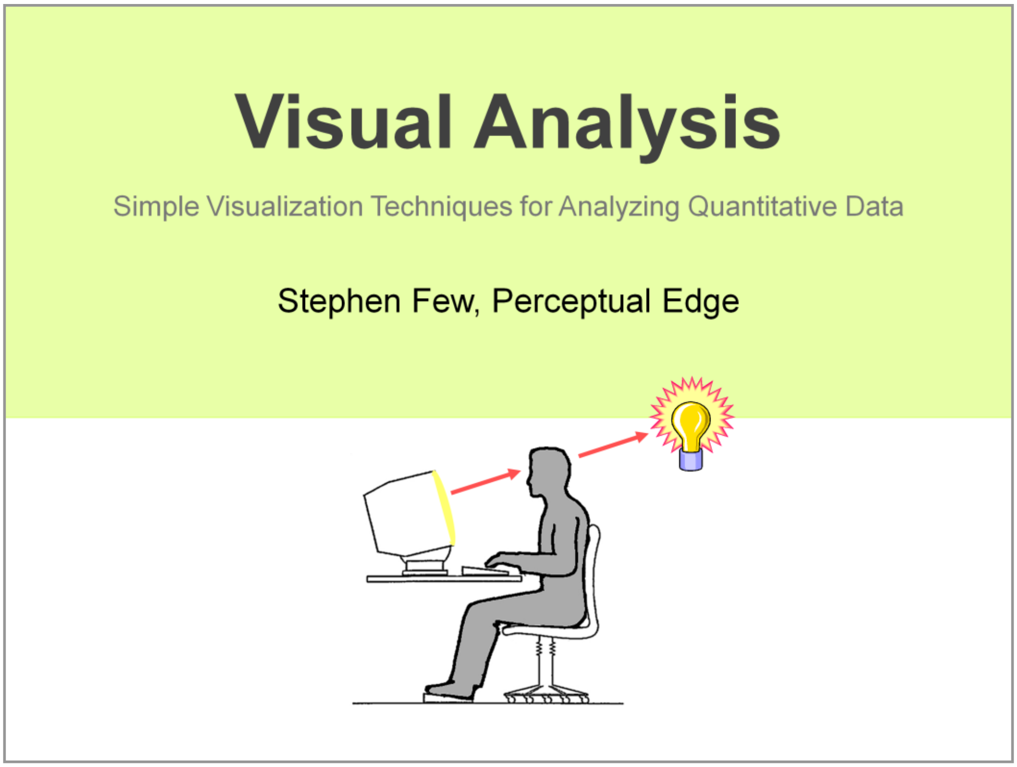
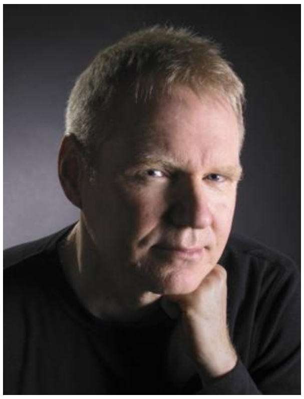


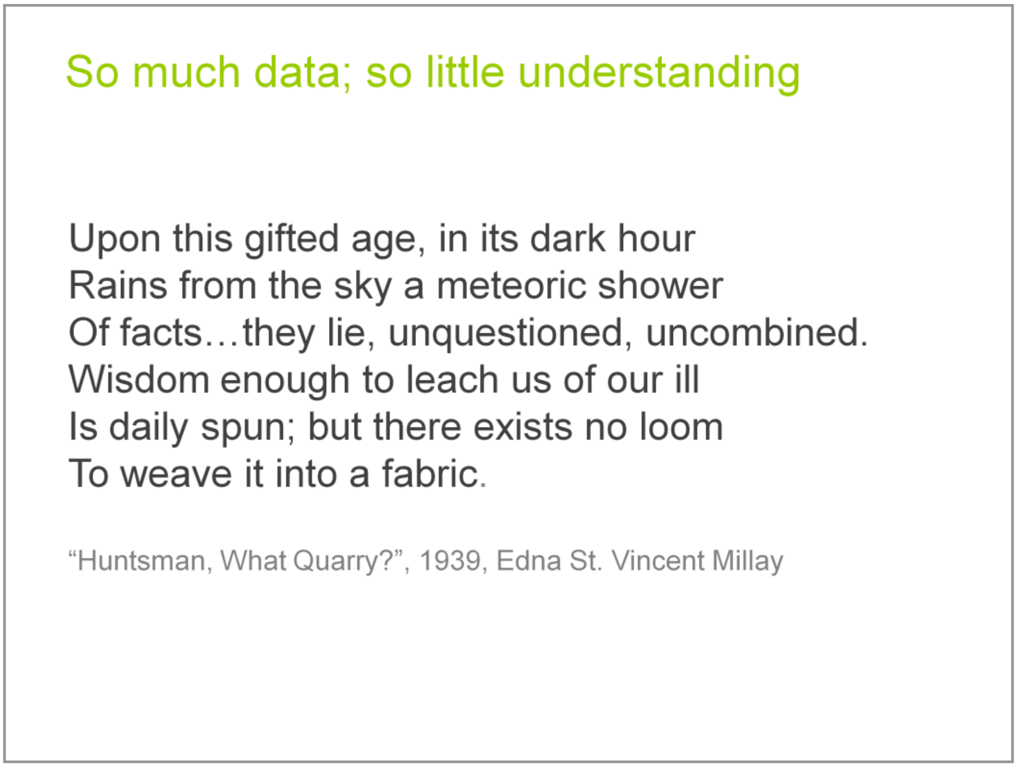
# Visual Analysis

Simple Visualization Techniques for Analyzing Quantitative Data

Stephen Few, Perceptual Edge



Stephen Few
Principal, Perceptual Edge
Author of the book
*Now You See It*



## So much data; so little understanding

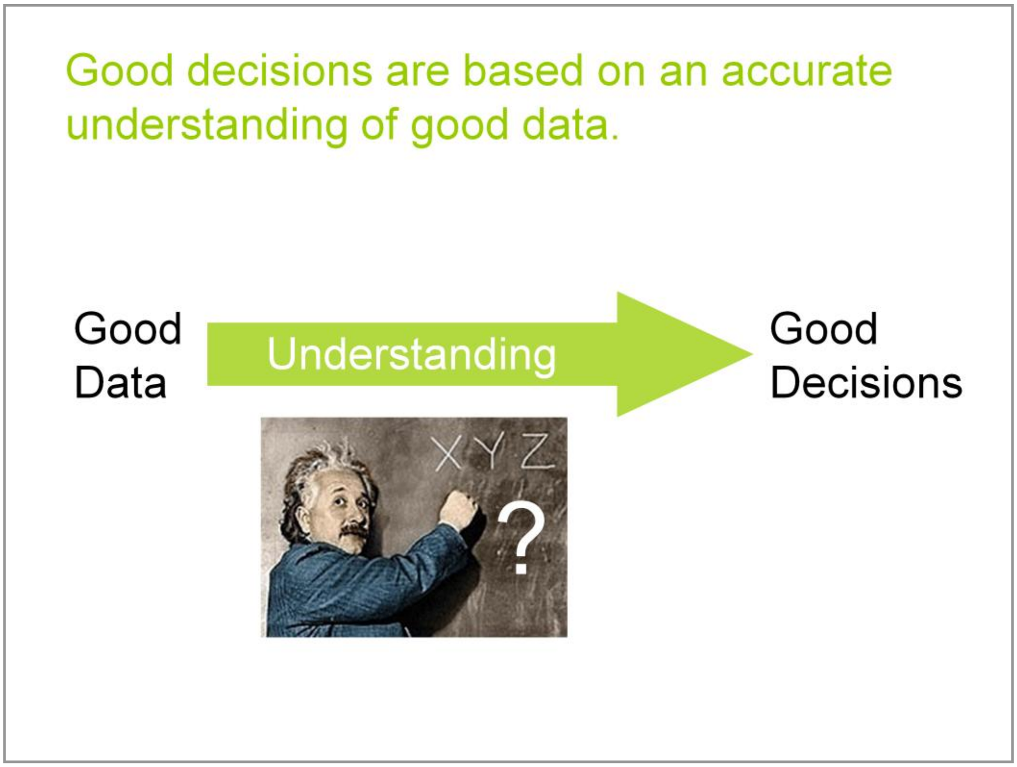
Upon this gifted age, in its dark hour
Rains from the sky a meteoric shower
Of facts…they lie, unquestioned, uncombined.
Wisdom enough to leach us of our ill
Is daily spun; but there exists no loom
To weave it into a fabric.

“Huntsman, What Quarry?”, 1939, Edna St. Vincent Millay

The amount of information that is available to businesses has increased dramatically in the last few years, but the ability to make use of it has increased little, if any.

> *Our networks are awash in data. A little of it is information. A smidgen of this shows up as knowledge. Combined with ideas, some of that is actually useful. Mix in experience, context, compassion, discipline, humor, tolerance, and humility, and perhaps knowledge becomes wisdom.*
>
> (*Turning Numbers into Knowledge*, Jonathan G. Koomey, 2001, Analytics Press: Oakland, CA page 5, quoting Clifford Stoll)

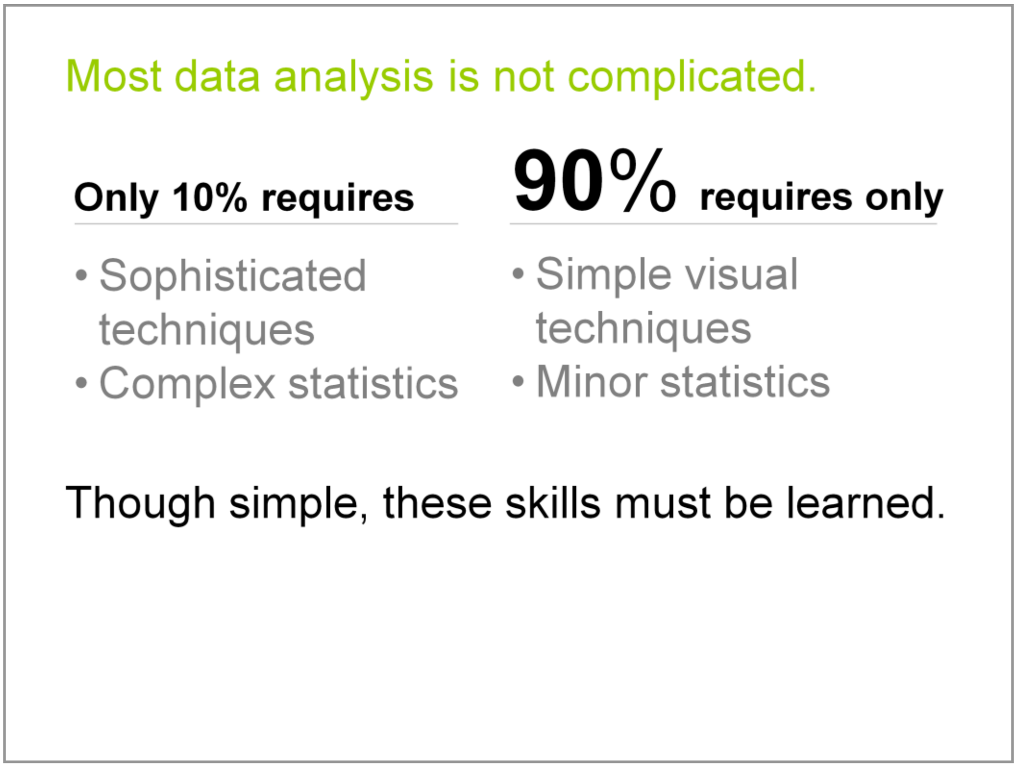
Most of us who are responsible for analyzing data have never been trained to do this. Knowing how to use Excel or some other software that can be used to analyze data is not the same as knowing how to analyze it.



Good decisions are based on an accurate understanding of good data.

Good Data → Understanding → Good Decisions

You are the key that opens the door for good data to result in good decisions. Software, no matter how sophisticated, is useless if you don’t possess the fundamental skills of data analysis. Data analysis is for one purpose: to enable good decisions. Do you need to be an Einstein to make sense of your business data?



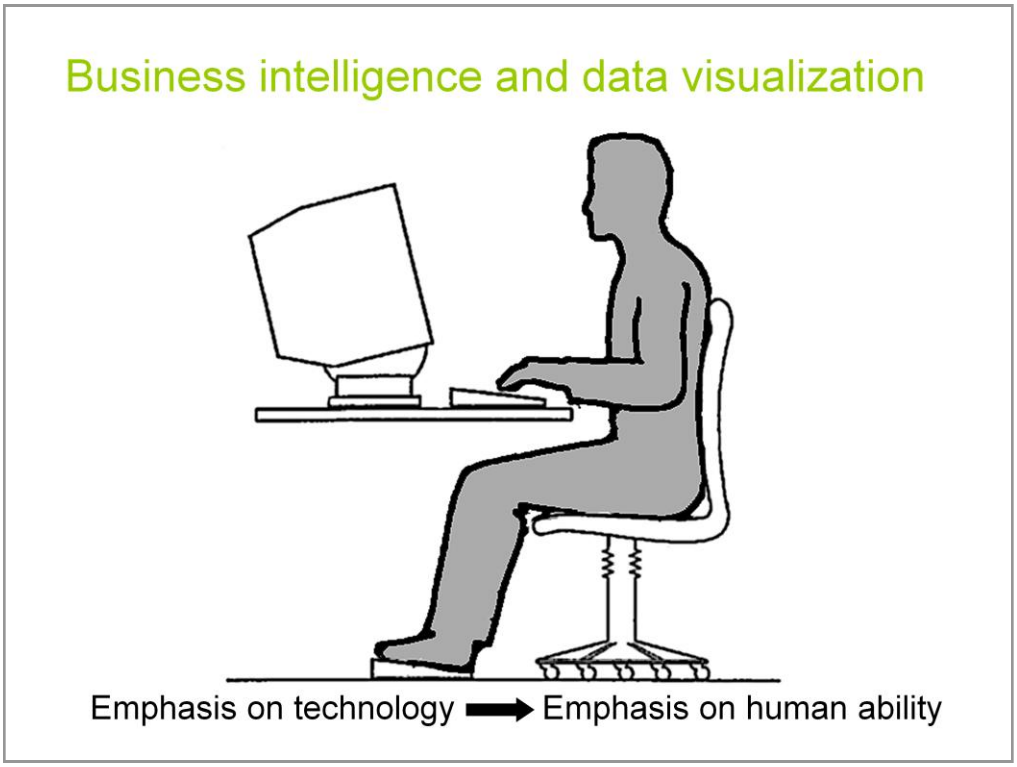
## Most data analysis is not complicated.

| Only 10% requires | 90% requires only |
|---|---|
| • Sophisticated techniques<br>• Complex statistics | • Simple visual techniques<br>• Minor statistics |

Though simple, these skills must be learned.

Most of the data analysis that is needed in the normal course of business requires relatively simple data visualization techniques, leaving little that requires the sophisticated techniques of statistical and financial analysis. If you search for resources that teach data analysis skills, you'll find many books and courses that present the sophisticated techniques needed by the few, but few resources if any that teach the simple techniques that most of us need to make sense of business data. The skills that most of us need to infuse our businesses with needed insights can be learned without a background in statistics, but these skills don't come naturally – they must be learned. You must develop expertise, but it is expertise that can be easily learned with the proper direction and practice. You must learn to see particular patterns in data that are meaningful.

> *People can learn pattern-detection skills, although the ease of gaining these skills will depend on the specific nature of the patterns involved. Experts do indeed have special expertise. The radiologist interpreting an X-ray, the meteorologist interpreting radar, and the statistician interpreting a scatter plot will each bring a differently tuned visual system to bear on his or her particular problem. People who work with visualizations must learn the skill of seeing patterns in data.*
>
> (*Information Visualization: Perception for Design*, Second Edition, Colin Ware, Morgan Kaufmann Publishers, 2004, page 209)



Business intelligence and data visualization

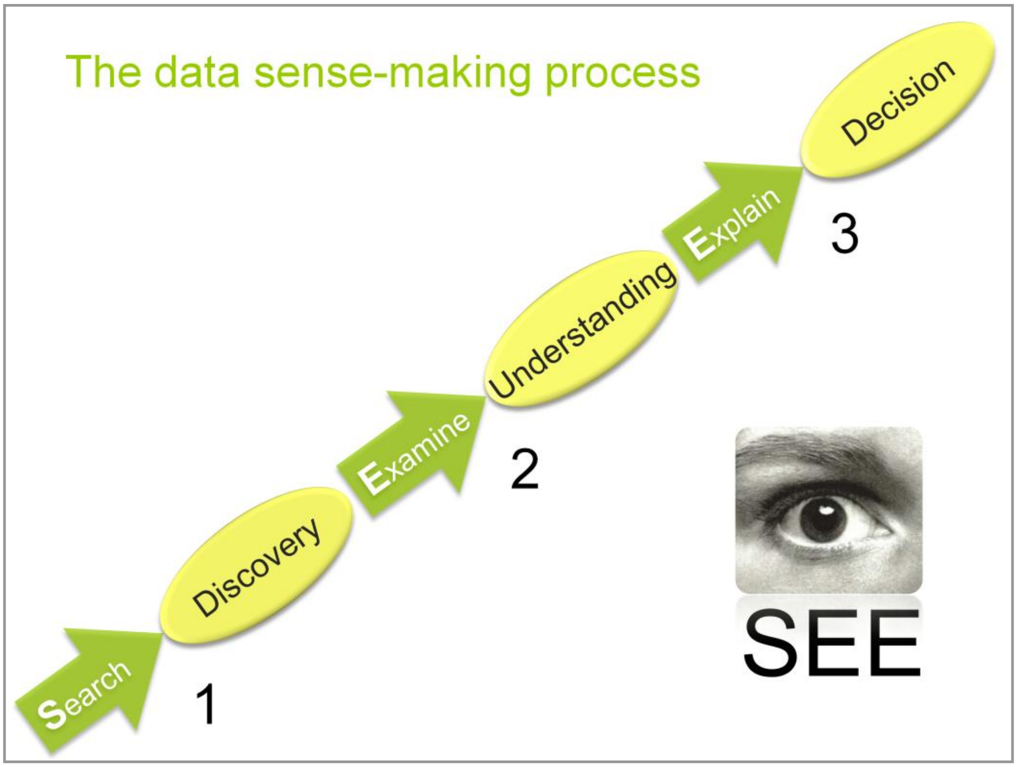
Emphasis on technology ➡ Emphasis on human ability

Something terribly important has been largely ignored in the realm of business intelligence (BI). This oversight has resulted in a great deal of waste and missed opportunities, even for organizations that have made a serious investment in business intelligence.

To date, business intelligence has focused on technology and project methodology, resulting in great advances. Today, we have huge and fast warehouses of information. Now it’s time to focus on the true essence of business intelligence – important, meaningful, and actionable information – and the most powerful resources for tapping into its value are those that engage the tremendous capacities of *human perception and intelligence* to make sense of and communicate information.

> *Many organizations aren't effectively analyzing the data they do have to improve their business.*
>
> *What's more troubling, perhaps, is many companies that purchase powerful analysis and business-intelligence tools don't use them effectively. Users of these products often generate the most basic and obvious reports and never get their hands dirty with the deep-analysis tools.*
>
> *By ignoring these products' deep-analysis capabilities, organizations could be missing important trends—information that might show, for example, where a company is losing business.*
>
> *Companies might also see that business decisions that were made based on basic information were wrong and ended up costing money in the long run.*
>
> (*eWeek,* “With Data Analysis, Less Isn’t More, Jim Rapoza, Ziff Davis Media Incorporated, June 7, 2004)

> *Today much of science and engineering takes a machine-centered view of the design of machines and, for that matter, the understanding of people. As a result, the technology that is intended to aid human cognition and enjoyment more often interferes and confuses than aids and clarifies.*
>
> *It will take extra effort do design systems that complement human processing needs. It will not always be easy, but it can be done. If people insisted, it would be done. But people don’t insist: Somehow, we have learned to accept the machine-dominated world. If a system is to accommodate human needs, it has to be designed by people who are sensitive to and understand human needs. I would have hoped such a statement was an unnecessary truism. Alas, it is not.*
>
> (*Things That Make Us Smart*, Donald A. Norman, Basic Books, New York, 1993, page s 9 and 227)



The data sense-making process

Search

1

Discovery

Examine

2

Understanding

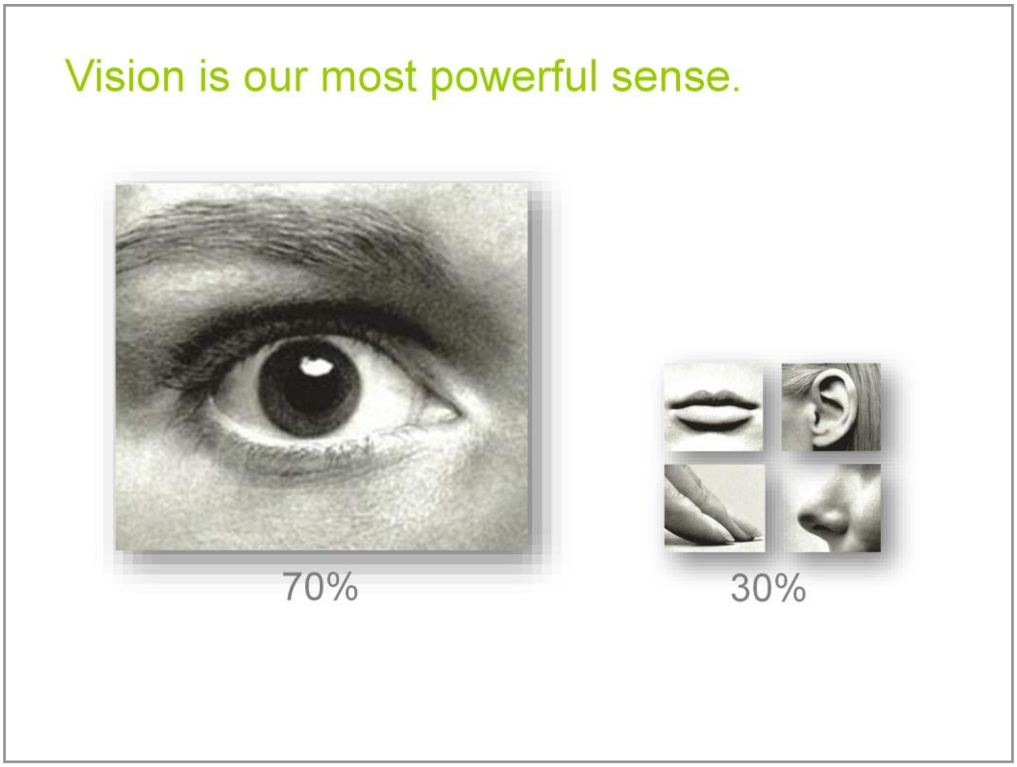
Explain

3

Decision

SEE

All business data analysis begins with (1) searching through the data to discover potentially meaningful facts, then involves (2) examining that data more closely to understand it, including what caused it to occur, so that you can then (3) explain what you've learned to those who can use that knowledge to make good decisions. Most of what we need to recognize and understand in our business data is not all that complicated.



Vision is our most powerful sense.

70%

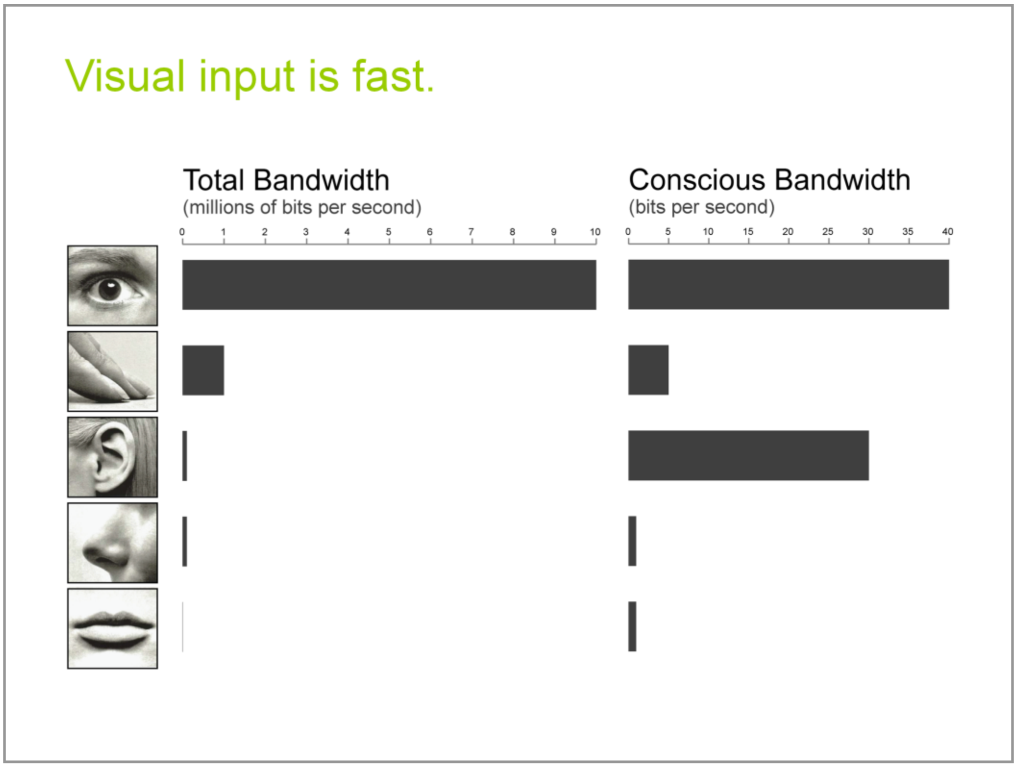
30%

Human perception is amazing. I cherish all five of the senses that connect us to the world, that allow us to experience beauty and an inexhaustible and diverse wealth of sensation. But of all the senses, one stands out dramatically as our primary and most powerful channel of input from the world around us, and that is vision. Approximately 70% of the body's sense receptors reside in the eye.

Perhaps the world's top expert in visual perception and how its power can be harnessed for the effective display of information is Colin Ware, who has convincingly described the importance of data visualization. He asks:

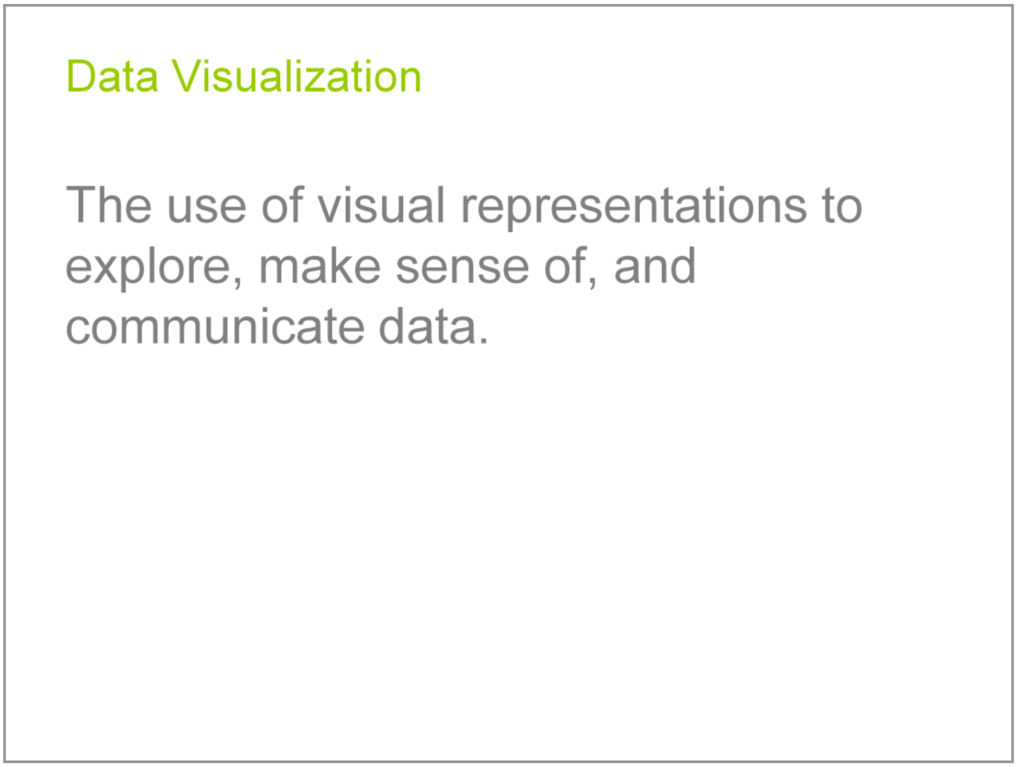
> *Why should we be interested in visualization? Because the human visual system is a pattern seeker of enormous power and subtlety. The eye and the visual cortex of the brain form a massively parallel processor that provides the highest-bandwidth channel into human cognitive centers. At higher levels of processing, perception and cognition are closely interrelated, which is the reason why the words 'understanding' and 'seeing' are synonymous. However, the visual system has its own rules. We can easily see patterns presented in certain ways, but if they are presented in other ways, they become invisible…The more general point is that when data is presented in certain ways, the patterns can be readily perceived. If we can understand how perception works, our knowledge can be translated into rules for displaying information. Following perception-based rules, we can present our data in such a way that the important and informative patterns stand out. If we disobey the rules, our data will be incomprehensible or misleading.*
>
> (*Information Visualization: Perception for Design*, Second Edition, Colin Ware, Morgan Kaufmann Publishers, 2004, page xxi)

Perhaps the best known expert in data visualization, Edward Tufte, says: "Clear and precise seeing becomes as one with clear and precise thinking." (*Visual Explanations*, Edward R. Tufte, Graphics Press: Cheshire, CT.1997 page 53)



Visual input is fast.

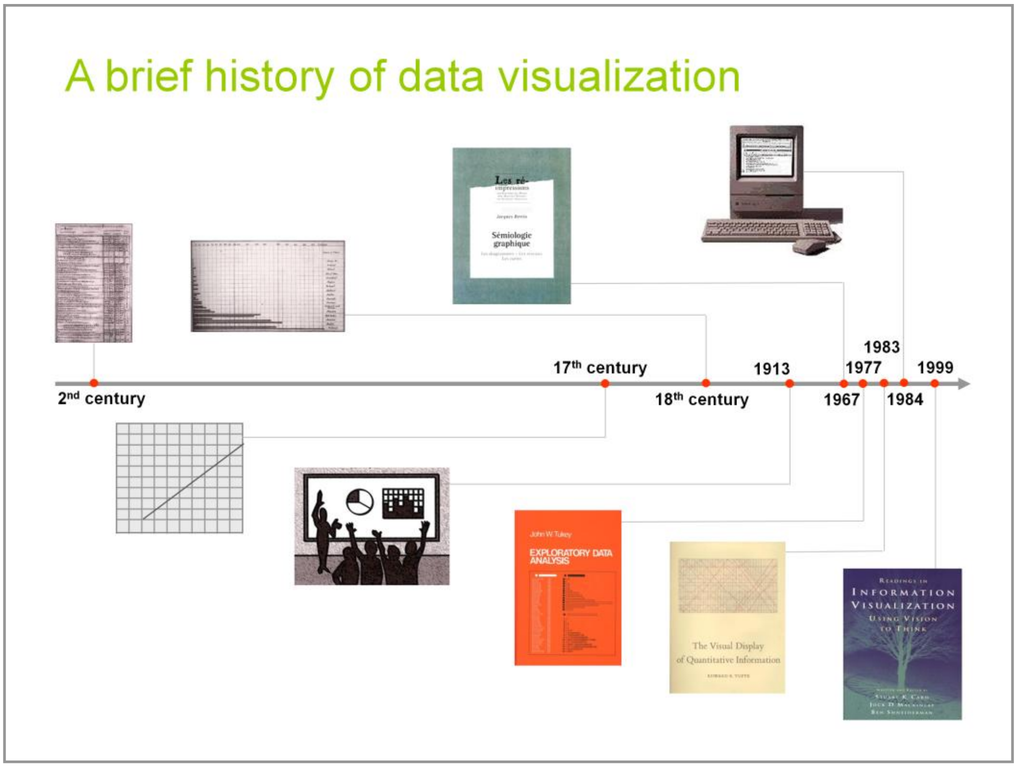
Stating that 70% of the bodies sense receptors reside in the eye perhaps tells an overly simplified story of what is going on. A slightly different picture unfolds when we compare the five senses in terms of total bandwidth (millions of bits of information per second), but vision still clearly dominates with 10 million bits per second versus 1 million for touch (via receptors in the skin), all the way down to only 1,000 bits per second of taste perception. To complete the picture of how each channel of sensory perception contributes to our overall awareness, however, we should consider only that part of perception of which we are conscious. From this perspective, vision still dominates with 40 bits per second of information, but hearing isn't far behind with 30, followed by touch with 5, and smell and taste with 1 each. Despite the reduced dominance of vision in the realm of consciousness, the fact that it dominates so dramatically in total sense receptors and bandwidths of perception means that it offers a vastly richer spectrum of what it can communicate tell us. (Source: Tor Norretranders, *The User Illusion: Cutting Consciousness Down to Size*, Viking Press, New York, 1998)



## Data Visualization

The use of visual representations to explore, make sense of, and communicate data.

I use the term *data visualization* as the umbrella term to describe all forms and uses of visual representations, which encode all types of data.



A brief history of data visualization

2nd century
17th century
18th century
1913
1967
1977
1983
1984
1999
Sémiologie graphique
John W. Tukey
EXPLORATORY DATA ANALYSIS
The Visual Display of Quantitative Information
Readings in Information Visualization Using Vision to Think

To fully appreciate what it offers, it is worthwhile to quickly trace the history of data visualization in general to see how far we've come and just how recent the opportunities introduced by information visualization are.

The tabular presentation of data has been with us since the 2nd century, when it was first used in Egypt to organize astronomical information and to aid navigation.

The representation of quantitative data in two-dimensional graphs didn't arise until much later, in the 17th century, when Rene Descartes, the French philosopher and mathematician famous for the words "Cogito ergo sum" ("I think therefore I am") invented the method, not originally for presenting data but for performing a type of mathematics based on a system of coordinates.

It wasn't until the late 18th and early 19th centuries that many of the graphs that we use today were invented or dramatically improved by a Scottish social scientist named William Playfair, including bar charts and pie charts.

Over a century passed, however, before recognition of the value of these techniques led to the first college course in graphical statistics, which was offered at Iowa State in 1913.

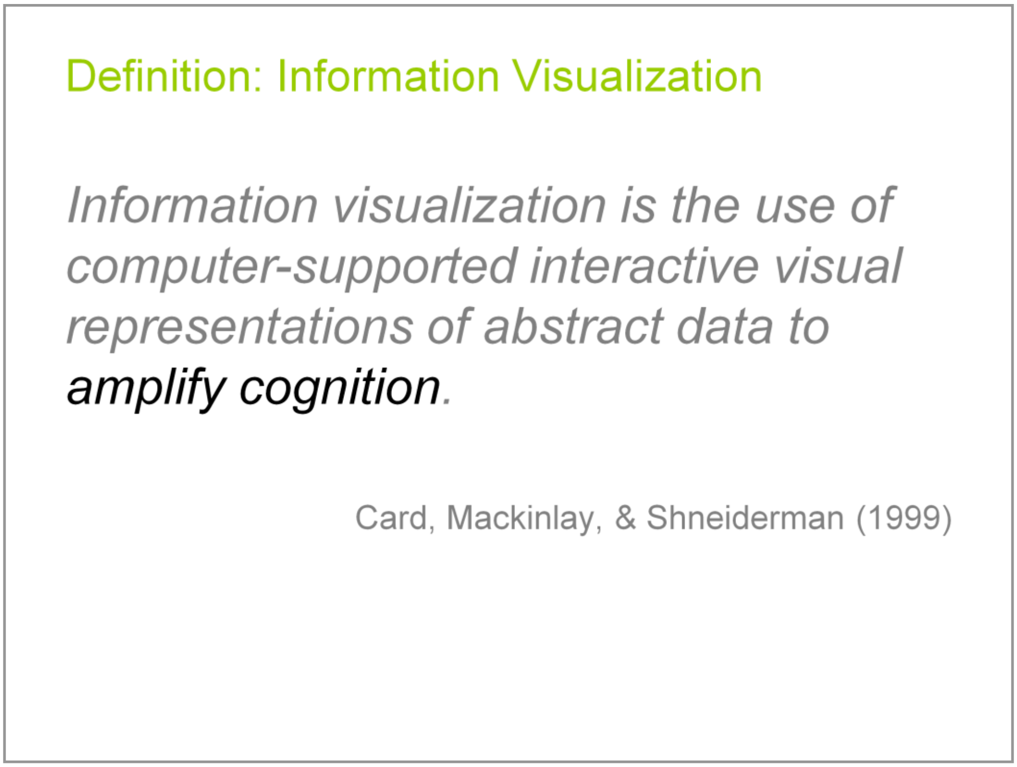
In 1967, with the publication of his book *Semiologie graphique*, Jacques Bertin introduced the notion of visual language—the fact that visual perception operated according to rules that could be followed to express information visually in ways that represented it intuitively, clearly, accurately and efficiently.

The person who really introduced us to the power of data visualization as a means of exploring and making sense of data was the statistics professor John Tukey of Princeton, who in 1977 gave form to a whole new approach to analyzing data called *exploratory data analysis*.

In 1983, data visualization guru Edward Tufte, came out with his groundbreaking book *The Visual Display of Quantitative Information*, which showed us that there were effective ways of displaying data visually and then there were the ways that most of us were doing it, which didn't communicate very well.

One year later, in 1984, while watching the Super Bowl, Apple Computer introduced us to the first popular and affordable computer that focused on graphics as a mode of interaction and display, a graphical user interface that was originally developed at Xerox PARC. This paved the way for the use of data visualizations that we could interact with directly on a computer.

Given the availability of affordable computers with powerful graphics, a new research specialty emerged in the academic world, which was coined "information visualization." In 1999 the book *Readings in Information Visualization: Using Vision to Think* collected this work into a single volume and made it accessible beyond the walls of academia.



## Definition: Information Visualization

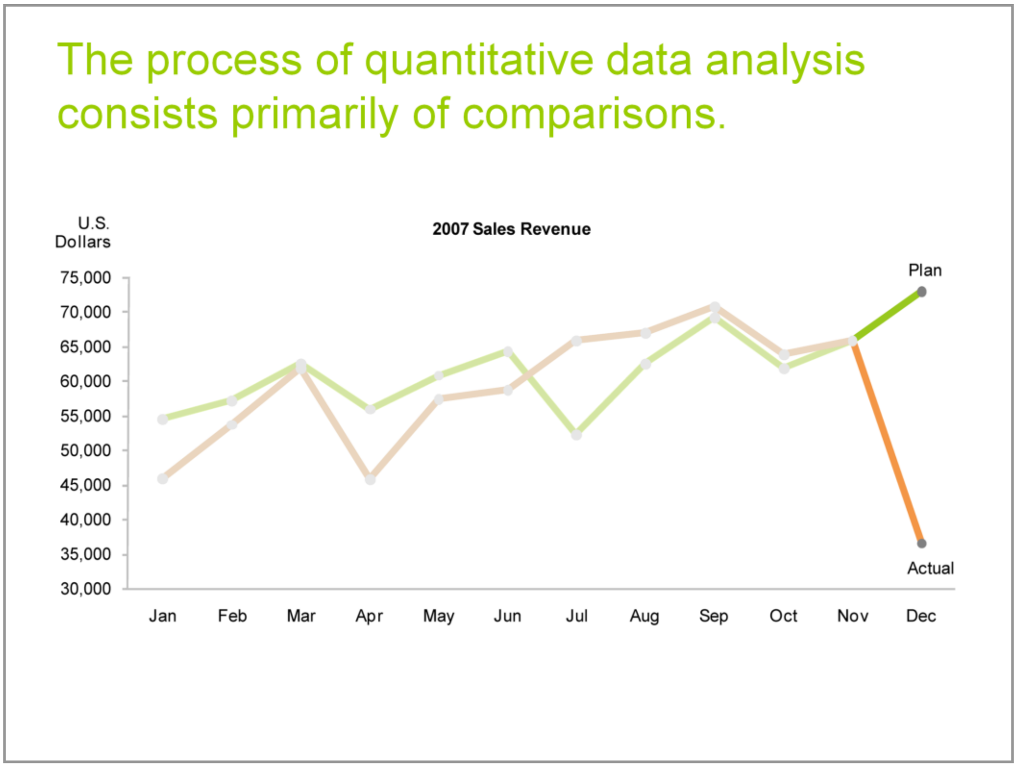
*Information visualization is the use of computer-supported interactive visual representations of abstract data to **amplify cognition**.*

Card, Mackinlay, & Shneiderman (1999)

The specialized sub-domain of data visualization called *information visualization* has been defined by the three authors of *Readings in Information Visualization: Using Vision to Think* (1999), Stuart Card, Jock Mackinlay, and Ben Shneiderman, as shown above.

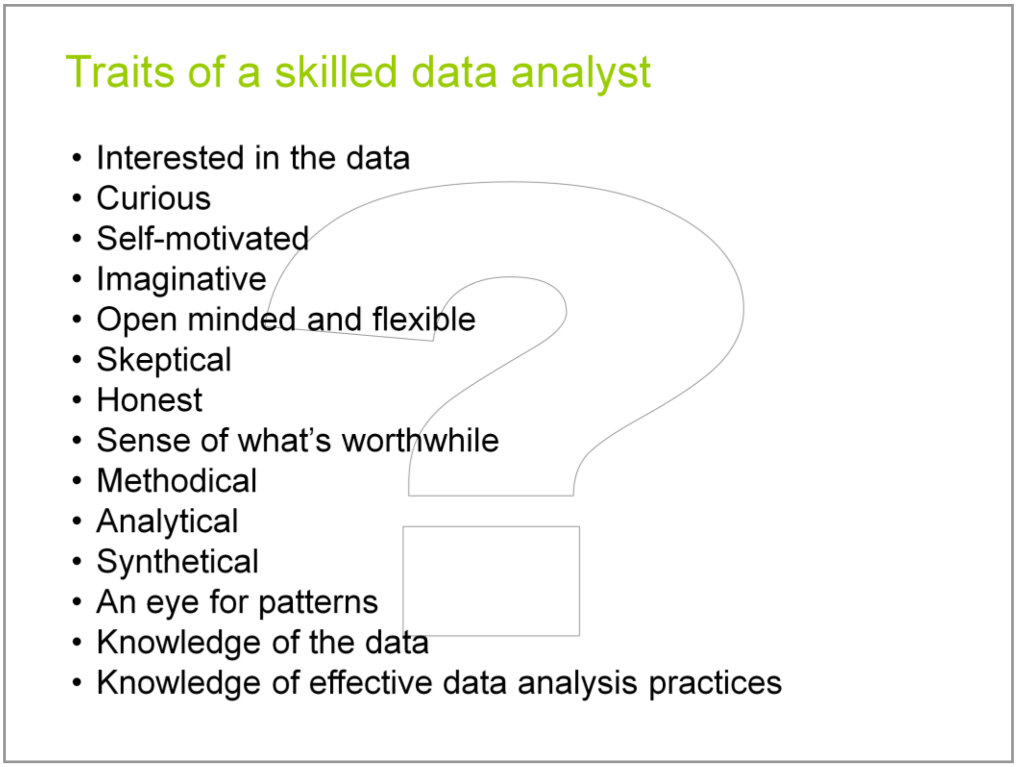
This definition features the following:

- *computer-supported* – The visualization is displayed by a computer, usually on a computer screen.
- *interactive* – The visualization can be manipulated directly and simply by the user in a free-flowing manner, including such actions as filtering the data and drilling down into details.
- *visual representations* – The information is displayed in visual form using attributes like location, length, shape, color, and size to make a picture of the data and thereby reveal patterns, trends, and exceptions that might not be seen otherwise.
- *abstract data* – Information such as quantitative data, processes, and relationships, as opposed to visual representations of physical objects, such as geography or the human body.
- *amplify cognition* – Interacting with these visualizations extends our ability to think by assisting memory and representing the information in ways that our brains can easily comprehend.



## The process of quantitative data analysis consists primarily of comparisons.

Visual analysis involves comparing the magnitudes of values, but not just one to another. We must compare many values. To do so, we must see how they relate to one another to form patterns. We not only compare the magnitudes of values; we also compare patterns formed by sets of values. We look for how they are similar and we look for how they are different, especially differences that appear to be dramatic departures from the norm. When we spot these visual characteristics in the data, we then interact with the data to find out why these things have happened.



## Traits of a skilled data analyst

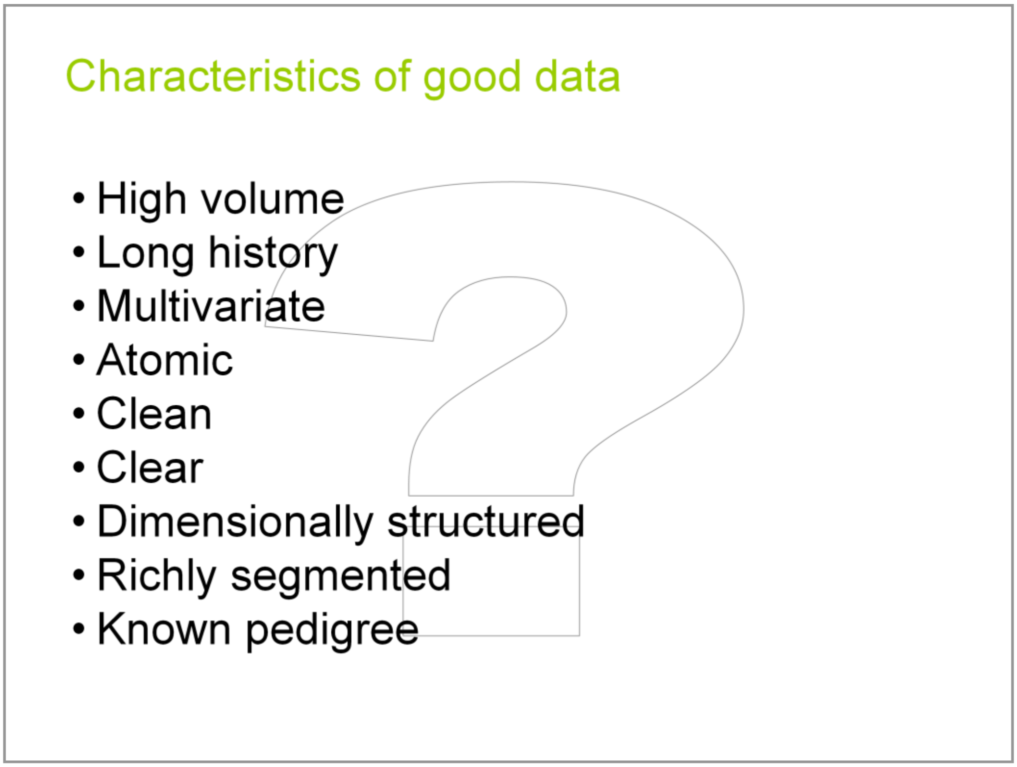
- Interested in the data
- Curious
- Self-motivated
- Imaginative
- Open minded and flexible
- Skeptical
- Honest
- Sense of what's worthwhile
- Methodical
- Analytical
- Synthetical
- An eye for patterns
- Knowledge of the data
- Knowledge of effective data analysis practices

Here are a few of the personal traits that support effective data analysis:

- Interested in the data
- Curious
- Self-motivated
- Imaginative
- Open minded and flexible
- Skeptical
- Honest
- Sense of what's worthwhile
- Methodical
- Analytical
- Synthetical (ability to see how pieces can fit together into a complex whole)
- Eye for patterns
- Knowledge of the data
- Knowledge of effective data analysis practices

The more that you naturally possess or work to acquire these attributes, the better you'll be at finding the important meanings in data.

> *Excellence in anything is the product of practice. That's especially true of quantitative reasoning, which doesn't come naturally to any of us.*
>
> (*What the Numbers Say: A Field Guide to Mastering Our Numerical World*, Derrick Niederman and David Boyum, Broadway Books, New York, 2003, page 5)

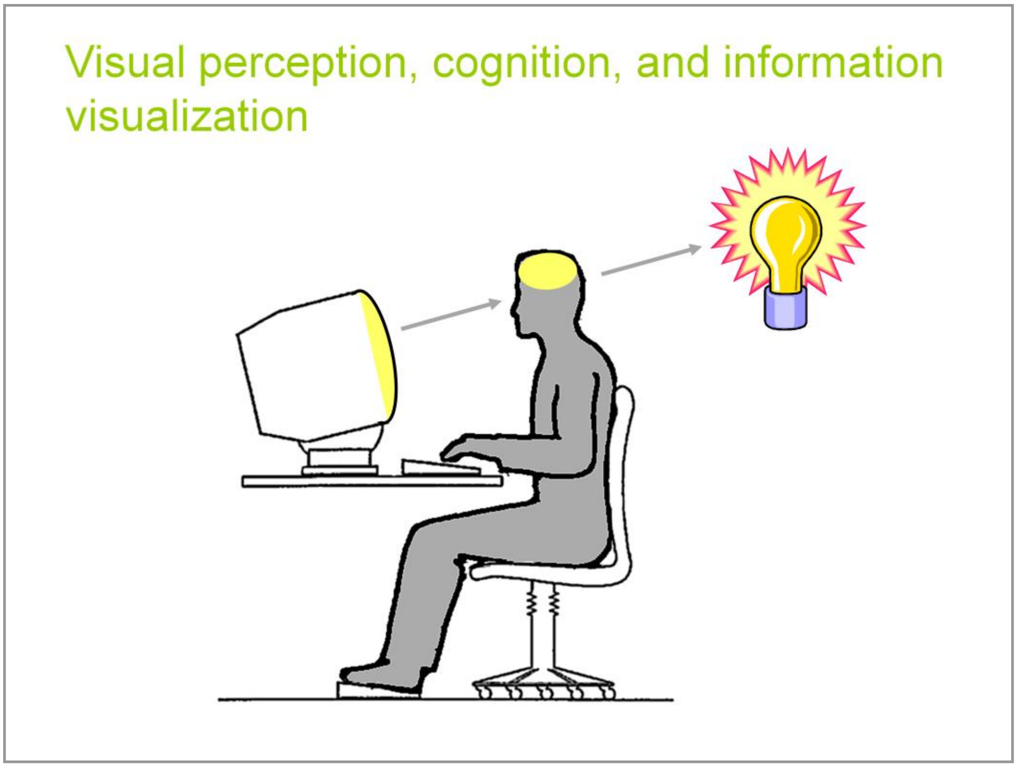


## Characteristics of good data

- High volume
- Long history
- Multivariate
- Atomic
- Clean
- Clear
- Dimensionally structured
- Richly segmented
- Known pedigree

Sometimes we have to work with the data we have, even though it isn't in ideal shape. If you have a say in the matter, however, these are the data characteristics that will provide the best potential for insight.

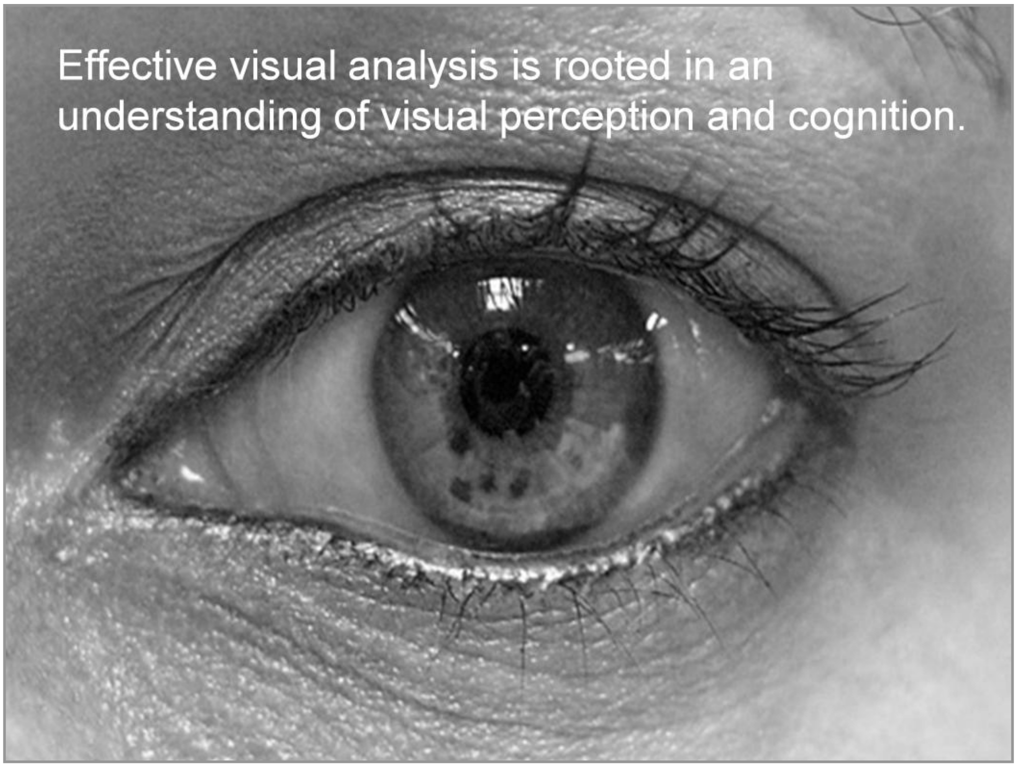
- **High volume**: Although you won't necessarily use it all, the more data that is available to you, the more chance there is that you'll have what you need.
- **Long history**: Much insight is gained from examining how data has changed through time. The more history that you have to examine, the more you can understand the meaning of those changes.
- **Multivariate**: Variables, both quantitative and categorical, are what we examine in the data. The more variables that are available to you, the richer potential for insight.
- **Atomic**: Most data analysis involves summarized data, but the ability to examine data at the lowest possible level of detail is sometimes needed  to understand what's going on.
- **Clean**: The quality of your analysis can never exceed the quality of the data. Data that is accurate, free of error, and complete is critical.
- **Clear**: If the data is expressed in cryptic codes that make no sense to you, it's meaningless. It's wonderful when someone else, like the data warehousing team, has already translated data that is difficult to interpret into clear terms.
- **Dimensionally structured**: Analysts' time is often eaten up trying to extract and pull data together from complex relationally-structured databases. If the data has already been structured according to good principles of dimensional data modeling, it is much easier to access and manipulate.
- **Richly segmented**: A great deal of useful analysis requires that sets of values for a particular variable (a.k.a., business dimension) be segmented into groups, such as customers by geographical regions.
- **Known pedigree**: It is important to understand your data's background, how it came to be, what system it came from, what calculations might have been used to create it, etc., in order to fully understand it.



## Visual perception, cognition, and information visualization

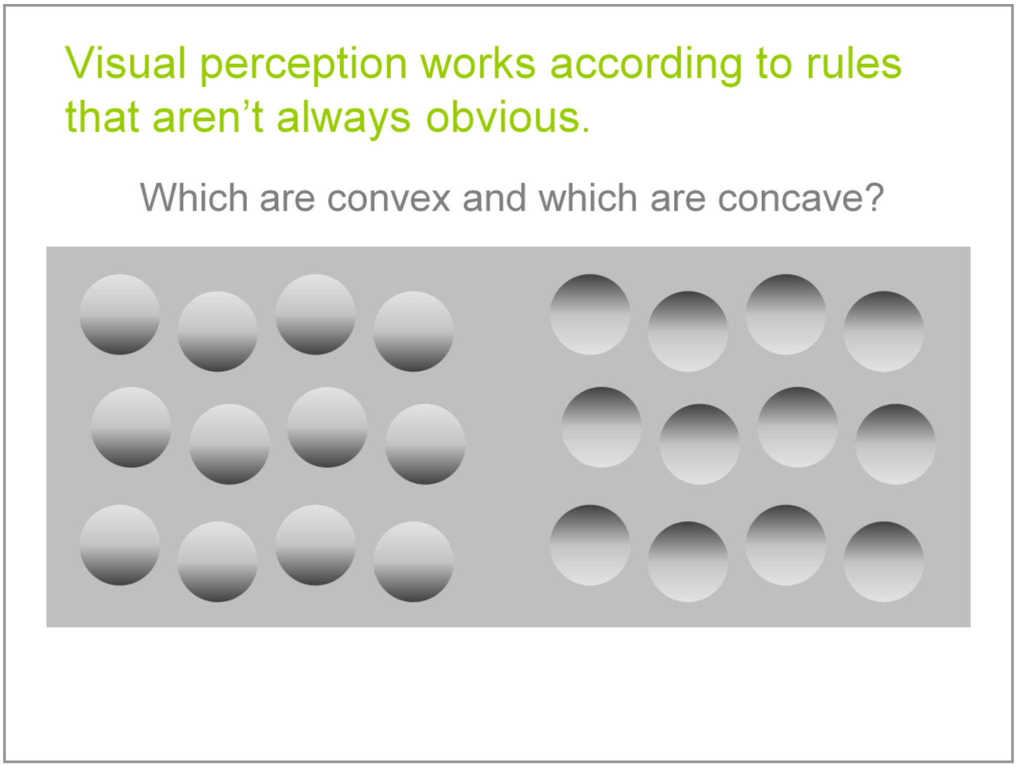
Even perfect data, however, can become difficult to understand if it is not presented to your eyes in a clear manner. Much of the software available for analyzing data does a poor job of presenting data. This is especially true of software that presents data visually in the form of graphs. If you know which graphs to use and how they should be designed to present your data clearly, however, you can make most software do the job with a little work.

> *When is one expression better than another for analysis? Basically, when the data are more simply described, since this implies easier and more familiar manipulations during analysis and, even more to the point, easier and more thorough understanding of the results.*
>
> (*The Collected Works of John W. Tukey*, John W. Tukey, Wadsworth, Inc.: Belmont, CA, 1988, page 12)



Effective visual analysis is rooted in an understanding of visual perception and cognition.

To know how to present information visually in an effective way, you must understand a little about visual perception – what works, what doesn’t, and why. We won’t delve into this deeply, but it is worth learning a bit about two aspects of visual perception that apply directly and powerfully to visual data analysis: the *pre-attentive attributes* of visual perception and the *limits of working memory*.

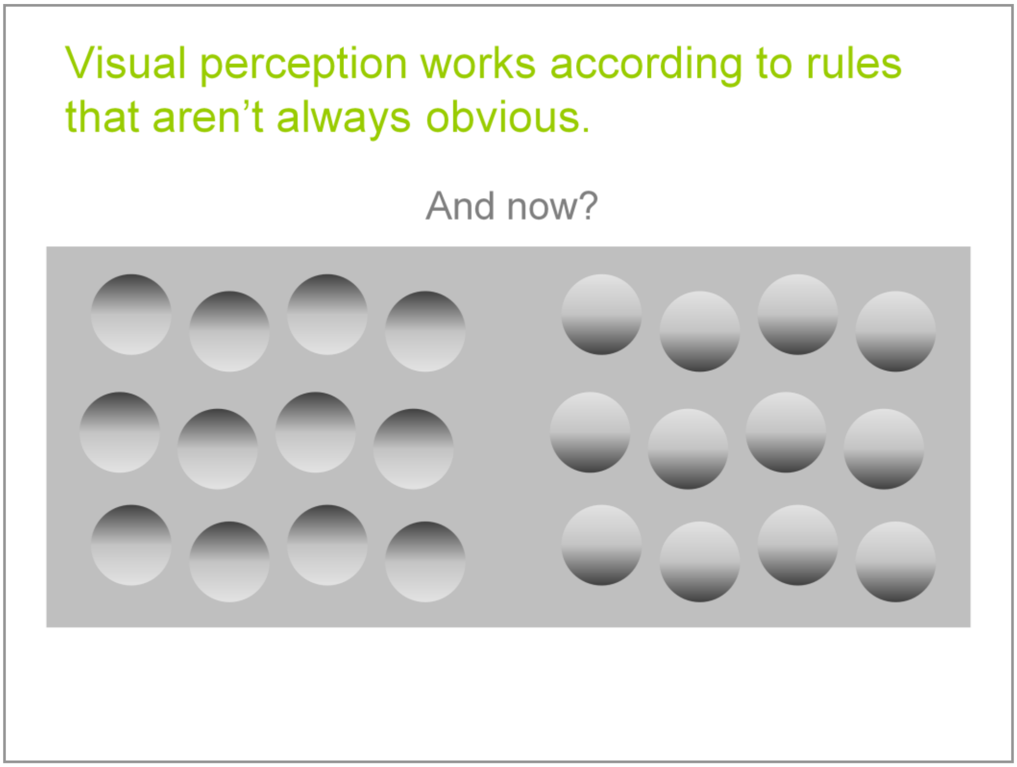


## Visual perception works according to rules that aren’t always obvious.

Which are convex and which are concave?

Many of the ways that visual perception work are not intuitive.

Looking at these two sets of objects, we naturally see those on the left as convex and on those the right as concave.



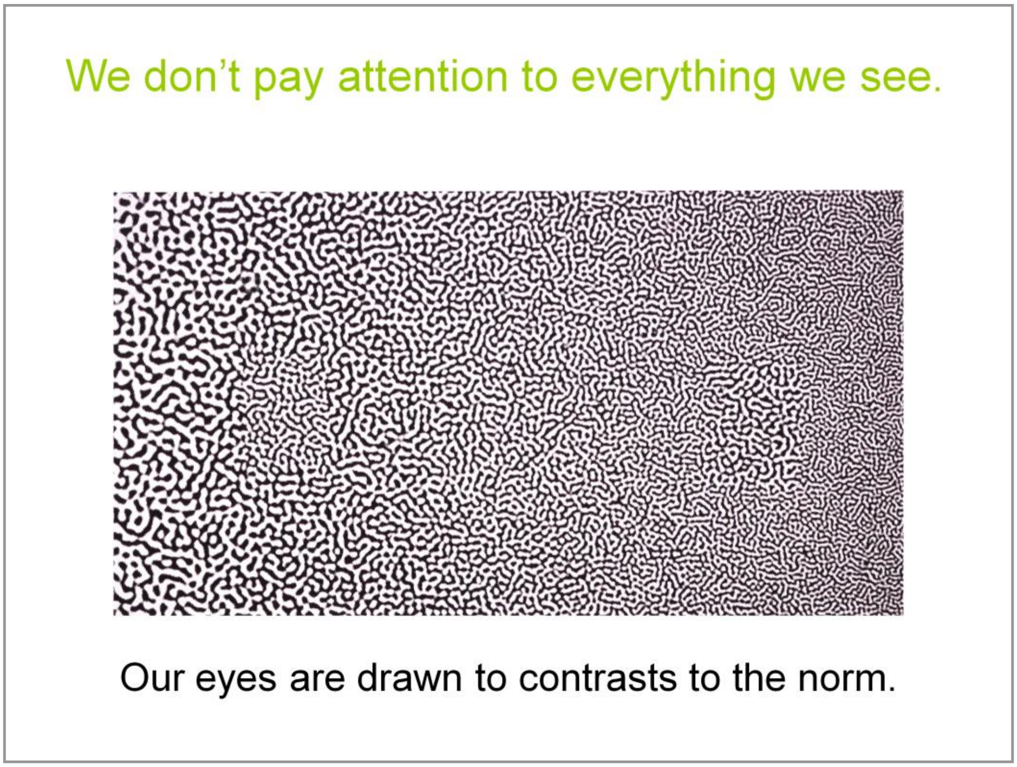
## Visual perception works according to rules that aren't always obvious.

And now?

The effect has now been reversed: we see the objects on the left as concave and those on the right as convex. All I did, however, was turn each sets of objects upside down—I didn't switch them. The reason that we now see those on the left as concave is because, through eons of evolution, visual perception learned to assume that light was shining from above, which causes us to see the objects on the left as concave, because the shadows are on the top, and those on the right as convex, because the shadows are on the bottom.

Remember the words of Colin Ware:

> *The visual system has its own rules. We can easily see patterns presented in certain ways, but if they are presented in other ways, they become invisible…The more general point is that when data is presented in certain ways, the patterns can be readily perceived. If we can understand how perception works, our knowledge can be translated into rules for displaying information. Following perception-based rules, we can present our data in such a way that the important and informative patterns stand out. If we disobey the rules, our data will be incomprehensible or misleading.*
>
> (*Information Visualization*, Second Edition, Colin Ware, Morgan Kaufmann Publishers, 2004, page xxi)



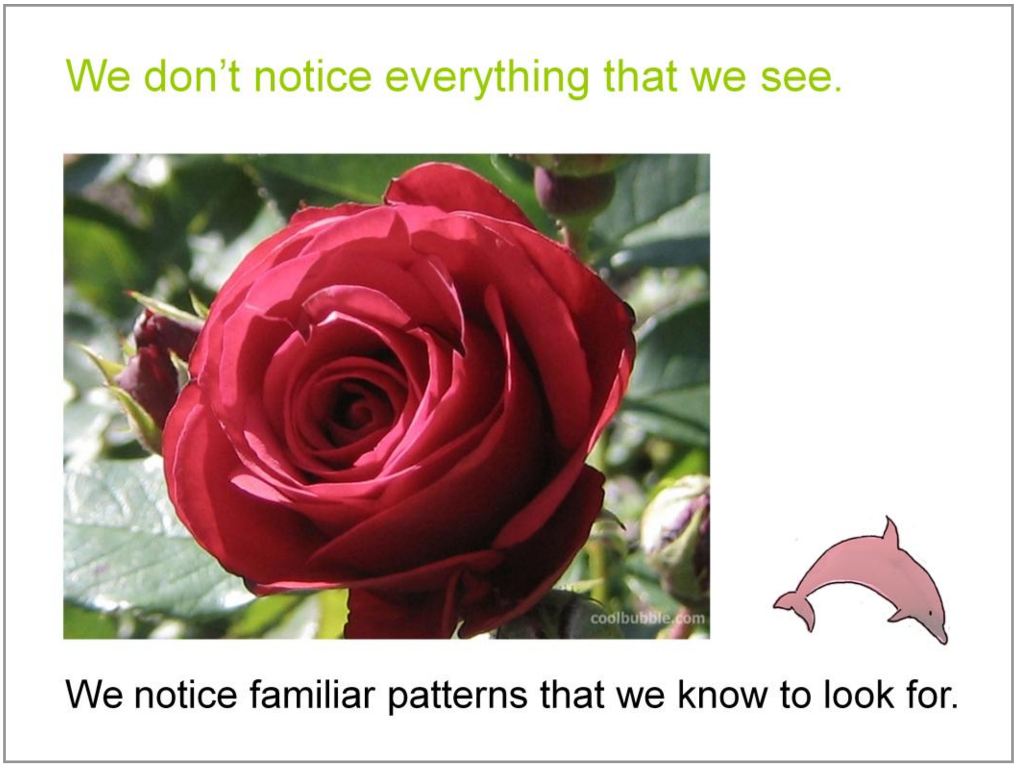
## We don't pay attention to everything we see.

Our eyes are drawn to contrasts to the norm.

We do not pay attention to everything in our field of vision. Visual perception is selective and must be, for an awareness of everything out there would overwhelm us. Attention tends to be drawn contrasts to the norm. For this reason, to successfully see meaning in the data, we must visually encode data in ways that allow what's interesting and potentially meaningful to pop out.

In the image above, the two sections of texture that stand out: one left of center and one right of center. What's not apparent is that these two regions in the image are exactly the same. They differ from what surrounds them because the lines that form the texture on the left are smaller than those that surround them, and those that form the texture on the right are larger than those that surround them.

(Note: This image appears in *Information Visualization: Perception for Design*, Second Edition, Colin Ware, Morgan Kaufmann Publishers: San Francisco, CA, 2004.)

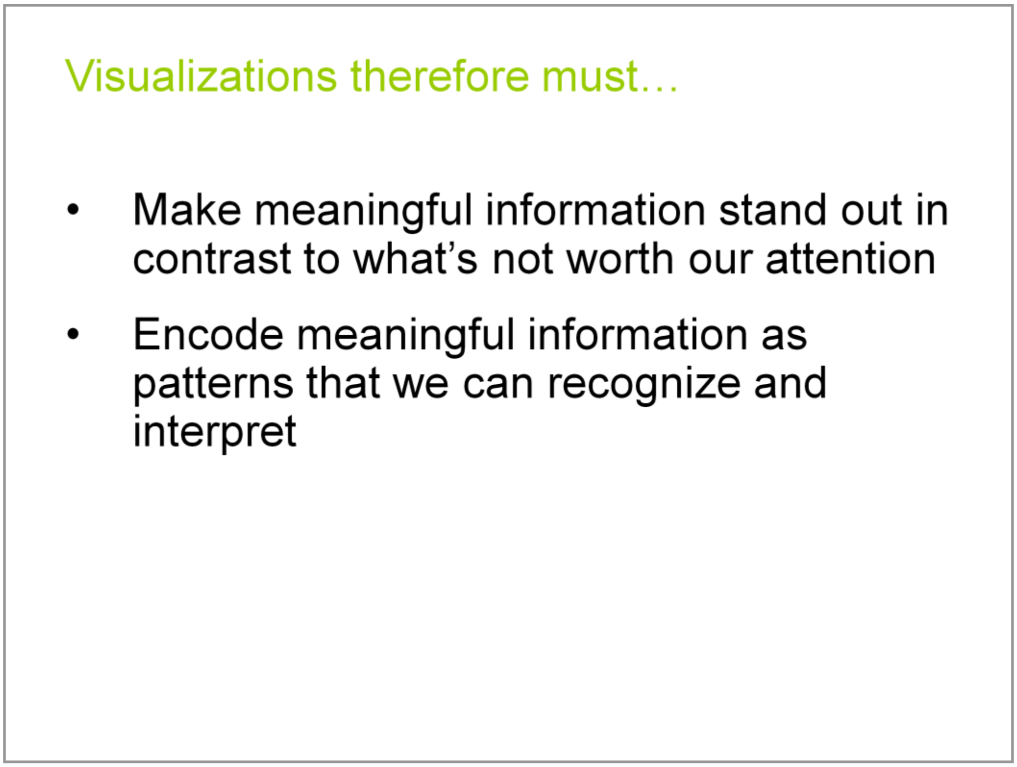


## We don’t notice everything that we see.

We notice familiar patterns that we know to look for.

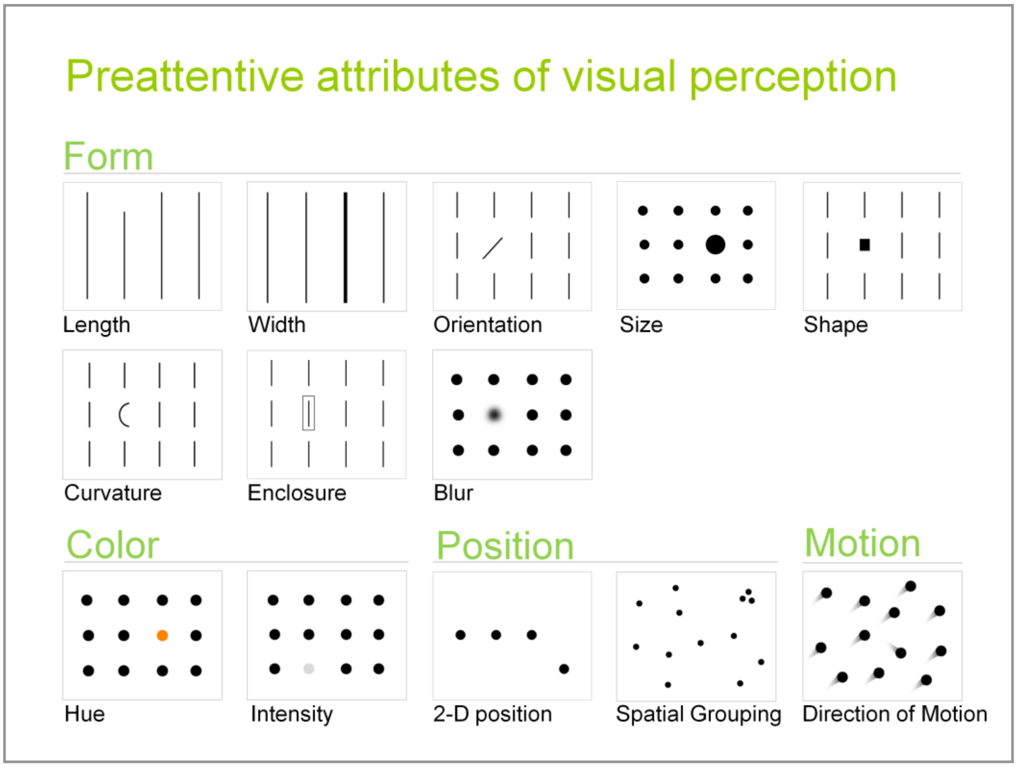
There is a distinct image that has been worked into the picture of the rose, which isn’t noticeable unless we know to look for it. Once primed with the image of the dolphin, however, we can easily spot it in the rose.

(Note: The image of the rose was found at www.coolbubble.com.)



## Visualizations therefore must…

- Make meaningful information stand out in contrast to what's not worth our attention
- Encode meaningful information as patterns that we can recognize and interpret



## Preattentive attributes of visual perception

### Form

Length
Width
Orientation
Size
Shape
Curvature
Enclosure
Blur

### Color

Hue
Intensity

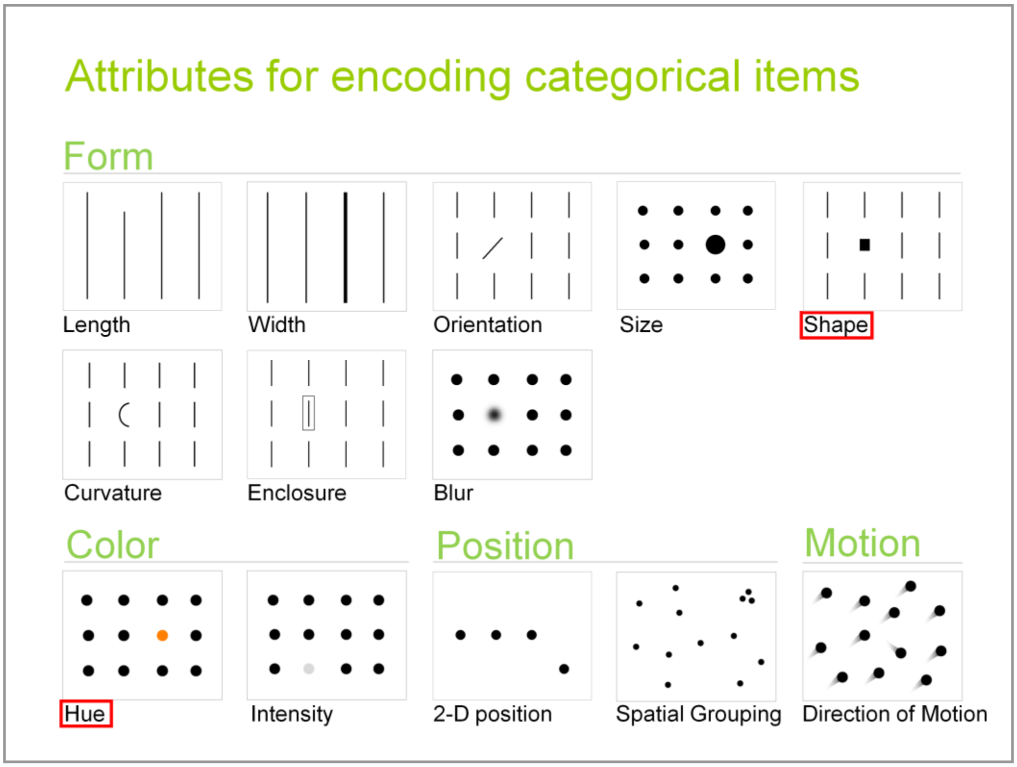
### Position

2-D position
Spatial Grouping

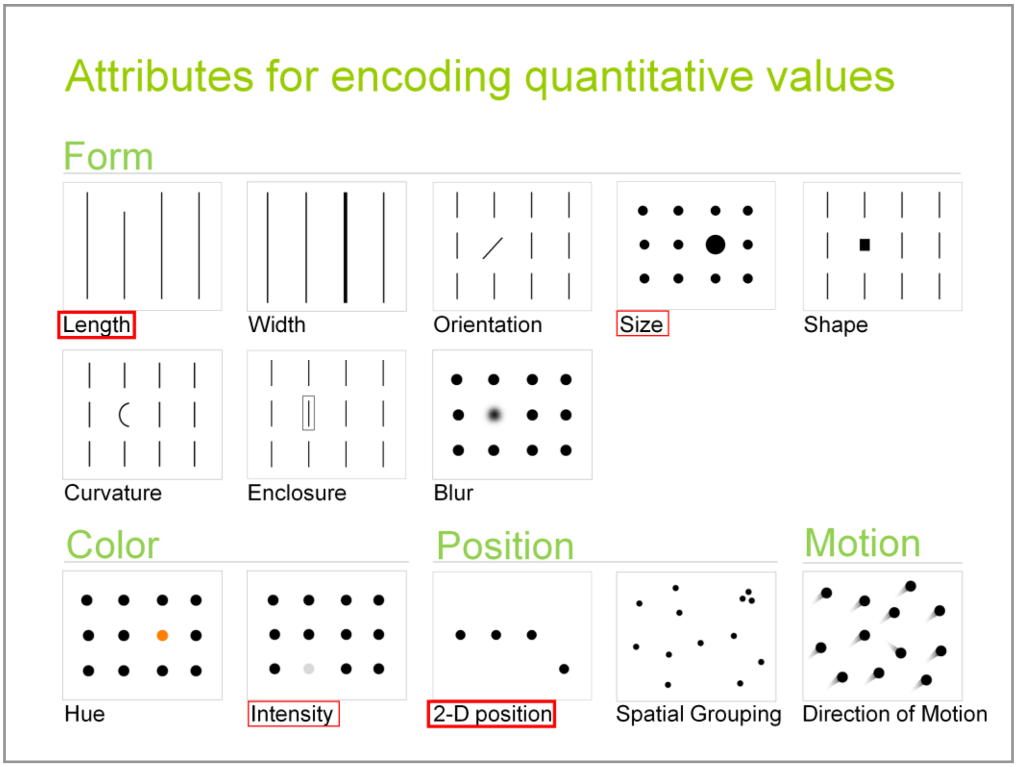
### Motion

Direction of Motion

The full list of visual attributes that we perceive preattentively is larger than the one above. These preattentive attributes, however, are the ones that are most useful for information visualization.



The preattentive attribute that provides the best means to visually distinguish data sets in a graph is hue. Simple shapes like squares, circles, triangles, diamonds, X's, and +'s (plus signs) can also be used, but they don't work as well as distinct colors. Distinct hues work best of all as long as you're not color blind.



Attributes for encoding quantitative values

Form

Length

Width

Orientation

Size

Shape

Curvature

Enclosure

Blur

Color

Hue

Intensity

Position

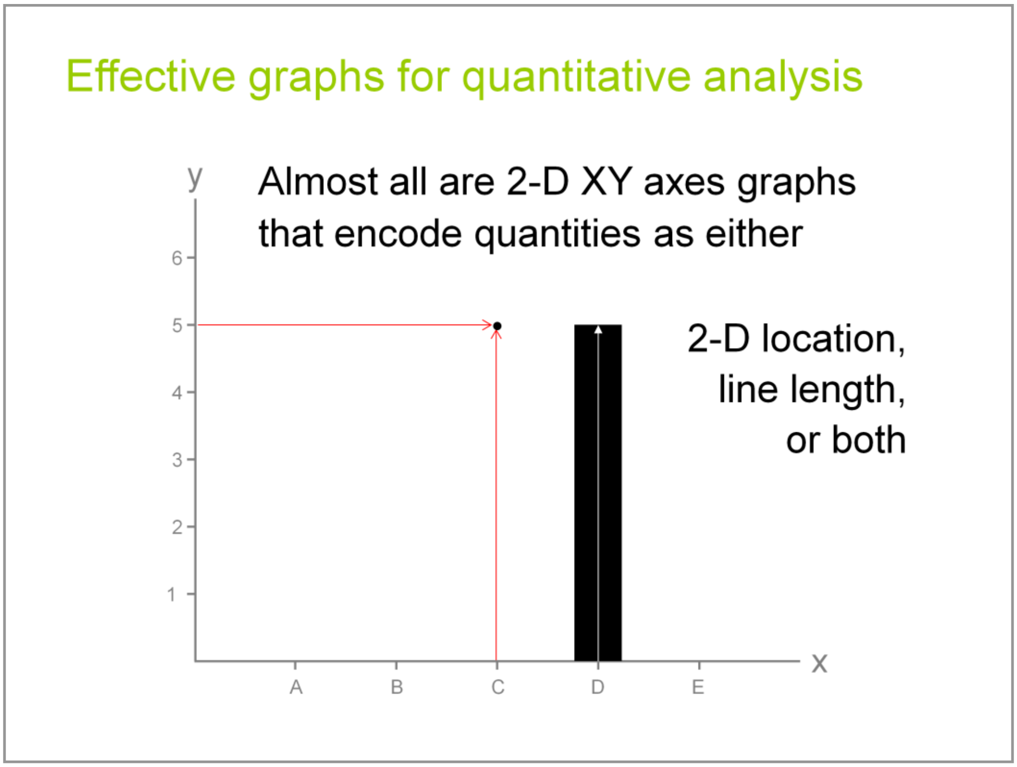
2-D position

Spatial Grouping

Motion

Direction of Motion

Some of these preattentive visual attributes are perceived quantitatively (i.e., some values are greater than others), which are marked with red boxes. The visual attributes that are marked with pale red boxes are perceived quantitatively but not as powerfully as length and 2-D position.

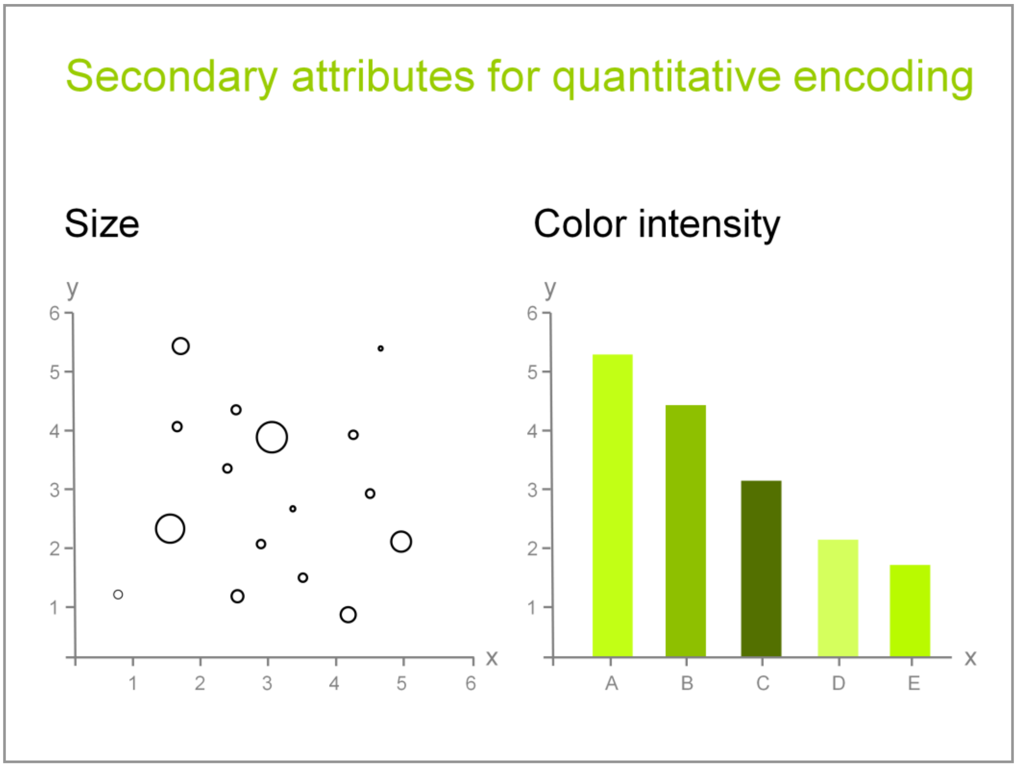


## Effective graphs for quantitative analysis

Almost all are 2-D XY axes graphs that encode quantities as either

2-D location,
line length,
or both

Despite the fact that software that is used for visual data analysis usually includes a broad assortment of graph types, only a few work well for the analysis of typical quantitative business data. All the useful graphs are of the 2-D XY type and use either the 2-D position of a data object (data points, data points along a line, and the endpoints of bars) or line length (length of bars) in relation to the quantitative access to encode quantitative values.



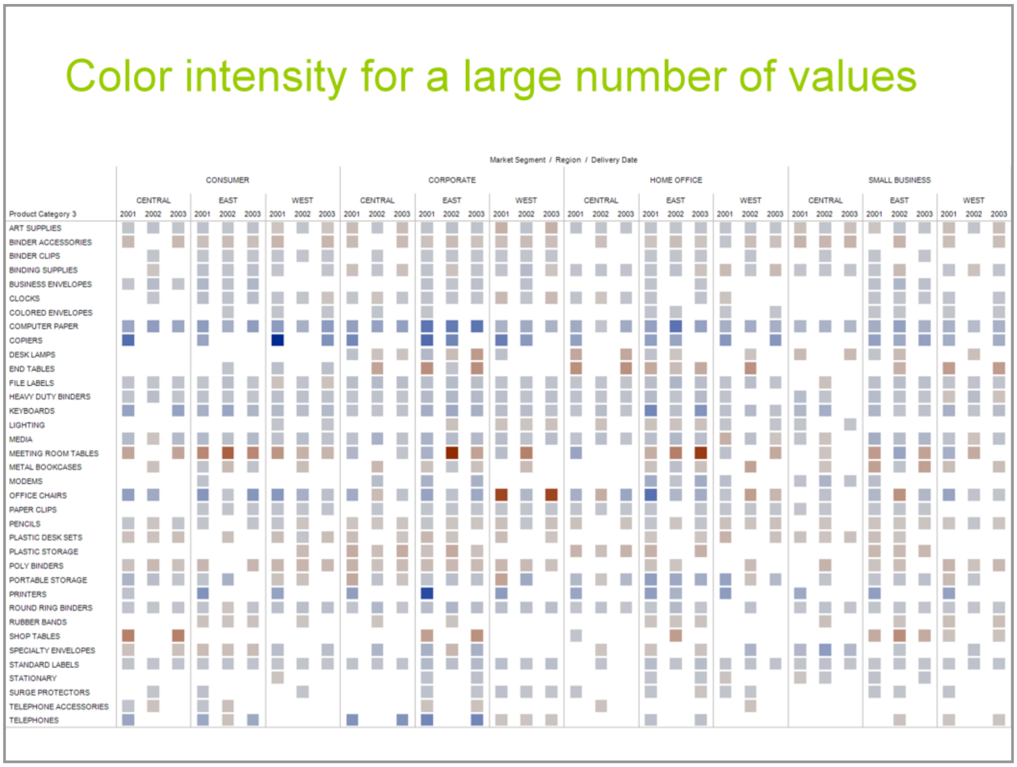
Secondary attributes for quantitative encoding

Size

Color intensity

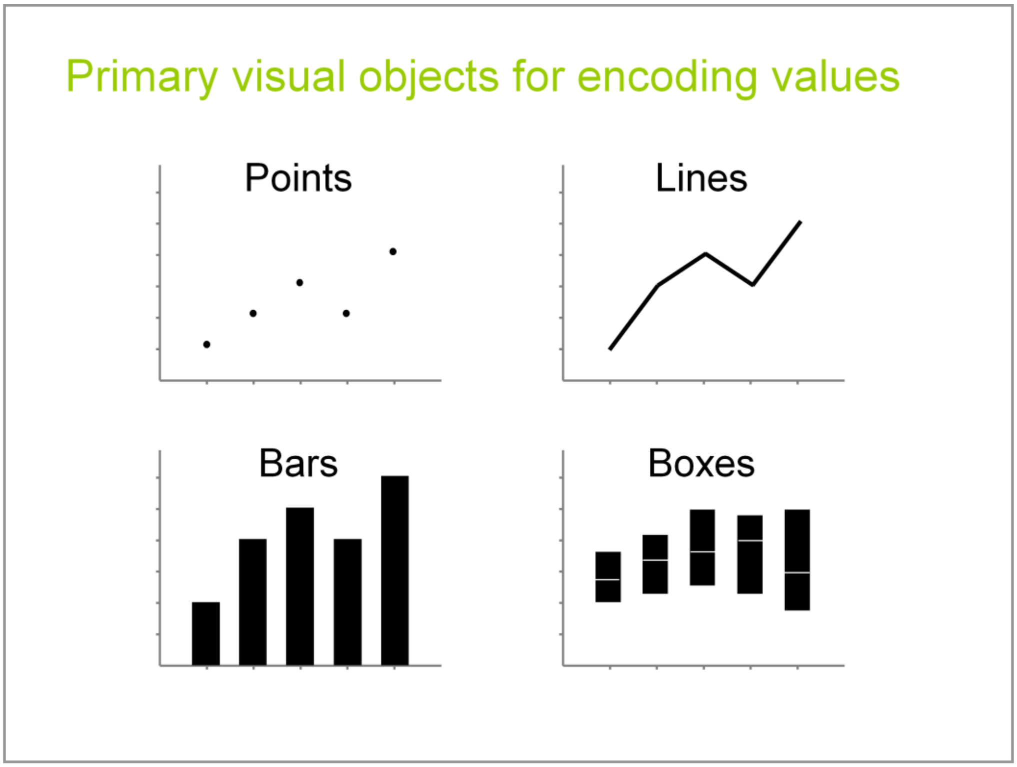
In the scatter plot on the left, the size of the data point is being used to encode a third quantitative variable, because 2-D position along the X and Y axes has already been used. As you can see, precise magnitude comparisons based on the sizes of these circles are not possible, but you can roughly perceive these differences in value.

In the bar graph on the right, two quantitative variables have been encoded in the bars: one as the heights of the bars in relation to the Y axis and a second as the color intensity of the bar ranging from light green for low values and dark green for high values. If you imagine that the heights of the bars encode revenues for various sales regions and that color intensity encodes profits, a graph such as this could be used to determine that the region with the greatest revenue falls somewhere near the center in terms of profits, whereas the third ranking region in revenues earns the highest profits. Precisely how much greater the profits are for region C compared to region A, however, you cannot determine, but if your objective doesn't require this precision, a graph of this type can be quite useful.



## Color intensity for a large number of values

When you need to get a birds-eye view of a large number of values to spot extremes, exceptions, concentrations, or gaps, heatmap matrices such as this (that is, a tabular arrangement of data using color intensity to encode the quantitative values) allow you to display a large number of values in the limited space of your computer screen.



## Primary visual objects for encoding values

Points

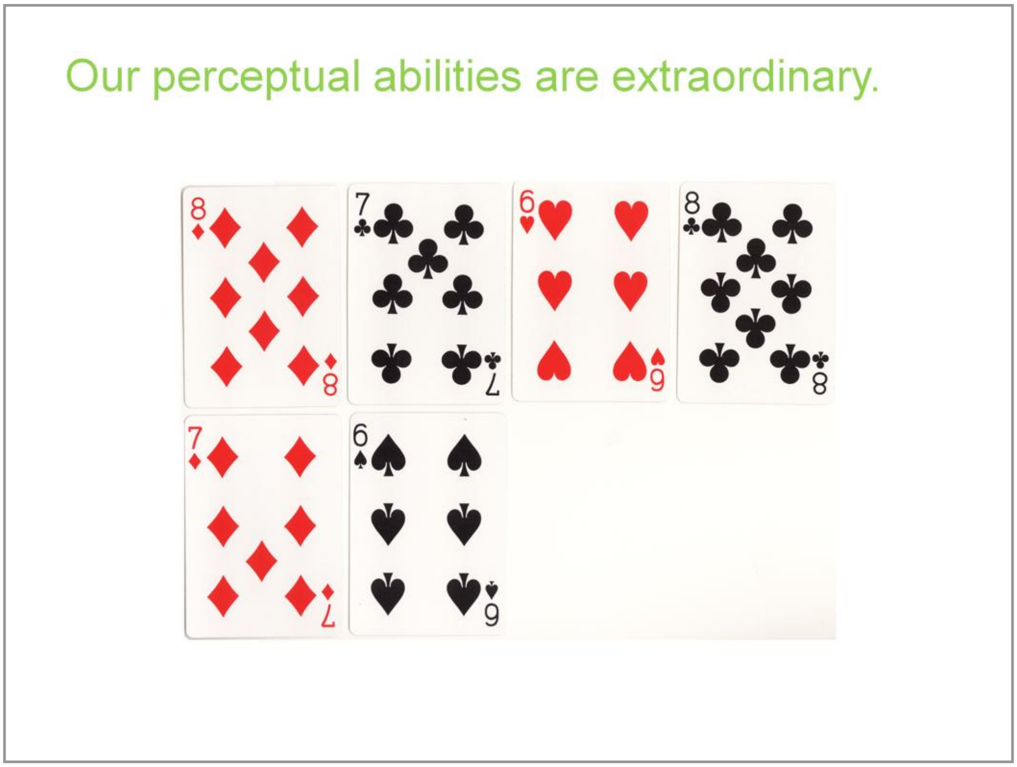
Lines

Bars

Boxes

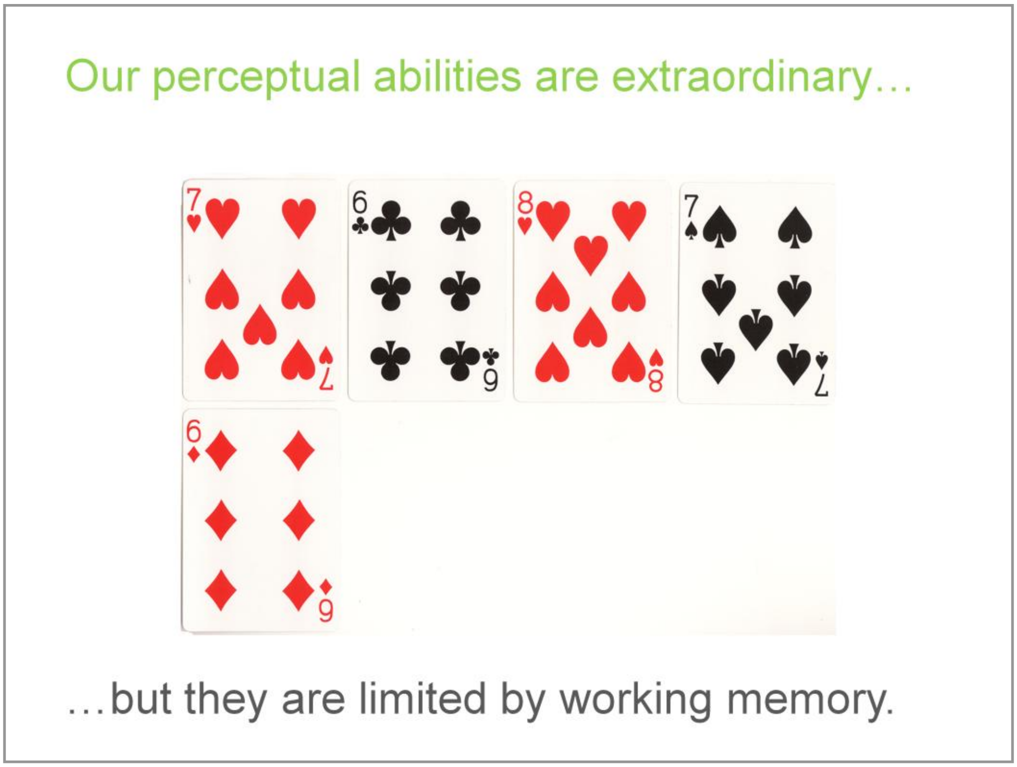
For typical quantitative data analysis, only four objects are needed to encode data in graphs:

- Data points (small circles, squares, triangles, etc.)
- Lines (with or without data points on them to mark the values)
- Bars
- Boxes (similar to bars, but used to display distributions of values from the lowest to the highest and usually points of interest in between)



## Our perceptual abilities are extraordinary.

From this set of six playing cards, select one and remember it. I will now identify and remove the card that you've selected, then rearrange those that remain. As I advance to the next slide, you'll discover that your card has been eliminated.

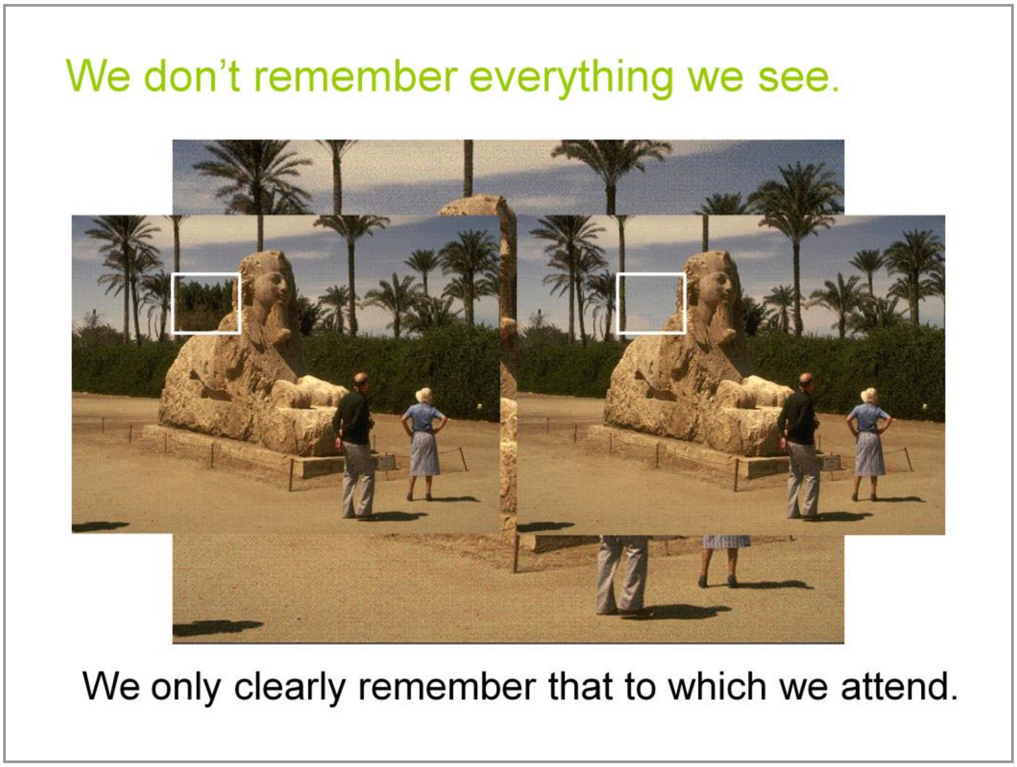


Our perceptual abilities are extraordinary…

…but they are limited by working memory.

Amazing. And I can do this again and again. If you go back to the previous slide and again pick a card, when you return to this slide you will see that I've once again eliminated it.

Actually, as I'm sure you realize, this card trick is an illusion that makes use of the limitations of working memory. None of the cards on the second screen are the same as the cards on the first screen, but you probably didn't notice this because you only remembered the card that you selected, not the others.



We don’t remember everything we see.

We only clearly remember that to which we attend.

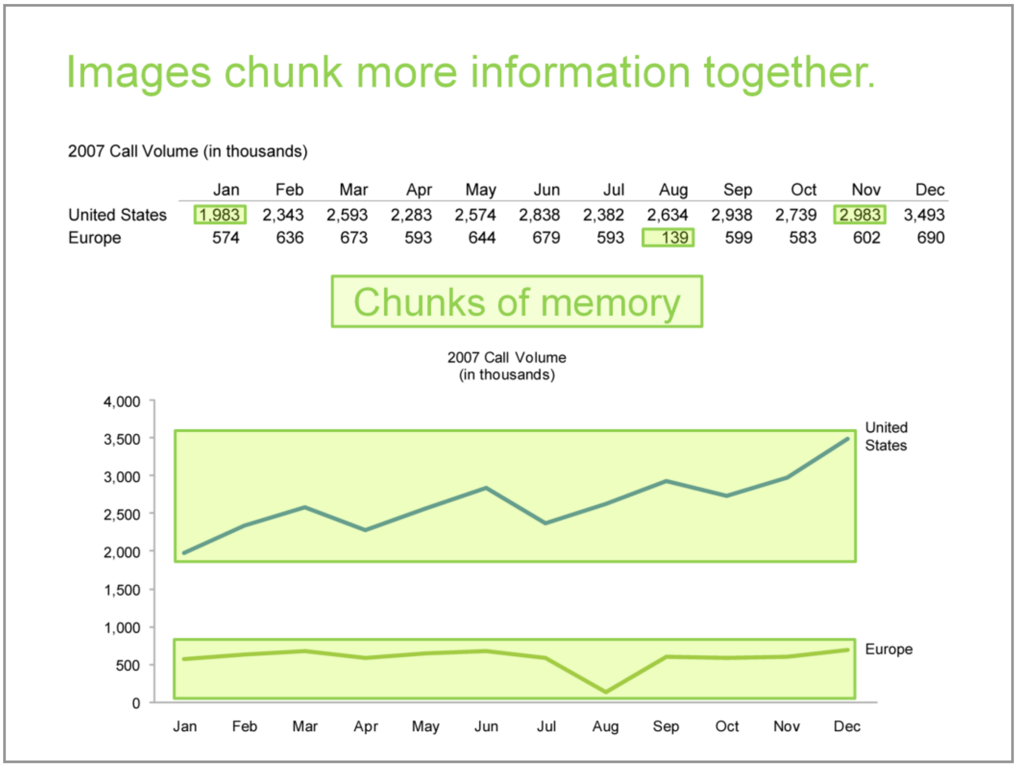
In addition to understanding visual perception, visual analysis tools must also be rooted in an understanding of how people think. Only then can they recognize and support the cognitive operations that are necessary to make sense of information.

Memory plays an important role in human cognition. Because memory suffers from certain limitations, visual analysis tools must be able to augment memory.

The example above illustrates one of the limitations of working memory. We only remember that to which we attend. Any part of this image that never gets our attention will not be missed when we shift to another version of the image that lacks that particular part. If we don’t attend to it, we might notice the change from one version of the image to the next, but only if the transition shift immediately from one to another, without even a split second of blank space between them.

In addition to not remembering, we also don’t clearly see that on which we don’t focus. To see something clearly, we must focus on it, for only a small area of receptors on the retinas of our eyes are designed for high-resolution vision.

(Source: This demonstration of change blindness was prepared by Ronald A. Rensink of the University of British Columbia. Several other examples of this visual phenomenon can be found at http://www.psych.ubc.ca/%7erensink/flicker/download/index.html.)



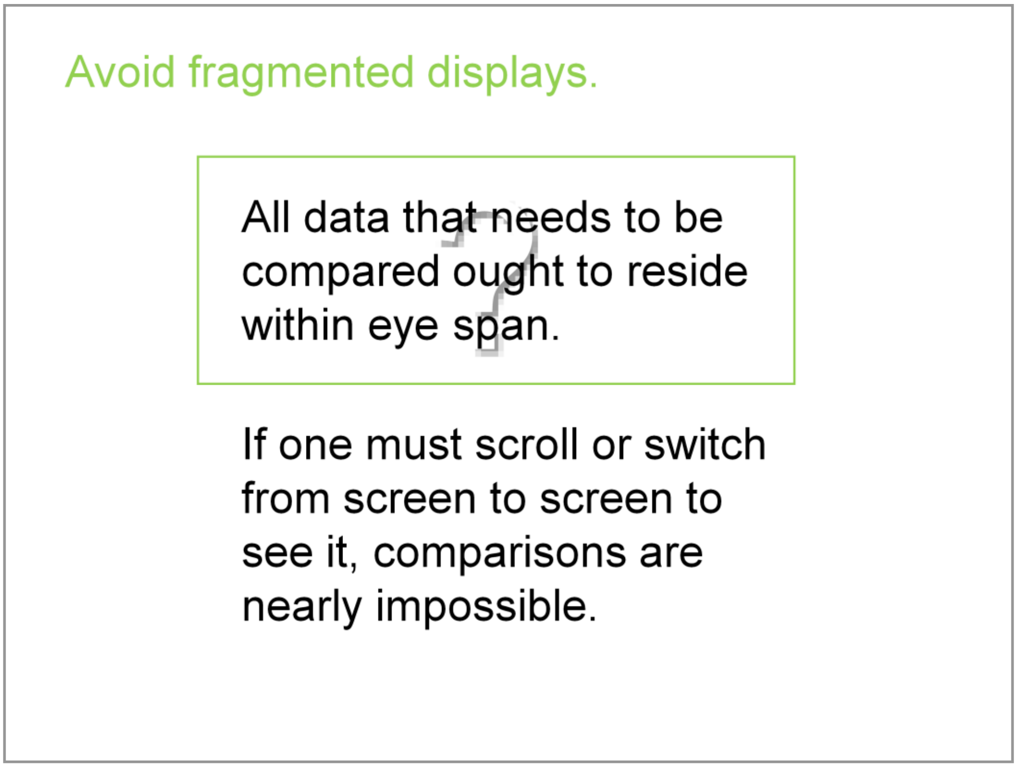
## Images chunk more information together.

2007 Call Volume (in thousands)

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| United States | 1,983 | 2,343 | 2,593 | 2,283 | 2,574 | 2,838 | 2,382 | 2,634 | 2,938 | 2,739 | 2,983 | 3,493 |
| Europe | 574 | 636 | 673 | 593 | 644 | 679 | 593 | 139 | 599 | 583 | 602 | 690 |

Chunks of memory

Memories are stored as chunks of information. A chunk can be as small as a single tiny fact (for example, revenue equals $56,384 for the quarter) or a larger set of facts that you've learned to think about as a single complex unit (for example, a trend line on a time-series graph that shows revenue increasing from month to month throughout the year.) The better you get at seeing and understanding meaningful patterns and relationships in data, the better able you are to store more data as a single chunk. Working memory  is where information is stored while we are thinking about something. It is like the working memory, or RAM, in a computer. Our brains are constantly swapping chunks of information in and out of working memory from either what we perceive in the outside world or from the more permanent storage of long-term memory. There is a limit to the amount of information that can be held in working memory at any one time, which is estimated by researchers to be about four chunks.

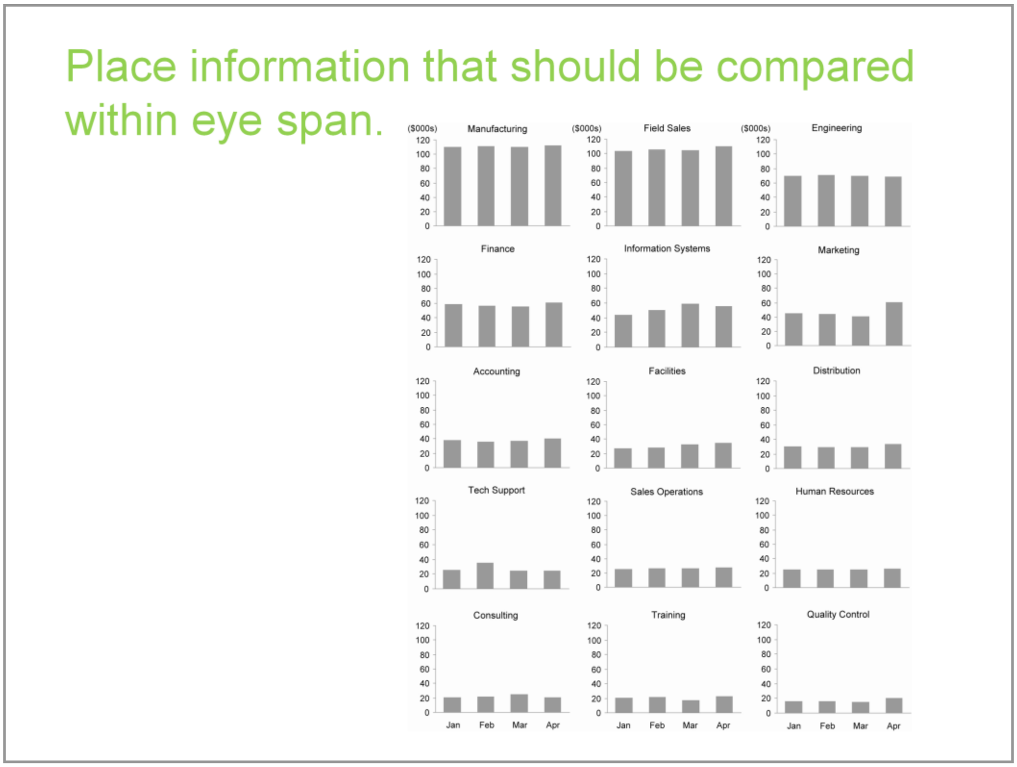


## Avoid fragmented displays.

All data that needs to be compared ought to reside within eye span.

If one must scroll or switch from screen to screen to see it, comparisons are nearly impossible.

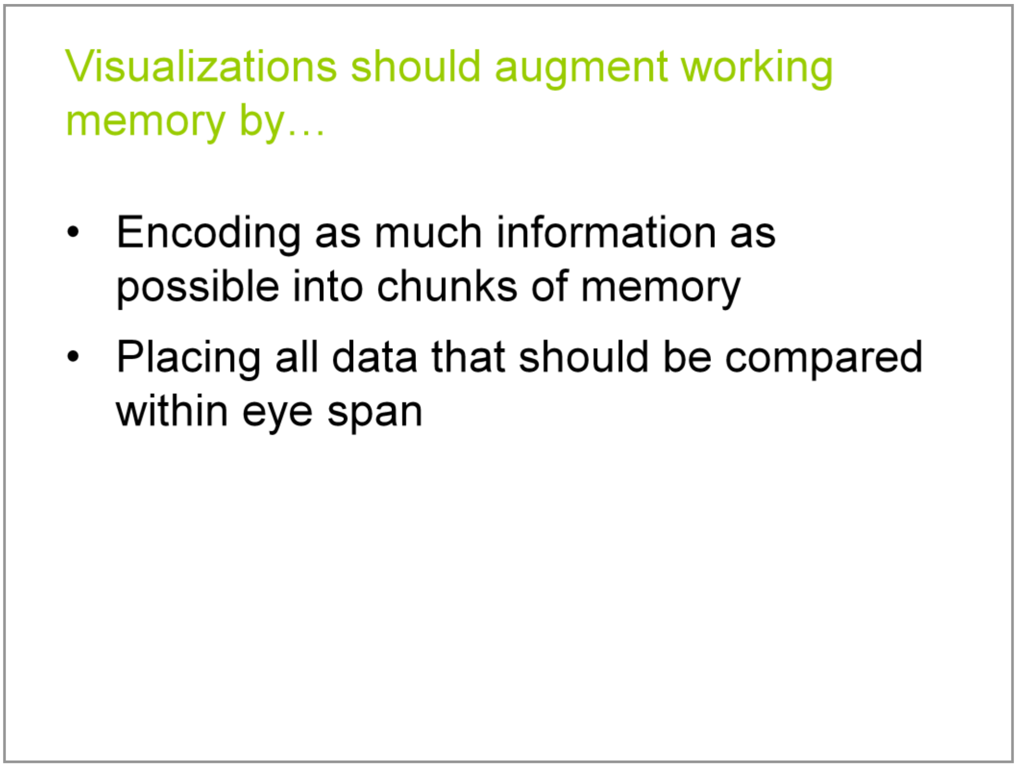
It is very difficult with most software to combine all of the information that you want to see together on a single screen without needing to scroll. You often end up bouncing from screen to screen to see separately what you would ideally like to see together in order to make comparisons and get a sense of the big picture.



## Place information that should be compared within eye span.

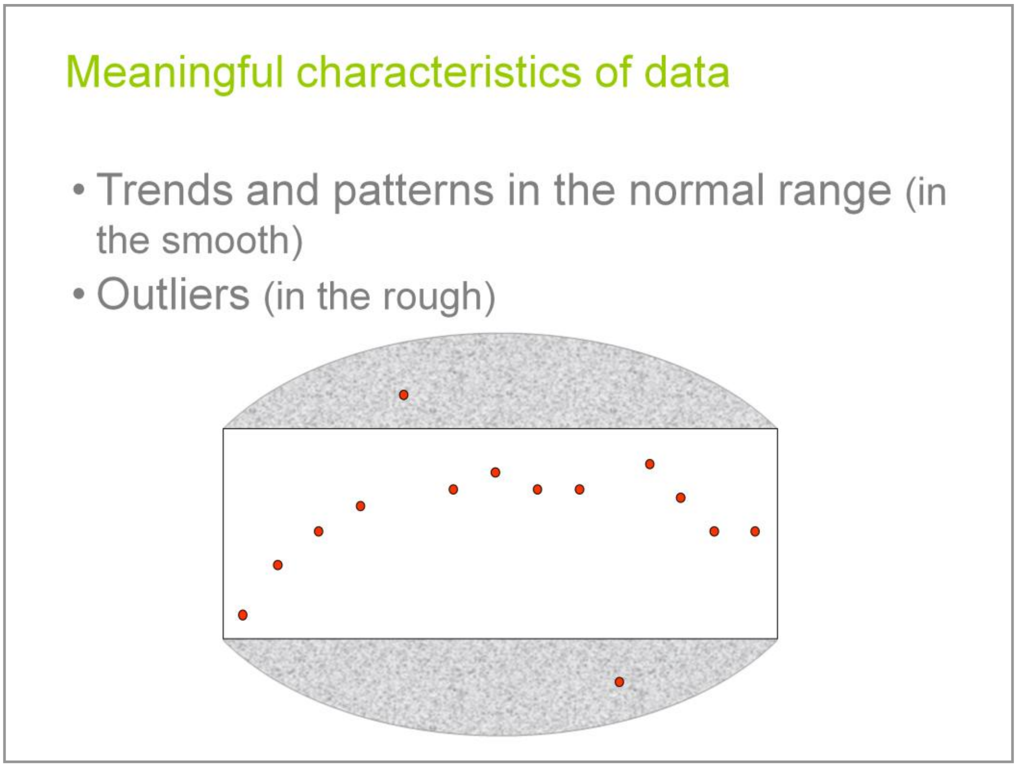
Now, however, with expenses for 15 separate departments visible at the same time, this display serves as an external aid to working memory, making it easy to make comparisons.

When exploring and examining data, it is important to place as much as possible within eye span. If you see patterns in a graph and then try to compare them to patterns in another graph on a different screen, you won’t remember everything that you were looking at previously. You’ll end up bouncing back and forth between displays, wasting time and getting very frustrated in the process.



## Visualizations should augment working memory by…

- Encoding as much information as possible into chunks of memory
- Placing all data that should be compared within eye span

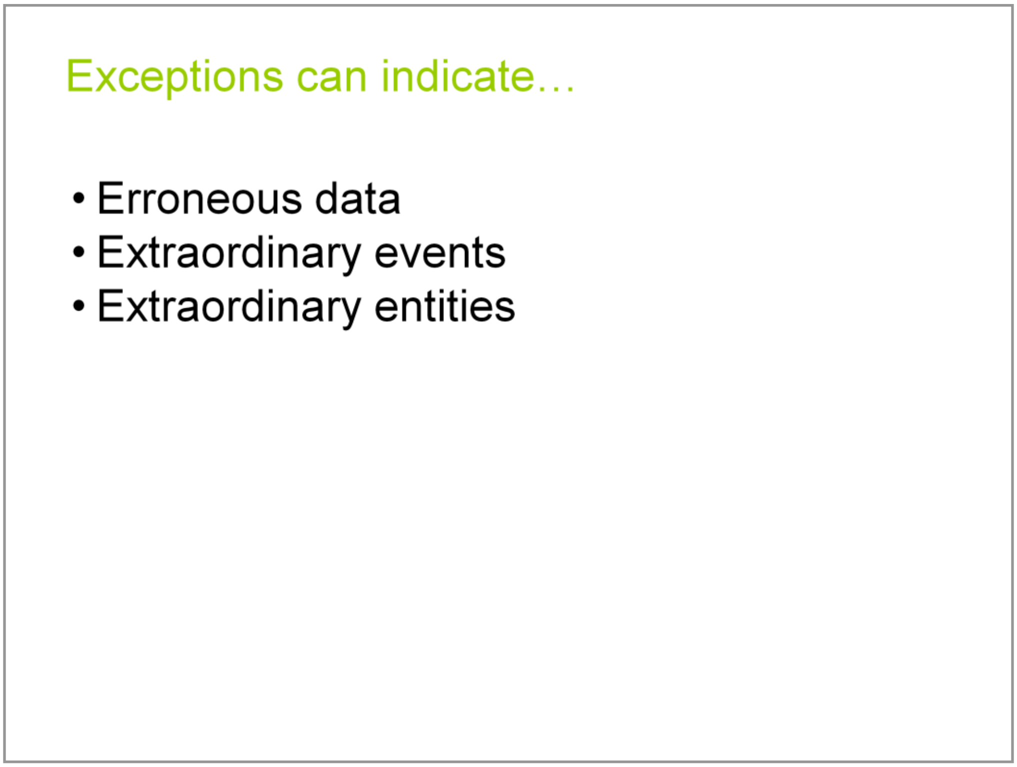


## Meaningful characteristics of data

- Trends and patterns in the normal range (in the smooth)
- Outliers (in the rough)

The terms *smooth* and *rough* come from the domain of *exploratory data analysis*. In the smooth refers to values in the normal range, the range that represents values that are typical. In the rough refers to values that reside outside the typical range, which are also called outliers.

> *The smooth is the underlying, simplified structure of a set of observations. It may be represented by a straight line describing the relationship between two variables or by a curve describing the distribution of a single variable, but in either case the smooth is an important feature of the data. It is the general shape of a distribution or the general shape of a relationship. It is the regularity or pattern in the data.*
>
> *Since the data will almost never conform exactly to the smooth, the smooth must be extracted from the data. What is left behind is the rough, the deviations from the smooth, the difference between the smooth and the observed data points.*
>
> (*Exploratory Data Analysis*, Frederick Hartwig with Brian E. Dearing, Sage Publications, Inc.: Thousand Oaks, CA, 1979, pages 10 and 11)
>
> *An outlier is a value which lies outside the normal range of the data, i.e., lies well above or well below most, or even all, of the other values… It is difficult to say at just what point a value becomes an outlier since much depends upon its relationship to the rest of the data and the use for which the data is intended. One may want to identify and set aside outlying cases in order to concentrate on the bulk of the data, but, on the other hand, it may be the outliers themselves on which the analysis should be concentrated. For example, communities with abnormally low crime rates may be the most instructive ones.*
>
> (Ibid., pages 27 and 28)
>
> *Outliers can…be described as data elements that deviate from other observations by so much that they arouse suspicion of being produced by a mechanism different than that which generated the other observations.*
>
> (“Summarization Techniques for Visualization of Large Multidimensional Datasets,” Sarat M. Kocherlakota, Christopher G. Healey, Technical Report TR-2005-35, North Carolina State University, page 4)



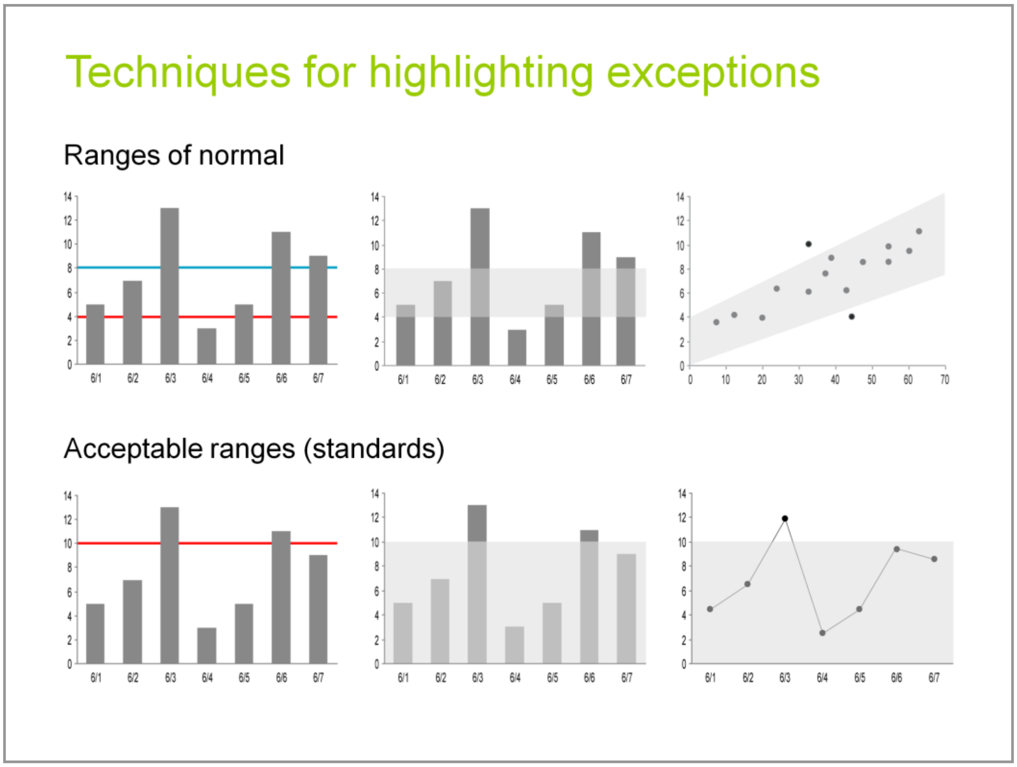
## Exceptions can indicate…

- Erroneous data
- Extraordinary events
- Extraordinary entities

I am using the term “exception” to refer to any abnormality in a set of data. Exceptions are sometimes called “outliers.” Technically, the terms outlier and exception do not share the exact same meaning. Both, however, are worth examining. Something is considered an exception anytime it falls outside of defined standards or expectations, and sometimes when it falls outside of a narrow definition of what is normal. Outlier is statistical term that refers to values that fall outside the norm based on a statistical calculation, such as anything beyond three standard deviations from the mean.

To identify exception, you must first define what is normal in a way that excludes only those values that are extraordinary. Exceptions can then be seen as anything well outside the range of normal.

Exceptions are often errors in the data caused by inaccurate data entry, inaccurate measurements, etc. Sometimes exceptions are the result of extraordinary events—something happened that caused behavior that is atypical, like a storm, the loss of a key employee, or the election of a new political leader. And finally, sometimes exceptions result from the behavior of a person, organization, or some other entity that itself falls outside of the norm, such as a customer from a country that rarely places orders or a person with highly unusual tastes.



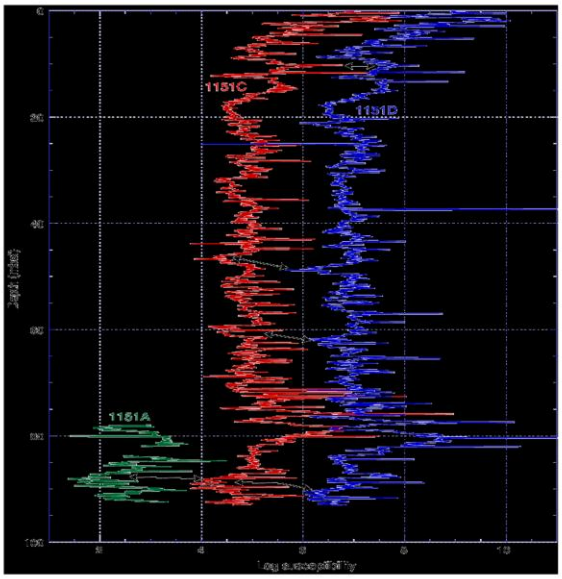
## Techniques for highlighting exceptions

Ranges of normal

Acceptable ranges (standards)

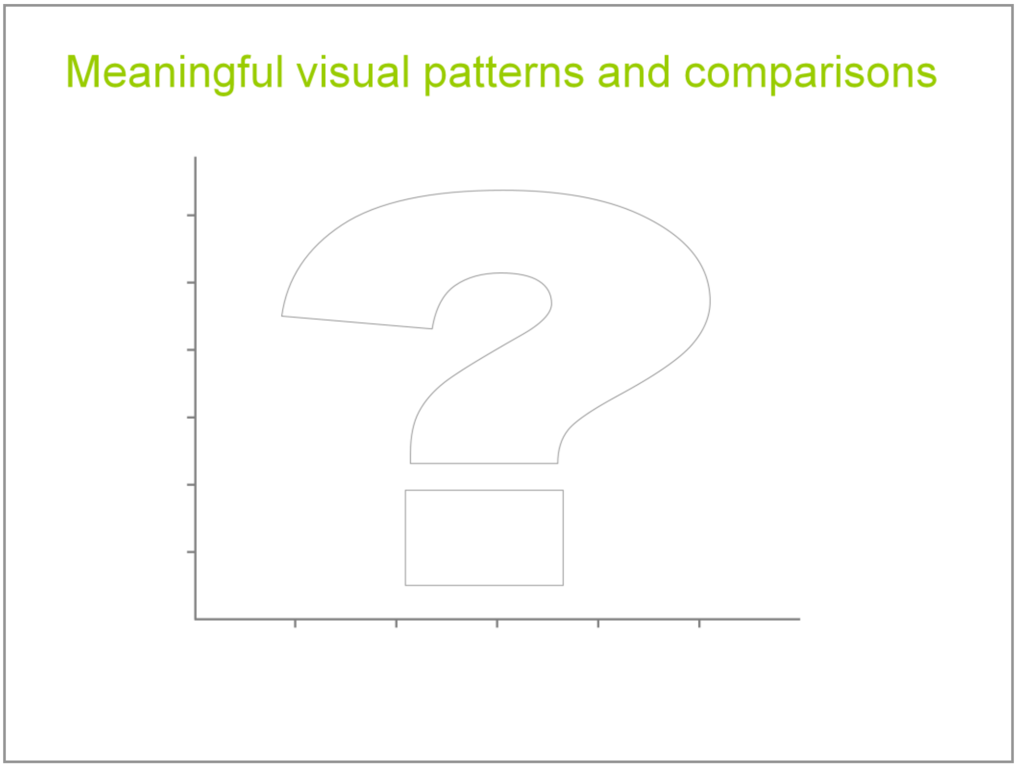
The upper and lower boundaries of what we define as normal can be easily shown on a 2-D xy graph as a line for each of the boundaries or as an area of fill color that either defines the range or defines the areas outside of the range. Standards can be displayed in a similar way. It is often valuable to use visual methods like this to clearly encode normal ranges so that exceptions will clearly stand out beyond these ranges.

## Not all patterns are meaningful.



The number of unique visual patterns that exist in the world is virtually without limit. The number of patterns that can represent meaningful quantitative information in 2-D graphs, however, is limited. If you learn to recognize these patterns, you can find them faster and more often, which will save you a great deal of time and effort.

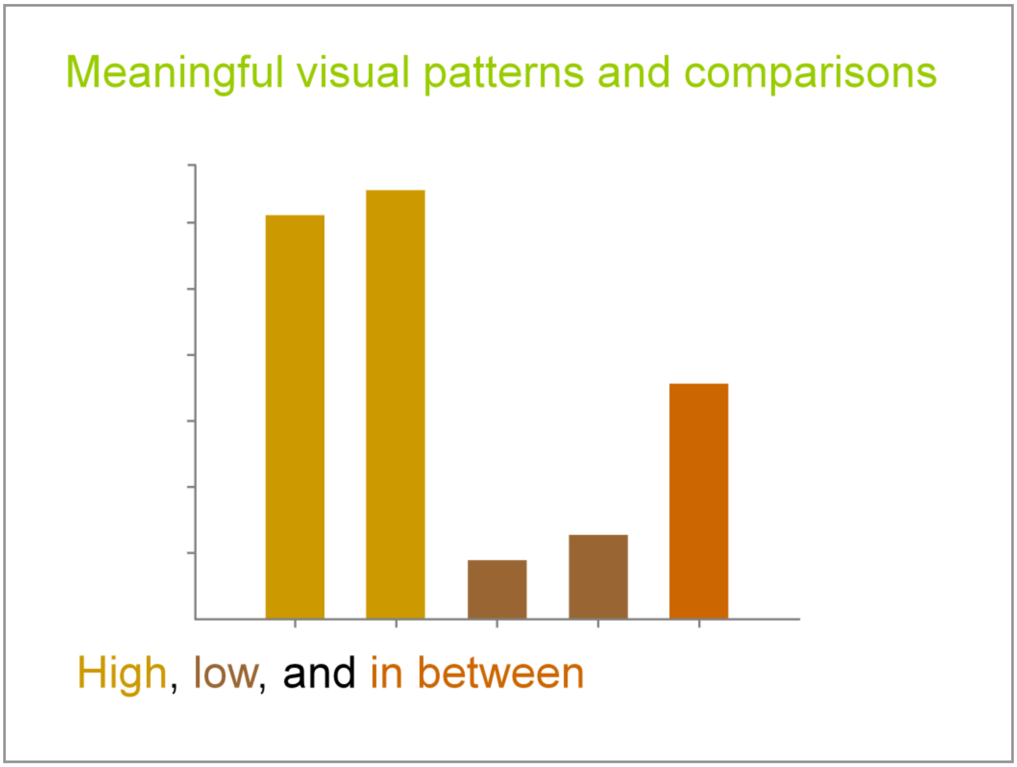
The example above looks overwhelming complex to most of us. To someone who has been trained to read this image, however, and has developed expertise in doing so through practice, this image is not overwhelming at all. Much of what appears in this image isn't important—it is visual noise—from which the meanings that matter must be extracted. Part of what you must learn to do is to quickly separate the meaning from the noise.



## Meaningful visual patterns and comparisons

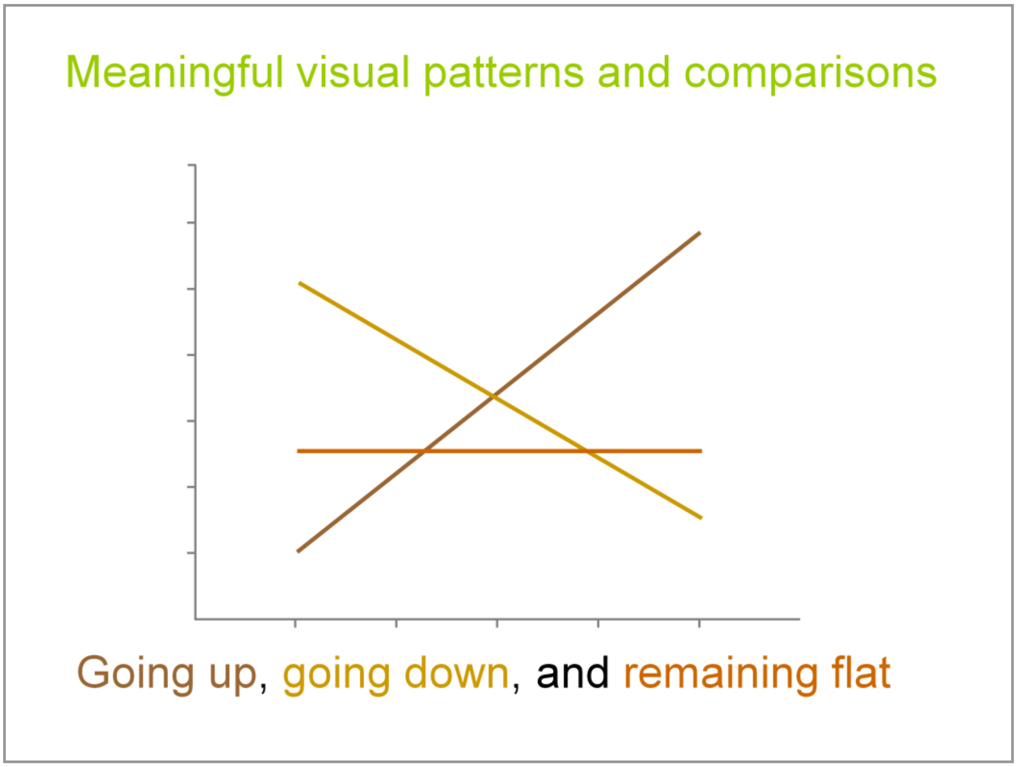
While looking at this blank 2-D graph, try to imagine all of the different visual patterns that can be displayed, which would represent quantitative values as a meaningful pattern or useful comparison. Take a minute to list as many as you can.

The next few slides illustrate visual patterns and comparisons that are often meaningful and useful for purposes of analysis.



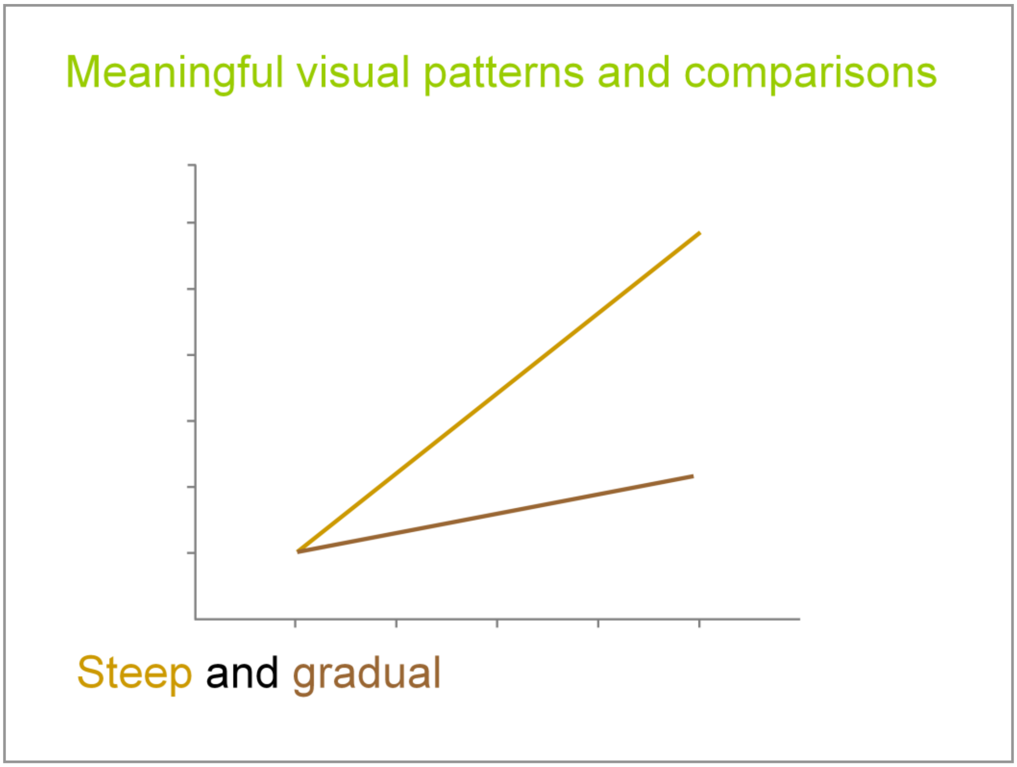
## Meaningful visual patterns and comparisons

High, low, and in between



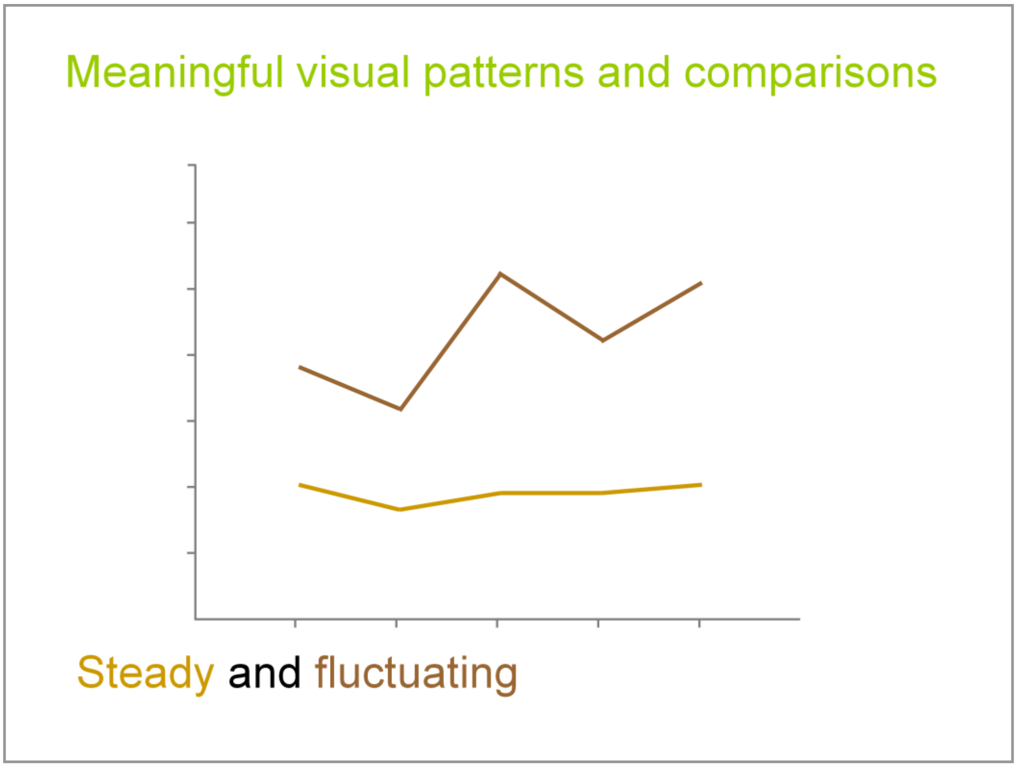
## Meaningful visual patterns and comparisons

Going up, going down, and remaining flat



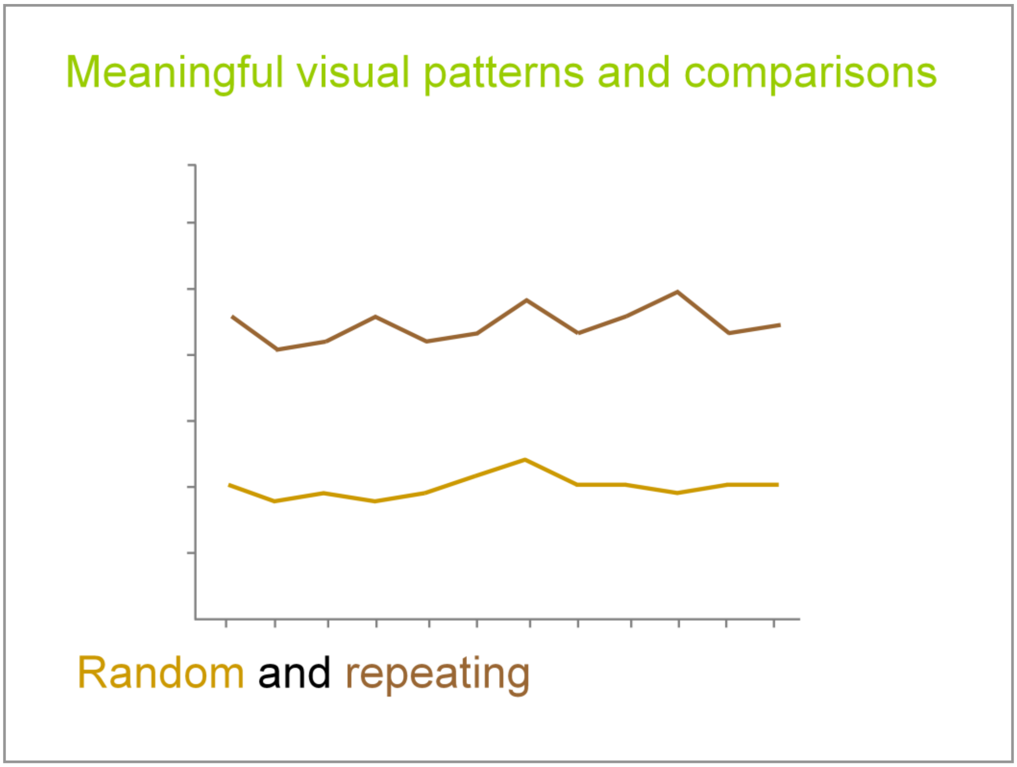
## Meaningful visual patterns and comparisons

Steep and gradual



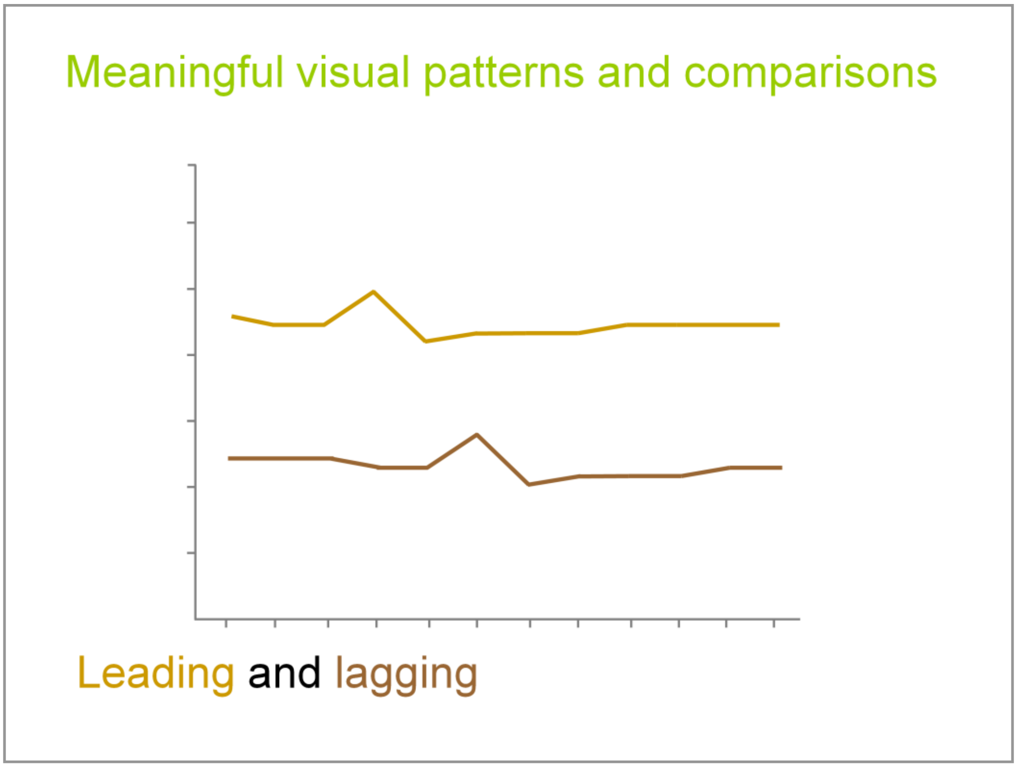
Meaningful visual patterns and comparisons

Steady and fluctuating



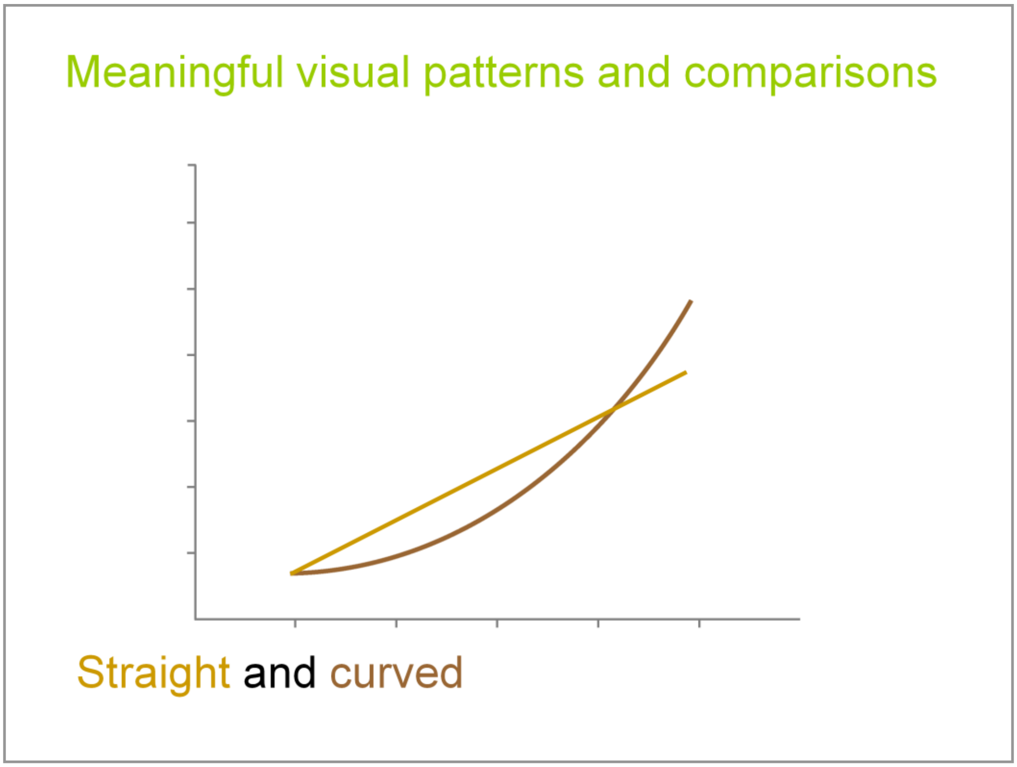
## Meaningful visual patterns and comparisons

Random and repeating



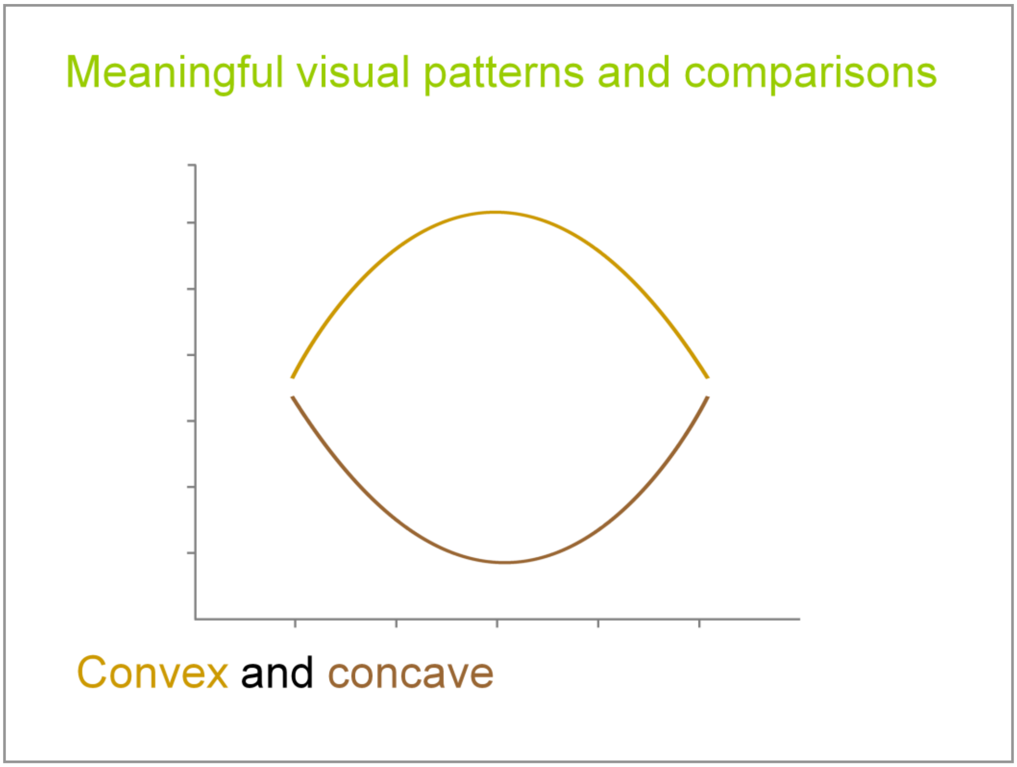
## Meaningful visual patterns and comparisons

Leading and lagging



## Meaningful visual patterns and comparisons

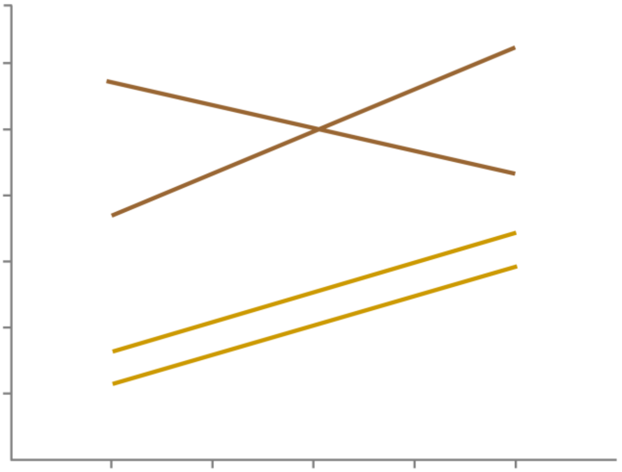
Straight and curved



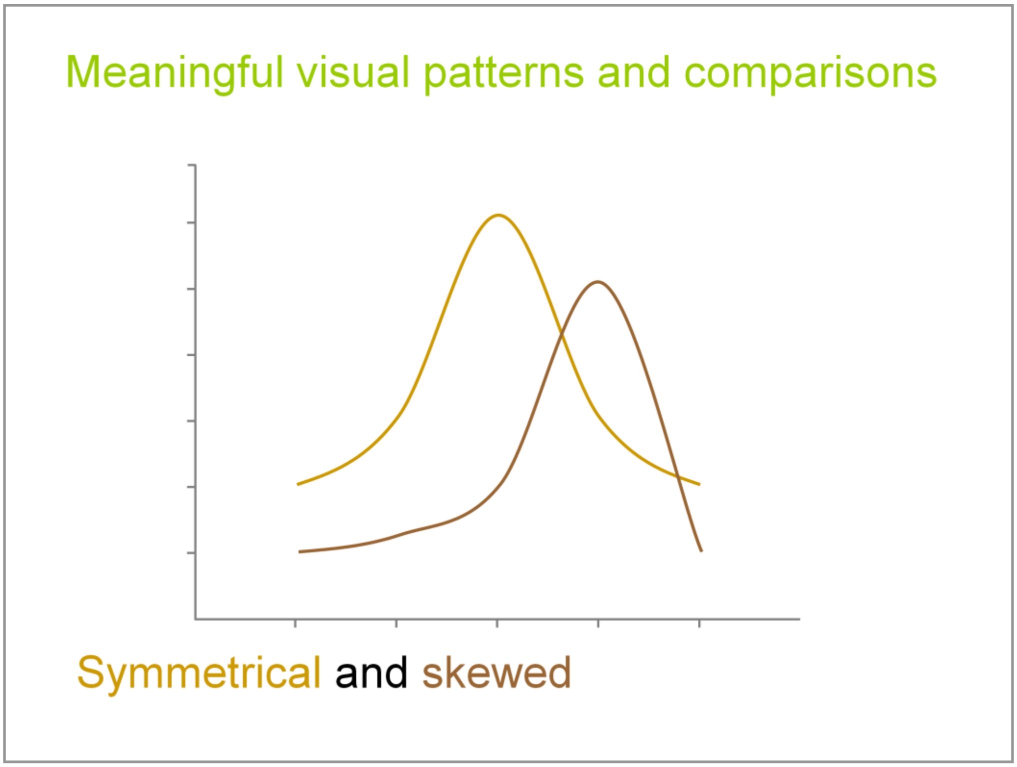
## Meaningful visual patterns and comparisons

Convex and concave

## Meaningful visual patterns and comparisons



Non-intersecting and intersecting

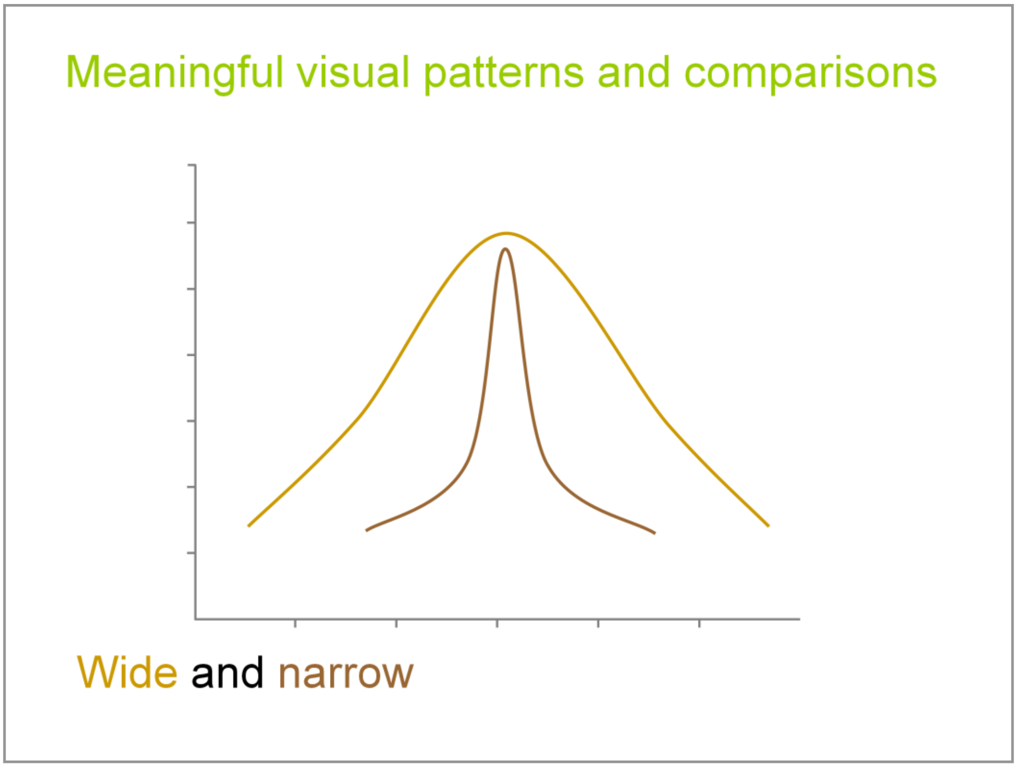


## Meaningful visual patterns and comparisons

Symmetrical and skewed

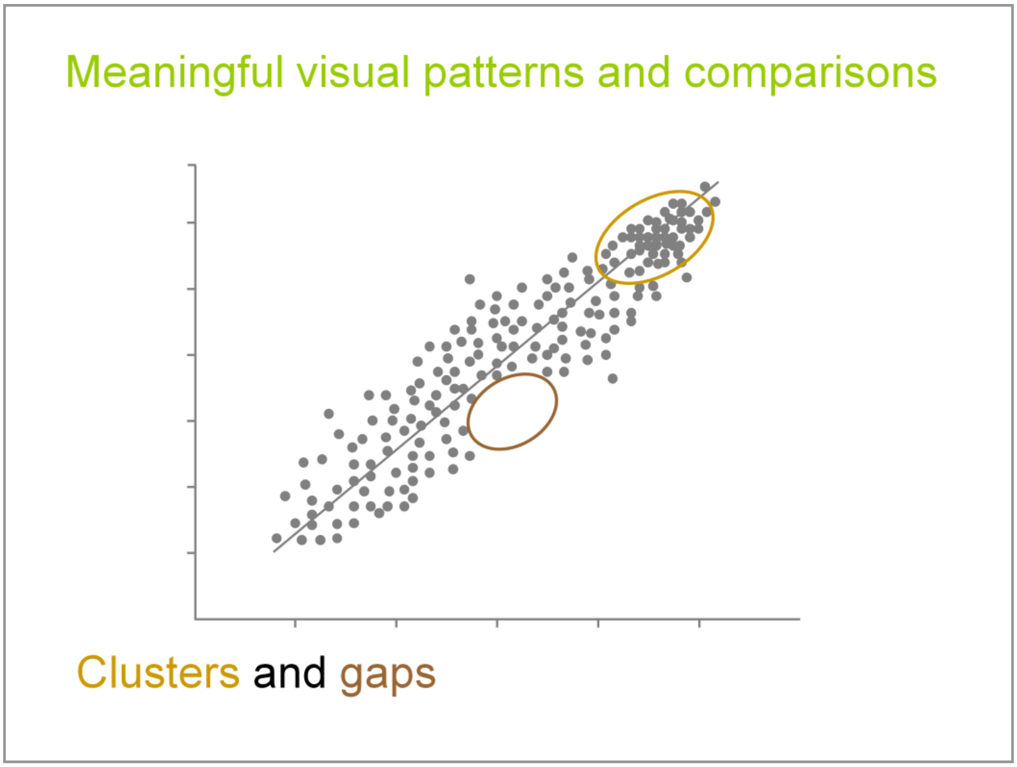
The two lines above each illustrate the shape of a set of values distributed across an entire range. The shape formed by the green line is called a *bell-shaped curve*, *normal curve*, or *Gaussian curve*. When the frequency of something's occurrence is measured for an entire population and broken into equal intervals of the full range of quantitative values that were measures from the lowest to the highest, and then graphed as a frequency distribution using a line to display the shape of the curve, it often looks something like this. Relatively few items fall into the low value ranges, but the number gradually increases around the middle ranges of values until it reaches its peak, then gradually decreases toward the higher values. This is a symmetrical distribution.

The orange line represents a distribution that is skewed to the left. The end of the distribution with the longer tail is the end toward which the distribution is said to be skewed. This particular shape tells us that more of the measures fall into the higher values than the lower values.



## Meaningful visual patterns and comparisons

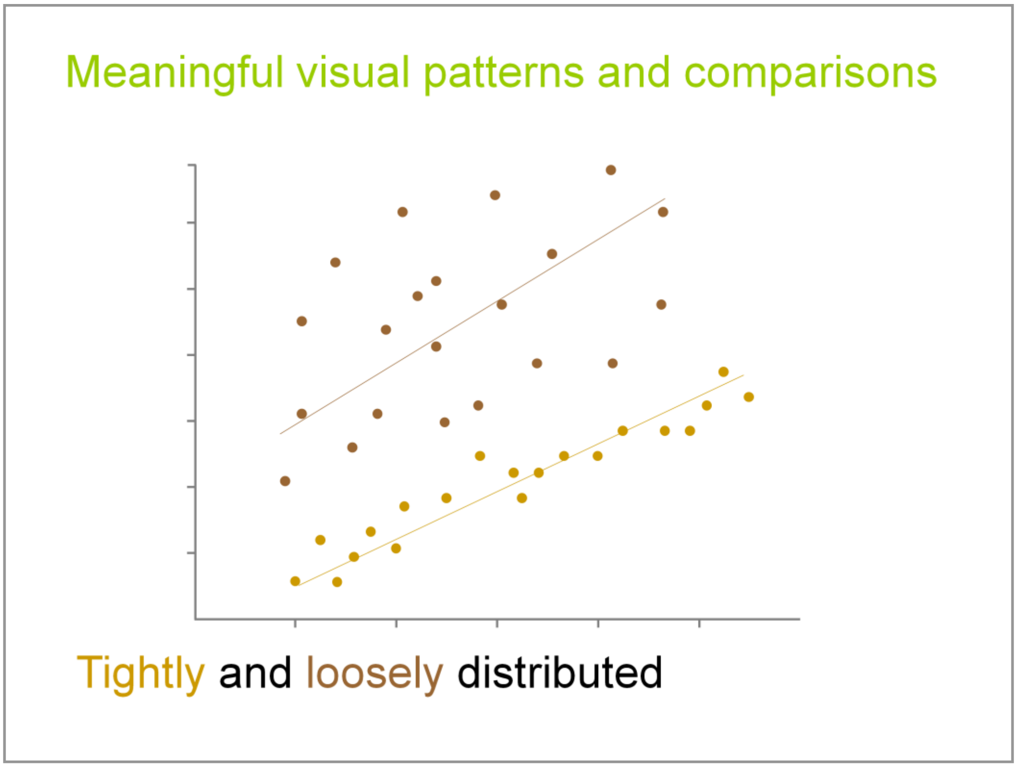
Wide and narrow



## Meaningful visual patterns and comparisons

Clusters and gaps

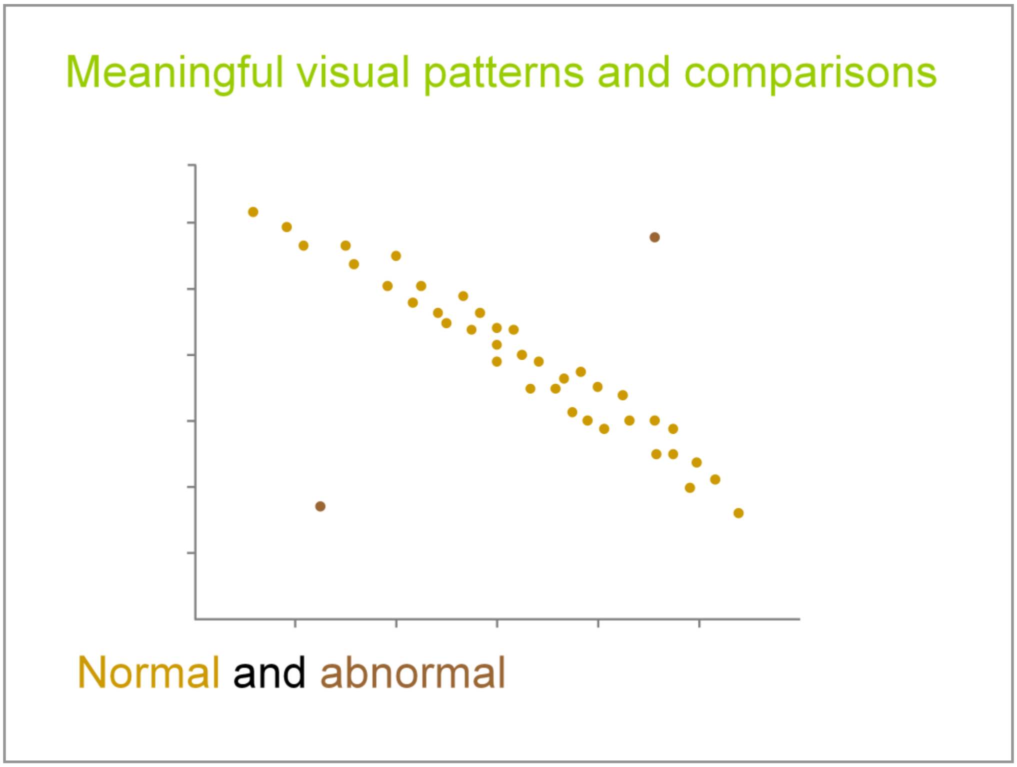
Groupings of data points in close proximity to one another and the conspicuous absence of data points in particular area is especially meaningful when using scatter plots to examine correlations between two paired sets of values.



## Meaningful visual patterns and comparisons

Tightly and loosely distributed

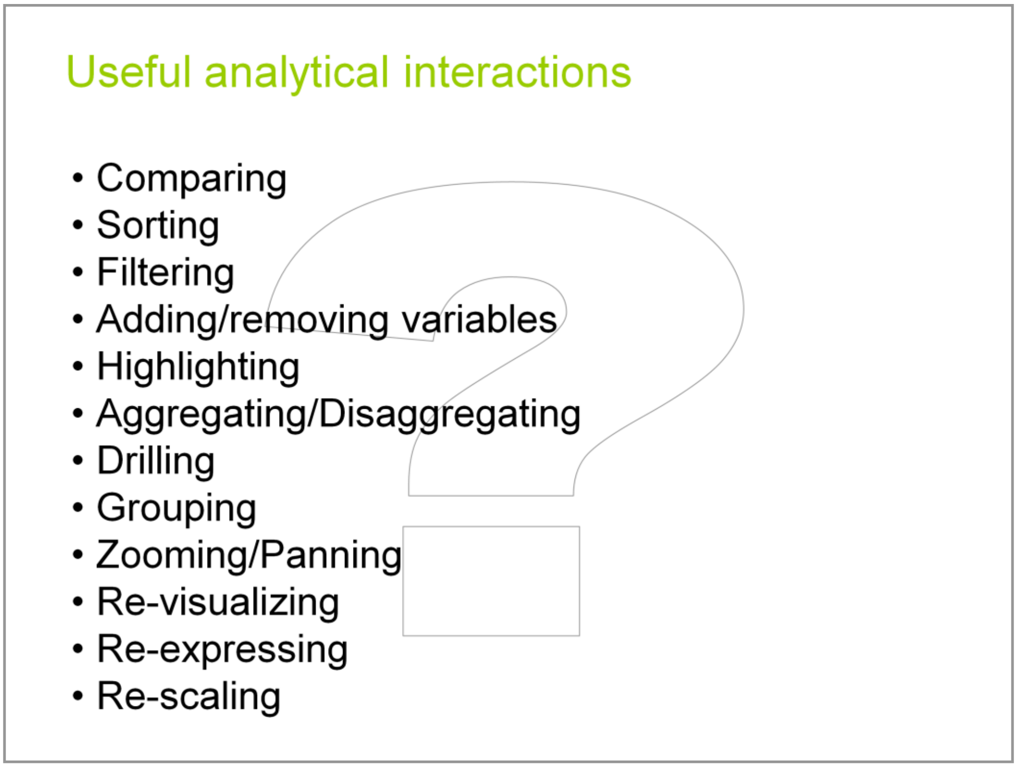
The tightness or looseness of a distribution of data points like those above is particularly meaningful when using scatter plots to look for correlations between two paired sets of values, like the ratings that people receive during their performance reviews and the salary raises that they receive.



Meaningful visual patterns and comparisons

Normal and abnormal

Whether you are using data points, lines, bars, or boxes to encode the values, most or all values fall within a common range, but sometimes a few fall far enough outside of the normal to be considered abnormal. These values are called exceptions or outliers. Much can often be learned from the investigation of abnormal values.

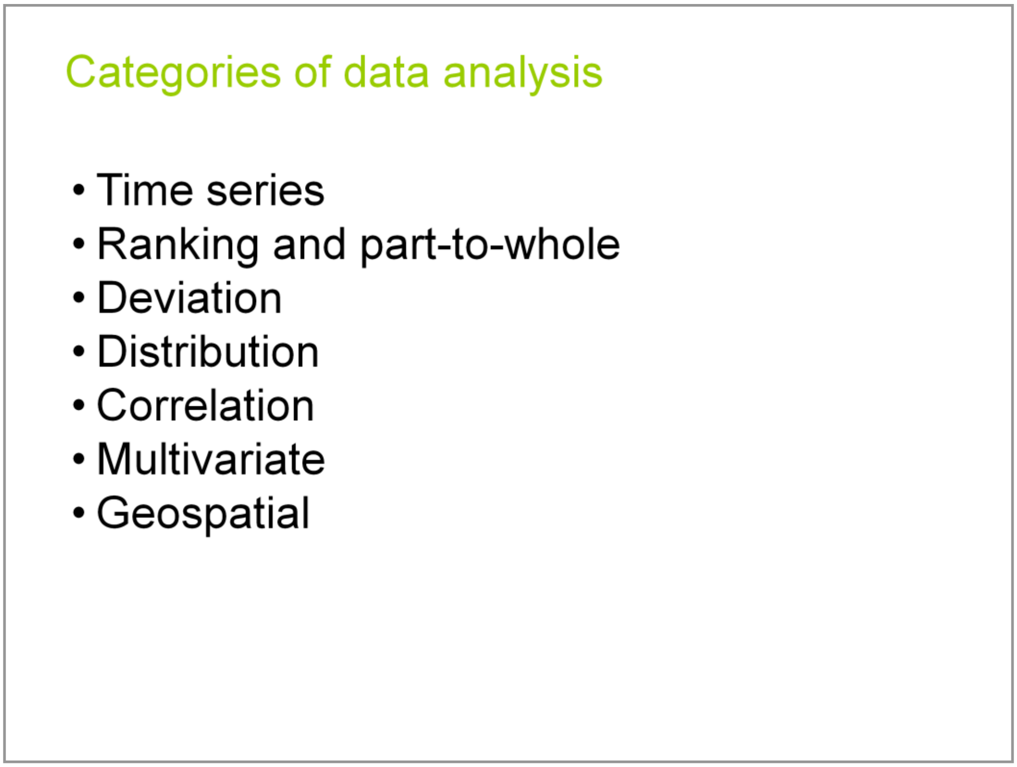


## Useful analytical interactions

- Comparing
- Sorting
- Filtering
- Adding/removing variables
- Highlighting
- Aggregating/Disaggregating
- Drilling
- Grouping
- Zooming/Panning
- Re-visualizing
- Re-expressing
- Re-scaling

The process of visual data analysis involves several common interactions with data to uncover what's meaningful. Here are some of the primary interactions:

- **Comparing**. No interaction is more basic or common than that of comparing values and patterns to one another. Tufte says the that basic question of data analysis is "compared to what?"
- **Sorting**. The act of sorting data, especially by the magnitude of the values from high to low or low to high, features the ranking relationship between those values and makes it easier to compare the magnitude of value to the next.
- **Adding/removing variables.** You might need to view different variable at different times during the analysis process, so it is common to add or remove field of data from view as necessary
- **Filtering**. When you want to focus on a subset of data, nothing makes it easier to do so than filtering—the removal from view of everything your not interested in at the moment.
- **Highlighting**. Sometimes you want to focus on a subset of information, but do so in a way that allows you to maintain a sense of how that subset relates to the whole. Rather than filtering out the data that falls outside your range of focus, you can simply reduce its visual salience or increase the visual salience of the data you wish to focus on. This allows you to focus on the subset with less distraction from the whole in a way that allow you to remain aware of the whole. This is one way of achieving what's called a focus+context view.
- **Aggregating/Disaggregating**. Analysis often requires that you examine data a different levels of detail. Aggregation involves viewing data at a higher level of summarization. Disaggregation involves viewing data at a lower level of detail.
- **Drilling**. Similar to disaggregation, drilling involves viewing data at a lower level of detail, but in a specific manner. Drilling also means that you are changing the view to the next level in a defined hierarchy, and excluding from view all data that is not directly related to the specific data value that you chose to drill into. For instance, if you drill into a particular product family, your next view only products that belong to that product family. In other words, a form of filtering is involved.
- **Grouping.** Sometimes it is useful to combine members of a variable together, treating them as a single member of the variable. This may take the form of combining some members and leaving others as they are, or of creating an entirely new variable that combines all members of an existing variable into a groups to form members of a higher level variable.
- **Zooming/Panning**. When a data visualization contains so much that it is difficult to clearly see all the data at once, it is useful to zoom in on that portion that you want to see more clearly. Panning involves moving around (for example, up, down, right, or left) in a zoomed view to focus on a different part of the larger visualization.
- **Re-visualizing**. No one visual representation of data can show you everything there is to see, so visual analysis involves shifting from one type of visualization to another to explore data from various perspectives.
- **Re-expressing.** Sometimes it is useful to express a quantitative variable as a different unit of measure, such as expressing dollars as percentages.
- **Re-scaling**. No single quantitative scale on a graph can serve every analytical need. Rescaling involves changing the range of the quantitative scale to make it easier to see particular patterns and sometimes even changing the nature of the scale, such as from a normal scale to a logarithmic scale.



## Categories of data analysis

- Time series
- Ranking and part-to-whole
- Deviation
- Distribution
- Correlation
- Multivariate
- Geospatial

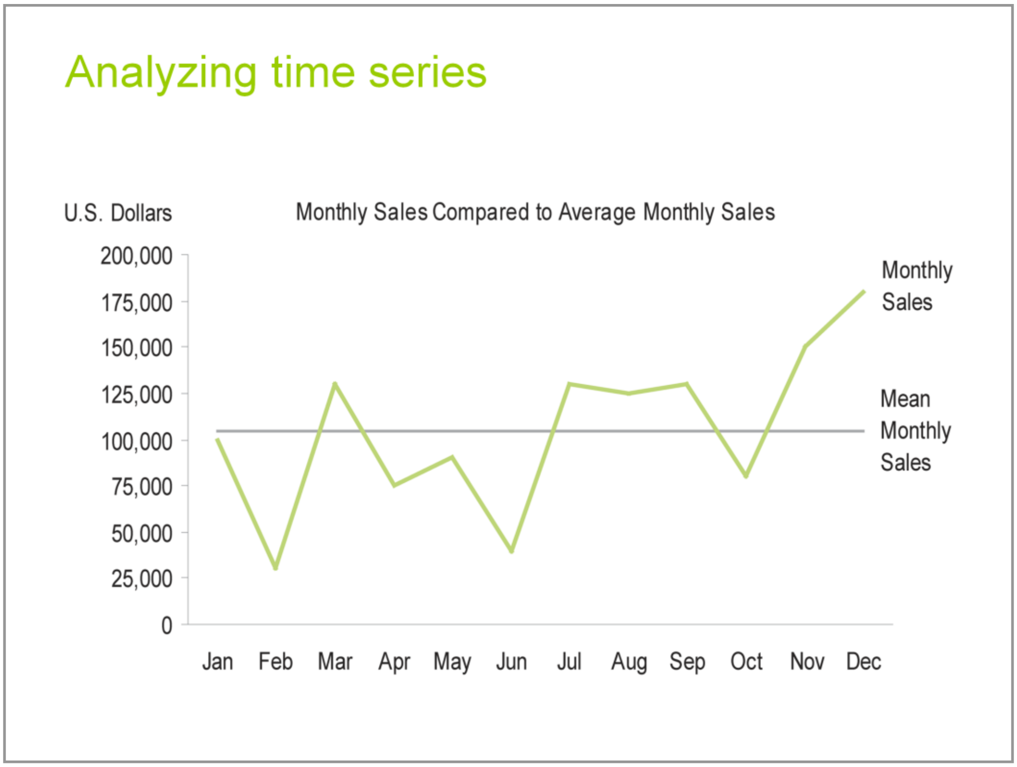
When we analyze quantitative data, we always focus on one or more relationships between the values. The patterns that we seek are only meaningful within the context of a particular relationship. The significance of a series of values that are increasing depends on the nature of the relationship that we are examining. When examining how a set of values changes through time an increase means something quite different than when examining how a set of values is distributed across a quantitative range unrelated to time.

The most common relationships that we examine when analyzing quantitative business data are the following:

- **Time Series:** How the values change through time.
- **Ranking and part-to-whole**: How the values associated with categorical items are sequenced by size and how those sizes compare to one another and the whole.
- **Deviation**: How two or more sets of values differ (e.g., actual vs. budgeted expenses).
- **Distribution**: How the values relate to one another as a matter of proximity (i.e., their distribution through the entire range of values).
- **Correlation**: How two sets of quantitative variables associated with a common set of entities behave in relation to one another.
- **Multivariate**: How multiple entities are similar or different based on a common set of variables.
- **Geospatial**: How the spatial positions of values (e.g., where they reside geographically) contribute to their meaning.

Each of these relationships can be best examined by using particular graphs and analytical techniques. For each of these types of analysis we will consider the following:

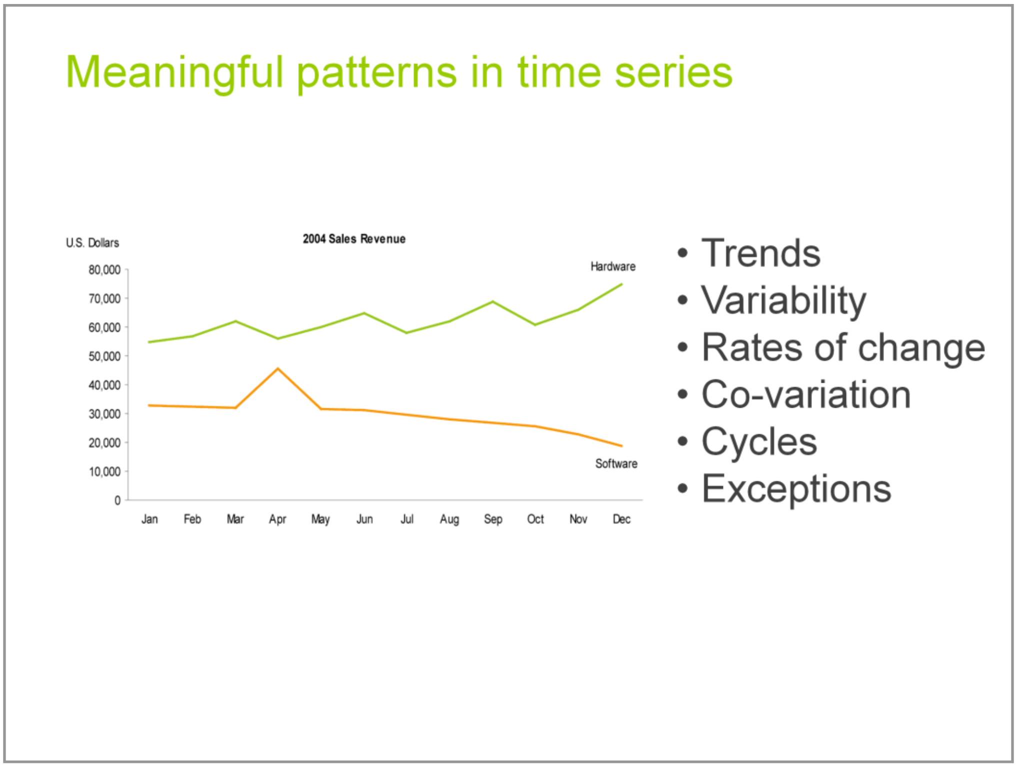
- Meaningful patterns
- The most effective graphs for viewing the patterns
- Useful analytical and graphical techniques for making sense of the patterns



## Analyzing time series

No quantitative relationship receives more attention in business analysis than the way values vary and relate to one another through time. “A random sample of 4,000 graphics from 15 of the world’s newspapers published from 1974 to 1989 found that more than 75% of all graphics were time series.” (“An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data,” Eamonn Keogh, Harry Hochheiser, and Ben Shneiderman)

We hope to see revenues increasing as time passes. We expect to see changes up or down in relation to influential events. Time is the most fundamental dimension of business data.



## Meaningful patterns in time series

- Trends
- Variability
- Rates of change
- Co-variation
- Cycles
- Exceptions

**Trend:** the overall tendency for a series of values to increase, decrease, or remain relatively stable during a particular span of time.

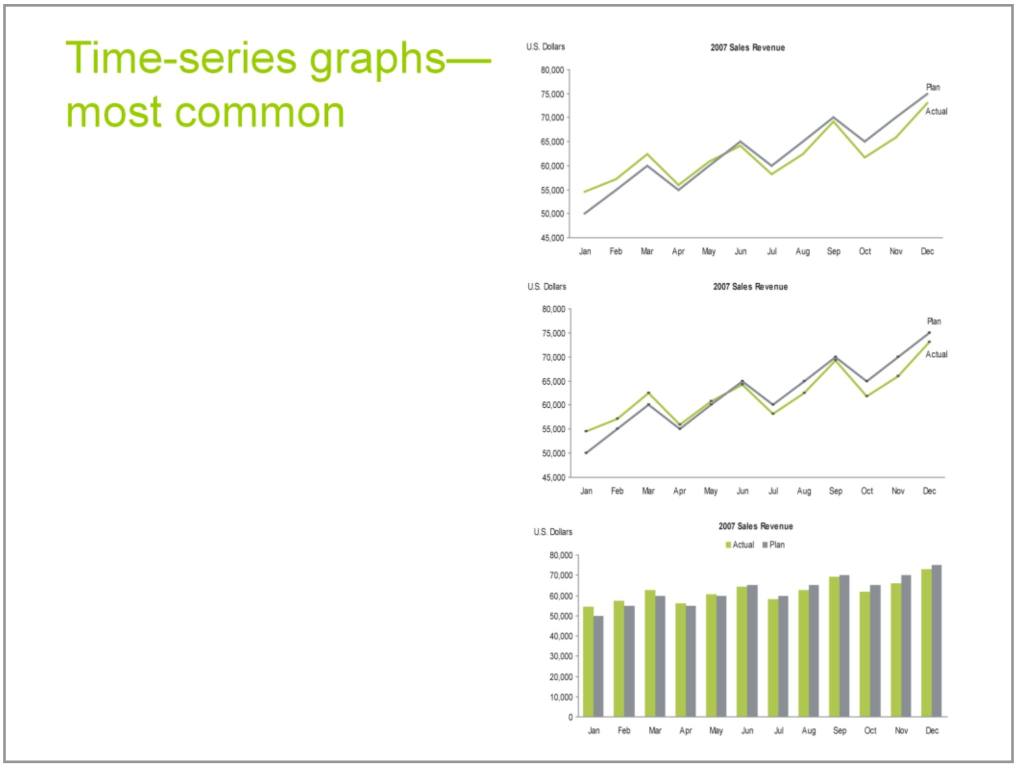
**Variability:** the average degree of change from one point in time to the next during a particular period of time.

**Rate of change:** when the amount of change from one value to the next is expressed as the percentage difference between the two, this measures the rate of change.

**Co-variation:** when two sets of time-series values (such a two lines in a line graph) relate to one another such that changes in one are reflected as changes in the other, either immediately or later, this is called co-variation.

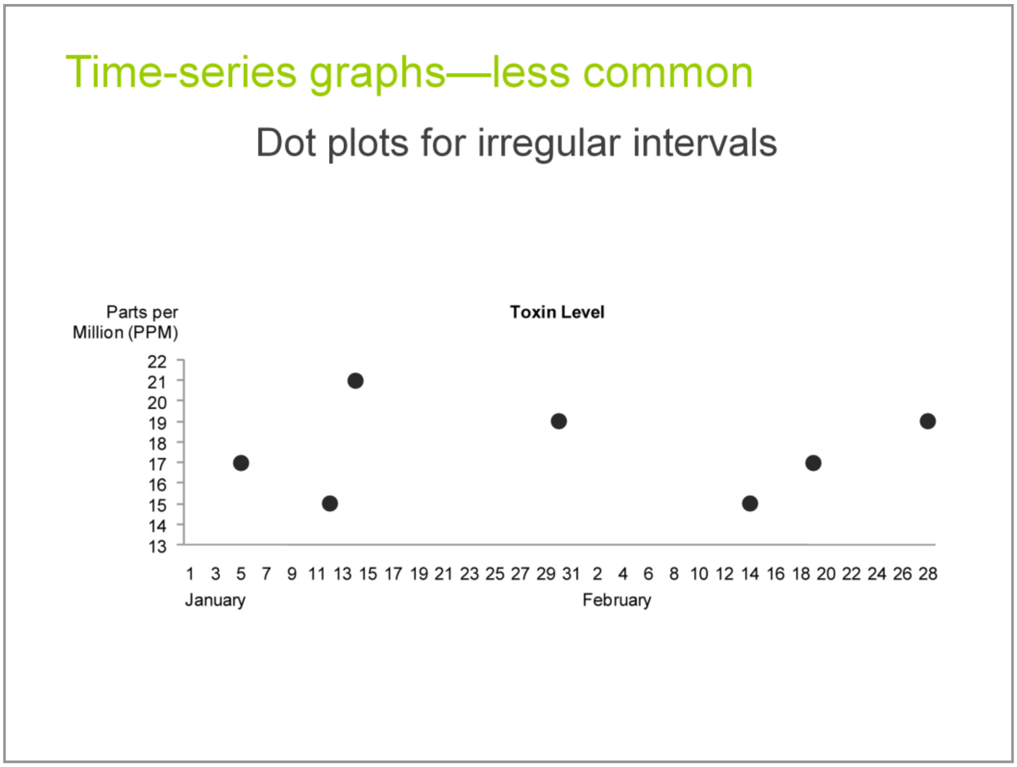
**Cycle:** patterns that repeat daily, weekly, monthly, quarterly, yearly, seasonally (winter, spring, summer and fall), or at some other regular interval of time.

**Exception:** values that fall outside the norm.



## Time-series graphs—most common

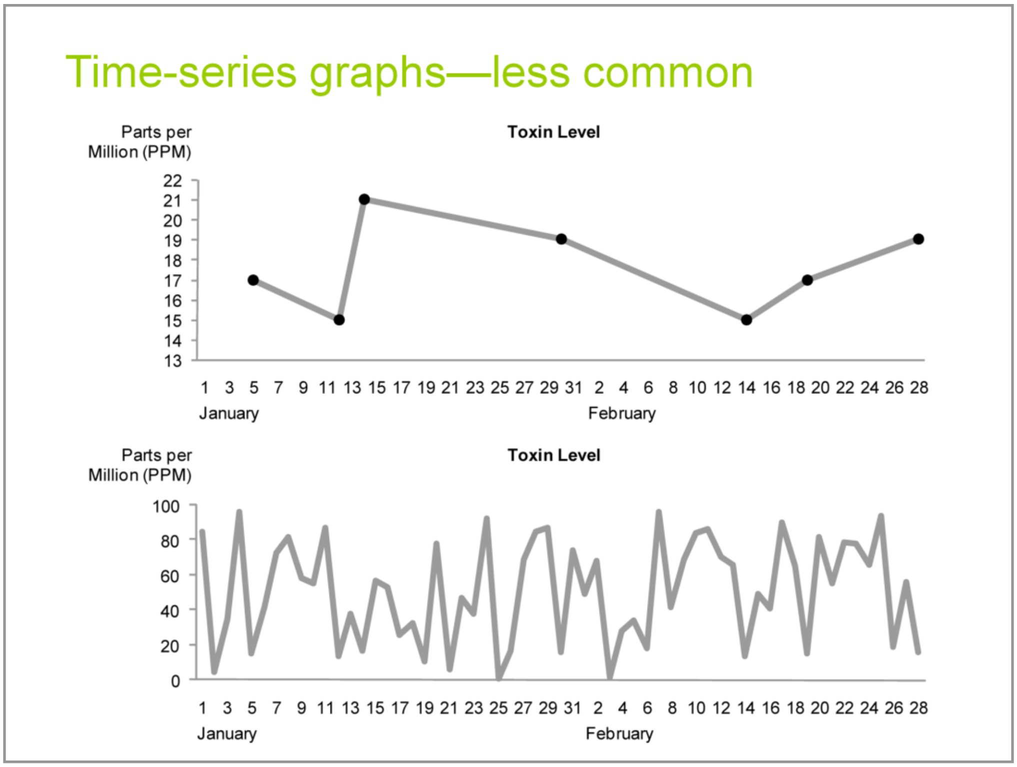
If your objective is to see how values vary in relation to time, nothing works better than a line graph or a combination line and point graph. Lines make visible the flow of and changes in values through time better than any other means. By their very nature they depict connectivity between values that visually presents the change that occurs from one value to next as the slope of the line. When it is your purpose to compare individual values at specific points in time, such as in a given month, then bar graphs do this best, but the overall shape of the values and how they change through time is revealed best by lines.



## Time-series graphs—less common

Dot plots for irregular intervals

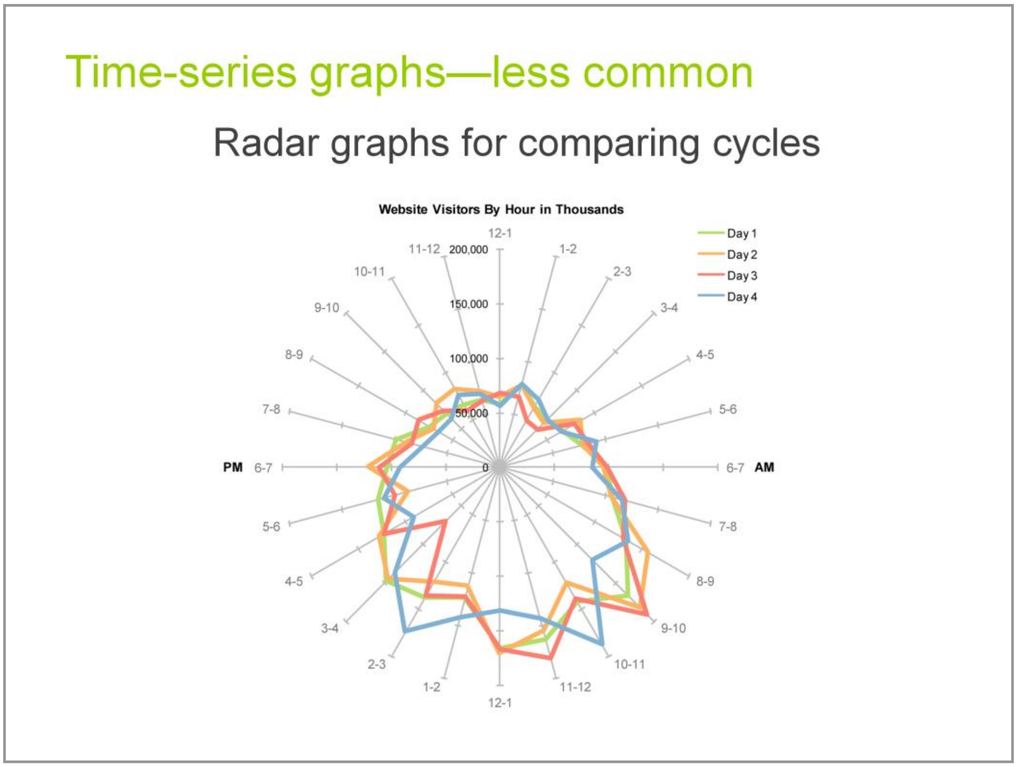
When analyzing values that are spaced at irregular intervals of time, don’t connect them with a line. By simply placing a point, such as a dot, to mark each value without connecting them with a line, we have what’s called a dot plot, as shown below. Dot plots discourage the misleading suggestion that there was a smooth linear transition from one value to the next.



## Time-series graphs—less common

Usually, when analyzing time-series data, each interval along the timeline contains one value for a given series. We might begin with a much larger set of individual values for the period of time that we're examining, but we reduce them by aggregating them and displaying one per interval of time (for example, one per month). Sometimes, however, the data that we're examining is not spread continuously throughout the period, but only exists sporadically, from time to time. For example, imagine that you're an inspector who measures the level of a particular toxin at a particular location in a river from time to time, rather than at regular intervals. If you used a line graph to display this data, it might look something like the one on the top above.

By connecting values that appear at irregular points in time with a line, the resulting slopes between those points suggest a smooth change in value from one to the next. This is a problem, because these smooth transitions might not at all correspond to what really happened. If the toxin levels had been measured every day, the picture of change might look quite different, such as in the graph above on the bottom.

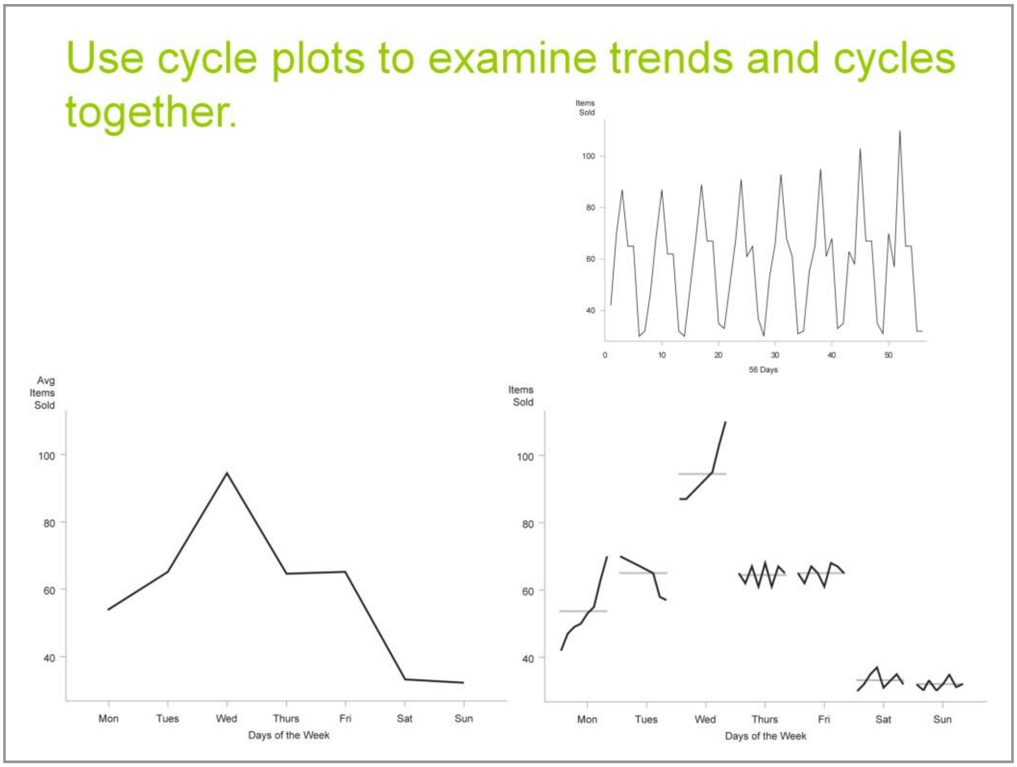


Time-series graphs—less common

Radar graphs for comparing cycles

In general, I'm not a big fan of radar graphs (also known as spider graphs), because their usefulness is limited to rare occasions and they're often used inappropriately, but they can play a useful role in time-series analysis. The circular shape of a radar graph can be used to represent the cyclical nature of time. For example, similar to how hours on a clock are sequentially arranged in a circle, the axes of a radar graph can be used to mark the hours of the day, as shown above.

The same data can be displayed using a line graph, which works just as well for analytical purposes, but if you prefer the way that radar graphs represent the cyclical nature of time—the minutes of the hour, hours of the day, or even days of the week or month, months of the year, and so on—you will find them useful.

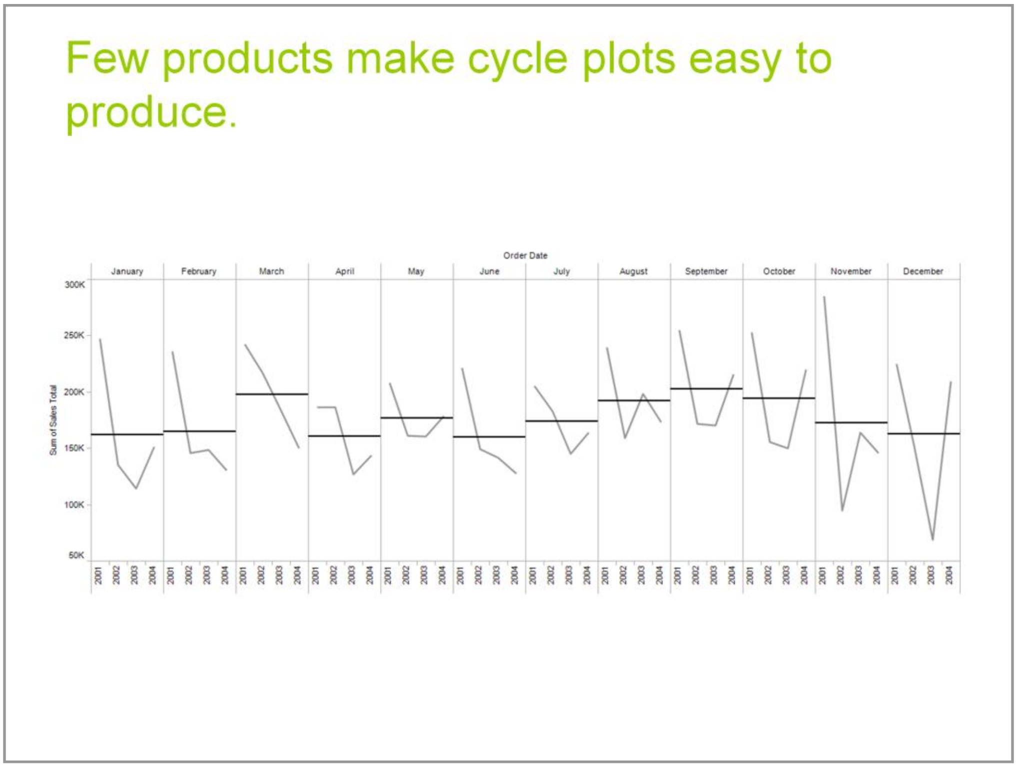


Use cycle plots to examine trends and cycles together.

Overlapping time scales, shown on the previous page, make it easy to compare cycles to one another, but do so in a way that sacrifices our ability to see trends that extend across multiple cycles.

The small graph at the top displays 56 days worth of sales. We can get a sense of weekly cycles, but to see the cycles clearly, we need a different view. The line graph on the bottom left displays the average sales per day of the week for these same eight weeks, but now we've lost sight of the variation that exists from week to week. On the right is the same data displayed as a cycle plot. Cycle plots (illustrated on the bottom right above) were developed by William Cleveland, Douglas Dunn, and Irma Terpenning at Bell Laboratories in the 1970s to give us a better view. They allow us to see two fundamental characteristics of time-series data in a single graph:

- The overall pattern across the entire cycle
- The trend for each point in the cycle across the entire range of time

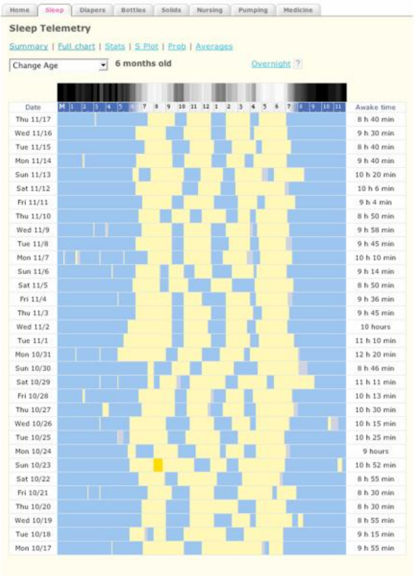


## Few products make cycle plots easy to produce.

The ability to summarize cycles and view longer trends without shifting from graph to graph can lead us to insights that might not otherwise occur. Some software products provide cycle plots as a special form of line graph. The example above was produced using Tableau. All that was necessary to shift from a normal line graph to this cycle plot was to reverse the order of the month and year fields when I constructed the graph, by placing month before year, which caused the years (2001-2004) to be grouped within the months. In other words, it took no longer to construct this cycle plot than it did to construct the normal line graph, and I could at any time switch back and forth between the two simply by reversing the order of the year and month levels of Order Date.

## Time-series graphs—less common

### Heatmaps for high-volume cycles and exceptions


Home
Sleep
Diapers
Bottles
Solids
Nursing
Pumping
Medicine
Summary | Full chart | Stats | S Plot | Prob | Averages
Change Age
Overnight ?
M 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11

Sleep Telemetry

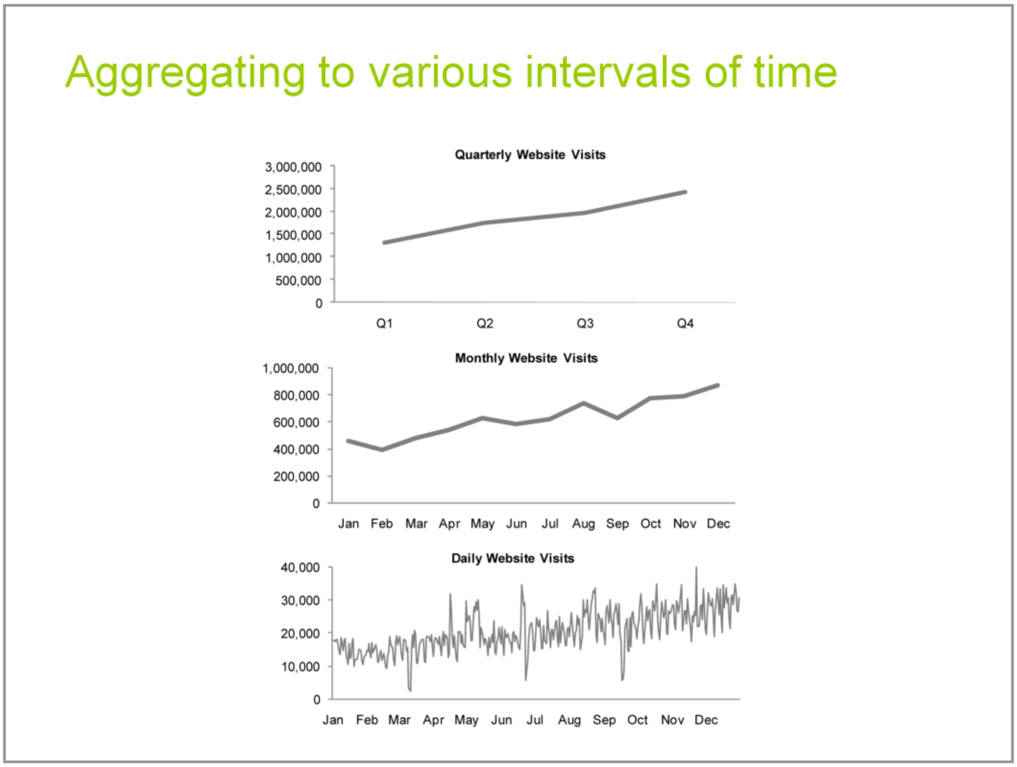
6 months old

| Date | Awake time |
|---|---|
| Thu 11/17 | 8 h 40 min |
| Wed 11/16 | 9 h 30 min |
| Tue 11/15 | 8 h 40 min |
| Mon 11/14 | 9 h 40 min |
| Sun 11/13 | 10 h 20 min |
| Sat 11/12 | 10 h 6 min |
| Fri 11/11 | 9 h 4 min |
| Thu 11/10 | 8 h 50 min |
| Wed 11/9 | 9 h 58 min |
| Tue 11/8 | 9 h 45 min |
| Mon 11/7 | 10 h 10 min |
| Sun 11/6 | 9 h 14 min |
| Sat 11/5 | 8 h 50 min |
| Fri 11/4 | 9 h 36 min |
| Thu 11/3 | 9 h 45 min |
| Wed 11/2 | 10 hours |
| Tue 11/1 | 11 h 10 min |
| Mon 10/31 | 12 h 20 min |
| Sun 10/30 | 8 h 46 min |
| Sat 10/29 | 11 h 11 min |
| Fri 10/28 | 10 h 13 min |
| Thu 10/27 | 10 h 30 min |
| Wed 10/26 | 10 h 15 min |
| Tue 10/25 | 10 h 25 min |
| Mon 10/24 | 9 hours |
| Sun 10/23 | 10 h 52 min |
| Sat 10/22 | 8 h 55 min |
| Fri 10/21 | 8 h 30 min |
| Thu 10/20 | 8 h 30 min |
| Wed 10/19 | 8 h 55 min |
| Tue 10/18 | 9 h 15 min |
| Mon 10/17 | 9 h 55 min |

Another way to visualize cyclical data, when there is a great deal of data that would suffer from over-plotting in a line or radar graph, is to use a *heatmap*. A heatmap is the generic name for any display that uses color to encode quantitative values. Weather maps are a familiar form; typically they use variations in color on a geographical map to display temperatures or levels of precipitation. Another form of heatmap involves a simple matrix (rows and columns) of cells, each color coded to represent a value. I found this example on the Web at www.trixietracker.com. It was used by parents to track and attempt to make sense of the sleeping patterns of their young child during the course of each day over a period of approximately one month.

Notice how easy it is to see the dominant patterns of awake time versus sleep time, especially in the way that it is summarized in the row of grayscale colors at the top. This heatmap provides a summary of the daily binary values (either on or off) of awake versus asleep throughout the day. Despite this example, values that appear in heatmaps are certainly not restricted to those that are binary.

## Time-series analysis techniques and practices

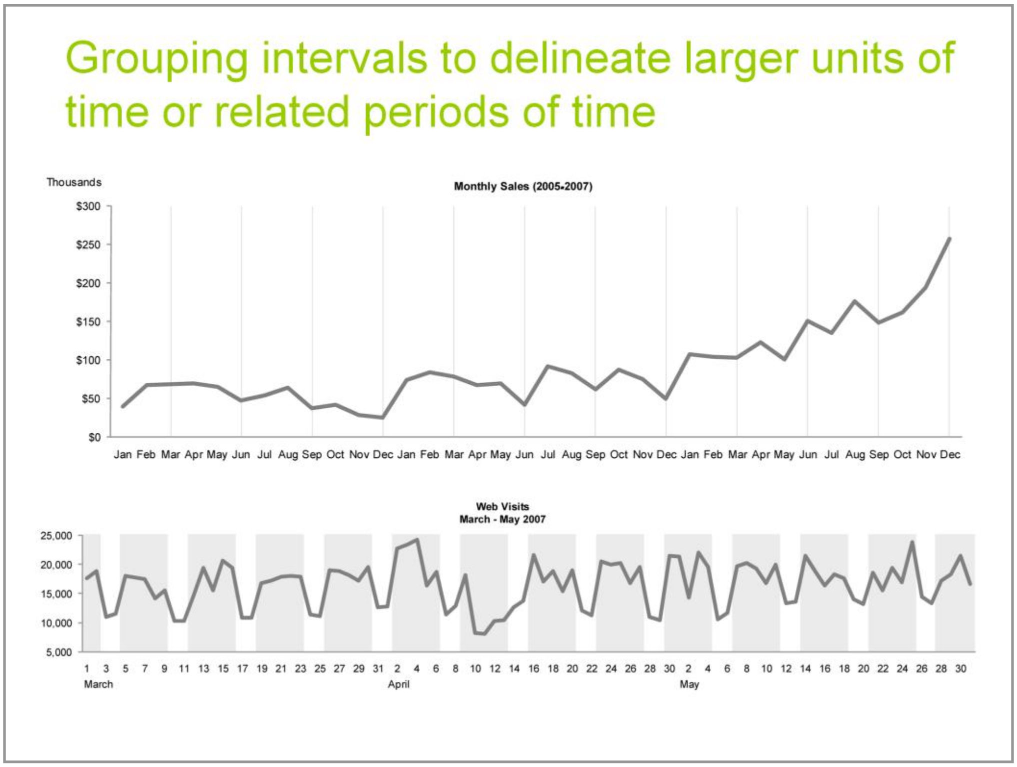


Aggregating to various intervals of time

Have you ever noticed that time-series data can look quite different if you change the interval of time that you've aggregated it to in a graph? For example, if you are examining a year's worth of visits to your website with one value per quarter (aggregated to quarterly intervals), and then switch to a monthly aggregation of the same data, and then switch to a daily aggregation, the patterns of change might look quite different.

All three versions of the graph are useful and correct, but the daily version reveals details that aren't visible when viewing the same data by month or quarter. On the other hand, the overall trend is difficult to discern from the daily view. One view isn't better than the other in general, but one is definitely better than the other when you have a specific analytical purpose in mind. For this reason, don't restrict your view of time-series data to a single interval of time, especially when looking for anything that seems interesting. Switch the level of aggregation from year to quarter, quarter to month, month to week, week to day, and so on—back and forth—to tease out the insights that can only arise when this is done.

Software products that allow you to quickly and easily switch between various intervals of time while viewing data graphically are what's needed to encourage this practice. The ability to switch time intervals with a mouse click or two, or by using something as simple as an interval slider control, will set you free to explore without distraction from onerous operations.

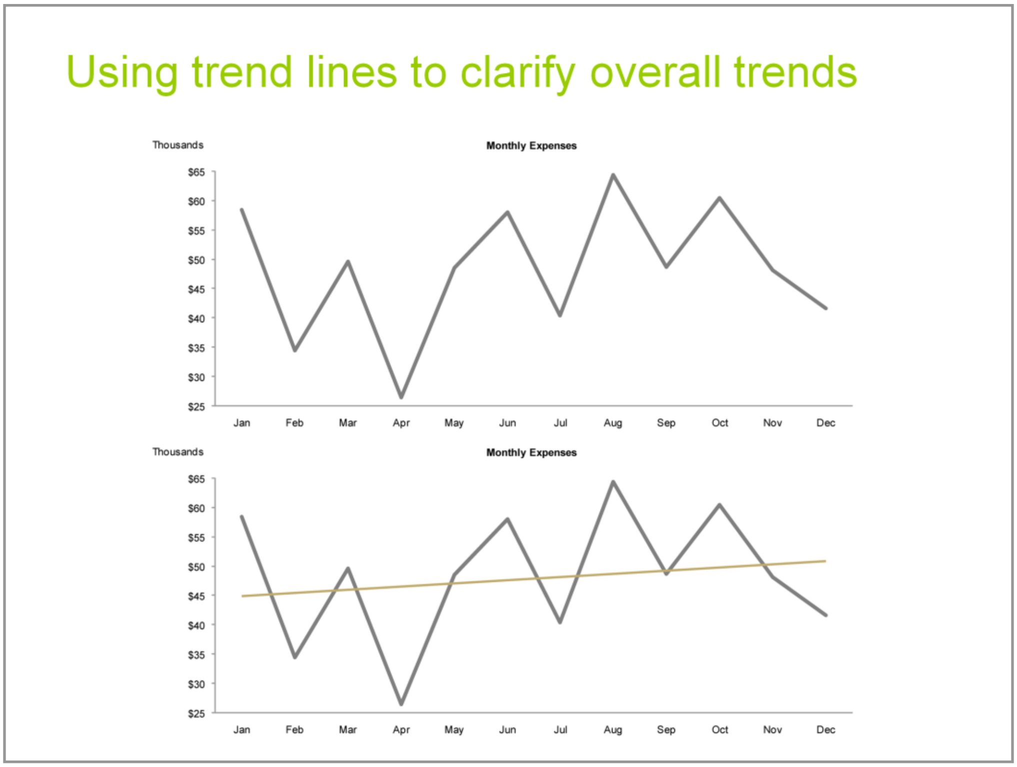


## Grouping intervals to delineate larger units of time or related periods of time

The interval of time to which values are aggregated is not always the only interval or the only means of grouping values that's worth seeing. For example, if we are examining  three year's worth of expenses by month, it would be useful to see those months delineated by year and perhaps also by quarter. As you can see below, the addition of vertical lines to divide the quarters makes it easier to examine and understand quarterly patterns.

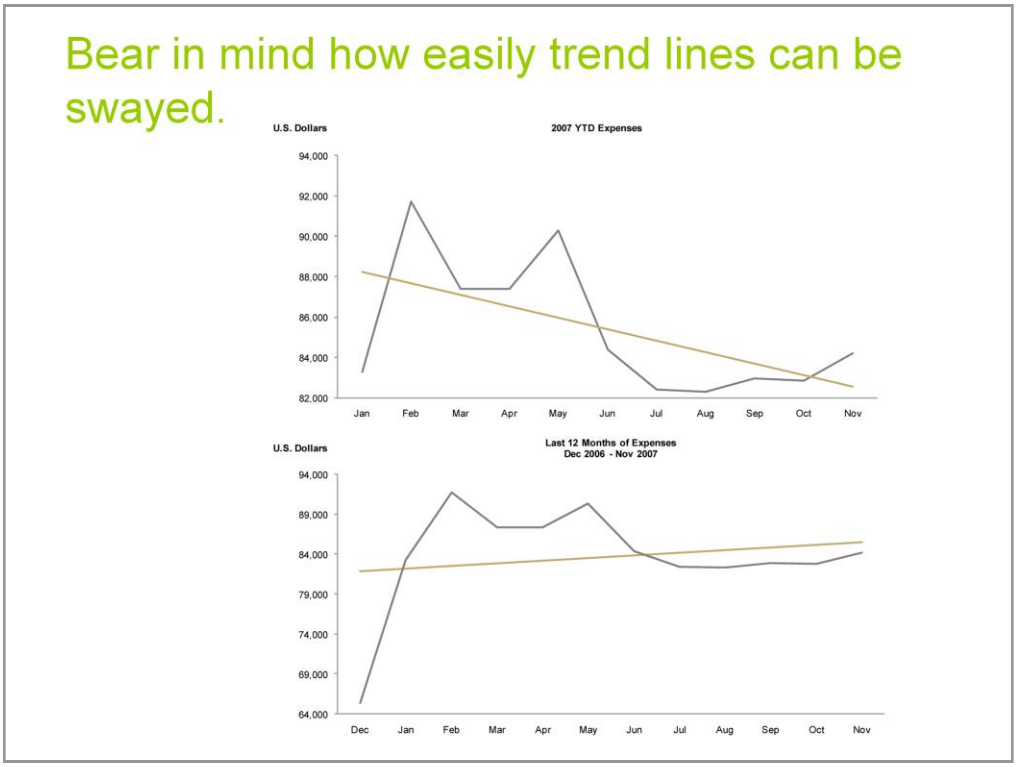
Another example of this, which I find useful, is also illustrated above. When viewing three months worth of daily website visits, it helps to clearly mark the weekdays, because this allows us to separate expected drops in Web traffic on weekends from drops that occur on weekdays when they are not expected and ought to be investigated more closely.

Unfortunately, few software products support the ability to delineate and groups periods of time in this manner. If yours does not, let your vendor know how useful this would be.



## Using trend lines to clarify overall trends

It is often difficult to discern the overall trend across a period of time, especially when values exhibit a lot of volatility. It's hard to imagine how the general pattern might look if you could smooth out a jagged line of time-series values, taking all the increases and decreases into account to discern what's happening on average. Trend lines can solve this problem for us, but must be used with caution.



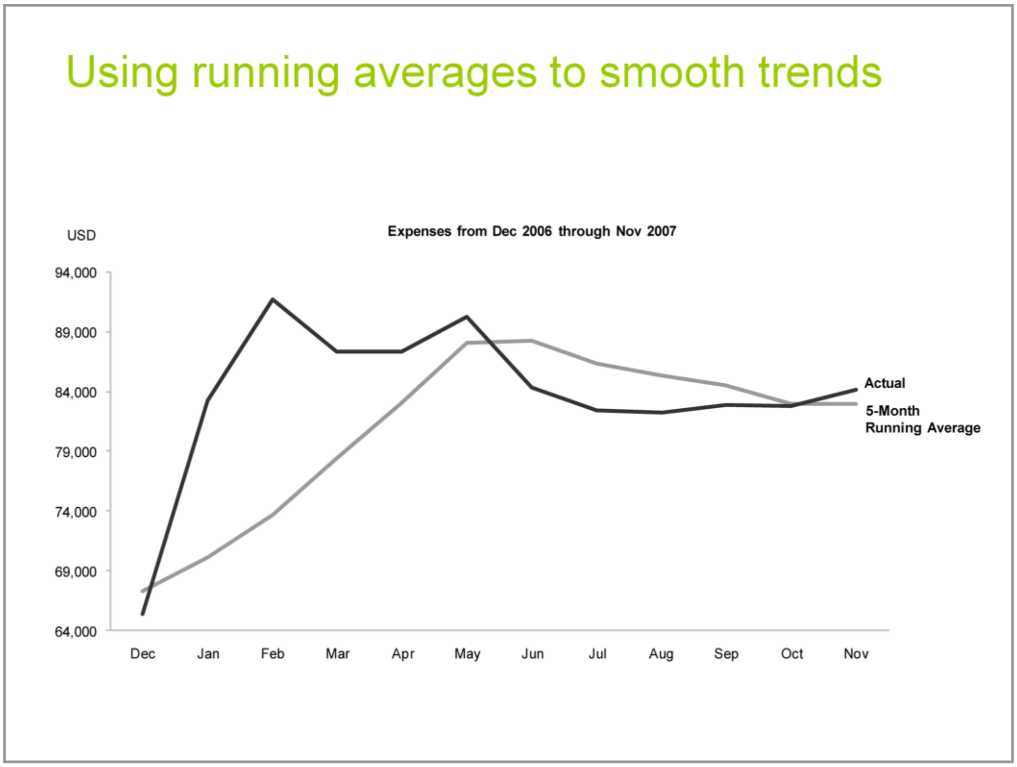
Bear in mind how easily trend lines can be swayed.

Bear in mind that whenever you take advantage of a software product's generous offer to draw a trend line for you, you are not only trusting it to do so accurately, you are asking it to display a trend in a particular stretch of time that will fail to consider what has happened before or after. In the top example above, I asked Excel to display the overall trend of expenses for the current year to date.

As you can see, the trend line indicates that expenses are trending downwards. Now look at how different the trend looks in the lower graph when I add a single month—December from the previous year—to include a full 12 months of expenses.

Quite a different trend, isn't it? Both graphs are accurate, based on the data they were asked to include when calculating the trend. When examining trends, be sure to examine data that falls outside the specified time period you're basing them on to make sure you haven't isolated a section that would trend quite differently if the period were longer.

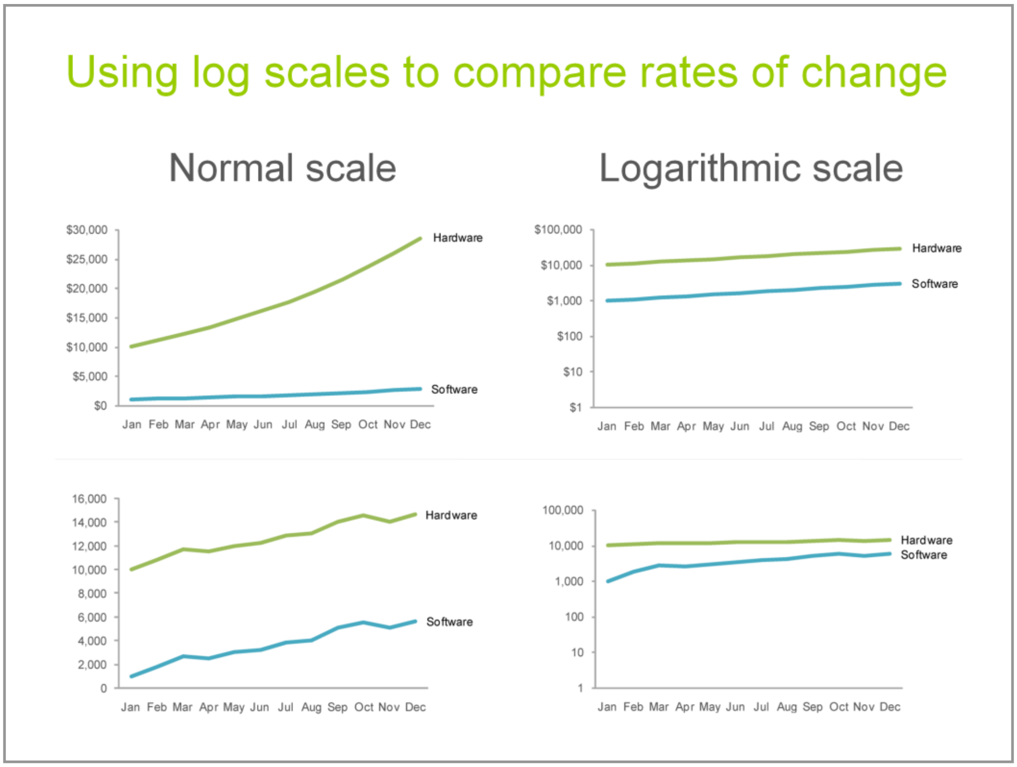
A straight line of best fit, which is what was appears in both examples above, is not the only type of trend line that can be used. This kind of trend line is based on a statistical calculation called a linear regression. It is determined by finding the straight line that passes through the full set of values in a graph from left to right such that the sum of the squares of the distance between each data point and the line is the least amount possible. Don't worry about the math involved—any software that displays trend lines will do the math for you.



## Using running averages to smooth trends

Variability in time-series values can be smoothed out to some degree by displaying, not the actual value for that period, but an average of that value and a few that precede it. For example, in the graph below we can see the pattern formed by taking the same values that appear in the two examples above and displaying each month's value as a five-month running average.

This example was created in Excel, simply by calculating each month's value as the average (mean) of that particular month and the four preceding it. It is often appropriate to examine time series from a smoothed (high-level) and an actual (low-level) value perspective at the same time, such as shown in the example below. Seeing both perspectives at once can help us avoid reading too much meaning into any one perspective.



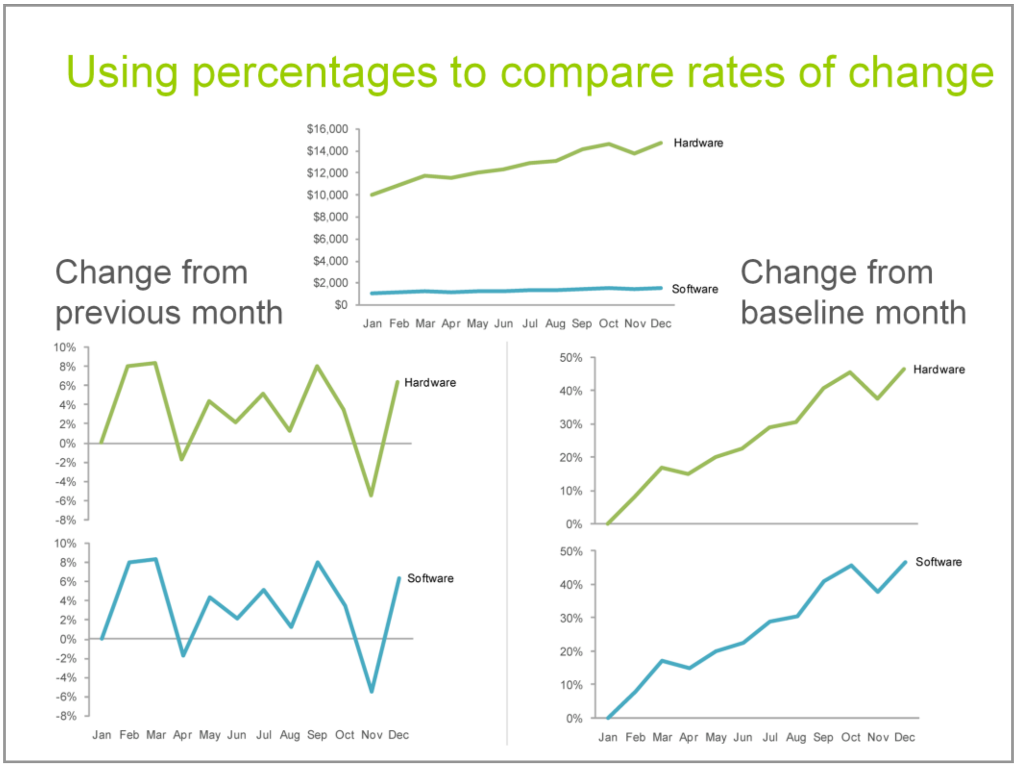
Using log scales to compare rates of change

Normal scale

Logarithmic scale

It is natural, when looking at a time-series graph like the one at the top on the left, to assume that the green line is increasing at a faster rate than the blue line, but in fact they are increasing at precisely the same rate. A 10% increase of $1,000 equals $100, while a 10% increase of $10,000 equals $1,000, and on a standard quantitative scale the slope of a line that increases by $100 is less steep than one that increases by $1,000. This does not hold true, however, for logarithmic scales. The same data appears in the graph at the top on the right, this time using a logarithmic scale. Equal rates of change on a logarithmic scale result in equal slopes, no matter how much the actual values are or the differences between them.

The second set of graph illustrates this from a slightly different perspective. Using a standard scale, the graph at the bottom on the left contains two lines that exhibit precisely the same visual patterns and slopes, which makes it look as if their rates of change are the same. The graph at the bottom on the right, however, uses a logarithmic scale to display the same data, which reveals that the rates of change for hardware and software were quite different.



Using percentages to compare rates of change

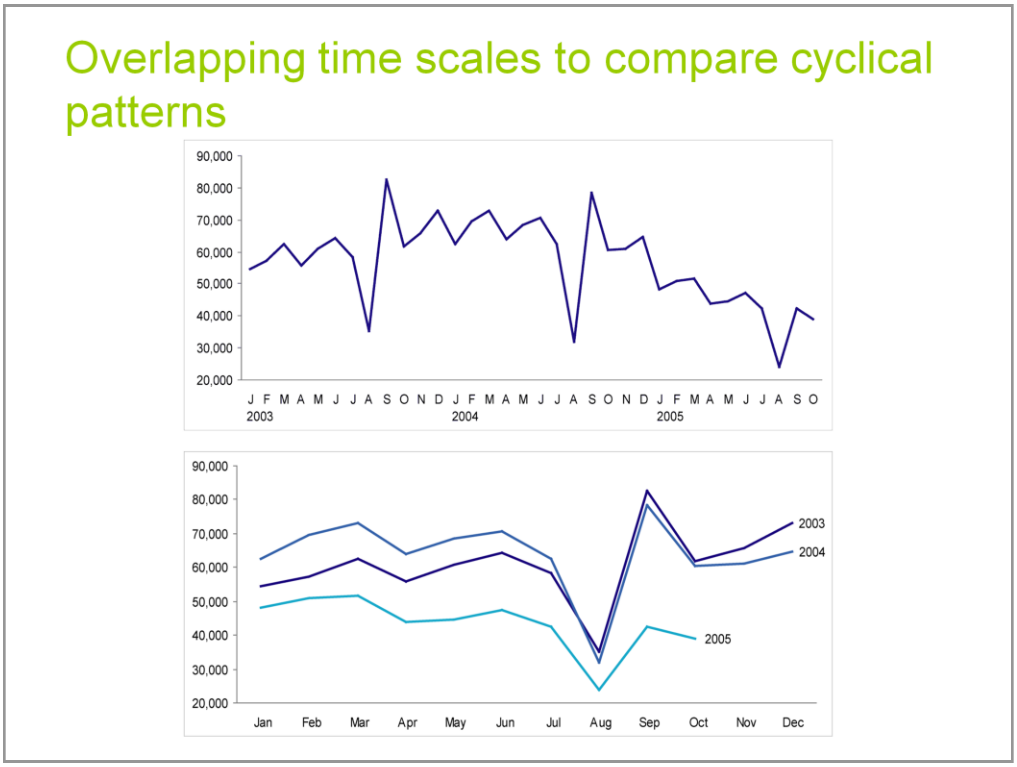
Change from previous month

Change from baseline month

Another way to make it easy to compare the rate of change is to graph the rate of change directly. Two methods are illustrated above:

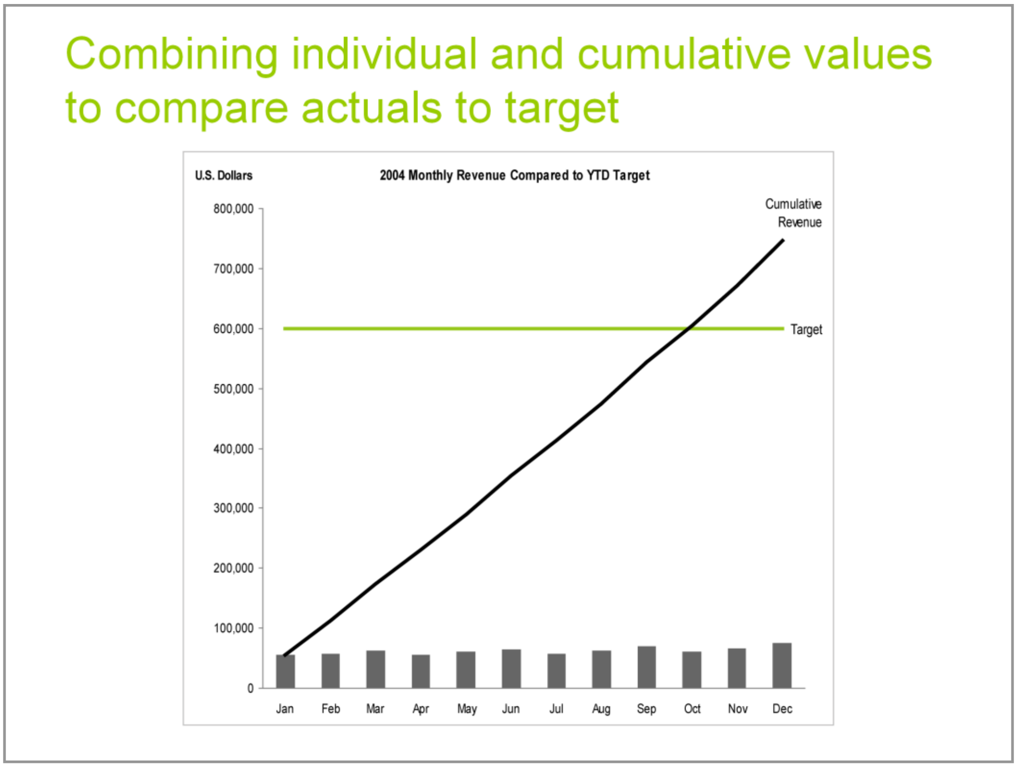
1. The one on the left displays each month as the percentage change compared to the prior month.
2. The one on the right displays each month as the percentage change from the first month, which in this case is January.

As you can see, when the rate of change from period to period is the same, the patterns and slopes that each of these methods display look precisely the same.



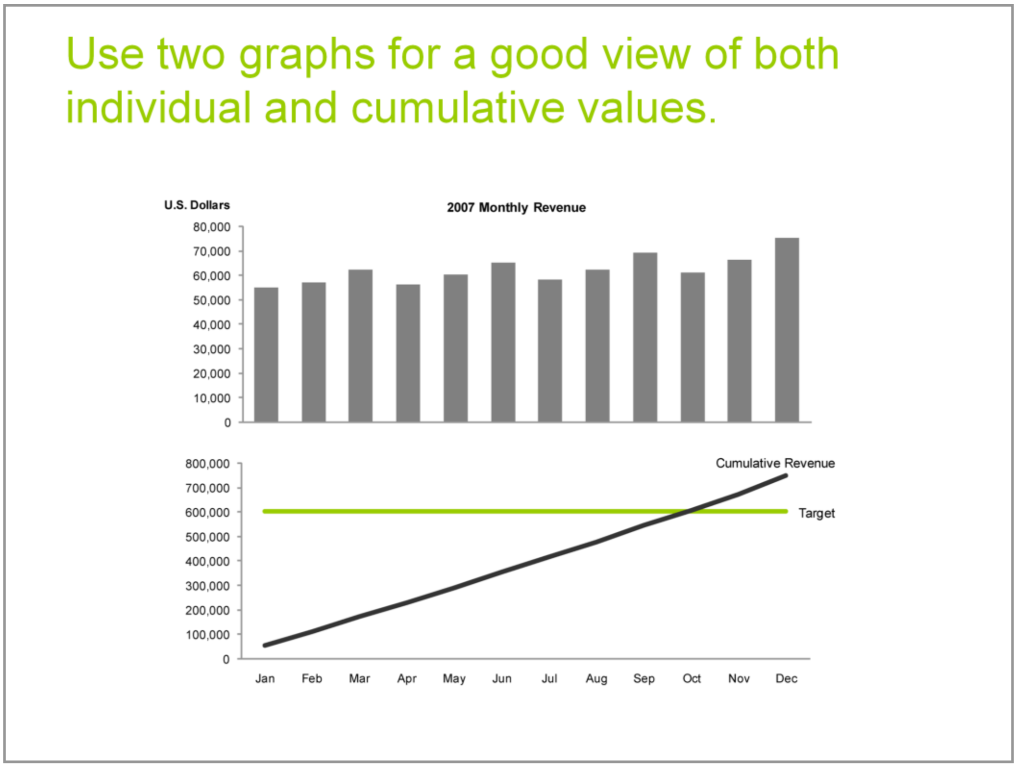
## Overlapping time scales to compare cyclical patterns

If you want to look for cyclical patterns in time-series data stretching across many cycles using a line graph, it is helpful to display each cycle as a separate line. As you can see in the graphs above, annual cyclical activity that can be detected to some degree in the top graph is much easier to see and compare in the bottom graph.



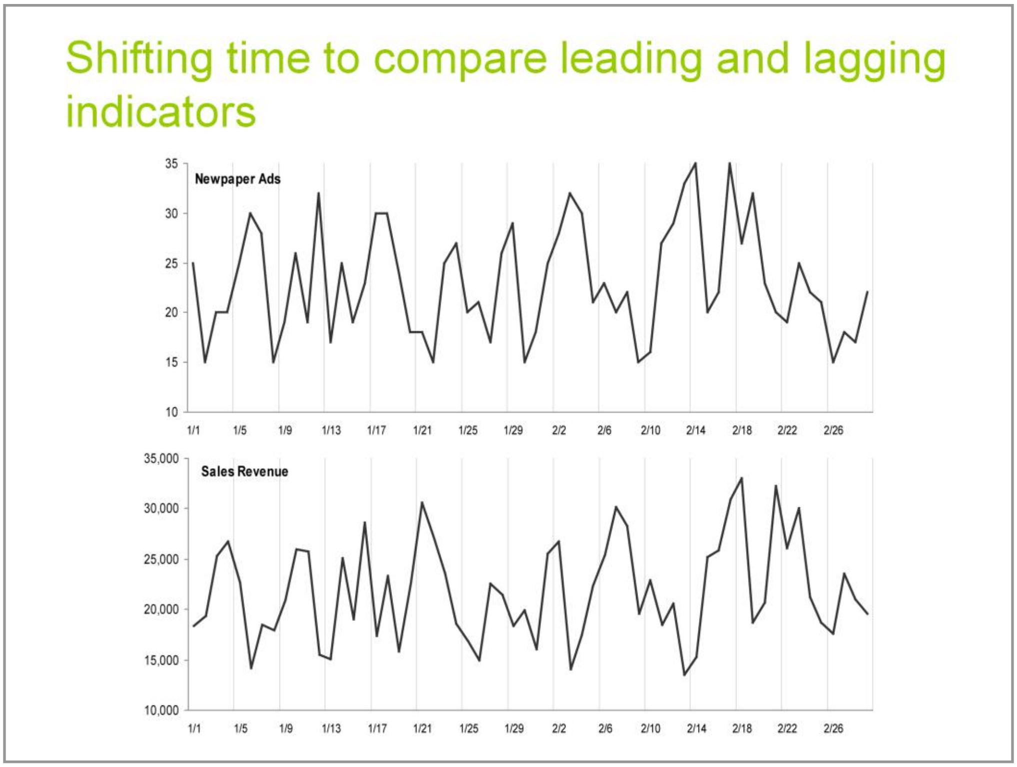
## Combining individual and cumulative values to compare actuals to target

Viewing cumulative values across units of time is often a good way to examine a measure's performance in relation to a target, such as a goal, when the target is cumulative in nature. For instance, you might want to view your approach to an annual target from month-to-month during the year. One way to display this is in the form of a graph that uses bars to encode the values for each month and a line to encode the cumulative values, with a separate line to encode the target.



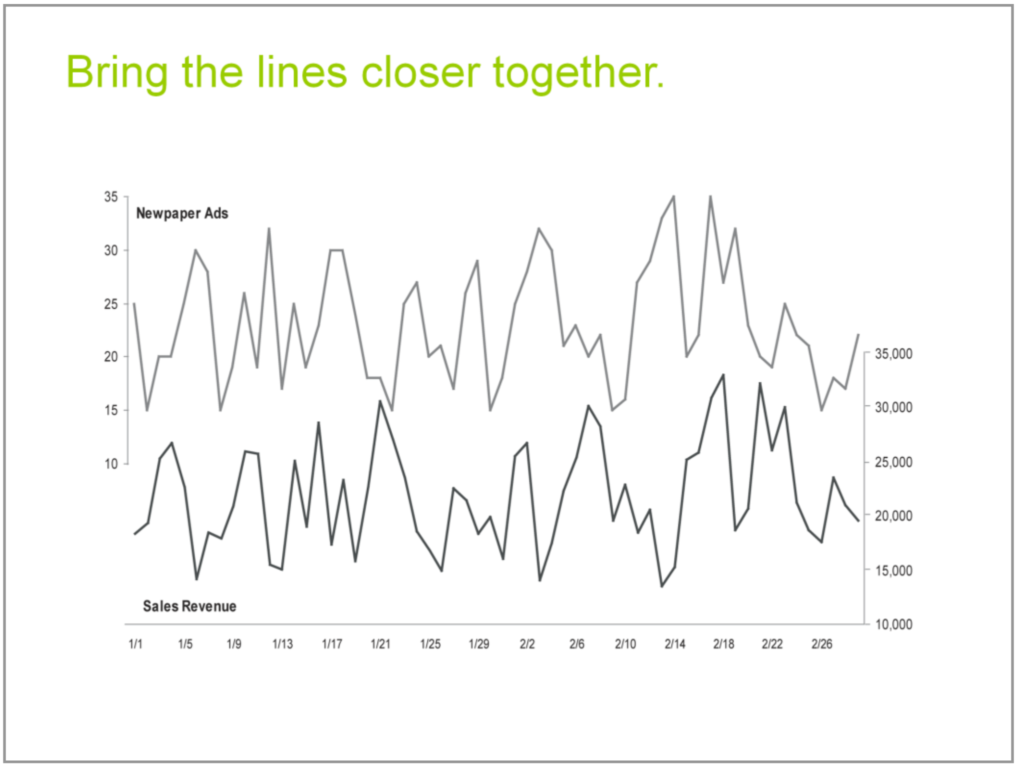
## Use two graphs for a good view of both individual and cumulative values.

The downside of combining individual and cumulative values in a single graph, however, is that by accommodating the quantitative scale required by the cumulative values, the individual values are squeezed down into a narrow range, which makes differences between them difficult to see. This problem can be solved by separating the individual and cumulative values into two graphs and positioning them closely, one above the other.



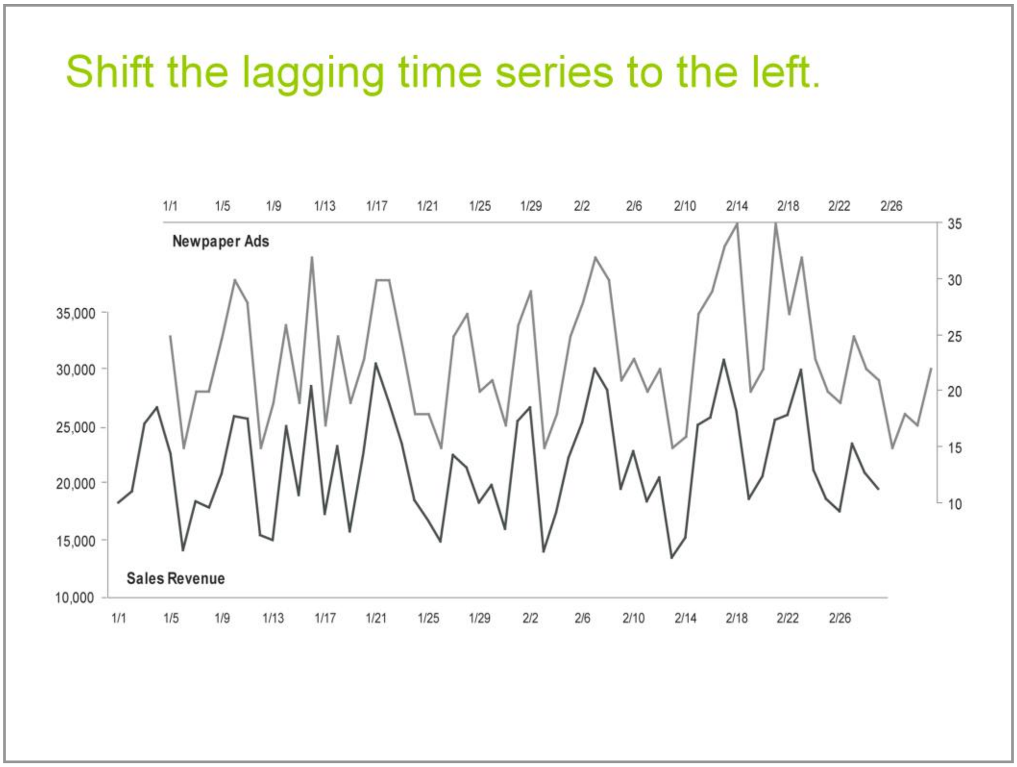
## Shifting time to compare leading and lagging indicators

It is often useful to examine the way one variable (independent variable) affects another (dependent variable). When we do so in the context of a time-series graph, it is sometimes difficult to see the result in the dependent variable as it relates to the independent variable when there is a lag in time between the cause and the effect.



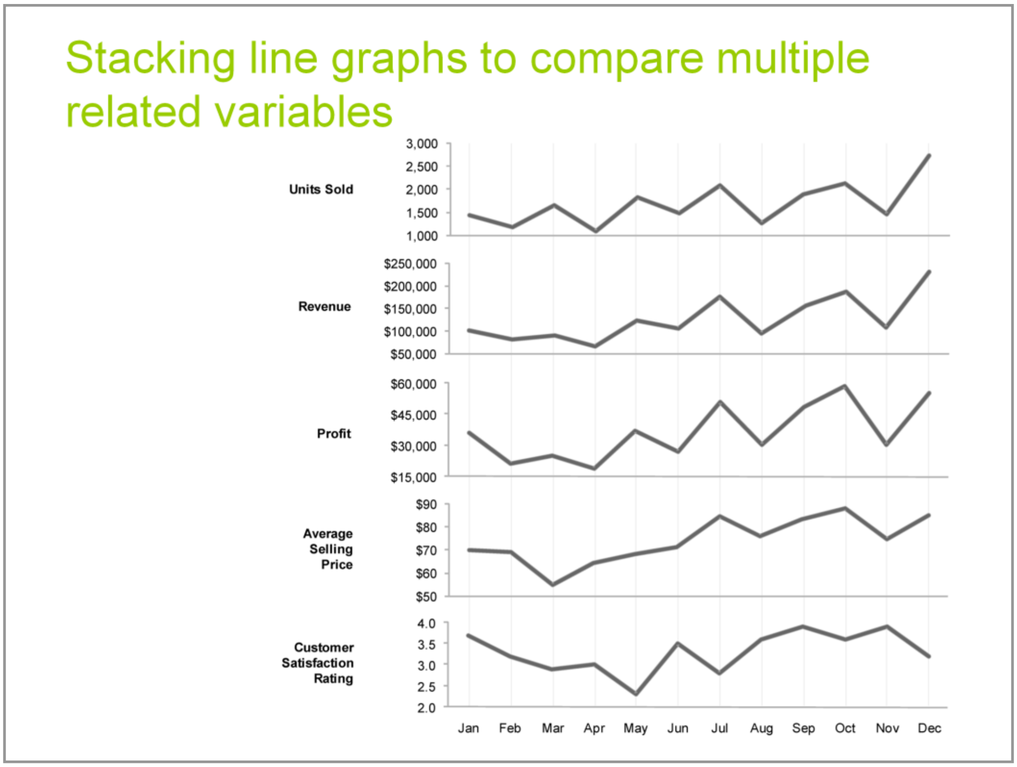
Bring the lines closer together.

To see this relationship more clearly, start by positioning the lines that represent each variable as close to one another as possible.



Shift the lagging time series to the left.

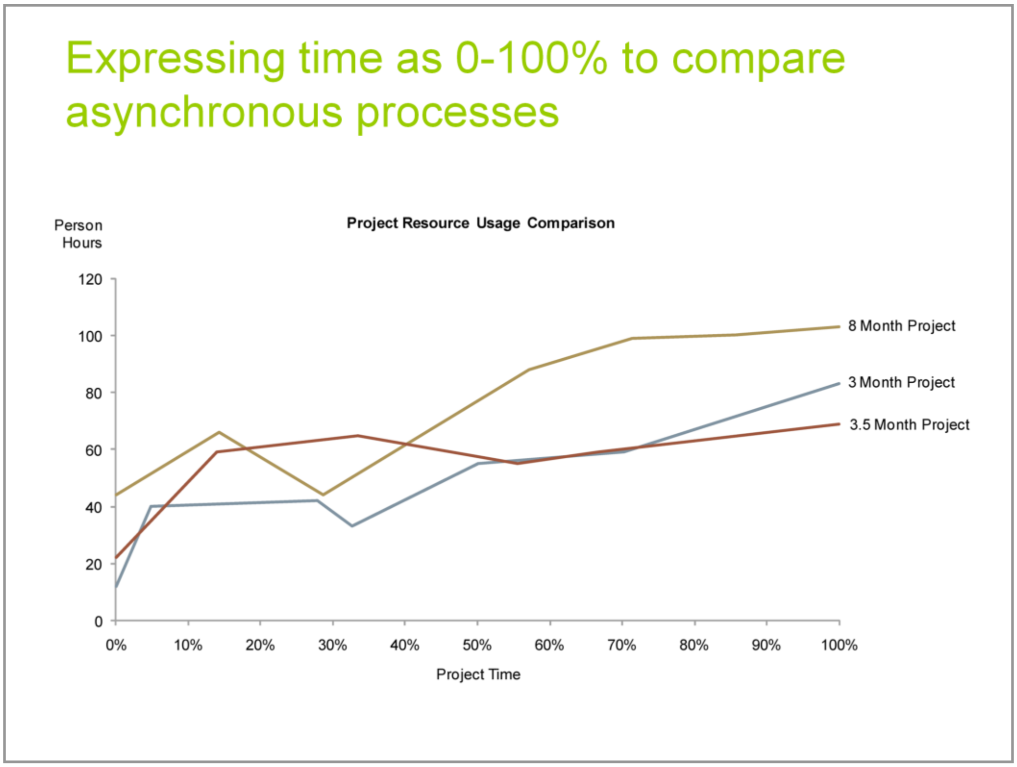
Next, shift the position of the graph that contains the dependent variable (lagging data) to the left to a degree that approximately equals the normal amount of time lag between the two events. This will align the causes and effects in time in a way that makes comparisons much easier.



Stacking line graphs to compare multiple related variables

It is often the case that data we need to compare when analyzing time series cannot all reside in a single graph, either because they are expressed in different units of measure, or there are huge differences in where they fall along the quantitative scale. For example, we cannot compare a product's sales revenues in U.S. dollars to the number of units sold in the same graph without using two scales, which can lead to confusion. Also, it is difficult to compare a product's profits to its average selling price in a single graph when monthly profits range in the millions of dollars and the average price per product is $99. Scaling the graph to accommodate values in the millions of dollars would cause the average selling prices to barely register as a straight line hugging the bottom of the plot area with no discernable pattern. But these are things that should be compared. This problem can be solved by using a series of graphs arranged vertically (stacked above and below one another) such that the same points in time in each graph are aligned.

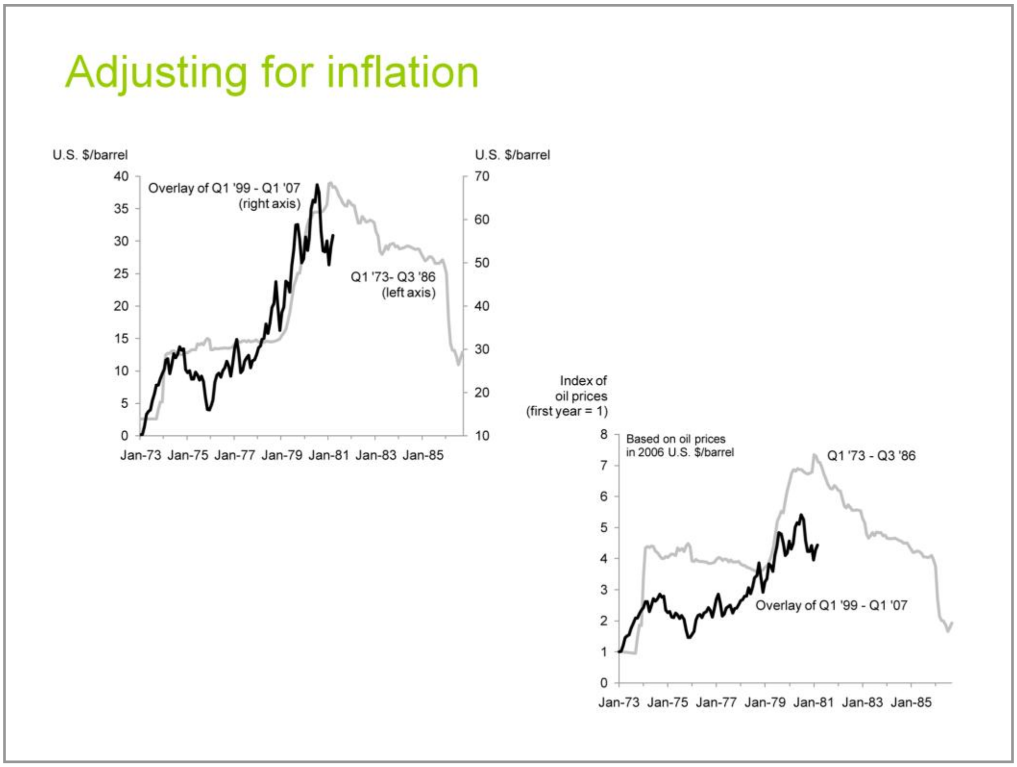
When graphs have different units of measure or even the same unit of measure, but their quantitative scales are not the same, you cannot compare the magnitudes of values in one graph to those in another, but you can compare patterns of change. This technique can be performed using Excel simply by creating separate graphs with the same time scale along their X-axes and arranging them vertically so that the same points in time are aligned. Other more powerful products are available for doing this, which reduce labor by arranging the graphs automatically.



Expressing time as 0-100% to compare asynchronous processes

Consider the following situation. You work in the Information Technology (IT) department of your company and you want to compare costs (in person hours) associated with 50 projects the department managed during the last five years. You're interested in finding out if project costs exhibit particular time-based patterns, such as high costs during the start-up phase, and increasing rate of costs near the end, or some other pattern that you can't even imagine. The problem that prevents you from analyzing these costs as you would any other time-series data is that the 50 projects did not all start at the same time and did not all last the same amount of time—they were asynchronous. What can you do?

One answer involves making time consistent for all projects—both starting time and duration. This can be done by expressing each project's duration as a percentage, beginning at 0% and ending at 100%, no matter when the project began or how long it lasted. This makes it possible to compare what's happening at the beginning of each process, at the end of each process, halfway through each process, 90% through each process, and so on, despite their asynchronous nature.



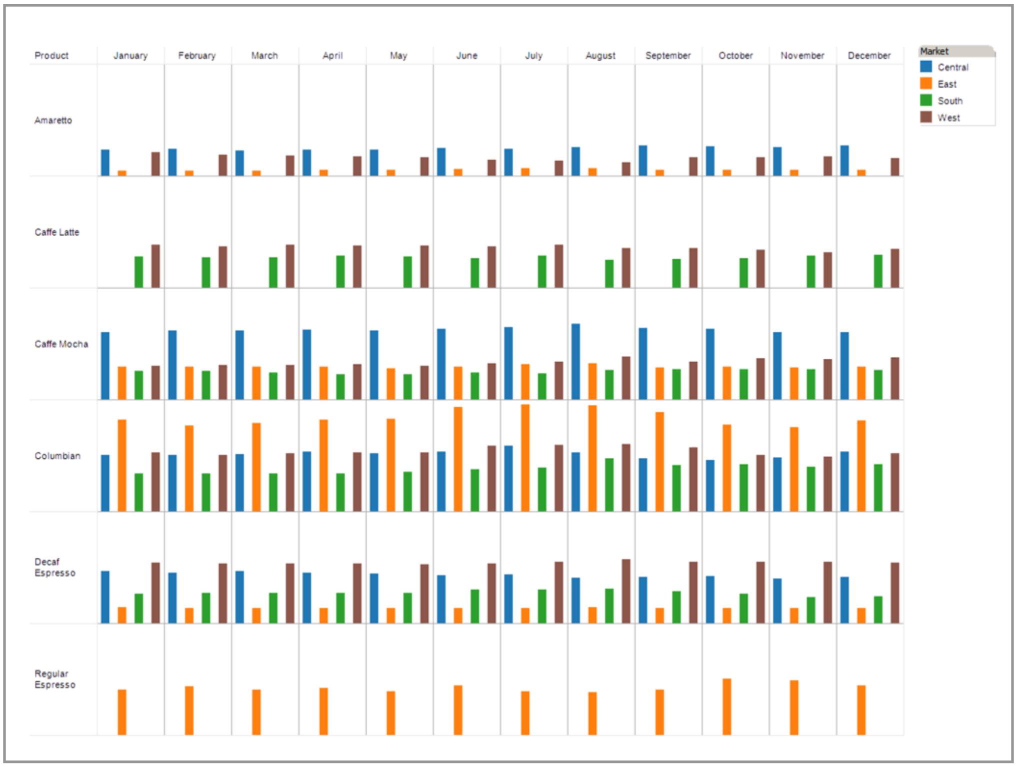
Adjusting for inflation

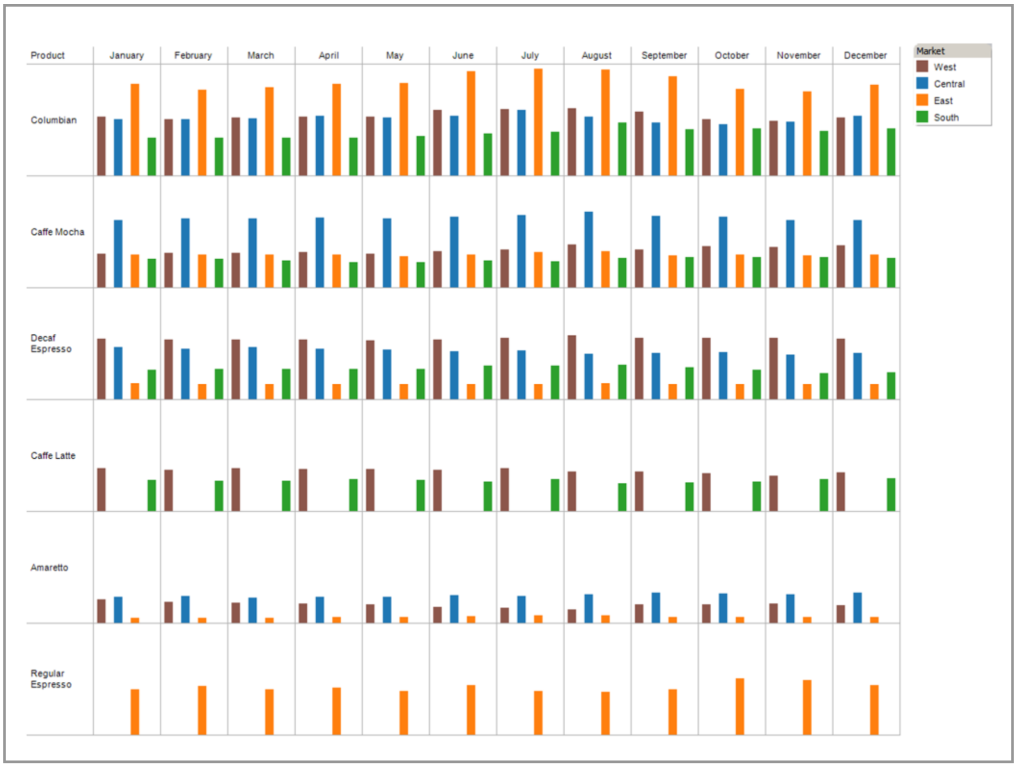
I've found that most analysis of business data involving money across multiple years fails to take inflation into account. When you wish to compare measures expressed as money across multiple years, comparisons will only be accurate if you take the changing value of money into account. Without doing so could cause you to conclude that the sales performance of a product is greater today than it was five years ago, when in fact performance has decreased.

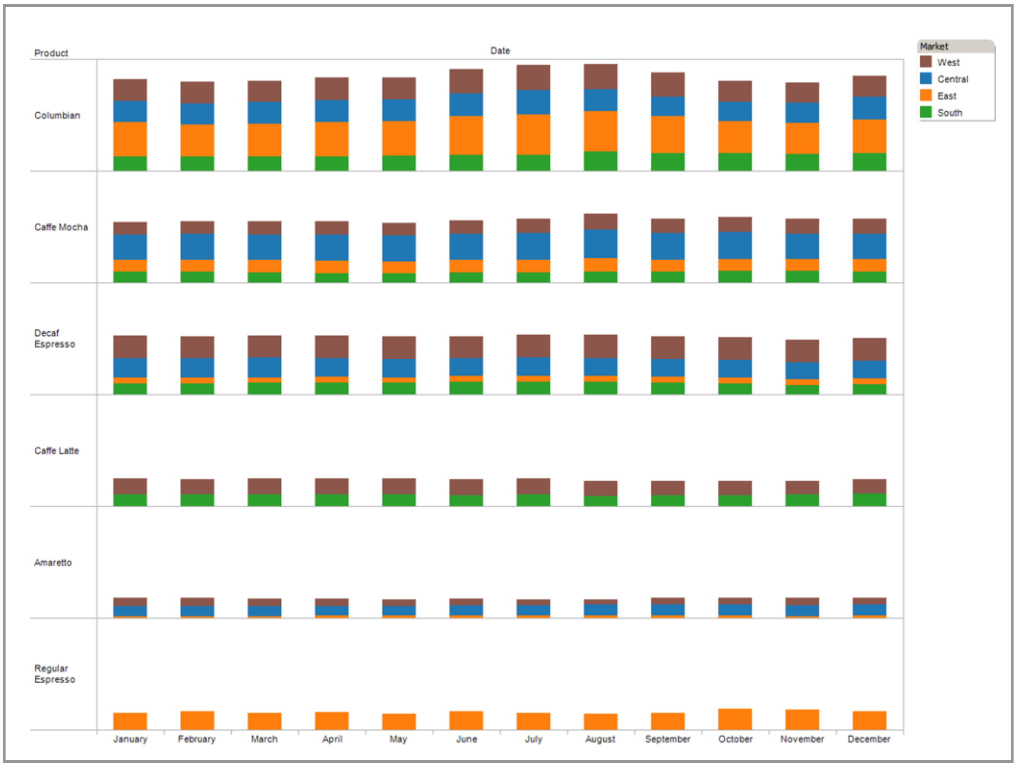
In September 2007, the investor newspaper *Barron's* published a graph like the one in the upper left, which compared the price of Saudi light crude oil from January 1973 through September 1986 to the price from January 1999 through March 2007. The reader was asked to compare two segments of the time series of oil prices and marvel at how similar they looked. "I've seen this movie before," said the person being interviewed about his predictions about future oil prices, implying that the similarity in the two graphs made a subsequent drop in oil prices likely. Regardless of the merits of that argument (and there are some) it is in this case based on false premises: It is simply incorrect to compare oil prices from two different time periods without correcting for inflation.

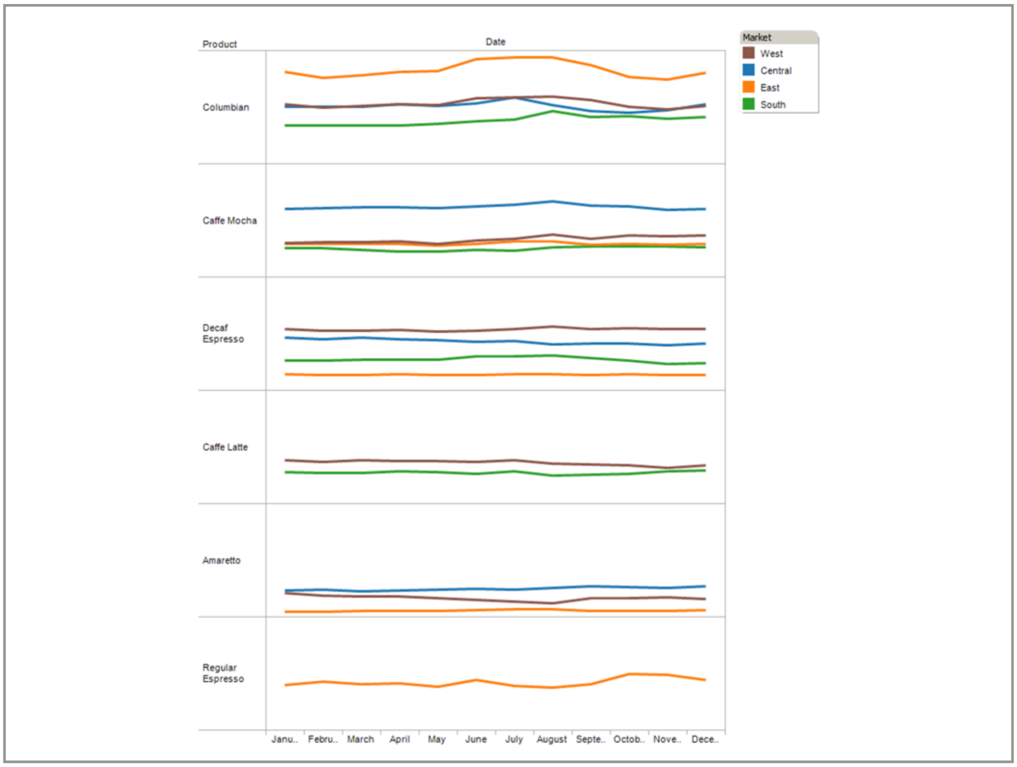
This graph was redesigned by Jonathan Koomey in an article titled "Inflation Matters", published in my *Visual Intelligence Newsletter*, April 2008. By adjusting for inflation, Koomey was able to show that the time-series patterns of change between these two periods of time were not nearly as alike as Barron's suggested.

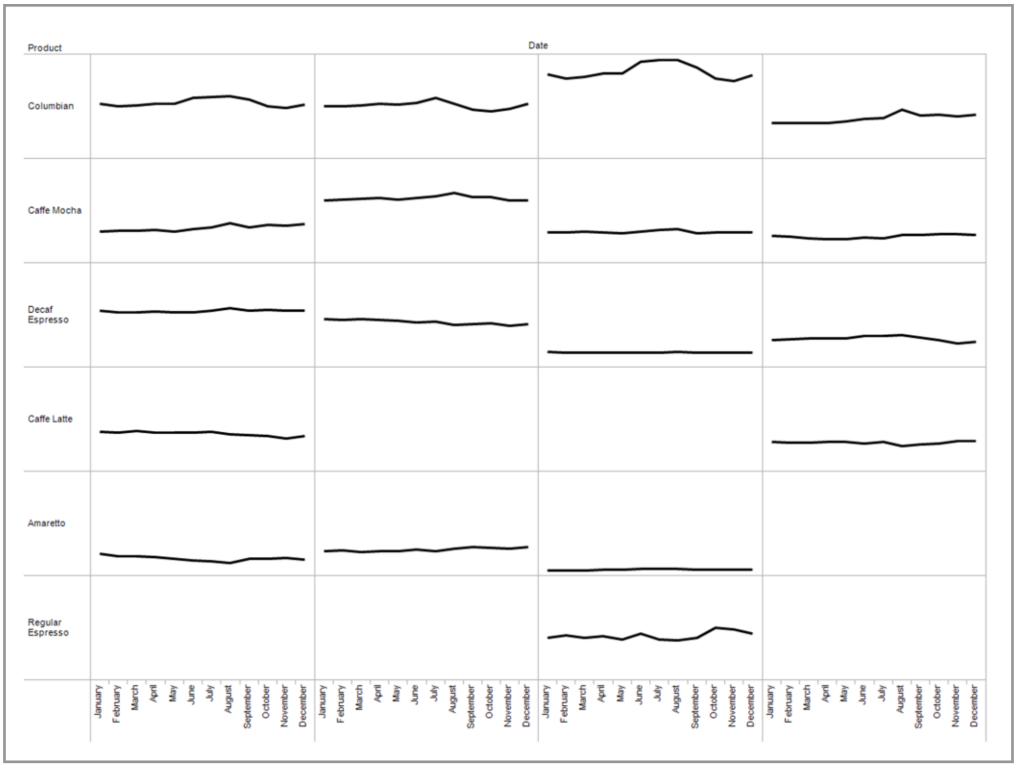
## Exercise: Analyzing time series

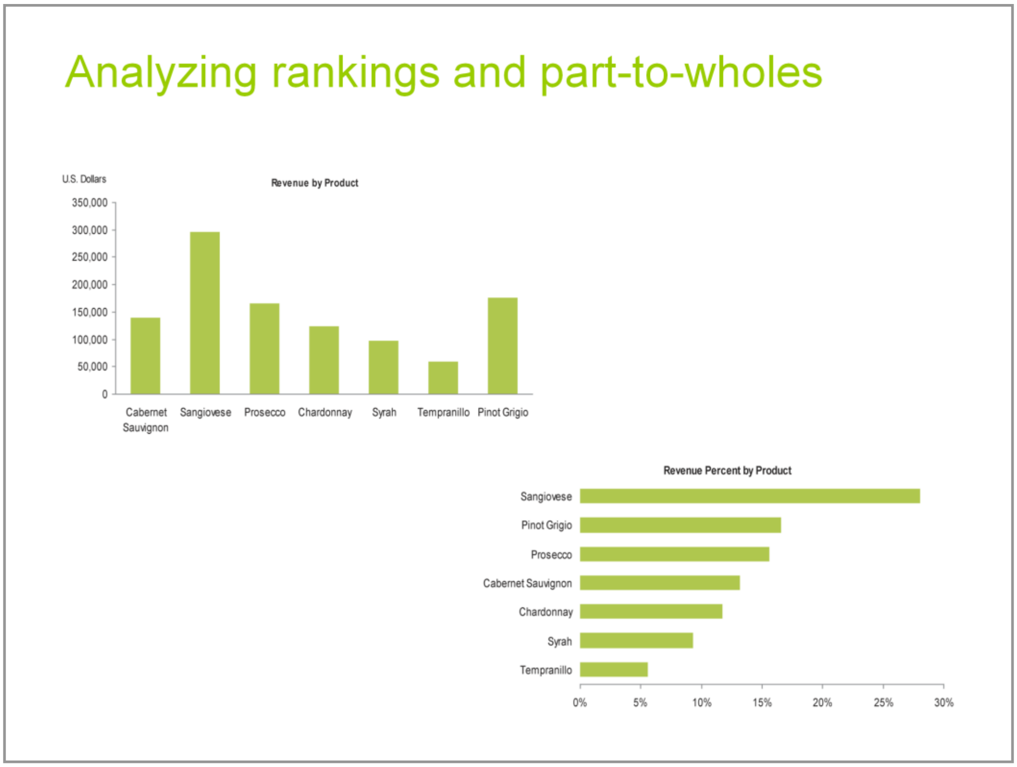

Product
January
February
March
April
May
June
July
August
September
October
November
December
Market
Central
East
South
West
Amaretto
Caffe Latte
Caffe Mocha
Columbian
Decaf Espresso
Regular Espresso


Product
January
February
March
April
May
June
July
August
September
October
November
December
Columbian
Caffe Mocha
Decaf Espresso
Caffe Latte
Amaretto
Regular Espresso
Market
West
Central
East
South


Product
Date
Market
West
Central
East
South
Columbian
Caffe Mocha
Decaf Espresso
Caffe Latte
Amaretto
Regular Espresso
January
February
March
April
May
June
July
August
September
October
November
December

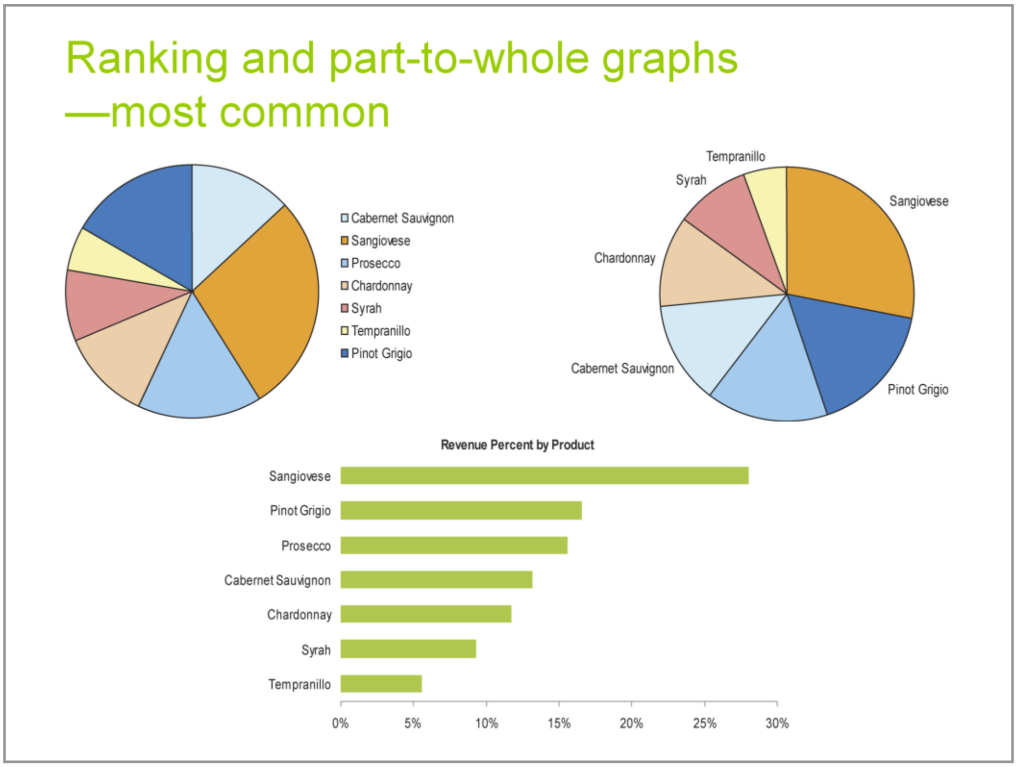

Product
Date
Market
West
Central
East
South
Columbian
Caffe Mocha
Decaf Espresso
Caffe Latte
Amaretto
Regular Espresso
Janu..
Febru..
March
April
May
June
July
August
Septe..
Octob..
Nove..
Dece..


Product
Date
Columbian
Caffe Mocha
Decaf Espresso
Caffe Latte
Amaretto
Regular Espresso
January
February
March
April
May
June
July
August
September
October
November
December



## Analyzing rankings and part-to-wholes

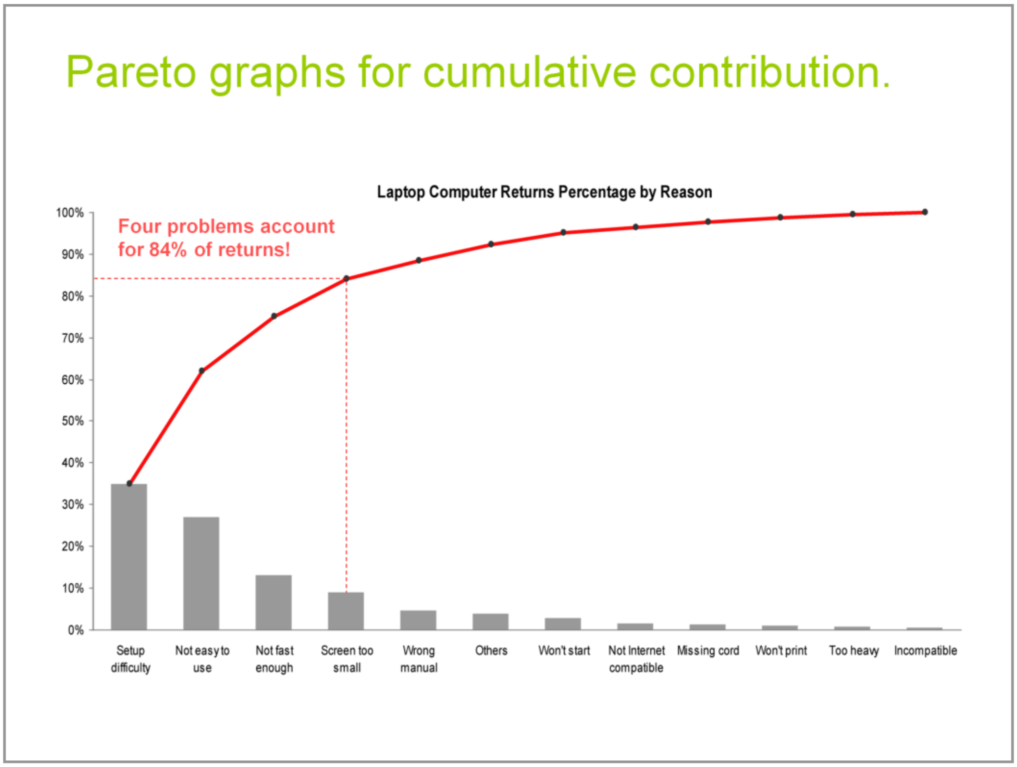
When you examine measures that have been divided into the parts of some larger grouping, such as individual products, regions, or departments, you often wish to see how these parts compare to another in size (ranking) and how they each compare to the whole (part-to-whole). Failing to sequence the items by size or to express them as percentages, as seen in the upper graph, makes these comparisons hard to make. Simple graphing techniques, like the lower graph, can be used to make these comparisons easy.



## Ranking and part-to-whole graphs —most common

Part-to-whole relationships are commonly displayed as pie charts. This is a problem, especially when you're trying to analyze the data, because pie charts rely on comparisons of the 2-D areas formed by slices of the pie, which we are not capable of doing easily or accurately. Pie charts are especially time-consuming to interpret when they use legends to label the slices, because you must constantly bounce your eyes back and forth between the legend and the pie to make sense of it. Even when slices are labeled directly, however, comparing them is difficult.

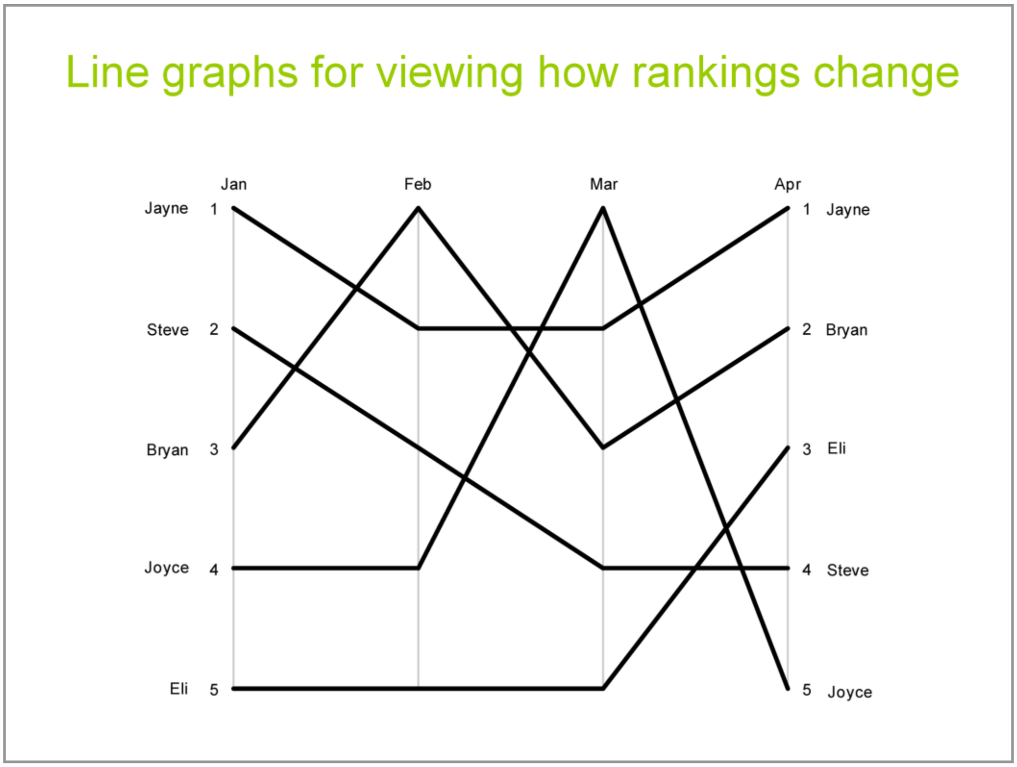
Bar graphs are much more effective for the analysis of ranking and part-to-whole relationships. What is difficult to see in the two pie charts is easy to see in the bar graph that displays the exact same data.



## Pareto graphs for cumulative contribution.

Sometimes when you examine values that have been ranked by size it is also insightful to examine the contribution of each part to the whole starting with the largest and working downwards in size. As illustrated in the above graph, it might be very useful to know that the top four of 12 total reasons that products are returned by buyers account for 84% of the total returns. Over half of the returns are due to only two reasons. The type of display illustrated in this graph is called a Pareto graph. To construct it, you simply rank the items by size, largest to smallest as one data set, and calculate cumulative values in the same order as a second data set, then graph the data with the individual values encoded as bars, the cumulative values encoded as a line, and express the values as percentages.

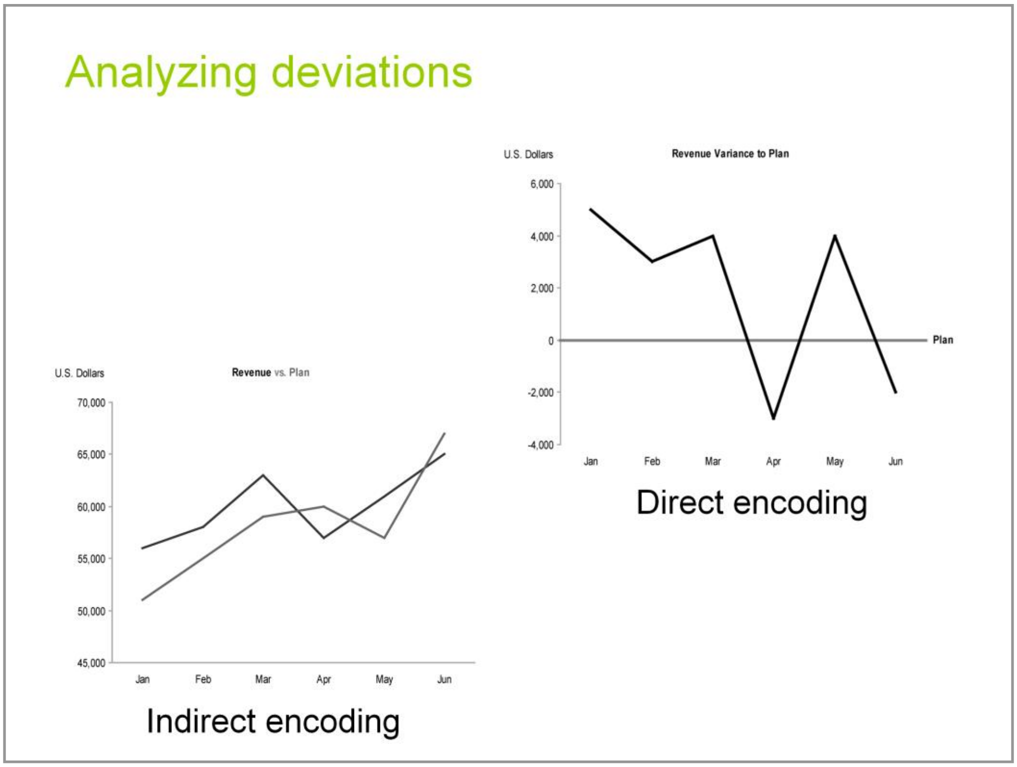
(Note: Technically, a Pareto chart does not require the line that encodes the cumulative value (this is optional), but including the line makes it much easier to see the cumulative contribution of the parts to the whole.)



Line graphs for viewing how rankings change

Although we would never use a line graph to display a single ranking relationship, when we want to see how a ranking relationship changes over time—a combination of a ranking and a time-series relationship—line graphs provide an excellent solution. This technique has been used for quite some time to display the results of rowing competitions using something called a *bumps chart*. Because collegiate rowing competitions in England traditionally take place on narrow rivers that prevent them from rowing side-by-side, crews are spaced with 1 ½ boat lengths between them at the start. When one crew rows so fast that it overtakes a crew ahead of it, the overtaken crew must pull over and allow the faster crew to pass. These races are called *bumps*, owing to the fact that a crew signals the fact that it has overtaken another by bumping it in some manner, often with an oar. Bumps competitions are spread across four days. The objective is to advance each day by overtaking one or more crews. On a bumps chart, each place where one line crosses another indicates that a boat has passed another. The line that slopes upwards represents the boat that overtook the other.

A line graph of a similar design can be used to display how the ranking relationship between a set of items, such as sales people ranked by sales performance, changes over time. The sole purpose in this case is simply to show changes in ranking, not the actual values associated with those changes, such as sales amounts. The slopes of the lines and their intersections provide strong visual cues for changes in rankings. This is a simple graph to construct, once you assign ranking positions (1, 2, etc.) to each of the items for each point in time. This example was constructed in Excel using a standard line graph.

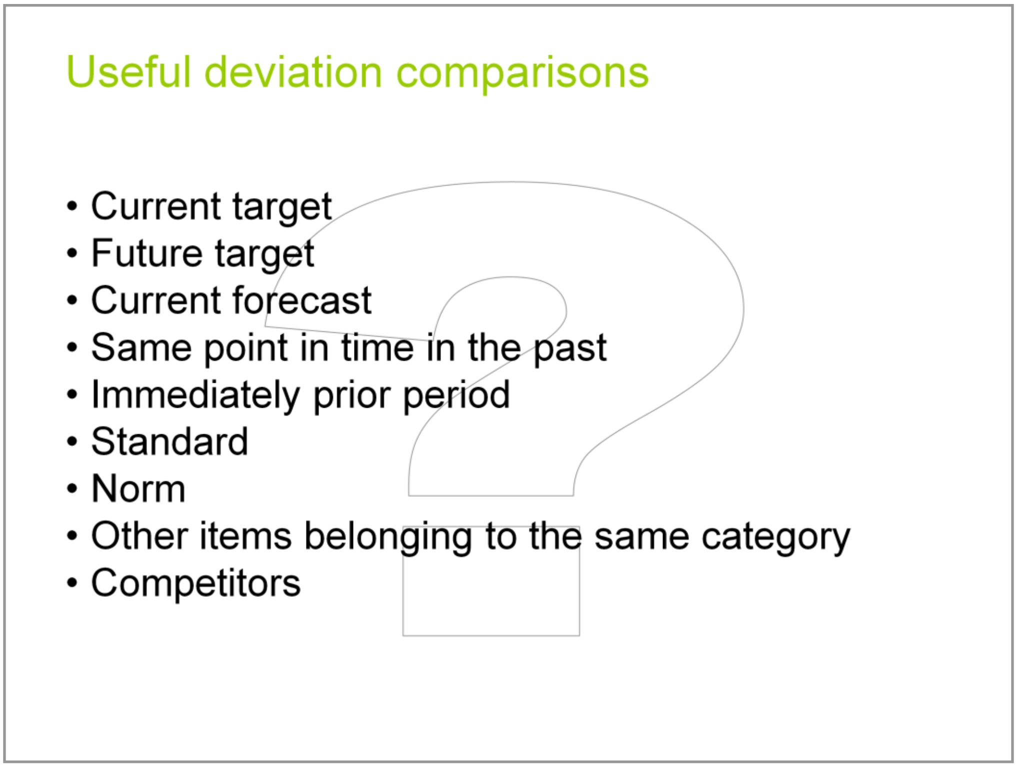


## Analyzing deviations

Direct encoding

Indirect encoding

Whenever you analyze quantitative data you are involved in making comparisons. Sometimes in doing so you focus on the differences between two sets of values. For instance, if you want to clearly understand the difference between revenues and expenses across time, displaying their actual values as separate lines as shown in the bottom graph is not the most direct way to reveal how they differ. Normalizing expenses to zero, 0%, or 100% across the entire span of time and then encoding revenues as a measure of the difference between them (plus or minus dollars in relation to zero dollars, plus or minus percentages in relation to 0%, or percentages compared to 100% expenses) results in a display that clearly and exclusively reveals the differences.

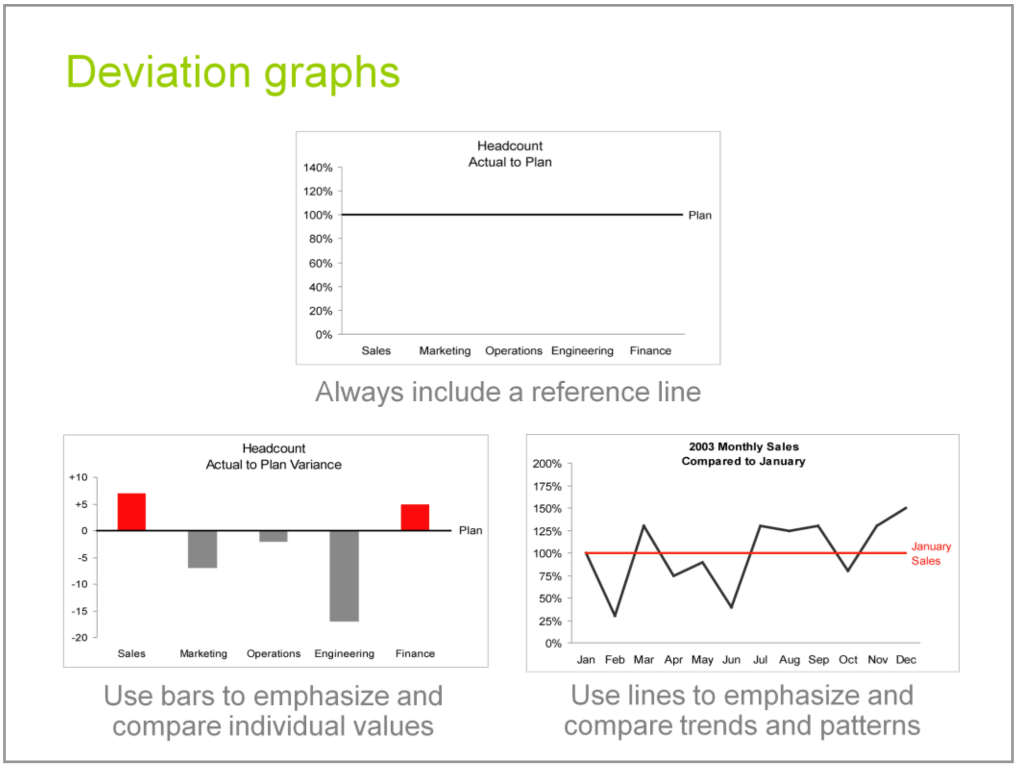


## Useful deviation comparisons

- Current target
- Future target
- Current forecast
- Same point in time in the past
- Immediately prior period
- Standard
- Norm
- Other items belonging to the same category
- Competitors

Deviation analysis can focus on a number of different meaningful comparisons, such as the following:

- Current target (such as a budget)
- Future target (such as percentage of annual goal)
- Current forecast (such as the sales forecast for the current quarter)
- Same point in time in the past (such as this day last year)
- Immediately prior period (such as the prior month)
- Standard (such as an acceptable number of defects)
- Norm (such as an average range)
- Other items belonging to the same category (such as different product)
- Competitors (other companies in the same market)



## Deviation graphs

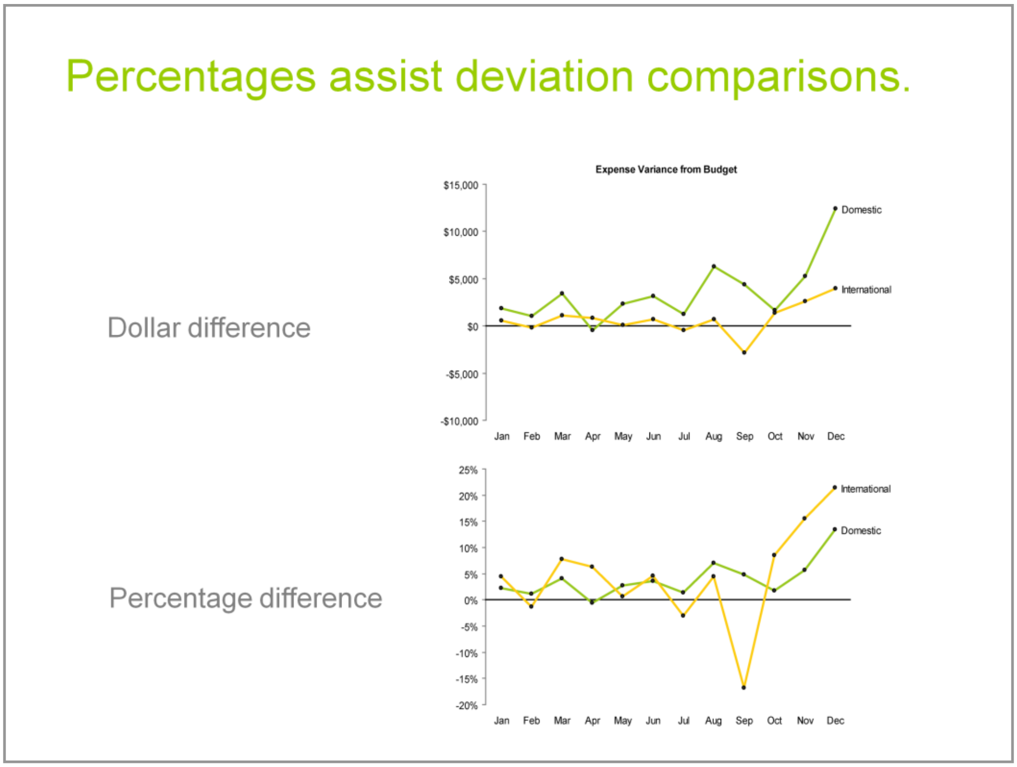
Always include a reference line

Use bars to emphasize and compare individual values

Use lines to emphasize and compare trends and patterns

The use of a reference line makes it clear that the main point of graphs like those pictured above is to display how one or more measures deviate from some point of reference. Bars can be used to encode data along nominal, ordinal, and interval scales. Lines must only be used to encode values along an interval scale, such as a time series, and are preferable to bars when you wish to focus on the overall shape of the data rather than each individual value or the comparison of individual values.

## Deviation analysis techniques and practices



## Percentages assist deviation comparisons.

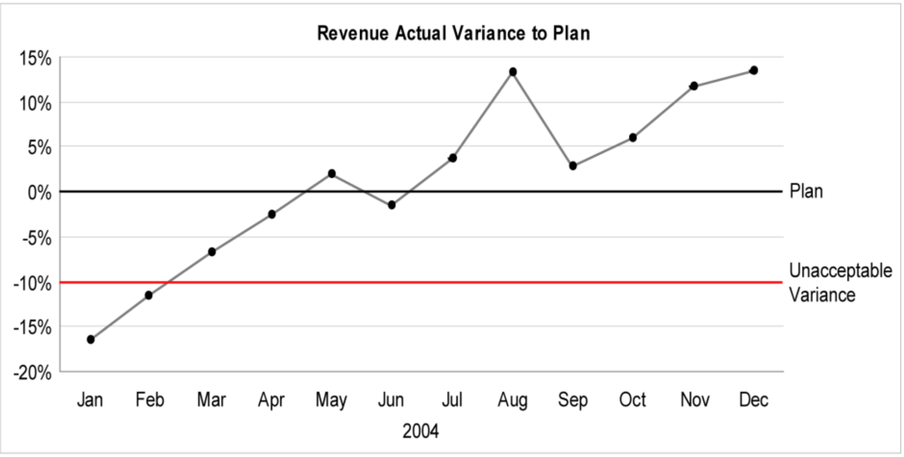
Dollar difference

Percentage difference

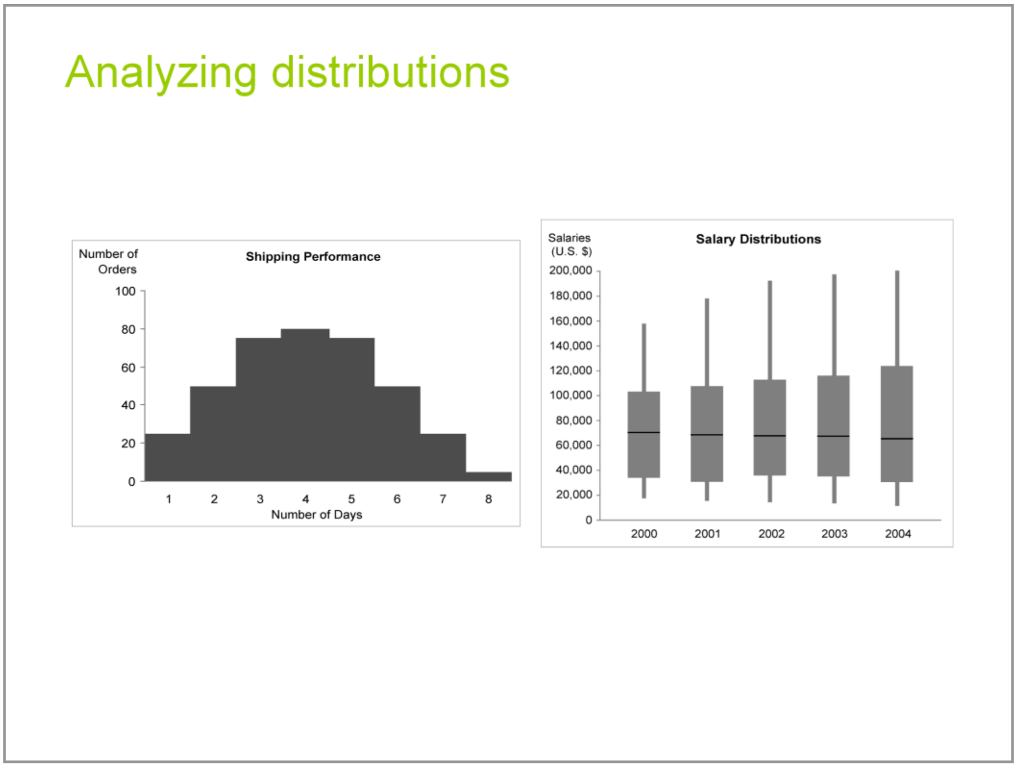
Deviations need not be expressed as percentages, but when they are, especially as plus or minus percentages, greater focus is given to the deviation, and when multiple data sets are being compared, the deviations are normalized in a way that makes comparisons easier.

There are times, however, when it is not practical to express deviations as percentages. For instance, if the values that you are comparing to a reference set of values fluctuate dramatically between small and large values, they could result in percentages that are huge, such as several hundreds or thousands of percent. Percentages work best when most of the values are less than or equal to 100% and exceptions do not exceed a few hundred percent.

## Deviations can be compared to acceptable ranges.

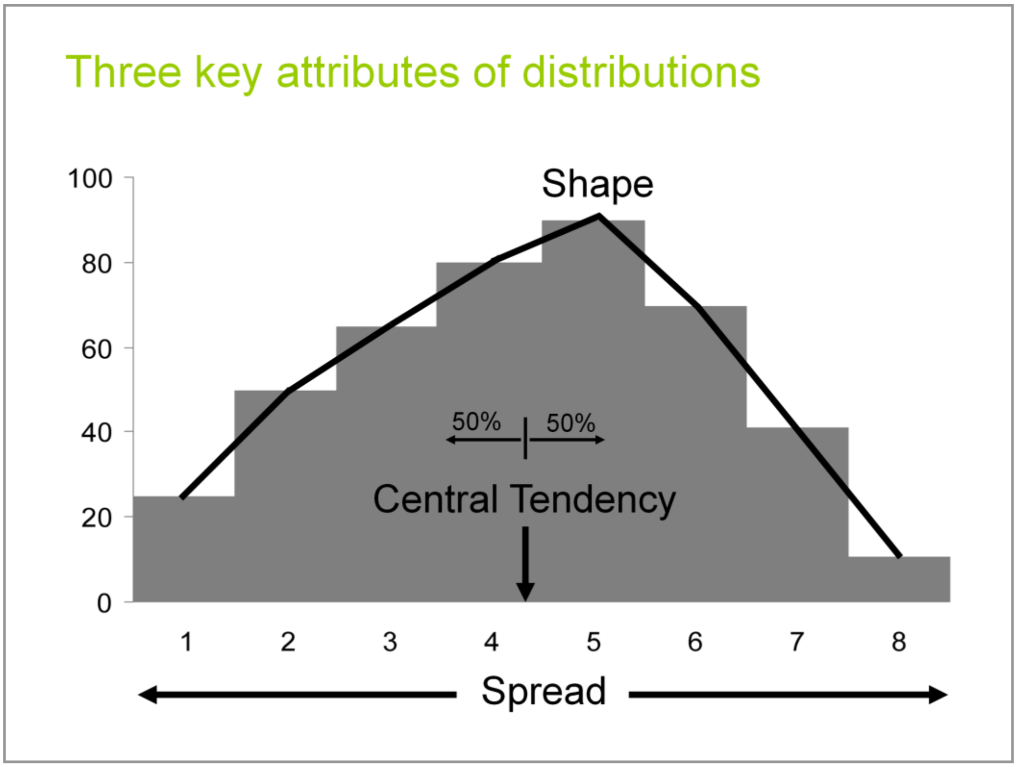


It is simple to add ranges of acceptability to your graphs, which makes exceptions clearly visible.



# Analyzing distributions

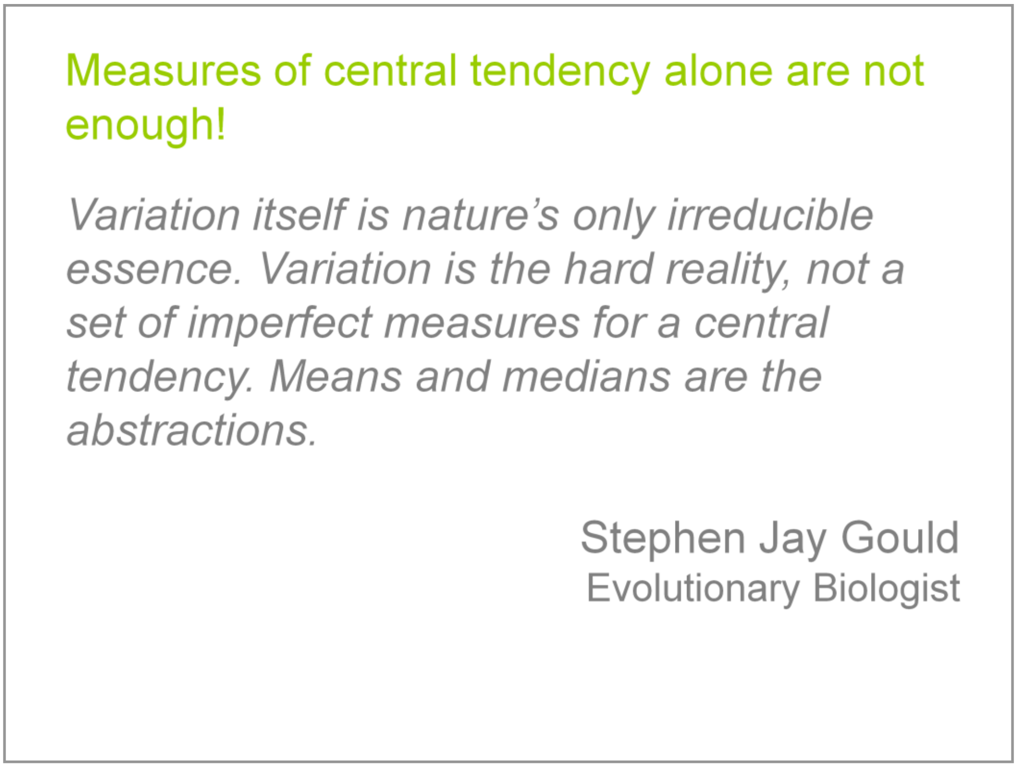
The distributions of one or more sets of values can tell compelling stories that are too often ignored by business analysts. Simply knowing that it takes an average (mean) of four days to ship orders doesn't tell you nearly as much as you might need to know. For instance, it could be taking five days to ship most of the orders or perhaps three days, but knowing the mean alone wouldn't reveal this. The mean salary earned by men and women in each pay grade could be close to the same and therefore suggest equity between them that masks the fact that a few women in each pay grade make extremely high salaries while the majority of women make much less than the majority of men. Knowing how to examine distributions is a useful skill.



## Three key attributes of distributions

> *[A] distribution can be described in terms of three important characteristics: location, spread, and shape. "Location" refers to the point at which the distribution is anchored, or located, on a continuum from the lowest to the highest possible value…To be effective, measures of location should identify the value most characteristic of a set of cases, the one value which best describes the entire set of values, or, in other words, the value around which the other values are distributed…The "spread" of the distribution refers to the variability or dispersion of cases, how wide the distribution is, how spread out the cases are…"Shape" is a bit more complicated, referring to the type of distribution, whether it is a bell-shaped normal distribution, symmetrical and single-peaked, whether it is skewed either to the right or the left, or multipeaked, whether it has outliers at the extremes or gaps within the distribution of the values, and so on."*
>
> (*Exploratory Data Analysis*, Frederick Hartwig with Brian E. Dearing, Sage Publications, Inc.: Thousand Oaks, CA, 1979, pages 13 and 14)

The term "location" is shorthand for the location of the central tendency in a distribution. Measures of central tendency come in two types: location and strength. In this workshop, when I refer to measures of central tendency, I am always referring to a measure of its location within the distribution. Rather than using the term "location" to describe the measure of a distribution's center, I prefer the term "central tendency," because it is clearer and more descriptive.



## Measures of central tendency alone are not enough!

*Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions.*

Stephen Jay Gould
Evolutionary Biologist

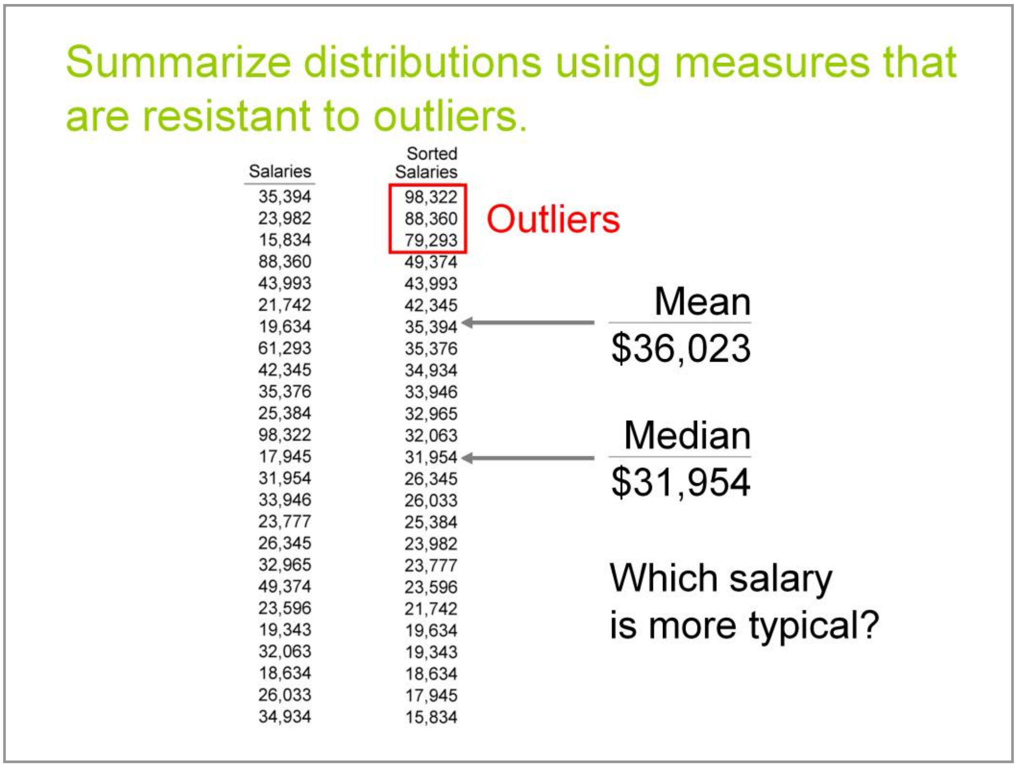
Measures of average are not enough. The eminent biologist, Stephen Jay Gould, learned a very personal lesson about the limited view that is contained in averages alone, which he wrote about in the article “The Median Isn't the Message.” In July 1982, Gould learned that he was suffering from *abdominal mesothelioma*, a rare and serious form of cancer, usually associated with exposure to asbestos. After surgery, he asked his doctor to recommend what he could read to learn about his condition, but was told that the literature wasn't very helpful. As soon as he could walk, he went immediately to Harvard's medical library to see for himself. After only an hour at the library, his doctor's attempt to discourage investigation became clear.

> *I realized with a gulp why my doctor had offered that humane advice. The literature couldn't have been more brutally clear: mesothelioma is incurable, with a median mortality of only eight months after discovery. I sat stunned for about fifteen minutes, then smiled and said to myself: so that's why they didn't give me anything to read. Then my mind started to work again, thank goodness.*

Most people, lacking an understanding of statistics, would have remained stunned and slipped into resignation, assuming that they had eight months at most to put their affairs in order. As someone who understood statistics, Gould realized that knowing that half the people with his condition survived only eight months or less was not enough, so he roused himself and continued his search for the full story. Later he wrote the following about our tendency to misinterpret measures of central tendency such as means and medians:

> *We still carry the historical baggage of a Platonic heritage that seeks sharp essences and definite boundaries…This Platonic heritage, with its emphasis in clear distinctions and separated immutable entities, leads us to view statistical measure of central tendency wrongly, indeed opposite to the appropriate interpretation in our actual world of variation, shadings, and continua. In short, we view means and medians as the hard “realities,” and the variation that permits their calculation as a set of transient and imperfect measurements of this hidden essence. If the median is the reality and variation around the median just a device for its calculation, the “I will probably be dead in eight months” may pass as a reasonable interpretation.*
>
> *Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions.*

His further investigation revealed a mortality distribution that was extremely skewed, extending to twenty years. After reading the conditions that favored longer survival, he realized that he was an ideal candidate for many more years of life. Revived by this hope, he in fact lived for 20 more very productive years and managed to publish his “Magnum Opus”, *The Structure of Evolutionary Theory,* just prior to his death in 2002.



## Summarize distributions using measures that are resistant to outliers.

| Salaries | Sorted Salaries |
|---|---|
| 35,394 | 98,322 |
| 23,982 | 88,360 |
| 15,834 | 79,293 |
| 88,360 | 49,374 |
| 43,993 | 43,993 |
| 21,742 | 42,345 |
| 19,634 | 35,394 |
| 61,293 | 35,376 |
| 42,345 | 34,934 |
| 35,376 | 33,946 |
| 25,384 | 32,965 |
| 98,322 | 32,063 |
| 17,945 | 31,954 |
| 31,954 | 26,345 |
| 33,946 | 26,033 |
| 23,777 | 25,384 |
| 26,345 | 23,982 |
| 32,965 | 23,777 |
| 49,374 | 23,596 |
| 23,596 | 21,742 |
| 19,343 | 19,634 |
| 32,063 | 19,343 |
| 18,634 | 18,634 |
| 26,033 | 17,945 |
| 34,934 | 15,834 |

Outliers

Mean
$36,023

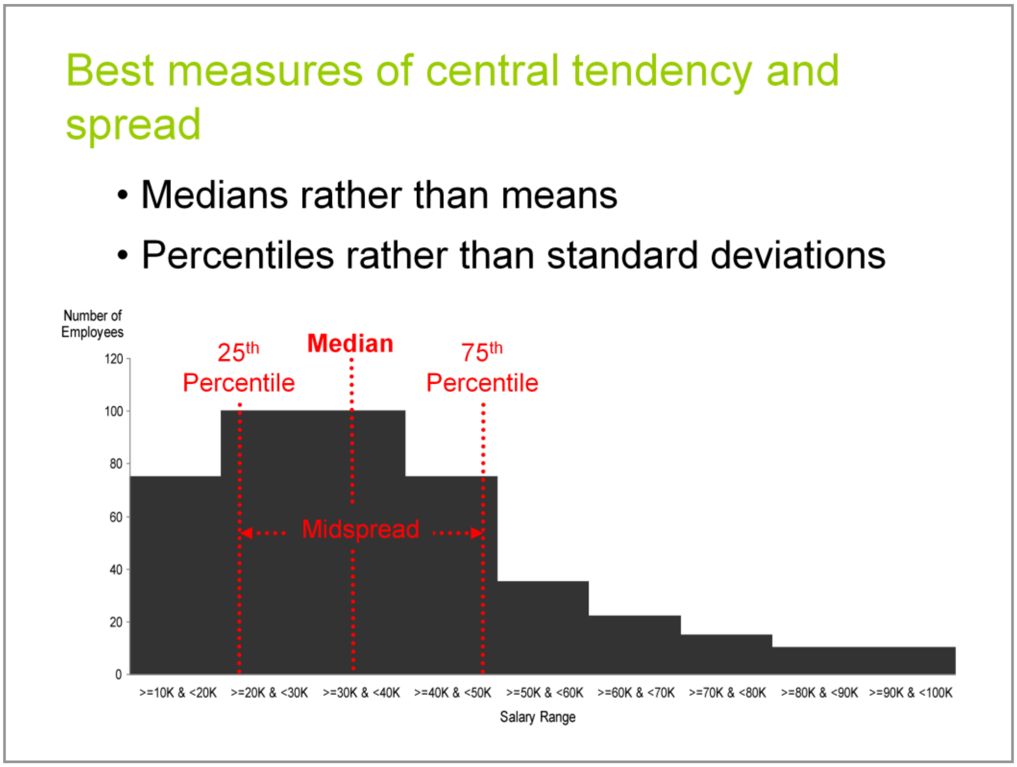
Median
$31,954

Which salary
is more typical?

When analyzing distributions of values, it is almost always necessary the summarize the distribution in some manner, rather than examining the entire set of individual values. Just knowing the highest and lowest values (the spread) is not enough because that tells you nothing about the shape of the values as their spread across that range or where most of the values are located. All summaries of a distribution revolve around some measure of its central tendency.

The *mean*, calculated by adding up the values in the entire set and dividing by the number of values in the set, is a measure of center that suffers from being highly influenced by outliers (extreme values). If it is your purpose to summarize the total financial impact of salary expenses per employee with a single measure of center, the mean works well. If you want to express the salary that is most typical, however, the mean isn't the best measure. The *standard deviation* is a measure of distribution around the mean, which, because it is based on the mean, is also heavily influenced by outliers. As a result, means and standard deviations are not the best measures of a distribution's center when you want to understand what is typical, because they are not resistant to abnormal values.

The *median*, determined by sorting the entire set of values by size and then selecting the value in the exact middle of the set, is very resistant to outliers, and thus a much better measure of central tendency when you wish to understand what is typical. Percentiles, as measures of distribution around the median are also resistant to outliers, because they are based on the median.



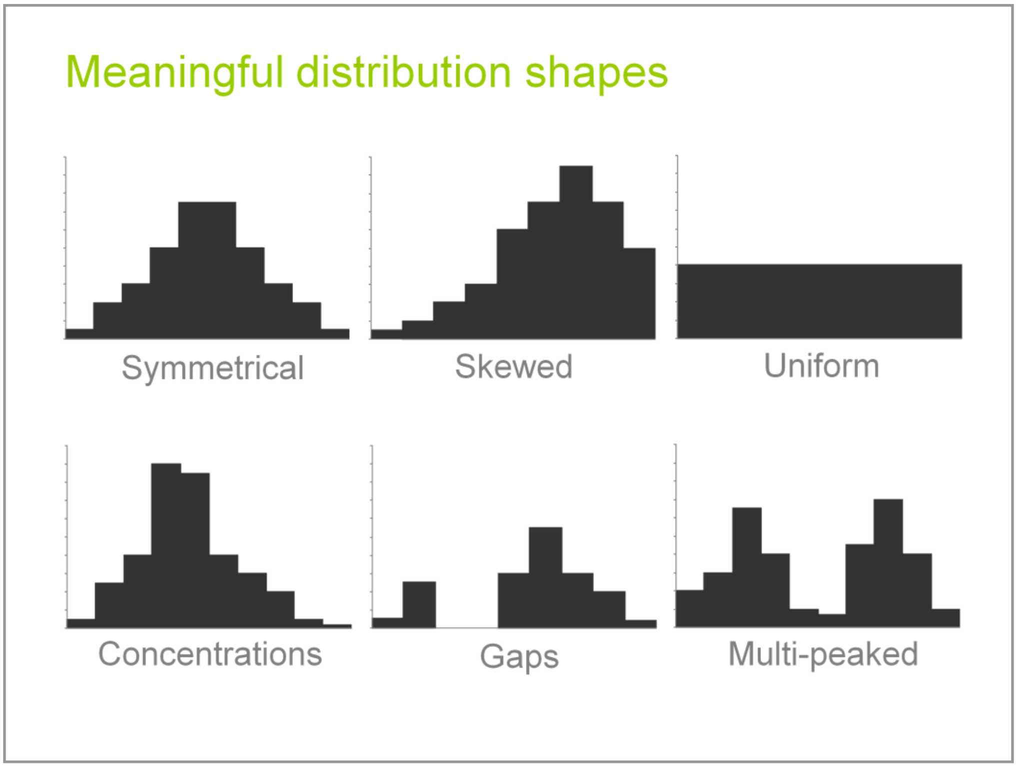
## Best measures of central tendency and spread

- Medians rather than means
- Percentiles rather than standard deviations

*A resistant measure of either location* [a.k.a., central tendency] *or spread is one that is relatively unaffected by or resistant to changes, no matter how large, in a small proportion of the total number of cases. Because statistics which lack resistance can be sensitive to a small number of values within a distribution, usually in the tails of the distribution where there are few cases, they may not accurately describe the bulk of the cases in the middle of the distribution…The standard deviation is a particularly nonresistant statistic and, in fact, is highly sensitive to a few extreme values…What makes the standard deviation so nonresistant is the fact that the deviations from the mean are squared.*

*"…the simple and most useful resistant measures are the order statistics, so called because they are based upon the rank order of the values in a distribution. The median, for example, is that value above which and below which fall one-half of the values in a rank-ordered list…The lower hinge is that point above which three-fourths and below which one-fourth of the values line (the bottom quartile), and the upper hinge is that point above which lie one-fourth of the values (the top quartile) and below which lie the other three-fourths. The distance between the lower hinge and the upper hinge is sometimes referred to by the imposing name 'interquartile range,' but midspread is both easier and more descriptive…A resistant summary of some of the major features of a distribution can be created by combining the median and hinges with the highest and lowest values. This produces a five-number summary…"*

(*Exploratory Data Analysis*, Frederick Hartwig with Brian E. Dearing, Sage Publications, Inc.: Thousand Oaks, CA, 197, pages 19-21)



## Meaningful distribution shapes

Symmetrical

Skewed

Uniform

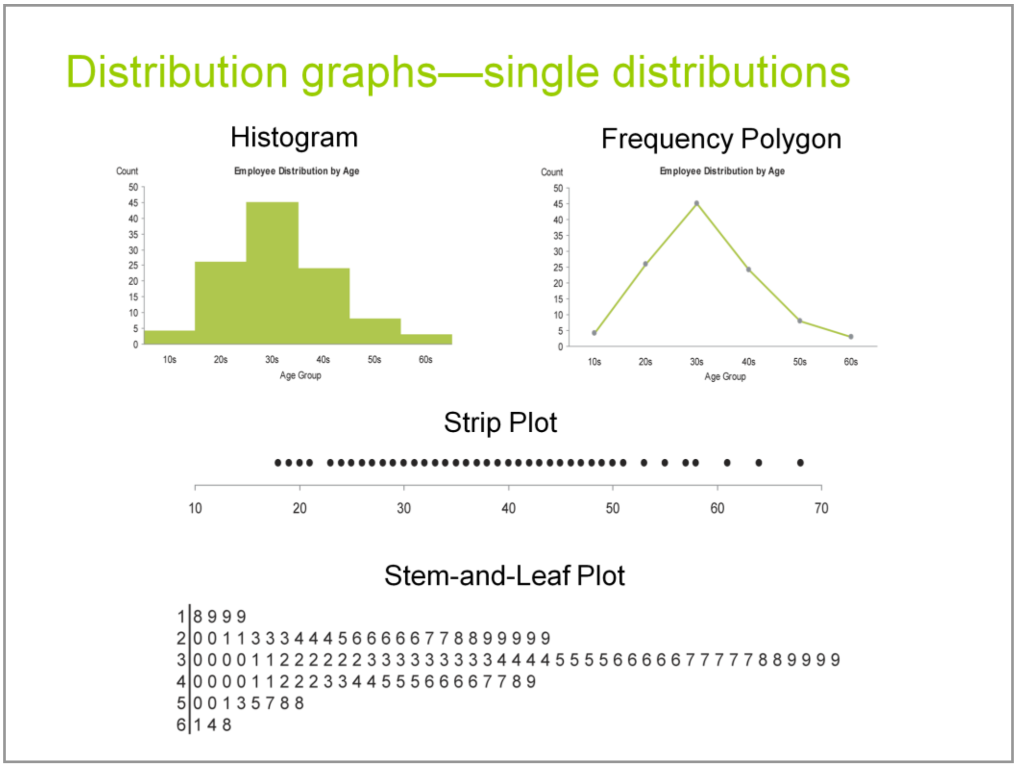
Concentrations

Gaps

Multi-peaked

The shape of a distribution can reveal several interesting patterns, including the following:

- Symmetrical
- Skewed (Rather than a symmetrical shape, most of the data are located to the right or the left of the distribution. If the long tail (as opposed to end where the values are greatest) extends to the right, the distribution is right skewed, and if to the left, it is left skewed, as in the example of a skewed distribution above.)
- Uniform
- Concentrations
- Gaps
- Multi-peaked



## Distribution graphs—single distributions

Histogram

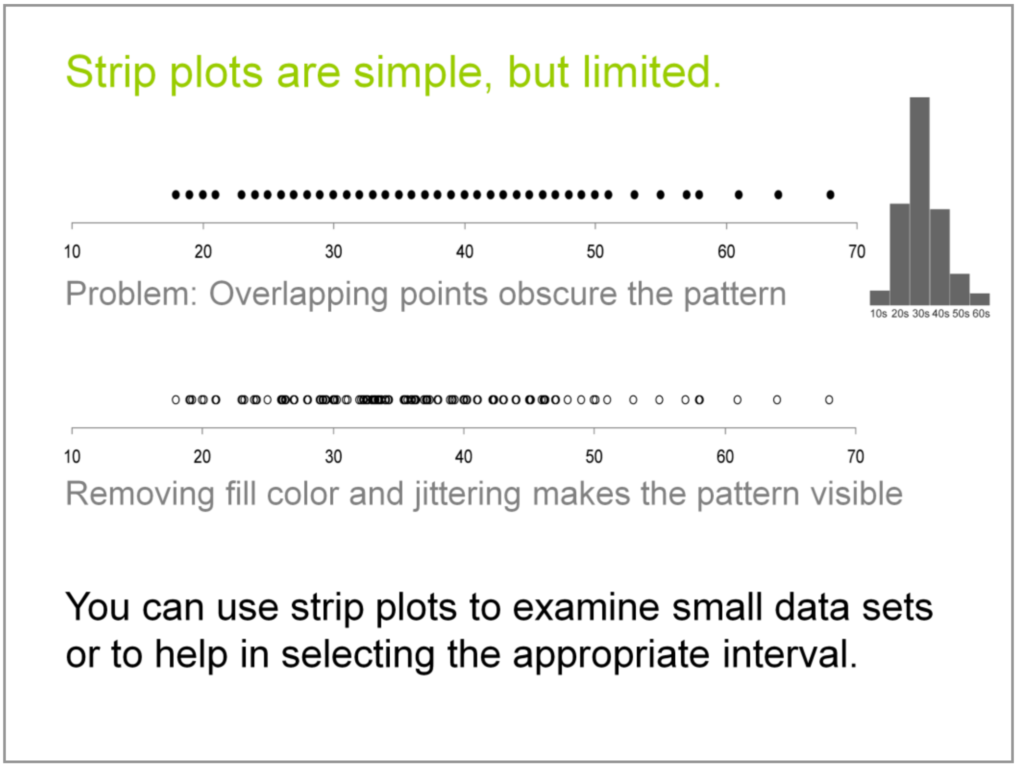
Frequency Polygon

Strip Plot

Stem-and-Leaf Plot

```
1|8 9 9 9
2|0 0 1 1 3 3 3 4 4 4 5 6 6 6 6 6 7 7 8 8 9 9 9 9 9
3|0 0 0 0 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 7 8 8 9 9 9 9
4|0 0 0 0 1 1 2 2 2 3 3 4 4 5 5 5 6 6 6 6 7 7 8 9
5|0 0 1 3 5 7 8 8
6|1 4 8
```

Four types of graph can be used to examine the distribution of a single set of values, depending on what you want to focus on the most. A histogram encodes the data as bars and works well for examining and comparing the values when grouped into intervals. A frequency polygon encodes the data as a line and excels as a means to see the overall shape of the distribution when grouped into intervals. A strip plot shows the details by displaying a point for every value, rather than by grouping them into intervals. A stem-and-leaf plot works like a histogram by grouping the values into intervals, but shows the details as well by displaying the individual values that make up each interval.



## Strip plots are simple, but limited.

Problem: Overlapping points obscure the pattern

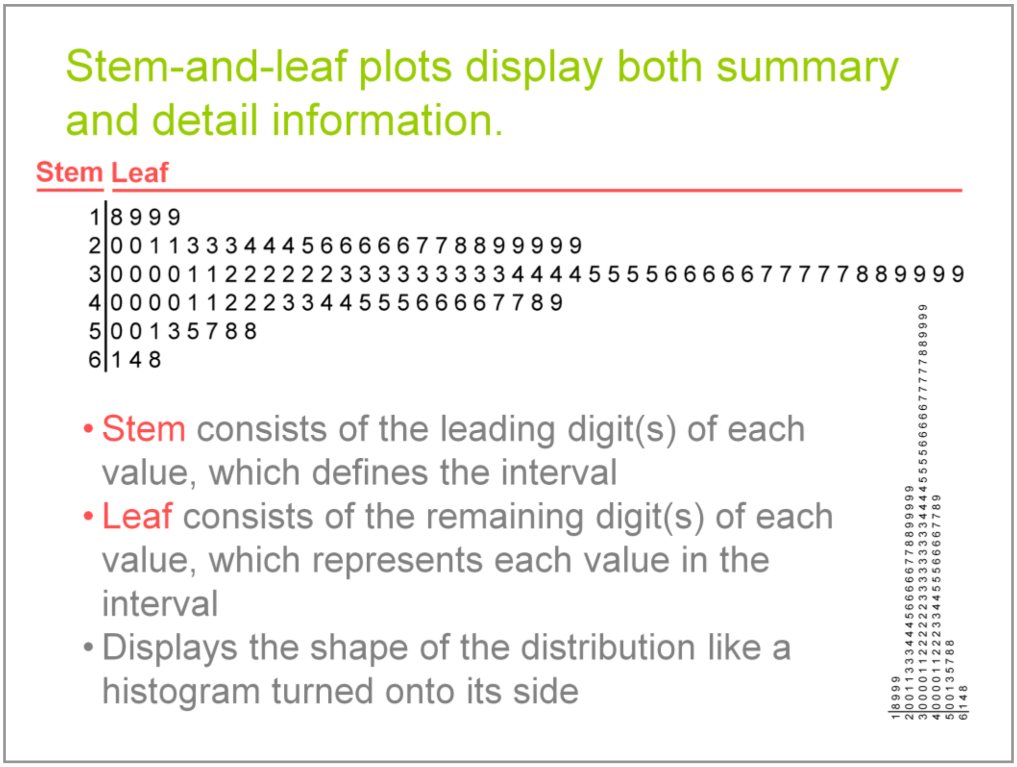
Removing fill color and jittering makes the pattern visible

You can use strip plots to examine small data sets or to help in selecting the appropriate interval.

Strip plots can give you a simple picture of a distribution that can often serve as a good place to start your analysis. They are especially useful as a means of anticipating problems associated with the selection of particular intervals for a histogram or frequency polygon.

When many values are exactly the same, however, which is the case in this distribution, strip plots can suffer from occlusion—the inability to see individual points because they are hidden behind one another. This problem can be alleviated by doing two things: 1) removing the fill color in the points, and 2) jittering the points. Jittering is the process of separating overlapping values by changing the values slightly so they don't occupy the same exact space on the graph.

Strip plots can help you to preview what will happen if you segment the data into intervals of a particular size. You want to make sure that the interval used to display the distribution as a histogram or frequency polygon does not fail to reveal meaningful patterns in the data, such as a big gap in the values that would not be visible if the interval you chose includes a bunch of values at the two ends, which would hide the gap in the middle.



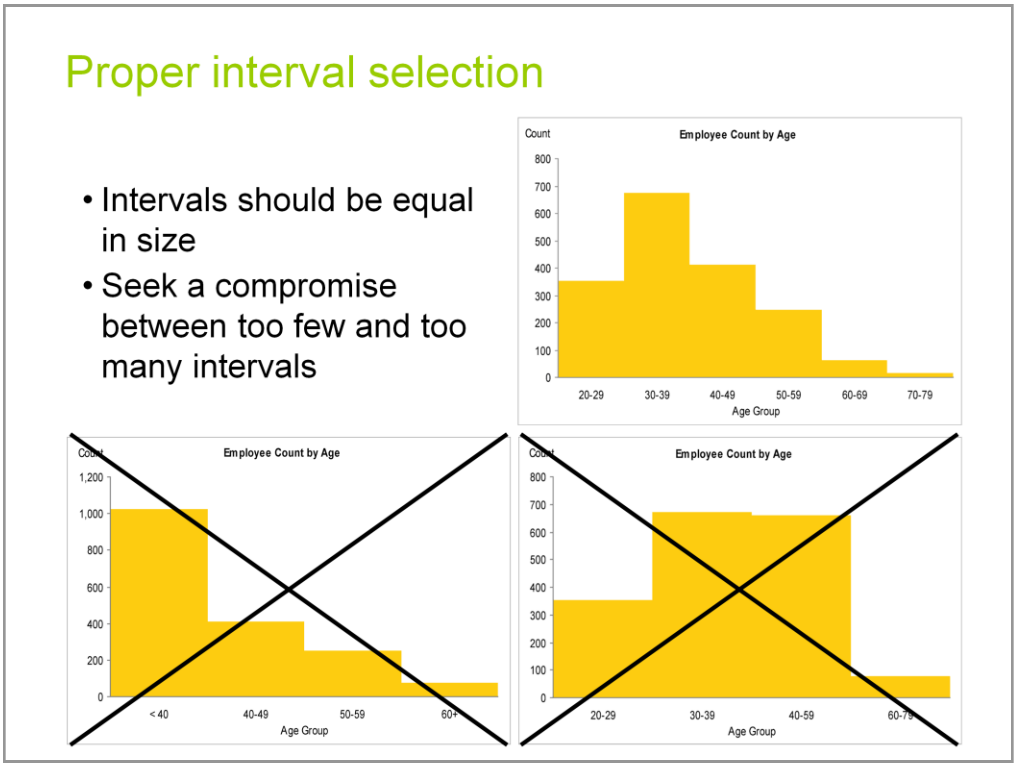
## Stem-and-leaf plots display both summary and detail information.

```
Stem Leaf
   1|8 9 9 9
   2|0 0 1 1 3 3 3 4 4 4 5 6 6 6 6 6 7 7 8 8 9 9 9 9 9
   3|0 0 0 0 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 7 8 8 9 9 9 9
   4|0 0 0 0 1 1 2 2 2 3 3 4 4 5 5 5 6 6 6 6 7 7 8 9
   5|0 0 1 3 5 7 8 8
   6|1 4 8
```

- **Stem** consists of the leading digit(s) of each value, which defines the interval
- **Leaf** consists of the remaining digit(s) of each value, which represents each value in the interval
- Displays the shape of the distribution like a histogram turned onto its side

Stem-and-leaf plots offer a nice combination of both summary and detail data. They are good for relatively small data sets, but, as you can imagine, become unwieldy with large data sets. Another advantage is that they can be easily constructed by hand. The stem-and-leaf plot was invented by John Tukey.

> *We prefer the stem-and-leaf display to the histogram because it retains the most significant digits of the data. This feature enables us*
>
> *•To see patterns in the data.*
>
> *•To see distribution of data values within an interval.*
>
> *•To go more easily from a value in the display to the datum that produced it.*
>
> (*Understanding Robust and Exploratory Data Analysis*, Hoaglin, Mosteller & Tukey, editors, John Wiley & Sons: New York, 1983, page 29)



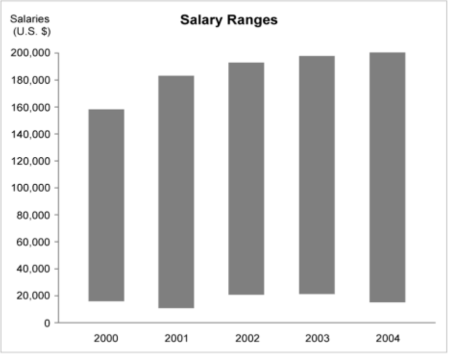
## Proper interval selection

- Intervals should be equal in size
- Seek a compromise between too few and too many intervals

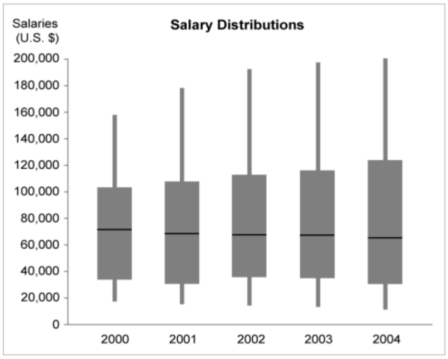
> *When a histogram is made, the interval width of the histogram is generally greater than the data inaccuracy interval, so accuracy is lost. As we decrease the interval width of a histogram, accuracy increases but the appearance becomes more ragged…In most applications it makes sense to choose the interval width on the basis of what seems like a tolerable loss in the accuracy of the data; no general rules are possible because the tolerable loss depends on the subject matter and the goal of analysis.*
>
> (*The Elements of Graphing Data*, William S. Cleveland, Hobart Press, 1994)

Too many intervals results in a ragged picture that is too complex for discerning a meaningful shape in the distribution. Too few intervals over-summarizes the data, resulting in a loss of meaningful variations in the distribution's shape.

## Distribution graphs—multiple distributions

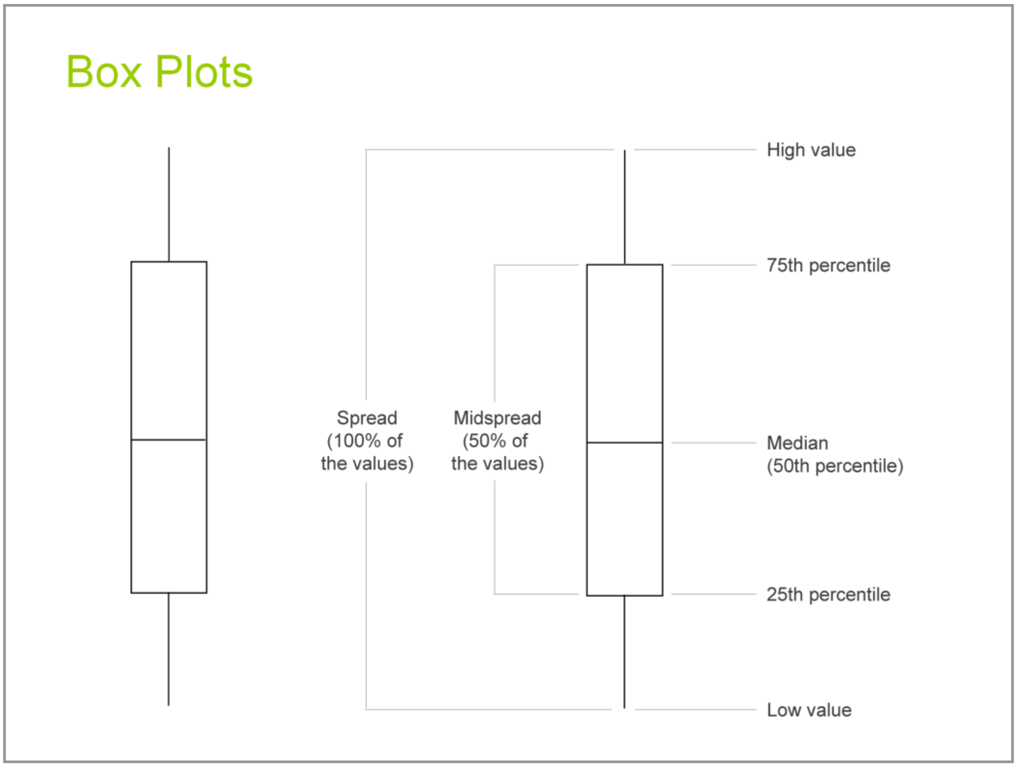


Range bars provide little information



Box plots provide rich information

When you need to compare multiple distributions to one another, such as the separate distribution for each year in the graphs above, different approaches from those we've already examined are required. The simplest possible means of graphing multiple distributions involves the use of range bars, but these are almost never adequate, because they only reveal the spread, but not the location and shape of a distribution. Box plots are ideal for examining and comparing multiple distributions and are easy to interpret with a little instruction and practice.

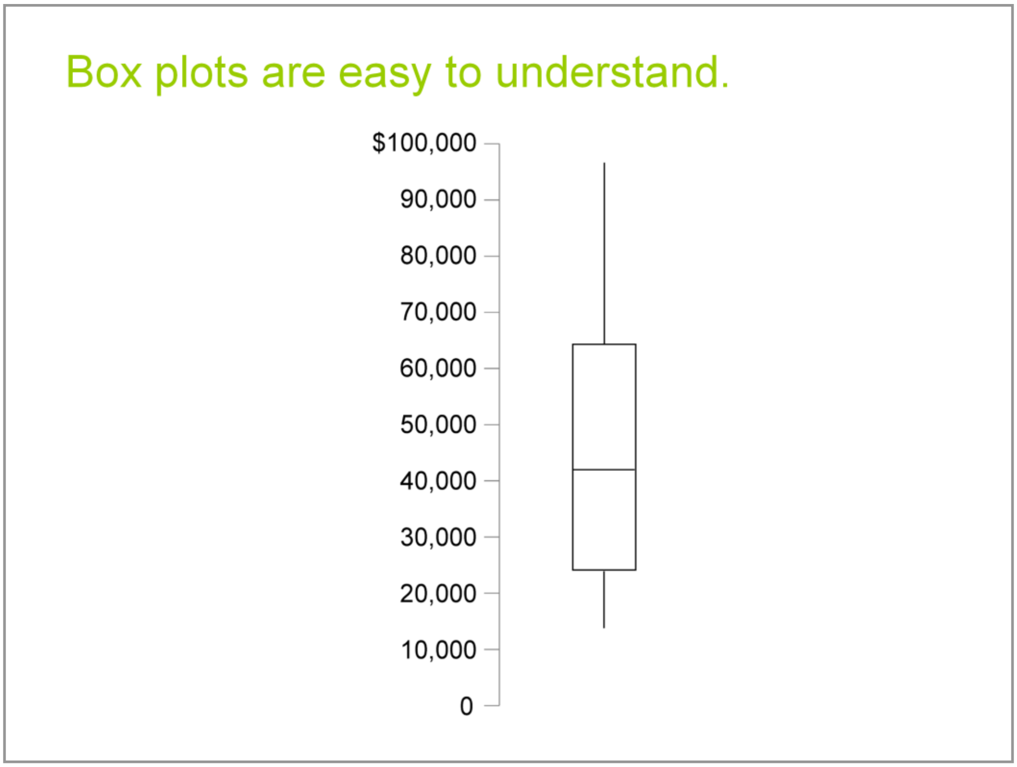


Box Plots

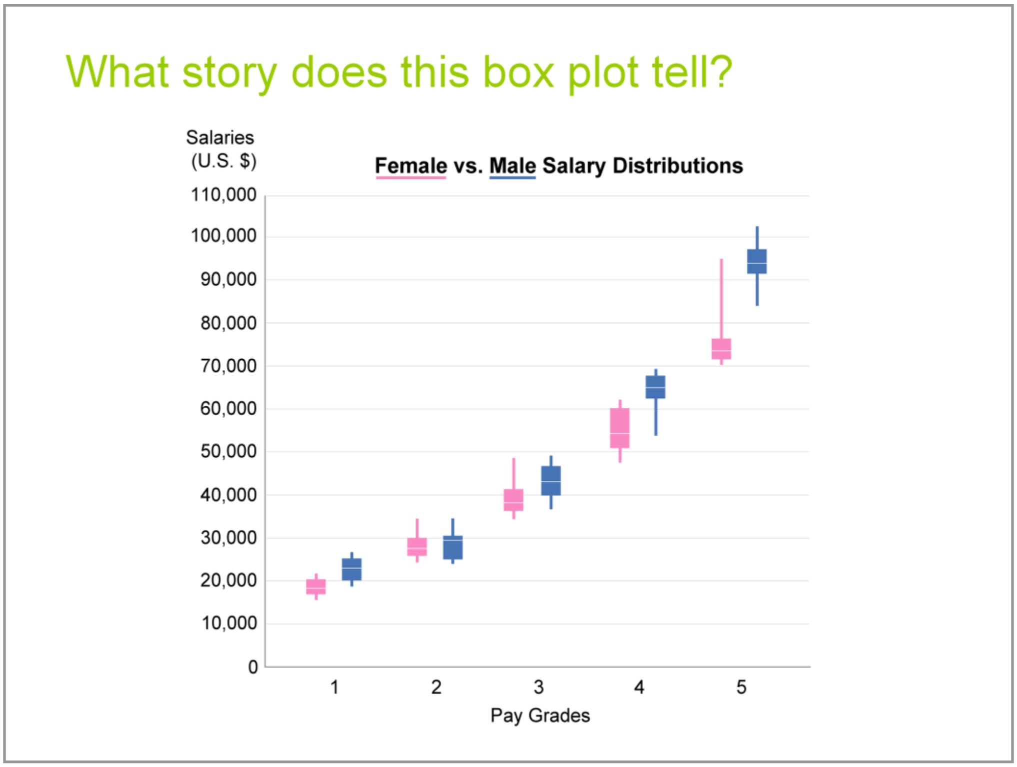
If box plots are foreign to you, and perhaps a bit intimidating, I guaranty that it will only take a moment to learn how to make sense of them. Given how much they can tell us about distributions of values, they are quite elegant yet simple in design. Here's a list of the separate facts that this box plot reveals:

- The highest value
- The lowest value
- The range of the values from the highest to the lowest, called the *spread*
- The center of the full set of values, which reveals the point above and below which 50% of the values reside, called the *median*
- The range of the middle 50% of the values, called the *midspread*
- The point above which 25% and below which 75% of the values reside, called the *75th percentile*
- The point above which 75% and below which 25% of the values reside, called the *25th percentile*

> *This compact visual display is especially useful for comparing several batches of data. By drawing a boxplot for each batch and arranging them in parallel, we can compare the batches with respect to location* [a.k.a, central tendency] *and spread, and perhaps also skewness and tail heaviness.*
>
> *(Understanding Robust and Exploratory Data Analysis, Hoaglin, Mosteller & Tukey, editors, John Wiley & Sons: New York, 1983, page 58)*



Assuming that this represents a distribution of salaries, the first thing this tells us is that the full range of salaries is quite large, extending from around $14,000 on the low end to around $97,000 on the high end. Secondly, we can see that more people earn salaries toward the lower rather than the higher end of the range. This is revealed by the fact that the median, encoded as the horizontal line in the middle of the rectangle (or box) at approximately $42,000, is closer to the bottom of the range than the top. Half of the employees earn between $25,000 and $65,000, which is definitely closer to the lower end of the overall range. The 25% of employees who earn the lowest salaries are grouped closely together across a relatively small $10,000 range of salaries. Notice how spread out the top 25% of employees are. This tells us that as we proceed up the salary scale there appear to be fewer and fewer people within each interval along the scale, such as from over $60,000 to $70,000, from over $70,000 to $80,000, and from above $90,000 to $100,000. In other words, salaries are not evenly spread across the entire range; they are tightly grouped near the lower end and spread more sparsely toward the upper end where the salaries are more extreme compared to the norm. This box plot offers a great deal more insight that a lone average, and even much more than an average complemented by the low and high salaries as well. Not bad for a simple box and three lines.

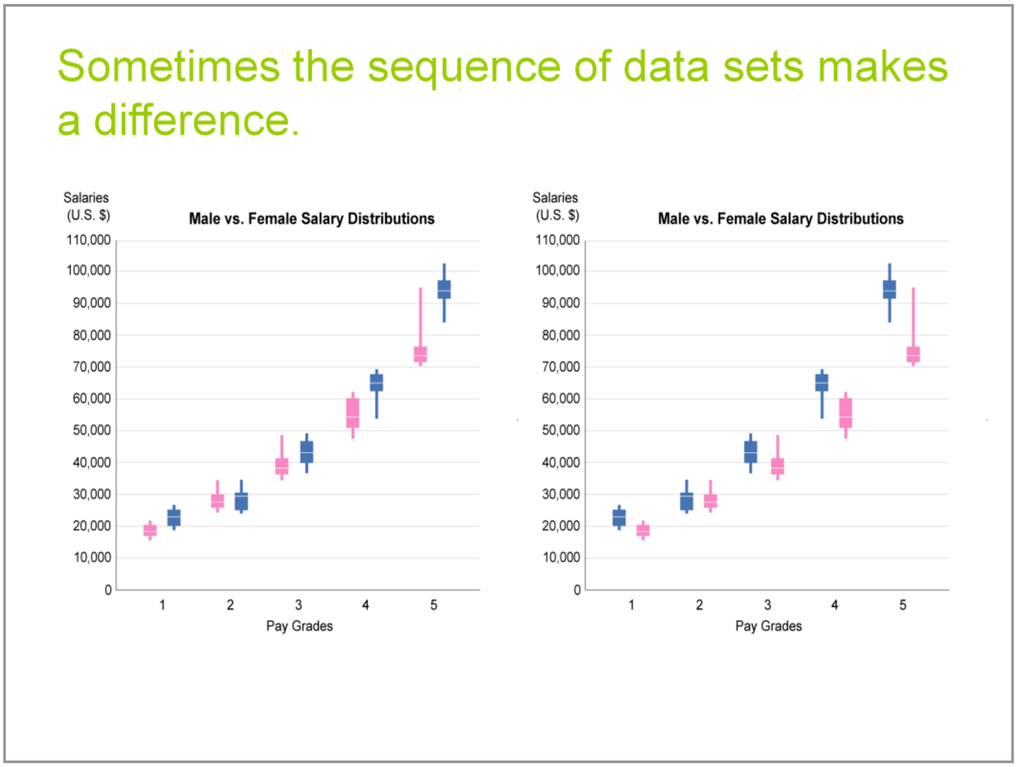


## What story does this box plot tell?

Now that you know how to read a box plot, test out your skills by trying to tease the story out of this one that compares female and male salaries among five pay grades.

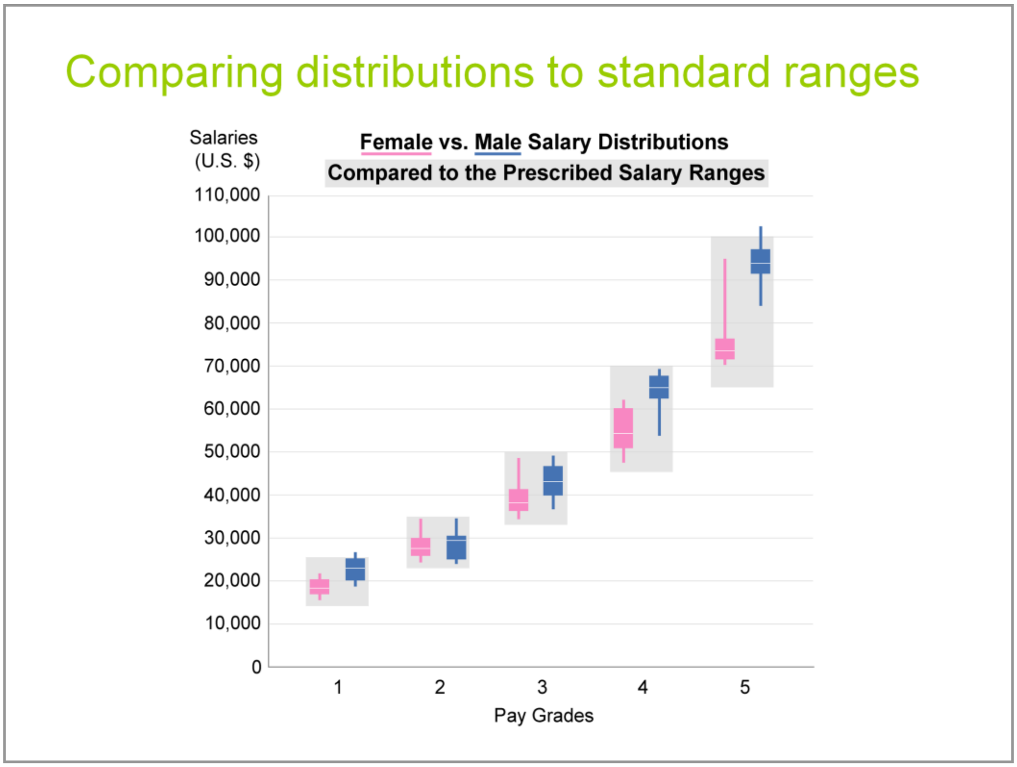
Here are some of the facts that I noticed:

- Women are typically paid less than men in all salary grades.
- The disparity in salaries between men and women becomes increasingly greater as one’s salary increases.
- Salaries vary the most for women in the higher salary grades.



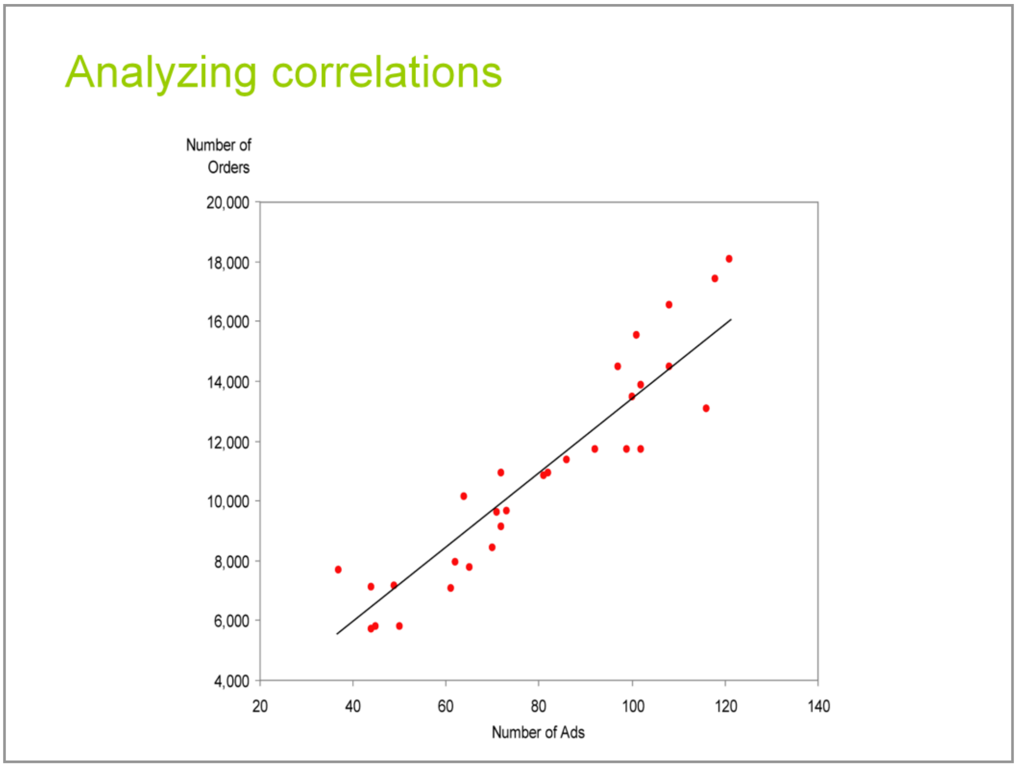
## Sometimes the sequence of data sets makes a difference.

The arrangement of data in the graph on the left, with female salaries followed by male salaries, presents a pleasing curve starting in the left-hand bottom corner and sweeping up to the right-hand top corner. Our eyes love nice continuous lines and curves. Because of this, the disparity between male and female salaries is not nearly as visible as it is in the graph on the right. These two graphs are exactly the same, except for the order of female and male salaries. When comparing sets of values that are encoded as boxes or bars, it is often useful to switch the order of the data sets to see if anything new jumps out that wasn't obvious before.



## Comparing distributions to standard ranges

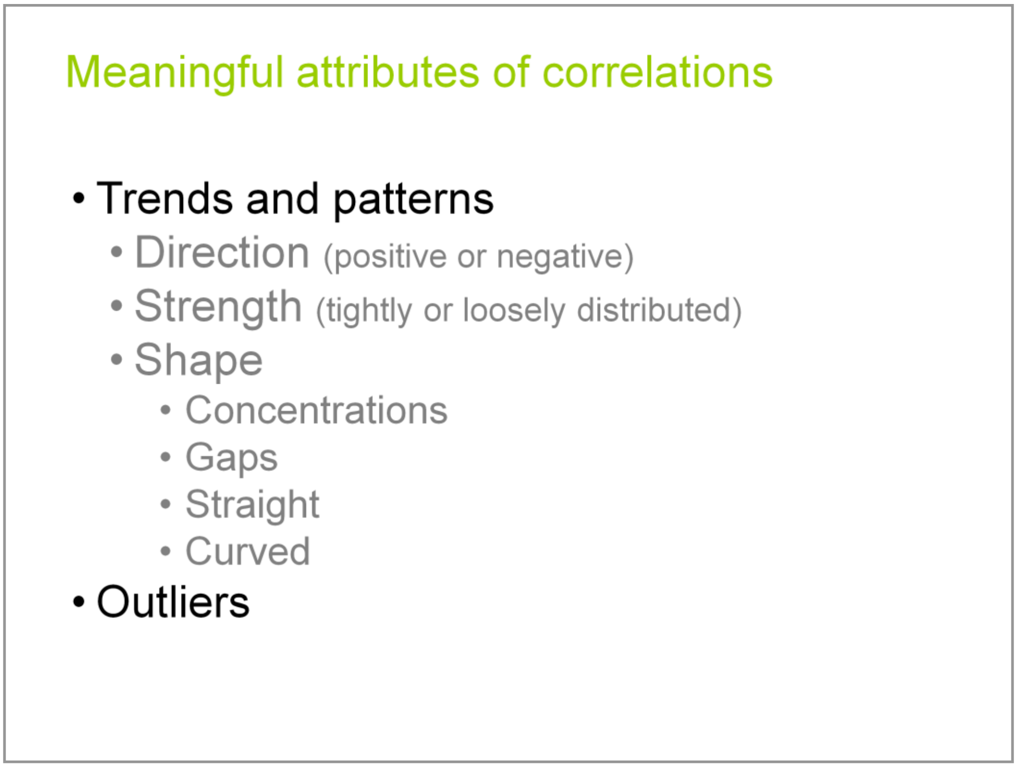
Just like with bar and line graphs, it is often useful to compare distributions in a box plot to ranges of the norm or defined standards, like the prescribed salary ranges for each pay grade in the graph above. This allows you to see that some men in pay grades 1 and 5 are being paid salaries that exceed the prescribed ranges.



## Analyzing correlations

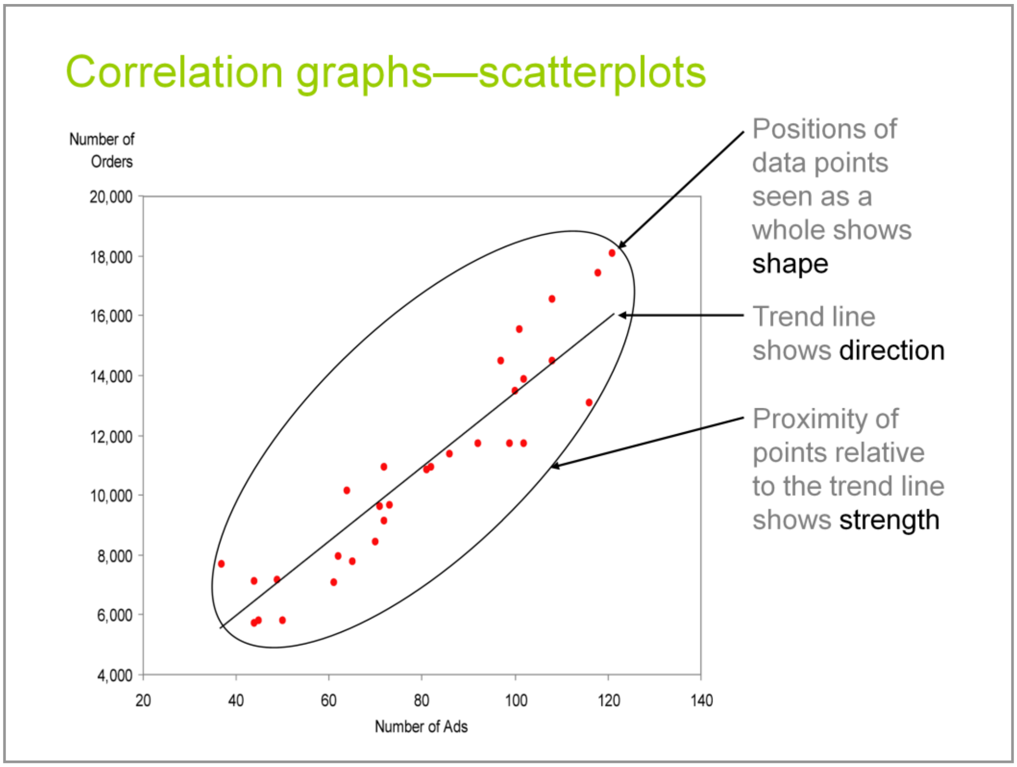
A common mistake that people make when analyzing correlations is to assume that the presence of a correlation proves causation. In fact, a correlation can indicate any of the following:

- Causation
- Common cause (known as a spurious correlation)
- Different expressions of the same data (for example, revenue and profit, because the calculation of profit is partly based on revenue)



## Meaningful attributes of correlations

- Trends and patterns
  - Direction (positive or negative)
  - Strength (tightly or loosely distributed)
  - Shape
    - Concentrations
    - Gaps
    - Straight
    - Curved
- Outliers



## Correlation graphs—scatterplots

Positions of data points seen as a whole shows **shape**

Trend line shows **direction**

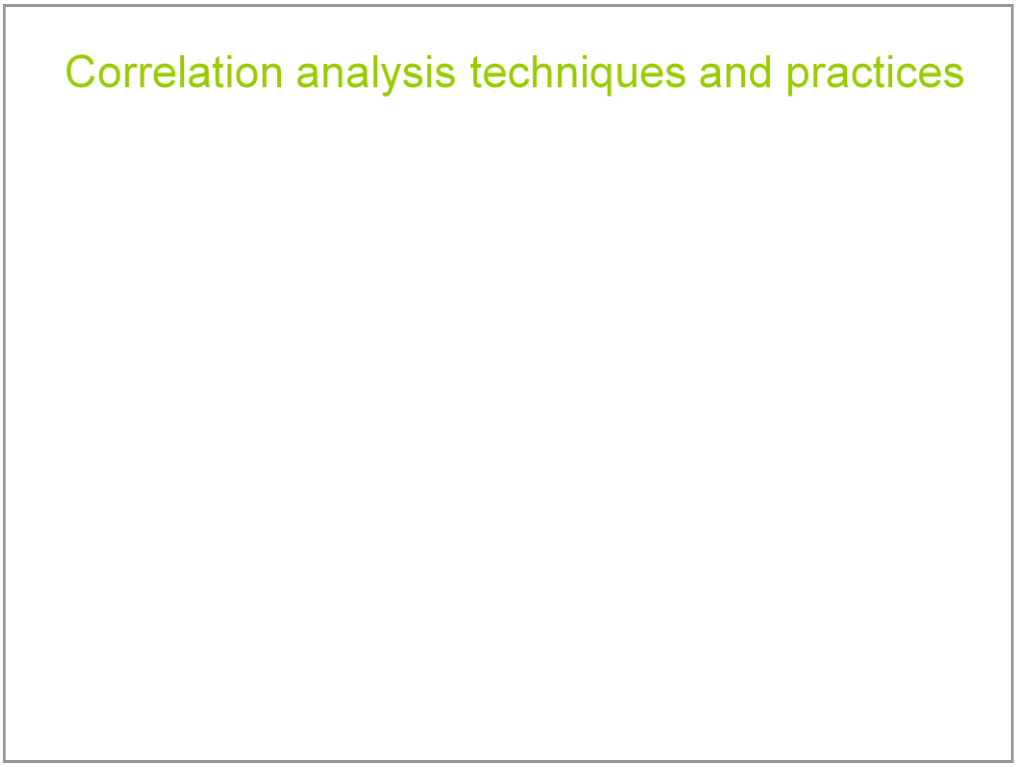
Proximity of points relative to the trend line shows **strength**

There are three key attributes that you should examine in scatterplots: direction, strength, and shape.

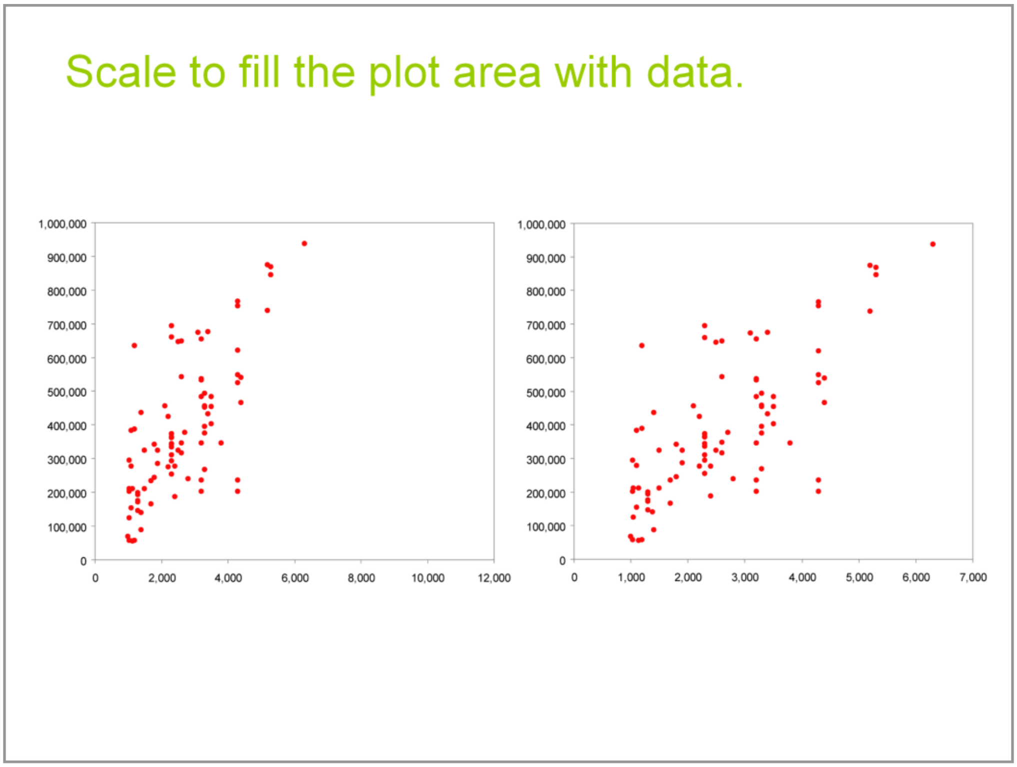
If the overall trend of the values is sloped upwards from left to right, the direction of the correlation is positive. This means that as values in the variable on the x-axis increase, values in the variable on the y-axis also tend to increase. If the overall trend slopes downwards from left to right, the direction of the correlation is negative. This means that as values in the variable on the x-axis increase, values in the variable on the y-axis tend to decrease.

If the values are closely grouped around the trend line, this means that the correlation is strong. With a strong correlation, you can predict with a fair degree of accuracy how much the dependent variable will increase or decrease in relation to specific increases or decreases in the independent variable. The more scattered the values are in relation to the trend line, the weaker the correlation.

By examining the individual data points and the shapes that they form, such as clusters or gaps, you can also discern behaviors worth investigating.



## Correlation analysis techniques and practices

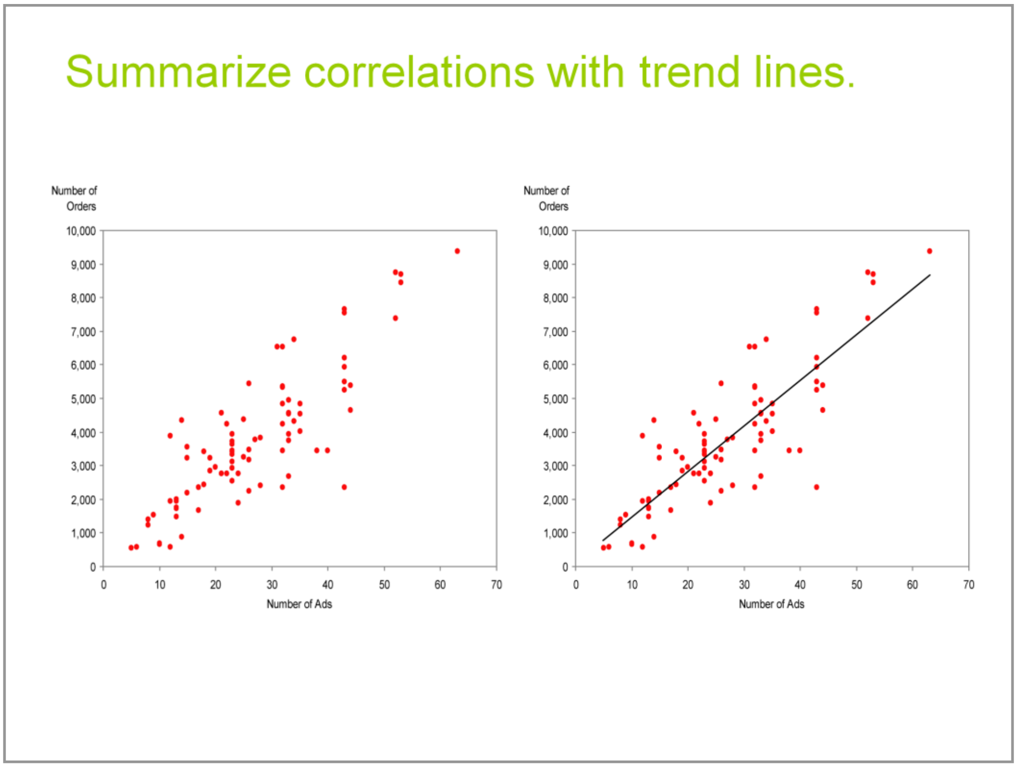


Scale to fill the plot area with data.

The shape of the data in a scatter plot is influenced by its aspect ratio (the relationship between the graph's height and width) and the range of the quantitative scales along the axes.

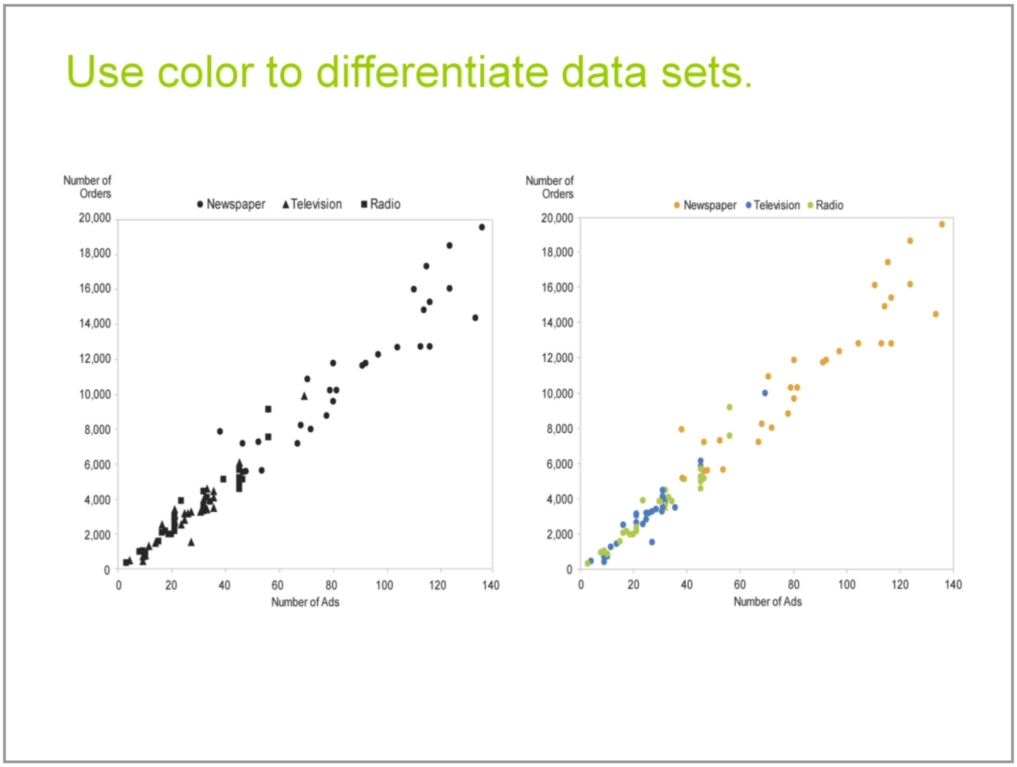
It is usually best to keep the height and width of the plot area approximately equal in scatterplots.

Even though the values look quite different in the two graphs above, both graphs contain the same exact data. The difference in appearance is because the quantitative scales in the graph on the right have been adjusted to fill the plot area with the data, while the scale along the x-axis in the graph on the left extends beyond the highest value, resulting in a large part of the plot area being empty of data. It is best to adjust the quantitative scales on both axes to extend just slight below and slightly above the highest values, which will cause the values to be distributed throughout the plot area, making is easier to see them clearly.



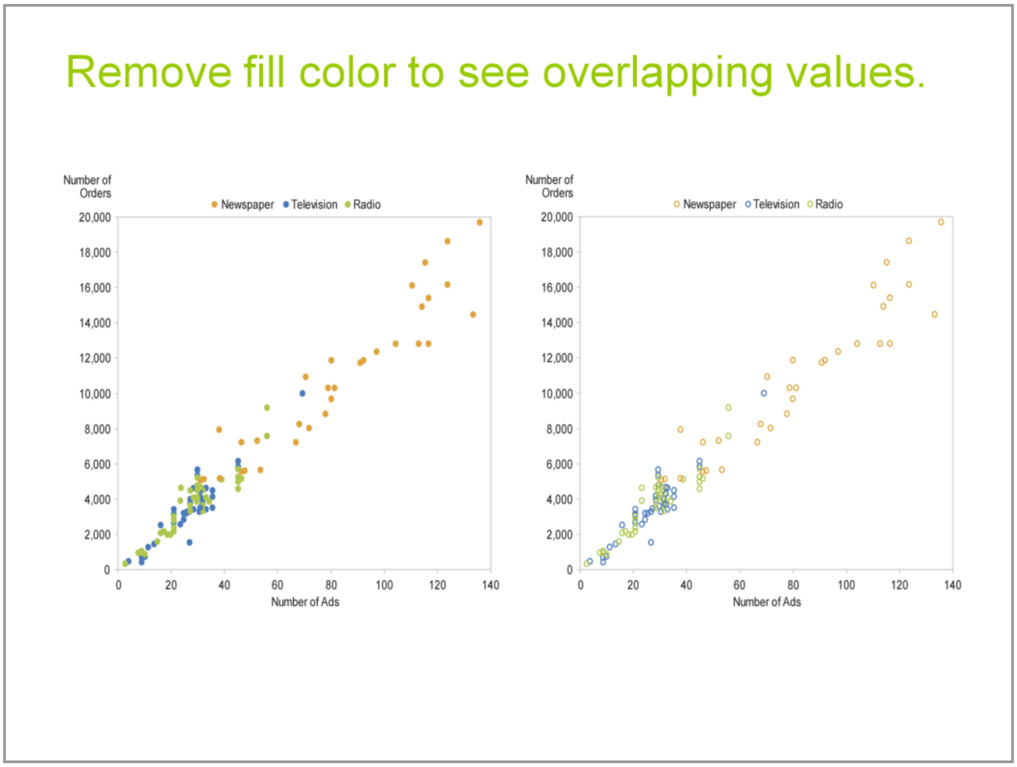
## Summarize correlations with trend lines.

Trend lines help you to focus on two of the important features of a correlation: its direction and strength. Trend lines are an attempt to draw a line through the center of the entire set of values. Fortunately, most graphing software will do this calculation for you.



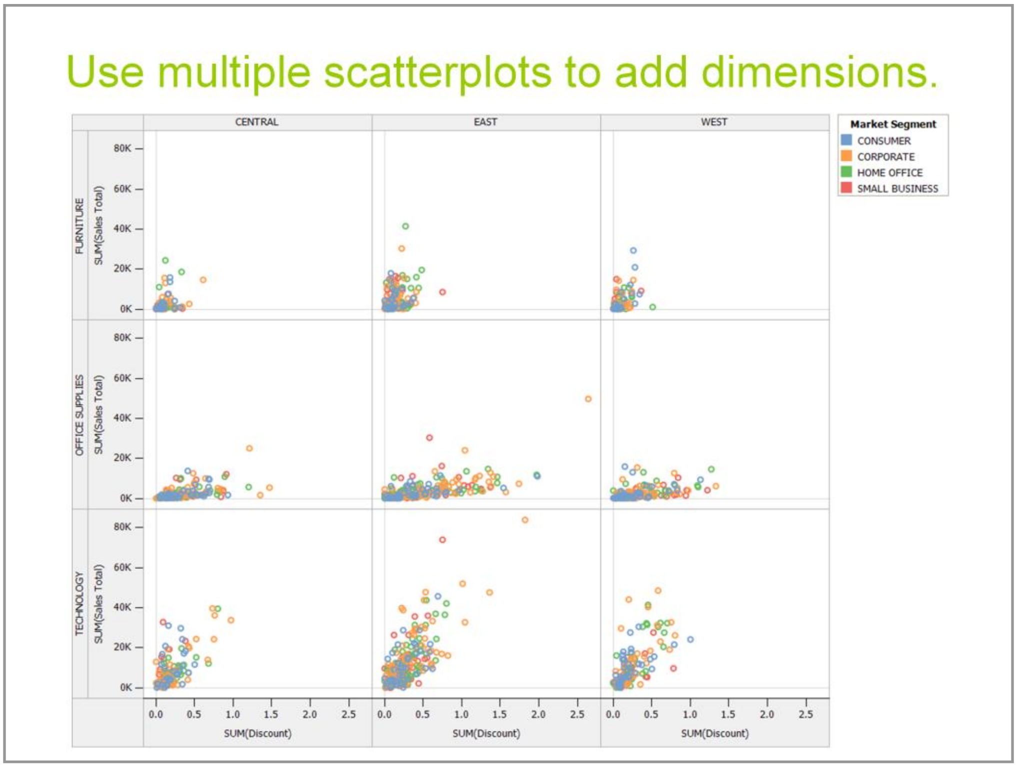
Use color to differentiate data sets.

When your scatterplot contains multiple data sets, such as those above, the best means to differentiate the data sets is by making each a different hue. Just make sure that the hues that you select are different enough from one another to be easily distinguished. If you are colorblind, you can either select colors that you can tell apart, or use different shapes (circles, squares, X's, etc.) instead.



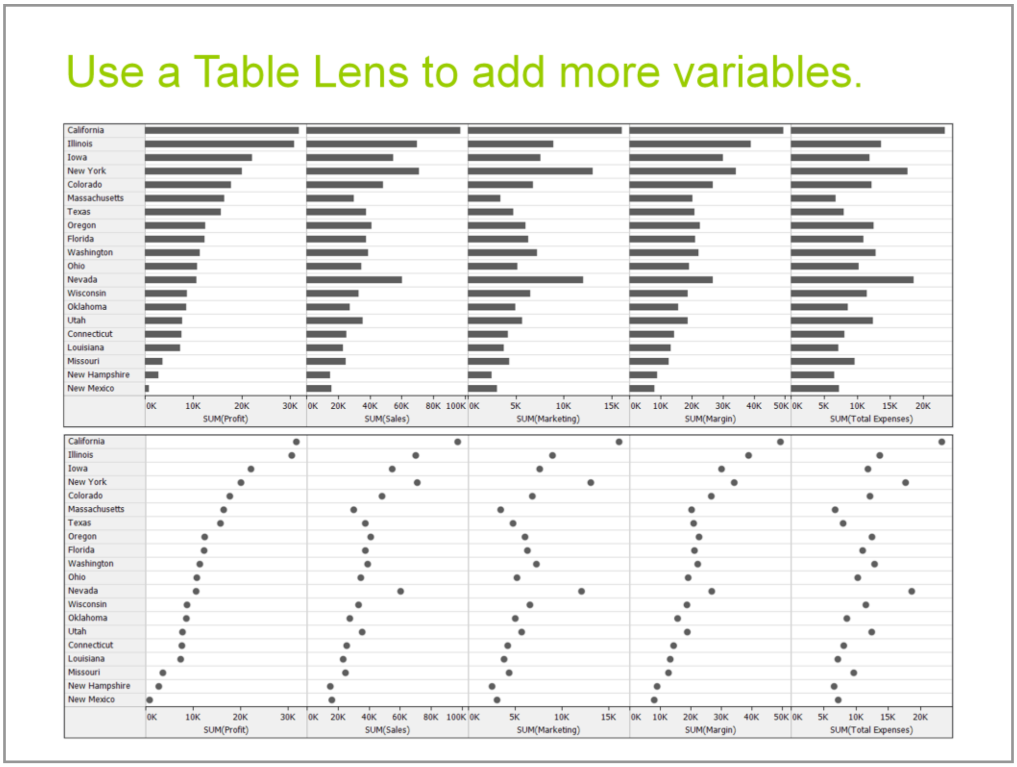
Remove fill color to see overlapping values.

If many data points are close together and overlap one another in the scatter plot, this problem can be easily reduced by removing the fill color from the points, leaving only their outlines. This makes it easier to see when data points overlap.



Use multiple scatterplots to add dimensions.

Small multiples can be used to see the relationship of one or more categorical variables to correlations, such as this example, created using Tableau Software, which correlates discounts and sales by regions (columns), product types (rows), and market segments (colors).

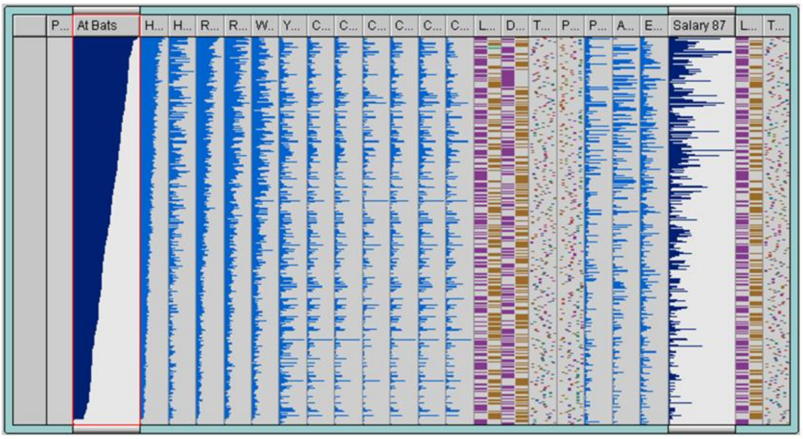


## Use a Table Lens to add more variables.

A Table Lens is a simple display that allows you to explore multivariate datasets by arranging data into tabular rows and columns. By sorting on all the columns based on the values in a single column, such as profits in this case (the leftmost column), you can easily see if other variables are correlated with profits to some degree, based on either a similar pattern consisting of bars of decreasing size (positive correlation) or a reverse pattern consisting of bars of increasing size (negative correlation). In this example, we can see that all of these variables are roughly correlated, with the strongest correlation to profit being margin ( the second column from the right).

Table Lens displays traditionally encode values as bars, but other encoding methods can be used, such as data points (dots) in the bottom example. Notice that with the dots, it is not as easy for your eyes to track across and compare the values in a single row (that is, a single state), but it is perhaps slightly easier to compare the shape of an entire set of values in a given column and the compare the overall shape of the values in one column to those in another.

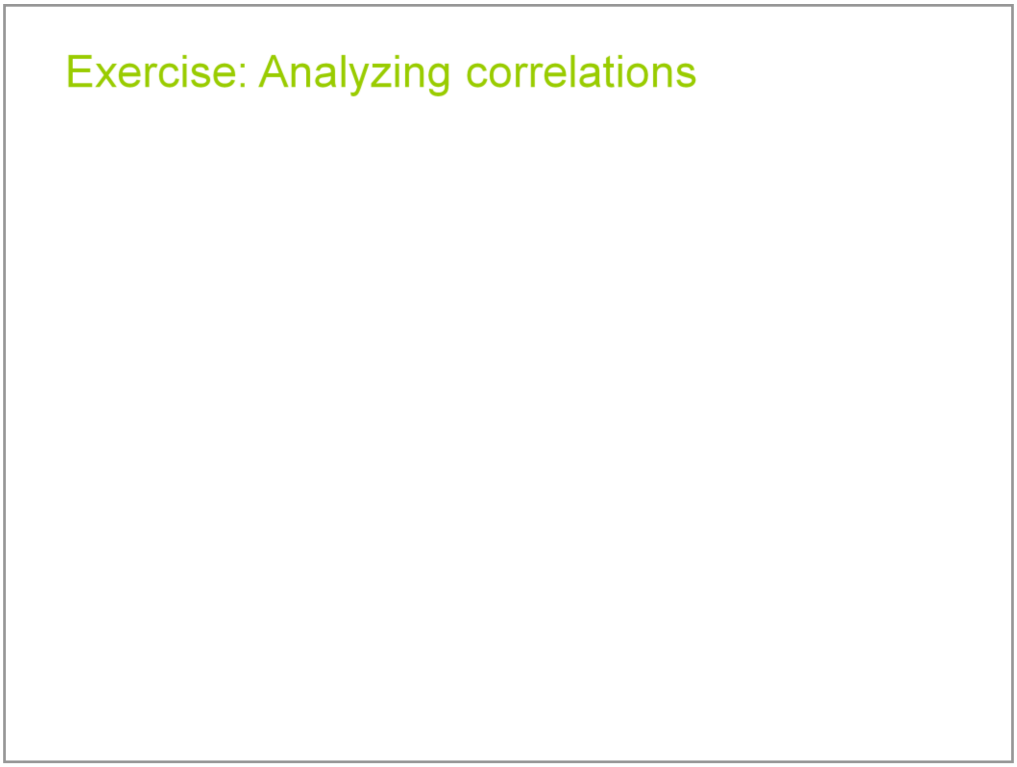
## Table Lens displays can include many variables.



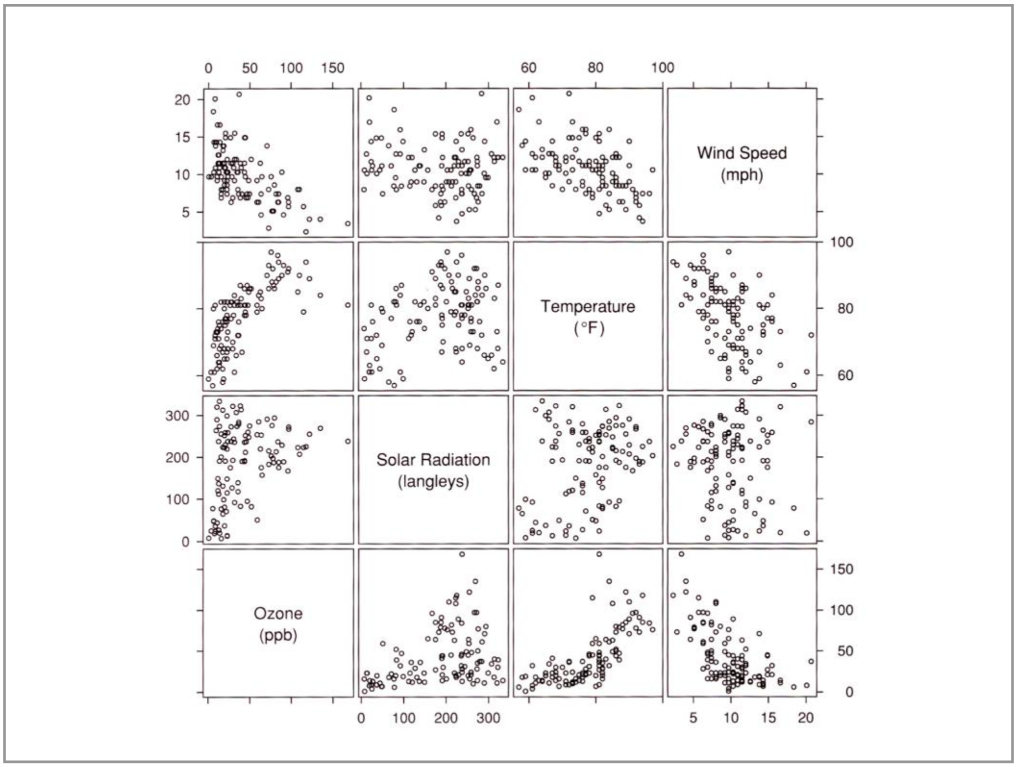
This Table Lens displays baseball hitting statistics from 1987 for 323 baseball players (the rows). It has 25 variables (the columns) for each player. The columns show quantitative statistics including *season and career at bats*, *hits*, *home runs*, and *RBIs* as well as categorical properties including *team* and *position* on the field. Quantitative variables are represented by graphical bars proportional in length to the represented values, and category values with a corresponding color and position within the field.

Sorting the table by different variables (in this case, At Bats) leads to a simple and intuitive way to understand the shape and spread of values for a given variable, as well as a means to spot correlations between variables. In addition to At Bats, other quantitative variables nearby also appear roughly sorted, which reveals a correlation between those variables. Salary also is very roughly correlated, though as would be expected, there must be other factors. A next question might be: Who is the guy that makes so much money?

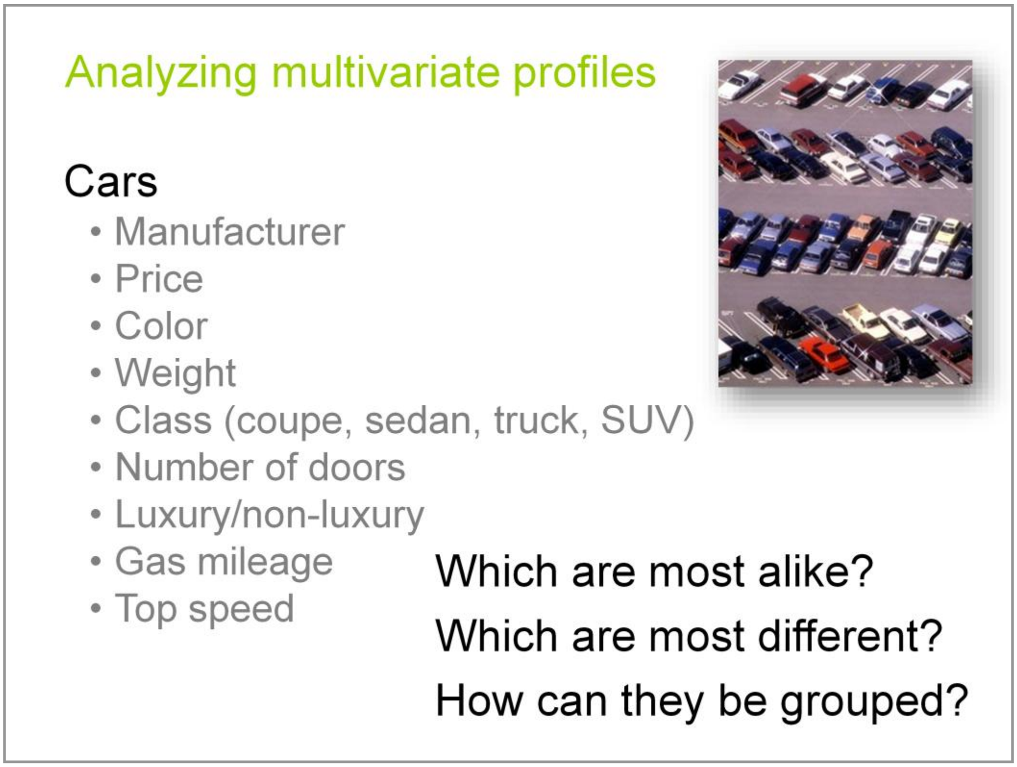
(Source: This example was taken from the product named Table Lens from Inxight Software.)



## Exercise: Analyzing correlations



(Source: This example appears in *The Elements of Graphing Data*, William Cleveland, Hobart Press, Summit, New Jersey, 1994, page 194.)



## Analyzing multivariate profiles

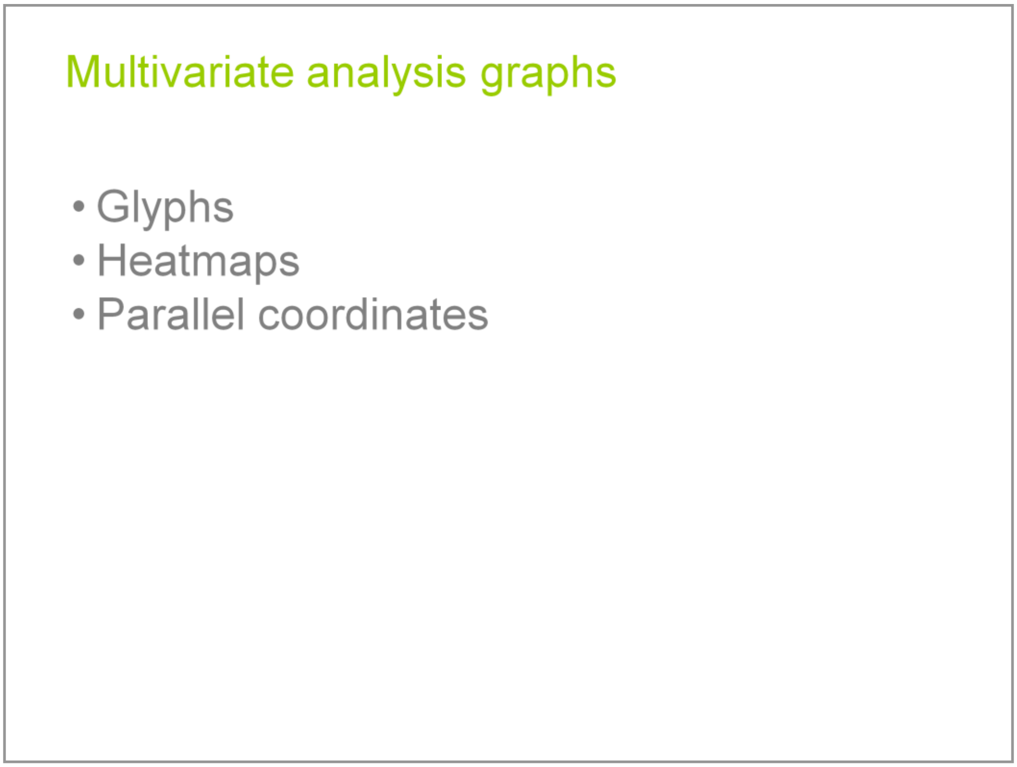
Cars

- Manufacturer
- Price
- Color
- Weight
- Class (coupe, sedan, truck, SUV)
- Number of doors
- Luxury/non-luxury
- Gas mileage
- Top speed

Which are most alike?

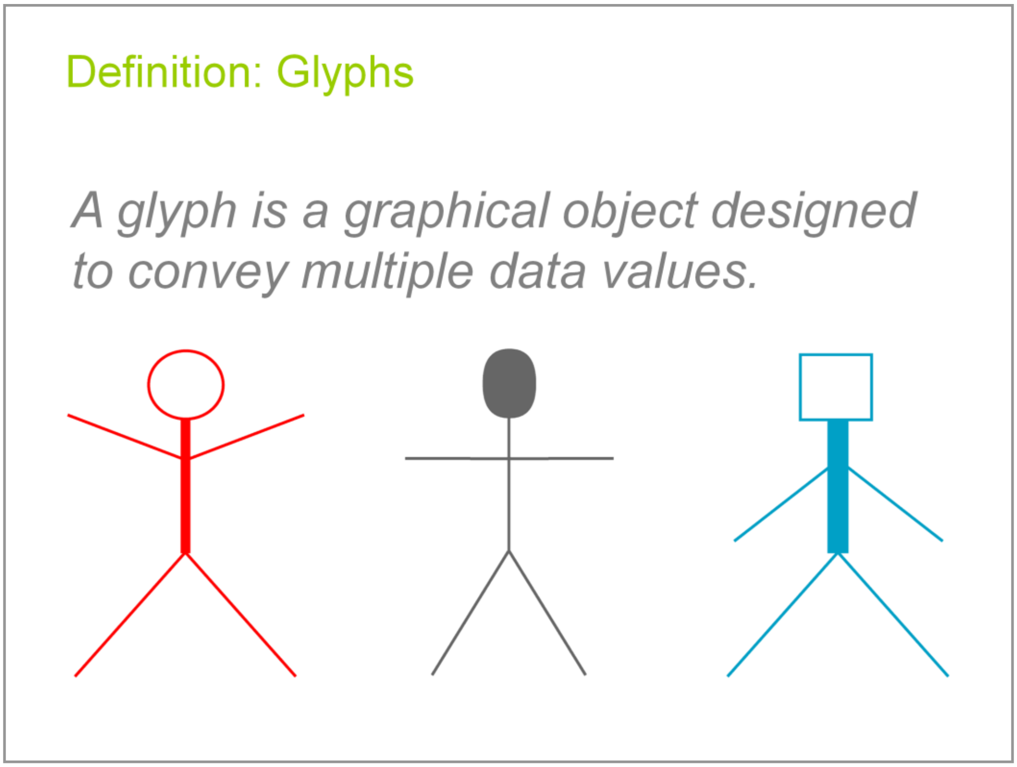
Which are most different?

How can they be grouped?



## Multivariate analysis graphs

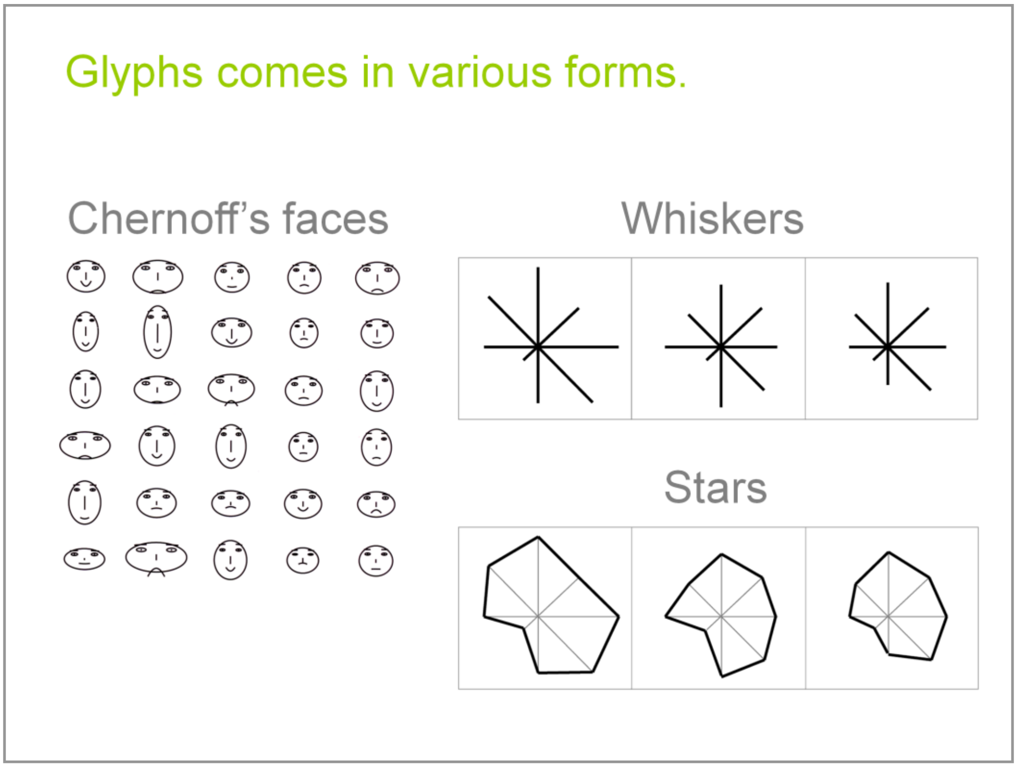
- Glyphs
- Heatmaps
- Parallel coordinates



## Definition: Glyphs

*A glyph is a graphical object designed to convey multiple data values.*

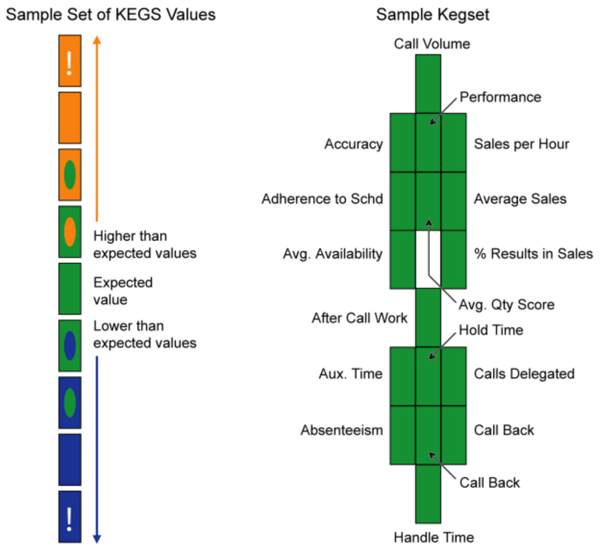
This definition comes from *Information Visualization: Perception for Design*, Second Edition, Colin Ware, Morgan Kaufmann Publishers: San Francisco, CA, 2004.



## Glyphs comes in various forms.

The best known example of a glyph was created by Herman Chernoff in 1972. He used simplified line drawings of the human face and mapped data variables to different parts of the face (size of the eyes, curvature of the mouth, shape of the head, etc.). He chose the human face because our visual perception has evolved to rapidly read and interpret different facial expressions. From early childhood we learn to recognize and respond to subtleties in facial expression. Whiskers and stars are two other popular glyph forms.

# Glyphs are often too complex.



Sample Set of KEGS Values

Higher than expected values

Expected value

Lower than expected values

Sample Kegset

Call Volume

Performance

Accuracy

Sales per Hour

Adherence to Schd

Average Sales

Avg. Availability

% Results in Sales

Avg. Qty Score

After Call Work

Hold Time

Aux. Time

Calls Delegated

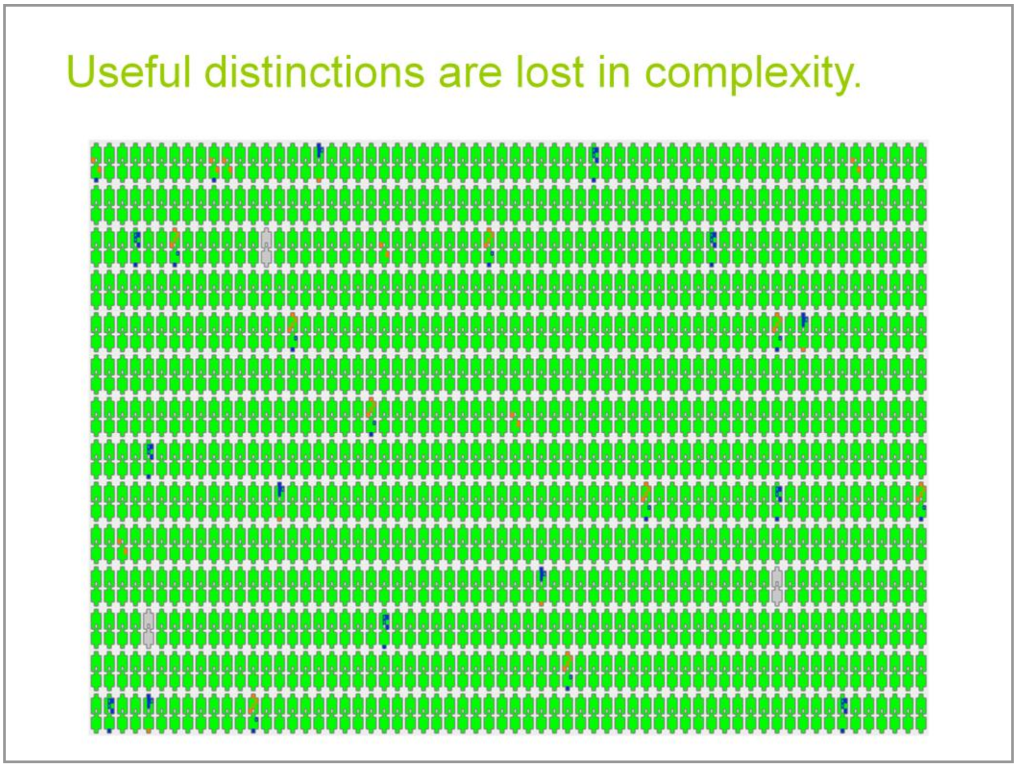
Absenteeism

Call Back

Call Back

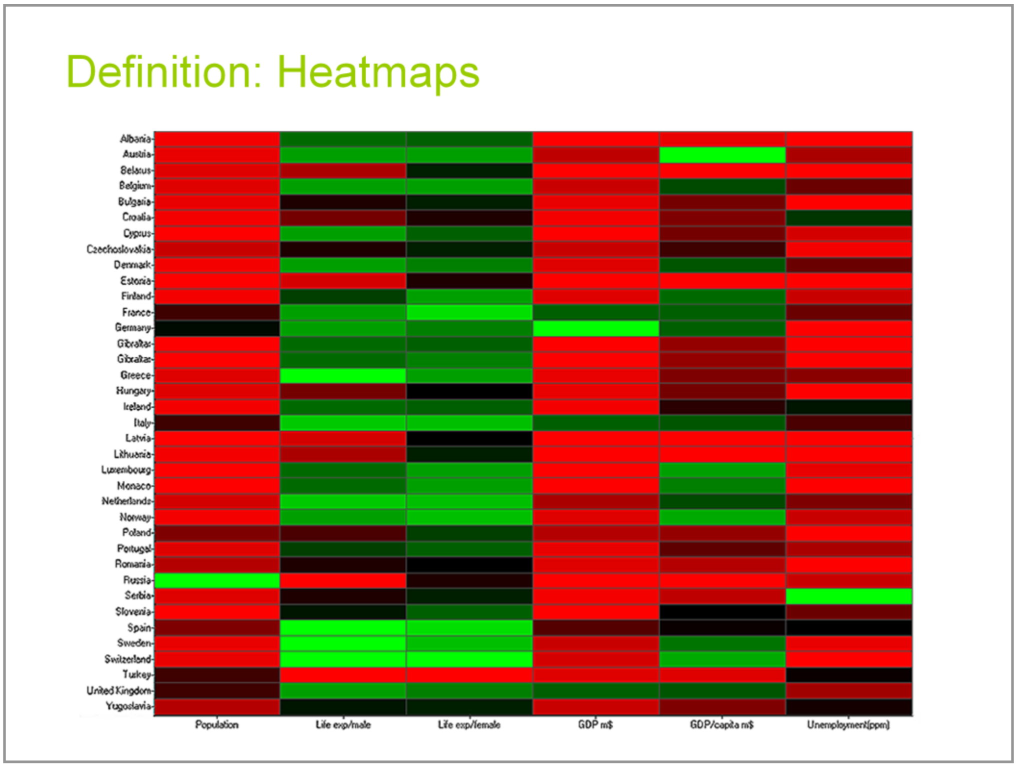
Handle Time

If a glyph is too complex, we can no longer perceive and compare the patterns that they form preattentively. You must limit the number of variables as well as the different values that can be expressed by a given variable. This example of a glyph from FYI Software is too complex for rapid perception.



## Useful distinctions are lost in complexity.

This is a full screen of the glyph that we saw on the previous slide. Because only a few variations from the norm of green rectangles appear, the glyphs with blue and orange pop out, as do the gray glyphs, which indicate missing data. If a greater range of variation (for instance, all nine color/shape values in all 17 rectangles) is being displayed in these glyphs, however, it isn't hard to imagine that this display will cease to be meaningful without a close, time-consuming study.



## Definition: Heatmaps

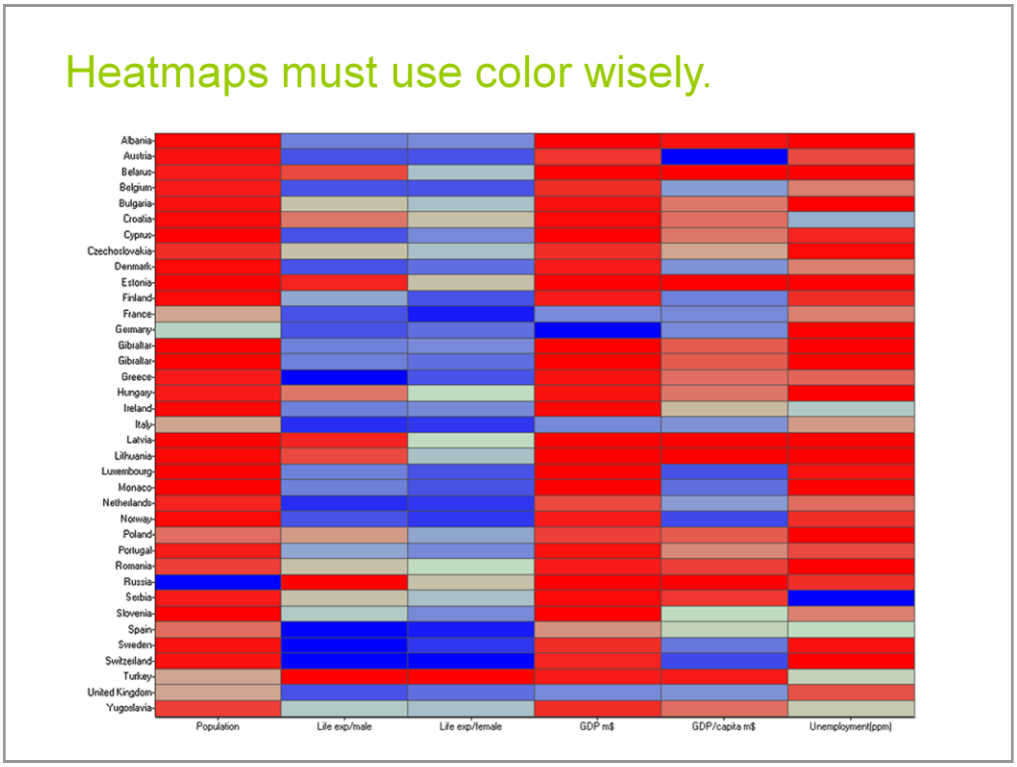
Heatmaps in the broadest sense are visual displays that encode quantitative values as color. Given this broad definition, an actual map that displays geography is a heatmap if sections, such as countries, are color coded to encode quantitative values, such as population.

Often, when we speak of heatmaps, however, we are referring to a matrix of columns and rows, similar to a spreadsheet, but instead displaying a quantitative values as text, they are encoded as color. Heatmaps, such as the above example, can be used to display multivariate data. In this case the entities that are being measured are countries (per row), and the variables are population, male life expectancy, female life expectancy, gross domestic product (GDP), per capita gross domestic product, and rate of employment (per column). The combination of colors across a single row represents a country's multivariate profile.

As you can see, heatmaps alone are difficult to use for discerning similar profiles, but they can be used to reveal exceptions, such as the single light green cell in the GDP column, which is associated with Germany.

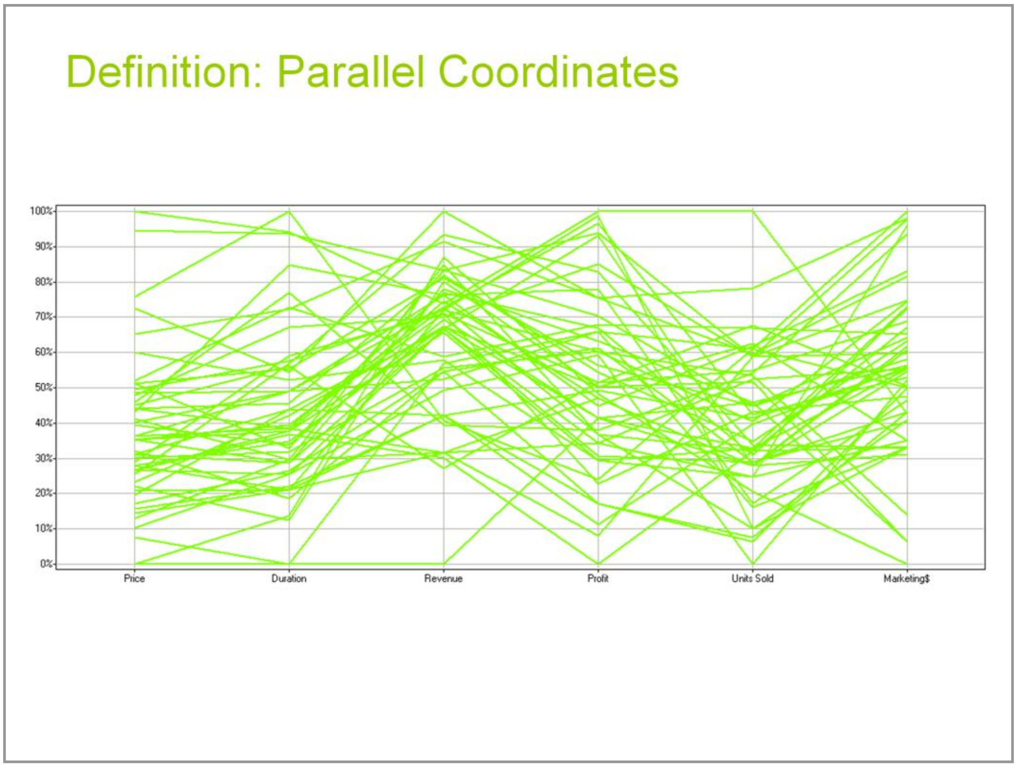
Whether they are being used to display multivariate data or for other purposes, heatmaps often suffer from poor color choices. This example illustrated three common problems:

1. The full set of values falls within a continuous range of positive values, yet three distinct hues have been used. Distinct hues would only be appropriate if there is some logical breakpoint in the values themselves, such as zero, with both positive and negative numbers. When no logical breakpoint exists in the data, varying intensities of a single hue encodes the data better.
2. The distinction between red and green cannot be seen by the 10% of males and 1% of females who suffer from the predominant form of color blindness.
3. When distinct hues are appropriate for encoding continuous values, such as positive and negative numbers, with a neutral hue between them to represent zero, a dark color such as black usually shouldn't be used, because we intuitively interpret darker colors as greater values.



## Heatmaps must use color wisely.

Assuming that the values range between positive and negative numbers, the colors used in this example work better. No form of color blindness prevents people from seeing the difference between red and blue. The light gray that has been used to represent numbers close to zero intuitively represents low values and grabs our attention less than the vibrant reds and blues that have been used to draw our attention to extremes on both ends of the continuum.

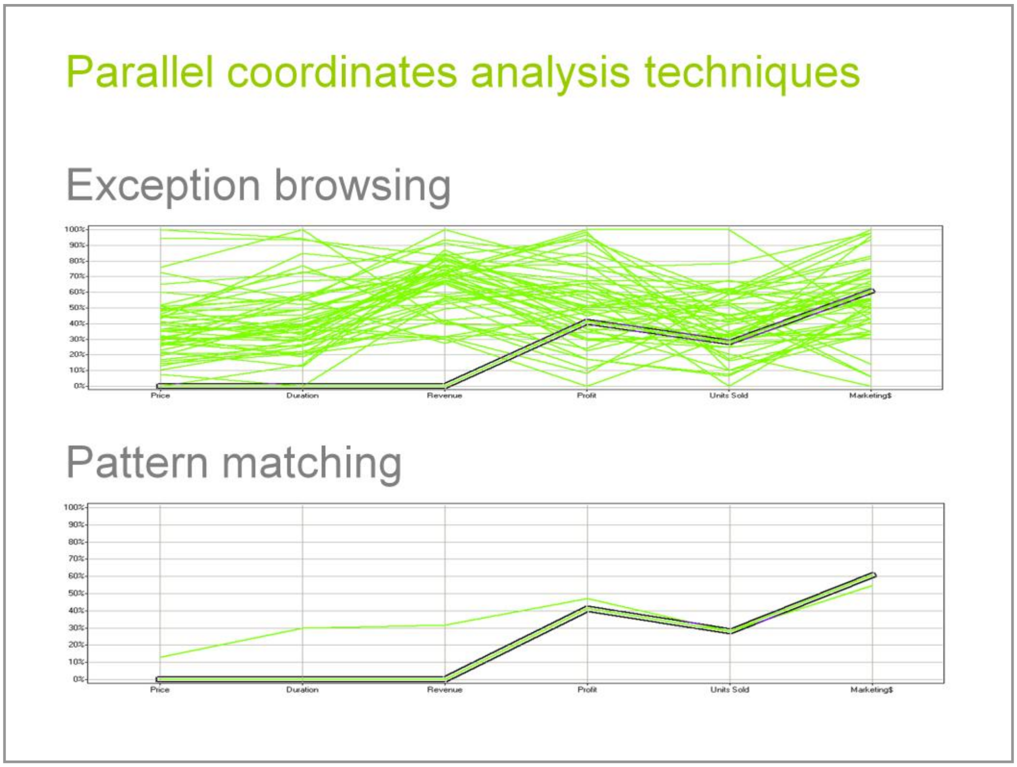


Definition: Parallel Coordinates

This graph displays six variables (market category, price, age, revenue, profit, units sold, and marketing expenses) for 50 products (one per line). The full range of values for each variable is expressed as a percentage scale that begins at 0% and extends to 100%. For example, price is expressed as a percentage, with the highest priced product at 100% and the lowest at 0%.

Unlike regular line graphs, which should only be used to connect values along an interval scale, such as time, parallel coordinates connect values that belong to entirely different variables. In this example, measures of six different variables (price, revenue, etc.) for a single product are displayed as a line that intersects each variable axis at the point where the value for that variable is located. It is appropriate to connect the values with a line in a parallel coordinates graph, because the line visually encodes the actual connection of various values associated with a single entity (product in this case). The shape formed by the line is meaningful, because it forms a multivariate pattern that represents the product’s multivariate profile. The patterns formed by various lines can be compared to determine the similarities or differences in the products’ multivariate profiles.

At first this looks like a jumbled mess of intersecting lines, but when used properly, insights can be gained about similarities and differences.



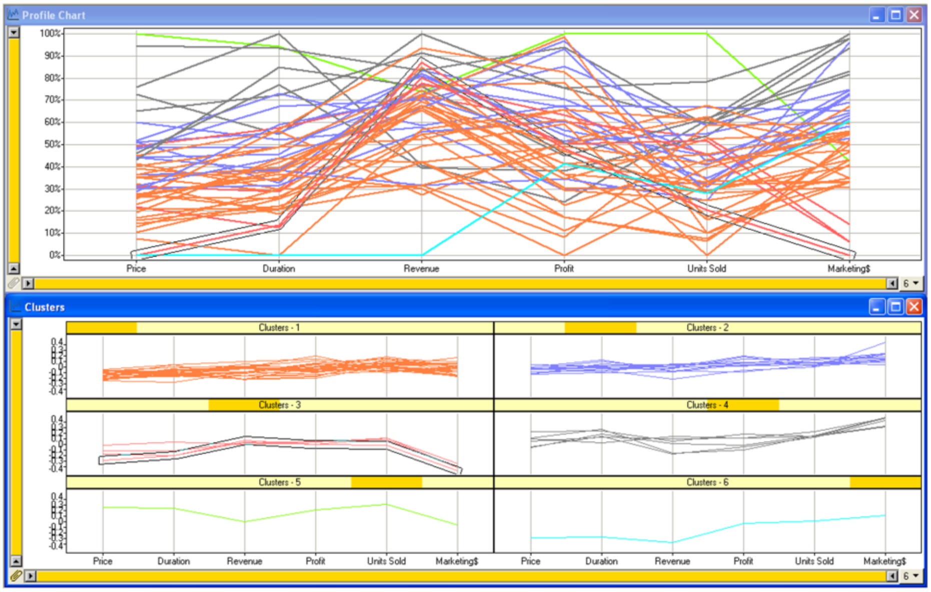
# Parallel coordinates analysis techniques

## Exception browsing

## Pattern matching

To get the full benefit from parallel coordinates, you need software that offers this kind of graph, but the simple task of identifying exceptions in a set of multivariate data can be simulated with tools as simple as Excel. By using a line graph with a line for each item and a value for each variable, you can create a simple parallel coordinates display. Because a normal line graph uses the same quantitative scale for all values, however, you must normalize the quantitative scale associated with each variable so they are the same. This can be done by converting each value of a particular variable to a percentage of the highest value associated with that variable. In this way, each variable will share a scale that ranges from 0-100%.

## Parallel coordinates can reveal more when organized into clusters.

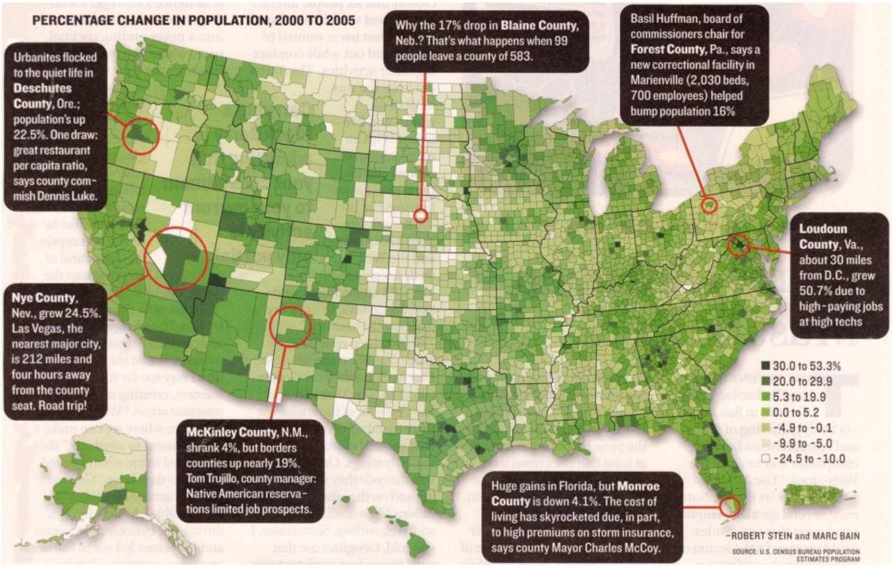


Visual data analysis for some tasks works best when paired with other techniques, such as statistical clustering algorithms that are able to break the data into groups based on mathematical similarities that would be difficult to discern with your eyes alone when viewing a busy display like this.

## Demo: Spotfire DecisionSite

Spotfire

## Analyzing geospatial data



PERCENTAGE CHANGE IN POPULATION, 2000 TO 2005

Urbanites flocked to the quiet life in **Deschutes County**, Ore.; population's up 22.5%. One draw: great restaurant per capita ratio, says county commish Dennis Luke.

Why the 17% drop in **Blaine County**, Neb.? That's what happens when 99 people leave a county of 583.

Basil Huffman, board of commissioners chair for **Forest County**, Pa., says a new correctional facility in Marienville (2,030 beds, 700 employees) helped bump population 16%

**Loudoun County**, Va., about 30 miles from D.C., grew 50.7% due to high-paying jobs at high techs

**Nye County**, Nev., grew 24.5%. Las Vegas, the nearest major city, is 212 miles and four hours away from the county seat. Road trip!

**McKinley County**, N.M., shrank 4%, but borders counties up nearly 19%. Tom Trujillo, county manager: Native American reservations limited job prospects.

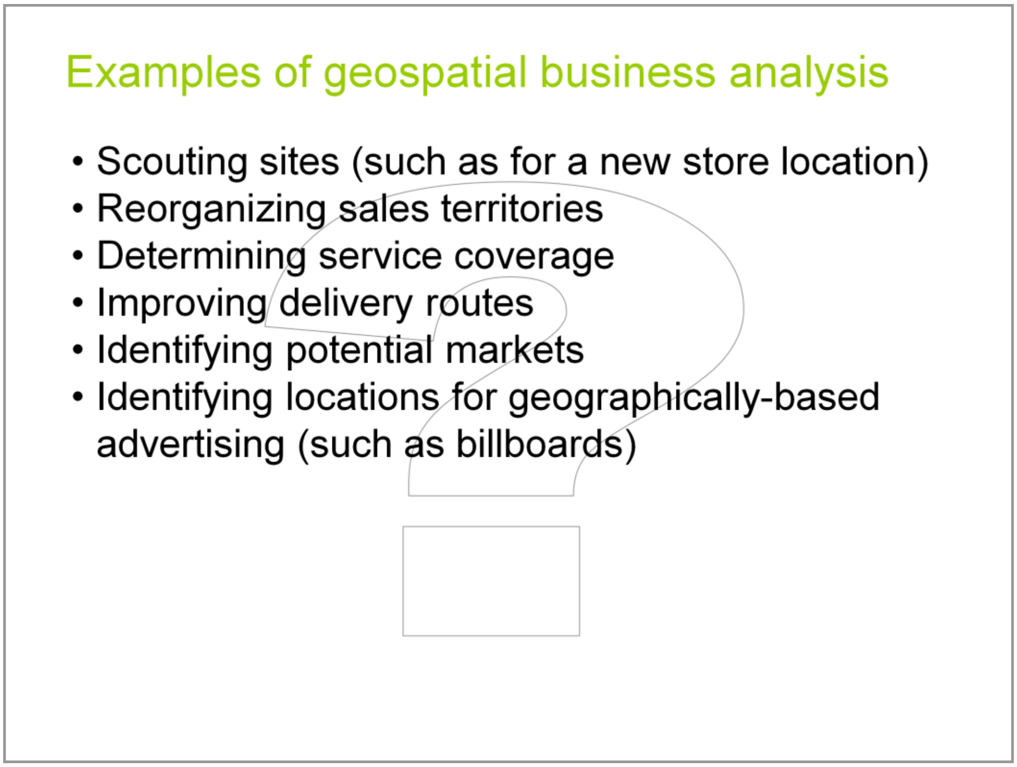
Huge gains in Florida, but **Monroe County** is down 4.1%. The cost of living has skyrocketed due, in part, to high premiums on storm insurance, says county Mayor Charles McCoy.

–ROBERT STEIN and MARC BAIN

SOURCE: U.S. CENSUS BUREAU POPULATION ESTIMATES PROGRAM

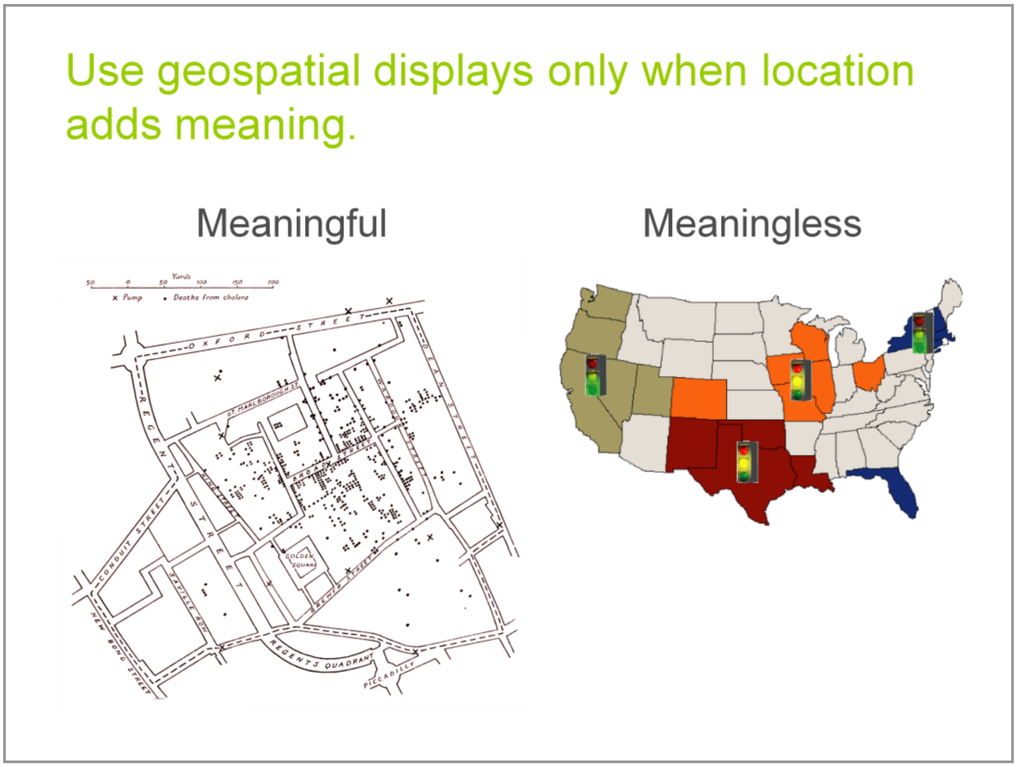
In recent years, systems that allow us to display data geospatially have increased in availability and decreased in cost. Geospatial displays are useful for data analysis when the locations of the things you are examining must be known to understand them. This isn't always the case, but when it is, the ability to see data arranged geospatially, such as on a map, is priceless.

(Source of this map: Robert Stein and Marc Bain, "Saying Goodbye to the Burbs", *Newsweek*, March 27, 2006.)



## Examples of geospatial business analysis

- Scouting sites (such as for a new store location)
- Reorganizing sales territories
- Determining service coverage
- Improving delivery routes
- Identifying potential markets
- Identifying locations for geographically-based advertising (such as billboards)



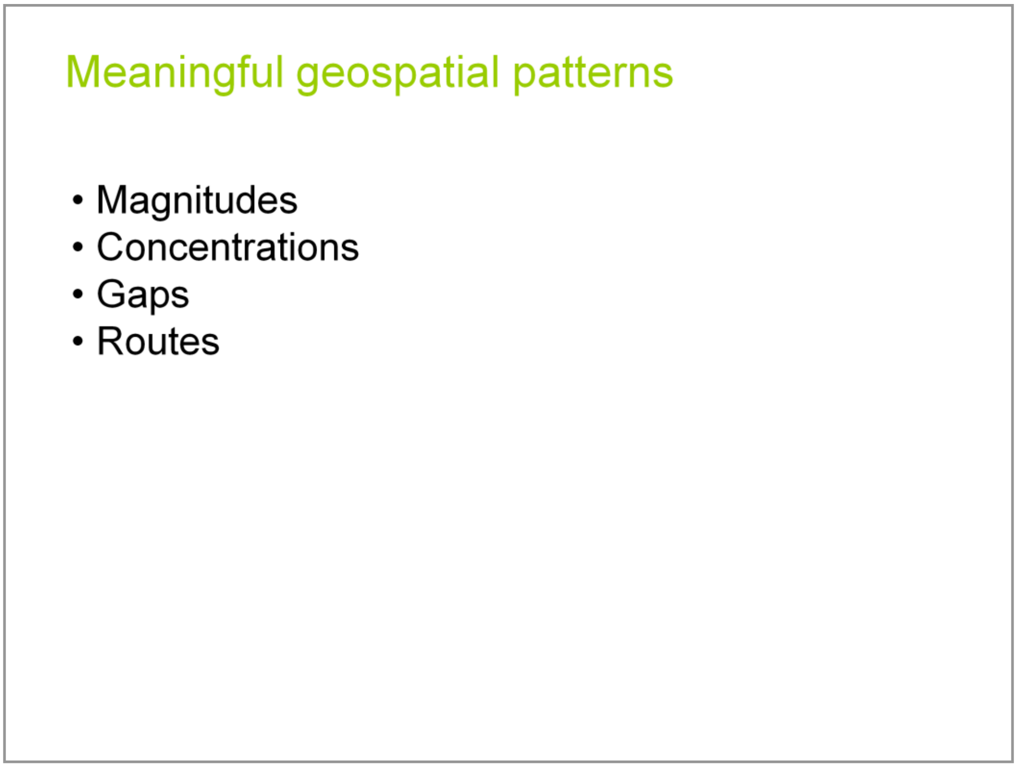
## Use geospatial displays only when location adds meaning.

Meaningful

Meaningless

The location of something or its proximity to something else is not always useful information when doing analysis. The example on the left nicely illustrates when the data couldn't be understood without seeing it displayed geospatially. Dr. John Snow created this display in an attempt to figure out the cause of a cholera epidemic in London in 1854. Each death is marked by a dot and each of central London's water pumps were marked by X's. By viewing the data in this way, Snow was able to determine that the Broad Street pump was probably the source of the disease (the one in the center of the map), and was able to confirm the fact by removing the pump's handle so it couldn't be used, which ended the epidemic that had taken over 500 lives. (E. W. Gilbert, "Pioneer Maps of Health and Disease in England," *Geographical Journal*, 124, 1958)

Placing traffic lights to indicate the state of sales in four regions on a map of the United States as shown in the right-hand example, however, doesn't seem to add any value. A simple table with a sales total for each region would have provided the information more directly.

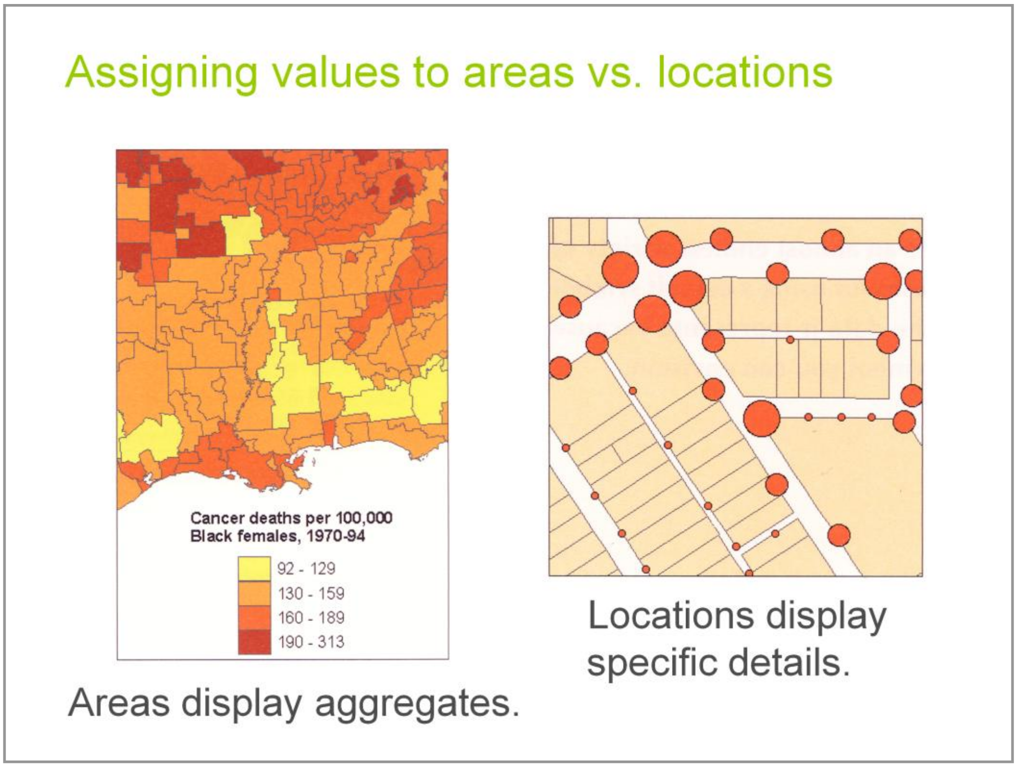


## Meaningful geospatial patterns

- Magnitudes
- Concentrations
- Gaps
- Routes

The following patterns are those that are most revealing in geospatial data:

- **Magnitudes** for simple comparisons of quantitative values, such as the population in various areas.
- **Concentrations** of values within a geographical area, which indicates locations where something exists or is happening to a greater degree, such as where more of a company's customers reside.
- **Gaps**, which indicate areas where something does not exist or occur, such as where no one has purchased a company's products.
- **Routes**, which show connections (paths) between various locations, such as that taken by most customers when traveling to a particular store.

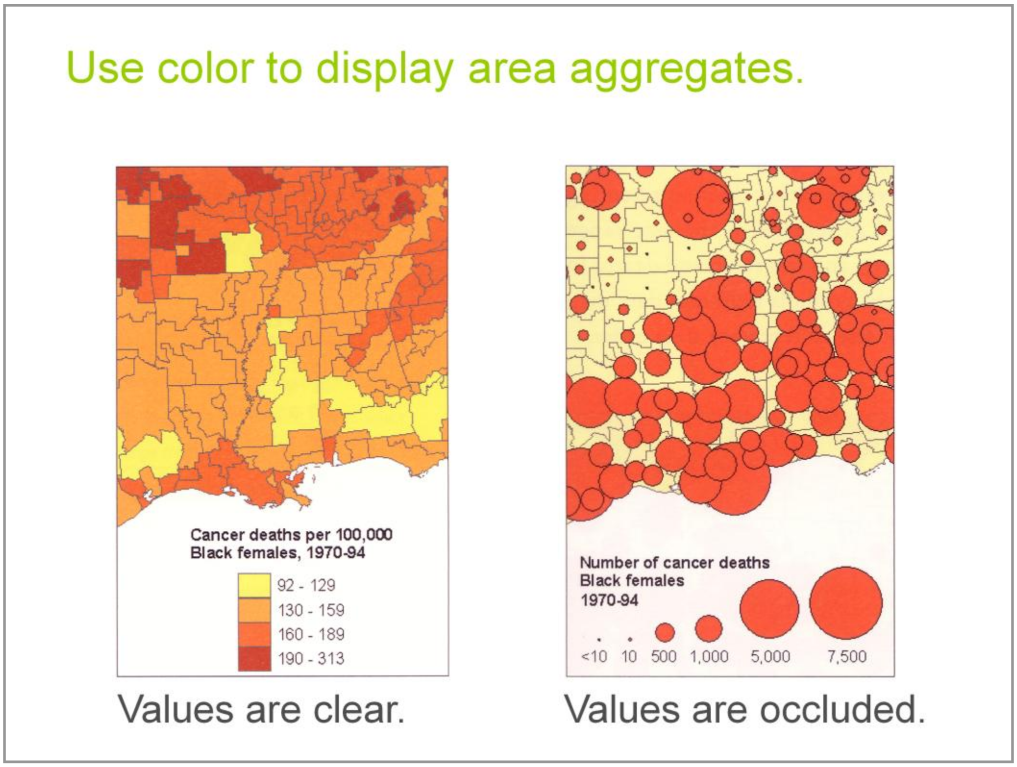


## Assigning values to areas vs. locations

Areas display aggregates.

Locations display specific details.

The essential difference between displays of areas versus displays of specific locations is the level of summary versus detail that they present—areas summarize a group of locations and locations show each item. Both are useful, depending on what you're trying to understand and the nature of the data you're examining. For example, the right-hand example above shows the precise location of street lights and the area that each illuminates, which could not be understood by summarizing the values at the area level. (Source: Cynthia A. Brewer, *Designing Better Maps: A Guide for GIS Users*, ESRI Press: Redlands, California, 2005.)

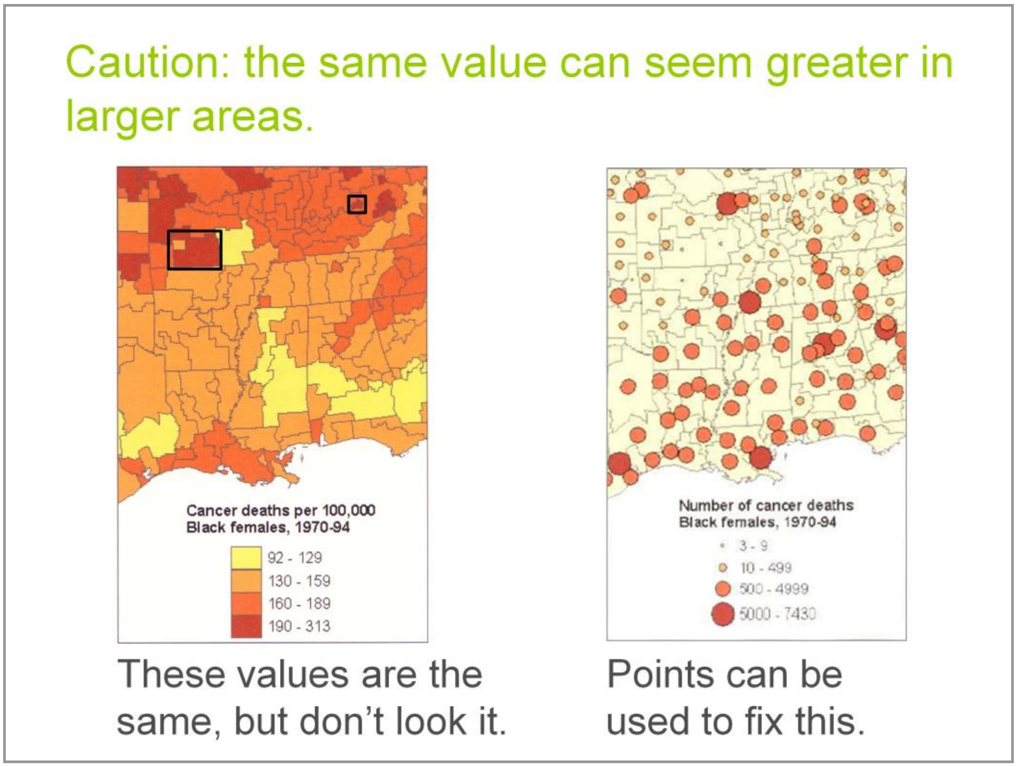


## Use color to display area aggregates.

Values are clear.

Values are occluded.

When you choose to display values summarized to the area level, color usually works best, because the areas remain clearly delineated and the problem of occlusion is eliminated, which results when using points that are centered in each area. (Source: Cynthia A. Brewer, *Designing Better Maps: A Guide for GIS Users*, ESRI Press: Redlands, California, 2005.)



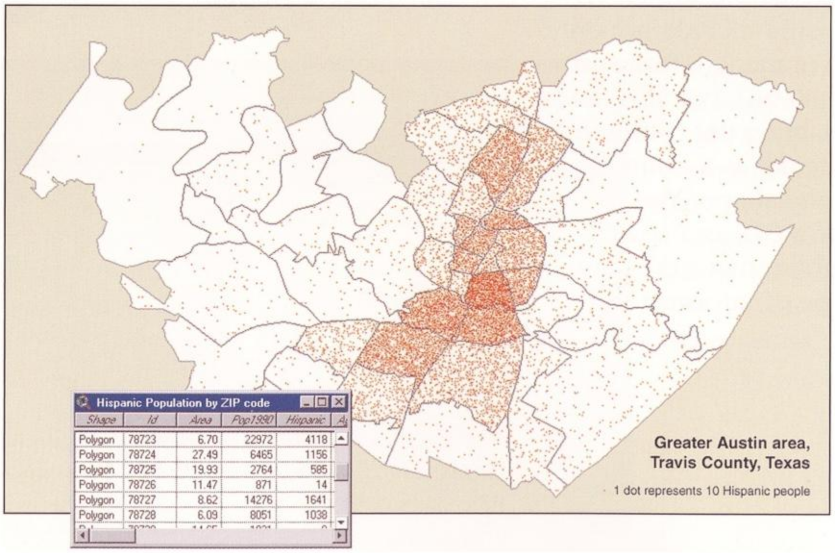
Caution: the same value can seem greater in larger areas.

These values are the same, but don't look it.

Points can be used to fix this.

Sometimes filling in entire areas, which vary considerably in size, results in a perceptual problem. Even when the color and thus the value is the same, larger areas, which display more of the color, naturally appear greater in value than smaller areas. This problem can sometimes be corrected by using points, such as circles, to encode the value for each area, as long as you keep them small enough to avoid occlusion. In fact, your can vary the size and color intensity of the points to both encode differences in values, as illustrated in the example on the right. (Source: Cynthia A. Brewer, *Designing Better Maps: A Guide for GIS Users*, ESRI Press: Redlands, California, 2005.)

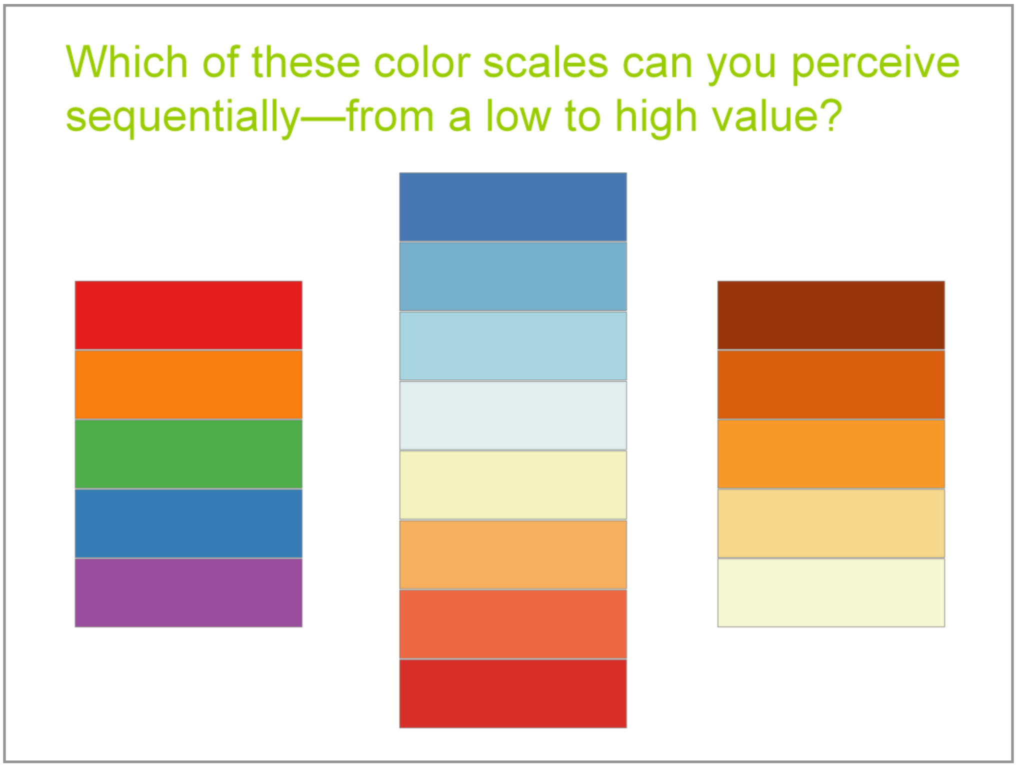
## Use individual points to display the details.



Hispanic Population by ZIP code

| Shape | Id | Area | Pop1990 | Hispanic |
|---|---|---|---|---|
| Polygon | 78723 | 6.70 | 22972 | 4118 |
| Polygon | 78724 | 27.49 | 6465 | 1156 |
| Polygon | 78725 | 19.93 | 2764 | 585 |
| Polygon | 78726 | 11.47 | 871 | 14 |
| Polygon | 78727 | 8.62 | 14276 | 1641 |
| Polygon | 78728 | 6.09 | 8051 | 1038 |

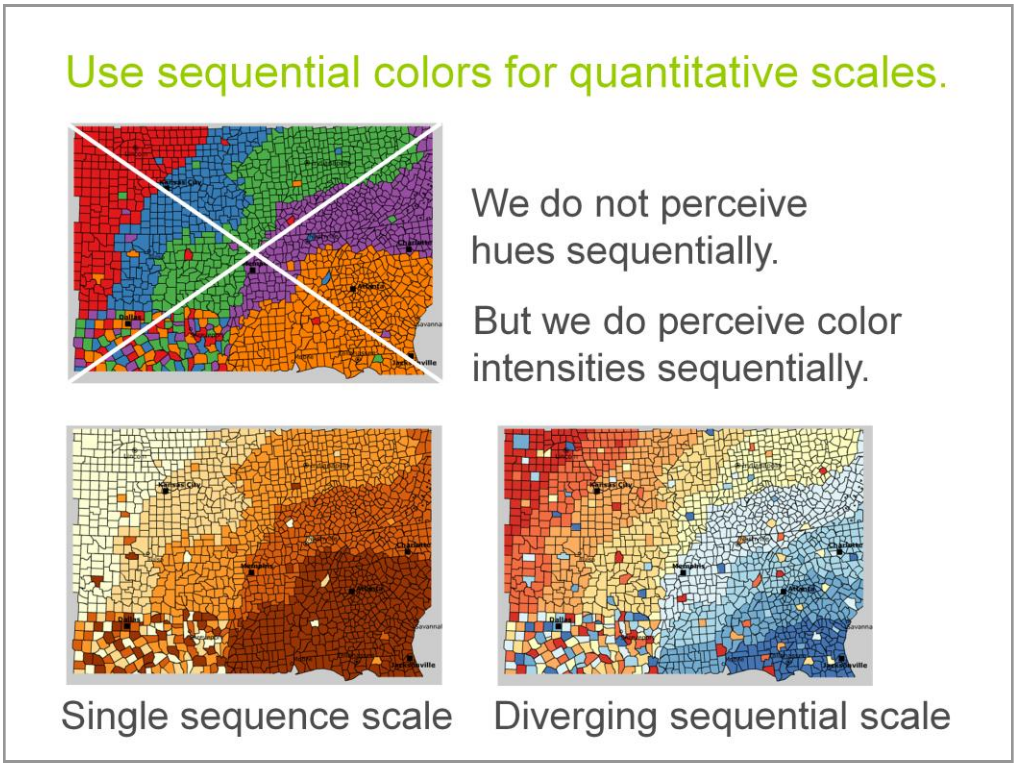
This example, which shows where people of Hispanic origin live in the city of Austin, is capable of revealing concentrations and gaps that would be lost if the data were summarized to the area level. (Source: David Boyles, *GIS Means Business, Volume Two*, ESRI Press: Redlands, California, 2002.)



Which of these color scales can you perceive sequentially—from a low to high value?

As you can see,  we do not perceive hues in a sequential manner, even when they are arranged according to their spectral values as the example on the left is. The best way to encode sequential values using color is to vary color intensity, especially through variations of lightness, from light to dark. This can be done by varying the intensity of a single hue or by varying hue to some degree as well, as shown in the example to the right, as long as you differentiate the colors primarily by intensity as well.

The color scale could be used to encode a single sequential series of values, from the lowest value at the bottom to the highest at the top. Notice that the color scale in the middle consists of two distinct hues (blue in the upper half and red in the lower half). The different expressions of each of these hues vary sequentially. The blue increase in intensity as they proceed upwards, and the reds increase in intensity as they proceed downwards. This is an example of a diverging scale, which is one that has a breakpoint somewhere in the middle, with one set of values the proceeds upwards and one that proceeds downwards from that breakpoint. A typical example of values that a diverging color scale can encode is a set that consists of both positive and negative values, with zero as the breakpoint in the middle. Another example might be values that are above the norm (for example, above average) and those that are below the norm.



## Use sequential colors for quantitative scales.

We do not perceive hues sequentially.

But we do perceive color intensities sequentially.

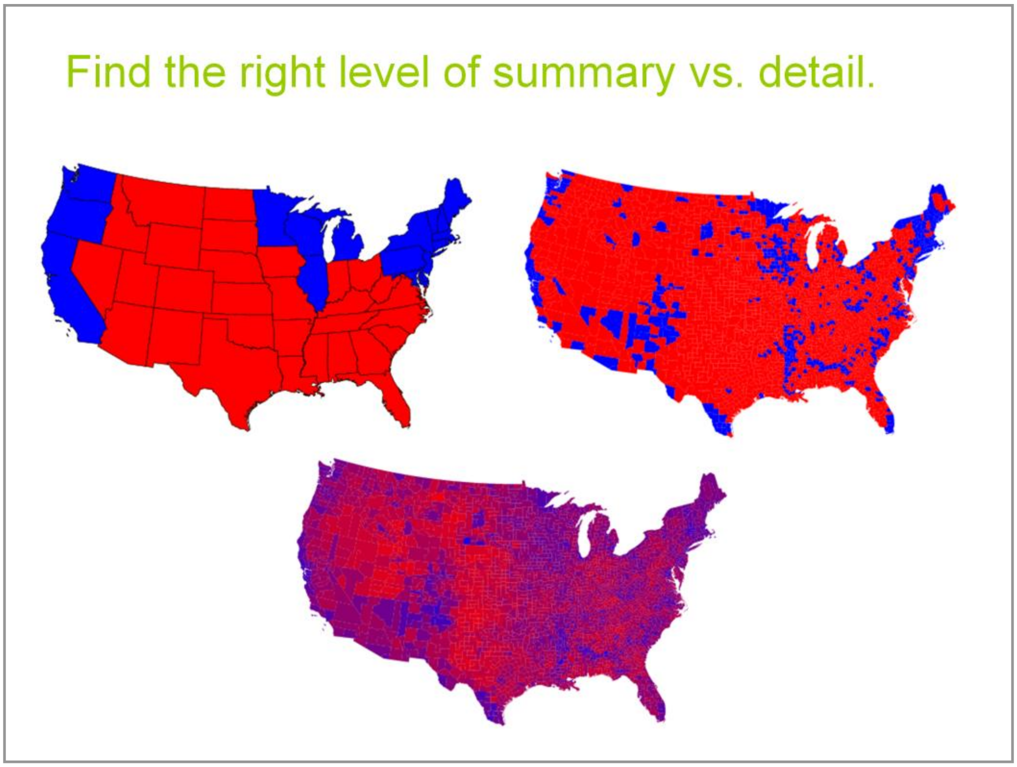
Single sequence scale

Diverging sequential scale

Here we see the same color scales that appear on the previous slide, but this time shown on a map to illustrate how they might be used for geospatial displays.

One of the experts in the use of color to encode data on maps, and about geospatial information design in general, is Cynthia Brewer of Penn State University. Dr. Brewer is a cartographer is the geography department. Take advantage of her free *Color Brewer* application at www.ColorBrewer.org. It is worthwhile as well to get a copy of her book *Designing Better Maps: A Guide for GIS Users*, published by ESRI Press in 2005.

As you can see in the top map, different hues work very effectively for assigning discrete categorical distinctions to data (see the top example), but not for displaying a range of quantitative values.



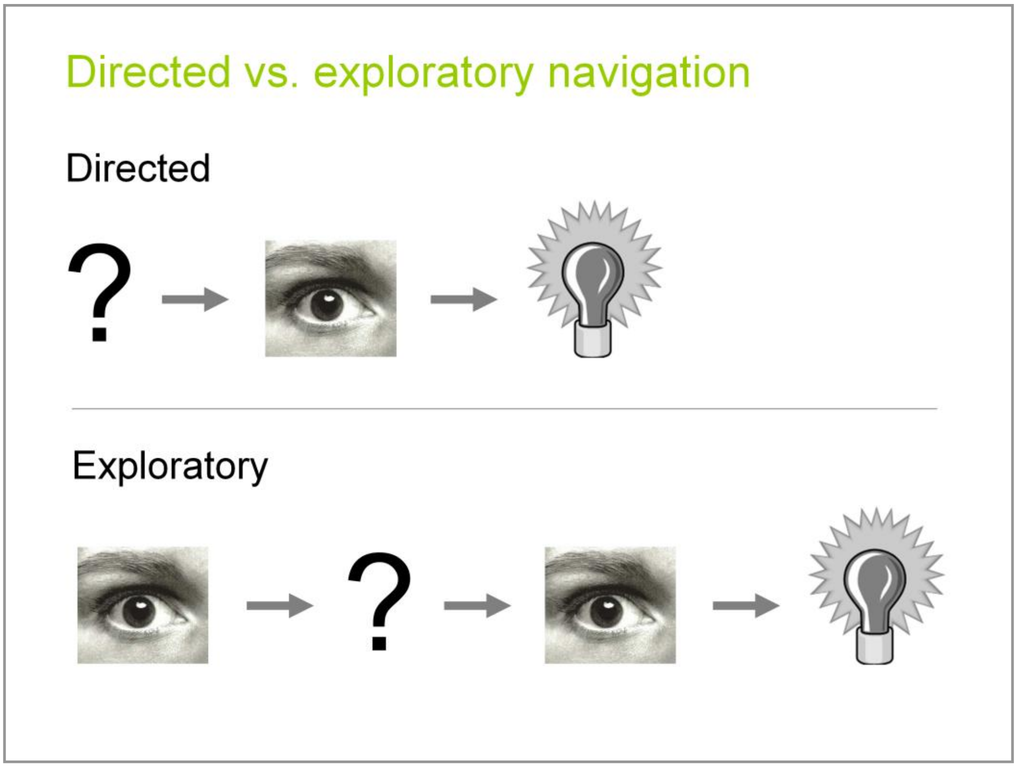
Find the right level of summary vs. detail.

Just as with any means visual display, the picture can differ considerably depending on the level to which you aggregate the values. You always need to find a level that isn't so highly summarized that you lose sight of meaningful details, but not so detailed that no patterns can be discerned. These three maps of the United States each display the same data—who won the 2004 presidential election (Bush = red and Kerry = blue)—but at different levels and in different ways. The top left map aggregates the data to the state level and displays it discretely (either Bush or Kerry), which makes it appear that Bush clearly dominated the election. The top right map displays the same data at the county level, again discretely, which reveals that Kerry won far more votes than was revealed at the state level. If you want a more accurate picture of the popular vote, however, the third map reveals that at the number of votes for each candidate were much closer. It does this by aggregating the data to the county level, but rather than doing so discretely (either red or blue), it blended the two colors based on the actual number of votes for each candidate, thereby revealing various shades of purple with some regions that were clearly dominated by Bush and an equal number that were dominated by Kerry. (Source: Michael Gastner, Cosma Shalizi, and Mark Newman of the University of Michigan, available at www-personal.umich.edu/~mejn/election/.)

## Analytical navigation

Visual data analysis is a process that consists of many steps and many paths to get us from where we start, knowing little, to where we need to be, understanding much. Some ways of navigating your path from step to step are more effective than others.

> *Data analysis, like experimentation, must be considered as an open-minded, highly interactive, iterative process, whose actual steps are selected segments of a stubbily branching, tree-like pattern of possible actions.*
>
> (*The Collected Works of John W. Tukey*, John W. Tukey, Wadsworth, Inc.: Belmont, CA, 1988, pages 5 and 6)



Directed vs. exploratory navigation

Directed

Exploratory

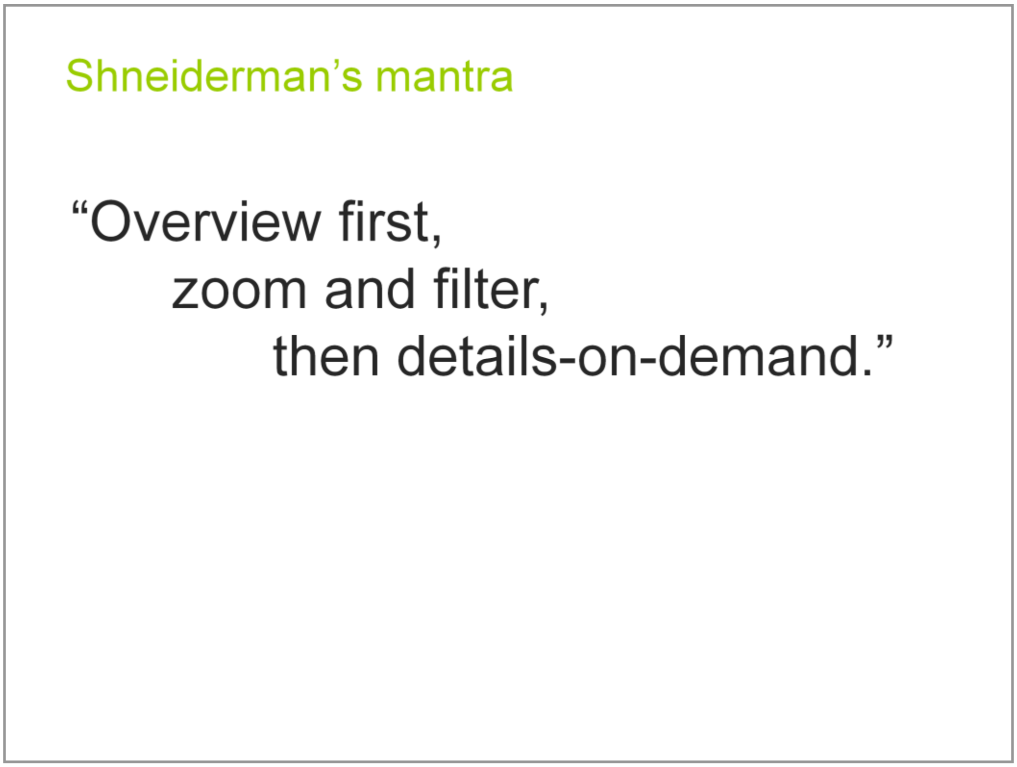
At the most fundamental level, analytical navigation can be divided between two approaches: directed and exploratory.

Directed analysis begins with a specific question that you wish to answer, proceeds to a search specifically for the data that will answer that question, such as a particular pattern, and hopefully results in finding the answer.

Exploratory analysis begins by looking at the data without predetermining what you expect to find, proceeds to noticing things in the data that are interesting and asking a question about it, then proceeding in a directed fashion in search of an answer to that question.

Both approaches are vital to data analysis. The main point I wish to make here is that comprehensive analysis requires that you sometimes start with a blank slate and let the data itself direct you to items worth examining.

> *Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data…To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient because the cost of intensive data analysis is typically very small compared with the cost of data collection.*
>
> (*The Elements of Graphing Data*, William S. Cleveland, Hobart Press, 1994, pages 8 and 9)

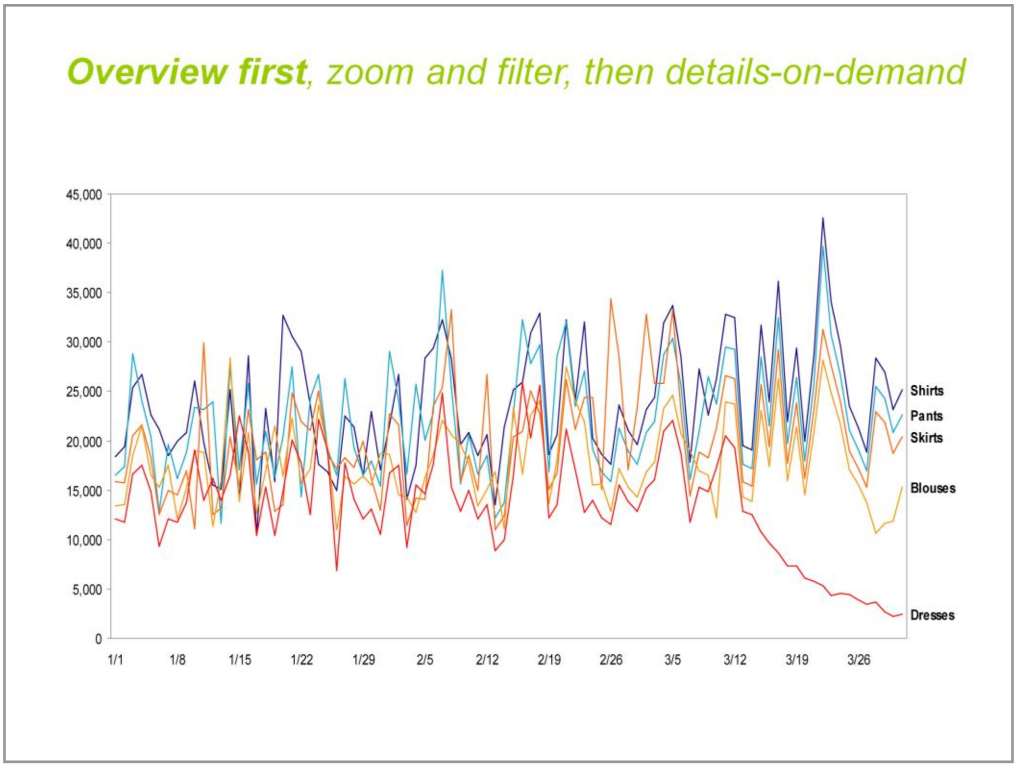


## Shneiderman’s mantra

“Overview first,
zoom and filter,
then details-on-demand.”

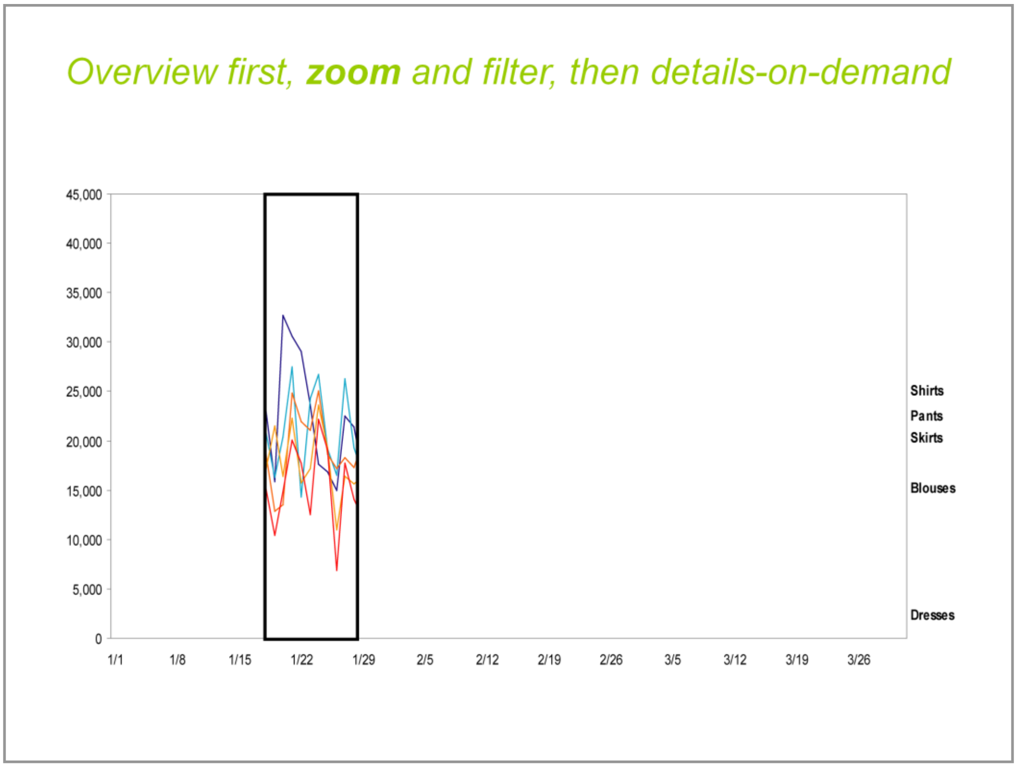
When new recruits by intelligence organizations are trained in spy craft, they are taught a method of observation that begins by getting an overview of the scene around them while being sensitive to things that appear abnormal, not quite right, which they should then focus in on for close observation and analysis.

> *A visual information-seeking mantra for designers: ‘Overview first, zoom and filter, then details-on-demand.’*
>
> (*Readings in Information Visualization: Using Vision to Think*, Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, Academic Press, San Diego, California, 1999, page 625)
>
> *Having an overview is very important. It reduces search, allows the detection of overall patterns, and aids the user in choosing the next move. A general heuristic of visualization design, therefore, is to start with an overview. But it is also necessary for the user to access details rapidly. One solution is overview + detail: to provide multiple views, an overview for orientation, and a detailed view for further work.*
>
> (*Ibid.*, page 285)
>
> *Users often try to make a ‘good’ choice by deciding first what they do not want, i.e. they first try to reduce the data set to a smaller, more manageable size. After some iterations, it is easier to make the final selection(s) from the reduced data set. This iterative refinement or progressive querying of data sets is sometimes known as hierarchical decision-making.*
>
> (Ibid., page 295)



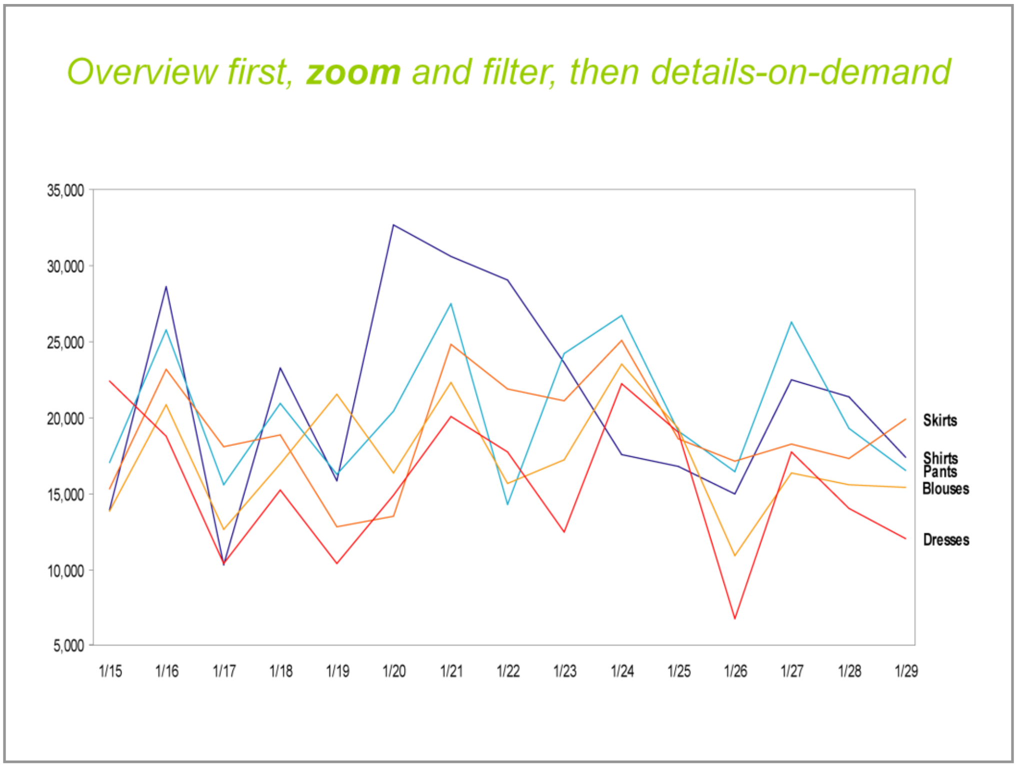
**Overview first**, *zoom and filter, then details-on-demand*

Shneiderman’s technique begins with an overview of the data – the big picture. Let your eyes search for particular points of interest in the whole.



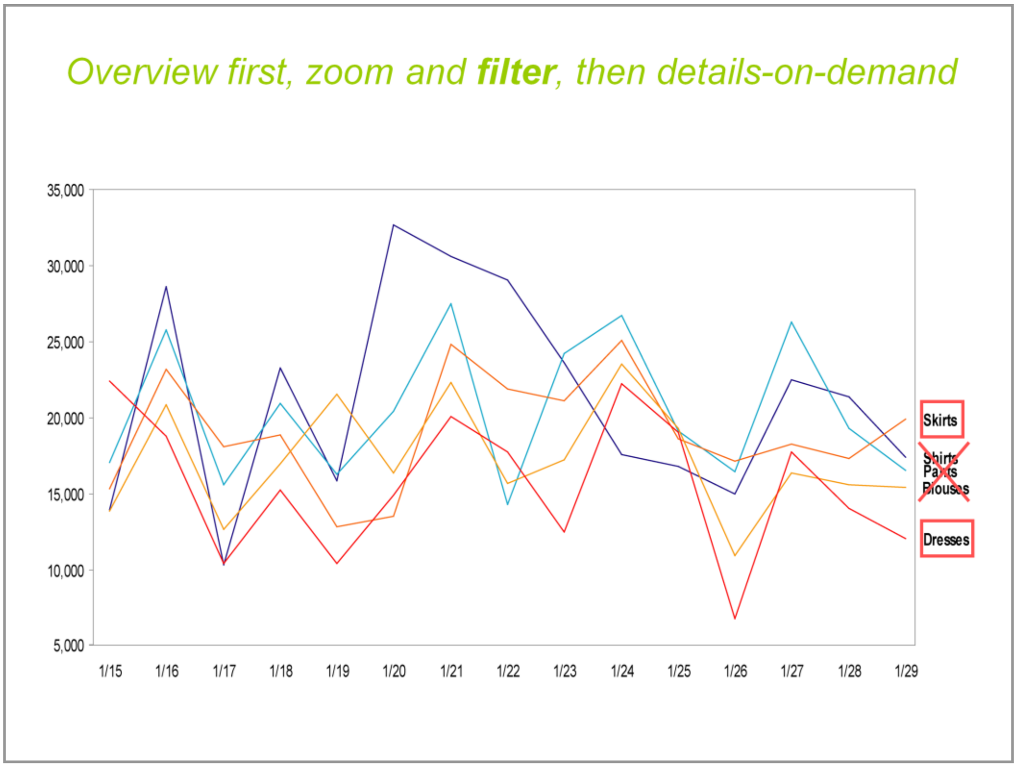
*Overview first, **zoom** and filter, then details-on-demand*

When you see a particular point of interest, then zoom in on it.



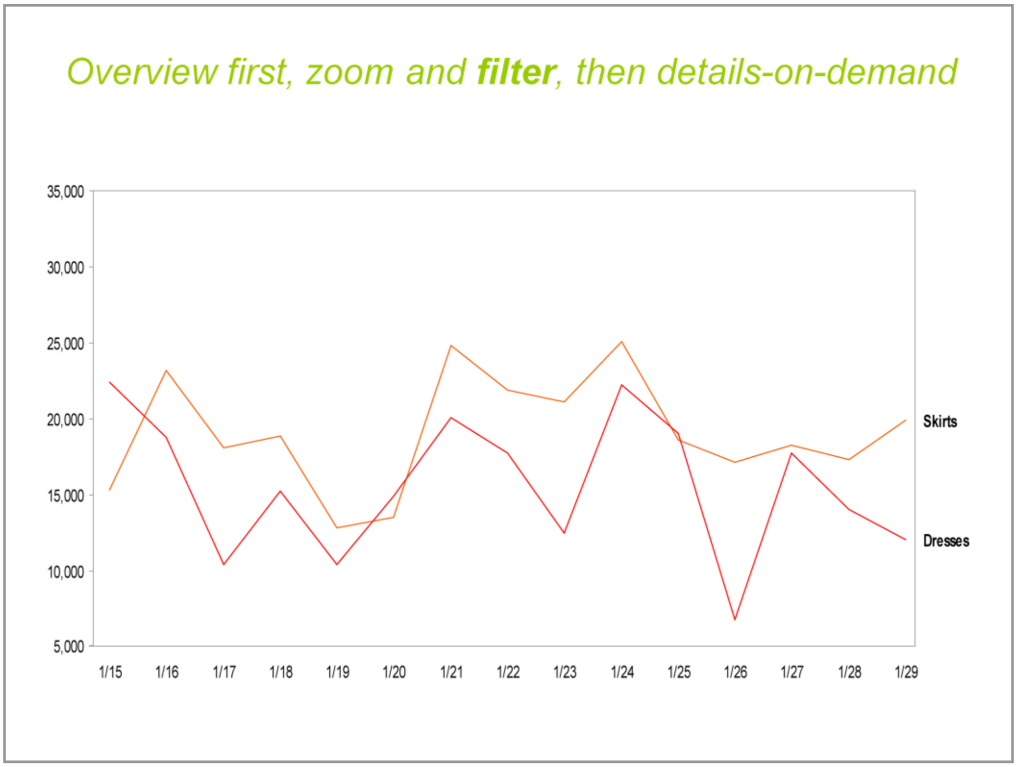
*Overview first, **zoom** and filter, then details-on-demand*

Once you've zoomed in on it, you can examine it more closely and in greater detail.



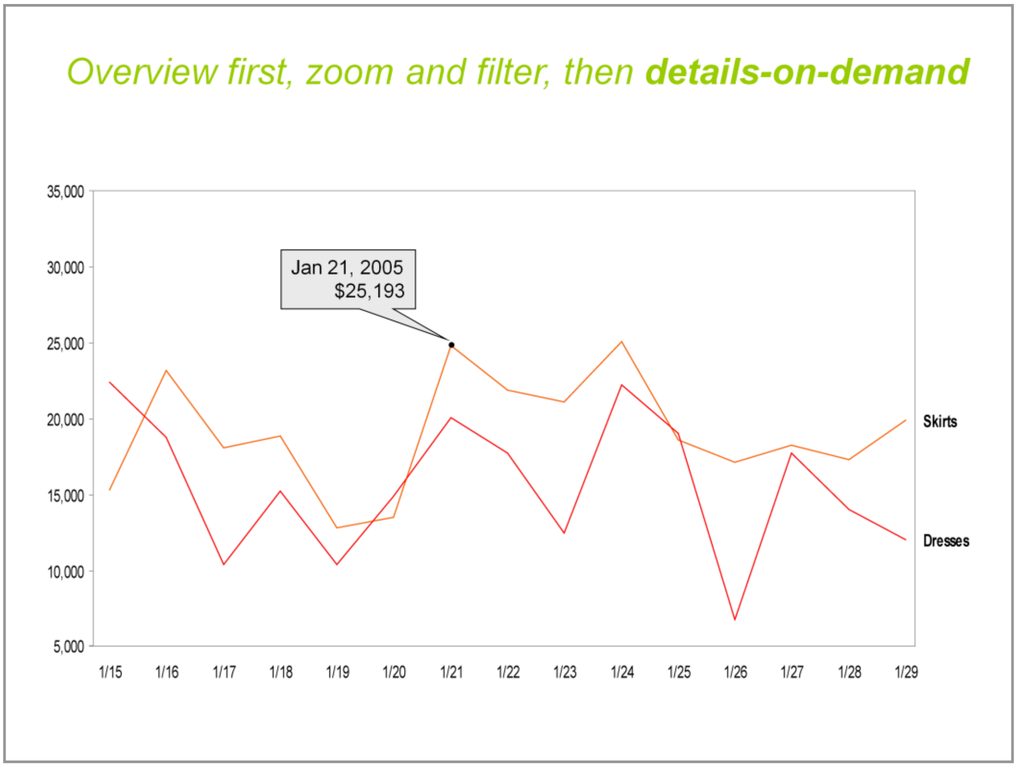
*Overview first, zoom and **filter**, then details-on-demand*

Often you must remove data that is extraneous to your investigation to better focus on the relevant data.



*Overview first, zoom and* ***filter****, then details-on-demand*

Filtering out extraneous data removes distractions from the data under investigation.



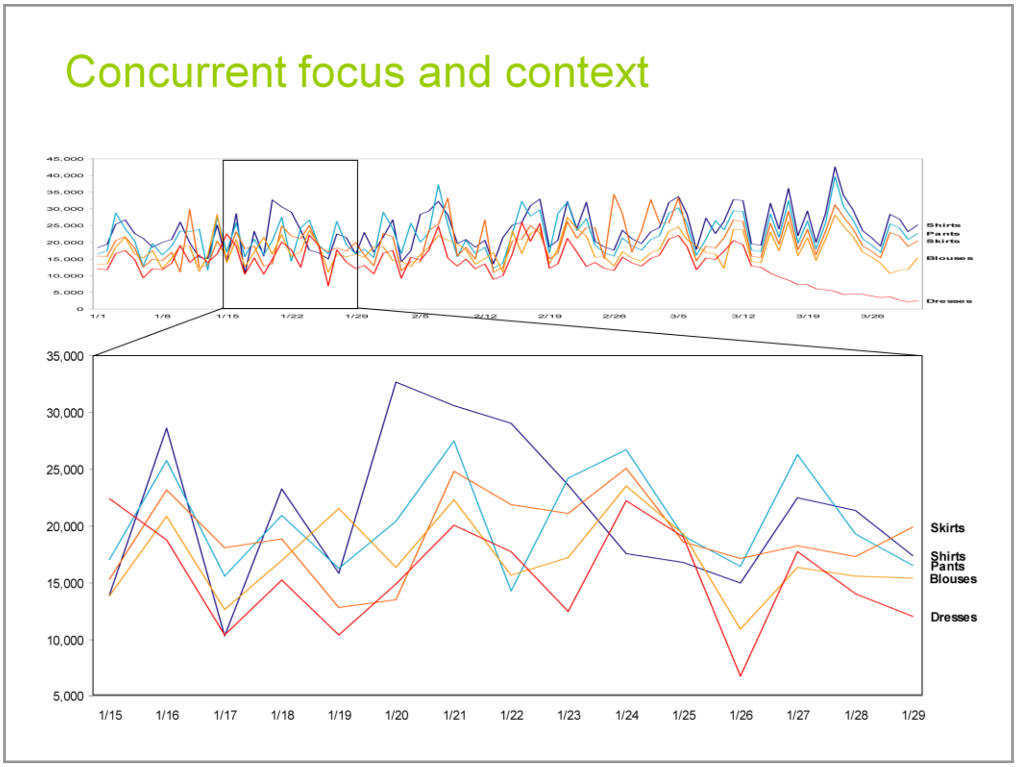
## *Overview first, zoom and filter, then* ***details-on-demand***

Jan 21, 2005
$25,193

Visual data analysis relies mostly on the shape of the data to provide needed insights, but there are still times when you need to see the details behind the shape of the data. Having a means to easily see the details when you need them, without having them in the way when you don't works best.

## Revealing analytical techniques

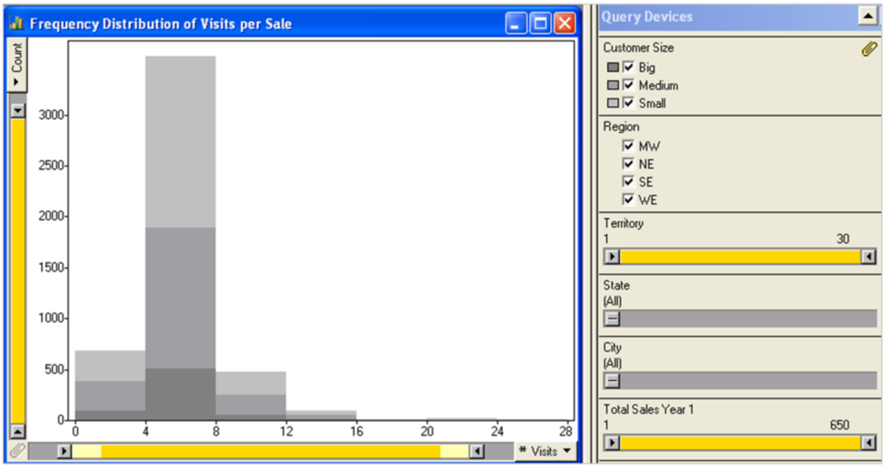
Information visualization research, which occurs mostly in the academic world, has contributed many important visualizations and visual techniques that improve our ability to see meaning in the data. These are eight of the most important, which we'll now examine one at a time.



## Concurrent focus and context

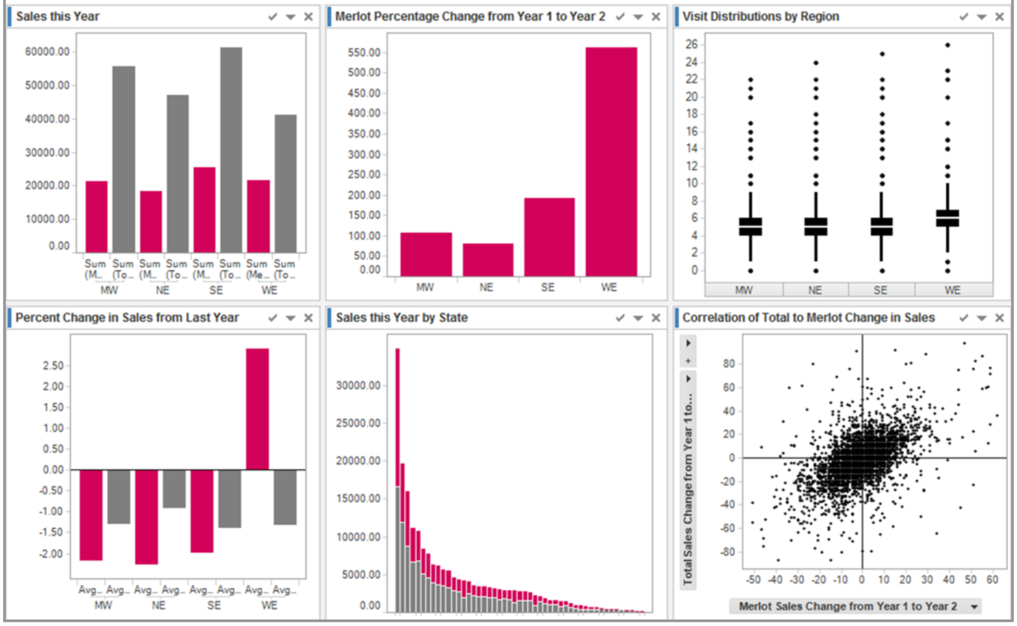
Concurrent focus + context views allow you to delve into details without losing a sense of where you are in the larger context of the whole. They keep you from getting lost among the trees no longer knowing where you are in the forest.

# Direct, dynamic data filters



Direct dynamic interaction with the data allows you to manipulate the data easily and immediately (such as by filtering it), without interrupting your stream of thought.

## Faceted analytical displays, enhanced by brushing



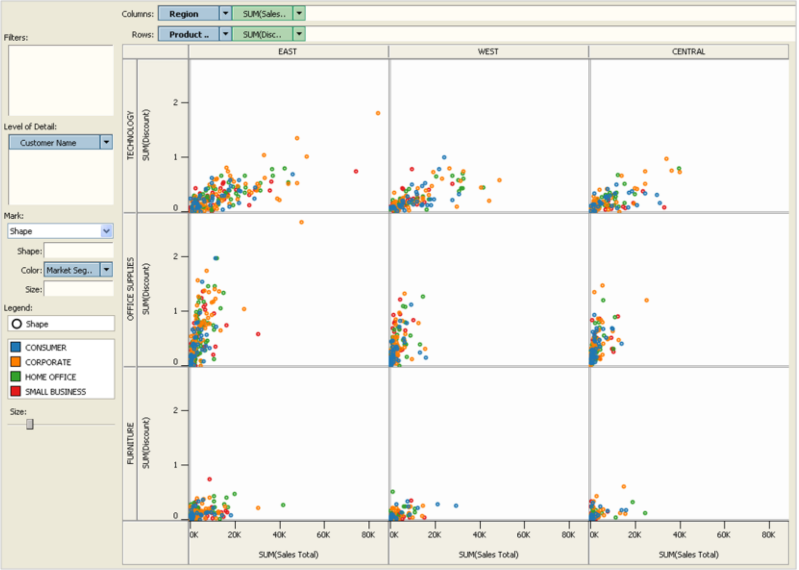
Faceted analytical displays allow you to see the data displayed in multiple ways simultaneously, which provides several perspectives at once and supports comparisons that could not be made otherwise.

When the same data appears in multiple concurrent displays, brushing allows you to highlight data in one display and see it automatically highlighted in the other displays as well.

## Demo: Spotfire DecisionSite

Spotfire

## Trellis and visual crosstab displays



Columns: Region, SUM(Sales..

Rows: Product .., SUM(Disc..

Filters:

Level of Detail: Customer Name

Mark: Shape

Shape:

Color: Market Seg..

Size:

Legend: ○ Shape

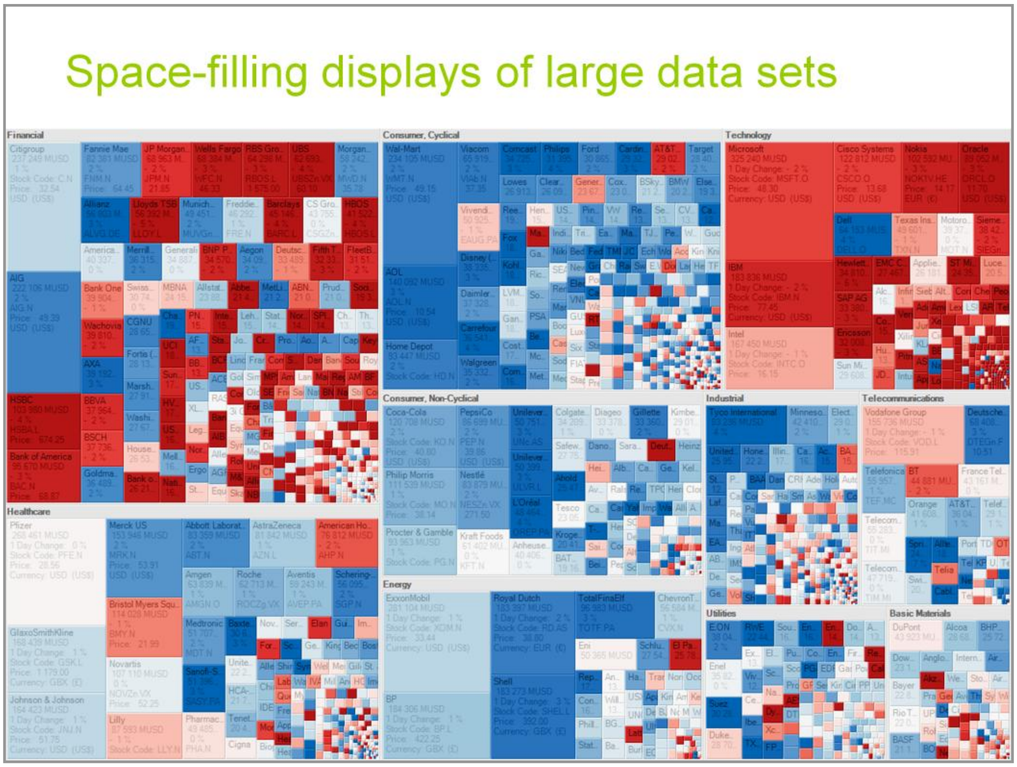
- 🟦 CONSUMER
- 🟧 CORPORATE
- 🟩 HOME OFFICE
- 🟥 SMALL BUSINESS

Size:

Small multiples provide a powerful means to examine more variables together without resorting to graphing methods, such as 3D, which suffer from perceptual problems.

## Demo: Tableau Software

TABLEAU SOFTWARE

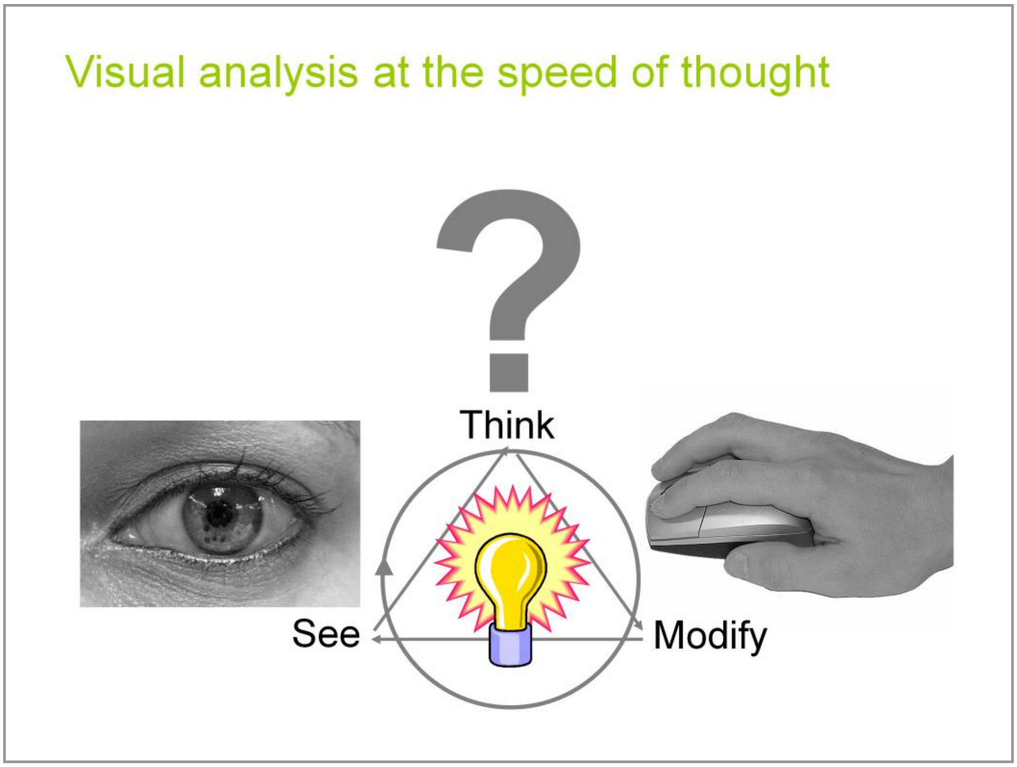


## Space-filling displays of large data sets

At times the volume of data that you must examine is too large for conventional visualizations like bar graphs. Special visualizations, such as tree maps (shown above) have been developed to make optimal use of the limited space of a computer screen to allow us to see a great deal of data at once. Visualizations of this type were not developed to enable precise comparisons between values, but rather to allow us to detect exceptions and significant patterns, which we can then explore with more precision using more conventional methods with subsets of the data.

## Demo: Panopticon Explorer .NET

panopticon



## Visual analysis at the speed of thought

Think

See

Modify

Direct dynamic interaction with the visualized data allows you to see something in the data visualization and interact with it directly to filter out what you don’t need, drill into details, combine multiple variables for comparison, etc., in a way that promotes a smooth flow between seeing something, thinking about it, and manipulating it, with no distracting lags in between.