SAND2011-8453C

Feature-Based Statistical Analysis of Combustion Simulation Data

Janine Bennett[1],Vaidyanathan Krishnamoorthy[2], Shusen Liu[2], Ray Grout[3], Evatt Hawkes[4], Jacqueline Chen[1], Jason Shepherd[1], Valerio Pascucci[2], and Peer-Timo Bremer[2,5]

[1]Sandia National Laboratories, [2]University of Utah, [3]National Renewable Energy Laboratory, [4]University of New South Wales, [5]Lawrence Livermore National Laboratory

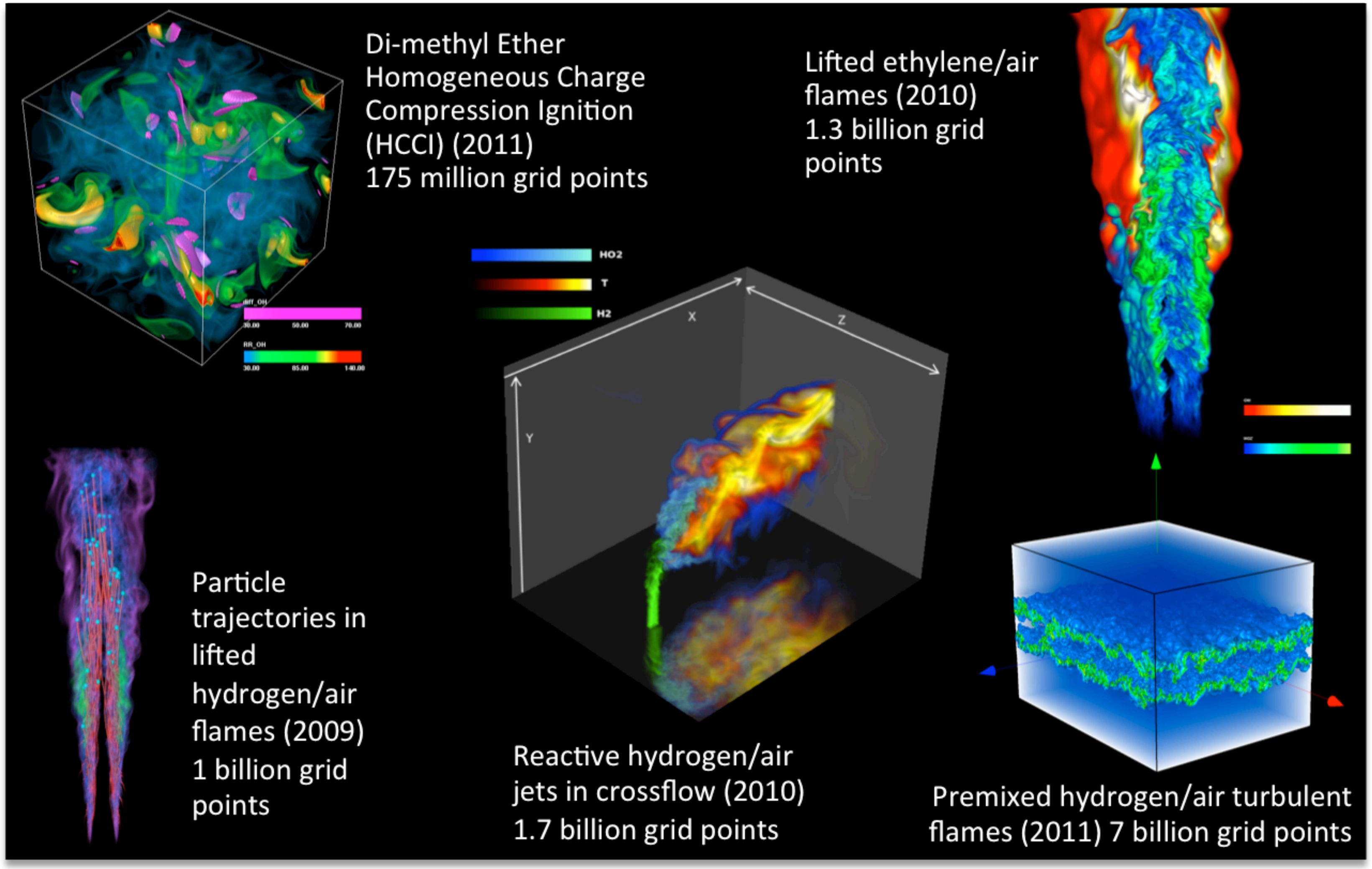
Motivation: state of the art simulations generate large-scale, complex data

- Increases in data size + complexity
  - Spatial resolution
  - Number of variables
  - Number of scales represented
- Our contribution: a feature-based analysis and visualization framework for large-scale data

Images courtesy of: National Energy Research Scientific Computing Center, Los Alamos National Laboratory, Argonne National Laboratory, and Oak Ridge Leadership Computing Facility.

Direct Numerical Simulations (DNS) are used to study fundamental turbulence-chemistry interactions

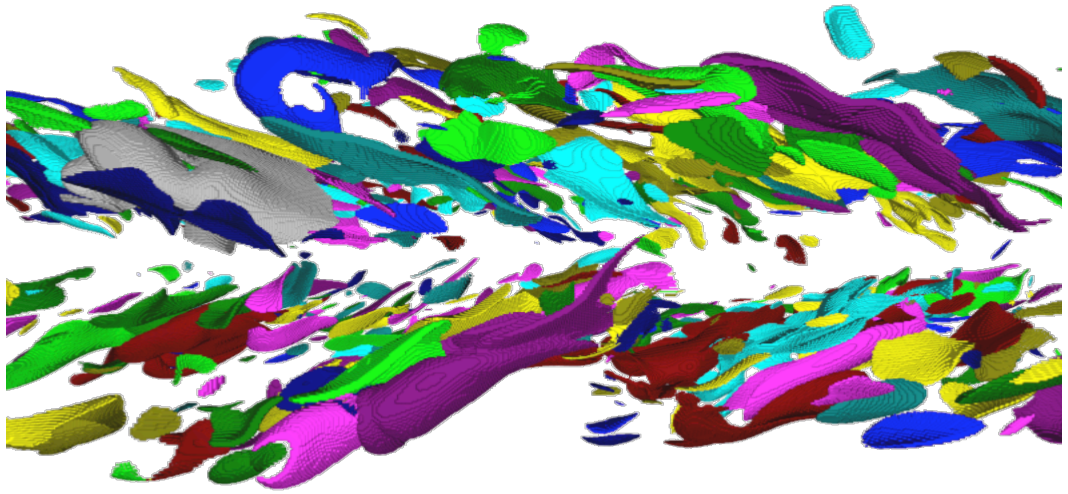
- DNS data is large and complex
- How do you define features?
  - Thresholds may vary locally
  - Thresholds may not be known *a priori*
- How many features are there?
- What is the behavior of other variables inside the features?

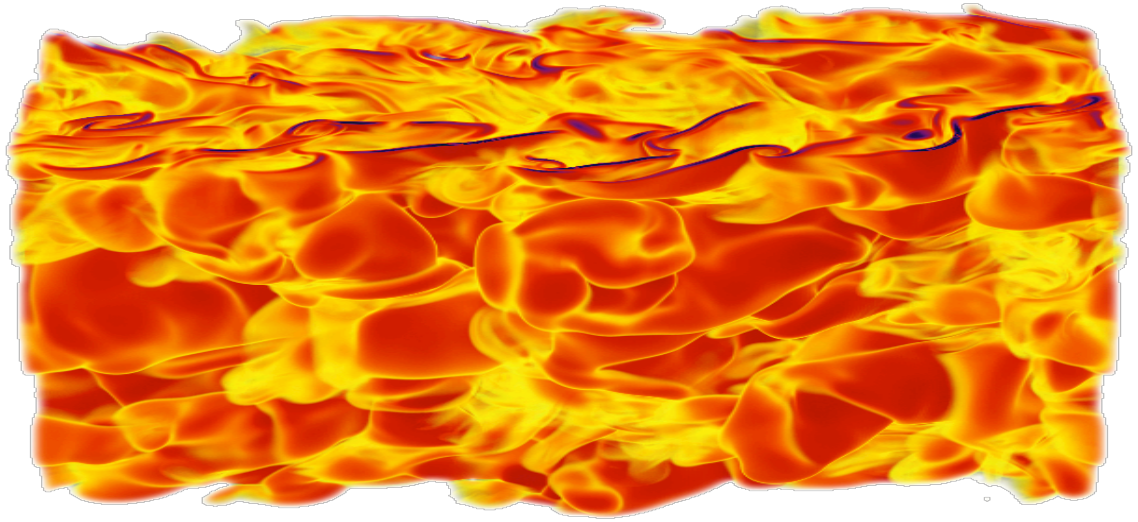


Recent DNS configurations performed using S3D, a DNS code written by Dr. Jacqueline Chen & her research group at the Combustion Research Facility, Sandia National Laboratories

Case study: characterizing the relationship between the mean temperature and thickness in regions of high χ

- Scalar dissipation rate, χ: rate of molecular mixing
- Goals:
  - Study relationship between mechanical strains & chemical processes
  - Compute feature-based statistical summaries

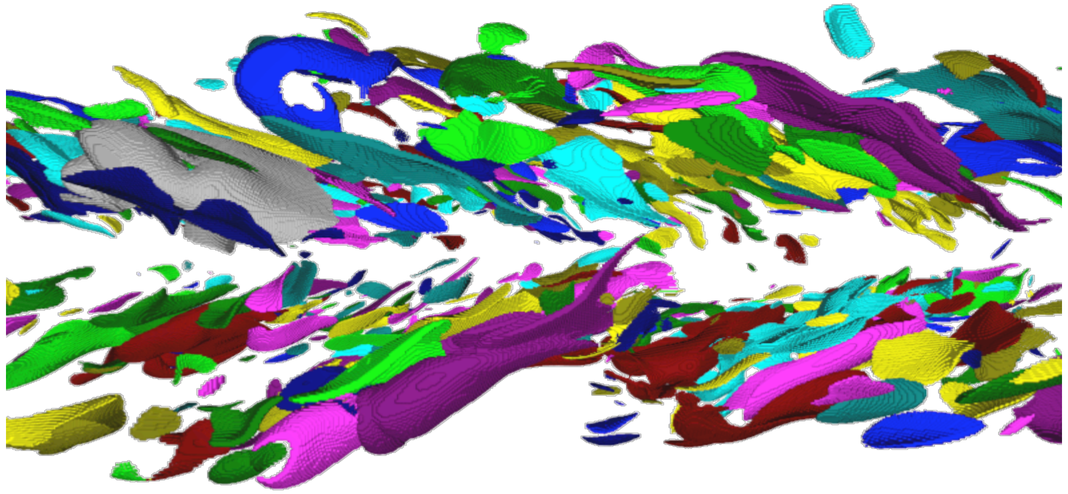


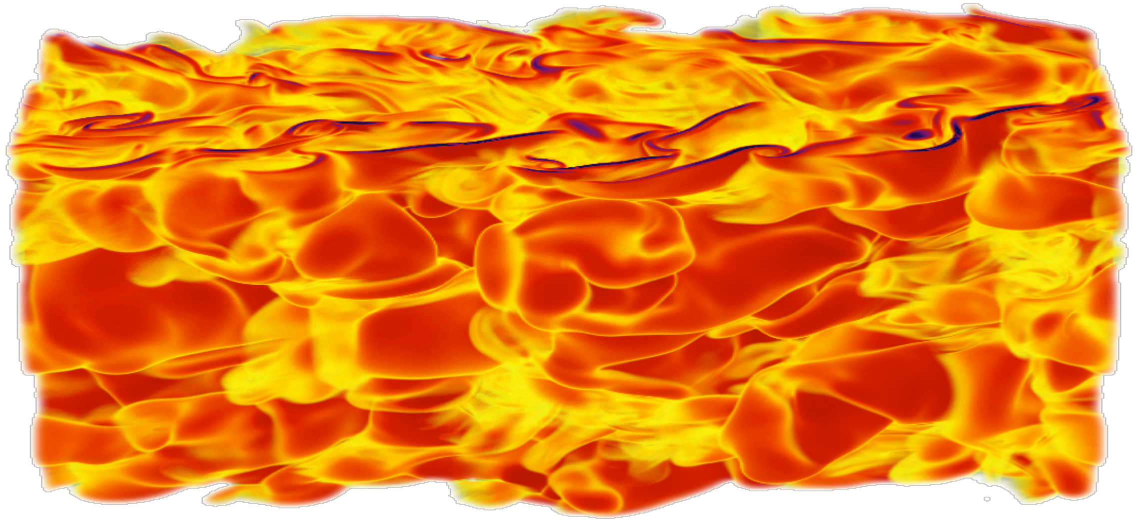


Case study: characterizing the relationship between the mean temperature and thickness in regions of high χ

Challenges:

- χ structures are defined by locally varying isovalues
- Sub-selection based on other criteria is important
- Visual feedback of the effect of parameter choices is desired
- Large data complicates matters



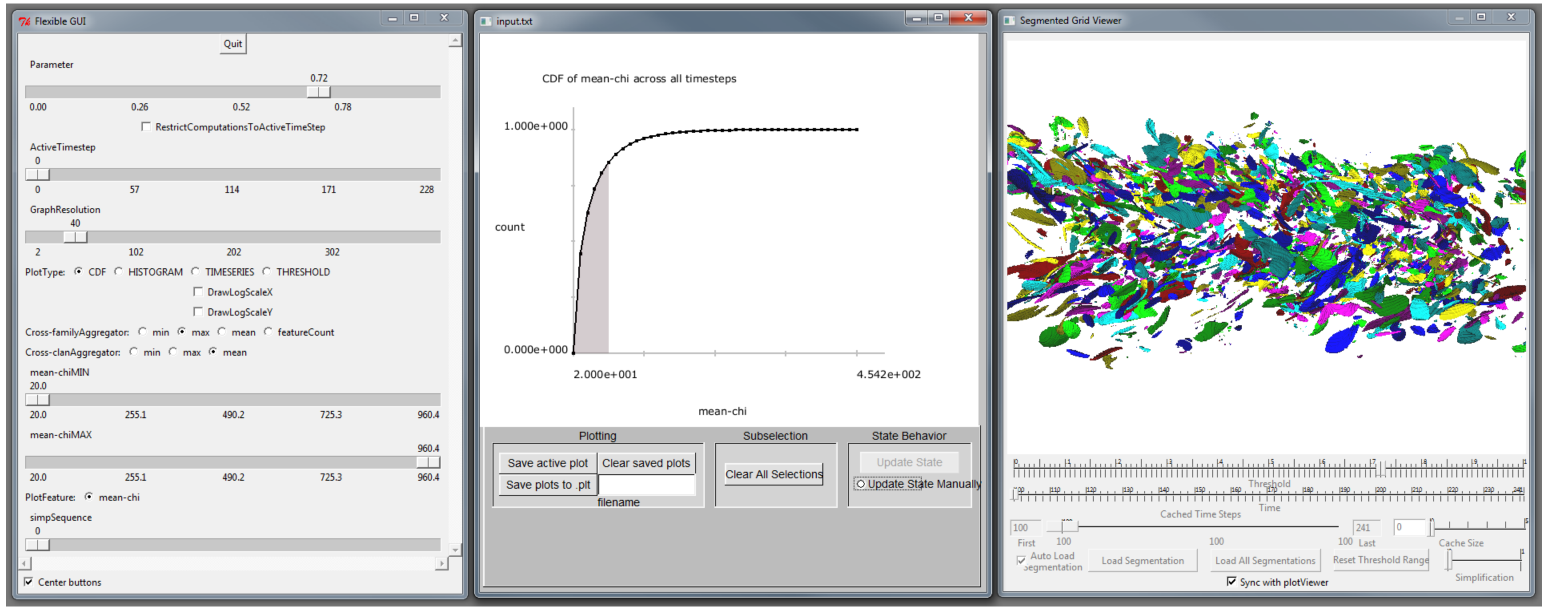


Related Work

- Conditional statistics
- Data warehouse technologies: *e.g.* FastBit [Wu, *et al.*]
  + Extracts and aggregates pre-computed information
  + Uses compressed bit map indices to provide efficient sub-selections
  – Regions cannot always be defined by range queries
  – Feature parameter thresholds not always known *a priori*
- Feature hierarchies
  - Merge Trees [Carr *et al.*, Pascucci *et al.*],
  - Morse-Smale Complex [Laney *et al.*, Bremer *et al.*, Gyulassy *et al.*]
  - Clustering methods [Hartigan]

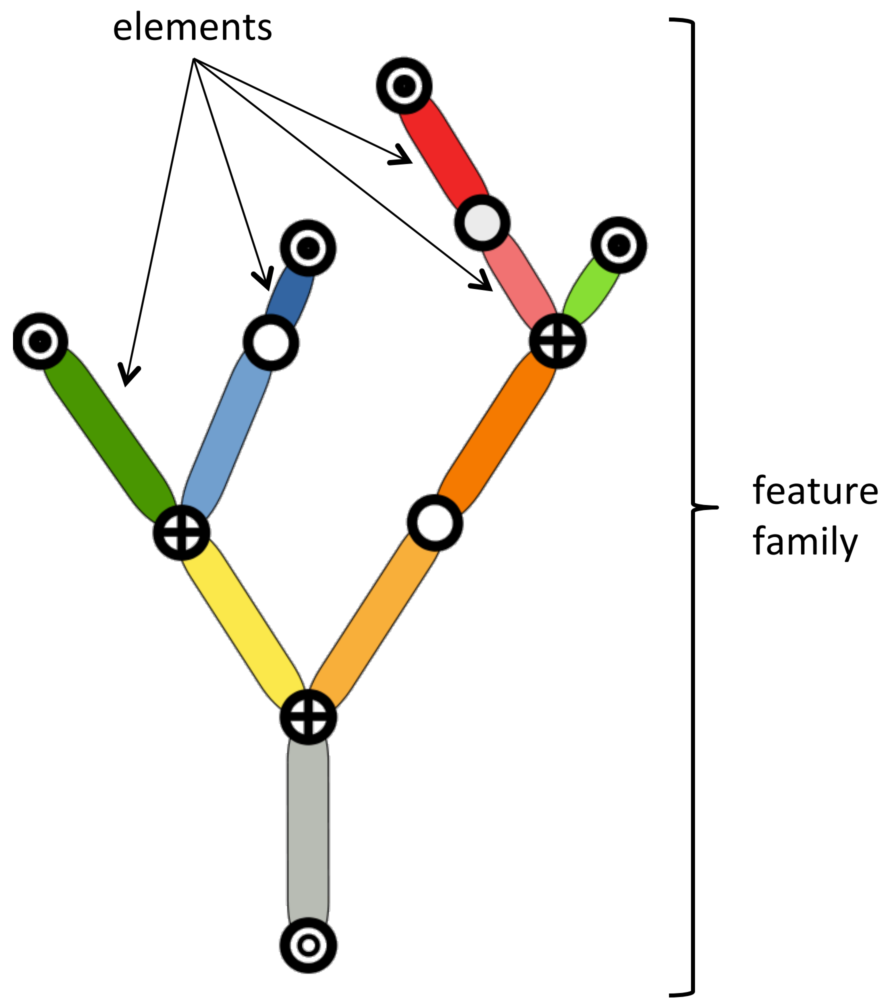
We have developed an integrated feature-based analysis & visualization framework to study large scientific data

- Pre-compute meta-data
  - Efficient encoding for multi-resolution hierarchies & statistics
  - Drastic data reduction
  - Preserves moments
- Interactive exploration
  - On the fly aggregation of feature-based spatial & temporal statistics
  - Creation of spatial & temporal statistical summaries
  - Linked view display of statistics & features



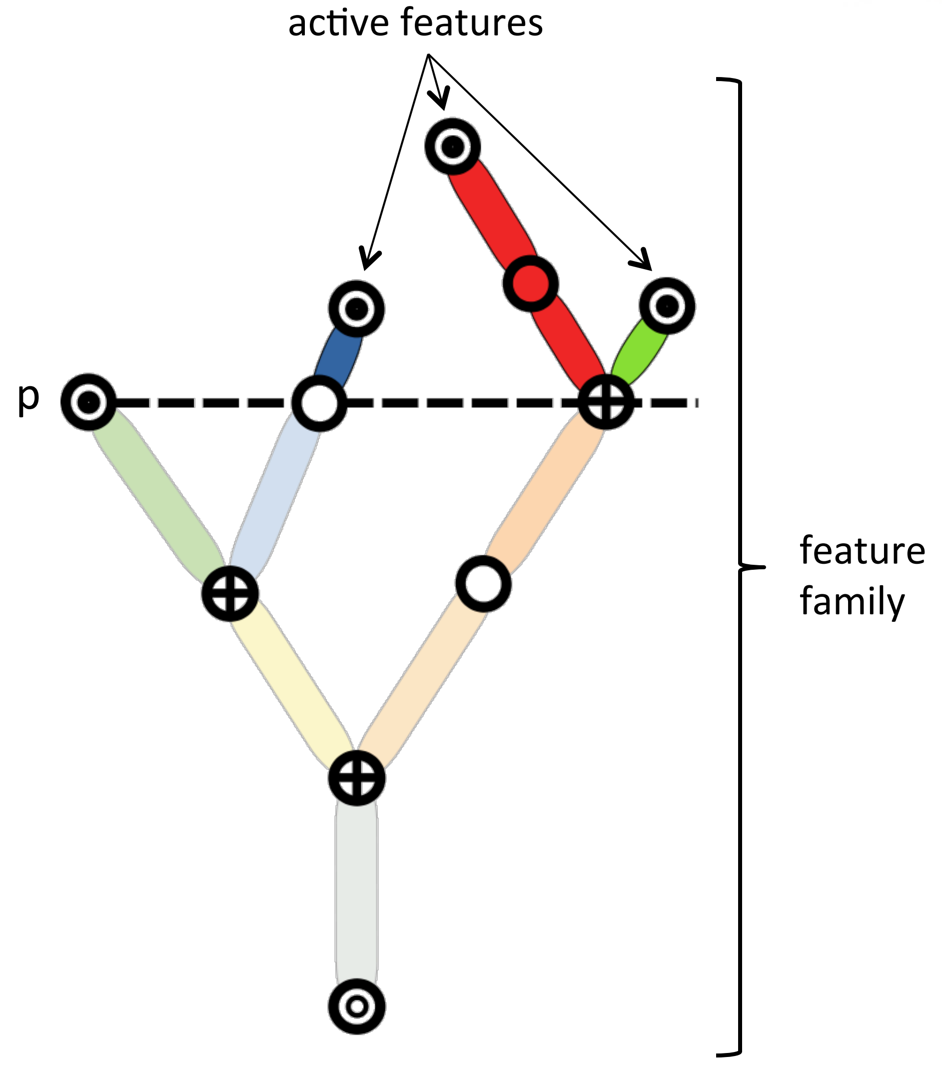
Augmented feature families form a compact data representation

- Element: spatial region of the input domain
  - Life span information
  - Parent/child information
  - Optional associated statistics
- Feature: collection of elements
- Feature Family: one-parameter family of features
  - Active features are identified by specifying a parameter value
- Clan: collection of feature families
  - *e.g.* across time steps

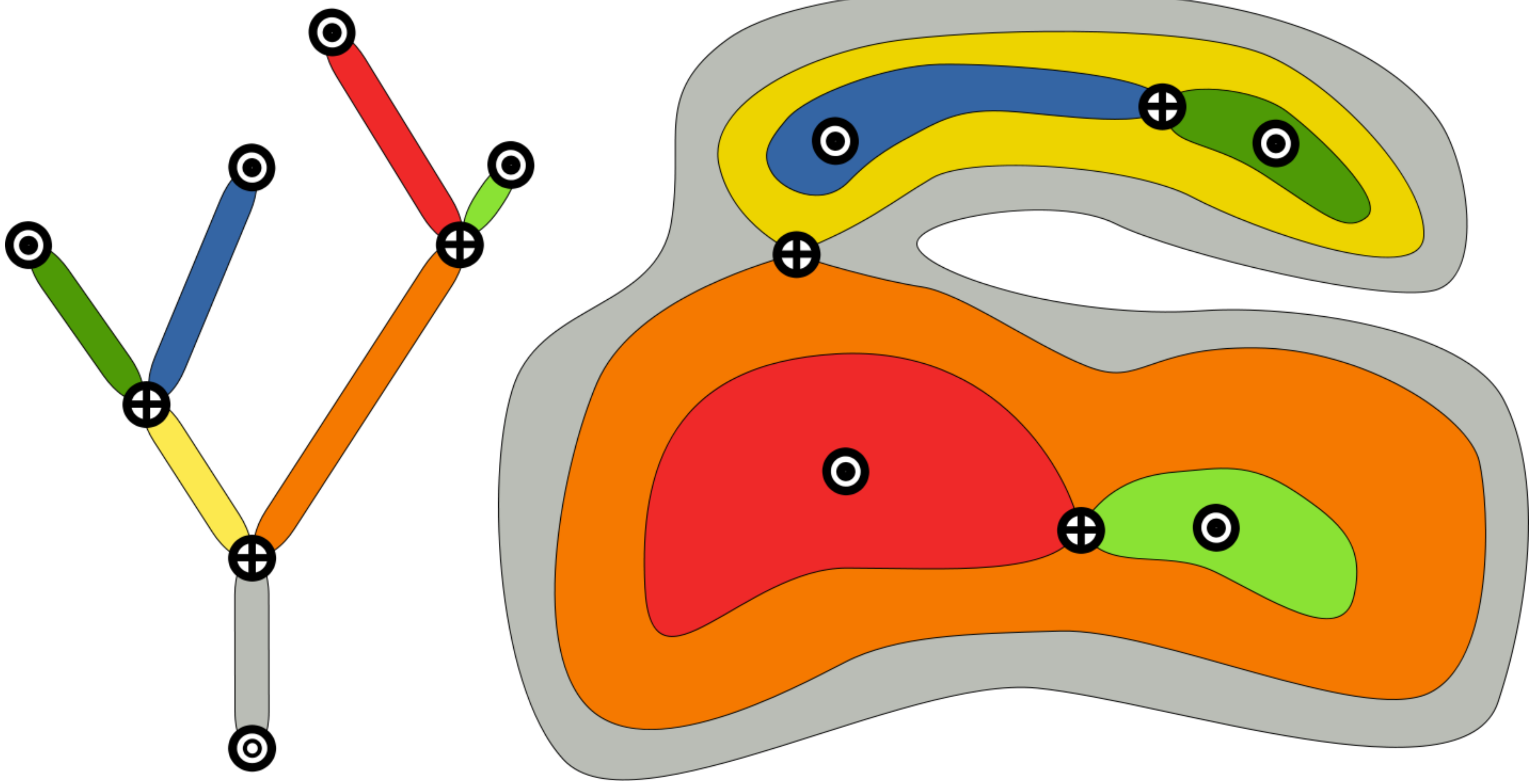


Augmented feature families form a compact data representation

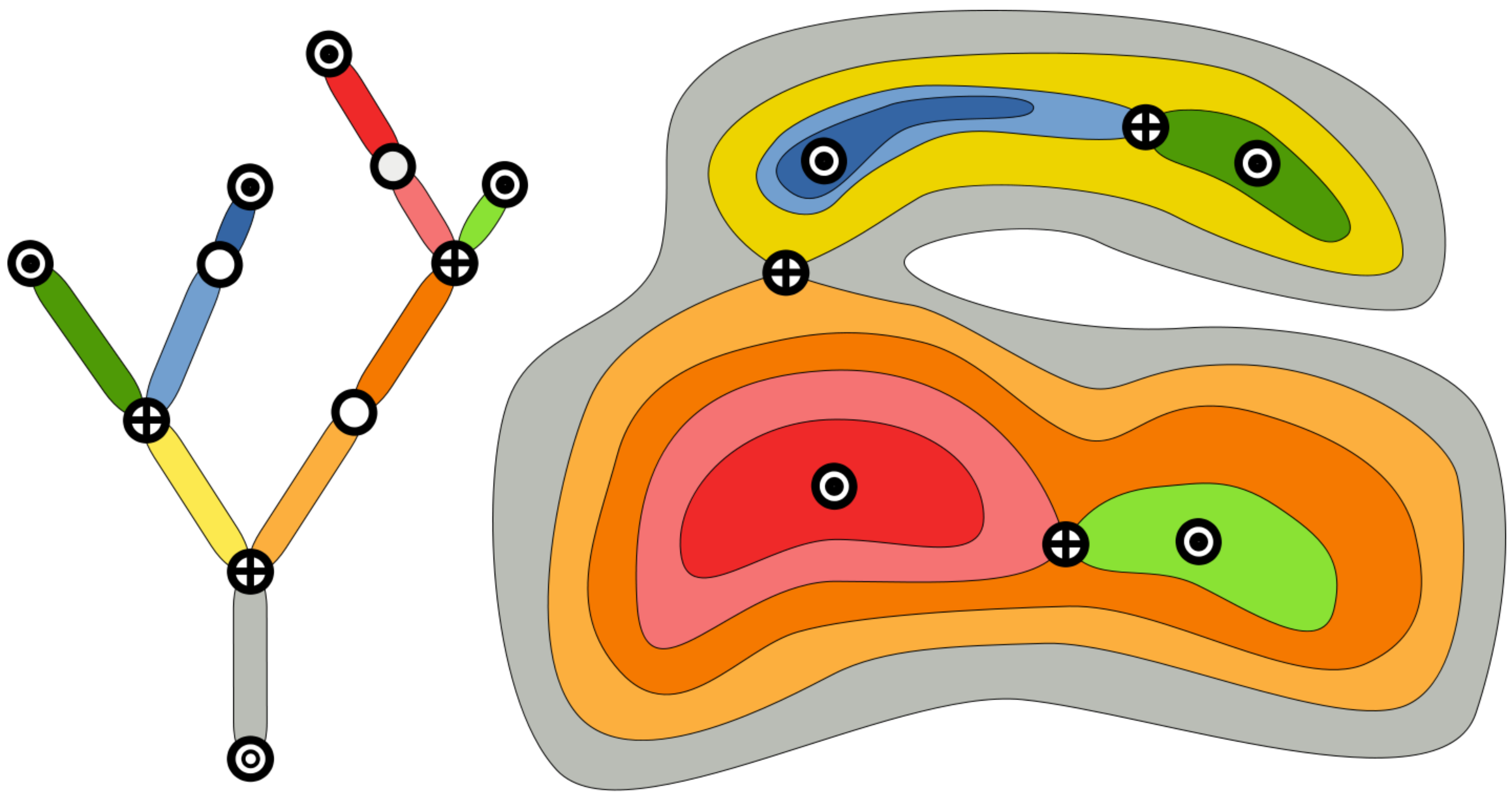
- Element: spatial region of the input domain
  - Life span information
  - Parent/child information
  - Optional associated statistics
- Feature: collection of elements
- Feature Family: one-parameter family of features
  - Active features are identified by specifying a parameter value
- Clan: collection of feature families
  - *e.g.* across time steps



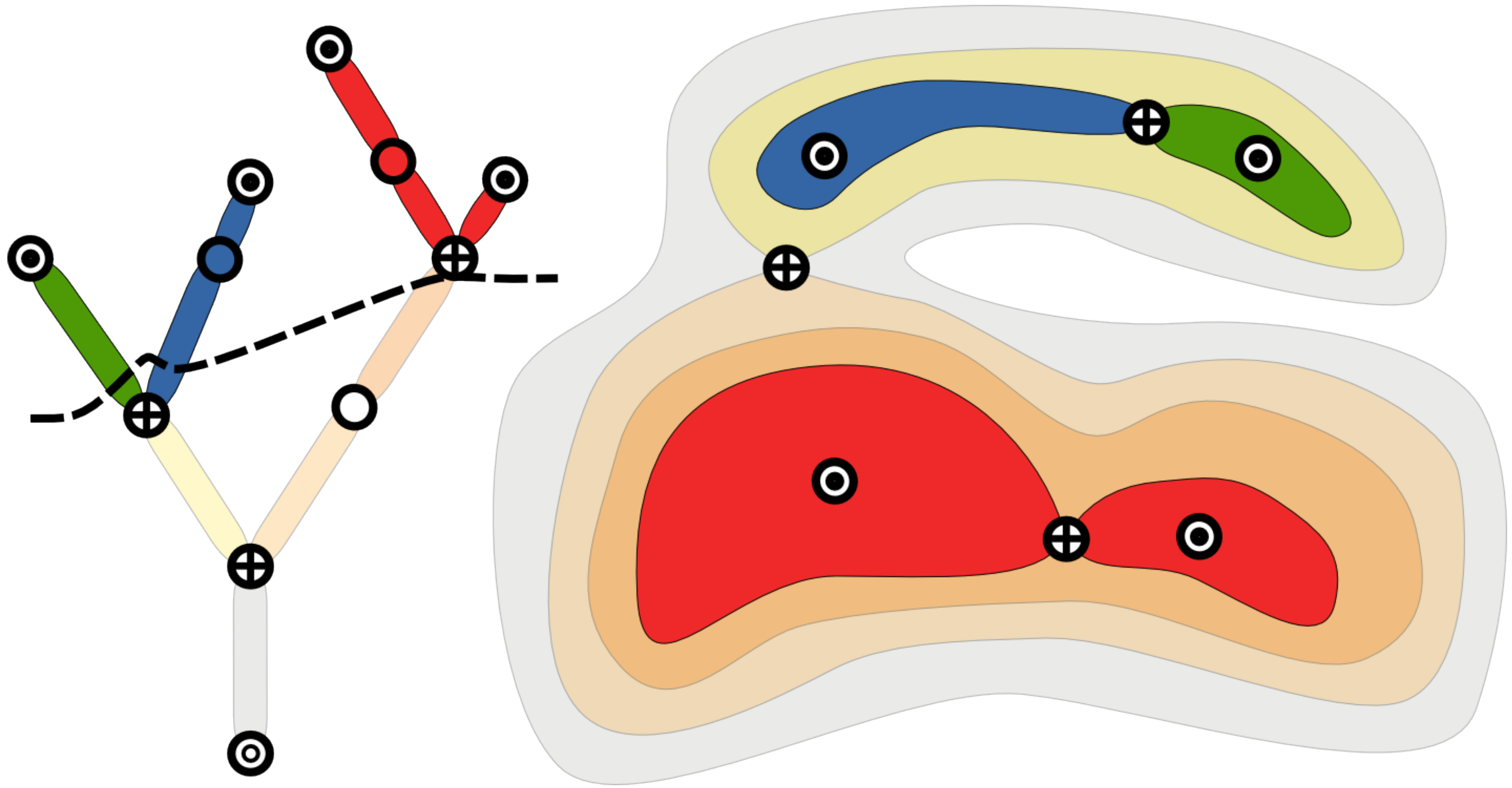
The merge tree segments a domain according to a function's level-set behavior



The resolution of the parameter space is increased by splitting long braches

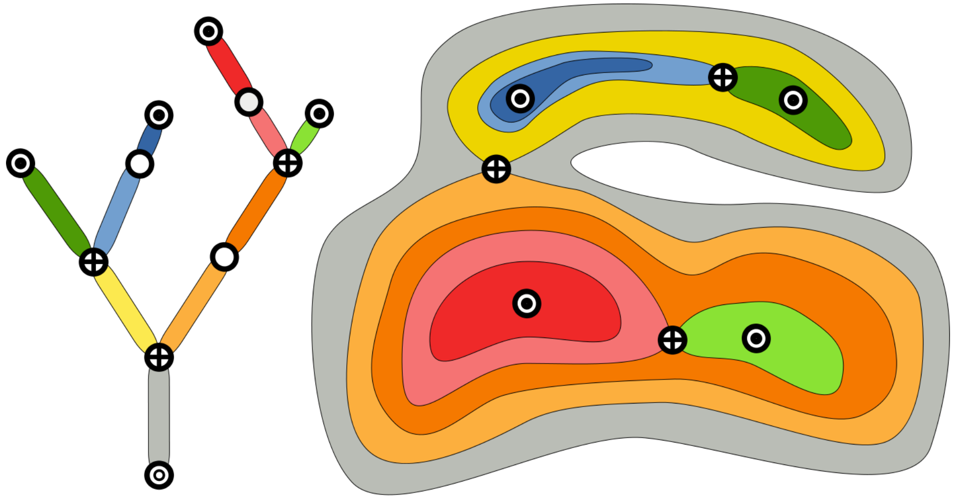


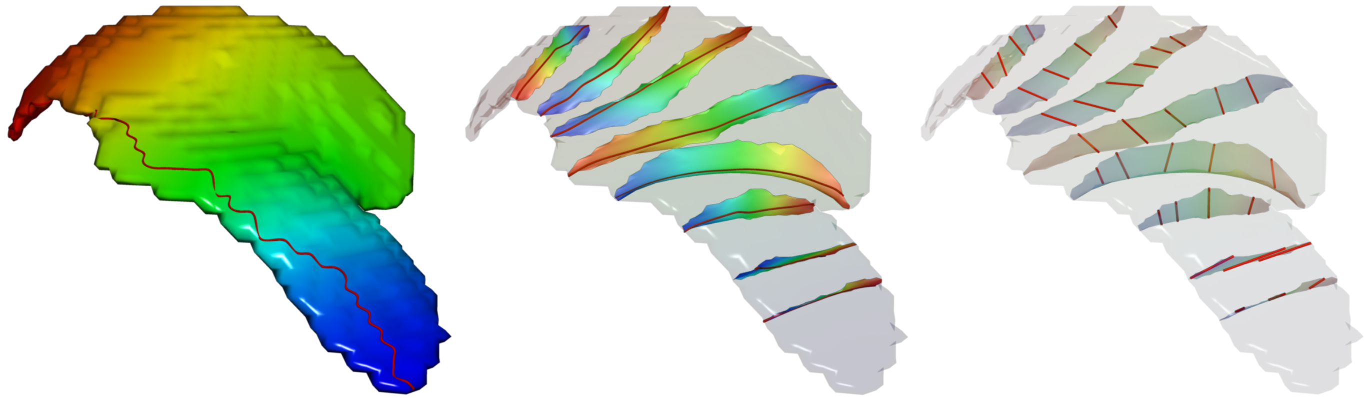
A relevance-based persistence measure is used to explore the augmented feature family



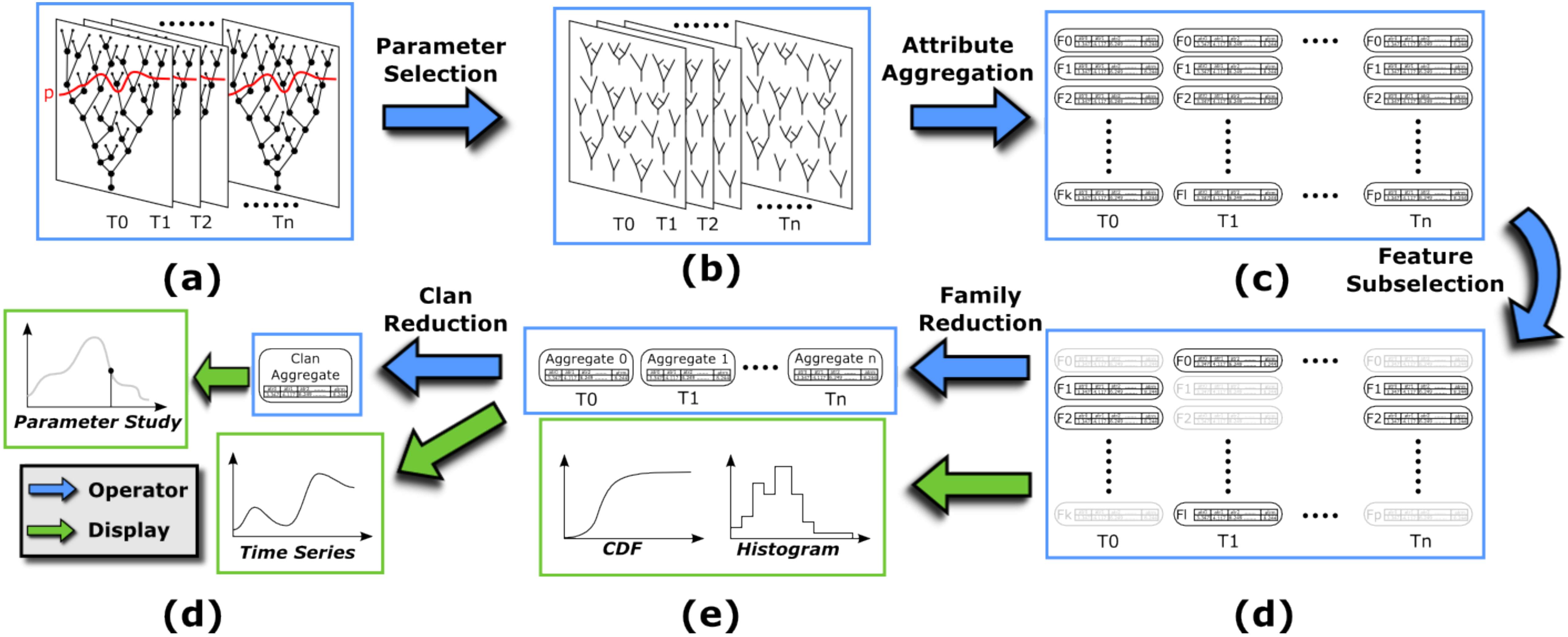
We aggregate feature-based statistics of interest & encode meta-data in a modular, extendable file format

- Descriptive statistics
  - Min/max
  - 1^{st}-4^{th} order moments
  - Sums
- Various length scales
  - Computed via a spectral technique

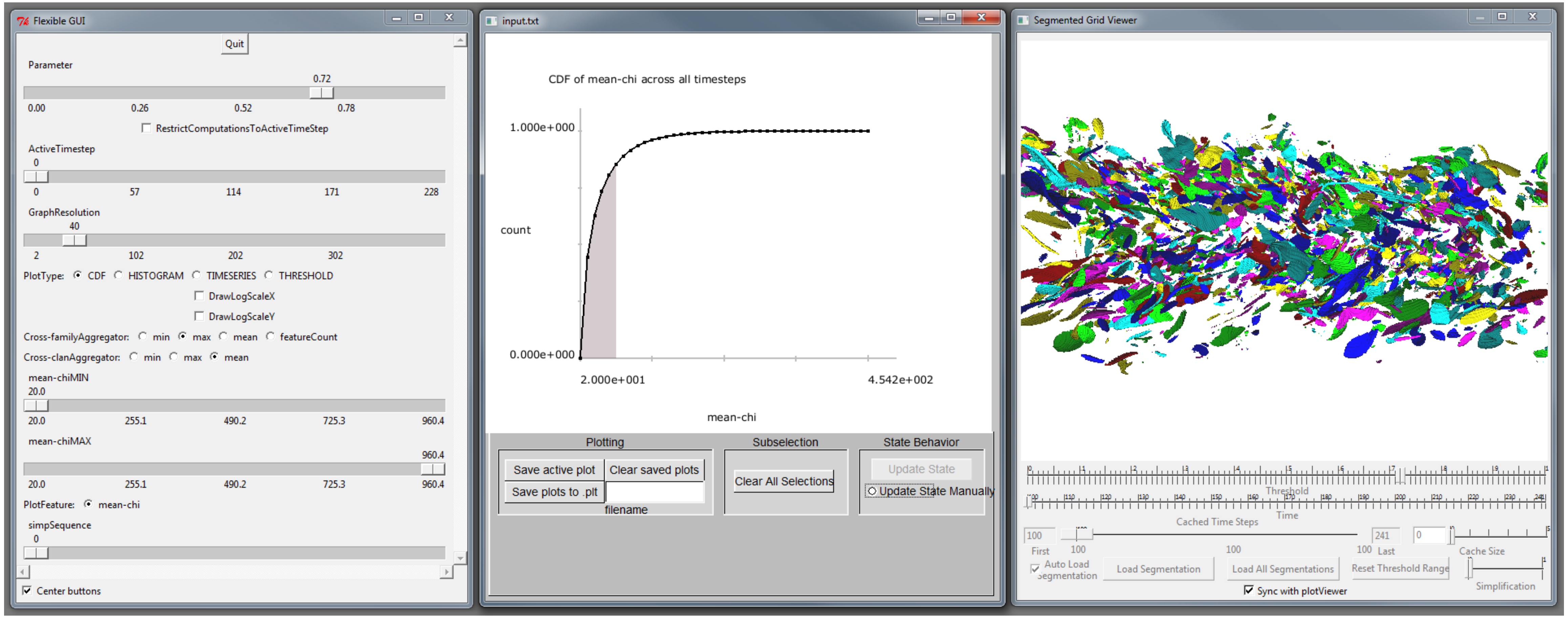




Our exploratory pipeline lets the user quickly explore a variety of statistical summaries

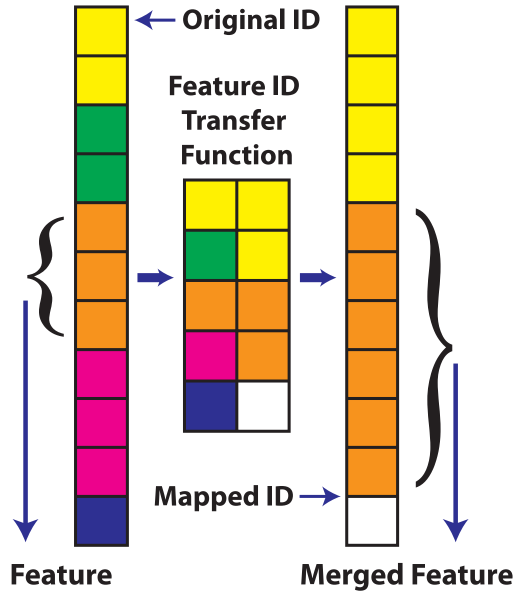


Cross-linked statistics & feature viewers provide insight into the effects of parameter selections

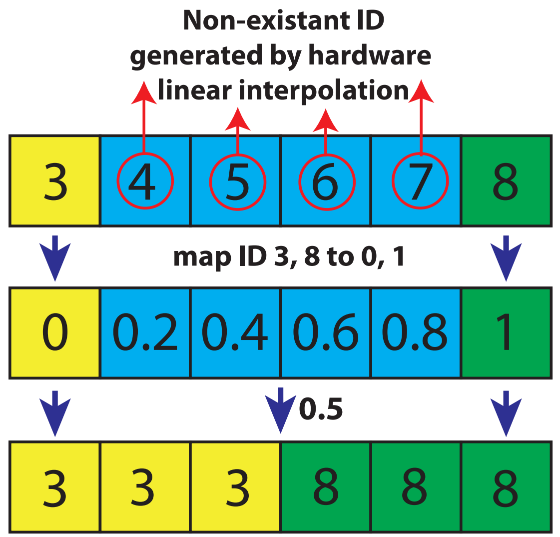


Visualizing a dynamic feature hierarchy poses challenges

- Feature: collection of elements
  - Elements store ids into regular grid
  - Binary segmented data
- Challenges
  - Identifying color of dynamically changing feature elements
  - Interpolation to smooth and light features in GPU
  - Features are dynamic



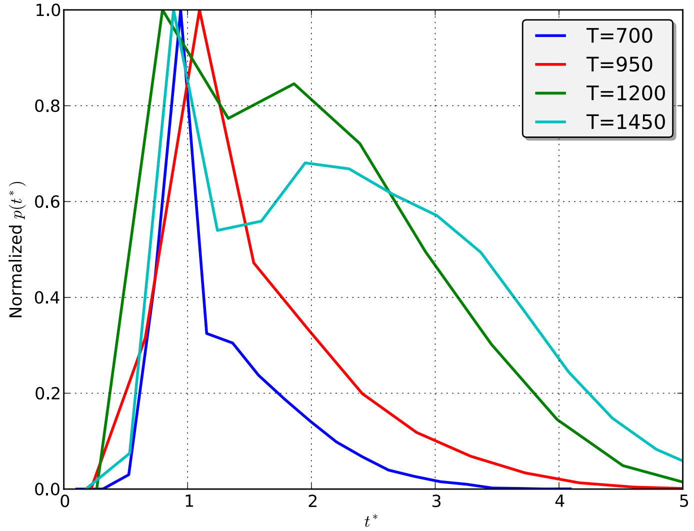
Feature transfer function mitigates cost of reloading feature id volume into GPU.



A 0-1 mapping approach [Hadwiger *et. al*] is used to address hardware linear interpolation issues.

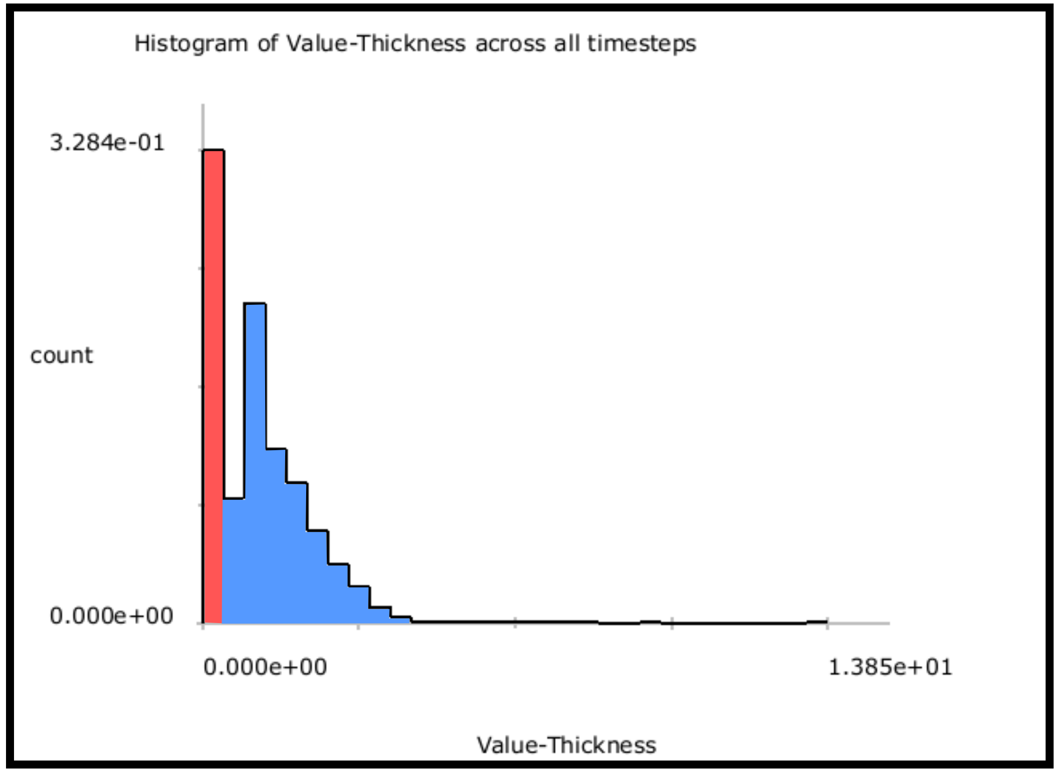
Case study results: efficient exploration of large-scale simulation data on commodity hardware

- Simulation has 0.5 billion grid points & 230 time steps
- Data reduction O(1 TB) → O(14GB)
- Building data:
  - In parallel on Lens: 32 node Linux cluster at Oak Ridge National Lab
  - Building merge tree & computing statistics: O(5 min)/time step
  - Length scales: O(90 minutes)/time step
- Exploring data:
  - Commodity hardware
  - Species distribution plots/time series: O(1 second)
  - Parameter studies: O(35 seconds)
  - Feature browser 12-25 frames/second

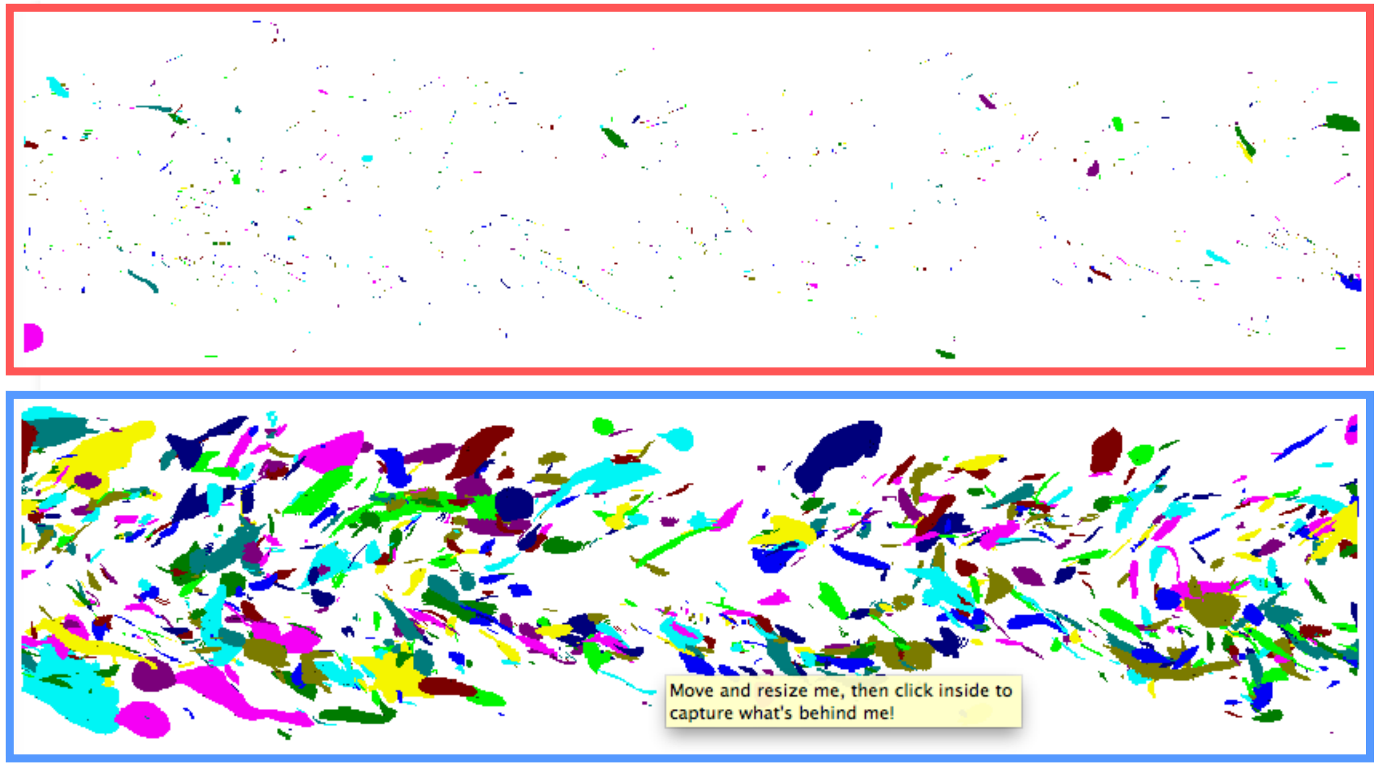


Distribution of χ thickness at relevance 0.85

Using our framework scientists can quickly diagnose issues with their analysis



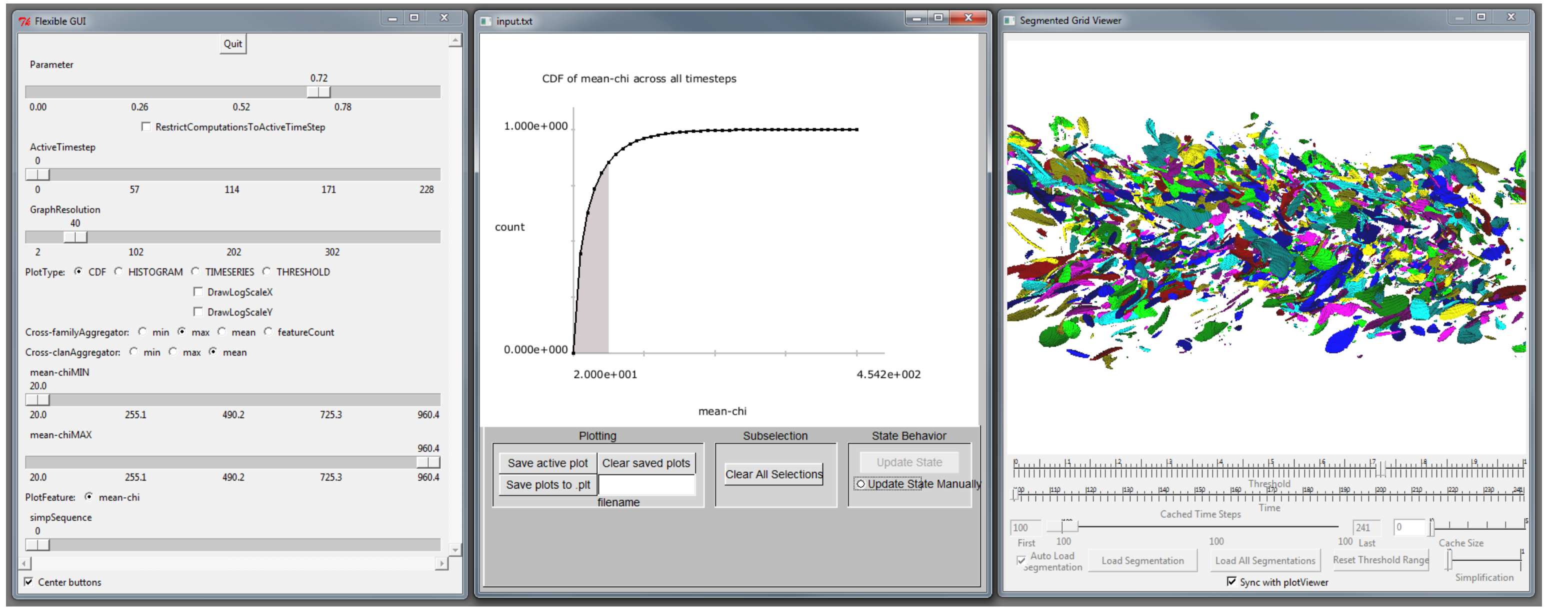
(a)



(b)

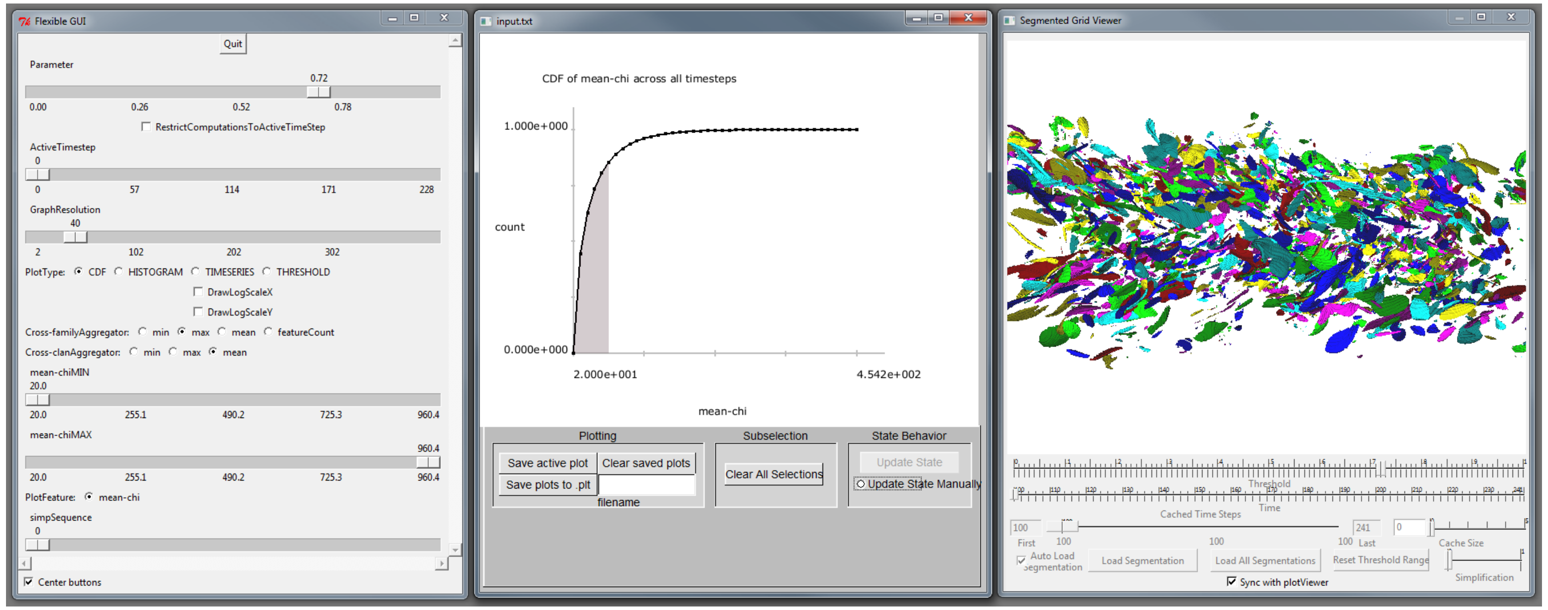
Conclusion

- Compact meta-data
  - Drastic data reductions
  - Maintains statistics of interest
  - Feature thresholds need not be known *a priori*
- Interactive linked view data exploration
  - Picking & highlighting
  - Runs on commodity hardware



Future Work

- Parallelize to support extreme-scale data
- Support additional reduction operators
- Support for alternate hierarchies



Questions?

Contact:

Janine Bennett

Sandia National Laboratories

jcbenne@sandia.gov

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

This work was performed under the auspices of the US Department of Energy (DOE) by the Lawrence Livermore National Laboratory under Contract Nos. DE-AC52- 07NA27344, LLNL-JRNL-412904L.