FACT DATA AND FORMATS

Kai A. Brügge
April 6, 2016

TU Dortmund, Physik E5b, Astroteilchenphysik

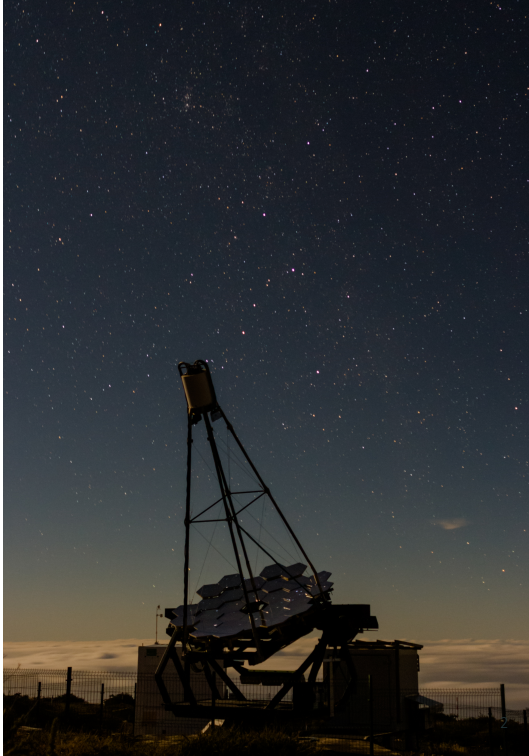
lehrstuhl physik e5

technische universität dortmund

The FACT Telescope

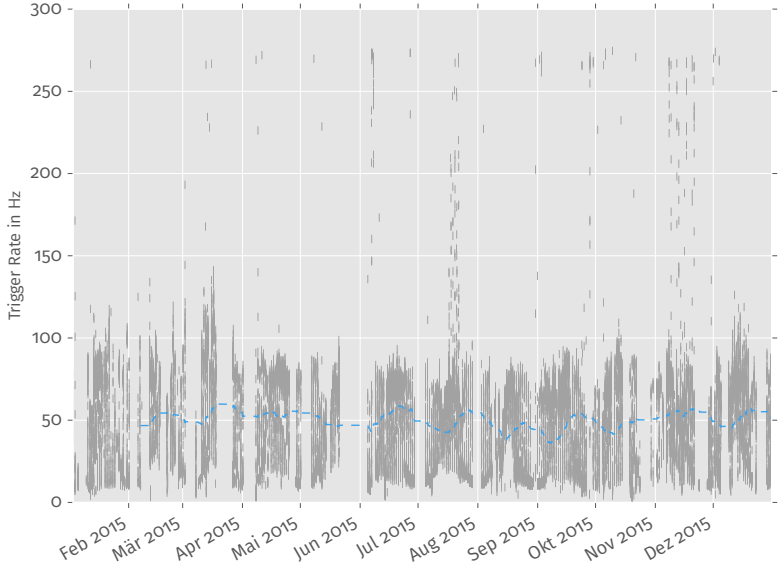
Small 4 m telescope on La Palma next to MAGIC.

FACT is a testbed for SiPMs and monitors bright gamma-ray sources for flares.



RAW DATA

Mean FACT trigger rate at 50 Hz.



One FACT event consists of 1440 pixel $\times$ 300 samples $\times$ 2 bytes $+$ overhead $\approx$ 900 kB.

File size for a typical FACT data run between 4 GB and 6 GB.

Data is stored as .fits or compressed as .zfits .fz files.

http://arxiv.org/pdf/1506.06045.pdf

RAW DATA FORMAT

Data is stored run wise in 5 min intervals.

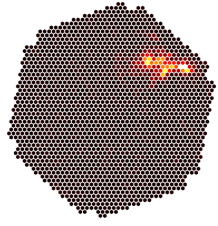
There is a database contianing weather and other meta information per run.

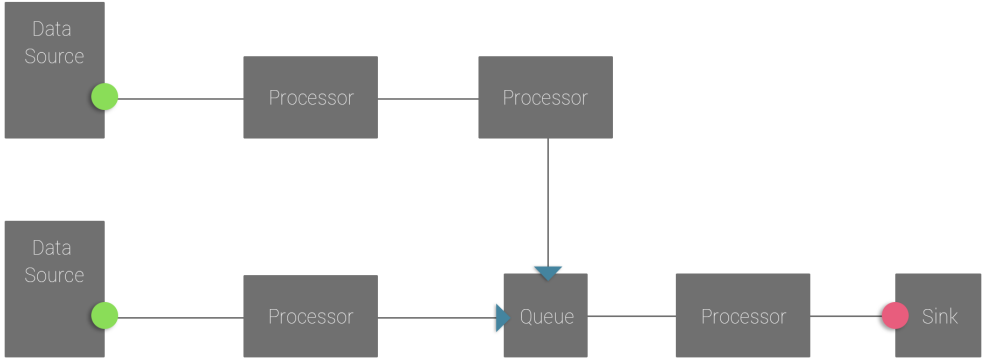
User select data by a single logical join operation.

FROM RAW DATA TO IMAGE PARAMETER

Typical steps performed when analyzing IACT data.

Preprocessing	Calibrate data and remove artifacts.
Extraction	Estimate number of photons and their arrival time.
Cleaning	Select the shower pixels.
Image Parameter	Calculate parameters for Signal / Background separation.
Classification	Use ML methods to suppress background.
Energy Estimation	Use regression method on the parameters to estimate the energy.




Data Source
Processor
Processor
Data Source
Processor
Queue
Processor
Sink

The usual FACT analysis produes between 40 and 60 parameters

Event timestamp	Pointing	Width	Length	more stuff ...
2015-01-24 00:34:12	23.4 84,7	40	60	...
2015-01-24 00:34:13	23.4 84,7	100	120	...
2015-01-24 00:34:15	23.4 84,7	10	12	...
2015-01-24 00:34:20	23.4 84,8	30	41	...
2015-01-24 00:34:21	23.4 84,8	15	64	...
2015-01-24 00:34:25	23.5 84,8	18	21	...

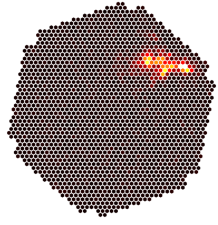
We have no common format on how to store these tables.

Many people (including myself) use the HDF5 file format.

One HDF5 file containing 90 hours of data has 6.8 GB.

SUPERVISED MACHINE LEARNING MODELS

Preprocessing	Calibrate data and remove artifacts.
Extraction	Estimate number of photons and their arrival time.
Cleaning	Select the shower pixels.
Image Parameter	Calculate parameters for Signal / Background separation.
Classification	Use ML methods to suppress background.
Energy Estimation	Use regression method on the parameters to estimate the energy.

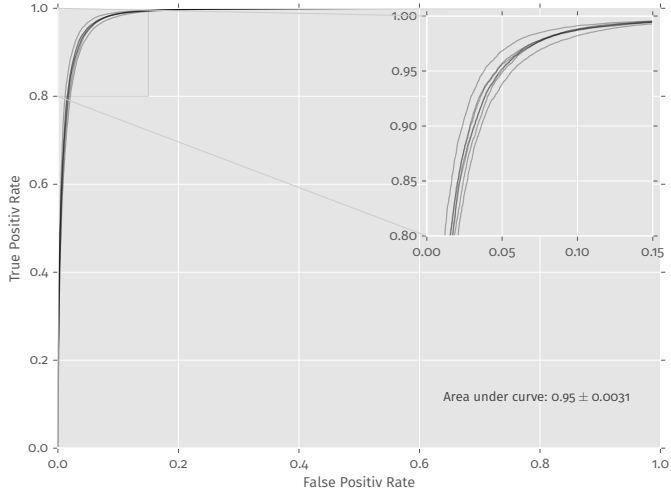


Supervised machine learning methods are used for event classification and energy estimation.

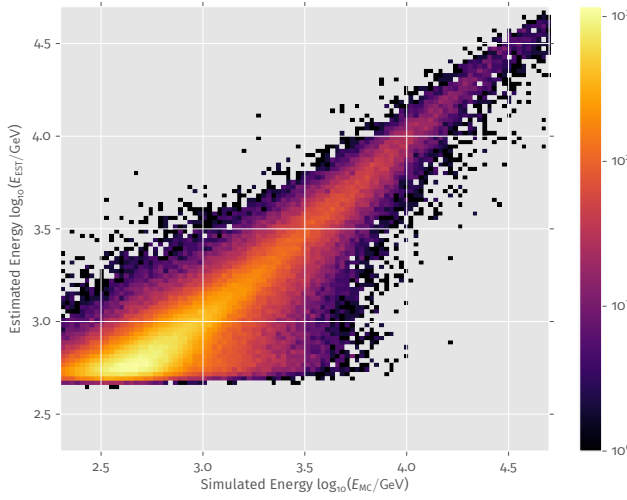
For both tasks ensembles of tree based learners have proven to be effective.

1. Use Python and scikit-learn for training models from simulated data.
2. Export models to PMML format.
3. Load PMML file into the `streams`-framework and apply per event.

Classifier performance measurements



Estimated Energy vs True Energy



Models stored as PMML have a size between 10s of MB and 1 - 2 GB

Stronlgy depending on the use case and complexity of the model.

Event timestamp	Pointing	Width	Length	P(Gamma)	Estimated Energy
2015-01-24 00:34:12	23.4 84,7	40	60	0.9	200
2015-01-24 00:34:13	23.4 84,7	100	120	0.6	250
2015-01-24 00:34:15	23.4 84,7	10	12	0.99	1200
2015-01-24 00:34:20	23.4 84,8	30	41	0.96	10
2015-01-24 00:34:21	23.4 84,8	15	64	0.5	5200
2015-01-24 00:34:25	23.5 84,8	18	21	0.23	2200

SOME THOUGHTS ON DATA STORAGE

FACT data summarized

- So far high level FACT Data is small enough to fit into memory.
- Our data is not publicly available (yet). Only a single user at a time.
- No complicated or nested querying operations on the data.
- The file format doesn't matter.

However FACT data becomes larger each night.

For sharing large data sets between users, files are not an option.

$\longrightarrow$ We need to use a database.

Any database technology works as long as the frontend code is backend agnostic.

Structure and schema of the data might change

$\longrightarrow$ Use a non schematic database. At least for the prototype.

Idea is to publish data in passes (similar to FERMI)

A list of Buzzwords for a VO?

- Apache Hive

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig
- Stinger

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig
- Stinger
- Apache Drill

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig
- Stinger
- Apache Drill
- Spark SQL

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig
- Stinger
- Apache Drill
- Spark SQL
- IBM BIGSQL

A list of Buzzwords for a VO?

- Apache Hive
- Apache Pig
- Stinger
- Apache Drill
- Spark SQL
- IBM BIGSQL
- Oracle BIGSQL

Heres another idea: