1 Lesson 4: Designing graphs and charts

Consider the problem of starting with a collection of data and picking the right graph or chart to illustrate information found in the table. We need to consider the various chart types available; we need to arrange the data to suit the chart we select; we need to format the chart to emphasize the things we want people to notice. Graphing data is most definitely not a one step process.

First, we can only work with the data we have. If we are simply given data, we may be limited by the form and the amount of the data we receive. If we collect the data ourselves, we need to think ahead to be sure that our results will suit our needs. Our first step is to organize the date, or at least identify the organization of the data we are given. Does the data have the information that we are trying to illustrate? Where does that data appear in the structure of each datum? Where are the numbers in the data and what do they represent?

1.1 Example 1

Let us consider our familiar examples of student grades:

Name	Test 1	Test 2	Test 3
April	55	71	64
Barry	63	67	63
Cindy	88	90	91
David	97	92	87
Eileen	58	55	75
Frank	90	89	96
Gena	88	100	85
Harry	71	70	71
Ivy	65	75	85
Jacob	77	70	65
Keri	75	88	85
Larry	88	92	92
Mary	95	95	100
Norm	86	82	80

We have discussed the various ways to interpret the organization of this data, but in this discussion we will consider each row as a datum. To simplify matters even further, we will rearrange the data by eliminating all but the top row of labels and one row of data.

For example, Mary:

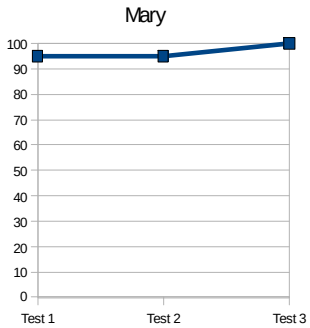
Name	Test 1	Test 2	Test 3
Mary	95	95	100

Mary's data set can be viewed as either a list of numbers $\{95, 95, 100\}$ or a table

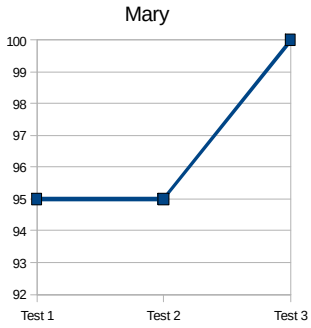
of pairs

Test 1	Test 2	Test 3
95	95	100

with the scores in a functional relationship with the tests. We might try representing this using a line chart (a scatter chart will probably be too sparse.)



or perhaps

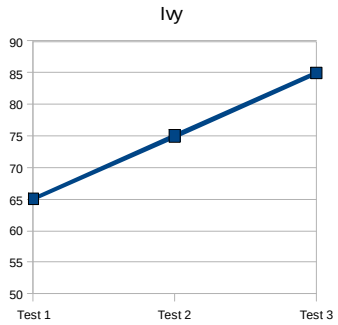


Notice that both these charts illustrate the same numerical information. The first has a scale from 0 to 100, and clearly illustrates that Mary has done quite well. The second has a scale set from 92 to 100. It make it easier to read Mary's exact test scores, but it make it look like her last test was a dramatic improvement over the previous two. Even though 100 is certainly better than 95, I think one would say that the main feature of Mary's work is that she has consistently done very well.

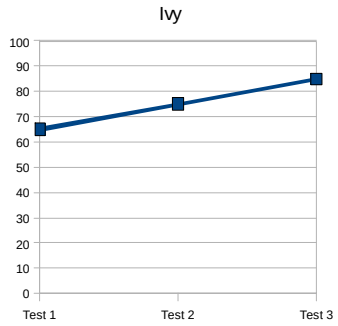
Consider another student, Ivy:

Name	Test 1	Test 2	Test 3
Ivy	65	75	85

Her scores can be clearly read from the graph



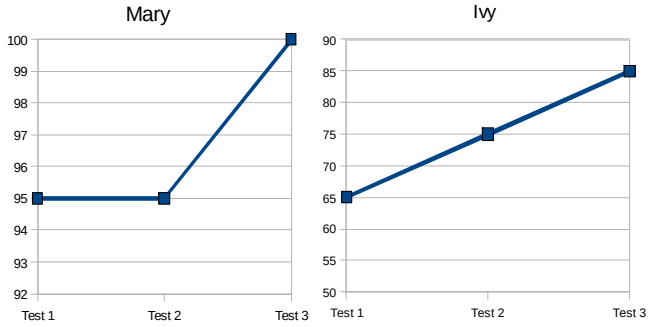
The graph also shows her steady improvement. This graph accurately illustrates Ivy's grades. On the other hand it does give the impression that Ivy is doing better than she is. The test scale only goes to 90 even though the tests are grades out of 100. Just like in Mary's case, a scale that is harder to read might be a better visual representation of Ivy's grades.



In this chart, we see Ivy's solid progress, and we see that her grades are not at the bottom of the scale. They are not quite at the top either.

It would also be wrong to compare the first graphical representation of Ivy's grades to either of the representations of Mary's. The graphs all have different scales, and these scales influence the way the data appears visually. If we are going to compare a graphic representation of Mary's scores with a graphic representation of Ivy's, the scales should be the same. Without that common scale, comparisons are still possible, but only after a much closer interpretation

of the two graphs:



The lesson to be learned from this is that the scale or scales used in a graph have a strong influence on the visual impact of the picture. The choice of the scale in a graph can have a major impact on the impression the graphic coveys. In the examples above, one presentation can highlight the fact that she consistently scores in the top range. Another graph can make her prefect score on the last test look even more impressive. One graph can emphasize Ivy's steady improvement but hide the fact that she still has room for much more. Another can give a more frank illustration of her overall grade standing that perhaps does not give her sufficient credit for her steady gains.

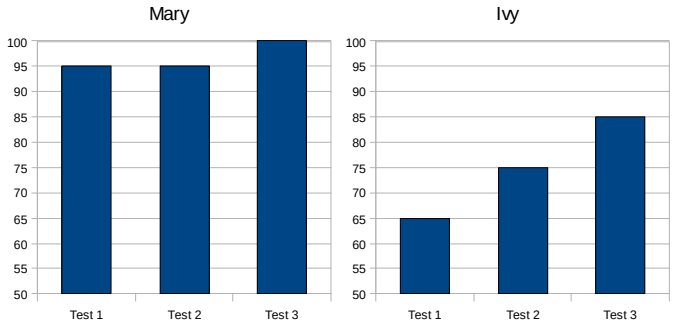
By using line graphs above, we have invited ourselves to view the three tests as sequential. By connecting the points in a scatter plot, we get the idea that the results occur in a particular order whether that is true or not. If we were to find another table of our data that looked like

Name	Math	English	History
April	55	71	64
Barry	63	67	63
Cindy	88	90	91
David	97	92	87
Eileen	58	55	75
Frank	90	89	96
Gena	88	100	85
Harry	71	70	71
Ivy	65	75	85
Jacob	77	70	65
Keri	75	88	85
Larry	88	92	92
Mary	95	95	100
Norm	86	82	80

we would realize that line graphs are not a good choice.

Column graphs can be used to present functional data without giving the impression that the results are sequential. Using columns, we can give Mary's

and Ivy's scores as



The scales are the same, so we can compare the two girls easily. Ivy appears further behind Mary than she actually is, but that is because of the scale, and it is not a fault of the column graph. We labeled the three tests with the old names. If the tests were sequential, then Ivy's steady improvement still shows. However, a better set of labels would easily eliminate this impression.

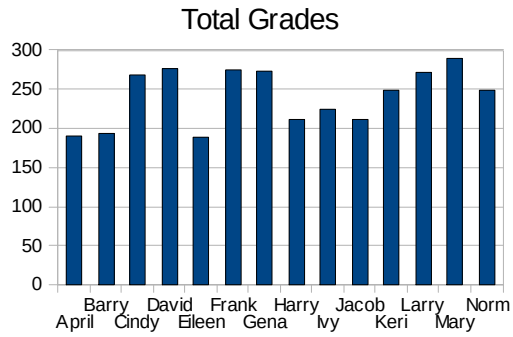
So far we have concentrated on creating graphs for each individual student. But our original data set is much larger than that. Can we illustrate the performance of the whole class. We could try to design a graph that illustrated all 42 numbers in the table, but the result would probably be a graph that is too crammed with information to be useful. It is a better idea to rearrange the data to create a more informative chart.

Let us add a column to out table that includes a total of each students grades. Remember, by doing arithmetic we are simply rearranging our data.

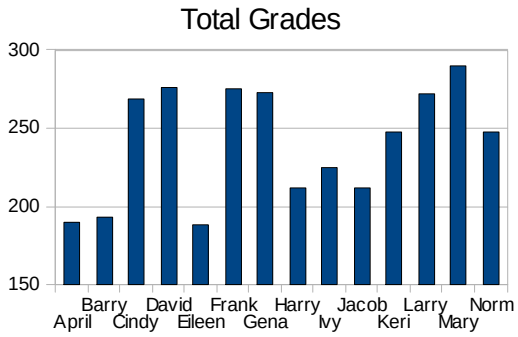
Name	Math	English	History	Total
April	55	71	64	190
Barry	63	67	63	193
Cindy	88	90	91	269
David	97	92	87	276
Eileen	58	55	75	188
Frank	90	89	96	275
Gena	88	100	85	273
Harry	71	70	71	212
Ivy	65	75	85	225
Jacob	77	70	65	212
Keri	75	88	85	248
Larry	88	92	92	272
Mary	95	95	100	290
Norm	86	82	80	248

Once we have this total column, we rearrange once again by ignoring all the other grade columns. With a functional relationship between the students and

their totals, we are ready to graph. Columns seems like the best choice:



As a teacher, you might be more interested in highlighting the differences in the different students Since everyone has scored at least 150 points, you might start the scale there.



You might also consider using a percentage scale instead of a total.

1.2 Example 2

We have already seen a table of data that gives word counts from three samples of Mark Twain's writings and word counts from letters attributed to Quintus Curtius Snodgrass. The table looked like:

Word Counts

Word length	Twain 1	Twain 2	Twain 3	Snodgrass
1	74	312	116	424
2	349	1146	496	2685
3	456	1394	673	2752
4	374	1177	565	2302
5	212	661	381	1431
6	127	442	249	992
7	107	367	185	896
8	84	231	125	683
9	45	181	94	465
10	27	109	51	276
11	13	50	23	152
12	8	24	8	101
13	9	12	8	61
Totals:	1885	6106	2974	13175

As usual, our first question is: how is this data organized?

First it helps to ignore any computed data in the table. Since that information is derived from the other data, it can always be re-derived later. At this point it is only a distraction. Next we identify and ignore any labels that appear in the table. We do not need to remove them since they will probably be useful later, but we do want to set them apart for now. So where are we? We are at least pretending that the "Totals" row is gone. Also the top line of labels should be gone. What about the word lengths? These are basically labels. They are number labels, and that is good, but they are labels and for now left out in identifying the organization of the data. We are left with a 13 by 4 table of numbers. Great:numbers are our friends.

The question now becomes: is each row a datum or is each column a datum? From a strict mathematical sense, it could be either. Still we have some ideas about what we want from this data, and those ideas must influence the way we identify the organization of the data. Is our primary interest the words or the authors? It is the authors. If this is the case, then the entries associated with the authors should be organized to be together. We do this by declaring that each column is a datum. Thus we have 4 pieces of data each made up of 13 numbers that count the number of words of various sizes.

At first this may seem unwieldy. But our next step is to arrange, and that is where we try to make things manageable. The trouble is that there are too many numbers, so maybe we should leave some out. Because the datum are complicated, maybe we should lose some of them, at least for now. The idea then becomes: look at each datum from the original as its own smaller data set.

For example, consider Twain 1:

Word length	Twain 1
1	74
2	349
3	456
4	374
5	212
6	127
7	107
8	84
9	45
10	27
11	13
12	8
13	9
Totals:	1885

Paired down like this, it is relatively easy to realize that this is a class and frequency chart for the writing sample. The classes are the word lengths. This itself suggests several possibilities, but let's not jump ahead. First we ask ourselves how this smaller data set is organized. In particular, is there a functional relationship we can use? Yes, the counts are in a functional relationship with the word lengths. Actually we expected when we realized that we had a class and frequency table.

Now it is time to look ahead. We are interested in this data so we can compare the writing styles taken from the samples. Eventually we expect to compare the charts from different samples. As we have seen, the charts we end up with should have the same scales. A look at our original table tells us that the total number of words varies considerably from table to table. Scaling a chart using the number of words would result in charts with different scales. To produce the same scales from different total counts, we can switch our frequencies from actual numbers to either percentages of the whole. Working

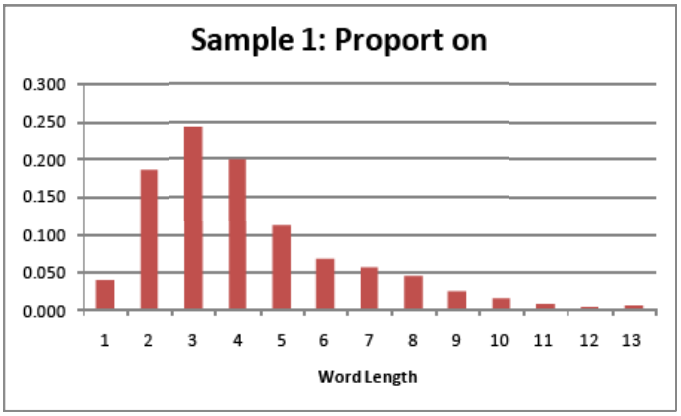
on sample 1, we rearrange our chart by adding two columns:

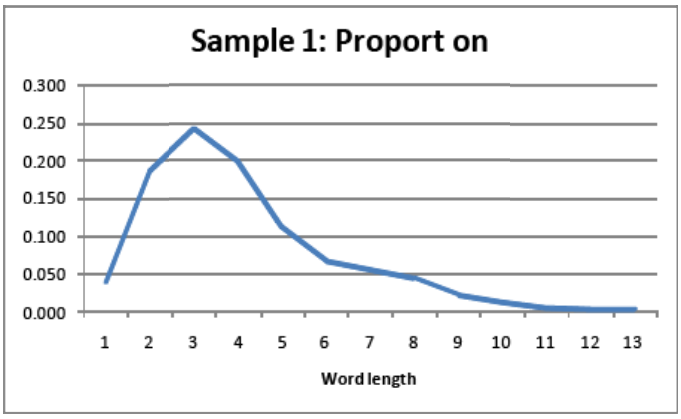
Word length	Twain 1	Proportion	Percentage
1	74	0.039	3.9%
2	349	0.185	18.5%
3	456	0.242	24.2%
4	374	0.198	19.8%
5	212	0.112	11.2%
6	127	0.067	6.7%
7	107	0.057	5.7%
8	84	0.045	4.5%
9	45	0.024	2.4%
10	27	0.014	1.4%
11	13	0.007	0.7%
12	8	0.004	0.4%
13	9	0.005	0.5%
Totals:	1885	1.000	100%

The choice between the proportion of the whole and the percentage is simply a matter of taste.

There are plenty of choices for the type of graph: scatter plot, column chart, line chart, pie chart, histogram, stem and leaf plot, or class and frequency plot. A scatter plot is probably going to be too sparse to give a good image. A pie chart will give us a pie with 13 pieces, but comparing it with another pie might not have the visual impact we want. The classes are already set and we have given up on the actual numbers in favor of ratios; so both a stem and leaf plot and its cousin the class and frequency plot are no longer appropriate. A histogram will end up looking like either a column graph with fat columns that touch or a line graph with the area under the line shaded in. We can postpone

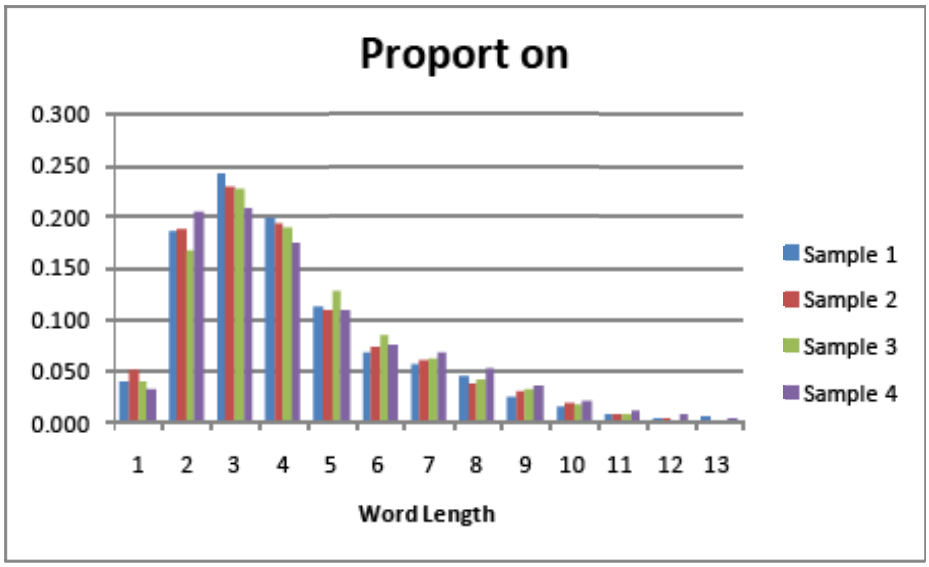
that choice until later We are down to line or column.

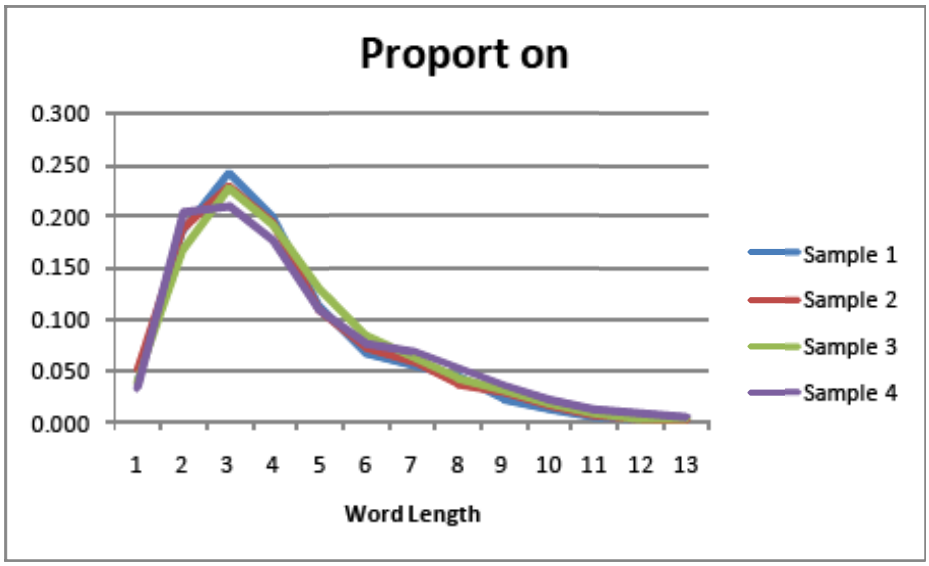




Both these give a good picture for the distribution of word lengths in sample 1. For the sake of comparison, we can draw similar graphs for the other samples. Of course, we use the same proportion scale through out. We could draw four separate graphs and place them side by side, or we can place all four in one

chart.





Neither of these graphs would have looked as good had we made them into histograms.

1.3 Example 3

The correct graph can illustrate information contained in data very clearly, and incorrect graph can be misleading. The following table contains a lot of

information, and finding the best way to illustrate can take some trial and error. The table contains the funding each of five states received from the Department of Education under the Math and Science Partnership program in thousands of dollars:

	2001	2002	2003	2004	2005	2006	2007	Total
Alabama	7,016	8,794	9,690	9,868	7,260	4,055	3,908	50,593
Alaska	2,250	3,075	3,214	3,304	2,405	1,317	1,317	16,885
Arizona	6,759	10,114	9,655	12,202	9,278	5,291	5,290	58,592
Arkansas	4,402	5,518	5,465	6,146	4,591	2,533	2,428	31,087
California	55,910	85,123	89,959	93,318	65,730	34,985	32,823	457,851

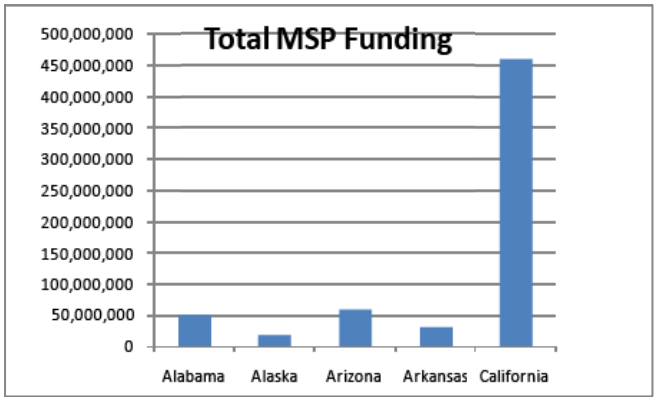
How is this data organized? The first row and the first column are labels. The last column contains the totals of the corresponding rows. For now, we will pretend that these are dropped from the table. The result is a table with 5 rows and 7 columns where all the entries are numbers. Should we consider each row as a datum or each column? If we are interested in the states, we keep their data together by considering each row a datum.

Now cramming 35 pieces of data into one chart is asking a lot from a picture. We will not give up on this, but for now we will try to cut down our expectations. If we are interested in showing how the funds were distributed among these five states, we can do this with a rearrangement of the data; Well, actually just keeping the table as we got it, with the column of totals, and then rearranging that by dropping all the columns except the labels and the totals.

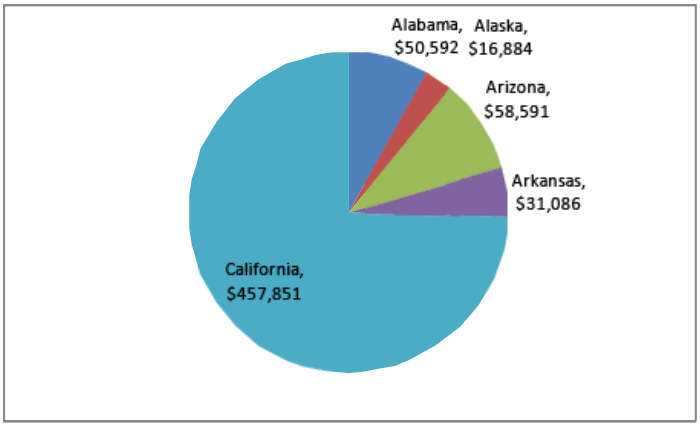
	Total
Alabama	50,593
Alaska	16,885
Arizona	58,592
Arkansas	31,087
California	457,851

The next question is what kind of chart should we use. Again, a scatter plot will be too sparse, but a line graph will give a sequential look to the graph by drawing arbitrary lines between the state totals. We probably do not want to

do that. A column chart is OK:

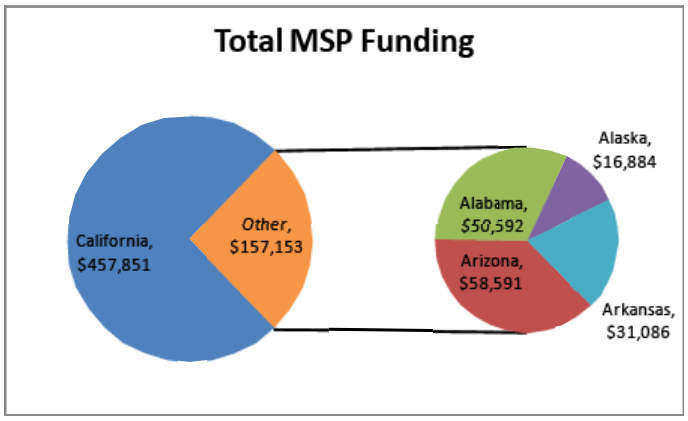


But when to comes to seeing the shares of the various states, nothing beats a pie chart:



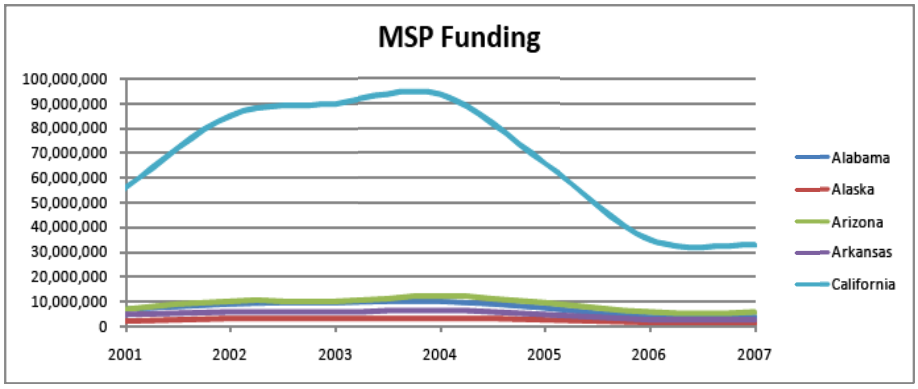
In both these charts, the California slice really stands out. The amount that California received dominates the other states so much, that the information about the other states almost is lost. There is not much we can do with the column chart to solve this, but there are tricks we can use with the pie chart.

We can try giving the smaller states their own little pie chart:



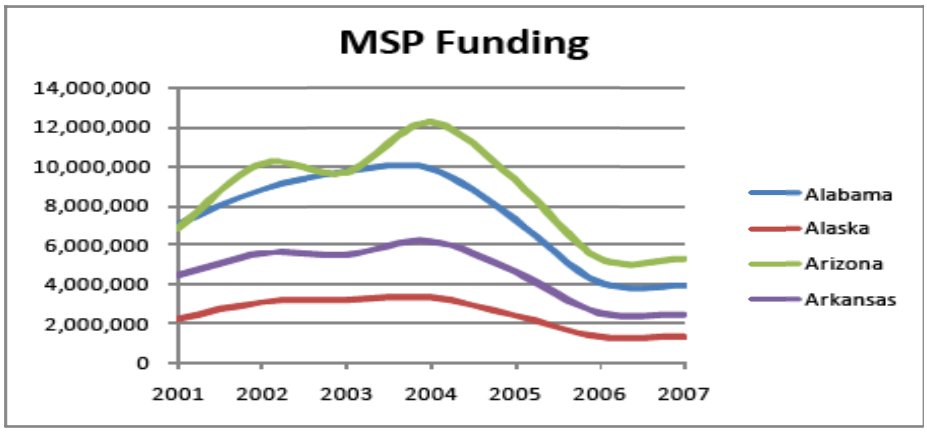
Thus a pie chart gives us more option for dealing with large differences in the data. A column chart is not as flexible.

Suppose, however, that we are interested in how the funds received varied over the years, a line graph gives a good picture of that information. The lines make sense because the years are sequential in time. This time we will smooth things out by curving the lines a bit. We can also try to put all five states on one graph



Unfortunately California is just too big to be on the same chart as the others.

Perhaps we should rearrange the data by leaving California out completely.



1.4 Example 4

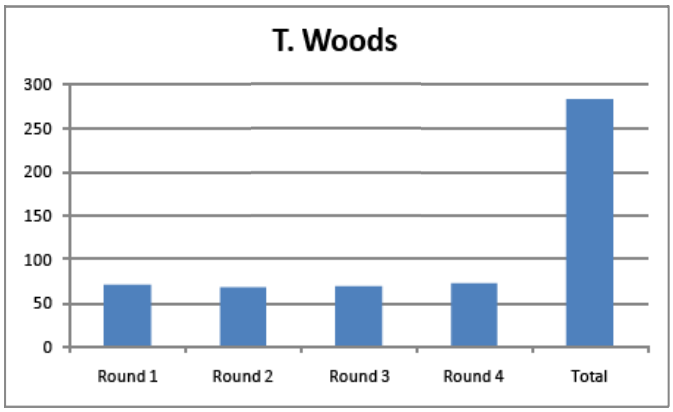
Consider another collection of data. In the 2008 United States Golf Associations's US Open was won by Tiger Woods in a playoff with Rocco Mediate. All the players played four rounds of 18 holes, but in the end Woods and Mediate were tied. Here are the results of the top 5 players in the regulation play that led to this tie.

Player	Round 1	Round 2	Round 3	Round 4	Total
T. Woods	72	68	70	73	283
R Mediate	69	71	72	71	283
L Westwood	70	71	70	73	284
R. Karlsson	70	70	75	71	286
D. Trahan	72	69	73	72	286

What is the best way to give a visual presentation of the match between these 5 players?

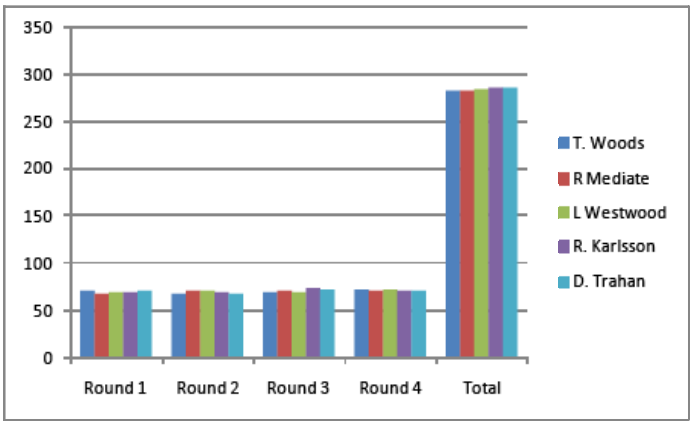
It seems reasonable to say that the data is organize so that each row is a datum. The makes sense, because we are probably more interested in looking at the players, and not the rounds of golf. Graphing one datum as its own data

set would give us a chart like:



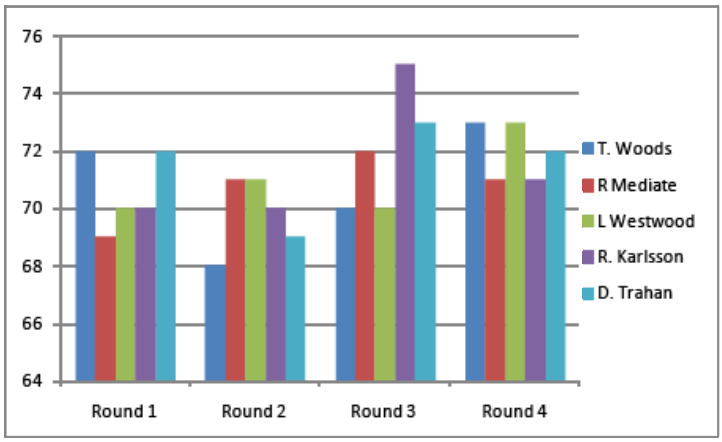
In this point of view, we have arranged the data in decreasing order of performance which means increasing total scores.

We can represent all the players data with one column chart:



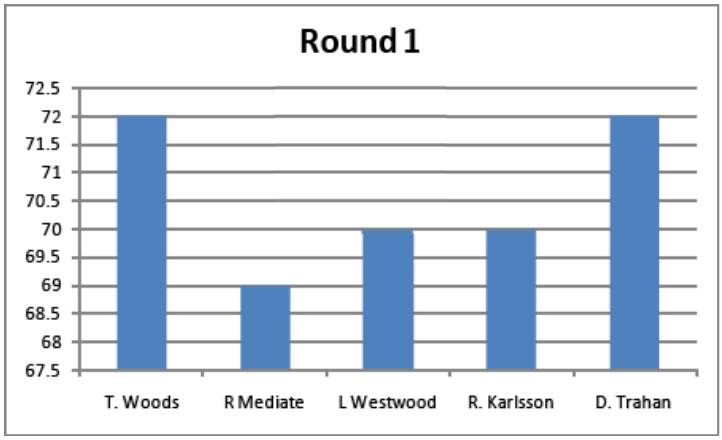
We are interested in how the match ended in a tie; so we included the total score. Unfortunately, this did not work out very well. By including the total score, we have made the differences in all the scores hard to read. It might be better to leave the total out. It is also a good idea to set the scale to exaggerate the differences. We know to do this; we set the bottom of the scale higher than

0, say 64.

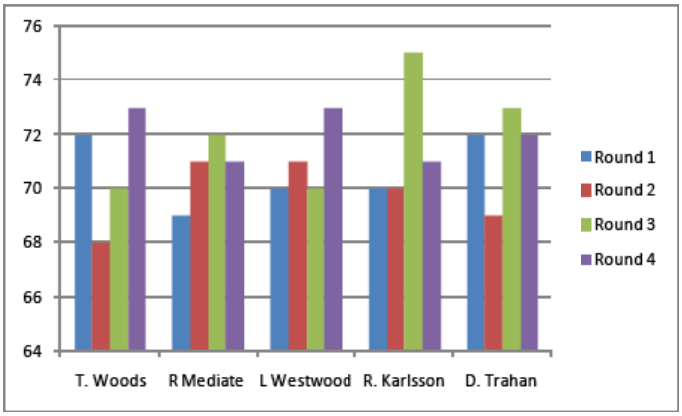


Now this shows the differences, but without the total, it is hard to see that it ended in a tie.

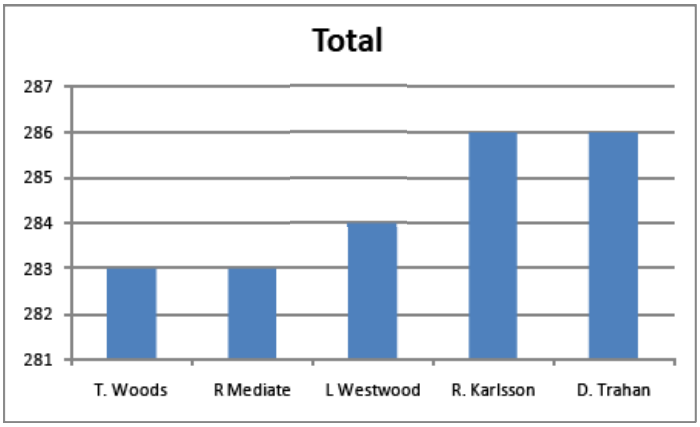
We can try a complete reorganization of the data. Instead of making a player's row a datum, we could make a column of scores in a round the datum. One round of data would look like



The match would look like

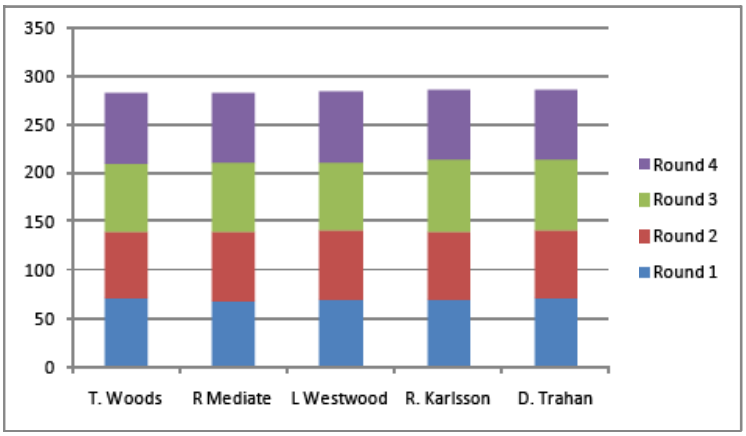


Another approach is just give a graph of the total, and again pick a scale that exaggerates the differences in score:



Here we clearly see that Woods and Mediate have tied in first place. (Of course, in golf the lowest score wins.) Still we have lost the rounds. Another trick is to graph the total by stacking the 4 rounds into one column for each player.

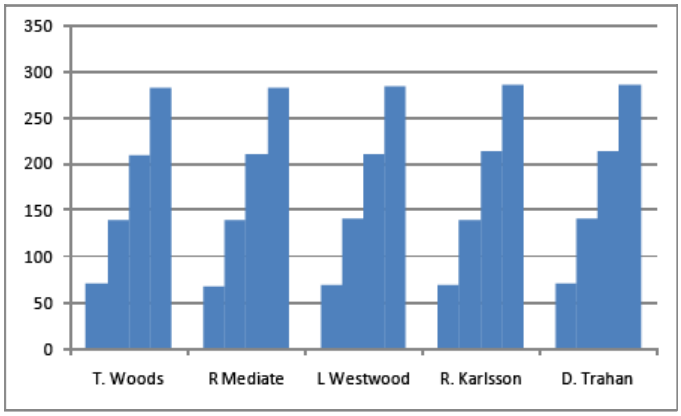
However, that precludes out trick of resetting the scale:



We can also try illustrating the round by round increase in scores for each player. First we rearrange the data to form a cumulation of scores as the rounds progress.

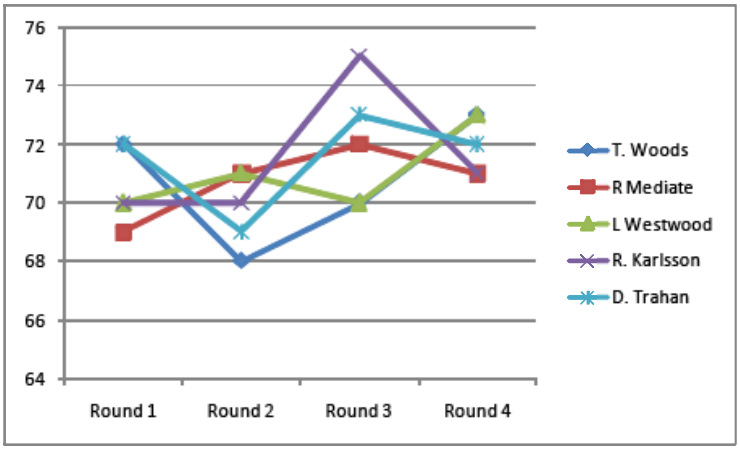
Player	Round 1	Rounds 1&2	Rounds 1-3	Rounds 1-4
T. Woods	72	140	210	283
R Mediate	69	140	212	283
L Westwood	70	141	211	284
R. Karlsson	70	140	215	286
D. Trahan	72	141	214	286

The next chart illustrates the progress through the rounds:

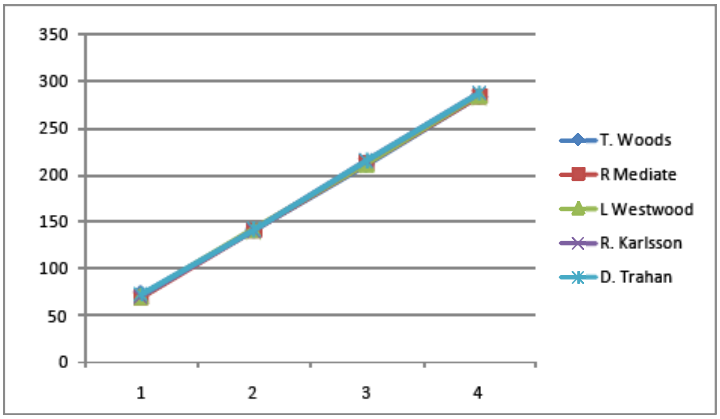


In the end, the only chart that illustrated the tie very well was the graph of the total. It may be that a column chart just will not do in illustrating the match..

We need to try other types of graphs. Consider the line chart



This shows the back and forth nature of the play, but again the final results have been lost. The column graph of accumulating scores came close to illustrating the progress of play through the rounds and the final outcome, perhaps it would make a better line chart?



Notice that the organization that look fine in a column chart has led to a very poor line chart.

In retrospect, we have not arranged our data correctly to be displayed well in any of these charts. It seems that the performances of the first five players in this tournament were so similar that the graphs lie one on top of the other. In retrospect, we should have realized that the differences in the scores was never more than 5. Since the total scores are between 283 and 286, these five players were never separated by more than 2% of the total.

We seem to be failing miserably to illustrate this golf match graphically because the numbers we are using are too large compared to the differences between them. The problem of finding the right graph to represent data is not

just a problem in mathematics; it is also a problem of aesthetics. Either way a bit of cleverness can be a great help.

Golfers know the problem we are dealing with quite well, and they have already found a solution. Instead of measuring their performance in absolute score, they measure it against a theoretical idea score known as par. On the course used in this match, par is 71 for each round. Thus Wood's first round of 72 was one over par, or simply +1. His second round of 68 is three below, or −3. Karlson's score in round 4 of 71 is measured as even par, or simply 0.

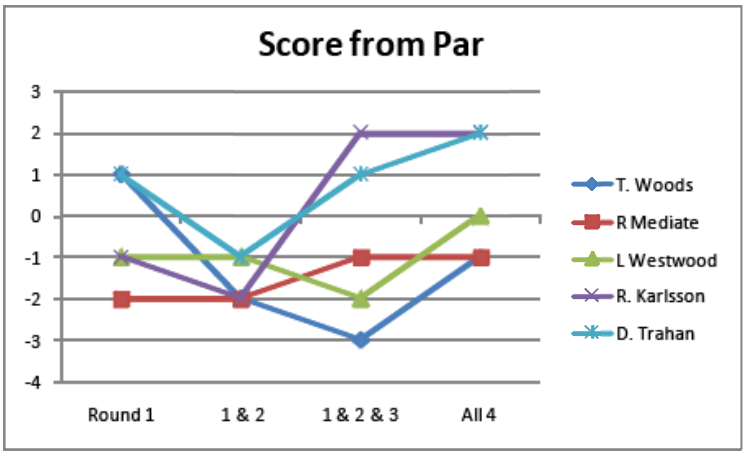
We should rearrange our data to fit this par scoring:

Player	Round 1	Round 2	Round 3	Round 4	Total
T. Woods	+1	−3	−1	+2	−1
R Mediate	−2	0	+1	0	−1
L Westwood	−1	0	−1	+2	0
R. Karlsson	−1	−1	+4	0	+2
D. Trahan	+1	−2	+2	+1	+2

Better still, we should rearrange this in running totals:

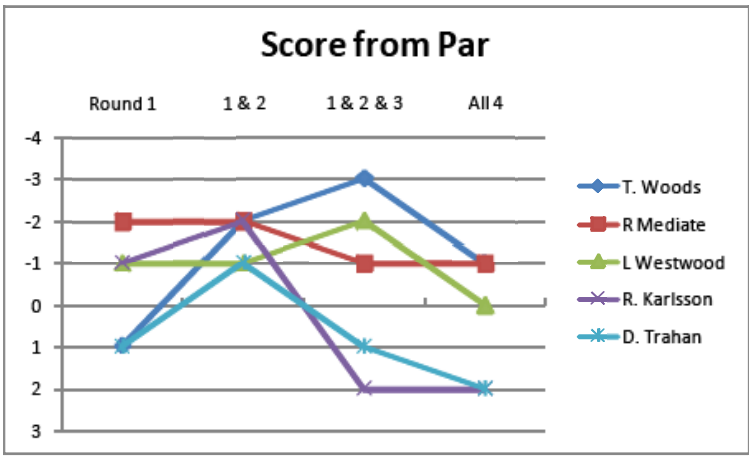
Player	Round 1	Rounds 1-2	Rounds1-3	Rounds 1-4
T. Woods	+1	−2	−3	−1
R Mediate	−2	−2	−1	−1
L Westwood	−1	−1	−2	0
R. Karlsson	−1	−2	+2	+2
D. Trahan	+1	−1	+1	+2

This leads to the line graph



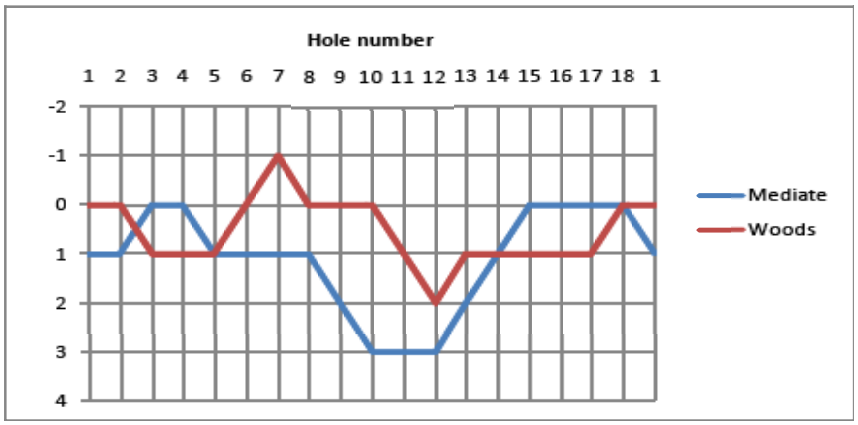
This finally shows the fight over all four rounds, and the tie. One last change

may lead us to the best picture possible



Notice that the scores are listed low to high, because the lowest score indicates the leader. This puts the winners at the top.

So the match ended with Tiger Woods and Rocco Mediate tied. These two played an 18 hole playoff, and the results are shown in the graph:



The graph shows Woods holding a lead from the 6-th to the 14-th hole, but Mediate made a rather dramatic comeback starting on 12. At the end of 18 holes, they were still tied. In a sudden death playoff, Mediate shot +1 on the first hole, but Woods shot par to win the tournament.

Selecting the best graph to represent the information found in data is almost an art form. But the steps are often more mathematical than artistic. The data must be organized correctly to concentrate on the right parts. The proper arrangement of the data is crucial. This may be a simple reordering of the list in alphabetic, numerical, or sequential order. It might require condensing the data by dropping parts completely. It may mean that some arithmetic is

in order: computing percentages, averages, totals or deviations from average. Selecting the right type of chart often takes experimentation. Final issues like shapes, colors and arrangement can lead to striking graphical results. It is worth the effort. A good picture can really clarify complicated ideas; a bad illustration can be very confusing.

Prepared by: Daniel Madden and Alyssa Keri: May 2009