

Query-Driven Visualization Accelerates Scientific Insight

12 October 2007
Computation and Complex Systems
MSRI, Berkeley CA, USA

E. Wes Bethel
<http://vis.lbl.gov/>
Lawrence Berkeley National Laboratory



The Take-Home Message

- ❖ Gaining insight from large and complex datasets is a major bottleneck.
- ❖ Thesis: effective software tools for gaining insight from such data must use a blend of technologies from:
 - Scientific Data Management
 - Visualization, Analytics
 - Computer Science
- ❖ Query-Driven Visualization is one such approach.



Outline

- ❖ Motivation and Background
- ❖ Query-Driven Visualization and Analytics
- ❖ Query-Driven Visualization Interfaces
- ❖ Case study: Cybersecurity data analysis



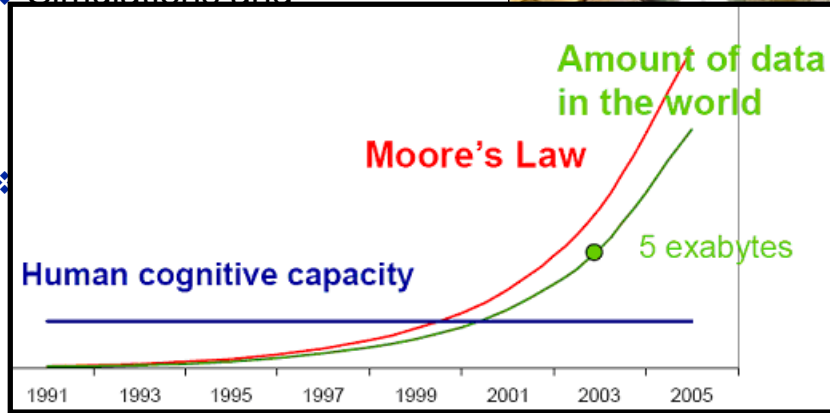
Motivation

- ❖ Simulations and experiments are generating data faster than it can be analyzed and understood.
- ❖ Science bottleneck: information analysis and understanding.



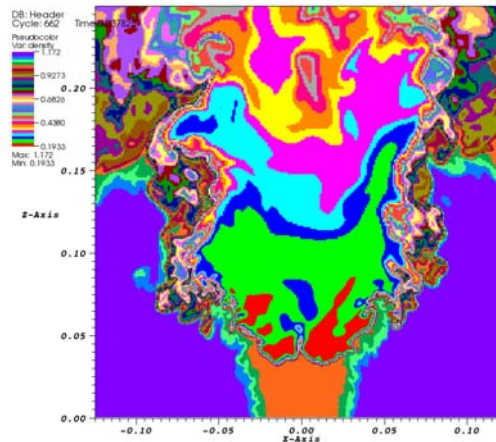
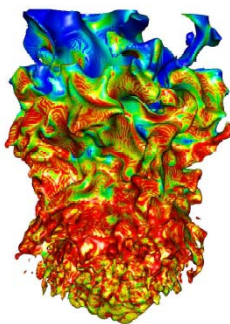
Motivation

❖ Simulations and



Cognitive Bandwidth Challenge

- ❖ Can't "see" a TB, much less a PB
 - Mesh cells map to subpixels
 - Depth complexity
 - Spatial complexity



Challenges and Approach

- ❖ Theses:
 - “Scalable visualization” solutions do not necessarily increase the likelihood of scientific insight.
 - More visualization output can cause more problems than it solves.
 - Increased depth complexity.
 - Increased cognitive workload.
 - Our approach to high performance visualization is to focus visualization processing on that subset of data deemed to be scientifically interesting.
 - Any tractable solution to “large data visualization” must address scientific data management issues.
 - Observation: 80% of the code in visualization applications does “data management.”
 - Use SDM technology for data management.



Outline

- ❖ Motivation and Background
- ❖ Query-Driven Visualization and Analytics
- ❖ Query-Driven Visualization Interfaces
- ❖ Case study: Cybersecurity data analysis

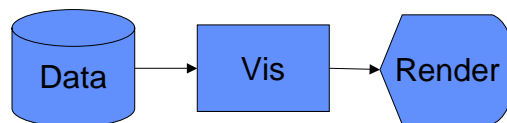


Query-Driven Visualization

- ❖ 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 $(T_1 < temp < T_2)$
- ❖ Quickly locate those data that are “interesting.”
- ❖ Pass results along to visualization and analysis pipeline.



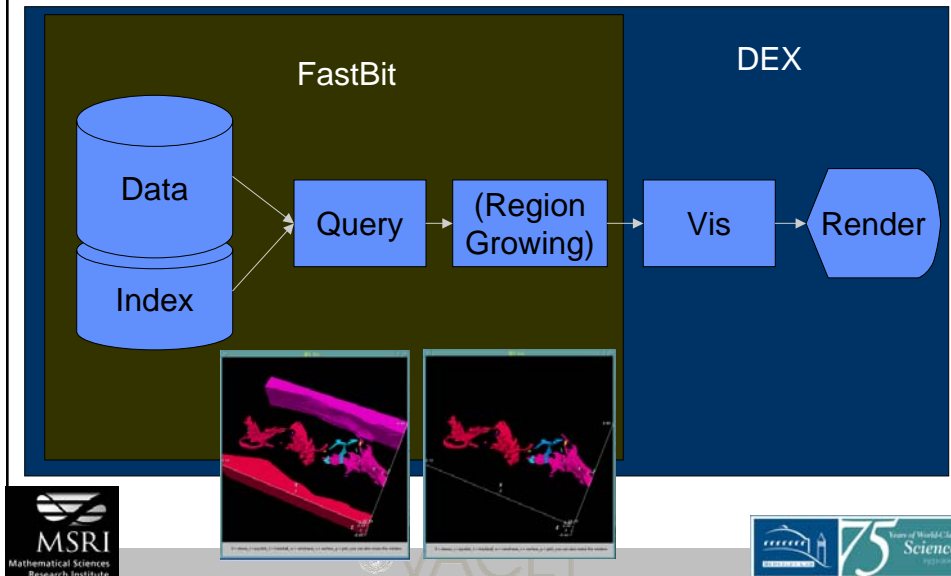
Query-Driven Visualization



The Canonical Visualization Pipeline



Query-Driven Visualization



Query-Driven Visualization

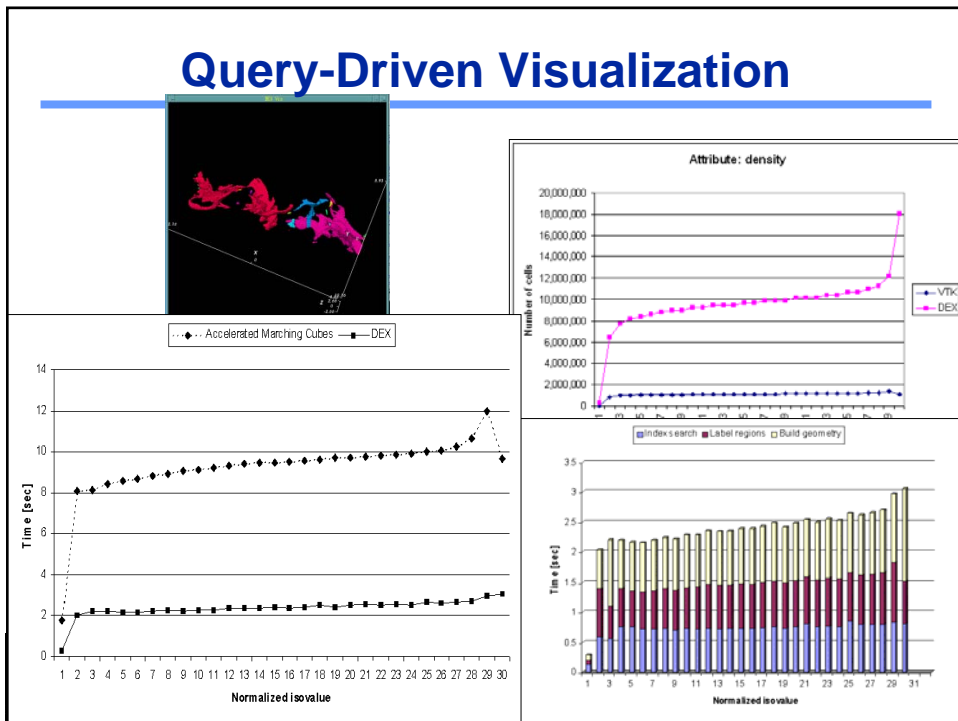
- ❖ $CH_4 > 0.3$
- ❖ $Temp < T_1$
- ❖ $CH_4 > 0.3$ AND $temp < T_1$
- ❖ $CH_4 > 0.3$ AND $temp < T_2$
 - $T_1 < T_2$

Query-Driven Visualization

- ❖ How fast is it?
 - Comparison: Isosurface algorithms:
 - Nice summary in: Sutton et. al., "A Case Study of Isosurface Extraction Algorithm Performance," *2nd Joint Eurographics-IEEE TCCG Symposium on Visualization*, May 2000
 - For n data values and k cells intersecting the surface:
 - Marching Cubes: $O(n)$
 - Octree methods: $O(k + k \log(n/k))$
 - Acceleration: pruning; sensitive to noisy data
 - Span-space methods:
 - NOISE: $O(\sqrt{n} + k)$
 - ISSUE: $O(\log(n/L) + \sqrt{n}/L + k)$
 - » L is a tunable parameter
 - Interval Tree: $O(\log n + k)$
- ❖ FastBit: $O(k)$ – the theoretical optimum.
 - Profound performance gain for Petascale visualization!



Query-Driven Visualization



Query-Driven Visualization

- ❖ What do these timing results mean?
 - In a one-sided matchup (DEX doing a lot more work), our performance results are markedly better for a given task than an industry-standard isocontouring implementation.
- ❖ These are single-valued queries.
 - FB capable of *n-dimensional* queries.
 - Tree-based indexing methods not practical for *n-dimensional* queries.
 - Trees: Curse of Dimensionality!
 - $O(N \cdot D)$ vs. $O(N^{**}D)$
- ❖ Why compare against isosurfacing?
 - Familiar to the visualization community.



Outline

- ❖ Motivation and Background
- ❖ Query-Driven Visualization and Analytics
- ❖ Query-Driven Visualization Interfaces
- ❖ Case study: Cybersecurity data analysis



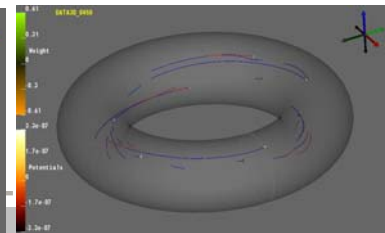
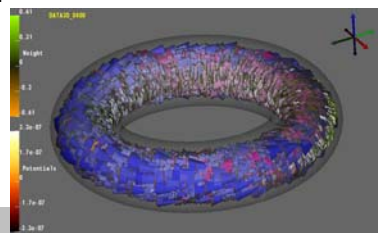
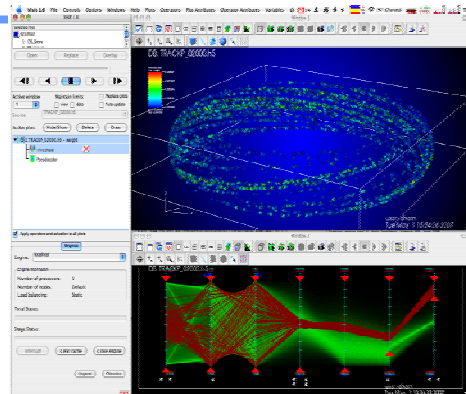
Motivation

- ❖ Usability
 - Don't want to force users to manually type in SQL-like queries:
 - `SELECT density WHERE (1100 <= temperature < 1200)`
 - Want to simplify use of potentially complex software machinery.
 - Want to optimize opportunity for leveraging human insight.

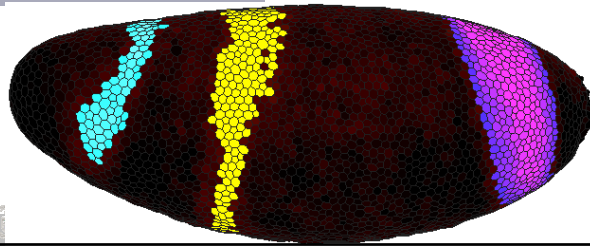
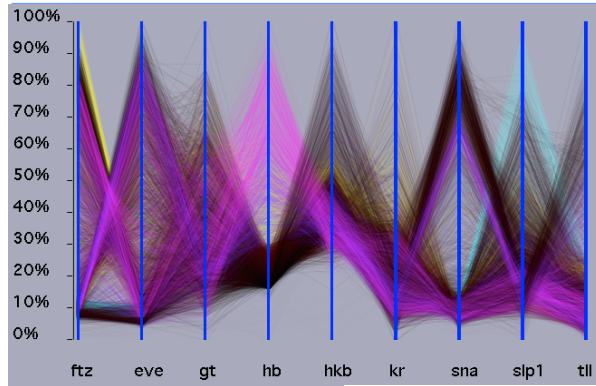


QDV Interfaces – Parallel Coordinates

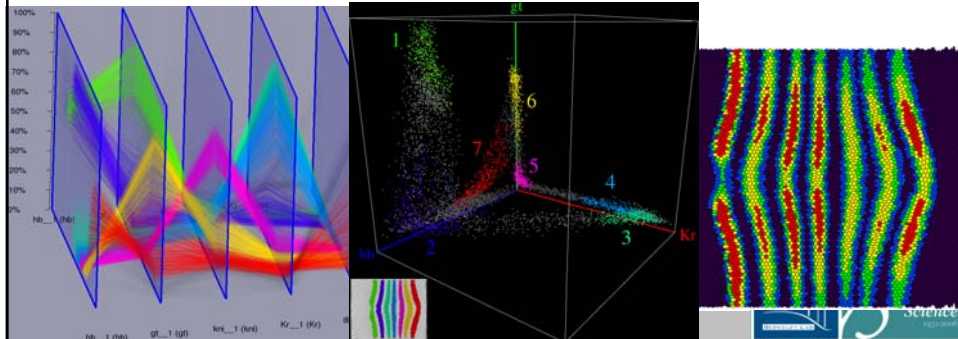
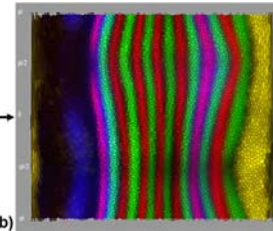
- ❖ Application: fusion microturbulence
- ❖ Objective: interactive multidimensional filtering to locate and analyze interesting phenomena.
- ❖ Result: capability in production software (VisIt).



QDV Interfaces – Multimodal, Linked Views



QDV Interfaces – Multimodal, Linked Views



QDV Cybersecurity Case Study

- ❖ The next sequence of slides discusses application of the work to a cybersecurity application – proof that the idea is generally applicable to large data visualization.
- ❖ The team:
 - NERSC Network Security
 - ESnet Network Engineers
 - Scientific Data Management Research
 - Visualization Research



QDV – Detecting Distributed Scans

- ❖ The problem:
 - One day's worth of traffic consists of tens of millions of individual connections.
 - Traffic increasing by an order of magnitude every 48 months.
 - ESnet monthly traffic levels now exceed 1 PB.
 - The Internet is a hostile environment, and it will get worse.
 - Objective: enable rapid forensic data analysis (network flow records).



QDV – Detecting Distributed Scans

- ❖ The data:
 - 42 weeks' of connection records from Bro (NERSC).
 - 281GB for raw data, 78GB for compressed bitmap indices.

- ❖ “Hero-sized problem”
 - No previous network analysis work has ever attempted to perform interactive visual analytics on data of this scale.
 - Result: what once took days (if at all possible) now takes seconds.



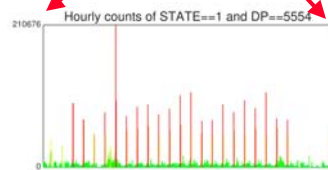
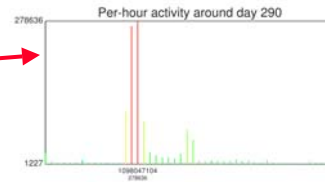
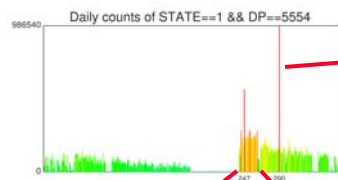
QDV – Detecting Distributed Scans

- ❖ The starting point:
 - You are a network security analyst
 - Your beeper goes off [at lunch, in the shower...]
 - You receive an alert that “something odd is happening with the network...IDS showing unusual levels of activity on port 5554”
 - Your job – answer questions:
 - What's going on now?
 - How long has this been happening?
 - Implications?



QDV – Detecting Distributed Scans

❖ The Case Study



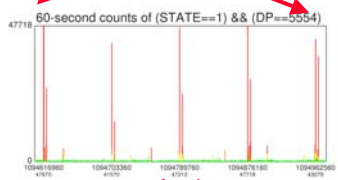
1. Query to produce a histogram of unsuccessful connection attempts over a 42-week period at one-day temporal resolution (upper left).
2. Drill into the data, query to produce a new histogram covering a four-week period at one-hour temporal resolution (lower left).
3. Generate a histogram of one-hour resolution over a two-day period around day 290 (upper right).



QDV – Detecting Distributed Scans



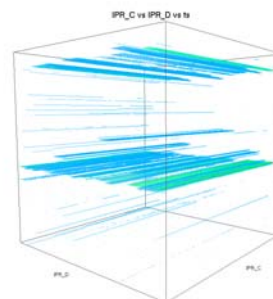
5. Query to generate a histogram of unsuccessful connection attempts over a five-day period sampled at one-minute temporal resolution (middle, left). Regular attacks occur at 21:15L, followed by a second wave 50 minutes later.



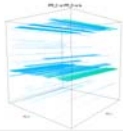
6. Query to generate histogram over a two-hour period at one-minute temporal resolution (lower left).



7. Query to generate a 3D histogram showing the coverage of attacks in destination address space (lower right).



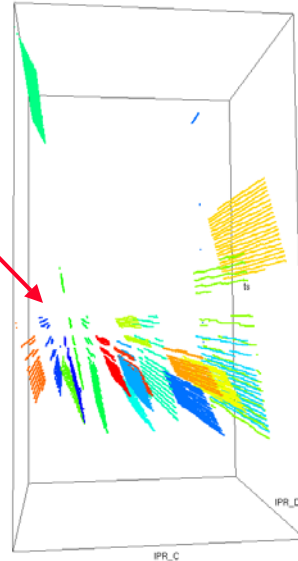
QDV – Detecting Distributed Scans



After establishing that (1) a temporally regular activity is occurring, and (2) that it is in fact a systematic probe (scan) of entire blocks of network addresses, the next task is to determine the set of remote hosts participating in the attack.

Working backwards, we isolate the A, B, C and D address octets of the hosts participating in the attack.

This image shows a 3D histogram of the destination address space being attacked by each of 20 different hosts. The vertical axis is time – a seven-minute window at one-second temporal resolution.



QDV – Detecting Distributed Scans

- ❖ Our analysis was performed in statistical space only.
 - We never accessed the raw data.
 - Our processing and visualization used only the index data.
- ❖ The same principles can be (and will be) applied to scientific data.
- ❖ Challenges:
 - How to define “interesting?”
 - Effective user interfaces for:
 - Support rapid interrogation, propagating query results from step to step in the analysis process.
 - Multivariate visualization
 - Drill-down (mining), linked/correlated views
 - Adapting, applying and deploying these principles to scientific data (e.g., AMR, distributed data/computing resources).



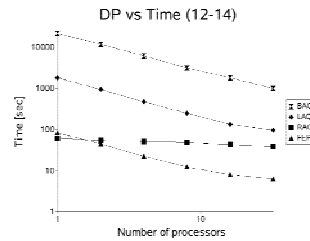
QDV – Detecting Distributed Scans

❖ How fast is it?

- 3 to 6 orders of magnitude faster than shell-based tools.
- 2 to 3 orders of magnitude faster than ROOT, the “gold standard” in the HEP community.
- Shows favorable scaling characteristics up to 32P.

1D Histogram, per-bin queries

PEs	Shell-based	ROOT/Projection Index	ROOT/FastBit
1	156381.14	1357.07	5.36
2	71835.32	600.05	3.72
6	21952.12	214.14	2.66
13	9389.96	113.88	2.58
21	2237.53	98.95	2.05



Conditional 2D histogram processing time

Query: (1000 < DP < 11000) AND (50 <= tsyday <= 350) AND (state==1) AND (12 <= tshour <=14), 10K total bins.



QDV Interface – Histogram “Cross-Products”

The screenshot displays the QDV software interface. At the top, there is a 'Query Window' with a text area containing a query: 'Q1 == DP == 1023) && (D042 == R_SIZE == 3416)'. Below the query window is a 'Current Query' section and buttons for 'Form Query' and 'Pose Histogram Query'. To the right is a 'Variable Browser Window' listing various variables like DP, R_SIZE, STATE, etc., with checkboxes for selection. The main area shows a 3D histogram plot with a red vertical line. Below this are two 2D histogram plots: 'R_SIZE' and 'DP'. The 'R_SIZE' plot shows a distribution of values with a vertical dashed line. The 'DP' plot shows a similar distribution with a vertical dashed line. At the bottom, there are control panels for each plot, including 'Selection Box' and 'Set' buttons. The system tray at the bottom right shows the time as 18:04.

QDV – Conclusion

- ❖ New capability: ability to focus visualization and analysis processing on interesting data.
 - Orders of magnitude faster than previous approaches.
 - Directly responsive to needs of scientific researchers.
 - Quantifiable reduction in data understanding duty cycle.
- ❖ Leverages state-of-the-art Scientific Data Management technology to accelerate searches.
- ❖ QDV concepts are general-purpose and scalable to 1000s of PEs.



The Take-Home Message

- ❖ Gaining insight from large and complex datasets is a major bottleneck.
- ❖ Thesis: effective software tools for gaining insight from such data must use a blend of technologies from:
 - Scientific Data Management
 - Visualization, Analytics
 - Computer Science
- ❖ Query-Driven Visualization is one such approach.



The End

