HUMAN-CENTERED DATA UNDERSTANDING

Chad A. Steed | http://csteed.com
Oak Ridge National Laboratory
Computer Science and Mathematics Division

CDA Group Meeting
July 18, 2018

Computers are faster,
Memory is cheap,
Data volumes are increasing,
And AI is advancing.

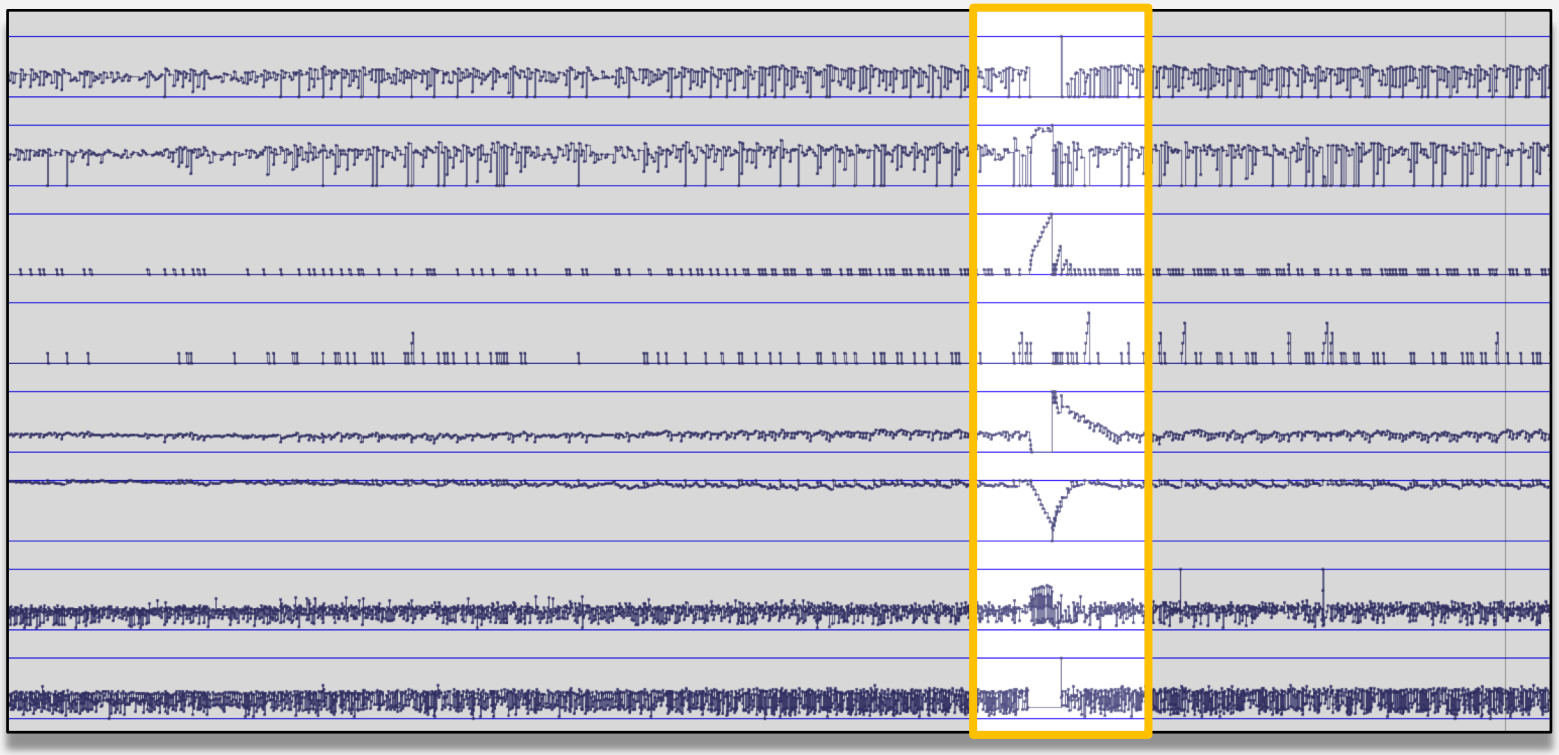
Do we need a human in the loop?

Colonel Stanislav Petrov
(image courtesy The New York Times)

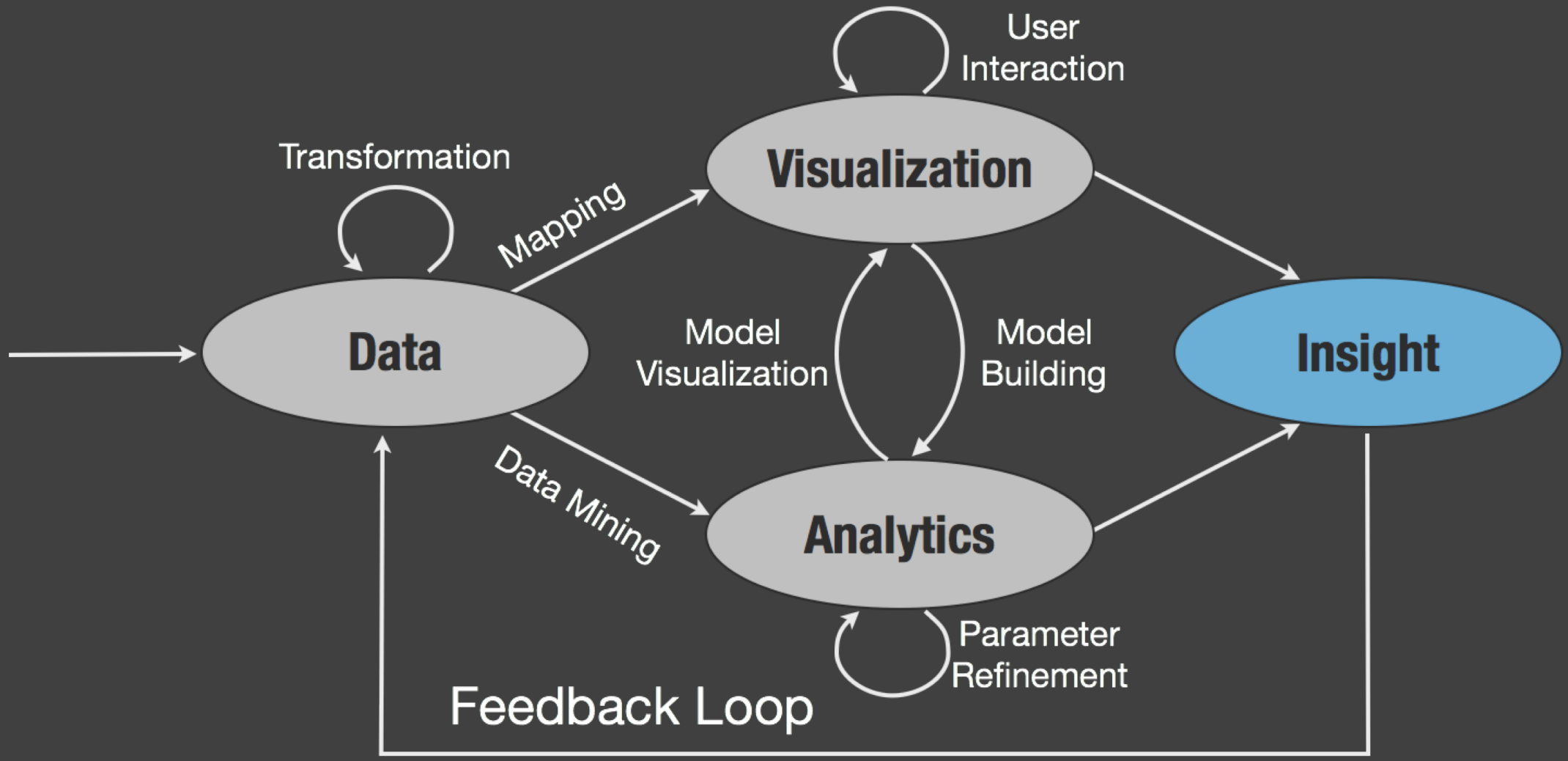
DATA

```
492258  2016-06-17 21:07:03.466|OPC.Temperature.ExtraTemp1|SuperUser (OPC)|9526586|618
492259  2016-06-17 21:07:03.466|OPC.Temperature.BottomTemperatureValidation|SuperUser
        (OPC)|9526586|618.19999999999982
492260  2016-06-17 21:07:03.486|Process.CathodeTuningControl.CathodePower|[OnChange(
        OPC.PowerSupply.Filament.VoltageFB)] Arcam.EBMControl.Process.
        CathodeTuningControl.OnCathodPowerChange() (Logic)|9526586|5.929202
492261  2016-06-17 21:07:03.516|Process.CathodeTuningControl.MeanCathodePower|[
        OnChange(Process.CathodeTuningControl.CathodePower)] Arcam.EBMControl.Process.
        CathodeTuningControl.MeasureMeanPower() (Logic)|9526587|5.905572
492262  2016-06-17 21:07:03.516|OPC.PowerSupply.HighVoltage.SafetySignal|[
        OnPositiveFlank(SafetySignalTimer.Timeout)] Arcam.EBMControl.Process.
        HighVoltageControl.OnTimeToSendSafetySignal() (Logic)|9526587|True
```

Insight



Visual Analytics



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

D. Donoho, “50 years of Data Science”, https:// dl.dropboxusercontent.com/u/ 23421017/50YearsDataScience. pdf

My Research Focus

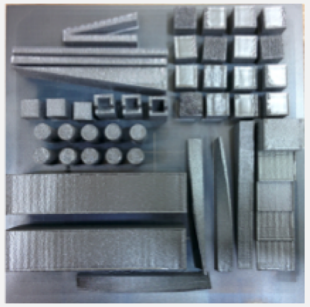
- **GOAL**: Improve human capabilities to discover, comprehend, and communicate insight in large and complex data through the design, implementation, and practical evaluation of visual analytics techniques

- **Specific Topics:**
  - **Graphical representations of data**
  - **Human interactions with visualizations**
  - **Integrating automated machine learning**

Agile, Interdisciplinary Data Science Research

Domain Experts
Dr. Ryan Dehoff
Additive Manufacturing

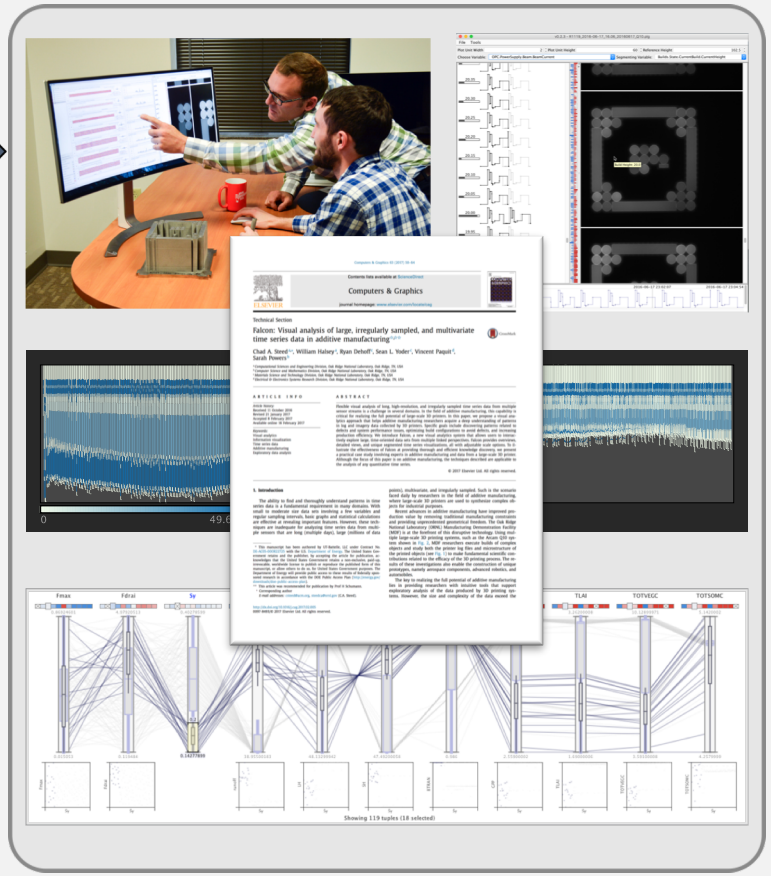
Scientific Domains
Large-scale 3D Printing



Real-world Data
Large, Complex, Unique

```
492258  2016-06-17 21:07:03.466|OPC.Temperature.ExtraTemp1|SuperUser (OPC)|9526586|618
492259  2016-06-17 21:07:03.466|OPC.Temperature.BottomTemperatureValidation|SuperUser
        (OPC)|9526586|618.19999999999982
492260  2016-06-17 21:07:03.486|Process.CathodeTuningControl.CathodePower|[OnChange(
        OPC.PowerSupply.Filament.VoltageFB)] Arcam.EBMControl.Process.
        CathodeTuningControl.OnCathodPowerChange() (Logic)|9526586|5.929202
492261  2016-06-17 21:07:03.516|Process.CathodeTuningControl.MeanCathodePower|[
        OnChange(Process.CathodeTuningControl.CathodePower)] Arcam.EBMControl.Process.
        CathodeTuningControl.MeasureMeanPower() (Logic)|9526587|5.905572
492262  2016-06-17 21:07:03.516|OPC.PowerSupply.HighVoltage.SafetySignal|[
        OnPositiveFlank(SafetySignalTimer.Timeout)] Arcam.EBMControl.Process.
        HighVoltageControl.OnTimeToSendSafetySignal() (Logic)|9526587|True
492263  2016-06-17 21:07:03.546|OPC.PowerSupply.SmokeDetector.Counts|SuperUser (OPC)
        |9526589|1
492264  2016-06-17 21:07:03.546|OPC.Temperature.ExtraTemp1|SuperUser (OPC)|9526589|598
492265  2016-06-17 21:07:03.546|OPC.Temperature.BottomTemperatureValidation|SuperUser
        (OPC)|9526589|598.09999999999991
492266  2016-06-17 21:07:03.666|OPC.Table.CurrentFeedback|SuperUser (OPC)|9526590|0.
        1153
492267  2016-06-17 21:07:03.666|OPC.Rake.CurrentFeedback|SuperUser (OPC)|9526590|0.
        1153
492268  2016-06-17 21:07:03.766|OPC.PowerSupply.SmokeDetector.Counts|SuperUser (OPC)
        |9526594|44
492269  2016-06-17 21:07:03.766|OPC.Temperature.ExtraTemp1|SuperUser (OPC)|9526594|544
492270  2016-06-17 21:07:03.766|OPC.Temperature.BottomTemperatureValidation|SuperUser
        (OPC)|9526594|544.40000000000009
492271  2016-06-17 21:07:03.786|OPC.InternalCooling.DifferentialPressureOverFilter|[
        OnChange(OPC.InternalCooling.PressureBeforeFilter_Unfiltered)] Arcam.
        EBMControl.Process.InternalCooling.UpdateDifferentialPressureOverFilter() (
        Logic)|9526594|0.02296734
```

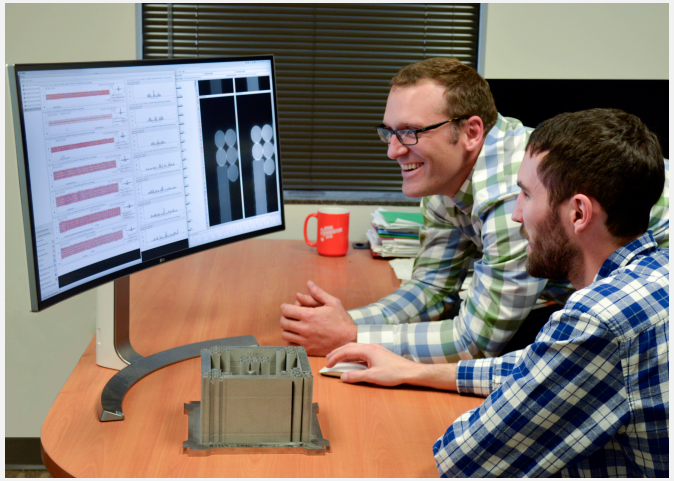
Results / Artifacts
Tools, Techniques,
Publications, Patents,
Open Source Software



Development Process

- **Domain experts are co-designers**
  - Participatory design
  - I spend time in learning domain challenges
  - They help design new features and evaluate releases
  - Domain experts keep me focus and they are invested in the process
  - We publish results together and share project artifact

- **Iterative development and evaluation cycle**
  - Design – Implement – Evaluate
  - Practical evaluations with real world scenarios
  - Proof of principle is new discoveries enabled by the techniques and transition to practice
  - Empirical evaluations are possible, but difficult at a national lab



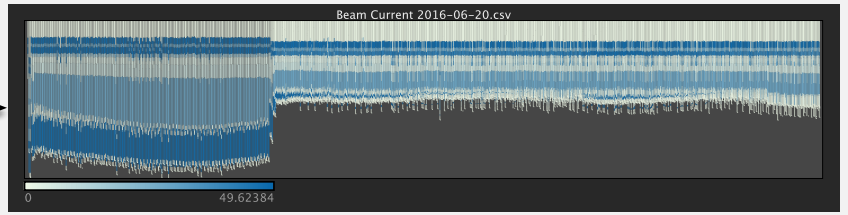
Falcon

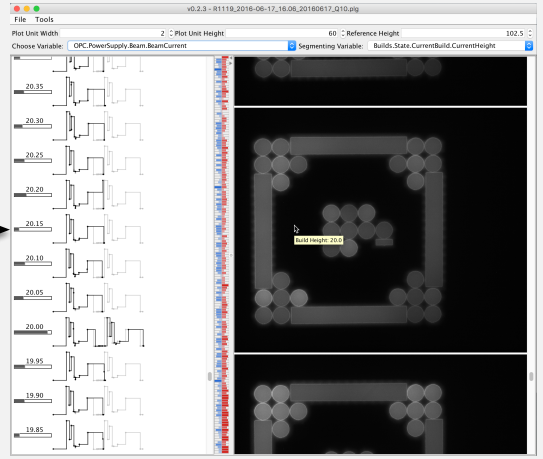
Visualizing Long, Irregularly Sampled Sensor Streams

Main Analysis Window

File / Variables Tree View

Waterfall Visualization





Segmented Time Series Visualization

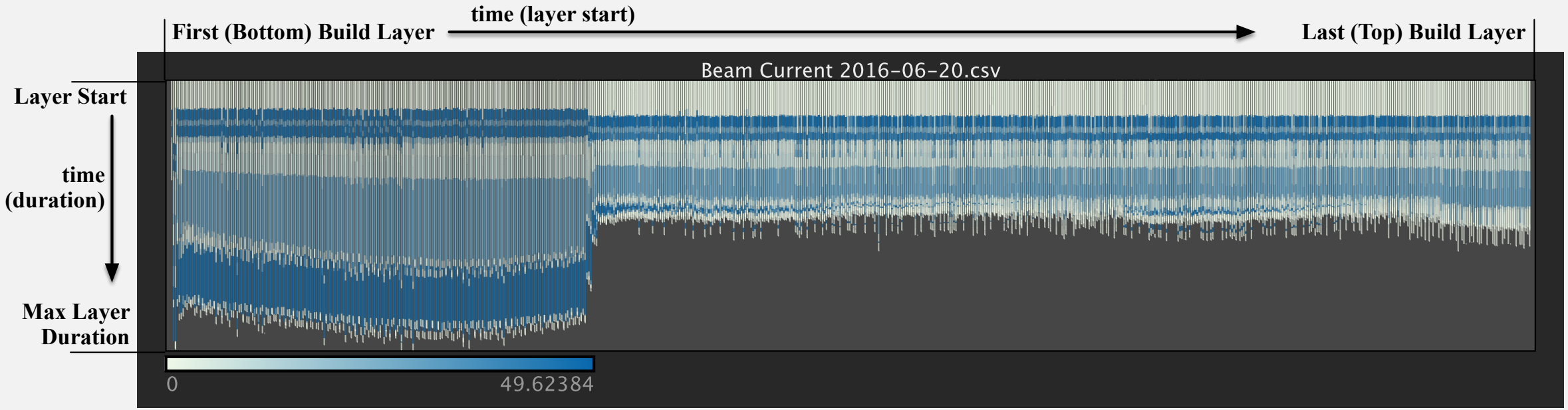
Settings Panel

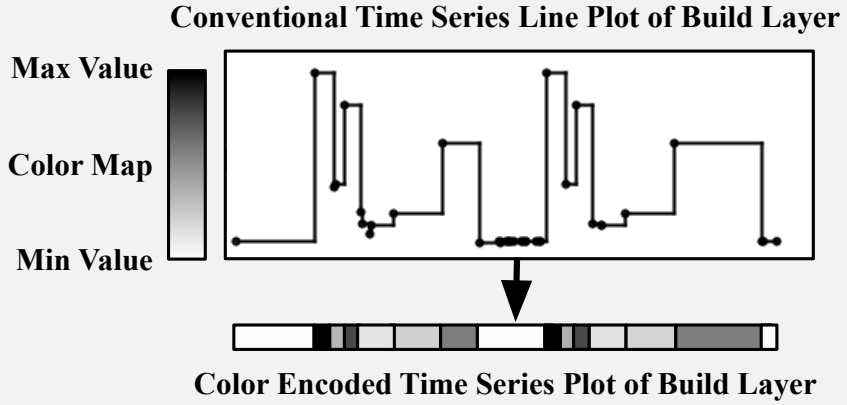
Variable Visualization Panel
(Left: detailed time series, Right: overview)

Selection Details Panel

Citation: "Falcon: Visual Analysis of Large, Irregularly Sampled, and Multivariate Time Series Data in Additive Manufacturing". C. A. Steed, W. Halsey, R. Dehoff, S. L. Yoder, V. Paquit, and S. Powers. *Computers & Graphics*, 63:50--64, 2017.

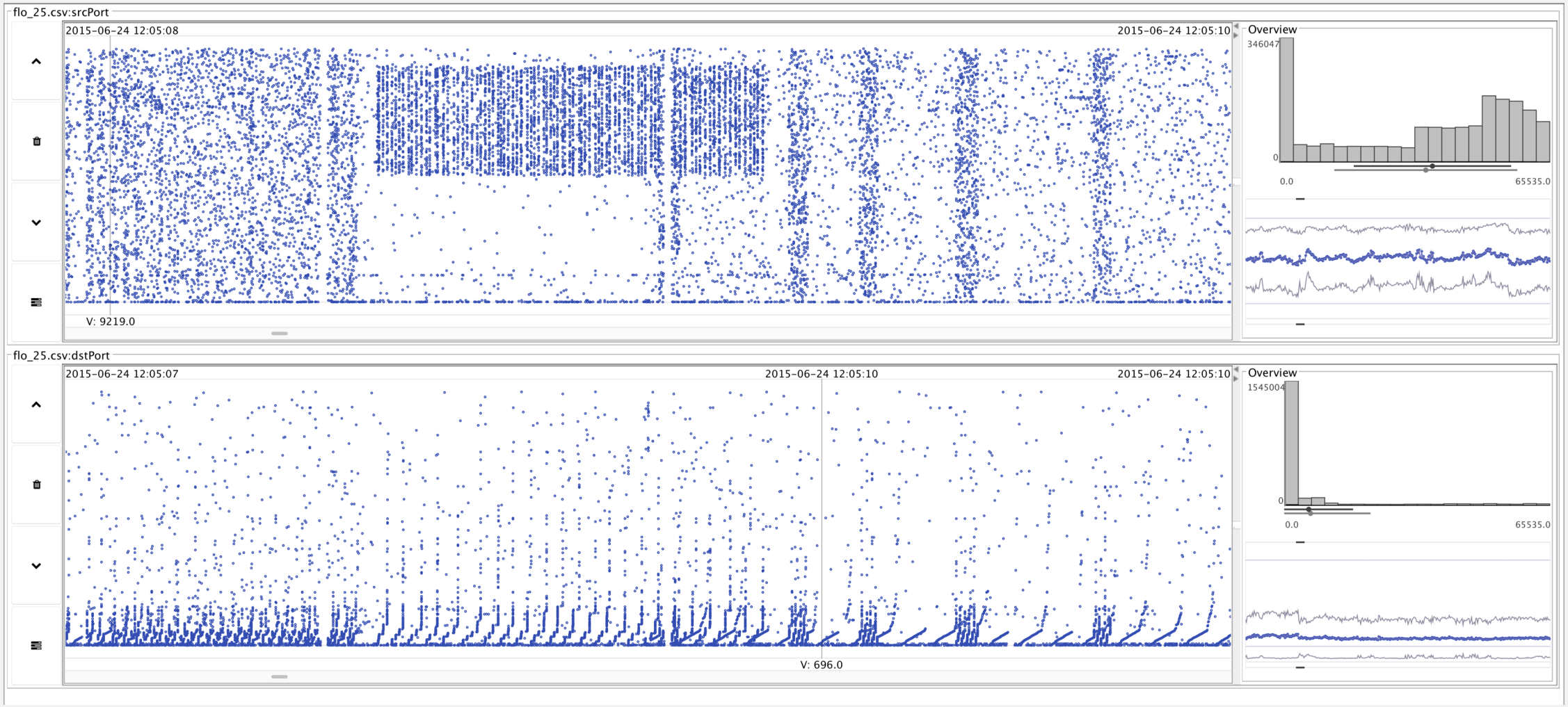
Falcon: Segmented Time Visualization





Citation: "Falcon: Visual Analysis of Large, Irregularly Sampled, and Multivariate Time Series Data in Additive Manufacturing". C. A. Steed, W. Halsey, R. Dehoff, S. L. Yoder, V. Paquit, and S. Powers. *Computers & Graphics*, 63:50--64, 2017.

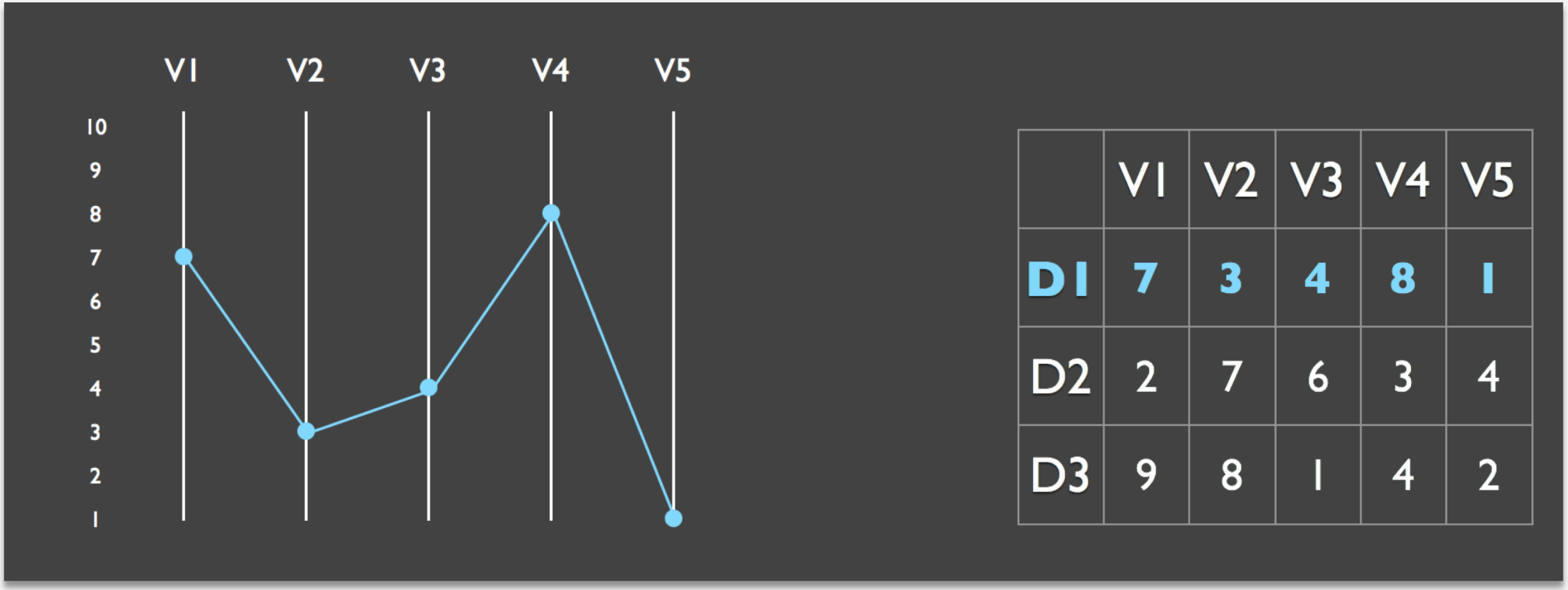
Falcon: Network Flow Information



Multivariate Data Visualization

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.

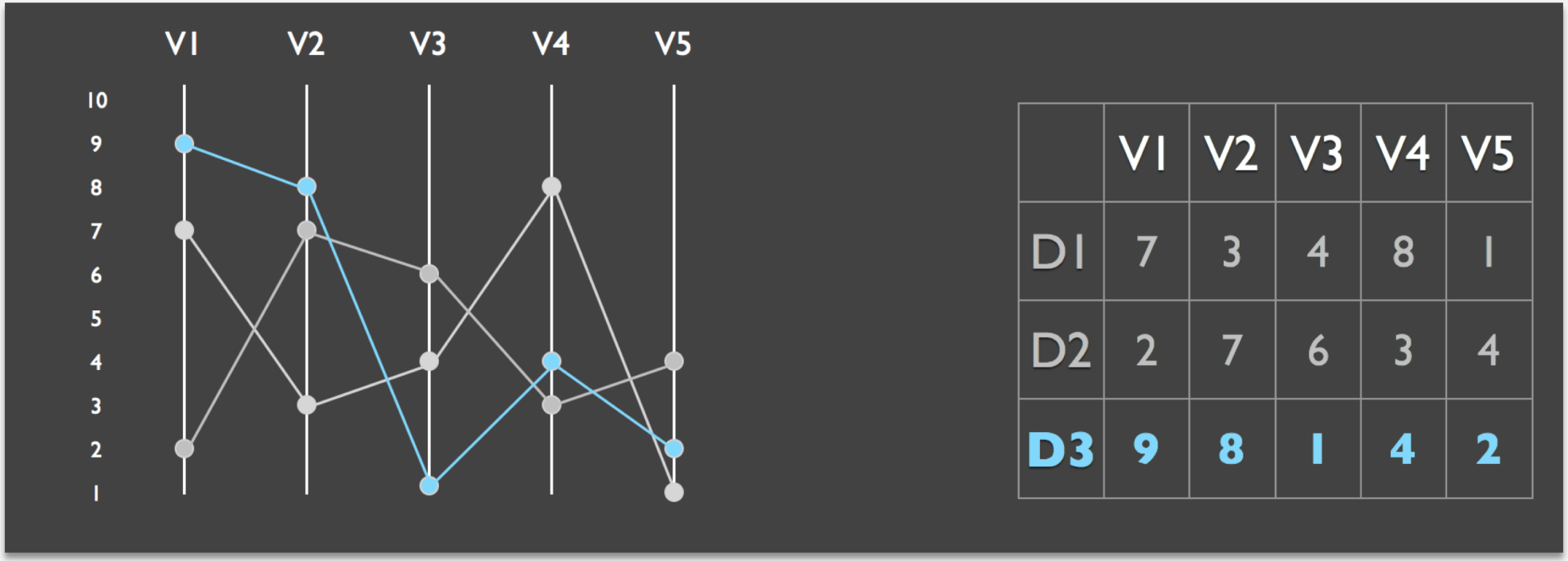


	V1	V2	V3	V4	V5
D1	**7**	**3**	**4**	**8**	**1**
D2	2	7	6	3	4
D3	9	8	1	4	2

Inselberg, 1985 | Wegman, 1990

Multivariate Data Visualization with Parallel Coordinates

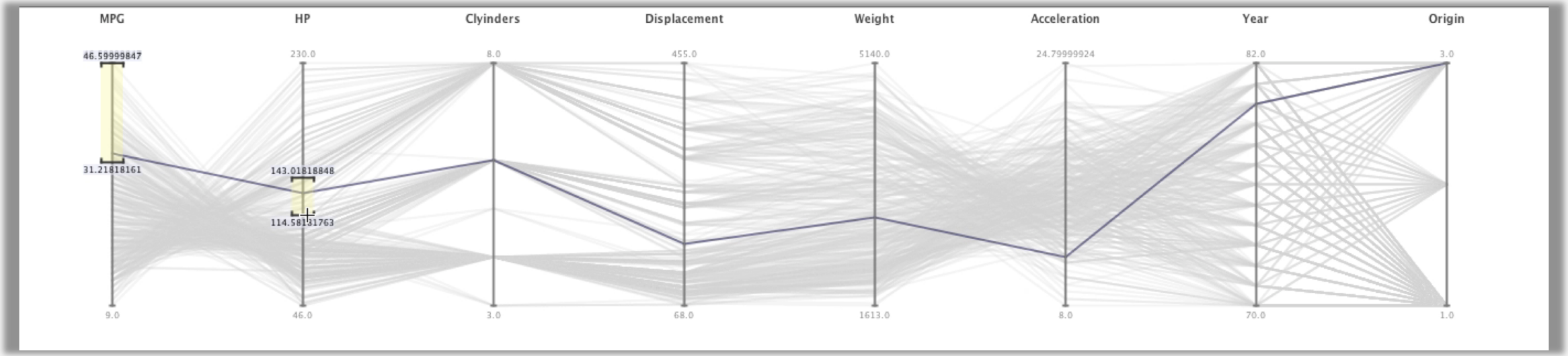
- Number of variables only limited by display resolution
- Two-dimensional view of high-dimensional data
- Avoids information loss -- no data reduction



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

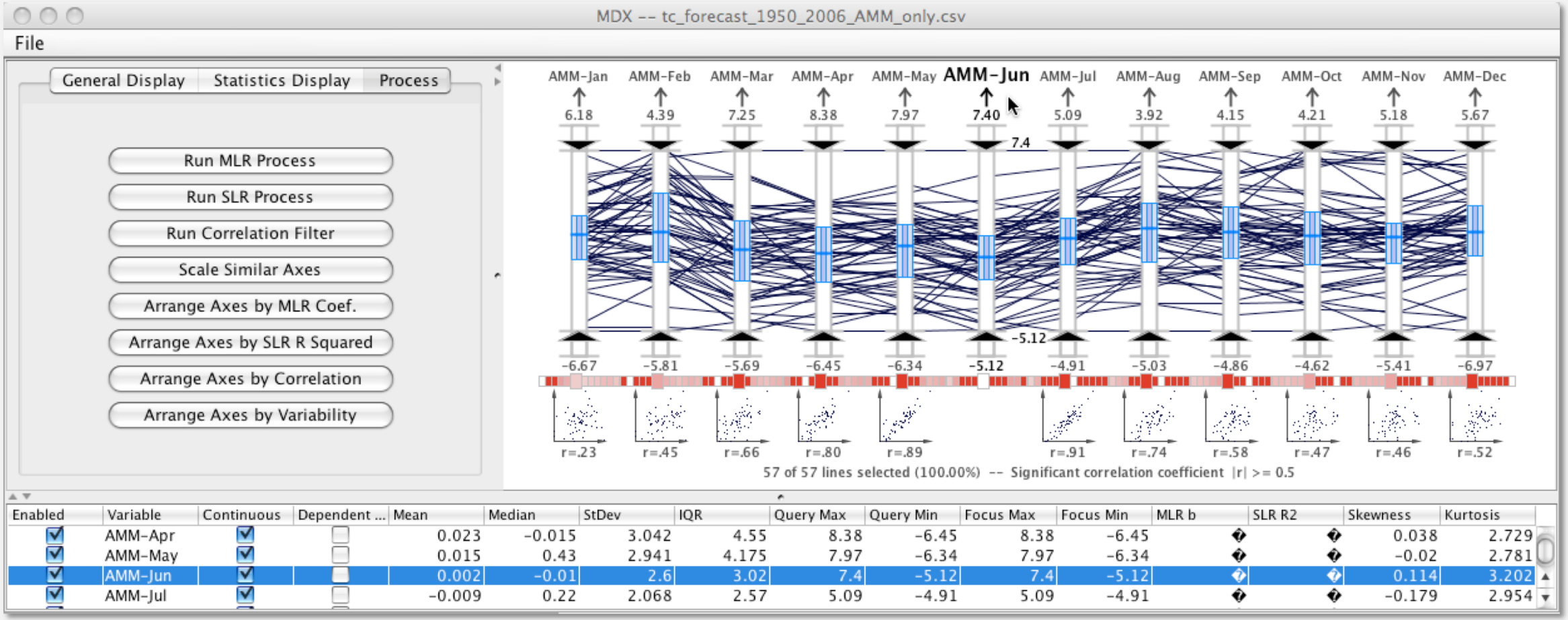
The Power of ||-coords

- Visual queries for asking questions of the data
- Correlations (line slopes), clusters, and trends visually revealed



Classical Parallel Coordinates Plot of the ‘83 ASA Cars Dataset
(Highlighted polyline shows Datsun 280ZX, 32.7 MPG & 132 HP)

Multidimensional Data eXplorer (MDX)



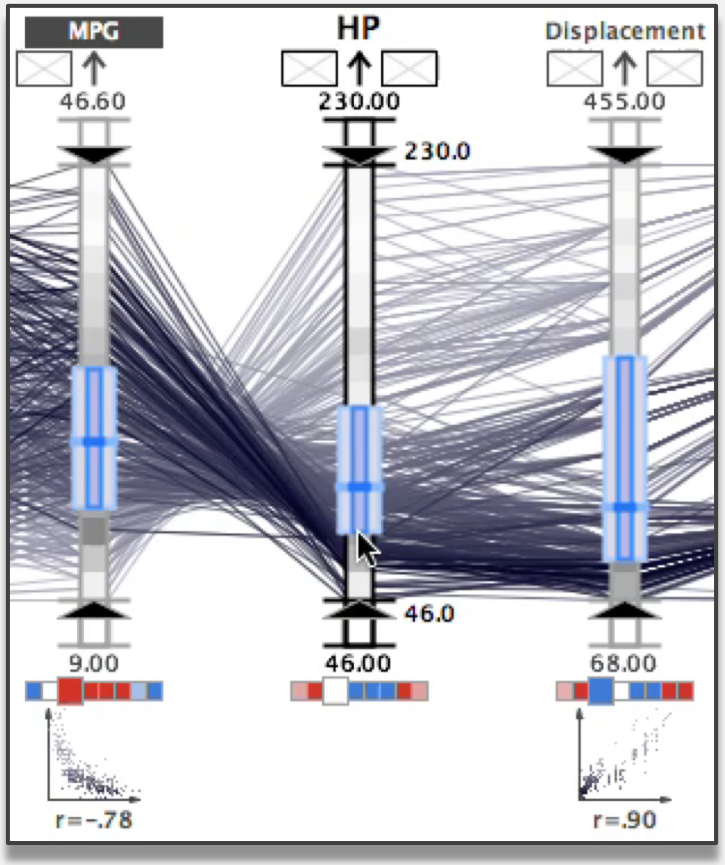
Enabled	Variable	Continuous	Dependent ...	Mean	Median	StDev	IQR	Query Max	Query Min	Focus Max	Focus Min	MLR b	SLR R2	Skewness	Kurtosis
☑	AMM-Apr	☑	☐	0.023	-0.015	3.042	4.55	8.38	-6.45	8.38	-6.45			0.038	2.729
☑	AMM-May	☑	☐	0.015	0.43	2.941	4.175	7.97	-6.34	7.97	-6.34			-0.02	2.781
☑	AMM-Jun	☑	☐	0.002	-0.01	2.6	3.02	7.4	-5.12	7.4	-5.12			0.114	3.202
☑	AMM-Jul	☑	☐	-0.009	0.22	2.068	2.57	5.09	-4.91	5.09	-4.91			-0.179	2.954

Citation: "Guided Analysis of Hurricane Trends using Statistical Processes Integrated with Interactive Parallel Coordinates". C. A. Steed, J. E. Swan II, T.J. Jankun-Kelly, and P. J. Fitzpatrick. Proceedings of the *Symposium on Visual Analytics Science and Technology (VAST)*, pp. 19–26, Oct. 2009.

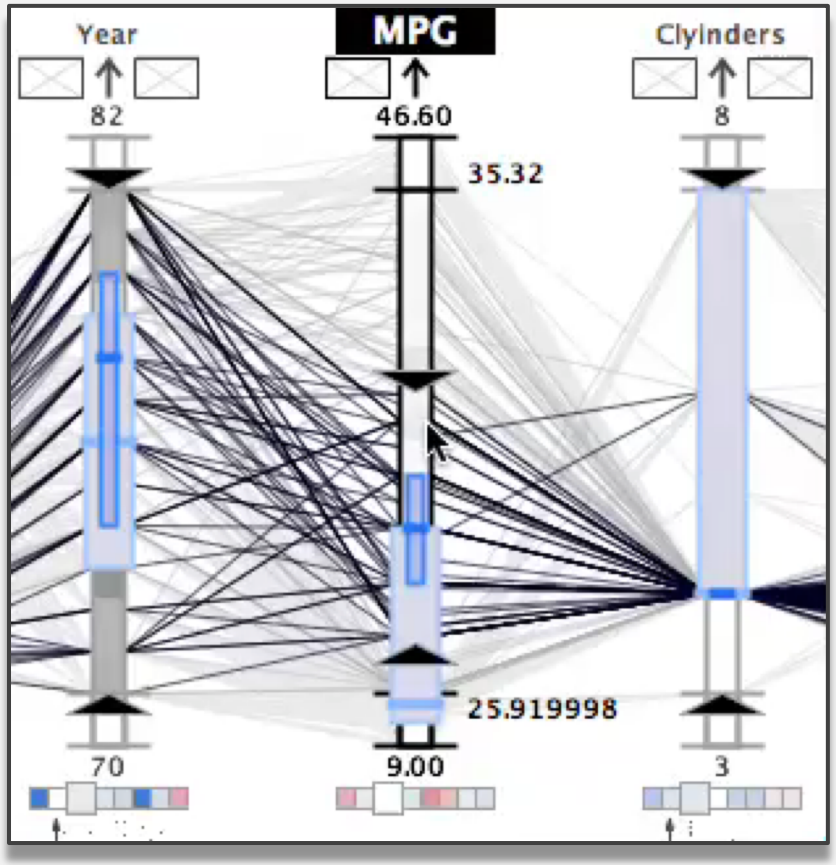
New Interaction Techniques for ||-coords

Dynamic Visual Queries

Aerial Perspective



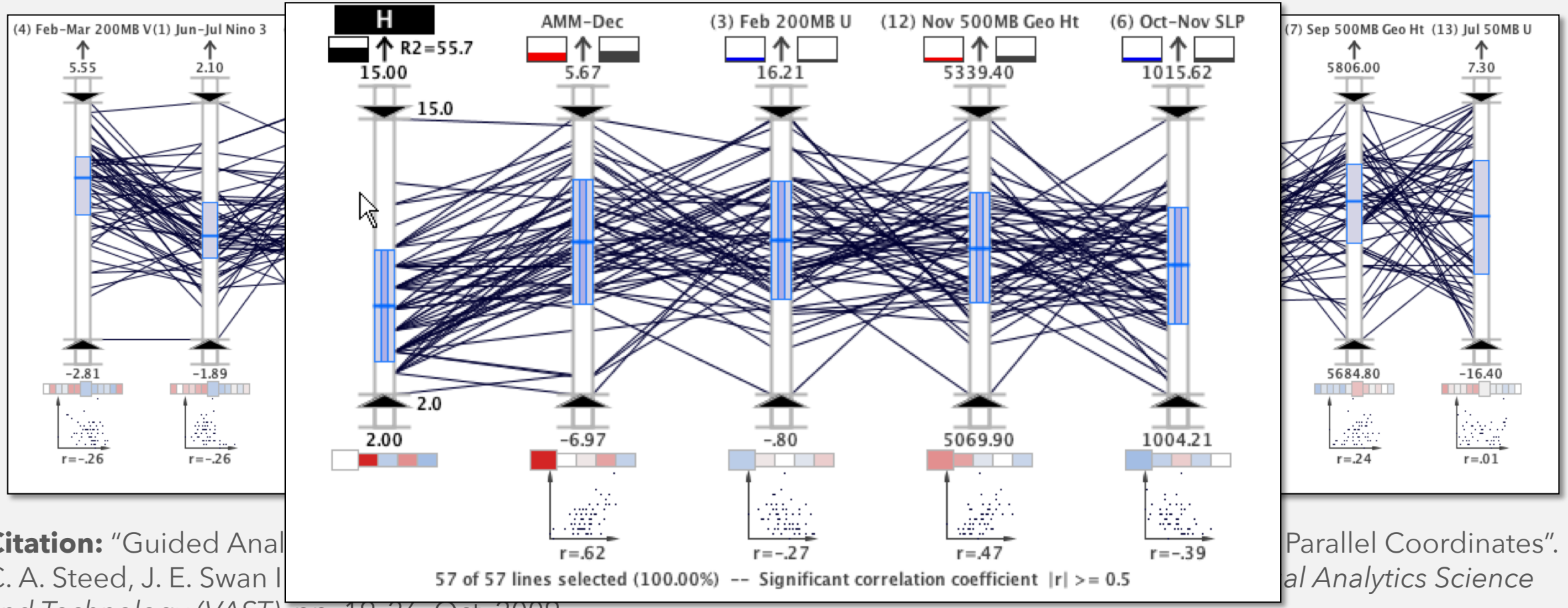
Dynamic Axis Scaling



Citation: "Guided Analysis of Hurricane Trends using Statistical Processes Integrated with Interactive Parallel Coordinates". C. A. Steed, J. E. Swan II, T.J. Jankun-Kelly, and P. J. Fitzpatrick. Proceedings of the *Symposium on Visual Analytics Science and Technology (VAST)*, pp. 19–26, Oct. 2009.

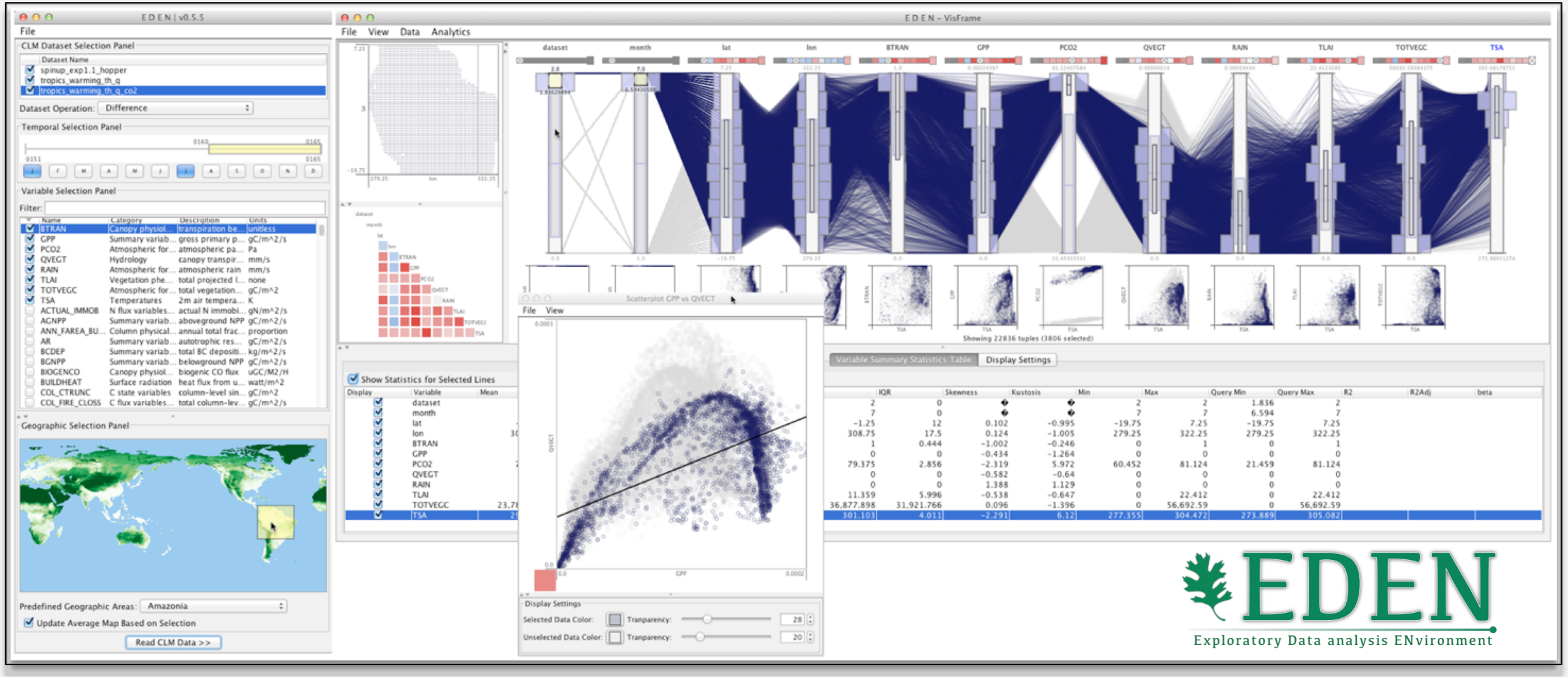
MDX: Regression and Filters

$$(y - \overline{y})/\sigma_y = \sum_{i=1}^{k} b_i(x_i - \overline{x}_i)/\sigma_i$$



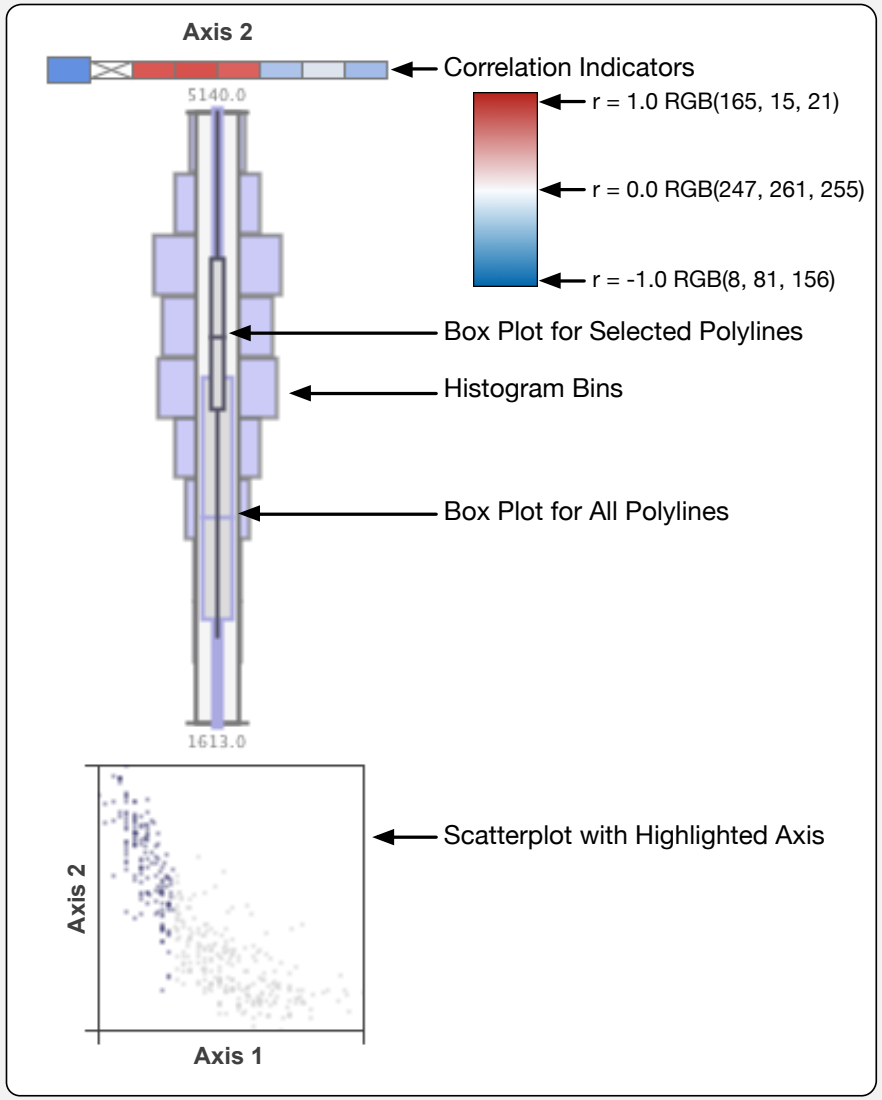
Citation: "Guided Anal Parallel Coordinates". C. A. Steed, J. E. Swan I *al Analytics Science and Technology (VAST)*, pp. 19-26, Oct. 2009.

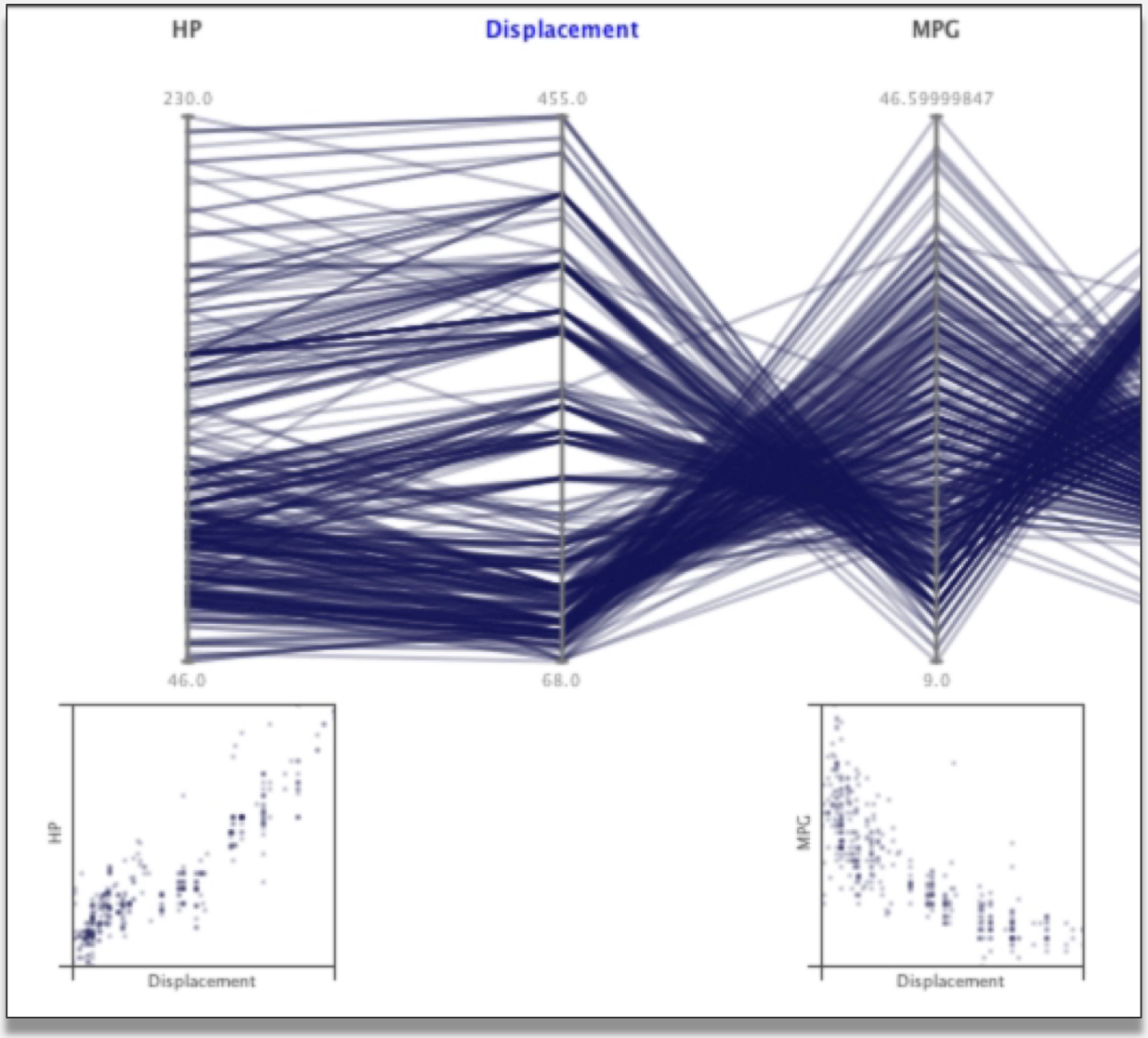
Exploratory Data analysis ENvironment



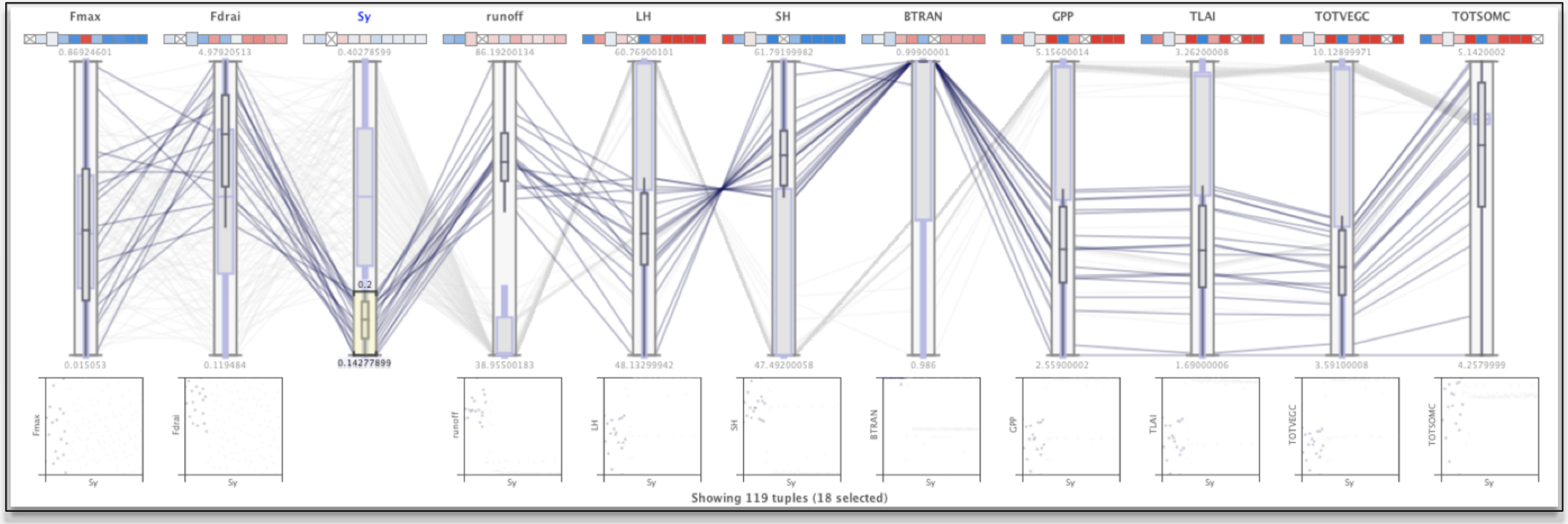
Citation: "Big Data Visual Analytics for Exploratory Earth System Simulation Analysis". C.A. Steed, D. M. Ricciuto, G. Shipman, B. Smith, P. E. Thornton, D. Wang, and D. N. Williams. *Computers & Geosciences*, 61:71-82, 2013.

Guided Analysis Using Information "Scent"



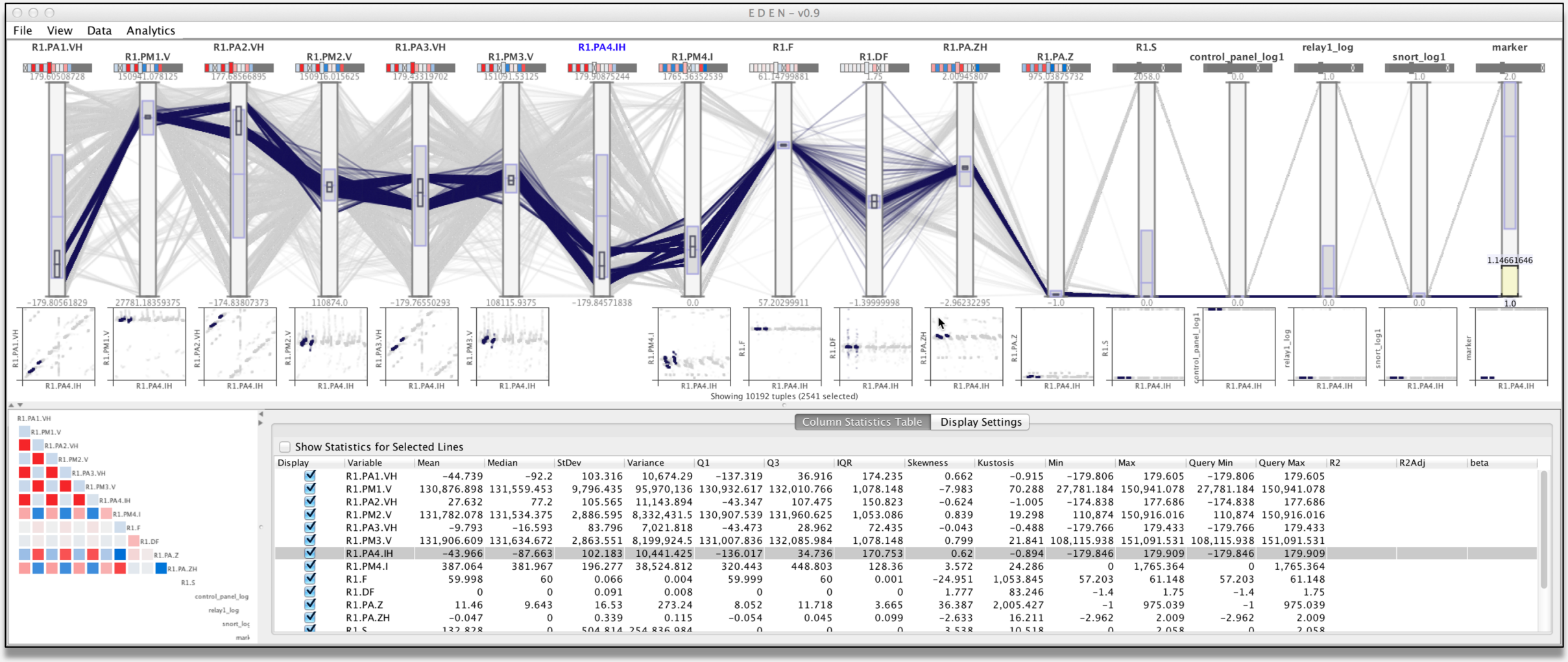


EDEN: Climate Ensemble Analysis



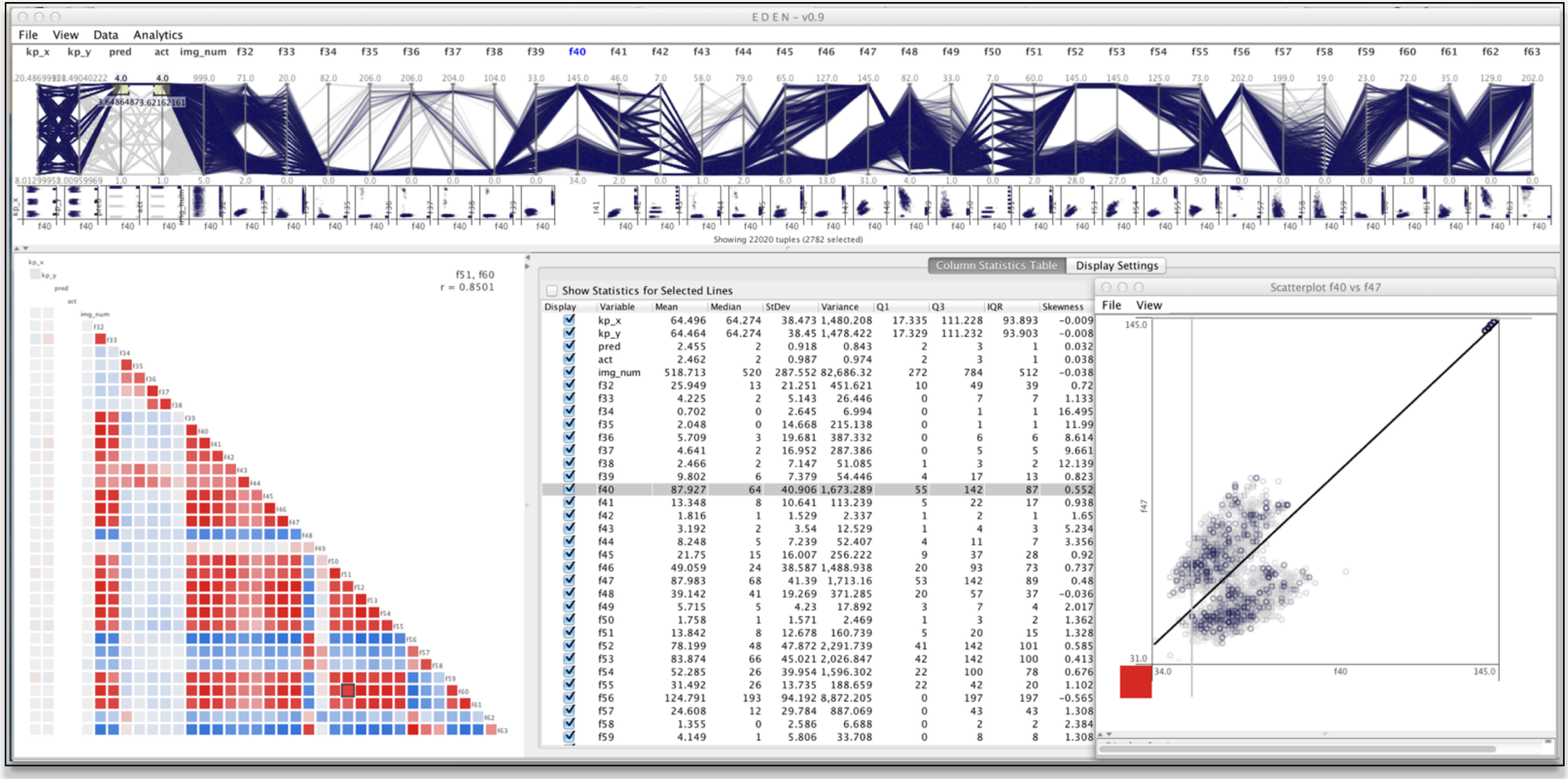
Citation: "Practical Application of Parallel Coordinates for Climate Model Analysis". Chad A. Steed, Galen Shipman, Peter Thornton, Daniel Ricciuto, David Erickson, and Marcia Branstetter. Proceedings of the *International Conference on Computational Science*, pp. 877–886, June 2012.

EDEN: Cyber Physical Data Analysis



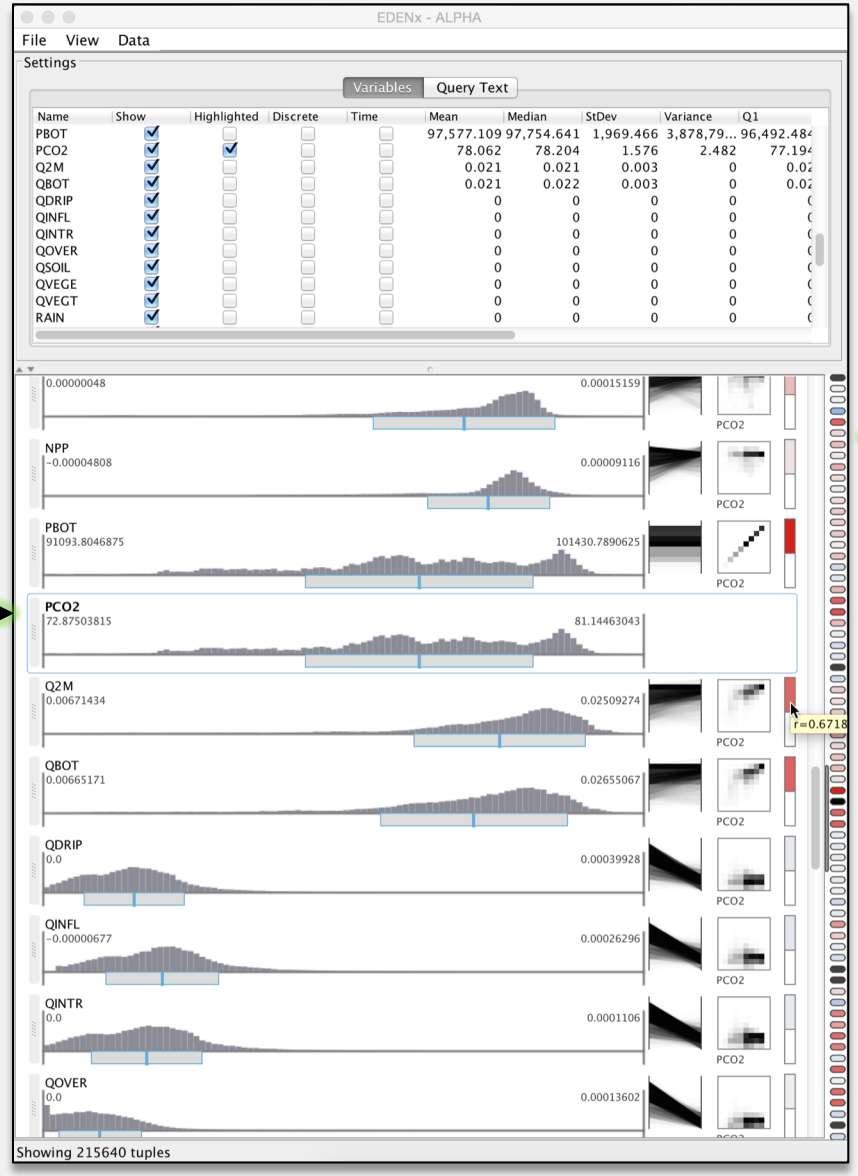
Display	Variable	Mean	Median	StDev	Variance	Q1	Q3	IQR	Skewness	Kustosis	Min	Max	Query Min	Query Max	R2	R2Adj	beta
☑	R1.PA1.VH	-44.739	-92.2	103.316	10,674.29	-137.319	36.916	174.235	0.662	-0.915	-179.806	179.605	-179.806	179.605			
☑	R1.PM1.V	130,876.898	131,559.453	9,796.435	95,970,136	130,932.617	132,010.766	1,078.148	-7.983	70.288	27,781.184	150,941.078	27,781.184	150,941.078			
☑	R1.PA2.VH	27.632	77.2	105.565	11,143.894	-43.347	107.475	150.823	-0.624	-1.005	-174.838	177.686	-174.838	177.686			
☑	R1.PM2.V	131,782.078	131,534.375	2,886.595	8,332,431.5	130,907.539	131,960.625	1,053.086	0.839	19.298	110,874	150,916.016	110,874	150,916.016			
☑	R1.PA3.VH	-9.793	-16.593	83.796	7,021.818	-43.473	28.962	72.435	-0.043	-0.488	-179.766	179.433	-179.766	179.433			
☑	R1.PM3.V	131,906.609	131,634.672	2,863.551	8,199,924.5	131,007.836	132,085.984	1,078.148	0.799	21.841	108,115.938	151,091.531	108,115.938	151,091.531			
☑	R1.PA4.IH	-43.966	-87.663	102.183	10,441.425	-136.017	34.736	170.753	0.62	-0.894	-179.846	179.909	-179.846	179.909			
☑	R1.PM4.I	387.064	381.967	196.277	38,524.812	320.443	448.803	128.36	3.572	24.286	0	1,765.364	0	1,765.364			
☑	R1.F	59.998	60	0.066	0.004	59.999	60	0.001	-24.951	1,053.845	57.203	61.148	57.203	61.148			
☑	R1.DF	0	0	0.091	0.008	0	0	0	1.777	83.246	-1.4	1.75	-1.4	1.75			
☑	R1.PA.Z	11.46	9.643	16.53	273.24	8.052	11.718	3.665	36.387	2,005.427	-1	975.039	-1	975.039			
☑	R1.PA.ZH	-0.047	0	0.339	0.115	-0.054	0.045	0.099	-2.633	16.211	-2.962	2.009	-2.962	2.009			
☑	R1.S	132.828	0	504.814	254,836.984	0	0	0	3.538	10.518	0	2,058	0	2,058			

EDEN: Machine Learning Feature Space



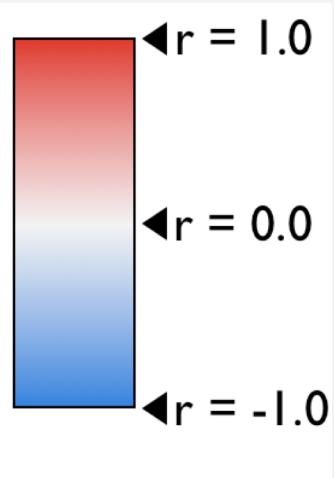
Display	Variable	Mean	Median	StDev	Variance	Q1	Q3	IQR	Skewness
☑	kp_x	64.496	64.274	38.473	1,480.208	17.335	111.228	93.893	-0.009
☑	kp_y	64.464	64.274	38.45	1,478.422	17.329	111.232	93.903	-0.008
☑	pred	2.455	2	0.918	0.843	2	3	1	0.032
☑	act	2.462	2	0.987	0.974	2	3	1	0.038
☑	img_num	518.713	520	287.552	82,686.32	272	784	512	-0.038
☑	f32	25.949	13	21.251	451.621	10	49	39	0.72
☑	f33	4.225	2	5.143	26.446	0	7	7	1.133
☑	f34	0.702	0	2.645	6.994	0	1	1	16.495
☑	f35	2.048	0	14.668	215.138	0	1	1	11.99
☑	f36	5.709	3	19.681	387.332	0	6	6	8.614
☑	f37	4.641	2	16.952	287.386	0	5	5	9.661
☑	f38	2.466	2	7.147	51.085	1	3	2	12.139
☑	f39	9.802	6	7.379	54.446	4	17	13	0.823
☑	f40	87.927	64	40.906	1,673.289	55	142	87	0.552
☑	f41	13.348	8	10.641	113.239	5	22	17	0.938
☑	f42	1.816	1	1.529	2.337	1	2	1	1.65
☑	f43	3.192	2	3.54	12.529	1	4	3	5.234
☑	f44	8.248	5	7.239	52.407	4	11	7	3.356
☑	f45	21.75	15	16.007	256.222	9	37	28	0.92
☑	f46	49.059	24	38.587	1,488.938	20	93	73	0.737
☑	f47	87.983	68	41.39	1,713.16	53	142	89	0.48
☑	f48	39.142	41	19.269	371.285	20	57	37	-0.036
☑	f49	5.715	5	4.23	17.892	3	7	4	2.017
☑	f50	1.758	1	1.571	2.469	1	3	2	1.362
☑	f51	13.842	8	12.678	160.739	5	20	15	1.328
☑	f52	78.199	48	47.872	2,291.739	41	142	101	0.585
☑	f53	83.874	66	45.021	2,026.847	42	142	100	0.413
☑	f54	52.285	26	39.954	1,596.302	22	100	78	0.676
☑	f55	31.492	26	13.735	188.659	22	42	20	1.102
☑	f56	124.791	193	94.192	8,872.205	0	197	197	-0.565
☑	f57	24.608	12	29.784	887.069	0	43	43	1.308
☑	f58	1.355	0	2.586	6.688	0	2	2	2.384
☑	f59	4.149	1	5.806	33.708	0	8	8	1.308

EDENx: Handling Larger Data Sets

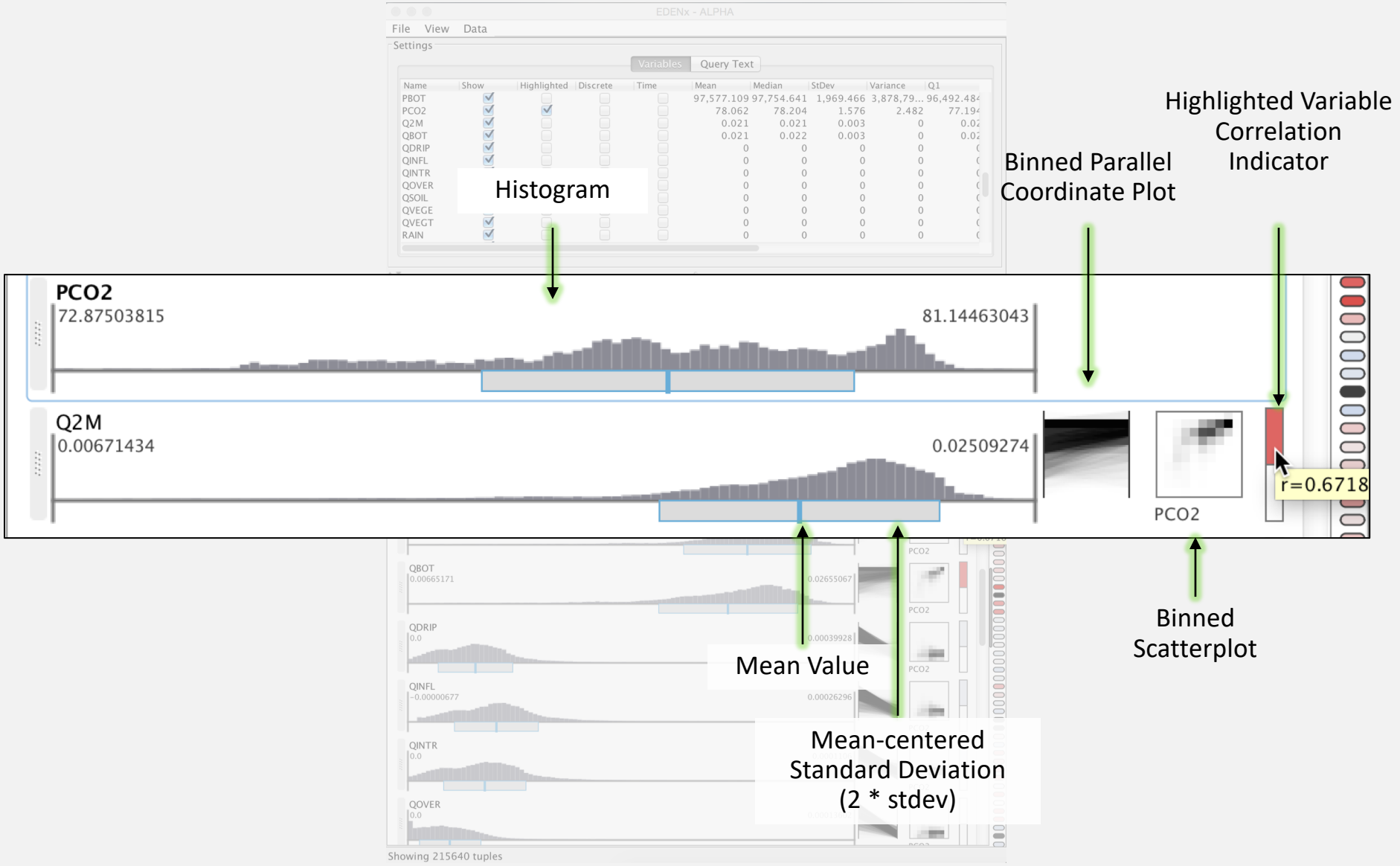


Highlighted Variable

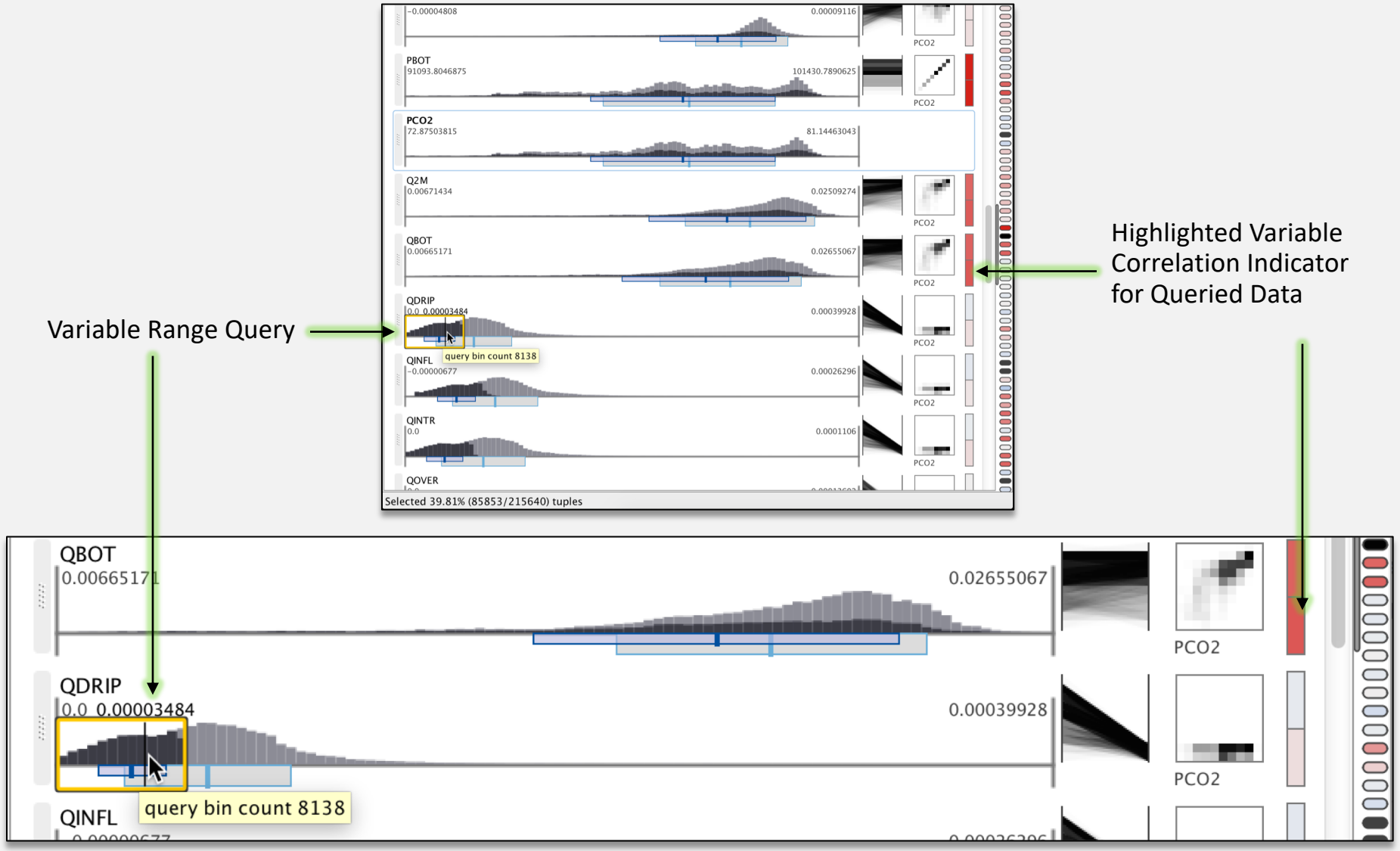
Correlation Indicators for all Variables



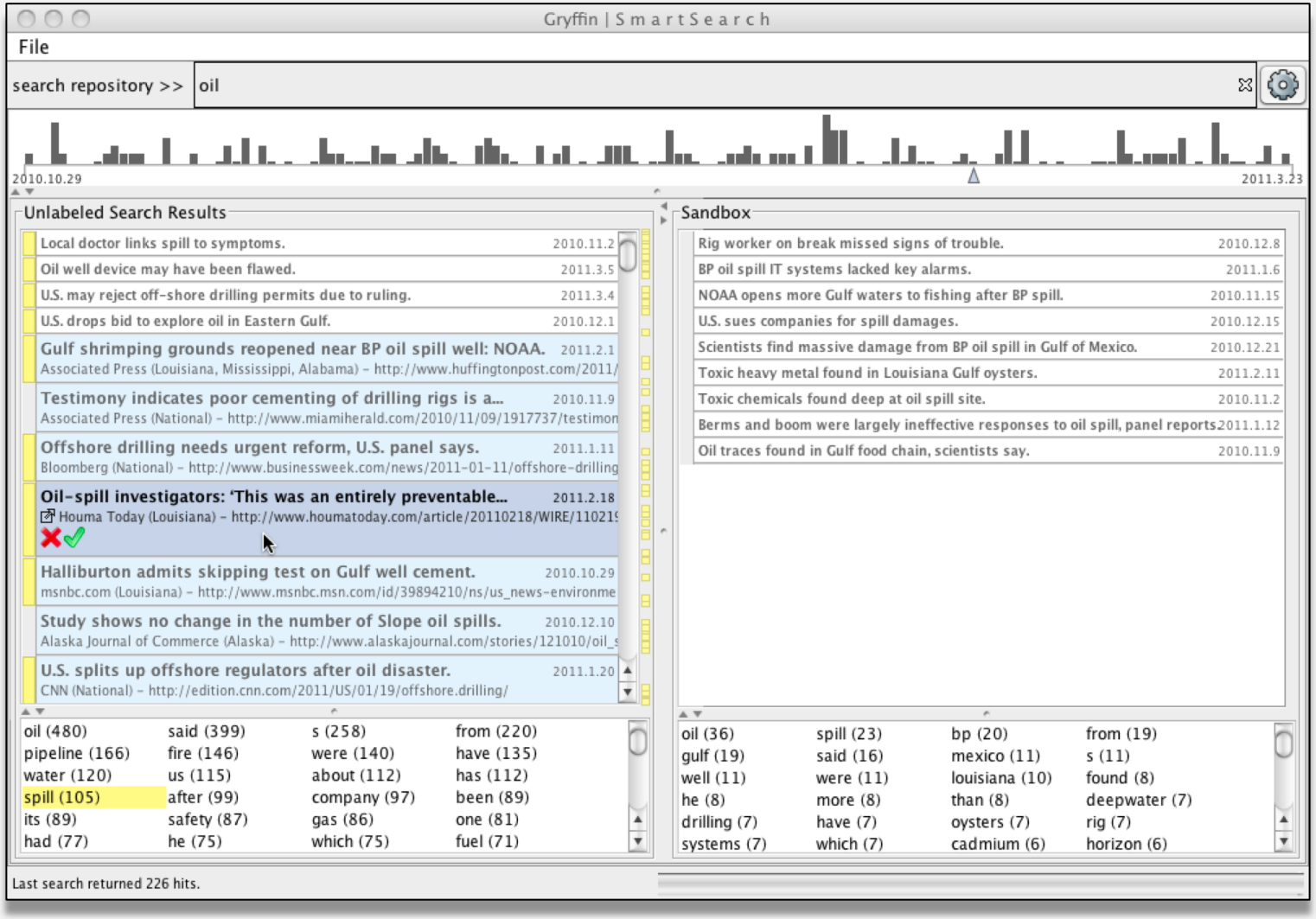
EDENx: Handling Larger Data Sets



EDENx: Handling Larger Data Sets



Gryffin: Guided Text Search

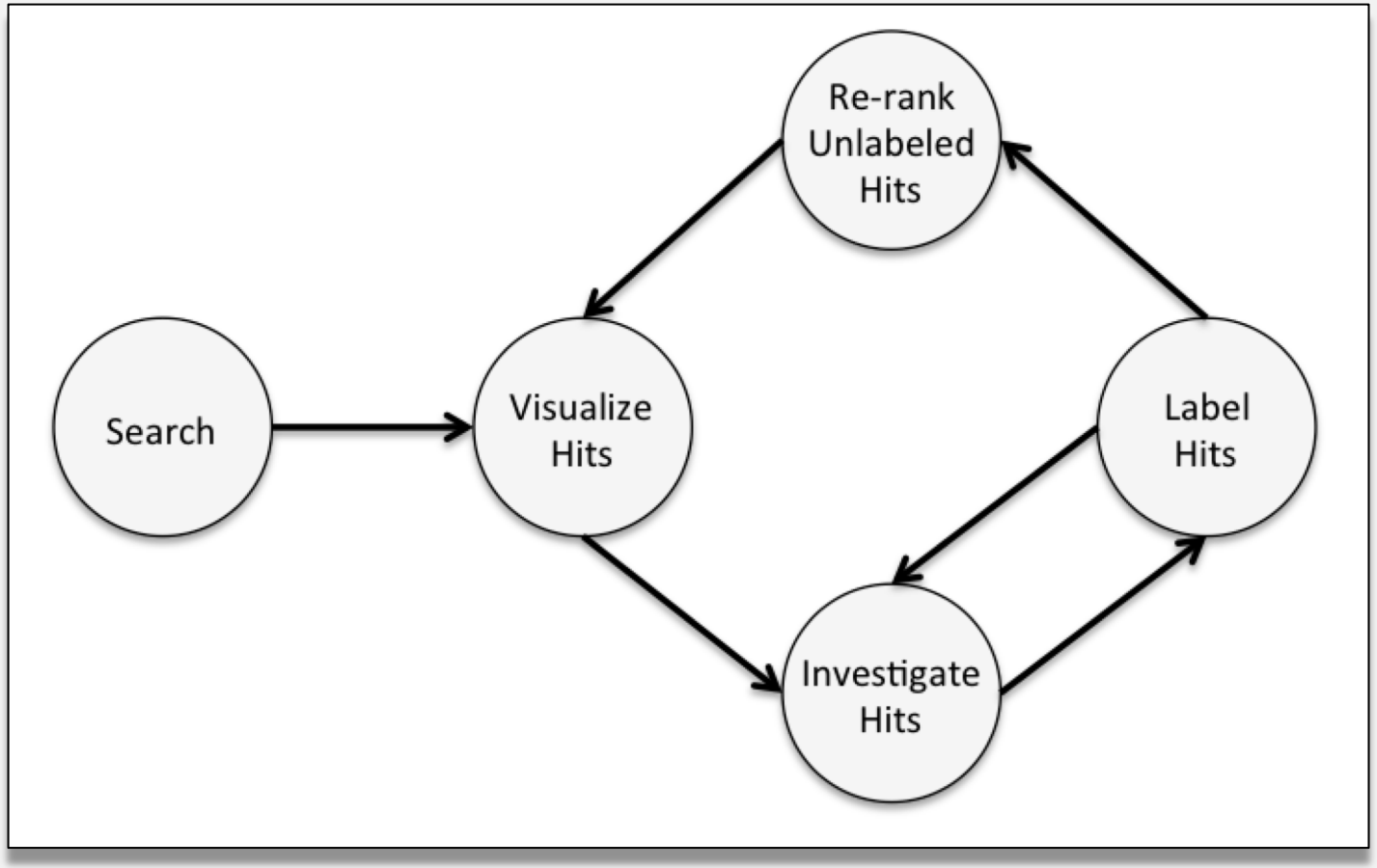


Citation: "Guided Text Analysis Using Adaptive Visual Analytics". C. A. Steed, C. Symons, F. DeNap, and T. E. Potok. Proceedings of the *Visualization and Data Analysis Conference*, Jan. 2012.

Gryffin: Intelligent User Interface

- **Workflow:** User focuses on the top of the list and labels documents as relevant / irrelevant

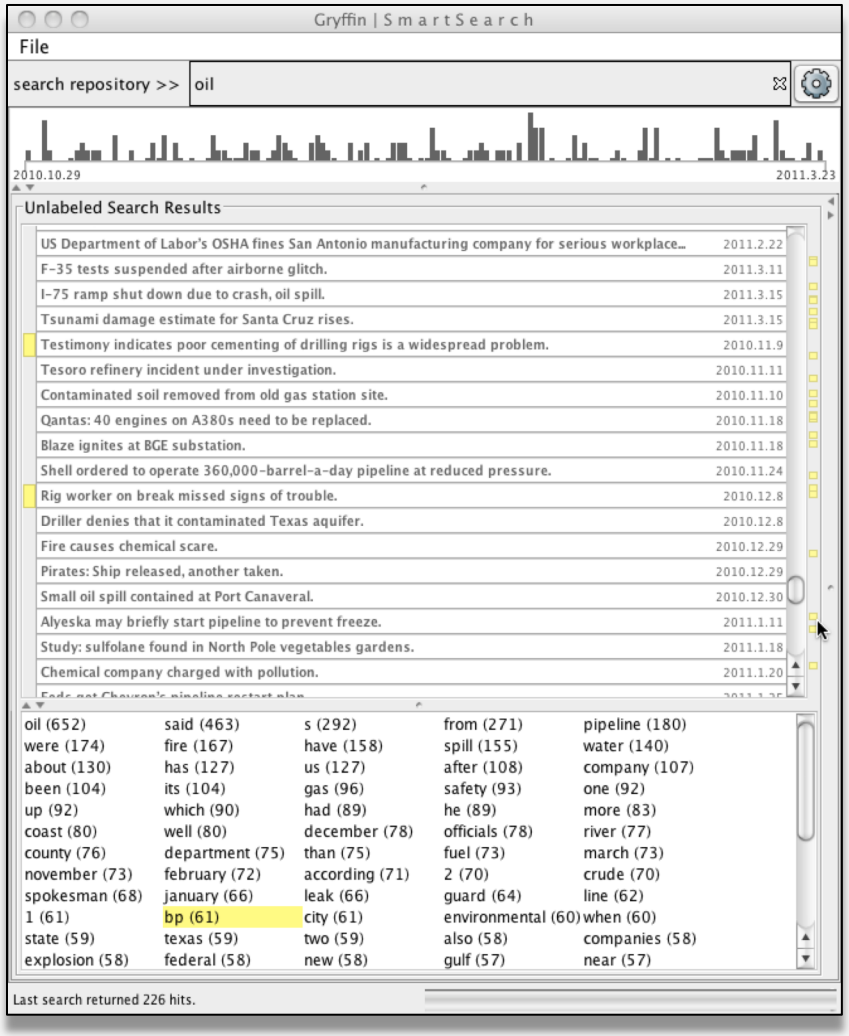
- A graph-based machine learning algorithm, developed by my colleague Chris Symons, uses labeled documents to re-rank list using document similarity



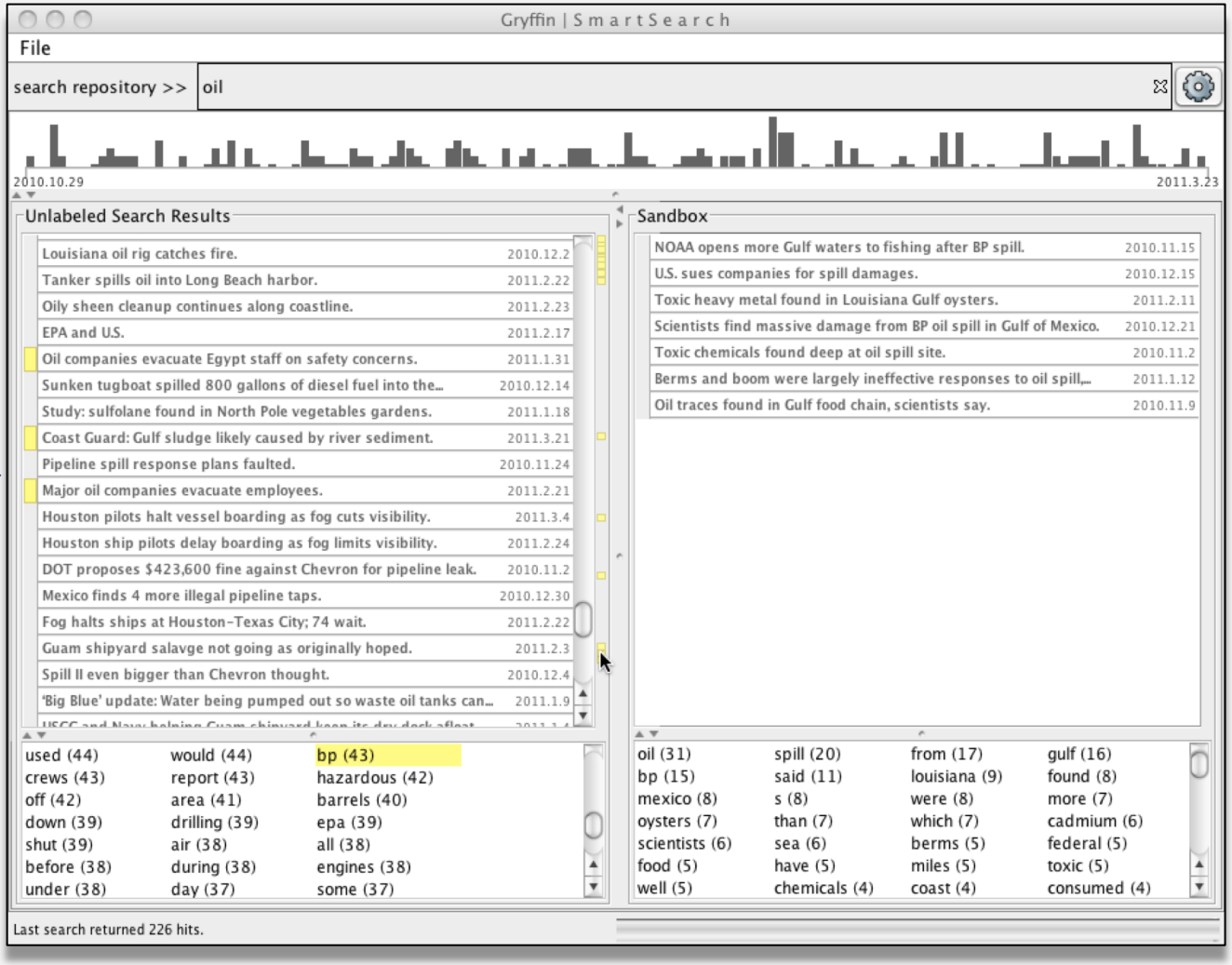
Citation: "Guided Text Analysis Using Adaptive Visual Analytics". C. A. Steed, C. Symons, F. DeNap, and T. E. Potok. Proceedings of the *Visualization and Data Analysis Conference*, Jan. 2012.

Gryffin: Deepwater Horizon Case Study

After Initial Search

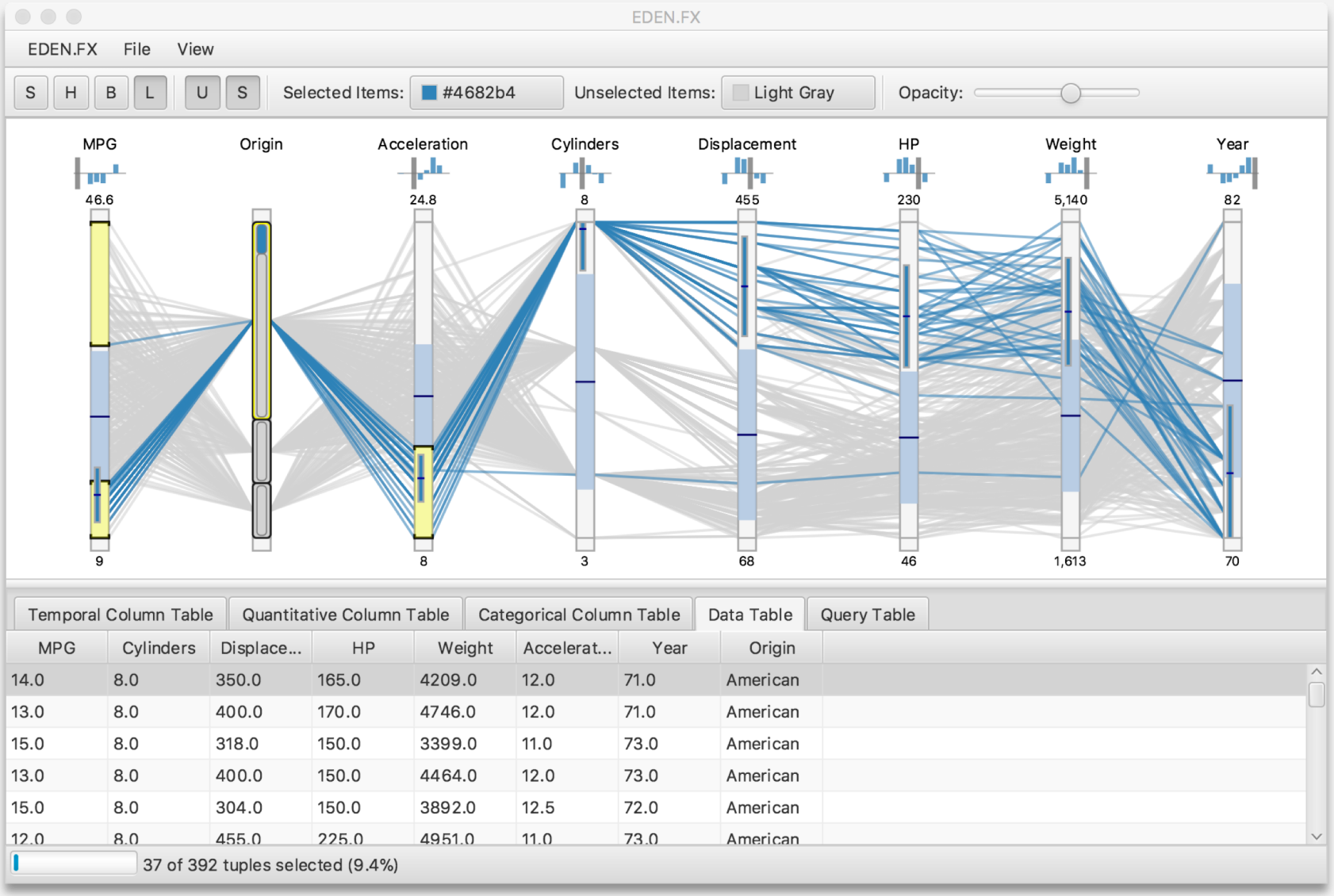


Label – Learn – Re-Rank



EDEN.fx: Scalable Multivariate Visual Analytics

- Combining EDEN, Falcon, Gryffin, and other systems
- Categorical, temporal axes
- Focus on scalability through level of detail rendering



EDEN.fx: Climate Simulation Analysis

Scientific Achievement

EDEN enables exploratory data analysis for new DOE E3SM climate simulation and observational data using techniques that combine interactive data visualization and statistical analytics.

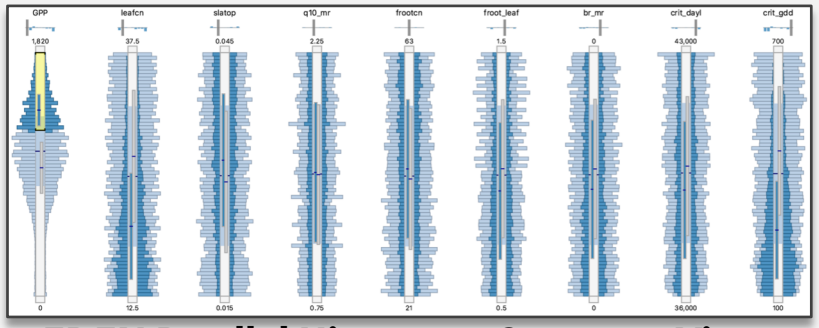
Significance and Impact

We are collaborating closely with Dan Ricciuto (ORNL) on his SciDAC BER application, "An Integrated System for Optimization of Sensor Networks to Improve Climate Model Predictions".

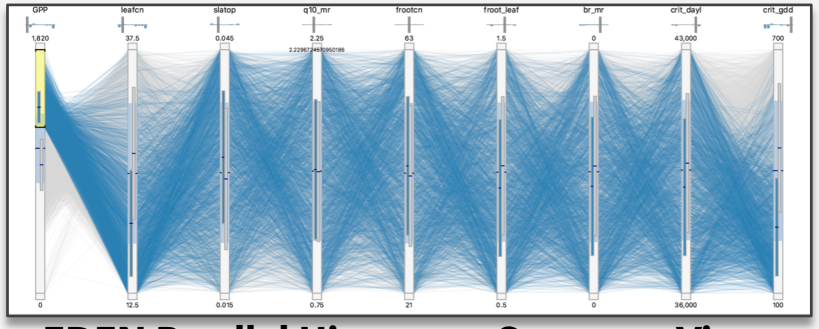
EDEN gives climate scientists the ability to consider more variables from large scale, land model parameter sensitivity analyses and ultimately improve DOE model accuracy.

Research Details

- The screenshots at left show one insight found during parameter sensitivity analysis for realistic values of GPP, a model output that measures photosynthesis in plants.
- The plots helped scientists see that high values of GPP are associated with low leaf carbon to nitrogen ratio values (leafcn) and low critical growing degree days (crit_gdd).
- Based on this insight, climate scientists will generate new ensembles covering smaller ranges of the leafcn and crit_gdd parameter space for more accurate surrogate models.



EDEN Parallel Histogram Summary View

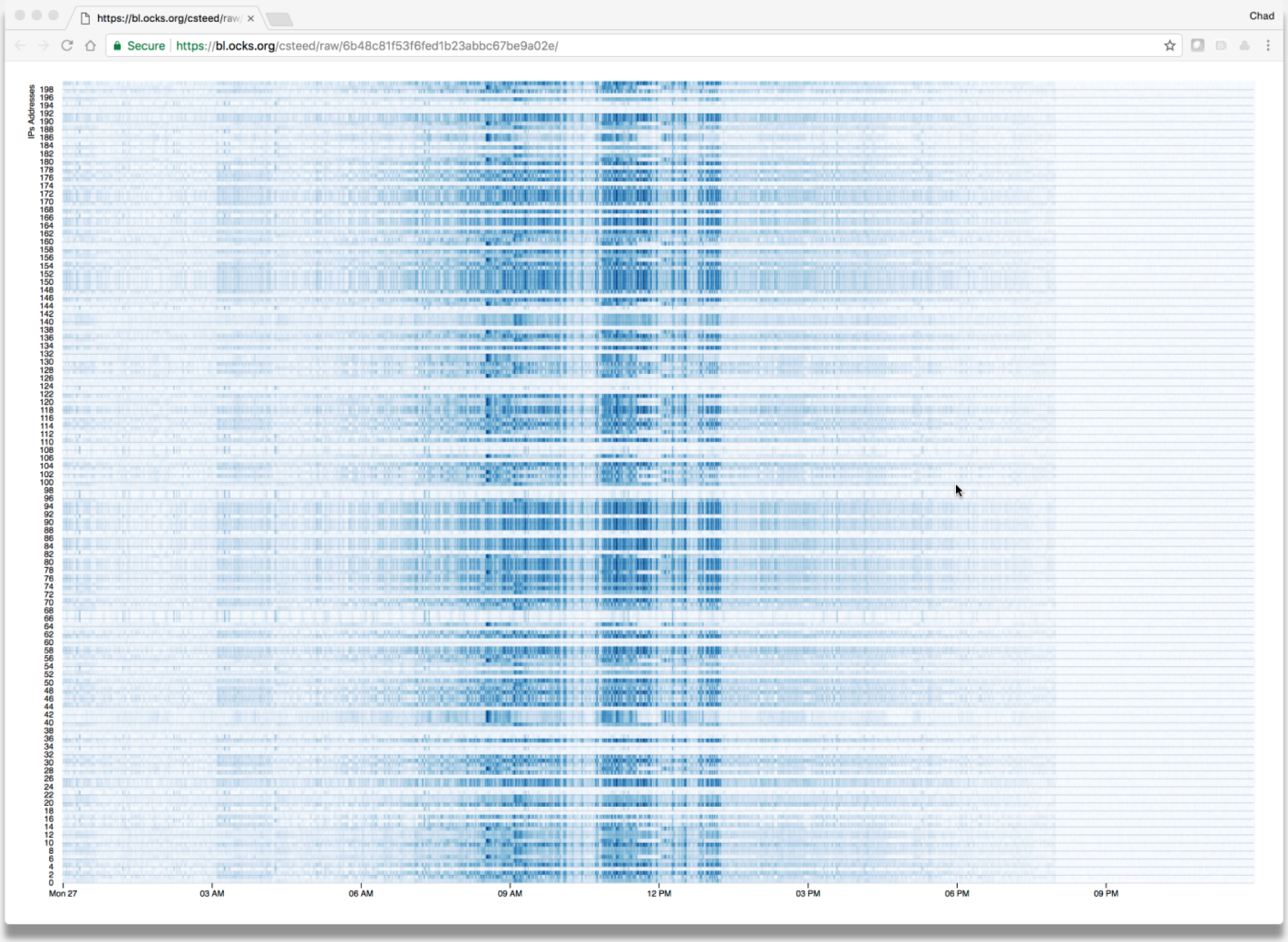


EDEN Parallel Histogram Summary View

The left axis, GPP, is a model output and the other 8 axes are parameters used to run model ensembles (each line is an ensemble run). In both views, the upper range of GPP values are selected revealing an association with low values of crit_gdd and leafcn. New model runs will be executed using more constrained values for these parameters.

Web-based Visualization: Overview IP/Stream Visualization

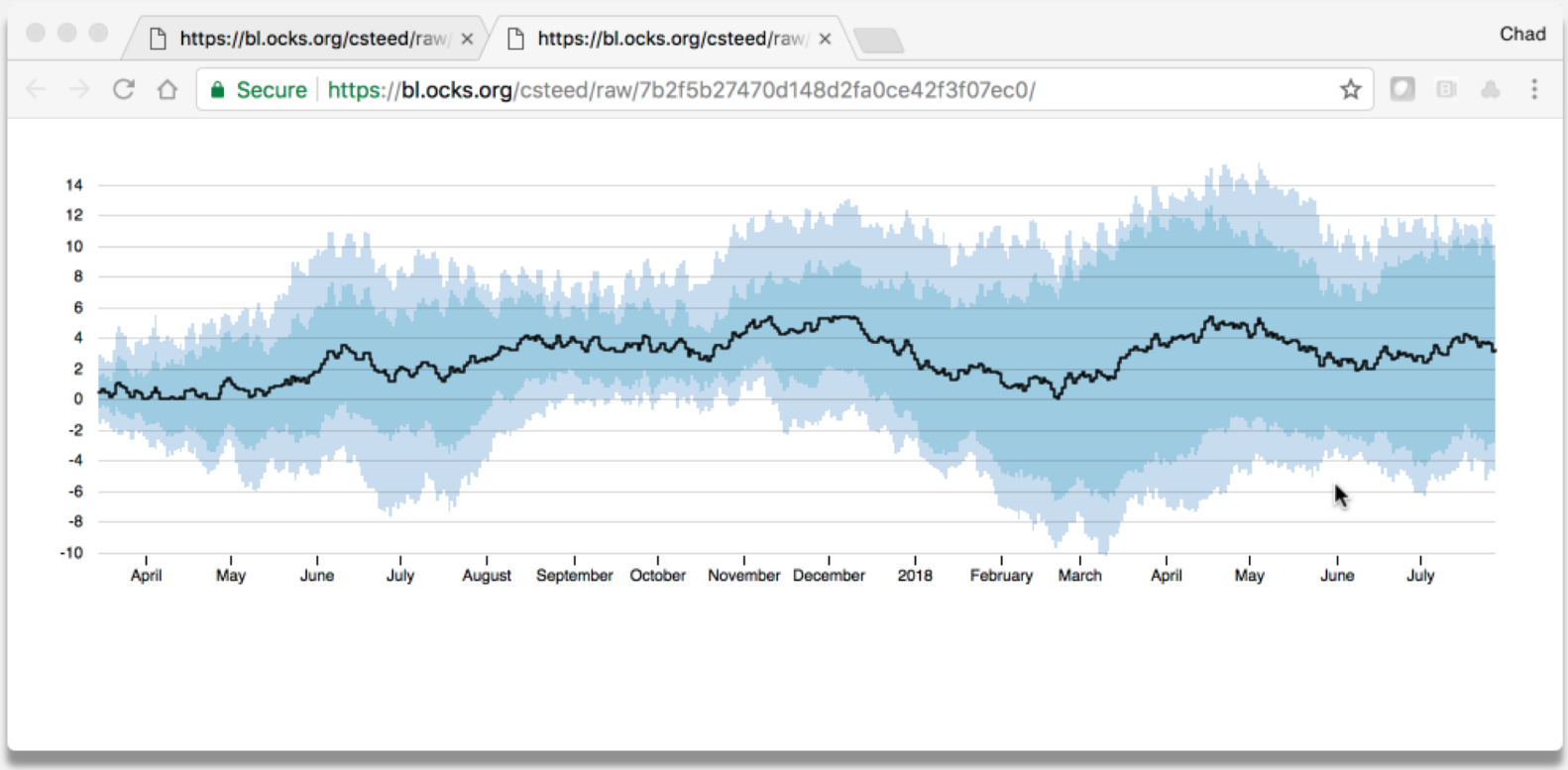
- **Goal**: Visualize aggregated time series data for hundreds of IP addresses
- Color scale encodes temporal bin statistics
- Support for zooming in to access more details (smaller time bins)



https://bl.ocks.org/csteed

Web-based Visualizations: Temporal Summary Visualization

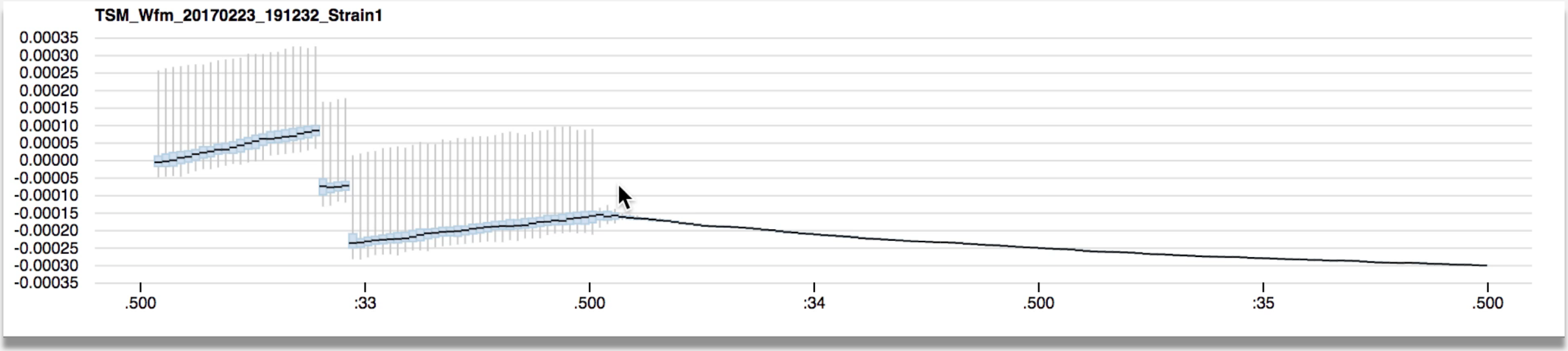
- **Goal:** Show time series typical values and variation in a single visualization
- Middle line encodes mean/median value
- Range polygons encode dispersion statistics
  - Standard deviation
  - Min/Max range
  - Interquartile range



https://bl.ocks.org/csteed

Web-based Visualizations: Segmented Time Series Data

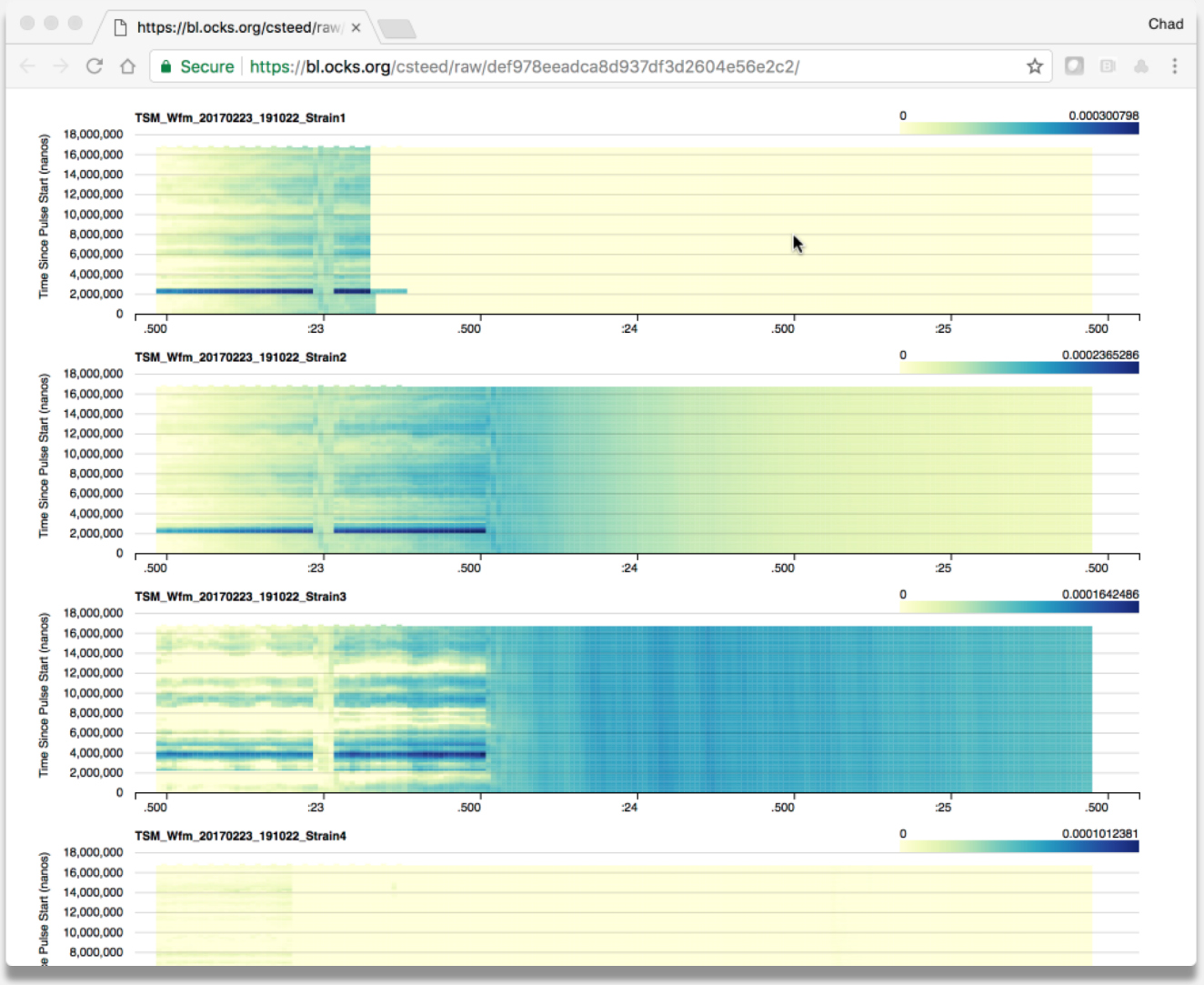
- **Goal:** Provide overview visualization of large scale sensor stream data as segments
- Data from strain sensors on ORNL SNS facility measuring neutron pulse activity
- Each vertical line represent range for pulse/segment
- Box represents standard deviation range or IQR for pulse/segment
- Black horizontal line represents mean or median for pulse/segment
- Interactively zoom in/out to see more/less details



https://bl.ocks.org/csteed

Web-based Visualizations: Temporal Summary Visualization

- **Goal:** Show detail time series visualizations for SNS pulse/segment data
- Mesoscale visualization between overview and raw data view
- Large scale data set at millisecond resolution binned and represented as color filled boxes



https://bl.ocks.org/csteed

Future Research Topics

- **Anticipatory Visual Analysis Models**
  - Human-in-the-loop
  - ML-based recommenders to guide analysis and reduce search space
- **Explainable Artificial Intelligence**
  - Data visualization to understand and fine-tune ML algorithms
- **New Data Visualization Techniques**
  - Focus on new ways to graphically represent abstract information (e.g., text, network flow, graphs, system performance)
- **Applications to Pressing National Challenges**
  - Cyber physical data analysis (e.g., streams, temporal + multivariate)
  - Analysis of network flow traffic (e.g., overview + detail)
  - Climate simulation and materials analysis