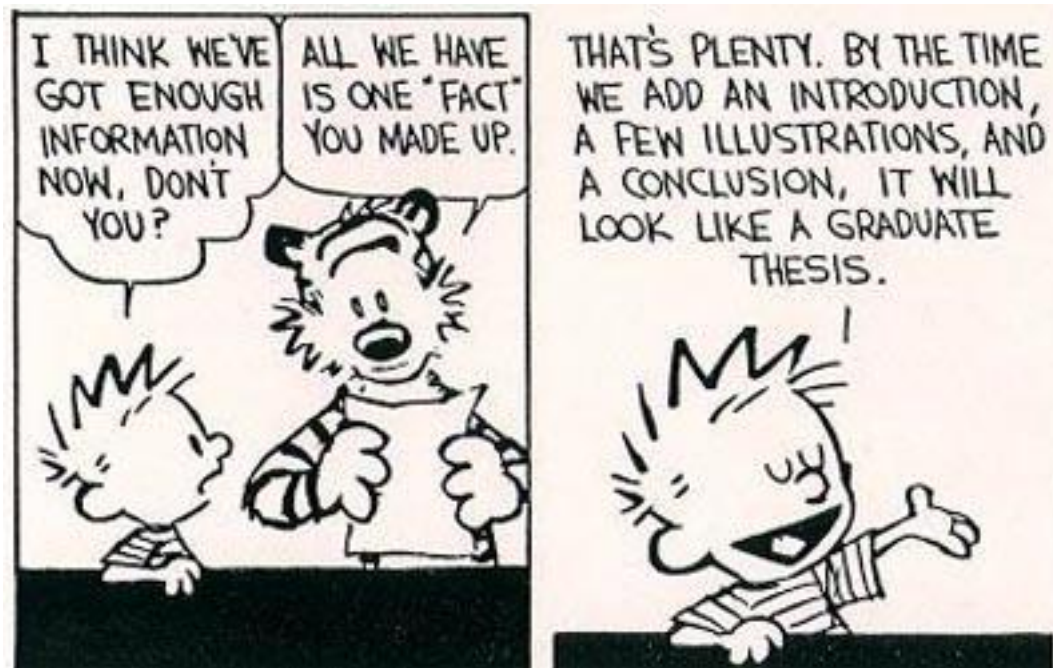


How to *Lie, Cheat, Manipulate, and Mislead* using Statistics and Graphical Displays.



(Also, how to do it right, and MOST IMPORTANTLY,
how to tell the difference!)

No, of course we'd NEVER do that...



Why are we being FORCED to listen to this (against our will) in CSE₃?

- How does Statistics and Graphical Displays (truthful or not) matter in a computer science class???
- Data and information are so prevalent in our lives today, that it is known as the “Information Age”
- Being literate today means not just being able to read, but being able to understand the massive amount of information thrown at us every day – much of it on the computer.



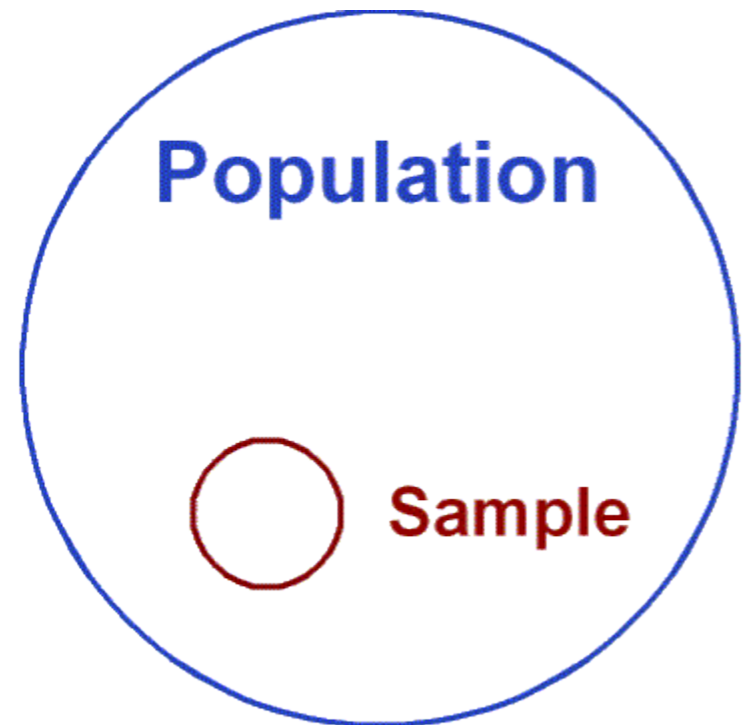
Let's start with Statistics...

- **Statistics** is the science of making effective use of numerical data.
- It deals with all aspects of this, including the
 - collection
 - analysis and
 - interpretation of data



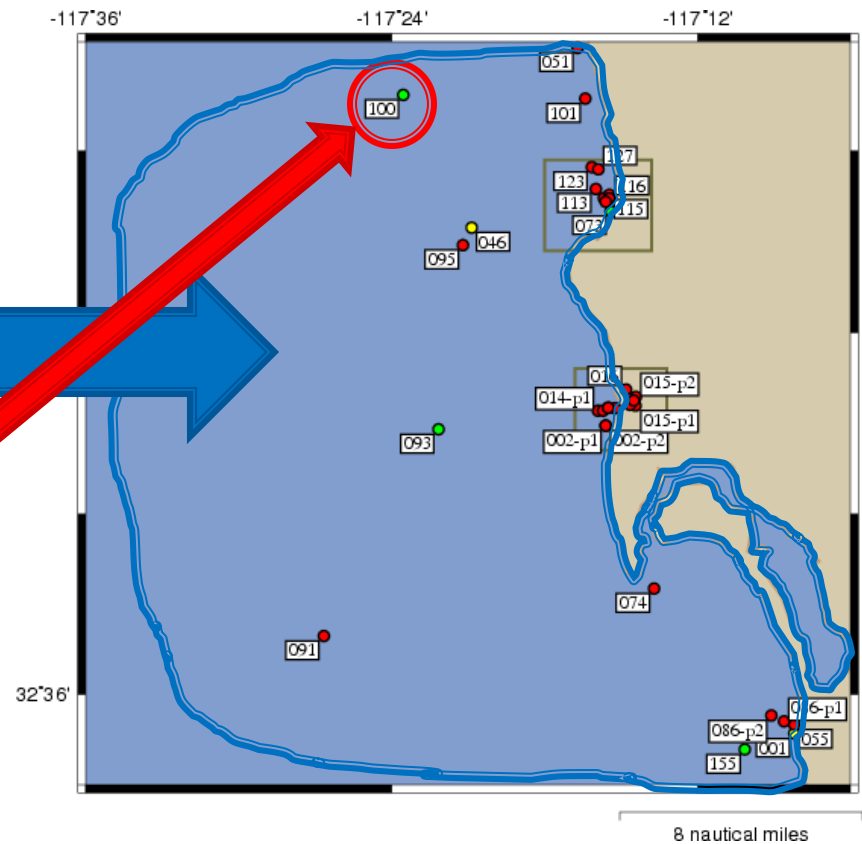
Collection of Data

- In order to analyze and interpret data, we must first collect it.
- The data that is collected is known as a **sample**.
- The sample is collected from a **population**.



Ocean Temperature Sampling

- We wanted to analyze San Diego ocean temperatures in CSE₃.
- Our **population** was the ocean off the coast of San Diego.
- Our **sample**, was the temperatures recorded by Buoy100 over the last 9 years.



How to Lie, Cheat, Manipulate, and Mislead through poor Sampling:

- If we were to claim that our results were representative of:
 - California coastal waters
 - Southern California coastal waters
 - San Diego coastal waters
 - Or even La Jolla coastal waters

That would be called

Biased Sampling

And we could use it to lie, cheat, manipulate, or mislead the general public.



Instead of taking all the bother and time to sample for data at all...



© Scott Adams, Inc./Dist. by UFS, Inc.

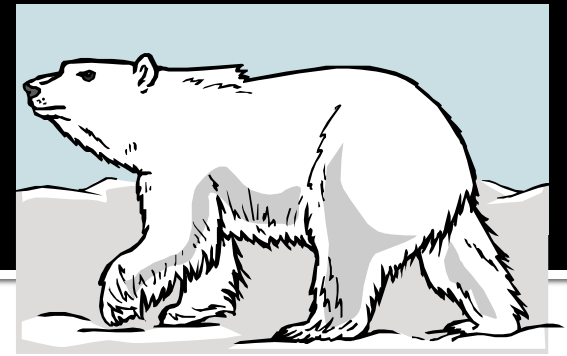
Biased Sampling

- There are many different types of sampling bias. Some examples include:
 - Area Bias
 - Self-Selection Bias
 - Leading Question Bias
 - Social Desirability Bias

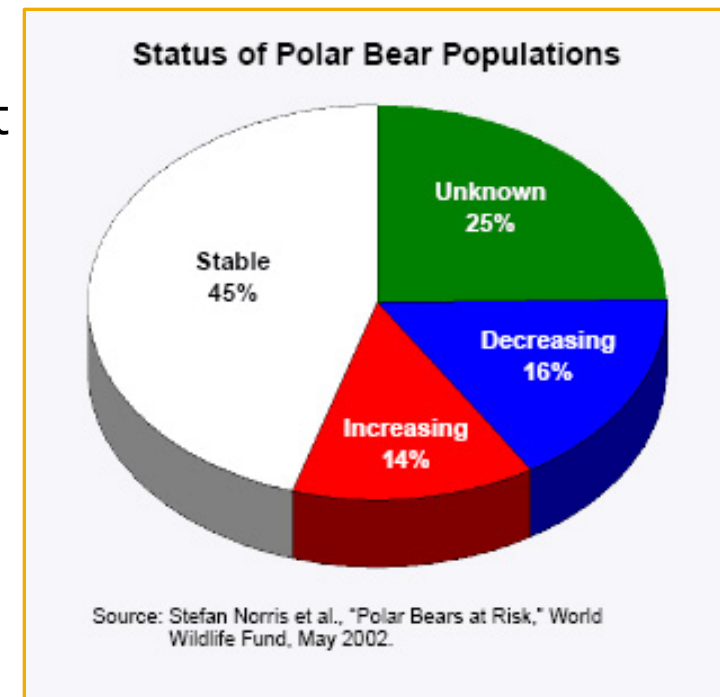
Area Bias

- If we were to claim that our findings were applicable to the entire California coast, or even just the San Diego coast, we would be guilty of perpetrating an **area bias**.
- The area of your sample needs to be representative of the study population.
- When reading news stories or scientific articles, make sure you verify that there is no **area bias** in the study.

Area Bias Example: Polar Bear Population



- The World Wildlife Fund (WWF) has written on the threats posed to polar bears from global warming.
- However, also according to them, about 20 distinct polar bear populations exist, accounting for approximately 22,000 polar bears worldwide.
 - Only 2 of the groups are decreasing.
 - 10 populations are stable.
 - 2 populations are increasing.
 - The status of the remaining 6 populations is unknown.
- If you only looked at the 2 groups that are decreasing, it would be easy to say that “Polar Bear Population is Decreasing”. You need to look at the whole picture to get the whole story.



Another example of Area Bias...



www.dilbert.com scottadams@aol.com



© 2004 Scott Adams, Inc./Dist. by UFS, Inc.



Self-Selection Bias

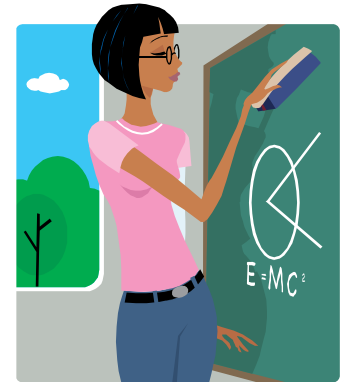
- In **Self-Selection Bias**, a participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample.



- For example:
If you were to set up a booth to ask people about their grooming habits...
The people who respond are more likely to be those who take more time to primp in the morning than those who just throw on something and head out the door.

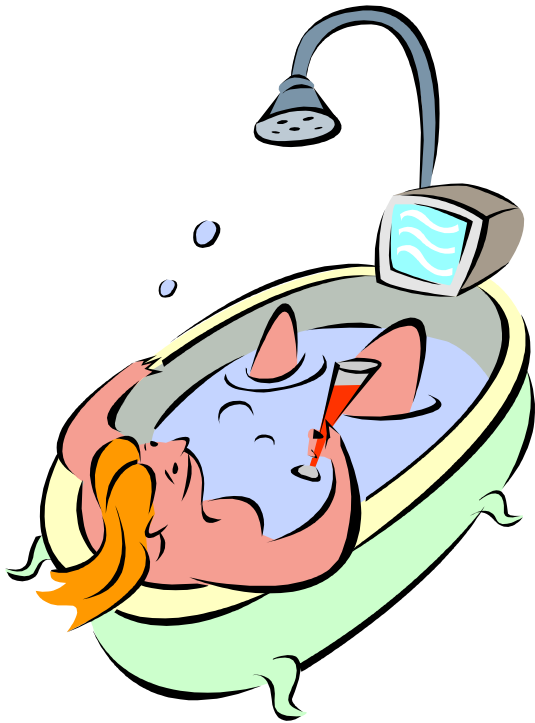
Leading Question Bias

- If you have a survey that asks:
 - Don't you think that CSE₃ TAs are paid too little?
 - A) Yes they should earn more
 - B) No they should not earn more
 - C) No opinion
- You are suggesting by the tone of the question what you believe the answer should be. That will **bias** your results (is it always bad?)



Social Desirability Bias

- If you ask people in a survey about how often they shower, or how often they recycle, your data is going to be biased by the fact that nobody wants to admit to doing something that is considered socially undesirable.



Don't Forget:

- Adding in a **Sampling bias** into your data collection is an important tool if you want to lie, cheat, manipulate, or mislead with your study results!



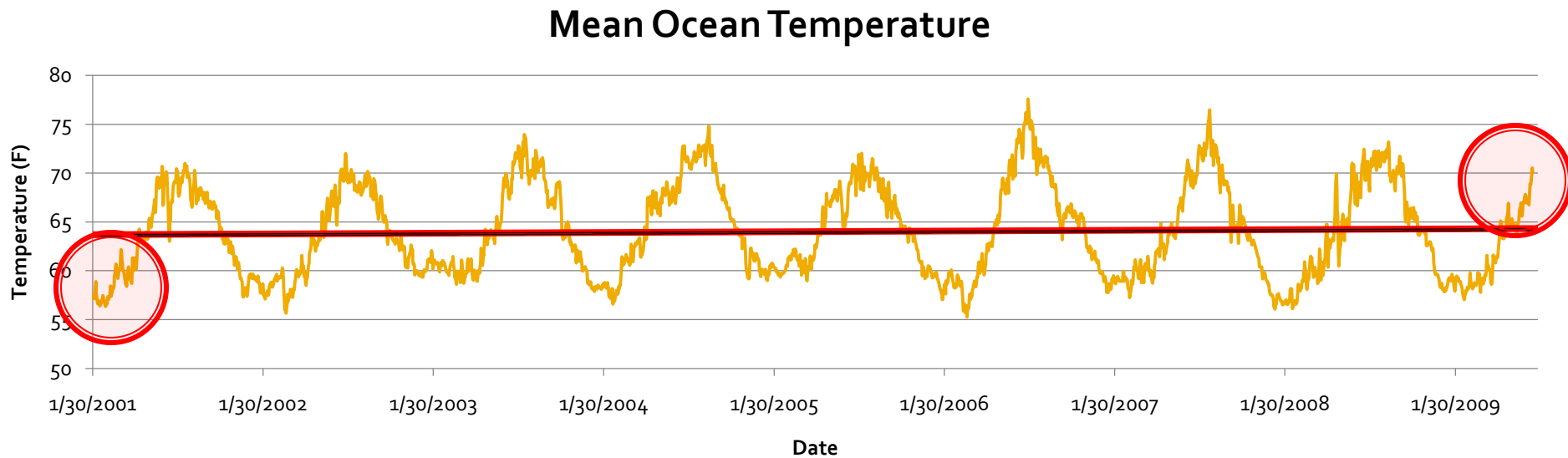
Analyzing Data

- **Data analysis** is a process of gathering, modeling, and transforming data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.



How to Lie, Cheat, Manipulate, and Mislead with poor Analysis:

- We saw in lab that if you weren't careful, you could accidentally miscalculate the trendline for the water temperature readings by including uneven data:



Other ways to Lie, Cheat, Manipulate and Mislead w/ Averages

- What if you were a real-estate agent and you were trying to convince people to move into a particular neighborhood.
- You could, with perfect honesty and “truthfulness” tell different people that the average income in the neighborhood is:



a) \$150,000

b) \$35,000

c) \$10,000

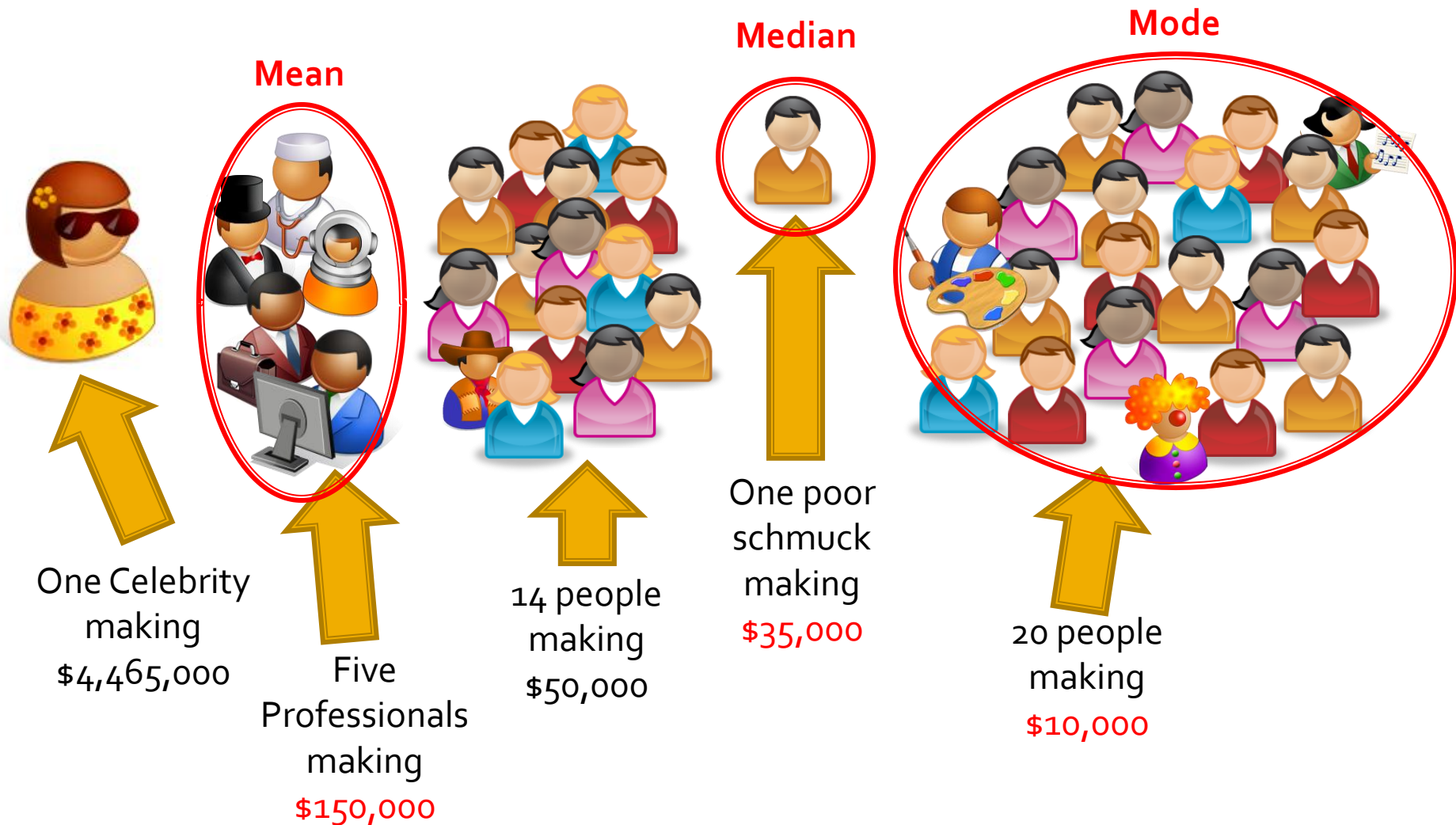
How? You ask with an evil grin...

- The \$150,000 figure is the arithmetic mean of the incomes of all the families in the neighborhood.
- The \$35,000 figure is the median.
- The \$10,000 figure is the mode.



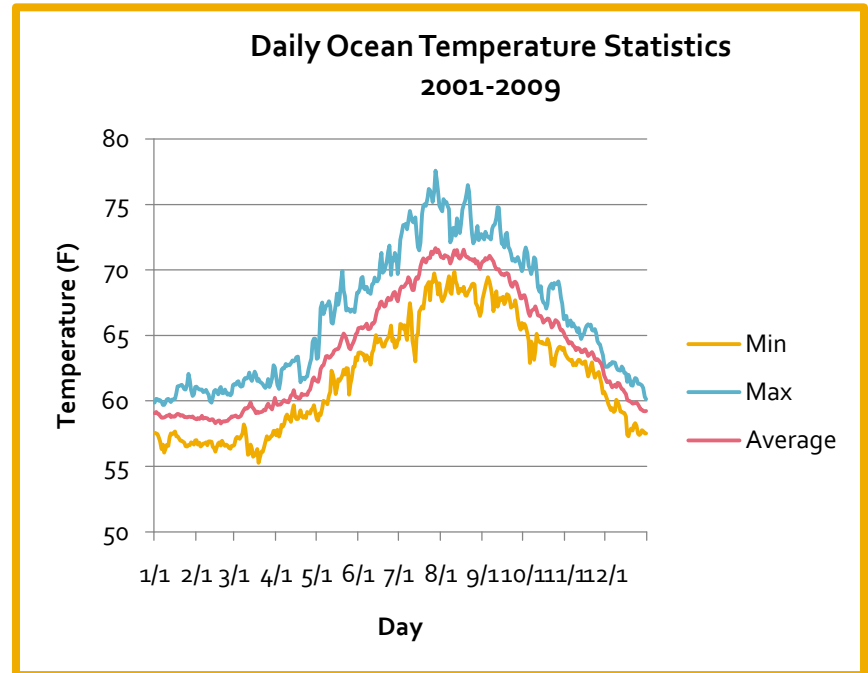
This particular neighborhood is lucky enough to be near a cliff... and the ONE home with an ocean view is a giant mansion on 50 acres that is owned by a Hollywood Star. With gates. And spikes. And security to keep out the riff raff of the rest of the neighborhood of poor people and the few middle class that live nearby.

Hmm... How does that work again?



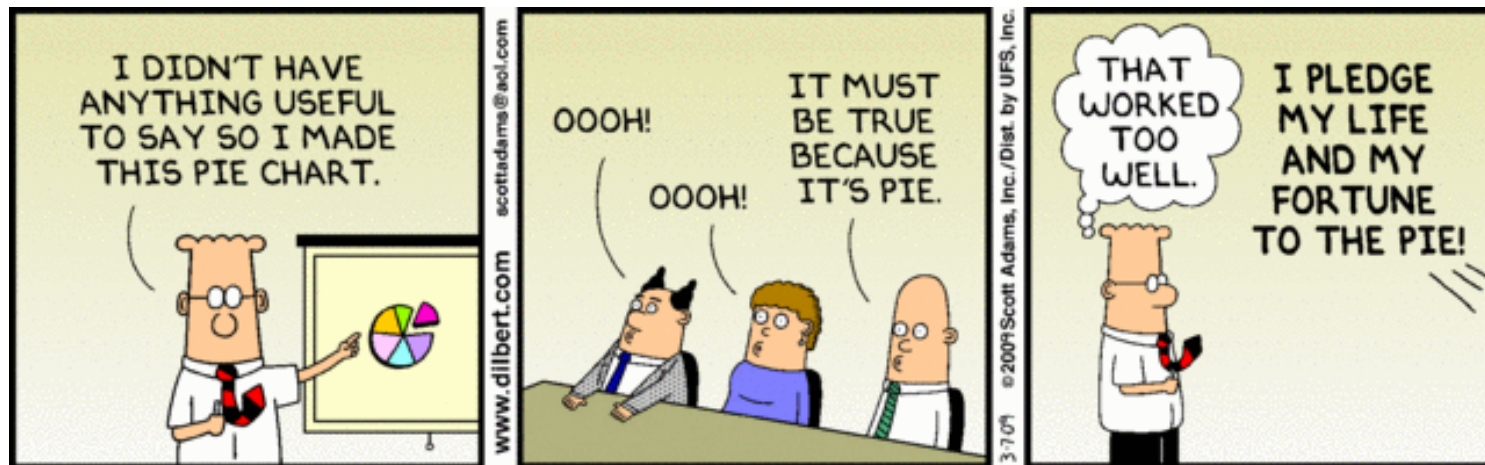
Interpretation of Data

- Interpreting data often involves displaying it in some useful way.
- To interpret our Water Temperature data, we created charts to visualize the information.
- Charts are a type of Graphical Display.



Graphical Displays of Information...

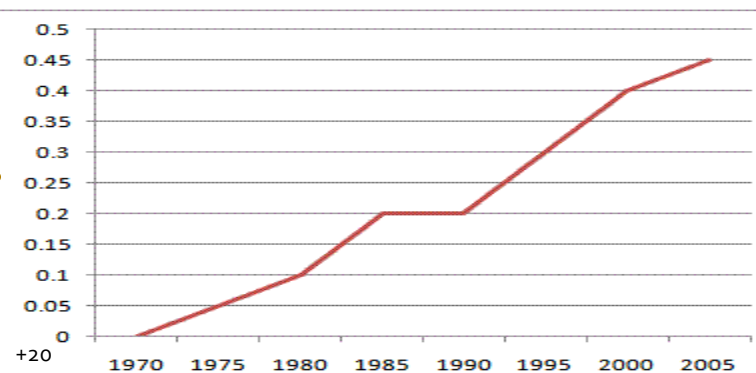
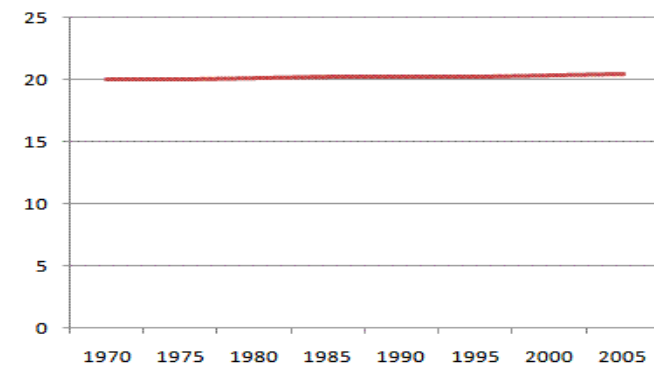
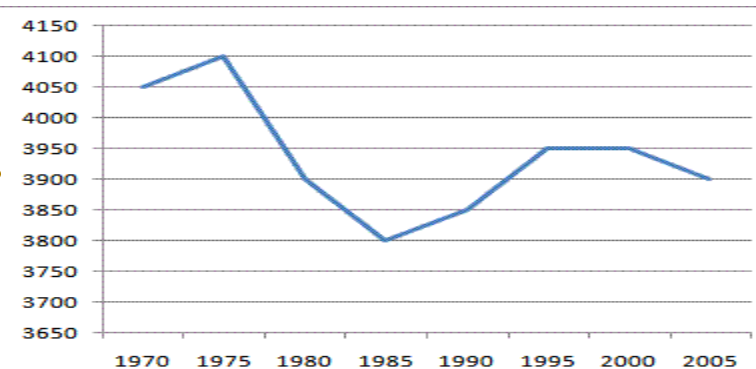
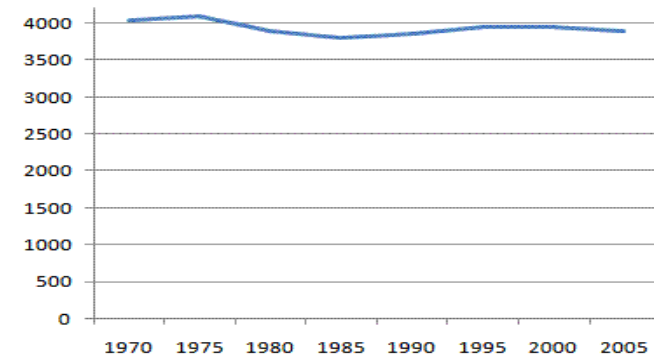
- If your goal is to lie, cheat, manipulate, or mislead, Graphical Displays are your friend...



How to Lie, Cheat, Manipulate, and Mislead w/ poor Graphical Displays:

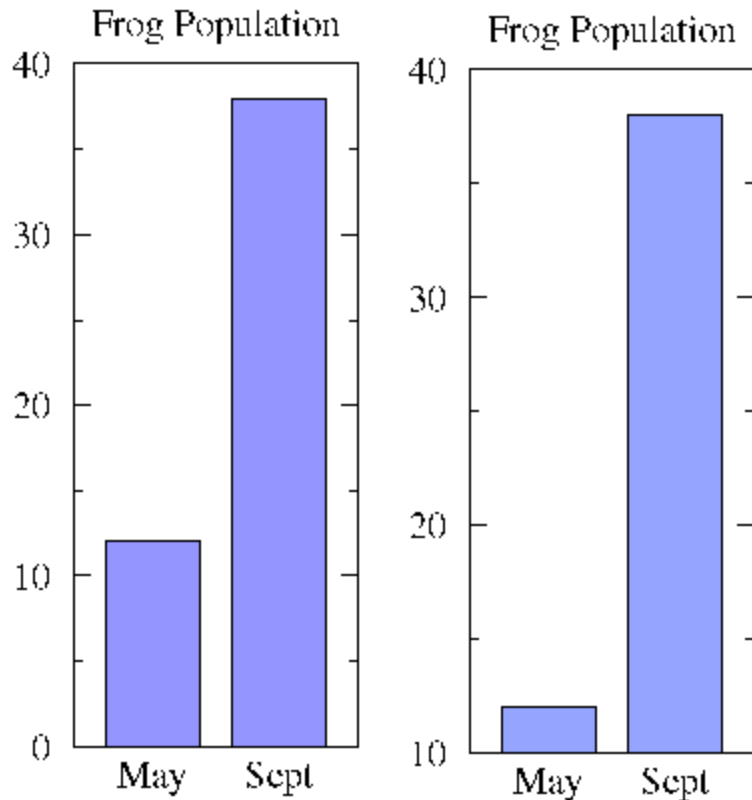
- The principals of Graphical Excellence (GE) are:
 - GE is the well-designed presentation of interesting data – a matter of **substance**, of **statistics**, and of **design**.
 - GE consists of complex ideas communicated with **clarity**, **precision**, and **efficiency**.
 - GE is that which gives to the viewer the **greatest number of ideas** in the **shortest time** with the **least ink** in the **smallest space**.
- To lie, cheat, manipulate, or mislead, do NOT follow this advice!!!

How to Lie, Cheat, Manipulate, and Mislead using Chart Adjustments



This is real data. The top graph shows the cosmic radiation rate in neutrons per hour. The lower is the temperature change since 1975 when it started. All from the BBC's website. They weren't trying to lie, cheat, manipulate, or mislead! No sirree.

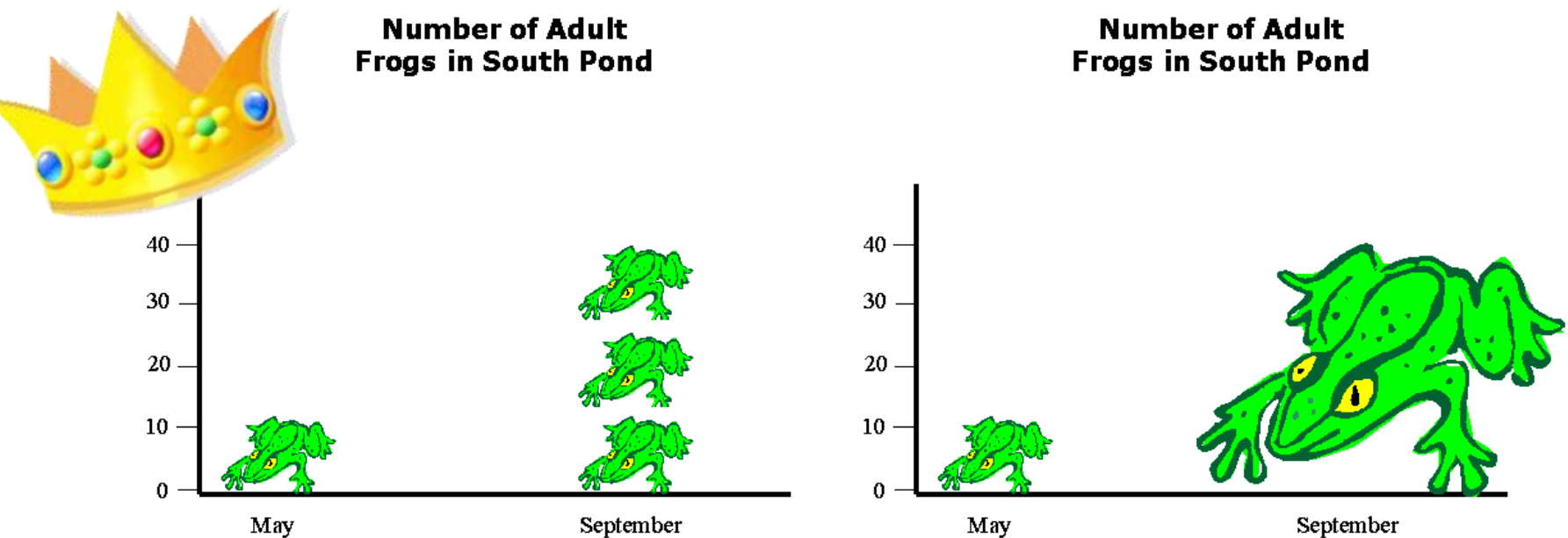
How to Lie, Cheat, Manipulate, and Mislead using Bar Charts



- Here, the data is the **same** but by changing the axis labels, someone was able to really suggest that the difference in population was much greater than it was.



It's even more dramatic to use volume instead of bars...

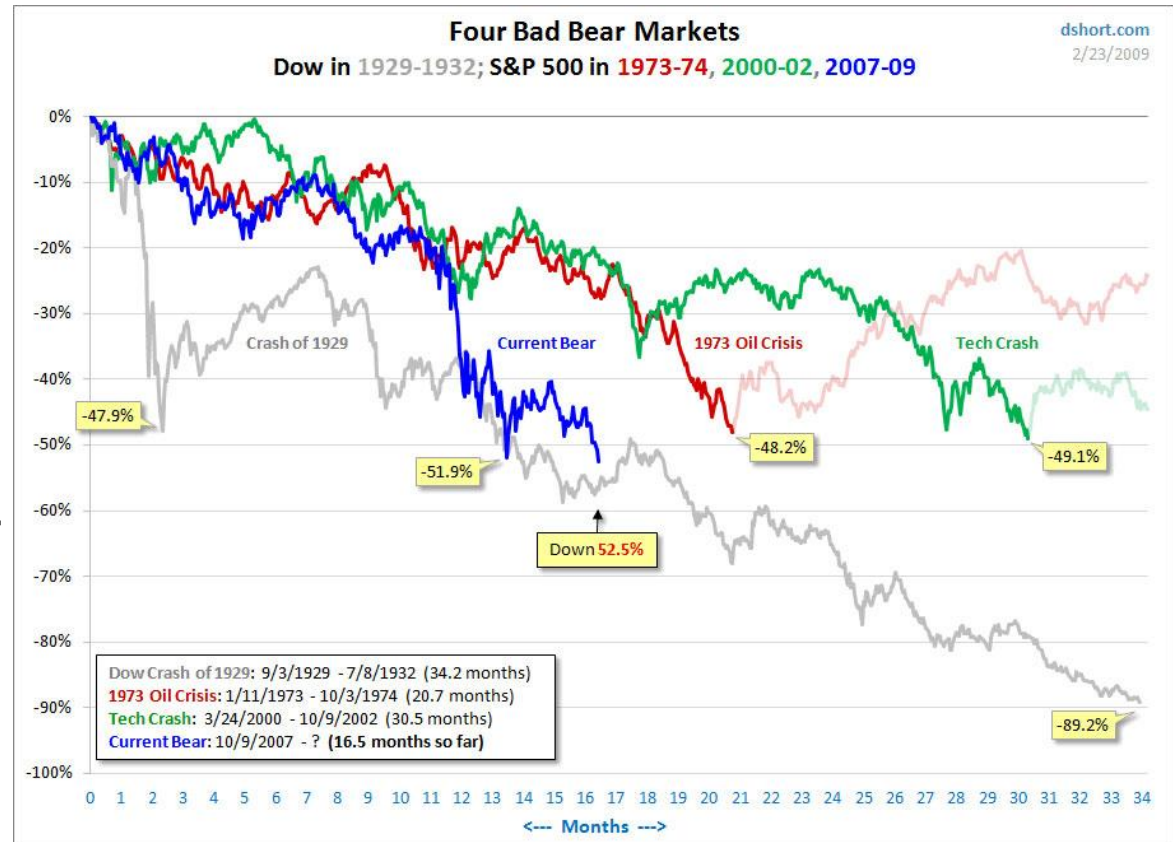


- Once again, both of these charts show the **same information** if you ONLY look at the HEIGHT of the frogs. The volume of an image is a great way to lie, cheat, manipulate, or mislead...

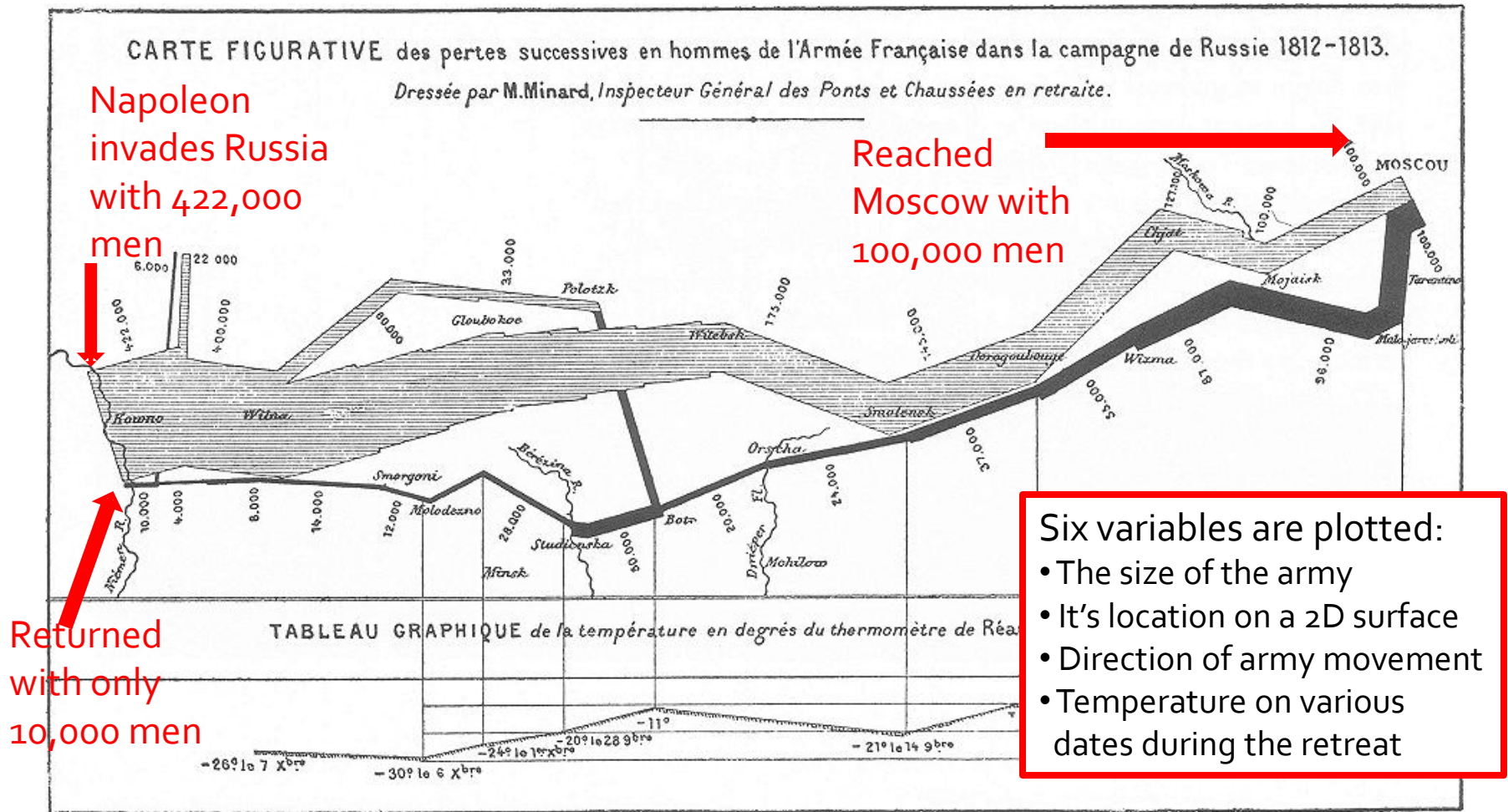


Some examples of GOOD Graphical Displays...

- Easy to read
- Lots of useful information
- Well labeled!
- Correct comparison of percentages rather than numbers!
- Scary results!



A very famous and excellent Graphical Display of information



So how do you tell the difference between good and bad information?

- Look at the sources. If none are given, do NOT trust the information.
- Check to see if there are any obvious sources of bias in the data. Look at how the data was collected and where it was collected from.
- Look very closely at the data axis and legend.
- And finally, do NOT believe everything you are shown just because it is “Science” and “Data”. Try to figure out if the source has some ulterior motive to manipulate your opinion.

Credits

- **How to Lie with Statistics**
 - By Darrell Huff, Norton, New York, 1954
- **The Visual Display of Quantitative Information**
 - By Edward Tufte, Graphics Press, February 1997