scikit learn

Claire Savard

CU Physics Department

22 November 2019

Overview

- Open source machine learning package for python
- BSD license
- Extension of SciPy
- Features various classification, regression, and clustering algorithms
- Largely written in python, parts in Cython (mix of python and C)
- Large, international community

History

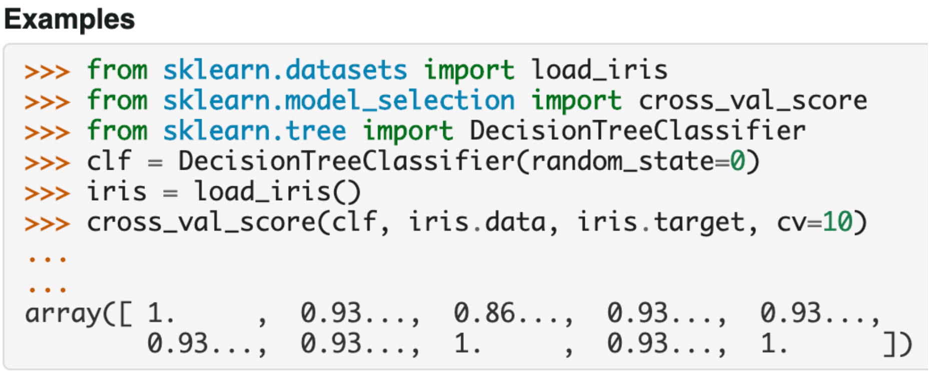
- Initially developed in 2007 by David Cournapeau for google summer of code
- Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort and Vincent Michel (🇫🇷) rewrote and published first release in 2010
- Since, funding from INRIA, google, telecom Paris, more
    - Finances managed by numFOCUS
- New releases every ~3 months

Community

- 19 authors (core contributors)
- >1500 contributors in lifetime
- ~25 issues per week
- Organized contributing rules stated on website
- Meritocracy, can move up in the food chain
- Most changes to code need 2+ developers to approve

Contribution

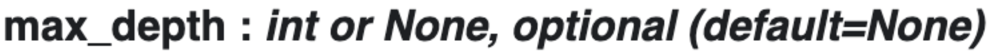
- Added examples to documentation of classes



Examples

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.model_selection import cross_val_score
>>> from sklearn.tree import DecisionTreeClassifier
>>> clf = DecisionTreeClassifier(random_state=0)
>>> iris = load_iris()
>>> cross_val_score(clf, iris.data, iris.target, cv=10)
...
...
array([ 1.     ,  0.93...,  0.86...,  0.93...,  0.93...,
        0.93...,  0.93...,  1.     ,  0.93...,  1.      ])
```

- Checked consistency of default values in doc with code and adding default values when they are not there

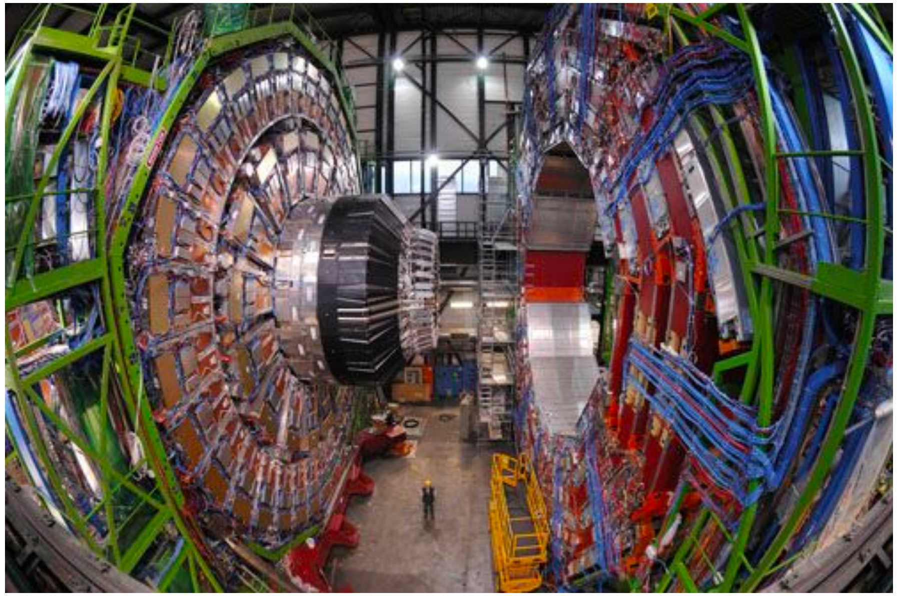


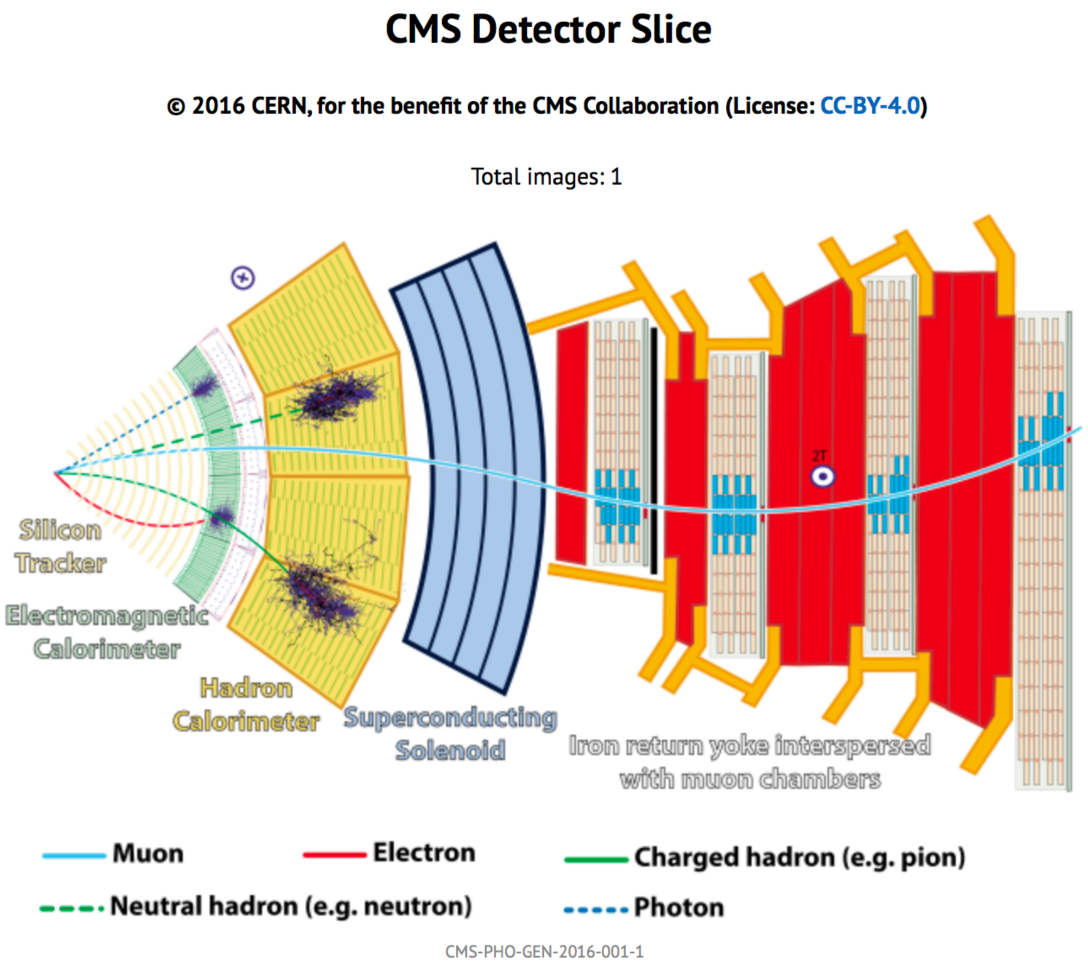
max_depth : ***int or None, optional (default=None)***

How I use scikit-learn

LHC physics

CMS detector





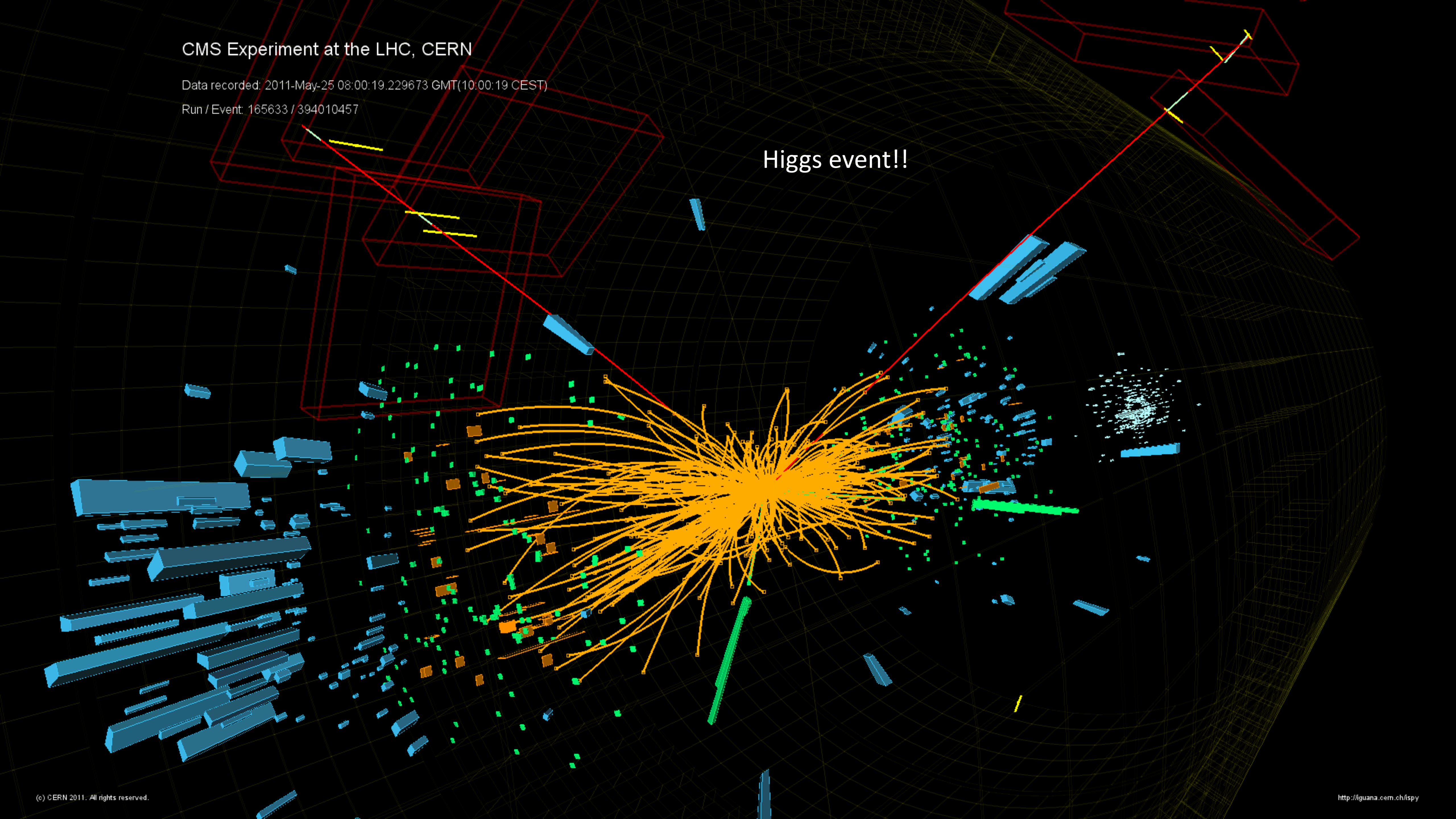
CMS Detector Slice

© 2016 CERN, for the benefit of the CMS Collaboration (License: CC-BY-4.0)

Total images: 1

Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with muon chambers
2T
Muon
Electron
Charged hadron (e.g. pion)
Neutral hadron (e.g. neutron)
Photon

CMS-PHO-GEN-2016-001-1


CMS Experiment at the LHC, CERN
Data recorded: 2011-May-25 08:00:19.229673 GMT(10:00:19 CEST)
Run / Event: 165633 / 394010457
(c) CERN 2011. All rights reserved.
http://iguana.cern.ch/ispy

Higgs event!!

Event selection

- Hundreds of terabytes of data produced by the collisions
- Need to throw out >99% of events deemed uninteresting
- Event selection: algorithms for selecting the few events that make the cut
    - These events then stored and fully reconstructed for physics analysis
- Interesting physics happens extremely rarely, important to select well
- How can machine learning improve event selection?

Fake vs. real particle tracks

- Real tracks = reconstructed particle track originating from an actual particle
- Fake tracks = reconstructed tracks resulting from error in the reconstruction process
- Use scikit-learn algorithms (and keras) to create a classifier for real and fake tracks
- Thus far, accuracy increased by ~10%!

Backup slides

Scikit-learn and HLS4ML

- HLS4ML = tool for translating python machine learning algorithms to C code interpretable by HLS
- Currently working on synthesizing sklearn and keras algos to FPGA-readable code with 3 goals:
  - Maintain high accuracy
  - Minimal resource usage
  - Total run time of a couple microsec