Machine Learning and Statistics in HEP

part 1

David Rousseau, IJCLab-Orsay

rousseau@ijclab.in2p3.fr @dhpmrou

ICFA 2023 Instrumentation School

Mumbai, Feb 2023

IJCLab Irène Joliot-Curie Laboratoire de Physique des 2 Infinis

cnrs IN2P3 Les deux infinis

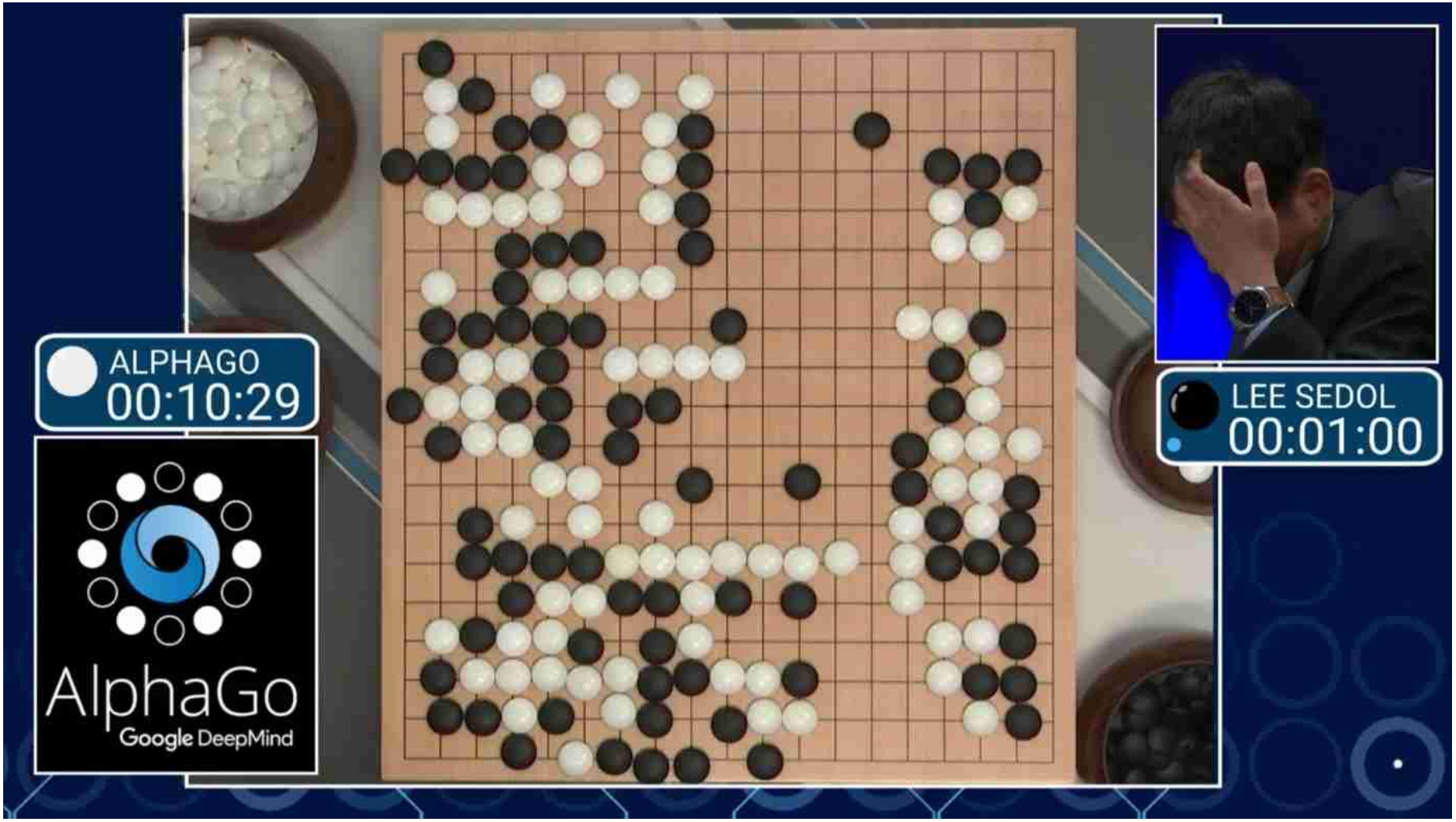
université PARIS-SACLAY

INSTITUT DATAIA Science des données, Intelligence & Société

Outline

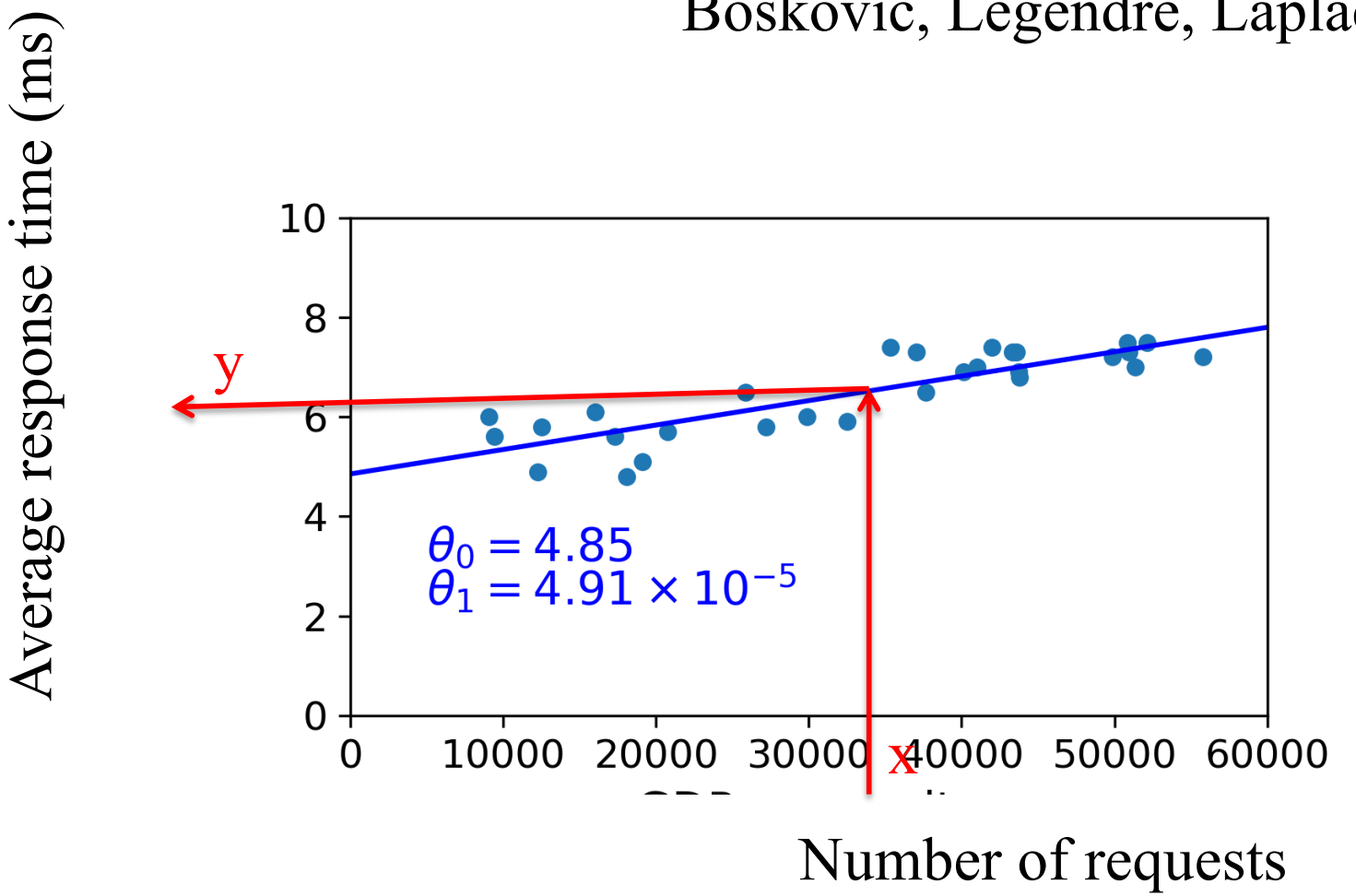
- Mostly Machine Learning with Statistics interludes
- Part 1 : Overview
  - What is Machine Learning?
  - Specificities of ML in physics
  - Useful concepts
- Part 2 : wider and deeper
  - NN on HEP data
    - various hammers and nails (including wrong ones)
  - Graph NN
  - Anomaly detection
- Part 3 : even wider and deeper
  - ML training tricks
  - Surrogate models
  - Recommendations for ML software and tools

See CERN Inter-Experiment Machine Learning workshop May 2022


ALPHAGO
00:10:29
AlphaGo
Google DeepMind
LEE SEDOL
00:01:00

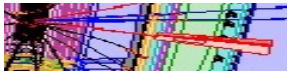
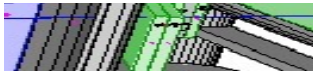
Linear Regression

Boskovic, Legendre, Laplace, Gauss



y=f(x)

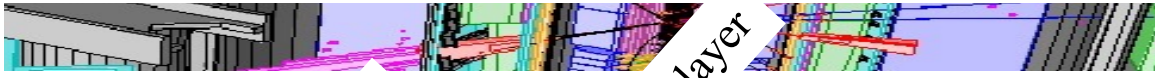
Given x_n, we want y_p

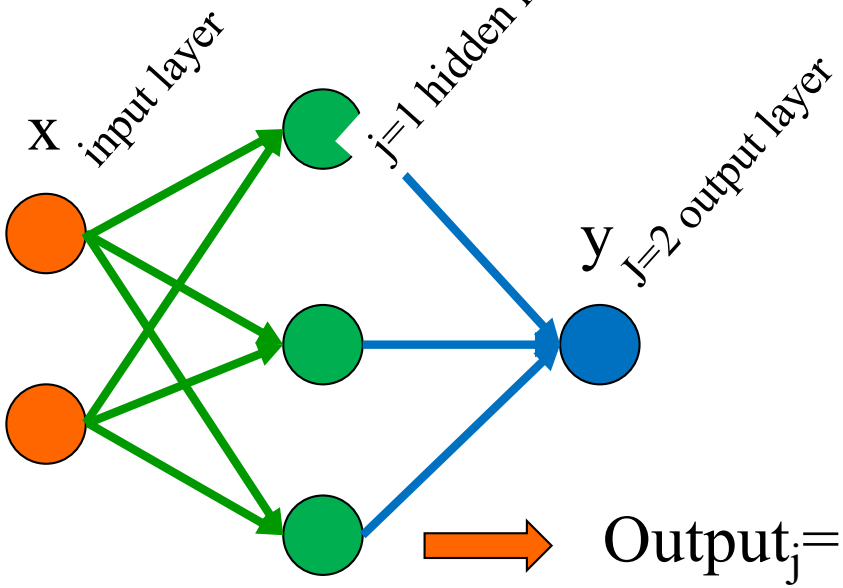

x_n

f

y_p

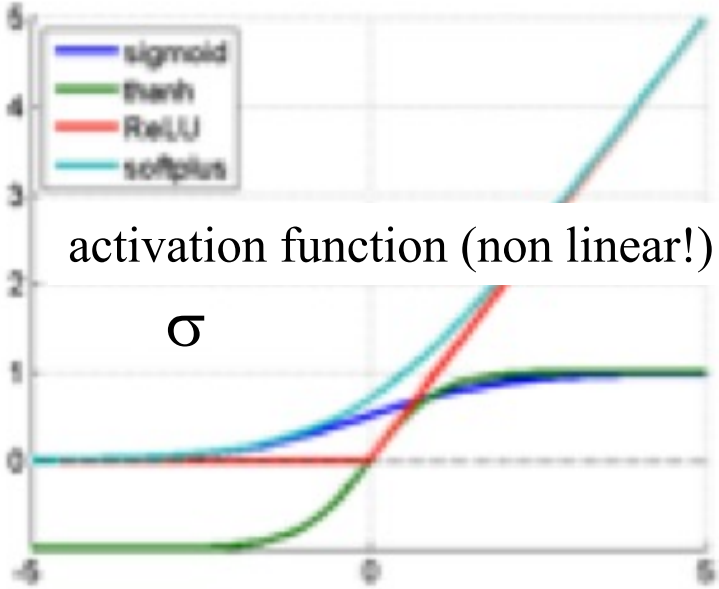

- Handwritten text ➔ text
- Picture ➔ Aashvi or Priyanka ?
- Image ➔ cat or dog ?
- « Comment ça va ? » ➔ ये कैसा चल रहा है?
- Speech ➔ text
- Chess board ➔ next move
- Camera + GPS ➔ car wheel action
- facebook data ➔ sponsored ads
- « Summarize this text » ➔ text summarized

Neural Network and Universal Theorem

Simple NN







$$\text{Output}_j = \sigma(b_j + \sum w_{ij} x_i)$$

$$h(x) = \sigma(b^2 + W^2 \sigma(b^1 + W^1 x))$$

Beware: superscript are layer indices!

Now with dimensions

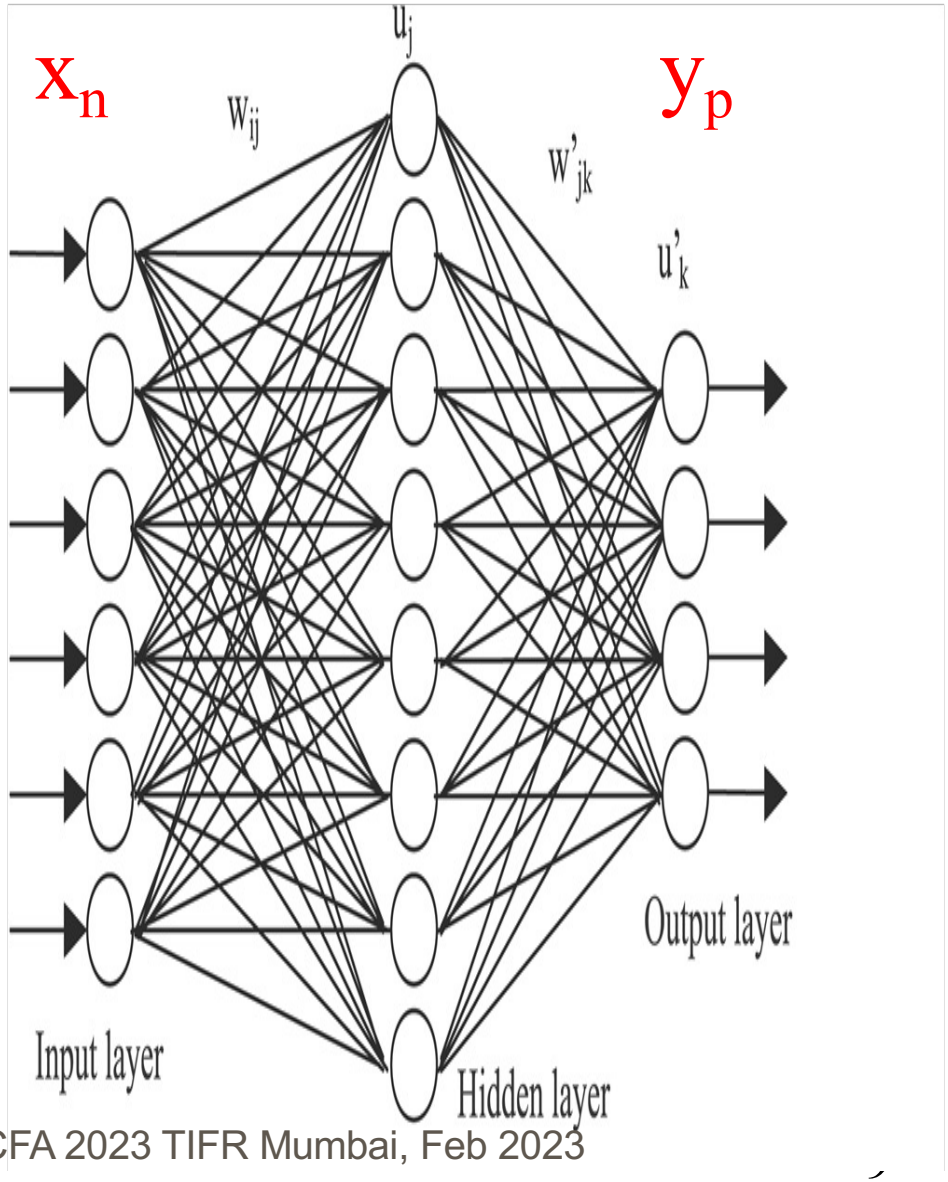
$$h(x_{(2)}) = \sigma(b^2_{(1)} + W^2_{(1,3)} \sigma(b^1_{(3)} + W^1_{(3,2)} x_{(2)}))$$

Biological neuron

- Artificial NN, very <> from biological neuron
- Build (somewhat) intelligent NN, very <> simulate the actual brain

Universal Approximation theorem

- Mathematical theorem 1991
  https://en.wikipedia.org/wiki/Universal_approximation_theorem
- Any continuous, bounded function $R^n \rightarrow R^p$
- … can be approximately sufficiently well (better than a given ε)
- … with a sufficiently large single hidden layer neural net
- But how to build it ?

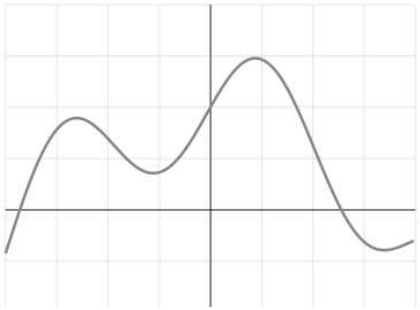


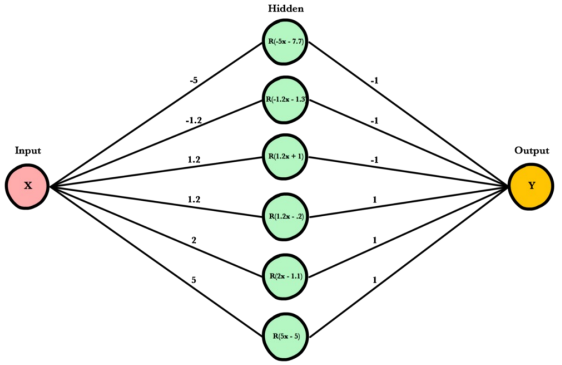
Addendum ResNet 1 neuron sufficient depth

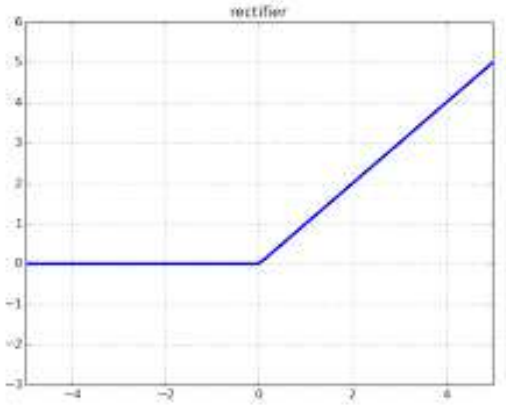
Universal Theorem at work

Universal approximation

We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions







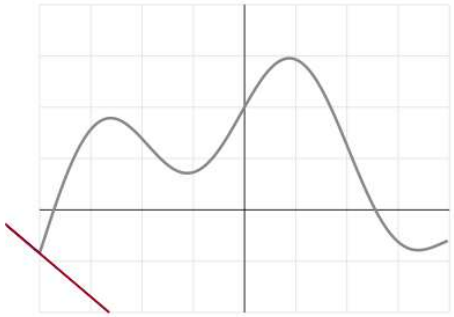
relu(x) = x if x>0 & 0 otherwise

$$y = \sum_i \mathrm{Relu}(\mathrm{a_i} \times \mathrm{x} + \mathrm{b_i})$$

Universal Theorem at work

Universal approximation

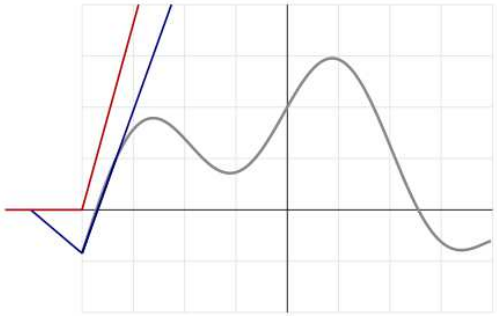
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

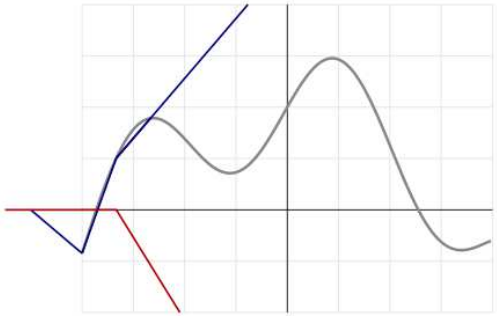
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

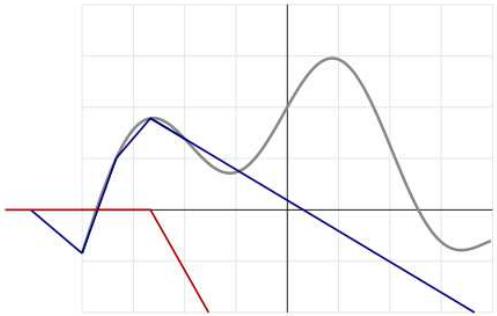
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

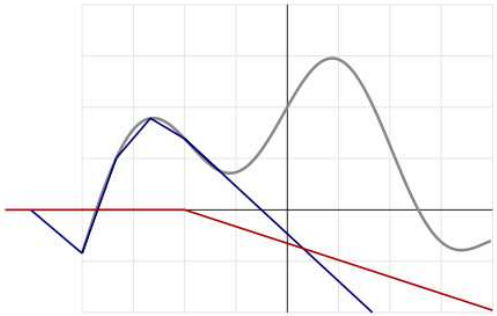
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

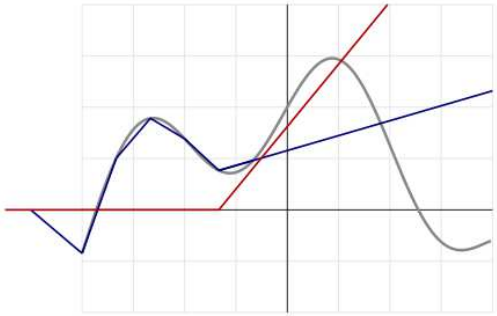
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

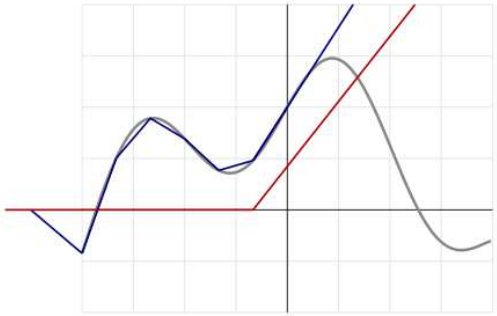
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

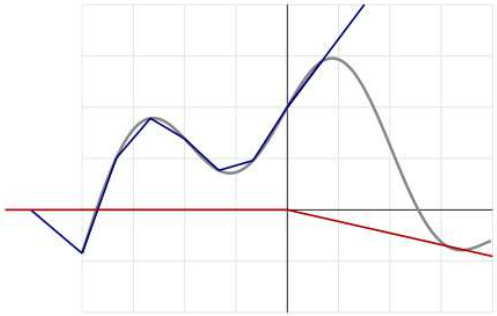
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

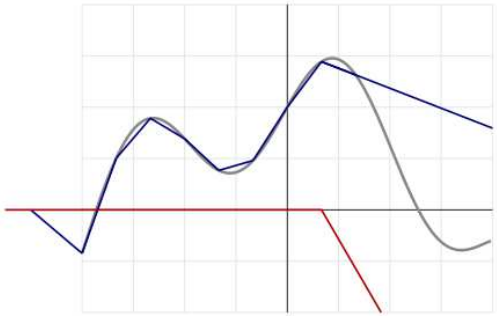
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

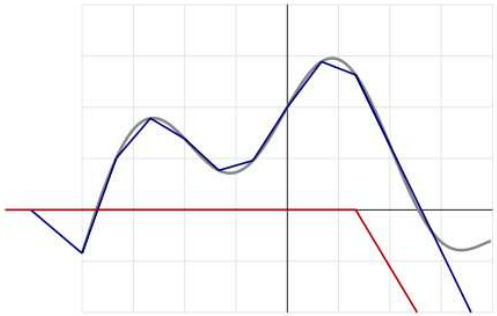
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

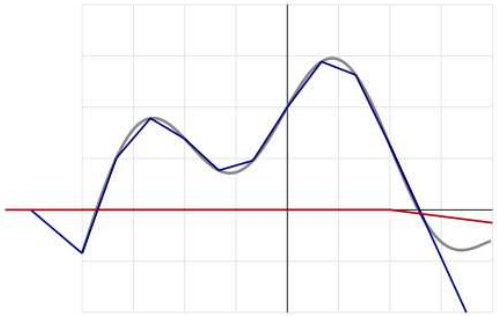
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

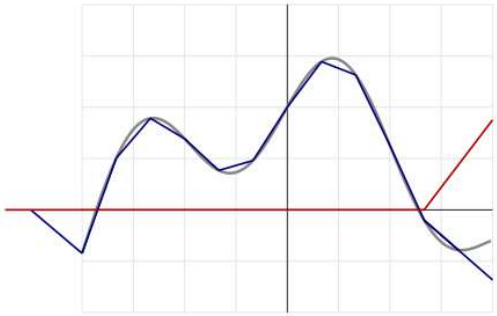
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

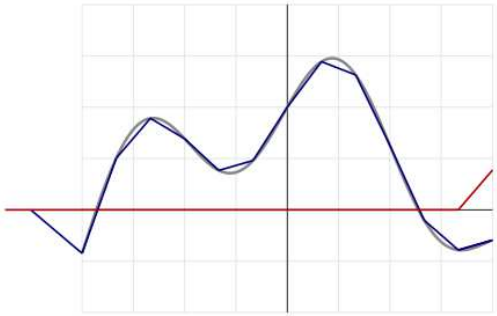
We can approximate any $f \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions



Universal Theorem at work

Universal approximation

We can approximate any $f \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions



$$\mathbb{R} \to \mathbb{R} \text{ generalised to } \mathbb{R}^n \to \mathbb{R}^p$$

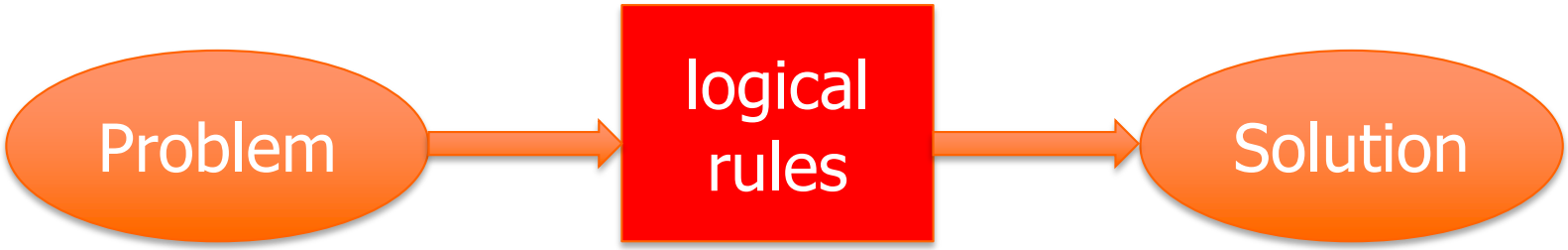
In short…

- If you can express your problem as finding a $\mathbb{R}^n \rightarrow \mathbb{R}^p$ function…
- A NN solving your problem exists…
- But how to build it ?

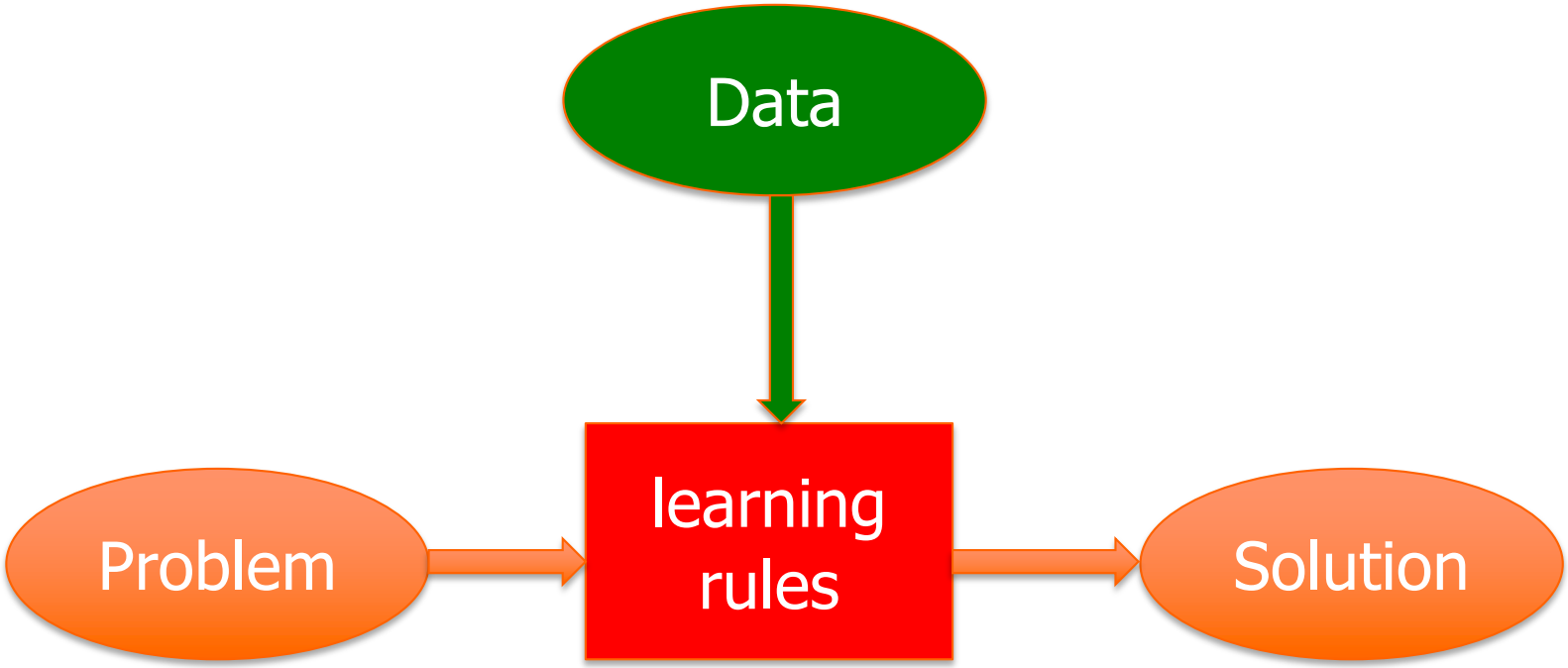
- Universal Theorem with a single hidden layer ? Why all the fuss about Deep Learning ?

Traditional Computing

…also called rule-based computing

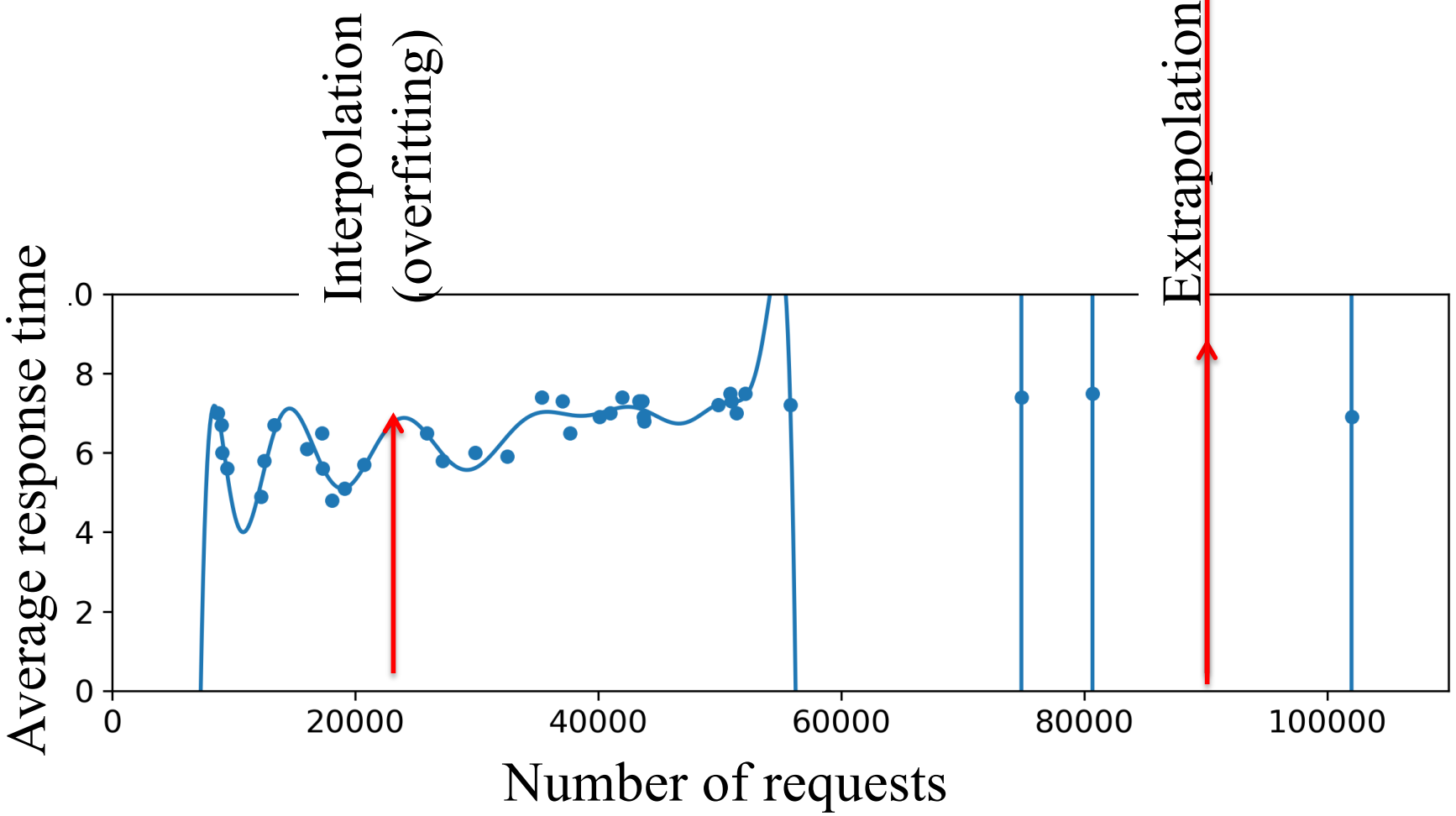


Machine Learning



Learning = optimise internal parameters of the algorithms: n=2 - trillions

Many pitfalls



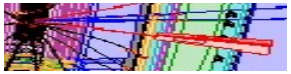
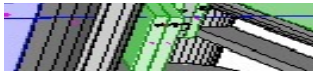
Classification

Given x_n, we want y_p

x_n f y_p

- Handwritten text ➔ text
- Picture ➔ Sofia or Sabrina ?
- Image ➔ cat or dog ?
- « Comment ça va ? » ➔ كيف حالك ؟
- Speech ➔ text
- Chess board ➔ next move
- Camera + GPS ➔ car wheel action
- facebook data ➔ sponsored ads

Given x_n, we want y_p

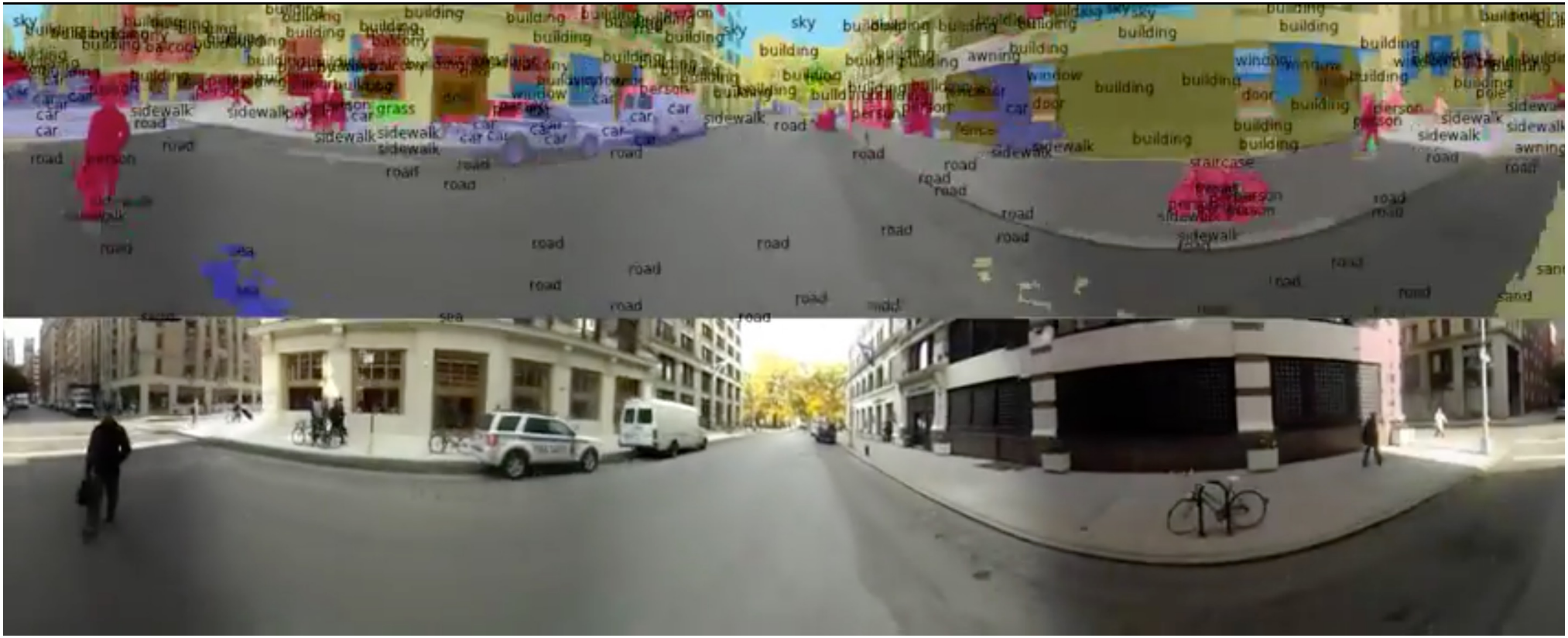
 x_n  f  y_p 

- Handwritten text ➔ text
- Picture ➔ Aashvi or Priyanka ?
- Image ➔ cat or dog ?

Classification

- « Comment ça va ? » ➔ आप कैसे हैं ?
- Speech ➔ text
- Chess board ➔ next move
- Camera + GPS ➔ car wheel action
- facebook data ➔ sponsored ads
- « Summarize this text » ➔ text summarized

Classification is everywhere

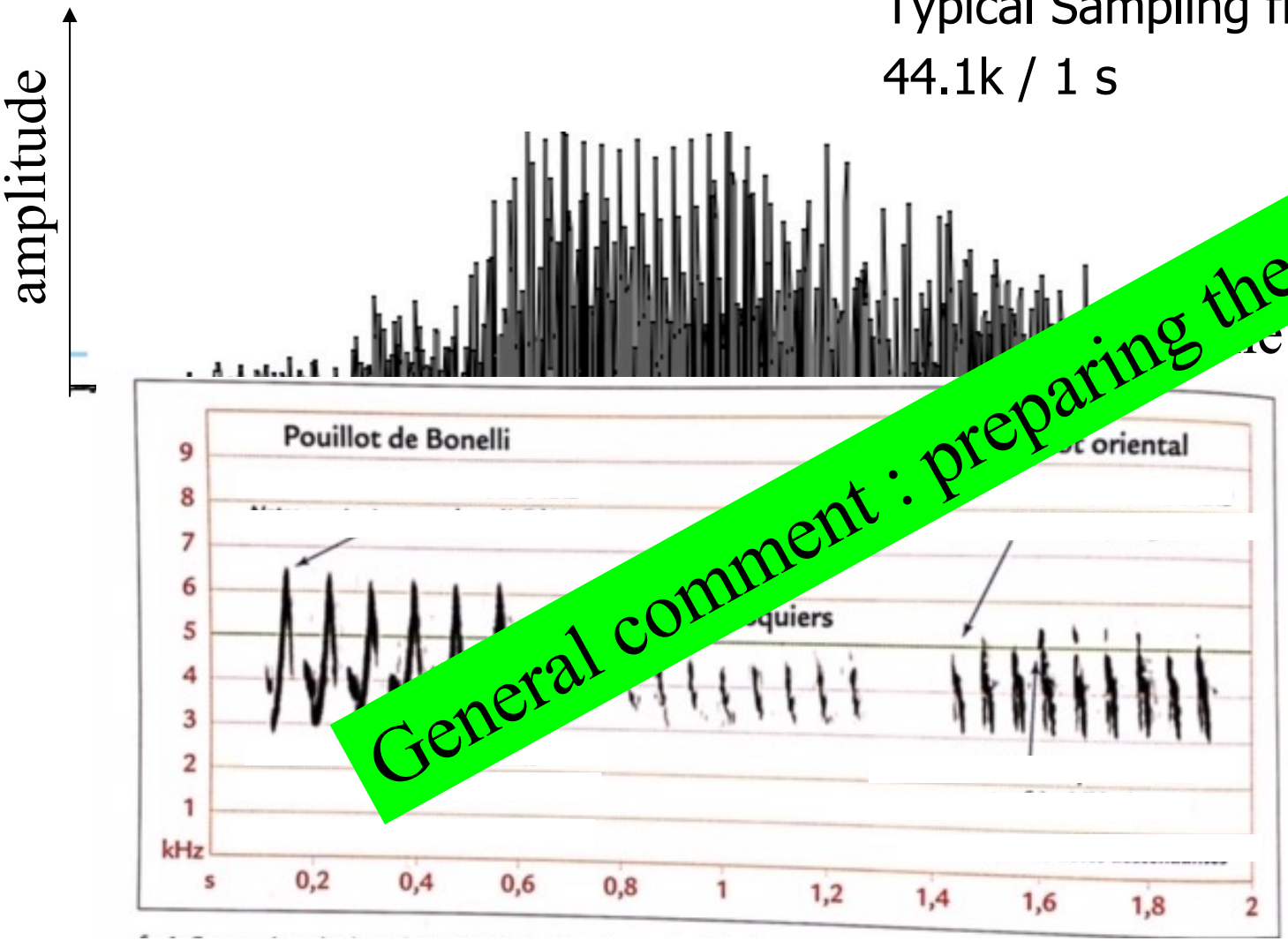


Inputs

Inputs

Typical Sampling frequency 44.1kHz

44.1k / 1 s



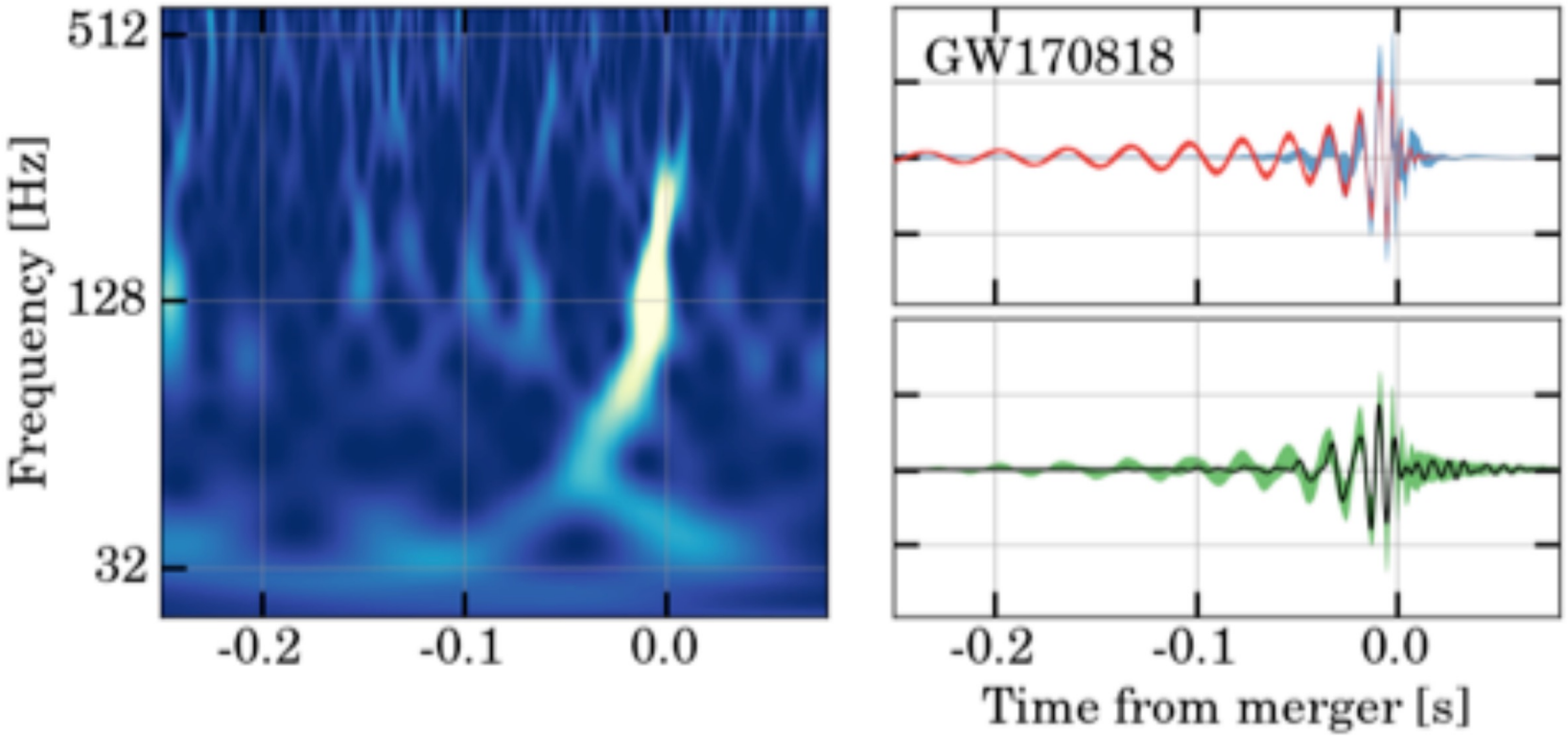
General comment : preparing the data

➔images

Time-frequency diagram

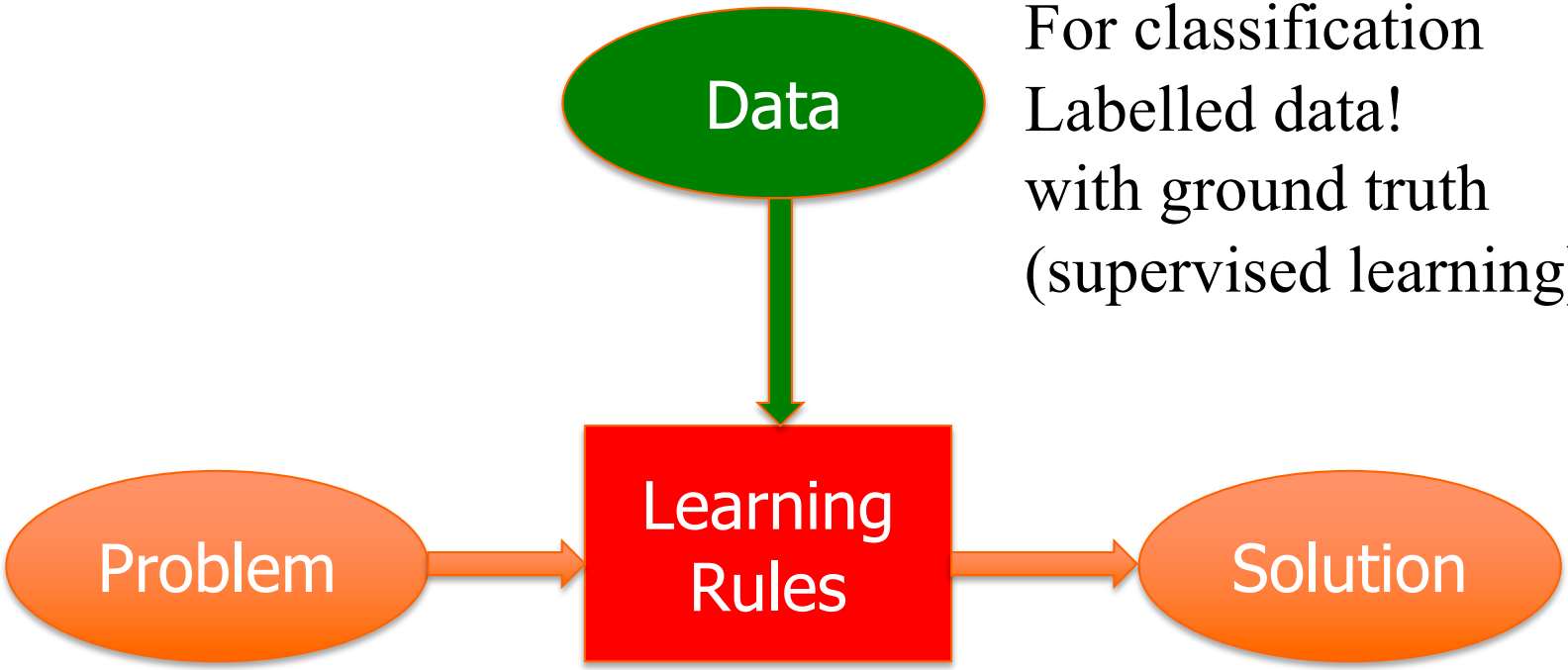
Stanislas Wroza

https://igfae.usc.es/igfae/ligo-and-virgo-announce-four-new-gravitational-wave-detections/



Chirp

Machine Learning



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

Data label example

Titanic dataset

```
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

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*

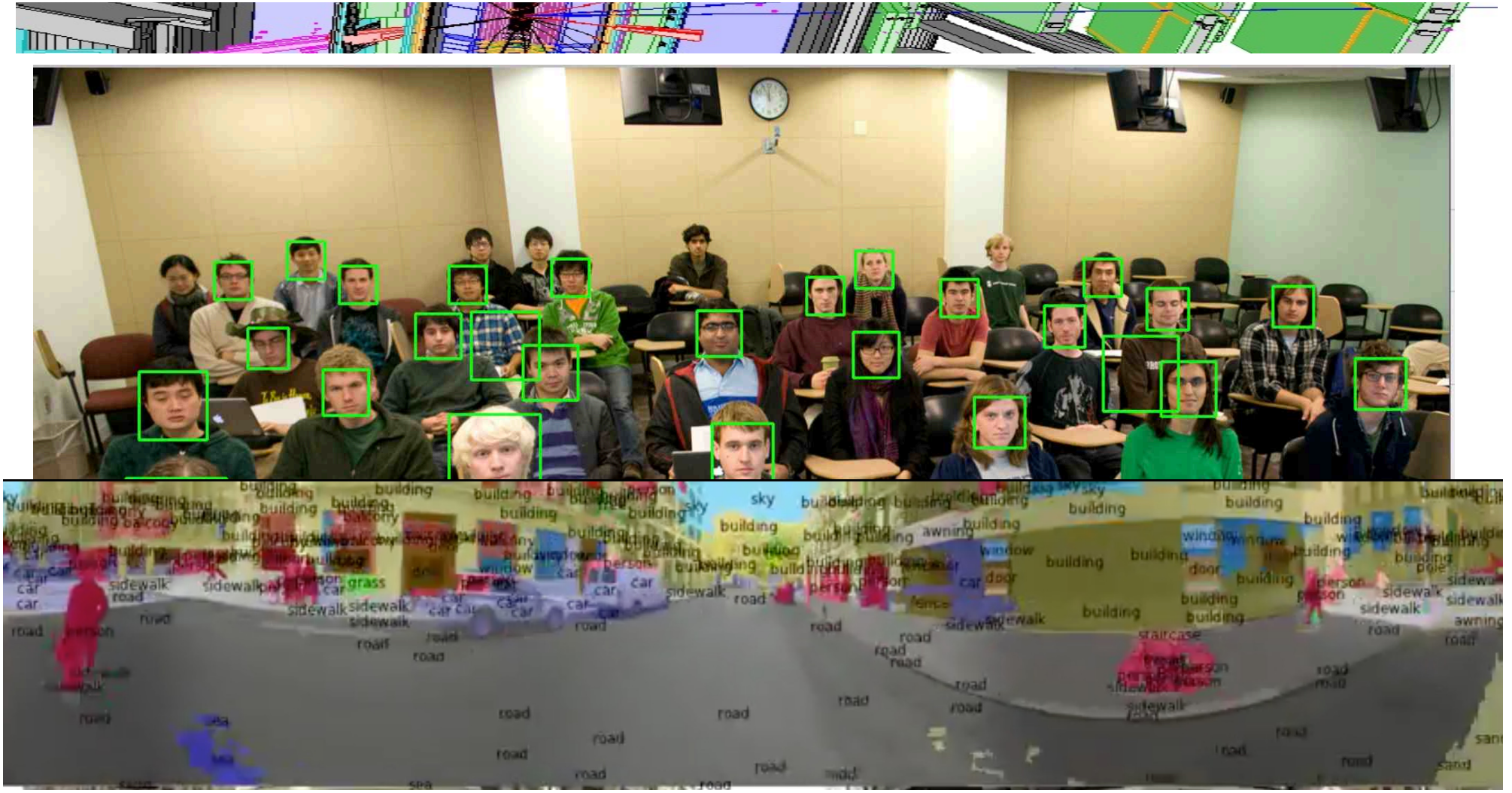
- Natural Language is not tabular
- Also features like : « Age of the children »
  - [],[3], [3,7,18]
- Closer to physics : « Energy of the jets in this proton collision »:
  - [],[120.5],[509.2,439.1,123.6,13.3]
- Special techniques to deal with these

Output label

- Two classes, usually :
  - y=0 ➔ background
  - y=1 ➔ signal
- N classes :
  - Nearly never used one output neuron with:
    - y=0➔cat
    - y=1➔dog
    - y=2➔rabbit
  - Rather use « one-hot vector », 3 output neurons:
    - y=[1,0,0]➔cat
    - y=[0,1,0]➔dog
    - y=[0,0,1]➔rabbit

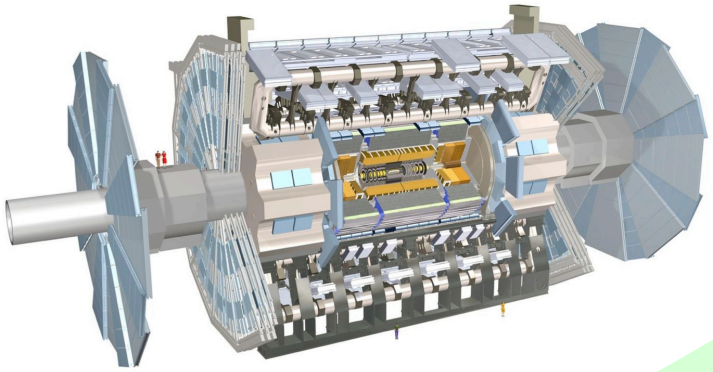
Training dataset

Data label example



➔most Computer Vision tasks need human labelling!

Data label example



Data

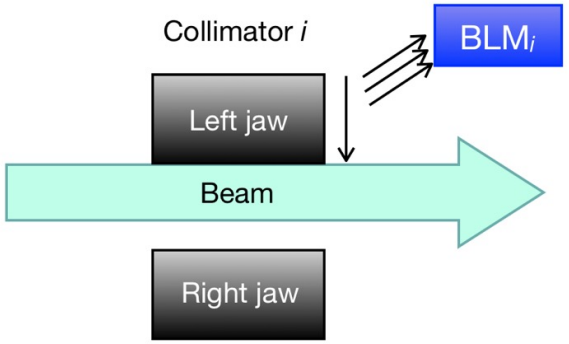
High Energy Physics specificity: we do have very accurate simulators, access to labelled data is usually not an issue

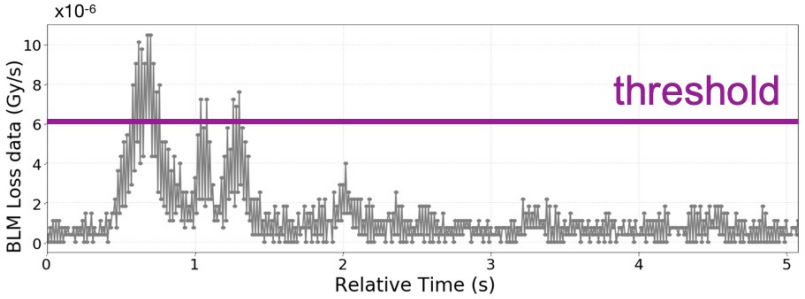
Simulator

Provide pseudo-data with labels

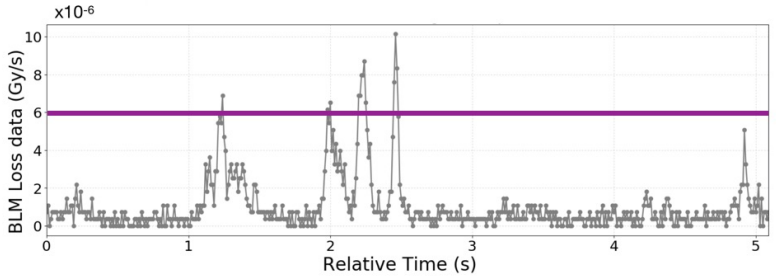
LHC spikes classification

G. Azzopardi, et al., NIM-A, 2019. Frederik Van Der Veken , EPS HEP 2019





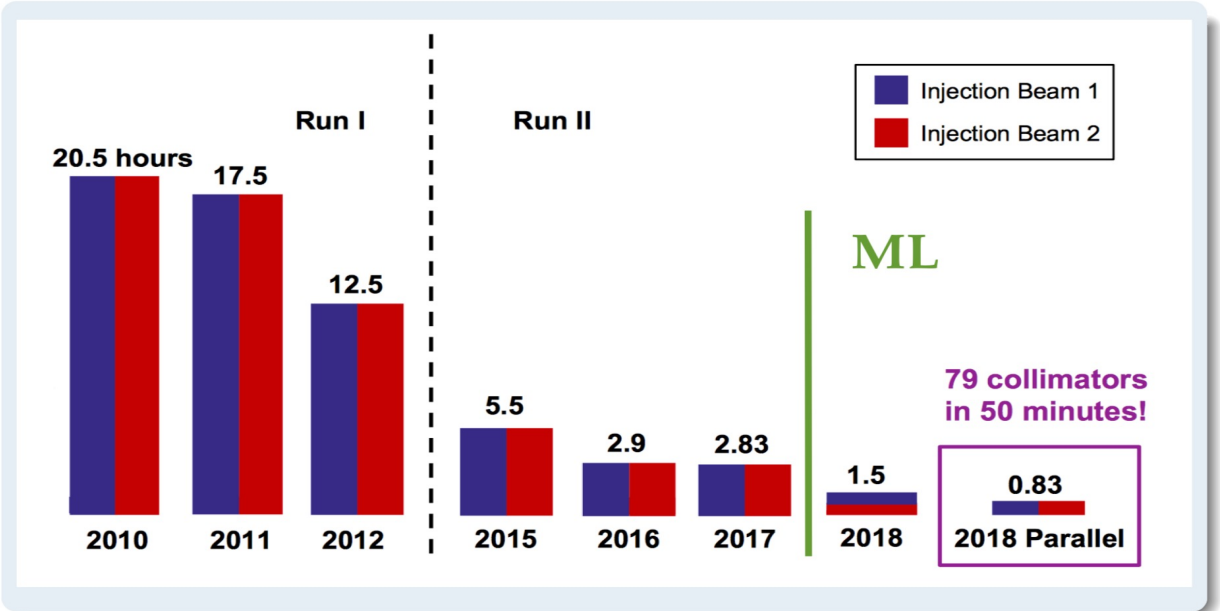
« signal »



« background »

Classifier trained from ~9000 real spikes labelled by hand (!) (rare in HEP)

Now used in production for automatic collimator alignment



Classification performance

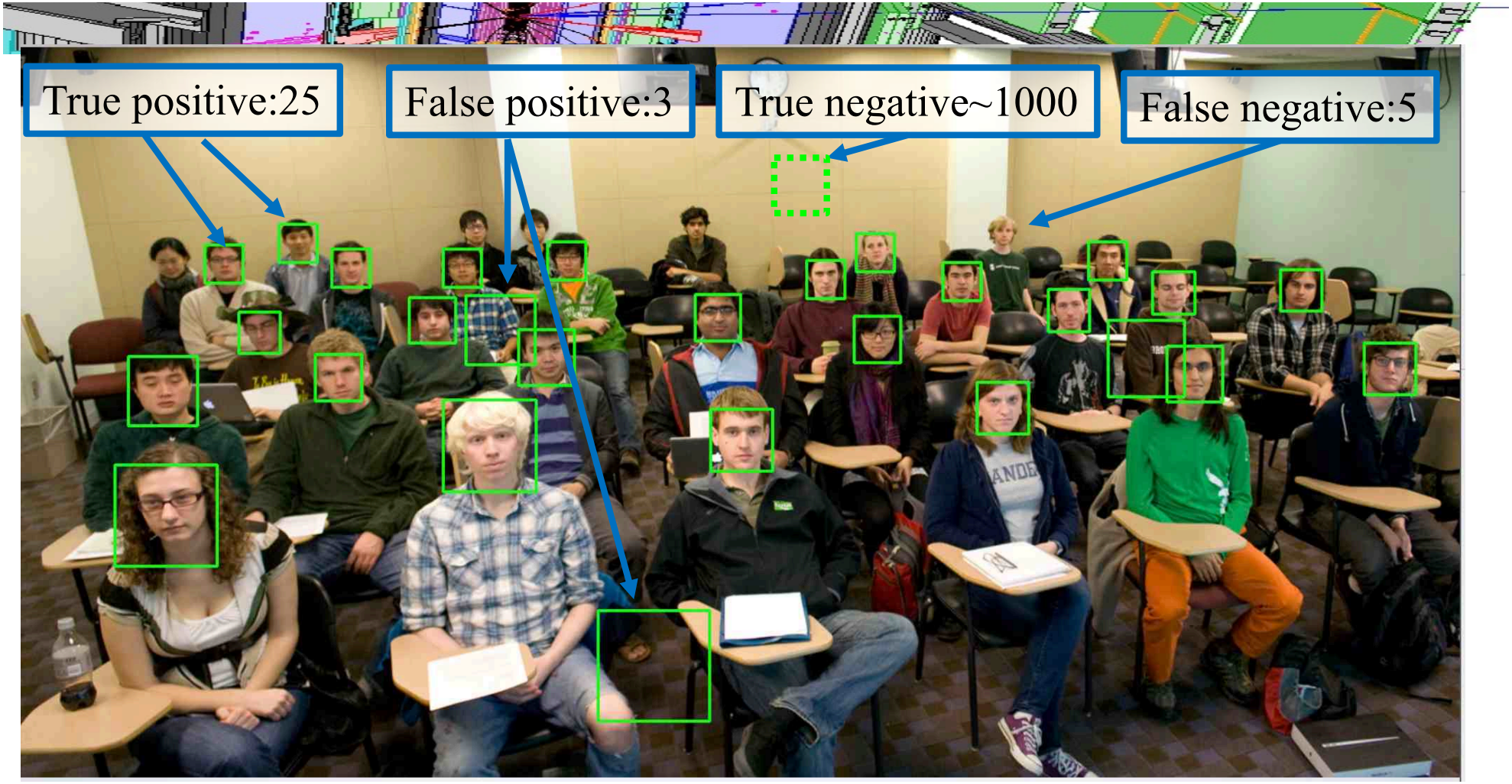
	Signal	Background
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

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)

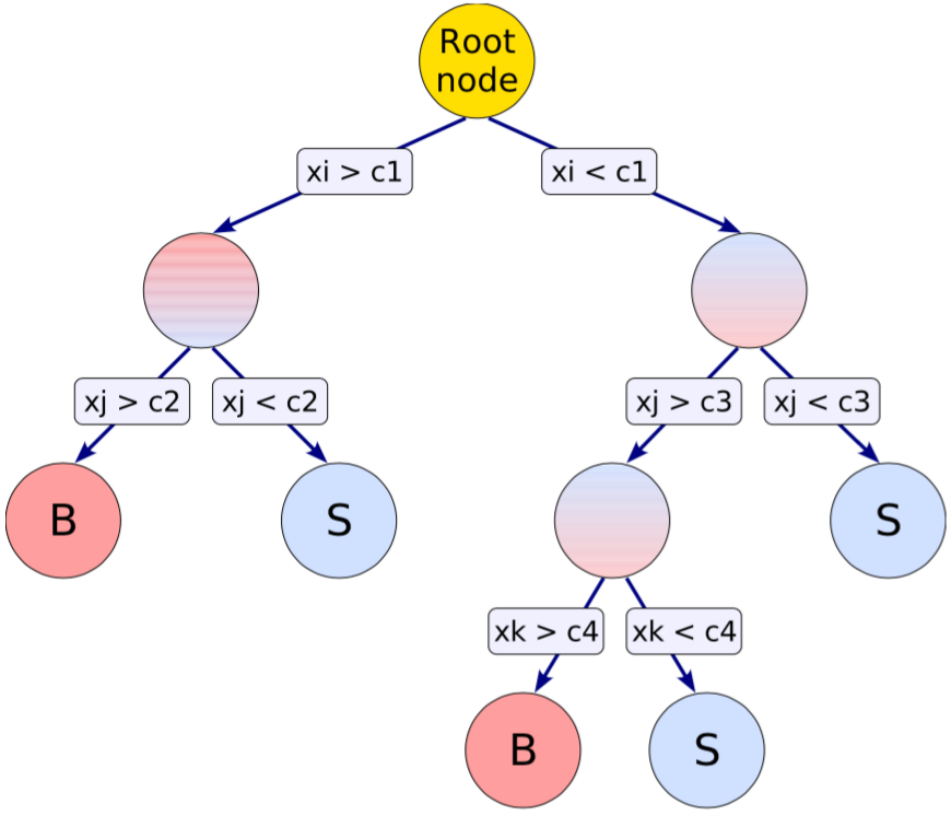
Purity==(ML) Precision=89%=25/(25+3)

In a nutshell

- "Classifier"
  - A model F(x)=y : with y=0 or 1 (e.g. 0=Background 1=Signal)
    - In general in physics we like to have a "critic" which outputs y ranking variable, the larger the more signal-like, on which we apply a treshold
  - 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

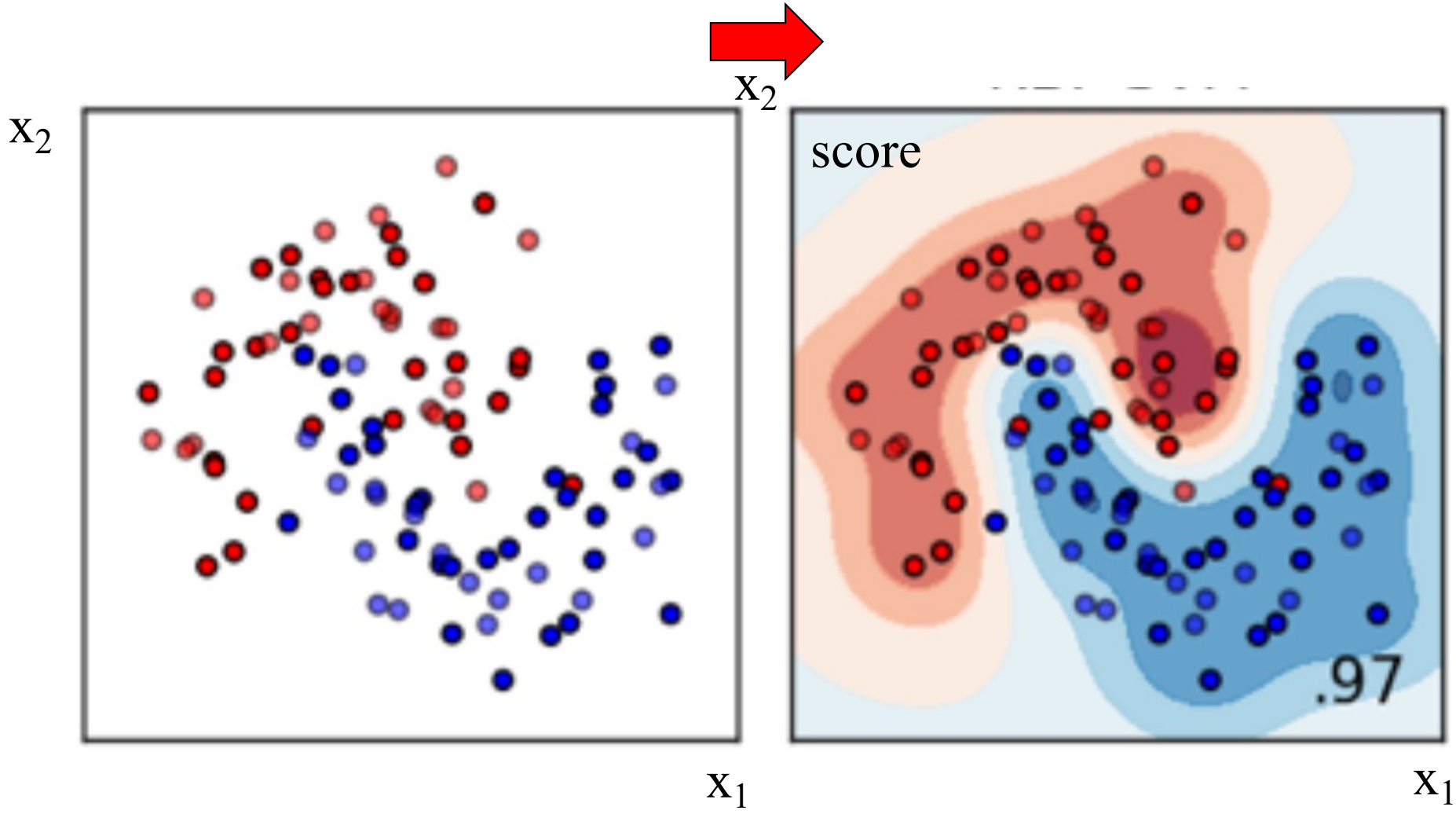
How does it work?

Boosted Decision Tree

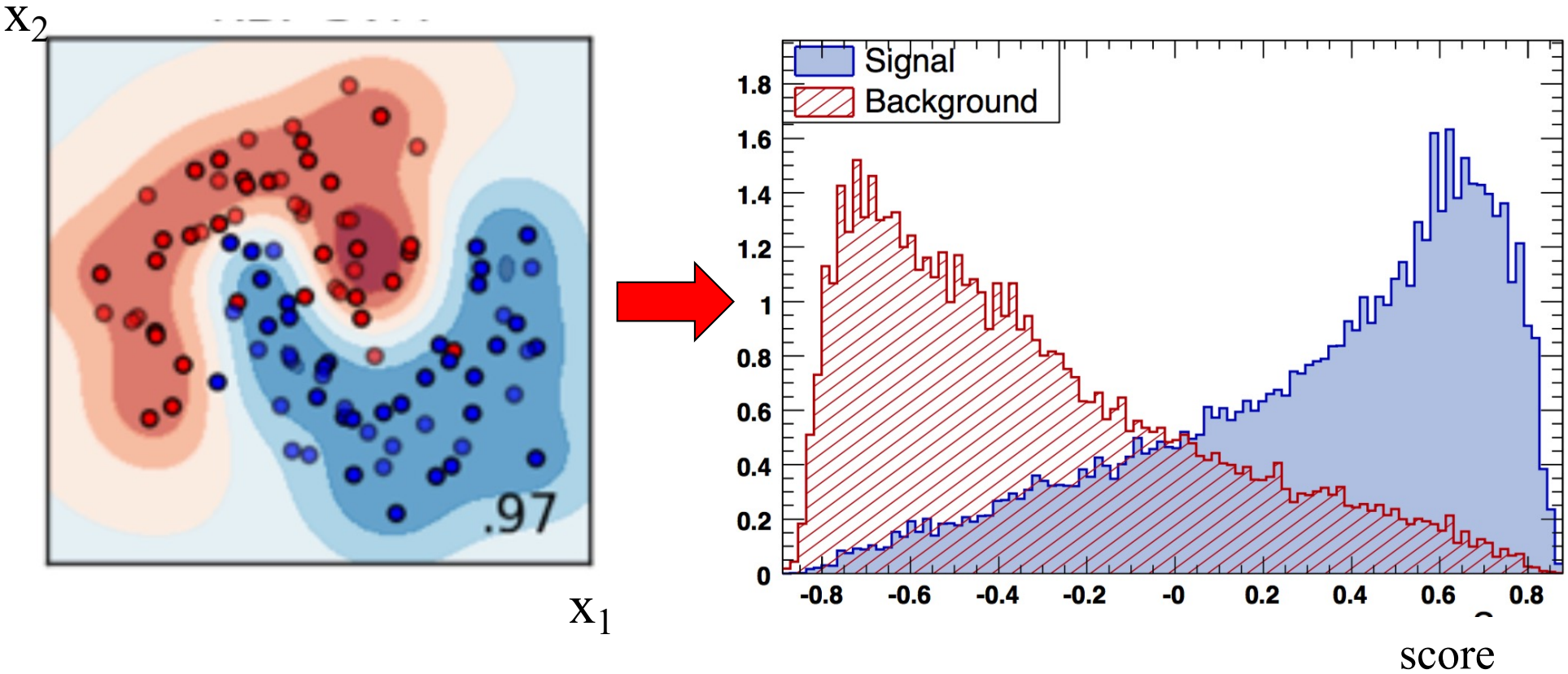


ROC curve, AUC and more

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

Cumulative Density Function

Random variable X

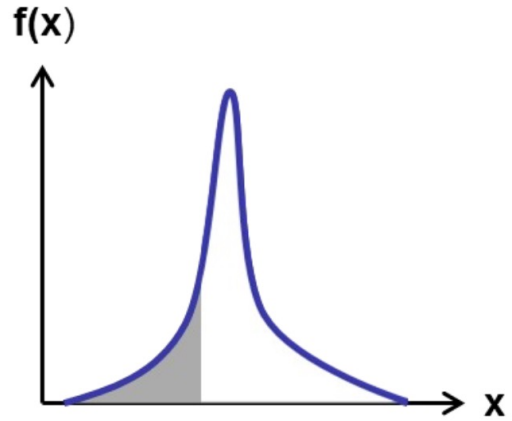
Julien Donini

Discrete random variable: result (realizations) $x_i \in \Omega$ with probability $P(x_i)$

→ **P** is the **probability distribution** and $\sum_i^N P(x_i) = 1$

For continuous variable: probability of observing x in infinitesimal interval

→ Given by the **probability density function** (p.d.f) f(x)



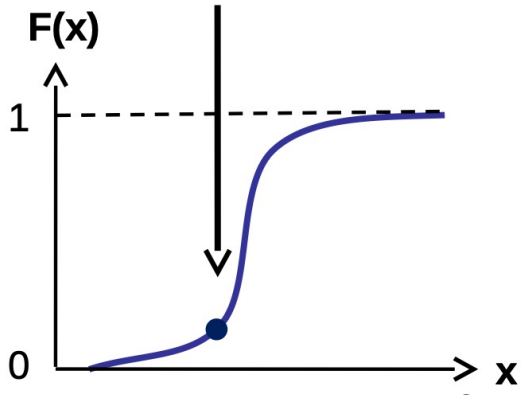
Probability of x in $[x, x+dx] = f(x)dx$

Probability of x in $[a,b] = \int_a^b f(x)dx$

with: $\int_\Omega f(x)dx = 1$

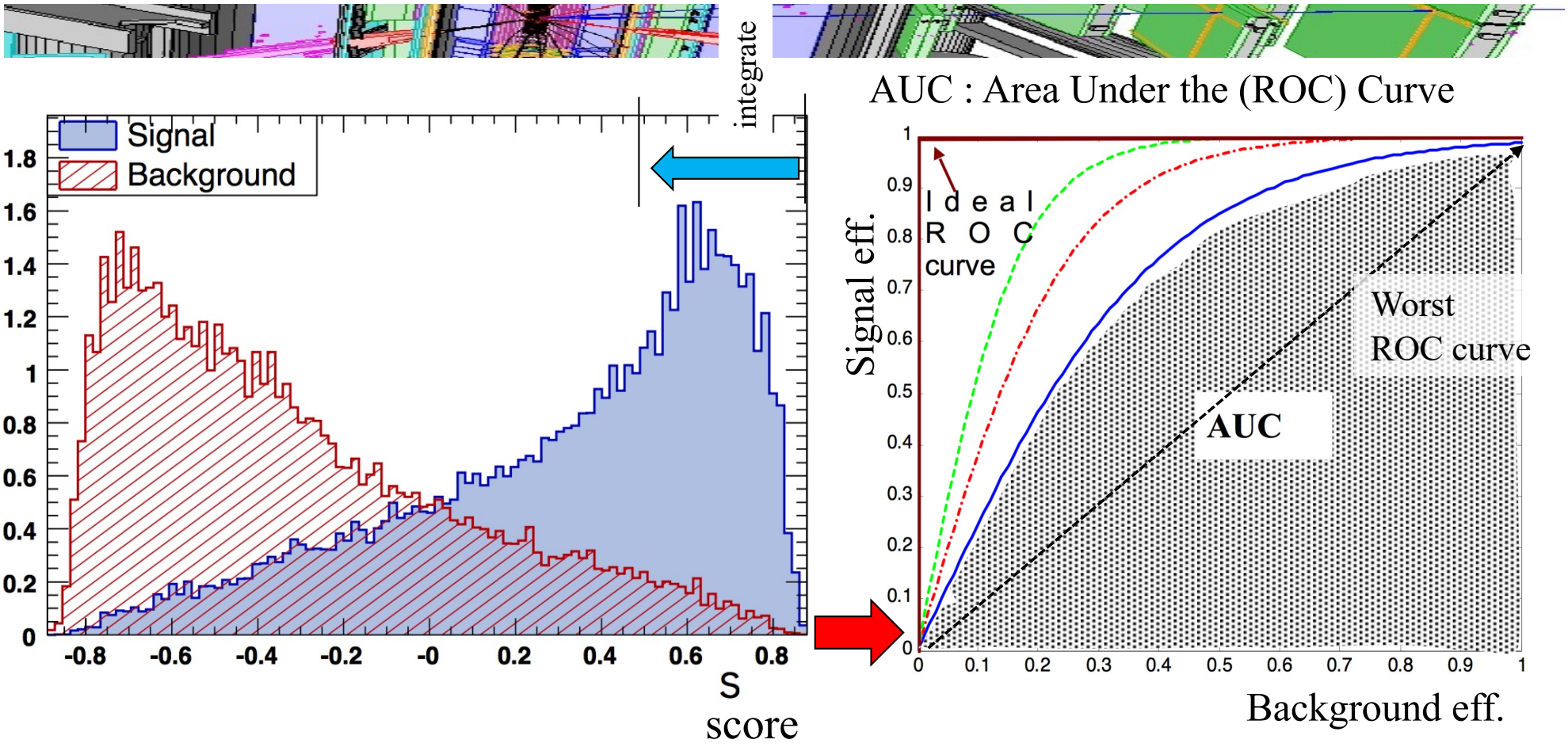
→ **Cumulative distribution** F(x): $F(x) = \int_{-\infty}^{x} f(x')dx'$

hence: $f(x) = \frac{dF}{dx}(x)$



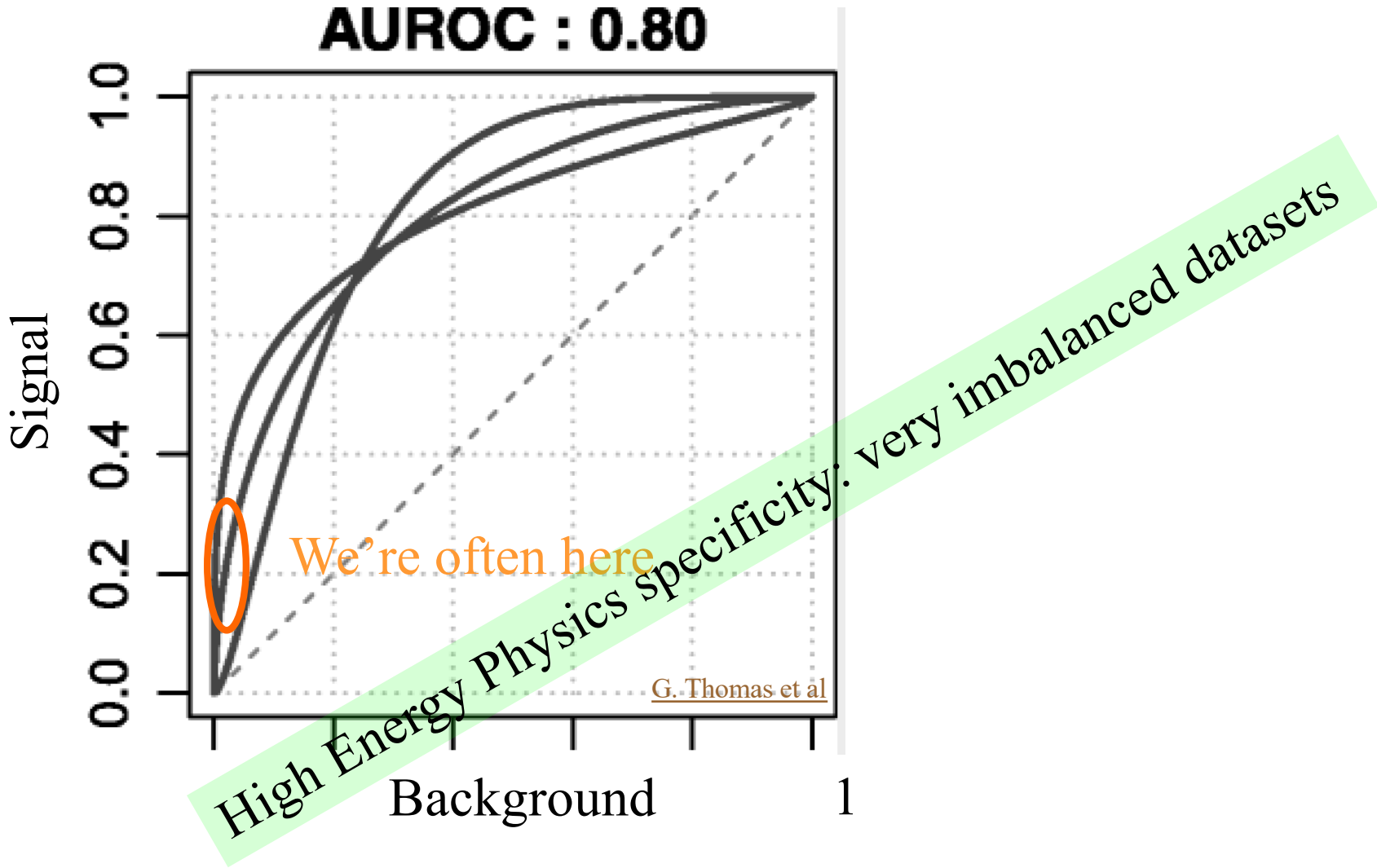
Also: « survival » distribution : 1-F(x)

ROC Curve



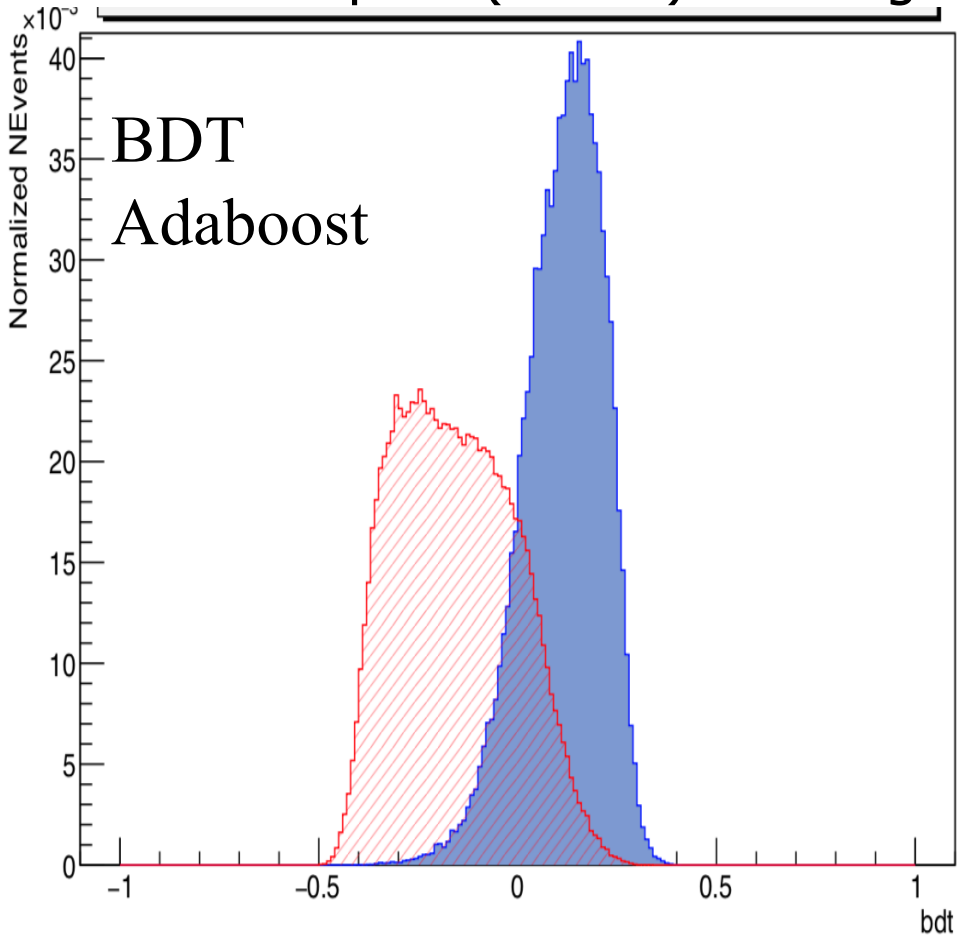
Ideal ROC curve : AUC=1
Worst ROC curve : AUC=0.5
The higher the AUC the better
However AUC not the full story

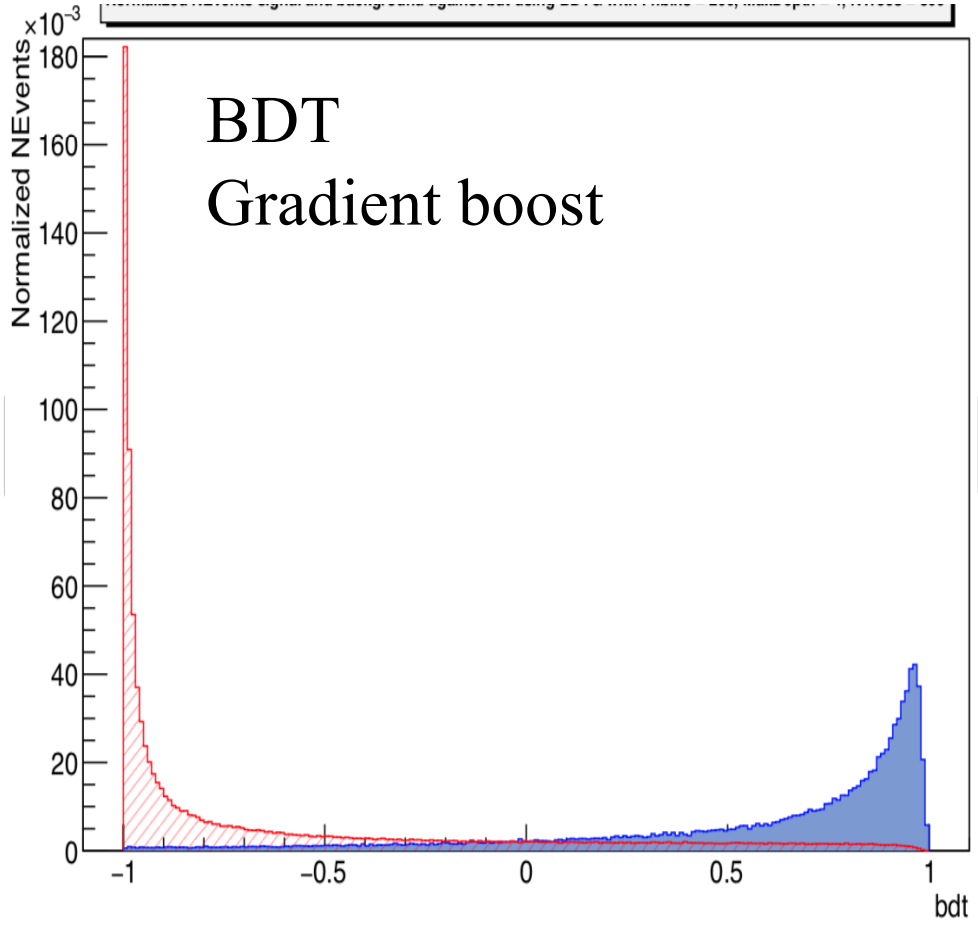
ROC curve pitfall



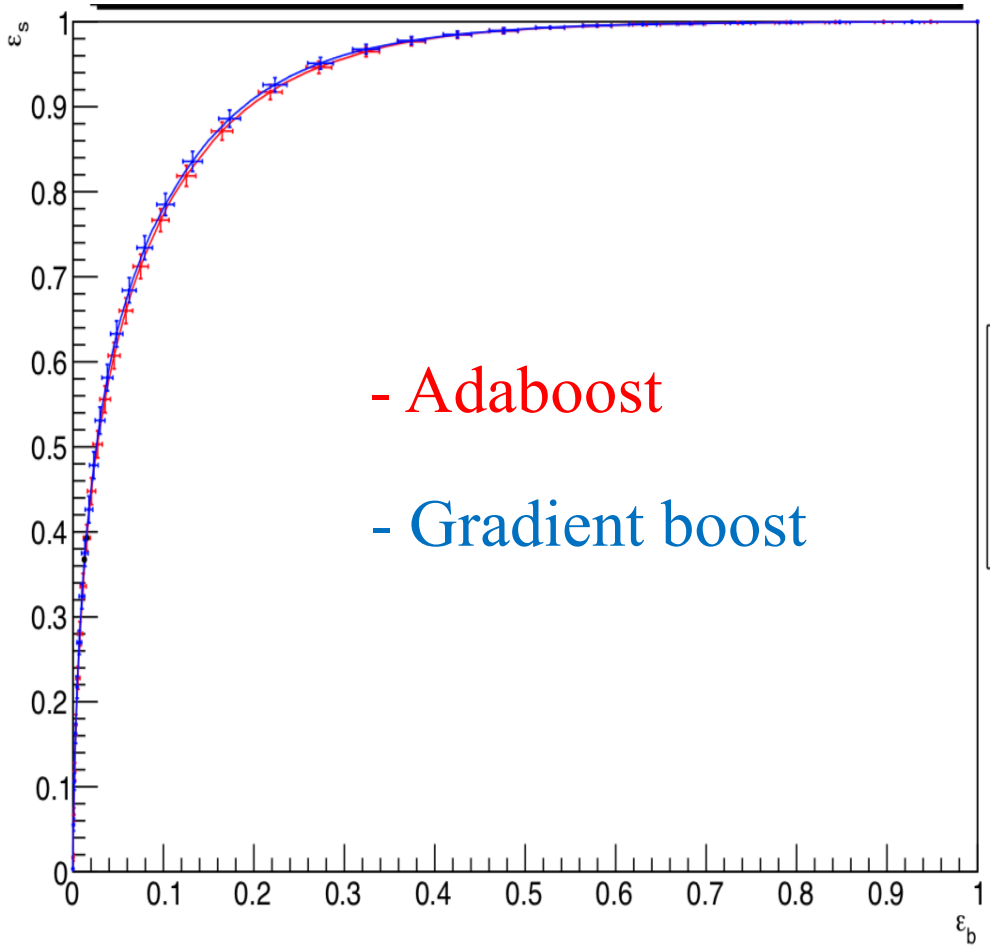
Score

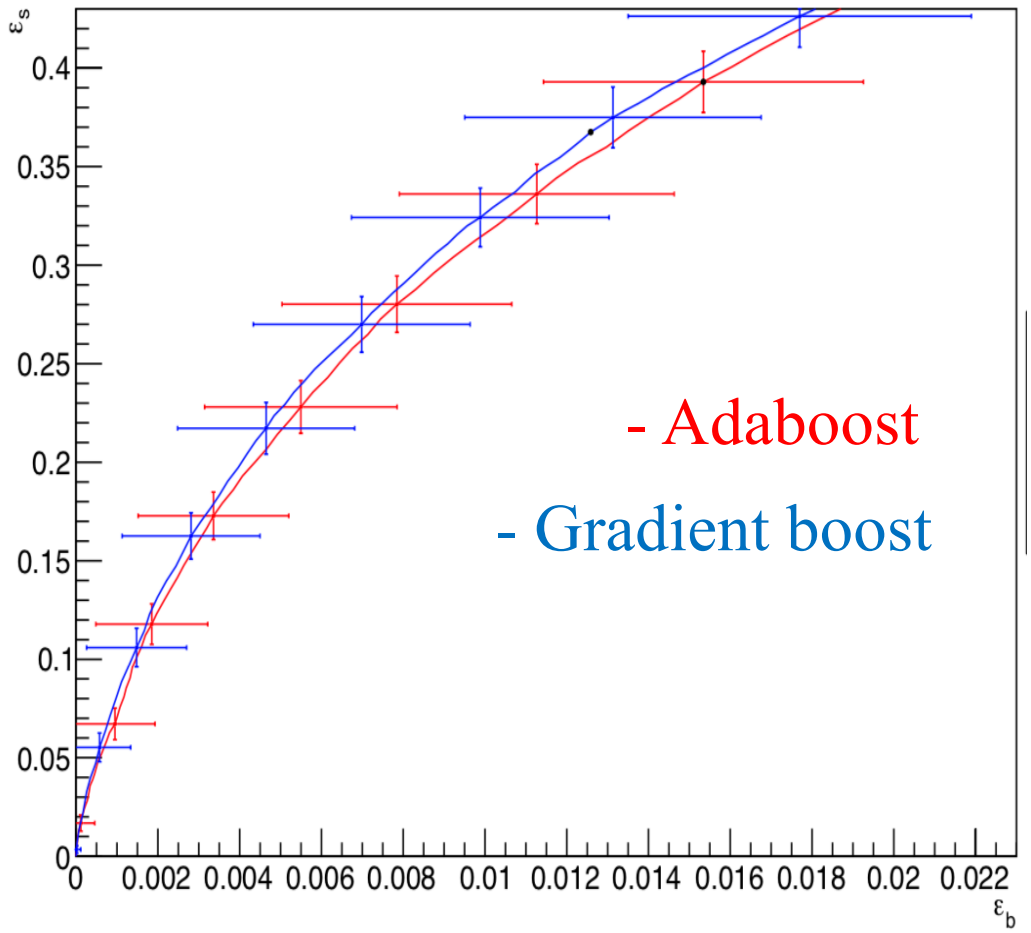
- Score output by algorithm is actually a ranking-variable
  - From most background like to most signal-like
  - Shape is (almost) meaningless





Score (2)


ε_s
ε_b
- Adaboost
- Gradient boost


ε_s
ε_b
- Adaboost
- Gradient boost

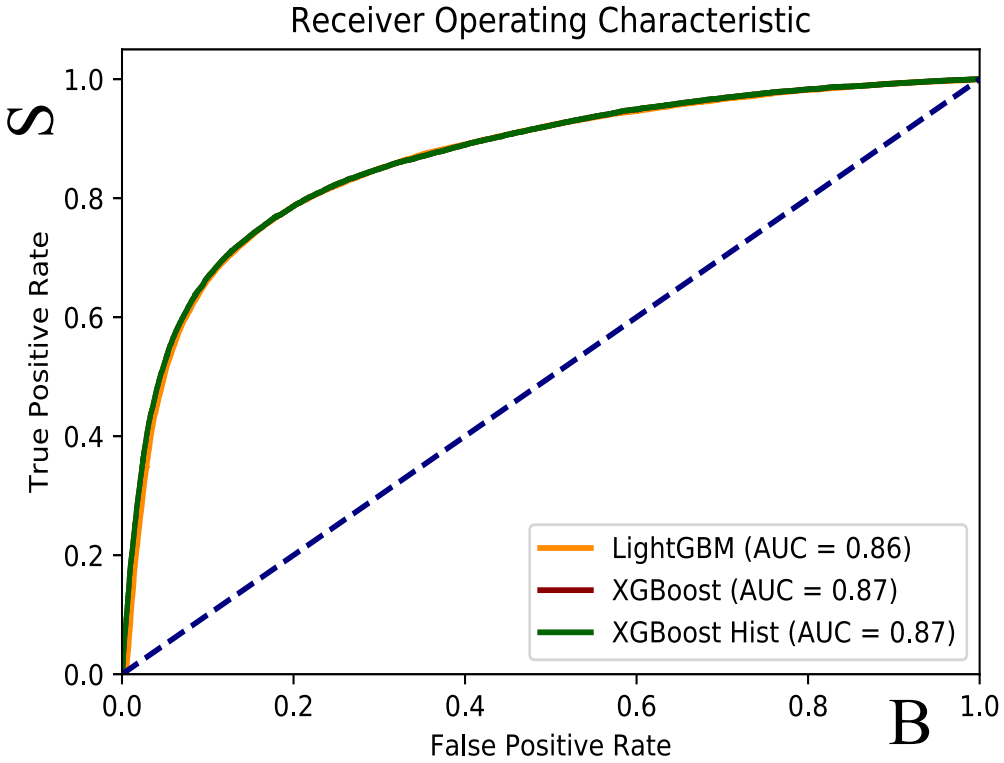
Significance

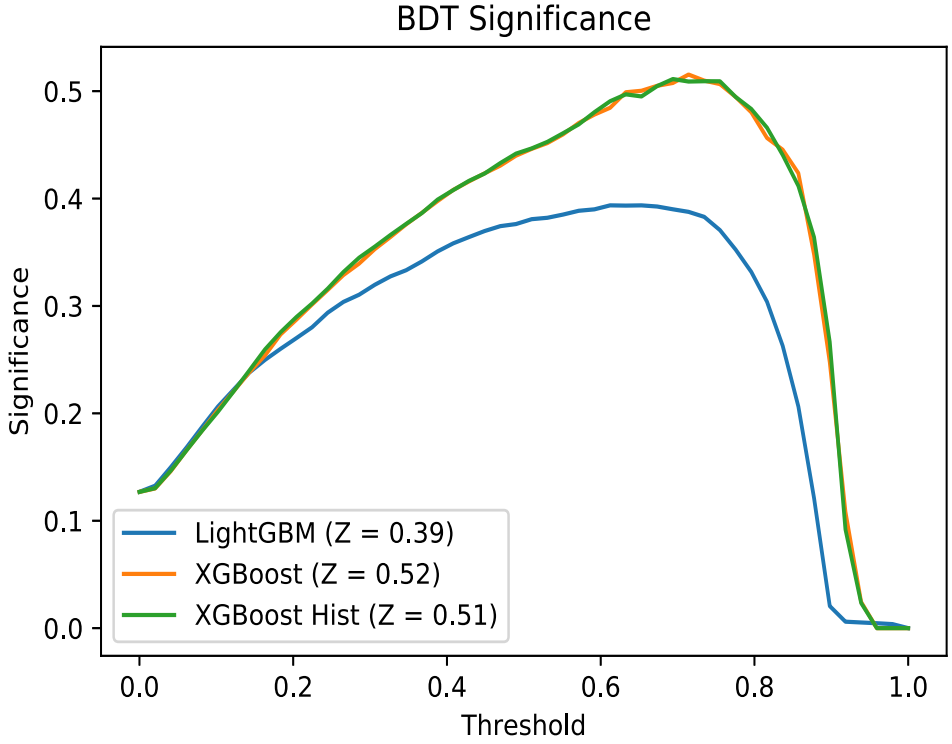
- Imagine counting experiment above threshold:
  - Number of expected Signal S Background B events
  - Expected « Asimov » significance:

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

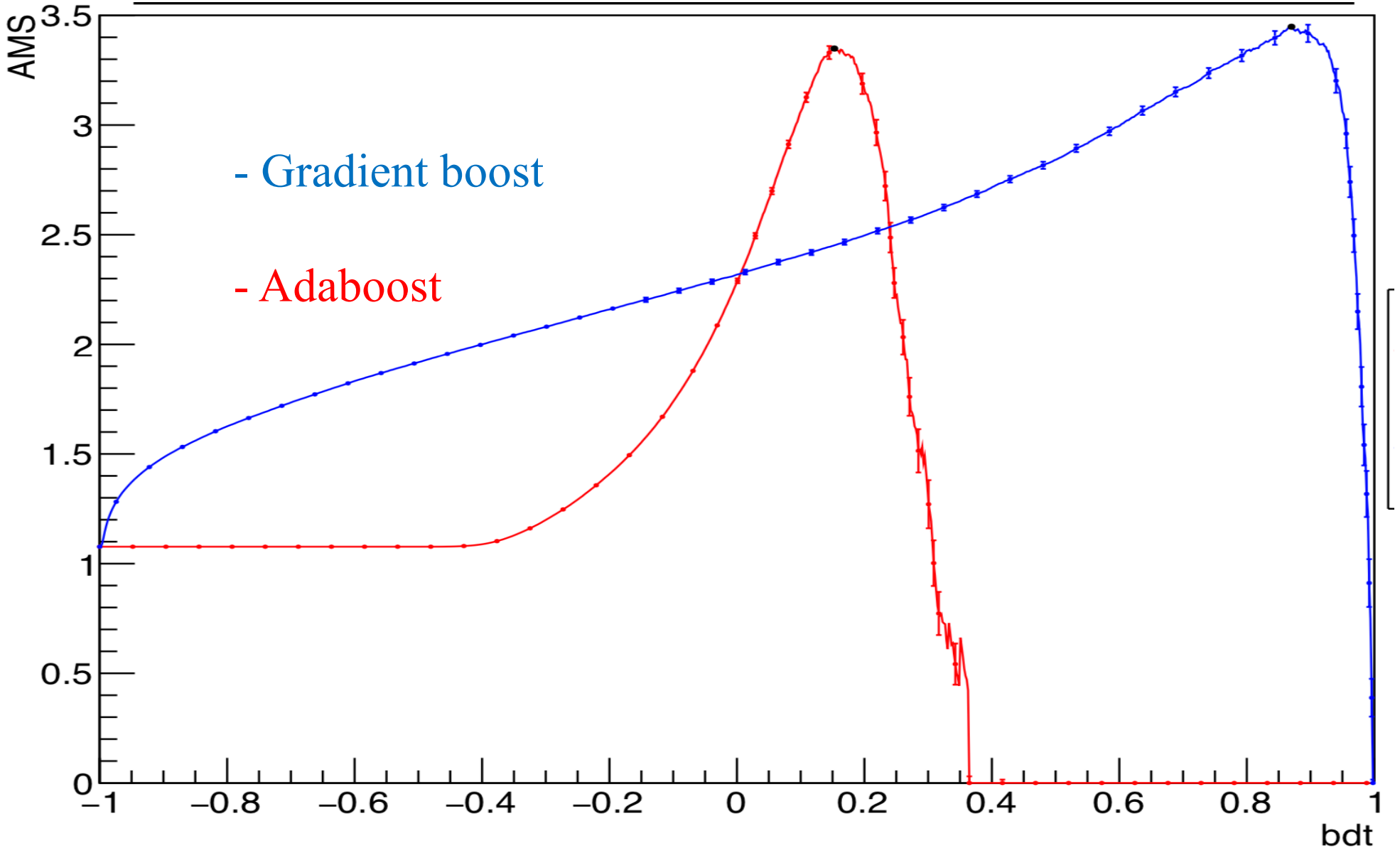
  - ~ $\frac{s}{\sqrt{b}}$ if s<<b Poisson statistics

Cowan et al arXiv:1007.1727 (eq. 97) (cited 8000 times!)

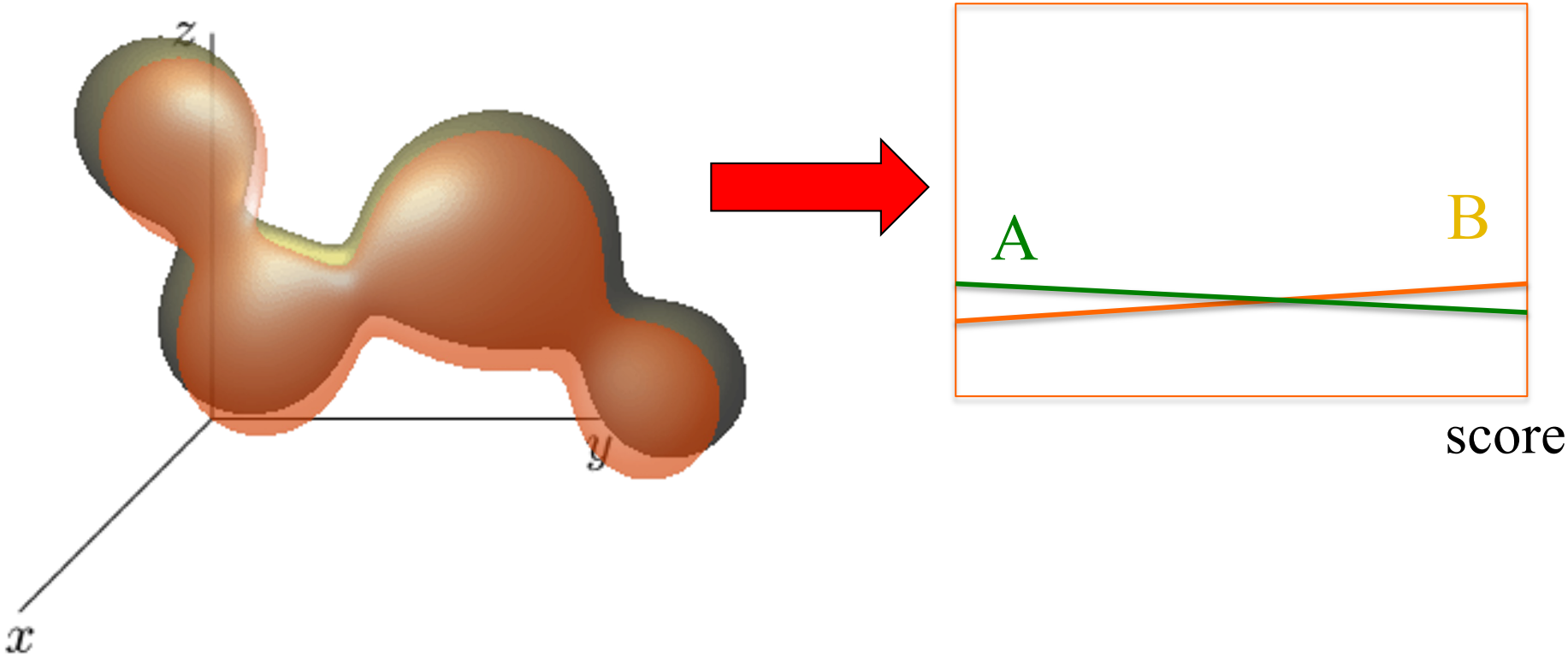




Significance (2)



What does a classifier do?

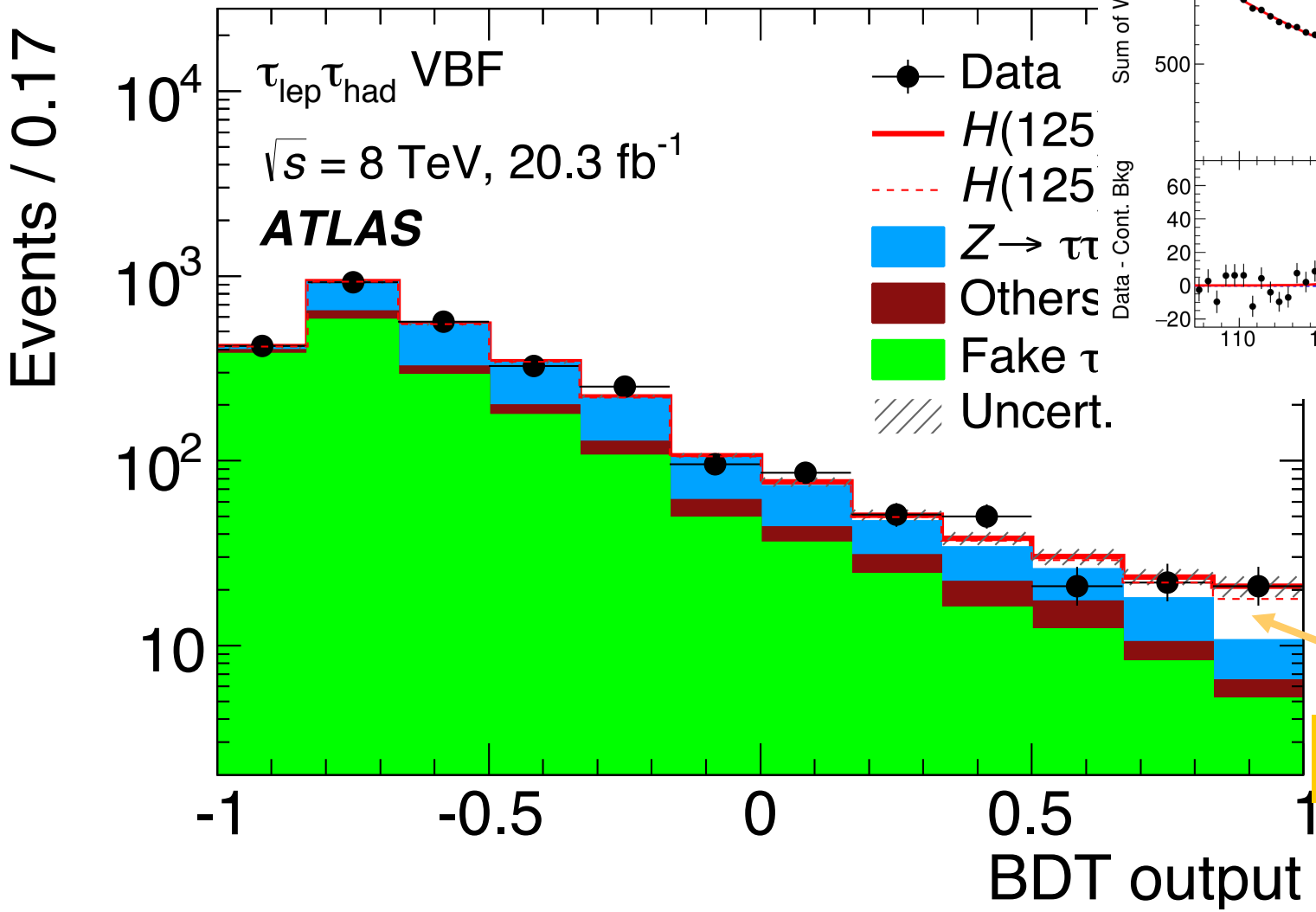


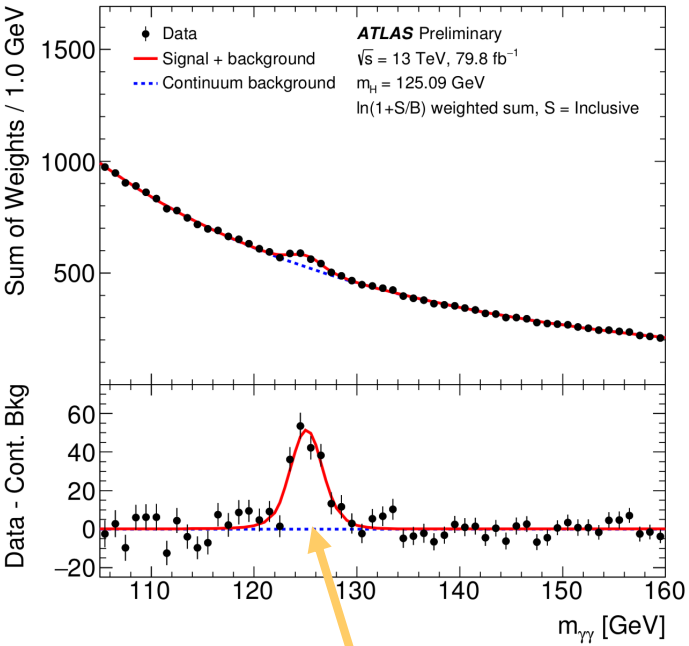
- The classifier "compresses" the two multidimensional "blobs" maximising the difference, without (ideally) any loss of information

Classifier in Higgs Physics

JHEP 04, 117 (2015) 1501.04943

BDT using ~dozen of high level variables

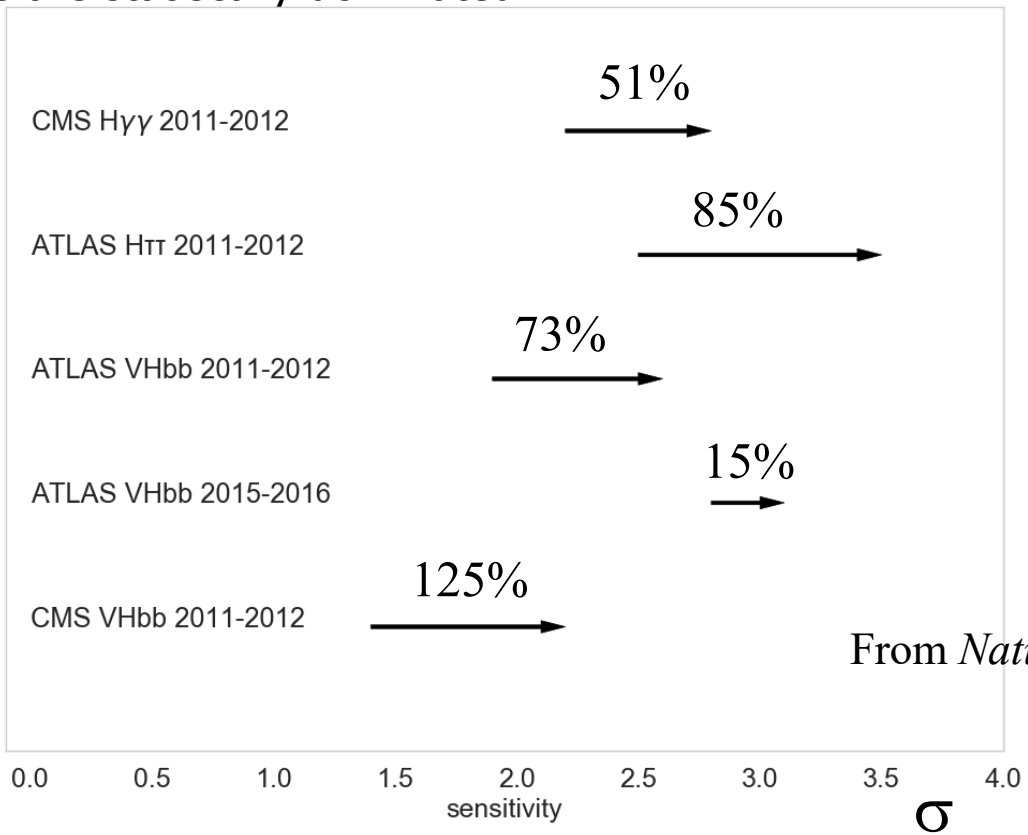




Higgs evidence

ML on Higgs Physics

- At LHC, Machine Learning used almost since first data taking (2010) for reconstruction and analysis
- In most cases, Boosted Decision Tree with Root-TMVA, on ~10 variables
- For example, impact on Higgs boson sensitivity at LHC, conservatively assuming measurements are statiscally dominated:

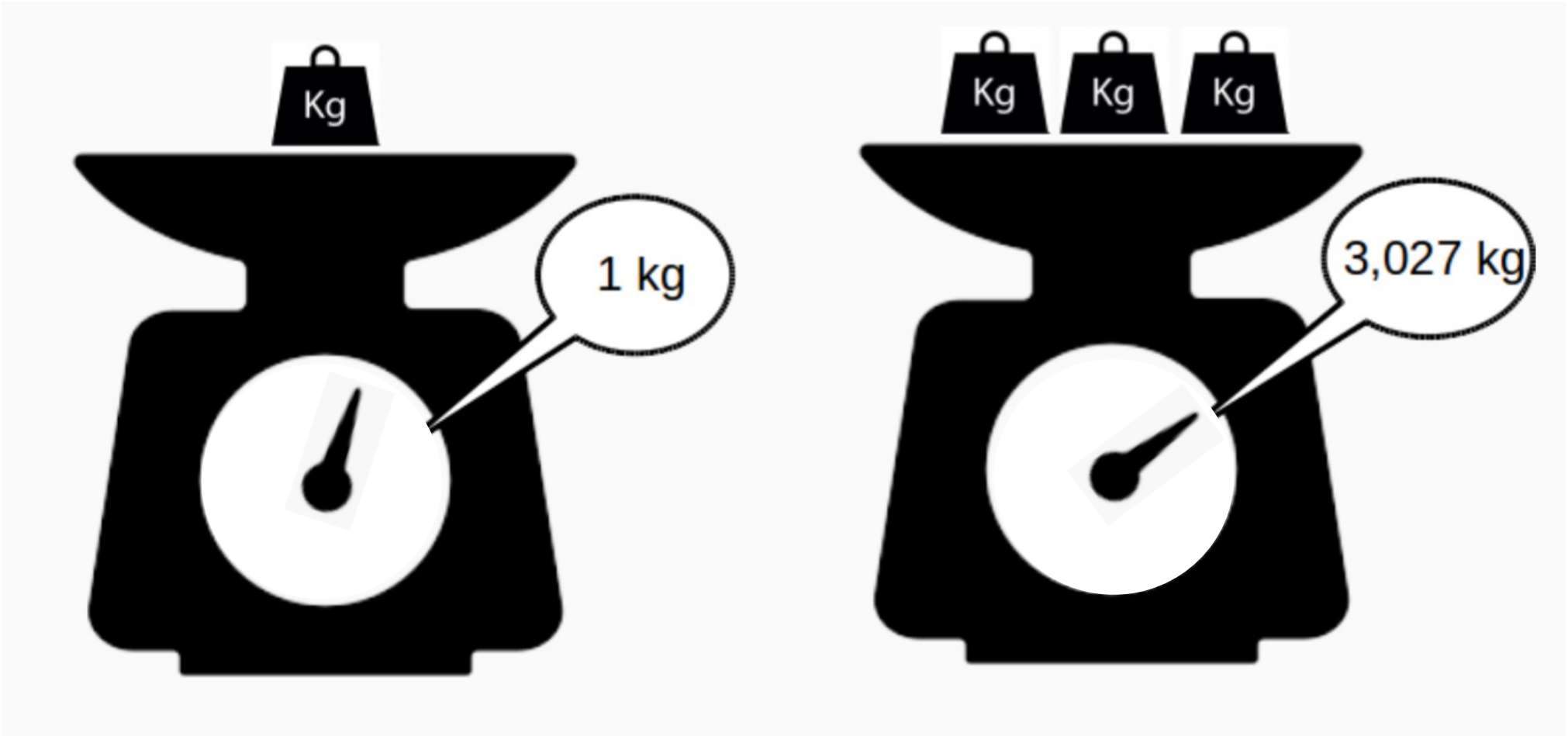


From *Nature* **560**, 41–48 (2018) |

➔~50% gain on LHC running

We’re doing science!

Experimental bias



Systematical effect

Example of impact of the angle on handwritten digits

Scientific Publication

Most complex measurement ever ?

Phys.Rev.Lett. 114 (2015)191803

Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments

(ATLAS Collaboration)[†]
(CMS Collaboration)[‡]

~6.000 signatures…

(Received 25 March 2015; published 14 May 2015)

A measurement of the Higgs boson mass is presented based on the combined data samples of the ATLAS and CMS experiments at the CERN LHC in the $H \to \gamma\gamma$ and $H \to ZZ \to 4\ell$ decay channels. The results are obtained from a simultaneous fit to the reconstructed invariant mass peaks in the two channels and for the two experiments. The measured masses from the individual channels and the two experiments are found to be consistent among themselves. The combined measured mass of the Higgs boson is $m_H = 125.09 \pm 0.21\,(\text{stat}) \pm 0.11\,(\text{syst})$ GeV.

Systematical uncertainties

>>>89-90: Where is this measurement documented in this note?
**That has now been finished and published, and references will be updated in the notes
JHEP 09 (2014) 079 arXiv:1407.5532

>>>116: Doesn't "3 GeV" contradict your previous statement that we cover a pT range "from 1 GeV..."
**No, those muons with pT~1 GeV and p~3 GeV are at high eta. But this is all irrelevant, as we only use muons with pT>4 GeV.

>>>116: I'm not sure what you mean when you say that they "require" the combined reconstruction. I would just say "Muons used in this analysis are reconstructed using a statistical combination of an MS track and ID track."
**Only the muons that have a successful statistical combination of an MS track and ID track are used in this analysis, the text is updated.

>>>128: what is fully efficient for signal candidates? It's not clear what the word "which" refers to.
**This related to the vertex quality criteria applied to the fit.

>>>142: Please mention the total number of di-muon candidates here.
**This is added to the text, 7.8M for 2011, and ~65M for 2012.

>>>171: could you provide more information about how the psi is assigned to a primary vertex? I'm surprised by your statement that "few" events contain multiple vertices; I thought pileup was a significant issue at 7 TeV.
** Not really. The only relevance the primary vertex has is to measure Lxy, which is measured in the transverse direction only, and hence is not changing much from one collision vertex to another. However, the determination of the primary vertex position depends on whether the two muon tracks were used in its fit or not, hence we need to know which vertex was it. But at 7 TeV there was not much ambiguity, the vertex which jpsi came from was almost always the main primary one. Studies from the 2011 Jpsi Phi analysis -- where vertex choice may have been an issued-- showed that there was no impact in the few cases of an incorrect choice of vertex.

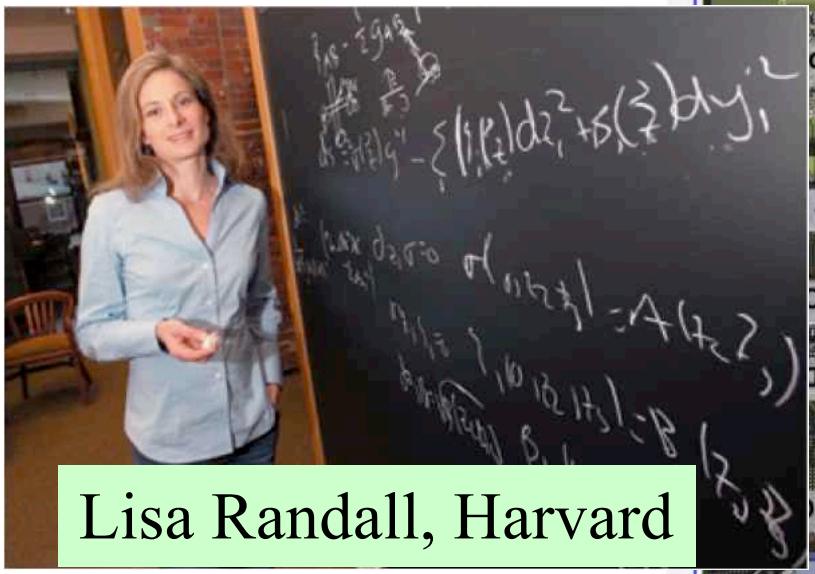
>>>Fig 1: Do you understand the eta dependence of this plot? What is the z-axis -- number of events per bin? The bins are extremely small -- what size bins did you choose?
** Yes. This is the scatter plot of dimuon candidates. The x-axis is the absolute rapidity "y" of the dimuon candidate, with the structure roughly reflecting the (smeared) geometry of the muon chambers, with dips near/around the cracks and edges.
The z-axis is just the candidate yield, the bin-width in |y| is 5e-3 and in pT is 320 MeV

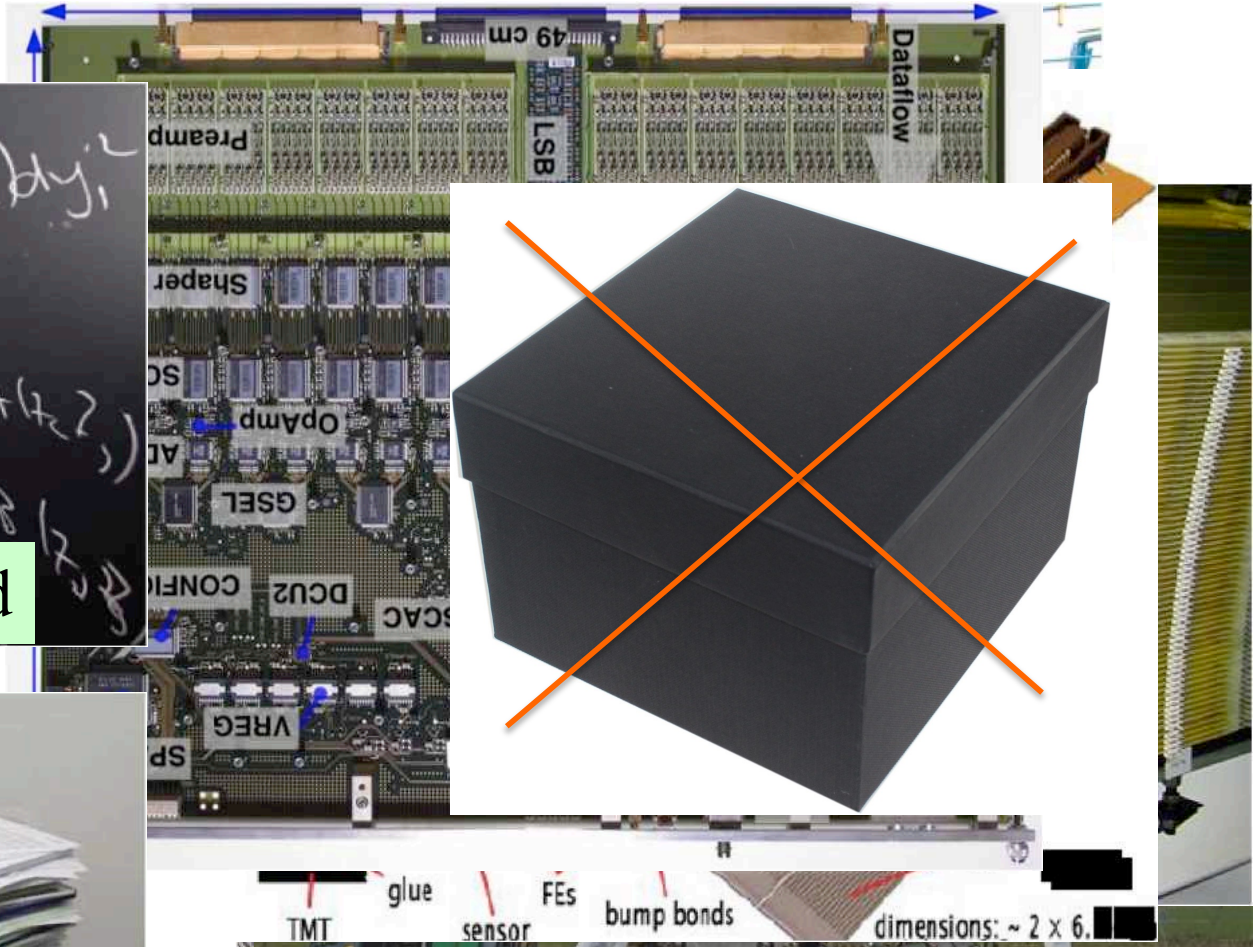
>>>228: The phi* definition is unclear to me, in particular what the "psi production" is. Do you mean the psi momentum vector?
**phi* is the angle between the psi production plane (defined by psi momentum and colliding proton

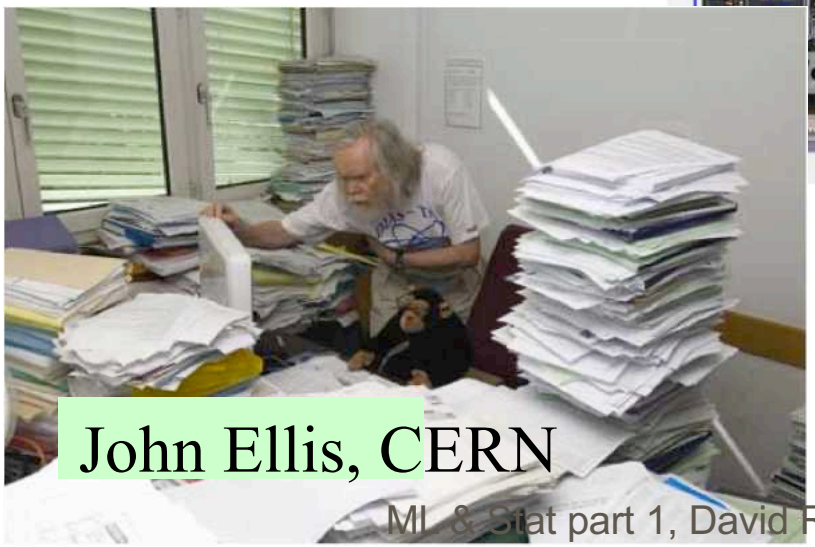
ue

More than 200 pages and weeks of Q&A before intenral approval!

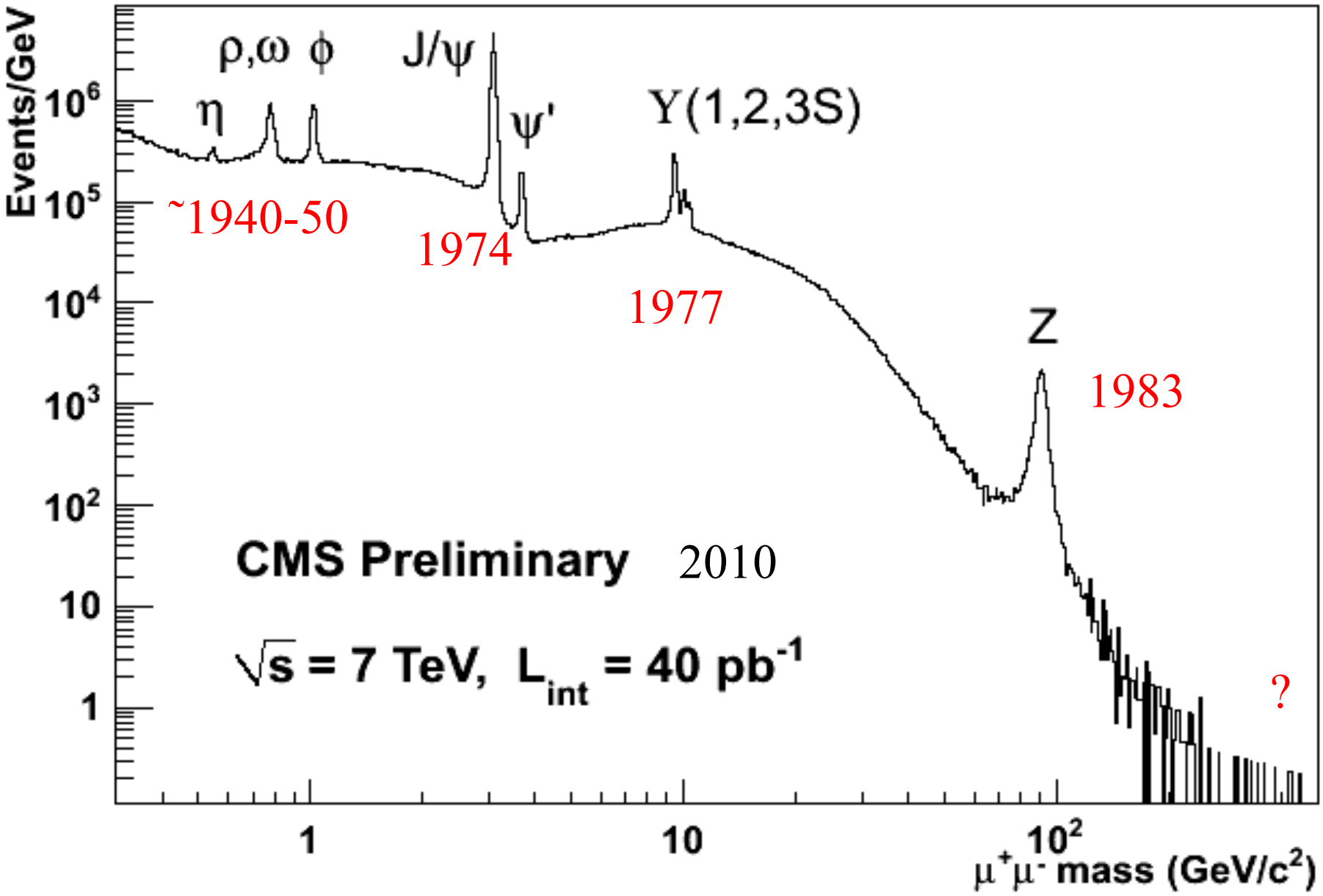
Trust but verify : Theory to experiment







Example of a cross-check Re-discovery of known particles



End of part 1