9] The LEP Collaboration. Combination of preliminary electroweak measurements and constraints on the standard model. *CERN-PH-EP/2004-069*, 2004.
- [10] V. Miransky, M. Tanabashi, and K. Yamawaki. Is the  $t$  quark responsible for the mass of the  $W$  and  $Z$  bosons? *Modern Physics Letters A*, 4:1043, 1989.
- [11] P. W. Higgs. Spontaneous Symmetry Breakdown without Massless Bosons *Physics Review* 145:1156, 1966.
- [12] M.M. Kado and C.G. Tully. Search for the Standard Model Higgs Boson at LEP *Physics Letters* , 565:61, 2002.
- [13] V.M. Abazov *et al.* Search for single top production at DZero using neural networks. *Physics Letters B*, 517:282, 2001.
- [14] D. Acosta *et al.* Measurement of the cross section for  $t\bar{t}$  production in  $p\bar{p}$  collisions using the kinematics of lepton + jets events. *Physical Review D*, 52:052003, 2005.
- [15] D.O. Whiteson and N.A. Naumann. Support vector regression as a signal discriminator in high energy physics. *Neurocomputing*, 55:251, 2003.
- [16] K. O. Stanley and R. Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2):99–127, 2002.

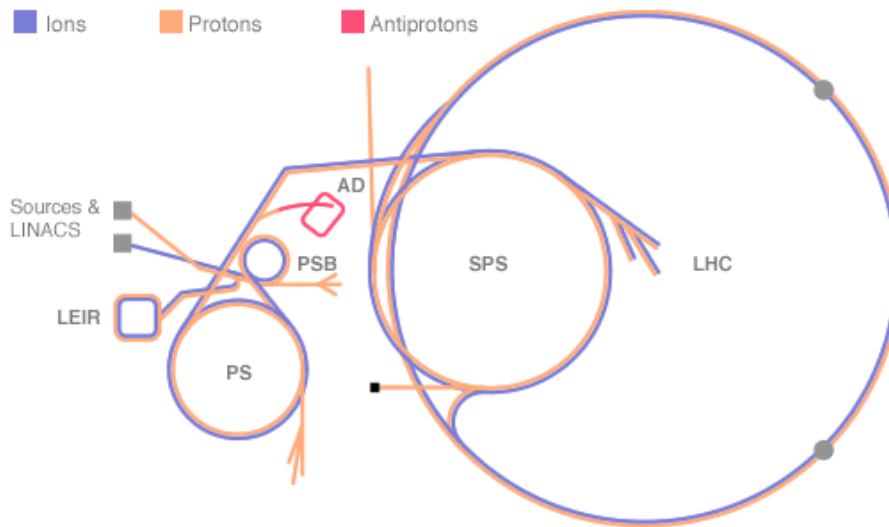


Fig. 20. The Large Hadron Collider (LHC), showing the 27-kilometer ring which serves as the final stage of the accelerator complex. Collisions of protons at 14 TeV center-of-mass energy will occur at two points around the ring.

- [17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, pages 318–362. 1986.
- [18] A. Abulencia *et al.* Measurement of the  $J/\psi$  meson and  $b$  hadron production cross sections in  $p\bar{p}$  collisions at  $\sqrt{s} = 1960$  gev. *Physical Review D*, 71:032001, 2005.
- [19] A. Abulencia *et al.* Measurement of the  $t\bar{t}$  production cross section in  $p\bar{p}$  collisions at  $\sqrt{s} = 1.96$  TeV using dilepton events. *Physical Review Letters*, 93:142001, 2004.
- [20] K. O. Stanley, B. D. Bryant, and R. Miikkulainen. Real-time neuroevolution in the NERO video game. *IEEE Transactions on Evolutionary Computation* 9(6): 653-668, 2005.
- [21] A. Abulencia *et al.* Top quark mass measurement from dilepton events at CDF II. *Physical Review Letters*, 96:152002, 2005.
- [22] T. Sjostrand *et al.* High-energy physics event generation with PYTHIA 6.1. *Computer Physics Communications*, 135:238, 2001.
- [23] S. Agostinelli *et al.* GEANT4. *Nuclear Instruments and Methods in Physics Research*, A506:250–303, 2003.
- [24] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.
- [25] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.
- [26] P. Domingos. Metacost: a general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155-164, 1999.
- [27] K. Cranmer and R. S. Bowman. PhysicsGP: a genetic programming approach to event selection. *Computer Physics Communications*, 167:165, 2005.
- [28] S. Whiteson and P. Stone. Evolutionary function approximation for reinforcement learning. *Journal of Machine Learning Research*, 7(May):877–917, 2006.
- [29] X. Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, 1999.
- [30] T. Aaltonen *et al.* Measurement of the top quark mass with dilepton events selected using neuroevolution at CDF. *Physical Review Letters* 102(152001), 2009.
- [31] J. Estrada, PhD Thesis, University of Rochester E. Brubaker, PhD Thesis, University of California, Berkeley E. Varnes, PhD Thesis, University of California, Berkeley A. Kovalev, PhD Thesis, University of Pennsylvania B. Jayatilaka, PhD Thesis, University of Michigan
- [32] T. Aaltonen *et al.* Cross Section Constrained Top Quark Mass Measurement from Dilepton Events at the Tevatron. *Physics Review Letters*, 100:062005, 2008.
- [33] S. Whiteson and D. Whiteson. Stochastic optimization for collision selection in high energy physics. In *Proceedings of the Nineteenth Annual Innovative Applications of Artificial Intelligence Conference*, 1819-1825, 2007.
- [34] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [35] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–137, 1997.
- [36] Shakhnarovich, Darrell, and Indyk (editors). Nearest-neighbor methods in learning and vision. The MIT Press, 2005.
- [37] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273-297, 1995.
- [38] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [39] F. Gomez, J. Schmidhuber, and R. Miikkulainen. Efficient non-linear control through neuroevolution. In *Proceedings of the European Conference on Machine Learning*, 2006.



- [40] Pedro and Lozano. (editors). Estimation of distribution algorithms: A new tool for evolutionary computation. Kluwer Academic Publishers, 2002.
- [41] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.
- [42] S. Whiteson, P. Stone, K. O. Stanley, R. Miikkulainen, and N. Kohl. Automatic feature selection in neuroevolution. In *GECCO 2005: Proceedings of the Genetic and Evolutionary Computation Conference*, 1225–1232, 2005.
- [43] T. Fawcett. Feature Discovery for Problem Solving Systems. PhD thesis, University of Massachusetts, Amherst, MA (1993)
- [44] P. E. Utgoff. Feature construction for game playing. In: *Machines that Learn to Play Games*. Nova Science Publishers (2001) 131–152
- [45] B. Taskar. Learning structured prediction models: A large margin approach. PhD Thesis, Stanford University, 2004.
- [46] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
- [47] L. Getoor and B. Taskar (editors). Introduction to relational statistical learning. MIT Press, 2007.
- [48] A. Cho. Aging atom smasher runs all out in race for most coveted particle. *Science*, 312:1302–1303, 2006.