Interactive Visual Analysis of Very Large Data

Tutorial: Interactive Visual Analysis of Scientific Data

Gunther H. Weber

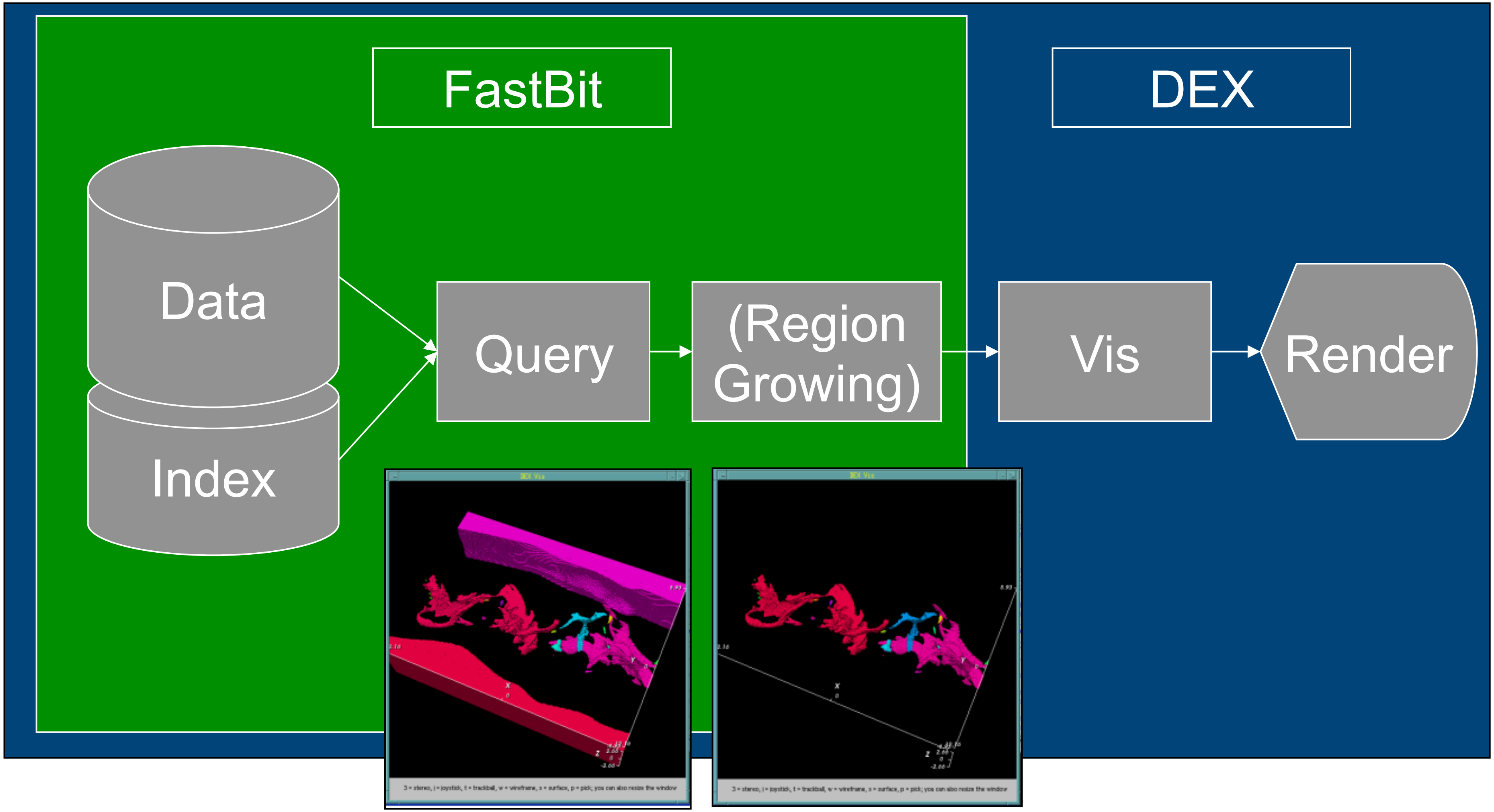
Motivation

- Apply interactive visual analysis to high performance computing (HPC) simulation results
- Example: Simulation of laser wakefield particle acceleration
  - 51 time steps
  - ~177 million particles per time step with 7 attributes (id, x, y, z, px, py, pz) ➔ 9.3GB per time step
  - Simulation performed in 2007/8, compute power and simulation result sizes continue to grow
- Due to data size interactive exploration impossible even for simple plots and operations
  - Parallel coordinates with 177 million lines?
  - Visual clutter makes results difficult to interpret

Lessons from Query-driven Visualization

- What is Query-Driven Visualization?
  - Find "interesting data" and limit visualization, analysis, machine and cognitive processing to that subset.
- One way to define "interesting" is with compound boolean range queries.
  - E.g., (CH_4 > 0.1) AND ($T_1 < temp < T_2$)
- Use index to quickly locate "interesting" data that
- Pass results along to visualization and analysis pipeline.
- Related to interactive visual analysis (consider query as brush), but queries often known *a priori*

Query-Driven Visualization



K. Stockinger, J. Shalf, K. Wu, W. Bethel. Query-Driven Visualization of Large Data Sets. In *Proceedings of IEEE Visualization 2005*, pp. 167-174. Minneapolis, MN., October 23-28, 2005.

Basic Bitmap Indexing

- First commercial version
  - Model 204, P. O'Neil, 1987
- Easy to build
  - Faster than building B-trees
- Efficient for querying: only bitwise
- logical operations
  - $A < 2 \rightarrow b_0$ OR b_1
  - $A > 2 \rightarrow b_3$ OR b_4 OR b_5
- Efficient for multi-dimensional queries
  - Use bitwise operations to combine the partial results
- Size: one bit per distinct value per row
  - Definition: Cardinality == number of distinct values
  - Compact for low (< 100) cardinality
  - Worst case: cardinality = N, index size: N*N bits

Data values	b_0 =0	b_1 =1	b_2 =2	b_3 =3	b_4 =4	b_5 =5
0	1	0	0	0	0	0
1	0	1	0	0	0	0
5	0	0	0	0	0	1
3	0	0	0	1	0	0
1	0	1	0	0	0	0
2	0	0	1	0	0	0
0	1	0	0	0	0	0
4	0	0	0	0	1	0
1	0	1	0	0	0	0

A < 2 (columns b_0, b_1)   2 < A (columns b_3–b_5)

Range Based Queries – FastBit

- Bitmap indexes
  - Sacrifice update efficiency to gain more search efficiency
  - Efficient for multi-dimensional queries (parallelizable)
  - Scale linearly with the dimension of a query
- Bitmap indexes may demand too much space
- FastBit solves the space problem by developing an efficient compression method that
  - Reduces index size, typically **30%** of raw data, compared to 300% for some common indexes
  - Improves operational efficiency
  - **10X speedup** relative to best known compressed bitmap index
  - Even higher speedup relative to conventional indexes

Interactive Visualization of Laser Wakefield Particle Accelerator Simulations

O. Rübel, Prabhat, K. Wu, H.R. Childs, J.S. Meredith, C.G.R. Geddes, E. Cormier-Michel, S. Ahern, G.H. Weber, P. Messmer, H. Hagen, B. Hamann and E.W. Bethel: *High Performance Multivariate Visual Data Exploration for Extremely Large Data.* In: Proc. Supercomputing SC08, Austin, TX, USA, Nov. (2008)

Laser Wakefield Particle Acceleration

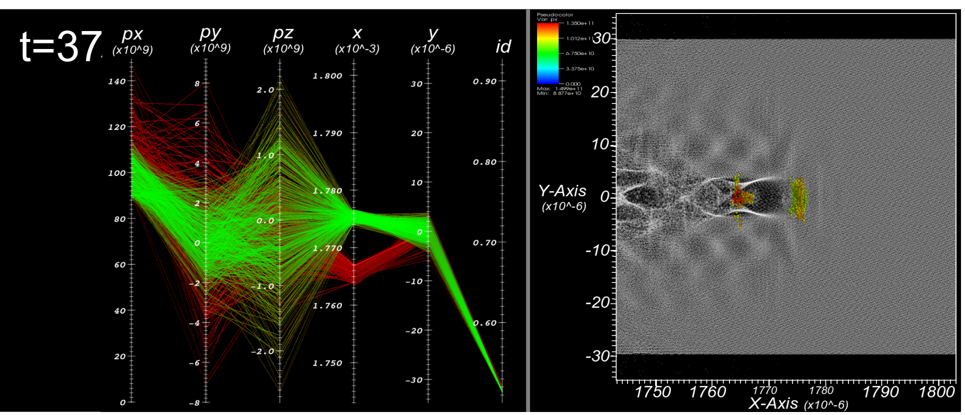
Advantage: Electric fields thousands of times stronger than in conventional accelerators → High acceleration in short distance

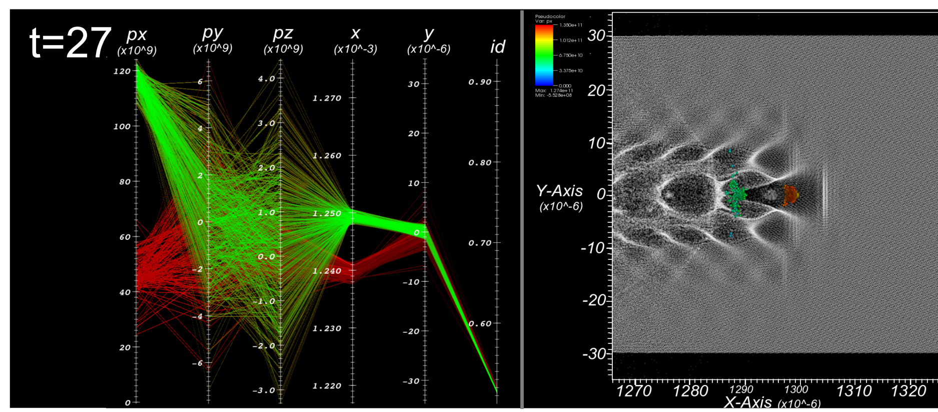
C.G.R. Geddes, C. Toth, J. van Tilborg, E. Esarey, C. Schroeder, D. Bruhwiler, C. Nieter, J. Cary, and W. Leemans. High-Quality Electron Beams from a Laser Wakefield Accelerator using Plasma-Channel Guiding, *Nature,* 438: 538-541, 2004

Analysis Task(s)

- Identify particles forming a beam
  - Interactive visual data exploration
  - Data sub-setting

- Track particles over time
  - Given particle IDs from a given time step,
  - Find those particles in all time steps
  - Subsequent visual data analysis



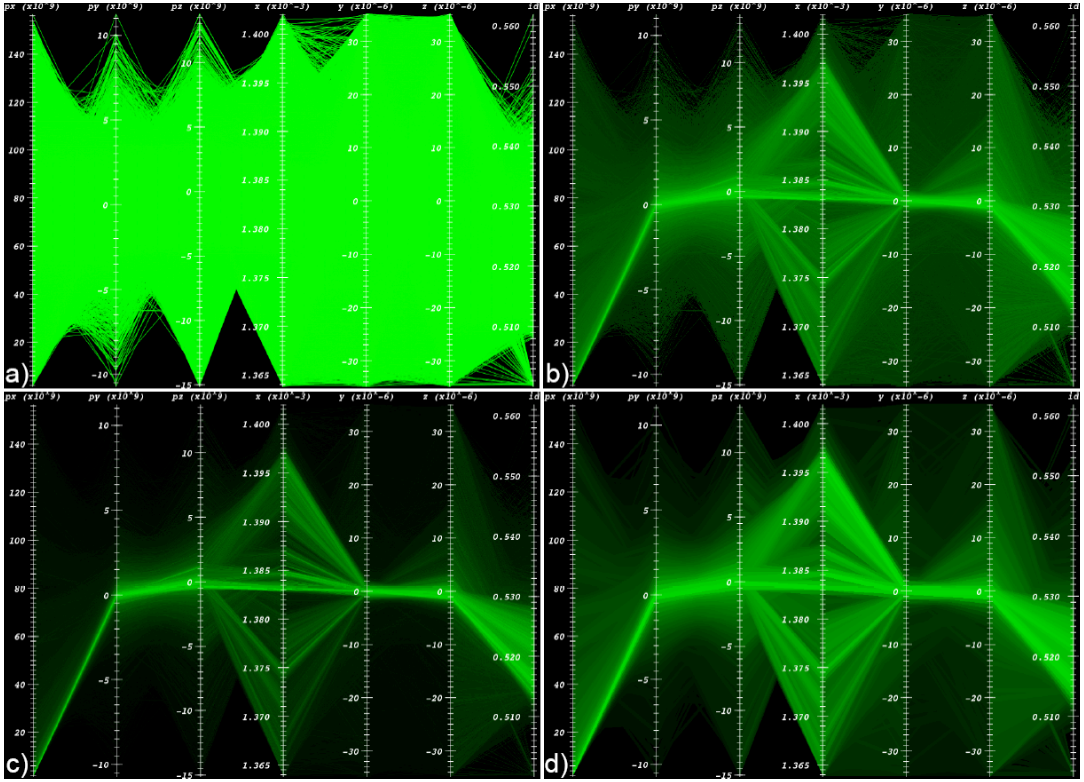


Fundamental Problem #1 – Interface

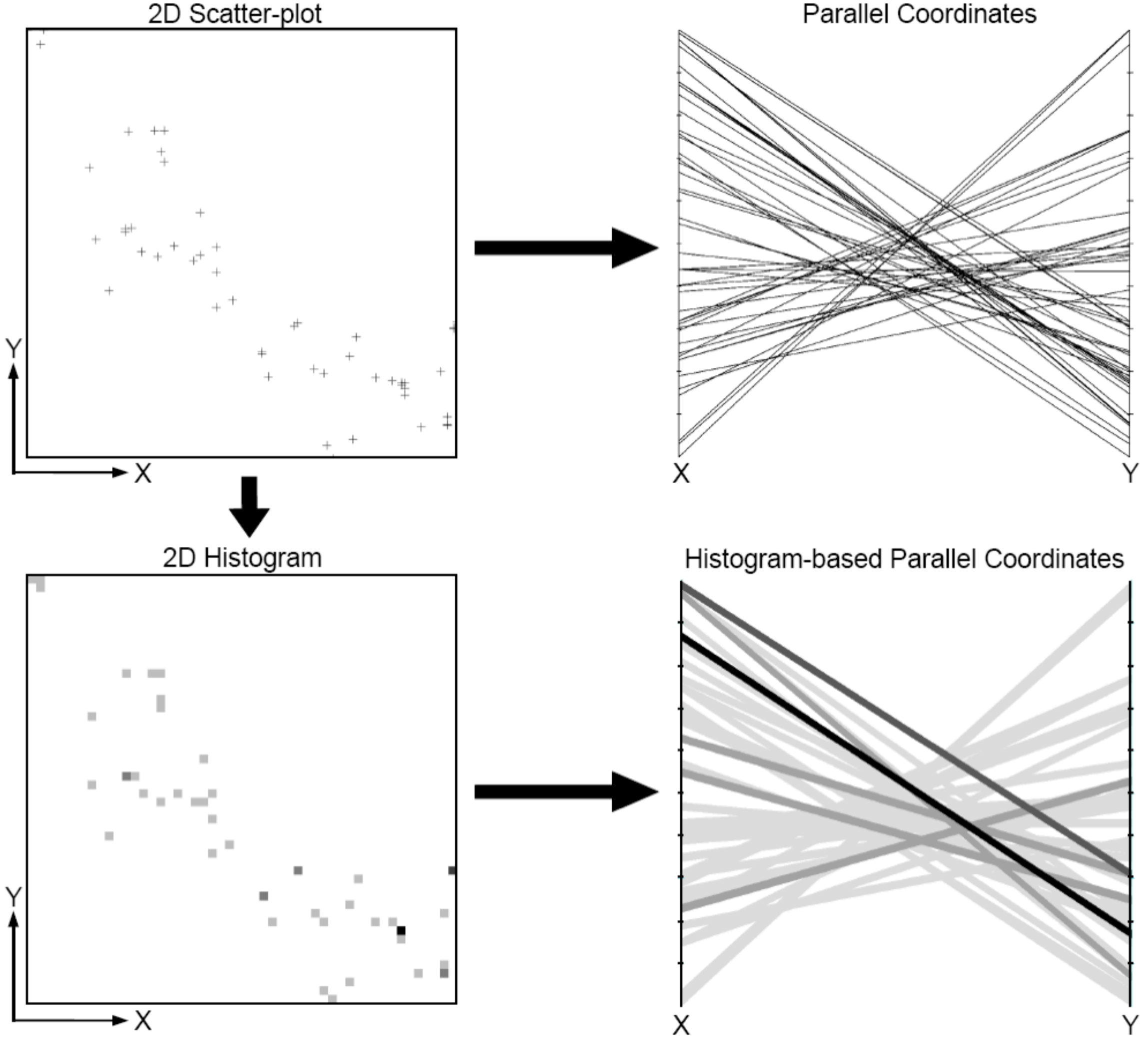
- Parallel coordinates
  - Mechanism for displaying multivariate data.
  - Interface for subset selection

- Problems with large data
  - Visual clutter
  - O(n) complexity

- Solution
  - Histogram-based parallel coordinates



Histogram-Based Parallel Coordinates

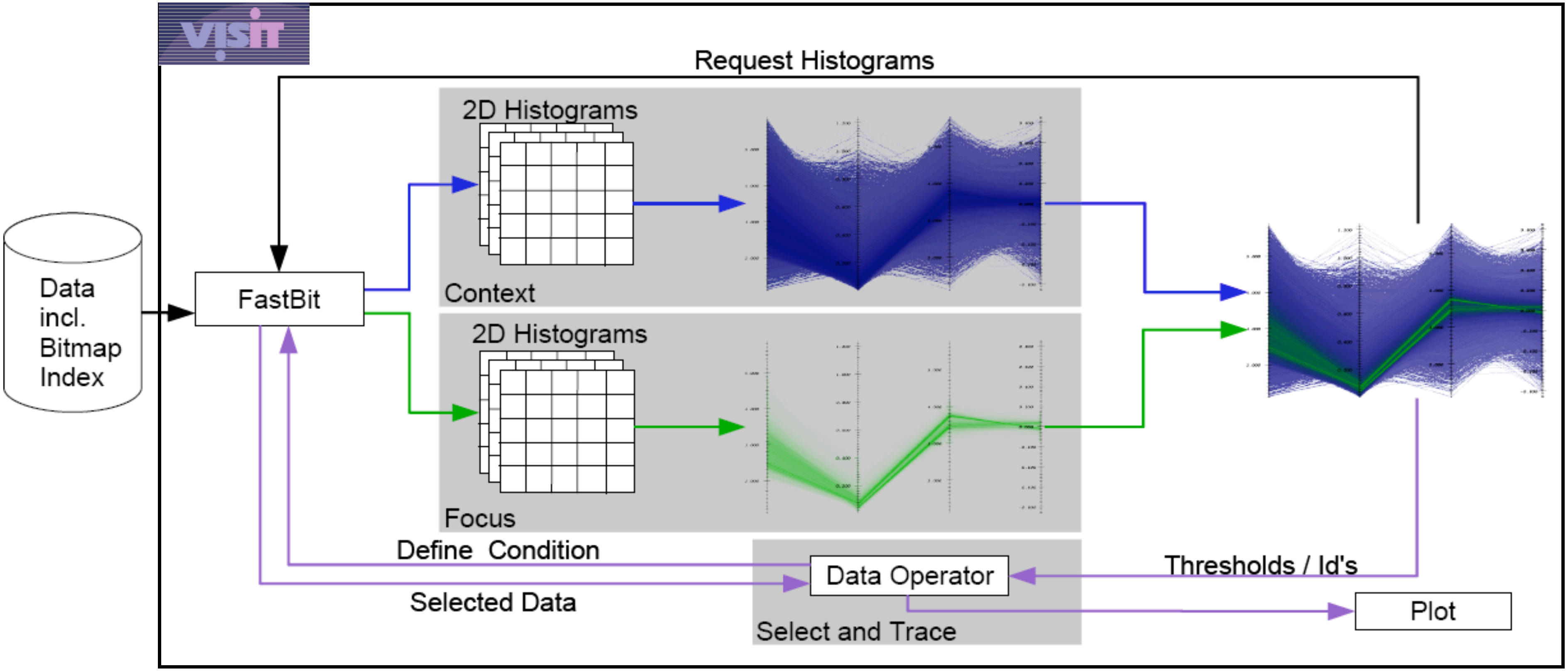


M. Novotny and H. Hauser, *Outlier-preserving Focus+Context Visualization in Parallel Coordinates*, IEEE TVCG, 12(5): 893–900 (2006)

Fundamental Problem #2 – Performance

- How to efficiently construct a histogram?
  - Naïve approach: O(n)
  - Better approach: "cheat" (use FastBit)
- How to efficiently do particle tracking?
  - Naïve approach: $O(n^2)$
  - Better approach: O(H*t) (use FastBit)

System Design



VisIt is available at https://wci.llnl.gov/codes/visit/

FastBit is available at https://codeforge.lbl.gov/projects/fastbit

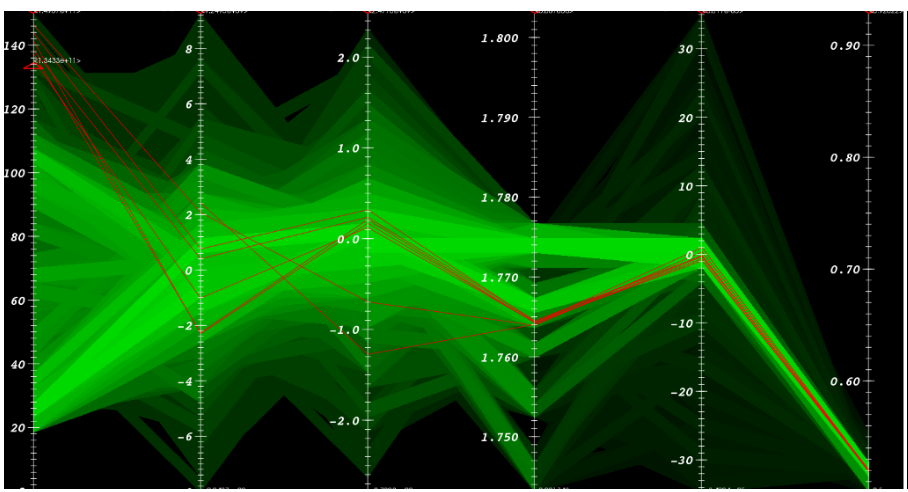
Histogram-based Parallel Coordinates

Histograms computed on request:

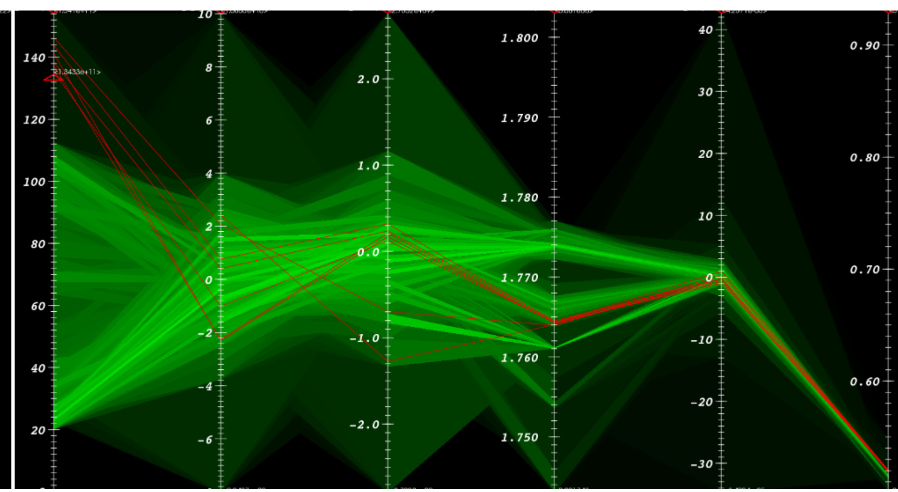
- Rendering of data subsets using histogram-based parallel coordinates
- Rendering with arbitrary number of bins
- Close zoom-ins and smooth drill-downs into the data

Adaptively binned histograms:

- More accurate data representation in lower-level-of-detail views

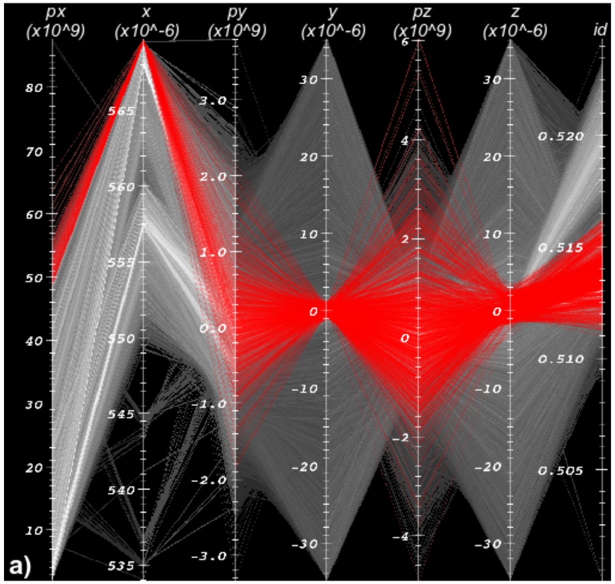


32x32 uniform binning

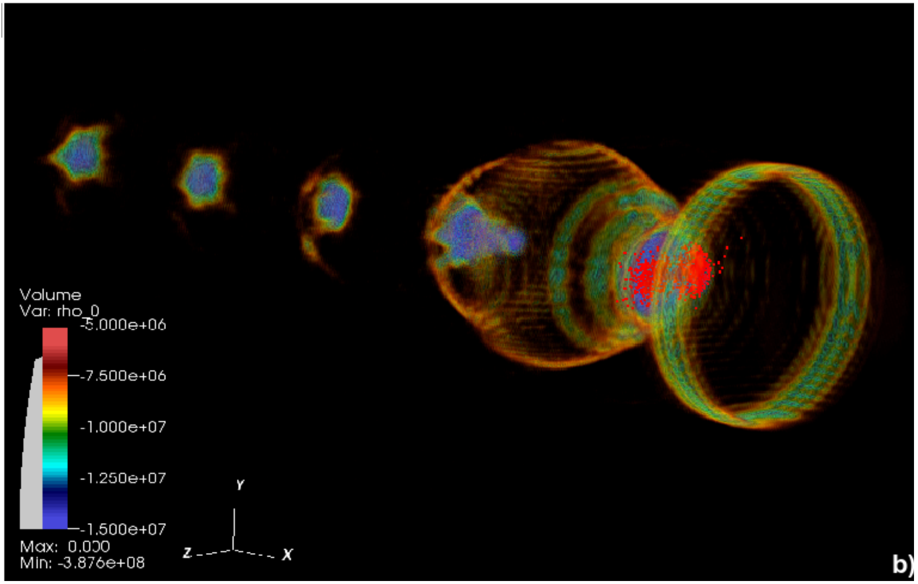


32x32 adaptive binning

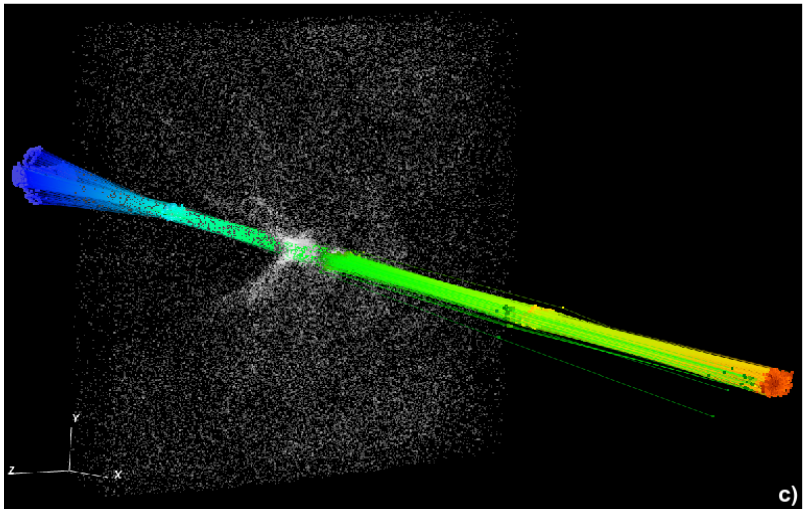
3D Analysis Example



Selecting particles of interest



Selected particles (red); volume rendering of plasma density



Traces of the the selected particle-bunch

Data Overview

- **Simulation:** VORPAL, 2D and 3D
- **Particle data** (scattered data):
  - x,y,z (location), px, py, pz (momentum), id (particle identifier)
  - Number of particles per timestep:
    - ~ $0.4*10^6$ – $30*10^6$ (in 2D)
    - ~$80*10^6$– **$200*10^6$ (in 3D)**
  - Total size:
    - ~1.5GB – >30GB (in 2D)
    - ~100GB – **>1TB (in 3D)**
- **Field data** (defined on regular grid):
  - Electric field, magnetic field, and RhoJ
  - Resolution: ~0.02-0.03µm longitudinally, and ~ 0.1-0.2µm transversely
  - Total size:
    - ~3.5GB - >70GB (in 2D)
    - ~200GB - **>2TB (in 3D)**

Cameron G.R. Geddes, ”Plasma Channel Guided Laser Wakefield Accelerator,” PhD-thesis, UC Berkeley, 2005

C. Nieter and J. R. Cary, “VORPAL: A Versatile Plasma Simulation Code,” J. Comput. Phys., 196(2):448–473, 2004

VisWeek 2012
SCIVIS • INFOVIS • VAST
BIOVIS • LDAV

Queries over Time: Interactive Visualization of Magnetic Fusion

A.R. Sanderson, B. Whitlock, O. Rübel, H.R. Childs, G.H. Weber, Prabhat, and K. Wu: *A System for Query Based Analysis and Visualization.* Proc. EuroVA 2012, Vienna, Austria, June 2012, pp. 25–31 (2012).

UNIVERSITAS BERGENSIS

BERKELEY LAB

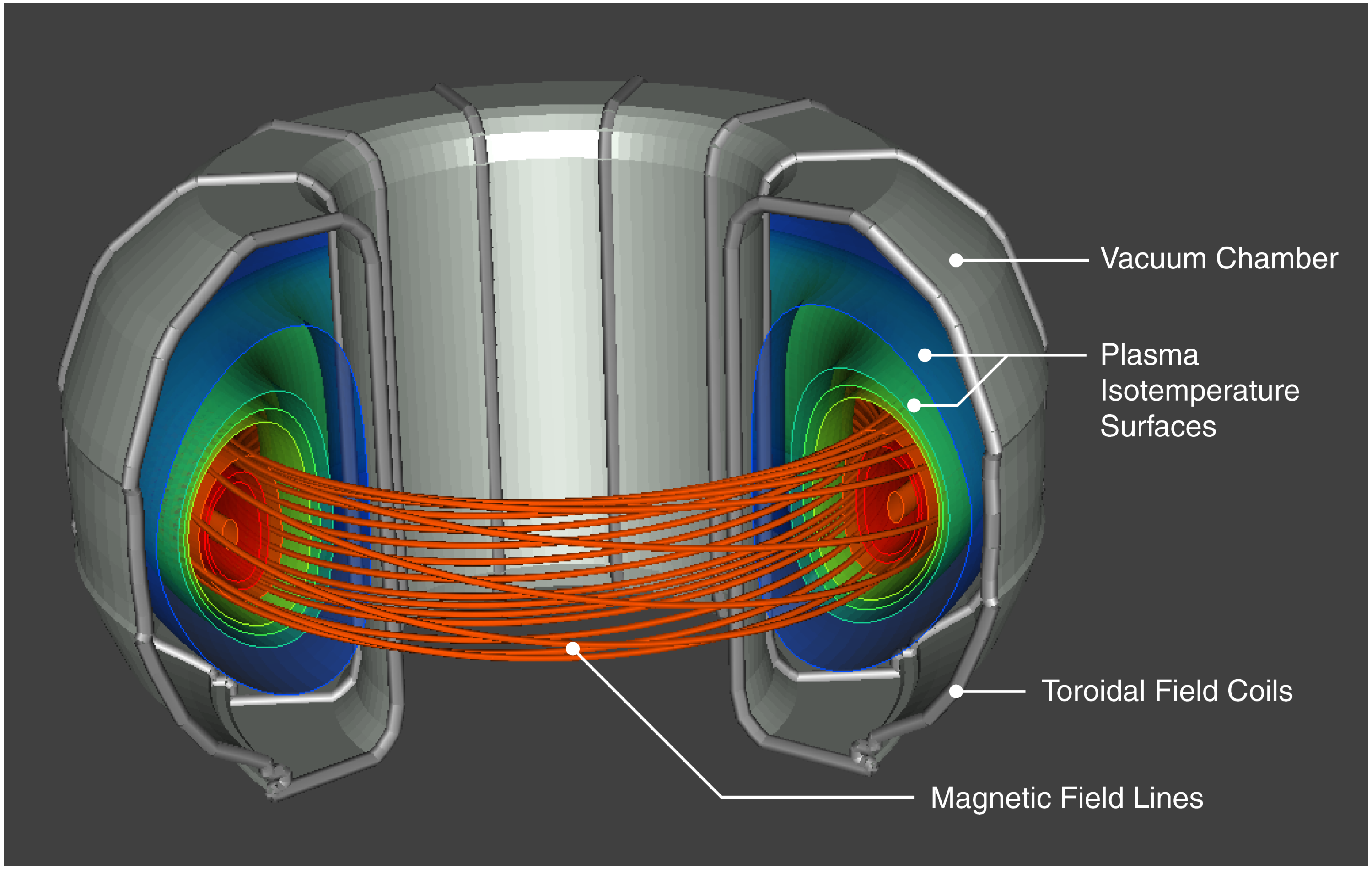
CD-adapco

VIS

Application: Magnetic Fusion

- Uses magnetic fields to confine the plasma which attempts to fuse light particles into heavier particles which then gives off energy,

- $E = mc^2$

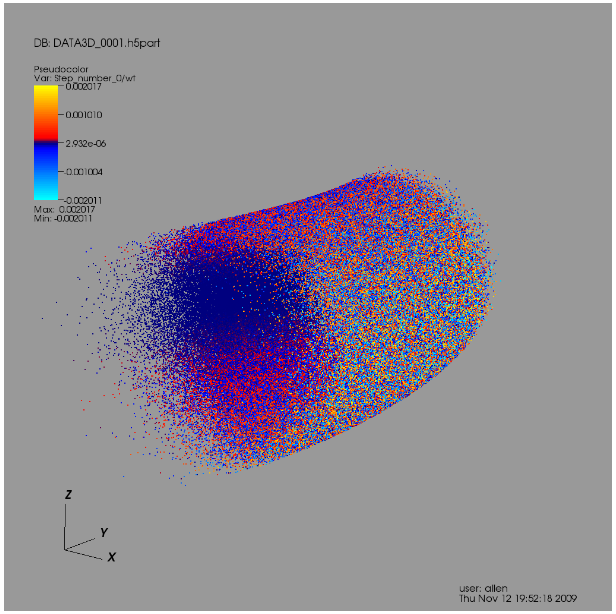


Application: Magnetic Fusion

- Turbulence in plasma studied via simulation of millions to billions of particles using Particle in Cell codes

- Visualizing large number of particles while interesting from a graphics point of view yields little domain knowledge

- More important to application scientists is finding anomalous particles and understanding mechanism for their radial diffusion



Analysis Goals

- Perform range based queries on large number of multivariate entries on an interactive basis.
- Range-based queries expressed in the context of threshold ranges, i.e. $101 \leq x \leq 205$
- Identify temporal features via intra- or inter-time step queries, e.g., ($wt_t > 0$) AND ($trapped_t$!= $trapped_{t+1}$)
- Accumulate results over all time steps, a.k.a. *cumulative queries*
- Refine results of cumulative queries

Single Time Step Queries

- Two types of queries:
  - **Range-based** (“create brush”):
    - Threshold ranges, i.e. $101 \leq x \leq 205$.
    - Logical combinations for multi-dimensional
    - Results are ID-based
  - **ID-based** (“store brush”):
    - Stored and managed by a central selection manager
    - Link multiple visualizations
    - Track selected data subsets over time.

- Each particle required to have unique identifier (ID)

Cumulative Queries

- Identification of temporal features that cannot be seen within single discrete time steps
- Intra- and inter-time step queries:

$(wt_t > 0)$ AND $(trapped_t \mathrel{!=} trapped_{t+1})$

intra-time *inter-time*

- Find all particles:

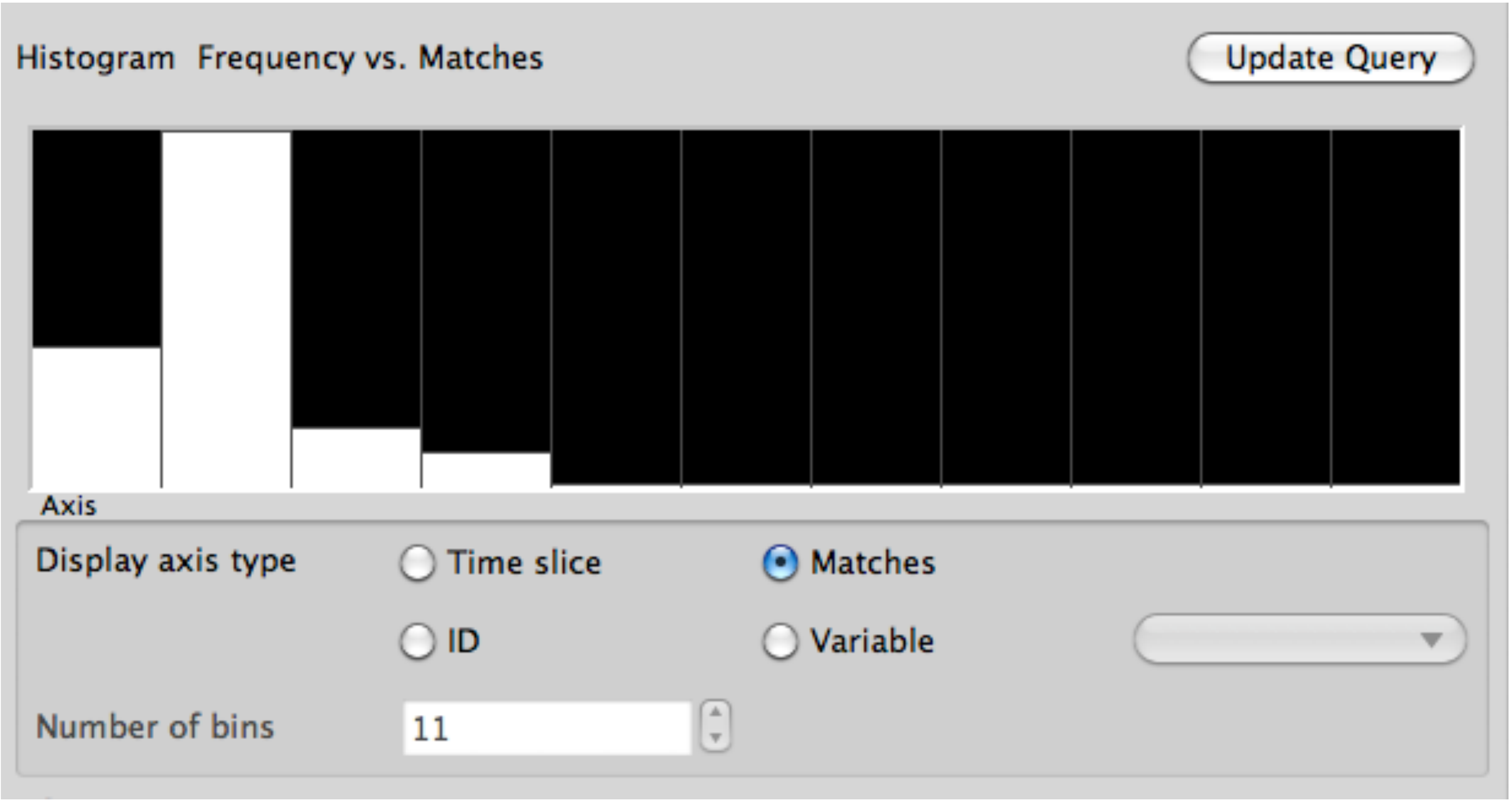
With weight greater than zero over all time steps

AND

Change state from trapped to passing

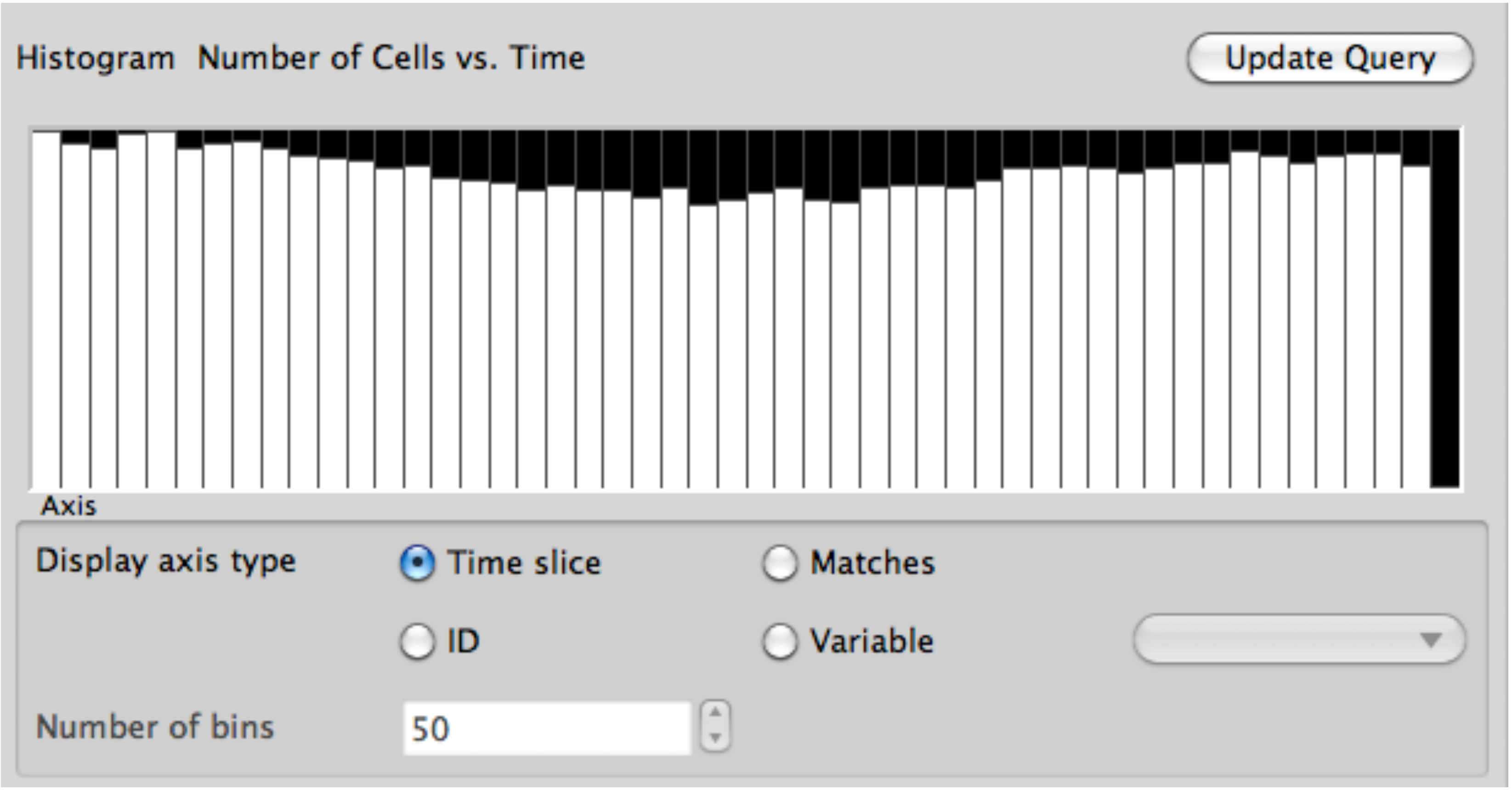
Cumulative Queries – Frequency of Matches

- Example 50 time steps with 500k particles:
  - Frequency of matches (1-11) in all time steps



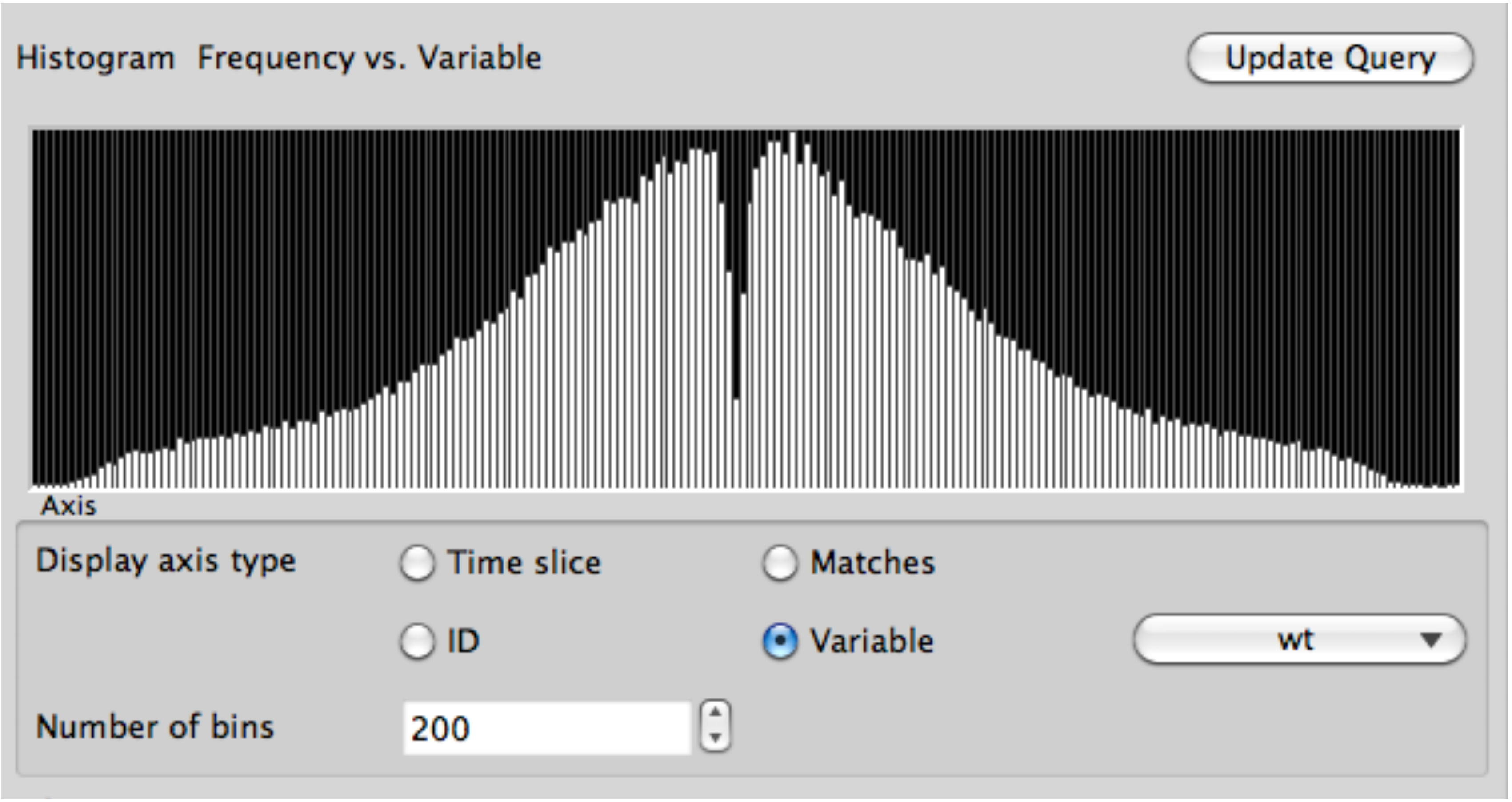
Cumulative Queries – Matches over Time

- Example 50 time steps with 500k particles:
  - Frequency of matches per time step



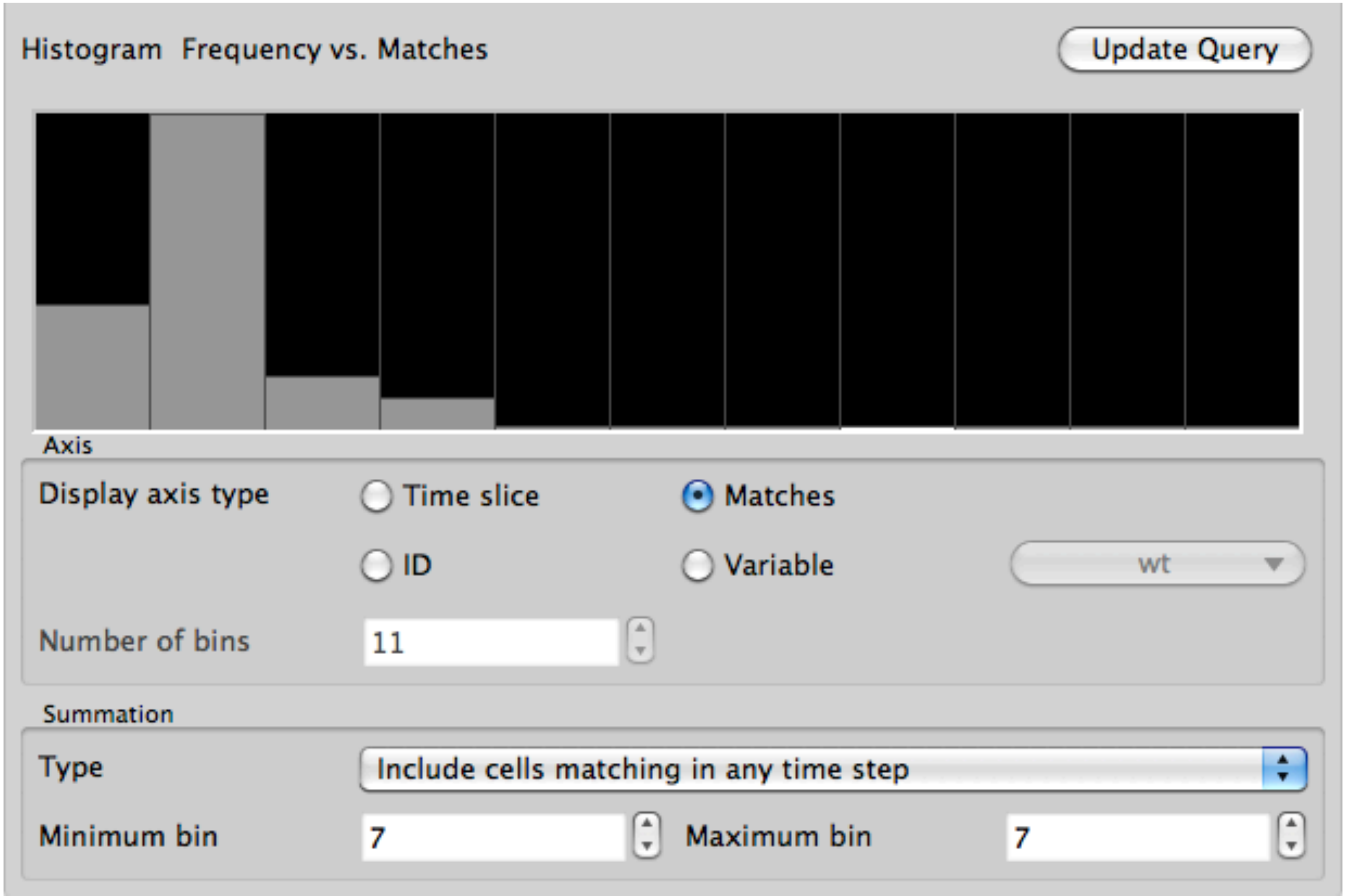
Cumulative Queries – Matches over Time

- Example 50 time steps with 500k particles:
  - Frequency of matches using variable



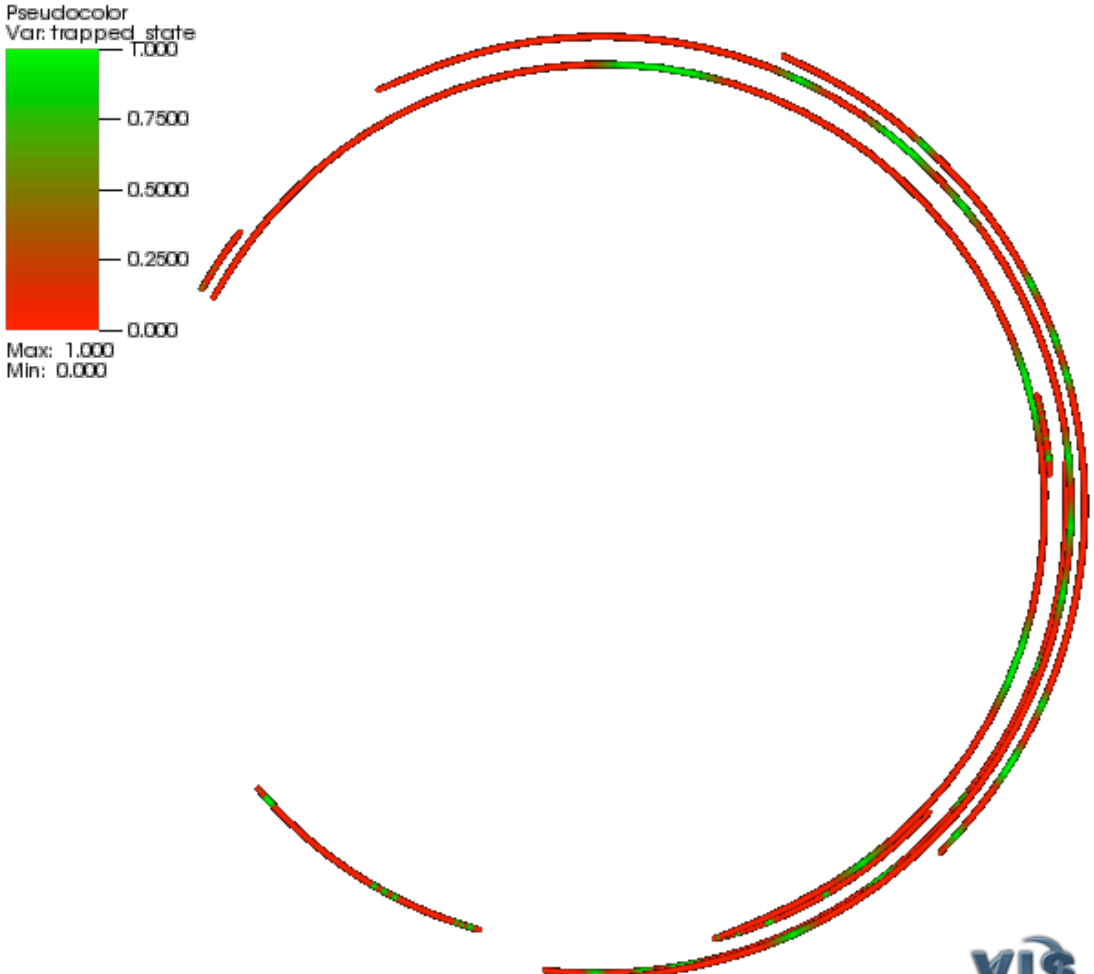
Cumulative Queries

- Sub-select bin 7 (i.e., the 8 particles matching in 7 time steps)



Cumulative Queries – Particle Paths

- Sub-selection bin 7 (8 out of 500k particles that matched in 7 time steps)
- Construct path from 50 time steps



Literature

- Allen R. Sanderson, Brad Whitlock, Oliver Rübel, Hank Childs, Gunther H. Weber, Prabhat and Kesheng Wu. A System for Query Based Analysis and Visualization, Third International Eurovis Workshop on Visual Analytics EuroVA 2012, K. Matkovic and G. Santucci (Editors), Vienna, Austria, June, 2012, LBNL-5507E.
- O. Rübel, Prabhat, K. Wu, H. Childs, J. Meredith, C.G.R. Geddes, E. Cormier-Michel, S. Ahern, G.H. Weber, P. Messmer, H. Hagen, B. Hamann and E.W. Bethel, "High Performance Multivariate Visual Data Exploration for Extremely Large Data." SC08, Austin TX, November, 2008. LBNL-716E.
- Oliver Rübel, Sean Ahern, E. Wes Bethel, Mark. D Biggin, Hank Childs, Estelle Cormier-Michel, Angela DePace, Michael B. Eisen, Charless C. Fowlkes, Cameron G. R. Geddes, Hans Hagen, Bernd Hamann, Min-Yu Huang, Soile V. E. Keränen, David W. Knowles, Cris L. Luengo Hendriks, Jitendra Malik, Jeremy Meredith, Peter Messmer, Prabhat, Daniela Ushizima, Gunther H. Weber, and Kesheng Wu. Coupling Visualization and Data Analysis for Knowledge Discovery from Multi-dimensional Scientific Data. In *Procedia Computer Science, Proceedings of International Conference on Computational Science, ICCS 2010*, May 2010. LBNL-3669E.
- K. Stockinger, J. Shalf, K. Wu, W. Bethel. "Query-Driven Visualization of Large Data Sets." In Proceedings of IEEE Visualization 2005, pp. 167-174. Minneapolis, MN., October 23-28, 2005. LBNL-57511.
- Kesheng Wu, Ekow Otoo, and Arie Shoshani. *Optimizing bitmap indices with efficient compression* ACM Transactions on Database Systems, v 31, pages 1-38, 2006.