Machine Learning Course 2 and tutorial introduction

David Rousseau

IJCLab-Orsay

david.rousseau@in2p3.fr @dhpmrou

CHACAL, Johannesburg, Jan 2024

Introduction to classification

Given x, we want y ➔ how to build f ?

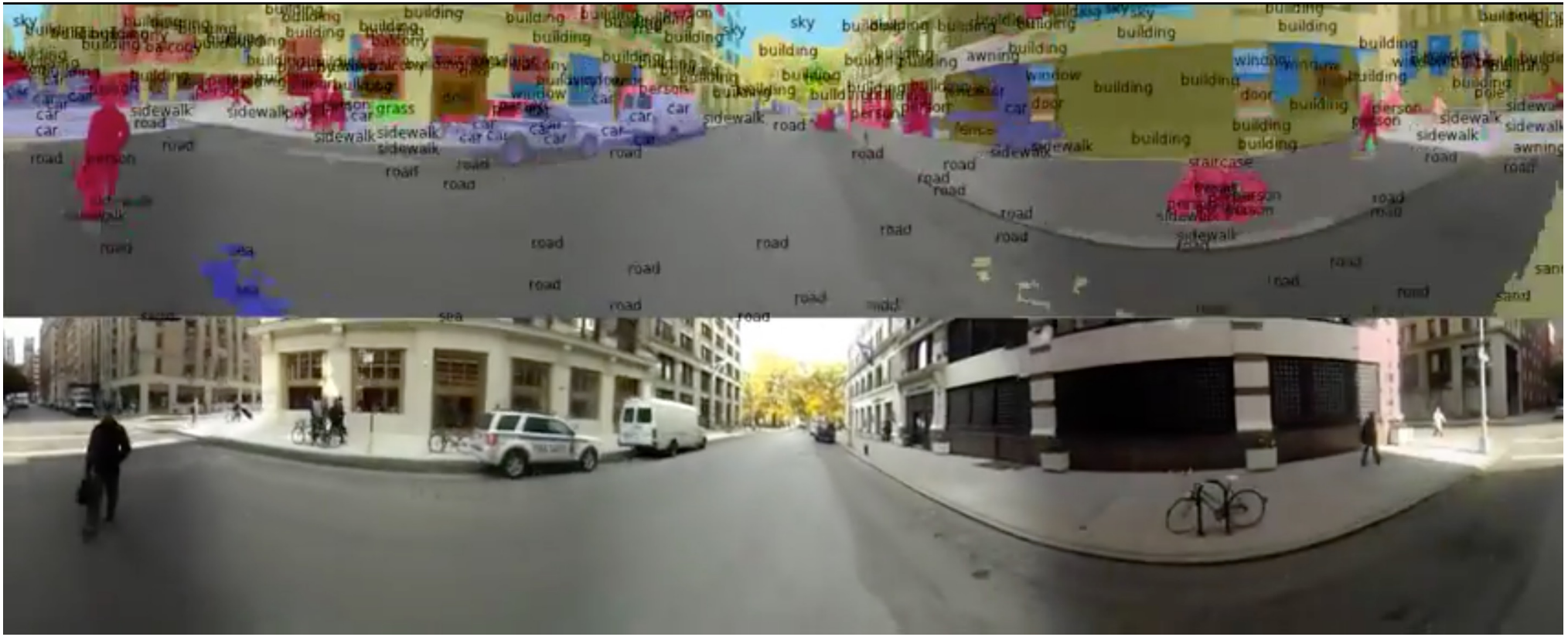
x f y

- Written text ➔ text
- Picture ➔ mom or granny?
- Image ➔ cat or dog?

Classification

- « Comment ça va ? » ➔ « Wie geht's ? »
- Speach ➔ text
- Stone positions ➔ next move
- Camera +GPS ➔steering action
- FB account details ➔ targetted ads

Classification is everywhere



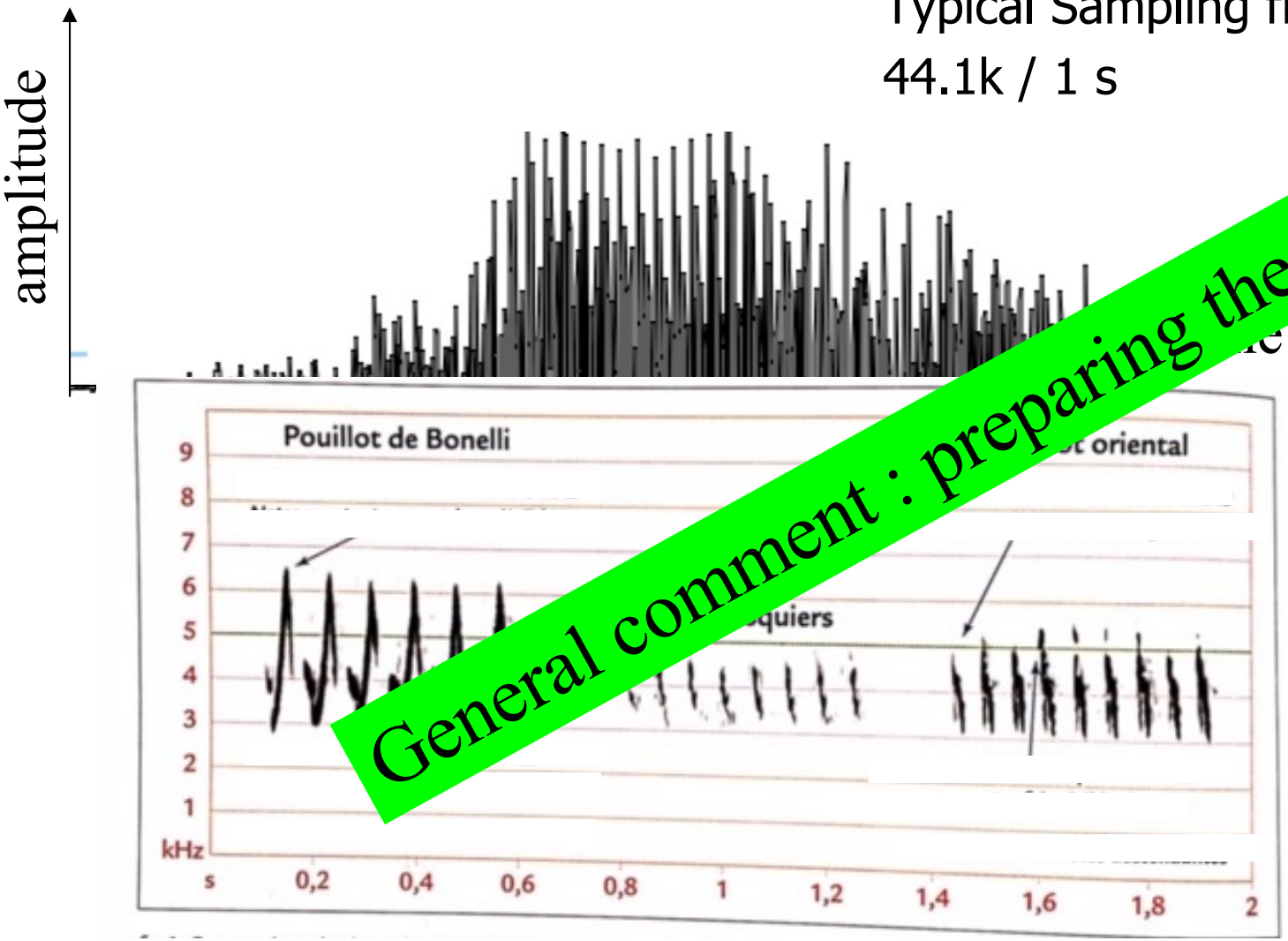
Inputs

Gokul Harikumar

Inputs

Typical Sampling frequency 44.1kHz

44.1k / 1 s



General comment : preparing the data

➔images

Time-frequency diagram

Stanislas Wroza

Tabular Datasets

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa

```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	[illegible]	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

	dijet_invmass	dijet_deltaeta	jet_pt_0	jet_pt_1	eta_zepp_ZZ	min_dR_jZ	pt4ljj_unconstrained	le
39	502.338403	3.694098	140.582153	43.095196	1.529987	1.799823	51.743145	
76	166.194427	0.426846	171.107452	81.588737	0.663560	1.020612	152.570358	
97	269.543396	2.568801	81.123795	64.938507	0.404464	0.050431	50.000000	
107	130.786301	0.119691	171.627014	31.095165	1.329497	0.539539	50.000000	
129	139.976868	2.145803	51.312862	37.323059	3.293238	0.423458	50.000000	

Tabular datasets (mostly) in this course

Not tabular datasets

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way - in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

Charles Dickens, *A tale of two cities*

- Natural Language is not tabular
- Also features like : « Age of the children »
  - [],[3], [3,7,18]
- Closer to physics : « Energy of the jets in this proton collision »:
  - [],[120.5],[509.2,439.1,123.6,13.3]
- Special techniques to deal with these, see ChatGPT (not in this course)

Output label

- Two classes, usually :
  - y=0 ➔ background
  - y=1 ➔ signal
- N classes :
  - Nearly never used: :
    - y=0➔cat
    - y=1➔dog
    - y=2➔rabbit
  - Rather use « one-hot vector »
    - y=[1,0,0]➔cat
    - y=[0,1,0]➔dog
    - y=[0,0,1]➔rabbit

Classification performance

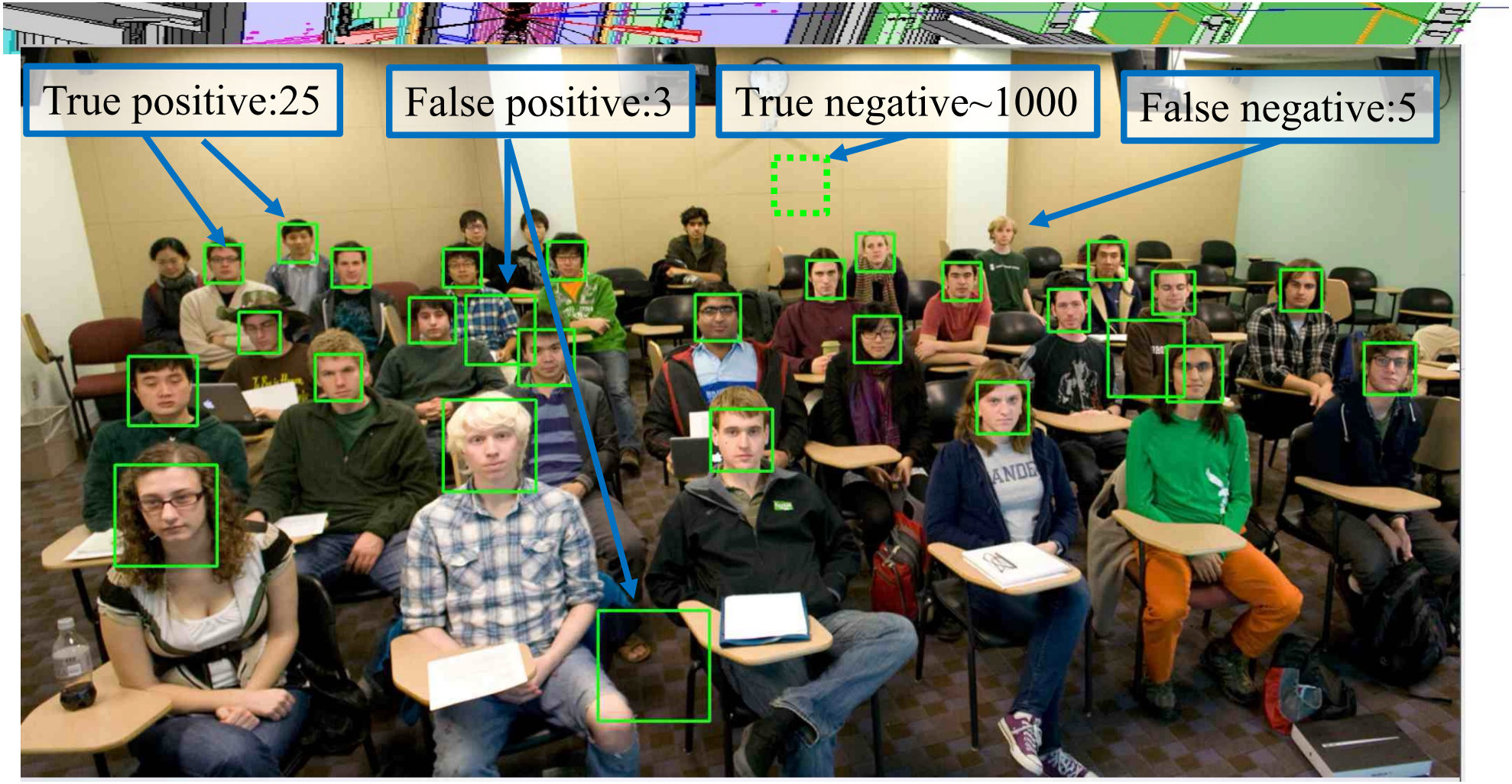
	Signal	Background
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

purity

efficiency

- Total Signal : TP+FN
- Total Background : FP+TN
- Performance numbers
  - (phys) Efficiency==(ML) Recall= TP / (TP+FN)
  - (phys) Purity==(ML) Precision = TP / (TP+FP)

Real-time face detection

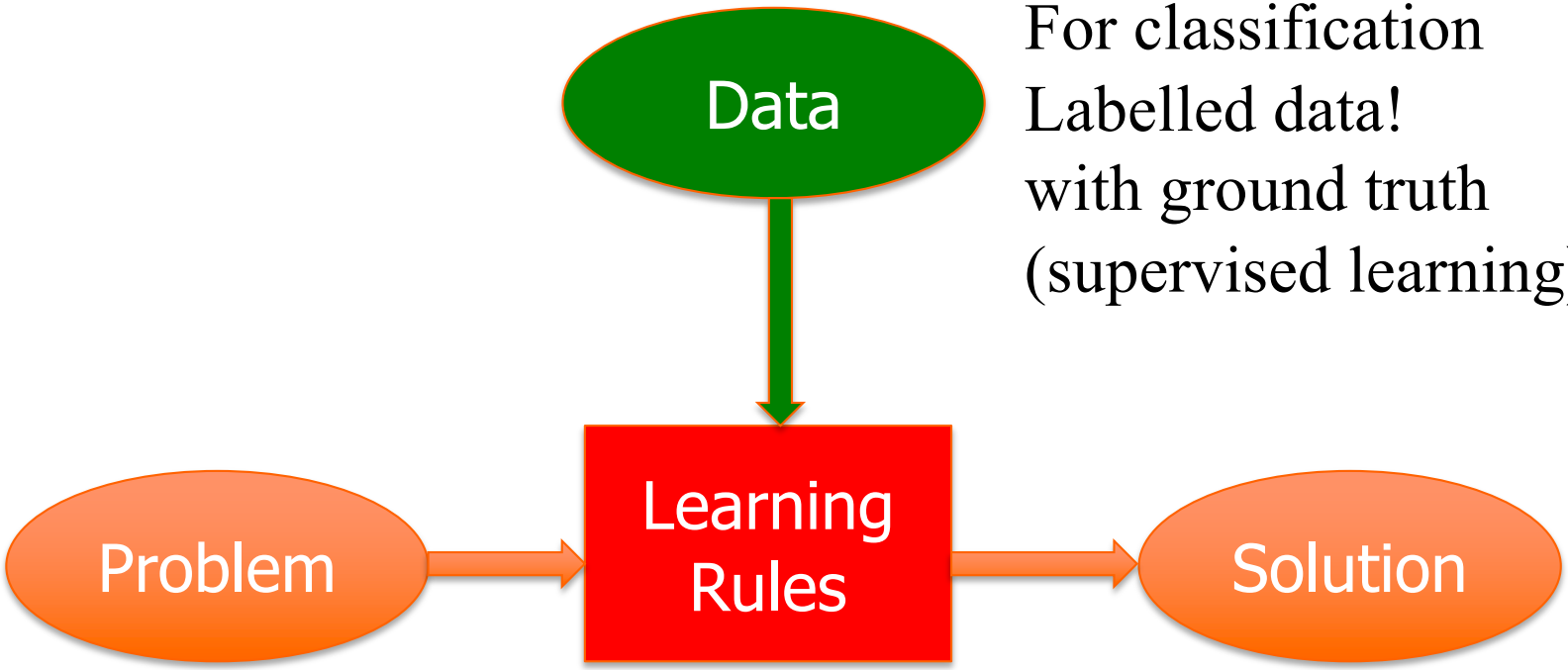


Efficiency==(ML) Recall=83%=25/(25+5)

Purity==(ML) Precision=89%=25/(25+3)

Training dataset

Machine Learning



For classification
Labelled data!
with ground truth
(supervised learning)

Data label example

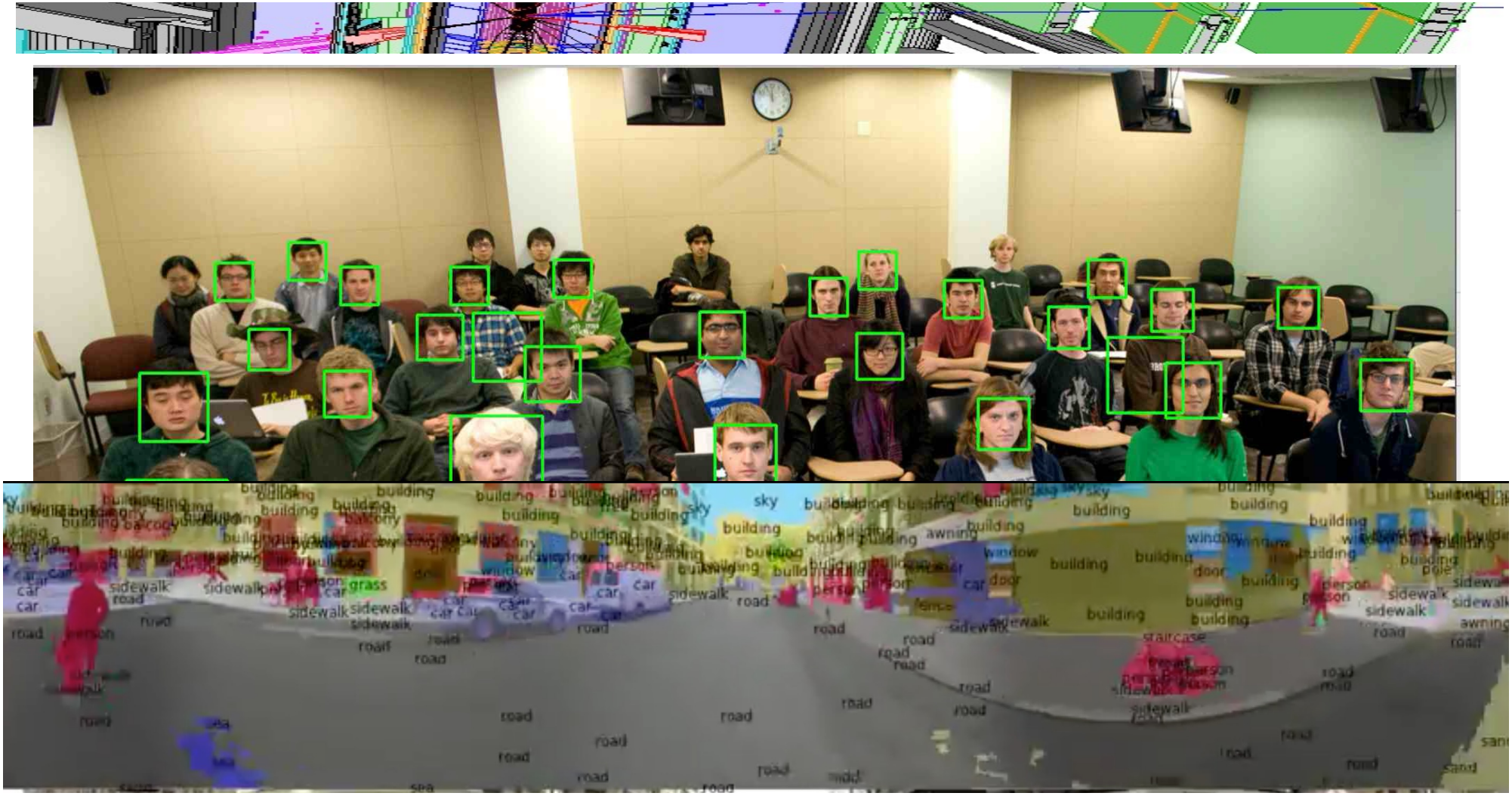
```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Data label example

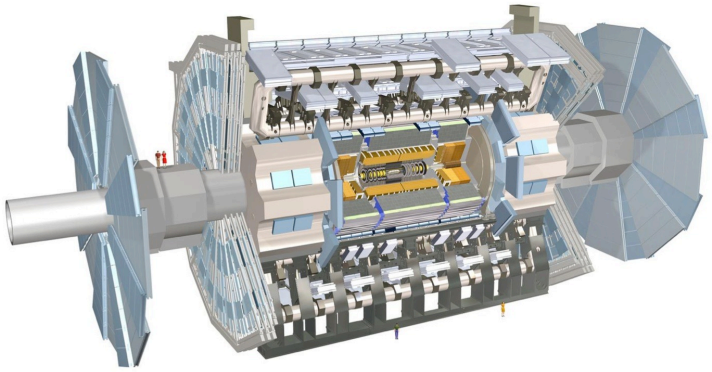
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Data label example

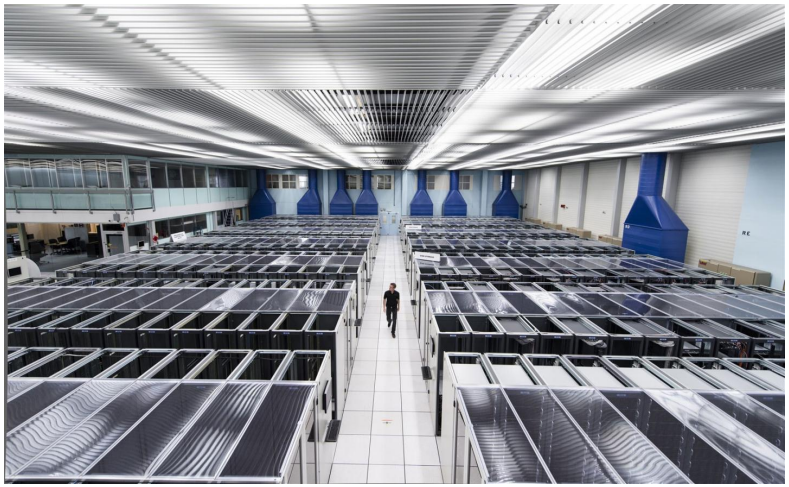


➔most Computer Vision tasks need human labelling!

Data label example



Data



Simulator

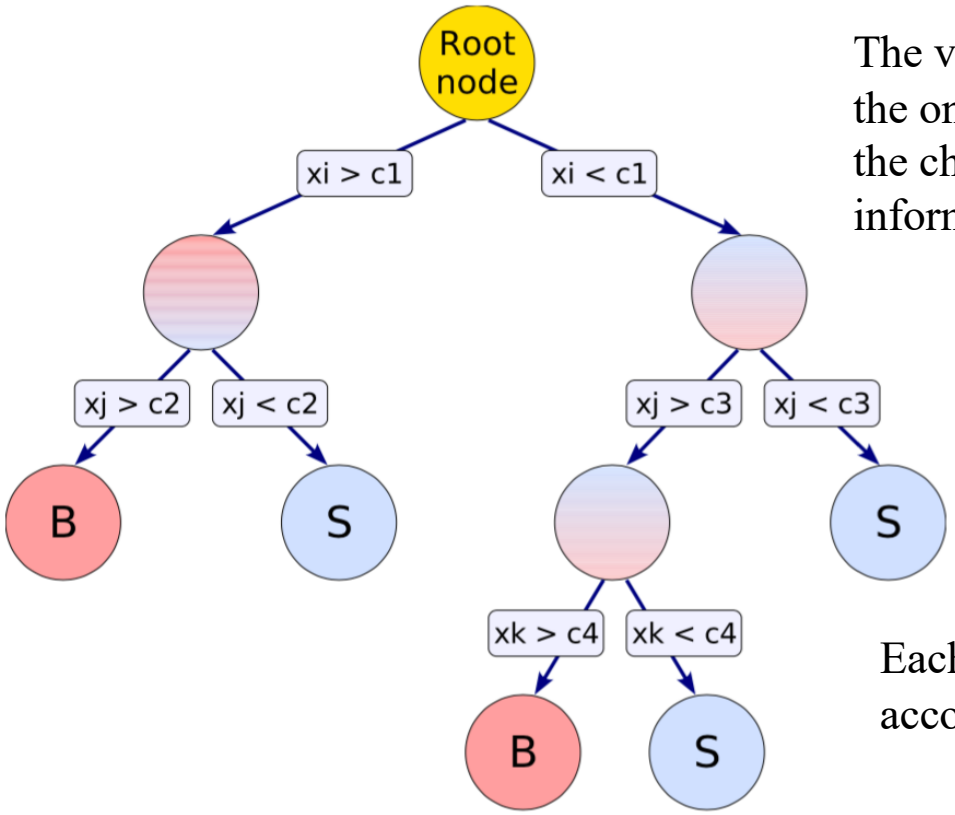
Provide pseudo-data with labels

In a nutshell

- "Classifier"
  - A model F(x)=y : with y=0 or 1 (e.g. 0=Background 1=Signal)
    - In practice : y ranking variable, the larger the more signal-like
  - If size (y)>1: (not in this course)
    - Multiclass : Σ y=1 : Cat or Dog or Elephant
- "Classes"==label : the different categories into which we want to classify. Two categories cases : (A,B), (Signal, Background), (sick, healthy),…
- "Features" == Variables (x)
    - Continuous
    - Discrete
- Classification performance, True/False Positive/Negative
  - Total Signal : TP+FN
  - Total Background : FP+TN
  - (phys) Efficiency==(ML) Recall= TP / (TP+FN)
  - (phys) Purity==(ML) Precision = TP / (TP+FP)
- Training dataset with ground truth : the « true » label==class

Decision Trees

Boosted Decision Tree

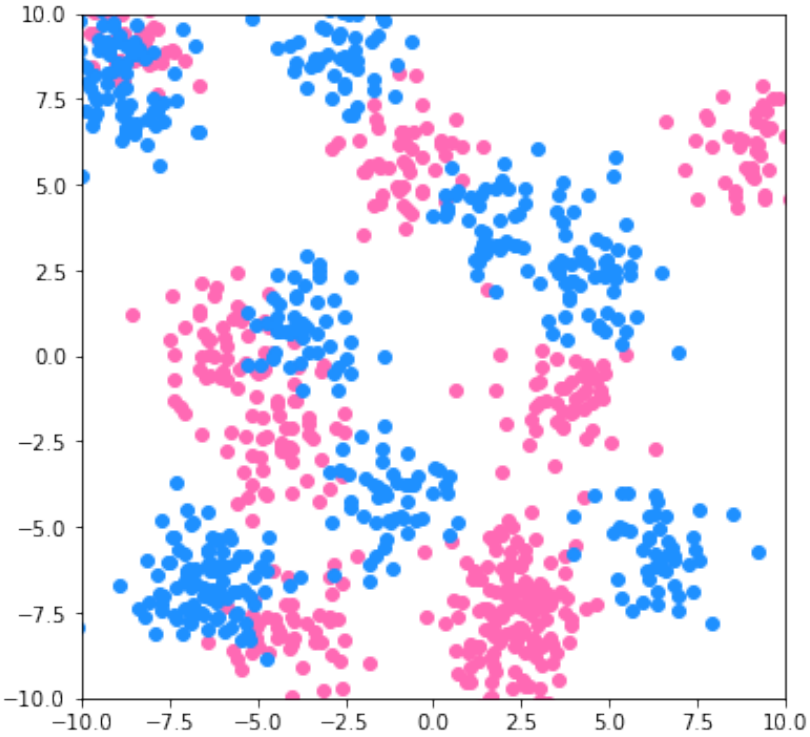


The variable used at each node is the one optimising a measure in the child node (min entropy, information gain…)

Each node given a score according to its purity

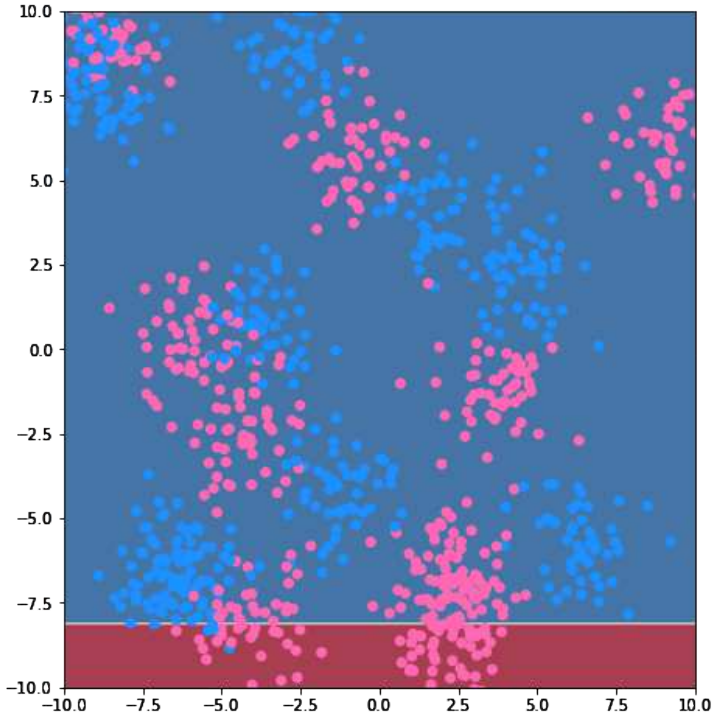
- Single tree (CART) <1980
- AdaBoost 1997 : rerun increasing the weight of misclassified entries ➔Boosted Decision Trees (Gradient BDT XGBoost, random forest…)

Trees at work



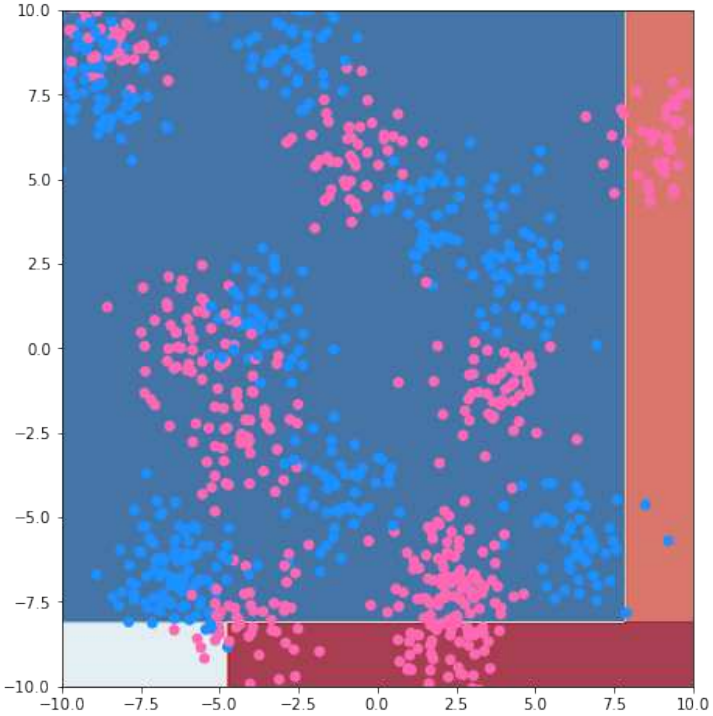
Trees at work

Decision tree, depth=1



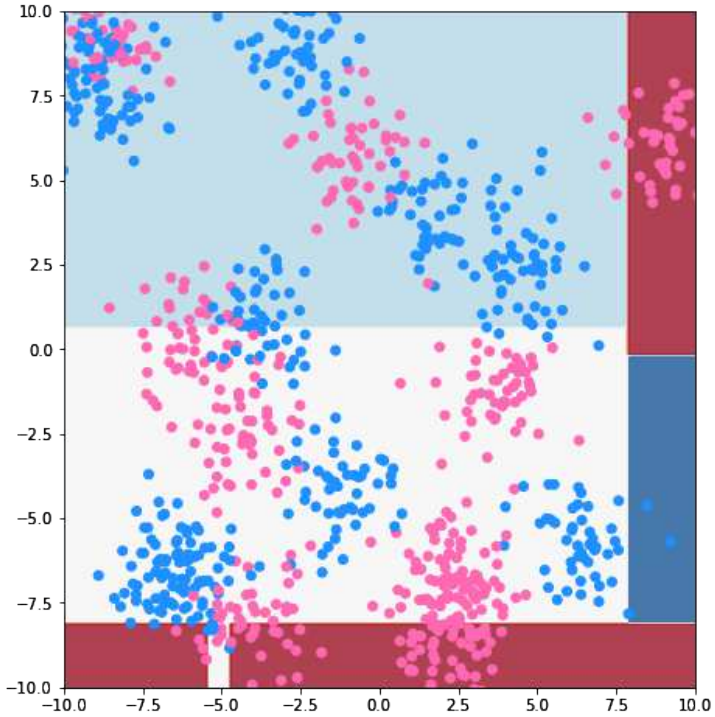
Trees at work

Decision tree, depth=2



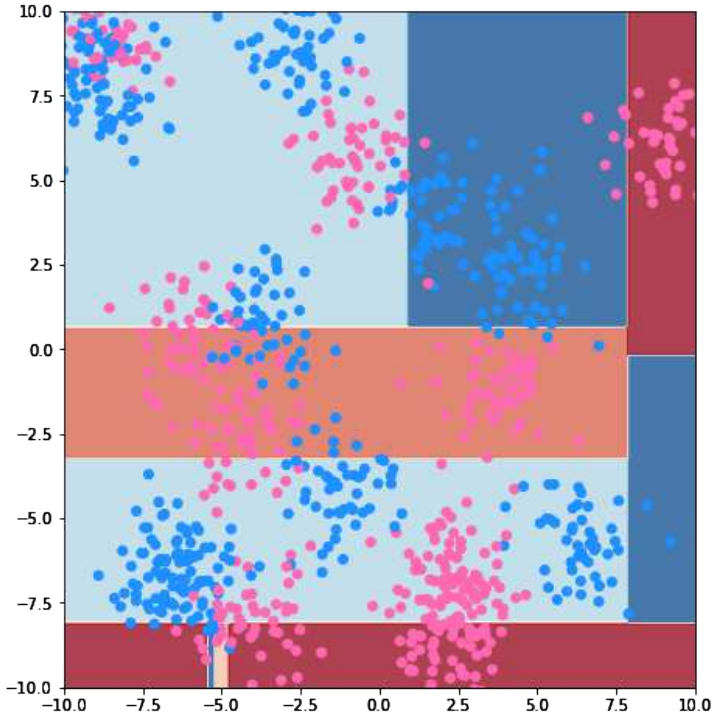
Trees at work

Decision tree, depth=3



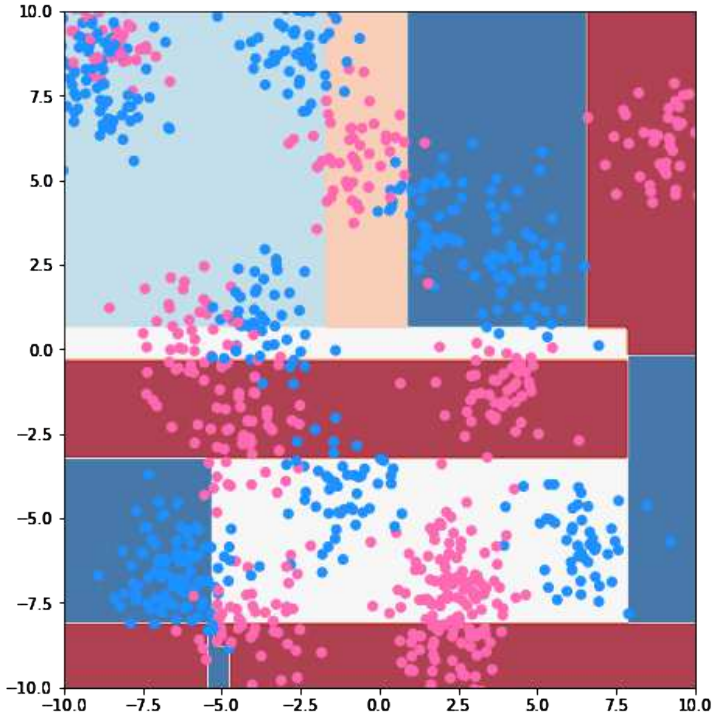
Trees at work

Decision tree, depth=4



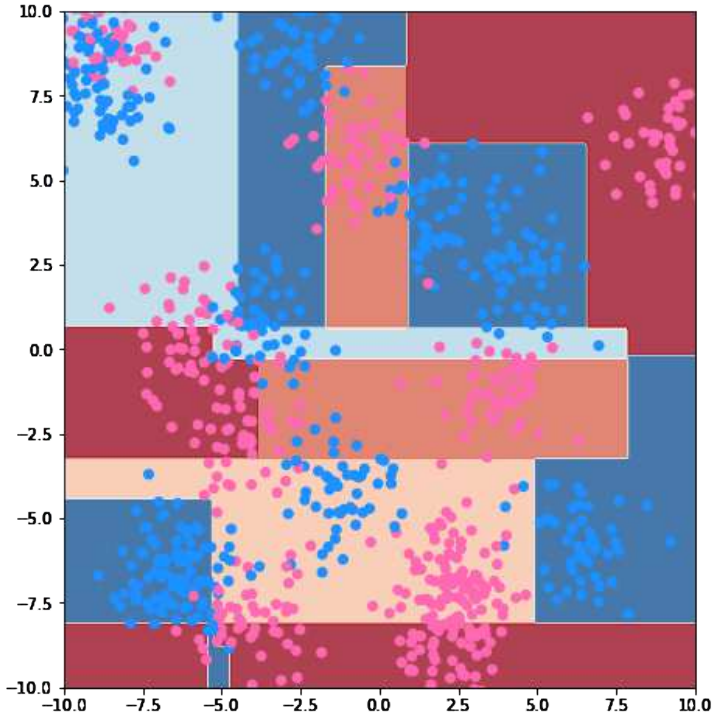
Trees at work

Decision tree, depth=5



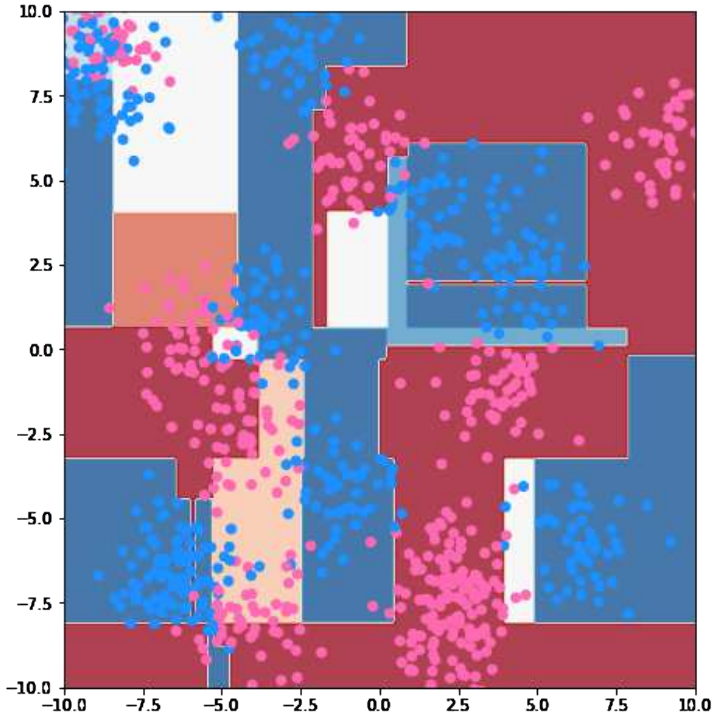
Trees at work

Decision tree, depth=6



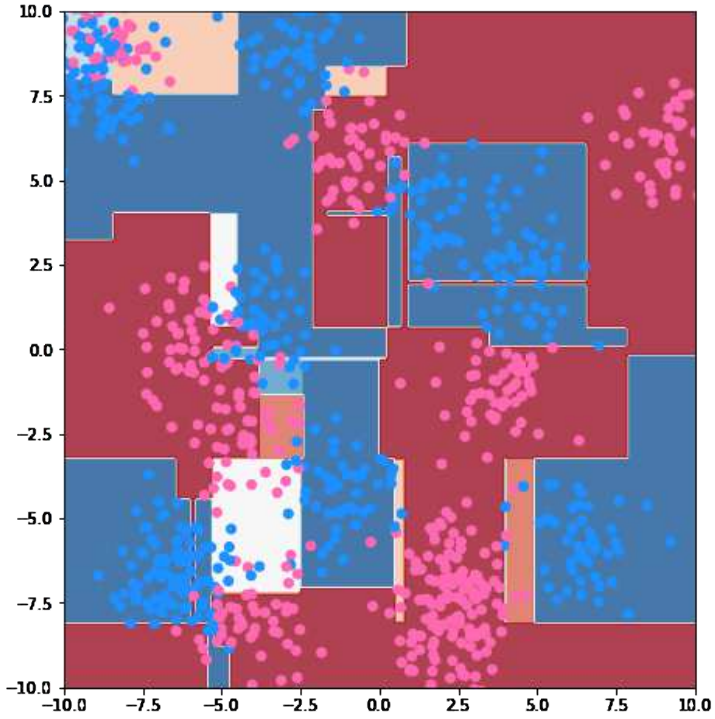
Trees at work

Decision tree, depth=8

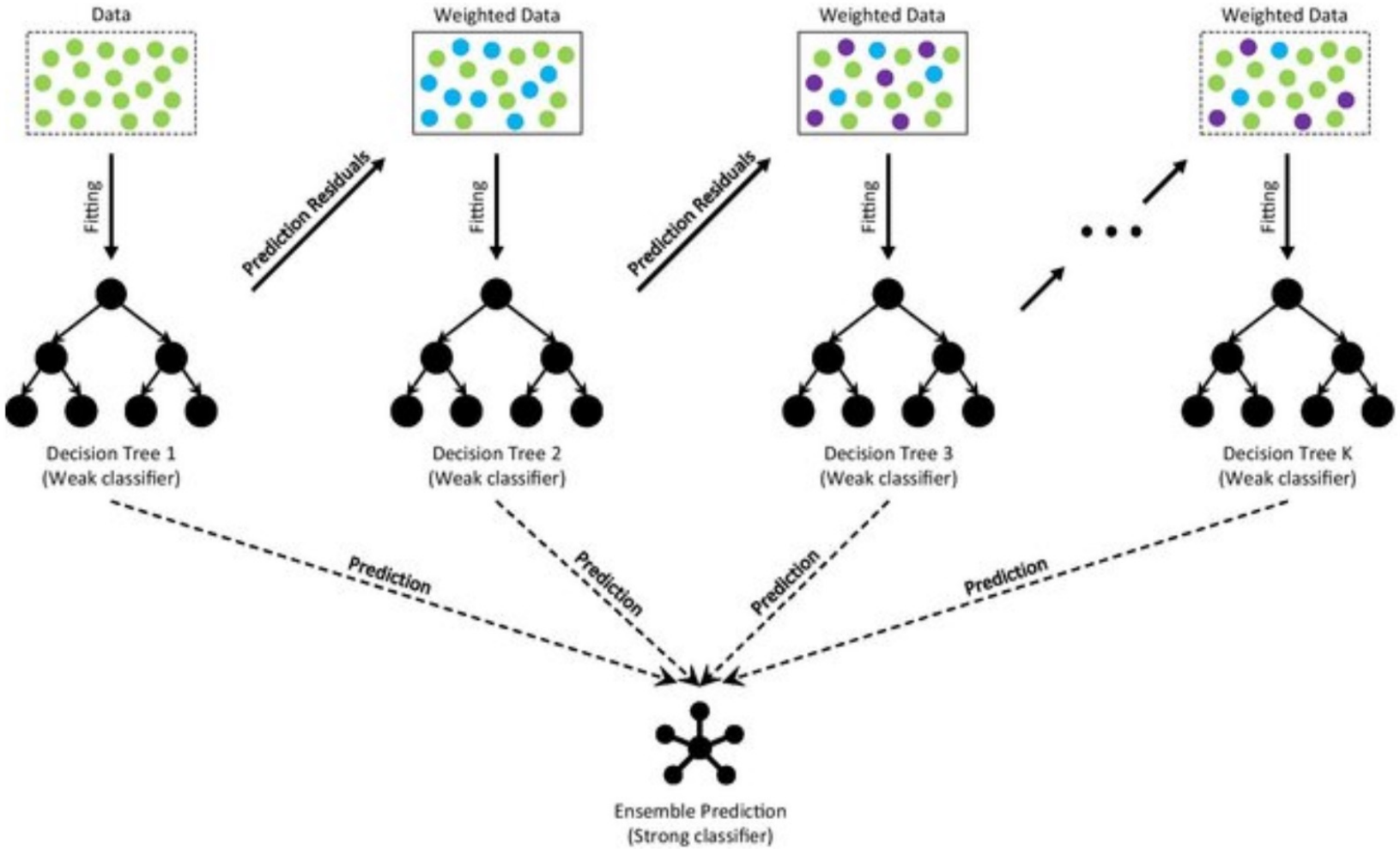


Trees at work

Decision tree, depth=9



Boosted Decision Tree



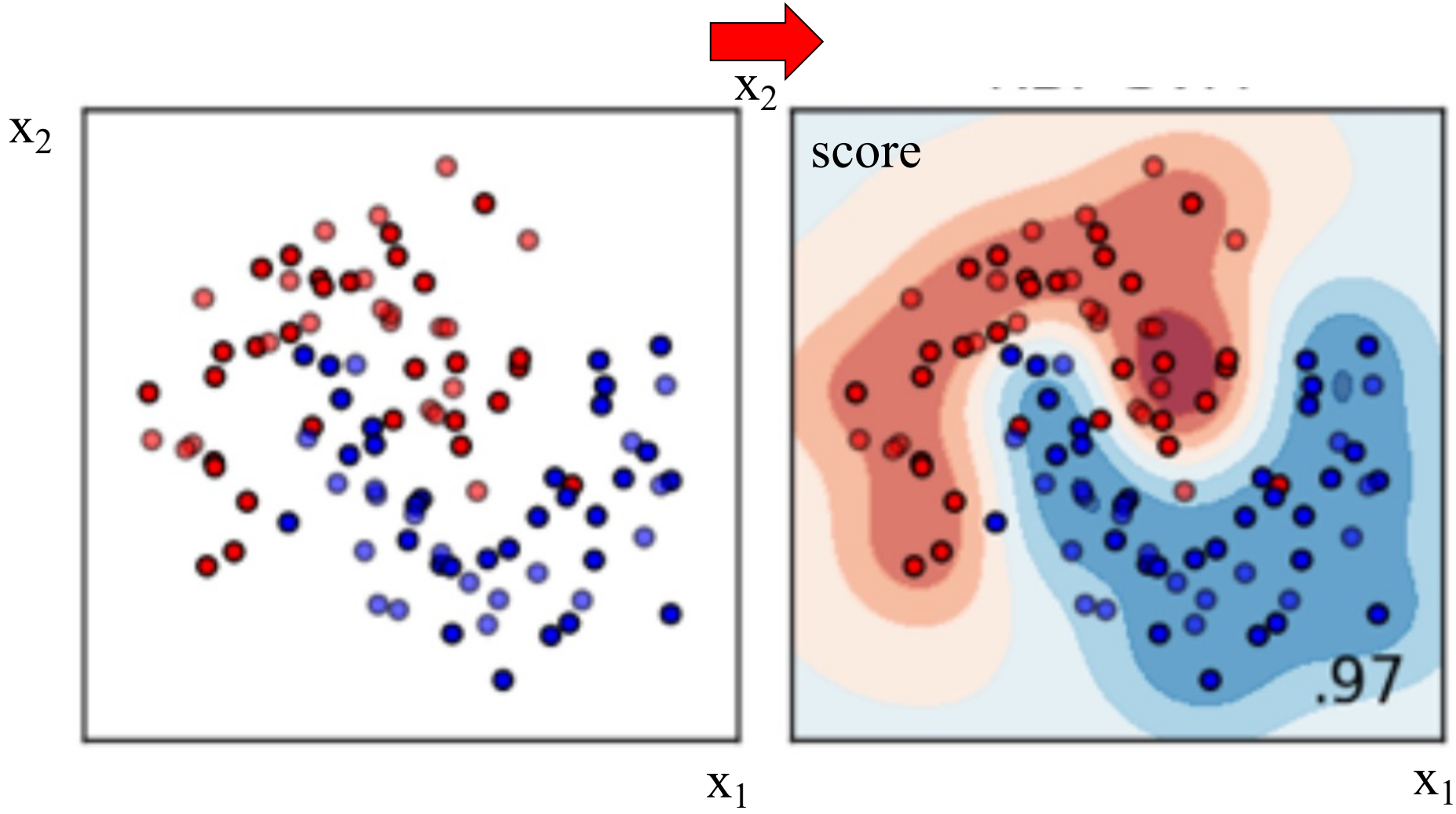
- Gradient Boosted Decision Tree chart

BDT software

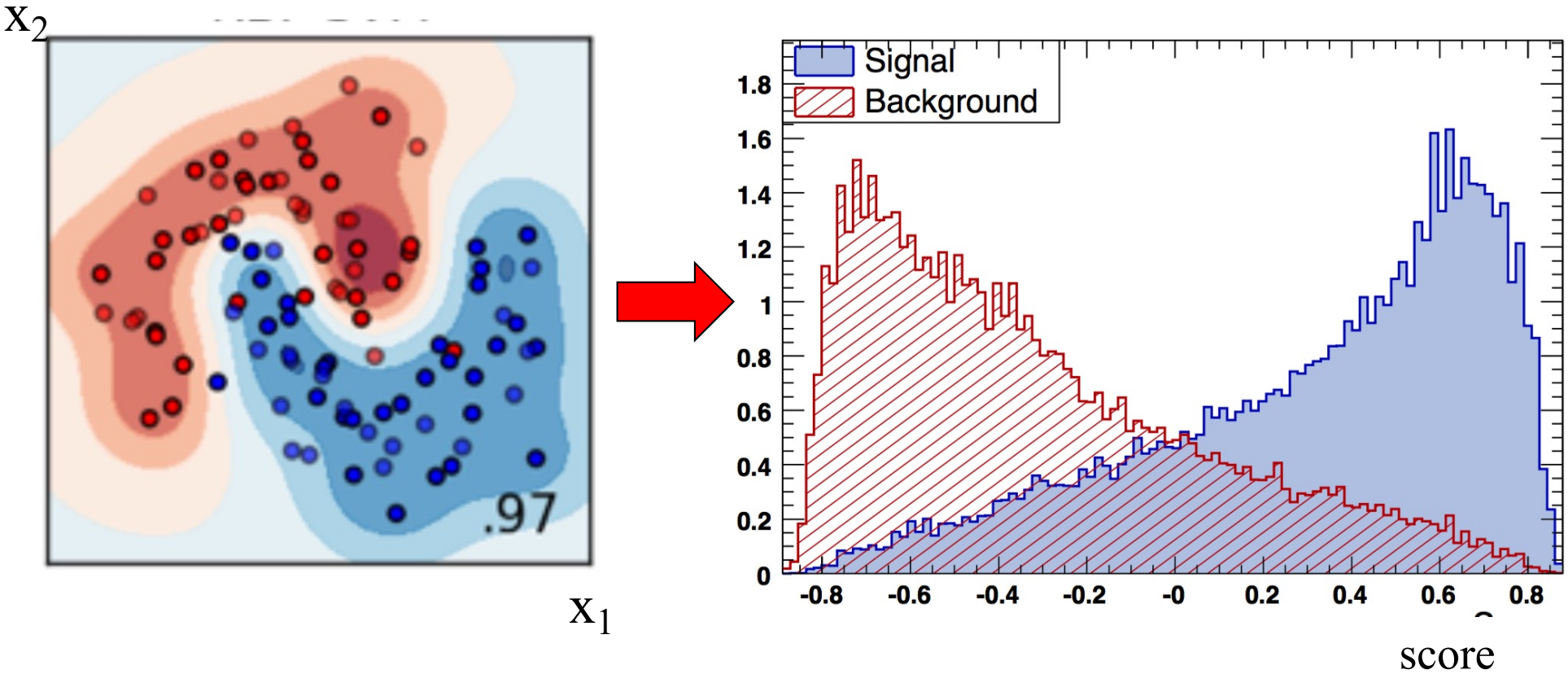
- BDT : first tool of choice for any supervised classification problem with <100 features
- XGBoost (with « hist » option)
- Lightgbm (Microsoft but free open source) (some issue with weighting)
- Sklearn DecisionTreeClassifier

ROC curve, AUC and more

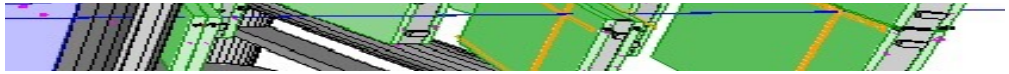
Score



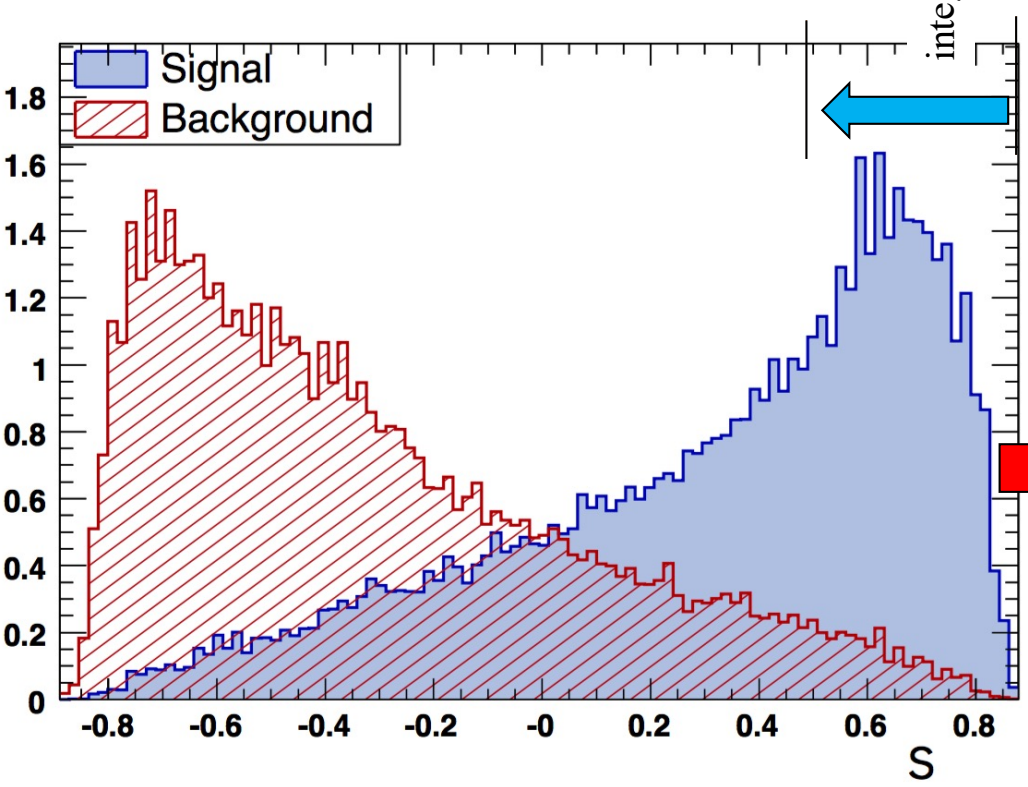
Score


x2
.97
x1
Signal
Background
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.8
-0.6
-0.4
-0.2
-0
0.2
0.4
0.6
0.8
score

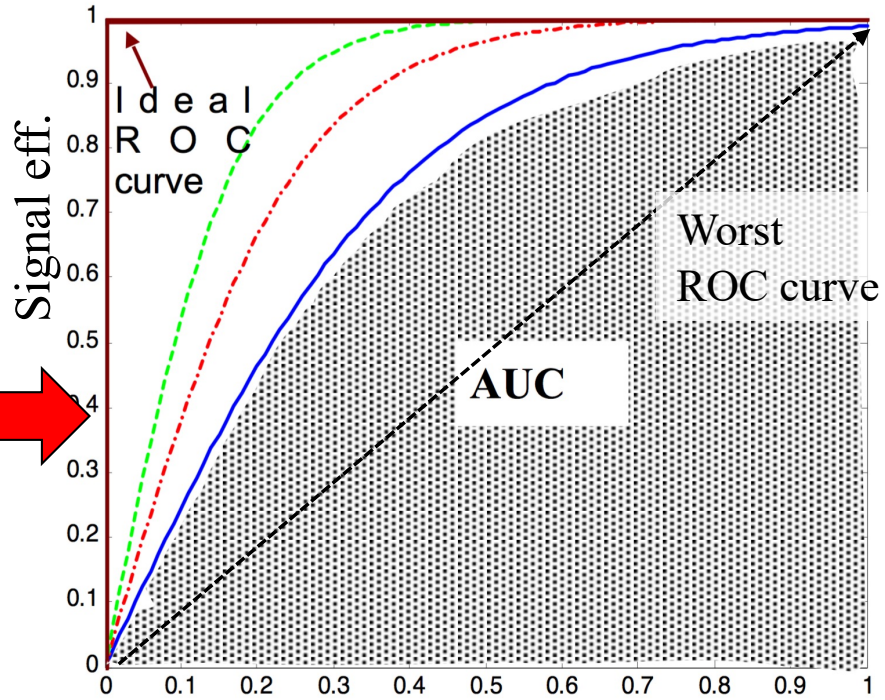
ROC Curve (BB)



integrate



AUC : Area Under the (ROC) Curve



Ideal ROC curve : AUC=1

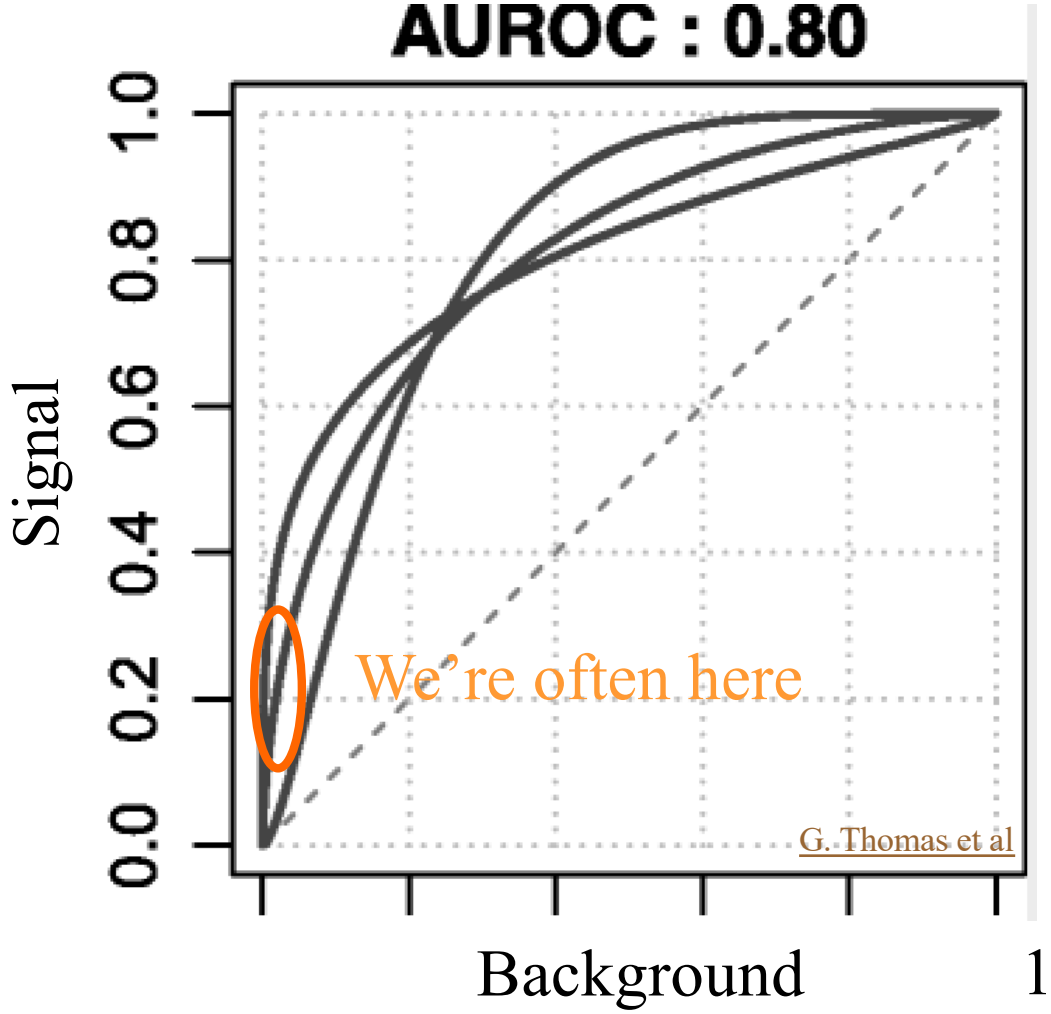
Worst ROC curve : AUC=0.5

AUC<0.5➔bug!

The higher the AUC the better

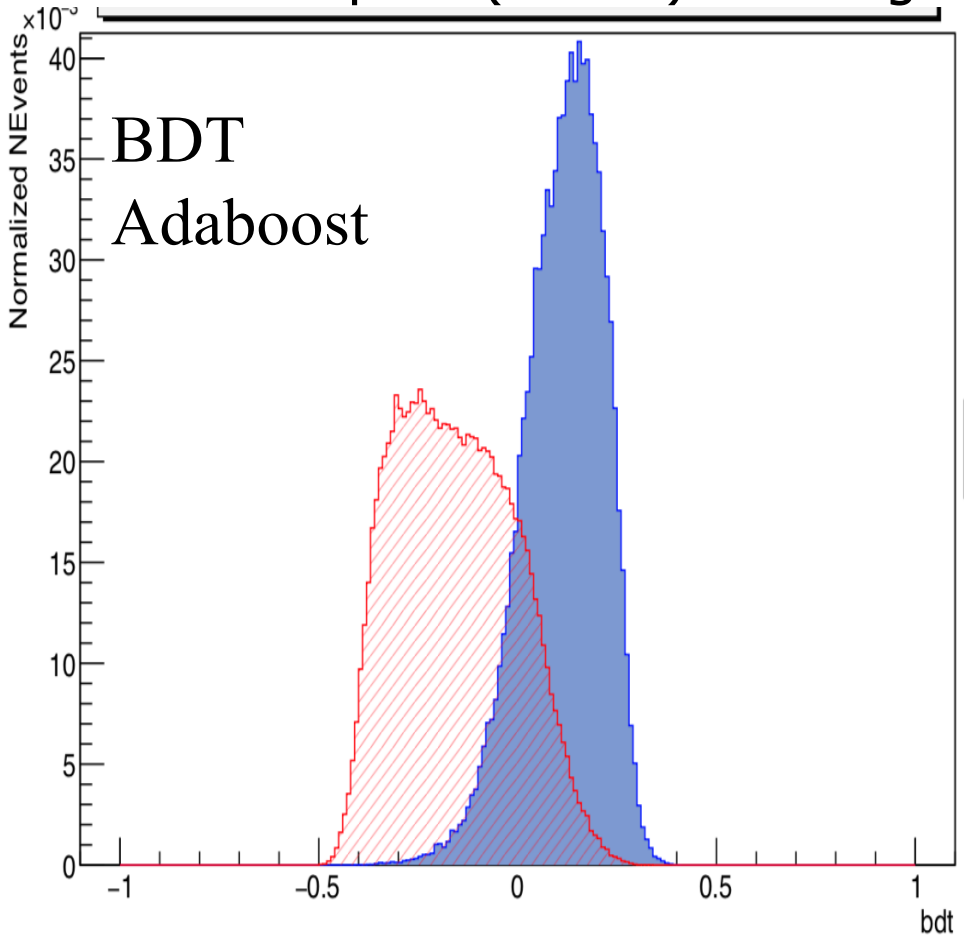
However AUC not the full story

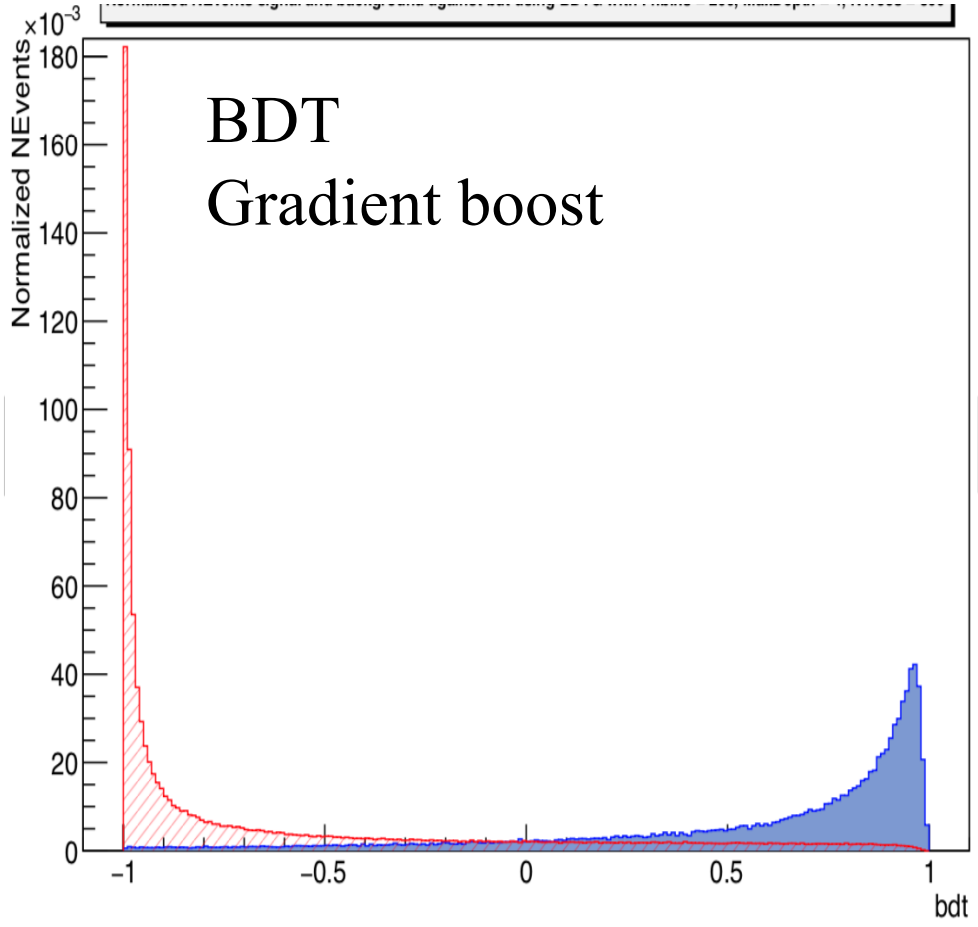
ROC curve pitfall BB



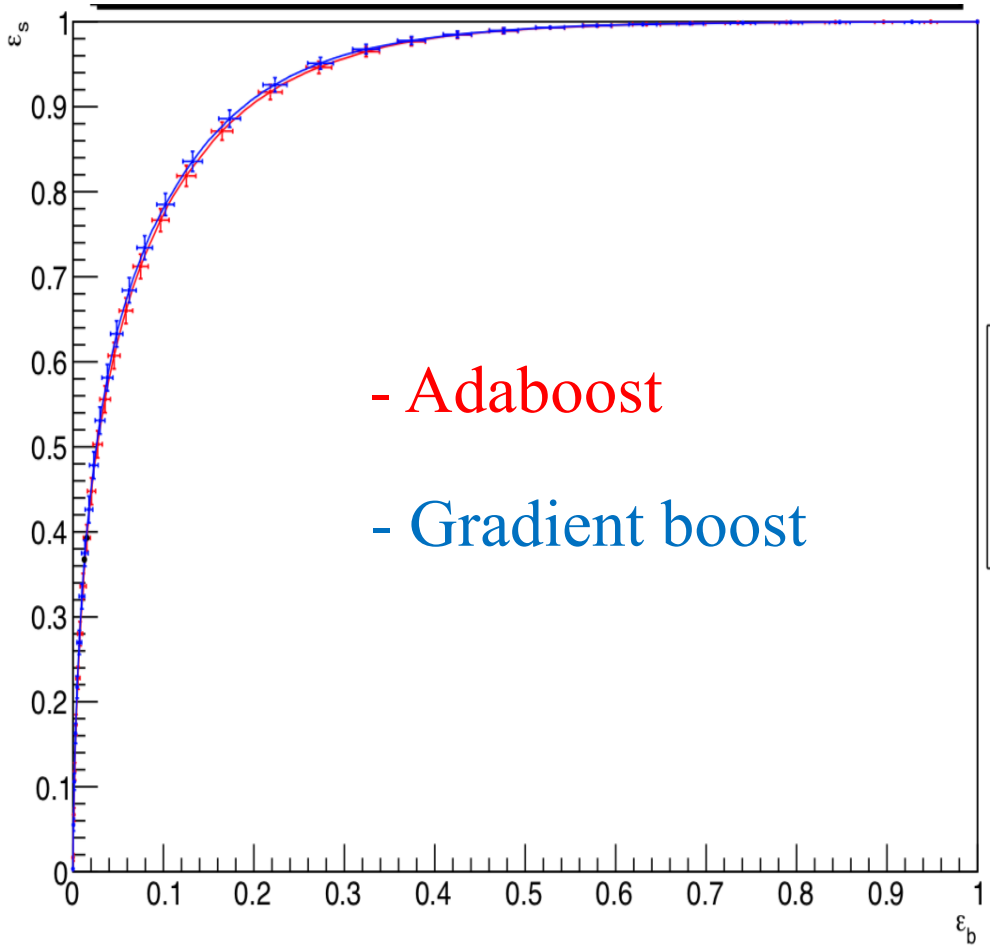
Score

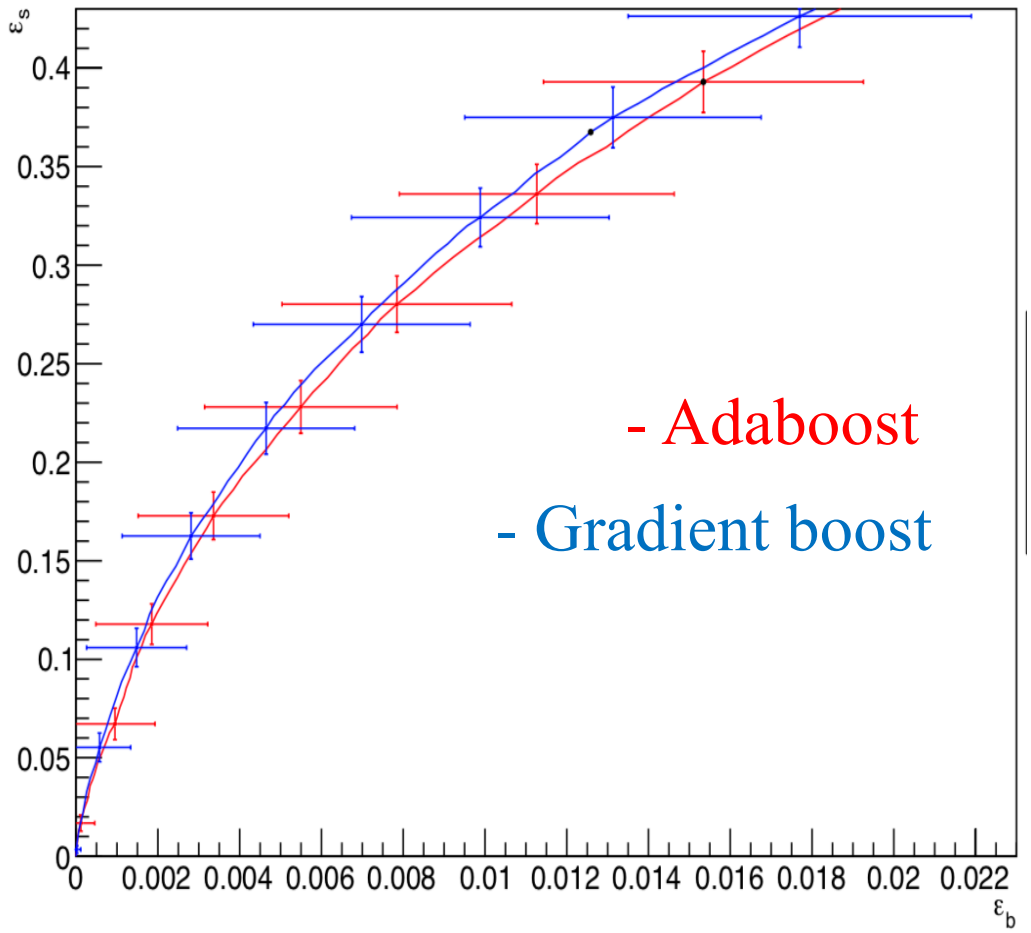
- Score output by algorithm is actually a ranking-variable
  - From most background like to most signal-like
  - Shape is (almost) meaningless



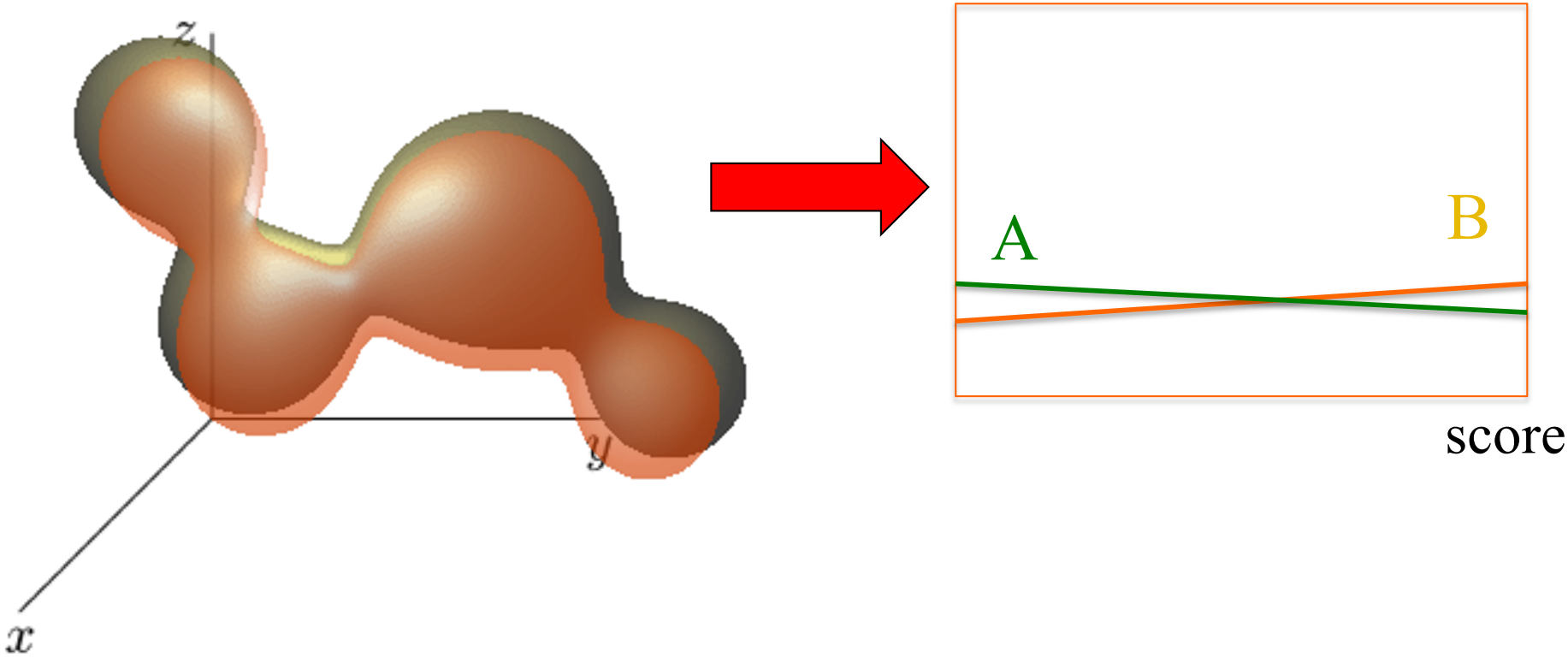


Score (2)


ε_s
ε_b
- Adaboost
- Gradient boost


ε_s
ε_b
- Adaboost
- Gradient boost

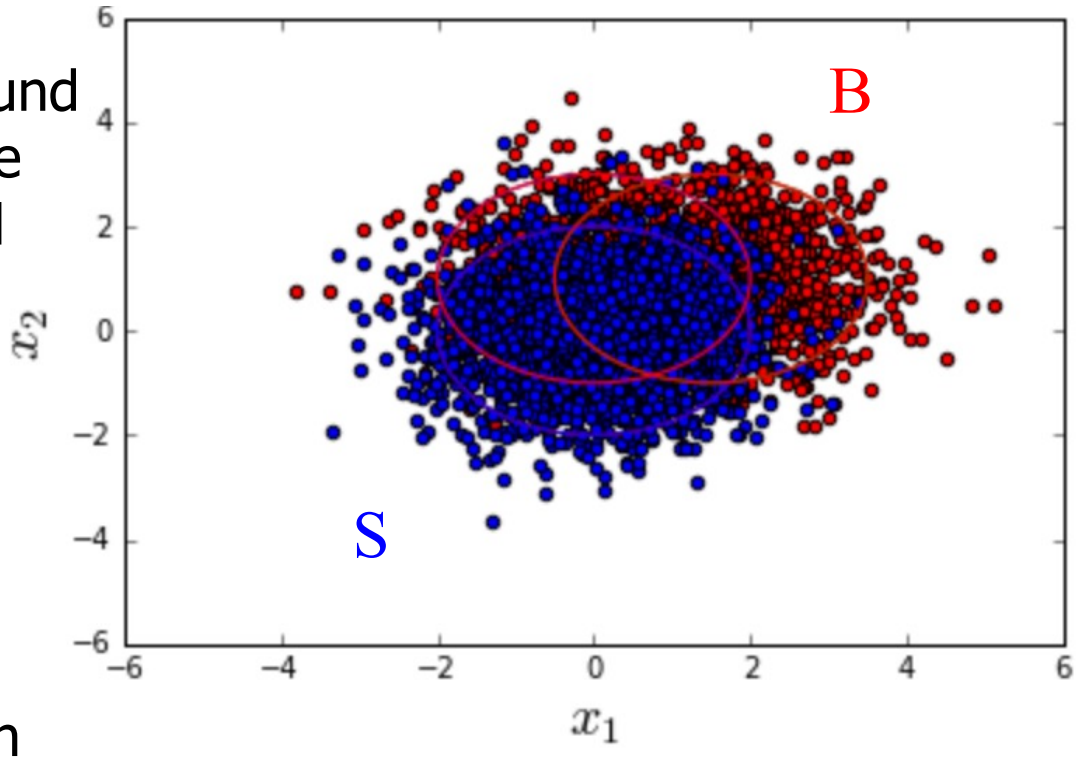
What does a classifier do?



- The classifier "compresses" the two multidimensional "blobs" maximising the difference, without (ideally) any loss of information

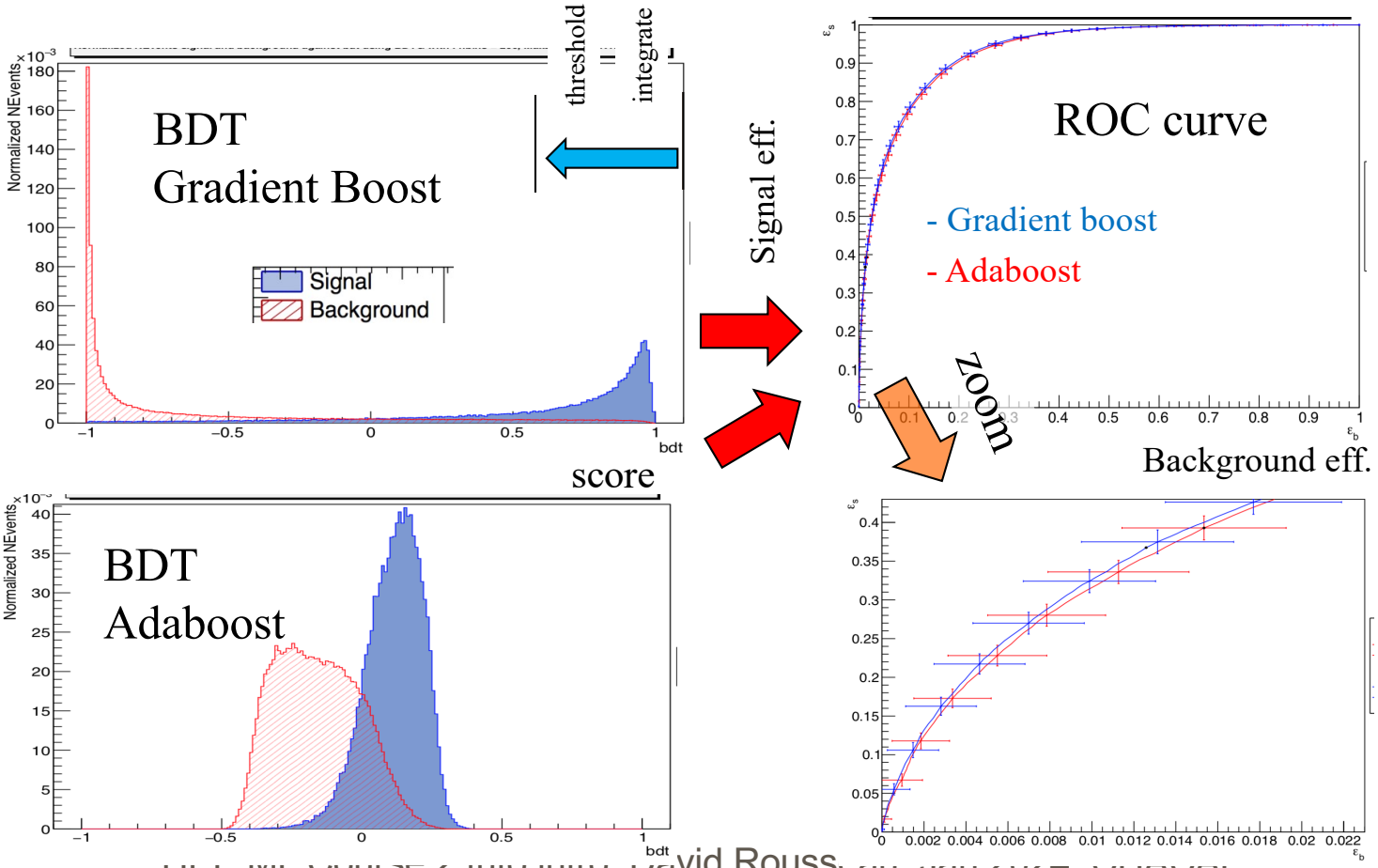
No miracle

- ML (nor Artificial Intelligence) does not do any miracles
- For selecting Signal vs Background and underlying distributions are known, nothing beats likelihood ratio! (often called "Bayesian limit"):
  - $L_S(x)/L_B(x)$
- OK but quite often L_S L_B are unknown
  - + x is n-dimensional
- ML starts to be interesting when there is no proper formalism of the pdf
- ➔mixed approach, if you know something, tell your classifier instead of letting it guess



Significance optimisation

Recall ROC Curve



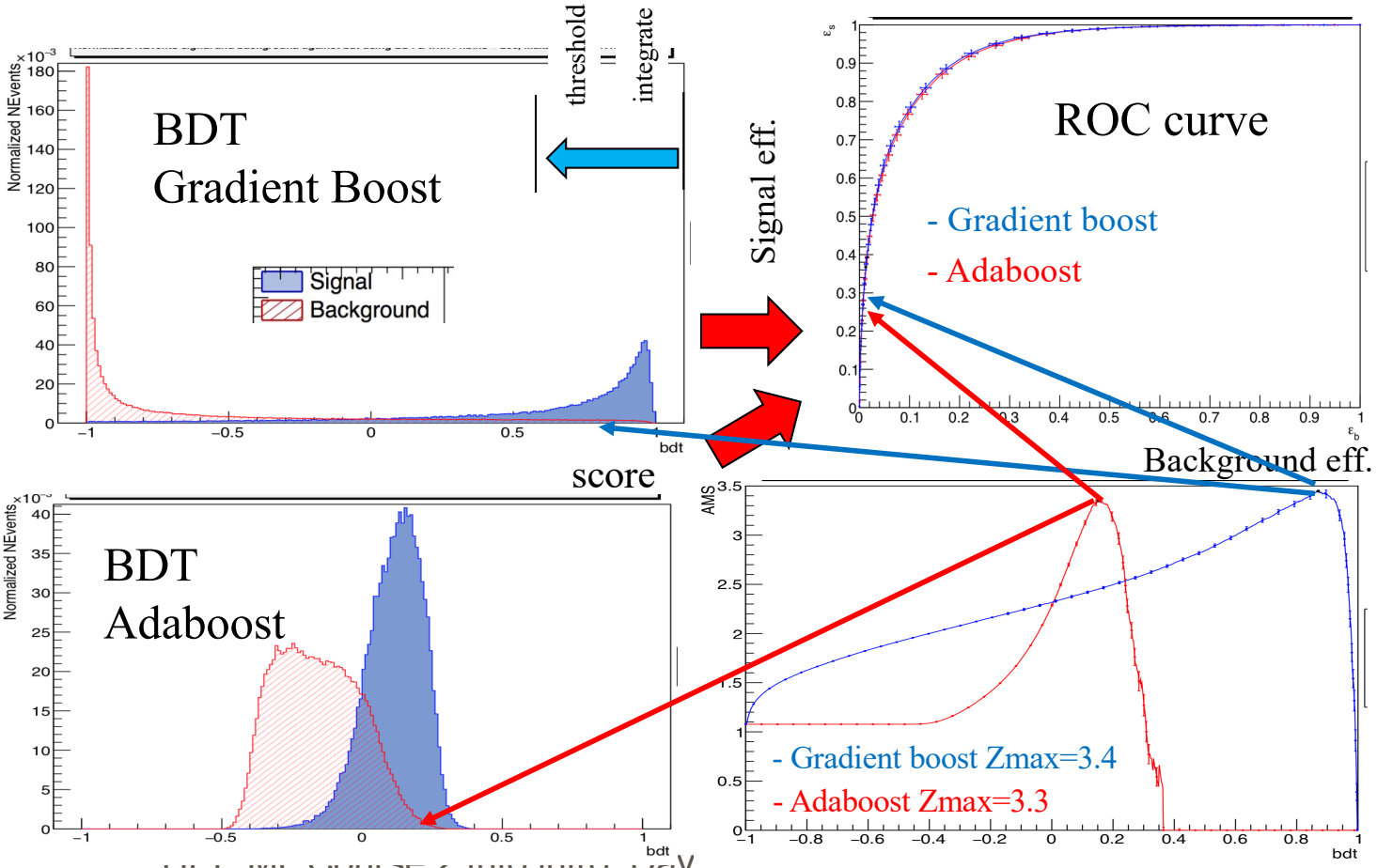
Significance

- Imagine counting experiment above threshold (see last slides of PP Course 4)

$$s = N_{sig}^{exp} = N_{sig}\epsilon_{sig}$$
$$b = N_{bkg}^{exp} = N_{bkg}\epsilon_{bkg}$$

$$Z = \sqrt{2\big((s+b) * \log(1+\frac{s}{b}) - s\big)} \sim \frac{\sqrt{s}}{b}$$

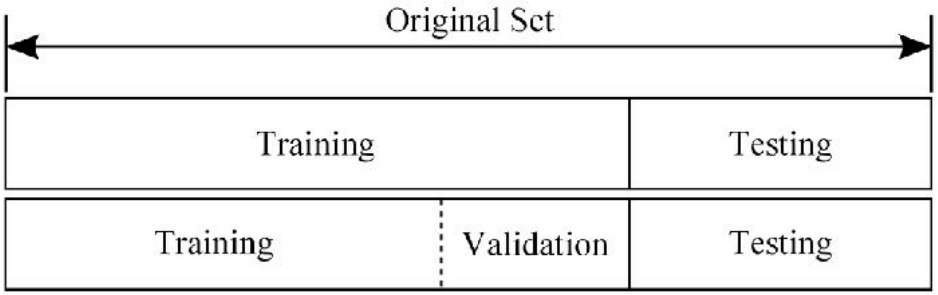
Significance curve



Training, Validation, Test

-

Divide the labelled set into training, validation and testing sets



* **Training set: used to train the classifier**
* **Validation set (optional): choose between different methods, finite-tune parameters,**
* **Testing set: predict the generalization error**

Ideally, look at it only once at the end

No cheat: do not use the test set to train your algorithm!

In practice, we'll only do Train/Test split in the TD

ML highest crime

Horror stories

See page

MIT Technology Review

Also some ChatGPT success stories

Intelligent Machines

Why and How Baidu Cheated an Artificial Intelligence Test

Machine learning gets its first cheating scandal.

by **Tom Simonite** Jun 4, 2015

The sport of training software to act intelligently just got its first cheating scandal. Last month Chinese search company Baidu announced that its

Also in physics…

Training on test set particularly bad because undetectable unless:

- training reproducible
- new i.i.d data

ML for Higgs physics

Using ML to see the Higgs Boson
Using Boosted Decision Tree first
Intro to BDT tutorial

Seeing the Higgs

Proton collisions

Young Einstein: $E=mc^2$

Two fundamental entities

- « Events » :
  - All measurements from one proton collision
  - List of particles with their properties
  - Derived quantities
  - =>ML to help select interesting events « Signal » with respect to « Background »
- « Particles »:
  - Extracted from an event
  - Jet, lepton, photon Missing ET

Before observation, all was known about the Higgs boson, except its mass

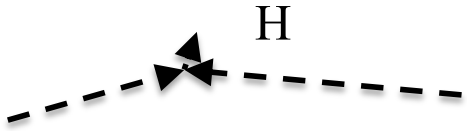
Probabilités de désintégration prédites pour une masse de 125 GeV

H → bb	58%
H → WW*	21%
H → τ+τ-	6.4%
H → ZZ*	2.7%
H → γγ	0.2%

$E=mc^2$

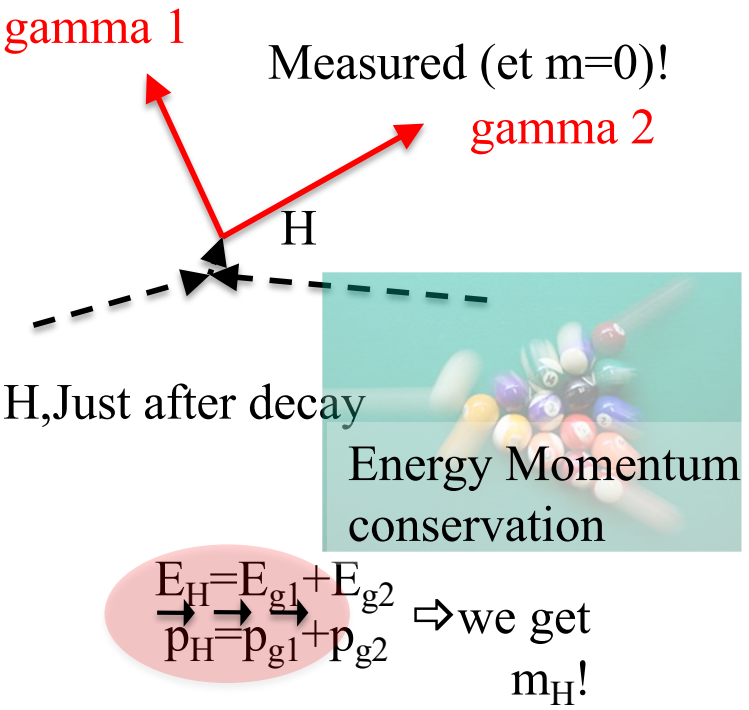
Einstein en 1905

$E^2=p^2+m^2$



H, before decay

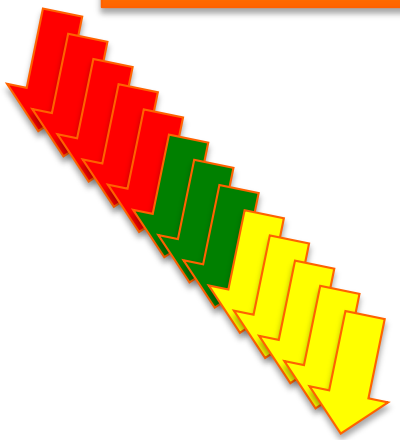
$m_H^2=E_H^2-p_H^2$



10^{14} collisions / year

Trigger: fast rough selection

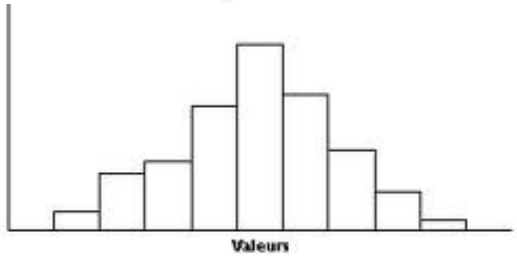
10^{9} events on disk

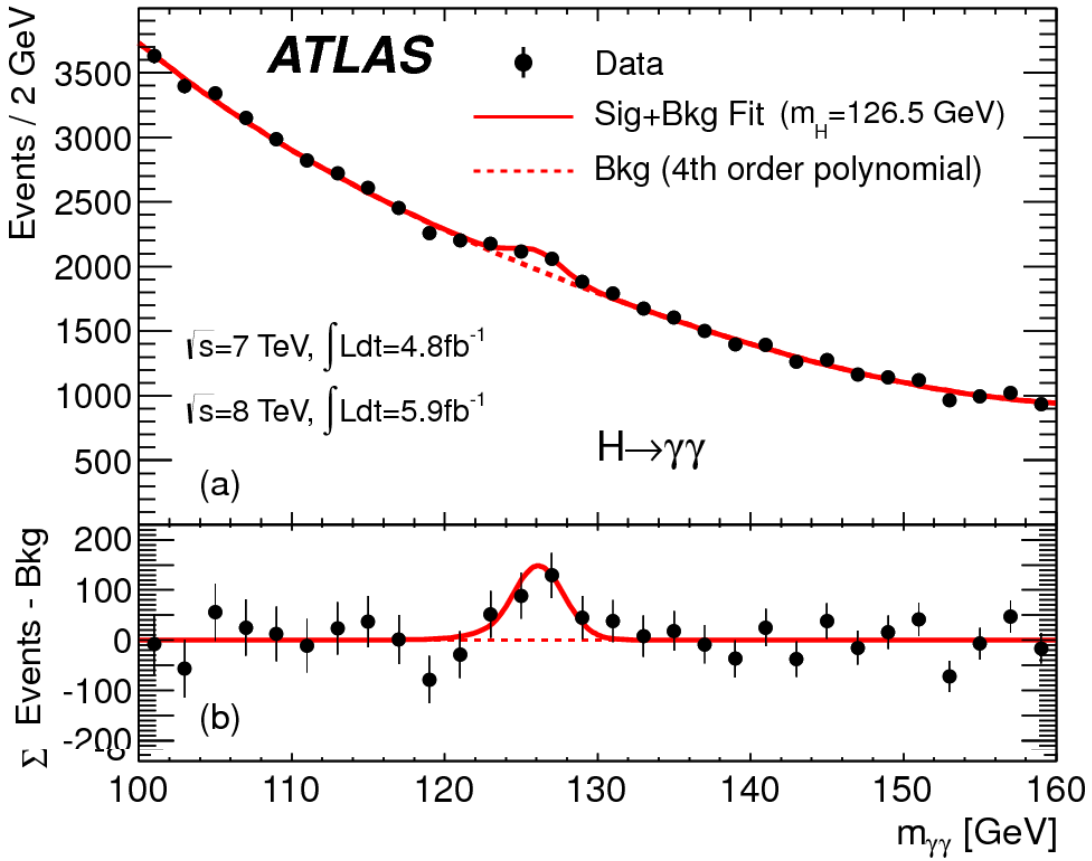


Tri précis

10^{5} events with 2 photons

Mass calculation
→histogramme

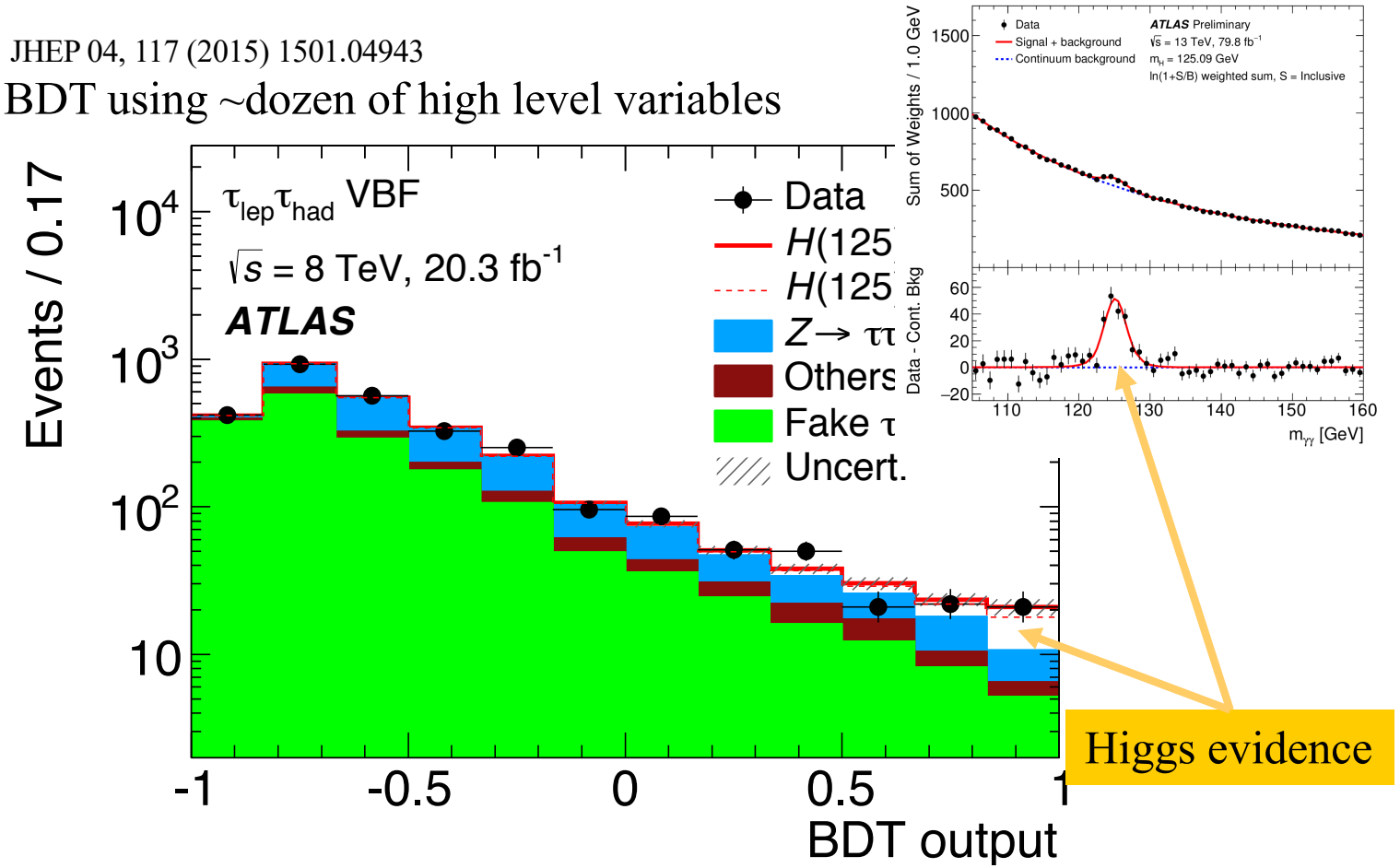



ATLAS
Data
Sig+Bkg Fit (m_H=126.5 GeV)
Bkg (4th order polynomial)
√s=7 TeV, ∫Ldt=4.8fb^-1
√s=8 TeV, ∫Ldt=5.9fb^-1
H→γγ
(a)
(b)
Events / 2 GeV
Σ Events - Bkg
m_γγ [GeV]

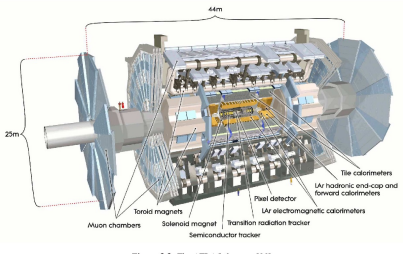
Classifier in Higgs Physics

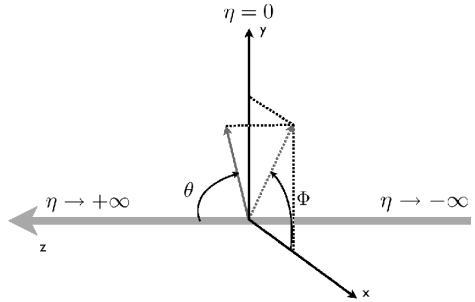
JHEP 04, 117 (2015) 1501.04943

BDT using ~dozen of high level variables



Coordinates





- P : momentum
- E : energy =sqrt(P^2+M^2)~P because P>>M
- Angles (cylindrical)
  - ϕ : azimuth angle]-π, +π]
  - θ : dip angle [0, +π]
  - η : eta, pseudo-rapidity = -log(tan(θ/2)), ~[-5,5]
- P_T : =P sin(θ) : transverse momentum
- ME_T : Missing Transverse Energy = $-\Sigma_{\text{all particles}} \vec{P_T}$: estimator of transverse momentum of neutrinos

H→WW

- One of the Higgs Discovery channel
- $H \rightarrow W^+(\rightarrow l^+ \nu)\ W^+(\rightarrow l^- \nu)$
  - ➔2 leptons of opposite charge
  - Neutrinos undetected ! => Missing Transverse Energy
  - No invariant mass peak!
- Background :
  - Other processes leading to $W^+(\rightarrow l^+ \nu)\ W^+(\rightarrow l^- \nu)$

tutorial

- Open Google Colab
- Search for my github repository: https://github.com/dhrou/HEPMLtutorials
- Open HEPML_HandsOn_BDT.ipynb
- Remember to save a copy first
- You can also run locally by switching COLAB=False, and downloading the dataset and fixing the path (under Load Events)