Extreme Scale Visual Data Science

Chad A. Steed, Ph.D.
csteed@acm.org
http://csteed.com

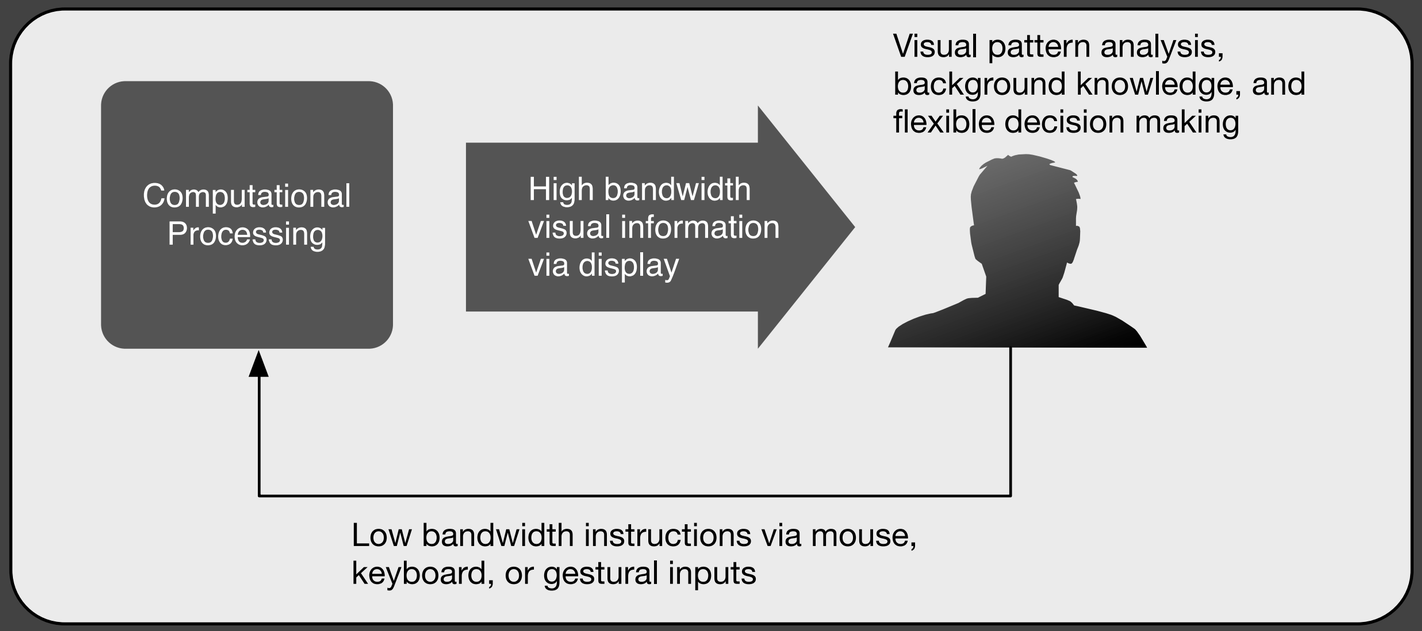
About Me

- **Visual Analytics Researcher at ORNL**
  - Computational Sciences & Engineering
- **Education**
  - Fine Art (Minor), Software Engineering (B.S.), Hydrography (M.S.), CS Vis and Computer Graphics (Ph.D.)
- **Professional**
  - Lockheed Martin (2 years), U.S. Naval Research Laboratory (9+ years), ORNL (~5 years)

Visual
Data Science

Why Visualization?

- **Visual Displays**
  - Highest bandwidth channel from computer to human
  - Flexible pattern matching capabilities
- **Visualization** is an external artifact supporting cognitive processing
- Information only becomes valuable when it is understood



What is Visualization?

"Transformation of the symbolic into the geometric"
[McCormick et al. 1987]

"…finding the artificial memory that best supports our natural means of perception"
[Bertin 1967]

"The use of computer-generated, interactive, visual representations of data to amplify cognition."
[Card, Mackinlay, and Shneiderman 1999]

The Goal:
Reveal Insight

Create interactive visualizations that efficiently reveal insight.

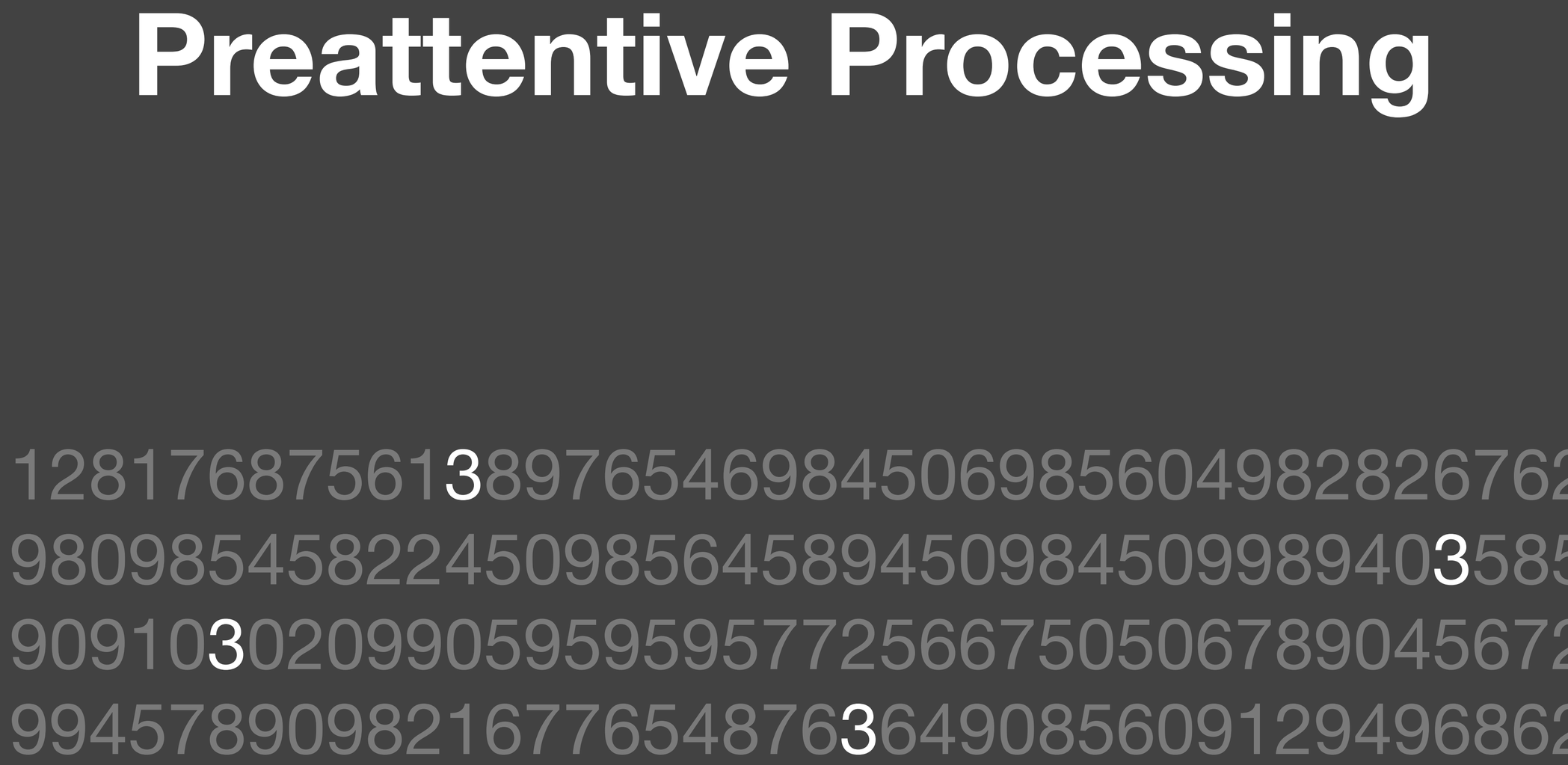
Joseph Berkson calls this
“interocular traumatic impact”.

“A conclusion that hits us between the eyes.”
[Cleveland, 1993]

Preattentive Processing

128176875613897654698450698560498282676298098545822450985645894509845099894035859091030209905959595772566750506789045672994578909821677654876364908560912949686

Preattentive Processing



The InfoVis Mantra

- Shneiderman proposed a taxonomy to understand what people do with visualizations
- Visual Information-Seeking Mantra:
  - “Overview first, zoom and filter, then details-on-demand”
- Design paradigm for InfoVis systems
- Useful starting point for designing advanced graphical interfaces

Ben Shneiderman
Visual Lang. Symposium ’96

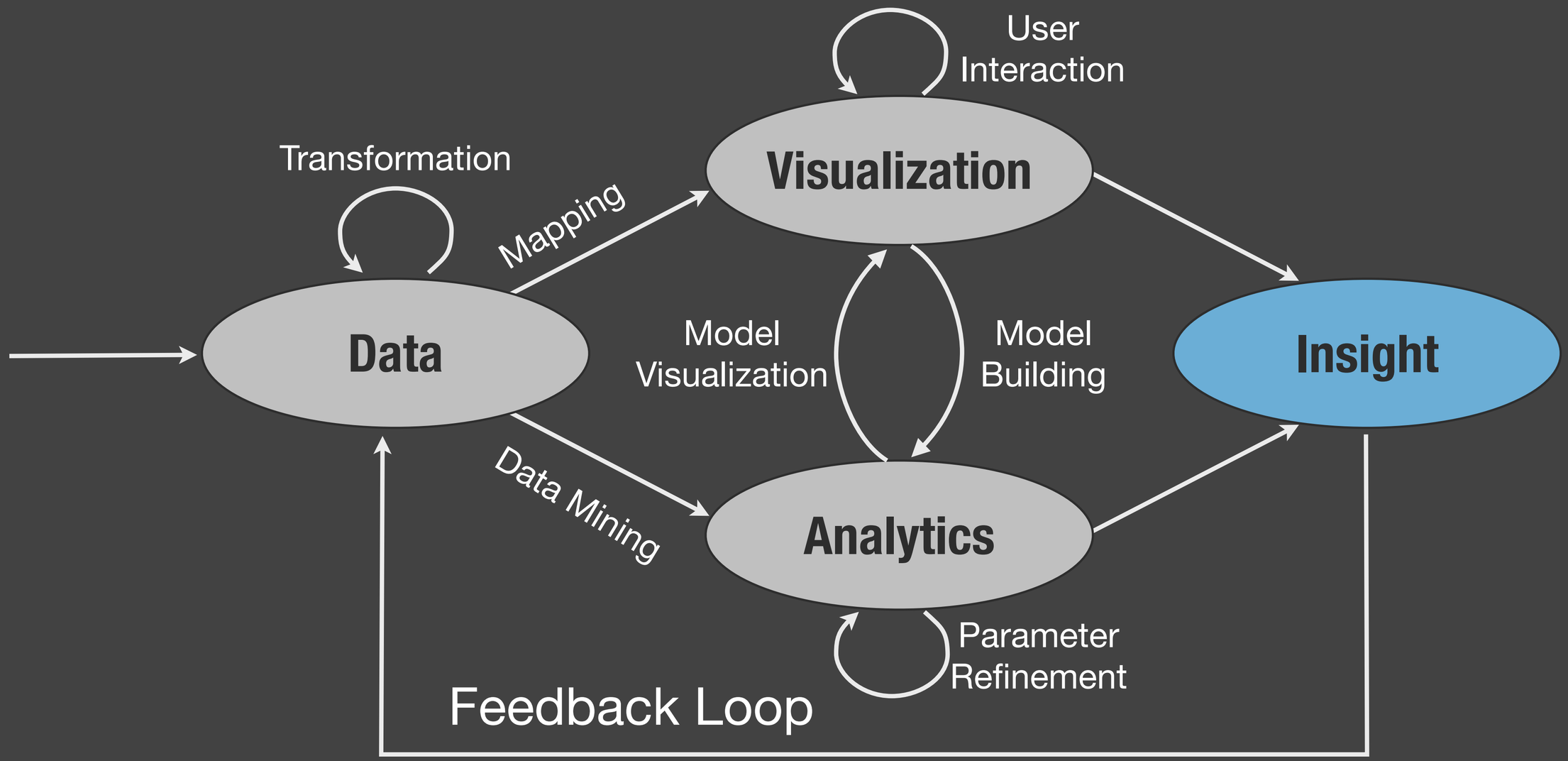
Can InfoVis Do More?

- Quest to provide more value by combining computational analysis with InfoVis (Shneiderman ’92)
  - No longer competitors — each has something to contribute
- Orchestration of computation with InfoVis leads to…

Visual Analytics

- “The science of analytical reasoning facilitated by interactive visual interfaces” -Thomas & Cook, *Illuminating the Path*
- **Combines:**
  - Data analysis, InfoVis, and analytical reasoning
- Grew from the view that InfoVis was neglecting these other aspects.

Visual Data Science



After Keim et al., Mastering the Information Age: Solving Problems with Visual Analytics, 2011

Visual Data Science for Climate Simulations

“Climate is what we expect, weather is what we get.”
-Mark Twain

Funded by DOE Office of Science:
Climate Science for a Sustainable Energy Future (CSSEF)
Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT)

Climate Data Deluge

- Current climate data analysis techniques are limited:
  - Increasing quality, quantity, and number of different parameters
    - Static visualizations and drastic data reductions
    - Limited linkage to decision support
- Methods to visualize and analyze data have not kept pace with computational advances.
- Scientists are forced to reduce the problem to fit the tools.

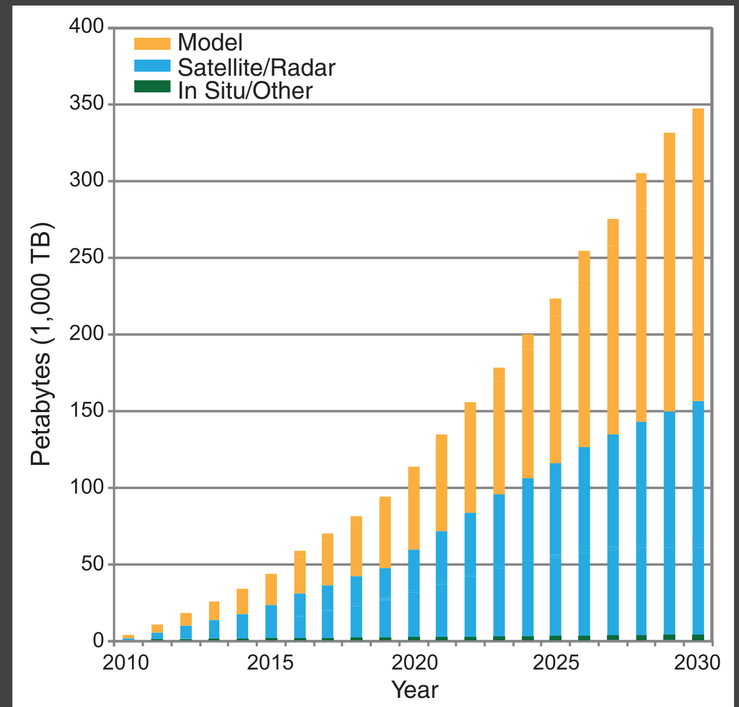
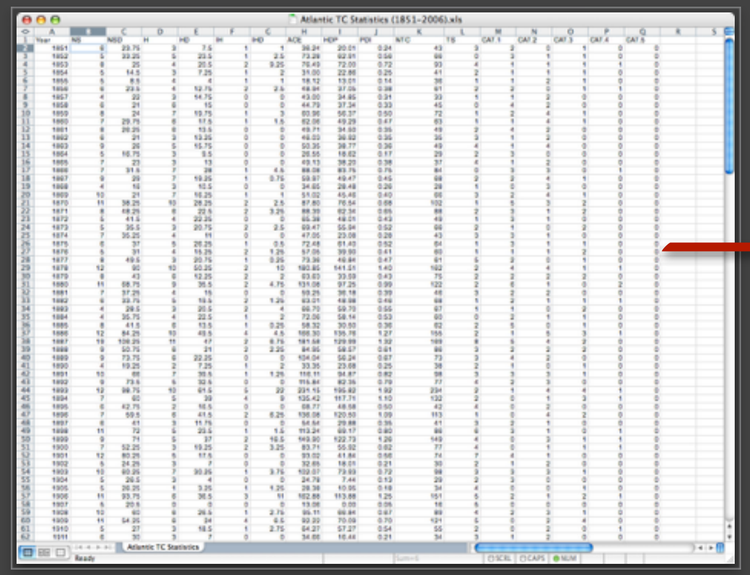
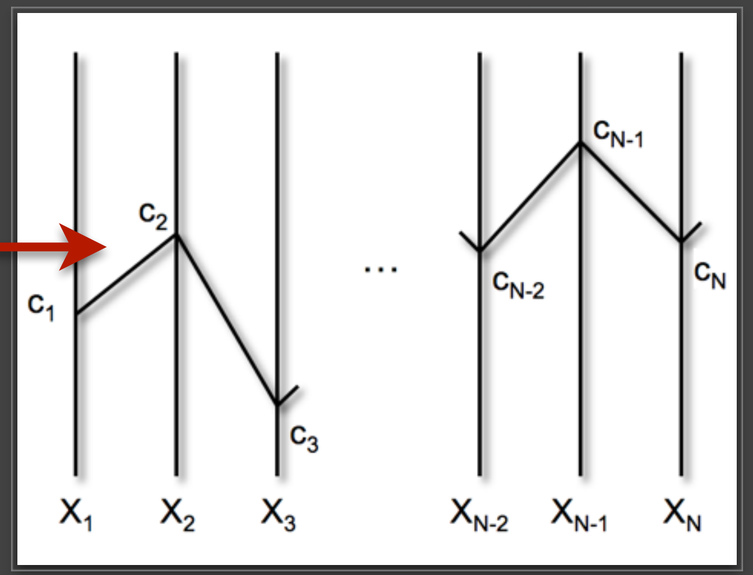


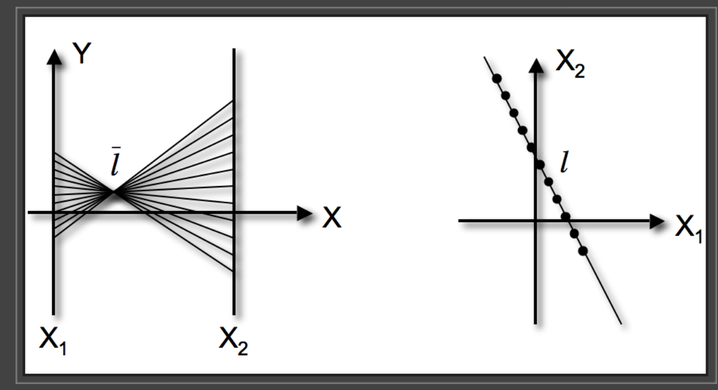
Fig. 2. The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you are not a climate scientist. The figure shows the projected increase in global climate data holdings for climate models, remotely sensed data, and in situ instrumental/proxy data.

Overpeck et al., Science 331, 700 (2011)

Parallel Coordinates







Point Line Duality

Parallel Coordinates

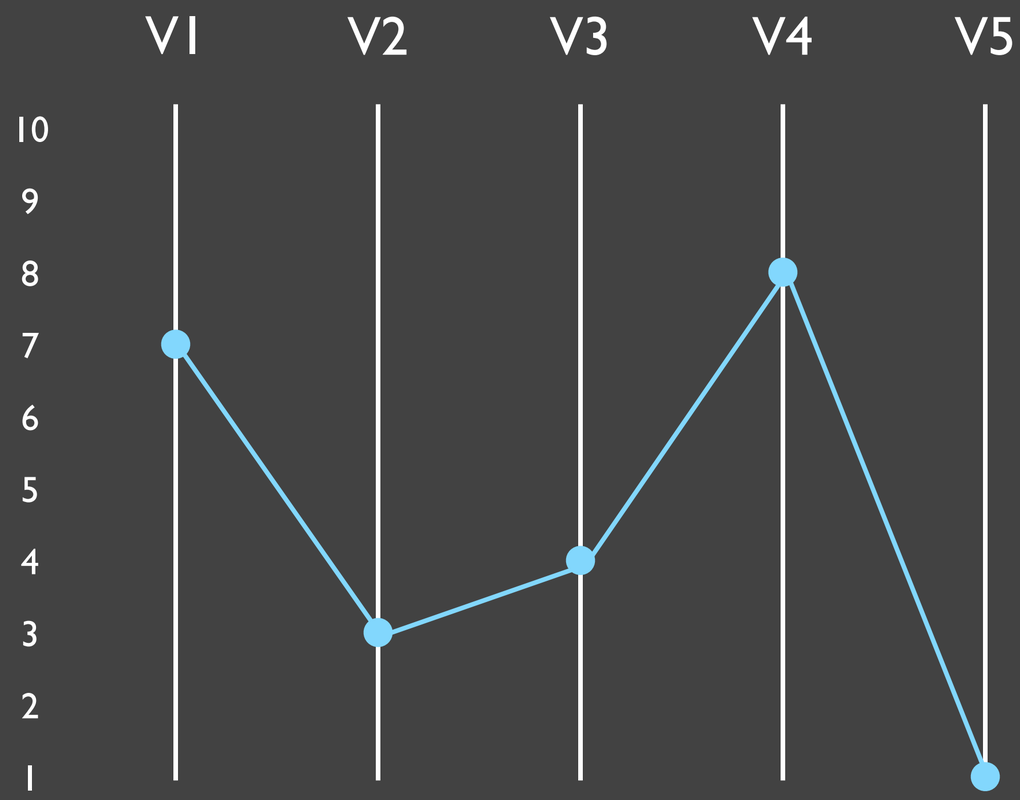
Two-dimensional representation of multidimensional data sets by representing the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ by points on N parallel axes which are joined with a polyline.

Inselberg, 1985 | Wegman, 1990

Parallel Coordinates

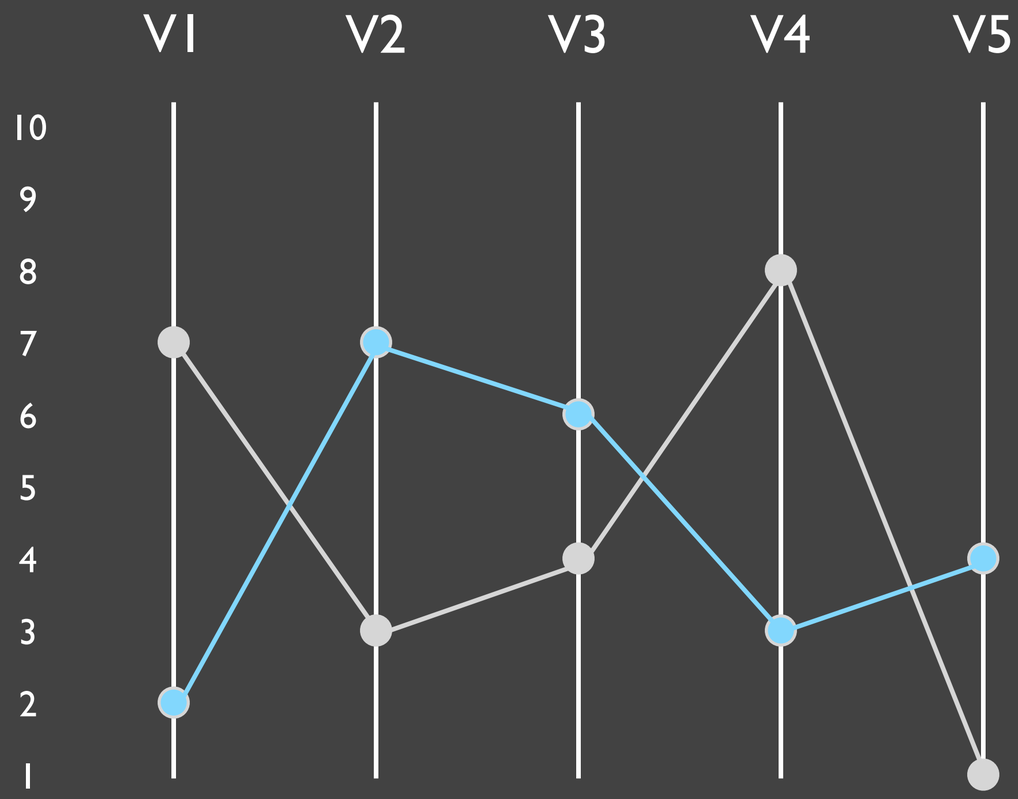
	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	2	7	6	3	4
D3	9	8	1	4	2

Parallel Coordinates



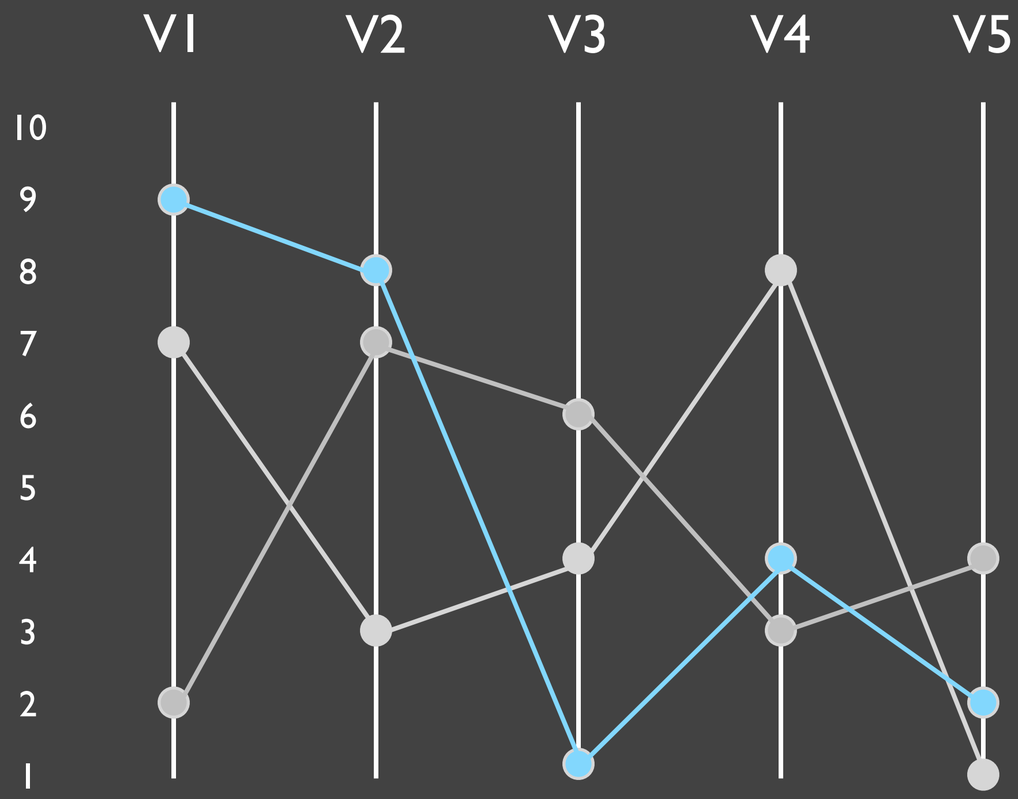
	V1	V2	V3	V4	V5
D1	**7**	**3**	**4**	**8**	**1**
D2	2	7	6	3	4
D3	9	8	1	4	2

Parallel Coordinates



	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	**2**	**7**	**6**	**3**	**4**
D3	9	8	1	4	2

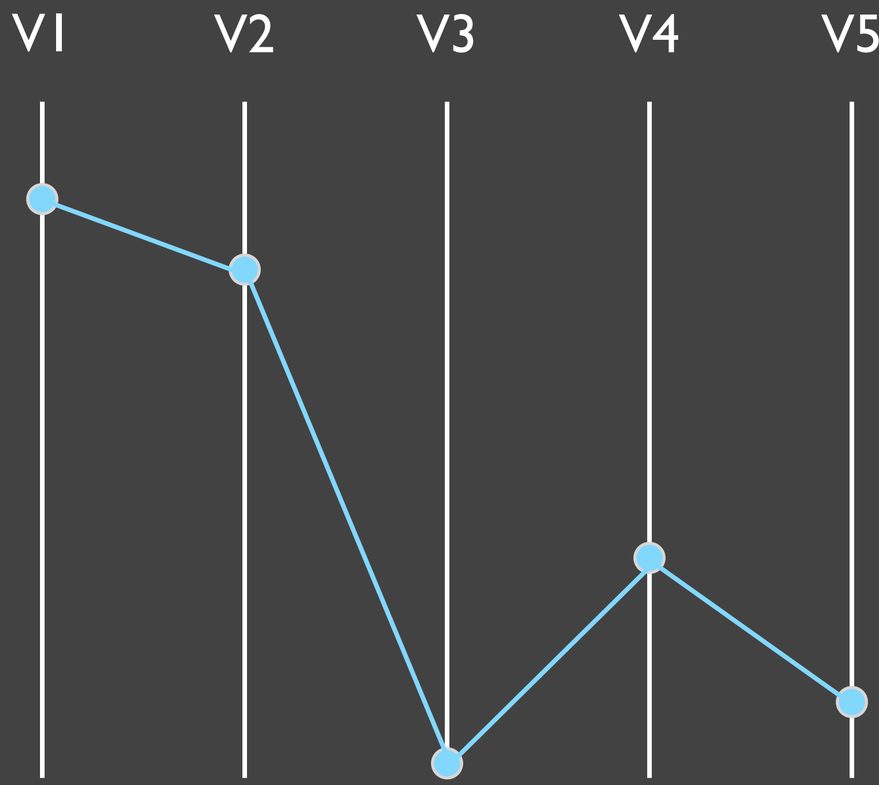
Parallel Coordinates



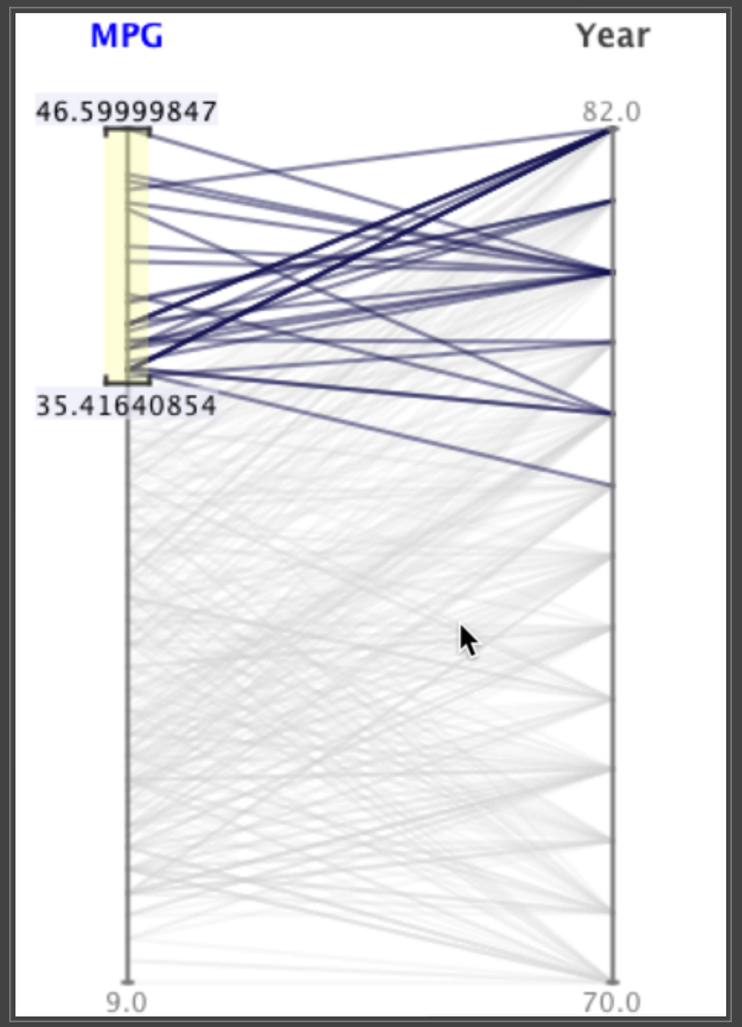
	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	2	7	6	3	4
D3	**9**	**8**	**1**	**4**	**2**

Parallel Coordinates

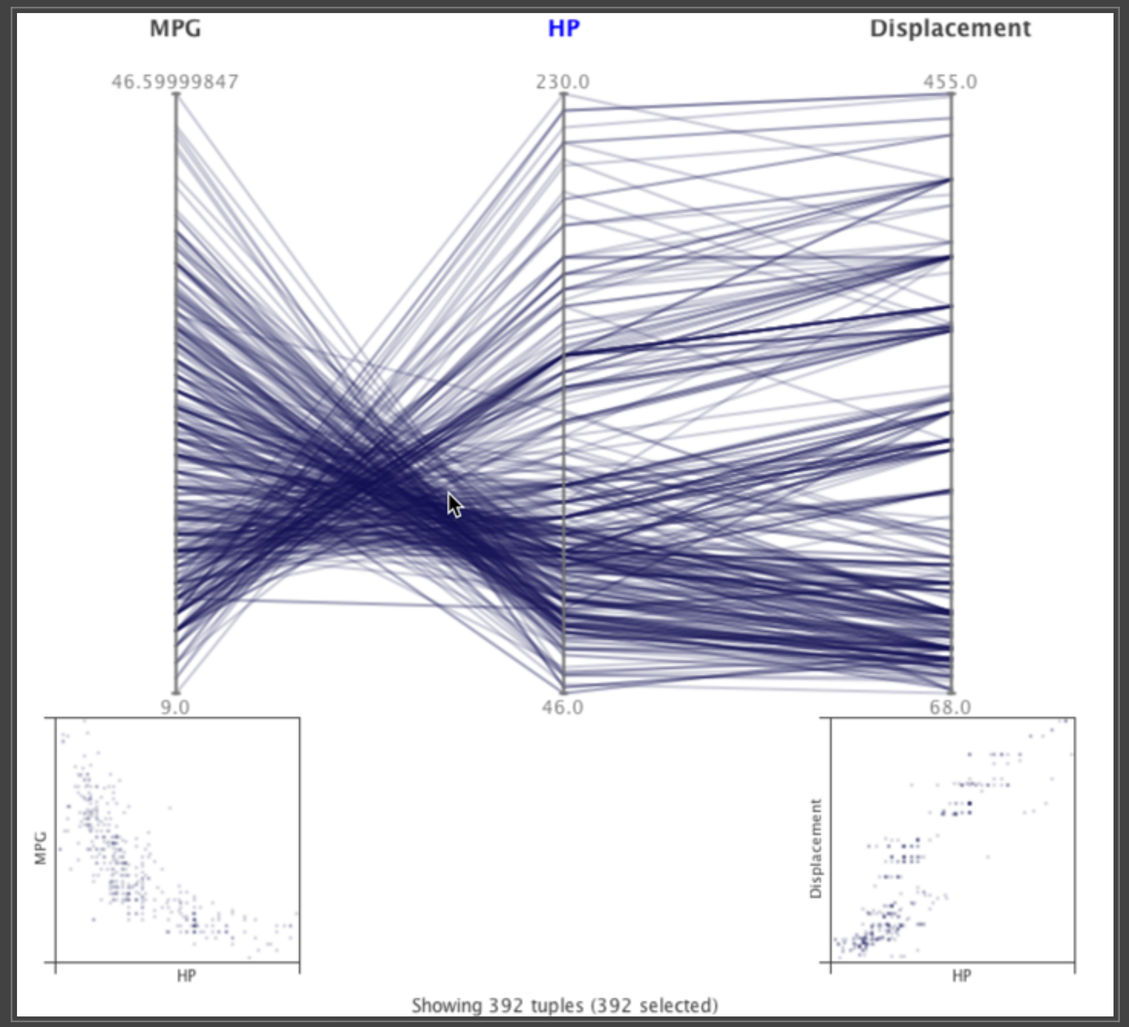
- Encode variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline
- Works for any hypervariate data



Visual Patterns in ||-coords

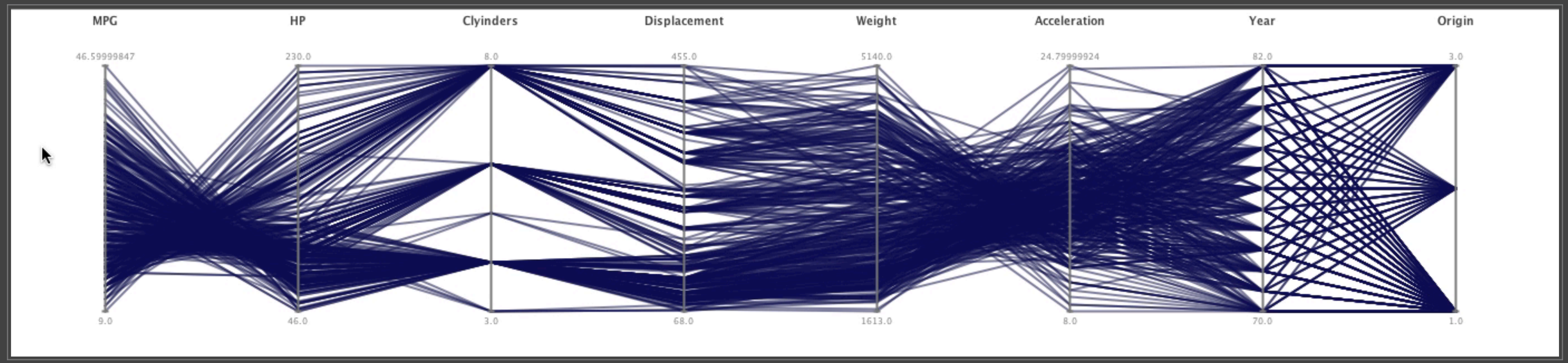


Trends



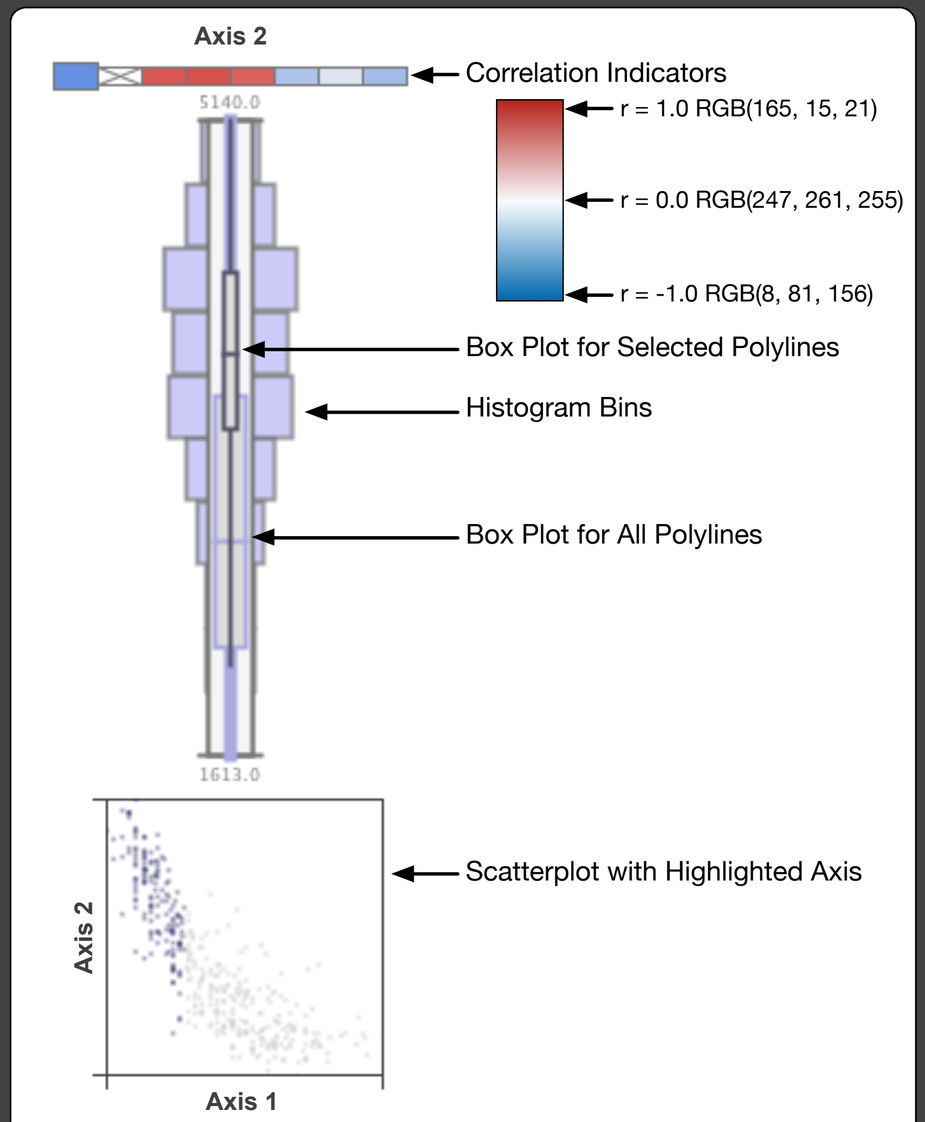
Correlations

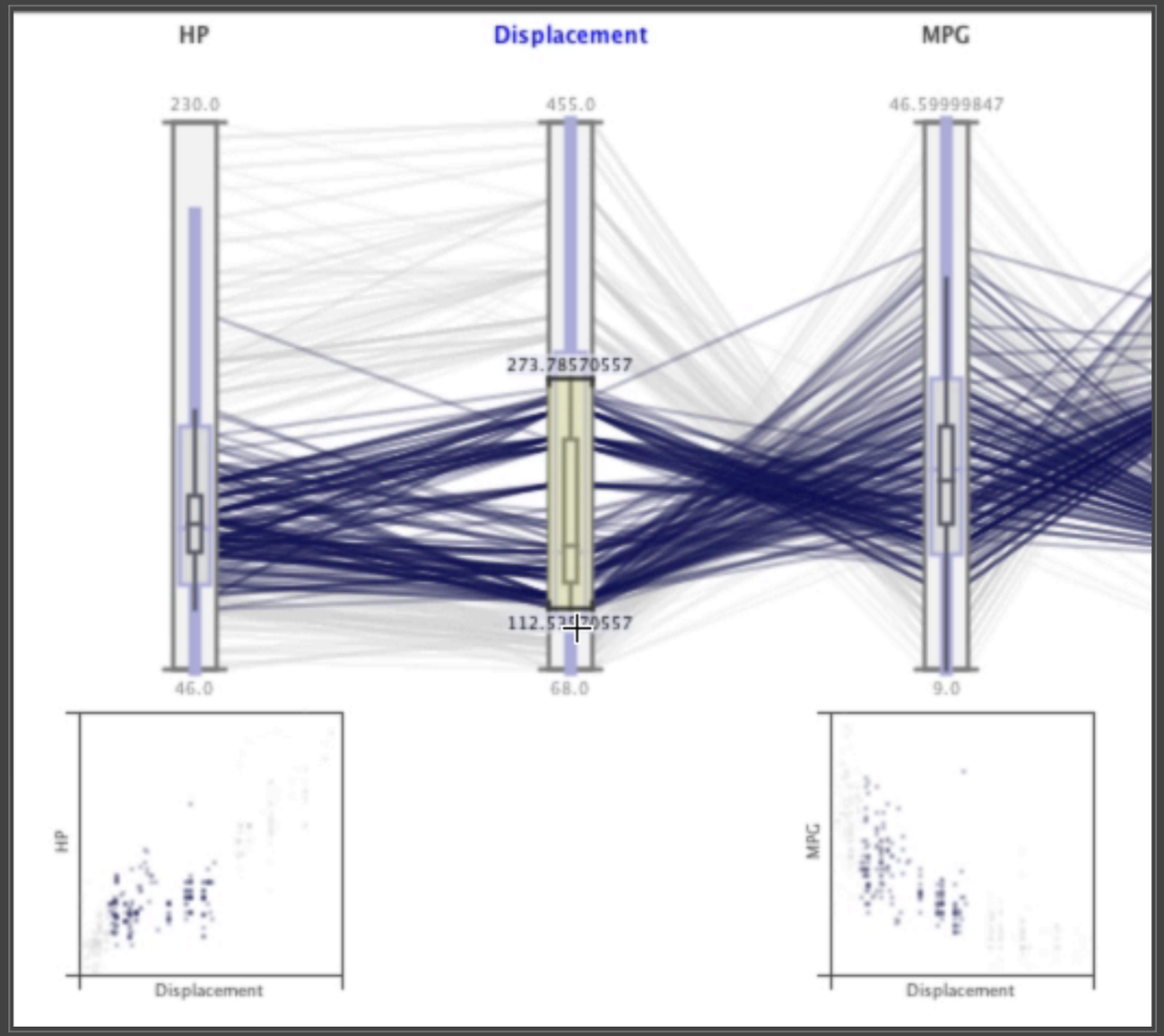
Interactive Queries in ||-coords



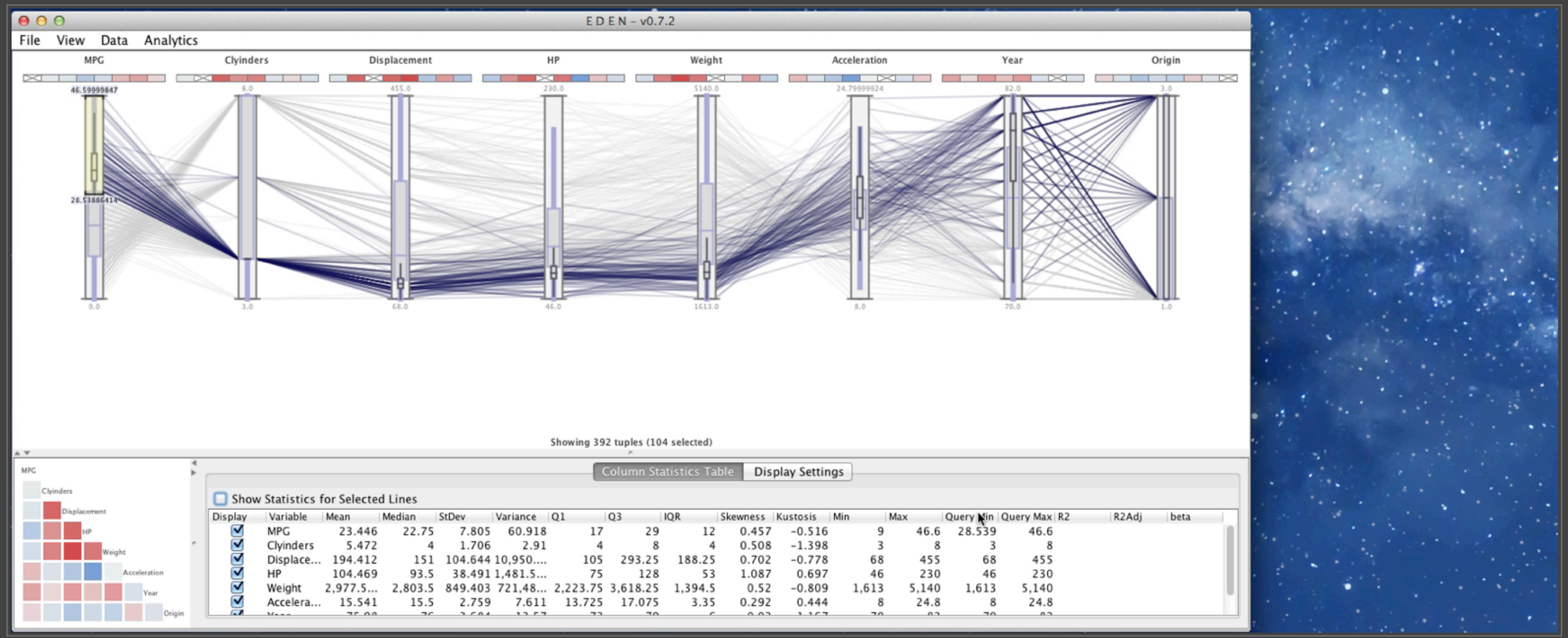
Hypervariate Queries | Axis Relocation | Focus+Context

“Scented” Axis Widgets



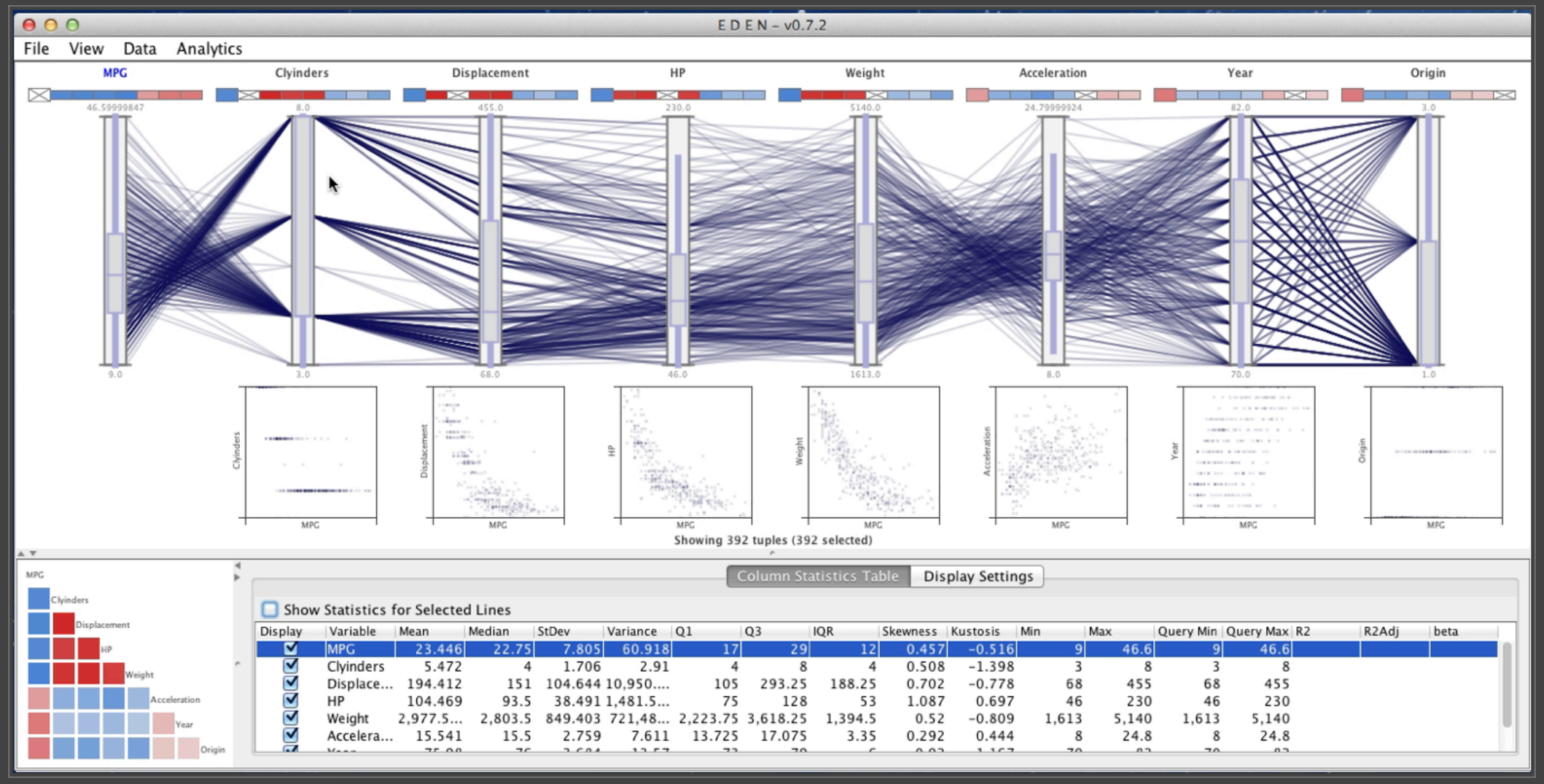


Coordinated Views and Drill Down



Coordinated Multiple Views | Details on Demand | Visual Properties

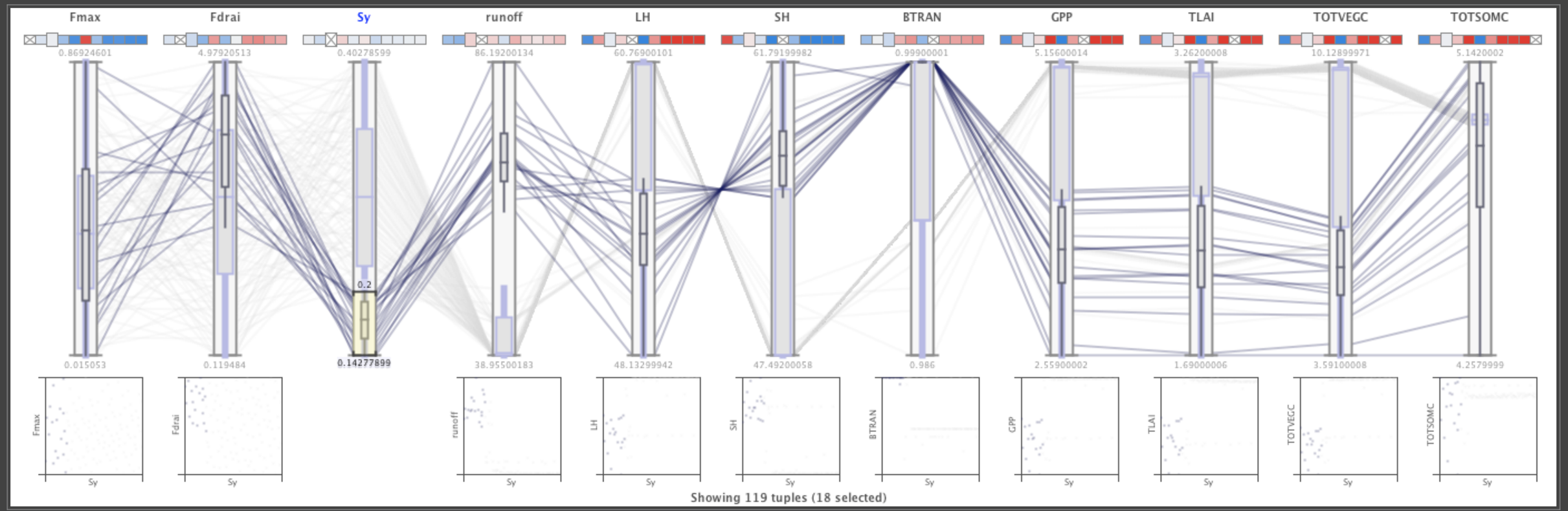
Correlation Mining and Auto Arrangement



Display	Variable	Mean	Median	StDev	Variance	Q1	Q3	IQR	Skewness	Kustosis	Min	Max	Query Min	Query Max	R2	R2Adj	beta
☑	MPG	23.446	22.75	7.805	60.918	17	29	12	0.457	-0.516	9	46.6	9	46.6			
☑	Clyinders	5.472	4	1.706	2.91	4	8	4	0.508	-1.398	3	8	3	8			
☑	Displace...	194.412	151	104.644	10,950....	105	293.25	188.25	0.702	-0.778	68	455	68	455			
☑	HP	104.469	93.5	38.491	1,481.5...	75	128	53	1.087	0.697	46	230	46	230			
☑	Weight	2,977.5...	2,803.5	849.403	721,48...	2,223.75	3,618.25	1,394.5	0.52	-0.809	1,613	5,140	1,613	5,140			
☑	Accelera...	15.541	15.5	2.759	7.611	13.725	17.075	3.35	0.292	0.444	8	24.8	8	24.8			

Correlation Matrix | Drill Down | Detail Table | Auto Arrangements

Exploratory Data analysis ENvironment



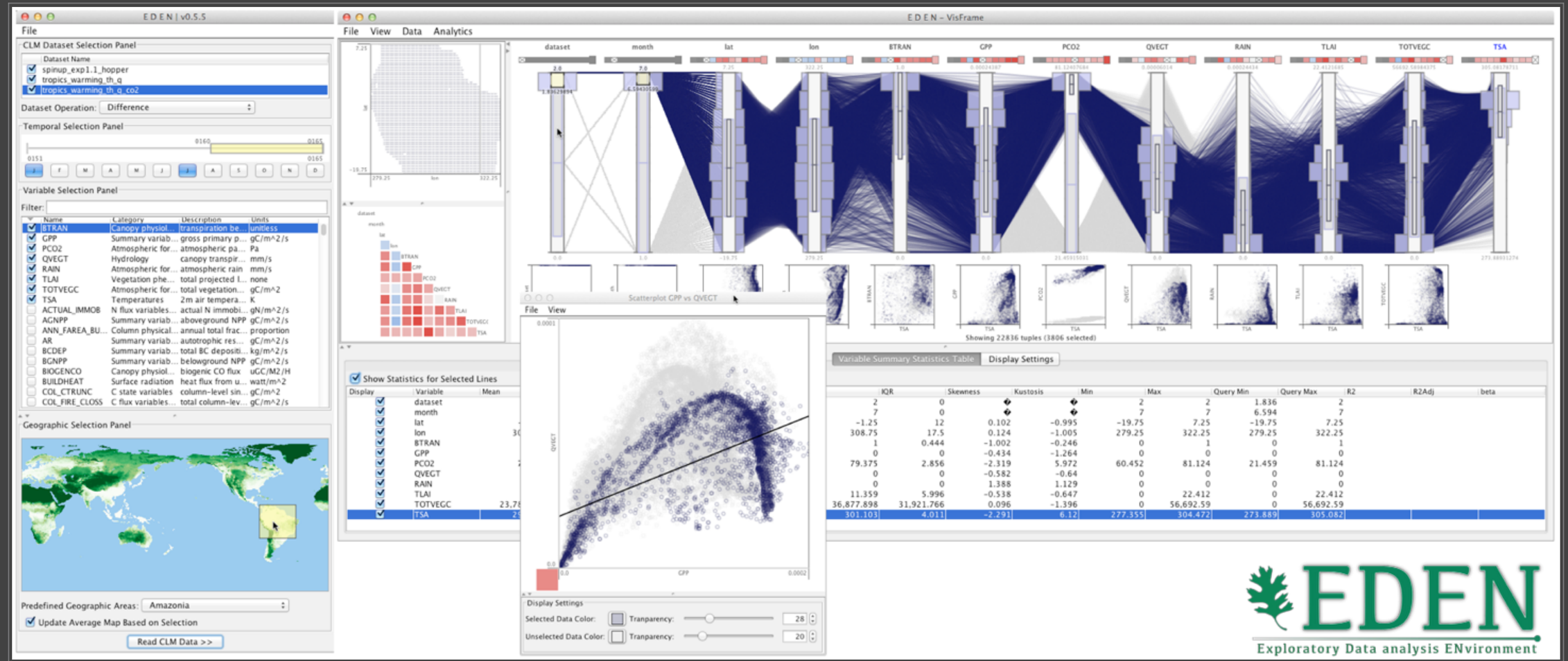
Enabling the discovery of **serendipitous** insight.

Citation: Chad A. Steed, Galen Shipman, Peter Thornton, Daniel Ricciuto, David Erickson, and Marcia Branstetter. "Practical Application of Parallel Coordinates for Climate Model Analysis." In *Proceedings of the International Conference on Computer Science*, June 2012, pp. 877-886. DOI: 10.1016/j.procs.2012.04.094.

EDEN
Exploratory Data analysis ENvironment

EDEN source code and executables are available at http://cda.ornl.gov/projects/eden/

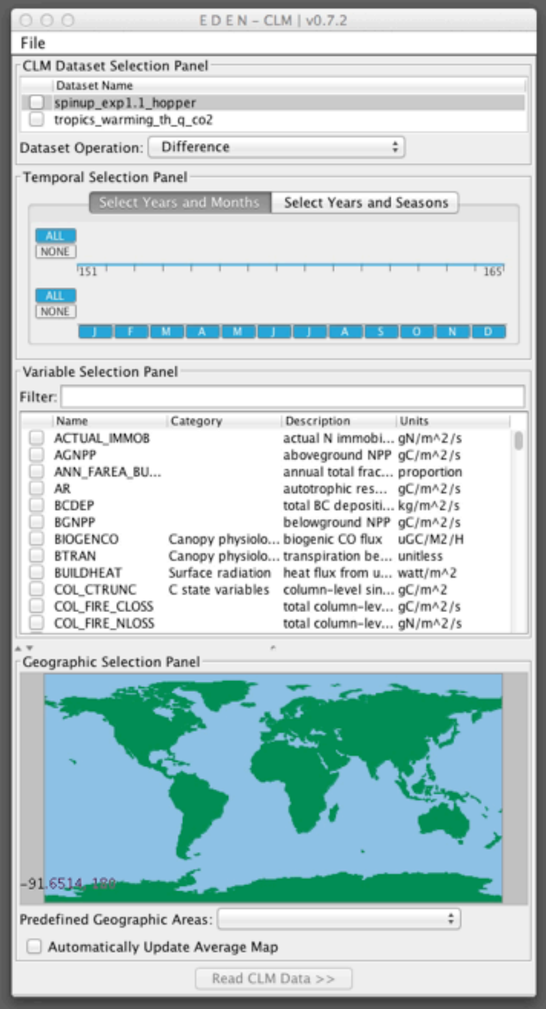
Exploratory Data analysis ENvironment

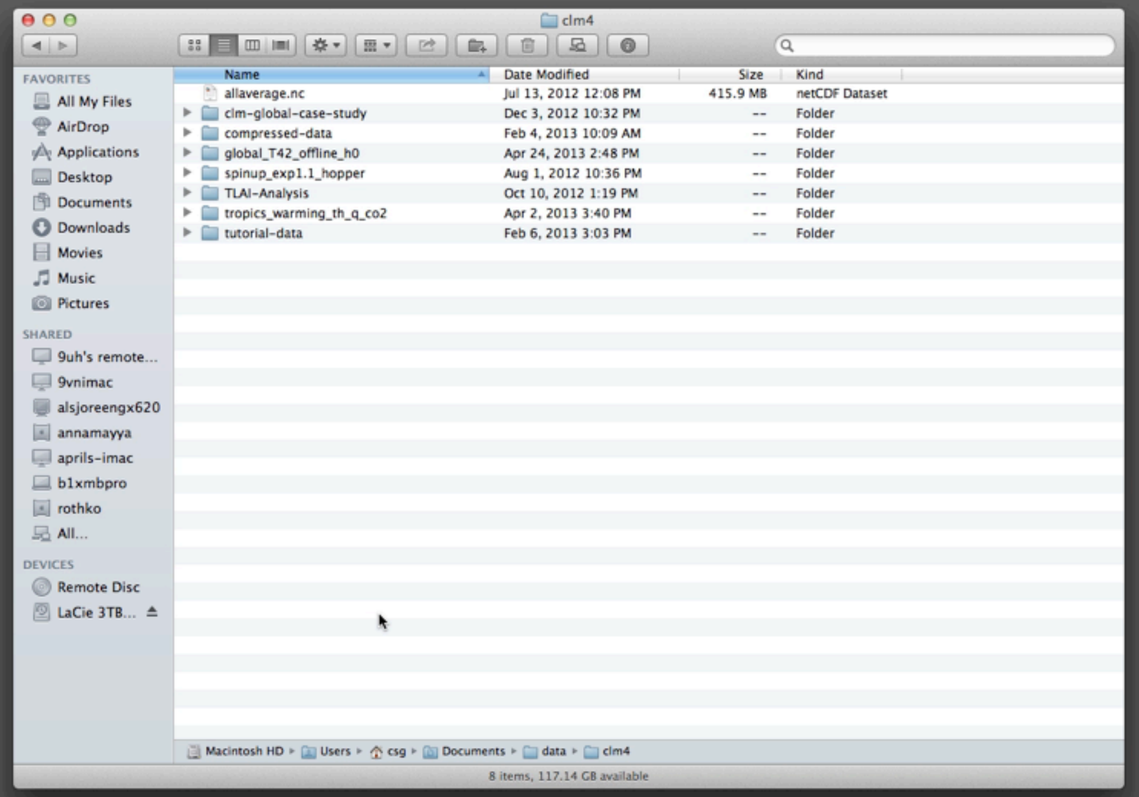


Visual analytics is necessary to realize the full potential of extreme scale computing.

Citation: Chad A. Steed, Daniel M. Ricciuto, Galen Shipman, Brian Smith, Peter E. Thornton, Dali Wang, and Dean N. Williams. "Big Data Visual Analytics for Earth System Simulation Analysis." *Computers & Geosciences*. vol. 61, Dec. 2013, pp. 71-82. DOI: 10.1016/j.cageo.2013.07.025.

Large Scale EDA for Climate Science





EDEN source code and executables are available at http://cda.ornl.gov/projects/eden/

Multi-Disciplinary Collaborations

ORNL Climate Scientists Collaborators
(Peter Thornton & Dan Ricciuto)

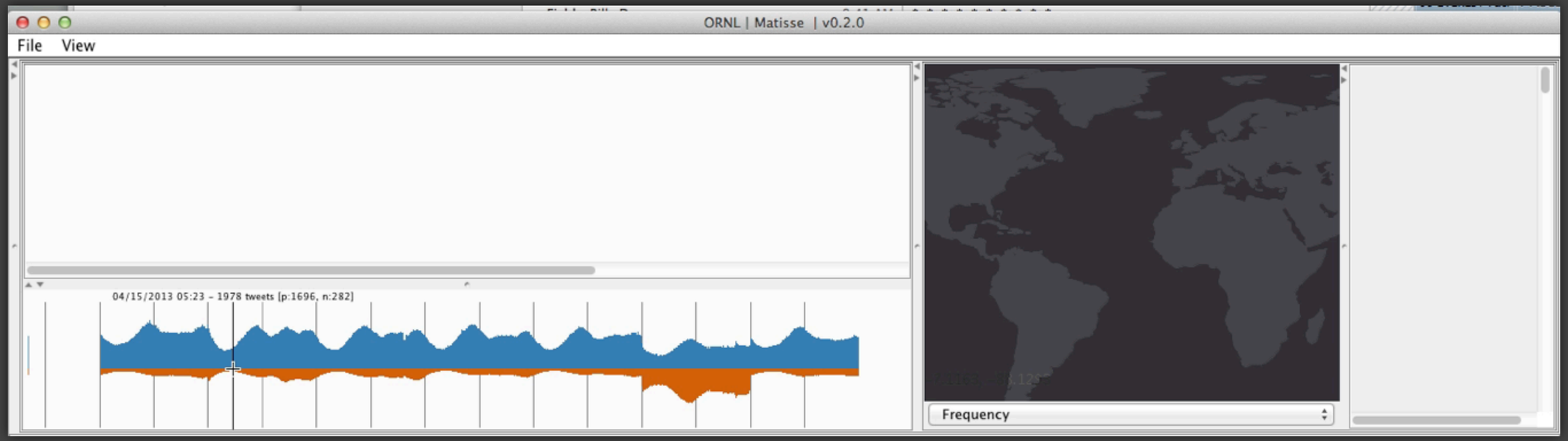
20 Years of Excellence in Computational Science
OLCF
Oak Ridge Leadership Computing Facility

Conclusions

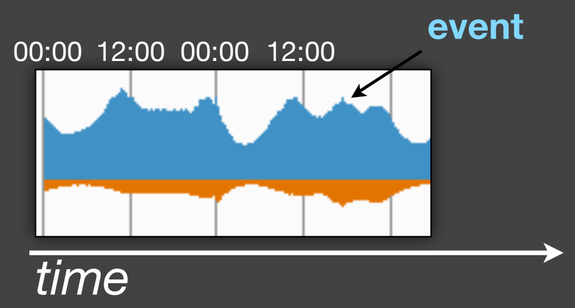
- Human-centered, exploratory analysis can help uncover surprise discoveries and capitalize on human intuition.
  - But difficult on extreme scale platforms that utilize machine-centered designs
- Statistical analytics can react to human interactions and guide analysis by making potentially significant information more visually salient
  - Seeking collaborations from math / statistics / data mining / HPC experts
- Close collaborations with domain experts is vital to cross-cutting data science endeavors
  - Integration in the design iterations can be a mutually beneficial activity.
  - Helps ensure response to actual needs of scientists
  - Manage tendency to minimize computer science focus
- Provenance/history capture is needed

Matisse

Interactive Social Media Visual Analytics



Week of Boston Marathon Bombing
14-20 April, 2013



Invention Disclosure: Title: An Interactive Visual Analysis Method for Dynamic Exploration of Streaming Textual Information. Docket Number 201303076. Inventors: Chad A. Steed, Thomas E. Potok, Paul Bogen, Christopher Symons, and Robert Patton.

TOTVEGC
TOTSOM
26729.30
26881.50
Questions?
http://csteed.com