

Taller de Aprendizaje Automático

Proyecto 1 - Bosón de Higgs

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería



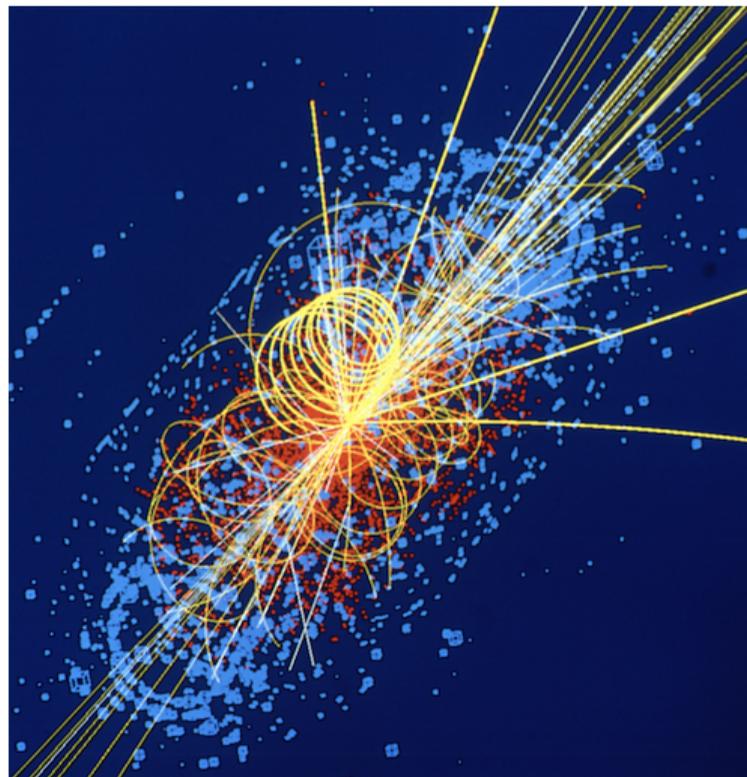
UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Montevideo, 2025

Introducción

Qué es el Bosón de Higgs ?

- Partícula de extrema importancia en la física de partículas.
- Propuesta por Higgs en 1964 para explicar por qué algunas partículas tienen masa.
- En 2012, la existencia de dicha partícula pudo ser confirmada experimentalmente en el CERN (Centro Europeo de Investigación Nuclear)

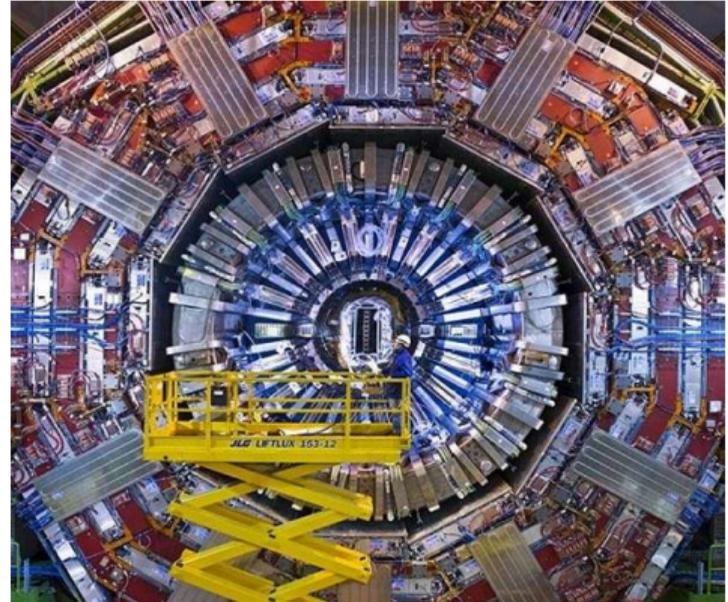


By CERN for the ATLAS and CMS Collaborations - <https://cds.cern.ch/record/1630222>, CC BY-SA 3.0
<https://commons.wikimedia.org/w/index.php?curid=29737816>

Introducción

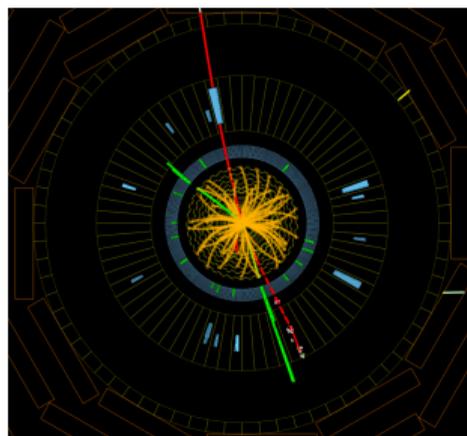
¿En qué consistió el experimento ?

- Dos partículas (protones) se aceleran en direcciones contrarias, hasta llegar casi a la velocidad de la luz.
 - Esta aceleración se realiza en el Large Hadron Collider: un anillo de 27 Km.
- Las dos partículas chocan casi a velocidad de la luz dentro de un detector, llamado ATLAS
 - ATLAS mide 46 metros de largo, 25 de alto y 25 de ancho, pesa 7000 toneladas y está ubicado en una caverna a 100 metros de profundidad.
- Las partículas se desintegran, produciendo decenas de nuevas partículas, que se disparan en todas direcciones



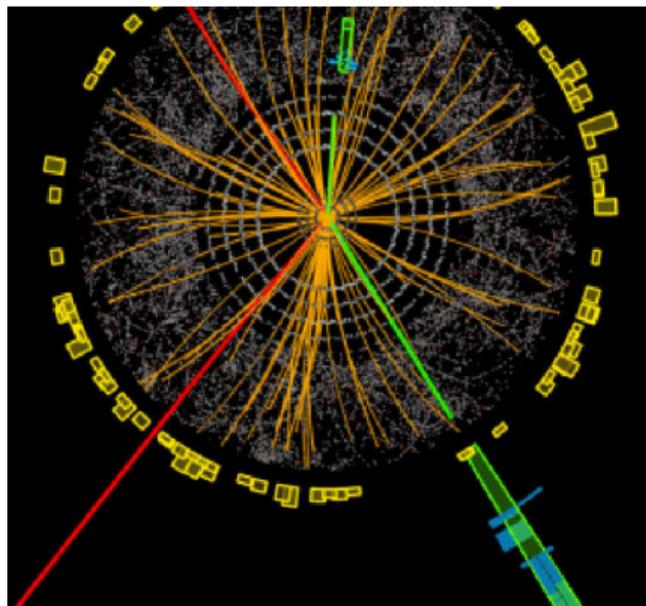
En que consistió el experimento ?

- Cada uno de estos "choques" se llaman "eventos". La cantidad de eventos que se generan es de aproximadamente 20 millones en un segundo. Cada evento contiene aproximadamente 10 partículas, que son "medidas" por el detector ATLAS en un conjunto de cientos de medidas.



Particularidades del experimento

- La mayoría de las partículas generadas por el choque NO son el Bosón de Higgs. Esto permite que se descarten la mayoría de los eventos, quedando unos 400 eventos por segundo para ser analizados.
- El Bosón de Higgs es altamente inestable: enseguida de crearse se descompone en otras partículas por lo cual su búsqueda es a partir de “rastros”.
- En el caso del experimento ATLAS, fue la partícula *tau* la que se observó y que indirectamente confirma la presencia del Bosón de Higgs.



Proyecto 1 - Bosón de Higgs

Objetivo del proyecto

El objetivo del proyecto es diseñar e implementar un clasificador que permita detectar los eventos en que aparece la partícula *tau* (llamados *signals*) de aquellos en que no aparecen (llamados *background*).

Particularidades del experimento

- Kaggle challenge: <https://www.kaggle.com/c/higgs-boson>
- Datos de training (training.csv): 250000 registros
- Datos de test (test.csv): 550000 registros
- Random submission: el formato para armar el envío
- HiggsBosonCompetition_AMSMetric.py: la función para calcular la métrica de evaluación

The screenshot shows the top section of the Kaggle competition page. At the top left, it says 'Featured Prediction Competition'. The main title is 'Higgs Boson Machine Learning Challenge' with a prize of '\$13,000'. Below the title, it says 'Use the ATLAS experiment to identify the Higgs boson'. There are '1,784 teams' and it was created '6 years ago'. A navigation bar includes 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', 'Rules', and 'Team'. On the right, there are 'My Submissions' and a 'Late Submission' button.

Data Description

File descriptions

- training.csv - Training set of 250000 events, with an ID column, 30 feature columns, a weight column and a label column.
- test.csv - Test set of 550000 events with an ID column and 30 feature columns.
- random_submission - Sample submission file in the correct format. File format is described on the [Evaluation](#) page.
- HiggsBosonCompetition_AMSMetric - Python script to calculate the competition evaluation metric.

For detailed information on the semantics of the features, labels, and weights, see the [technical documentation](#) from the [LAL website](#) on the task.

Some details to get started:

- all variables are floating point, except PRL_jet_num which is integer
- variables prefixed with PRI (for PRimitives) are "raw" quantities about the bunch collision as measured by the detector.
- variables prefixed with DER (for DERived) are quantities computed from the primitive features, which were selected by the physicists of ATLAS
- it can happen that for some entries some variables are meaningless or cannot be computed; in this case, their value is -999.0, which is outside the normal range of all variables

Detalles de los datos

- Cada evento está representado por un vector de 30 características.
- Algunas de estas características son obtenidas mediante mediciones directas en los sensores de ATLAS, mientras que otras características son derivadas a partir de las primeras.
- Todas las variables son en general en punto flotante.
- Todos los ángulos son en radianes entre $[-\pi, \pi]$
- El valor -999.0 indica una variable que no pudo ser medida o calculada, y es un valor por fuera del rango de todas las variables.

Detalles de los datos

- Cada evento está representado por un vector de 30 características
- Las variables que comienzan con "DER" son medidas derivadas
- Las variables que comienzan con "PRI" son medias directas

	EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet
0	100000	138.470	51.655	97.827	27.980	0.91
1	100001	160.937	68.768	103.235	48.146	-999.00
2	100002	-999.000	162.172	125.953	35.635	-999.00
3	100003	143.905	81.417	80.943	0.414	-999.00
4	100004	175.864	16.915	134.805	16.405	-999.00

ling_eta	PRI_jet_leading_phi	PRI_jet_subleading_pt	PRI_jet_subleading_eta	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label
	0.444	46.062	1.24	-2.475	113.497	0.002653	s
	1.158	-999.000	-999.00	-999.000	46.226	2.233584	b
	-2.028	-999.000	-999.00	-999.000	44.251	2.347389	b
	-999.000	-999.000	-999.00	-999.000	-0.000	5.446378	b
...	-999.000	-999.000	-999.00	-999.000	0.000	6.245333	b

Detalles de los datos

Las variables "EventId", "Weight" y "Label" no deben ser usados como datos de entrada para el entrenamiento o clasificación.

	EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet		
0	100000	138.470	51.655	97.827	27.980	0.91		
1	100001	160.937	68.768	103.235	48.146	-999.00		
2	100002	-999.000	162.172	125.953	35.635	-999.00		
3	100003	143.905	81.417	80.943	0.414	-999.00		
4	100004	175.864	16.915	134.805	16.405	-999.00		
ling_eta	PRI_jet_leading_phi	PRI_jet_subleading_pt	PRI_jet_subleading_eta	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label	
	0.444	46.062	1.24	-2.475	113.497	0.002653	s	
	1.158	-999.000	-999.00	-999.000	46.226	2.233584	b	
	-2.028	-999.000	-999.00	-999.000	44.251	2.347389	b	
	-999.000	-999.000	-999.00	-999.000	-0.000	5.446378	b	
...	-999.000	-999.000	-999.00	-999.000	0.000	6.245333	b	

Detalles de los datos

- training.csv - Training set of 250000 events, with an ID column, 30 feature columns, a weight column and a label column.
- test.csv - Test set of 550000 events with an ID column and 30 feature columns.

Links de interés

- Kaggle challenge: <https://www.kaggle.com/c/higgs-boson>
- Detalle del challenge:
https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf