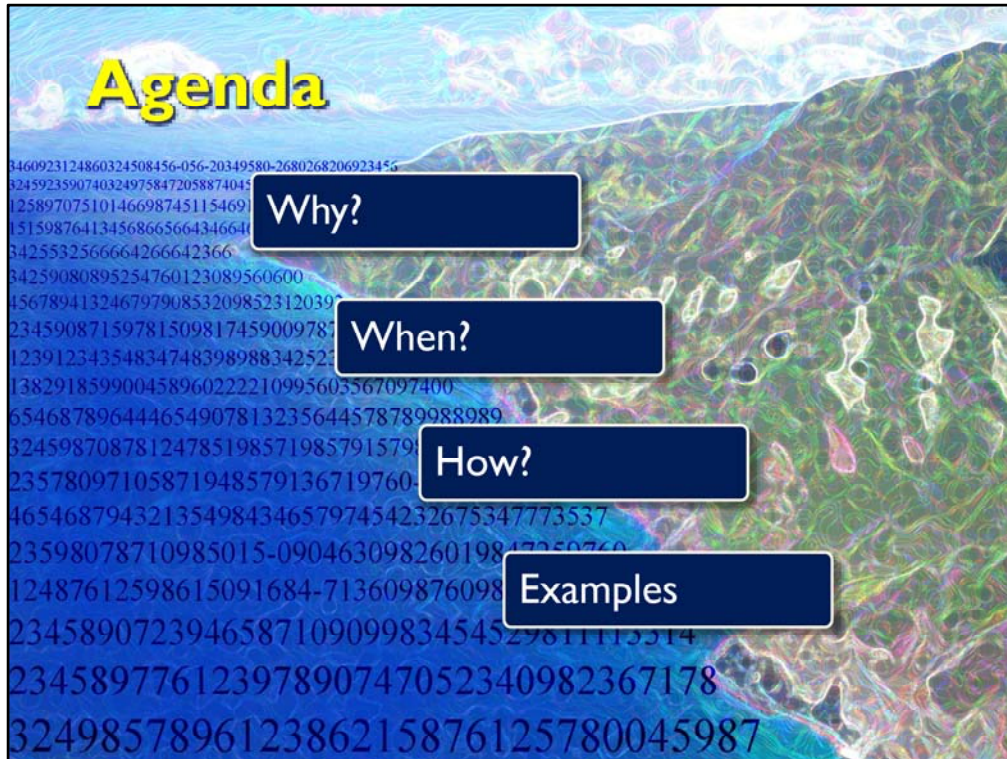
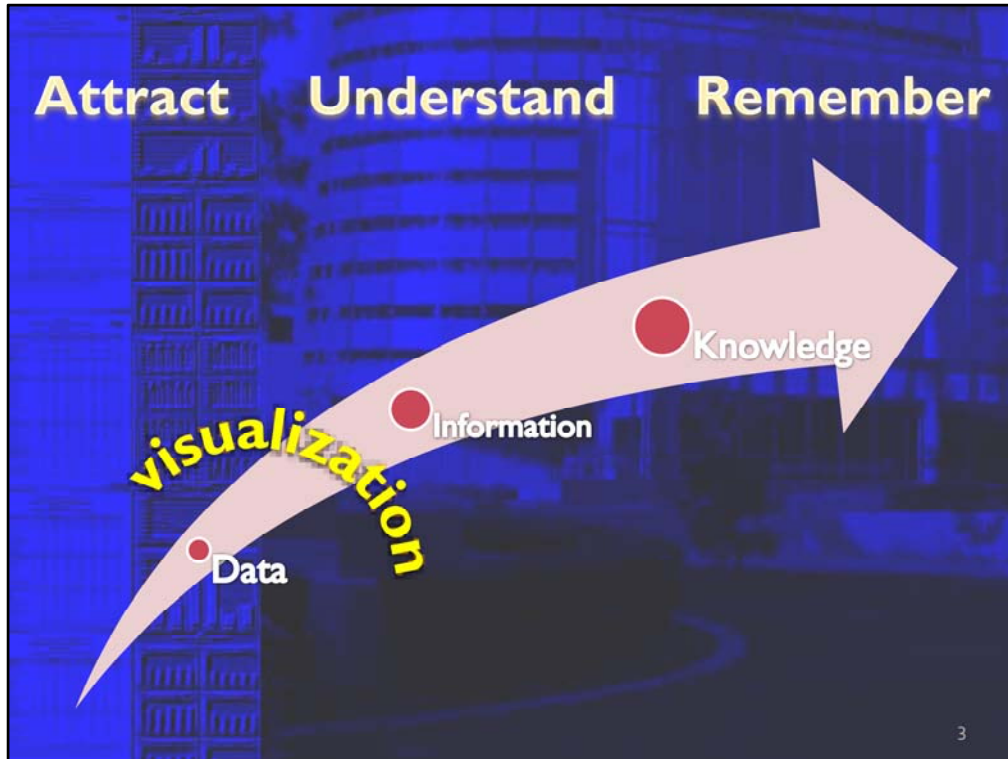




My presentation is about data visualization. How to use visual graphs and charts in order to explore data, discover meaning and report findings. The goal is to show that visual displays can be very effective and efficient, but one has to put some thought into design choices.



I am going to talk about why I think visualization can be an effective tool in analysis and reporting of statistical data. When graphs and charts might do a better job than tables or narrative. How to build graphic designs that help to understand data easily and quicker, rather than mislead or distort data. I will show examples that we used in our office in various projects.



Visualization : to draw our attention to data, to help us to understand it and remember it. By exploring and discovering meaning through visualization we move from data to information, and by remembering information we can use it as our knowledge.

# Visual Databook

*"...It helps to paint a picture of Fresno State.*

*...It brings data to life.*

*Easy to interpret – easy to use.*

*...This format is awesome!*

*...This was really cool.*

*I enjoyed it.*

*...Wow!!"*

IRAP Visual Data Book

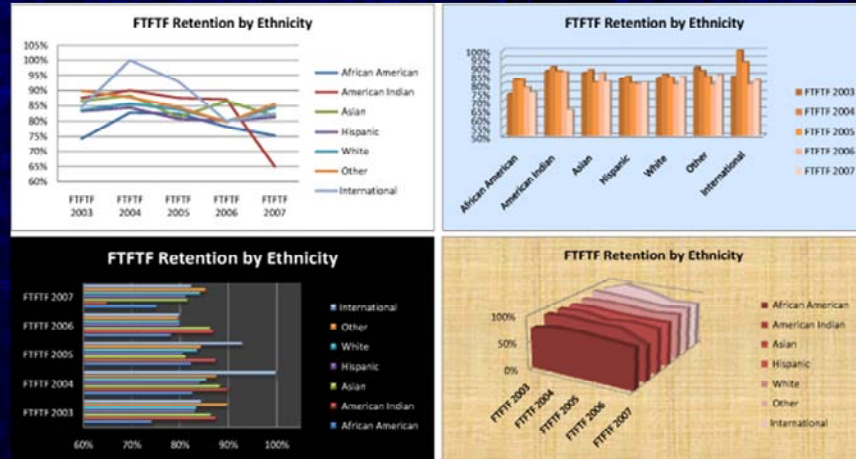


[http://www.csufresno.edu/irap/documents/2005\\_databook/Visual\\_Databook.pdf](http://www.csufresno.edu/irap/documents/2005_databook/Visual_Databook.pdf)

Two years ago we were working on our annual data book. While producing tables I was trying to identify patterns and trends and found that it was not an easy task just by looking at rows and columns of numbers. I put together a few charts and found them not only visually appealing but also helping me to think about the data and discover patterns and trends. I thought, if we add charts and graphs to the tables in our data book people might find it interesting and useful too. Thus we created the visual data book and published it on our website. Basically it's a pdf copy of our printed version of datebook, but each table had a link to a corresponding chart.

The feedback from faculty and staff was very positive. They found the Visual Databook very useful and easy to use. But it did take a lot of time and work to make one. This is when I really started thinking about how to produce charts and graphs that are clear and understandable for everyone. The excel has lot's of chart options, but which one works better for certain data? I wondered whether there are any rules and guidelines in graph designing.

## Which one presents data the best?



A picture can be worth a thousand words – but only if you can decipher it. (Kosslyn, 1994)

5

For example, here are four different charts that present the same data on first-time full-time freshmen retention across five years by ethnic groups. Which chart do you think presents the data the best?...

As one of the statisticians and authors on graph designing, Stephen Kosslyn, said “A picture can be worth a thousand words – but only if you can decipher it” ...I will show my version of the graph later on in the presentation.

## Principles of Graphical Excellence

- Well-designed presentation of interesting data
- Complex ideas communicated with clarity, precision, and efficiency
- Gives the viewer the greatest number of ideas with the least ink in the smallest space
- Nearly always multivariate
- Requires telling the truth about the data

*(Tufte, 2001)*

6

Indeed, after reading various books on data visualization including a classic by Edward Tufte, using charts and graphs extensively in various projects for analysis and presentation of data and getting feedback from viewers, I found that there are certain principles and rules that one should remember when designing graphics.

Tufte came up with five principles of graphical excellence:

- Well-designed presentation of interesting data – a matter of substance, of statistics, and of design
- Complex ideas communicated with clarity, precision, and efficiency
- Gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Nearly always multivariate.
- Requires telling the truth about the data

The principles don't need to be applied rigidly. But we do want to aim for "Clear portrayal of complexity, not the complication of the simple".

# To Graph or Not to Graph?



## Tables:

- |   |                       |
|---|-----------------------|
| 1 | Exact Values          |
| 2 | Small Data Sets       |
| 3 | Localized Comparisons |

7

Not everything needs to be graphed. A lot of times, tables or text are more efficient ways of communicating data. If you want to know exact values, if you have small data sets or if you need to make localized comparisons consider using tables.

However, if you want to show relations, trends or patterns, a graph might be a very efficient and effective medium of display.

# How?

Know your Question

Know your Data

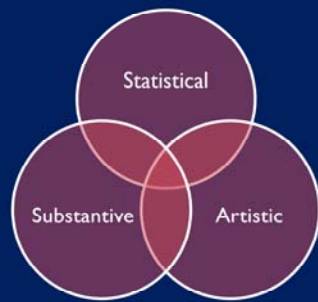
Know your Audience

8

Before you can determine how to effectively present the message you, first, need to know what the message is. Then consider your data: line graphs work better with ratio scales and time-series, bar-graphs with categorical or nominal data, scatter-plots in correlations. Finally, know your audience. Do they have an expertise to read your chart, how much detail they need, or what they want to focus on. Present the data needed to answer specific questions (no more or less) and use concepts and displays that are familiar to the audience.

# Graphical Elegance

Simplicity of design and  
complexity of data



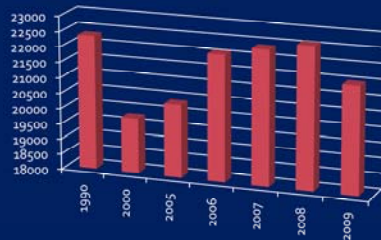
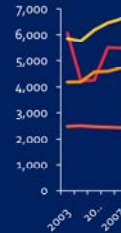
(Tufte, 2001)

9

A well-designed graph shows complex data in a simple to understand form. In order to build one, we need to have a combination of three skills: substantive, statistical and artistic, If we miss one, we can get a misleading or confusing message.

# GOOD DESIGN

- Non-zero baseline in bar graphs
- Unequal intervals
- 3-D
- All capitals, fancy fonts
- Cluttered displays
- Too narrow or too wide



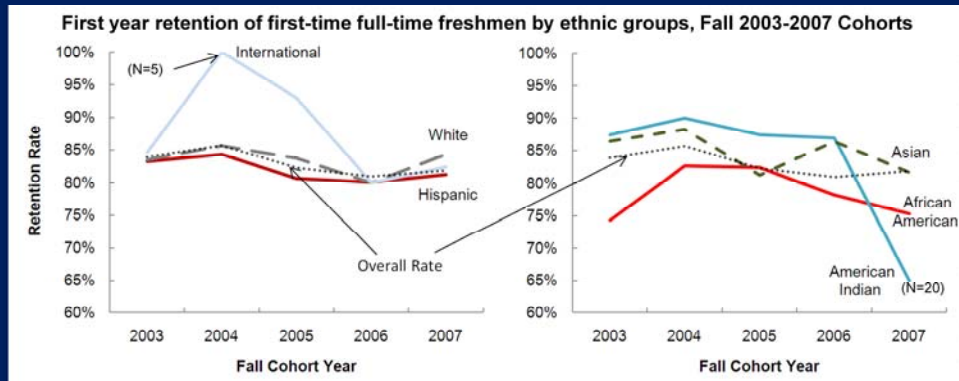
10

Some common mistakes in designing graphs include using:

- non-zero baseline in bar graphs. The length of a bar encode the value
- unequal intervals in time-series
- designing elements that distract from actual data, such as colored background, fancy fonts, 3-D format, objects
- cluttered displays, for example when you have multiple lines placed on one chart
- color choices that don't work with black-and white prints or can be hard to read for color-blind people. For example, red and green
- Frame proportions (Y-X aspect ratio) too narrow or too wide, when it distorts the display

It's always good to put in some thoughts on what design will work better instead of using default options.

## Line Graphs - Trends



- No more than 4 lines per graph
- Labels next to lines
- Place time on the horizontal axis
- Title reflects the question
- Consider line style and color

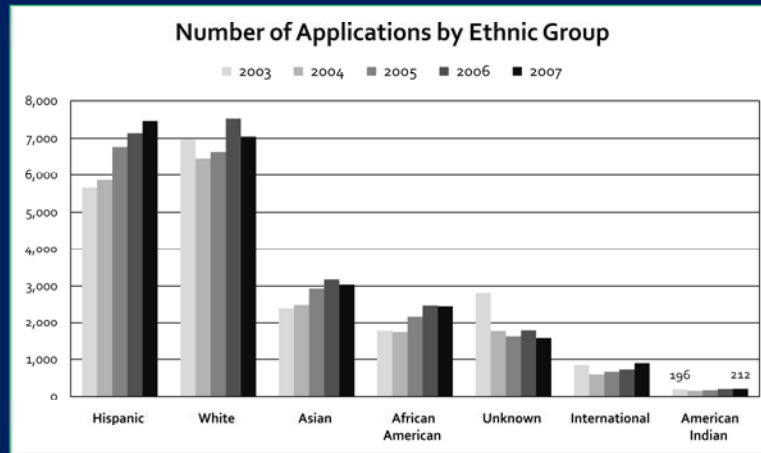
11

So here is my solution. Since we have time series data, I am using line graphs, which are particularly good in showing changes over time. Because we have quite a few ethnic groups, I decided to split them between two panels, so we can clearly see trends for each of the groups. I put ethnic groups that had their retention rates going up in the last fall semester in the left panel (White, Hispanic and International). Ethnic groups that had their retention rate decreased displayed on the right side. The dashed line showing the overall rate is displayed in both panels so we can easily compare each group against average. Placing labels next to lines help to identify lines quicker rather if we would use the color legend. I tried to use colors and line styles that would be easy to distinguish if printed in black and white. Non-zero baseline is okay in this case, since we are not using bar charts. By moving the scale base to 60% we can see changes in retention better.

The graph shows that Whites have slightly higher rate than Hispanics, but both are very close to the overall rate. Which is not surprising since both groups have the largest representation in the student body. Jumps in rates of international students and American Indians explained by low group sizes.

I think that this graph is superior in communicating the message than the previous versions built using the default options.

## Vertical Bars - Categorical Data



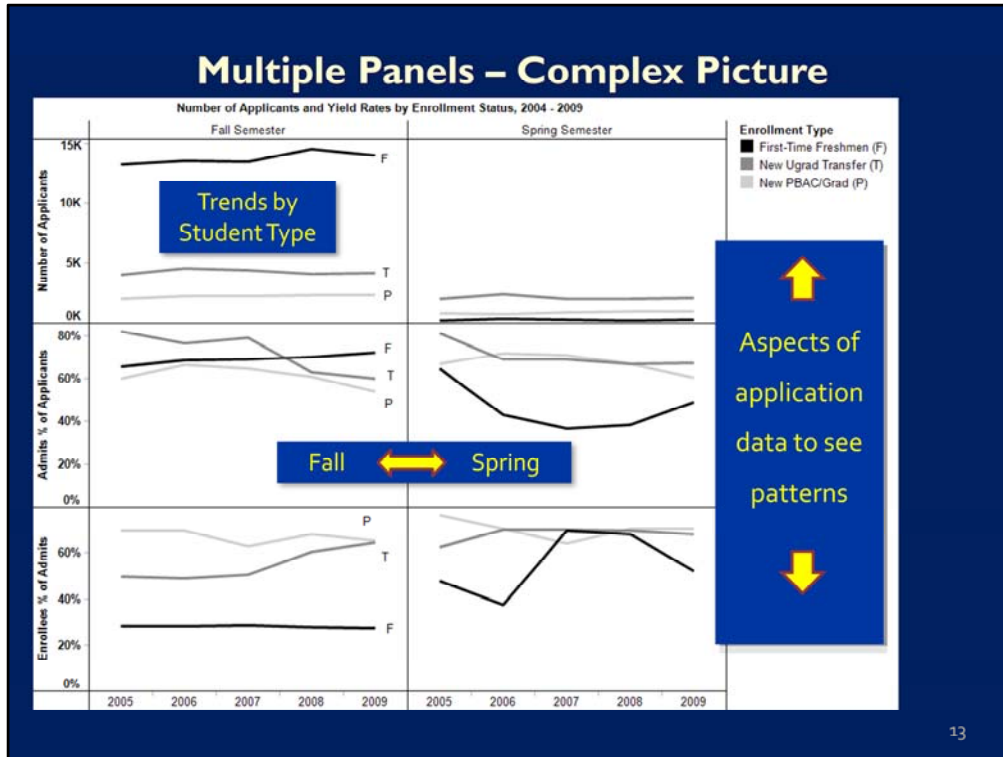
- Use bars to emphasize individual values
- Sort to highlight high-low values
- Shades of black instead of colors

12

In the next example I used vertical bars to compare number of applications across five years among ethnic groups. This example is similar to the previous one, but I decided to go with vertical graphs instead of lines, since I wanted to emphasize individual values, instead of showing trends.

When you use bars, your baseline must be zero. This is because the length of the bars encode their values. By sorting values, I can easily see that the largest number of applications comes from Hispanics and Whites. Since American Indians had a very small number of applicants compared to other groups and the change across years is difficult to see on the current scale, I placed the first and last values above the bars. Now we can see that the number of applicants for American Indians increased slightly from 196 – 212 over the five years.

If we use default option, we usually get bright color bars. In this case, I replaced them with variations of gray, with the darkest one showing the latest year. This way, this graph will work fine if printed or photocopied in black and white

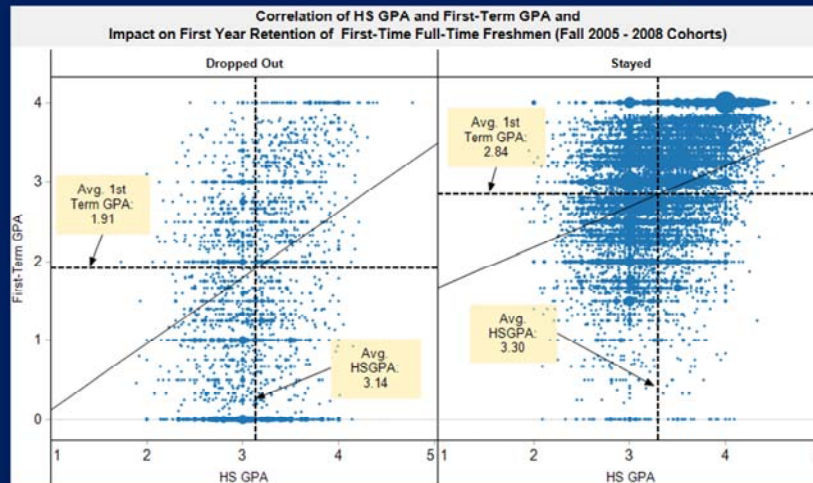


If you want to show a complex picture or in some way tell a story unveiled by numbers, consider using multiple panels,

Here I am using multiple panels to show various aspects of application data and compare trends among three student types (FTF, transfer, and New PBAC). The horizontal axis displays five fall semesters from 2005 to 2009 on the left and spring semesters on the right. The vertical axis include three sections: the top section showing number of applicants, the middle one— percentage of admits of applicants, and the bottom section - percentage of enrollees of admits. This matrix display allows us to compare trends among student types and between fall and spring semesters. It's clear that the largest number of applications come from freshmen in fall semester, however in spring we get more transfers and PBACs. The trends in number of applications are relatively stable across semesters. Percentage of FTF admits is going up in spring, while percentage of transfer and Pbac admits is going down. Pbac have the highest enrollment yield, while FTF the lowest (around 30%)

I am using first letters of the student levels in addition to shades of gray so we can easily identify lines on the graph. Also notice that the horizontal label showing the years is only shown on the very bottom of the graph to avoid duplication and clutter.

## Scatter Plots - Relationships



- Two related variables
- Trend and average lines
- Shape and density of the cloud
- Don't include grid

14

If you have two related measures (interval or ratio scale), consider using scatter plots to show the relationship

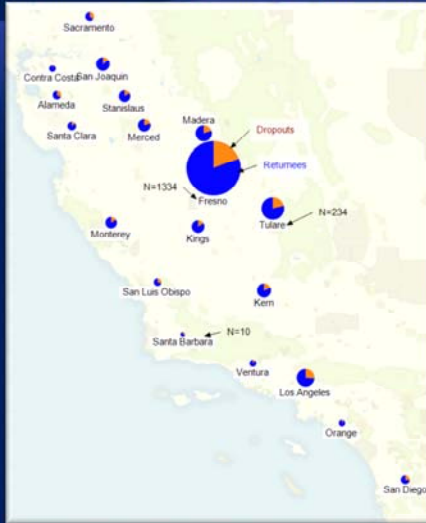
This graph using scatter plots shows the relationship between first-term gpa and high school gpa and impact of those on the first-year retention. The horizontal axis shows HS GPA, the vertical axis shows First-Term GPA. Left side of the graphs displays values for drop outs, right side – for those who stayed after the first year of enrollment,

At first sight, this graph might look overwhelming. But what you want to look at is the shape and density of the cloud. It's clear that for those who stayed, the cloud is more dense in the upper right corner meaning they had higher first-term gpa and high school gpa. For those who dropped out, the cloud is dispersed and we see a lot of values at the bottom left corner. The average lines prove that students who stayed had higher first-term and high school gpa. Slope of the trend line shows that there is a positive correlation between both gpas.

Don't include grid (the point is to convey an overall trend, not individual values) For individual values use table.

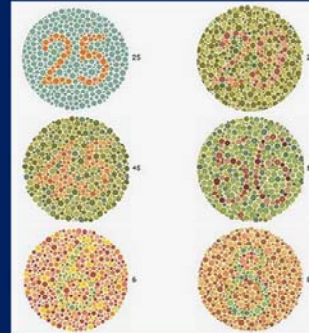
# Pie Graphs - Proportions

First Year Retention of First-Time Full-Time Freshmen by County of Residence, Fall 2008 Cohort



**Color Alert!**

8% of males are color-blind



15

To show geographic location or distance maps are invaluable.

This map with pie charts displays retention statistics by county of residence. Pie chart typically is not the best choice for graphic displays since it's hard to make comparison of angles. In this graph, I think it works. It's relatively easy to compare the proportions of the orange wedge. The size of the whole pie varies based on the count of students. I displayed the max and min counts to give an idea of sizes.

Consider colors carefully. Avoid red and green to define a boundary, because about 8% of the male population have trouble distinguishing these colors. Also when printed in black and white, you might not be able to distinguish the elements. Use colors that are well separated in the spectrum. Warm colors (red, orange) work better for a foreground.

# Reporting with Visualization

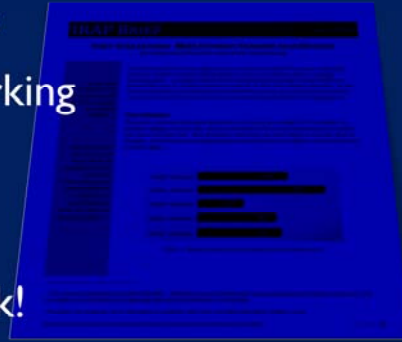
**Executives:** key elements

**Policy Makers:** benchmarking

**Faculty:** tailored info

**Experts:** details

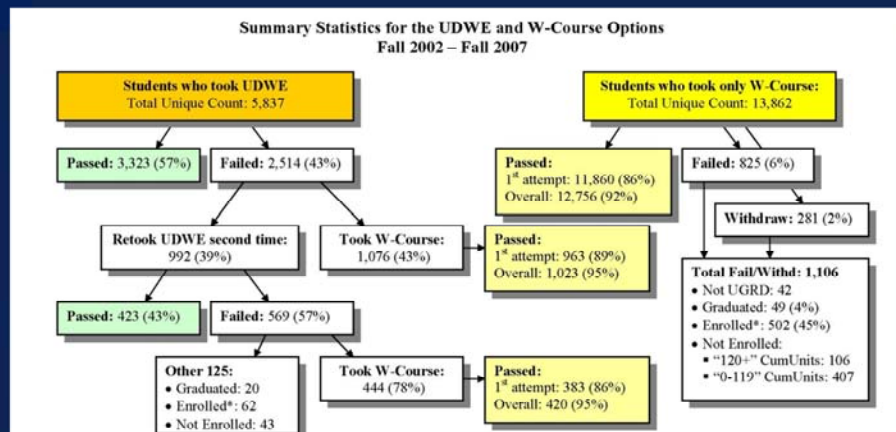
**Combination:** Good Luck!



16

When using graphs for reporting findings in specific projects it's important to know your audience, You need to know what you want your readers to focus on and how they will use the information.

## How students satisfy writing requirement

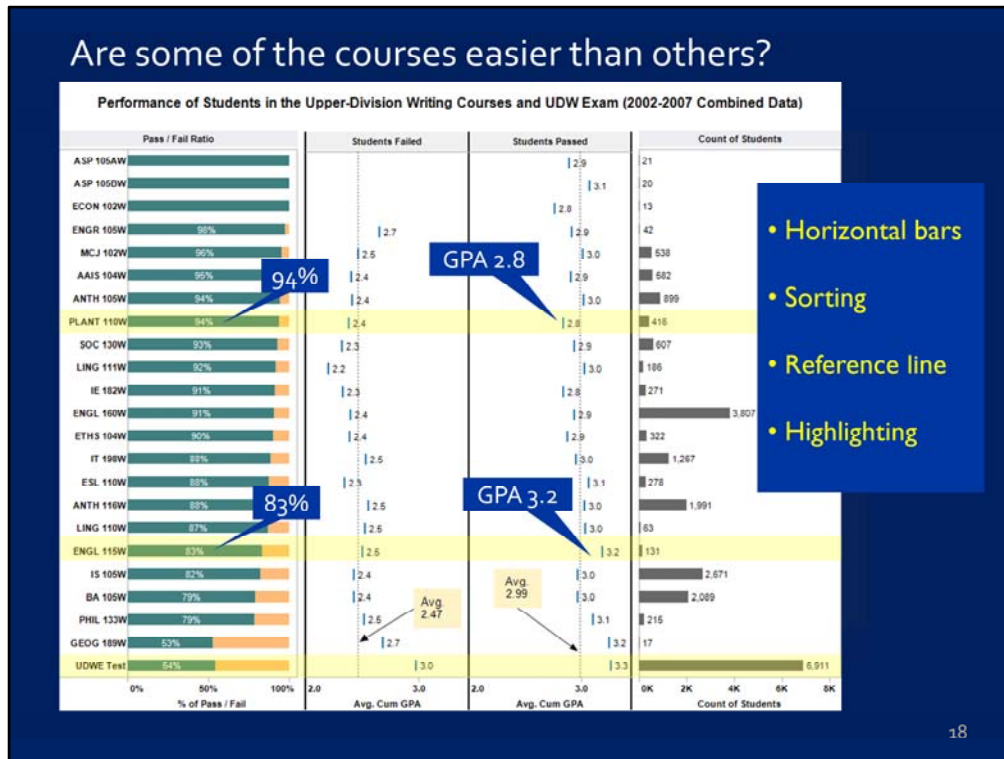


- Flow charts for qualitative data
- Colors and shading helps

17

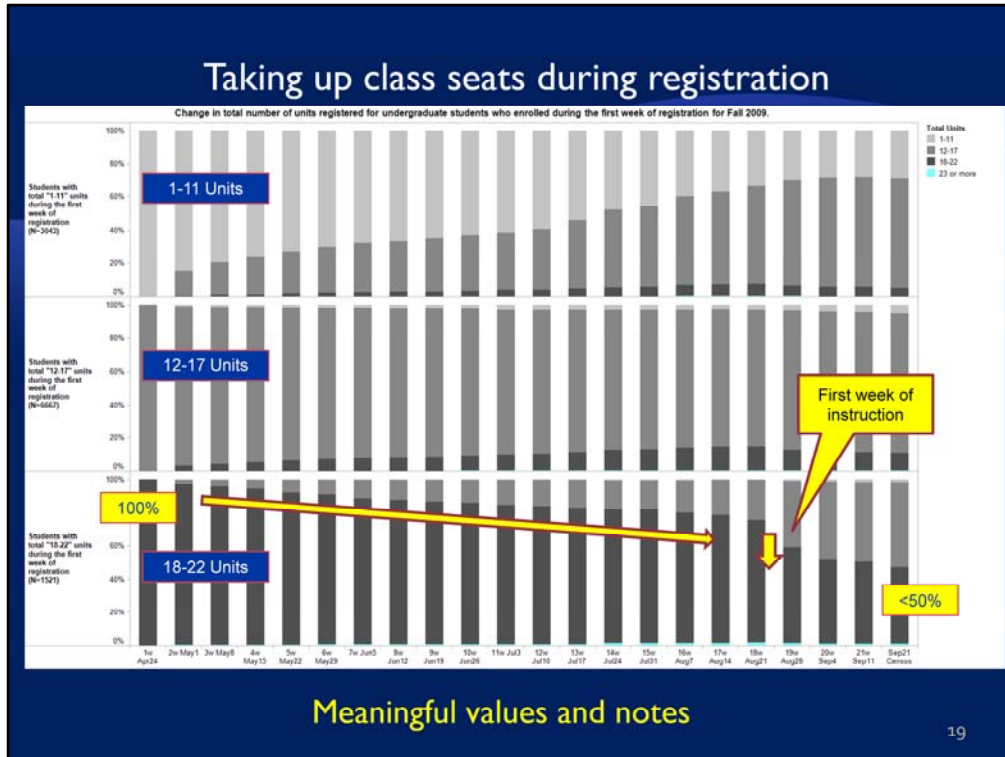
Charts and diagrams to display qualitative data.

## Are some of the courses easier than others?



In the same project, we used a graph to show comparison of W-courses against each other and against the UDW exam. Because of a large number of courses I used horizontal bars. Sorting by Pass/Fail ratio made it easy to identify courses with the highest and lowest passing rates and to compare among courses (see left part of the graph). The right side of the graph displays counts of students in each of the course. By using bars I can clearly see what courses are more popular than others (such as ENGL 160, IS 105, and BA 105). At the same time, enrollment in some of the courses was very low (lower than 30 students), which makes comparison of these courses to other not practical. However, I left them on the graph, as well as I displayed all individual values, because I knew that the requester was interested in seeing the whole picture and capable of reading the data.

In the middle part of the graph, I showed average GPA of the students who failed or passed courses. I used reference line (average GPA) to compare GPA values against each other. Highlighting helps to communicate the findings. First, the writing exam is a more difficult option in satisfying the requirement than w-courses – average GPA of those who failed the test is at the same level as average GPA of students passing W-courses. Second, some courses (PLANT 110W) seemed relatively easier than others (ENGL 115, PHIL 113) – high passing rates, but low students' GPA.



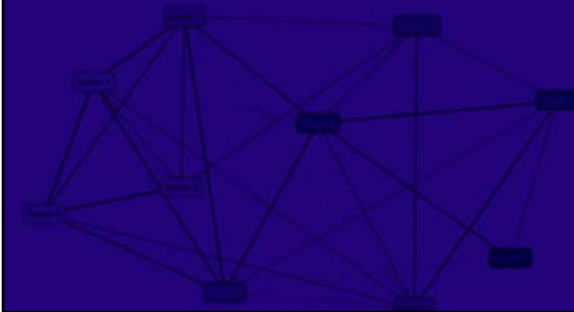
We had a request from Registrar office to analyze students registration activity. They had a feeling that some students register for the maximum of units during the first week of registration, hold up the seats until the instruction starts and then drop the classes they don't want. Thus, other students are not able to register for the classes, since all seats are taken. So, is it the case?

The graph seems to prove it. The vertical bars represent proportions of units registered. The horizontal axis shows weeks of registration. By looking at proportion changes we can see how students drop or add more classes during the whole period of registration. The top section of the chart shows students who registered for 1 to 11 units during the first week of registration. The bottom section is of a particular interest in our case: students who registered for 18-22 units. We clearly see that at census date less than half of those students keep the same course load (18-22 units). The major drop occurred during the first week of instruction.

# Visualization for Analysis

Insight and understanding are best achieved not through statistical tests but by creating pictures with data.

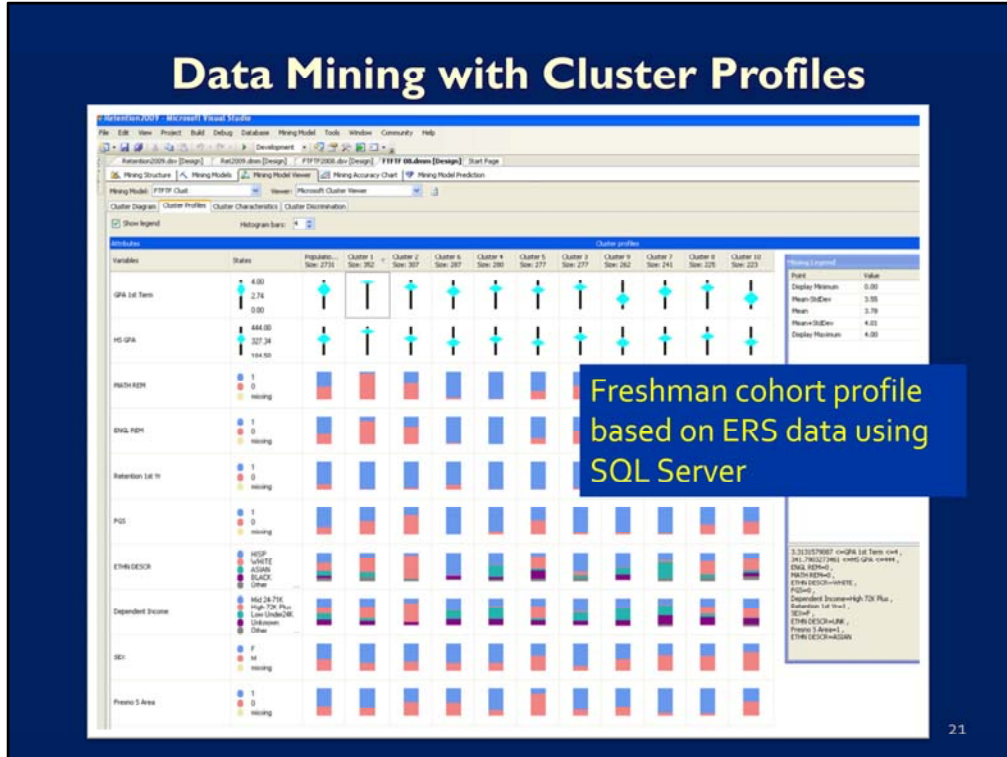
John W. Tukey



20

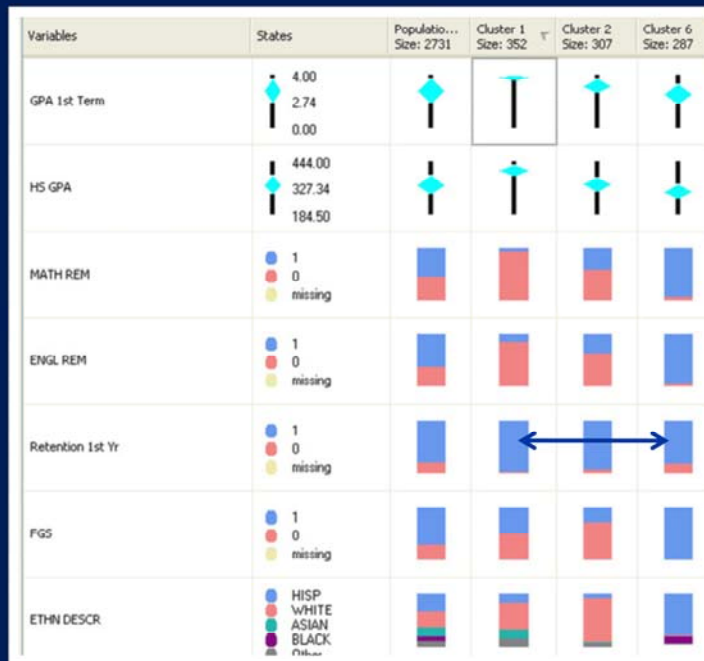
I use visualization not just for reporting but also for exploring and analysis of the data. A lot of time, visual displays make it easier to identify patterns in large data sets.

# Data Mining with Cluster Profiles



This data mining output is a cluster profile of the first-time full-time freshmen cohort. I input ERS data, students GPA and first-year retention outcome into the SQL Server data mining clustering algorithm.

# Data Mining with Cluster Profiles



The process split the whole cohort into 10 different clusters, which each has a set of students with similar characteristics. For example, cluster #1 has 352 students with first term gpa from 3.3 to 4.0, HS GPA from 3.4 to 4.4, no English or math remediation need, mainly whites, with almost 100 percent retention rate. Vertical stack bars allow easy and quick comparison between the clusters.

# Summary

Visualization efficient and effective  
But can mislead and confuse  
Know your question and audience  
When designing graphs

“above all else show the data”

(Tufte, 2001)

## References

- **Stephen Kosslyn** (1994). *Elements of graph design*.
- **Stephen Few** [www.perceptualedge.com](http://www.perceptualedge.com)
- **Liz Sanders & Joseph Filkins** (2009). *Effective Reporting*.
- **Edward Tufte** (2001). *The visual display of quantitative information*

# Questions, Suggestions?

Dmitri Rogulkin  
drogulkin@csufresno.edu



We all learned first from pictures...

<http://www.csufresno.edu/irap/staff/documents/cairpresentation.pdf>

25

Here is my little son. They said that looking at those various lines stimulates brain development at that age. And this is how he discovers this world - through pictures; he doesn't know the numbers yet!