Using the Purdue DB Technology to build simple on-demand data exploration tools

Michael Grobe
Pervasive Technology Institute
Indiana University

The general idea

This presentation describes some of the work that came out of our collaboration with Ann-Christine Catlin's group at Purdue. The primary goal involved making data from Connie Weaver's Nutrional Science Lab at Purdue available on the Indiana CTSI Hub using the tools developed by Ann-Christine's group, which we took to calling the "Purdue database technology toolset."

At the time we were working on this project, the Purdue database technology was focused around several tools that are **very** useful for collecting and displaying data, and therefore for building integrated data base interfaces:

Dbparse: allows users to upload CSV files to backend database tables

DataView: displays contents of backend database tables

Photo_Gallery and the **Collections manager**: maintains and displays collections of images which may be displayed on-their-own or from within DataView views

com_form: a component for building web-based forms for data collection and display

The general idea (continued)

To use the DataViewer for data display, programmers simply prepare configuration files that identify database tables to be displayed, and define display appearance, including suitable plots.

However, for non-programming users, it seemed likely that a small set of ad hoc data exploration tools might be useful for getting a quick look at the data.

In particular, the following capabilities seem to constitute a basic package of capabilities:

- frequency tables,
- cross tabulations,
- simple correlations/regression plots, or
- simple correlation matrices.

The prototype and a question

The com_form and DataViewer components were used to provide a **PROTOTYPE** version of a new HubZero component that performs these functions (com_datashape).

com_form was used to build forms requesting and validating user input, and

com_dataview was used to display frequency charts and regression plots as directed at run-time. (Something of a mutation of the DataViewer model.)

So today's presentation will:

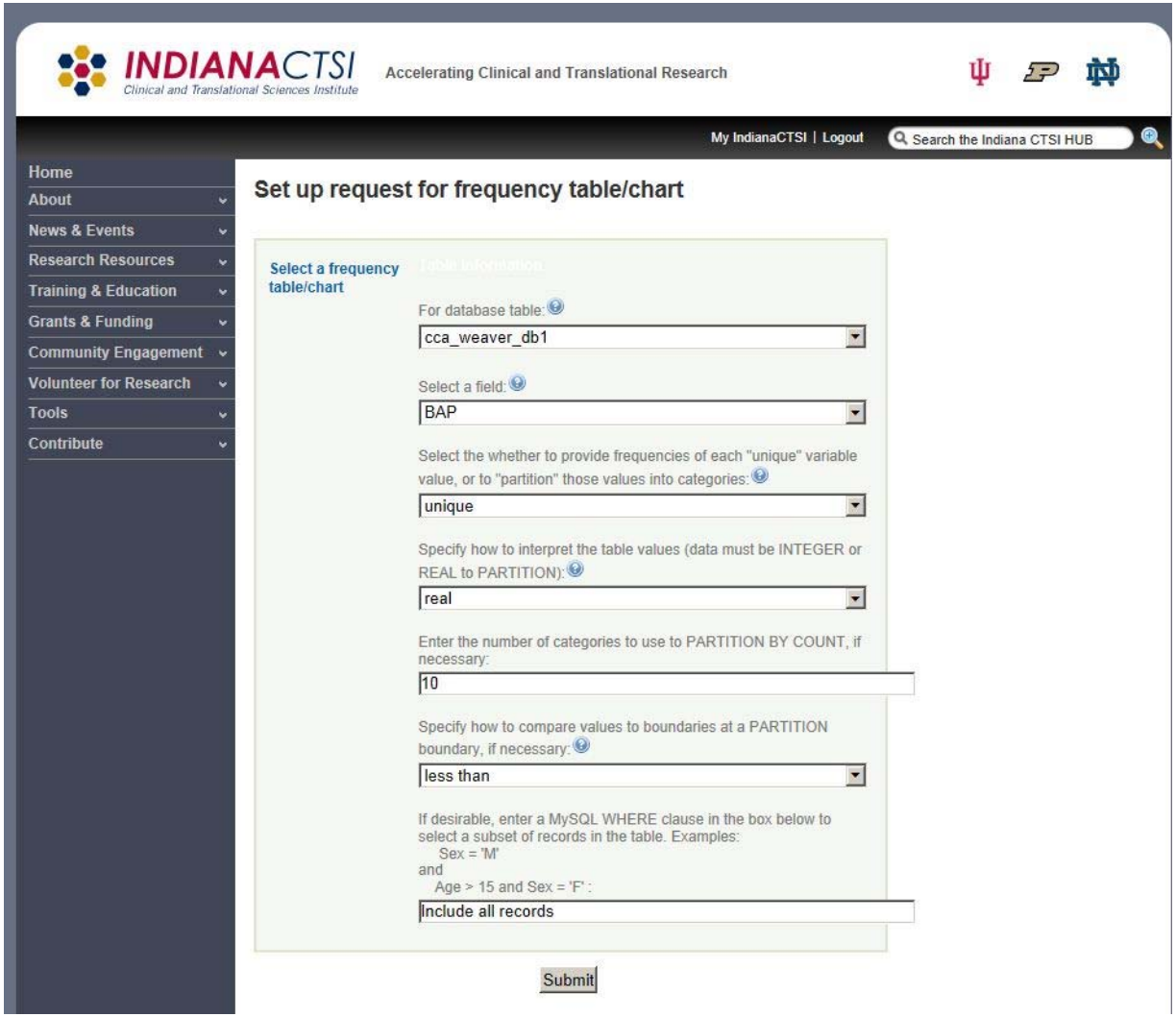
1) **illustrate** how these components can be used together, and
2) help **gauge user-interest** in having such capabilities proceed past the prototype phase.

(It will not however, provide a tutorial on actual coding techniques.)

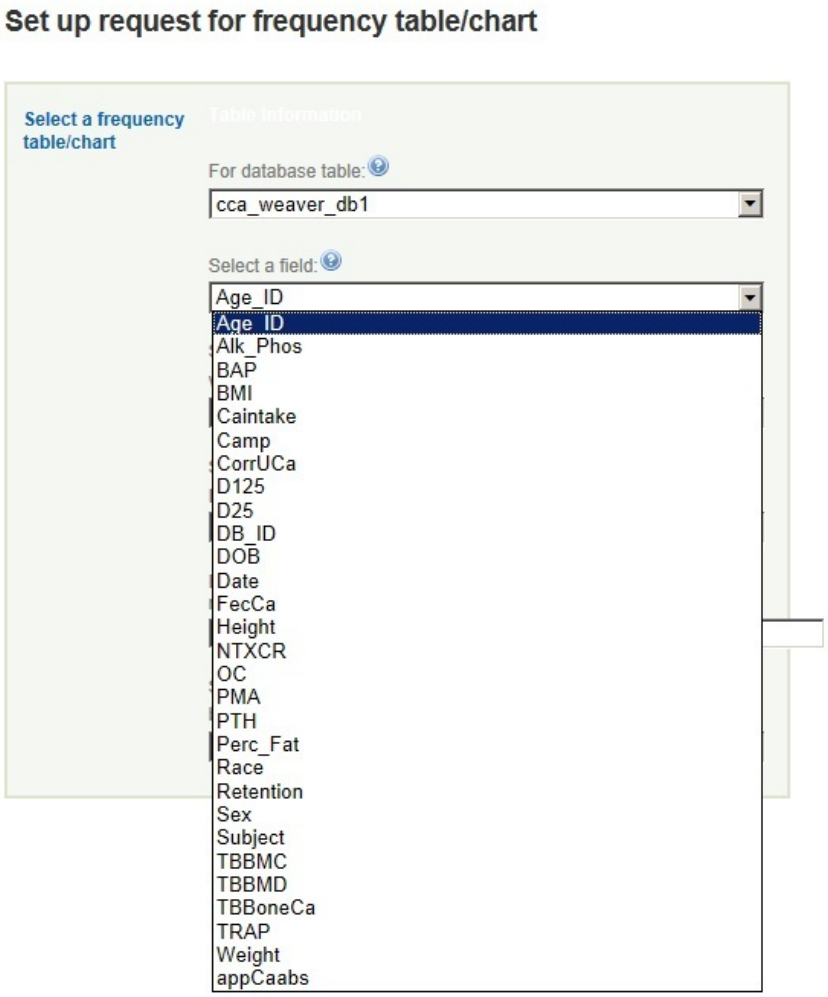
A starting form to create queries on the fly.

(http://dev1.indianactsi.org/form?proj_id=menu&form_id=form0)

Request a frequency table/chart



The field name pulldown menu

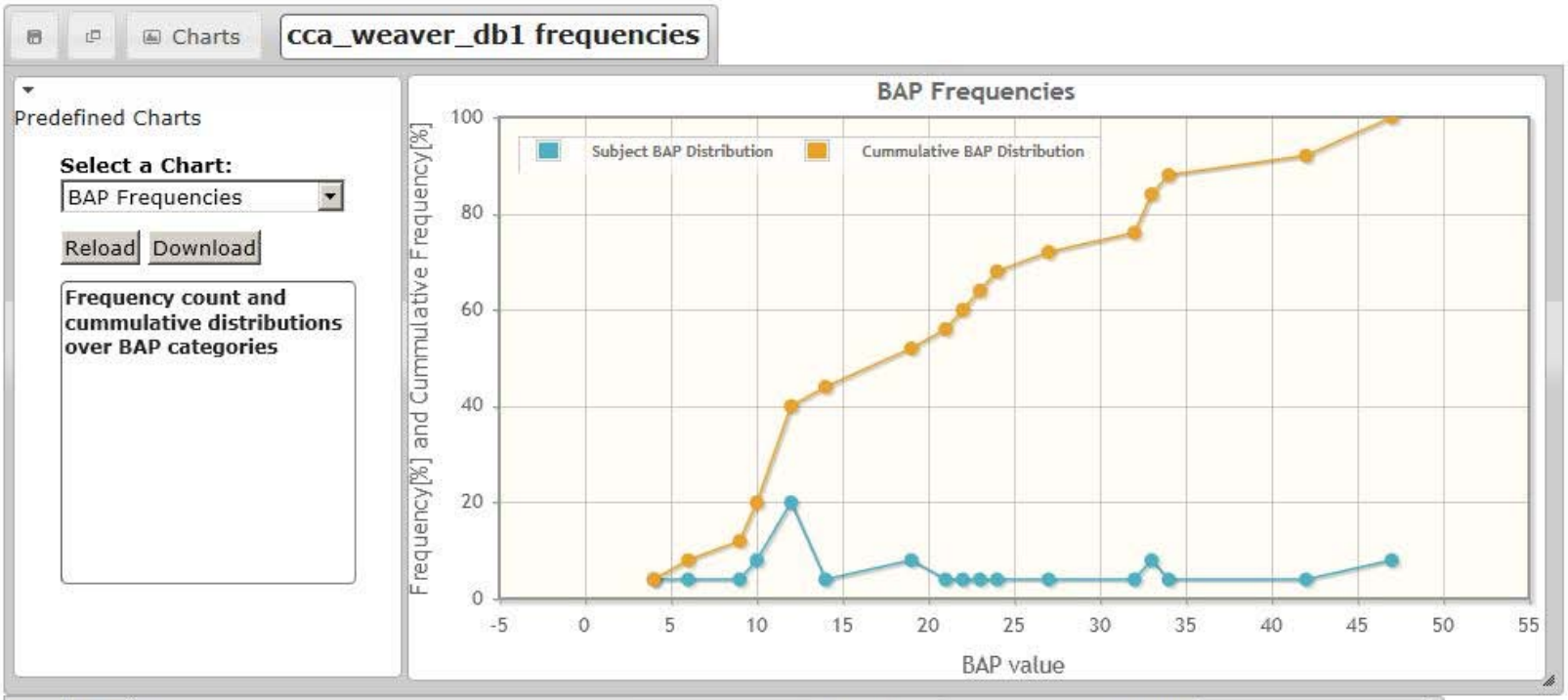


The resulting frequency chart

The data table is: cca_weaver_db1

All records in the table will be included.

For field "BAP": N: 25 Mean: 21.9861 Median: 47.8695 Standard deviation: 12.4821



Show 10 entries First Previous 1 2 Next Last Search:

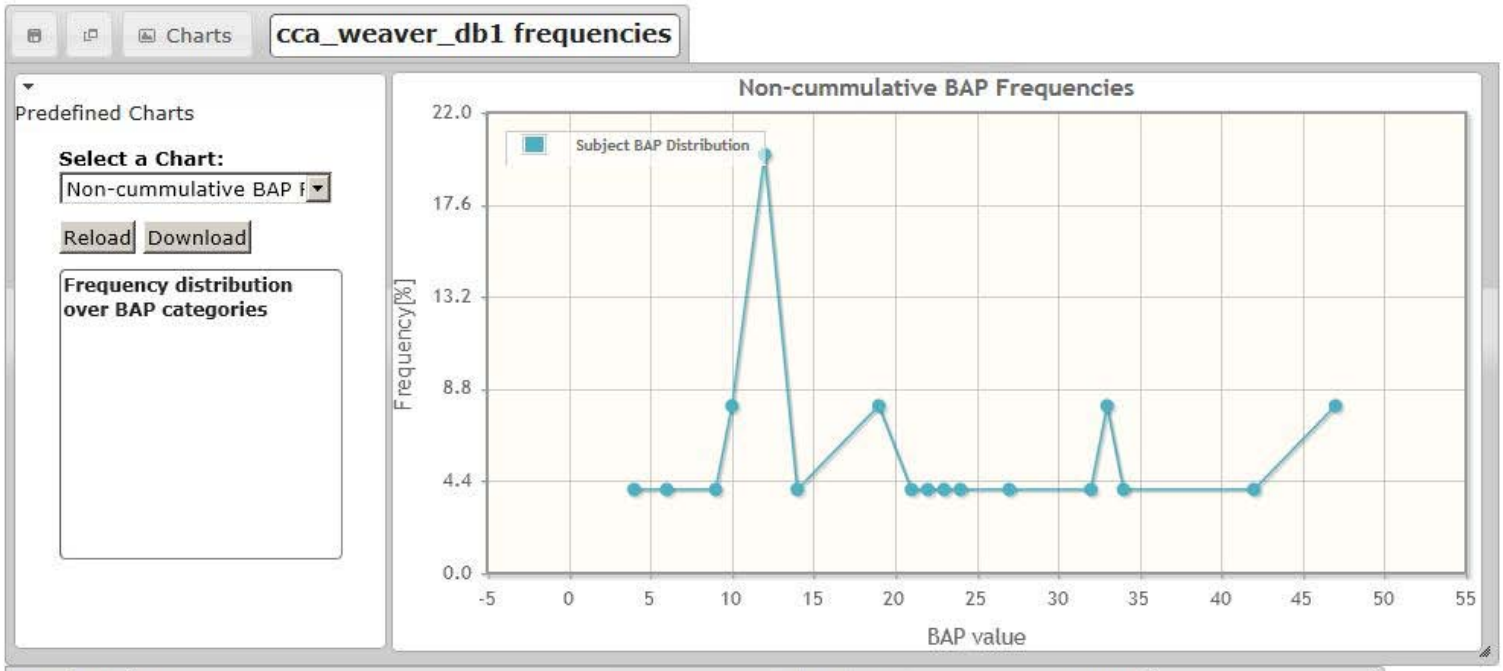
Id	Name	Value	Freq	Freq Per Cent	Cumm	Cumm Per Cent
1	BAP:real-unique::	4	1	4.00	1	4.00
2	BAP:real-unique::	6	1	4.00	2	8.00
3	BAP:real-unique::	9	1	4.00	3	12.00

The cumulative frequency chart

The data table is: cca_weaver_db1

All records in the table will be included.

For field "BAP": N: 25 Mean: 21.9861 Median: 47.8695 Standard deviation: 12.4821



Show 10 entries First Previous 1 2 Next Last Search:

Id	Name	Value	Freq	Freq Per Cent	Cumm	Cumm Per Cent
1	BAP:real-unique::	4	1	4.00	1	4.00
2	BAP:real-unique::	6	1	4.00	2	8.00
3	BAP:real-unique::	9	1	4.00	3	12.00

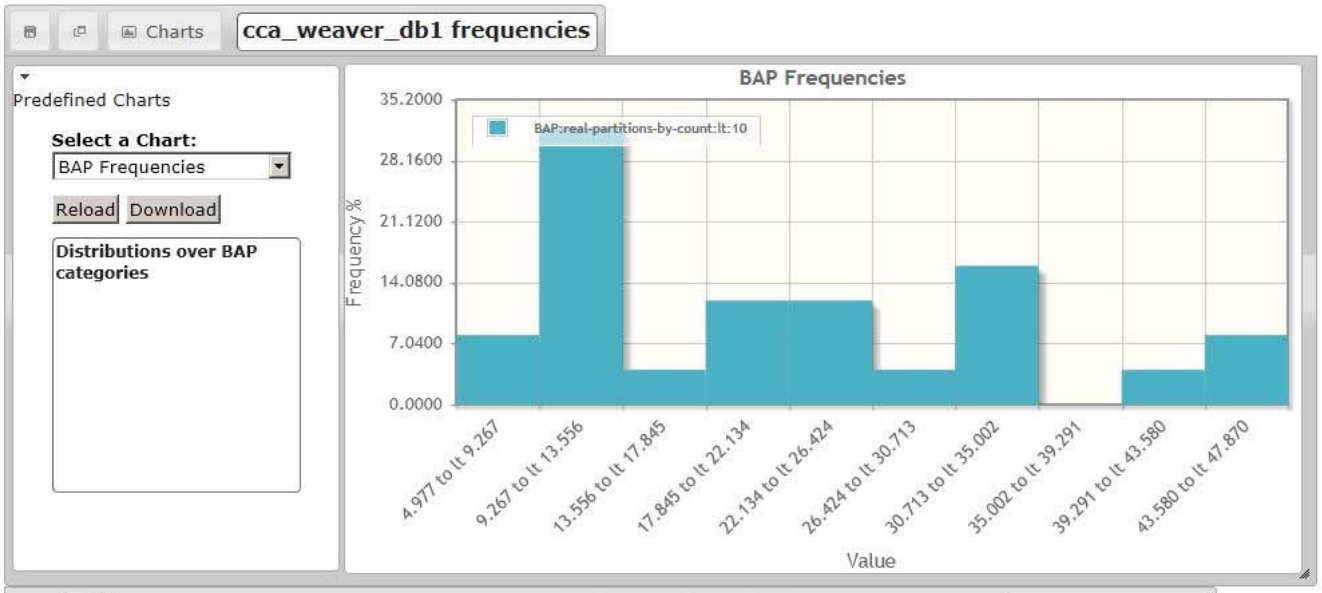
Request a chart of partitioned field values

Resulting chart showing partition frequencies

The data table is: cca_weaver_db1

All records in the table will be included.

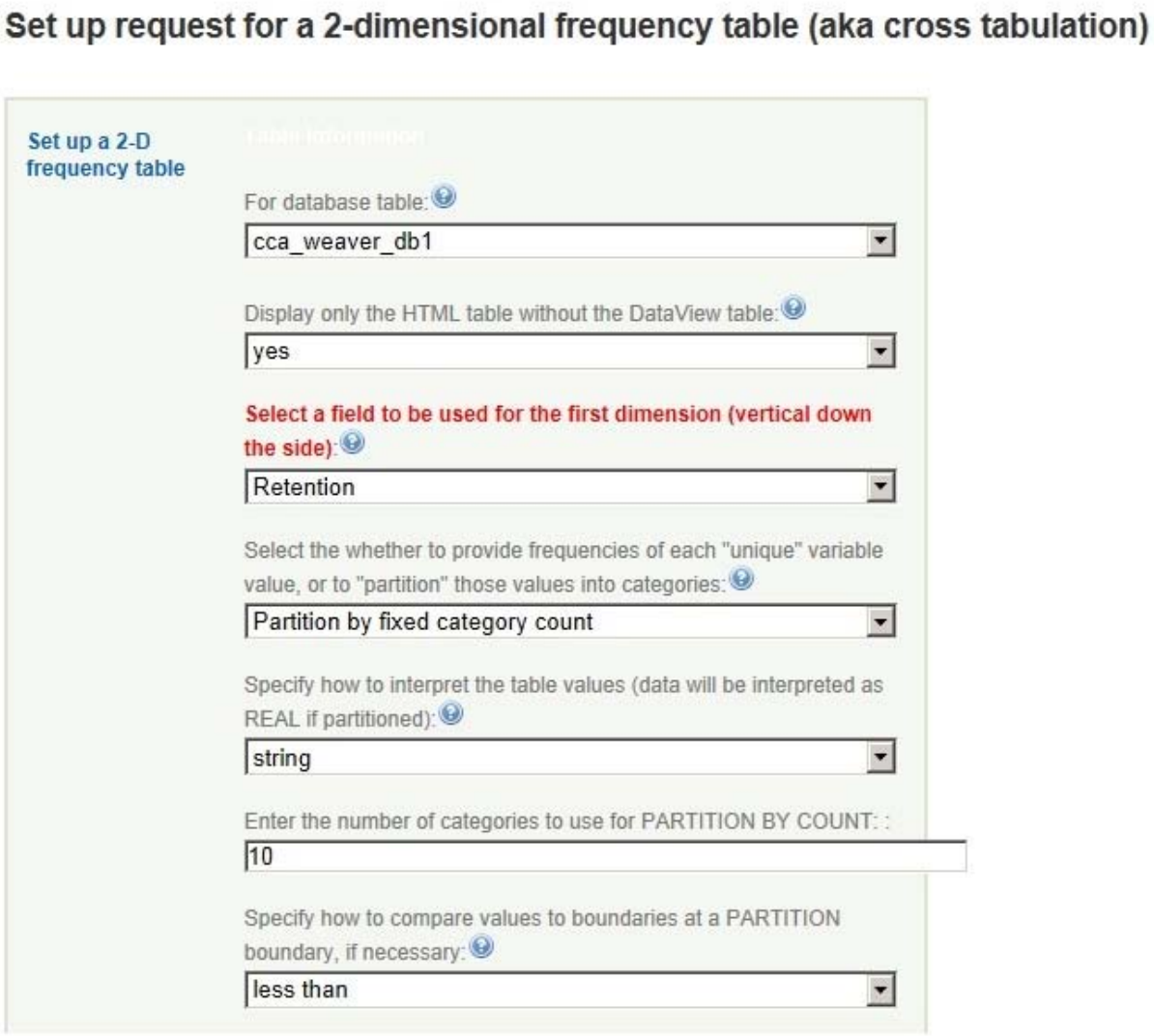
For field "BAP": N: 25 Mean: 21.9861 Median: 47.8695 Standard deviation: 12.4821



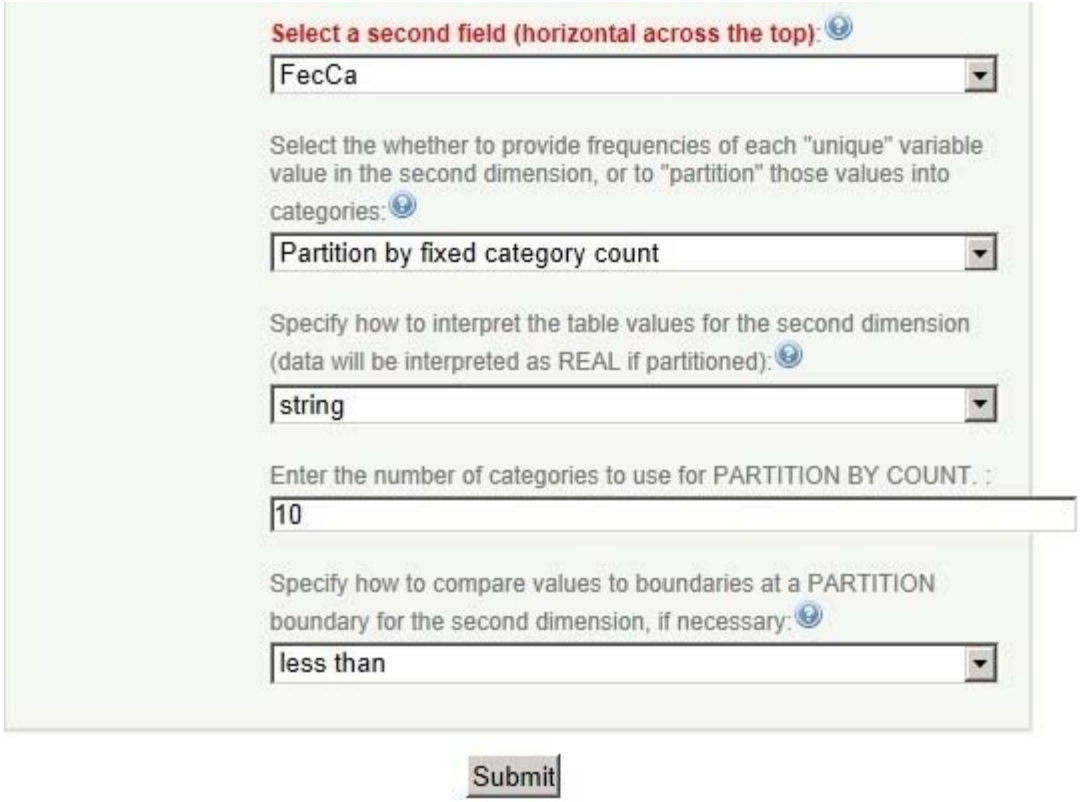
Show 10 entries First Previous 1 Next Last Search:

Id	Name	Value	Freq	Freq Per Cent	Cumm	Cumm Per Cent
1	BAP:real-partitions-by-count:lt:10	4.977 to lt 9.267	2	8.00	2	8.00
2	BAP:real-partitions-by-count:lt:10	9.267 to lt 13.556	8	32.00	10	40.00
3	BAP:real-partitions-by-count:lt:10	13.556 to lt 17.845	1	4.00	11	44.00
4	BAP:real-partitions-by-count:lt:10	17.845 to lt 22.134	3	12.00	14	56.00
5	BAP:real-partitions-by-count:lt:10	22.134 to lt 26.424	3	12.00	17	68.00

Cross tabulation on 10 partitions of 2 fields



Cross tabulation (second half)



Resulting tabulation
(10 partitions of each variable)

The data table is: cca_weaver_db1

All records in the table will be included.

For field "Retention": N: 25 Mean: 215.8800 Median: 220.0000 Standard deviation: 161.6736 Minimum value: -106.0000 Maximum value: 512.0000

For field "FecCa": N: 25 Mean: 971.4800 Median: 950.0000 Standard deviation: 150.0217 Minimum value: 614.0000 Maximum value: 1292.0000

	FecCa values or categories									
Retention values or categories	614.000 to < 681.800	681.800 to < 749.600	749.600 to < 817.400	817.400 to < 885.200	885.200 to < 953.000	953.000 to < 1020.800	1020.800 to < 1088.600	1088.600 to < 1156.400	1156.400 to < 1224.200	1224.200 to <= 1292.000
-106.000 to < -44.200										2
-44.200 to < 17.600						1				
17.600 to < 79.400								2		
79.400 to < 141.200										
141.200 to < 203.000				1	3		2		1	
203.000 to < 264.800					1		2			
264.800 to < 326.600				3	2					
326.600 to < 388.400							1			
388.400 to < 450.200			1	1						
450.200 to <= 512.000	1			1						

User specified partitions

Users may explicitly specify partitions.

In the next example, FecCa partitions were specified as:

700 | 800 | 900 | 1000 | 1100 | 1200

and the Retention partitions were specified as:

-200 | -100 | 0 | 100 | 200 | 300 | 400 | 500

Real partition values may be specified using either standard decimal notation or scientific notation; they must of course be in ascending order.

These are strict "partitions" rather than category "boundaries", so the first and last categories extend to negative or positive infinity, respectively; they are "unbounded", though that may not always be explicitly represented in the headers.

Resulting tabulation
(user-defined partitions of each variable)

The data table is: cca_weaver_db1_1_1996

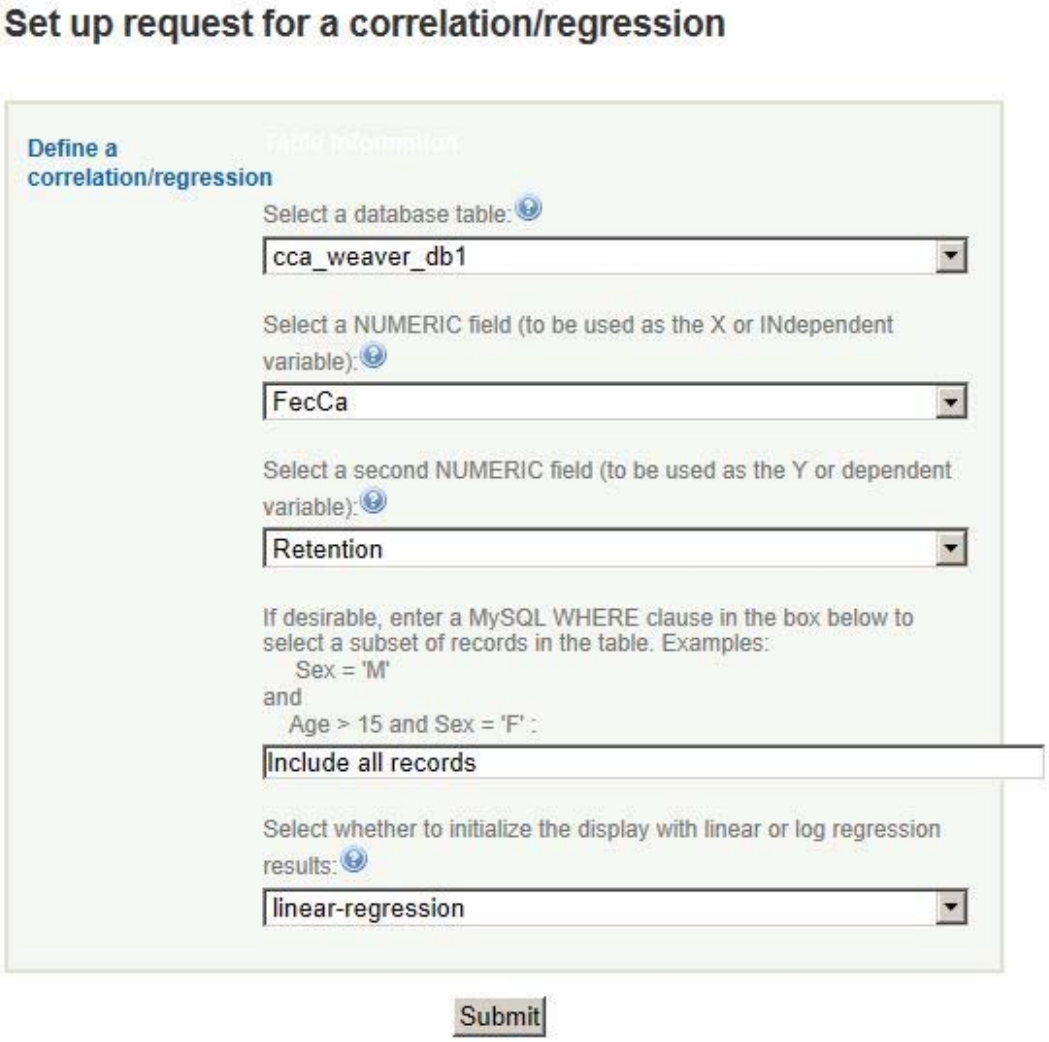
All records in the table will be included.

For field "Retention": N: 25 Mean: 215.8800 Median: 220.0000 Standard deviation: 161.6736 Minimum value: -106.0000 Maximum value: 512.0000

For field "FecCa": N: 25 Mean: 971.4800 Median: 950.0000 Standard deviation: 150.0217 Minimum value: 614.0000 Maximum value: 1292.0000

	FecCa values or categories						
Retention values or categories	< 700	700 to < 800	800 to < 900	900 to < 1000	1000 to < 1100	1100 to < 1200	>= 1200
< -100							1
-100 to < 0				1			1
0 to < 100						2	
100 to < 200			1	3	2	1	
200 to < 300				1	2		
300 to < 400			4	1	1		
400 to < 500			3				
>= 500	1						

Request a simple correlation/regression



Resulting regression info

Summary statistics

Data field name	N	Mean	Median	Standard Dev.	Min value	Max value
FecCa	25	971.4800	944.0000	150.0217	614.0000	1292.0000
Retention	25	215.8800	210.0000	161.6736	-106.0000	512.0000

Correlation	Slope	Intercept	Pearson correlation value
Raw X against Raw Y	-0.880	1071.248	-0.817
Raw X against log(Raw Y)	-0.00557	14.27704	-0.646

Charts | **Regression results**

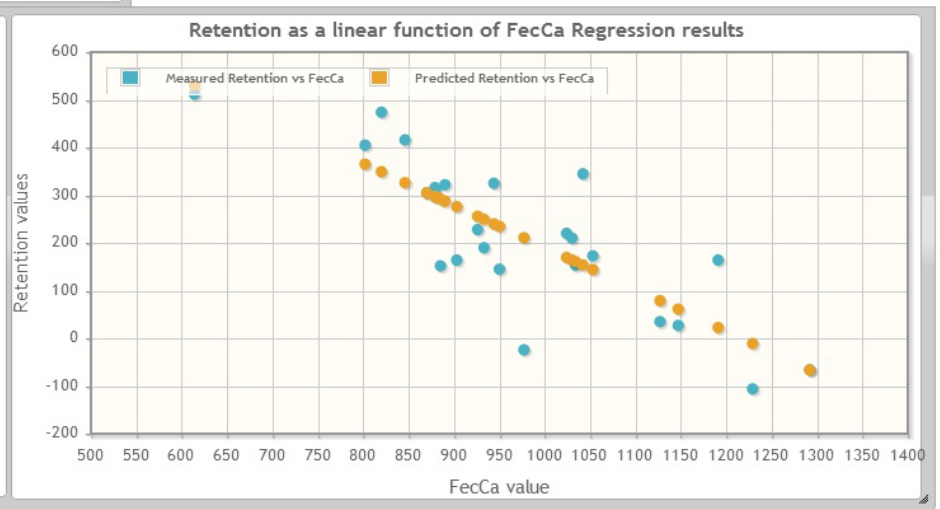
Predefined Charts

Select a Chart:

Retention as a linear fu

Reload Download

Predicted and measured values of FecCa => Retention



Show 10 entries First Previous 1 2 3 Next Last Search:

Id	FecCa	Retention	Y Predicted Linear	Linear Residual	Y Predicted Log	Log Residual
1	614.0000	512.0000	530.6339	18.6339	51921.4062	51409.4062
2	1024.0000	220.0000	169.6372	-50.3628	5203.8516	4983.8516
3	846.0000	416.0000	326.3626	-89.6374	14196.3037	13780.3037
4	926.0000	228.0000	255.9242	27.9242	9056.3779	8828.3779
5	890.0000	322.0000	287.6215	-34.3785	11089.3623	10767.3623

A correlation_matrix

The data table is: cca_weaver_db3_2_1997

All records in the table will be included.

Correlation matrix:

Notes: Correlation values less than 0.5 or greater than 1.0 are omitted; click on links for detailed information

	DB_ID	Camp	Date	Caintake	CorrUCa	FecCa	Retention	Age	Height	Weight	BMI	Tanner	PMA_set_to_zero
DB_ID	1.00	-0.76	-0.76										
Camp	-0.76	1.00	1.00										
Date	-0.76	1.00	1.00										
Caintake				1.00		0.93	0.50						
CorrUCa					1.00								
FecCa				0.93		1.00							
Retention				0.50			1.00						
Age								1.00				0.54	0.64
Height									1.00	0.53			
Weight									0.53	1.00	0.93		
BMI										0.93	1.00		
Tanner								0.54				1.00	0.52
PMA_set_to_zero								0.64				0.52	1.00

Request user-defined displays via URLs

You can invoke these applications via statically coded URLs. These examples will only work (at present) if you are logged in to the target CTSI hub instance in the same browser.

- Cross tablulate Retention with "Age category" (Teen or Adult) (1996) (show only the HTML table to avoid formatting imposed by the DataViewer display)
- Cross tabulate Retention with "Age category" (Teen or Adult) (1996) (show BOTH the HTML and DataView tables)

- Cross tabulate 10 Alk Phos categories with 10 Retention categories (HTML table only)
- Cross tabulate unique Retention values with unique Retention values (both cases interpreted as integers) (HTML table only)
- Retention cross-tabulated with FecCa (Fecal Calcium)
- Age by **Pre-specified** Age Month categories (There are some missing Age values.)

- Correlate and regress Retention as a function of Post Menarcheal age (1996)

- Correlation matrix composed of all numeric fields in the table cca_weaver_db1_1_1996 showing only values greater than .5

Summary

Prototypes for 4 tools for (very) simple data exploration were built over 2 existing Purdue DB technology tools. (Correlation matrix tool can now run independently of the DataViewer.)

These provide on-demand frequency charts, cross-tabulations, and simple correlation/regression output.

These tools can help users get a “feel” for the data, but are not suitable for serious statistical analysis.

Can these tools be helpful in the Hub context?

Additional information

- Tutorial: Setting up a spreadsheet for use with the cceHub DataViewer
- Tutorial: Using the Web form component to build scripts to modify databases
- Tutorial: Using the Web form and DataViewer components for to request user-specified frequency tables/charts, cross-tabs, and regressions
- Putting a Joomla! component into SVN for the development cycle