**Data Visualization is not entirely new!**

The history of data visualization goes way back to 2nd century C.E. Initially, data was arranged into rows and columns, and it evolved to the initial quantitative representations in the 17th century.

They say 'Knowledge is Power'. But how do we make knowledge powerful? Especially when that knowledge comes in the form of data. Lots of data. How do you find the meaning? Tell the story? Share the story? - "Infographics" where data meets design.

What makes good data visualization? Take a few seconds to count 7's in this number set. How many were there? Not sure? Now try.

A simple colour change makes comprehension almost instant. Colour is one of several pre attentive attributes - like size, orientation, flicker visual clues that the human brain processes within 250 milliseconds.

Now imagine that we're not looking for specific numbers, but patterns. We can use colour to show correlation, size to show quantity, orientation to show trends and not just data.

The power of design can be used to better communicate all sorts of information, processes, hierarchy, anatomy, chronology and better communication through innovation. Because, your message is only as good as your ability to share it.

**Origin of Data Visualization**

Let us understand the origin of data visualization, how nature of visualization has changed with changing scale of data. Let's take a look at some modern data visualizations to get some inspiration.

Now, in one sense visualization is nothing new. Data Visualization can be seen in some ways as nothing more than a more elaborate version of statistical graphing; like the pie charts and the line graphs that we're familiar with. On the other hand there is so much more to it than the simple and often simple minded pie charts and graphs that we see.

It turns out that most kinds of statistical graphs have a long history and they all have one important thing in common. In the words of **John Tukey**, the spiritual father to the modern field of data visualization, "**The greatest value of a picture is when it forces us to notice what we never expected to see**." You can find this happening in one of the first known graphics, which is this chart of planetary movements from the 10th Century. It's pretty rough by modern standards. But, even though it's over a thousand years old, it's pretty easy to tell that it's showing the positions of several celestial bodies over time, making it the earliest version of what can be called a **Multiple Time Series Chart**. It's also pretty clear that some of the bodies seem to move around a lot more than others, which can lead to some interesting theories about the nature of the solar system.

Moving forward 800 or 900 years, **two people** made particularly important contributions to statistical graphics. The first was **William Playfair**, who was a Scottish engineer and economist. In this graph from **1786**, Playfair created what is considered by many to be the ***first bar chart***, which is still available right in Excel. In another graph of trade data, he also created a Time Series Line Chart, which was not terribly unlike the early planetary chart, but which was more clearly labelled and showed excellent use of colour. Fifteen years later in 1801, Playfair came out with another innovation that's still with us today, for good or evil, ***the pie chart***. The pie slices are easiest to see in the second pie to the left and the fifth one to the right. I should mention that despite Playfair's innovation, there are a lot of good reasons for not making pie charts and we'll talk about those in a later movie.

While Playfair's graphs were significant for getting the visualization ball rolling, they existed primarily as an early form of annual report fodder, that they simply communicated information and tried to do so in a clear and attractive way. It was the English nurse **Florence Nightingale** who was the first to use statistical graphics as compelling tools for persuasion and policy change. Her best-known chart was the 1858 Diagram of the Causes of Mortality in the Army in the East, which depicted causes of death among soldiers in the Crimean war in Turkey. This particular chart which is a variation on Playfair's pie chart is called a ***Polar Area*** Diagram or sometimes in her honor, a ***Nightingale Rose diagram***. Although Nightingale herself called it a ***coxcomb***. As a result of her presentation, Queen Victoria appointed a sanitary commission that came to Turkey and removed dead animals from the water, got rid of rotten floors and improved ventilation. As a result, the mortality rate there dropped from 52% to 20%, making this perhaps the graph that saved more lives than any others.

**Playfair's and Nightingale's graphs also served to illustrate one half of a potentially important distinction in visualization, which is the difference between what is called Information Visualization and Data Visualization**. Now this is far from a hard and fast distinction, but at least it can help focus thinking about graphics. Essentially, **Information Visualization refers to graphics that are created to communicate information that is already understood by at least some people.** This was the case for both, Playfair's and Nightingale graphs and it's true for most of the Infographics, trotted out in today's newspaper.

**Data visualization on the other hand can be thought of as a graphics that are designed to help researchers find the patterns in the first place.** One of the great historical examples of this kind of pattern searching comes from the 1854 cholera outbreak in the Soho district of London, which eventually, claimed over 600 lives. The predominant theory of cholera at the time was, it was passed by "bad air." However, the physician John Snow charted each case of cholera and found an epicenter at the public water pump on Broad Street. This led to the removal of the pump handle, which may have contributed to the steep decline in cholera at the time, that is some say, it declined because they took the handle off, others say, it was already going down, but either way it was great detective work and an excellent low-tech solution to a serious problem.

But really, Data Visualization is not identical to these graphs from 150-200 or 1100 years ago. Despite the fact that you can still produce most of these graphs in any spreadsheet program, a lot has changed since then, and the nature of visualization has changed significantly. One important change is the ability to automate analyses and graphics with computers, which is something that none of these pioneers had. But perhaps the most important change is the scale of data available and how the modern deluge calls for new and different methods. When a dataset has hundreds or thousands of variables and possibly millions or billions of cases, it's simply not possible to go through manually and do one variable or one correlation or one pie chart at a time.

**Basic Visual Elements of Design**

**Visual aspects** - Depending on the type of data or information being conveyed, some design elements are better than others when used in graphics.

**Colour** should be used to ***create an aesthetically pleasing visual*** and should relate to or at least ***not take away the meaning of officials*** in your overall design. Note that Colour is one of the **most effective ways to convey nominal or category based information in charts**. For example in a line chart comparing two things, if those lines and points are different Colour, the audience will easily be able to identify which is which. Use of contrast when comparing items has the effect of clearly delineating one aspect of comparison with the second.

**Position** is the most ***effective and accurate method of conveying quantitative, numerical and ordinal ordered data in charts and impacting graphics***. **Shape** like square, circle, triangle is somewhat ***effective when used to communicate nominal data***. It is not very effective or accurate at conveying quantitative or ordinal data. **Size and length** are ***very effective and accurate in conveying quantitative data.*** Less so for ordinal and nominal data. Position and size - when viewed as a whole, the composition says a lot about relative importance.

**It's all about perception** and we perceive differences in quantitative values very accurately when size or length is used. **Headers are prominently displayed. Relatively large and at the top of the visual. Less relevant information is best located at the side and relatively small in size.**

Colour is powerful, but has its nuances. **Colour schemes are subjective choice for the overall design of the visualization**. One exception in subjectivity is when ***creating a work-related visualization***. In this case, you should follow your ***company's branding using shades*** and complementary Colour to complete your palate. **Bold Colour** can **convey** a **message**. **But**, could clash or be garish. **Don't overuse bold Colour. Colour theory provides a good starting point when designing a visualization**.

In charts, **shape is less effective than Colour for differentiating data**. Have a look at this scattered chart

**Edward Tufte's Principles**

So first, let's talk about ‘**Chart Junk’**. Wikipedia defines chart junk as “**All visual elements in charts and graphs that are not necessary to comprehend information that's represented on the graph or that distract the viewer from this information**”. In order to make this a little more clear for, I actually made a few over-the-top examples of what this could entail. For example, check the text “Texture and Shading”, it's a size that even I can't read it. It seems so small and has weird font that you would want to avoid. Basic thing you want to avoid is anything to distract the audience from focusing on the data. In other words, it’s representing simplicity in design and simplicity in function. For example, in your data visualization, you don't want to have weird fonts or colors or bolded fonts. You just want a simple standard text and the same with their data labels too. You want to standardize font, no coloring, no bolding and no size distractions.

Hopefully, the idea of chart junk is not a new concept for you guys. I am thinking that for some of you guys, the idea of Data Ink and Data Ink Ratio probably will be new. Just because it really has its home here in data visualization.

What **data ink** is, and here's what he says. He says, “**Large share of Ink on a graphic should present data information, the ink changing as the data change. Data Ink is a non-erasable core of a graphic, the non-redundant Ink arranged in response to variation in the numbers represented.**” Basically, that's kind of a mouthful. But what he's getting at is that, in any particular graphic design, say we have a bar chart, there is a minimal amount of ink or substance in a chart that needs to be there that really defines the core of the data itself and the core of your message. But there's also a lot of opportunity to add distracting items and excess information that you do not need. We're going to go over a couple of examples of those.

Finally we have Edward Tufte's “**Focus on Substance over Design**”. Borrowing from Tufte himself again, in his book he says, “**Induce the viewer to think about the substance rather than about methodology, graphic design, the tech of graphic production, or something else**”. So, in terms of a high impact presentation, Edward Tufte really understood that at the end of the day, the purpose of data visualization and the purpose of representing our data in bar charts, lines is to really display a message. It's not to figure out who can come up with the most creative way of displaying information. It is a way of effectively and simplistically displaying information for another audience and that's what makes the work of Edward Tufte so great. It’s his focus on simplicity throughout his career and throughout his data visualization efforts.

**Principles of Gestalt Psychology**

The Principles of Gestalt Psychology will provide techniques that could be used in design. This will make it easily perceivable for a user and help avoid some common mistakes.

So, these eight principles are: **Proximity, Similarity, Closure, Symmetry, Common Fate, Continuity, Good Gestalt and Figure and Ground**.

So let's start with **Proximity**. For example, you can see here that all the dots are having equal distance. So you think that this is actually a single object. But in the right hand side, you can see that these dots are divided into three parts. The first part is the first column and second column of the dots, then the second part is the third column in the fourth column of the dots and the third part is the fifth column and the sixth column of the dots. So what we analysed from this fact that our mind perceives these objects which are closer to each other as forming a group. So basically, it may not be a group but we think it as a group because of the proximity or the distance they are in. In summary, **the law of proximity states that, when an individual perceives an object, they perceive objects that are closer to each other forming group now.**

Let's move on to **Law of Similarity**. Now, this law states that, **if you perceive objects which are similar to each other, then you think that they are actually grouped together**. For example, in this diagram, you can see that there are black circles and then there are white circles. You can see that there is a horizontal line forming from the black circles and a horizontal line forming from the white circles. You will not see the vertical lines because, then you will have the pattern white black, white black - which is not similar objects. But you can see all the white circles and the black circles as they are similar, they seems to form a group. So this is Law of Similarity.

Moving on to the next law, which is **Law of Closure**. To understand the Law of Closure, if you see the image on the right hand side, you can see that there are three circles which seems to form white triangle. But in reality, there is no white triangle. This is something which we perceive, because they are positioned in certain way. So, our perception is actually filling the visual gap. Even if you see the triangle which is in the background, which we think as a triangle, is not a triangle at all. It is actually three arc’s which are just positioned like three vertex of the triangle and we think that the white triangle is actually on top of the background triangle. So, we are not able to see the entire triangle. But in reality, we have only three arcs and 3 circles. In the picture, on the left hand side also you can see that there is no circle and rectangle drawn. But we perceived that there is a circle and then there is a rectangle. To summarize the Law of Closure, **it states that individual actually perceives that object has shapes, letters, pictures, etc. as being whole when they are really not.**

So moving on to **Law of Symmetry**. Now, what you see here is, a square bracket first, then a curly bracket group and then again a square bracket. So we tend to observe these pairs of symmetrical brackets rather than six individual brackets. It is because we see two brackets as a whole and not as individual things or entities. So basically, the Law of Symmetry states that the **mind perceives the objects as being symmetrical and forming around a central point. It is perceptually pleasing to divide object into an even number of symmetrical pattern**. So, it’s very interesting fact to know that our mind is actually very happy to divide the objects into even number of some symmetrical parts.

So now moving on to the **Law of Common Fate**, to understand the Law of Common Fate, I have drawn few circles. Now, if I do a little animation here, can you see a few of the circles moved to the right hand side? Now, what you will think of those circles which moved to the right hand side? They are actually a group. But, in reality it can be or it can be not. So, let me do that again. So, if I come back to the left and then I go back to right again. So the Law of Common Fate tells us that **if an object or multiple objects are moving in a single line or a single part, we perceive that they are in a group**.

Moving on to the **Law of Continuity** - Law of continuity tells us that **we perceive an object to be made up of continuous objects**. Any point we see that the continuity is breaking, we don't consider that. So, here we see that the cross is made up of a greater than and lesser than sign.

Moving on to the **Law of Good Gestalt**, we see an image which is made up of a rectangle, a triangle and circle. So, basically what we are seeing is three different objects, instead of one single object. Law of Good Gestalt says that, **elements of an object seem to be perceived are grouped together if they form a pattern that is regular, simple or ordered.** For example, if you are presenting a date and your data is showing a pattern which is actually shown in this image, the user will actually see the pattern in three different distinct sets. The one that's a rectangle and other one is triangle and the third one as a circle. This gives us a good idea about what the user will think instead of what we want to show them.

So moving on to the last law, which is the **Law of Past Experience**. So, the law of past experience **implies there are certain circumstances, when visual stimuli are characterized according to past experience**. So, in this two images, you can see one on the left hand side and one on the right hand side. The elements used are similar, because of their position. If I put this two filled circles outside of my main circle, then you will really not see anything important here. But, once I put them inside the biggest circle with this arc, you will recognize from your past experience that this is forming a smiley.

**Introduction to Visualization Techniques**

Visualizations have become increasingly vital to communicate real-time, actionable insights to support decision-making. Failing to choose appropriate visualizations leaves the audience wondering what they should do. Choosing the right type of visualization depends on a number of factors. It depends on what you need to show, what information the audience needs and the level of detail they need.

**Right Type of Visualization**

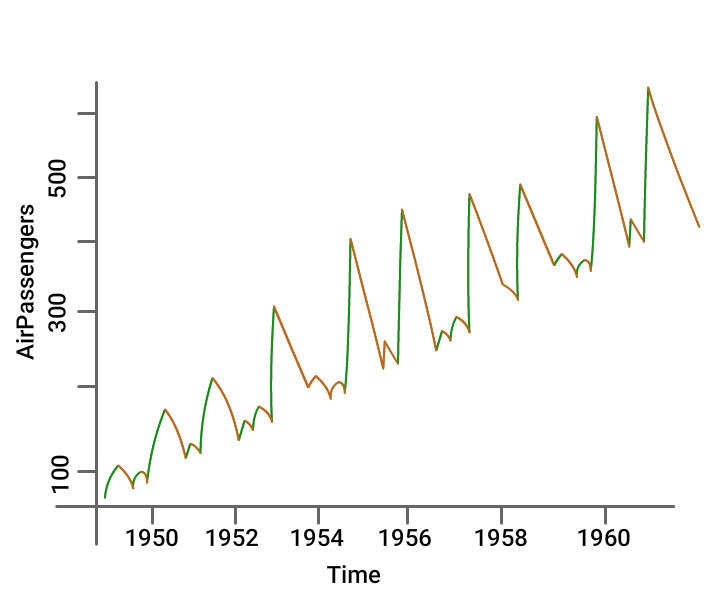
Choosing the type of visualization for the data. Quite often text and numbers by themselves lack impact and do not express the story behind the data or at least the story that you're trying to tell. They may also not convey the data well, if there's a pattern or some other aspect of the data that you're trying to communicate. Numbers and text alone in this case just are not adequate. You need to choose graphics and visuals that match the data and look pleasing to the audience for maximum impact.

Simple effective graphs used two axes to depict the data using X and Y coordinate. Typically at least one coordinate is a quantitative value. The data must map to two variables relating to your X and Y coordinates. In this example, amount and time. The amount is on the y axis and it is in the degrees Fahrenheit and the time across the x-axis is a particular day. Graphs are particularly powerful for conveying a trend or pattern in this type of data. For example, in this graph we can infer that, there is a warming trend occurring from day 1 to day 6.

Certain diagrams work particularly well with particular information or content types. For information comprised of changing data over a time progression, a time line diagram is most effective. When the information comprises a guide or a plan to be completed, a template diagram based on the guide or plan is typically called for. When you have an ordered set of instructions to visualize, a flow chart is very effective and provides a clear coherent and structured order that the audience can follow. If the content type consists of tasks to be crossed off as they are completed, a checklist is a good candidate for visualizing the information.

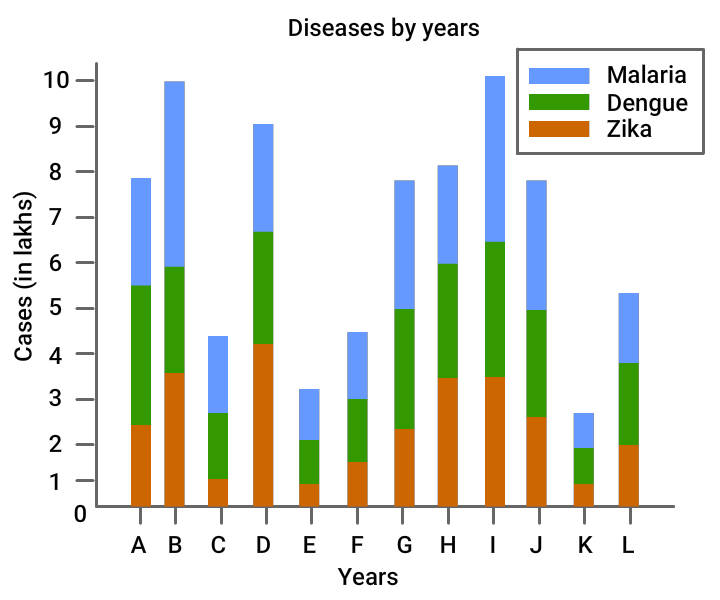
***A mind map is a dynamic technique used to visually present the linking of ideas. A mind map is typically generated around a single topic, with the main topic being central to the diagram. Major topics related to the main topic are connected directly to the main with subtopics and other concepts radiating outward. Mind maps can incorporate words, images, numbers and color. It can present an overview of a central topic with large amounts of related information***.

**Basic types of Charts and Plots**



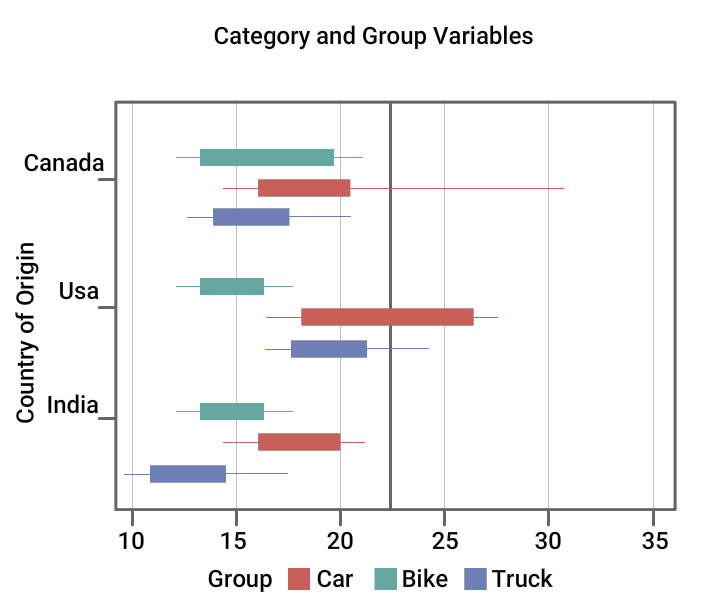
**Line Charts** are commonly preferred while **analysing a trend over a period**.

Further, the line plot is also suitable where there is a need to **compare relative changes** in quantities across some variable like time.



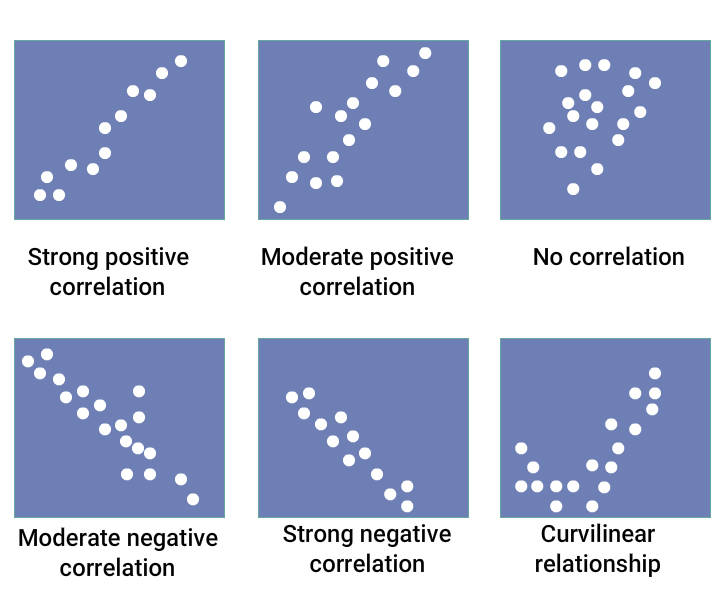
**Bar Plots** are suitable to depict a **comparison between cumulative totals** across several groups.

**Stacked Plots** are used for bar plots for various categories.



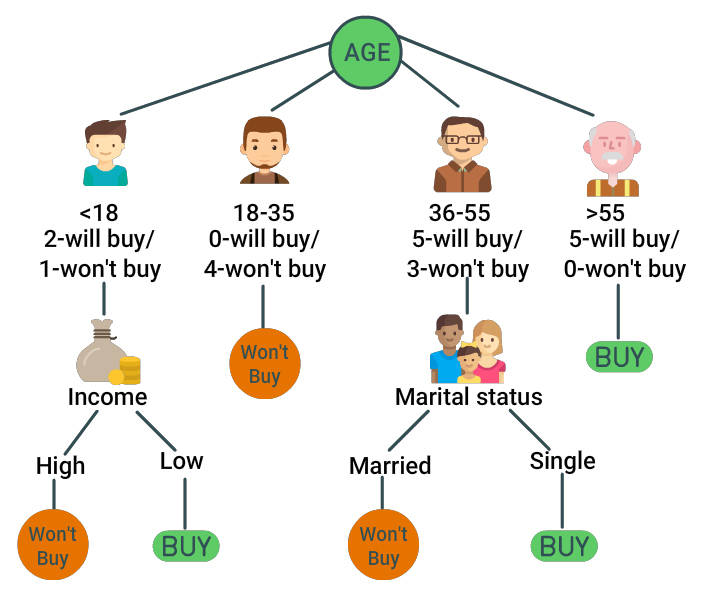
**Box Plot** shows **five statistically significant numbers - the minimum, the 25th percentile, the median, the 75th percentile and the maximum**.

It is useful for visualizing the **spread of data and for deriving inferences** accordingly.



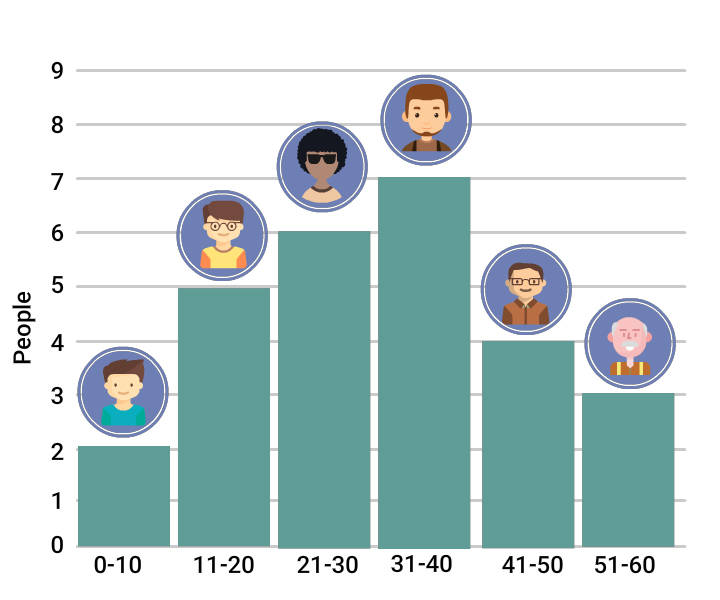
**Scatter plots** help in inspecting multiple variables simultaneously by **color coding**.

Scatter plots reveal the **relationship or association between two variables**; the **extent to which one variable is affected by another.**



**Decision Trees** are excellent tools that help in choosing the right action among several courses of actions.

They provide a **highly effective structure to lay out options and investigate the possible outcomes of choosing those options**.



**Histograms** are used to plot **quantitative data,** and the **ranges of the data are grouped into bins or intervals.**

***Histograms show distributions of variables while bar charts compare variables***.

The human eye cannot visualize circular distances as accurately as linear distances.

Simply put, anything that can be expressed in a pie chart is better represented as a line graph.

**Interactivity in the Data Visualizations**

So far, we've seen a few commonly used visualizations that enable users to interpret data. To make data interactive, data visualizations use technology to drill down data into charts and interactively change what data the user sees and how they process it. This will enable users to explore data sets on their own and find patterns in data that were not evident before.

Common Types of Interactivity

Let's talk about interactivity. Interactive data visualizations allow users to manipulate some aspect of the data set or the presentation of the results, allowing those users to gain insight from the data at a level they require. For example, I can select a different region in this visualization and zero in on exactly the results that I am looking for. This type of control over the result set effectively allows users to get answers to different questions that they might have about the data.

By providing interactive data visualizations over the web, we can achieve - Scale, because the web is a massively scalable platform. We can achieve - Reach, because we have the potential to reach a more extensive audience. And being web based, allows us to take advantage of the maturity of the web as a delivery platform. The ability to employ rich controls and interface elements using a wide array of tools, libraries and other resources. A good example is the free D3 JavaScript library. This cool website offers a diverse and powerful JavaScript library, for manipulating data driven documents.

Benefits of interactivity include: ‘Immediacy’ - the data changes in response to user input. Who can then, explore the data and the result in changes immediately. The convenience of ‘Self-Control’ - the consumer has the ability to manipulate the visualization on their own. It also opens the doors to ‘Collaboration’, then their ‘Simplicity’.

Data visualizations can be designed to be simple with controls that are easy to use and recognize. The consumer does not have to perform any technical or programming activities like redesigning a database query. In this example, I don't see the word that I'm looking for in the word list already. So, I will add that word to the add word or phrase dialogue. So I will add that word to the add word or phrase text box. It gets added to the query and I can see the results immediately. I didn't have to change any database query or anything of the sort.

Now let us talk about some Common types of interactivity.

1. Select - Common forms of selection within visualizations include mouse hover, mouse click and region selections. Here, I click on an icon to change my resulting query.
2. Filter - Filtering is important to the visualization process, because those interested in the data rarely want to visualize the entirety of the dataset at once. In this example, I can grab one of the handles of my dataset across the timeline and compress it and change the result set by filtering only a subset.
3. Brushing and linking - Brushing and linking is the process of selecting or brushing items in one display to highlight or hide corresponding data that is linked in some data dimension in another view. Here is a link series of scattered plot views of data on Iris flowers. Here is a link series scatterplot views of data on iris flowers. Brushing some of the data in one view, highlights the corresponding data in another view.
4. Zooming and Scaling - Zooming and Scaling offers a way to navigate data. Here is a line chart comparing temperatures. In this example, I can choose a region and zoom in on it for more detail.

So, in this video you saw how to identify data interactivities in data visualizations.

Parallel Coordinate Plots

In this learning module, we will consider the ways in which Spotfire parallel coordinate plots may be used to visualize data, including their utility in evaluating multivariant data, which occupies vastly different data ranges. We will consider the normalization, which is applied in order to make this possible, because the dynamic nature of these calculations plays a big role in the interpretation of the patterns you see displayed in the parallel coordinate plots.

Note that you could consider representing your multivariant data in the form of a line chart, where increasing values for each dimension are represented along the Y-axis of the line chart and the items described represent individual lines, or patterns – in this example, the composition of different food items. As you can see, soybeans are high in protein and mustard seeds are high in fat. However, take a look at the comparison of fibre quantities between these different food items. The range of values is much narrower. So, what happens when the items you are comparing are represented by measures on extremely different scales?

For example, let’s say we are in the market to buy a new car. If we were to compare a three other cars (labelled car A, car B, and car C) to our current car, in a line chart of the values for price, mileage, passenger space, cargo space and engine power. The numerical range of the price values would overshadow all other numbers, making it impossible to reasonably compare all dimensions of the data.

Instead of a line chart, if we insert a parallel coordinate plot, the data is normalized to scale between 0 and 100%, based on the minimum and maximum value for each measure, making an equally weighted comparison of each measure possible. Now we can select our new car, based upon the desired magnitude of values along the range of each of the metrics we have available. If we are focused on fuel efficiency and low price, car C may be the one for us. However, we may have to sacrifice space and engine power in that choice.

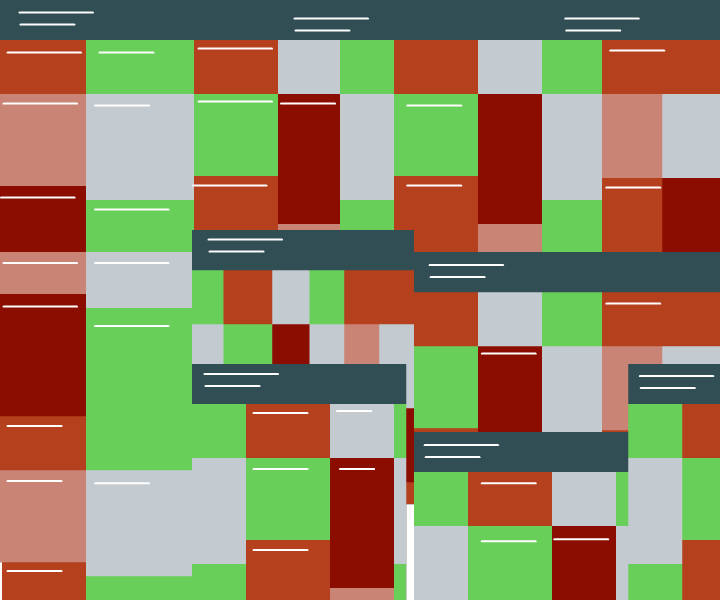
Now, let’s take a look at a demonstration, which highlights the features of several parallel coordinate plot visualization examples. Use Open File, to navigate to the example with Interpreting Parallel Coordinate Plots in the title. Click Open, and agree to the terms and conditions of use. The data table which supports the parallel coordinate plots presented on the next two pages, contains purchase amounts, made in different departments, collected from a group of 754 loyalty program customers across our chain of 84 Mega Mart stores. On the next page, this parallel coordinate plot has been configured to display each customer’s purchase pattern across the seven departments as an individual line. You will notice that the Y-axis scale reads 0 to 100%; therefore, each of these points with values of 100% represent the highest purchase amount in that department. As we hover our mouse over a line, it will be outlined or highlighted, and reveal the relative purchase amounts that the customer made in other departments. For example, this male customer made relatively high purchase amounts in Furniture, Grocery, and Clothing departments, relatively low purchases in Toys and Garden departments and, just on the low side of mid-range in Electronics department purchases.

If we wanted details for the ‘big spenders’ in these first five departments, we could drag a rectangle to mark those patterns. Note that, crossing any portion of a line with this marking area will mark the entire pattern within a parallel coordinate plot visualization. And, we can now evaluate the data displayed in the Details-on-Demand panel. While the Y-axis displays the normalized values of 0 to 100%, the actual values in each departmental column for the marked customers, supplied by the underlying data table, are displayed in the Details-on-Demand table.

If we wanted to see the scale for these values displayed on the parallel coordinate plot, we could click on the department column label of interest, like shoes, and the range of values for that column will be displayed along the right-side of the plot. While the shoe purchase amounts for all 754 customers range on the scale of roughly 0 to 300, clicking on the Furniture label on this axis, reveals the scale for that department, which ranges roughly between 0 and 2200. As with other visualization types, these scales will adjust with filtering. For example, if we remove Los Angeles customers from the visualization, the Furniture scale changes, because some of the largest Furniture purchase amounts were filtered away. Note that the Y-axis scale still reads 0 to 100%, indicating that the normalization calculations are dynamic, and will adjust when maximum or minimum values are filtered out from the plot. So, if we filter out customers in Urban store settings, the patterns adjust to the new maximum and minimum values in each column. This is a key element for understanding and interpreting parallel coordinate plots. The patterns displayed by each line, represent the values relative to the other lines available in the plot. We could illustrate this relative comparison in an extreme situation, by filtering to show only two customers. In this case, one customer must have the maximum purchase amount in each column and the other must have the minimum purchase amount in each column. If we reset all filters, you may find that, due to the dynamic nature of the normalization calculation, trellising parallel coordinate plots (for example, dragging and dropping store location on the final trellis target) is a better method for isolating patterns of interest, due to the fact that the lines are not removed from the plot, and therefore the patterns remain the same, even though the lines appear in separate panels.

If we click where there is nothing in order to unmark data and move on to the next page, you can see that categorical values have been applied to the trellis properties of these two parallel coordinate plots. The plot on the left, has patterns separated into panels based upon assignment to store location and store setting values. While the plot on the right, displays a panel for each of the 84 stores. Within each of those 84 panels are individual lines for each customer’s purchase pattern across departments. Note that in the plot on the left, each department column label is preceded by an aggregation. The legend reveals that, these lines are not configured to display a line for each customer, or row in the underlying data table. Instead, the Line-by property has been assigned based upon Store Number. Therefore, there are 84 lines in this parallel coordinate plot, each representing the relative total purchase amounts made in each department in that store. So, we can see that this rural New York store, Store Number 24 according to the tooltip information box, has some of the highest Shoe and Toy purchase totals. And, marking that store, will mark the lines for the individual customers who contributed to this summary pattern. Even higher Toy purchase totals can be seen in this urban New York store, store number 32. And, marking that pattern allows the customer patterns, which contribute to that store’s relatively high toy purchase totals, to stand out.

Tree Maps



Treemap is a 2D visualization technique for quickly analyzing large, hierarchical data. Each data element in a treemap is represented as a cell. The cell arrangement, size, color are each mapped to an attribute of that element, and these cells can be grouped by common attributes. Treemap provides users with the ability to see both the high-level overview and finer details of data.

In this learning module, we will consider the ways in which Treemap’s may be used to visualize data, including their suitability for evaluating data which has been organized into a hierarchy. As we will see, a series of nested rectangles represents this data organization and the size and color of each rectangle can be configured to represent values from the underlying data table.

The first thing you should know about treemaps is that, you must have hierarchically organized data in order to make this visualization type useful. The rectangles in the treemap represent the hierarchically-organized groupings of data. In this example, four city locations are the highest level of the hierarchy, individual store numbers are organized within each location, and departments within each store are represented by colored rectangles. The color and size of rectangles are then set to represent the magnitude of a continuous variable or variables for the purpose of relative comparisons.

In order to appreciate the interactive features of treemaps, let’s open a configured analysis example in Spotfire. Use Open File and navigate to the file with Interpreting Treemaps in the title. Click Open, and agree to the terms and conditions of use. This data table represents values collected from the sales of products at an office supplies company. Note that sales are organized into different regions and assigned to one of 76 different sales associates. For each order, there is information about the price and profit for the products in those orders, as well as a discount amount which is left to the discretion of each sales associate. We have, over a thousand products which are organized into three primary product categories, seventeen secondary product categories, and sixty tertiary product categories. And, those columns have been organized into a custom hierarchy named: Product Hierarchy. It is that information from the underlying data table which is useful for interpreting the treemap visualizations which have been created on the pages which follow.

On this page, a treemap has been configured which relates the Average Sale Price as a color gradient, and the sum of Profit values as a size gradient. There are currently only three rectangles, and both the total profit and average sales are highest in the Technology product category. Interestingly, though the Furniture product category represents the second highest average sales price, according to its pink color; however, the total profit derived from those sales is the lowest of the three categories, based upon the small size of the rectangle.

As indicated, the true value of the treemap is its ability to represent hierarchically-organized data. Therefore, let’s slide the hierarchy slider to the right in order to evaluate the second product category level. Notice that the three primary product categories are still represented by the same-sized rectangles which allowed us to compare total profit values at that level; however, the color-by variable, representing the average sale price, has shifted to the lowest level of the current hierarchy displayed.

In addition, the size of these nested rectangles reflects the total profit for products grouped at the secondary category level. You can see that while products assigned to the Copiers and Fax category represent our highest average sale price, the products assigned to Telephones and Communication represent our highest total profit values. Moving the hierarchy slider to the tertiary product category level creates a new ‘lowest-level’ color assignment, for comparison of sales price averages, and nests differentially-sized rectangles within rectangles within rectangles in order to allow a comparison of total profit values at different levels of the hierarchy.

Sliding to the final level of the hierarchy, where individual products define each rectangle, the treemap becomes increasingly crowded with nested rectangles. While this may be a valuable ‘high-level’ overview of our entire product line, you may also wish to zoom-in on areas of interest within the treemap. Each of the ‘parent’ rectangle labels is an active link which, when clicked upon, will zoom in to that level of the hierarchy. Once you are zoomed in to the area of interest, labels on rectangles, identifying the current ‘lowest-level’ of the hierarchy, become easier to read.

We are also presented with a header across the top which indicates our zoom path down to this subset of the hierarchy. We can click on any of these parent levels in order to zoom out to that level, or we can click on the “(All)” link in order to zoom out completely. So, a hierarchy slider, if presented, and the ability to zoom in and zoom out are the navigational tools which will allow you to interpret the information related in a treemap visualization.

Let’s proceed to the next tab, where another treemap has been configured to represent different information. The first thing you may notice about this treemap is that, different categorical variables have been assigned to the hierarchy property. If we pop up the legend, and widen the dialog in order to read each of the hierarchy selector labels, you can see the different columns of data which were added, and the order in which they were chosen to represent this hierarchy. However, because these were added as individual columns and not a custom hierarchy, there is no slider available to adjust the level of the hierarchy displayed. Labels for the lowest level of the hierarchy, Order ID, were turned off, due to the number and size of those rectangles. Note that, rectangles change size based upon the sum of sales, and the color-by variable represents an average of the discount amount applied to the order. The color by property has been configured using a segmented color mode and a number of points were added, in order to define color thresholds for each segment between the minimum and maximum at values of 2, 5, and 10%.

With that understanding of the treemap configuration provided in this high-level overview of our sales activities, if we want to examine the details for any of the areas in the treemap where a preponderance of red rectangles suggests that the sales associate in that region may have been over-applying discounts, hovering our mouse over a rectangle will present select values in the tooltip information box.

Future of Data Visualization

So far, we've seen a glimpse of Interactive Visualization Techniques. Now, take a look at this video on what future of Data Visualization looks like.

In popular conversation, discussions of data visualization often invoke examples such as - Custom Bespoke Graphics, often intended to convey a story and hand crafted by skilled visualization designers. Over my years working in visualization research, my students and I have sought to create tools that enable these kinds of sophisticated graphics. Giving rise to popular tools that we've developed such as Protovis, vega and d3. While we've been very happy with the success these tools have had, I think they're only a small part of the larger ecosystem of visualizations.

If you think more broadly, the vast majority of visualizations are not hand-coded, but rather built with end user tools. Often leading to visualizations that look like this, or applications that look like this. While well-intentioned, many of today's end user tools either have lacked consideration of perceptual principles or fall short of fully supporting the process of visual analysis and exploration. When I think about the future of data visualization and how we chart a path forward, one consideration to think about is - how do we move from tools that work well for designers to those that really enable analysis and decision makers to advance their causes, to enable better decision-making across industry and government.

For example, how can we advance the state of the art? I think, one way is to begin to bake more Design smart into our visualization tools. To give you a sense of what I mean, let's just conduct a quick experiment.

I'll show you some shapes. I want you to compare them. Don't yell out the answer, will take a quick poll on your response. So, here's two circles and I want you to compare their area. How much larger is the larger circle that the smaller circle? Now, raise your hand, if you think the big circle is 4 times bigger.

Take a look around…

Five? Six? Seven? Eight? Nine? Ten? - A lot of people Eleven or more? Those others?

Ok, now let's try another example. Compare the length of these two bars. How much larger is the big bar than the smaller? Raise your hand if you think it's 4 times bigger.

No one… Five? Six? Seven? Even more? - A lot of you. Nine? - Much fewer. Ten? Eleven or more? - Almost no one.

It turns out, the answer in both cases is the same. It’s seven times larger. But if you looked around the room, what you have noticed is that, despite the same difference in areas, bars were actually much more accurate and less variance overall for comparing length than area.

This was hardly a scientific poll. My group and others have conducted experiments like this in controlled settings, to compare the accuracy of different encodings. For example, you see that comparing things such as position, length, angle and area, and by combining the results of these experiments, we can actually build rankings of visual effectiveness for different encodings for different types of perceptual tasks.

When comparing quantities, position and length outperforms things like angle and area, which roundly outperform color encodings. This is useful to guide and not dictate what human designers do, but also very helpful for guiding algorithms which might automatically recommend effective and useful graphics.

At Trifacta, our product includes support for visual profiling of large data sets. Here we see a visualization of a geographic field showing where political contributions come from, using a common color encoding across states. Do you notice the problem?

Well, as we just discussed, color is one of the least accurate encoding channels for comparing quantitative data. There is even larger problem with this graphic, which is, our perception of value is correlated with the size and shape of the state's themselves. So, we can't even see what's happening in Washington DC. Which as it turns out as a source of many contributions.

So, what other visual encodings might we consider? For example, we'd like to maintain the spatial context of the map, so positions already spoken for, but we could move to something more accurate like area, to enable more effective comparisons. Thus, I also mentioned position is one of the most effective encodings.

We also augment this display with bar charts on the left, so we can make comparisons and then link these views together through interaction that's enabling more exploration, allowing us to see more patterns and make comparisons more accurately.

So, baking in better design smarts into our tools, I think is one way we move forward. Now, it's just the beginning. I think, more broadly it's time to rethink some of our basic user interfaces for data visualization. For example, most tools involve a process of specifying charts. How do we move this to a richer process of rapid exploration?

Common end user tools will have us select data subsets that we are interested in, and then choose from a set of chart types or specific visual encodings to manually build up a display. But this can be overly tedious. Also, seems like there's a lost opportunity to recommend interesting views.

For example, in the Trifacta visual profiler, we don't have people build charts explicitly. We provide them. For example, on the left, you see overviews where we automatically present summaries of all the dimensions within a data set, then users can drill down for more detail.

This panel on the right is showing when political contributions occur. This is shown both as an overall timeline, as well as summary displays for time periods such as, day of the week, day of the month, etc. By doing this, we can see a number of patterns. We see that, contributions occur with increasing regularity as we get near an election and they're also more likely on weekdays or at the end of a month. We learned something useful immediately without having to specify the charts. But probably the most important point here is that these displays were chosen automatically based on the data.

There's very little variation among things like hours, minutes and seconds. So we don't show those charts, to avoid distracting the user wasting their time. Simultaneously, in my interactive data lab at the University of Washington, we are exploring new end-user exploration tools. For example, our data Voyager system actually searches over a space of thousands of visualizations and ranks them according to statistical and perceptual measures to provide recommended visualizations for a data set.

For example, you can steer the recommendations and update display based on indicating fields of interest. Again trying to change the way we interact and explore with data to enable a broader and more rapid exploration. There are a lot of challenges that come up, while moving from specification to exploration, that I'm not covering. To do this, we will actually require the combined talents of the Strata community, drawing on visualization design, statistics and machine learning as well as Big data systems.

I'm very excited about the challenges that are presented to us. But I also think, we have to consider the ways in which these tools will really augment analysis in the most productive ways. To illustrate that, I'd like to share with you a dataset that I give to students in my visualization course. The classic data set, originally published in the 1950’s, comparing the effectiveness of antibiotics which were at that time - new Wonder drugs, against a variety of bacterial strains. I give this data set to the students and ask them to you explore the data, create a graphic that answers an interesting question.

So, here's some of the students submissions. Already you can see a great variety and designs that they explore. So, that's at first really interesting. But one thing that pops out as you start to dig in and you really inspect these graphics, is you realize that while they have lots of design variation, they're actually all addressing the exact same question – ‘Which antibiotics should one use?’. This fixation on this one question reminds me one of my favorite Maxim's from the visualization expert - Edward Tufte, who advises us to show data variation not just design variation. The idea is that, while it can be useful to explore multiple representations of the same subset of data, it is often even more vital to explore different slices and transformations of the data that might spur new questions or address different hypotheses. Rephrasing this, I think we should consider, how our tools might better spur us to exercise skepticism about data and consider new questions.

Now, consider this alternative visualization, it's fairly simple. What we're seeing is the different bacteria in a scatter plot. The bacteria are colored by genus and are plotted according to the effectiveness of two antibiotics - neomycin and penicillin. In this case, the lower left corner is an area of very high resistance and the upper right corner is a very low resistance. The idea here is that, just by suggesting this relatively simple visualization, we may still be prompted to consider other questions.

By putting the bacteria front and center, we we might ask instead, “What does antibiotic response reveal about the biology of bacteria?”. Flipping the question, we look back at this chart, you might have noticed something interesting. There are different clusters of bacteria, but they span genus and perhaps surprising ways like, wouldn't a bacteria of same family be more likely to group together?

Well, you'd be right to be skeptical here. Because, there's actually errors in the data. The scientific community was originally wrong and has two misclassifications in this data set. It actually took multiple decades after the original publication of this data for the scientific establishment to overturn these airs and yet the initial evidence was here all along, if we had thought to look at the data, through this particular lens.

It's interesting to think how our tools might prompt us to consider our data, more broadly. As we move from specification to exploration, we should also keep in mind - how this best enables analysis. For example, enabling data variation over design variation.

In conclusion, I'd like to look forward to a future in which our tools don't just help us build visualizations, but help us much more richly explore the data. In the end, hopefully leading to better insights and better decisions. So now I look forward to all of us building the future together. Thank you!

Data Visualization Tools

The first on the list is – Qliktech, which has Qlikview and Qliksense in its portfolio. The main selling point of Qliktech is, it has a very strong customer base, a very excellent cloud presence and a very strong scripting capabilities. It is a pioneer in data visualization and analytics area.

Second on our list is – Tableau, it's one of the leading data visualization and analytic tools. The main selling factor for Tableau is, it has a very flexible GUI and excellent mobile application capabilities and it's very easy to use. No coding experience is really required. It has a very strong integration with Alteryx and that's one of the features which is keeping Tableau ahead of a lot of other software’s. Apart from this, it also has a strong cloud presence.

Third on our list is - IBM, it has Cognos and they recently launched Watson Analytics. As you know, IBM Cognos and other reporting tools from IBM are pioneers, and they are well established for last 10 to 15 years and more. They have a very big customer and developer bases, so they are very prevalent in the market. Recently they have launched IBM Watson Analytics, which may turn out to be a game-changer. Don't know yet, yet to be proved.

Fourth on our list is - the dark horse of the 2014, which is really trying to realize its potential in this year, the last year and 2016 – is Alteryx. It's a relatively new player, but it's really growing fast. It has an excellent data mining, analytical and dashboarding capabilities. The biggest-selling factor for Alteryx at this point in time, is it has a very strong integration with Tableau.

Number five on our list is - Panorama software, this is very new software, like 3 or 4 years old. We consider it as the dark horse and it's also looking to expand its base really fast. It's a very decent data discovery tool and an awesome cloud capabilities. It also has good mobile capabilities and it's very easy to understand and learn. No coding experience necessary here too.

Having said that these are the tools in the market at this point in time, which are really ahead of the game. Qliktech which is Qlikview and Qliksens, Tableau, IBM Alteryx, SAS, Microstrategy, Tibco, Panorama, Fusion Charts and Maptive. These are the front runners, based on the analysis of the market we have conducted over past few months.

Quick recap – Qliktech, followed by Tableau, IBM Cognos, Alteryx and Panorama software. Thank you for watching!

Most widely used Data Visualization Tools

A few other powerful Data Visualization tools include D3.JS, R Charts (ggplot2 package), Pentaho, SAP Lumira, TIBCO Spotfire, QlikView, JasperSoft, and Microstrategy.