Data Science for Space Science

A Database, Data-driven Modeling and Visualization Approach

James Ahrens

April 2019

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

Charge – Discuss Data Science Capabilities for Visualization and Databases for Space Science

Data Science Capabilities

1. Visualization
2. Databases

Goal: Specific recommendation of open source tools

Use Cases

1. Large spatial/temporal simulations
2. Ensembles of simulations and experiments
   - Materials database

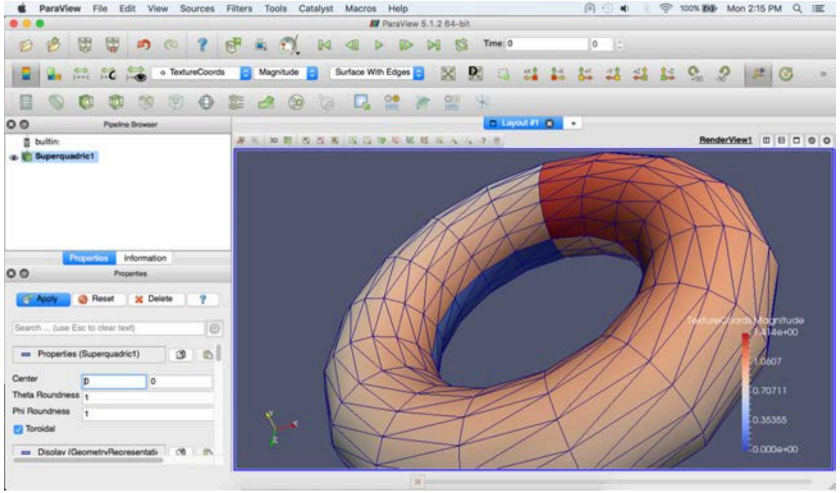
Use Cases	Data Science Capabilities	Open Source Tool
Large spatial/temporal simulations	Visualization	Paraview.org
Large spatial/temporal simulations	Database	HDFgroup.org
Ensembles of simulations and experiments	Visualization	Cinemascience.org
Ensembles of simulations and experiments	Database	SQLite.org

Visualization of large spatial/temporal simulations

Responding to the Trends: ParaView

- An open-source, scalable, multi-platform visualization application
- Support for distributed computation models to process large data sets
  - Billions of AMR cells, Scaling test over 1 Trillion cells
- Used by academic, government and commercial institutions worldwide
  - Downloaded ~100K times per year
  - Developed by Kitware, LANL, SNL…
- Originally designed to support a post processing workflow
  - Simulations save data to storage and scientist interactive visualizes results

- *The ParaView User's Guide*
  - http://www.paraview.org/paraview-guide/
- Tutorials
  - http://www.paraview.org/tutorials/
- The ParaView web page
  - www.paraview.org
- ParaView mailing list
  - paraview@paraview.org



Time: 0 | 0

Center 0 0
Theta Roundness 1
Phi Roundness 1
☑ Toroidal

Visualization of Pressure anisotropy in ParaView (Morley)

Pressure anisotropy in the equatorial plane, and the streamlines are traced through the 3D magnetic field

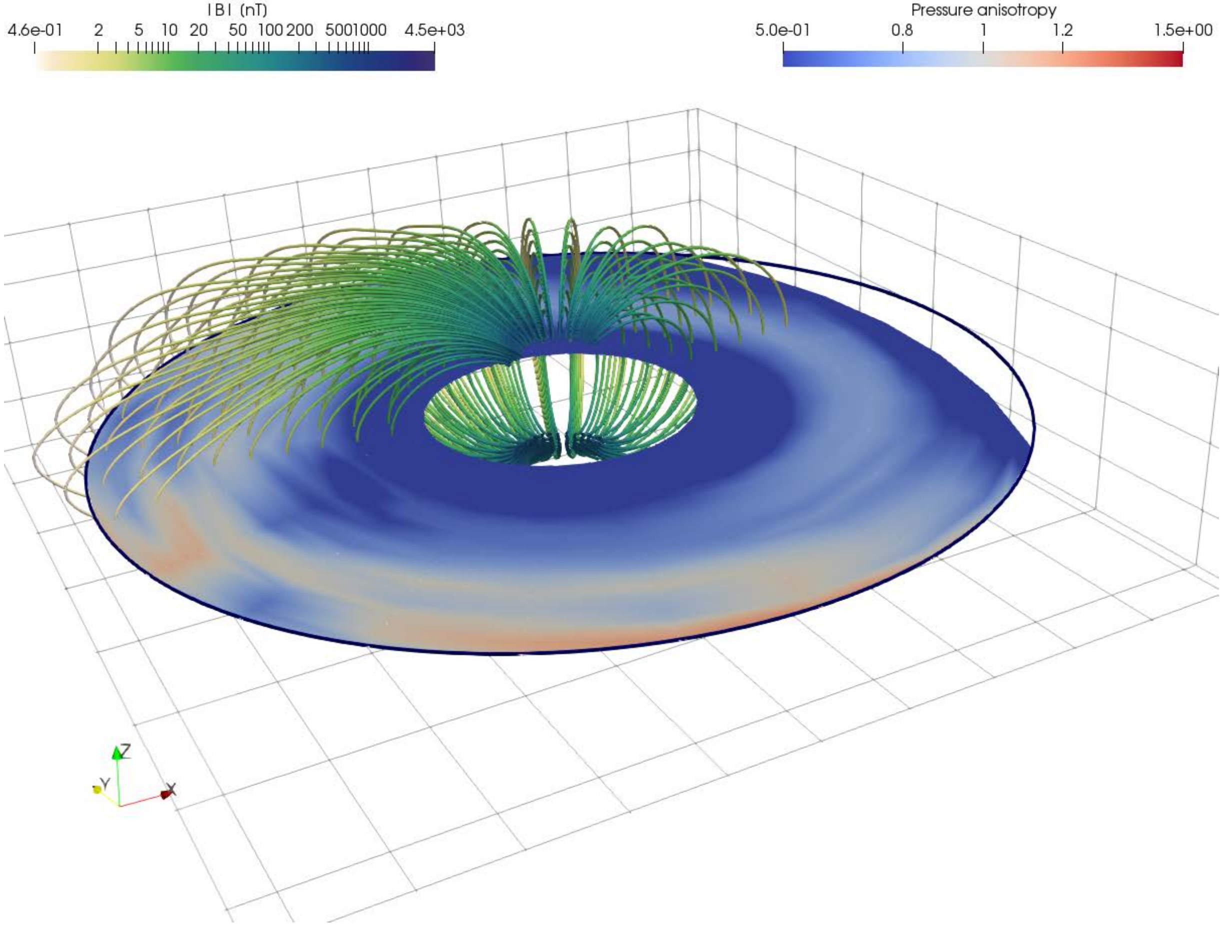
- calculated to be in pressure-balance with the plasma

The streamlines are traced from a set of points from a disc in the equatorial plane, applied a clip to cut away half of the field.

The black circle marks the outer boundary of the RAM model domain.

The bit that's missing was a cylindrical clip to exclude the region where the magnetopause (outer edge of Earth's magnetic field) had moved inside the model domain.

The X-direction is towards the Sun.



Database of large spatial/temporal simulations

- **Definition of database – organized collection of data, typically in digital form**
  - Codd’s *relational data* model - Based upon relational algebra and tuple relational calculus.
  - SQL (Structured Query Language) is a domain-specific language used in programming and designed for managing data held in a relational database management system.

- **Task-based approach**
  - – Querying of simulation data
    - Spatial and temporal understanding
  - – Storage and accessing of data & meta-data

Particle		
X_location	Type	Density
345	Smith	100
587	Jones	350
219	Smith	200

```
SELECT Type, X_location FROM Particle
WHERE Density > 150 ORDER BY
X_location;
```

- Extremely successful database model
- SQL table, Spreadsheets, Panda data frames, numpy array, R, Machine learning.

Database recommendation - SQLite

- SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine.
  - SQLite is the most used database engine in the world. There are over 1 trillion SQLite databases in active use.
- The SQLite file format is stable, cross-platform, and backwards compatible and the developers pledge to keep it that way through at least the year 2050.
- SQLite source code is in the public-domain and is free to everyone to use for any purpose.
- SQLite is serverless / file-based.

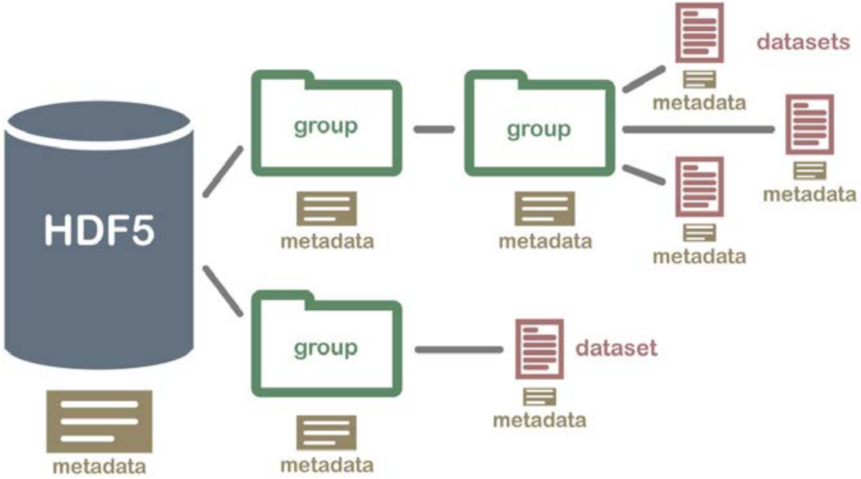
If your data elements are very large (>10s GBs) they need to be stored (not in a database) on a filesystem

Use HDF5 (hdfgroup.org):

HDF5 is a data model, library, and file format for storing and managing data.

- It supports an unlimited variety of datatypes,
- is designed for flexible and efficient I/O and for high volume and complex data.
- HDF5 is portable and is extensible, allowing applications to evolve in their use of HDF5.

You can still work metadata in relational database…



However:

- Realize that relational database is a huge market with drivers to make queries on massive data fast
  - For example, Hardware for SQL - Yellowbrick

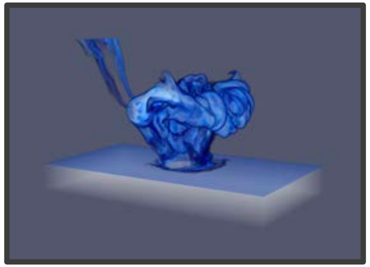
Visualization of ensembles of simulations

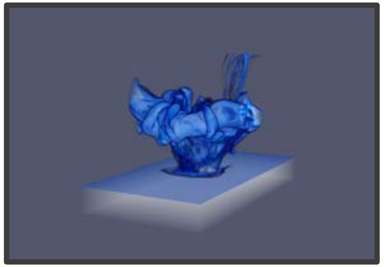
Cinema is a LANL-invented *ecosystem* for interactive analysis and exploration of databases

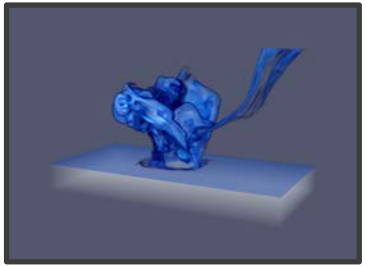
Original Use Case: Extreme Scale Scientific Simulations

Original insight: large data can be explored as a set of images

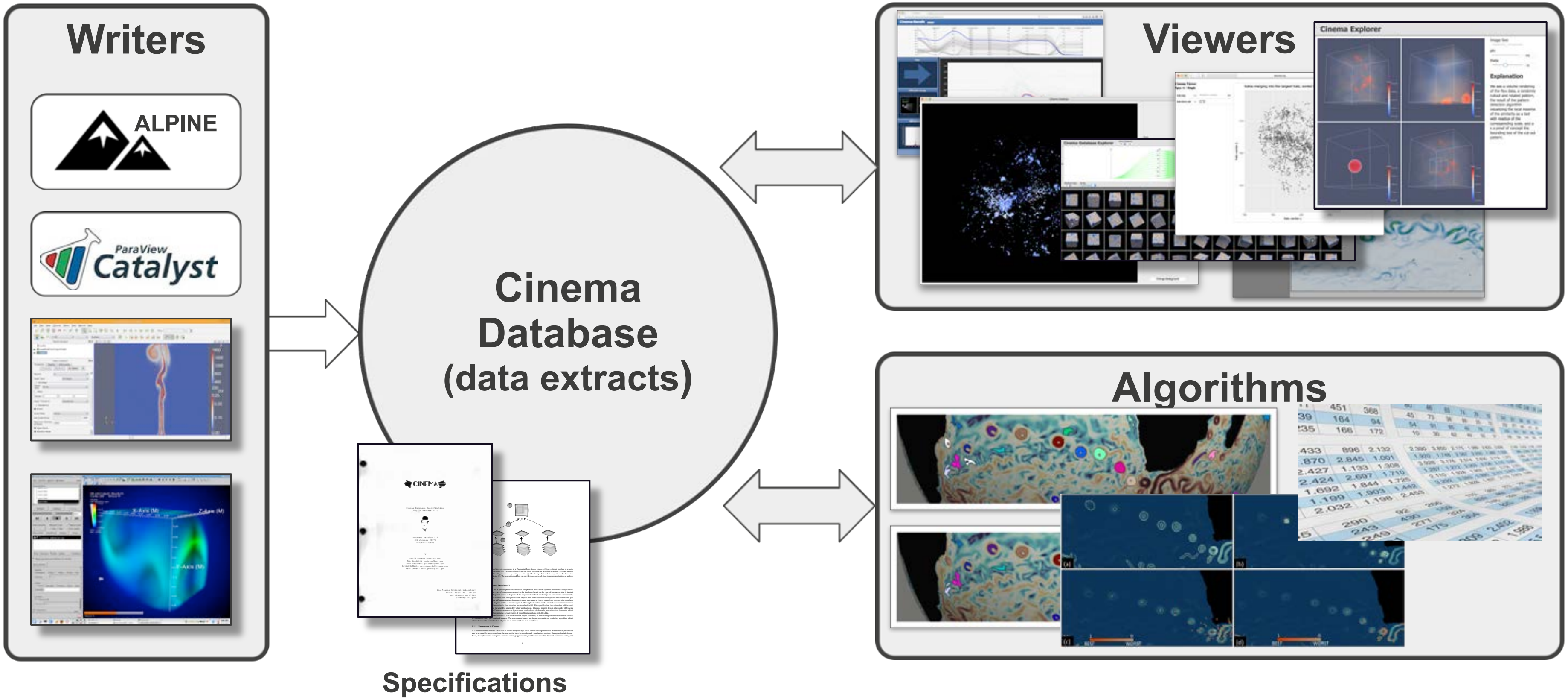
- Pre-render all images (via cartesian product of parameters) needed to explore data
  - ‘short circuit’ for normal visualization workflow
- Make database entries for parameters and associated image







CINEMA



Writers

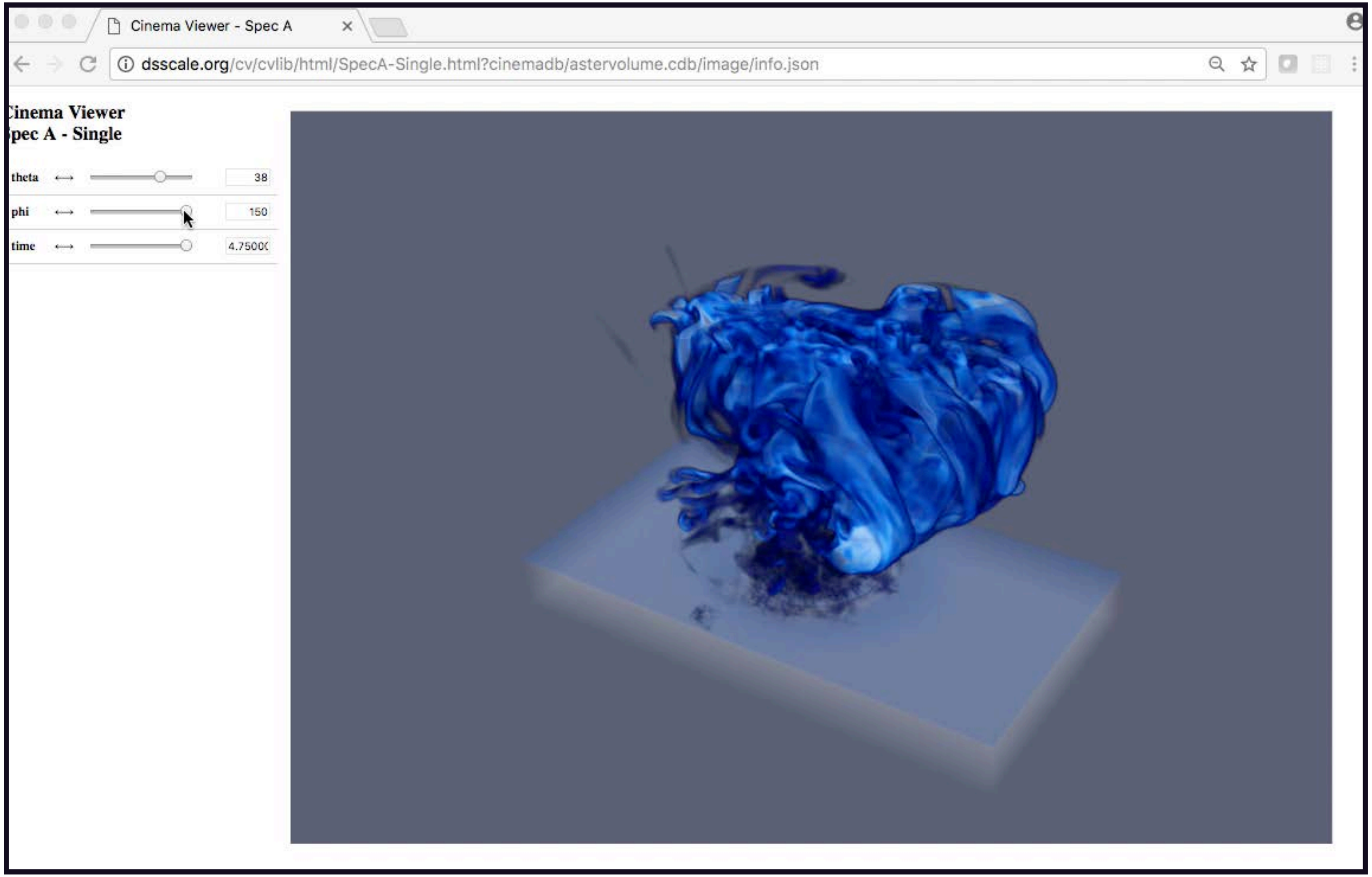
Cinema Database (data extracts)

Specifications

Viewers

Algorithms

Example: Explore Large Data Through Images



Cinema Viewer - Spec A

dsscale.org/cv/cvlib/html/SpecA-Single.html?cinemadb/astervolume.cdb/image/info.json

Cinema Viewer
Spec A - Single

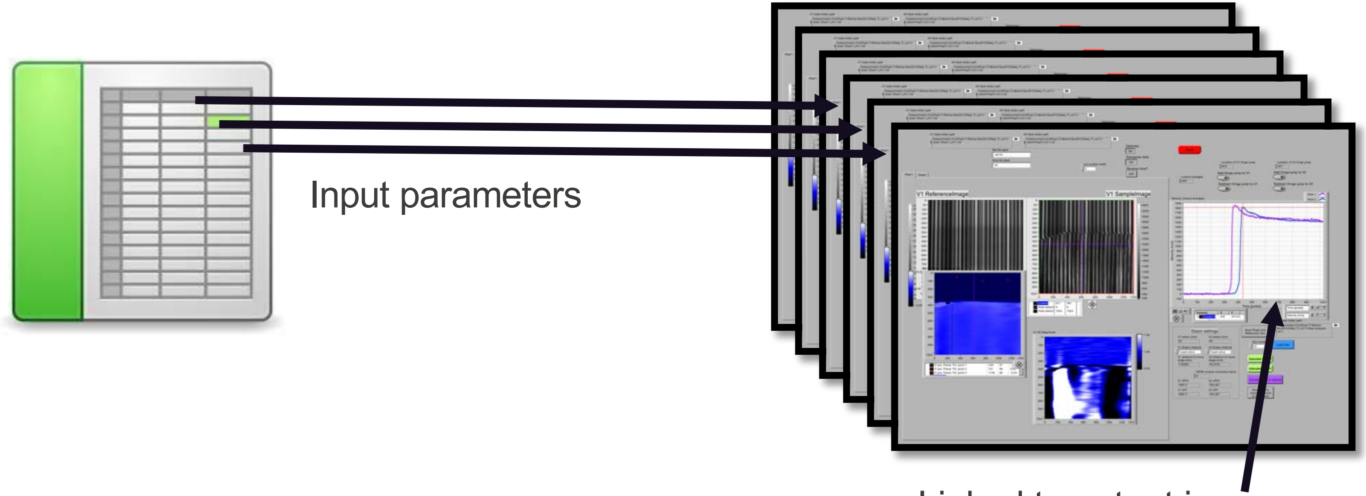
theta ⟷ 38

phi ⟷ 150

time ⟷ 4.7500

Visualization of ensembles of simulations

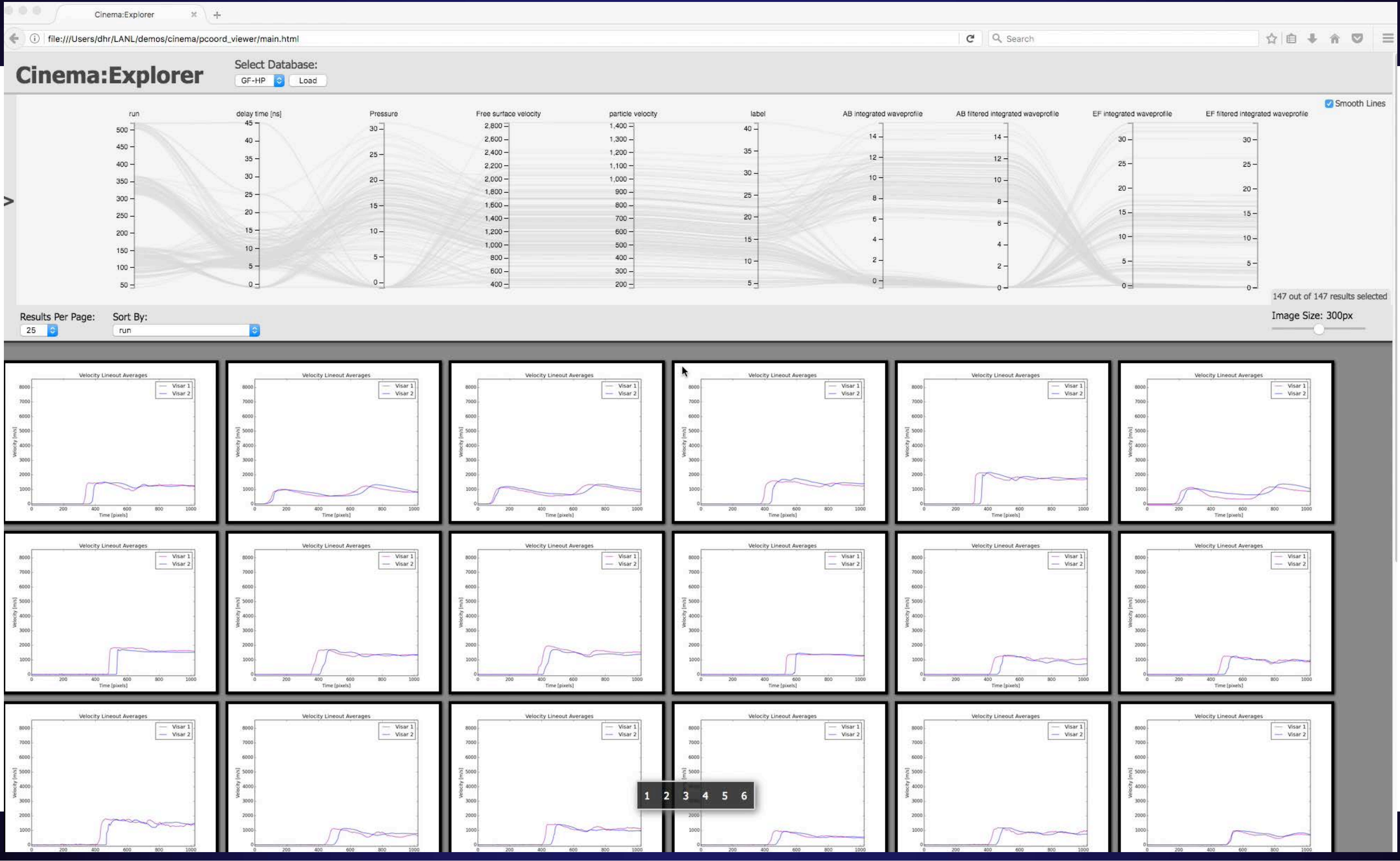
- This also works for a set of non-cartesian product entries
  - For example: Latin hypercube sampling



Input parameters

Linked to output images or data

Cinema:Explorer designed for high dimensional relational database visualization



file:///Users/dhr/LANL/demos/cinema/pcoord_viewer/main.html

Cinema:Explorer

Select Database: GF-HP

☑ Smooth Lines

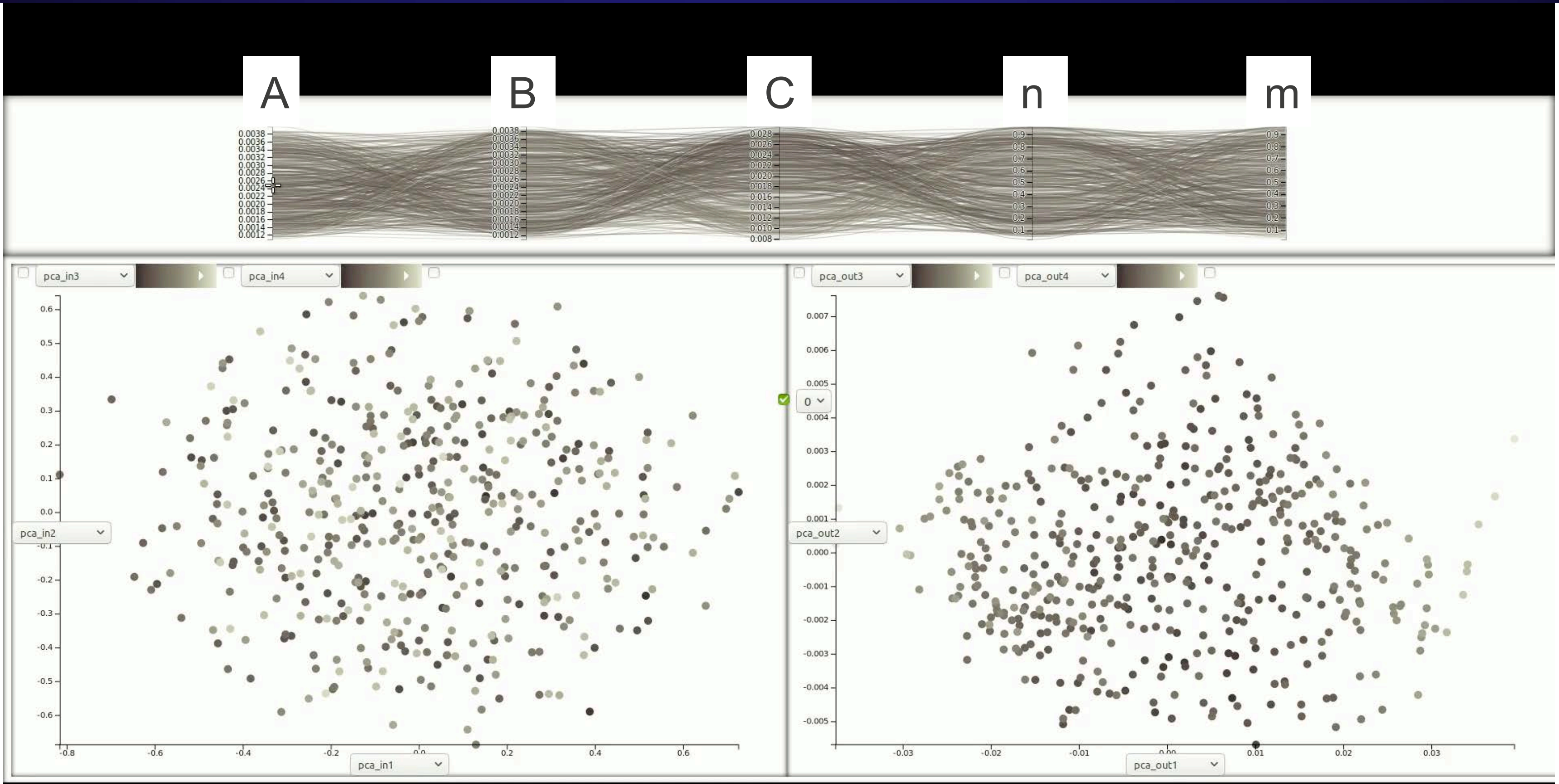
147 out of 147 results selected

Results Per Page: 25

Sort By: run

Image Size: 300px

Cinema input/output high dimensional analysis



PCA mapping of input parameters (A,B,C,n,m) to 2D

PCA mapping of output velocimetry to 2D

Collaborators

- Divya Banesh, John Barber, Derek Bingham, Ayan Biswas, Chris Biwer, Cindy Bolme, Marc Cawkwell, Soumya Dutta, Devin Francom, Anthony Fredenburg, Aditi Krishnapriyan, Earl Lawrence, Darby Luscher, Kelly Moran, Dan Orban, Arun Ramanathan, Kyle Ramos, David Rogers, Richard Sandberg, Christine Sweeney, Cameron Tauxe, Ash Tripathi, Sven Vogel, David Walters (LDRD)
- David Rogers (Team Lead) Curtis Canada, Patricia Fasel, Li-Ta (Ollie) Lo, John Patchett, Christopher Sewell, Jonathan Woodring, Roxana Bujack, Pascal Grosset (Data Science at Scale Team)
- Garrett Aldrich, Ayan Biswas, Jesus Pulido, Max Zeyen, Daniel Orban, Divya Banesh, Cameron Tauxe, Soumya Dutta, Anne Berres, Daniel Ben Naim (Students and Postdocs)
- Berk Geveci, Patrick O’Leary, Dave DeMarle, Sebastian Jourdain (Kitware)
- Colin Ware (UNH), Francesca Samsel, Greg Abram, Terry Turton (UTexas) (Stardust)