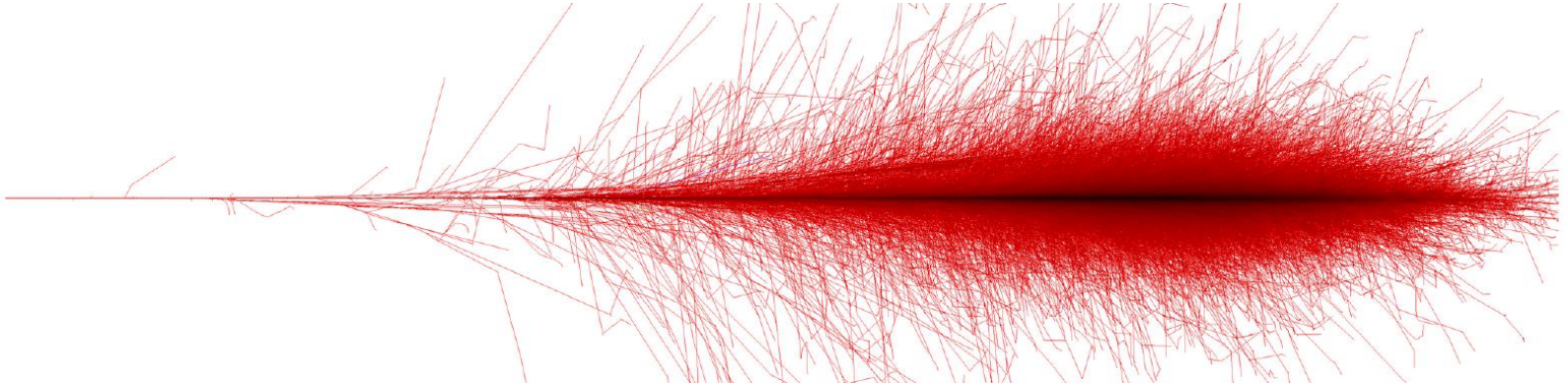
Karlsruhe Institute of Technology

Output Format

Tim Huege (KIT & VUB)



What we have

- Mixture of YAML and Parquet
  - Some stuff that could have gone into YAML is in Parquet (e.g. longitudinal profile)
- Corresponding Python library
- Some stuff not integrated (geneaology goes into .npy)

Downsides of Parquet

- Apache Arrow as a dependency?
- Tabular data does not cover all use cases
- Not well-known in our community
- Problems with random access within huge files (memory mapping)?
- Ok for a first release, but we might want to look for something better

Contenders

- ROOT
  - Powerful file format
  - Well-known in our community
  - But hesitant about having the framework as dependency
    - Development workflow can probably ensure that it does not creep into the framework

Contenders

- HDF5
  - Widely used in many areas of physics
  - Maintenance ensured? (consortium responsible)
  - According to Lukas maybe order of magnitude slower than ROOT, but might depend on test cases and how exactly set up

Contenders

- Compressed ASCII (e.g. inside a .tar.gz) with transparend access
  - If restricted to reasonable number of digits might be not much bigger than binary format
  - But probably access to filtered data not possible, also not processing huge files that cannot be loaded into memory at once

IO library

- We are not really using our own IO library at the moment
- We should do so, though, as it provides a layer of abstraction which will allow us to change the underlying file format without user-code having to change
- In C7 everybody wrote their own code because there was not a common solution
- We should avoid this in C8, by making sure we provide a well-tested, convenient IO library along with examples how to use it for the very first release