

# 3 Essential steps for better data visualization

Dr. Abby Benninghoff

UtahStateUniversity

## *3 Essentials*

- 1 Use the right kind of plot for your data**
- 2 Figures should elucidate information, not obfuscate**
- 3 Apply principles of effective formatting**

UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



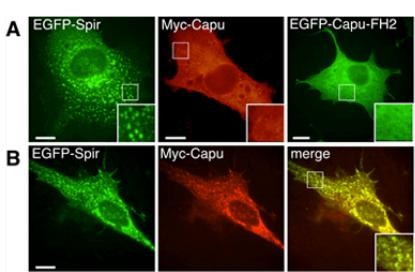
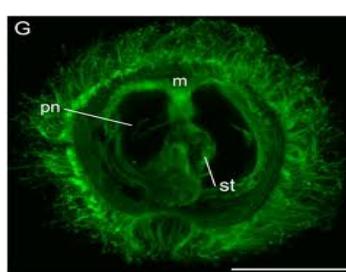
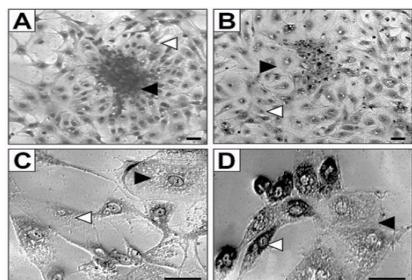


# 1 Use the right kind of plot for your data

UtahStateUniversity

## Pictures as figures

- Never assume that anyone knows what is in a picture
  - Use arrows, markers to identify features
  - Include scale bar
  - Specify meanings of colors
  - Include key explanations in figure legend

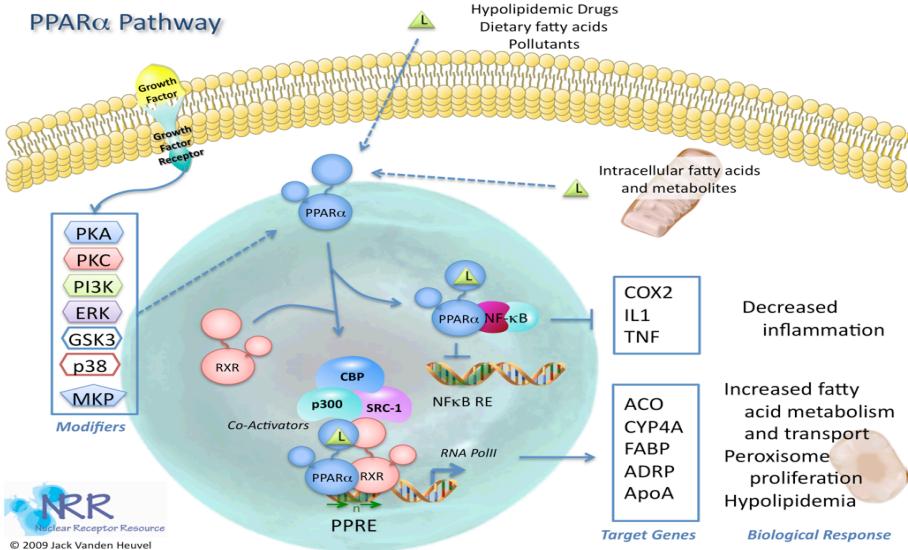


UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



# Use diagrams to convey complex ideas



## This works.

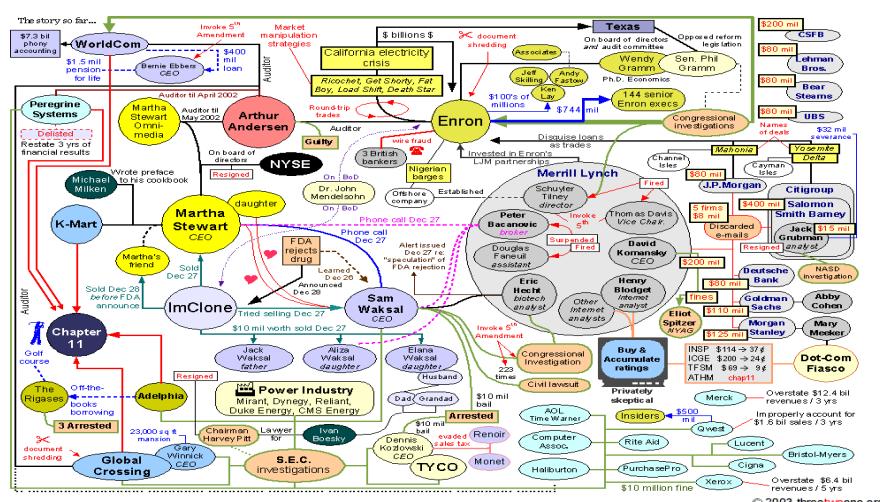
Particularly as an overview in an introduction section or a review paper.

But, probably too much for a quick slide in a presentation.

Utah State University

*3 Essential steps for better data visualization. January 20, 2016*

## Use diagrams to convey complex ideas



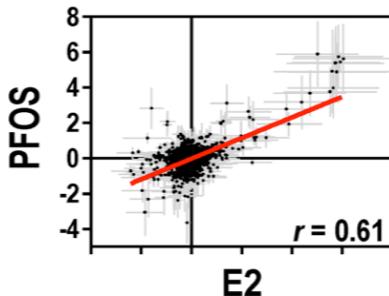
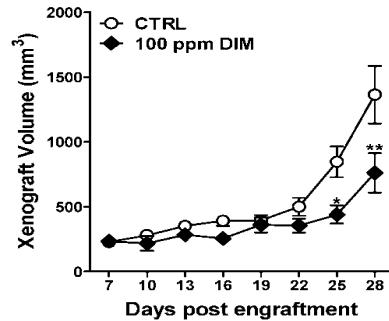
This. not so much.

Utah State University

3 Essential steps for better data visualization. January 20, 2016

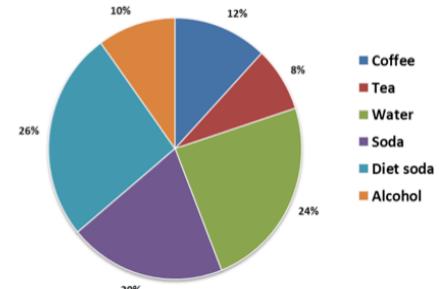
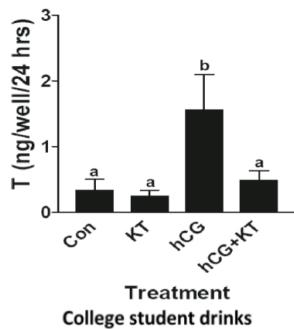
## What type of graph?

- **Line graphs** for dynamic comparisons
  - Dependent variable changes with respect to time, dose, etc.
  - Do not over-crowd with multiple (>4) lines
  - Use curve-fitting when appropriate (provide curve fit model)
- **Scatter plot** for correlations
  - Most commonly made in two dimensions
  - Line of best fit – linear regression
  - Show  $r$  statistic, confidence interval lines or other indication of robustness of the data



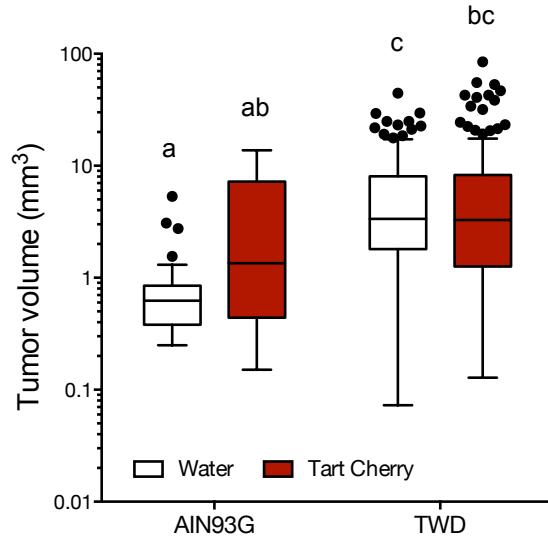
## What type of graph?

- **Bar graph** for subdividing and comparing data
  - Preferred as vertical rather than horizontal
  - Used if no continuum of data points that would work as a line graph
  - Make bars the same width, space between is one-half a bar
- **Pie chart** to compare parts of a whole
  - Useful for visually showing how much subgroups contribute to the whole
  - Problematic if too many groups are shown
  - Problematic if color not allowed (shading okay, but patterns look bad)



# What type of graph?

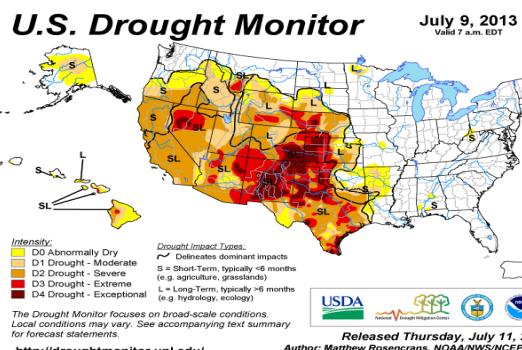
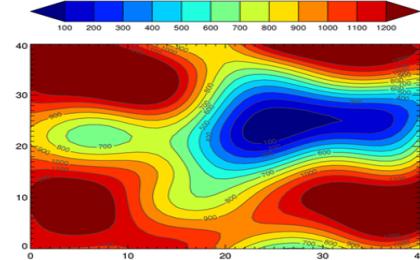
- **Box plot** useful for descriptive statistics
  - Will highlight differences in data set populations
  - Box outline shows 25<sup>th</sup> to 75<sup>th</sup> percentile
  - Line shows the mean
  - Error bars show a defined distance set by user (min to max, 10 to 90<sup>th</sup> percentile)
  - Dots indicate possible outliers



# What type of graph?

## Contour plot

- Useful for multi-dimensional data
- Contour lines indicate regions that fit within a defined scale
- Shows scaled values in relationship to two experimental factors or other reference measurements on XY axes or other coordinate system
- Color is very effective in these plots



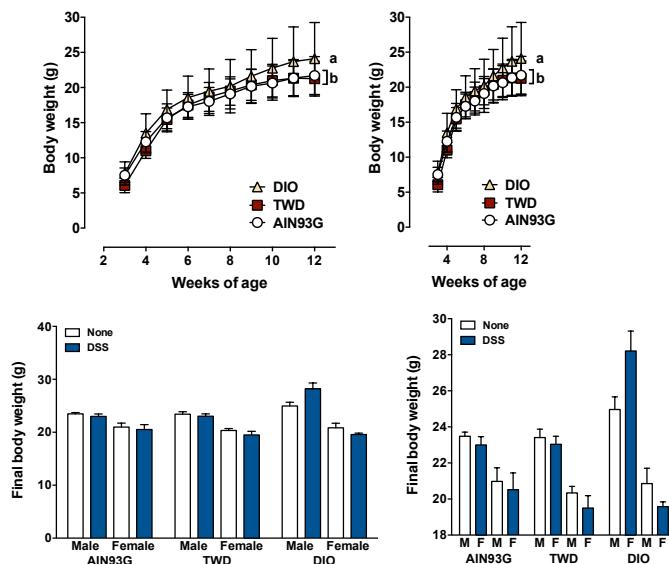


## Graphs should elucidate information, not obfuscate

UtahStateUniversity

### Do not manipulate to (de)emphasize observations

- Do not manipulate the figure to sway the reader or inappropriately emphasize or disguise trends
- Do not start axes at midpoint in scale without visual cue (e.g., axis break)



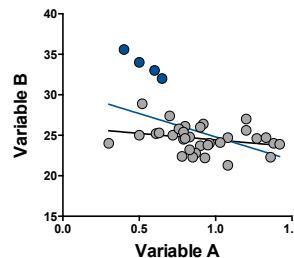
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



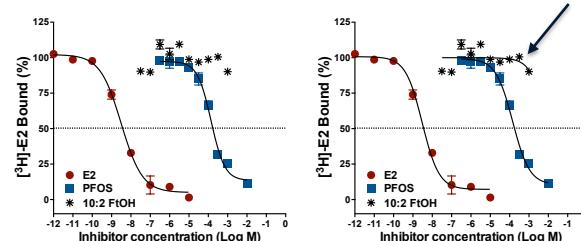
# Including/excluding data

- Do not arbitrarily delete data points without clear justification
    - Experimental error documented in your laboratory notebook
    - Statistical test for outliers
  - Don't "massage" line fits or change parameters post-hoc to best fit your data



*Consider how the four points colored blue influence the regression fit.*

Are these outliers? Do they have an outsized impact on the apparent trend?



*A switch from 4-parameter to 3-parameter curve allows for fit of 3<sup>rd</sup> data set, but is this post-hoc change in analysis appropriate?*

3 Essential steps for better data visualization. January 20, 2016

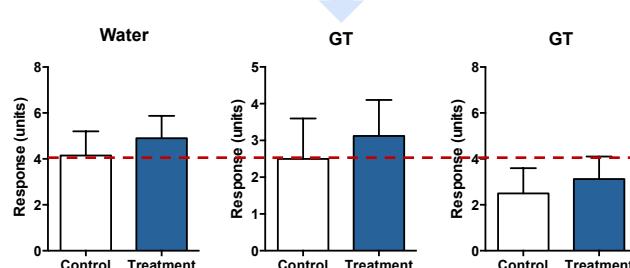


Utah State University

# Presenting multiple figures

- Similar data presented jointly should be shown at the same scale (most of the time) to enable appropriate comparisons
  - Data groupings should infer what type of analysis was performed.
    - This presentation suggests two separate Student's  $t$ -tests. But what if a two-way ANOVA were performed?

The center panel uses a different scale for the same type of data as in the left panel. Clearly, an comparison between "Water" and "GT" is intended. But the different scales obfuscate the comparison.



The right panel corrects this problem. Now you can see more easily that the response was overall a bit lower for both control and treatment in the GT group compared to the Water group. The dashed red line added helps you visualize the impact of the selected scale.

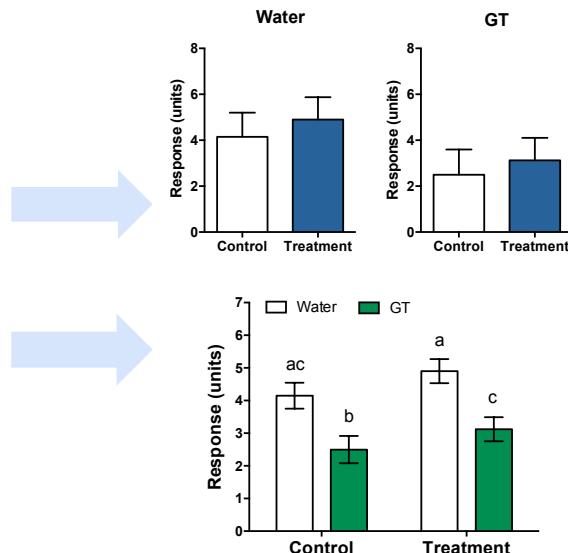
Utah State University

3 Essential steps for better data visualization, January 20, 2016



## Graph style should reflect experiment design

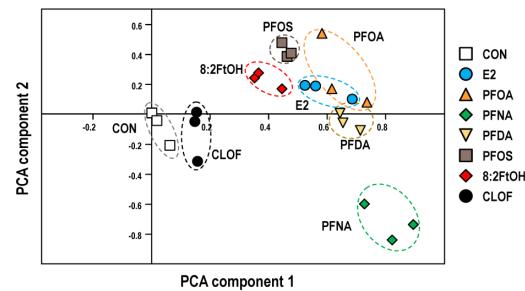
- Data groupings should reflect the experiment design and the type of statistical analysis performed.
  - This presentation suggests two separate Student's *t*-tests. But what if a two-way ANOVA were performed?
  - This presentation groups the data by control/treatment and by Water/GT, appropriate for this 2x2 experiment design.
  - Note that the first approach did not allow for the most interesting comparison, water vs. GT!



## Apply principles of effective formatting

# Characteristics of well-designed figures

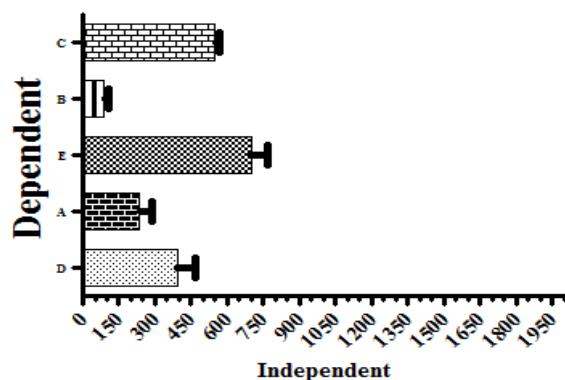
- **Neatness** – Image is clean and sharp, invites attention
- **Readability** – Eye can discern important information easily, quickly
- **Font** – Text is large enough to be read, placed well and limited
- **Size** – Graph is sized appropriately for anticipated reduction during printing.
- **Aesthetics** – Balanced graphs, good use of white space, eyes drawn to most important features
- **Use of color** – Distinguishes important content, pleasing to eye
- **Consistency** – Similar graphs have same stylistic scheme (line width, font type and size, labeling, scales, etc.)



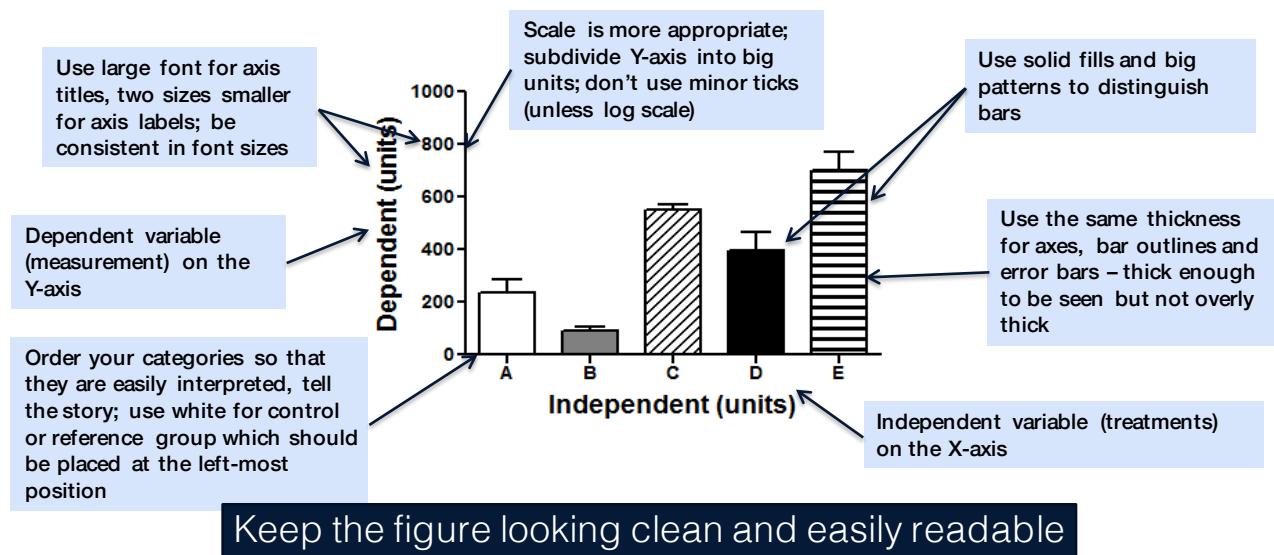
## A poorly formatted graph

Let me count the problems

1. No units
2. Scale isn't appropriate
3. Axis, error bars and bar outline thickness
4. Font on Y-axis too small, serif font
5. Axis title fonts uneven
6. Dependent and independent variables not on the right axes
7. Minor ticks not helpful
8. Too many X-axis labels
9. Patterns are rather obnoxious, difficult to distinguish
10. Categories on Y-axis are not in any useful order



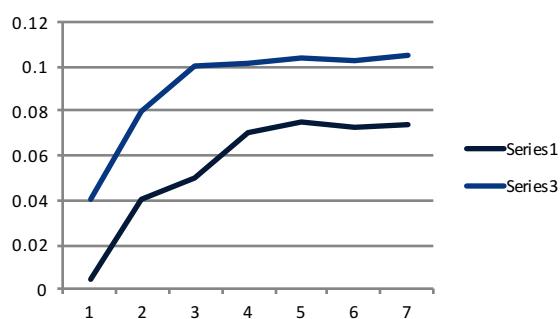
# How can I fix it?



## Poorly formatted line graph

Straight from Excel with no tinkering.  
What is missing? What are the problems?

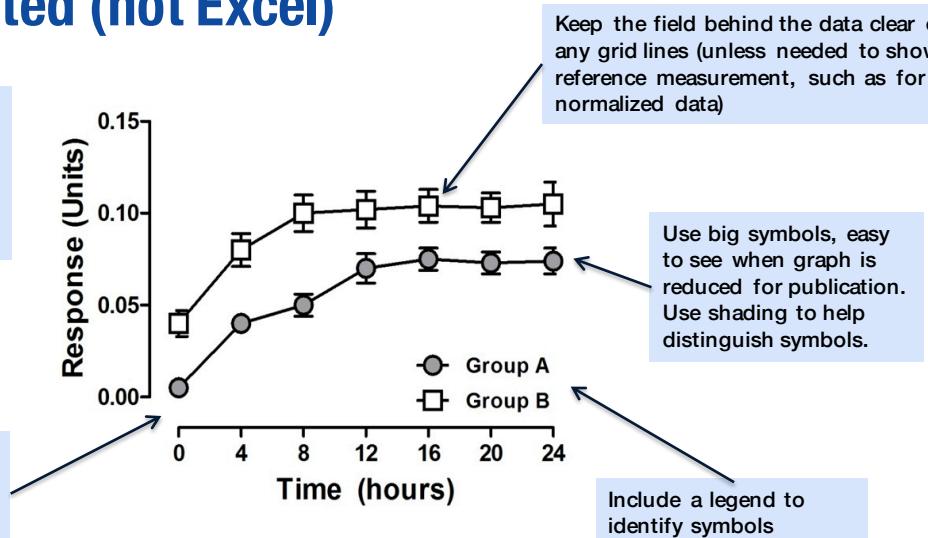
1. Grid lines are distracting
2. No axis labels
3. Category labels are not informative
4. Series aren't labeled
5. Position of data points not clear (no symbols)
6. No error bars
7. Fonts too small



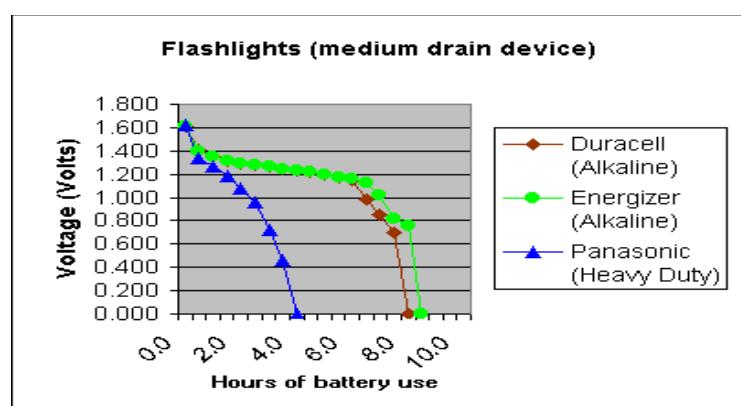
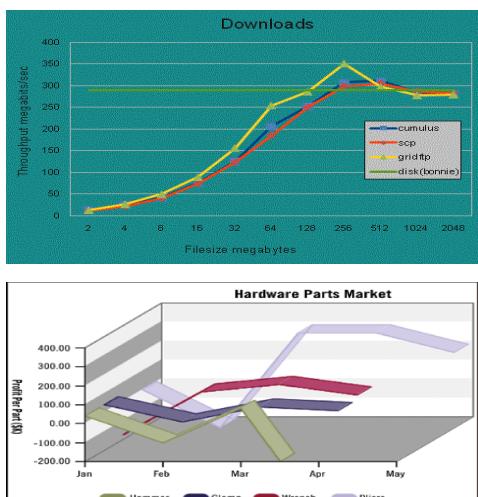
## Reformatted (not Excel)

Follow previous recommendations for font sizes, axes thickness, etc. Also use the same thickness for symbol outline.

If the first data point falls on top of the Y-axis, then offset the axes for clarity.

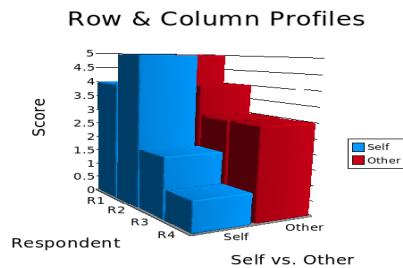


## Examples of poorly done graphs

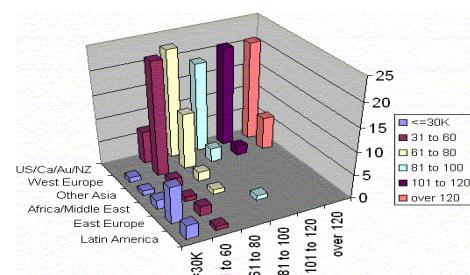
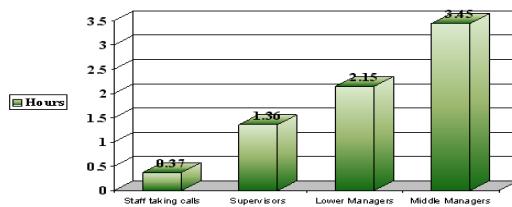


## The 3D temptation

- Do not use 3D unless absolutely necessary.
- Many other ways to show 3 levels of data without using a rotated 3D graph



How many hours surfing the internet per day?

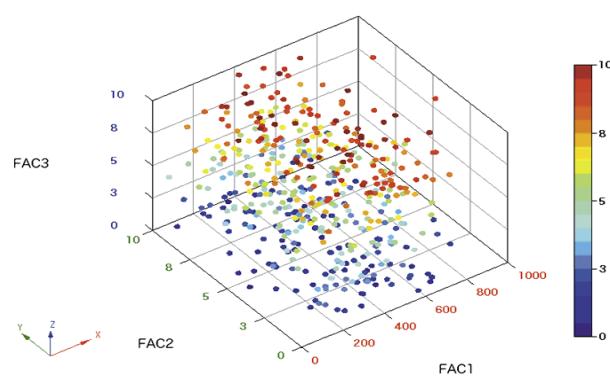
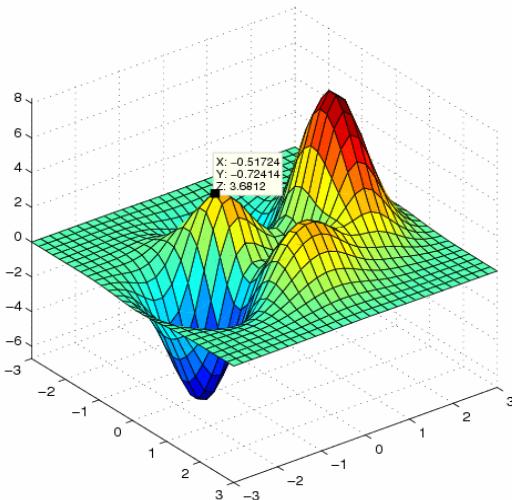


UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## When 3D works



(Actually, these plots incorporate 4 levels of information; 3 different axes and color!)

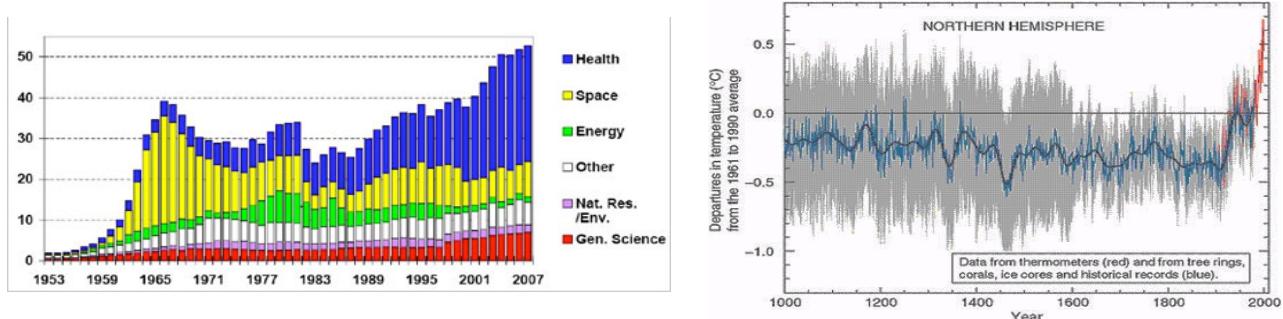
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Exceptions to the “rules”

- Sometimes, breaking the rules of figure formatting is necessary to make the graph more clear and to emphasize the key findings

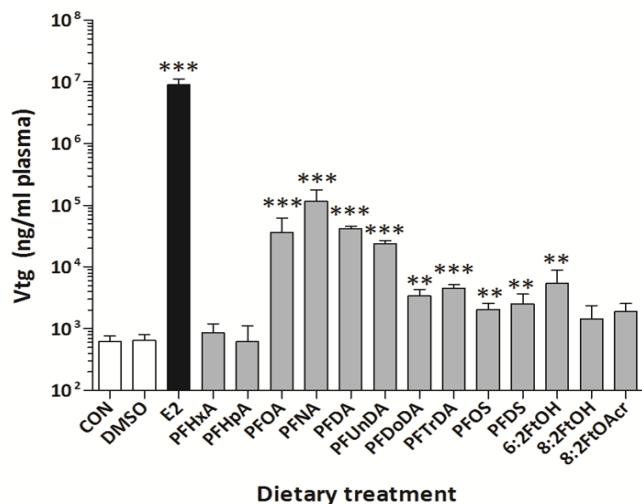


- Using horizontal lines to emphasize data trends, deviation from baseline
- (Note that these graphs have other design problems)



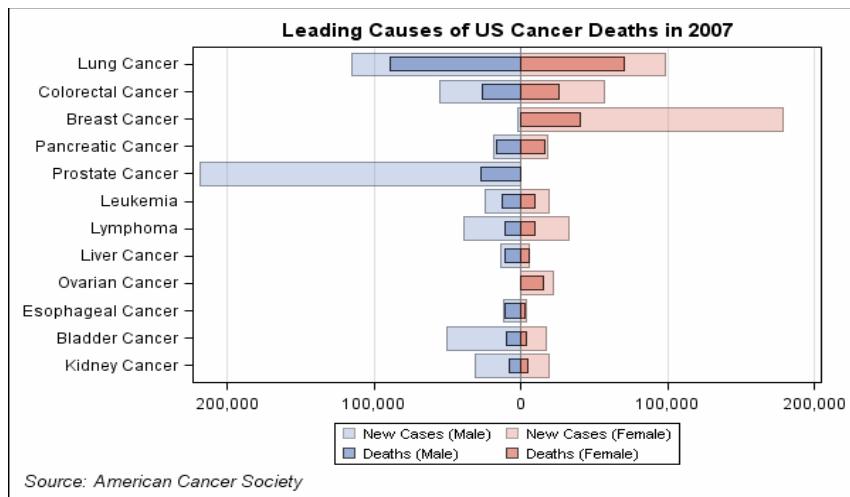
## Exceptions to the “rules”

- Use minor ticks to distinguish logarithmic scales
- A signal to the reader that the scale is not linear



## Exceptions to the “rules”

- Horizontal graphs are okay, if they aid in data interpretation



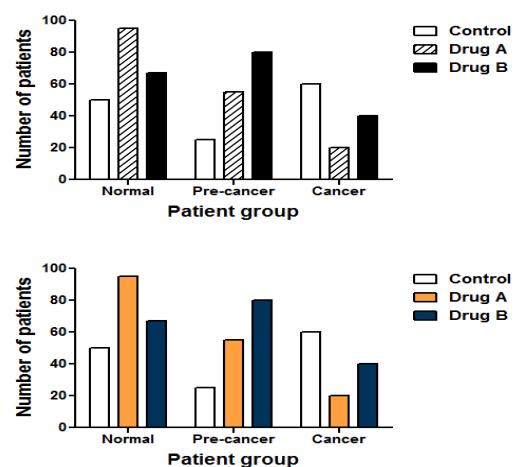
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## (Judicious) Use of color in publications

- Use color if it is necessary to emphasize your point
- Do not use it to merely jazz up the paper, color is typically costly for publication
- Suggest keep your control/reference white, color the other groups.
- Use colors with lots of contrast that go well together; yellow or orange with blue; primary colors; shades within a color.
- Avoid red + green
- All this said, this example graph works just fine as B&W for print publication with the right formatting

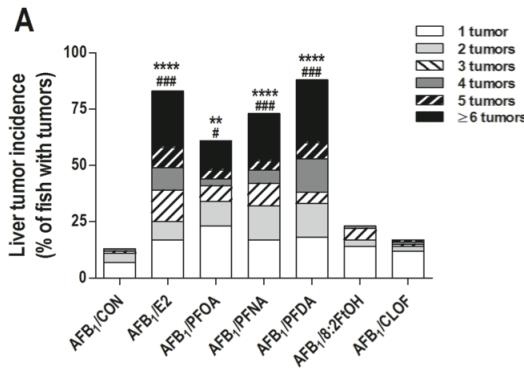


UtahStateUniversity

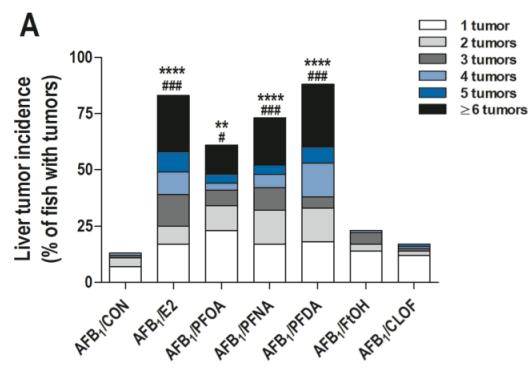
3 Essential steps for better data visualization. January 20, 2016



## Black & White



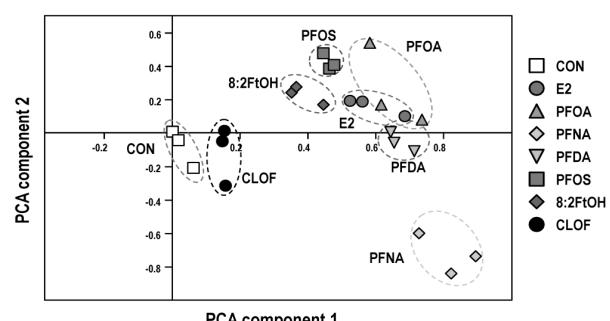
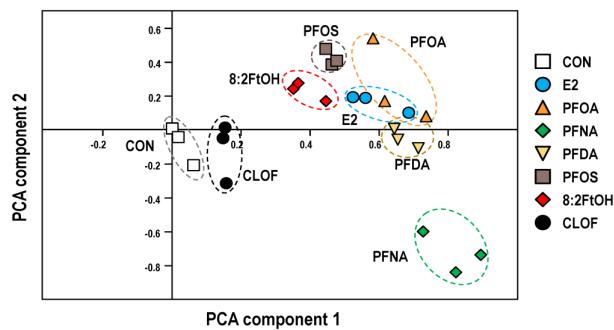
## Color



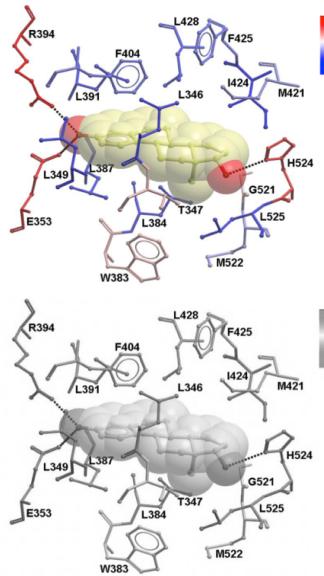
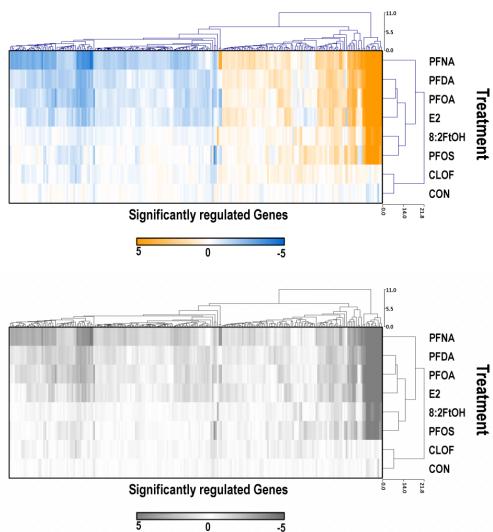
- Color costs money in hard copy publication! So, be prepared to have both B&W and Color versions of your data.
- Definitely use color in your presentations



## Some figures are clearly better in color



## Some figures require color

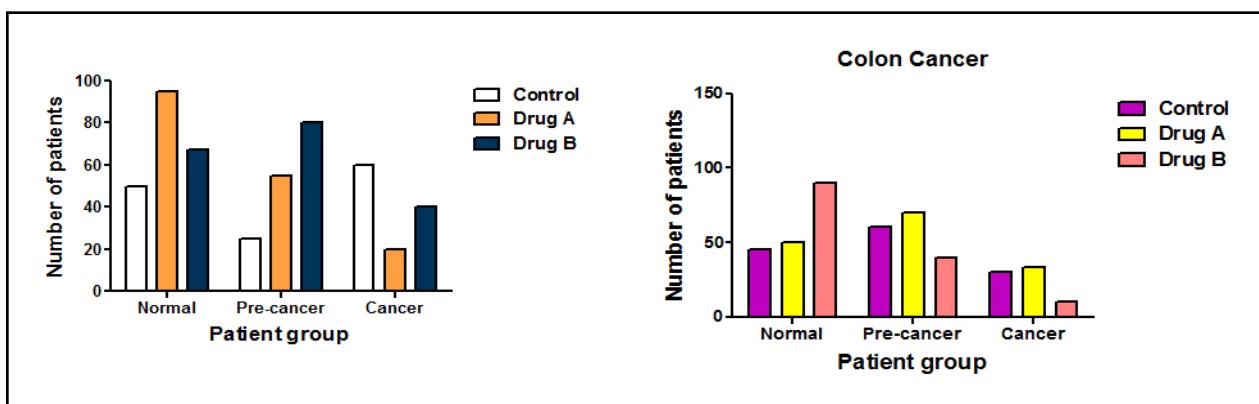


UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Poorly designed multi-panel figure



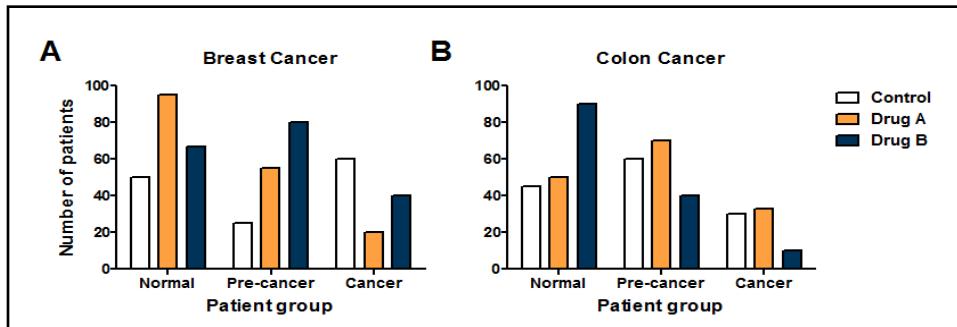
What's wrong with this multi-panel figure?

UtahStateUniversity

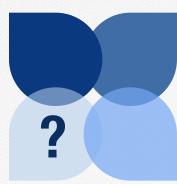
3 Essential steps for better data visualization. January 20, 2016



## Better design



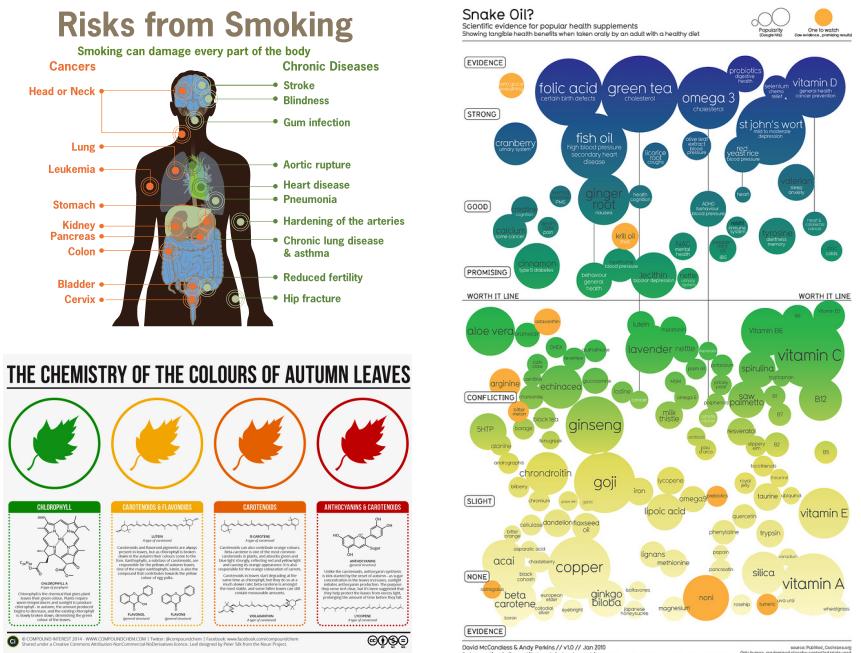
- Only need to show the legend once since it clearly applies to both panels
- A graph title is optional (depends on reviewer); use the letters to identify the panels in the figure legend
- Mirror formatting, scale, colors, etc. between the panels when appropriate
- Don't have to repeat Y-axis title for panel B since it is the same measurement



## Can I use trendy infographics?

# What is an infographic?

- A visual image such as a chart or diagram used to represent information or data
- **Highly shareable!**
- Makes information accessible to non-experts



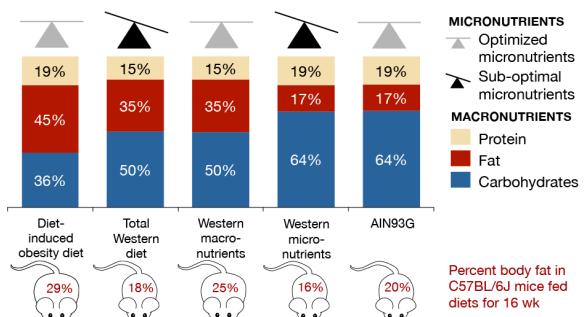
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## When to use infographic style?

- Excellent for oral presentations, poster presentations
  - Especially for **coordinating complex ideas** presented in introduction or conclusion
- Excellent for public presentation
  - Infographics excel at **distilling complex data into a simple visual format**
- Not generally appropriate for professional journals
  - Insufficient detail, often no statistics shown



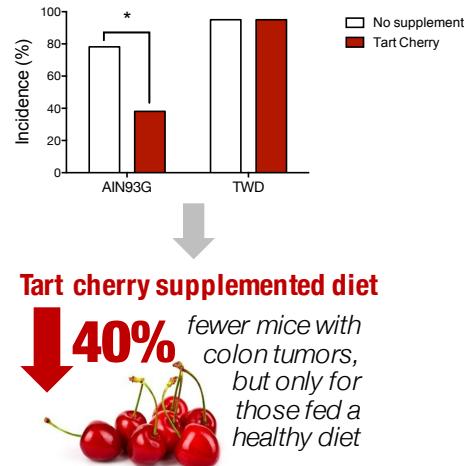
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Use elements of the infographic style

- Just as with regular figure, infographic must be easily discernable within allowed the time frame for viewing
- Create new data charts using different design that focuses less on details
- Avoid excessive detail, as seen in some infographic examples
- You may want to work with professional graphic designer to achieve desired look and style



For more information on infographics, see [http://www.slideshare.net/IQ\\_Agency/5-rulesinfographicssuccess](http://www.slideshare.net/IQ_Agency/5-rulesinfographicssuccess)

UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## What software do you recommend?

UtahStateUniversity

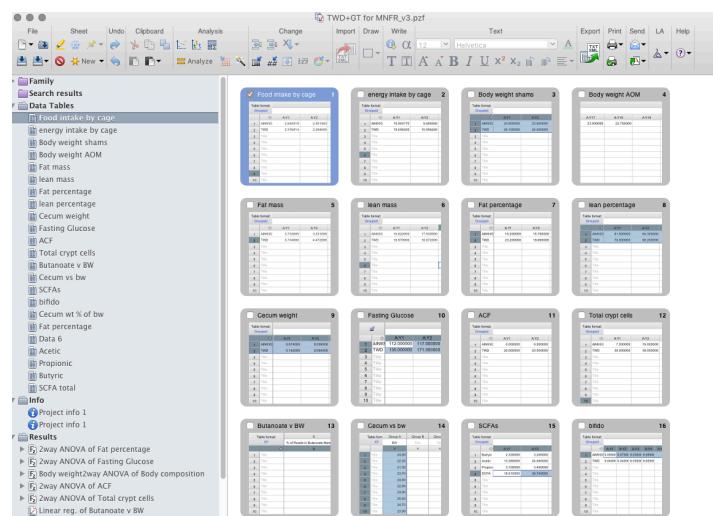
# Software options

- Excel is just awful for making science graphs
  - Especially bad for multi-panel figures
  - Thinks too much for you
- Try out other software that specializes in scientific data presentation
  - SigmaPlot <http://www.sigmaplot.com/>
    - Haven't worked with this one in years
    - Advanced graphics with somewhat steep learning curve
  - **GraphPad Prism** <http://www.graphpad.com/>
    - Easy to use interface
    - Integrates statistics with graphing
    - Drawback – pricey for individuals (\$100/yr); bulk licensing available
    - 30-day trial available



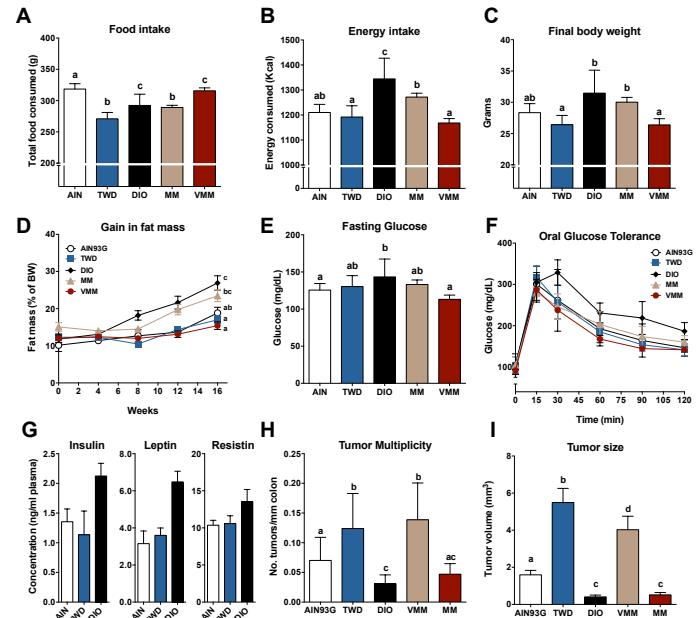
## Primer on Prism

- Organization of sheets by folder structure
  - Data tables
  - Notes
  - Analyses
  - Charts
  - Layouts
- Each data table is linked to other “family” sheets



# Advantages of Prism

- Very customizable
- Create your own “template” and then use that for other charts
- **Excellent** for multi-panel figures



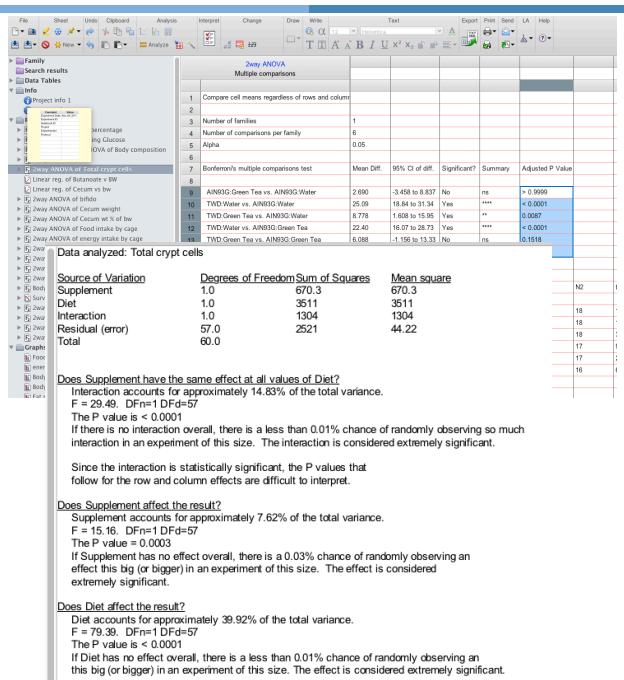
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



# Advantages of Prism

- Integrates basic statistics
  - Students *t*-test
  - One-way and two-way ANOVA
  - Histogram analyses
  - Contingency analysis
- Includes interpretation of results to help students understand analysis results
- Excellent for curve fitting



UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Other tools for data visualization

- Create diagrams or simple infographics using MS PowerPoint
- Create cartographic data using Tableau
  - Free public platform available for non-protected data
- Work with digital image data using Adobe PhotoShop or other photo editing software\*
  - Follow discipline-specific rules about editing image-based data



UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016

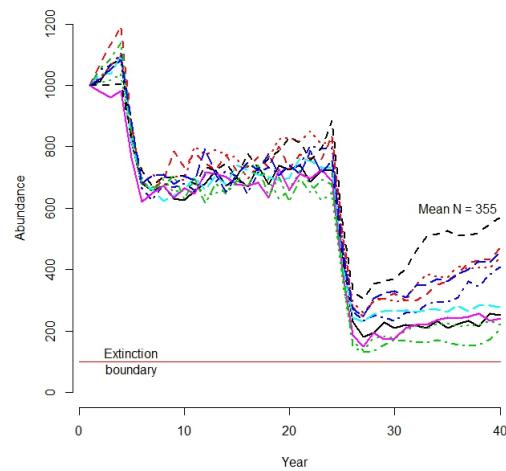


## Comments and revisions of sample figures

UtahStateUniversity

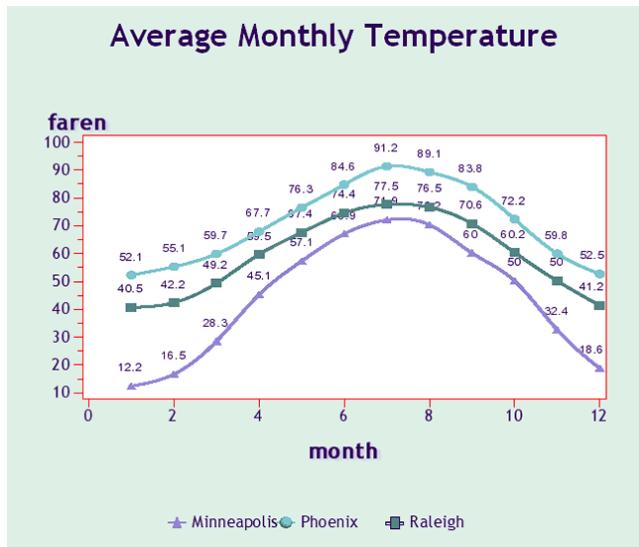
## Comments on sample figures

- Patterns of lines difficult to discern
  - Is it important to be able to distinguish these lines?
  - Some colors are quite similar to each other. I see two different green dotted lines that are quite hard to distinguish
- Need a legend to define each of the traces
- Label “Mean N = 355” doesn’t make much sense in context of its location. Does one of the lines indicate the mean response?
- Labels need to be bold, much larger in size. This figure is likely to be reduced for publication.

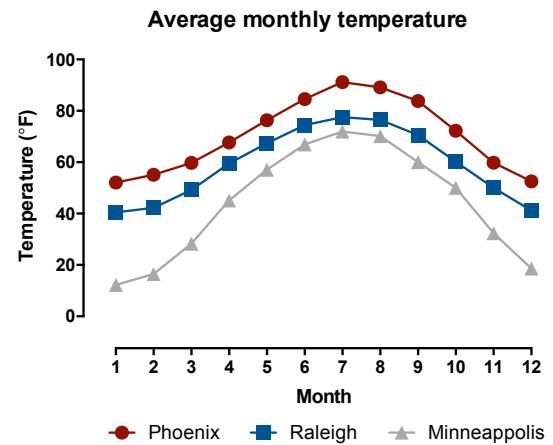
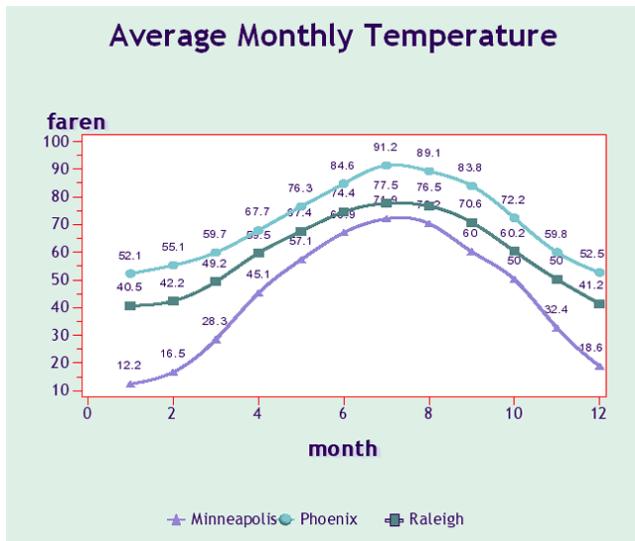


## Comments on sample figures

- Not necessary to label each data point. Crowds the graph.
- Mixture of colors is confusing.
  - Avoid red border around image, especially on the light green background
  - Use more contrast between colors or the different locations. Phoenix and Raleigh are quite similar.
- Place Y axis label along the axis, not at the top
- Not necessary to use minor ticks; makes figure look crowded.



## Revised example



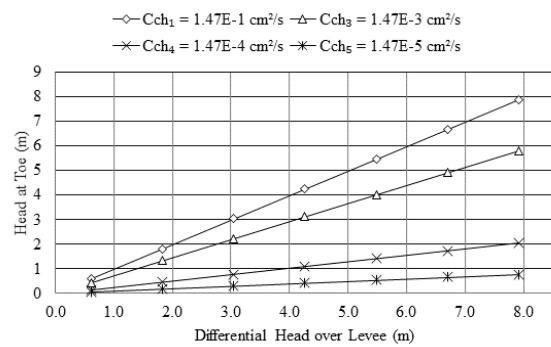
UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Comments on sample figure

- The grid behind the data is unnecessary, crowds the graph
- If you remove the grid, then make sure the X and Y axis have ticks at each of the number labels
- Font size for the X and Y axes labels should be a bit bigger
- Use a sans serif font, such as Arial or Helvetica
- Not entirely clear whether lines represent connections between symbols or regression analysis
- Why a rectangular plot? Scales seem very similar for X and Y axes. Perhaps square plot is needed to show relationships between variables more precisely.
- Symbols need to be easily distinguished. X and the asterisk are too similar, especially right near each other.

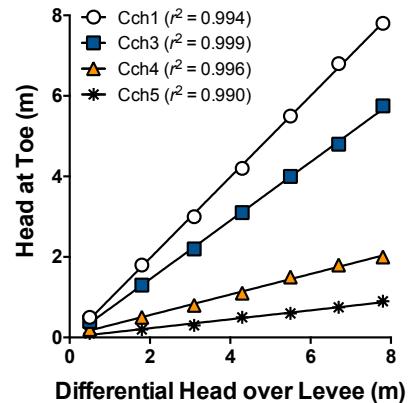
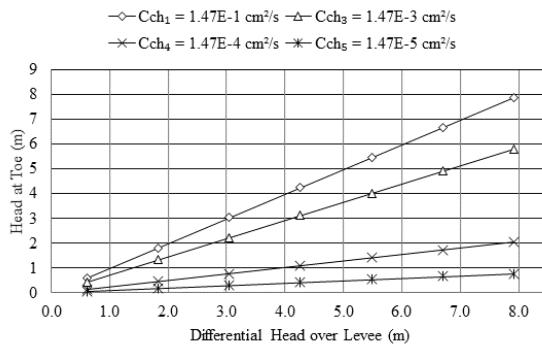


UtahStateUniversity

3 Essential steps for better data visualization. January 20, 2016



## Comments on sample figure



I assumed connecting line was a regression in this example, and showed the  $r^2$  value  
I also shortened the legend text, as the  $1.47E-1 \text{ cm}^2/\text{s}$  can easily go into the figure legend text.  
But, if showing the figure as a presentation, the author may want to keep that info.



## Contact for more information or questions

**Abby D. Benninghoff**

Department of Animal, Dairy and Veterinary Sciences  
abby.benninghoff@usu.edu  
435-797-8649