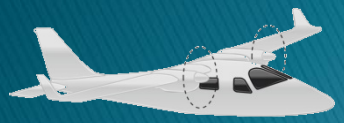
Neue Herausforderungen in der Prozessierung und Verwaltung von Massendaten

– wie bekomme ich Zugang zu TopoBathymetrischen Datensätzen?

Dipl.-Ing. Marcel Ritter, Dr. Werner Benger

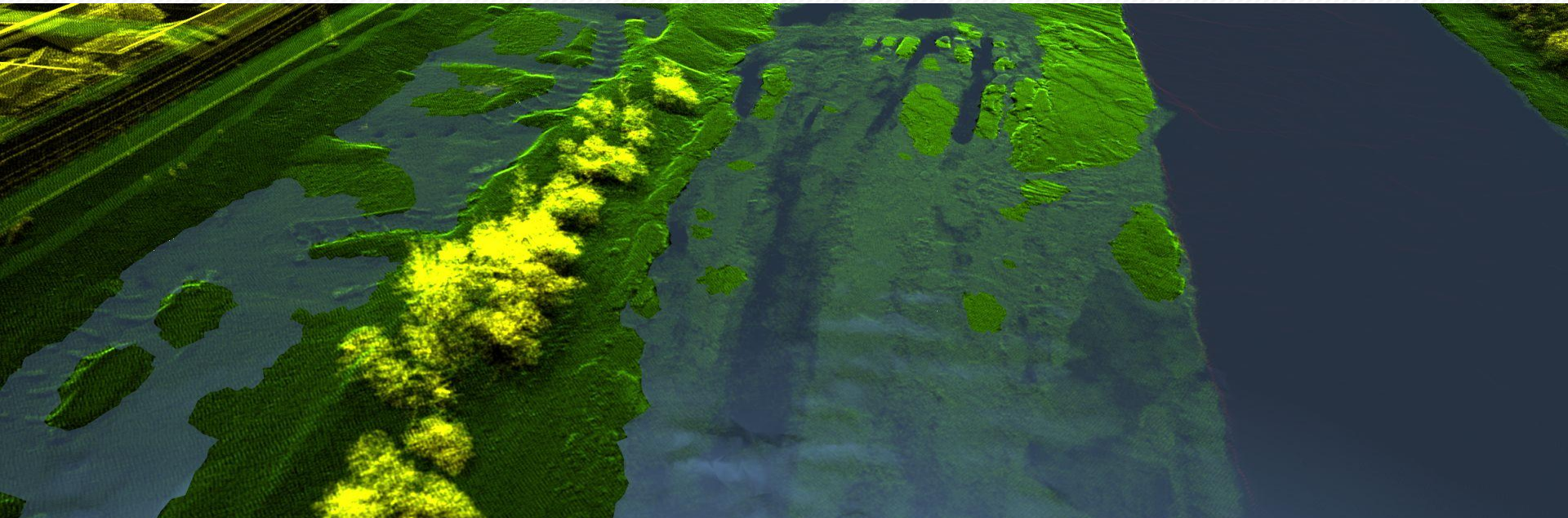
AHM

1. Workshop zum Comet K – Projekt „Alpine Airborne Hydromapping“
23.2.2015



Overview

- Data
  - Challenges/Motivation
  - HDF5
    - Remote Access
  - F5
    - Example



Challenges

data visualization

what you see is what you understand

data investigation

non-limited data investigation and analysis

data management

handling the mass of data

data processing

topo-bathymetric lidar datasets

data storage

long lasting data accessibility

filtering & classification

new methods and algorithms

data conversion

convert/processing of items but keeping all data

fluent workflow

from raw data to end-user

Challenges – TopoBathy vs ALS

- much **larger** lidar datasets and higher point densities
- you need **foreign mass datasets** for processing (RGB values, echosounder data, gauges)
- data needs to be **manipulated by input information** to get the correct data
  - (not everything is coming from the sensor)
- data might need to be manipulated strip wise, **day wise, weather wise, …**
- **different water surface models** might be needed for refraction and runtime correction (need of calculated data)
- water dependent influences within water column
- different algorithms and correction approaches needed for topo- and bathymetric data **within one dataset**
- ……

Motivation

Higher amount of data processing – Easier to make mistakes during processing

- **Visualization** for processing and quality control
- Keeping all processing steps and data to **retrace processing**
- Data conversion and export possibilities **without loosing information**
- **Merged datasets** of different data sources

-> **F**lexible **F**ile **F**ormat **F**or the **F**uture (**F5**)

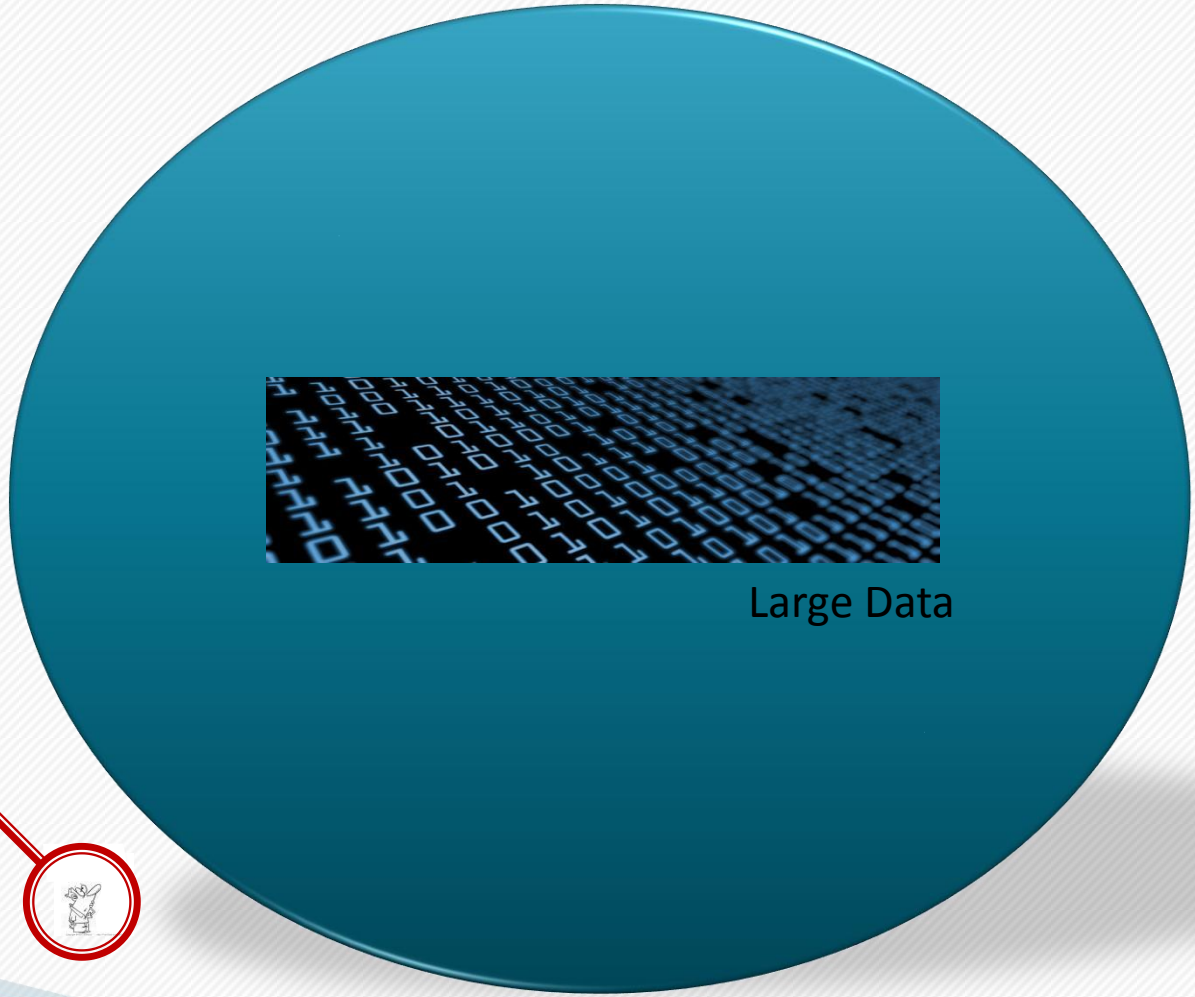
Data

- Before doing data processing and visualization one has to deal with data
  - Big data sets
  - Many different kinds
  - Many formats

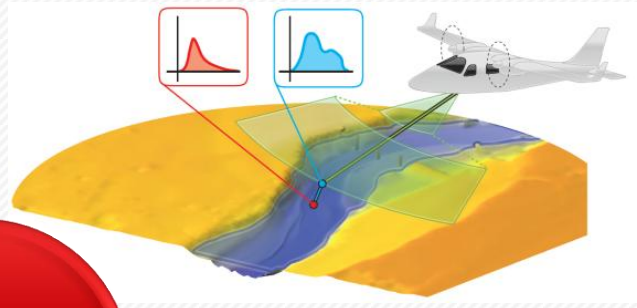
- Data management and handling is important

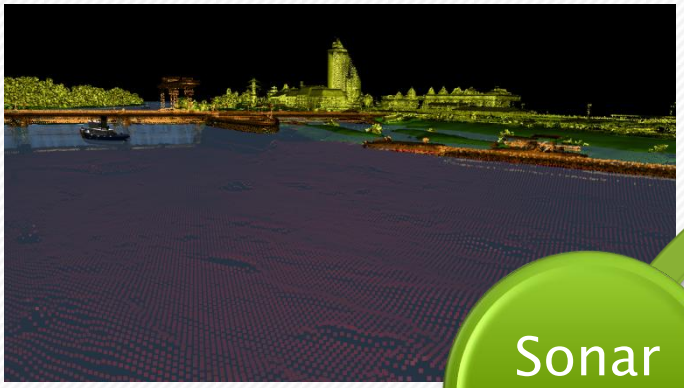
Large Data

→Surveying produces big data sets (>100 GB to TBs)



Different Kinds





LIDAR

Sonar Data

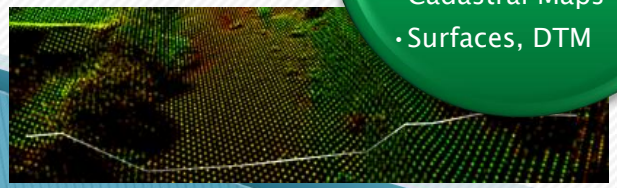
Diverse Data

Satellite Images

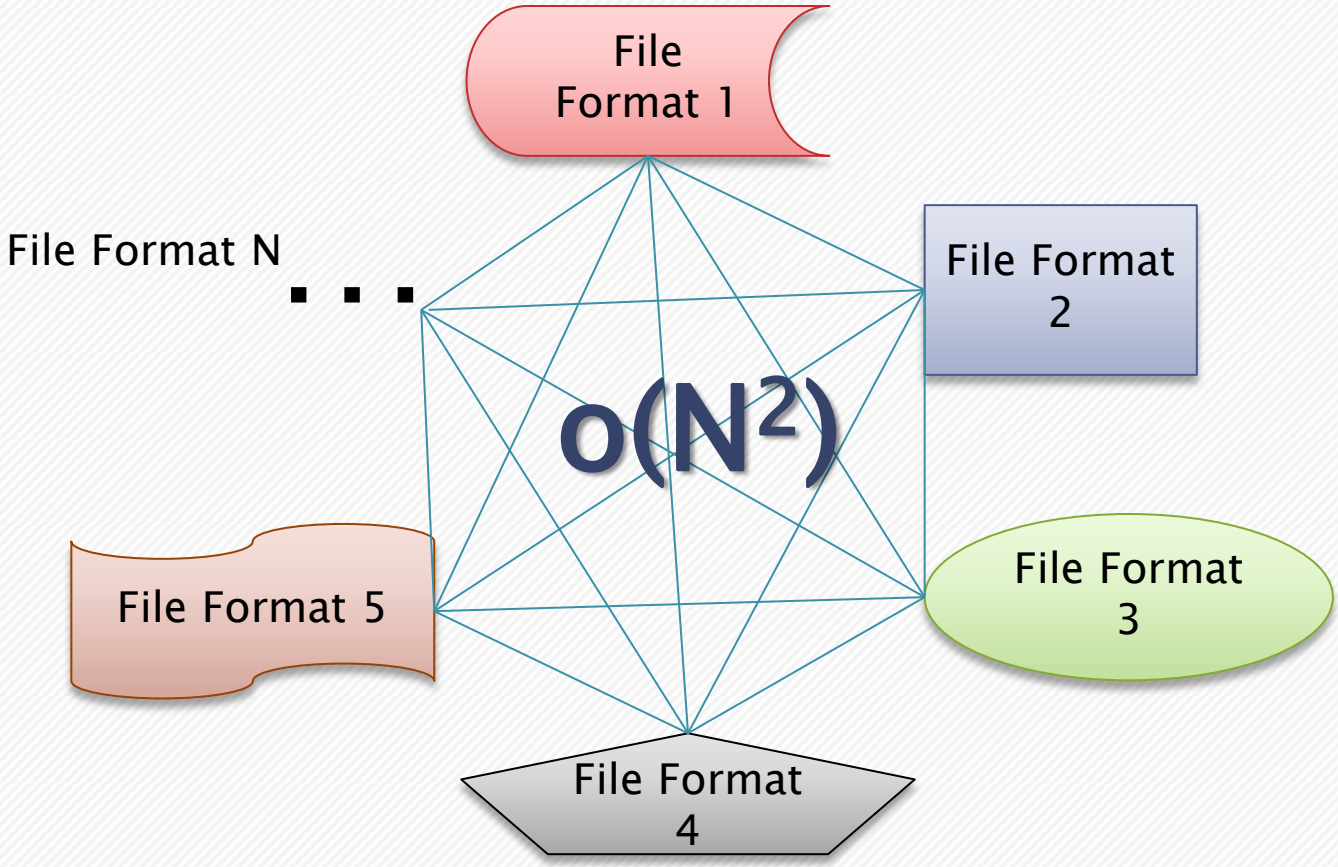
Standard Surveys

- Cross Sections
- Cadastral Maps
- Surfaces, DTM

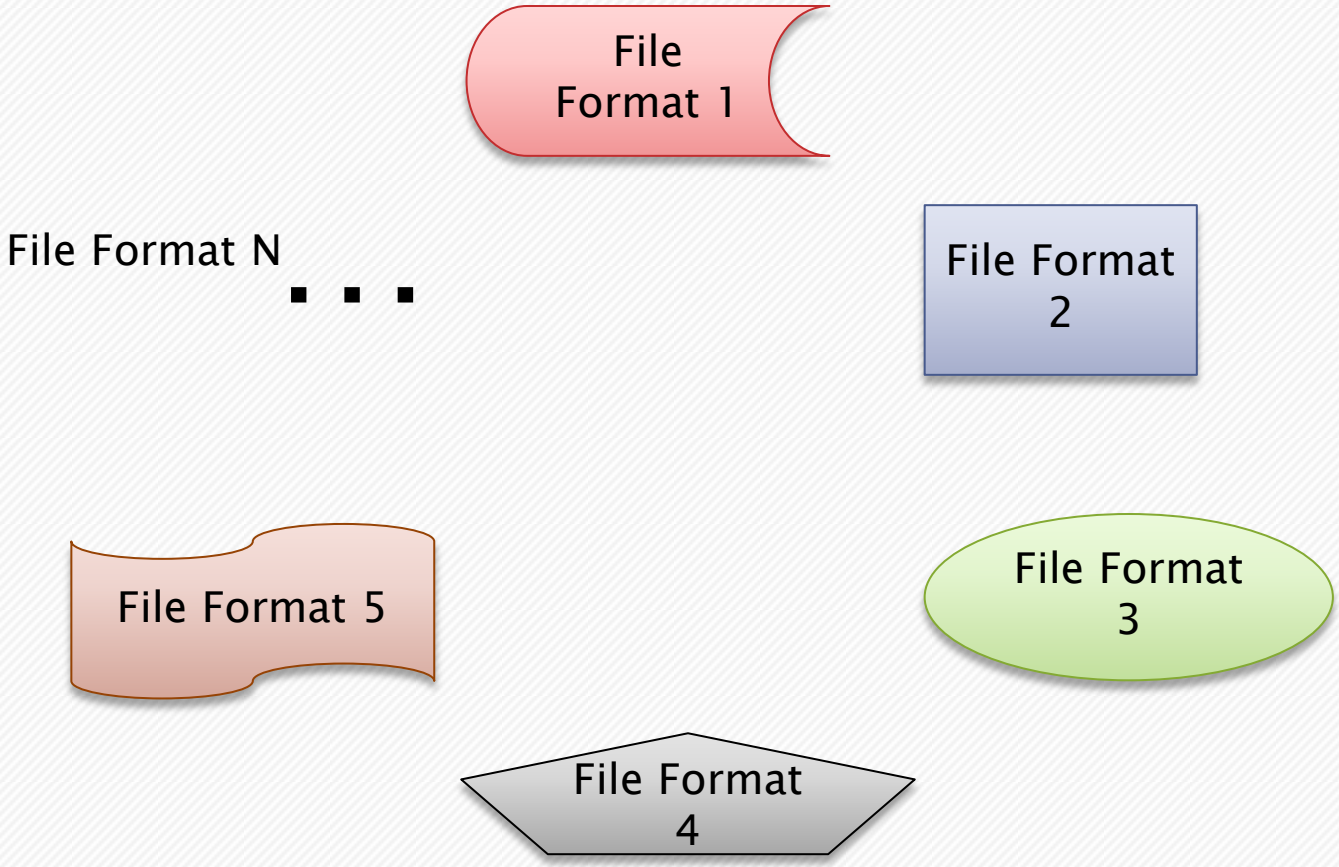
Arial Images



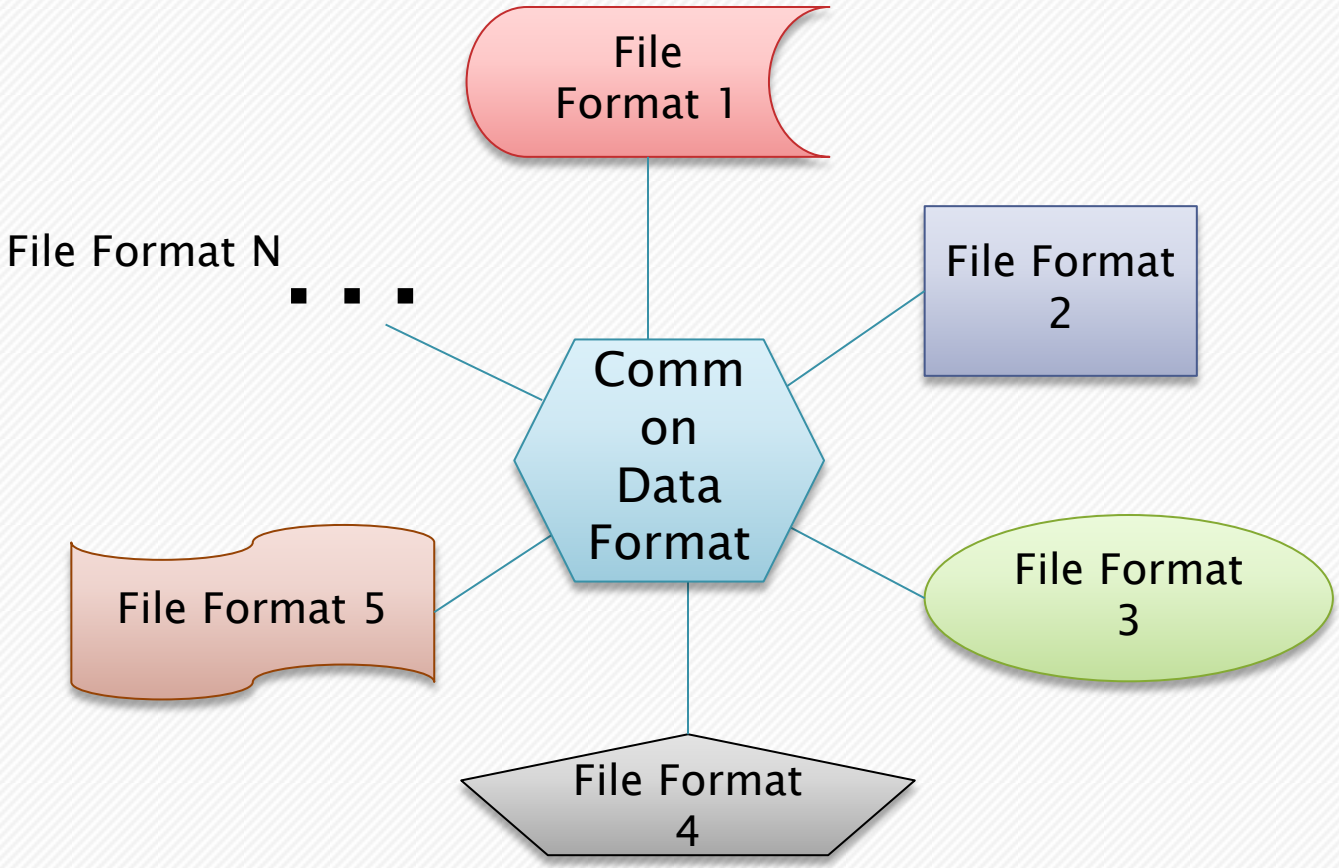
Many File Formats



Many File Formats



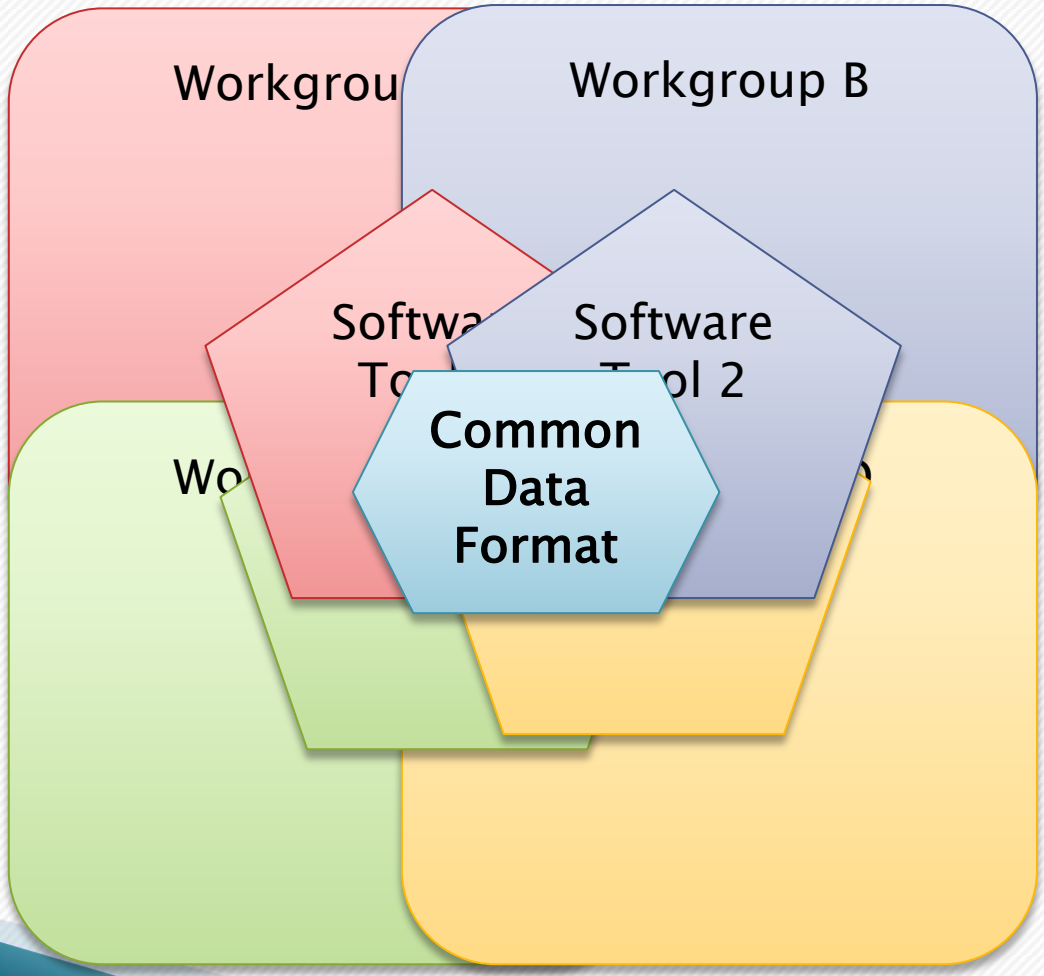
Many File Formats



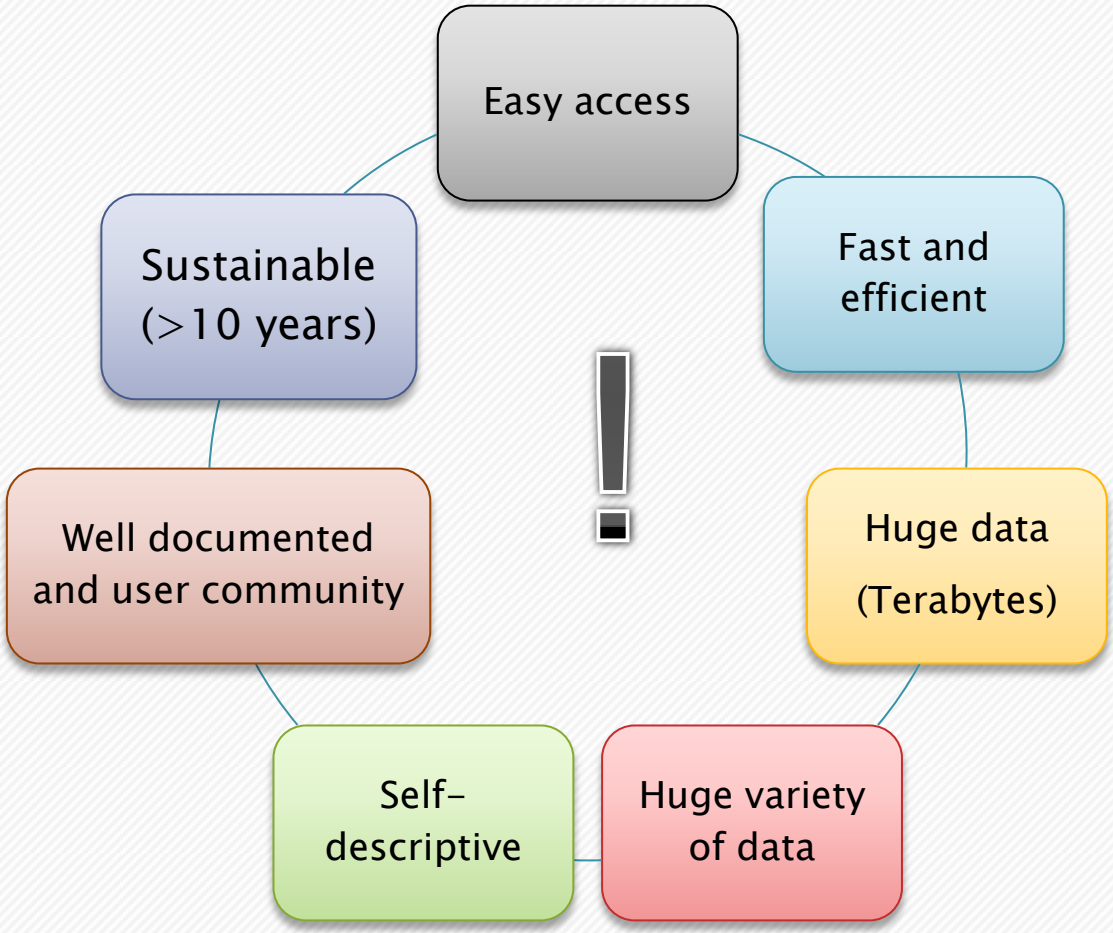
Less Implementation Effort o(N)

Many File Formats

Easier collaboration
More time for science



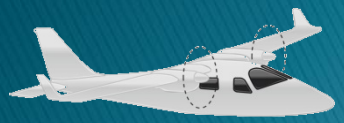
Requirements on a Data Format



HDF5

Hierarchical Data Format 5

http://www.hdfgroup.org/HDF5



HDF5 – A few Analogies

File system

(in a file)

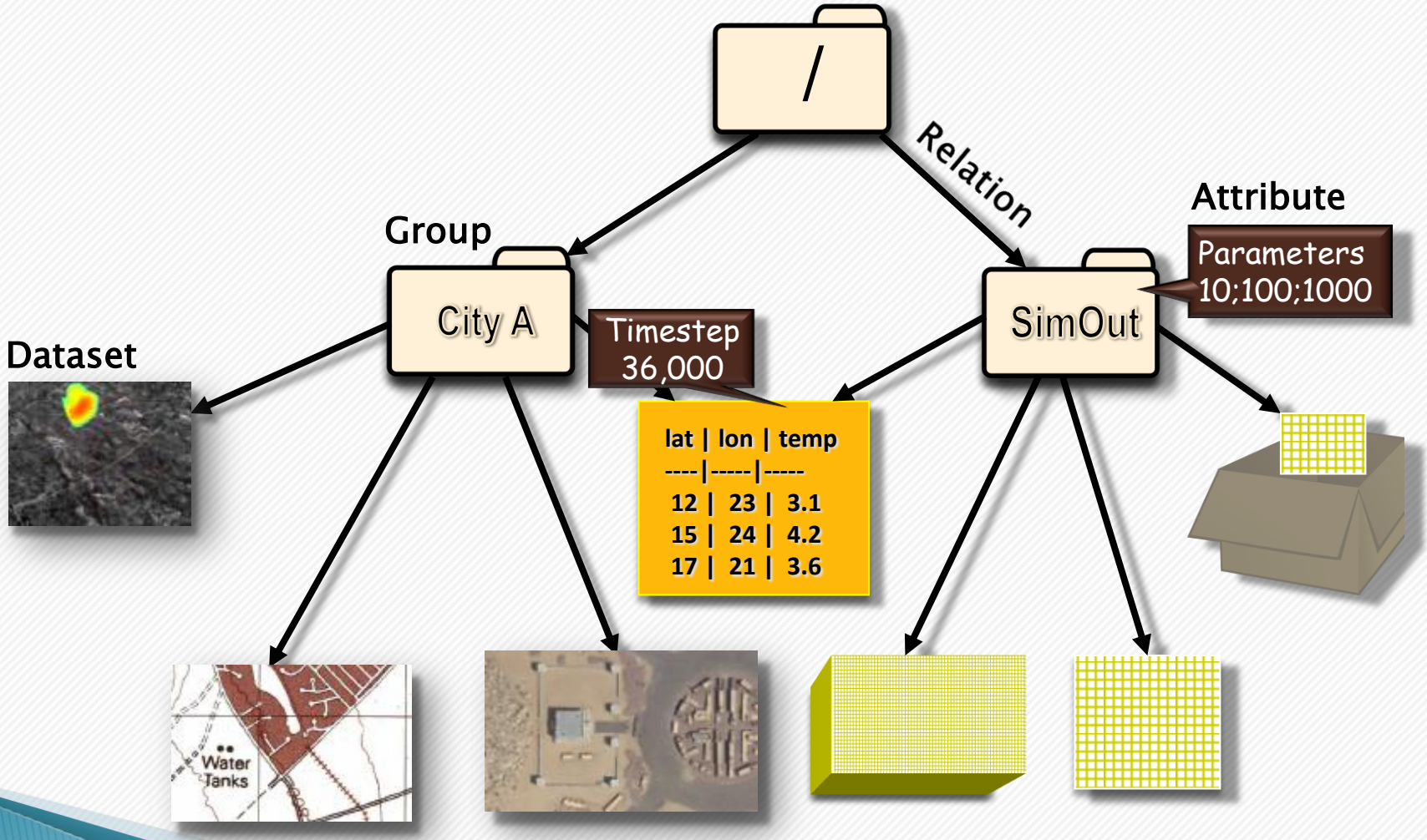
Binary **XML** file

PDF for numerical data

Database

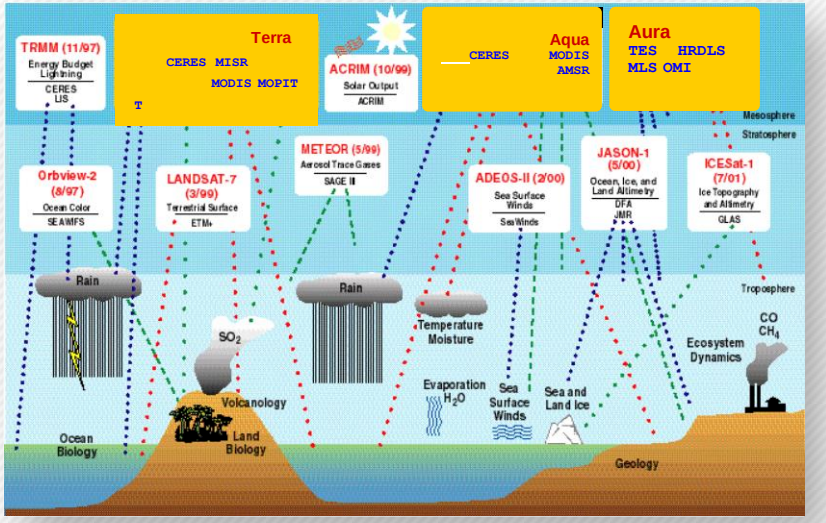
(container for array variables)

HDF5 – Relationships

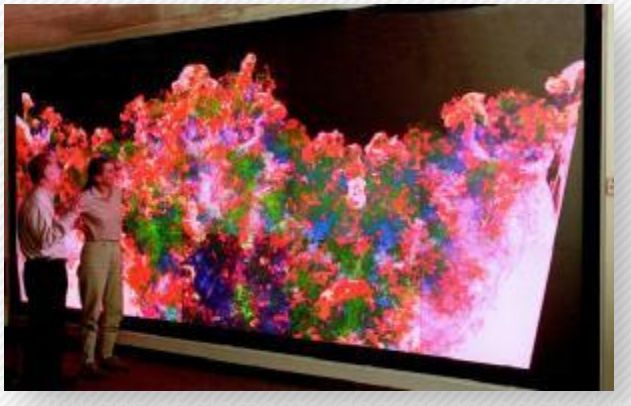


HDF5 – Applications

Earth Science (Earth Observing System)



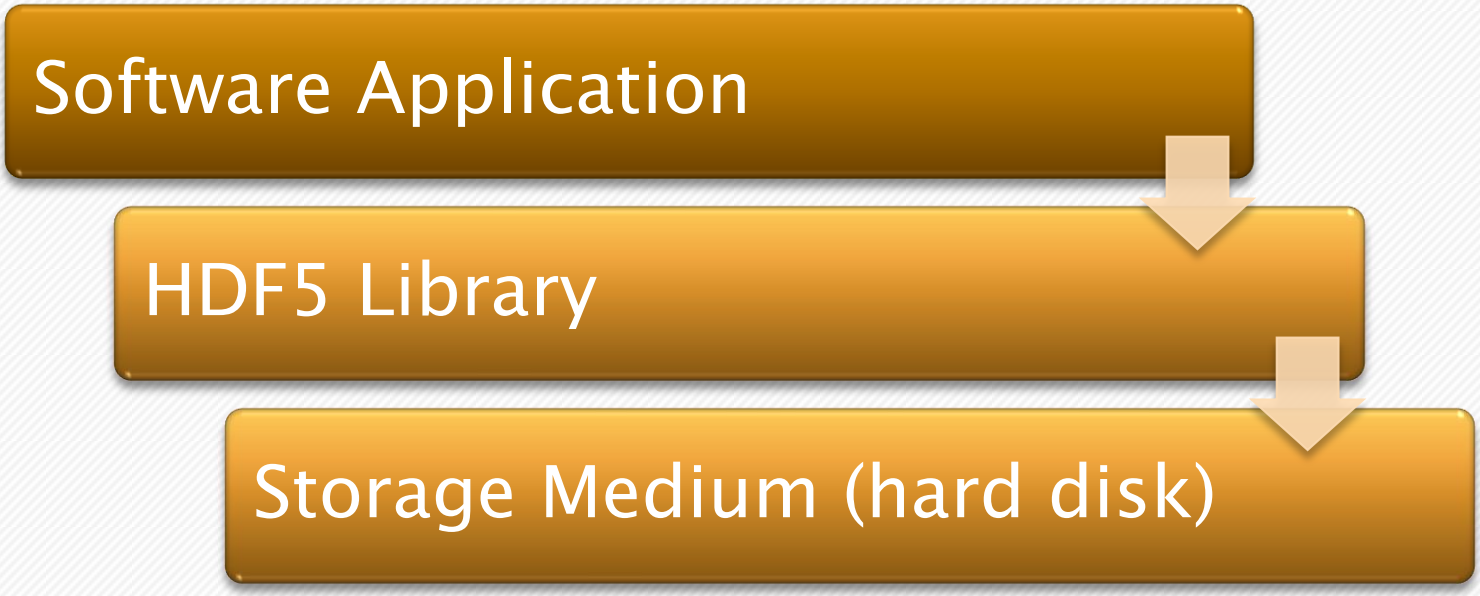
Big simulations



Movie Making

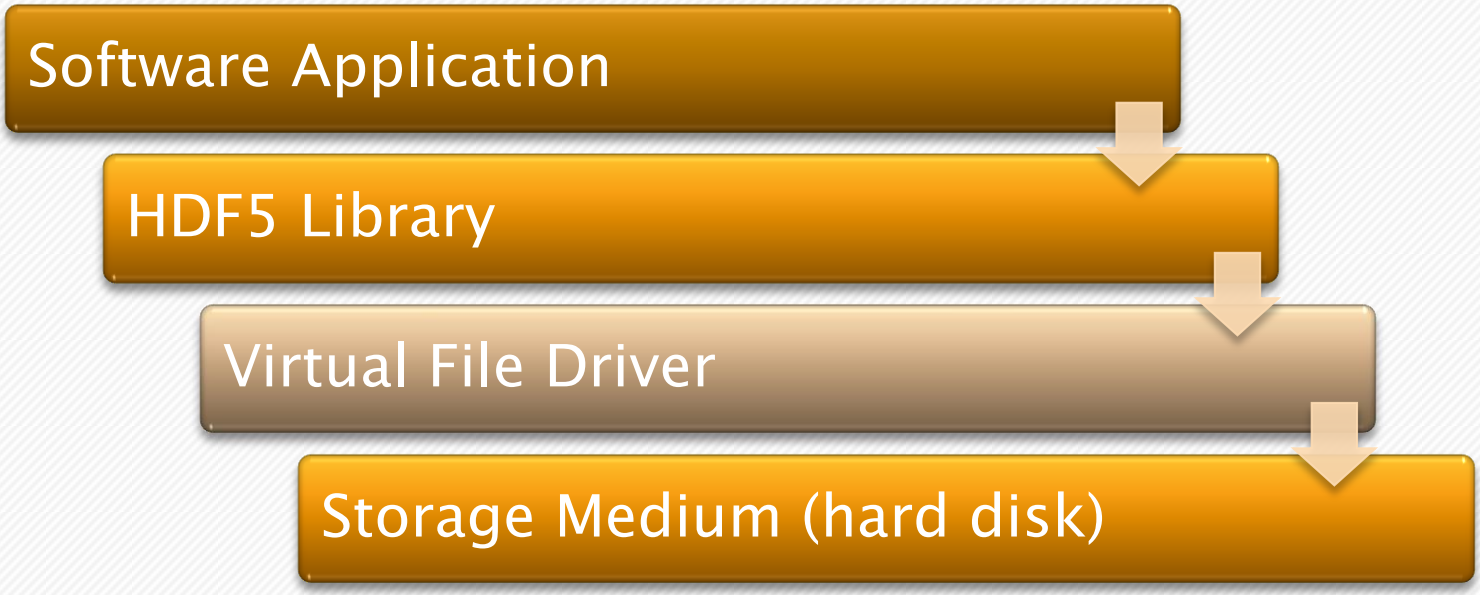
HDF5 – Software Library

- HDF5 is a software library to handle the HDF5 file format

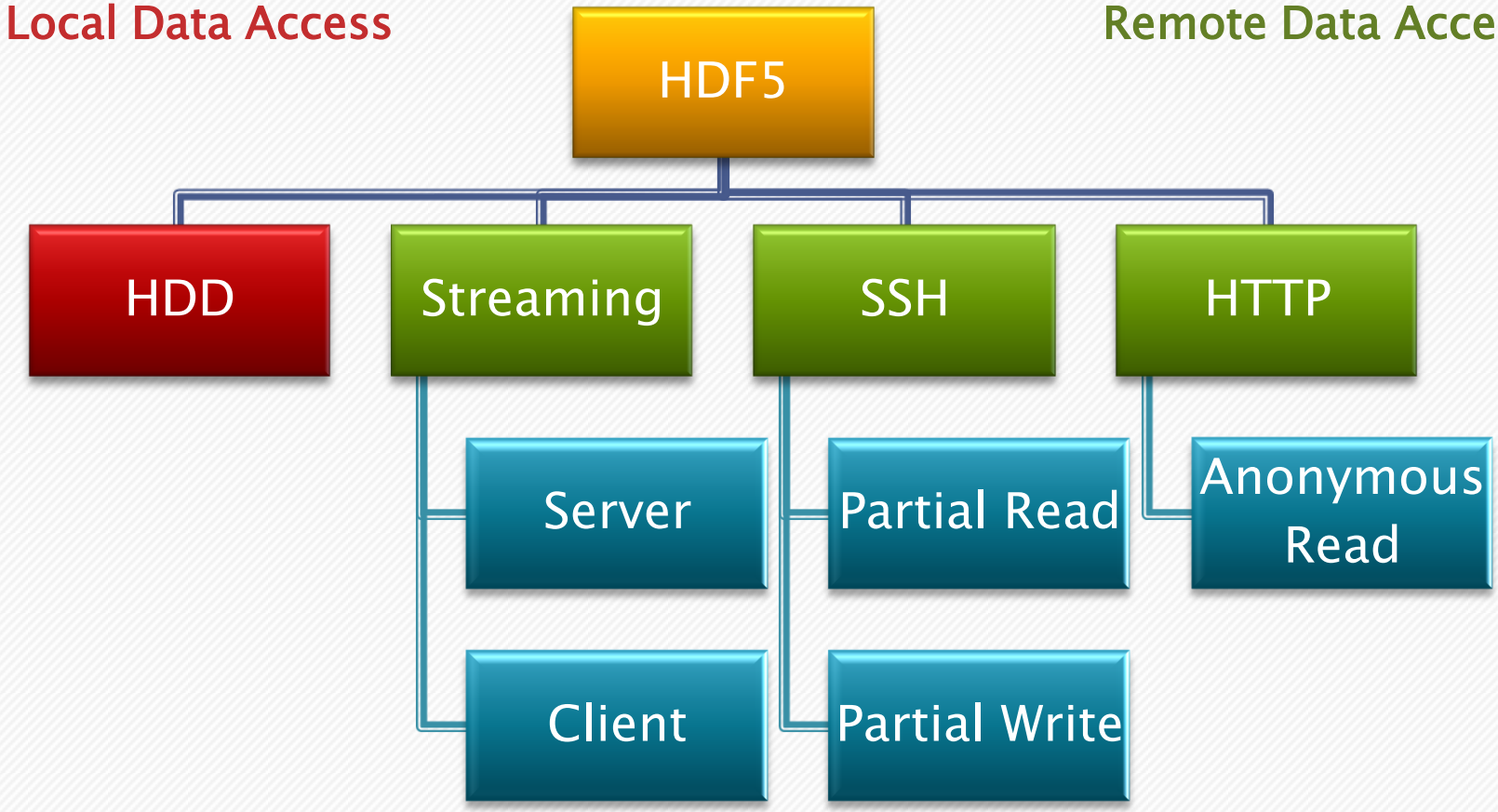


HDF5 – Virtual File Drivers

- HDF5 VFD allows to replace the storage device access routines

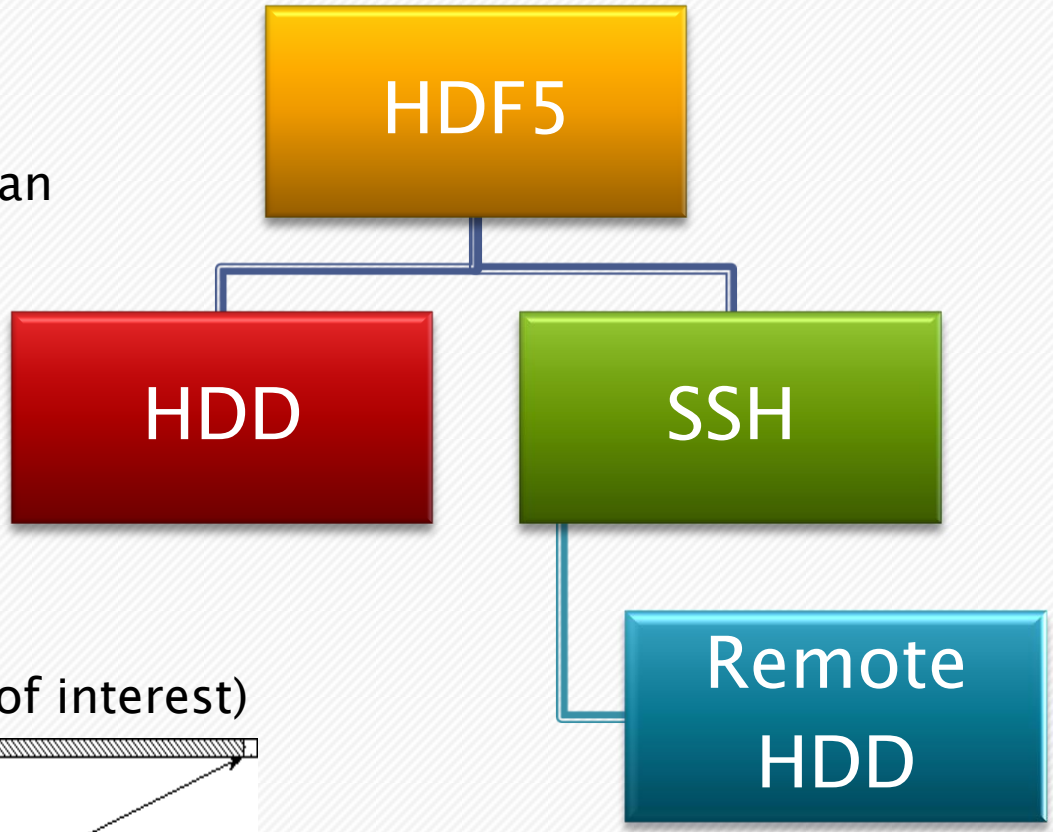


HDF5 VFD Overview

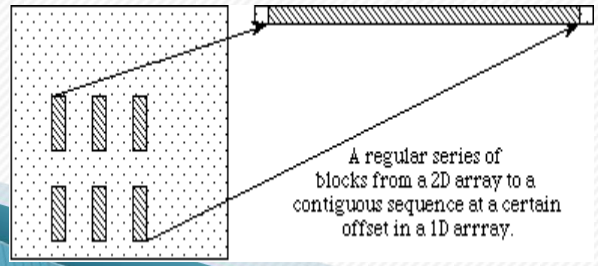


HDF5 SSH VFD (work in progress)

- Standard SSH server
  - No special server required, just place HDF5 file where it can be accessed via scp
- Allow *partial* file acess
  - Does **NOT** copy the entire file, e.g.:



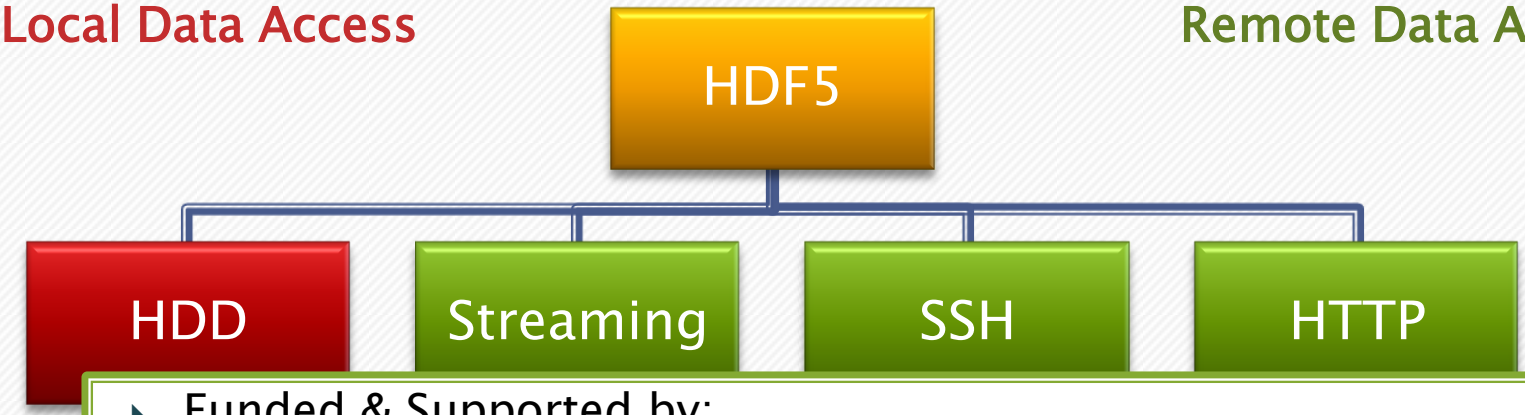
Local (region of interest)



HDF5 VFD Overview

Local Data Access

Remote Data Access



- Funded & Supported by:

- US National Science Foundation Award #1251137
- BIGDATA: Small: DCM: Collaborative Research: An efficient, versatile, scalable, and portable storage system for scientific data containers

- Stony Brook University, New York
- Brandeis University, New York
- Louisiana State University, Baton Rouge

HDF5 – What Users Get

- A multi-platform **library** and **tools** built on over 10 years experience in large data handling from the high performance computing community (HPC).

- A capability that:
  - Lets them organize large and/or complex collections of data
  - Gives them efficient and scalable data storage and access
  - Lets them integrate a wide variety of types of data and data sources

  - Guarantees long-term data integrity and preservation
  - Supports remote file access (work in progress)

HDF5

More than a container such as ZIP or TAR

Also allows to describe the structure of the contents of a file

Provide remote data access via the virtual layer.

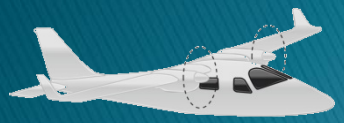
HDF5

How to store different *kinds* of data sets *consistently* in HDF5?

F5

Fiber Bundle Data Model

http://www.fiberbundle.net



AHM

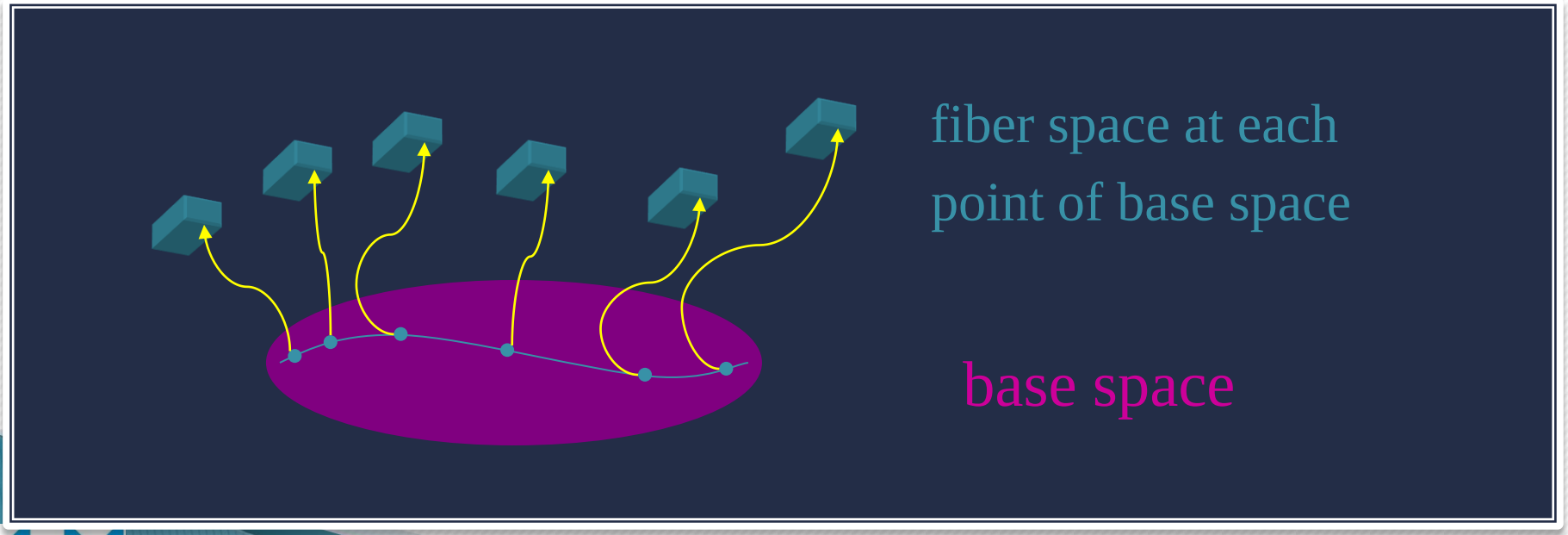
Initial Approach

The proper abstractions for scientific data are known. We just have to use them.

A visualization model based on the mathematics of fiber bundles, David M. Butler and M. H. Pendley, Computers in Physics, Sep/Oct 1989, p 45-51, 3(5)

In Words

- The total space E can be written locally as a product space B×F of the base space B with some space F.



F5

Designed for Scientific Data

C Library

Provides High Level Functions

- Allows to write complicated data by a few function calls

Originates from application in numerical relativity

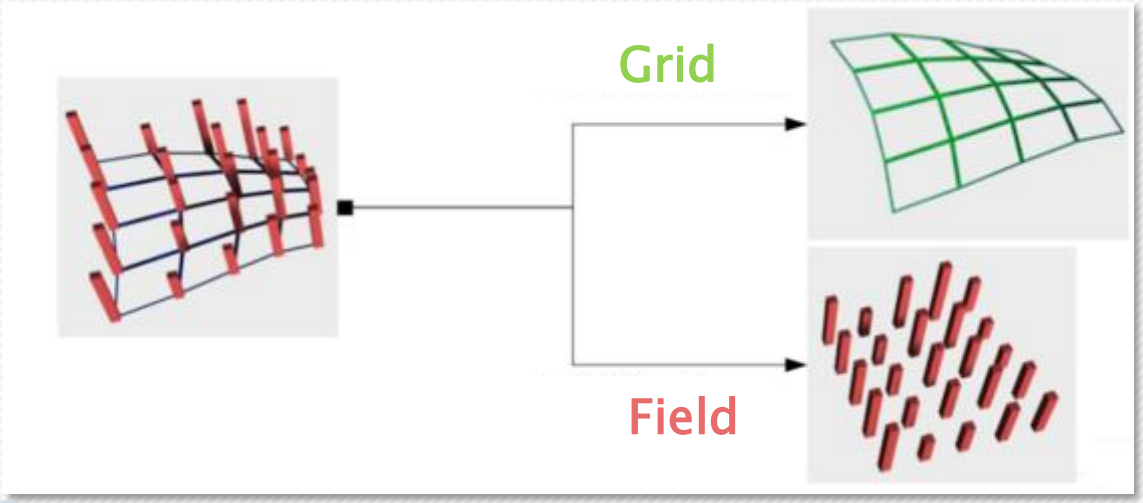
F5

Based on HDF5

Inspired by concepts of:

Topology	Differential Geometry	Geometric Algebra

Separation of **Grids (Topology, Geometry, Coordinate Systems)** and Data-**Fields**

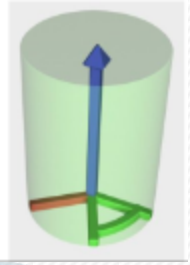
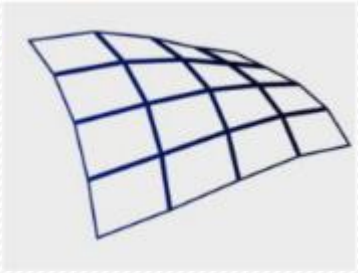
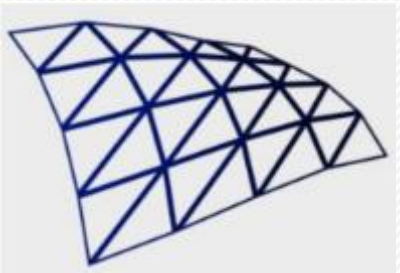
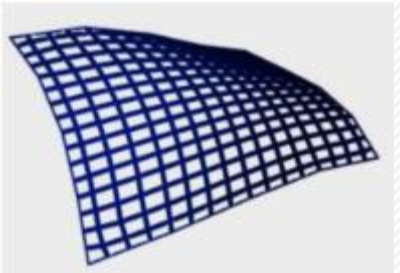


F5

Grid

the base space

- Manifold describing the **base space**
  - Topology
  - Refinement level
  - Coordinate representation
  - Vertex positions in representation
  - Neighborhood

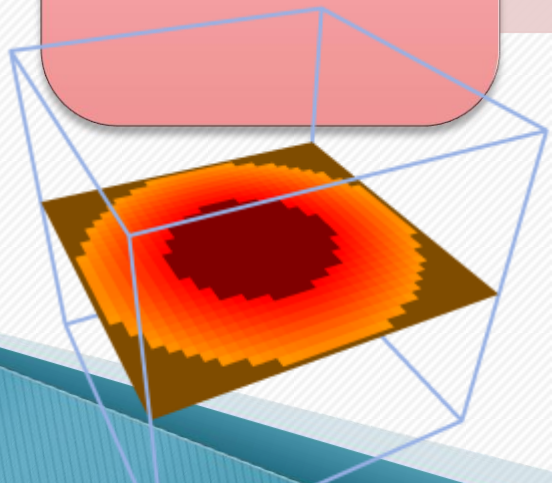
F5

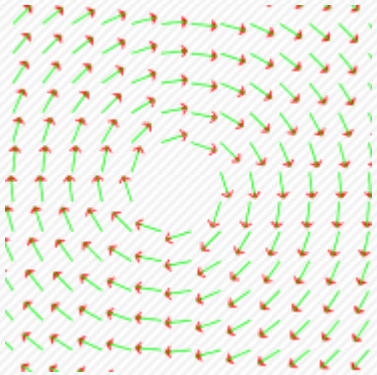
Field

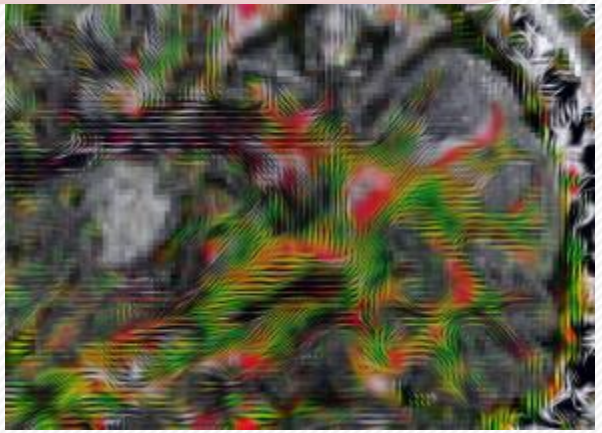
the fiber space

- Dataset holding numerical data
  - per k-cell on the grid (vertex, edge, triangle, cell, …)

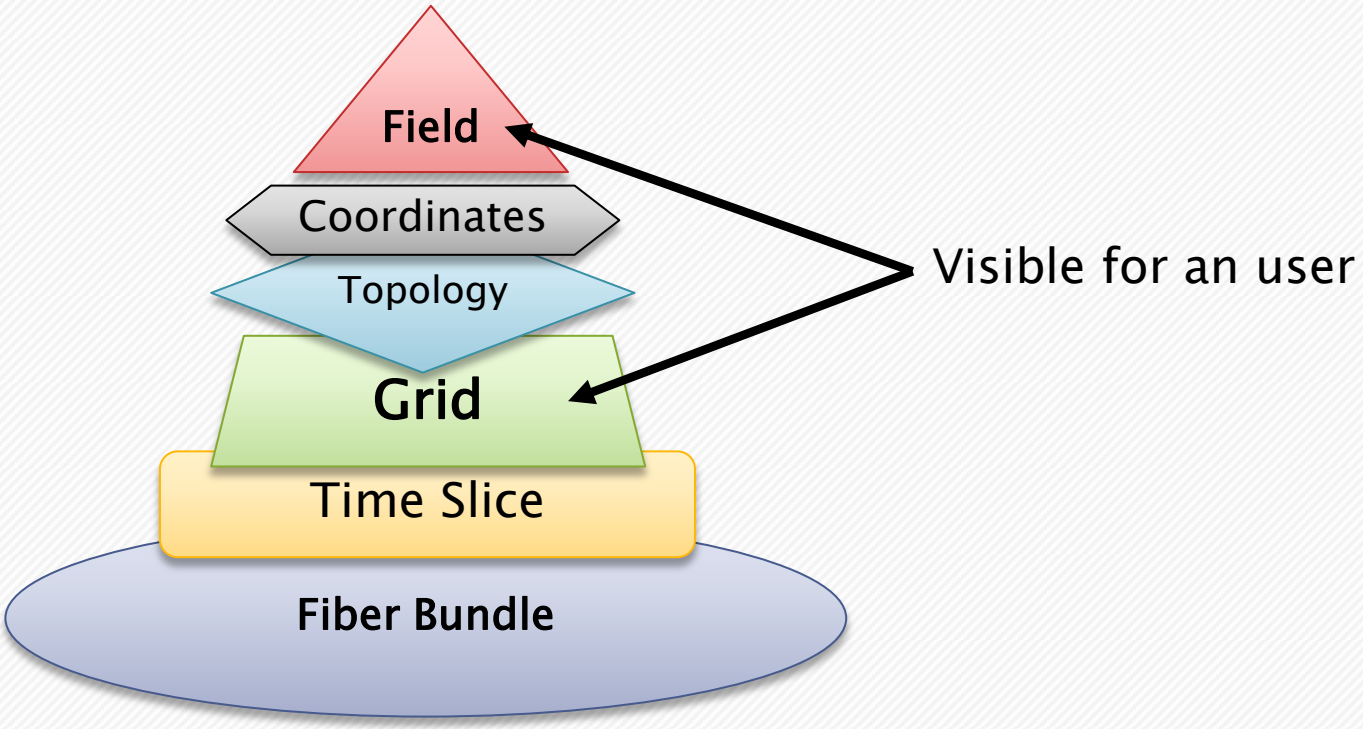
- Array of arbitrary type, for example:
  - Scalar
  - Vector, BiVector, …
  - Tensor
  - Any other user defined type





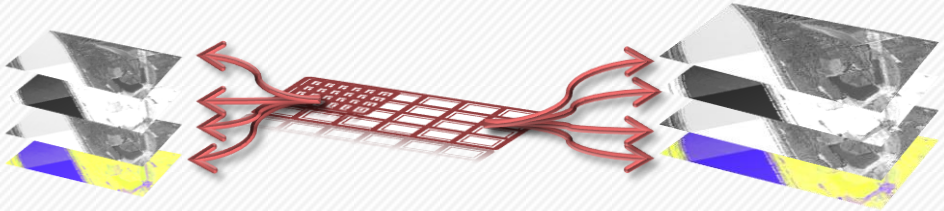


F5 – Hierarchy

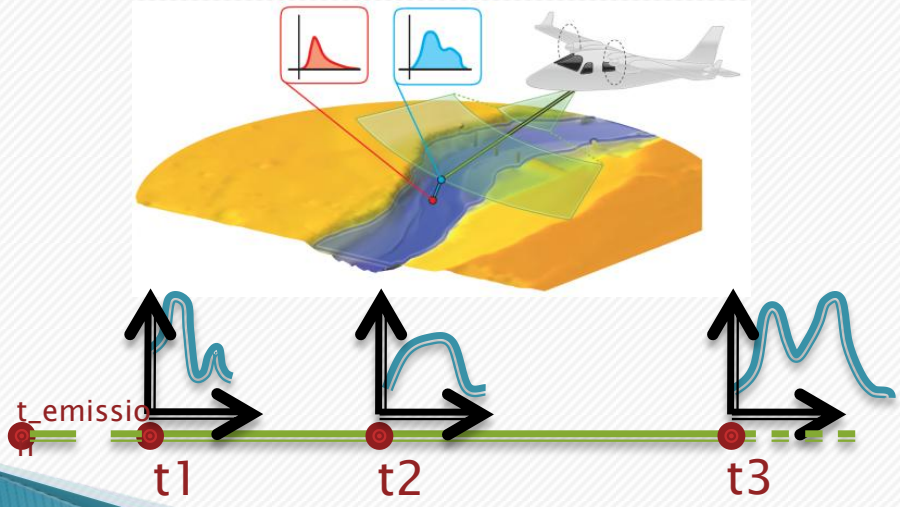


GeoScience – Examples

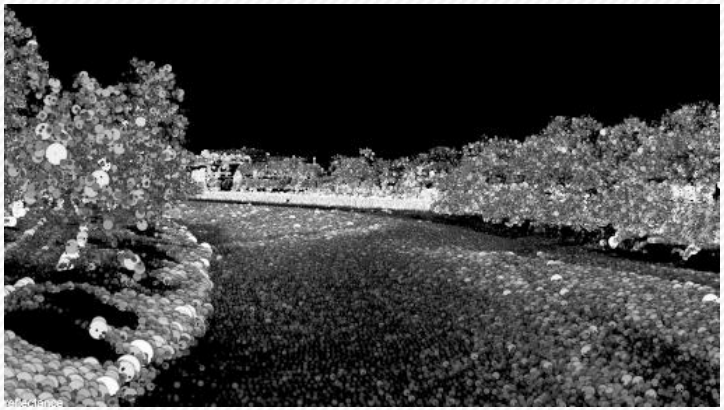
- Multi Channel – Multi Resolution Images:



- Full Waveform LIDAR:

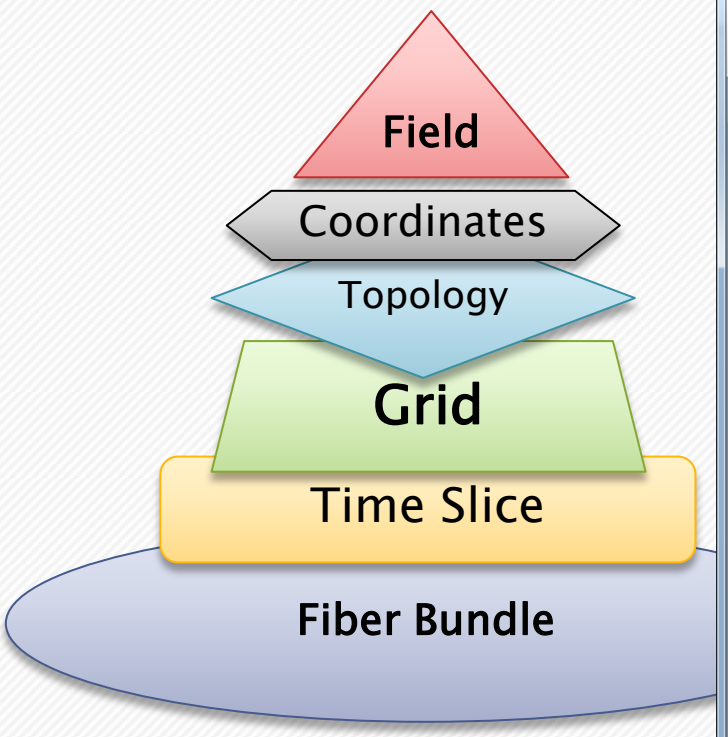


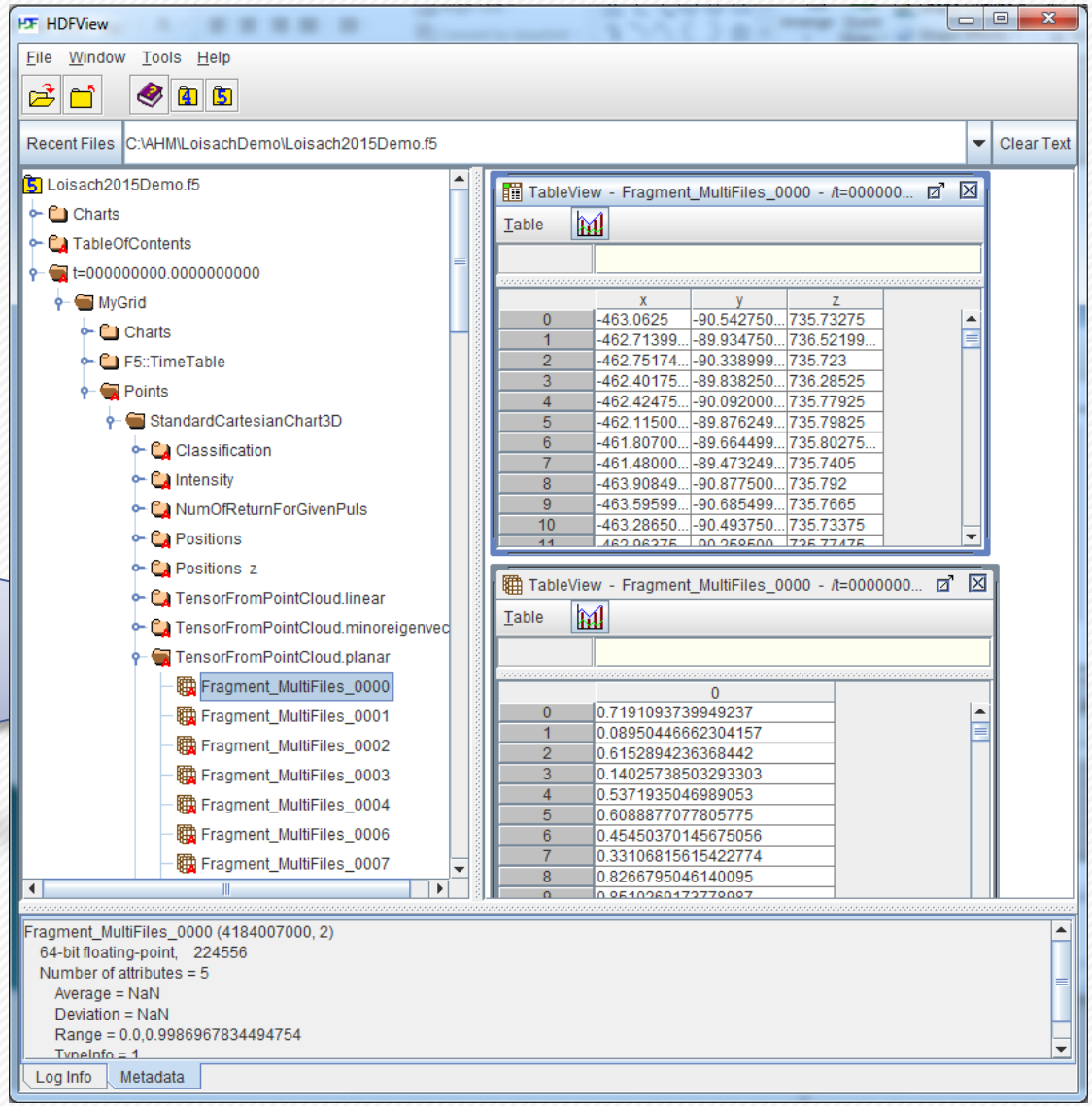
- Point Cloud Data



- Shape Data

F5 -Example


Field
Coordinates
Topology
Grid
Time Slice
Fiber Bundle


HDFView
File Window Tools Help
Recent Files
C:\AHM\LoisachDemo\Loisach2015Demo.f5
Clear Text
Loisach2015Demo.f5
Charts
TableOfContents
t=000000000.0000000000
MyGrid
Charts
F5::TimeTable
Points
StandardCartesianChart3D
Classification
Intensity
NumOfReturnForGivenPuls
Positions
Positions z
TensorFromPointCloud.linear
TensorFromPointCloud.minoreigenvec
TensorFromPointCloud.planar
Fragment_MultiFiles_0000
Fragment_MultiFiles_0001
Fragment_MultiFiles_0002
Fragment_MultiFiles_0003
Fragment_MultiFiles_0004
Fragment_MultiFiles_0006
Fragment_MultiFiles_0007
TableView - Fragment_MultiFiles_0000 - /t=000000...
Table
x y z
0 -463.0625 -90.542750... 735.73275
1 -462.71399... -89.934750... 736.52199...
2 -462.75174... -90.338999... 735.723
3 -462.40175... -89.838250... 736.28525
4 -462.42475... -90.092000... 735.77925
5 -462.11500... -89.876249... 735.79825
6 -461.80700... -89.664499... 735.80275...
7 -461.48000... -89.473249... 735.7405
8 -463.90849... -90.877500... 735.792
9 -463.59599... -90.685499... 735.7665
10 -463.28650... -90.493750... 735.73375
TableView - Fragment_MultiFiles_0000 - /t=0000000...
Table
0
0 0.7191093739949237
1 0.08950446662304157
2 0.6152894236368442
3 0.14025738503293303
4 0.5371935046989053
5 0.6088877077805775
6 0.45450370145675056
7 0.33106815615422774
8 0.8266795046140095
Fragment_MultiFiles_0000 (4184007000, 2)
64-bit floating-point, 224556
Number of attributes = 5
Average = NaN
Deviation = NaN
Range = 0.0,0.9986967834494754
Log Info
Metadata

F5 – Commandline Tools

HDF5

h5ls

- Good for overview and data viewing

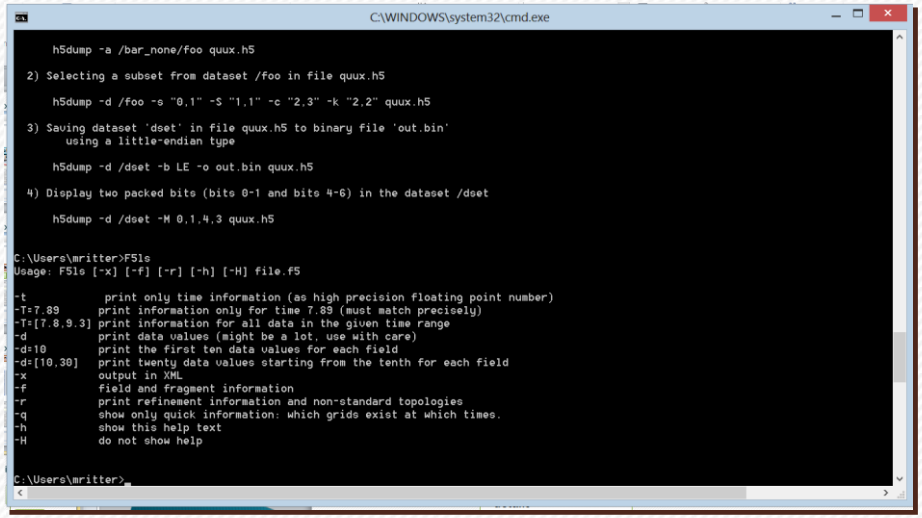
h5dump

- Good for inspecting details
- Xml conversion

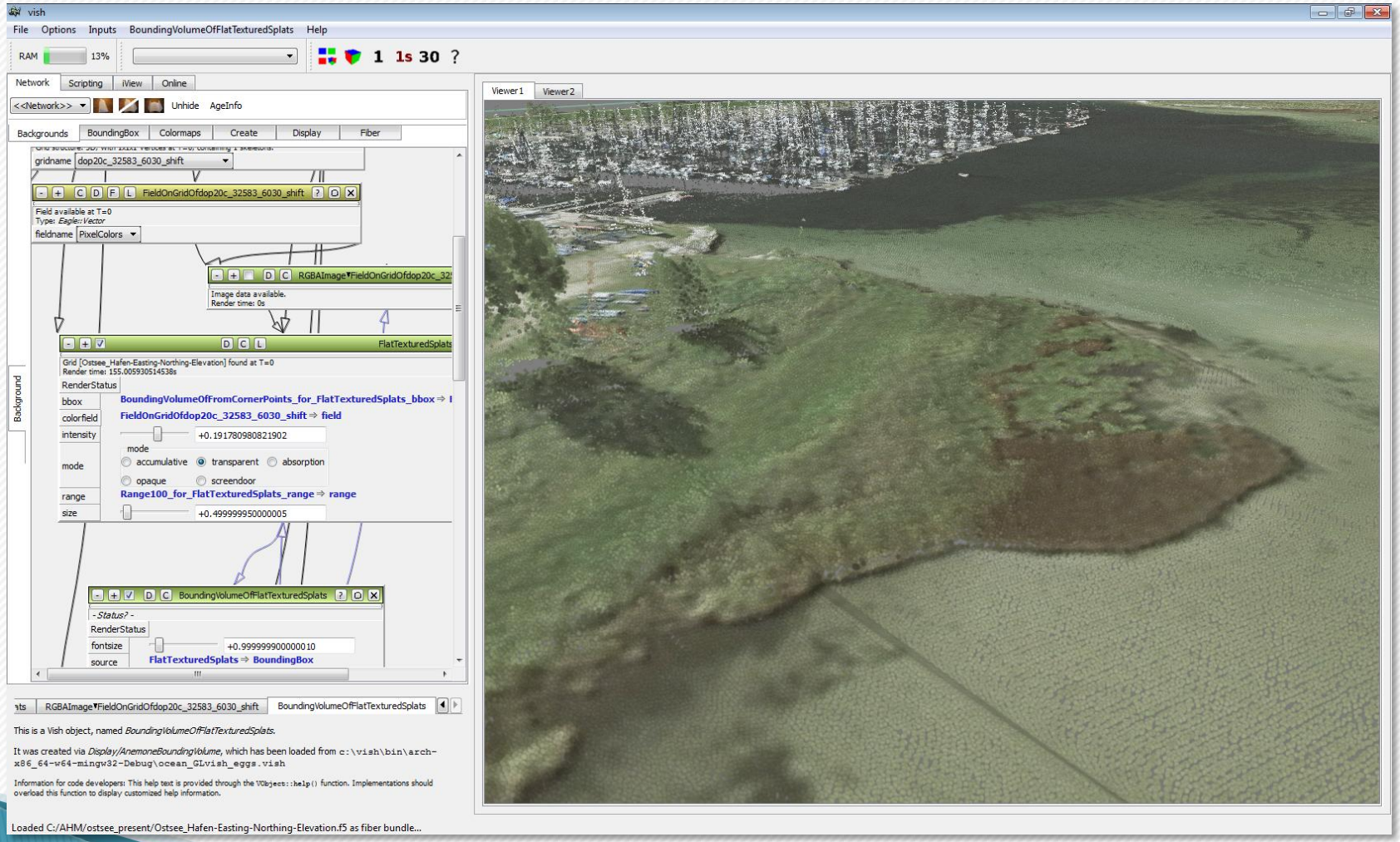
F5

F5ls

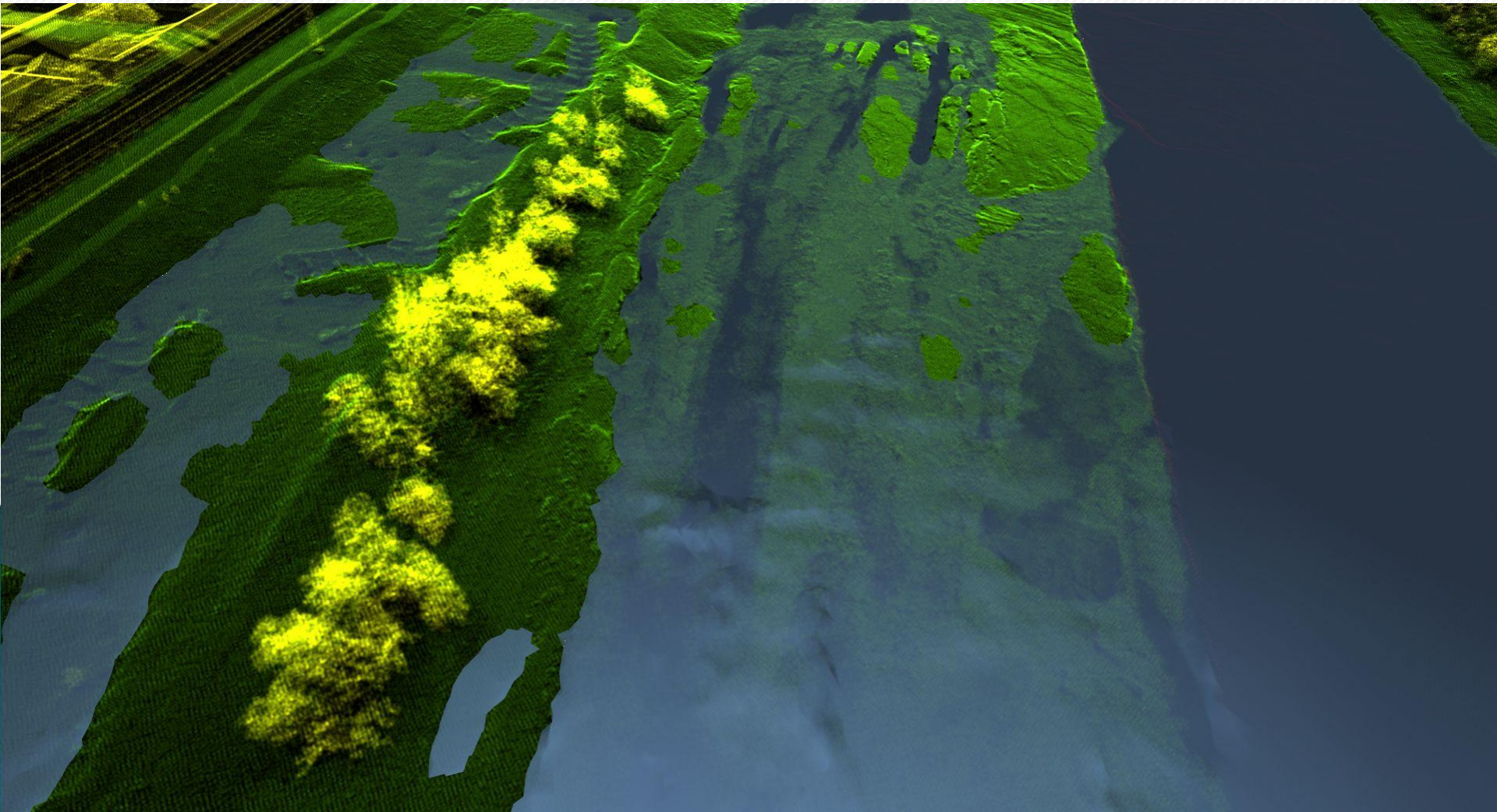
- Fiber bundle organized based data viewing



F5 access via HydroVish



Thank you!



F5

Supported **Grid** types:

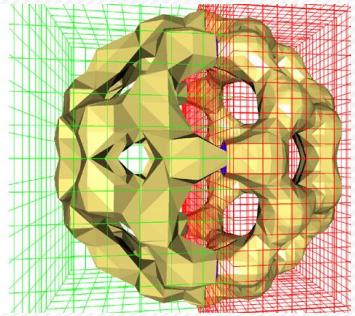
- Uniform Grid
- Curvilinear Grid
- Rectilinear Grid
- Adaptive Mesh Refinement Grid (AMR)
- Point Cloud
- Lines
- Triangular/Quad and Mixed Surfaces

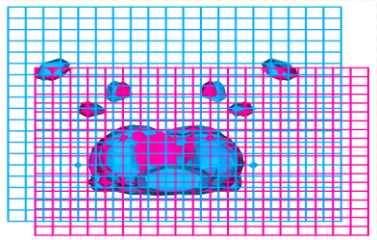
Grids can be fragmented (Blocks) having Ghost Zones

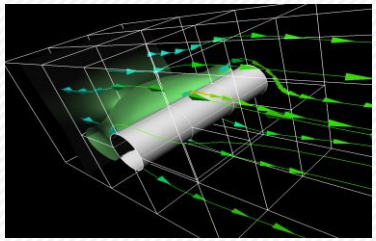
Grids can have refinement levels

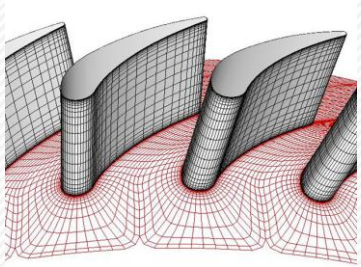
Work in progress:

- Hexahedral Grid
- FEM Grid
- Connected Graph Data
- Full Waveform LIDAR Laser Data





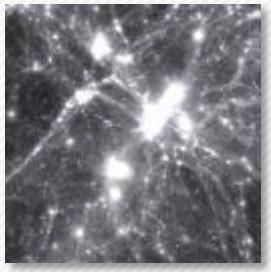


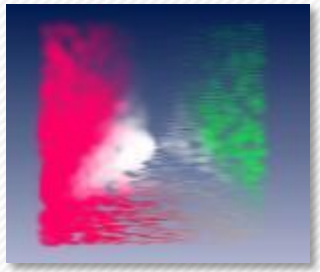


Fiber: 0D 1D 3D 6D

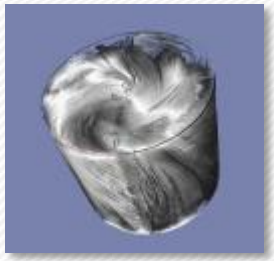
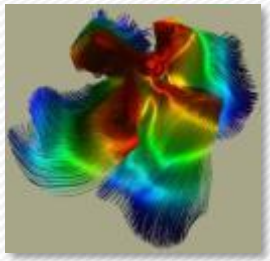
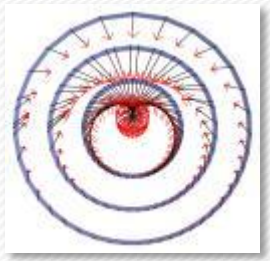
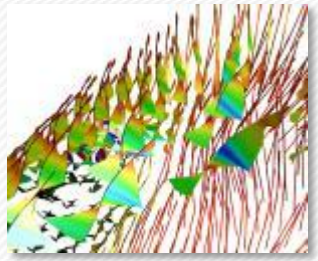
BASE:

0D

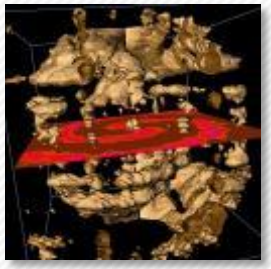
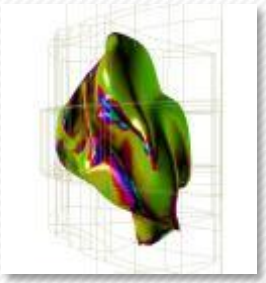
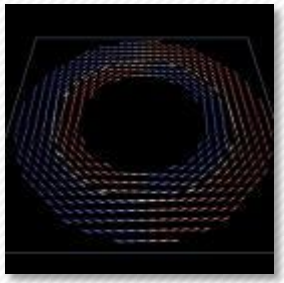
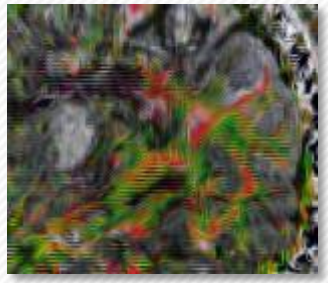


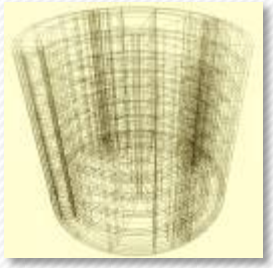
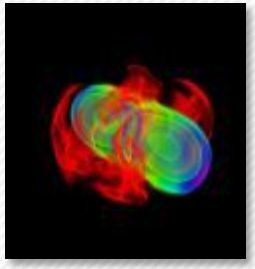
1D

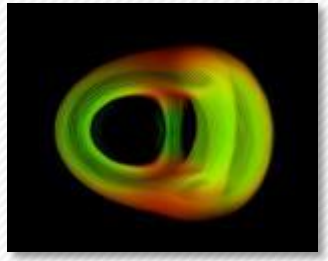





2D

3D



Definition: "K-cell"

- A subset $c \subset X$ of a Hausdorff space X is a k-cell if it is homeomorphic to an open k-dimensional ball in $\mathbb{R}^n$.

0-cells are called vertices, 1-cells are edges, 2-cells are faces or polygons, 3-cells are polyhedra.

F5 – as a library

Comes as a C-library
(free academic, free personal)

Provides high-level functions for data reading and writing

Comes with several data converters

- RDBtoF5, LidarToF5, ShapeToF5, F5toDXF, …

Comes with several F5-file based command-line tools

- F5ls, F5Copy, F5shift, …

Future: HDF5 – Virtual Object Layer (VOL)

- HDF5 VOL allows to replace high-level API operations
  → More **efficient** bundling of operations