

Published on *CERN openlab* (<http://test-static-05.web.cern.ch>)

[Home](#) > Random Decision Forests on Apache Spark

Random Decision Forests on Apache Spark ^[1]

Date:

Tuesday, 12 April, 2016 - 15:00 to 16:00

Location:

[IT Amphitheater](#) ^[2]

Apache Spark continues to gain momentum as the new processing paradigm for Apache Hadoop, and for the data scientist, it has a lot to like: natively distributed, REPL, Python APIs in addition to native Scala, and a library of machine learning algorithms, MLlib.

Spark includes an implementation of random decision forests, an important and popular ensemble classifier/regressor algorithm. This talk will introduce Spark and random decision forests to the curious, and demonstrate the process of analyzing a real-world data set with them. The session will cover loading data and understanding the data set, and introduce ideas like training and test set evaluation, ensemble methods, feature types, and supporting concepts like impurity and entropy.

Indico or other event webpage:

[For more information about the event](#) ^[3]

- [Visit Us](#)
- [RSS Feeds](#)

DISCLAIMER: This Web page contains pointers to material related to the management of CERN openlab in the Information Technology Department at the European Organization for Nuclear Research (CERN). Their use and distribution are regulated by the [CERN copyright notice](#).



Source URL: <http://test-static-05.web.cern.ch/news/random-decision-forests-apache-spark>

Links

[1] <http://test-static-05.web.cern.ch/news/random-decision-forests-apache-spark>

[2] <https://maps.cern.ch/mapsearch/https://maps.cern.ch/mapsearch/mapsearch.htm?n=%5B%2731/3-004%27%5D>

[3] <https://indico.cern.ch/event/516771/>