

# Mining Your Data: An Easy Intro to a Tough Topic

Lynn Cherny

Ghostweather Research & Design, LLC

Boston Mini-UPA 2010

# Topics I'll Touch On

- Basic Data Description Using...
  - Simple Excel tricks
  - Useful Excel plugins
  - R and Rattle
- Text Data
  - “Cleaning” it in Excel, shell scripts
  - Online Resources like Many Eyes
  - Lexical analysis tools like Concordance software
- Cluster Analysis (and R)
  - Card Sort Data
  - Using Rattle for easier data exploration/clustering
  - Text Mining and dendrograms

# A Cautionary Tale

4	4	4	4	4	4							4	4		
4	4	4	4	4	4							4	4		
im 5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im		
im 5 - very im	5 - very im	5 - very im	5 - very im	5 - very important	5 - very important	5 - very im	5 - very im	5 - very im	5 - very important			2 5 - very im			
2	2 5 - very im	5 - very im		4	4							1 - not imp	1 - not imp	3	4
4	4	5 - very im	5 - very im	5 - very im	5 - very important									4	4
im 5 - very im	5 - very im	5 - very im	5 - very im		2			5 - very important						5 - very im	5 - very im
im	4 5 - very im	4 5 - very im	4 5 - very im	4 5 - very im	5 - very important									4	4
im 5 - very im	5 - very im		4	4 5 - very important				5 - very important						4	4
4	4	4	4	4	2			5 - very important				1 - not imp	1 - not imp	1 - not imp	
4	4		4	4	2			5 - very im	5 - very im	5 - very important		5 - very important			
3	2 5 - very												1 - not imp		3
4	4													3	4
4	4													4	5 - very im
im 5 - very im	5 - very im	5 - very im	5 - very im	5 - very important	5 - very important	5 - very im	5 - very im	5 - very im	5 - very important			5 - very im	5 - very im	5 - very im	5 - very im
4 5 - very im		4	4	4	4	5 - very im	5 - very important							3	4
im 5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very important	5 - very important		5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im
3	2	3	3	4	3 5 - very important									3	4
2	2 5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	1 - not imp	5 - very im	5 - very im	1 - not imp	1 - not imp	1 - not imp			4 5 - very im	1 - not imp
4	4 5 - very im	5 - very im	5 - very im	5 - very im		2 1 - not important					1 - not important			3 5 - very im	5 - very im
4	4	4	4	4	4									4	4
4	3 5 - very im		4	4	2	1 - not important					1 - not important			2	4 1 - not imp
4	4	4	4	3	3									3	4 5 - very im
2	4	4	4	5 - very im	3									2	4
3	3	4	4	4	2	5 - very important								4	4
im 5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very important			5 - very important						3	4
im	4	4 5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very im	5 - very important					4	5 - very im
3	2	4	2	2	4							5 - very im	5 - very im	3	3
4	4	4	4	4										4	4
3	3	4	3	3	1 - not imp	1 - not important						5 - very important		2	2 1 - not imp

# Simple Data Basics in Excel

- Transform Data to Plottable – text to numeric, etc.
- Histograms with Pivot Tables
- Explore visually – anything odd?
  - Outlier removal

# DEMO Excel Tricks

# Excel 07 Bar Defaults Defects

A		B		C		D	
	5		67		14.6		45
	3		0		15.2		-10
	2		45		14		33
	0		46.271		16.2		34
	8		985		13.67		12
must be min of 10% of the bar even if 0				odd default scale			
gradient				negatives are = 0, or 10% of cell width			

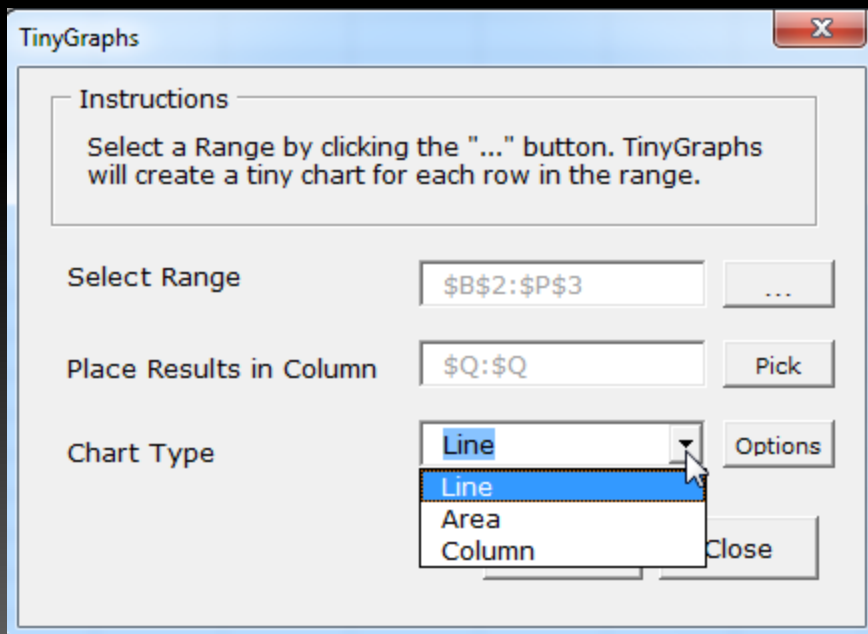
Beware of the built-in bars!

*Examples courtesy of Alex Kerin*

# Sparklines in Excel: Addins Till '10

Tiny Graphs: <http://www.spreadsheetsml.com/products.html>  
Sparklines for Excel: <http://sparklines-excel.blogspot.com/>  
Sparkmaker: <http://www.bissantz.com/sparkmaker/>  
Microcharts: <http://www.bonavistasystems.com/>

*Ref list courtesy  
Alex Kerin*



O	P	Q
2009	2010	Trend
2691873	2744838	
778015	780764	

# Reminders: Basic Data Description

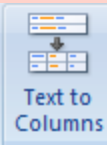
- “Gotchas” to look out for –
  - Distributions, outliers, and data integrity!

1. Rate this...					
	Very important				Not important
Fun	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Intelligent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interesting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Handsome	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- Advanced descriptive moves: R, Principal Components/Factor Analysis et al.



# Useful Excel Text Functions

Goal	Function	Use Cases
Extract first word in cell	<code>=IF(LEN(A1)=0,"",IF(ISERR(FIND(",A1)),A1,LEFT(A1,FIND(",A1)-1)))</code>	<i>"SW2009 SP1 x64" → "SW2009"</i>
Extract first words after a specific word, delim by spaces	Find the word after "P/N" <code>=MID(A11,SEARCH("P/N ",A11)+4,SEARCH(" ",MID(A11&amp;" ",SEARCH("P/N ",A11)+4,200))-1)</code>	<i>"P/N 432435" → "432435"</i>
Count words in a string	<code>=IF(LEN(TRIM(A1))=0,0,LEN(TRIM(A1))-LEN(SUBSTITUTE(TRIM(A1),CHAR(32),""))+1)</code>	<i>"I really hate your product, let me tell you why... it can be summarized in 3 paragraphs"</i>
Concatenate strings in different cells (w/ space)	<code>=CONCATENATE(B1, " ", A1)</code>	<i>"Frank"   "Charles" → "Charles, Frank"</i>
Text to Columns in Data Tab		<i>"SW2009 SP1" → SW2009   SP1</i>
Transform a string containing an odd element	<code>=IF(NOT(ISERR(FIND("Overclock",L2))),CONCATENATE(SUBSTITUTE(L2,"Overclock at ",""),"(oc)",L2)</code>	<i>"Overclock at 3 GHz" → "3 GHz (oc)"</i>

# Text Data Manipulations

- Cleaning with Excel – Splitting, stripping, functions to help (we saw some of this...)
- (With lots of files) Faster to use shell scripting tools (awk, sed, cygwin/unix) (quick code samples)
- Python examples for the more serious: Why, and how to store it for use later or output for Excel again

# Shell Ops for Data Munging

**First:** Install Cygwin on windows (or use a linux system command line)

- Useful commands – *grep*, *cat*, *>* (redirect output to a file), *cut*, *paste* (for field manipulations)

Ex: *grep [word/pattern] file[s] > output.txt*

This will find all lines with [word/pattern] in a set of files and save them to the single file “output.txt”

- *Awk*, *sed* for complex pattern manipulations

# Demo cygwin unix file munging

# What did I just do...

- Find all the lines containing just <graphics> in a bunch of embedded directories, save in one file
- Process that file to remove all but the graphics lines
- Then cut out JUST the refresh rate column
- Count occurrences of a particular line in file
- Convert spaces to commas (or tabs...)

# Move on....

## Text Data – Issues and Challenges

- It can be messy – mixed case, punctuation, misspellings, ungrammatical, long, fragmentary...
- Time-consuming to analyze: Most quant companies ignore it in surveys or do a bad job.
- Hard(er) to visualize/compare
- Requires judgments about what's “important”

I bought Adobe because I think of them as the industry standard when it comes to anything creative.

Ulead is very simple to use and cheap in comparison - it offers a lot for that. But there it's also very lim

I often feel looked down on because I use Movie Maker, but I love it. It does everything I need it to do

This hobby is ridiculously cost-prohibitive for the average non-professional user.

Magix Movie Edit was the first one I used, so I stick with it. However, I am proficient in using and do

I know that I need to get better software, but I don't have the money to do so and I don't feel inclined

I've used a lot of different software over the years and it really doesn't matter which one you use for th

I'm currently trying to get After Effects though from the trial it still seems pretty complicated.

I use a variety of other tools. Riverpast Video Perspective to force 16x9 into a 4x3 box, the Llama Enc

Weeeeell I did want to switch to a newer version, but the newer one ate RAM like candy so I am back

Final Cut Pro is generally amazing, but very fussy about the formats it will accept, and of course, it cos

We don't use as much software as other people; we capture straight into Final Cut (and into Premiere

I wish that clips 1 & 3 would stay in place when I delete or move clip 2.

I wish I had a better understanding of the tech that I use. I know I've only scratched the surface, and I

I barely scratch the surface of what FCP can do because I do a lot of straight cutting. Some things the

Am learning to use Motion, which has a \*terribly\* counterintuitive interface but can do amazing things

Lots of memory is the best suggestion I can give. It makes all the difference. I wish FCP were cheaper

I am constantly amazed by how powerful these tools are. After 8 years, I still feel like I'm just scratchi

The only thing that I don't like

# Tools for Text Data Exploration

- Wordles – clever text clouds
- Many Eyes
- Concordance/lexical analysis tools
- Write your own in python/perl or R



# Wordles & Issues

<http://www.wordle.net/>

Buffy the Vampire Slayer	Doctor Who/Torchwood	Heroes
Pandora Hearts	Shugo Chara	Final Fantasy
Highlander	Star Trek (TOS)	Supernatural
Invisible Man	Joan of Arcadia	Multi-media
Hornblower	Supernatural	Xena
Battlestar Galactica	Supernatural	Firefly/Serenity
Firefly	Buffy the Vampire Slayer/Angel	Tru Calling
Battlestar Galactica (2)	Other fandoms (Heroes, Pushing Daisies, lost, Buffy)	
Firefly	Buffy the Vampire Slayer	House
Supernatural	Rent	Stage musicals
Queer as Folk (oh, god)	Buffy the Vampire Slayer/Angel	Six Feet Under, Veronica
The Tudors	Robin Hood (BBC)	Merlin (BBC)



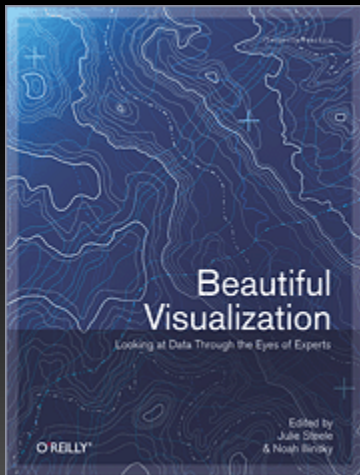
Issues – Raw data has spaces inside strings and needs normalizing/cleaning; Wordle strips out common “stopwords” (like “the”) which you may not want removed.

# FYI: Excellent Articles on Wordles

Table 4: Percentage of respondents who did not know what font size means in a Wordle

	Male	Female
<b>Under 20</b>	<b>35</b>	<b>49</b>
20-30	12	18
Above 30	19	31

Participatory Visualization With Wordle –  
Viegas, Wattenberg, Feinberg  
[http://www.research.ibm.com/visual/papers/wordle\\_final2.pdf](http://www.research.ibm.com/visual/papers/wordle_final2.pdf)



Chapter by Jon Feinberg himself:  
[http://static.mrfeinberg.com/bv\\_ch03.pdf](http://static.mrfeinberg.com/bv_ch03.pdf)

# Many Eyes Visualizations of Text

- Tag Clouds / Wordles too
- Phrase Networks
- Word Trees
  
- And/Or use concordance software...

# Demo Many Eyes and Concordance software

# Workflow to try on Many Eyes

- Wordle or Tag Cloud – what's common?
- Look at those word(s) in networks and trees to understand the context

The image shows a word cloud with various words in different colors and sizes. Two pop-up boxes are overlaid on the cloud, providing context for specific words. The first box is for the word 'long' and shows six occurrences from a text source. The second box is for the phrase 'secret earth' and shows three occurrences from a different text source.

**long**  
Showing 6 of 27 occurrences

- ...rise and fall All day LONG above the fire!" Red...
- ...lay me down in this LONG grass And close my eyes...
- ...I passed. All my life LONG Over my shoulder have I...
- ...fain would lie in this LONG grass And close my eyes...
- ...Cat birds call Through the LONG afternoon, and creeks at dusk...
- ...passionate eye can reach, And LONG, ah, long as rapturous eye...

**secret earth**  
Occurrences: 3

- ...your big eyes In the SECRET EARTH securely, Your thin fingers, and...
- ...some way, surely, From the SECRET EARTH shall rise; Not for these...
- ...shall the chemistry Of the SECRET EARTH restore. All your lovely words...

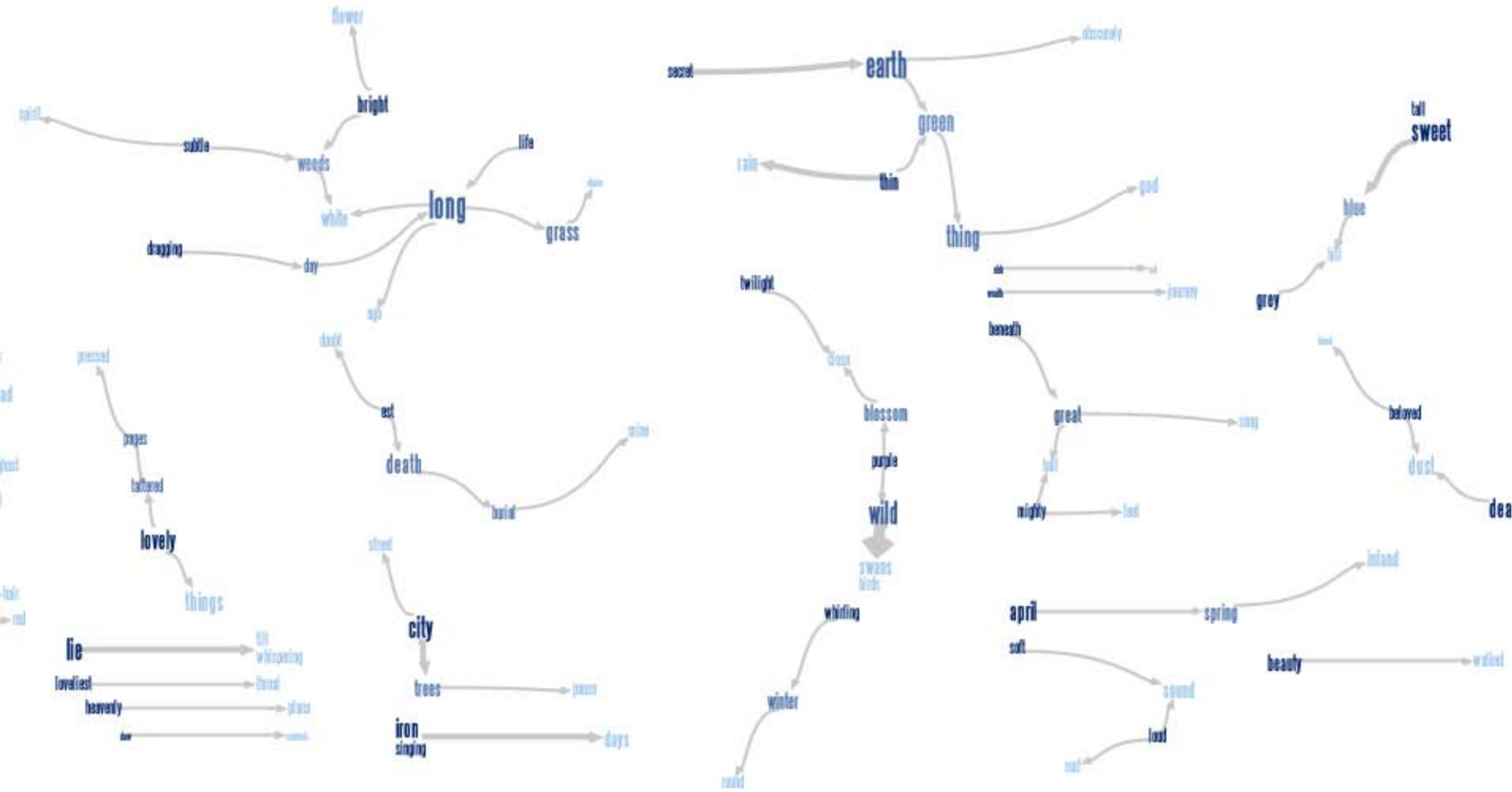
*Tag cloud shows context on rollover*

(When you get frustrated, download Concordance software.)

# Beware the settings...

word1 [space] word2  
or enter your own  
\* \* Submit

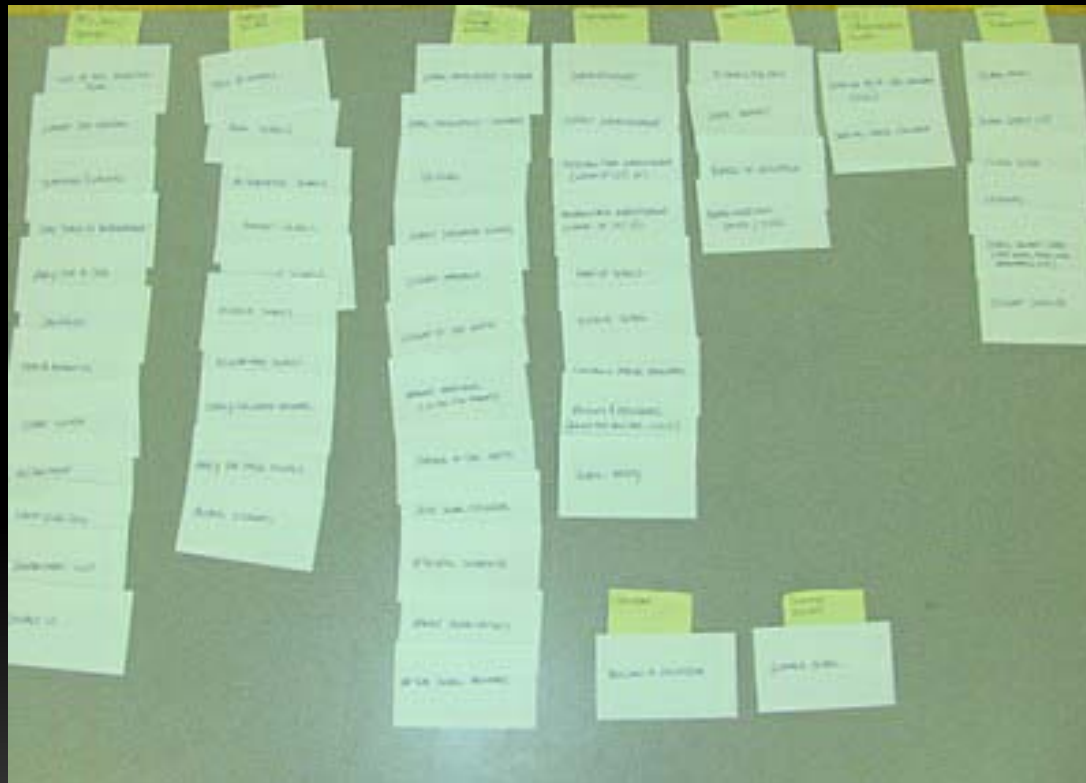
Filters  
Show top: 150  
Hide common words



# Cluster Analysis

- Intuitive idea – group related items in terms of “distance”
- Example Uses:
  - Cluster cards based on multiple subjects’ groupings
  - Find similarities among different ratings in survey data
  - Input to persona definitions, based on appropriate survey question data
- Use for “grouping” related text comments/docs

# Card Sorting...






# Steps for Card Sort Data Clustering in R

- Create binary matrix in Excel: Cards as col1, ALL groups created as other columns (with 1 or 0 for each card)
- Import as tab or csv delim text in R
- Use only the groups for a distance matrix calculation (calculates relationship between cards)
- Cluster on the distances
- Print a tree, and figure out how many “groups” you care about! Label, output, etc.

# Format: Matrix of Cards...

All cards given  
out

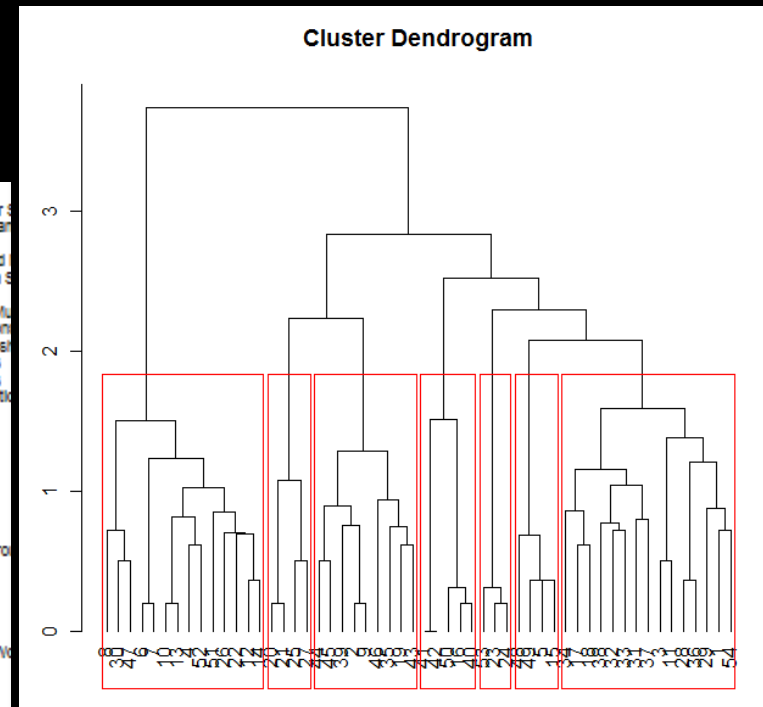
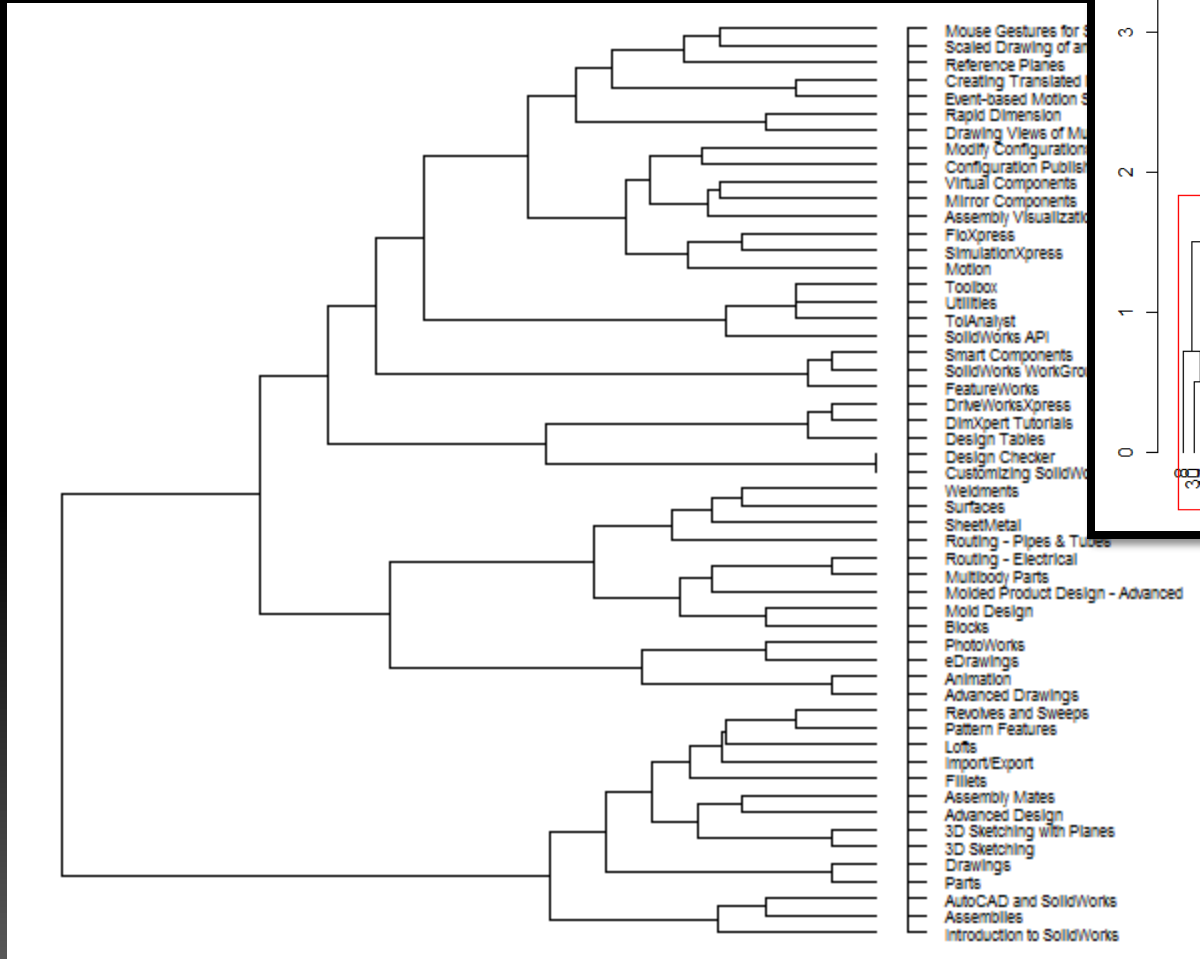
Unique column headers for all groups created at card  
level



	Group1_by_user A	Group2_by_user A	Group1_by_user B	Group2_by_user B
CardName1	1	0	1	0
CardName2	0	0	1	0
CardName3	1	0	0	1
CardName4	0	1	1	0
CardName5	0	1	0	1
CardName6	0	1	0	0
CardName7	0	0	0	0

*Note: The cluster is about groups, not about users' label input. I understand Donna Spencer's spreadsheet tool will create this for you.*

# Results: Decide How Many You Want



# R Code (in case you want to try)

The distance matrix between all the cards:

```
> d <- dist(cards[,2:107], method="binary")
```

The clustering:

```
> fit <- hclust(d, method="ward")
```

Plot it vertically:

```
> plot(fit, hang=-1, labels=cardnames)
```

Make 7 groups from the cards:

```
> groups.7 <- cutree(fit, k=7)
```

Outline the 7 in red:

```
> rect.hclust(fit, k=7, border="red")
```

Make horizontal plot with card labels:

```
> dend <- as.dendrogram(fit)
```

```
> plot(dend, horiz=TRUE, leaflab="none")
```

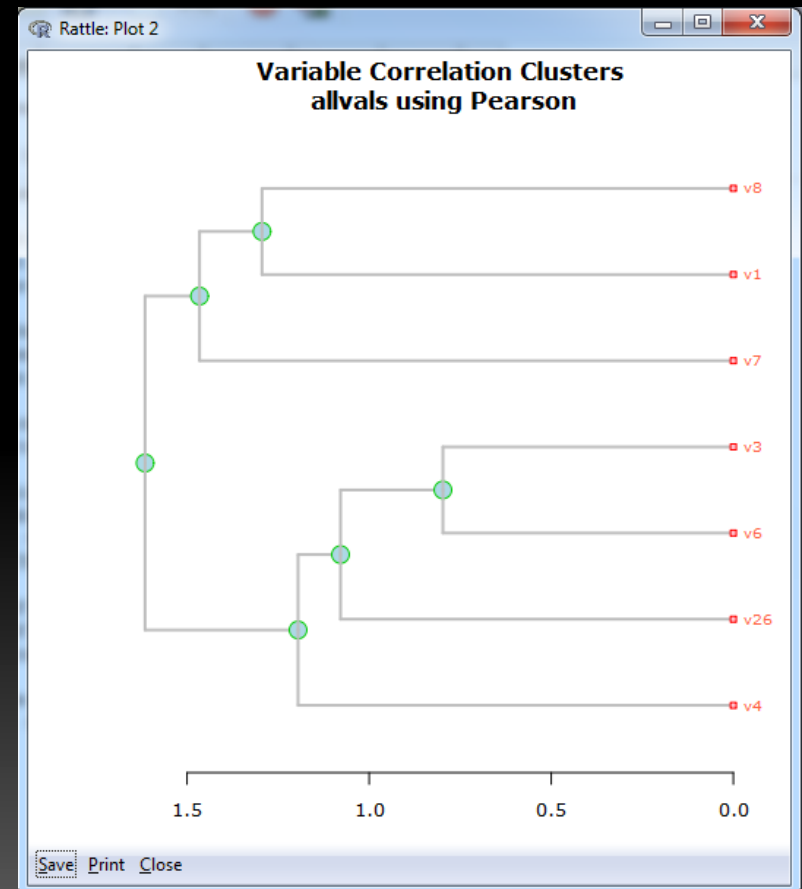
```
> axis(4, at=c(1:54), labels=cardnames, cex.axis=.5, las=2)
```

Get the first groups' cards labels:

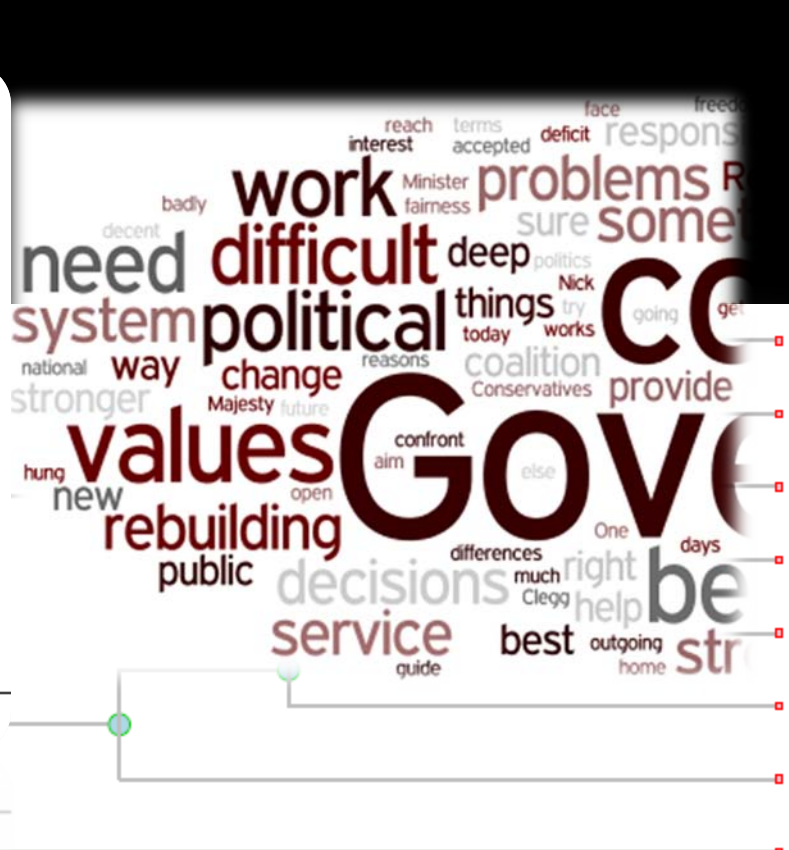
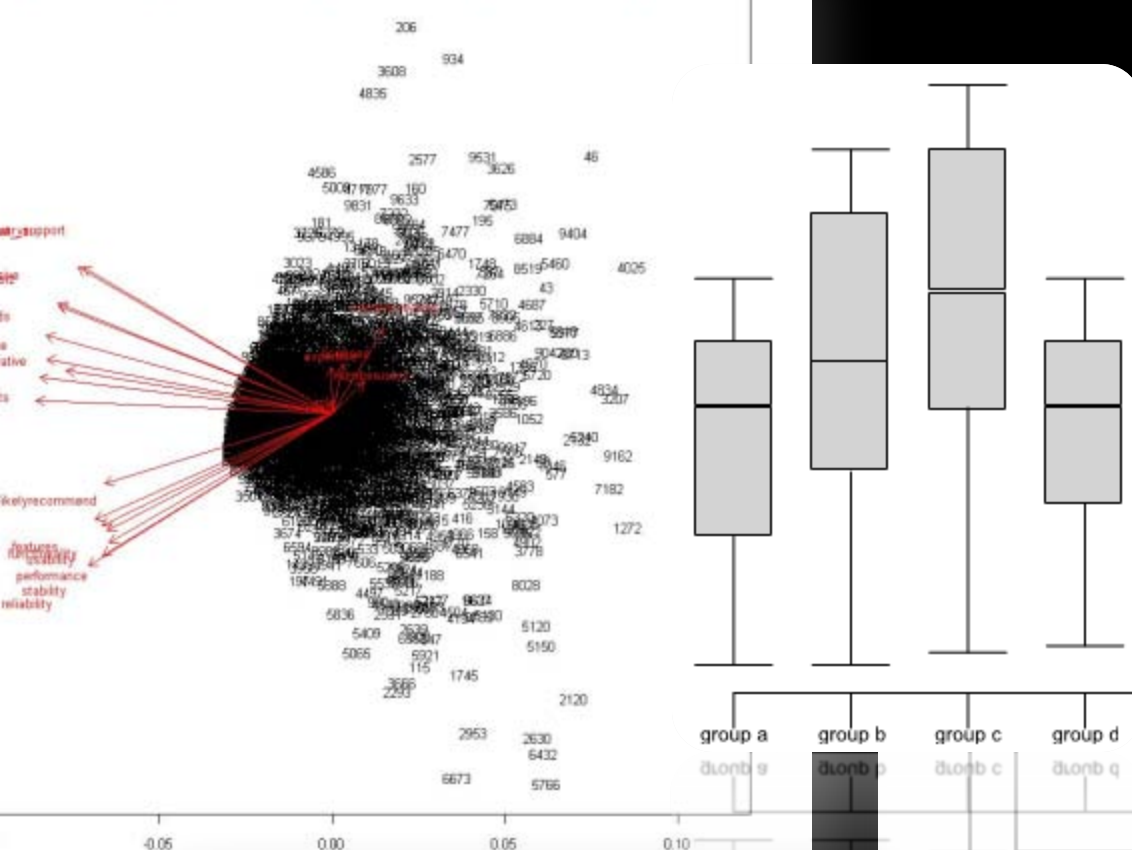
```
> names(groups.7[groups.7==1])
```










# Related: Cluster Numeric Survey Responses

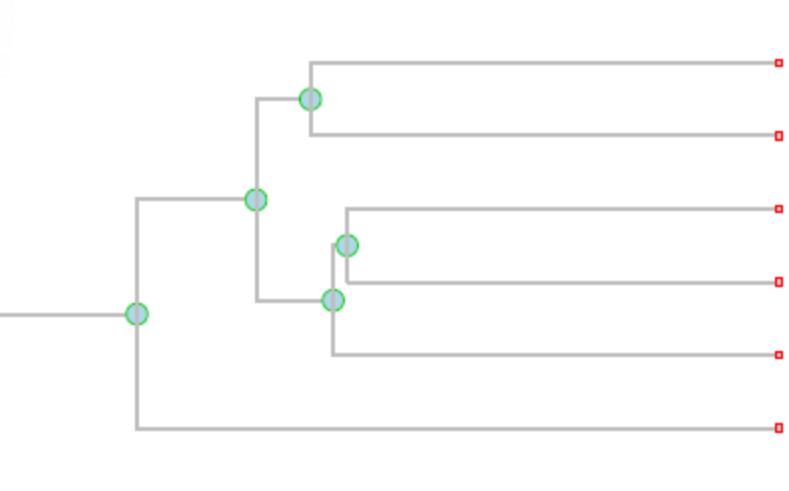
Easiest to use  
Rattle, an R tool with  
a (weird) GUI for  
data exploration



# Demo Rattle – R gui tool



	54	27	33	82
	82	61	50	70
	12	77	65	22
	54	27	33	82
	82	61	50	70
	12	77	65	22
	70	50	70	62
	12	77	65	22
	12	77	65	22



Thanks for attending...  
(references lists follow)

Questions or comments?    Need consulting on your  
data?

[Lynn @ ghostweather.com](mailto:Lynn@ghostweather.com)

Ghostweather Research & Design, LLC



# A Few Excel References

Daily Dose of Excel: [www.dailydoseofexcel.com/](http://www.dailydoseofexcel.com/)

Text Functions from MVPS [www.mvps.org/dmccritchie/excel/strings.htm](http://www.mvps.org/dmccritchie/excel/strings.htm)

Advanced Excel: Spreadsheet Analytics: <https://sites.google.com/a/usfca.edu/business-analytics/>

Juice Analytics on more excel tricks for in-cell graphs: [www.juiceanalytics.com/writing/more-on-excel-in-cell-graphing/](http://www.juiceanalytics.com/writing/more-on-excel-in-cell-graphing/)

Resources, training, consulting on excel (I took his Advanced Dashboard Design class co-taught with Alex Kerin and loved it.) [peltiertech.com/](http://peltiertech.com/)

A good book on Excel tricks and data presentation: **Excel 2007 Dashboards & Reports For Dummies**, by Michael Alexander (This is actually a fairly advanced book (it influenced a lot of the Peltier-Kerin Excel course I recently took.)

Sparkline addins - free or cheap (but sparklines will be built in to Excel 2010):

Tiny Graphs: [www.spreadsheetsml.com/products.html](http://www.spreadsheetsml.com/products.html)

Sparklines for Excel: [sparklines-excel.blogspot.com/](http://sparklines-excel.blogspot.com/)

Sparkmaker: [www.bissantz.com/sparkmaker/](http://www.bissantz.com/sparkmaker/)

Microcharts: [www.bonavistasystems.com/](http://www.bonavistasystems.com/)

Juice Analytics' Excel Training Worksheet: [www.juiceanalytics.com/writing/excel-training-worksheet/](http://www.juiceanalytics.com/writing/excel-training-worksheet/)

Excel Pivot Table Tutorials, especially for tips on grouping data: [www.contextures.com/xlPivot07.html](http://www.contextures.com/xlPivot07.html)

Tag Cloud VBA for Excel: [chandoo.org/wp/2008/04/22/create-cool-tag-clouds-in-excel-using-vba/](http://chandoo.org/wp/2008/04/22/create-cool-tag-clouds-in-excel-using-vba/)

# Unix, R, Many Eyes, Etc.

Cygwin for using command line unix commands, or use a linux installation shell window: [www.cygwin.com/](http://www.cygwin.com/)

Linux Reference card for useful commands:

[www.cfa.harvard.edu/~jbattat/computer/linuxReferenceCard.pdf](http://www.cfa.harvard.edu/~jbattat/computer/linuxReferenceCard.pdf)

More detailed, with links to explanations/samples: [www.perpetualpc.net/srtd\\_commands\\_rev.html](http://www.perpetualpc.net/srtd_commands_rev.html)

Wordles or Many Eyes: [manyeyes.alphaworks.ibm.com/manyeyes/](http://manyeyes.alphaworks.ibm.com/manyeyes/), wordle.net

Concordance Software for text analysis - not free, but very cheap and has a 30-day trial.

[www.concordancesoftware.co.uk/](http://www.concordancesoftware.co.uk/)

Simple Graphs with R: [www.harding.edu/fmccown/R/](http://www.harding.edu/fmccown/R/)

Rattle, a GUI for data exploration in R: [rattle.togaware.com/](http://rattle.togaware.com/)

Cluster Analysis in R: [www.statmethods.net/advstats/cluster.html](http://www.statmethods.net/advstats/cluster.html)

Online courses in statistics and R, including topics that cover cluster analysis: [www.statistics.com/](http://www.statistics.com/) (I have taken at least 4 of their courses now... beware, they tend to be advanced!)

For card sorting: Use this spreadsheet tool to generate the columns of 1/0's for each card and group, or do it by hand:

[http://www.boxesandarrows.com/view/analyzing\\_card\\_sort\\_results\\_with\\_a\\_spreadsheet\\_template](http://www.boxesandarrows.com/view/analyzing_card_sort_results_with_a_spreadsheet_template)

Then go into R with this matrix, or email me.