Turning Data into Information – Tools, Tips, and Training

A Summer Series Sponsored by Erin Gore, Institutional Data Council Chair; Berkeley Policy Analysts Roundtable, Business Process Analysis Working Group (BPAWG) and Cal Assessment Network (CAN)

Overview of this Session: Infographics – Tools to Present a Lot of Data in a Condensed Space

Presented by Pamela Brown and Russ Acker, Office of Planning & Analysis

As we learned from *Tufte in Twenty Minutes* in our first session of this IDMG summer series, "Rows of numbers do not have any visual impact….A chart is more memorable than a table of numbers."[1] Or perhaps you prefer, "The graphical method has considerable superiority for the exposition of statistical facts over the tabular. A heavy bank of figures is grievously wearisome to the eye, and the popular mind is as incapable of drawing any useful lessons from it as of extracting sunbeams from cucumbers."[2] In any case, one of the most common data visualization techniques that you'll use is displaying a table of values as a picture of some sort. What that picture looks like, though, depends to a large extent on the type of table that you have.

For instance, a very common type of data that analysts here at the university work with is a time series, which shows how a value changes over time. Another table, however, might show hierarchical data, such as the current number of students in a college, department, and major. Or you might have a combination of the two, not to mention the fact that tables can range in size from tiny to enormous. The major point to keep in mind here is that different charts have different strengths and weaknesses for displaying different kinds of tabular data.[3]

In this session, we'll discuss tools for creating several chart types that can help you make sense of large tables, including:

- Sparklines, for integrating charts directly into tables;
- Treemaps, for displaying very large tables of hierarchical data;
- Heatmaps, for highlighting certain values within a table;
- Scatter plots, for highlighting clusters of data points that show the relationships between two numeric values;
- Circos diagrams, for highlighting the flows of data that define the relationships between rows and columns.

---

[1] Dona Wong, *The Wall Street Journal Guide to Information Graphics* (New York: W.W. Norton & Company, 2010), pp. 82-83.

[2] A.B. Farquhar and H. Farquhar, *Economic and Industrial Solutions* (New York: G.B. Putnam's Sons, 1891), p. 55, as quoted in Stephen Few, *Now You See It* (Oakland: Analytics Press, 2009), p. 3.

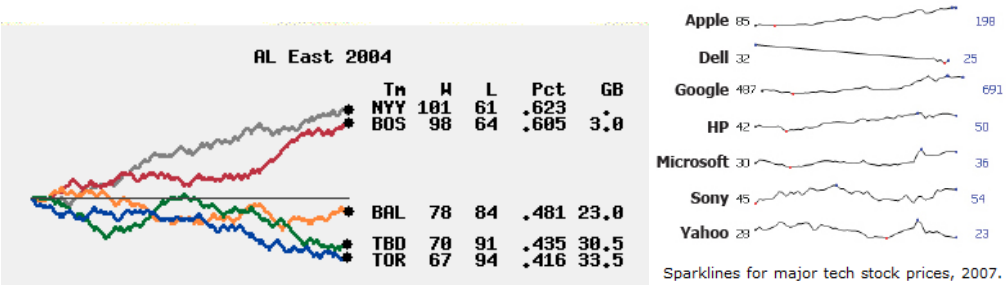
[3] Stephen Few provides a handy graph selection reference at http://www.perceptualedge.com/articles/misc/Graph_Selection_Matrix.pdf.

Turning Data into Information – Tools, Tips, and Training

Sparklines

Overview of the Tool

Sparklines, as described by Edward Tufte, are "small, high-resolution graphics usually embedded in a full context of words, numbers, [and] images. Sparklines are *datawords*: data-intense, design-simple, word-sized graphics."[4] The general idea here is that sparklines allow you to integrate charts into text or tables, at the point where the data's being discussed. And you may have already seen sparklines on the business, sports, or weather pages in your local newspaper.



Sparklines for major tech stock prices, 2007.

Sparklines are particularly good at adding trend information to tables of numbers. Since sparklines are integrated directly into the table, rather than as a separate graph, readers don't have to keep switching back and forth between the table and visualization; everything's in one place.

How to Get the Tool

Sparkline generators are all over the place, but here are a few of the better ones:

- If you have Excel 2010 (Windows) or Excel 2011 (Mac, but not yet released as of this writing), three different kinds of sparklines are now built-in. Yay!
- If you use Excel 2003/2007 (Windows), Excel 2004 (Mac), or just want a larger set of sparklines to choose from, try the Excel add-ins at http://sparklines-excel.blogspot.com/.
- If you use Open Office Calc, there's a sparkline extension at: http://extensions.services.openoffice.org/project/eurooffice-sparkline.
- If nothing else works (because, for instance, you have Excel 2008 for Mac), you can generate sparklines online at http://sparklines.bitworking.info/ and then copy them into your document. You can also sort of make your own sparklines in Excel by creating a really small chart and removing all the window dressing from it, but they usually don't look quite as good as sparklines generated by the tools noted above.

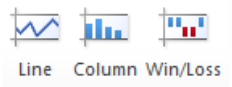
[4] Edward Tufte, *Beautiful Evidence* (Cheshire, CT: Graphics Press LLC, 2006), p. 47.

How to Use the Tool / Examples

Below is a table with offerings and enrollments for a subset of foreign language courses. There are ways to improve this tabular data (e.g., sorting it by total enrollment, presenting the percent change); for some, however, it will still be a table of numbers. So if your audience glazes over and the precise figures aren't necessary to advance the discussion, sparklines provide an alternative for presenting trends and relationships.

Select Foreign Language Course Offerings & Enrollments

	2000-2001	2001-2002	2002-2003	2003-2004	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009	TOTAL
NUMBER OF SECTIONS										
CLASSICS	19	21	19	19	18	19	18	17	19	169
EAST ASIAN LANGUAGES & CULTURES	103	118	111	115	132	135	144	143	145	1146
ENGLISH	1	1		1	1			1	1	6
FRENCH	69	74	71	74	73	71	72	75	70	649
SLAVIC LANGUAGES & LITERATURES	32	35	35	35	41	43	43	40	41	345
SPANISH AND PORTUGUESE	88	82	88	84	80	74	76	72	65	709
NUMBER OF ENROLLMENTS										
CLASSICS	200	272	263	252	252	234	255	259	248	2235
EAST ASIAN LANGUAGES & CULTURES	2028	2197	2323	2173	2490	2731	3030	2892	2739	22603
ENGLISH	7	7		8	13			12	6	53
FRENCH	1145	1149	1136	1174	1187	1120	1197	1229	1204	10541
SLAVIC LANGUAGES & LITERATURES	252	293	308	366	346	348	359	337	385	2994
SPANISH AND PORTUGUESE	1568	1498	1692	1716	1604	1549	1535	1472	1343	13977



Excel 2010 has the option to insert three kinds of sparklines: line, bar, and win-loss charts. You begin by specifying the destination cell, selecting insert sparkline, and then inputting the data range (e.g., table data). Below is an example of the results for sparklines (using the bar chart). As you can see, the sparklines allow you to present a lot of information in a small space, and in this example, an easier way to compare offerings and enrollments. For example, while East Asian Languages and Cultures offerings have gone up, enrollments have started to drop. One of the criticisms of sparklines is that the lack of scaling will overemphasize trends and that the sizes are not comparable among sparklines. So it can be useful to insert end points as numerical references (e.g., averages, percent changes) to provide additional information on scaling and trends.

Select Foreign Language Course Offerings & Enrollments

	Numbers of Sections Offered			Number of Enrollments		
	2000-2001	9 Year Trend	2008-2009	2000-2001	9 Year Trend	2008-2009
EAST ASIAN LANGUAGES & CULTURES	103		145	2028		2739
SPANISH AND PORTUGUESE	88		65	1568		1343
FRENCH	69		70	1145		1204
SLAVIC LANGUAGES & LITERATURES	32		41	252		385
CLASSICS	19		19	200		248
ENGLISH	1		1	7		6

TIP: *When your data has blank values (e.g., English), it is much better to use a bar chart instead of a line chart. Also, while Excel 2010 seems to cleanly convert sparklines to a pdf, there have been some troubles with the line charts (not bar charts) generated by the Sparklines-Excel add-in, where the sparklines seem to slip on the page after conversion to pdf.*

Excel 2010 has several other sweet features for formatting sparklines. You can tag high and low points, first and last points, and negative points. You can specify the color of your sparklines to help you highlight certain units. And to address sparkline critics, you can also provide common scaling for the sparklines as seen in the chart below.

	Numbers of Sections Offered			Number of Enrollments		
	2000-2001	9 Year Trend	2008-2009	2000-2001	9 Year Trend	2008-2009
EAST ASIAN LANGUAGES & CULTURES	103		145	2028		2739
SPANISH AND PORTUGUESE	88		65	1568		1343
FRENCH	69		70	1145		1204
SLAVIC LANGUAGES & LITERATURES	32		41	252		385
CLASSICS	19		19	200		248
ENGLISH	1		1	7		6

The final results can be sent out in your Excel document, saved as a pdf file, or cut and pasted into a Word document (as seen in this handout).

References/Other Resources

Edward Tufte's book, *Beautiful Evidence* (Cheshire, CT: Graphics Press LLC, 2006), pp. 46-63, contains the most extensive of several sparkline discussions in his publications. His blog also has a very lengthy thread on the topic at http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR.

Treemaps

Overview of the Tool

Treemaps provide a way to view very large amounts of hierarchical data. In most cases, they work better "live" rather than printed, since hierarchical situations generally encourage users to drill down into data. With relatively small datasets and relatively shallow hierarchies, however, printed versions look fine.

A treemap shows a hierarchy through the use of nested boxes. The size of each box represents one numeric value, while the color of each box represents another.

How to Get the Tool

The tool we'll demo today is the Microsoft Research Treemapper, which can be used either stand-alone or as an add-in to Excel. Although this free tool is Windows-only, it has the advantage of being an installable product that doesn't require creating websites or publicly sharing data. All of the treemappers noted in the References/Other Resources section below, however, create treemaps that look more or less the same.

You can download the Microsoft Research Treemapper at: http://tinyurl.com/mstreemap

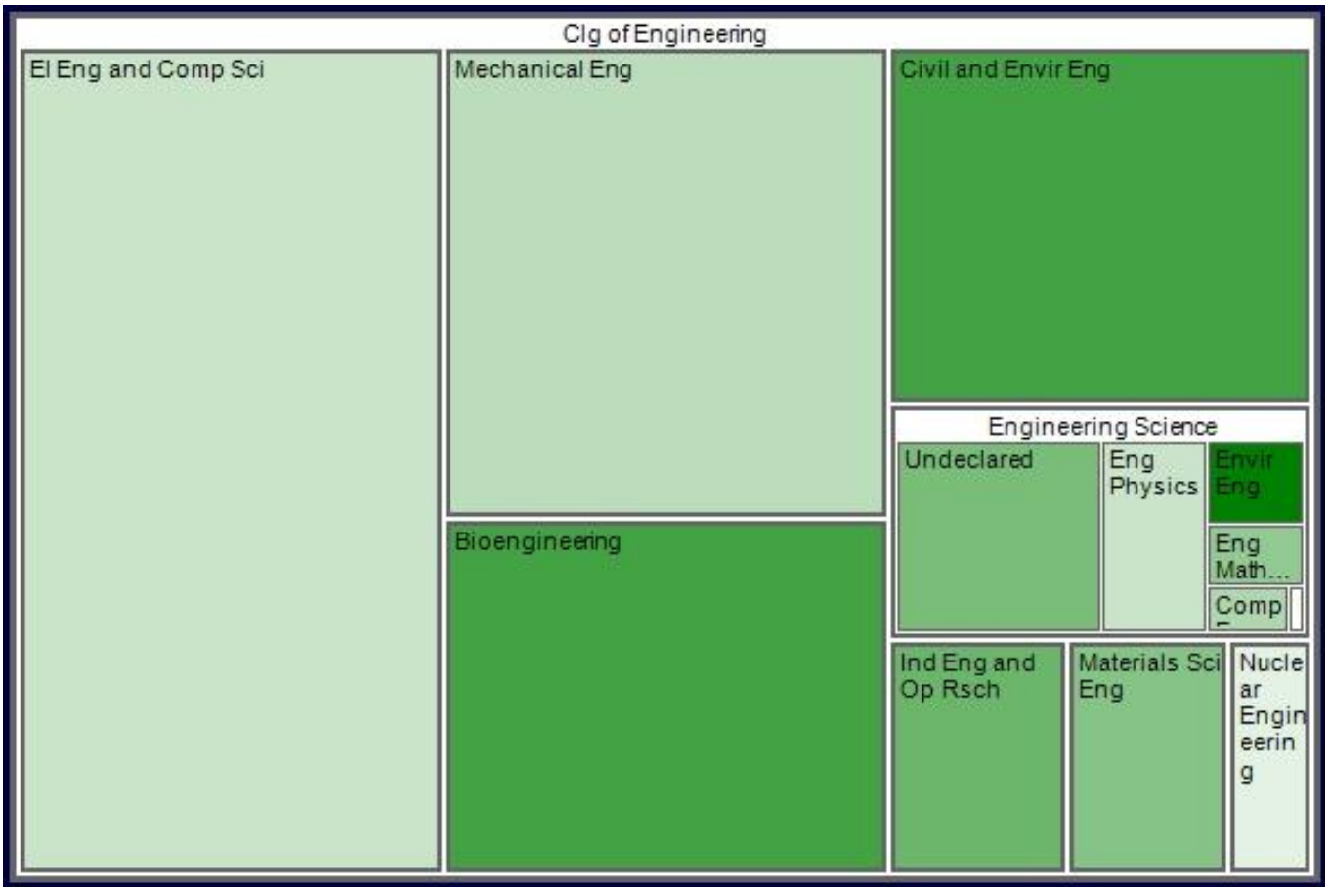
How to Use the Tool / Examples

Let's say you want to create a treemap that shows the number of 2009-2010 undergraduate students in the College of Engineering, by department and with an indication of what percentage are women.

The file layout that the Microsoft Treemapper uses expects to see the size metric, then the color metric, then the highest hierarchy level, then the second highest, etc. (there doesn't seem to be a limit, other than your computer's memory and your eyesight, to the number of hierarchy levels that you can include). So, your input spreadsheet, saved as a .csv file, might look something like this (we've included column headers for clarity, but they shouldn't be in the file):

Size Metric (Total Undergrads)	Color Metric (% Women)	Hierarchy Level 1	Hierarchy Level 2	Hierarchy Level 3
424	0.35	Clg of Engineering	Bioengineering	
404	0.35	Clg of Engineering	Civil and Envir Eng	
944	0.11	Clg of Engineering	El Eng and Comp Sci	
70	0.11	Clg of Engineering	Engineering Science	Eng Physics
137	0.25	Clg of Engineering	Engineering Science	Undeclared
21	0.21	Clg of Engineering	Engineering Science	Eng Math/Stat
28	0.47	Clg of Engineering	Engineering Science	Envir Eng
13	0.15	Clg of Engineering	Engineering Science	Comp Eng
112	0.27	Clg of Engineering	Ind Eng and Op Rsch	
100	0.22	Clg of Engineering	Materials Sci Eng	
567	0.13	Clg of Engineering	Mechanical Eng	
50	0.07	Clg of Engineering	Nuclear Engineering	

Once you've created the input file, you just do a File/Open… in the Microsoft Treemapper. It will then display a treemap like this:



Using the View/Options… menu, you can then change quite a few things about the treemap's appearance, including the colors, the labels, and the tooltips that appear when you move your mouse over the map.

Using the File/Save As… menu, you can save the treemap in a variety of formats suitable for inserting in Word or PowerPoint documents.

References/Other Resources

Stephen Few's book *Now You See It: Simple Visualization Techniques for Quantitative Analysis* (Oakland: Analytics Press, 2009), pp. 87-91, has a nice discussion covering the appropriate uses of treemaps.

And here are some other treemapping tools you may want to try:

- **IBM ManyEyes Treemap**
  http://manyeyes.alphaworks.ibm.com/manyeyes/page/Treemap.html
  IBM's ManyEyes site provides a very nice data visualization toolkit, including two different kinds of treemappers. These tools work entirely inside a web browser, so you can use them on either a PC or Mac. However, you do have to sign up for a free account, and any data that you upload to the site is completely public.
- **Data Applied Treemap**
  http://www.data-applied.com/
  The Data Applied website is much like IBM's ManyEyes in that everything happens inside a web browser, but here you have the option of keeping data private (although it's still stored on their server). The free version of this site requires a sign in and limits you to a single data file at a time.
- A quick search will return dozens of other treemapping tools, from some written in JavaScript to the version included in the Portfolio library of R, the open source statistical tool. So have fun!

Turning Data into Information – Tools, Tips, and Training

Heatmaps

Overview of the Tool

Heatmaps (also known as colormaps or highlight tables) allow you to apply a scale of colors or shapes to values within a table, so that a viewer can quickly determine relative values. Heatmaps are very easily done in Excel, particularly in versions 2007 and 2010 in Windows and 2011 on the Mac.

How to Get the Tool

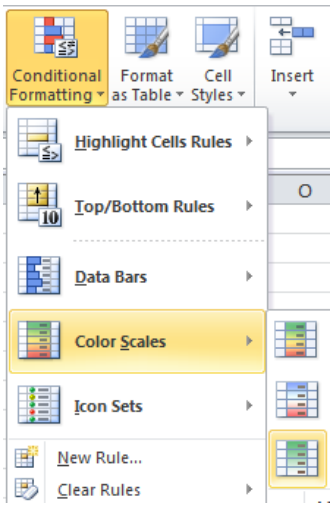
The Conditional Formatting feature of Excel allows you to create a heatmap over a table of data. In newer versions of Excel, there are built-in menu options called Data Bars, Color Scales, and Icon Sets within Conditional Formatting. In older versions of Excel, you can still do most of this through Conditional Formatting, but you have to set the rules up yourself.

How to Use the Tool / Examples

We'll start with this table of registered undergrads as an example:

Name	Fall 05	Fall 06	Fall 07	Fall 08	Fall 09
Business	564	573	594	583	609
Chemistry	797	820	804	768	795
Env. Design	665	657	668	651	609
Physical Sci.	481	429	506	541	570
Other L&S	815	752	696	832	1018

To turn this table into a heatmap in Excel 2007/2010, first highlight the spreadsheet cells that contain the data. Then go to Home/Conditional Formatting and choose the type of heatmap you want to use; for this example, we'll go with a Green-White Color Scale, as shown in the screenshot to the right.



Notice all the other options available on this menu, including different color scales, data bars, and icon sets. Keep in mind that several of the color scales and icon sets would be indistinguishable to people with color blindness, and that several icon sets imply trends that may not exist.

Applying this color scale turns the table into a heatmap, with lower values indicated by white and higher values indicated by darker green, as shown below. It now takes only a quick glance at the table to find the high, middle, and low values, as well as to see trends within each row.

Name	Fall 05	Fall 06	Fall 07	Fall 08	Fall 09
Business	564	573	594	583	609
Chemistry	797	820	804	768	795
Env. Design	665	657	668	651	609
Physical Sci.	481	429	506	541	570
Other L&S	815	752	696	832	1018

References/Other Resources

There are dozens of free Excel training classes available through e-Learn (for example, "Manipulating and Formatting Data and Worksheets in Excel 2007") and the UC Learning Center (for example, "Microsoft Excel 2007: Beyond the Basics"), all of which you can access through the Blu portal (http://blu.berkeley.edu) in the Self-Service area.

As a university employee, you also have free online access to many books on Excel through e-Learn's Books 24x7 (using the Blu portal at http://blu.berkeley.edu as noted above) or through the Safari Books Online service provided by the UC Berkeley Library (http://www.lib.berkeley.edu/find/types/electronic_resources.html).

Finally, you can view or print an excellent set of quick reference cards for various Microsoft Office products at http://customguide.com/quick_references.htm.

Turning Data into Information – Tools, Tips, and Training

Scatter Plots

Overview of the Tool

Finally, we'll be looking at a couple of charts that let you focus on the relationships between two variables. The first of these is just a good old scatter plot, which, according to Michael Friendly and Daniel Denis in the *Journal of the History of the Behavioral Sciences* (Vol. 41(2), p. 103, Spring 2005) "is arguably the most versatile, polymorphic, and generally useful invention in the history of statistical graphics." Wow.

How to Get the Tool

Since the scatter plot was invented at approximately the same time as air and dirt, almost any reasonably business-like software can create one (even Word, as I see while typing this). For illustration purposes, however, we'll just stick with Excel.

There is, however, a major deficiency in Excel's implementation of scatter plots: It can't display a categorical variable value (i.e., a label that's actually useful) next to each point. Fortunately, you can find anything on the internet, including the AppsPro XY Chart Labeler (http://www.appspro.com/Utilities/ChartLabeler.htm), a free Excel add-in. So, we'll be using that, too.

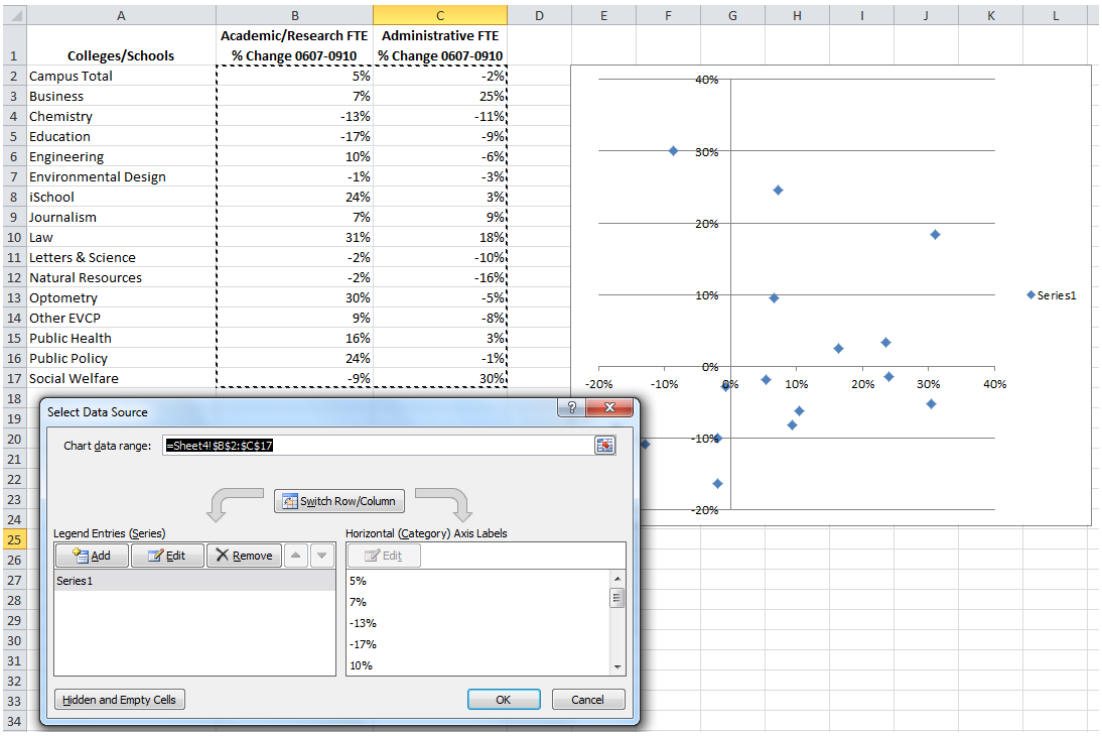
How to Use the Tool / Examples

For this example, we'll look at a table containing data on the percent change of academic/research and administrative FTE for each college and school, from 2006-07 to 2009-10:

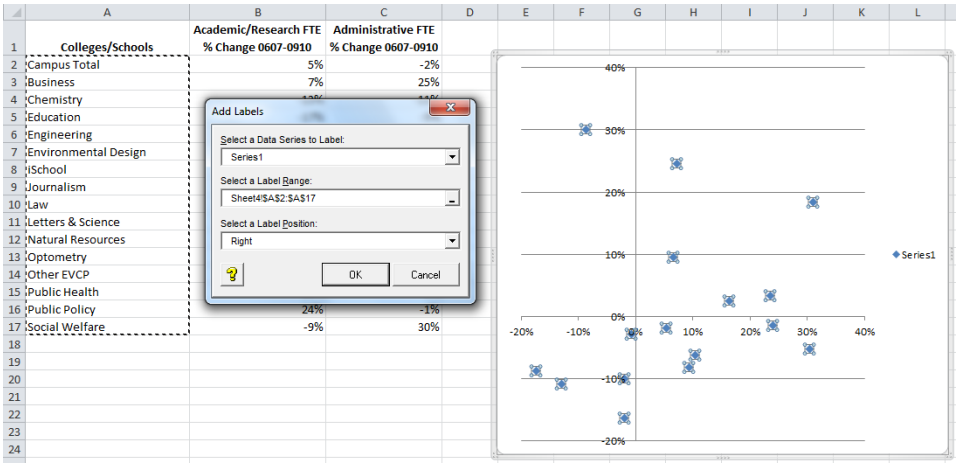
Colleges/Schools	Academic/Research FTE % Change 0607-0910	Administrative FTE % Change 0607-0910
Campus Total	5%	-2%
Business	7%	25%
Chemistry	-13%	-11%
Education	-17%	-9%
Engineering	10%	-6%
Environmental Design	-1%	-3%
iSchool	24%	3%
Journalism	7%	9%
Law	31%	18%
Letters & Science	-2%	-10%
Natural Resources	-2%	-16%

Optometry	30%	-5%
Other EVCP	9%	-8%
Public Health	16%	3%
Public Policy	24%	-1%
Social Welfare	-9%	30%

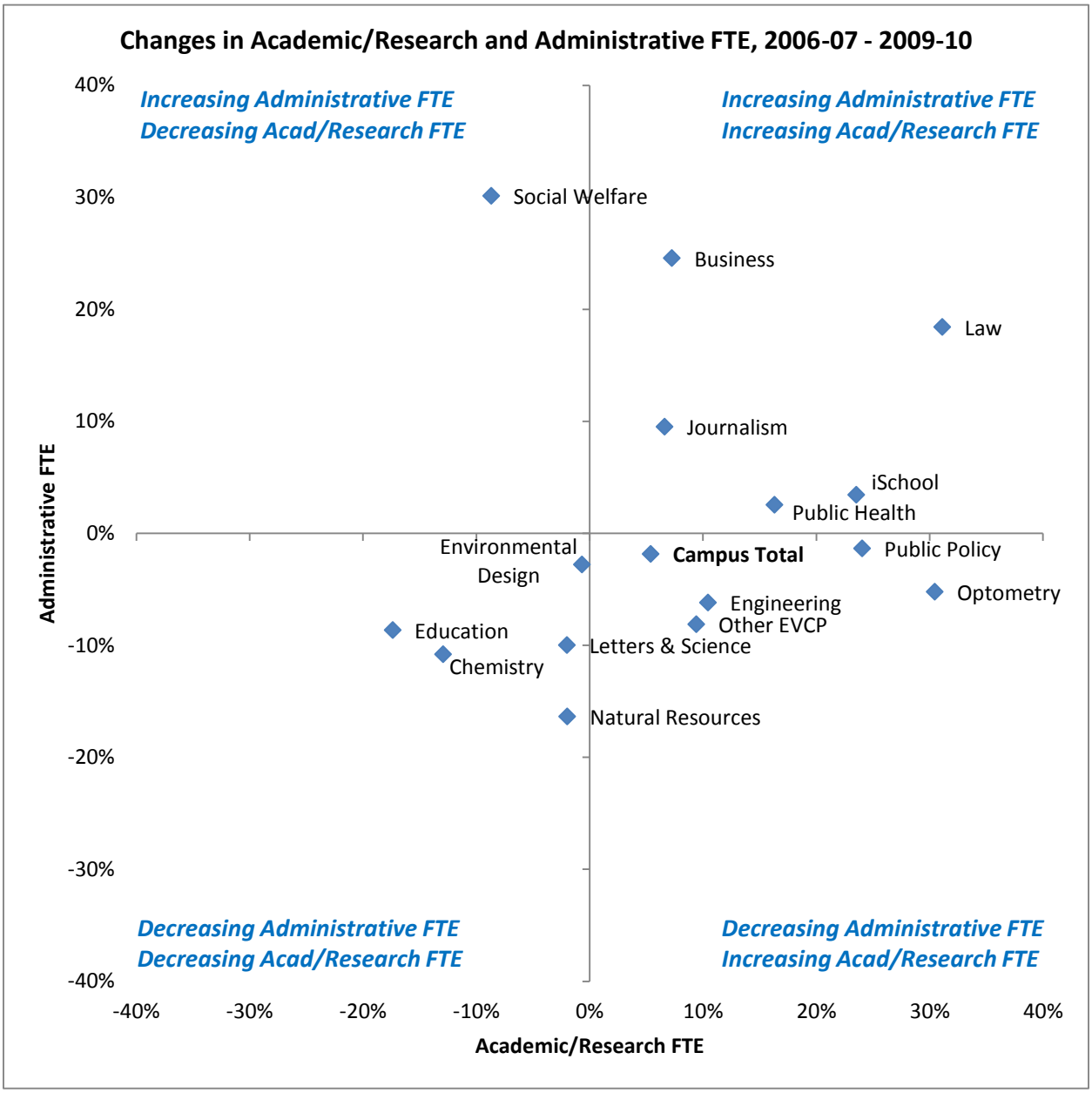
To create an Excel scatter plot from this table, just insert a scatter chart, then select only the numeric values in the table as the data source, as shown below:



To add the chart labels, use our newly installed XY Chart Labeler add-in to select the college/school names, as shown here:



After making numerous formatting changes to the chart, including resetting the axes, getting rid of gridlines, removing the legend, moving the point labels to make them more readable, adding some titles, and adding a textbox to identify each quadrant, we wind up with something like this:



References/Other Resources

As noted above, there are dozens of free Excel training classes available through e-Learn and the UC Learning Center, all of which you can access through the Blu portal. University employees also have free online access to books covering hundreds of tech topics, through either e-Learn or the UC Berkeley Library. And quick reference cards are available at http://customguide.com/quick_references.htm.

Turning Data into Information – Tools, Tips, and Training

Circos Diagrams

Overview of the Tool

The last kind of chart that we'll look at today is useful for highlighting the relationships between rows and columns in a table, particularly with data that has a progression from one state to another (although it works fine with any simple table). This tool is called the Circos Tableviewer,[5] which was created and is maintained by Canada's Michael Smith Genome Sciences Centre in Vancouver. You can either download the tool and install it on your PC or Mac (which requires some knowledge of Perl), or just use the online version (which is reasonably easy to use).

How to Get the Tool

I'd recommend using the online version if possible, but if you want to download Circos and install it locally, go to: http://mkweb.bcgsc.ca/circos/software/download/

For everyone else, the online version is at: http://mkweb.bcgsc.ca/circos/tableviewer/

How to Use the Tool / Examples

Using the online Circos Tableviewer involves just a couple of easy steps. First, you need to create a data file in the appropriate format. As the website notes, "appropriate format" means:

1. Your file must be plain text.
2. Your data values must be non-negative integers.
3. Data must be tab or space-separated.
4. Maximum number of rows/cols is 15.
5. No two rows or two columns may have the same name.
6. Things look best when rows/columns have a one character name (e.g. A, B, C, 1, 2, 3, ...).[6]

For this example, we'll use some data that shows the entry and exit colleges/divisions for new freshmen who started at Cal in Fall 2003. The spreadsheet looks like this (apologies for the truncated titles, but this is a wide table that doesn't fit here very well):

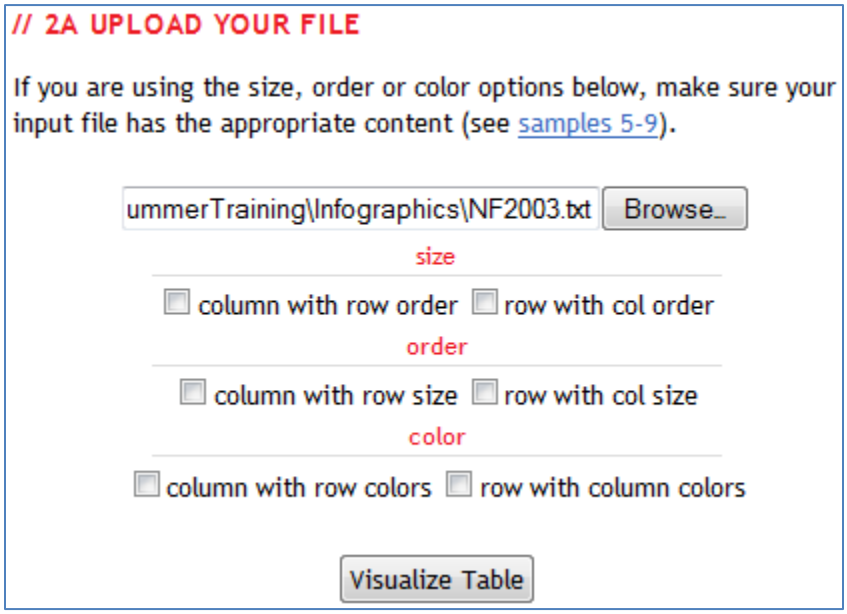
[5] Krzywinski, M. et al. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Res* (2009) 19:1639-1645.

[6] Data format rules slightly paraphrased from http://mkweb.bcgsc.ca/circos/tableviewer/visualize.

	Exit	CNR	Haas	Chem	COE	CED	LS-Other	LS-UGIS	LS-BS	LS-A&H	LS-PS	LS-SS	Not Grad
Entry	**Label**	**A**	**B**	**C**	**E**	**G**	**I**	**N**	**P**	**Q**	**T**	**V**	**Y**
CNR	**A**	105	7	4	3	3	19	10	15	15	2	27	30
Chem	**C**	4	0	97	10	0	2	2	8	5	3	8	20
L&S	**D**	79	182	49	45	26	219	335	480	288	88	678	279
COE	**E**	2	4	6	437	2	5	4	8	7	15	18	31
CED	**G**	0	0	2	1	91	0	4	4	4	0	9	11

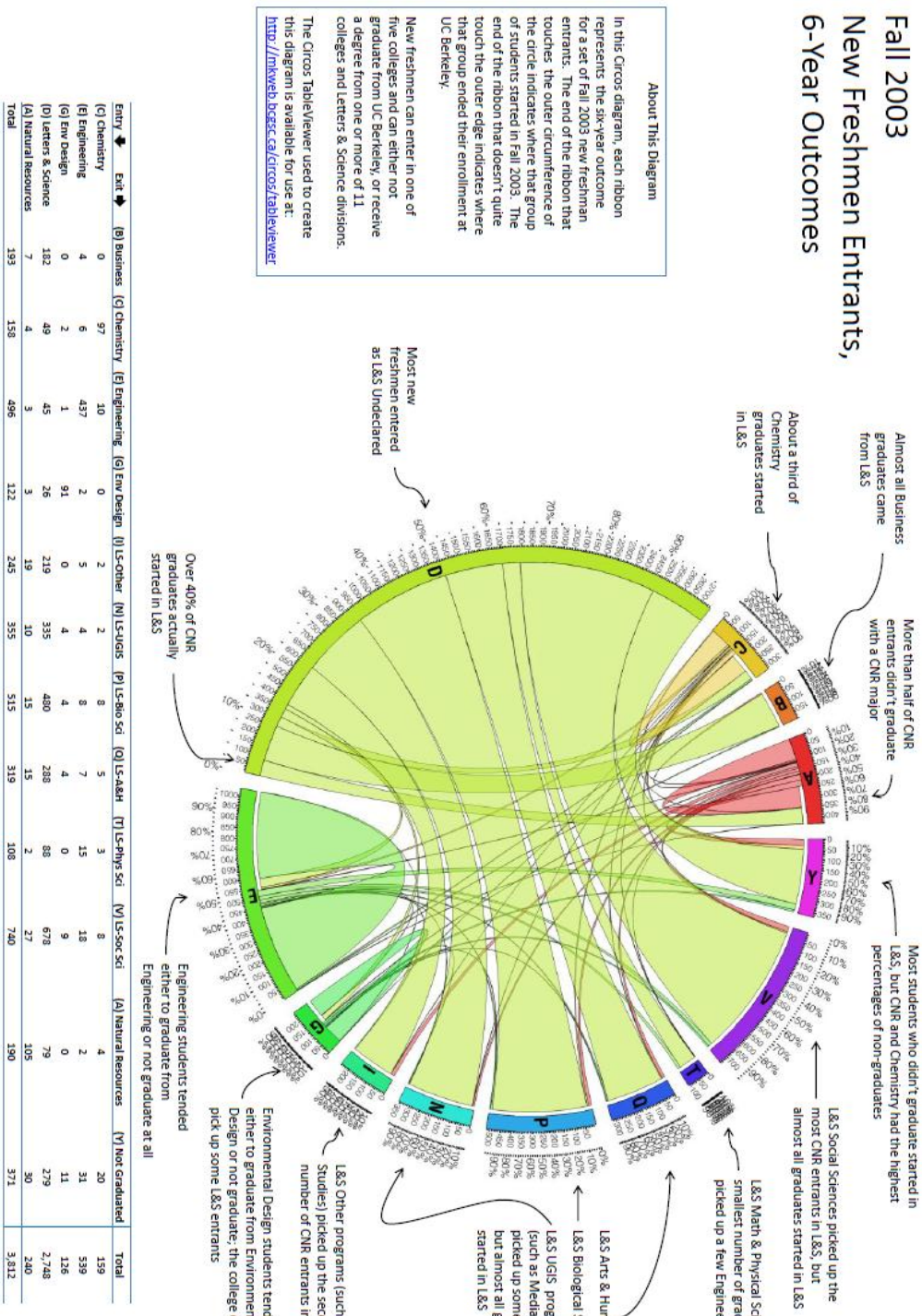
As you can see, we've assigned one-character codes to each of the entry and exit units, per rule #6 above. This means that when you use a Circos diagram in a report or supergraphic, you'll need to add some annotations to explain what's what.

Once you have a table like this in Excel, you can easily make a Circos-compatible file by just copying everything except the first column and top row, then pasting it into a text file (using Notepad, for instance). This makes a tab-delimited file that you can save and upload to Circos, which you do at http://mkweb.bcgsc.ca/circos/tableviewer/visualize:



There are lots of settings that you can play around with in Circos, but I usually just use the defaults, with the exception of turning off the contribution tracks. Once you've selected your input file, just click the Visualize Table button and wait a few seconds for the result. You can then enlarge the diagram by clicking on it, after which you can copy or download as you wish. (Note that because the file you uploaded here only has one-character codes and a bunch of numbers, privacy is much less of an issue than with most online tools.)

After adding some explanations and annotations (most of which aren't visible in this heavily squooshed version of a supergraphic; but it looks great at 11x17), you might wind up with something like this:



Entry ⬇ Exit ⬇	(B) Business	(C) Chemistry	(E) Engineering	(G) Env Design	(I) LS-Other	(N) LS-UGIS	(P) LS-Bio Sci	(Q) LS-A&H	(T) LS-Phys Sci	(V) LS-Soc Sci	(A) Natural Resources	(Y) Not Graduated	Total
(C) Chemistry	0	97	10	0	2	2	8	5	3	8	4	20	159
(E) Engineering	4	6	437	2	5	4	8	7	15	18	2	31	539
(G) Env Design	0	2	1	91	0	4	4	4	0	9	0	11	126
(D) Letters & Science	182	49	45	26	219	335	480	288	88	678	79	279	2,748
(A) Natural Resources	7	4	3	3	19	10	15	15	2	27	105	30	240
Total	193	158	496	122	245	355	515	319	108	740	190	371	3,812

References/Other Resources

I'm not aware of another tool that does the same thing as Circos, although there probably is one somewhere. If you decide to use Circos, however, the website (http://mkweb.bcgsc.ca/circos/guide/) has fairly extensive documentation on all aspects of the software.

For More Information About the Summer Series

http://idmg.berkeley.edu/summerseries.htm