高能物理实验：机器学习方法
Machine Learning in High Energy Physics

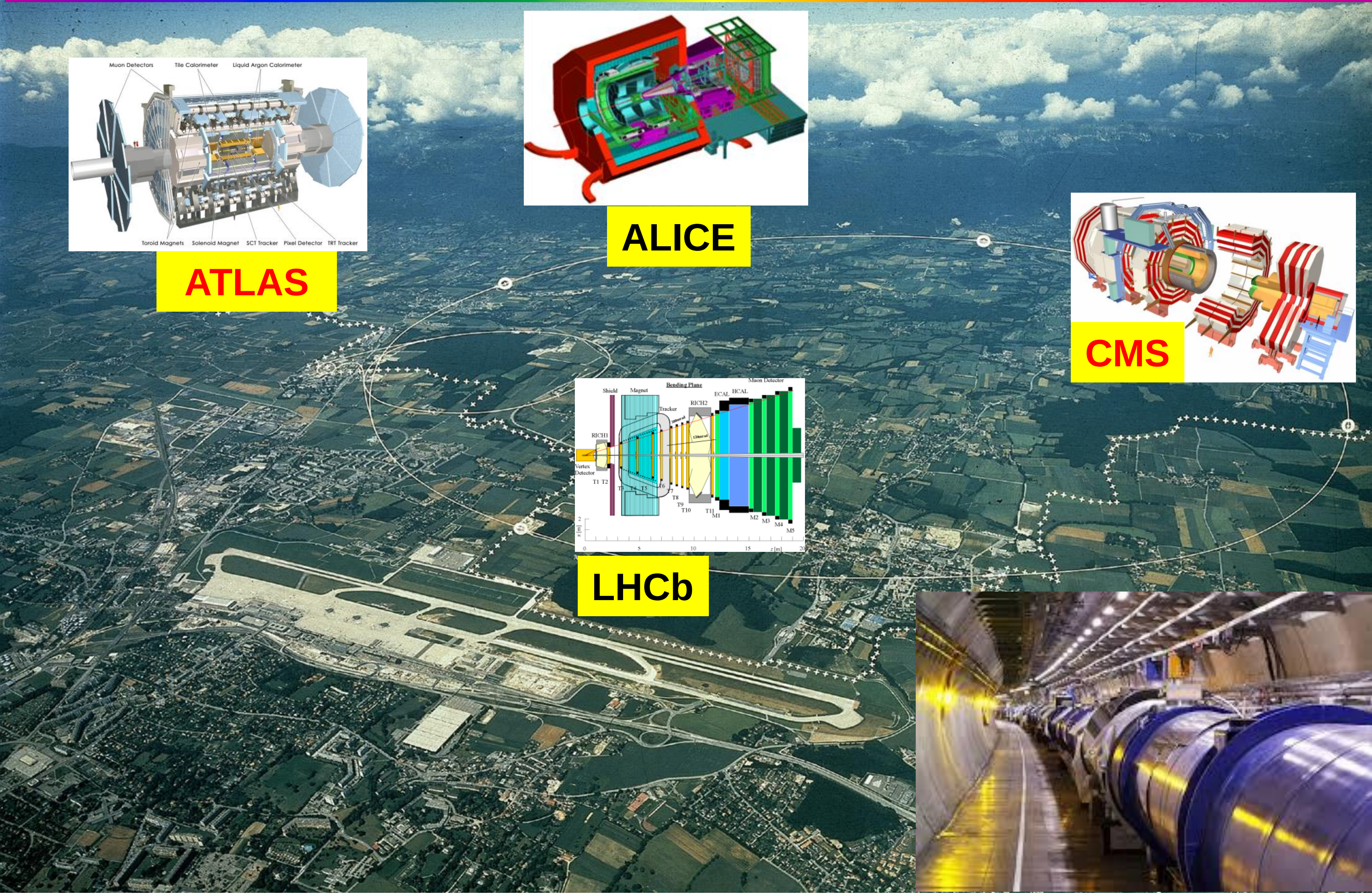
杨海军（上海交通大学）

The International Summer School on
TeV Experimental Physics (iSTEP)
武汉大学
2018年7月18日

报告大纲

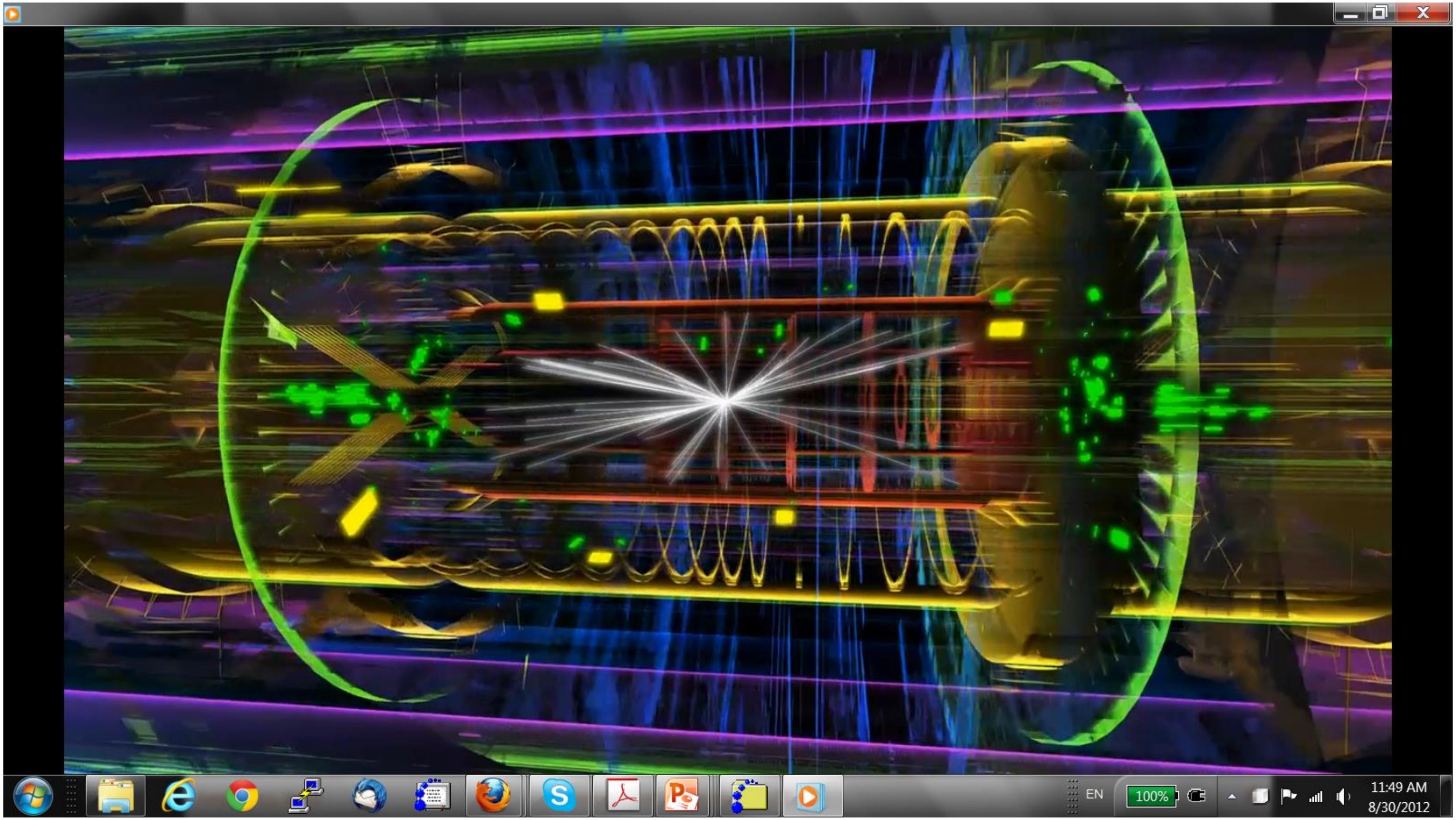
- 高能物理实验（譬如大型强子对撞机实验LHC）
- 提高粒子探测效率的重要性
- 常用的机器学习方法
  - Artificial Neural Networks (ANN)
  - Boosted Decision Trees (BDT)
- 基于Root的多变量分析软件包TMVA

CERN大型强子对撞机LHC

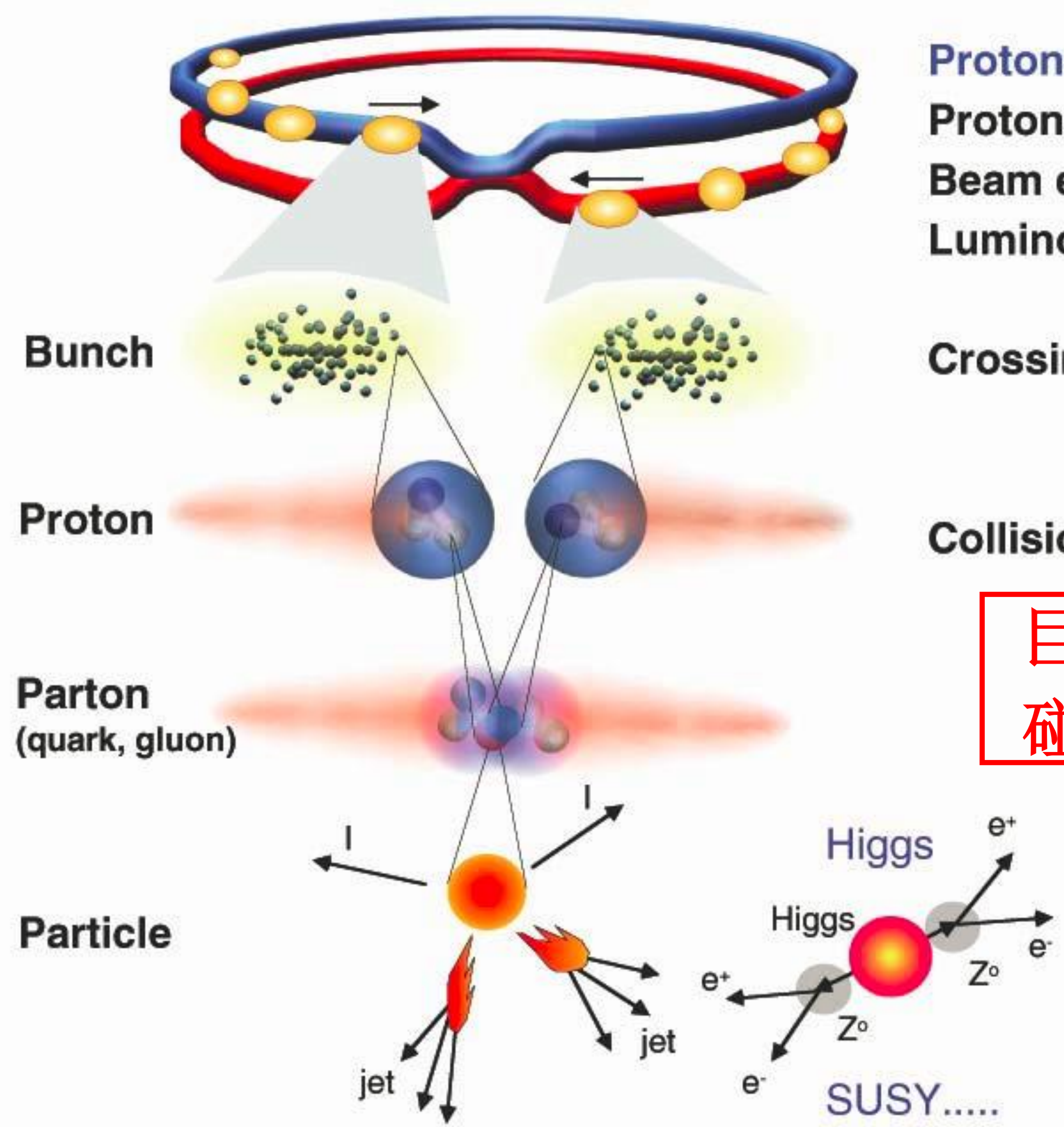


大型强子对撞机 LHC：质子对撞

- 质子-质子对撞示意图



大型强子对撞机 LHC：质子对撞

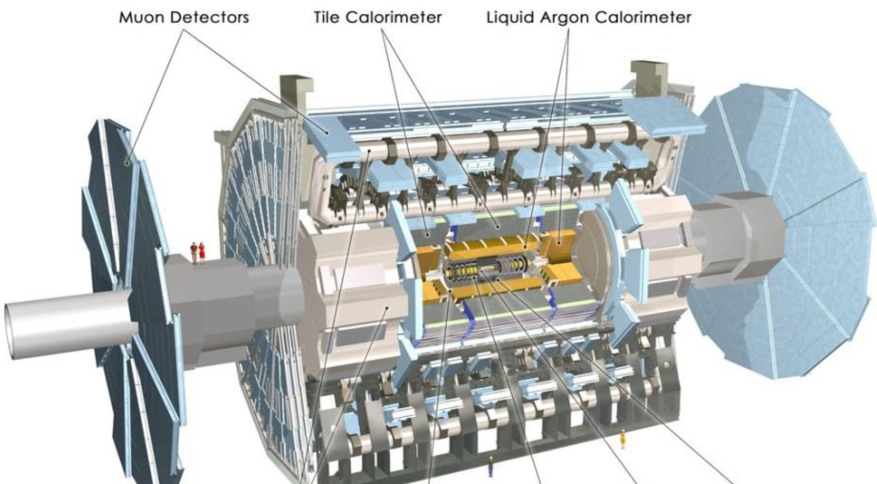


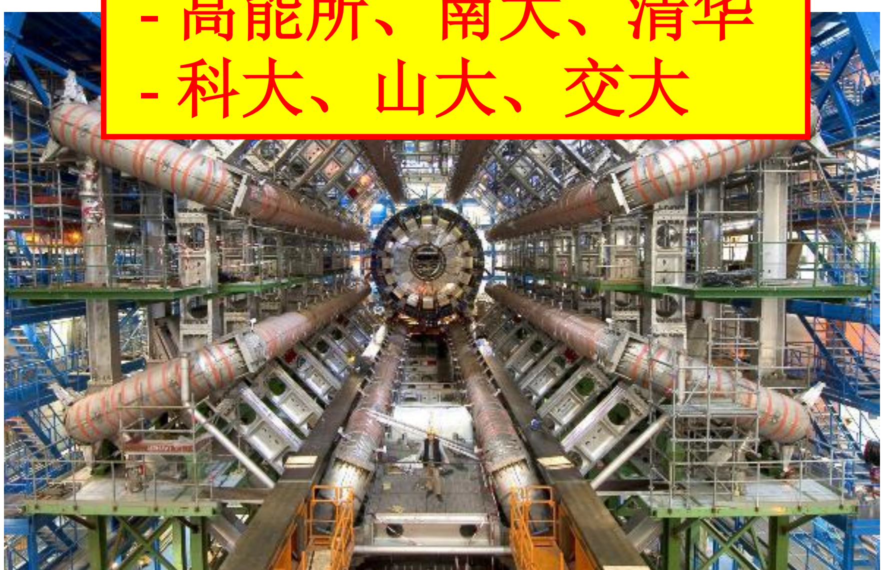
Proton-Proton	2835 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	10^{34} cm^{-2} s^{-1}
Crossing rate	40 MHz
Collisions ≈	10^7 - 10^9 Hz

巨大挑战：从10万亿次碰撞中挑选一个希格斯

Selection of 1 in 10,000,000,000,000

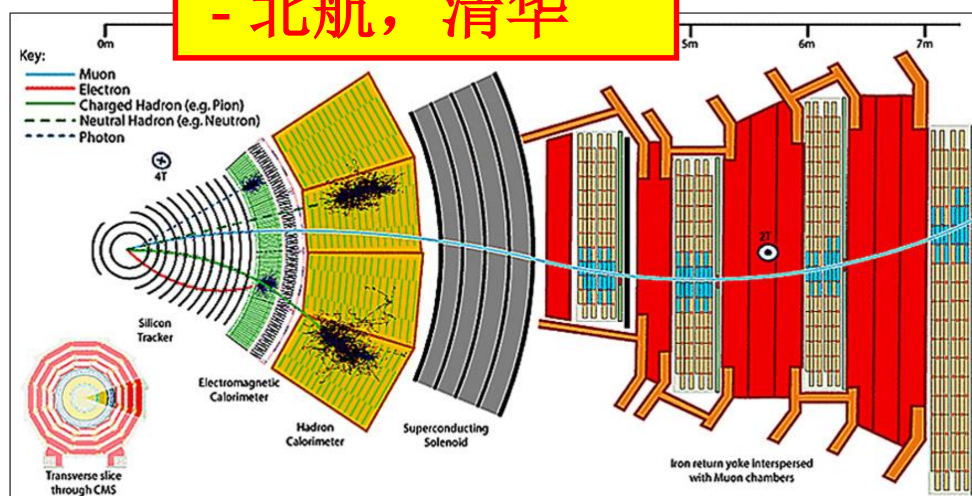
ATLAS 和 CMS 探测器



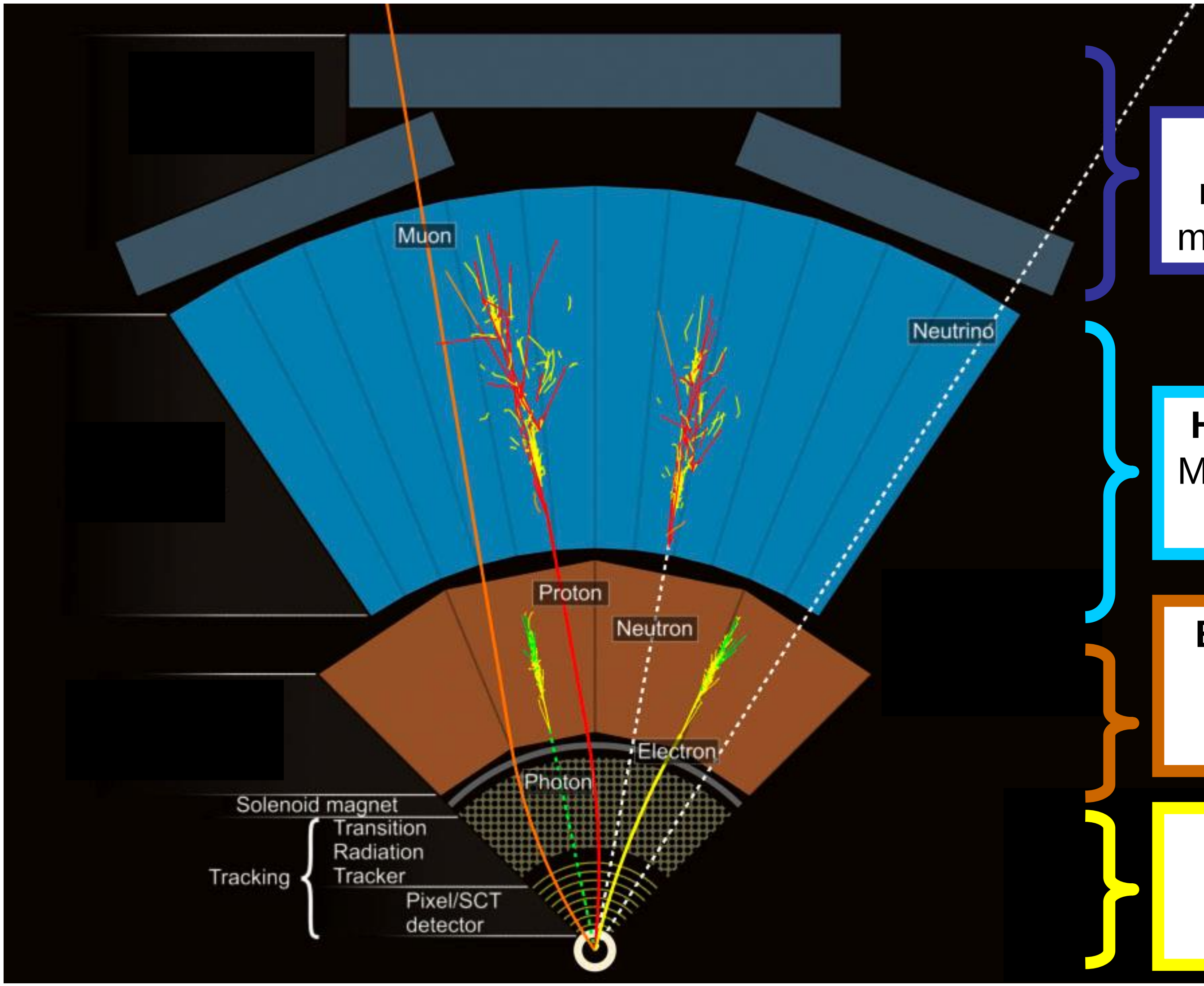


ATLAS-中国组
- 高能所、南大、清华
- 科大、山大、交大

CMS-中国组
- 高能所，北大
- 北航，清华



探测器：粒子鉴别



Muon Spectrometer:
muon identification and momentum measurement

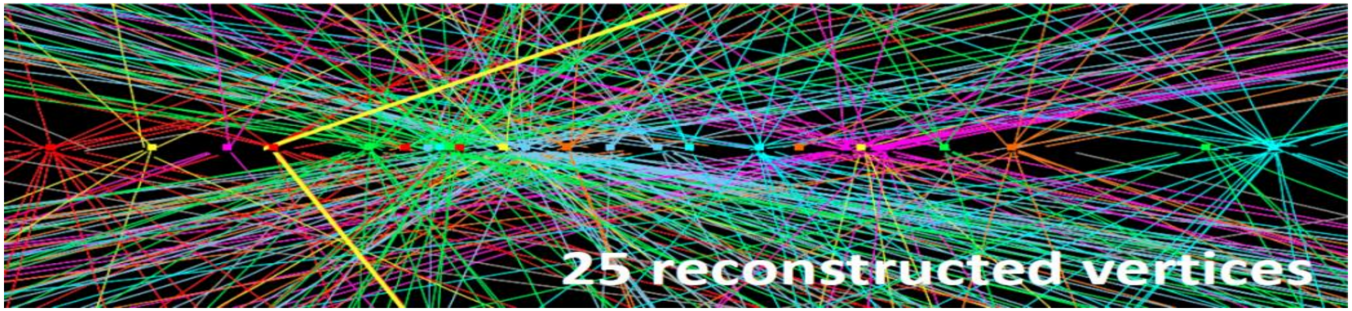
Hadronic calorimeter:
Measurement of jets and missing energy

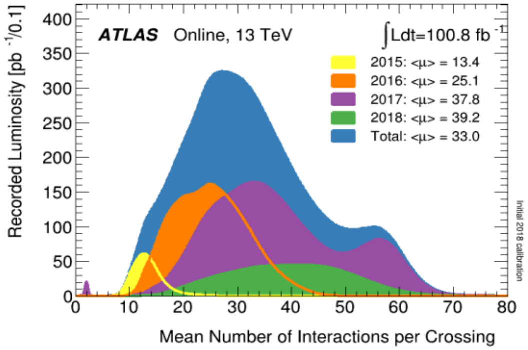
Electromagnetic calo:
e/γ identification and energy measurement

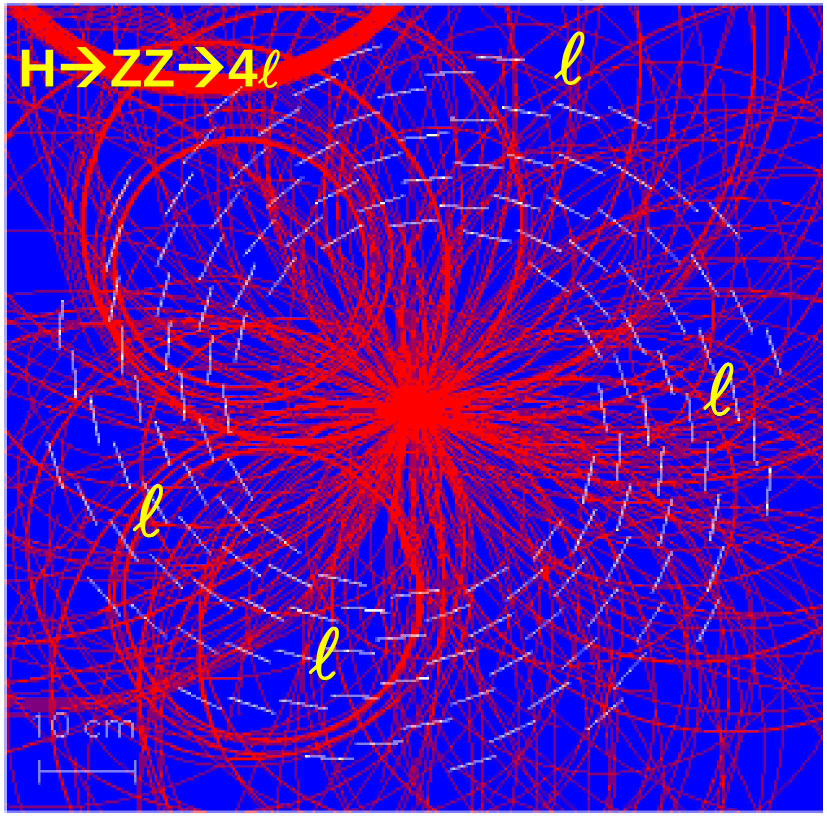
Tracking system:
Charged particle momentum, vertexing

LHC实验面临的挑战

LHC 亮度提高 （2×10^{34} $cm^{-2}s^{-1}$ ）➔ 造成大量事例堆积 (~ 50)





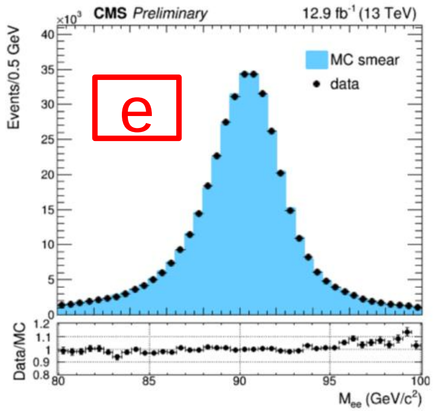


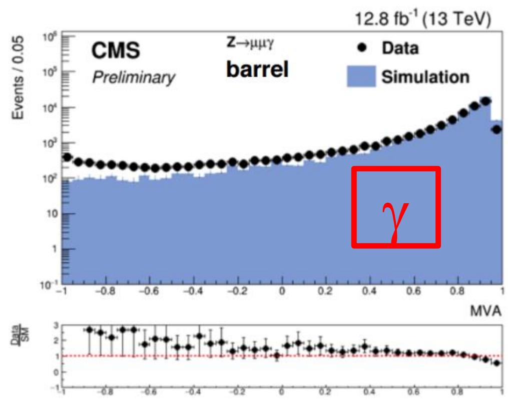
大量事例堆积对探测器信号读出、粒子重建和鉴别造成巨大的挑战 ！

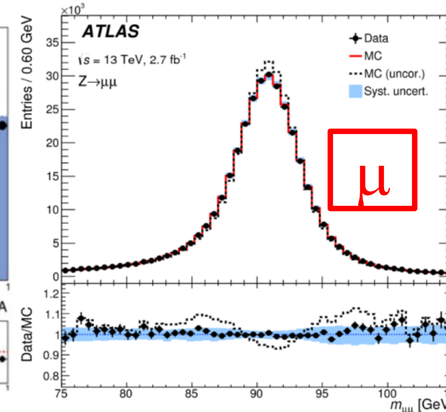
广泛采用多变量方法提高粒子鉴别效率和事例识别

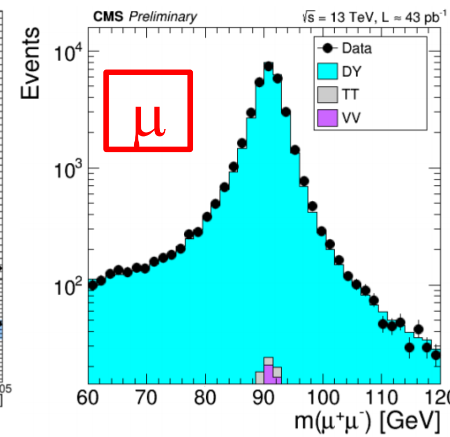
NIMA543 (2005) 577-584
NIMA555 (2005) 370-385
arxiv:0703039.pdf

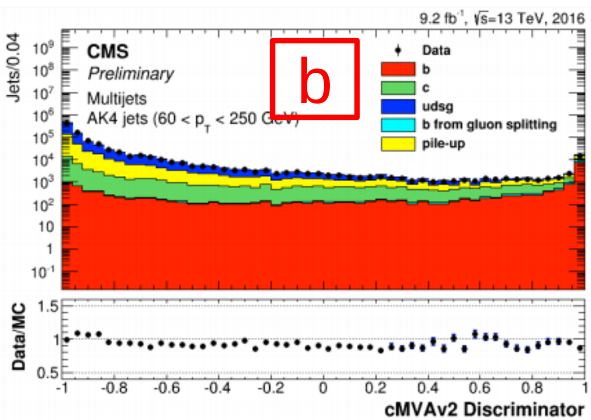
BDT方法应用于LHC实验

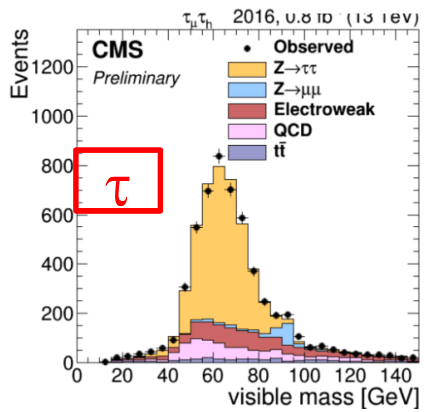




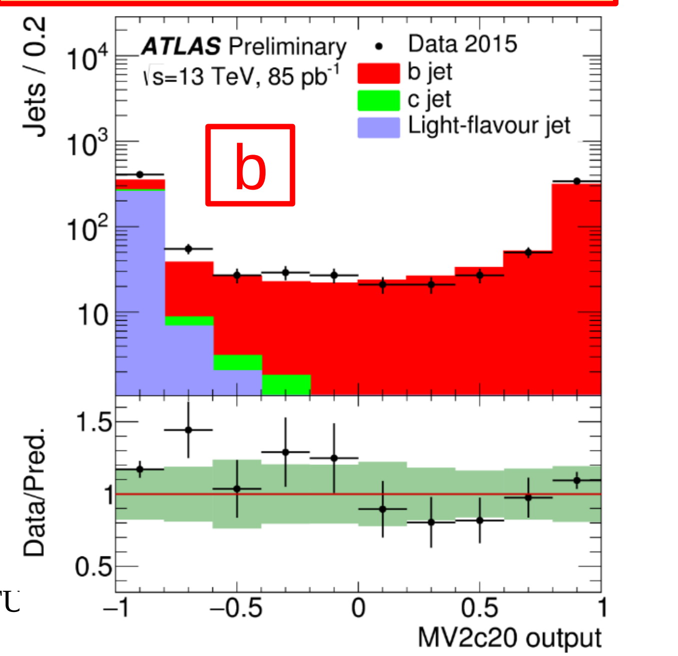


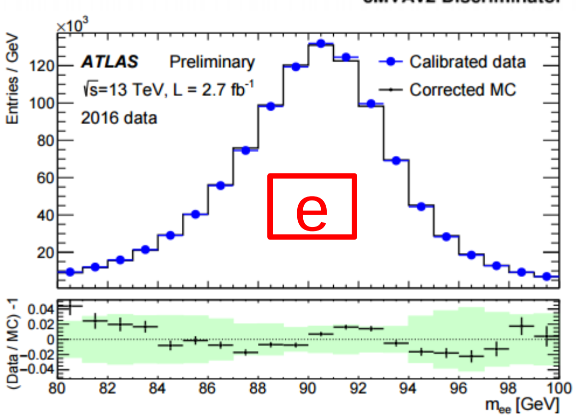


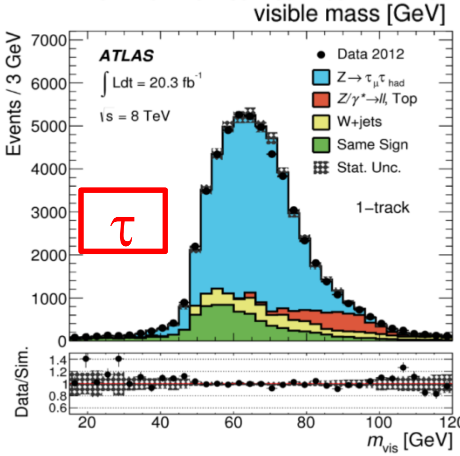




e, γ, τ, b粒子的鉴别采用多变量分析- BDT方法。





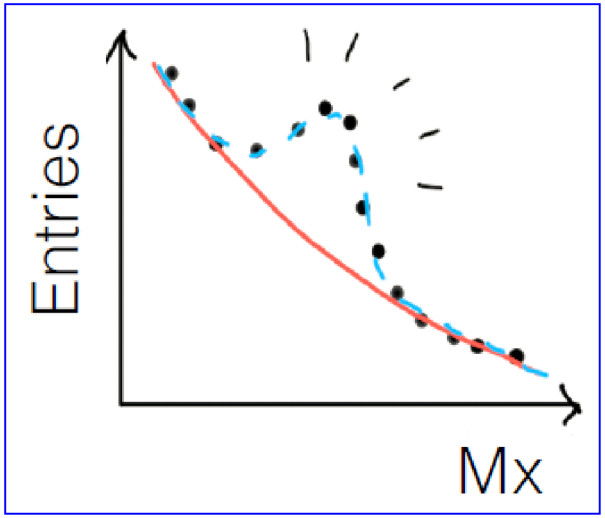


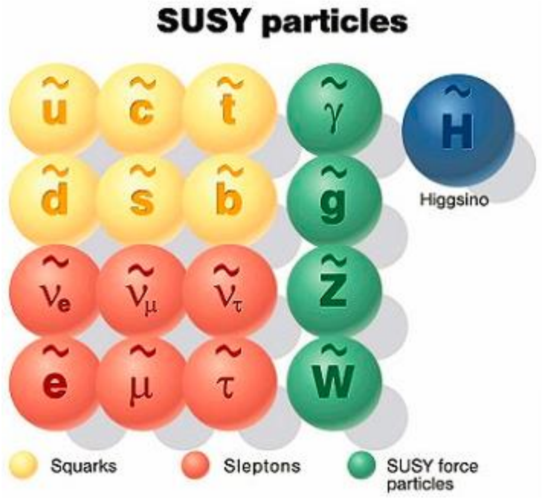
g (SJTU

粒子物理实验：寻找新粒子

寻找新粒子

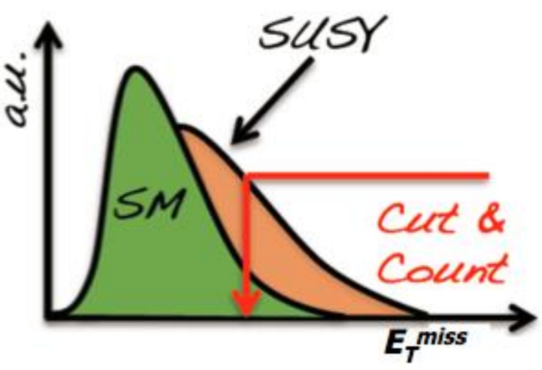
- ✓ X → di-photon
- ✓ X → di-boson
- ✓ X → Z+photon
- ✓ X → di-lepton
- ✓ X → di-jets
- ✓ X → hh





- ✓ SUSY Particles
- ✓ Dark Matter
- ✓ Heavy Quarks
- ✓ Majorana neutrino
- ✓ Long Lived Particle

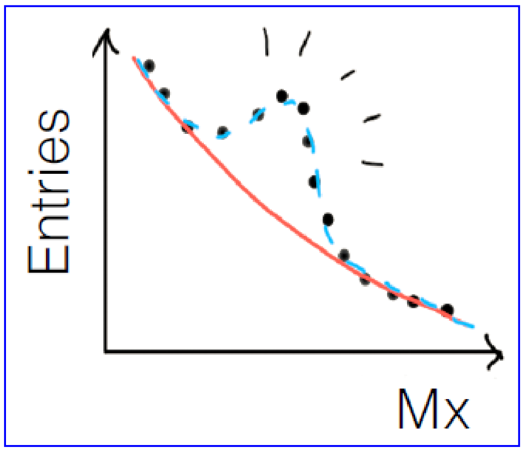
…… ……



如何寻找新粒子？

高能物理实验寻找新粒子：

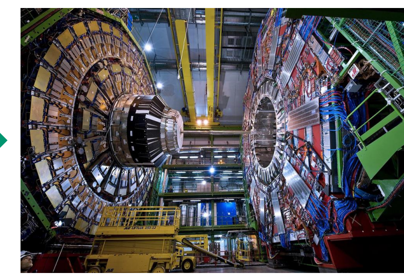
- 信号很小，背景本底很多
- 信噪比低，难以在实验中发现



➔提高统计量：
提高对撞机亮度
增加运行时间

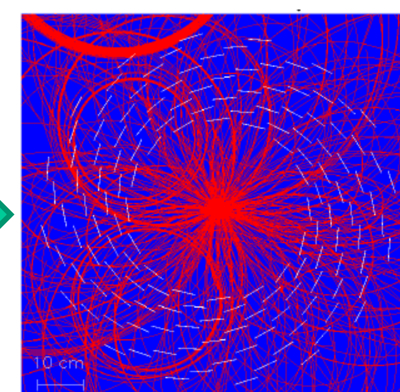
➔提高探测器精度

➔提高粒子探测效率



采用先进的机器学习方法：

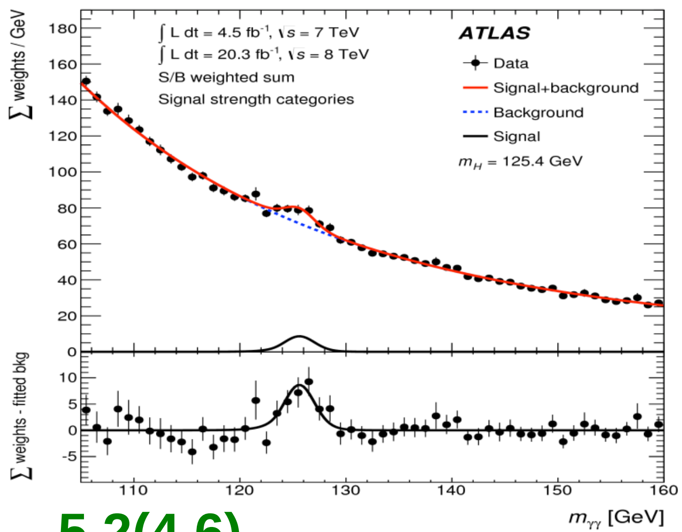
- 提高粒子鉴别效率
- 提高信噪比
- 提高新粒子探测的灵敏度



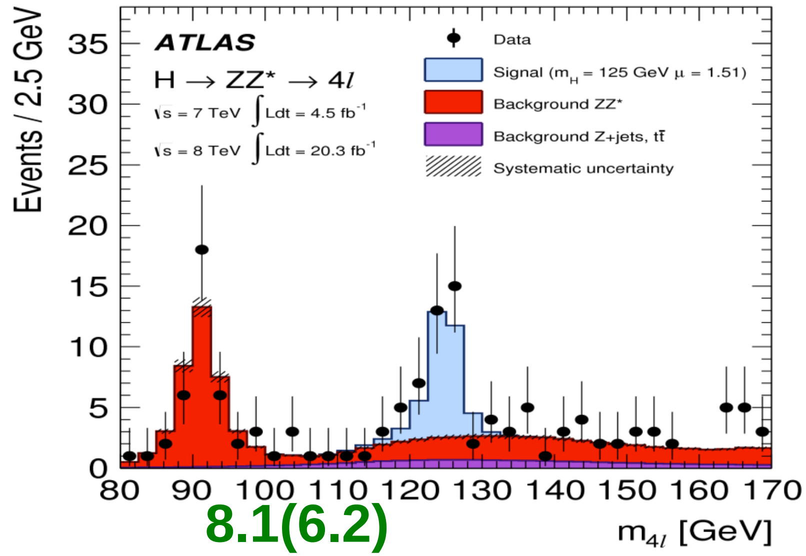
希格斯粒子的发现（2012）

ATLAS

obs(expected) significance

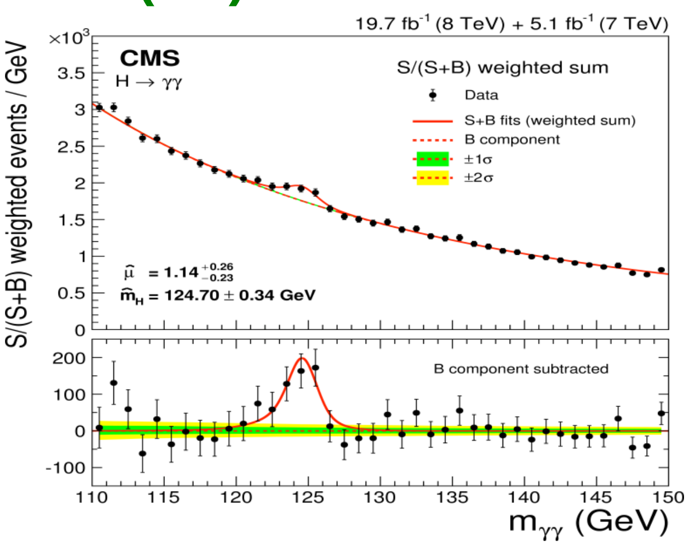


5.2(4.6)

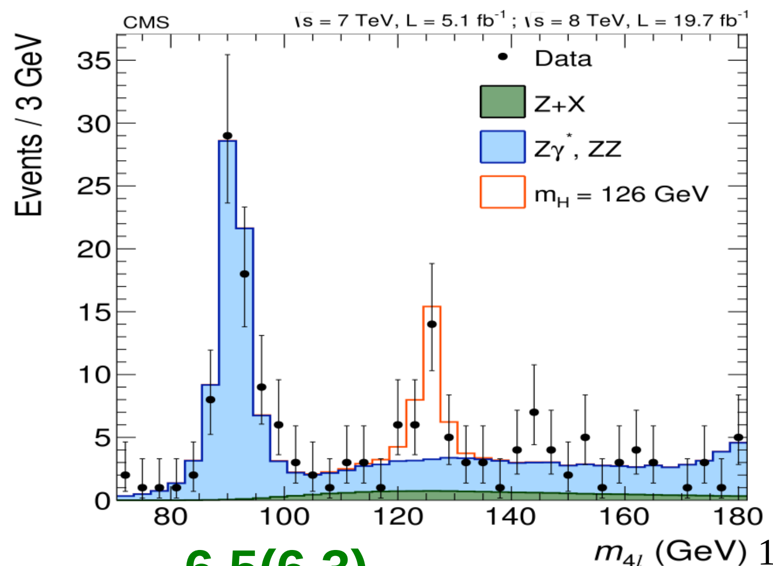


8.1(6.2)

CMS

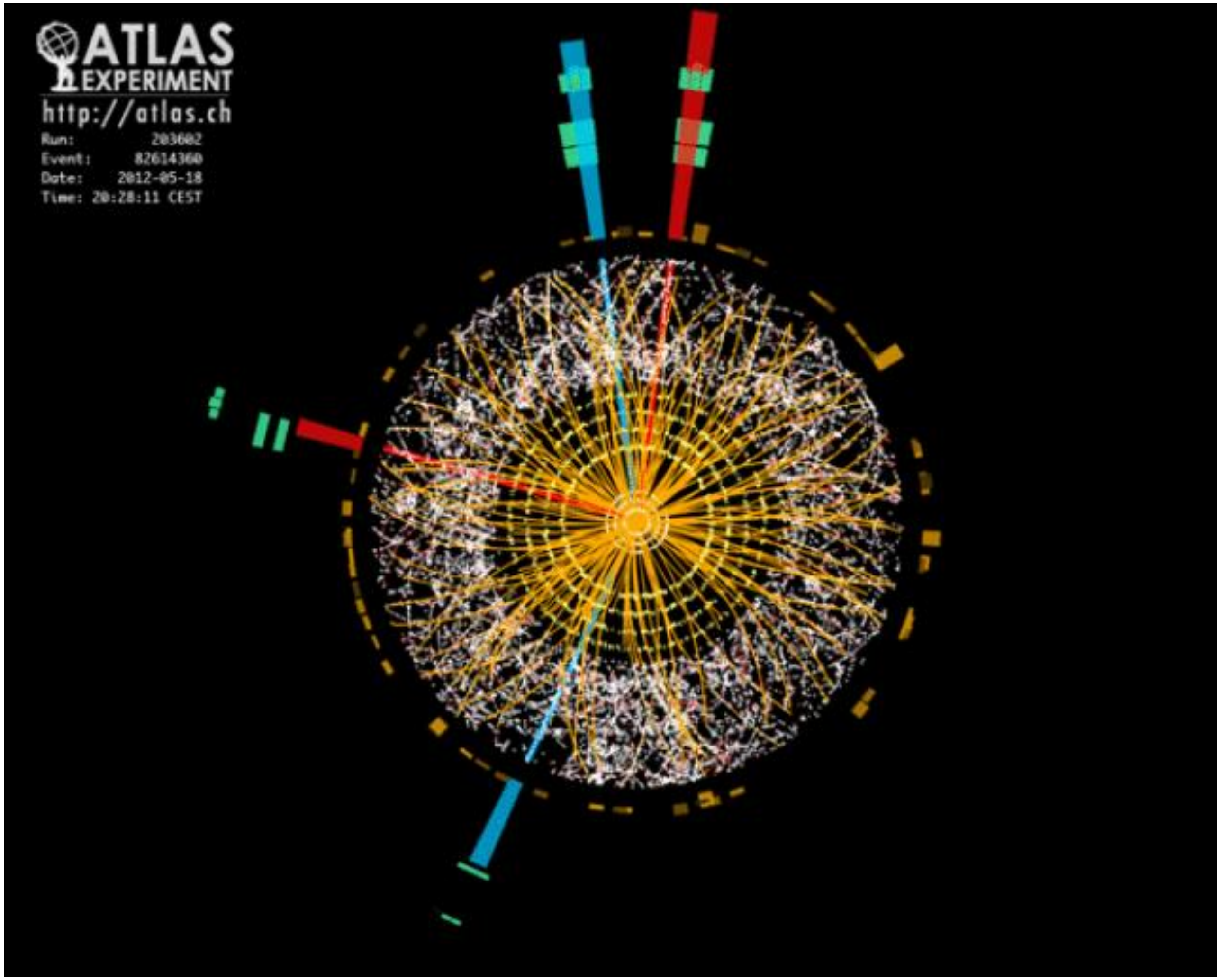


5.6(5.3)

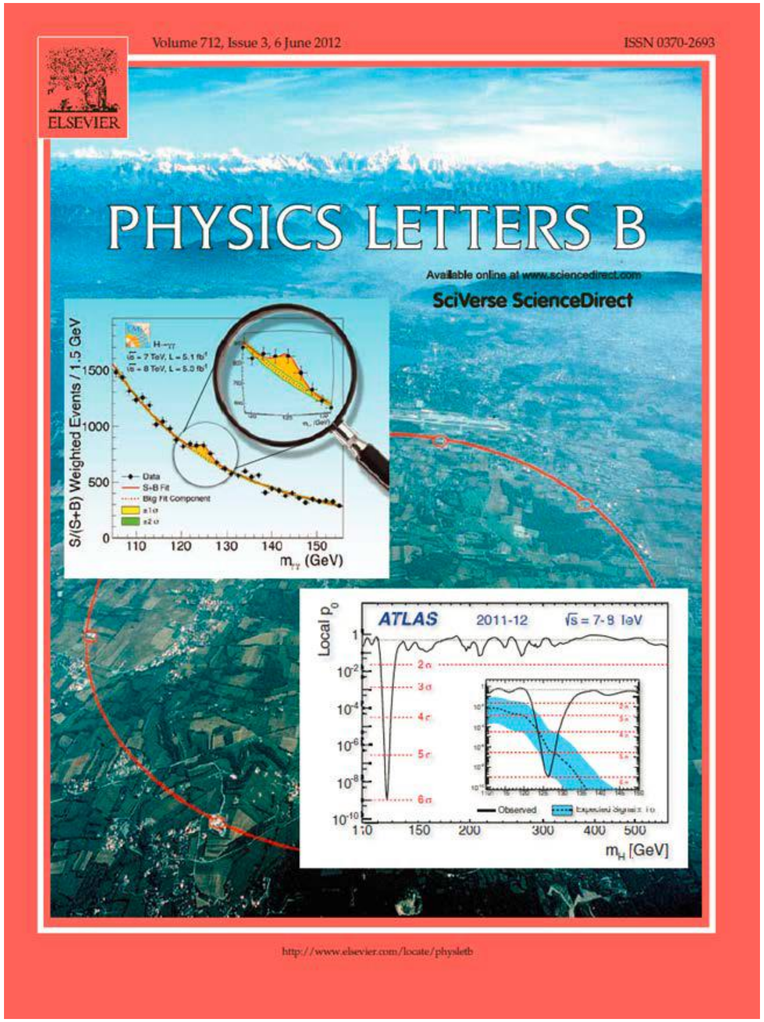


6.5(6.3)

Higgs →ZZ* →4ℓ 的发现（视频）



希格斯粒子的发现（2012）

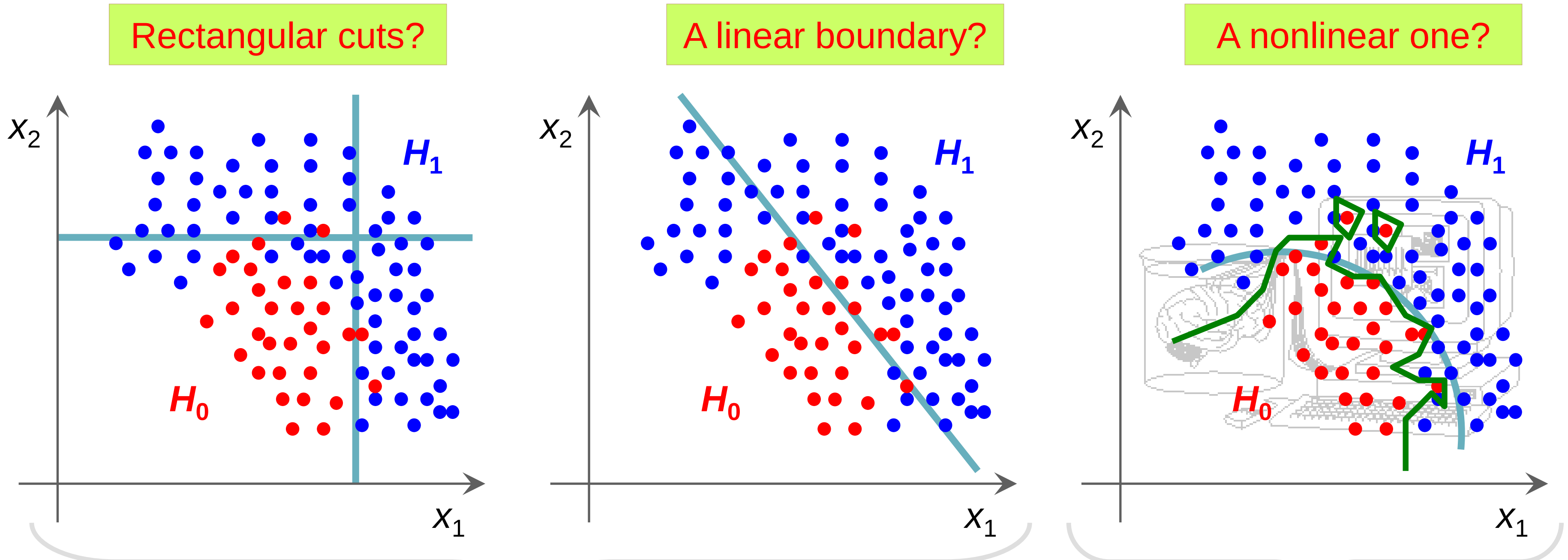


Phys. Lett. B 716 (2012) 1-29 (ATLAS)
Phys. Lett. B 716 (2012) 30-61 (CMS)
每篇超过8000次引用

http://www.sciencemag.org/site/special/btoy2012/
Science 338 (2012) 1576-1582
Science 338 (2012) 1569-1575

如何挑选事例？

- 假定有两组数据，对应不同的事例: H_0(B), H_1(S)
  - We have found discriminating input variables x_1, x_2, …
  - What decision boundary should we use to select events of type H_1 ?

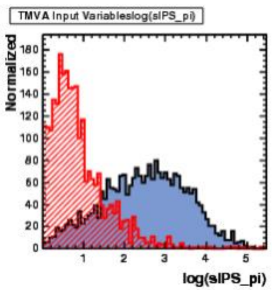


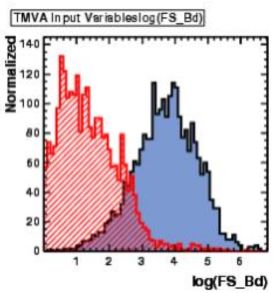
- TMVA helps to decide on the model and finds the "optimal" boundary!

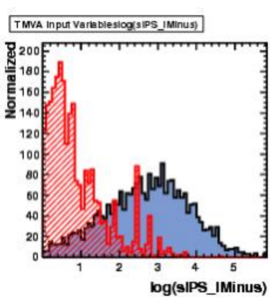
Low variance (stable), high bias methods

High variance, small bias methods

如何挑选事例？



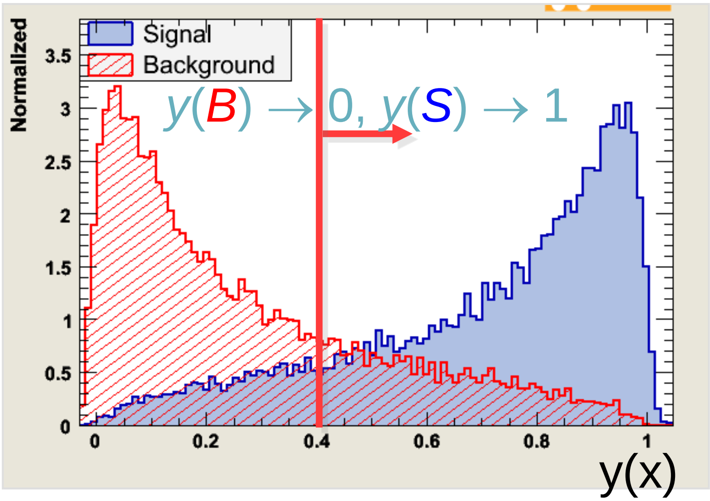




D

"feature space"

Each event, **Signal** or **Background**, has "D" measured variables.



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

y(x): "test statistic" in D-dimensional space of input variables

Distributions of y(x): $PDF_S(y)$ and $PDF_B(y)$

Used to set the selection cut!

→ Efficiency and purity

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

y(x) = const: surface defining the decision boundary.

Overlap of $PDF_S(y)$ and $PDF_B(y)$ affects separation power, purity

如何挑选事例？

➔ Decide to treat an event as "Signal" or "Background"

Type-1 error: （本底误判为信号事例）

classify event as Class C even though it is not

(accept a hypothesis although it is not true)

(reject the null-hypothesis although it would have been the correct one)

→ loss of purity (in selection of signal events)

Type-2 error: （信号误判为本底事例）

fail to identify an event from Class C as such

(reject a hypothesis although it would have been true)

(fail to reject the null-hypothesis/accept null hypothesis although it is false)

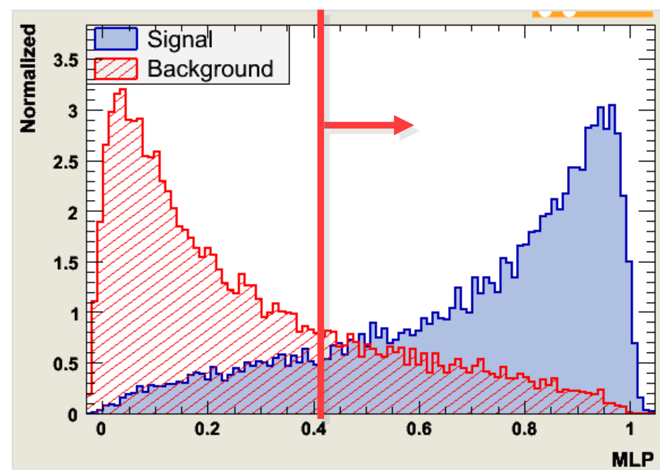
→ loss of efficiency (in selecting signal events)

How to choose "cut"? → need to know prior probabilities (S, B abundances)

- Measurement of signal cross section: maximum of S/√(S+B)
- Discovery of a signal : maximum of S/√(B)
- Precision measurement: high purity (p)
- Trigger selection: high efficiency (ε)

Trying to select signal events:
(i.e. try to disprove the null-hypothesis stating it were "only" a background event)

truly is: \ accept as:	Signal	Back-ground
Signal	☺	Type-2 error
Back-ground	Type-1 error	☺

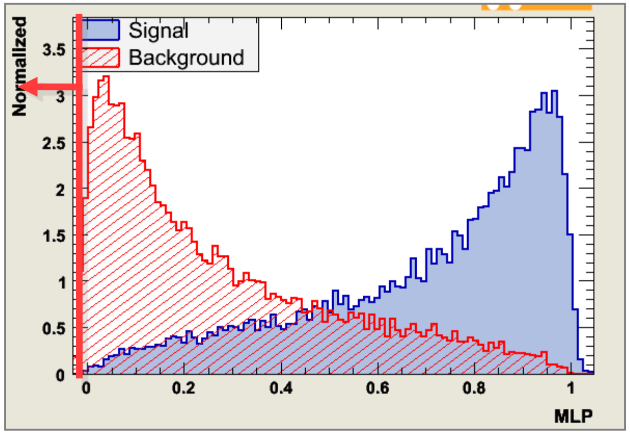


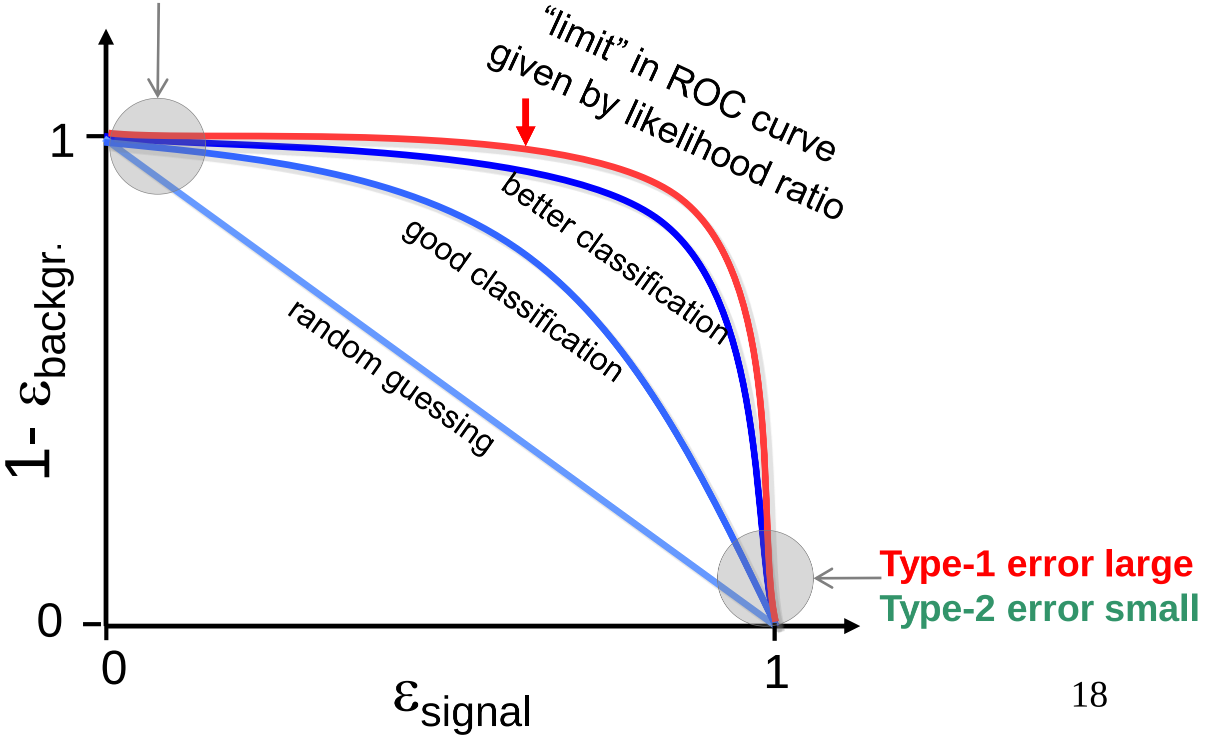
如何挑选事例？

接收操作特征（Receiver Operating Characteristic，ROC）曲线，即通常所讲的ROC Curve，是机器学习领域中常用的分类性能评估曲线。

Type-1 error:（本底误判为信号事例）
Type-2 error:（信号误判为本底事例）

$$y(x) = \frac{\mathrm{P}(x\,|\,S)}{\mathrm{P}(x\,|\,B)}$$





机器学习: Machine Learning

- **机器学习**是人工智能的一个分支。人工智能的研究是从以“推理”为重点到以“知识”为重点，再到以“学习”为重点，一条自然、清晰的脉络。机器学习在近30多年已发展为一门多领域交叉学科，涉及概率论、统计学、逼近论、计算复杂性理论等多门学科。

- **机器学习**已广泛应用于数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别、战略游戏和机器人等领域。

- **机器学习的定义：**是对能通过经验自动改进的计算机算法的研究，用数据或以往经验来学习和优化计算机程序的性能。

机器学习的类型

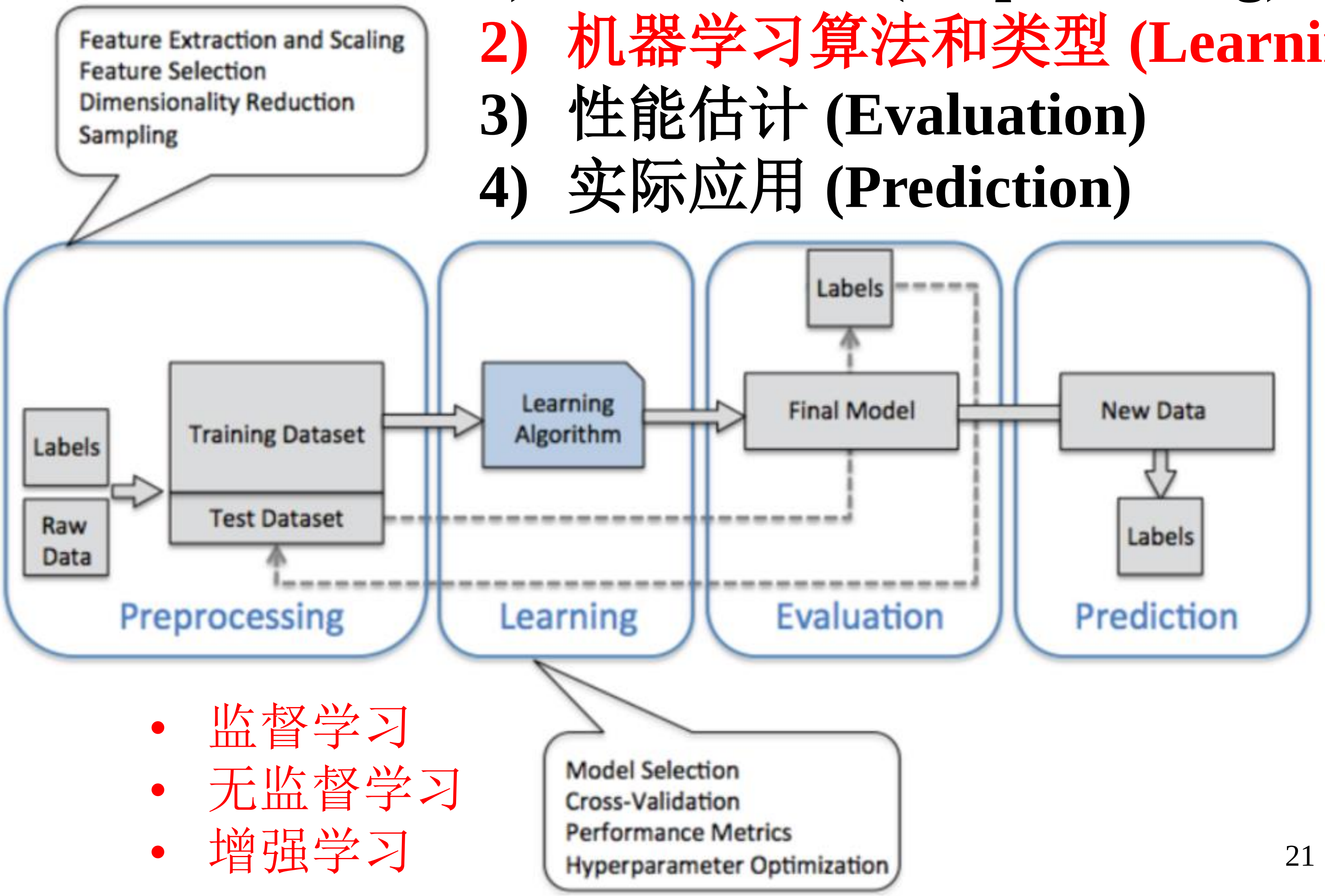
- **监督学习（Supervised Learning）**从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求是包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。常见的监督学习算法包括回归分析和统计分类。**（➔粒子鉴别）**

- **无监督学习（Unsupervised Learing）**与监督学习相比，训练集没有人为标注的结果。常见的无监督学习算法有聚类。

- **增强学习（Reinforcement Learning）**通过观察来学习和试错的方式来获得最佳策略。每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断。增强学习在很多领域已经获得成功应用，比如自动直升机，机器人控制，市场决策，工业控制，高效网页索引等。

如何构建机器学习系统

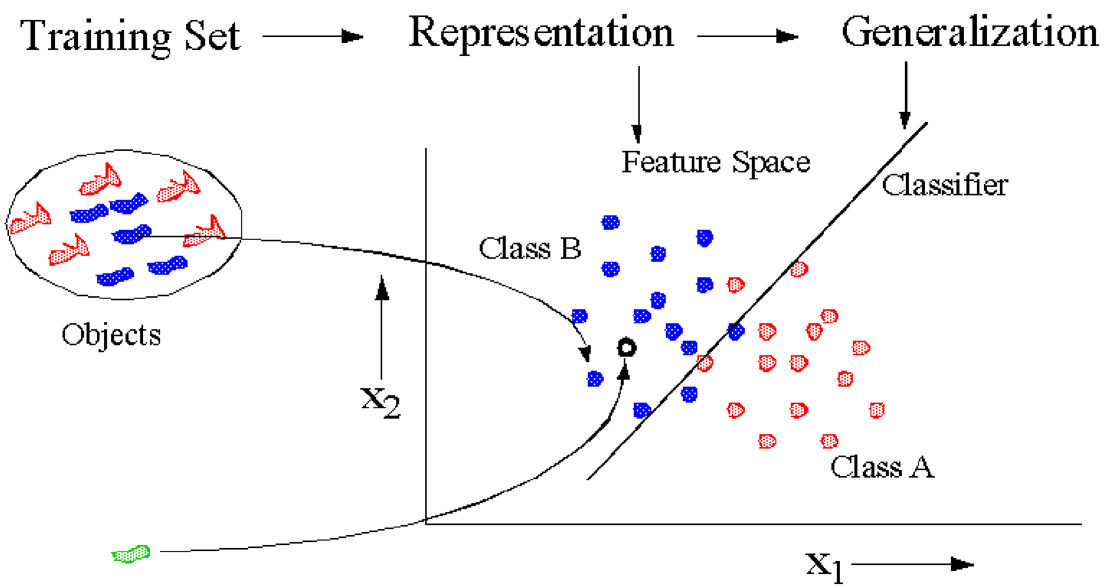
1) 数据预处理 (Preprocessing)
2) 机器学习算法和类型 (Learning)
3) 性能估计 (Evaluation)
4) 实际应用 (Prediction)

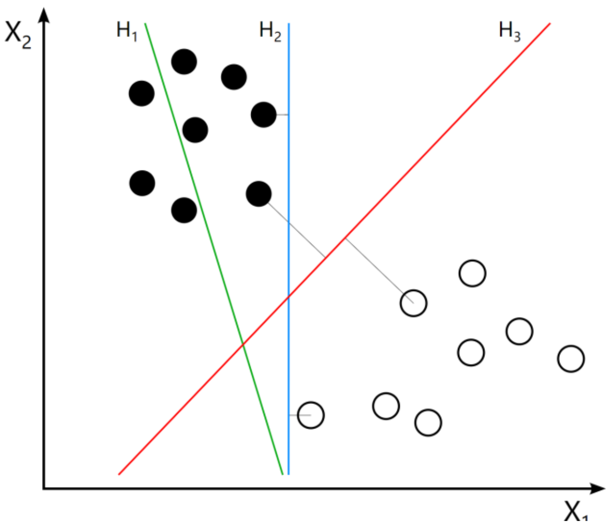


- 监督学习
- 无监督学习
- 增强学习

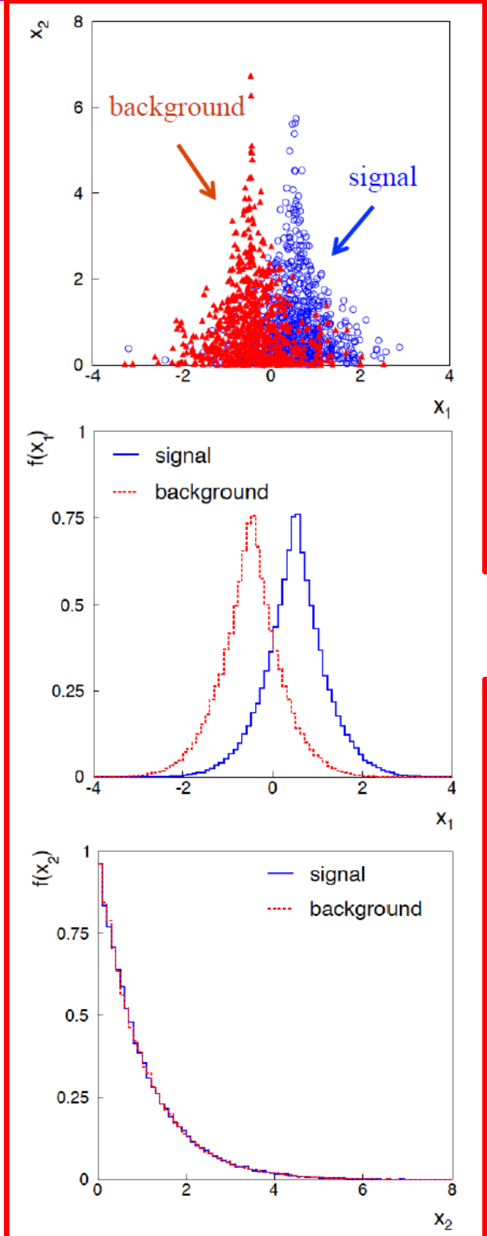
机器学习算法：模式识别

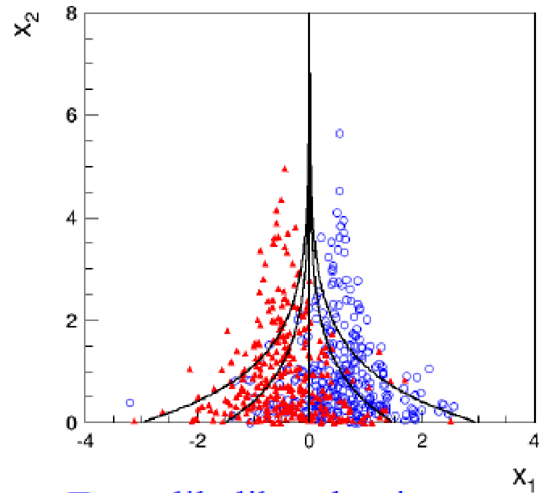
- 人工神经网络 (Artificial Neural Networks)
- 决策树 (Decision Tree, Boosted Decision Trees)
- **深度学习 (Deep Learning**, 多隐层人工神经网络)
- **支持向量机 (Support Vector Machines**, 支持向量机在高维中构造超平面或超平面集合用于分类, 使得分类边界距离最近的训练数据点越远越好。)
- **费舍尔的线性鉴别方法 (Fisher Discriminant,** 使用统计学和模式识别方法，试图找到两类物体特征的一个线性组合，以能够区分它们。)



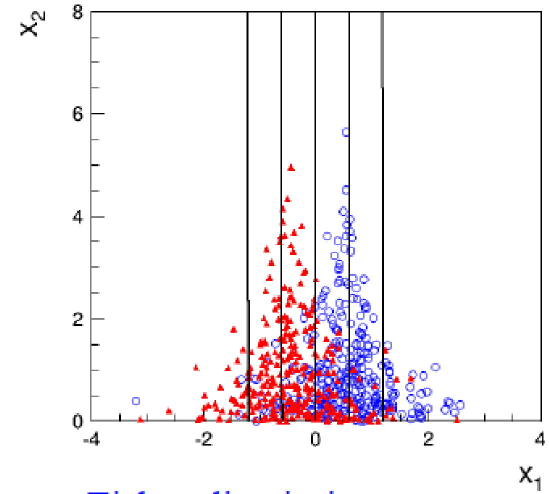


各机器学习算法区分策略

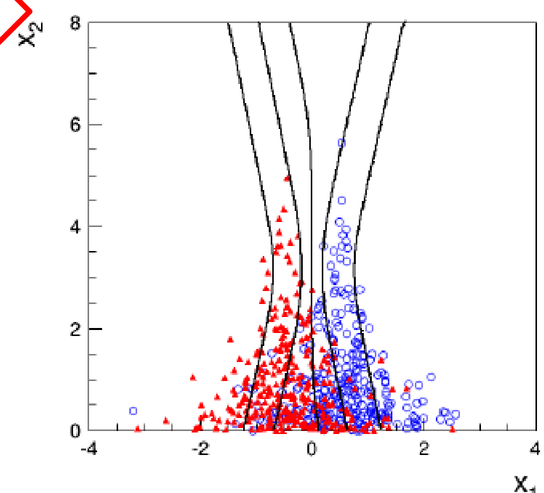




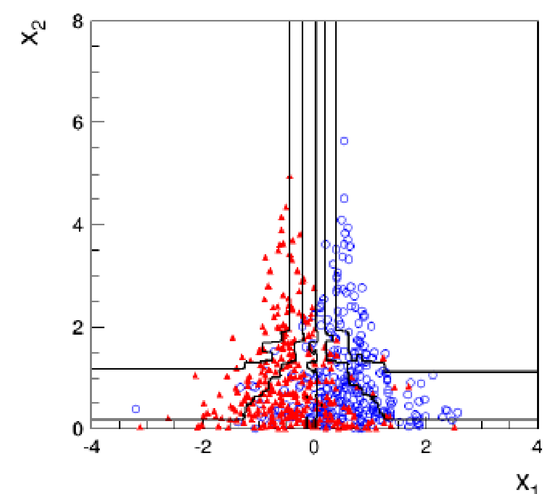
Exact likelihood ratio



Fisher discriminant

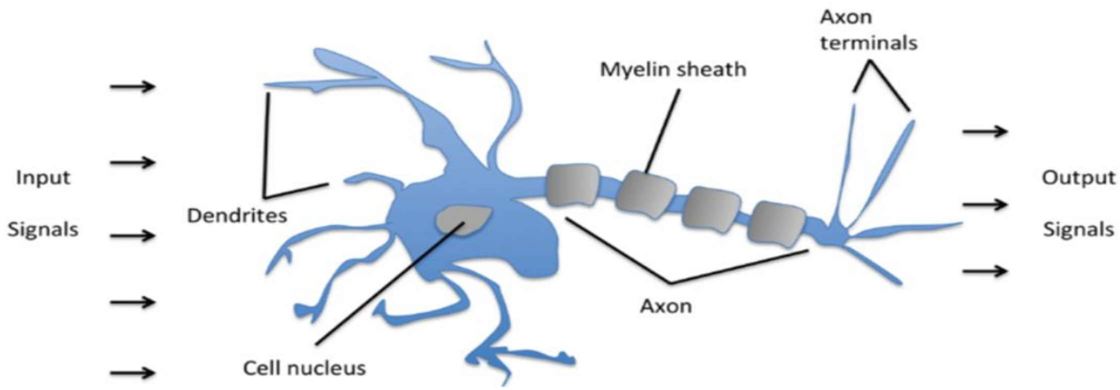


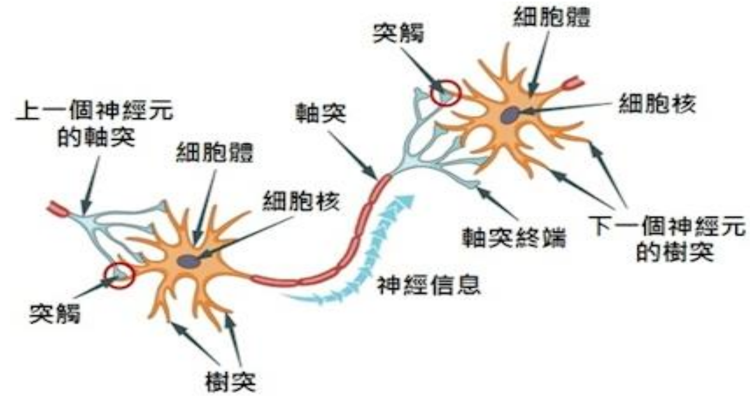
Multilayer Perceptron
1 hidden layer with 2 nodes



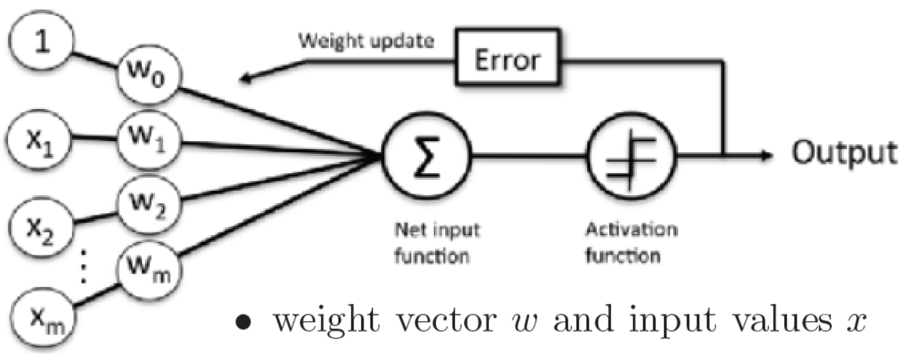
Boosted Decision Tree
200 iterations (AdaBoost)

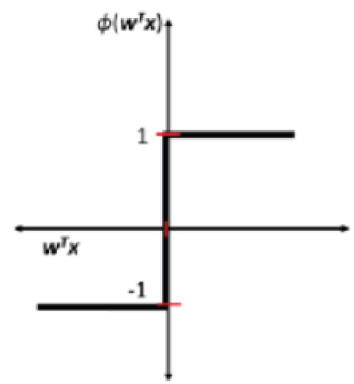
人工神经元：MCP Neuron

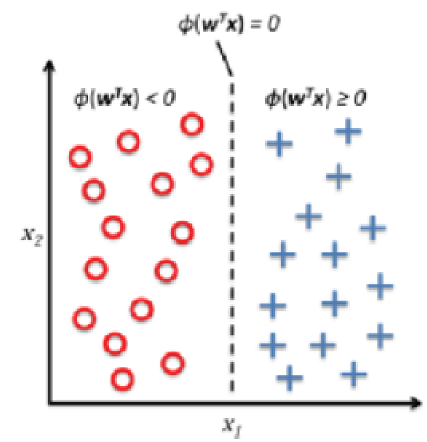




单个人工神经元







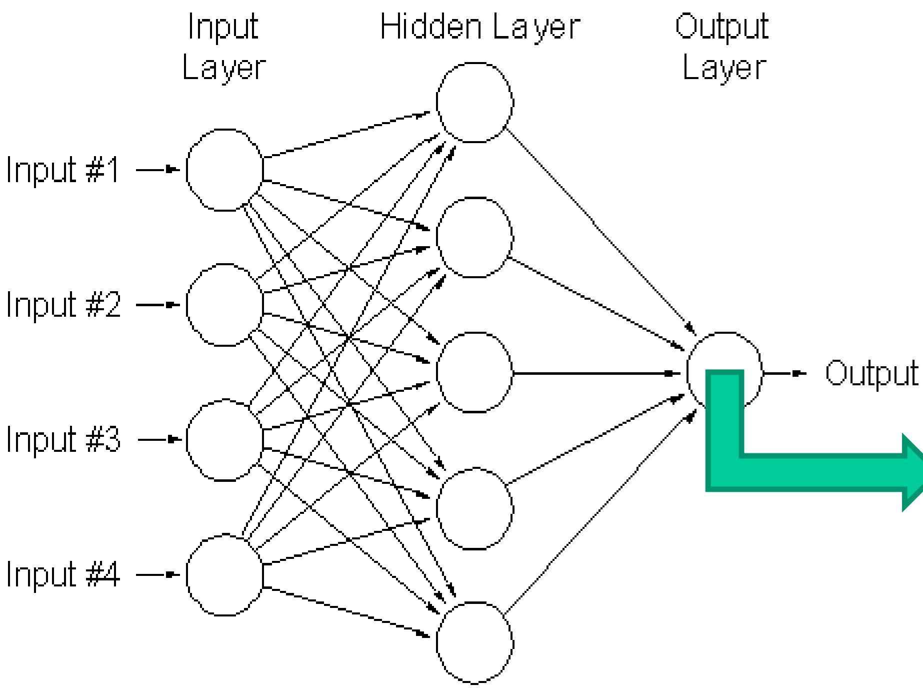
- weight vector w and input values x
- net input $z = w^T x$
- activation function $\phi(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{otherwise} \end{cases}$

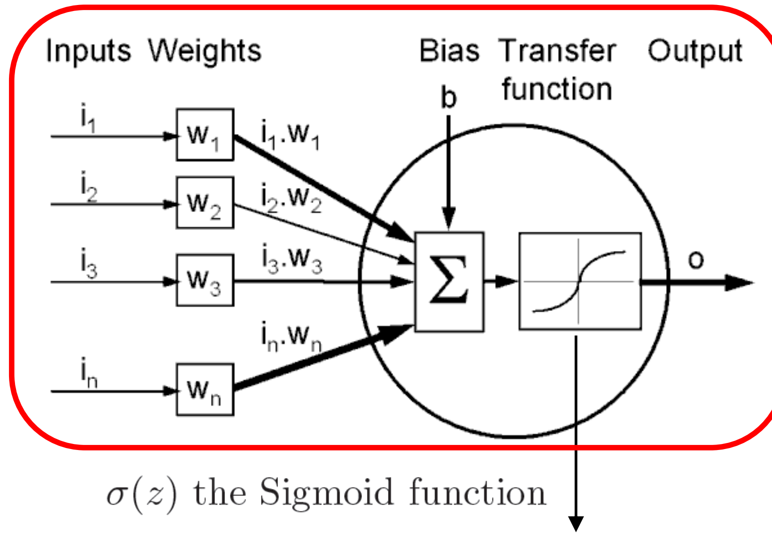
* W. S. McCulloch and W. Pitts. *A Logical Calculus of the Ideas Immanent in Nervous Activity*. The bulletin of mathematical biophysics, 5(4):115–133, 1943

* F. Rosenblatt, The Perceptron, a Perceiving and Recognizing Automaton. Cornell Aeronautical Laboratory, 1957

人工神经网络 (ANN)

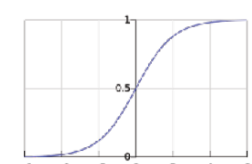
多个人工神经元 ➔ 人工神经网络





$\sigma(z)$ the Sigmoid function as the activation function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



➔用训练样本优化各节点之间的权重(w_i)和阈值(b_i)！

人工神经网络 (ANN)

- **监督学习：** Suppose signal events have output 1 and background events have output 0.
- **Mean square error E:** for given N_p training events with desired output
  - $o_i = 0$ (for background) or 1 (for signal)
  - ANN output result t_i .

$$E = \frac{1}{2N_p} \sum_{p=1}^{N_p} \sum_{i} (o_i^{(p)} - t_i^{(p)})^2$$

人工神经网络 (ANN)

- **Back Propagation Error to Optimize Weights**

$$w_{t+1} = w_t + \Delta w_t,$$

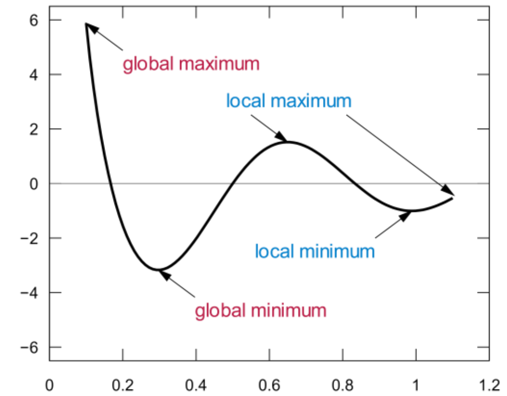
where

$$\Delta w_t = -\eta \frac{\partial E}{\partial w}$$

$$+\alpha \Delta w_{t-1}, "momentum\_term\_to\_stabalize"$$

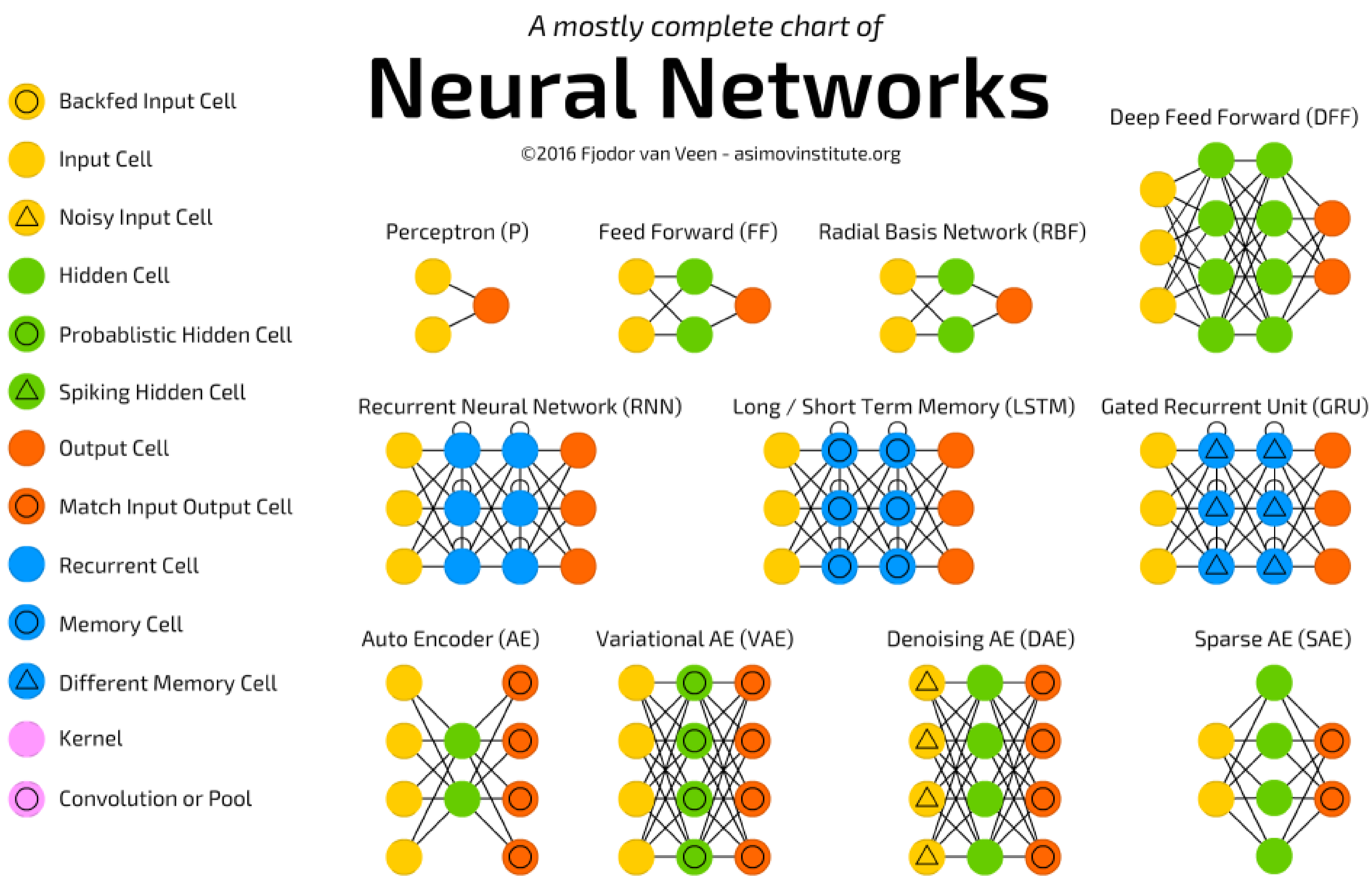
$$+\sigma, "noise\_term\_to\_avoid\_local\_minima"$$

ANN 典型参数
η = 0.05 (learning rate)
α = 0.07 (momentum)
σ (noise)



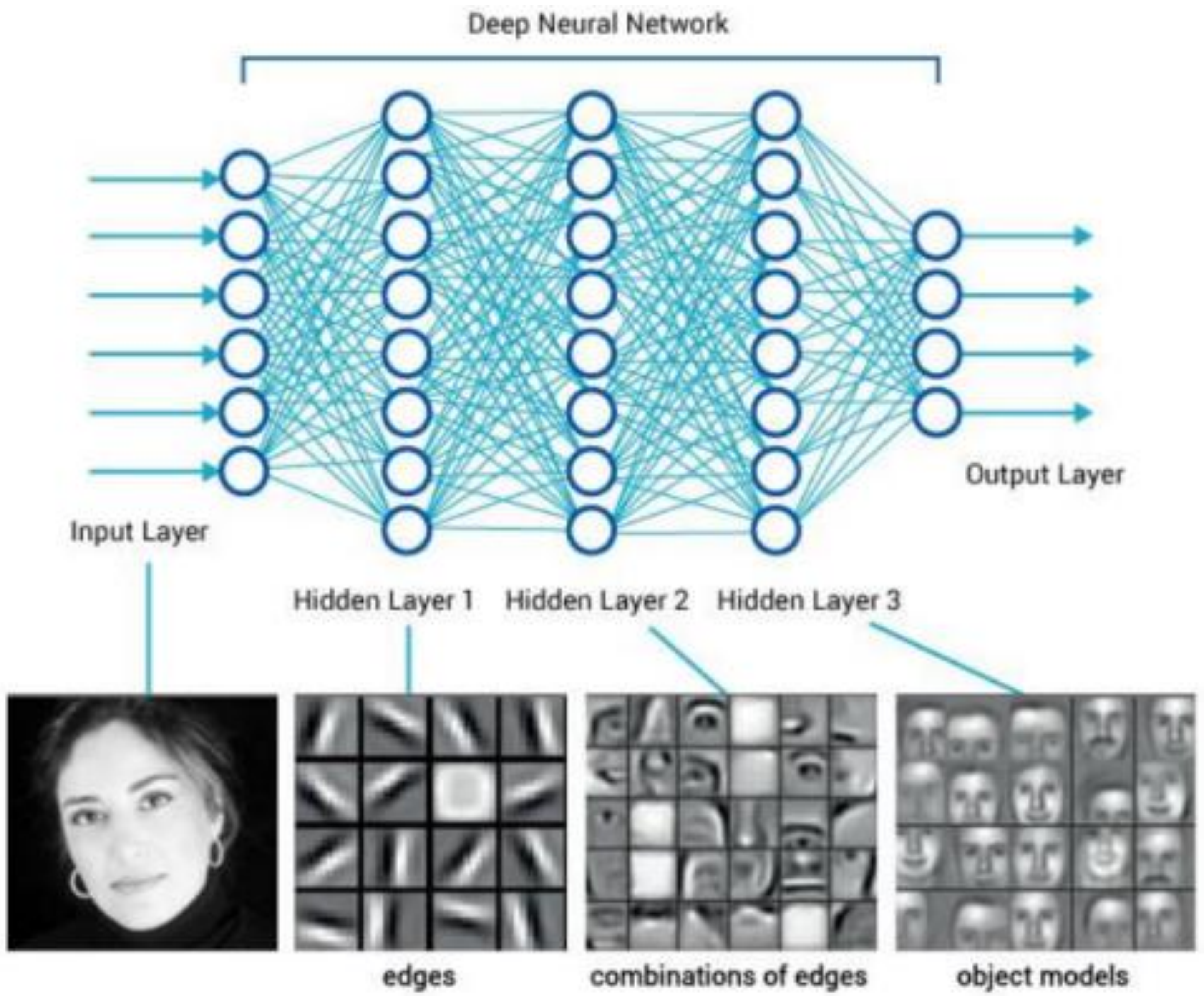
- **Three layers: ANN**
  - Input Layer # input nodes(= # input variables)
  - Hidden Layer # hidden nodes (= 1~2 × # input variables)
  - Output Layer # 1 output node

人工神经网络类型



A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)

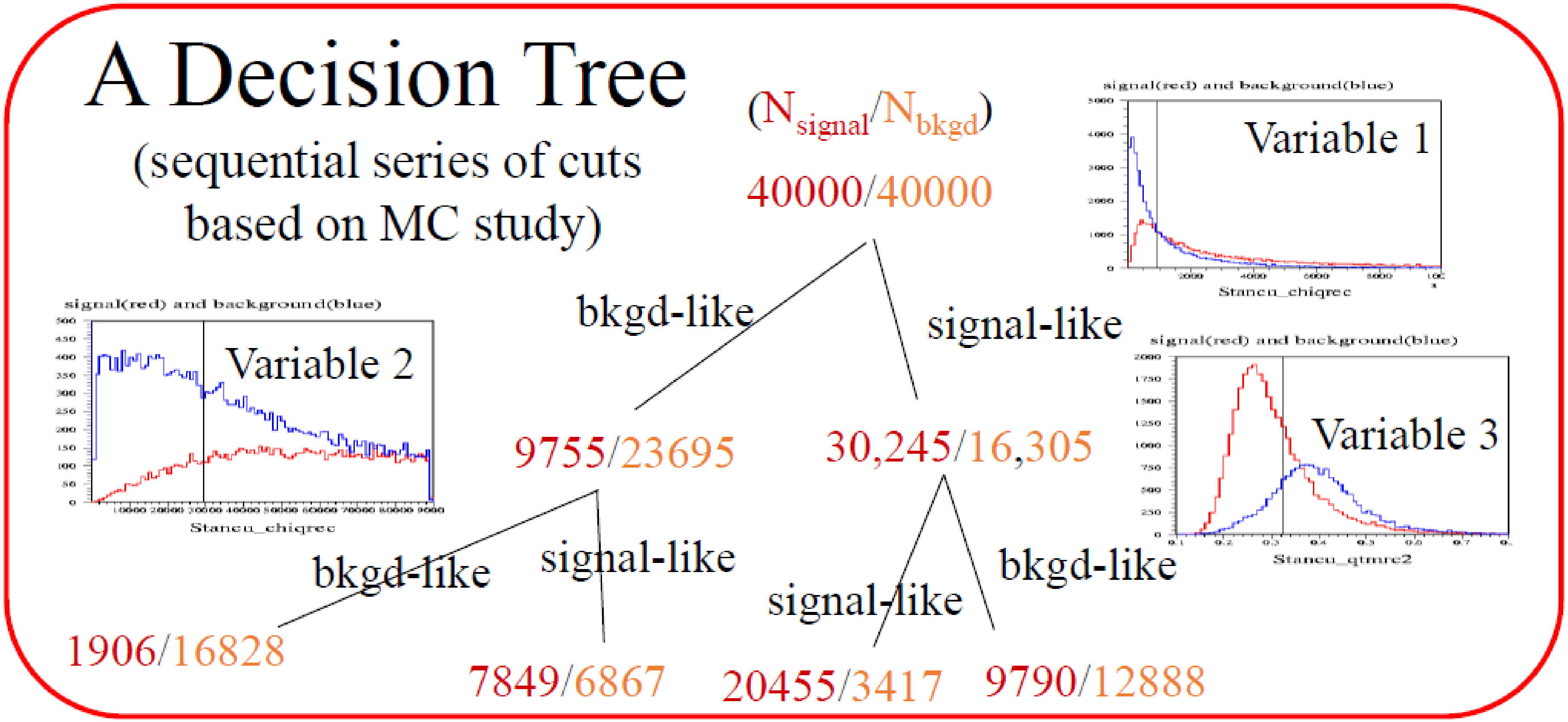
多层人工神经网络 ➔ 深度学习



决策树：A Decision Tree

➔Decision Trees have been available about three decades, they are known to be powerful but unstable, i.e., a small change in the training sample can give a large change in the tree and results.

L. Breiman, J.H. Friedman, R.A. Olshen, C.J.Stone, "Classification and Regression Trees", Wadsworth, 1983.



如何挑选最佳变量

- **Purity** ***P***, is the fraction of the weight of a node (leaf) due to signal events.
- **Gini Index**: Note that Gini index is 0 for all signal or all background.

$$Gini = \left(\sum_{i=1}^{n} W_i\right) P(1-P)$$

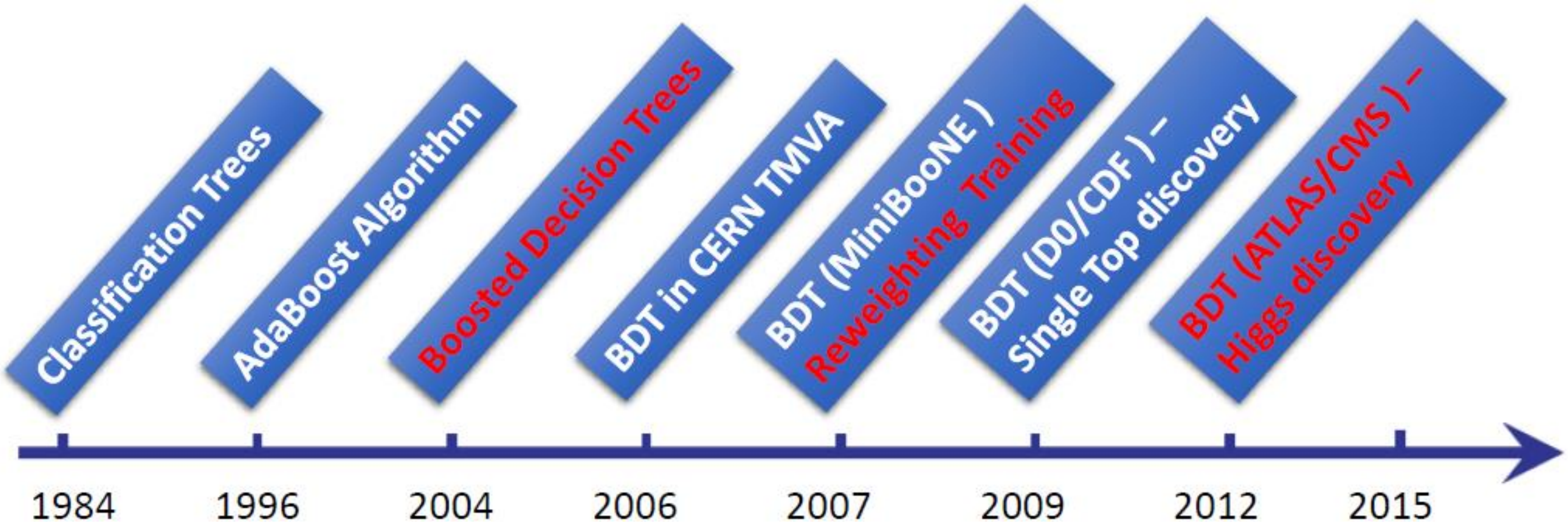
- **The criterion** is to minimize

  Gini_left_node + Gini_right_node.

- Pick the node to maximize the change in Gini index. Criterion =

  $\text{Gini}_{\text{parent\_node}} - \text{Gini}_{\text{right\_child\_node}} - \text{Gini}_{\text{left\_child\_node}}$

Boosted Decision Trees (BDT)



➔1984. L. Breiman, J.H. Friedman, R.A. Olshen, C.J.Stone, “Classification and Regression Trees”, Wadsworth, 1984. **(首次提出 Classification Trees概念)**

➔ 1996. Ref: Y. Freund, R.E. Schapire, “Experiments with a new boosting algorithm”, Proceedings of COLT, ACM Press, New York, 1996, pp. 209-217. **(首次提出AdaBoost算法)**

➔2004. 本人和Byron P. Roe, Ji Zhu首次把Boosting算法和Decision Trees结合，提出Boosted Decision Trees (BDT)，作为通讯作者发表4篇论文，为BDT应用于粒子物理实验数据分析做出了开创性的贡献。BDT广泛应用于希格斯粒子的发现和性质测量及新物理寻找等，如ATLAS, CMS, LHCb, MiniBooNE, CDF, D0, BarBar, BESIII, AMS, IceCube, PandaX 等等.

Boosted Decision Trees (BDT)

CERN TMVA 软件包收入，http://tmva.sourceforge.net/

Available online at www.sciencedirect.com
SCIENCE DIRECT
Nuclear Instruments and Methods in Physics Research A 543 (2005) 577–584
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

Boosted decision trees as an alternative to artificial neural networks for particle identification

Byron P. Roe[a], Hai-Jun Yang[a,*], Ji Zhu[b], Yong Liu[c], Ion Stancu[c], Gordon McGregor[d]

[a] Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
[b] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
[c] Department of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35487, USA
[d] Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Available online at www.sciencedirect.com
SCIENCE DIRECT
Nuclear Instruments and Methods in Physics Research A 555 (2005) 370–385
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

Studies of boosted decision trees for MiniBooNE particle identification

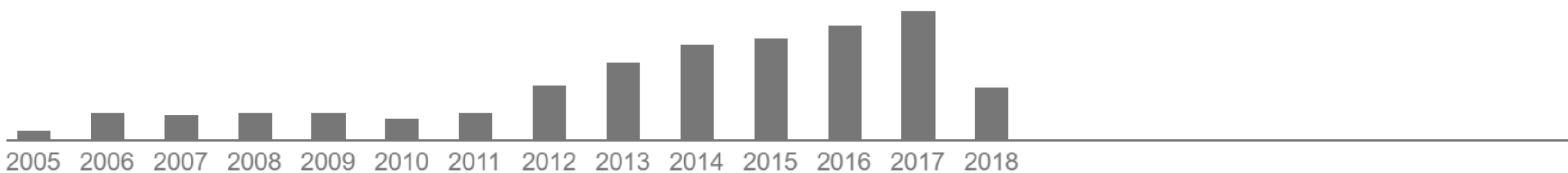
Hai-Jun Yang[a,c,*], Byron P. Roe[a], Ji Zhu[b]

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

Received 8 August 2005; received in revised form 12 September 2005; accepted 16 September 2005
Available online 4 October 2005

Abstract

Boosted decision trees are applied to particle identification in the MiniBooNE experiment operated at Fermi National Accelerator Laboratory (Fermilab) for neutrino oscillations. Numerous attempts are made to tune the boosted decision trees, to compare performance of various boosting algorithms, and to select input variables for optimal performance.



has been widely used in data analysis of High Energy Physics experiments in the last decade. The use of the ANN technique usually gives better

MiniBooNE experiment [1] at Fermi National Accelerator Laboratory. The MiniBooNE experiment is designed to confirm or refute the evidence for $\nu_\mu \to \nu_e$ oscillations at $\Delta m^2 \simeq 1\,\mathrm{eV}^2/c^4$ found by the LSND experiment [2]. It is a crucial experiment which will imply new physics beyond

*Corresponding author.
E-mail address: yhj@umich.edu (Hai-Jun Yang).

Initial comparisons of these techniques with artificial neural networks (ANN) using the MiniBooNE MC

*Corresponding author. Tel.: +1 734 764 3407; fax: +1 734 936 6529.
E-mail address: yhj@umich.edu (H.-J. Yang).

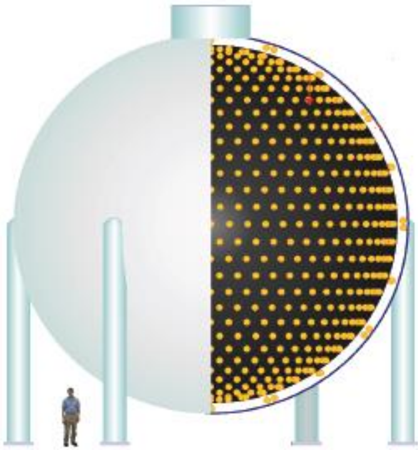
0168-9002/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2005.09.022

procedure that combines many "weak" classifiers to achieve a final powerful classifier. In the present work numerous trials are made to tune the boosted decision trees, and comparisons are made for various algorithms. For a large number of discriminant variables, several

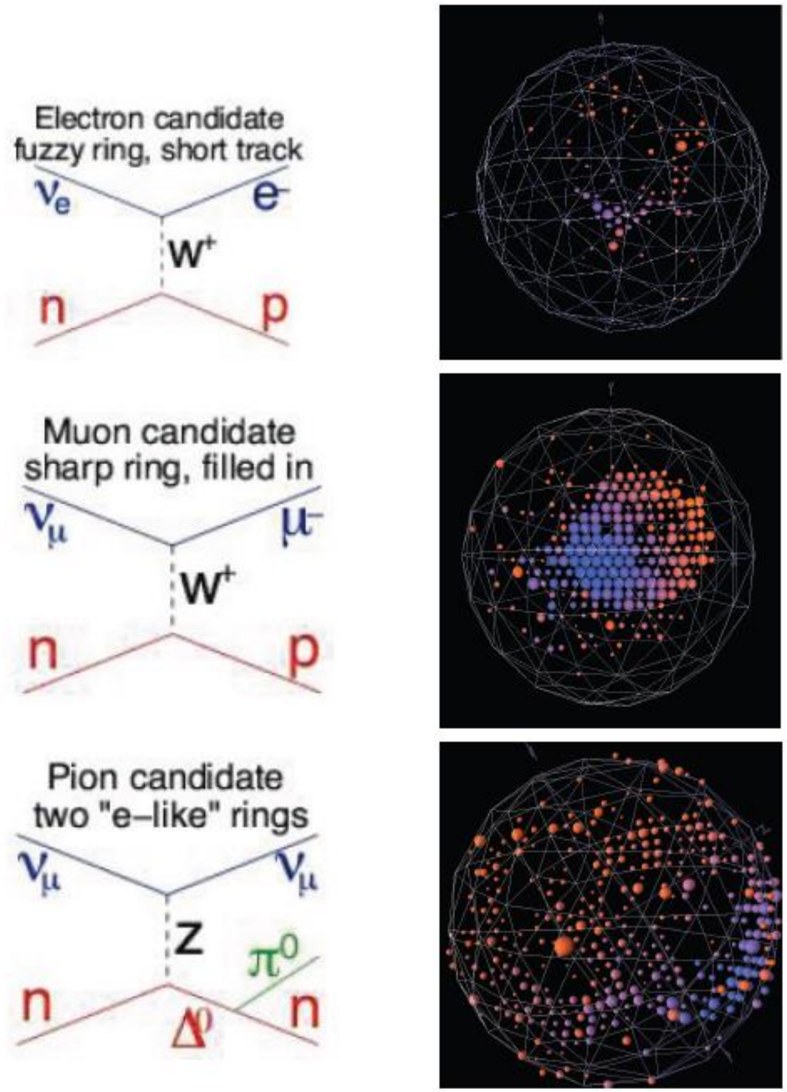
早期应用于MiniBooNE实验

Detector is a 12-m diameter tank of mineral oil exposed to a beam of neutrinos and viewed by 1520 photomultiplier tubes:

MiniBooNE Detector

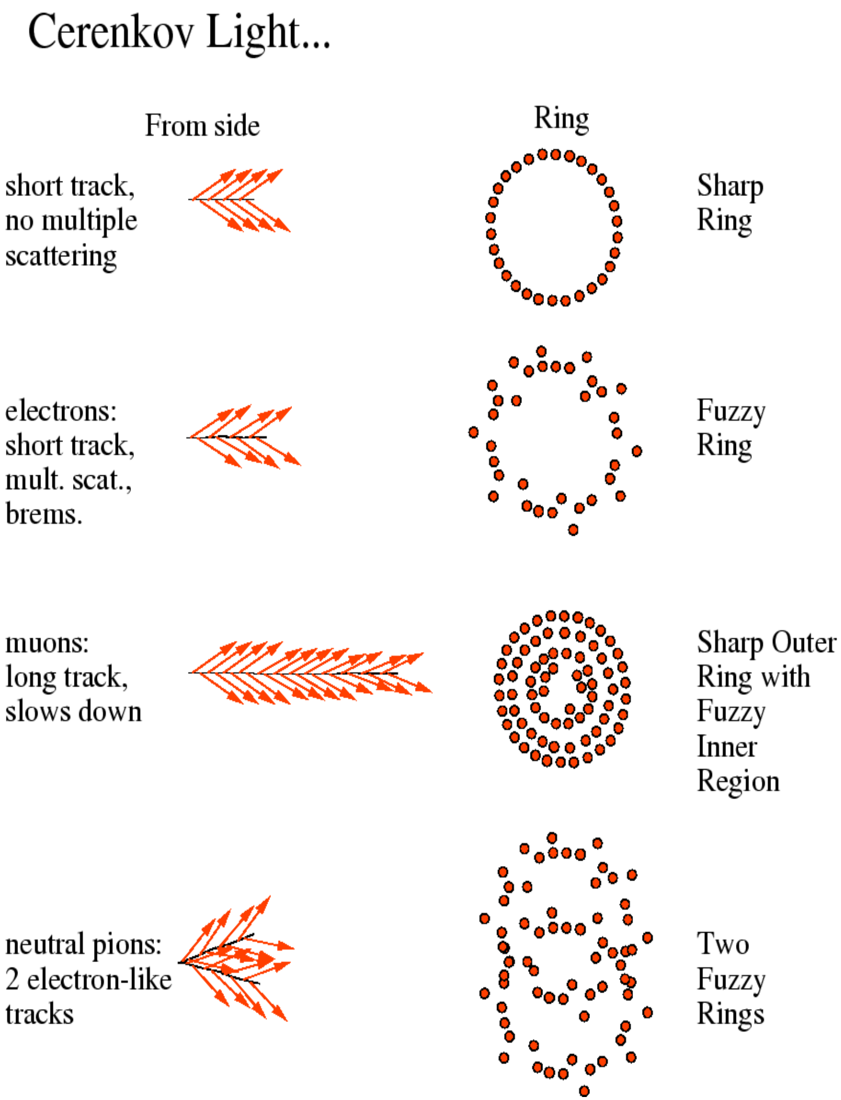


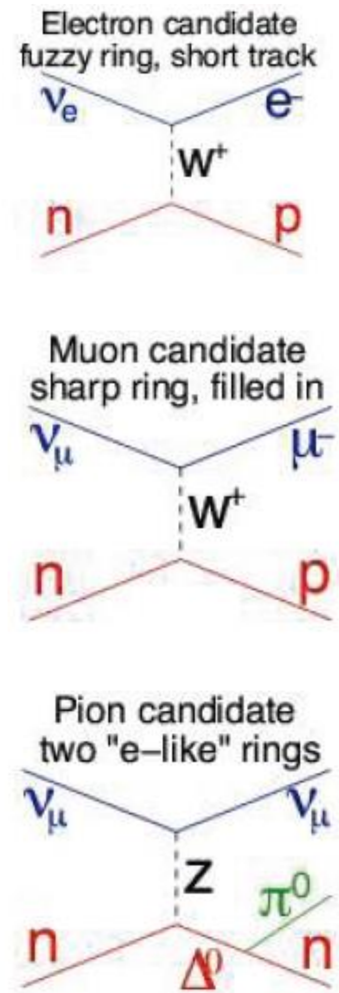
Search for ν_μ to ν_e oscillations required particle i.d. using information from the PMTs.

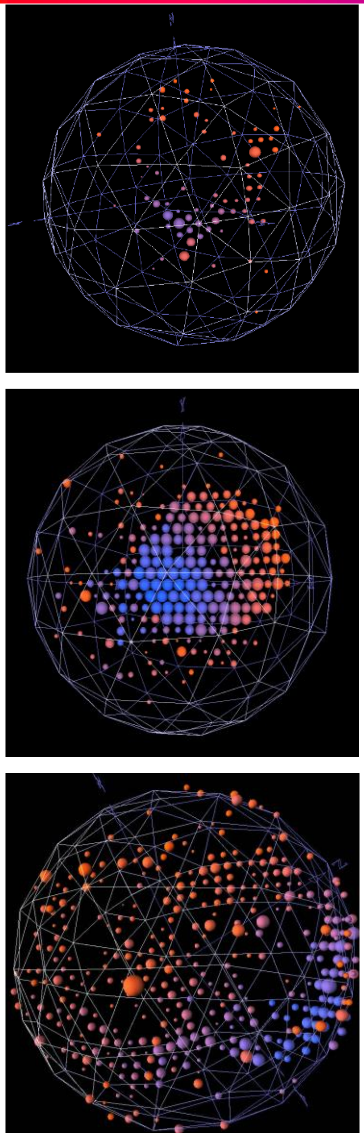


H.J. Yang, MiniBooNE PID, DNP06

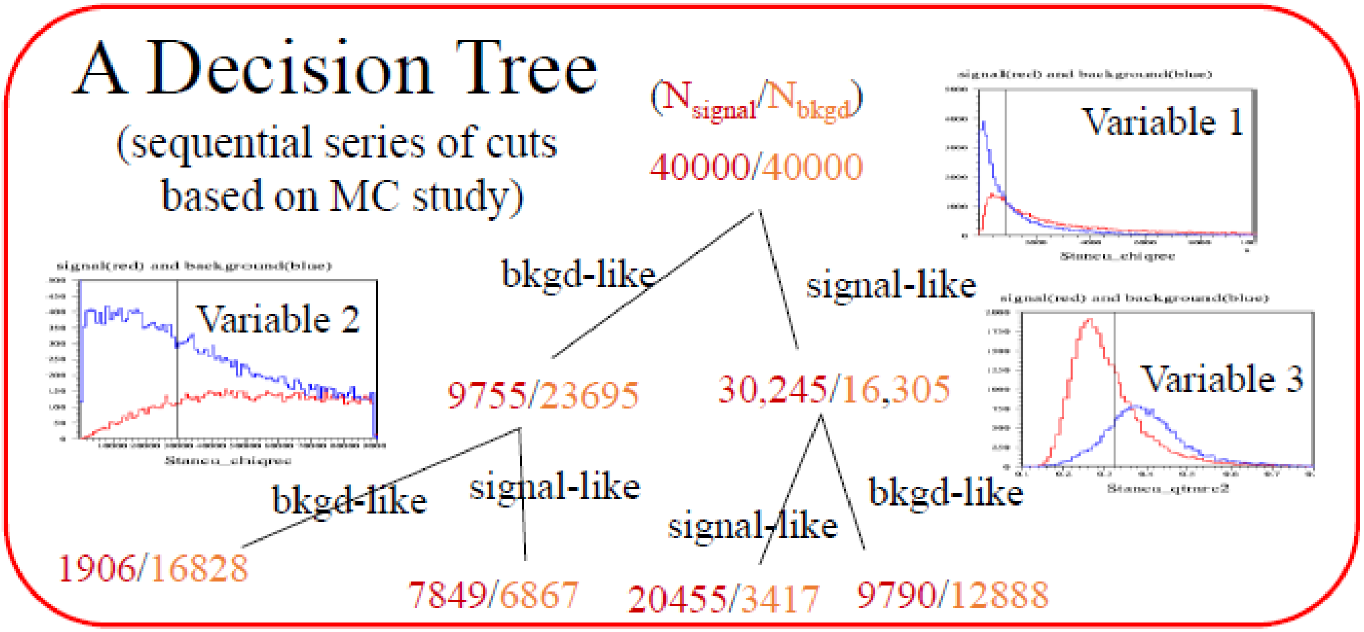
模式识别：MB事例特征


Cerenkov Light...
From side
Ring
short track,
no multiple
scattering
Sharp
Ring
electrons:
short track,
mult. scat.,
brems.
Fuzzy
Ring
muons:
long track,
slows down
Sharp Outer
Ring with
Fuzzy
Inner
Region
neutral pions:
2 electron-like
tracks
Two
Fuzzy
Rings

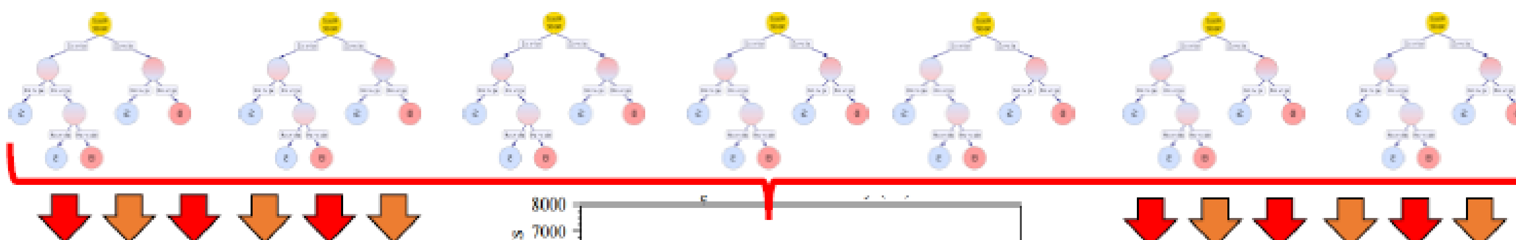

Electron candidate
fuzzy ring, short track
ν_e
e^-
W^+
n
p
Muon candidate
sharp ring, filled in
ν_μ
μ^-
W^+
n
p
Pion candidate
two "e–like" rings
ν_μ
ν_μ
Z
π^0
n
Δ^0
n



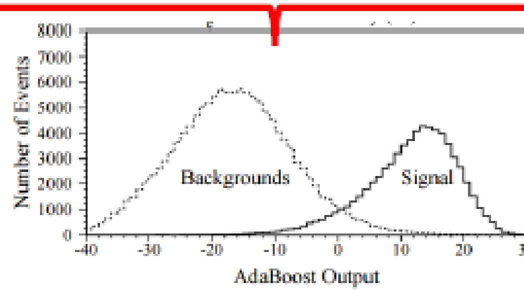
Boosted Decision Trees (BDT)



通过Boosting 算法不断提高误判事例的权重，产生一系列Decision Trees



把每个事例在所有Decision Trees获得的积分累加，通过“Majority vote”方法提高性能和稳定性。



通过Boosting不断提高误判事例的权重，使得这些难以区分的事例在后续的Decision Trees获得的正确区分，提高效率。

Boosting Algorithms

➔ 1996. Ref: Y. Freund, R.E. Schapire, "Experiments with a new boosting algorithm", Proceedings of COLT, ACM Press, New York, 1996, pp. 209-217.

- AdaBoost Algorithm:

1. Initialize the observation weights $w_i = 1/n$, i = 1, 2,..., n
2. For m = 1 to M:
   - 2.a Fit a classifier $T_m(x)$ to the training data using weights w_i
   - 2.b Compute

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq T_m(x_i))}{\sum_{i=1}^{n} w_i}$$

I = 1, if a training event is misclassified; Otherwise, I = 0

   - 2.c Compute $\alpha_m = \beta \times log((1 - err_m)/err_m)$
   - 2.d Set $w_i \leftarrow w_i \times exp(\alpha_m I(y_i \neq T_m(x_i)))$, i=1, 2,...,n
   - 2.e Re-normalize $w_i = w_i / \sum_{i=1}^{n} w_i$
3. Output T(x) $= \sum_{m=1}^{M} \alpha_m T_m(x)$

- ϵ−boosting Algorithm:

1. Initialize the observation weights $w_i = 1/n$, i = 1, 2,..., n
2. For m = 1 to M:
   - 2.a Fit a classifier $T_m(x)$ to the training data using weights w_i
   - 2.b Set $w_i \leftarrow w_i \times exp(2\epsilon I(y_i \neq T_m(x_i)))$, i=1, 2,...,n
   - 2.c Re-normalize $w_i = w_i / \sum_{i=1}^{n} w_i$
3. Output T(x) $= \sum_{m=1}^{M} \epsilon T_m(x)$

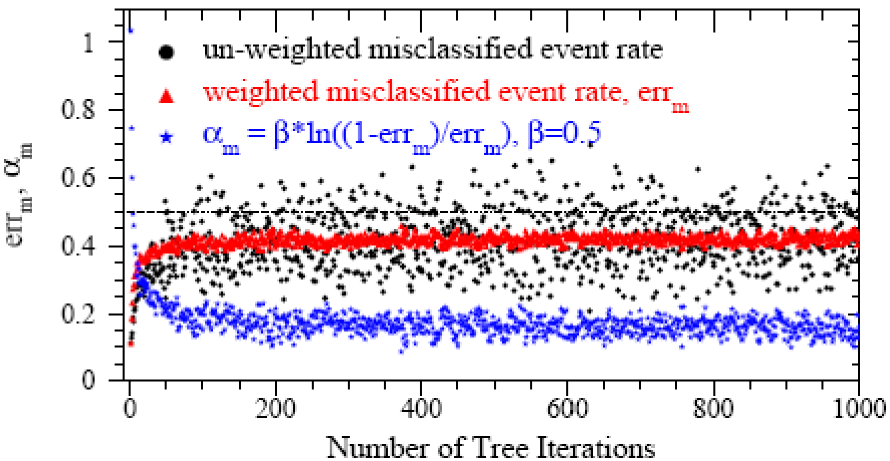
Boosting Algorithms

- **AdaBoost: the weight of misclassified events is increased by**
  - error rate=0.1 and $\beta = 0.5$, $\alpha_m = 1.1$, $\exp(1.1) = 3$
  - error rate=0.4 and $\beta = 0.5$, $\alpha_m = 0.203$, $\exp(0.203) = 1.225$
  - Weight of a misclassified event is multiplied by a large factor which depends on the error rate.
- **ε–boost: the weight of misclassified events is increased by**
  - If $\varepsilon = 0.01$, $\exp(2*0.01) = 1.02$
  - If $\varepsilon = 0.04$, $\exp(2*0.04) = 1.083$
  - It changes event weight a little at a time.

➔ AdaBoost converges faster than ε-boost. However, the performance of AdaBoost and ε–boost are very comparable with sufficient tree iterations.

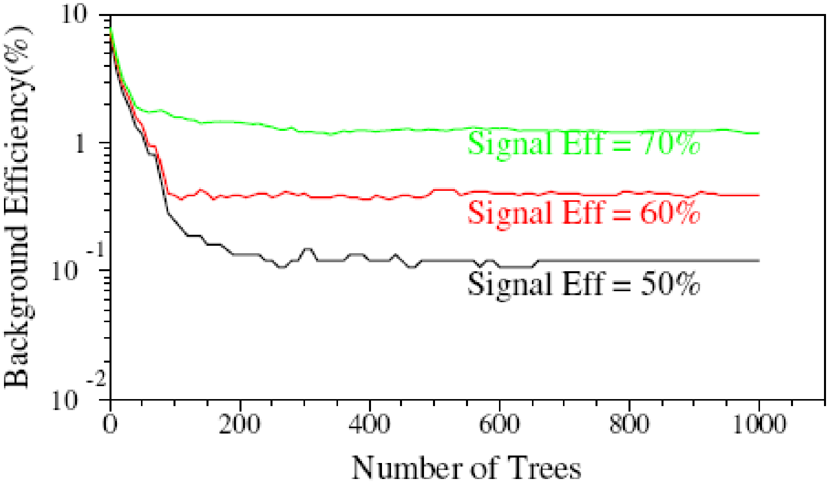
弱分类器 ➔ 强分类器

Decision Tree ➔ Boosted Decision Trees

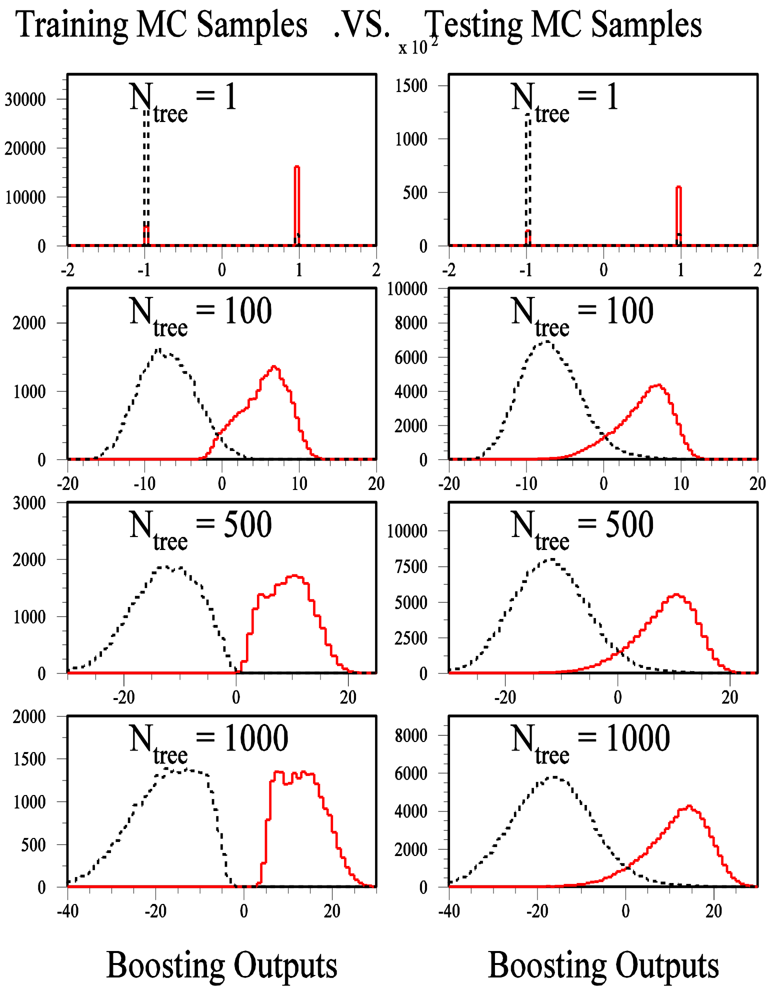


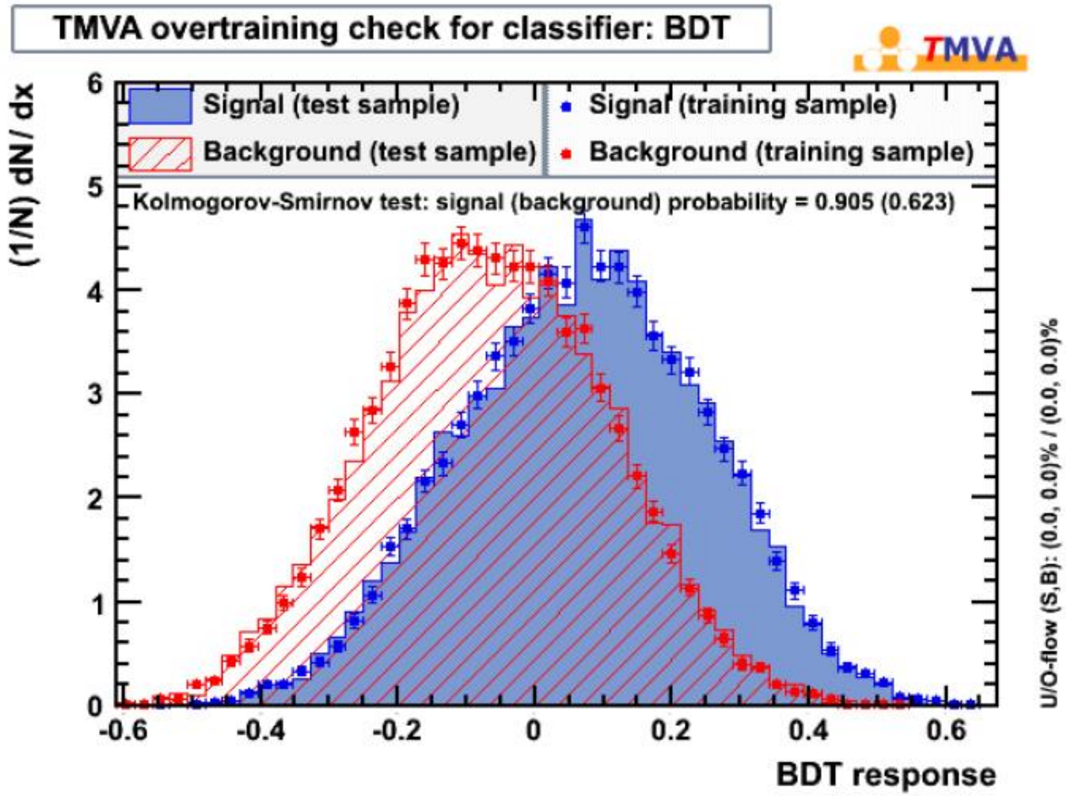
➔ **BDT逐渐提高误判事例的权重，经过几十次后，单个决策树的分类效果(misclassification event rate)比较差，接近40-45%，属于弱分类器。**

➔**BDT的优势是把所有的决策树整合在一起，即把弱分类器 ("weak" classifiers) 变成强分类器。整合大约200个决策树后，BDT 的性能趋于稳定。**



性能估计：训练样本 vs 测试样本





➔ ANN / BDT 性能估计必须用统计独立的样本测试！

➔ 训练时要避免过度训练！

TMVA 软件包

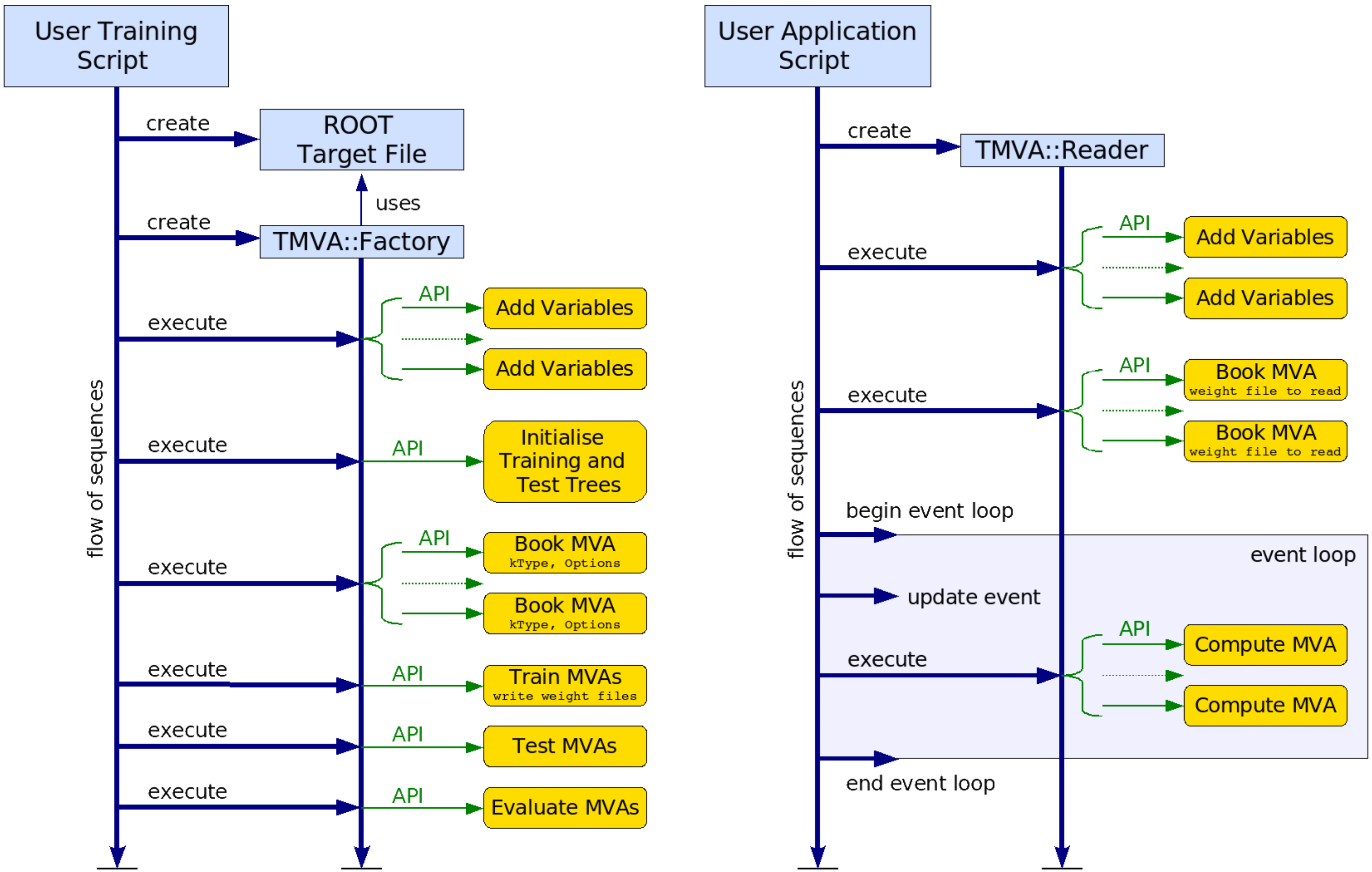
- 基于CERN ROOT的多变量数据分析软件包

  http://tmva.sourceforge.net/， https://root.cern.ch/tmva

- 软件包内的机器学习算法包括：

- Rectangular cut optimisation
- Projective likelihood estimation (PDE approach)
- Multidimensional probability density estimation (PDE - range-search approach)
- Multidimensional k-nearest neighbour classifier
- Linear discriminant analysis (H-Matrix and Fisher discriminants)
- Function discriminant analysis (FDA)
- Artificial neural networks (three different implementations)
- Boosted/Bagged decision trees
- Predictive learning via rule ensembles (RuleFit)
- Support Vector Machine (SVM)

TMVA 训练和分析应用流程



ROOT script for Training

```
void TMVAnalysis( )
{
  TFile* outputFile = TFile::Open( "TMVA.root", "RECREATE" );

  TMVA::Factory *factory = new TMVA::Factory( "MVAnalysis", outputFile,"!V");

  TFile *input = TFile::Open("tmva_example.root");

  factory->AddSignalTree        ( (TTree*)input->Get("TreeS"), 1.0 );
  factory->AddBackgroundTree ( (TTree*)input->Get("TreeB"), 1.0 );

  factory->AddVariable("var1+var2", 'F');
  factory->AddVariable("var1-var2", 'F');
  factory->AddVariable("var3", 'F');
  factory->AddVariable("var4", 'F');

  factory->PrepareTrainingAndTestTree("", "NSigTrain=3000:NBkgTrain=3000:SplitMode=Random:!V" );

  factory->BookMethod( TMVA::Types::kLikelihood, "Likelihood",
                                "!V:!TransformOutput:Spline=2:NSmooth=5:NAvEvtPerBin=50" );

  factory->BookMethod( TMVA::Types::kMLP, "MLP", "!V:NCycles=200:HiddenLayers=N+1,N:TestRate=5" );

  factory->TrainAllMethods();
  factory->TestAllMethods();
  factory->EvaluateAllMethods();

  outputFile->Close();
  delete factory;
}
```

create *Factory*

give training/test trees

register input variables

select MVA methods

train, test and evaluate

ROOT script for Application

```
void TMVApplication( )
{
  TMVA::Reader *reader = new TMVA::Reader("!Color");

  Float_t var1, var2, var3, var4;
  reader->AddVariable( "var1+var2", &var1 );
  reader->AddVariable( "var1-var2",  &var2 );
  reader->AddVariable( "var3", &var3 );
  reader->AddVariable( "var4", &var4 );

  reader->BookMVA( "MLP classifier",  "weights/MVAnalysis_MLP.weights.txt" );

  TFile *input = TFile::Open("tmva_example.root");
  TTree* theTree = (TTree*)input->Get("TreeS");

  // … set branch addresses for user TTree
  for (Long64_t ievt=3000; ievt<theTree->GetEntries();ievt++) {
    theTree->GetEntry(ievt);

    var1 = userVar1 + userVar2;
    var2 = userVar1 - userVar2;
    var3 = userVar3;
    var4 = userVar4;

    Double_t out =  reader->EvaluateMVA(  "MLP classifier"  );

    // do something with it …
  }
  delete reader;
}
```

create *Reader*

register the variables

book classifier(s)

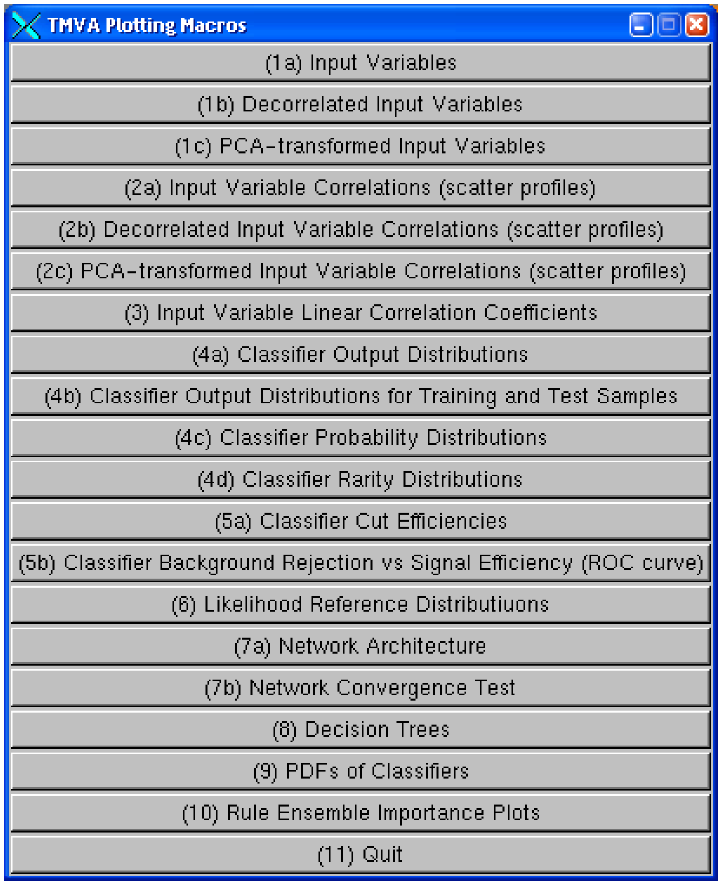
prepare event loop

compute input variables

calculate classifier output

MVA Evaluation Framework

- After training, TMVA provides ROOT evaluation scripts (through GUI)



Plot all signal (S) and background (B) input variables with and without pre-processing

Correlation scatters and linear coefficients for S & B

Classifier outputs (S & B) for test and training samples (spot overtraining)

Classifier *Rarity* distribution

Classifier significance with optimal cuts

B rejection versus S efficiency

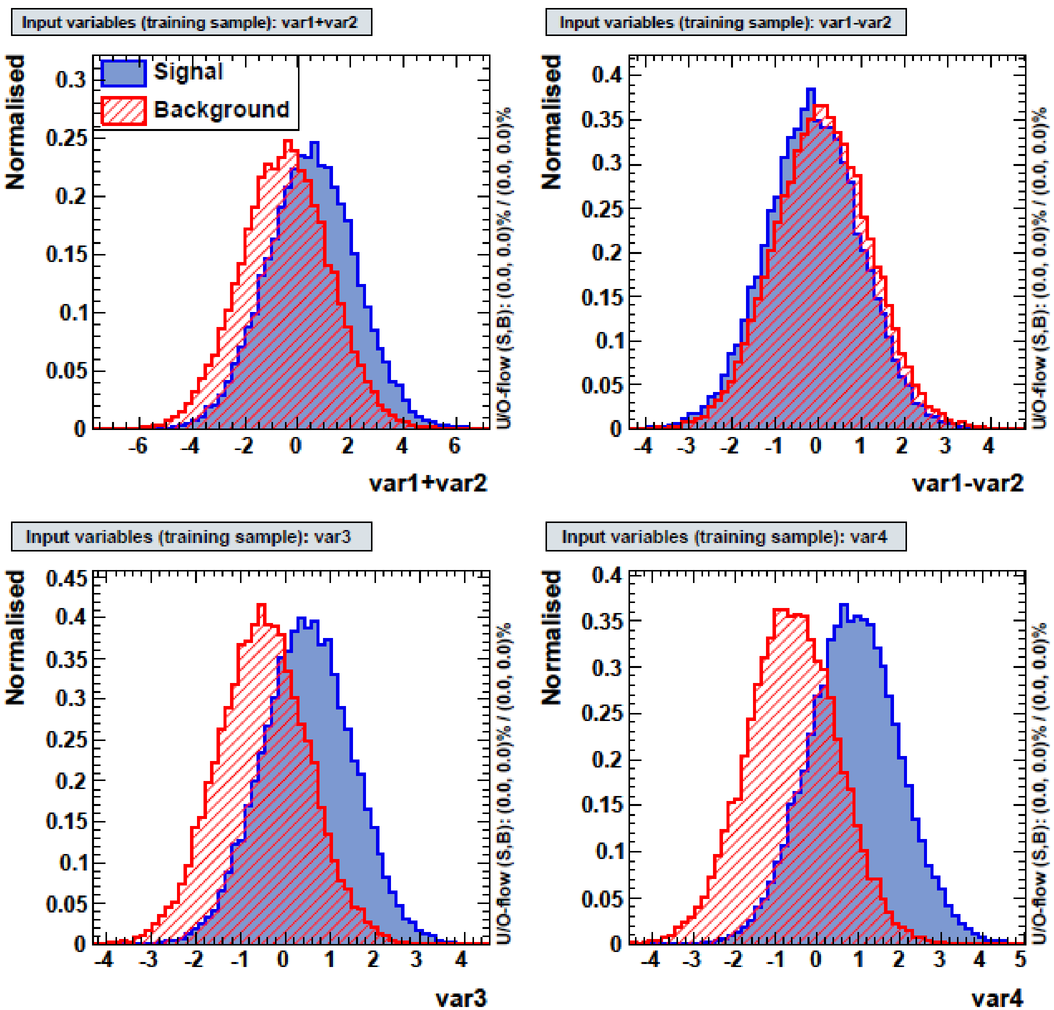
Classifier-specific plots:

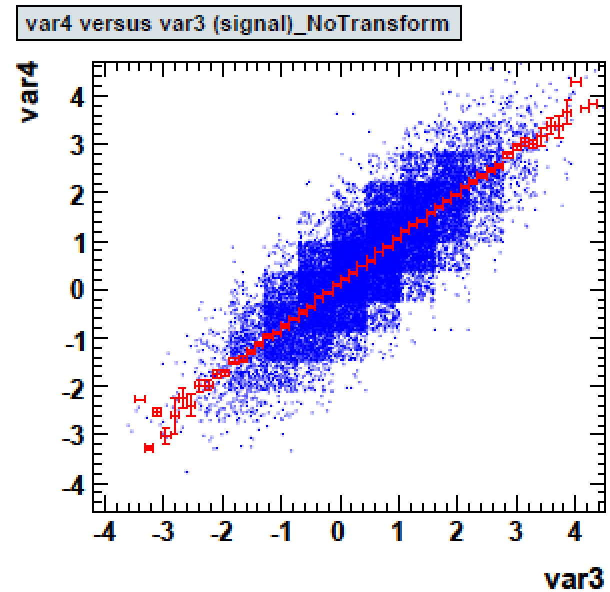
- Likelihood reference distributions
- Classifier PDFs (for probability output and Rarity)
- Network architecture, weights and convergence
- Rule Fitting analysis plots

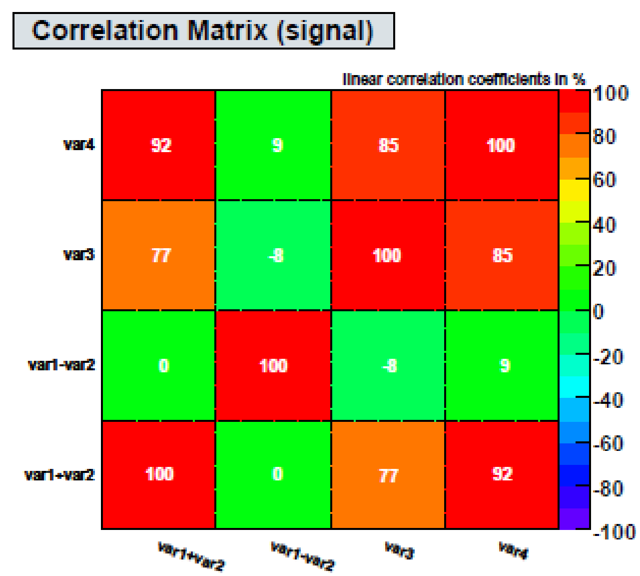
- Visualise decision trees

TMVA应用例子

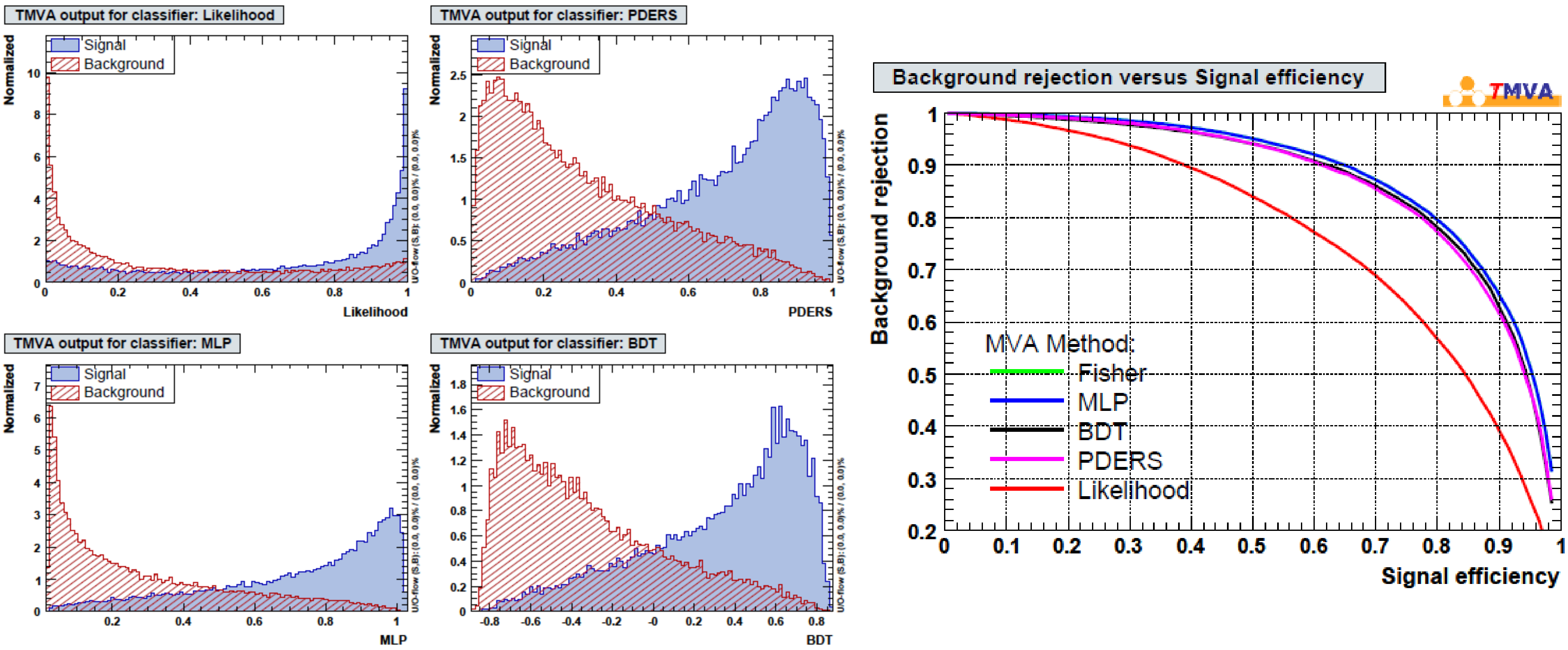
四个输入变量　　变量关联 ➔







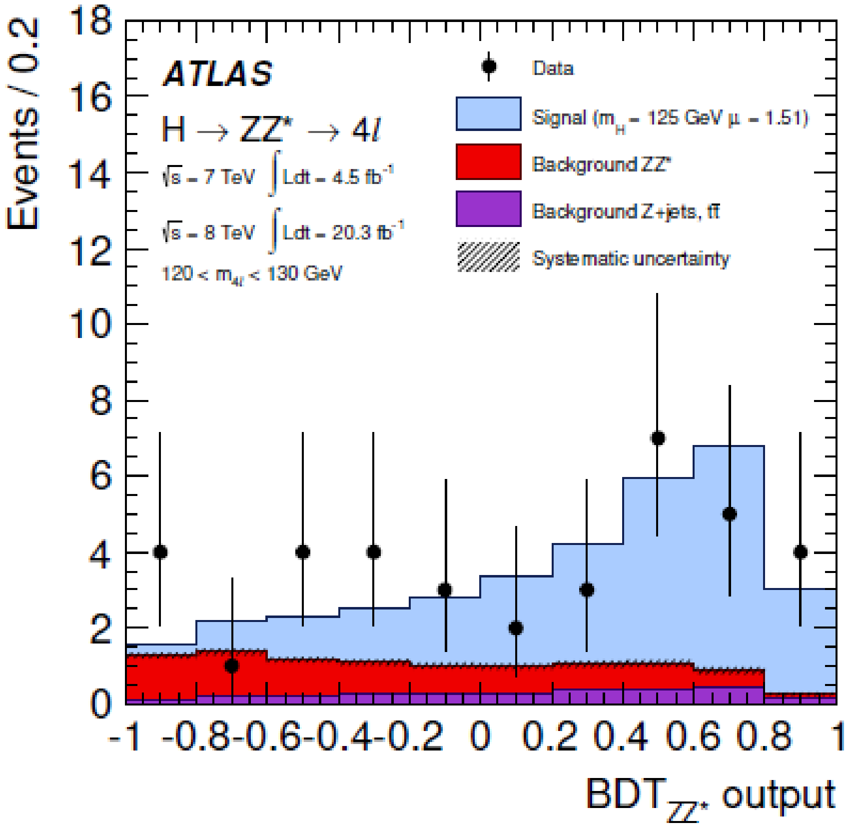
TMVA应用例子



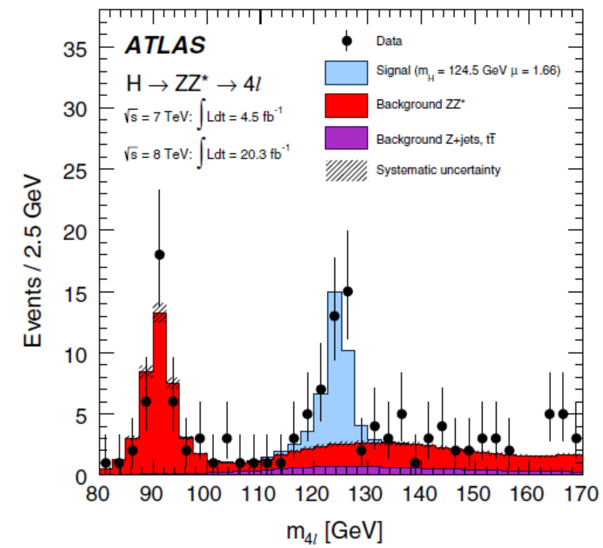
TMVA output for classifier: Likelihood
Signal
Background
Normalized
Likelihood
TMVA output for classifier: PDERS
PDERS
TMVA output for classifier: MLP
MLP
TMVA output for classifier: BDT
BDT
Background rejection versus Signal efficiency
TMVA
Background rejection
Signal efficiency
MVA Method:
Fisher
MLP
BDT
PDERS
Likelihood

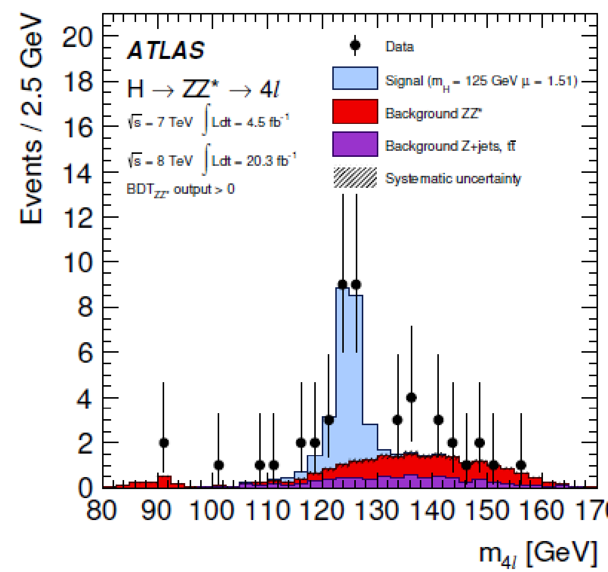
提高希格斯粒子的发现潜力

BDT 方法广泛应用于希格斯粒子的发现质量、自旋和宇称、截面等性质测量



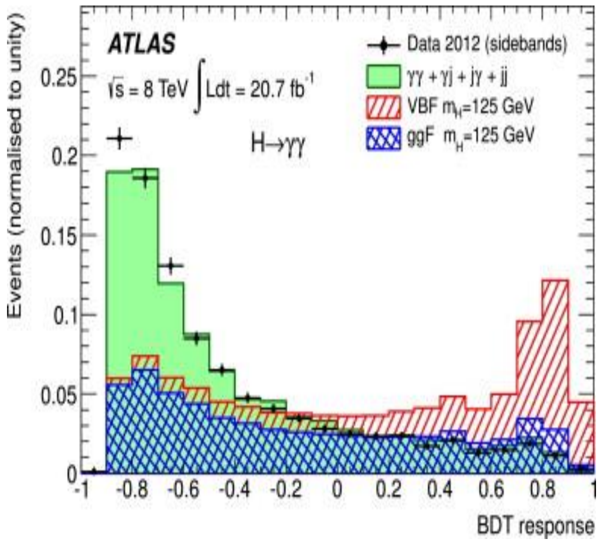
$BDT_{ZZ} > 0$

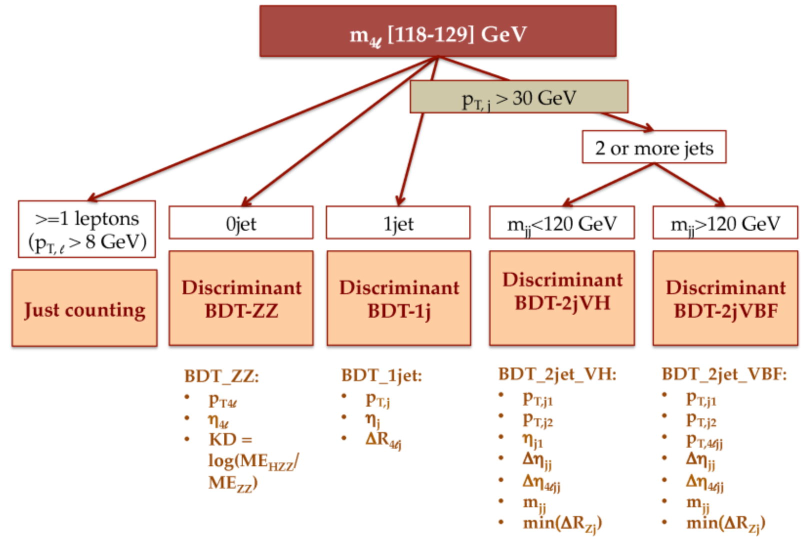


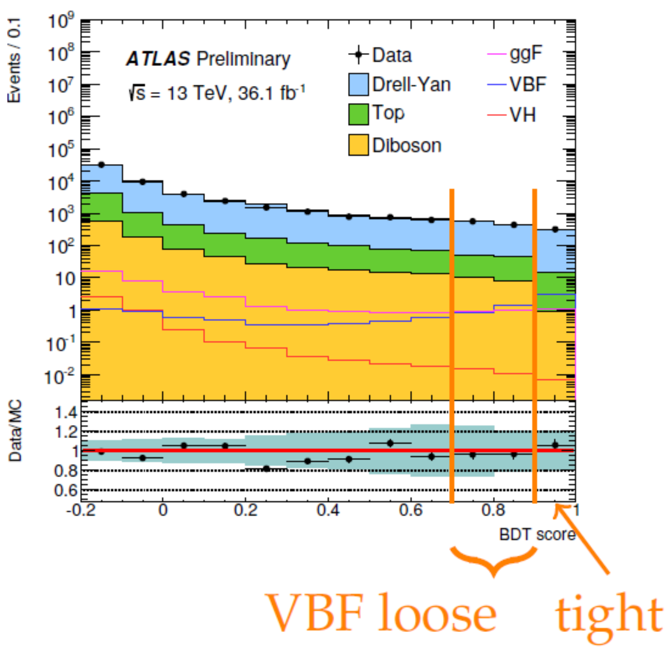


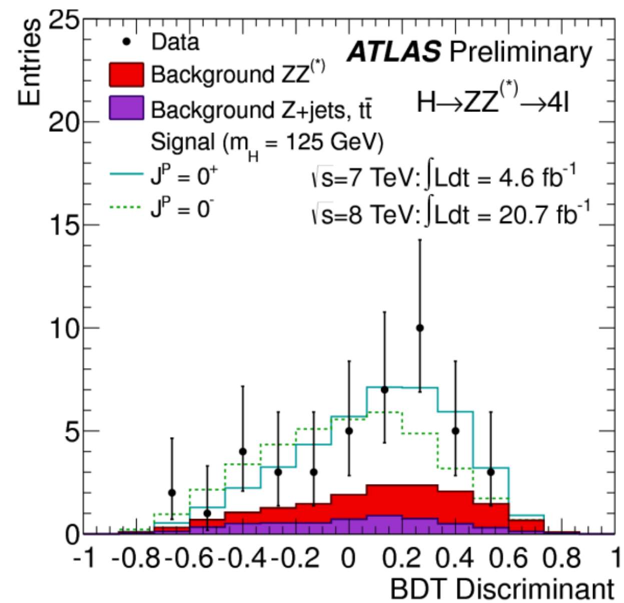
希格斯的探测灵敏度显著提高 ($5\sigma \rightarrow 6.2\sigma$)

提高希格斯粒子性质的测量精度









Tutorial

- ssh –Y inpactest@bl-1-1.physics.sjtu.edu.cn
- passwd: inpac123456
- cd Erec_tuto/tmva
- less TMVARegression.C
- Please follow the ‘tutorial_Erec.pdf’

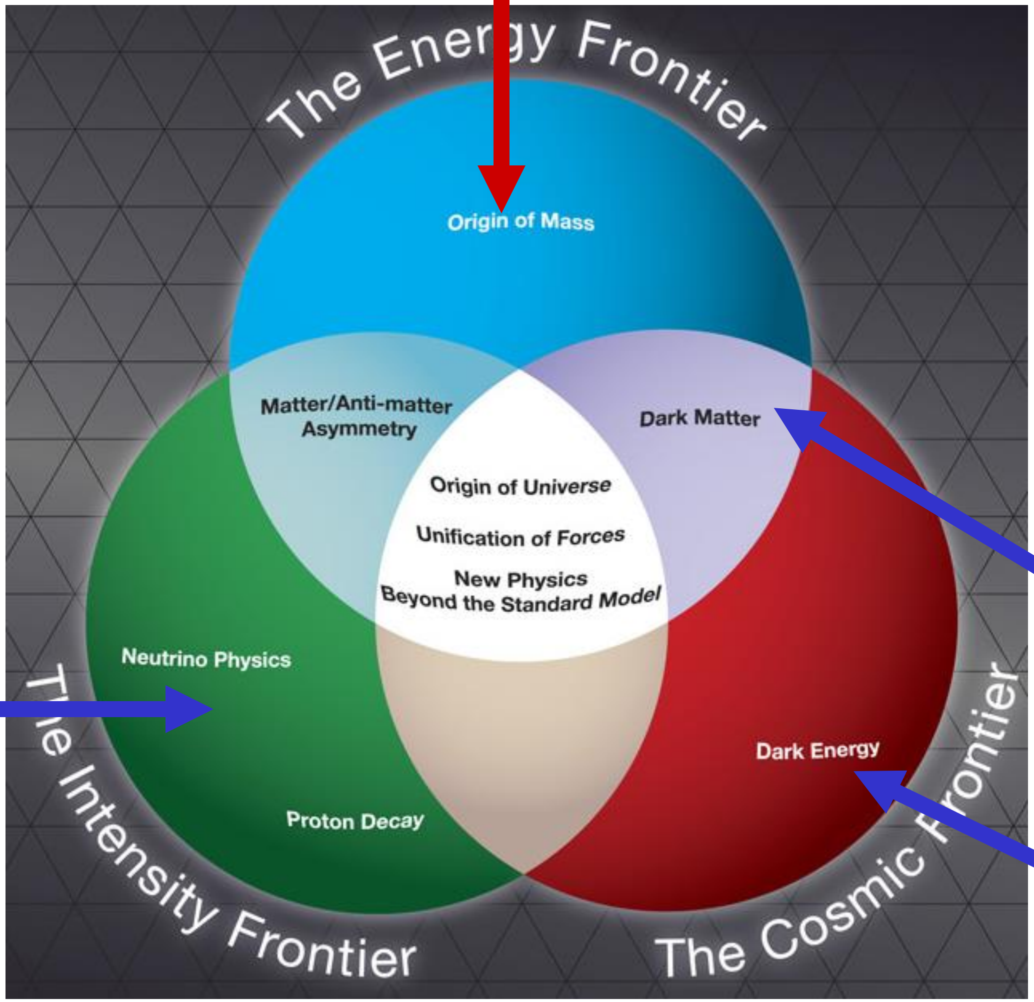
机器学习方法有非常广泛的应用！

粒子物理数据分析、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别、机器人等领域

谢谢大家！

研究物质世界的三大前沿

高能量前沿 (交大粒子所参与LHC/ATLAS实验)



交大粒子所主导
PandaX暗物质实验

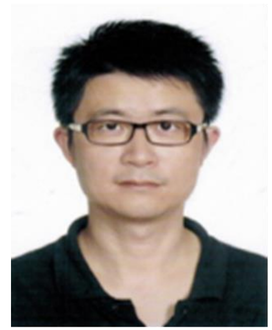
高强度前沿

交大粒子所参与
大亚湾和江门
中微子实验

交大天文系
宇宙学前沿
研究

交大对撞机实验团队成员

杨海军 教授、青千（所长）
瑞士联邦理工大学和高能所
联培博士，美国密西根大学
博士后和研究科学家
ATLAS，BESIII，CEPC
电话：15800893756
haijun.yang@sjtu.edu.cn



李亮 副教授、青千
美国威斯康辛大学博士
美国加州大学博士后
和研究科学家
ATLAS，Muon g-2，CEPC
电话：18016387798
liangliphy@sjtu.edu.cn

郭军 副教授、青千
纽约州立大学石溪分校博士
美国哥伦比亚大学博士后
ATLAS，CEPC
电话：18217103176
jun.guo@sjtu.edu.cn

周宁 副教授、青千
美国哥伦比亚大学博士
美国加州大学博士后
ATLAS，PandaX
电话：13918560945
nzhou@sjtu.edu.cn

杨勇 副教授、青千
美国加州理工学院博士
瑞士苏黎世大学博士后
ATLAS，PandaX
电话：13918787204
yong.yang@sjtu.edu.cn

李数 副教授
法国马赛大学和科大博士
美国杜克大学博士后
ATLAS，CEPC
电话：15010392595
shu.li@cern.ch

Boosted Decision Trees (BDT)

- ➔ **2004/8/30, arXiv:physics/0408124, [Nucl.Instrum.Meth. A543 (2005) 577-584]**
  Byron P. Roe, **Hai-Jun Yang*,** Ji Zhu, Yong Liu, Ion Stancu, Gordon McGregor,
  "Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification"
- ➔ **2005/8/8, arXiv:physics/0508045, [Nucl.Instrum.Meth. A555 (2005) 370-385]**
  Hai-Jun.Yang*, Byron P. Roe, Ji Zhu,
  "Studies of Boosted Decision Trees for MiniBooNE Particle Identification"
- ➔ **2006/10/31, arXiv:physics/0610276, [Nucl. Instrum. & Meth. A 574 (2007) 342-349]**
  Hai-Jun Yang*, Byron P. Roe, Ji Zhu,
  "Studies of Stability and Robustness for Artificial Neural Networks and Boosted Decision Trees"
- ➔ **2007/8/27, arXiv:0708.3635, [JINST3:P04004,2008]**
  Hai-Jun Yang*, Tiesheng Dai, Alan Wilson, Zhengguo Zhao, Bing Zhou,
  "A Multivariate Training Technique with Event Reweighting"

希格斯新产生模式的发现（2018）

- ❑ New production mode – ttH, H→ γγ
- ❑ ttH产生态双光子测量道的动画演示