**DATA VISUALIZATION TECHNIQUES**

**(Professional Elective – III)**

**UNIT – I**

**INTRODUCTION AND DATA FOUNDATION**

* 1. **Introduction and Data Foundation**

Introduction and data foundation form the fundamental building blocks of any data-related project. In this context, introduction refers to the initial phase of a project where the objectives, scope, and purpose of the project are defined. Data foundation, on the other hand, involves the collection, organization, and preparation of data for analysis and interpretation.

The introduction stage is crucial as it sets the direction and goals for the project. It involves understanding the problem or question at hand, identifying the target audience or stakeholders, and defining the desired outcomes. This phase typically includes conducting a needs assessment, defining the project scope, and establishing success criteria. By clearly articulating the objectives and scope, the introduction phase ensures that the subsequent data-related activities align with the project's goals.

Data foundation focuses on the collection, organization, and preparation of data to facilitate analysis. It involves identifying relevant data sources, acquiring or collecting the necessary data, and transforming it into a usable format. This process may include data cleaning, data integration, data formatting, and data quality assurance. The goal is to ensure that the data is accurate, complete, consistent, and properly structured for subsequent analysis.

Additionally, data foundation often involves establishing a data governance framework, which includes defining data ownership, access controls, and data security protocols. This ensures that data is managed responsibly, adhering to privacy regulations and safeguarding sensitive information.

A solid data foundation is essential for accurate and reliable analysis. It provides a strong basis for making informed decisions, deriving insights, and identifying trends or patterns in the data. Properly collected, organized, and prepared data sets the stage for various analytical techniques and visualization methods to be applied effectively.

* 1. **Relationship between Visualization and Other Fields**
* **What is Data Visualization?**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Something as simple as presenting data in graphic format may seem to have no downsides. But sometimes data can be misrepresented or misinterpreted when placed in the wrong style of data visualization. When choosing to create data visualization, it’s best to keep both the advantages and disadvantages in mind.

**Advantages**

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

Some other advantages of data visualization include:

* Easily sharing information.
* Interactively explore opportunities.
* Visualize patterns and relationships.

**Disadvantages**

While there are many advantages, some of the disadvantages may seem less obvious. For example, when viewing visualization with many different data points, it’s easy to make an inaccurate assumption. Or sometimes the visualization is just designed wrong so that it’s biased or confusing.

Some other disadvantages include:

* Biased or inaccurate information.
* Correlation doesn’t always mean causation.
* Core messages can get lost in translation.
* **Why data visualization is important?**

The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise.

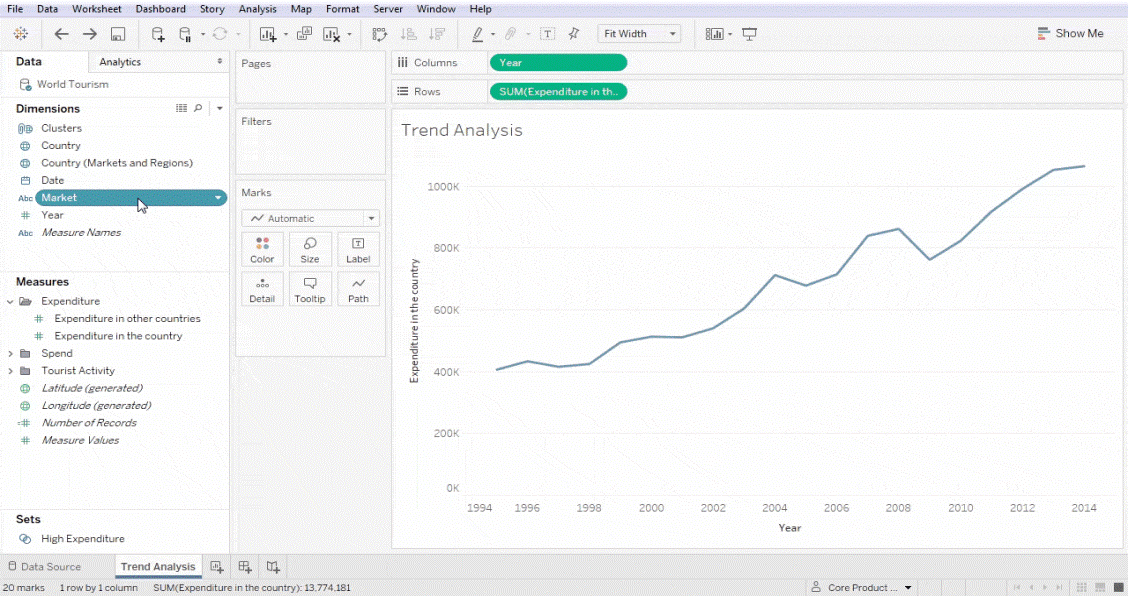
It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every STEM field benefits from understanding data and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

While we’ll always wax poetically about data visualization (you’re on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it’s also one of the most useful professional skills to develop. The better you can convey your points visually; whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how.

While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

**Different types of visualizations**

When you think of data visualization, your first thought probably immediately goes to simple bar graphs or pie charts. While these may be an integral part of visualizing data and a common baseline for many data graphics, the right visualization must be paired with the right set of information. Simple graphs are only the tip of the iceberg. There’s a whole selection of visualization methods to present data in effective and interesting ways.



**Types of Visualizations:**

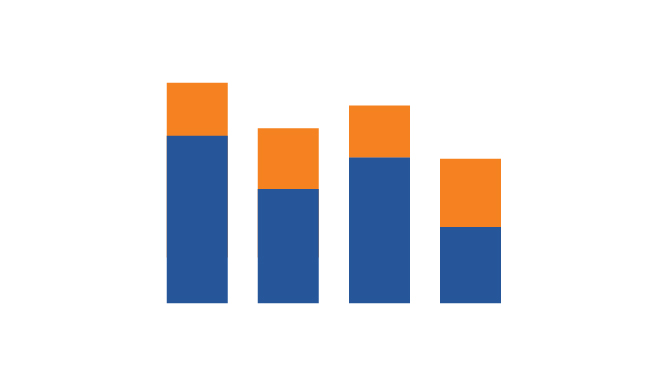
* **Chart:** Information presented in a tabular, graphical form with data displayed along two axes. Can be in the form of a graph, diagram, or map.
* **Table:** A set of figures displayed in rows and columns.
* **Graph:** A diagram of points, lines, segments, curves, or areas that represents certain variables in comparison to each other, usually along two axes at a right angle.
* **Geospatial:** A visualization that shows data in map form using different shapes and colors to show the relationship between pieces of data and specific locations. Learn more.
* **Infographic:** A combination of visuals and words that represent data. Usually uses charts or diagrams.
* **Dashboards:** A collection of visualizations and data displayed in one place to help with analyzing and presenting data.

**More specific examples**

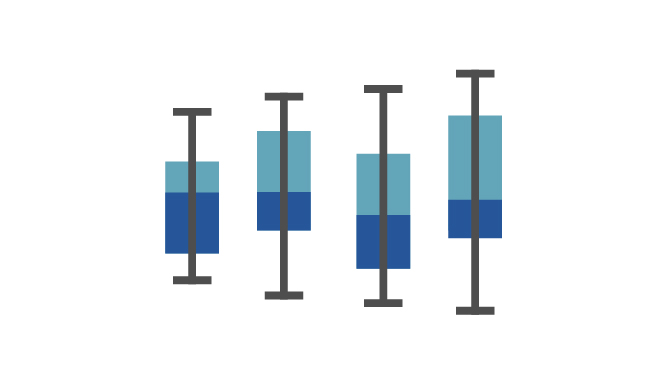
* **Area Map:** A form of geospatial visualization, area maps are used to show specific values set over a map of a country, state, county, or any other geographic location. Two common types of area maps are choropleths and isopleths.



* **Bar Chart:** Bar charts represent numerical values compared to each other. The length of the bar represents the value of each variable.



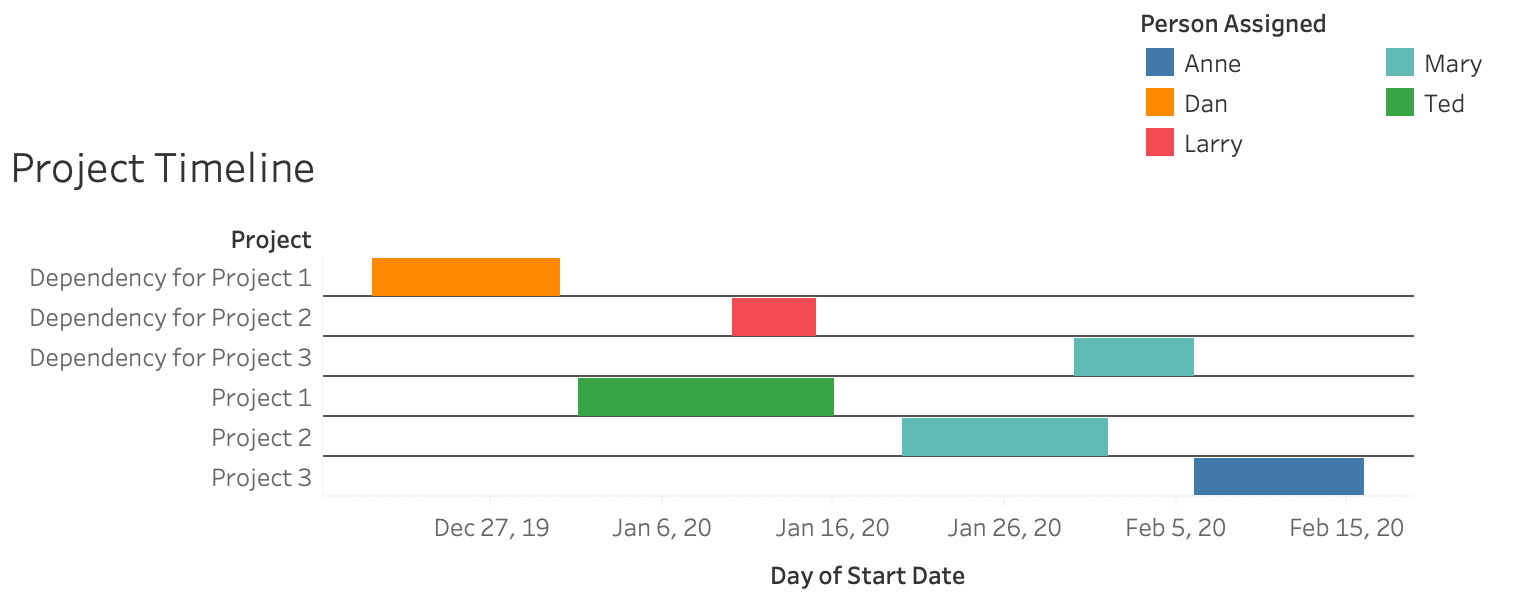
* **Box-and-whisker Plots:** These show a selection of ranges (the box) across a set measure (the bar).



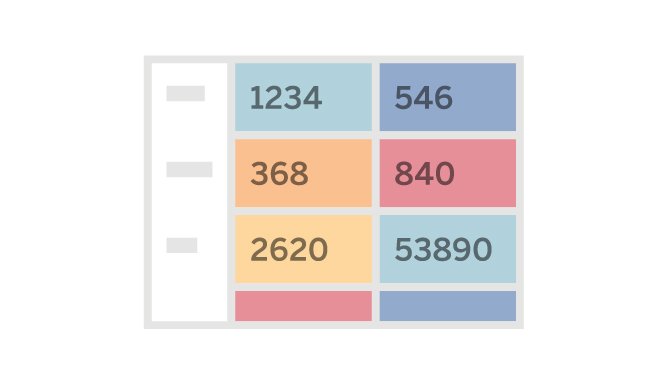
* **Bullet Graph:** A bar marked against a background to show progress or performance against a goal, denoted by a line on the graph.



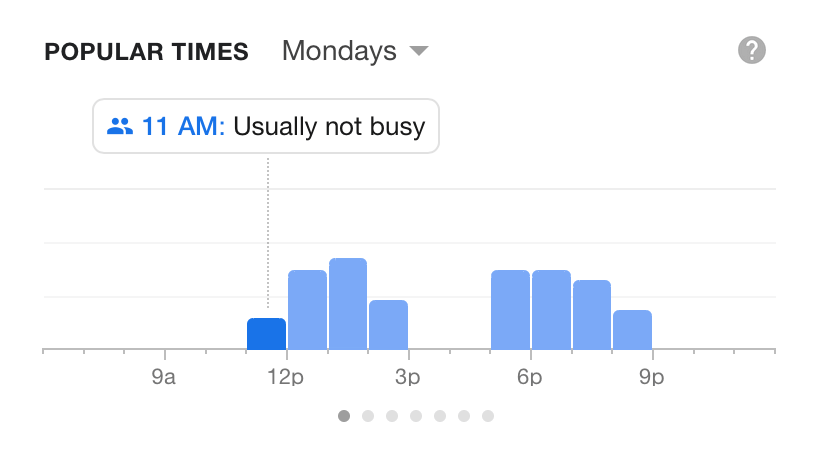
* **Gantt Chart:** Typically used in project management, Gantt charts are a bar chart depiction of timelines and tasks.



* **Heat Map:** A type of geospatial visualization in map form which displays specific data values as different colors (this doesn’t need to be temperatures, but that is a common use).



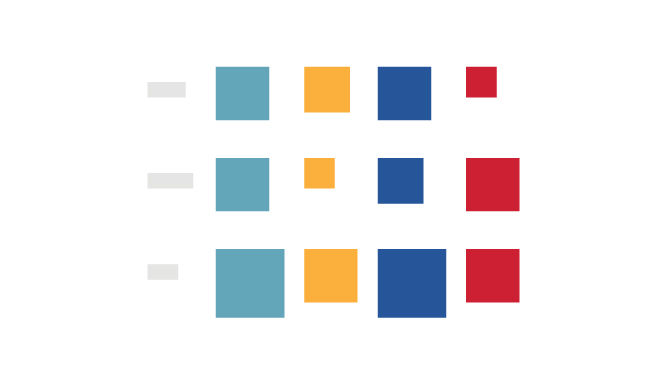
* **Histogram:** A type of bar chart that split a continuous measure into different bins to help analyze the distribution.



* **Pie Chart:** A circular chart with triangular segments that shows data as a percentage of a whole.



* **Treemap:** A type of chart that shows different, related values in the form of rectangles nested together.



* 1. **Visualization Process**

The visualized process comprises a number of nodes on a graph that represents the activities in the process. The nodes are linked by graph edges and these reflect the transitions that are defined in the process definition.

People around the world have been using various kinds of visualization techniques, meditations, and prayers for centuries. However, visualization often gets a bad reputation as being something mystical or woo-woo that isn’t grounded in reality.

But the truth is, you don’t need to be spiritual to benefit from visualization techniques. Visualization has been studied by psychologists to understand how it works.

And nowadays, everybody, from professional athletes to CEOs, are benefiting from visualization techniques.

Let’s take a look at what visualization is and why it’s important. Then we’ll outline some visualization techniques that you can start using yourself.

* **What is visualization?**

Visualization is the practice of imagining what you want to achieve in the future. As if it were true today.

It involves using all five senses of sight, smell, touch, taste, and hearing. The process of visualizing directs your subconscious to be aware of the end goal you have in mind.

It reminds you on a consistent basis. And it trains your brain to respond as if that outcome were true in the present moment.

There are two types of visualization:

(1) Outcome visualization: envisioning the desired future end-point, and

(2) Process visualization: envisioning every step toward that desired outcome. With all the senses engaged.

When the two visualizations are used in tandem, they yield the best results.

* **Why is visualizing important?**

We live in a world of digital overload. There’s a constant cycle of news and entertainment. Plus, remote work is available with the click of a button. It’s easy to become distracted.

It's easy to believe that there’s not enough time in the day. You just want time to step back and focus on the things that matter to you the most.

In his presentation Unleash Your Super Brain to Learn Faster, Jim Kwik offers a great set of metaphors.

You can choose to be either a thermostat or a thermometer. The thermostat responds to the environment. The thermometer sets the temperature and creates the desired condition.

Visualization is a technique that allows you to set the parameters to make your future vision a reality.

In creative visualization, you direct your brain to focus on what matters the most to you. And to engage in a process called selective attention.

Have you ever bought a car and then noticed that everyone else seems to be driving that same car? We see the things that we choose to focus on. This concept of selective attention is exemplified in the classic video Test Your Awareness: Do the Test.

What you focus on and take deliberate steps toward are more likely to become a reality.

That’s true in both the positive and the negative.

Have you ever imagined the worst possible outcome? Along with all the visceral fear, and then it actually came true?

What if you changed that to the best possible imagined outcome. With the related elation and full sensory awareness?

For the answer, we can look to cognitive behavior theory. Visualization is built on the foundation it teaches that thought precedes action.

Visualization is a technique that takes that idea one step further. In imagining the future state in full sensory detail, your brain’s neuroimagery records the future state as if it's true today.

* **What are visualization techniques for?**

Visualization can be used to motivate you to focus on, and work toward, your future ideal state.

It can be used to build your self-confidence.

Imagine yourself presenting in front of a large audience and hearing the resounding applause. You’re likely to believe, and act in a way, to make it happen.

Visualization can also be used as a form of mental rehearsal. Through process visualization, you increase selective attention. And engage the mental pathway that helps you to fine-tune the movement before you even step onto the stage.

Your movements are envisioned. And, in a way, programmed to take the pre-planned actions. You can add these actions to a longer-term plan 5-year plan.

Visualization techniques can also reduce anxiety. They direct your attention back to the details of your desired outcome. In the process, minimizing the noise of other (less relevant) distractions.

* **How do you visualize what you want?**

First, you have to do the pre-work.

An excellent place to start is the Balance Wheel exercise.

Check in on your satisfaction levels with the various areas of life.

Is your life wheel balanced, or does an area need attention?

Next, become clear of the values that act as your North Star.

Do you value family, career, and money? Or travel, adventure, and freedom?

How do those values impact the decisions you make?

Finally, take some time to do some positive thinking. Use your imagination through reflective writing prompts in the Best Possible Self exercise.

Who'd you want to be when you were little? Who are you today? Is it the same or a different version?

Also, consider when you were energized and felt most engaged and alive. What was it about that time that invigorated you? What mattered about that moment?

Take a piece of that moment and imagine the future state you want. Try to get to that same feeling of excitement and emotional intensity.

Imagine where you are and what you are doing. Who's around you? How do you feel? What smells are in the air, and what tastes are on your tongue? What are your thoughts at that moment? If you believe your thoughts, it's possible. If you don’t, it’s not.

With an understanding of what matters to you most, and why, you're ready to begin the practice of visualization.

* **Visualization techniques and tools**: learn how to practice visualization

Learning visualization can be tricky. Here are five tools and techniques you can use to learn how to practice visualization successfully:

1. Create a vision board

This visualization tool is a collection of images and words that inspire you and represent your goals. It serves as a visual representation of what you want to achieve.

Place your board somewhere you'll see every day. That way, you'll be constantly reminded of what you're working towards.

2. Listen to a guided visualization meditation

Youtube is full of free guided meditation videos. An interactive visualization can help you to relax and set some time aside to focus on your goals. Guided imagery helps give you something to focus on.

3. Use index cards

Do you remember using flashcards as a kid? Maybe you used them to learn math or words. As an adult, we can use index cards in a similar way as a tool for visualization.

Make a list of 10 or 20 goals that you're currently working toward. Write each one on an index card and keep it near your bed.

Every morning and every night, go through the stack of cards. One at a time. Read each card, then close your eyes and visualize yourself completing that goal.

4. Picture and describe

The more detail you use when visualizing, the more real it'll feel to you.

It's one thing to get a quick thought that you want a slice of pizza. It's entirely different to really allow yourself to obsess about it.

Think how much more real it seems if you take a few minutes to imagine how it'd taste, feel, and smell to eat a slice of pizza right now.

Create as specific of a mental image as possible. Try to provide your brain with as many details about your goal as possible. More detail means a better solution.

5. Utilize exposure

Exposing yourself to things related to your goals can help make it more real for you.

Maybe one of your goals is to go parachuting. If you've never done it before, it's hard to imagine exactly what it's like.

Watch some videos of other people parachuting. Read accounts people have written about parachuting online. Or talk to people who have tried it. All of these things will increase your knowledge about it and make it feel more real for you.

Try out a visualization technique from the list above. Your selective attention will be directed toward fulfilling your vision.

You'll be the thermostat creating the condition that's just right for you.

* **Use visualization to achieve your goals**

Visualization can help you make better decisions and get what you want out of life.

It’s hard to make progress toward your goals if you don’t take some time to think about them.

Set some time aside to use the techniques we’ve shared. They'll provide you with actionable insights into achieving your goals.

BetterUp can help you find your focus and purpose. Discover your strengths and achieve your life goals with a personalized BetterUp coaching experience.

* 1. **Pseudo code Conventions**

Pseudocode often uses structural conventions of a normal programming language, but is intended for human reading rather than machine reading. It typically omits details that are essential for machine understanding of the algorithm, such as variable declarations and language-specific code.

This document explains the conventions and common functions used in some of my articles that use pseudocode.

**Symbols**

In addition to the familiar +, -, \* (multiplication), and / (division) operators, other symbols are defined below.

pi is the constant π, the ratio of a circle's circumference to its diameter.

nothing indicates the absence of a value. It corresponds to null in Java, C#, and JavaScript, nil in Ruby, and None in Python.

true and false are the two Boolean values.

== means "is equal to".

!= means "is not equal to".

A minus before a variable means 0 minus that variable. For example, -a means (0 - a).

The << operator in the pseudocode is a bitwise left shift, with both sides of the operator being integers. If each side is 0 or greater, it is the same as multiplying the left-hand side by 2n, where n is the right-hand side.

The >> operator in the pseudocode is a bitwise right shift, with both sides of the operator being integers. If each side is 0 or greater, it is the same as dividing the left-hand side by 2n, where n is the right-hand side, and discarding the fractional part of the result.

The | operator in the pseudocode is a bitwise OR operator between two integers. It combines the bits of both integers so that each bit is set in the result if the corresponding bit is set on either or both sides of the operator.

**Loops**

Pseudocode may use while loops, which are self-explanatory.

Pseudocode may also use for loops, defined as follows:

for X in Y...Z; [[Statements]] ; end is shorthand for X = Y; while X < Z; [[Statements]]; X = X + 1; end.

for X in Y...Z: [[Single-Statement]] is shorthand for X = Y; while X < Z; [[Single-Statement]]; X = X + 1; end.

**List and Files**

In the pseudocode, lists are indexed starting with 0. That means the first item in the list has index 0, the second item in the list has index 1, and so on, up to the last item, whose index is the list's size minus 1.

In this context, a list is to be understood as a resizable array of items, not as a linked list.

A list can be expressed by wrapping items in brackets; for example, [0, 1, 2] is a three-item list.

NewList() or [] creates a new empty list.

AddItem(list, item) adds the item item to the list list.

size(list) returns the size of the list list.

list[k] refers to the item at index k of the list list.

GetNextLine(file) is a method that gets the next line from a file, or returns nothing if the end of the file was reached.

**Functions**

* sqrt(a) is the square root of a, and is equivalent to pow(a, 0.5).
* ln(a) is the natural logarithm of a.
* exp(a) is the inverse natural logarithm of a. Also known as the base of natural logarithms raised to the power a, so that exp(1) is the base of natural logarithms commonly denoted e.
* pow(a, b) is the number a raised to the power b.
* sin(a), cos(a), and tan(a) are the sine, cosine, and tangent of the angle a, respectively, where a is in radians.
* atan2(y, x) is—
* the inverse tangent of y/x, in radians, if x > 0,
* π plus the inverse tangent of y/x, in radians, if y >= 0 and x < 0,
* −π plus the inverse tangent of y/x, in radians, if y < 0 and x < 0,
* −π divided by 2 if y < 0 and x == 0,
* π divided by 2 if y > 0 and x == 0, and
* 0 if y == 0 and x == 0.
* abs(a) is the absolute value of a; it makes negative numbers nonnegative.
* min(a, b) is the smaller of a and b.
* max(a, b) is the larger of a and b.
* floor(a) is the highest integer that is less than or equal to a.
* rem(a, b) is the part of b that does not divide evenly into a, where the result has the sign of b. This operation is equivalent to a - floor(a / b) \* b.
* ceil(a) is the lowest integer that is greater than or equal to a. This function is equivalent to (0 - floor(0 - a)).

**Notes:**

* The inverse sine, in radians, of a is equivalent to atan2(a, sqrt(1.0 - a \* a)).
* The inverse cosine, in radians, of a is equivalent to atan2(sqrt(1.0 - a \* a), a).
* An integer n is odd if rem(n, 2) is 1.
* An integer n is even if rem(n, 2) is 0.

**Pseudocode Notes**

In the pseudocode:

* Divisions do not round to an integer. (In some programming languages, division of two integers results in an integer, which may be rounded differently depending on the language. For instance, Python's and Ruby's integer division does a floor rounding on the result of division, while Java's discards the fractional part of the result of division.)
* The pseudocode shown is not guaranteed to cover all error handling, such as recovery from overflows, out-of-bounds memory accesses, divisions by zero, unexpected infinity values, and other errors that might happen in a particular implementation.
* The pseudocode shown is not guaranteed to yield high performance in a particular implementation, either in time or memory.
* In general, computer implementations of the operators and functions above risk numerical errors, since computers generally can't operate "exactly" on real numbers. (This is less of an issue if the implementation uses an arbitrary-precision rational number format and the pseudocode uses only rational arithmetic and inputs. In addition, an implementation can work with symbolic representations of real numbers instead of those numbers.)
  1. **Scatter plot**

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram)[3] is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

**Overview**

A scatter plot can be used either when one continuous variable is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval. For example, weight and height would be on the y-axis, and height would be on the x-axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the dots' pattern slopes from lower left to upper right, it indicates a positive correlation between the variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation. A line of best fit (alternatively called 'trendline') can be drawn to study the relationship between the variables. An equation for the correlation between the variables can be determined by established best-fit procedures. For a linear correlation, the best-fit procedure is known as linear regression and is guaranteed to generate a correct solution in a finite time. No universal best-fit procedure is guaranteed to generate a correct solution for arbitrary relationships. A scatter plot is also very useful when we wish to see how two comparable data sets agree to show nonlinear relationships between variables. The ability to do this can be enhanced by adding a smooth line such as LOESS.[5] Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

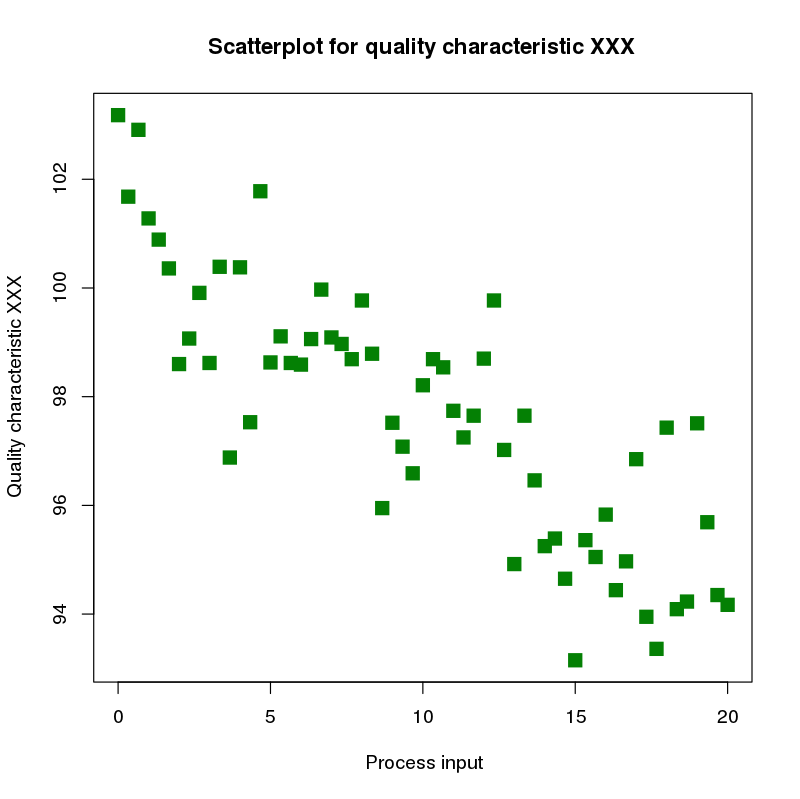


Figure: Scattor Plot

**1.6 Data Foundation:**

Data is defined as facts or figures, or information that's stored in or used by a computer. An

example of data is information collected for a research paper. An example of data is an email.

**Types of Data**

Understanding the different types of data (in statistics, marketing research, or data science) allows you to pick the data type that most closely matches your needs and goals. Whether you are a businessman, marketer, data scientist, or another professional who works with some kinds of data, you should be familiar with the key list of data types. Why? Because the various data classifications allow you to correctly use measurements and thus to correctly make decisions.

Data science is all about experimenting with raw or structured data. Data is the fuel that can drive a business to the right path or at least provide actionable insights that can help strategize current campaigns, easily organize the launch of new products, or try out different experiments.

All these things have one common driving component and this is Data. We are entering into the digital era where we produce a lot of Data. For instance, a company like Flipkart produces more than 2TB of data on daily basis.

In simple terms, data is a systematic record of digital information retrieved from digital interactions as facts and figures. Types of statistical data work as an insight for future predictions and improving pre-existing services. The continuous data flow has helped millions of organizations to attain growth with fact-backed decisions. Data is a vast record of information segmented into various categories to acquire different types, quality, and characteristics of data, and these categories are called data types.

When this Data has so much importance in our life then it becomes important to properly store and process this without any error. When dealing with datasets, the category of data plays an important role to determine which preprocessing strategy would work for a particular set to get the right results or which type of statistical analysis should be applied for the best results. Let’s dive into some of the commonly used categories of data.

**Qualitative Data Type**

Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes. It means that this type of data can’t be counted or measured easily using numbers and therefore divided into categories. The gender of a person (male, female, or others) is a good example of this data type.

These are usually extracted from audio, images, or text medium. Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on. All this information can be categorized as Qualitative data. There are two subcategories under this:

Must read: Data structures and algorithms free course!

**Nominal**

These are the set of values that don’t possess a natural ordering. Let’s understand this with some examples. The color of a smartphone can be considered as a nominal data type as we can’t compare one color with others.

It is not possible to state that ‘Red’ is greater than ‘Blue’. The gender of a person is another one where we can’t differentiate between male, female, or others. Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

Nominal data types in statistics are not quantifiable and cannot be measured through numerical units. Nominal types of statistical data are valuable while conducting qualitative research as it extends freedom of opinion to subjects.

Read: Career in Data Science

**Ordinal**

These types of values have a natural ordering while maintaining their class of values. If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large. The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.

These categories help us deciding which encoding strategy can be applied to which type of data. Data encoding for Qualitative data is important because machine learning models can’t handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.

For nominal data type where there is no comparison among the categories, one-hot encoding can be applied which is similar to binary coding considering there are in less number and for the ordinal data type, label encoding can be applied which is a form of integer encoding.

**Quantitative Data Type**

This data type tries to quantify things and it does by considering numerical values that make it countable in nature. The price of a smart phone, discount offered, number of ratings on a product, the frequency of processor of a smart phone, or ram of that particular phone, all these things fall under the category of Quantitative data types.

Also read: Learn python online free!

The key thing is that there can be an infinite number of values a feature can take. For instance, the price of a smart phone can vary from x amount to any value and it can be further broken down based on fractional values. The two subcategories which describe them clearly are:

**Discrete**

The numerical values which fall under are integers or whole numbers are placed under this category. The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.

Discrete data types in statistics cannot be measured – it can only be counted as the objects included in discrete data have a fixed value. The value can be represented in decimal, but it has to be whole. Discrete data is often identified through charts, including bar charts, pie charts, and tally charts.

**Continuous**

The fractional numbers are considered as continuous values. These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores.

Unlike discrete data types of data in research, with a whole and fixed value, continuous data can break down into smaller pieces and can take any value. For example, volatile values such as temperature and the weight of a human can be included in the continuous value. Continuous types of statistical data are represented using a graph that easily reflects value fluctuation by the highs and lows of the line through a certain period of time.

**Importance of Qualitative and Quantitative Data**

Qualitative types of data in research work around the characteristics of the retrieved information and helps understand customer behavior. This type of data in statistics helps run market analysis through genuine figures and create value out of service by implementing useful information. Qualitative types of data in statistics can drastically affect customer satisfaction if applied smartly.

On the other hand, the Quantitative data types of statistical data work with numerical values that can be measured, answering questions such as ‘how much’, ‘how many’, or ‘how many times’. Quantitative data types in statistics contain a precise numerical value. Therefore, they can help organizations use these figures to gauge improved and faulty figures and predict future trends.

* 1. **Data Preprocessing**

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and missing values, etc.

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Examples of data preprocessing include cleaning, instance selection, normalization, one hot encoding, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

Data preprocessing may affect the way in which outcomes of the final data processing can be interpreted. This aspect should be carefully considered when interpretation of the results is a key point, such in the multivariate processing of chemical data (chemometrics).

**Tasks of data preprocessing**

**Data cleansing:** Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting or a data quality firewall.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code), or with fuzzy or approximate string matching (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross-checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information. For example, appending addresses with any phone numbers related to that address. Data cleansing may also involve harmonization (or normalization) of data, which is the process of bringing together data of "varying file formats, naming conventions, and columns", and transforming it into one cohesive data set; a simple example is the expansion of abbreviations ("st, rd, etc." to "street, road, etcetera").

**Data editing:** Data editing is defined as the process involving the review and adjustment of collected survey data. Data editing helps define guidelines that will reduce potential bias and ensure consistent estimates leading to a clear analysis of the data set by correct inconsistent data using the methods later in this article. The purpose is to control the quality of the collected data. Data editing can be performed manually, with the assistance of a computer or a combination of both.

**Data reduction:** Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The purpose of data reduction can be two-fold: reduce the number of data records by eliminating invalid data or produce summary data and statistics at different aggregation levels for various applications. Data reduction does not necessarily mean loss of information. For example, the body mass index reduces two dimensions (body and mass) into a single measure, without any information being lost in the process.

When information is derived from instrument readings there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, encoding, sorting, collating, and producing tabular summaries. When the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are often needed. The data reduction is often undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined.

An example in astronomy is the data reduction in the Kepler satellite. This satellite records 95-megapixel images once every six seconds, generating dozens of megabytes of data per second, which is orders-of-magnitudes more than the downlink bandwidth of 550 kB/s. The on-board data reduction encompasses co-adding the raw frames for thirty minutes, reducing the bandwidth by a factor of 300. Furthermore, interesting targets are pre-selected and only the relevant pixels are processed, which is 6% of the total. This reduced data is then sent to Earth where it is processed further.

Research has also been carried out on the use of data reduction in wearable (wireless) devices for health monitoring and diagnosis applications. For example, in the context of epilepsy diagnosis, data reduction has been used to increase the battery lifetime of a wearable EEG device by selecting and only transmitting EEG data that is relevant for diagnosis and discarding background activity.

**Data wrangling:** Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.

The process of data wrangling may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use. It is closely aligned with the ETL process.

**Data Set:**

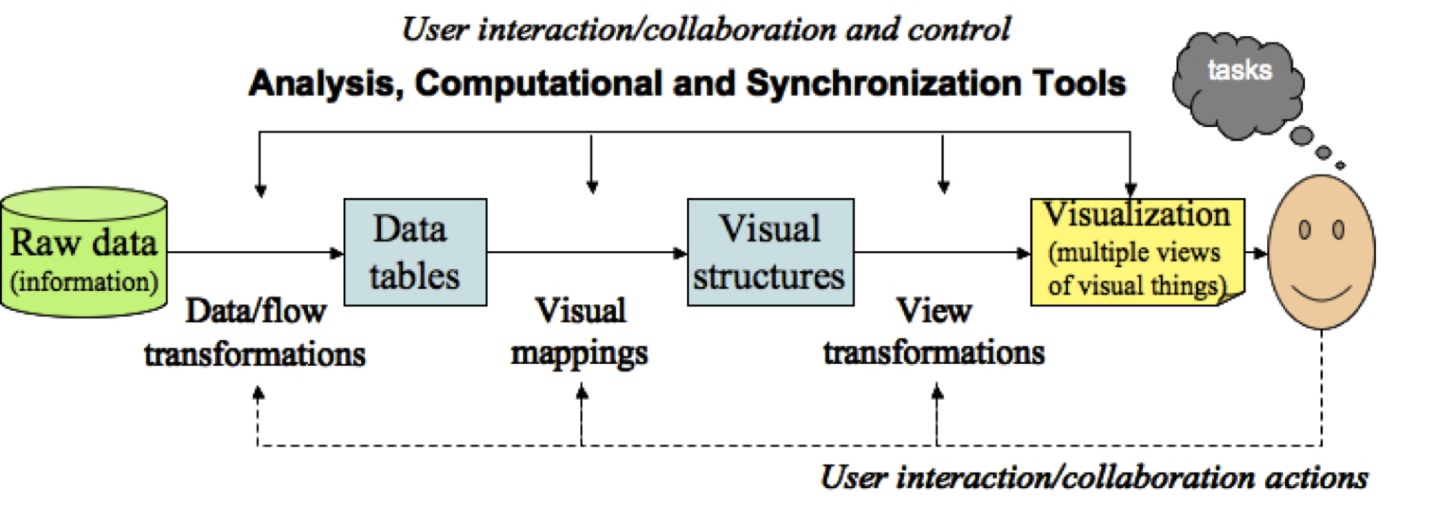
A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files.

In the open data discipline, data set is the unit to measure the information released in a public open data repository. The European data.europa.eu portal aggregates more than a million data sets

**UNIT –II**

**FOUNDATIONS FOR VISUALIZATION**

**2.1** **Foundations for Visualization: Visualization stages**

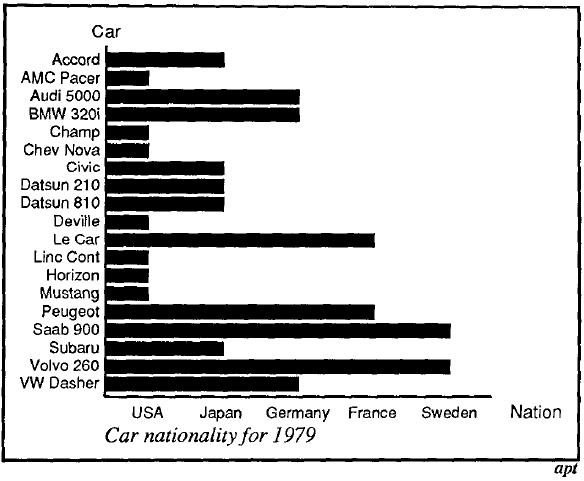
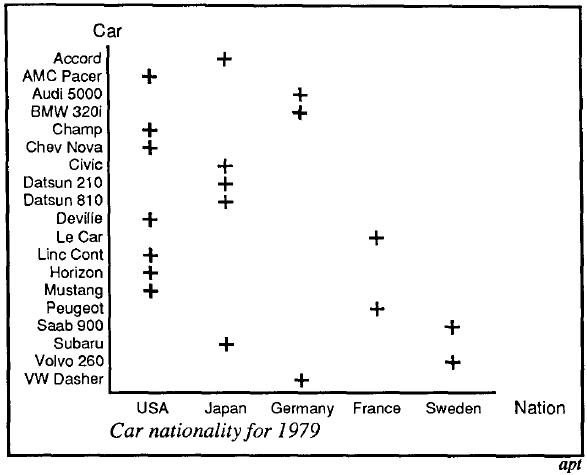


### 1. Data preprocessing and transformation (review)

* map data to data types and attribute names
* deal with missing values (e.g., interpolate with average, remove records)
* deal with errors in input (similar considerations as missing values)
* deal with too large a data set (sample, filter, aggregate or partition)

### 2. Mappings for visualization

* choose most appropriate visualization from options within your toolset , e.g. scatter plot versus bar chart (depends on the data type of the axes).
* **why is right hand chart better than left below?**

### 3. Rendering tools.

See table of tools in text and syllabus.

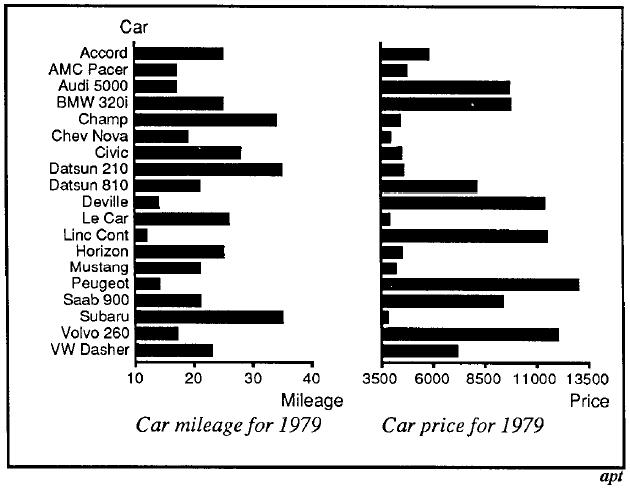
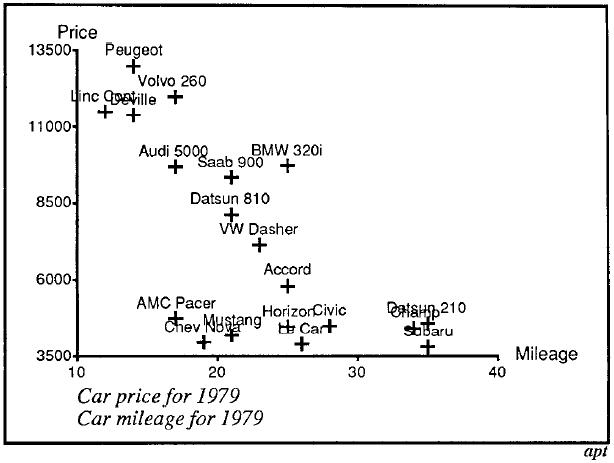
We want ***expressiveness***, ***effectiveness***, ***accuracy*** and ***maximization*** of data displayed.

**Expressiveness** is presenting the information in the data and no more

* measured as ***Expressiveness*** = "info displayed" / "info want to present"
* GOAL: ideally***Expressiveness***= 1
* if <1, we are not displaying enough
* if >1, then TMI--too much information! which may be wrongly interpreted or interferes with what we want to present

**Effectiveness** is interpreting accurately and quickly with cost-effective rendering (efficiency)

* measured as ***Effectiveness*** = 1 / (1+ costToInterpret + costToRender)
* a value between 0 and 1
* GOAL: large, closer to one is better, that is, costs to interpret and render are near zero



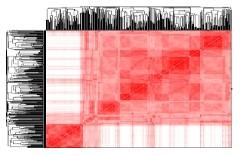
## Semiology of graphical symbols

Semiology = study of signage

A symbol must be easily recognized. E.g, a red octagon is universally understood as a stop sign.

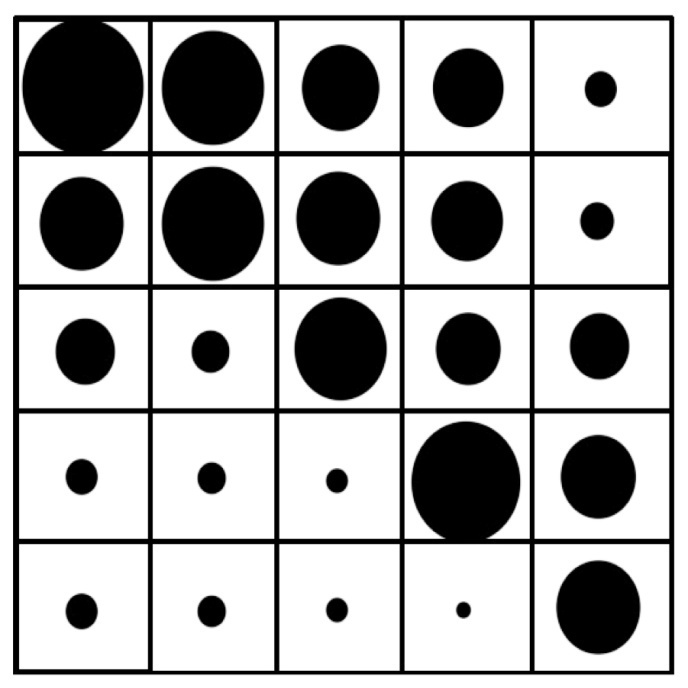


If you use a symbol that requires a key-to-meaning mapping or reference, interpretation will be slowed and therefore efficiency is reduced.



Symbol positions should also be easily understood: requiring that axes of a graph are easily understood.

Graphics can have three or more dimensions. E.g.



## Eight Visual Variables/Features

The symbols or marks are considered as part of the visualization primitives.

Bertin, presented earlier, identified 6 of the 8, omitting brightness and motion

### 1. Positionposition and mark

* on a one-dimension line, two-dimension x-y graph, or a three-dimensional rendering

**Scaling**

* linear or logarithm scale
* scaling refers to adjustment of values to fit the graphic,
* ensure the units between tick marks are powers of 10 multiples of 1, 2 or 5.

**Projection**

* maps need to choose from different projections
* 3D to 2D rendering is a projection too

### 2. Mark

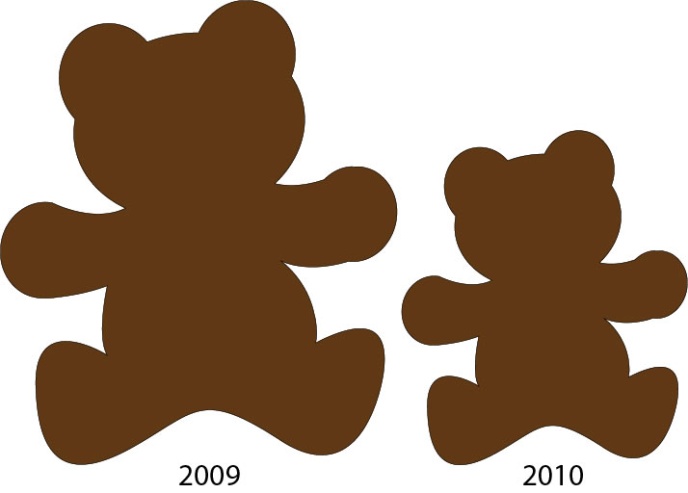
* point, shape, or "glyph"
* use for categorization or position
* see Horsepower X Highway MPG to right

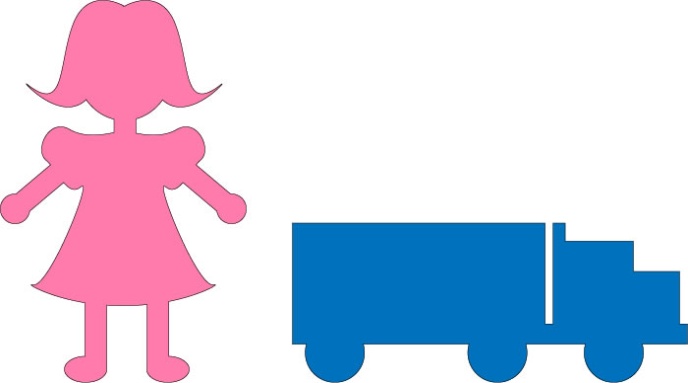
### 

### 3. Sizeareas hard to distinguish

* length, area, volume
* careful with combination of shape as areas and volumes become hard to distinguish
* area and volume can be used to misrepresent data

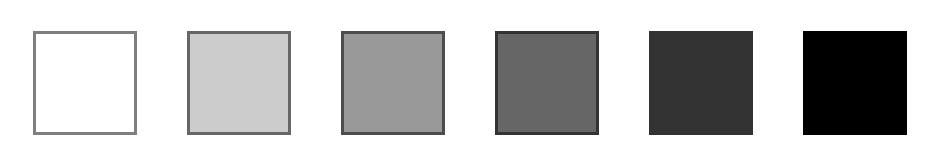
The reason consistent glyphs are used--so we can validly compare size





### 4. Brightness or luminance

* good for large interval and continuous data
* limited ability to distinguish among a lot of levels of brightness unless adjacent



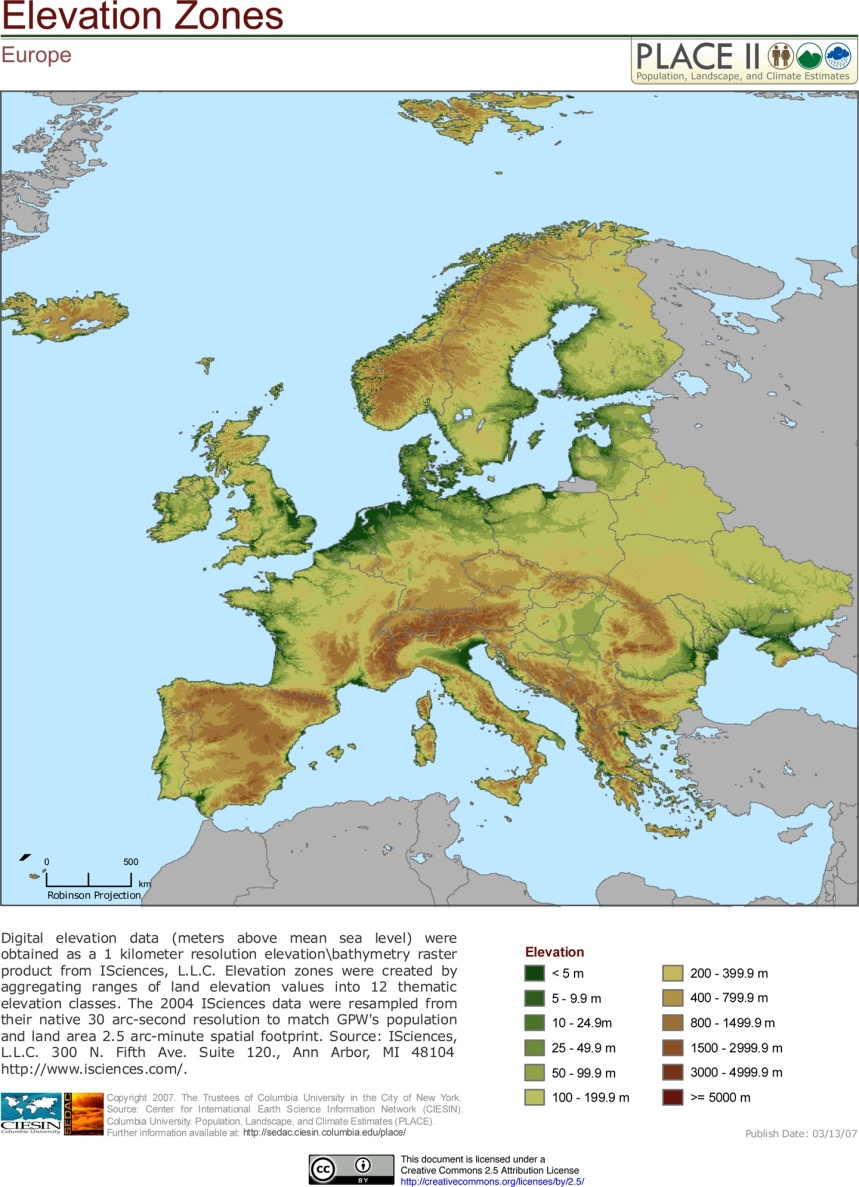
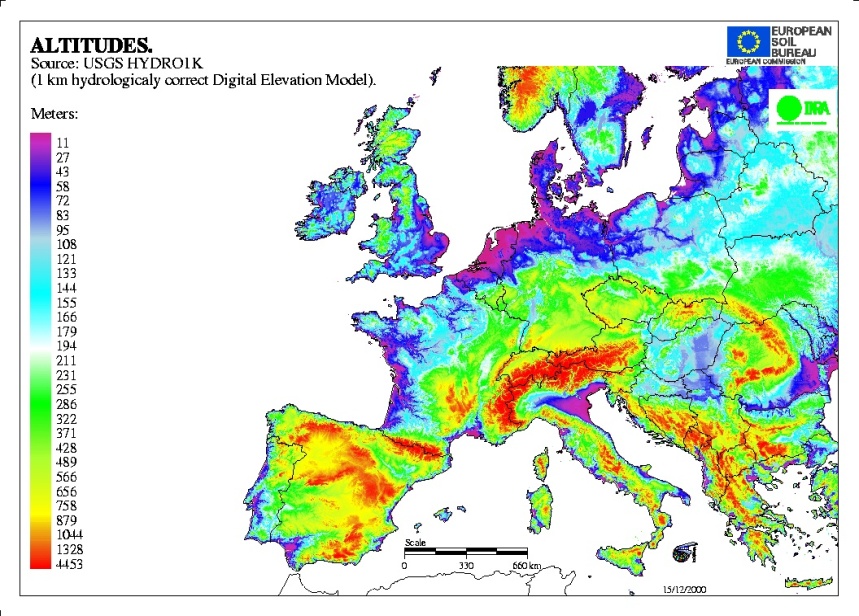
### 5. Color

* hue is the color from the spectrum
* saturation is the hue relative to gray

Color maps can range between two colors through gray as the middle, or black to white with a color in the middle.

color map

https://jcsites.juniata.edu/faculty/rhodes/ida/images/Figure_4_17c.jpg



### 6. Orientation

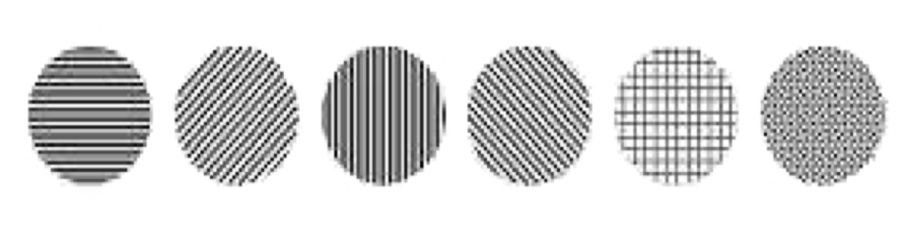
* The mark has to have a clear orientation over the 360 degree rotation or just 90 degrees depending on the mark.

orientation

### 7. Texture

Texture can provide visual dimensions of varying degrees:

* regularity
* directionality
* contrast
* size
* coarseness
* limited to a few values before it becomes interpreted as a form of brightness or orientation



### 8. Motion

Or animation. Objects ("glyphs" or "boids") can be distinguished through preattentive perception by

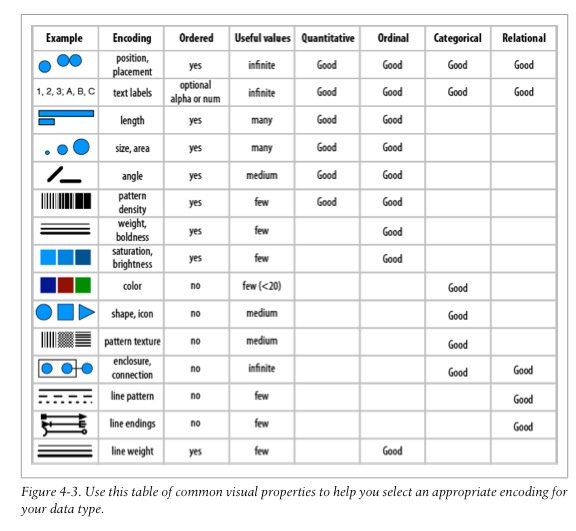
* flicker
* direction
* speed

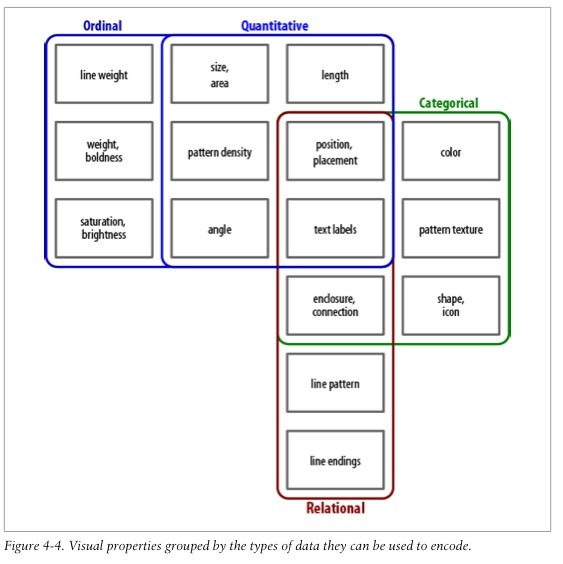
Motion can be applied to the above 7 variables. Especially useful if the visual is interactive, or active.

* position changes
* size changes
* brightness changes
* orientation changes
* color
* texture and shape limited
* rates of change

## Feature hierarchy

Because of the preattentive perception, we can use these different features to represent multidimensional (hypervariate) visualization. Preattentive perception is in play here as well as conveying additional information tasks (association, selection, order, quantity)





**2.2 Experimental Semiotics based on Perception**

Experimental semiotics based on perception is an interdisciplinary field that combines elements of linguistics, semiotics, psychology, and cognitive science to study how perception influences the interpretation and understanding of signs and symbols.

Traditional semiotics is the study of signs and their meanings, encompassing both linguistic signs (e.g., words and sentences) and non-linguistic signs (e.g., images, gestures, or sounds). It examines how signs convey meaning and how people interpret those meanings within a cultural context.

In experimental semiotics, researchers use empirical methods and experiments to investigate how perception plays a role in the interpretation of signs. This field aims to understand the cognitive processes involved in understanding signs and symbols and how these processes might vary across different individuals or cultures.

Key areas of study in experimental semiotics based on perception may include:

Perceptual Encoding of Signs: How signs and symbols are visually or auditorily processed by individuals. This includes understanding the mechanisms involved in recognizing and encoding signs into mental representations.

Cross-Modal Perception: Examining how signs in one modality (e.g., visual) may influence the perception and interpretation of signs in other modalities (e.g., auditory).

Cultural Differences in Perception: Investigating how culture shapes the perception and interpretation of signs. Different cultures may attach different meanings or significance to the same sign.

Semiotics of Non-Linguistic Communication: Studying the semiotic systems used in non-linguistic communication, such as gestures, facial expressions, or even artistic symbols.

Semiotic Priming: Analyzing how exposure to certain signs can prime individuals to interpret related signs or symbols in a particular way.

Multimodal Communication: Exploring how multiple modes of communication (e.g., verbal language, gestures, and images) interact to convey meaning.

Experimental Methodologies: Developing and utilizing experimental techniques and paradigms to test hypotheses related to semiotic perception.

Neurosemiotics: Investigating the neural basis of semiotic processing using brain imaging techniques, such as fMRI or EEG.

**2.3 Gibson‘s Affordance theory**

Gibson's Affordance Theory, developed by the psychologist James J. Gibson in the 1970s and 1980s, is a fundamental concept in ecological psychology. This theory focuses on how individuals perceive and interact with their environment, emphasizing the relationship between an organism's abilities and the properties of the environment.

The central idea of Gibson's Affordance Theory is that the environment contains "affordances" that are opportunities for action available to an individual. Affordances are the action possibilities that the environment offers, based on the physical properties of objects and the organism's capabilities. In other words, they are the potential actions that an individual perceives they can perform with respect to objects or elements in their surroundings.

Here are some key points of Gibson's Affordance Theory:

Perception-Action Coupling: According to Gibson, perception and action are closely linked and inseparable. Perception is not just about passive observation; instead, it is intimately connected to an individual's potential actions in the environment. The environment is perceived in terms of what actions it allows and invites.

Direct Perception: Gibson rejected the idea that perception involves complex cognitive processing or mental representations. Instead, he proposed that perception is direct and immediate. When individuals perceive the affordances in their environment, they do so effortlessly and without the need for conscious decision-making.

Relation to the Observer: Affordances are relational properties, meaning they depend on the interaction between the individual and the environment. For example, a chair affords sitting for a human but not for a bird. Affordances are not intrinsic to objects; they emerge from the relationship between the observer's abilities and the properties of the environment.

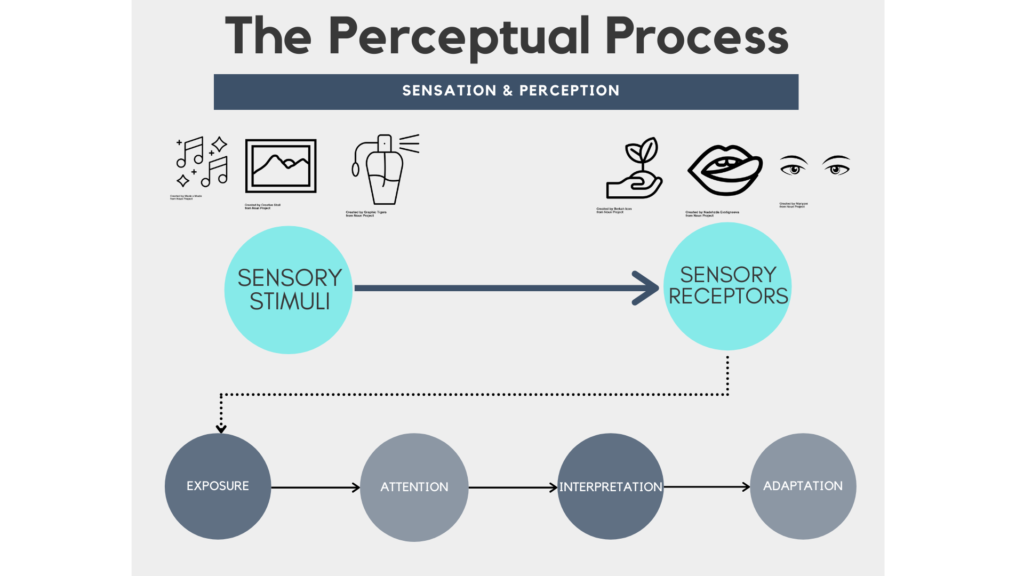
Information for Action: The environment provides rich and meaningful information that guides an individual's actions. Perception is considered functional as it provides the necessary information to carry out specific actions successfully.

Perception of Scale: Gibson's theory also accounts for the perception of spatial layout and distance. Affordances, such as the size and shape of objects, inform individuals about how to interact with and navigate through their environment.

Gibson's Affordance Theory has had a significant impact on various fields, including psychology, cognitive science, design, and human-computer interaction. It highlights the importance of studying perception and action in a holistic manner, acknowledging the inseparability of the individual from their environment and the affordances it presents. The theory has inspired research into topics such as visual perception, motor control, and the design of user-friendly environments and interfaces that consider the natural interaction between humans and their surroundings.

**2.4 A Model of Perceptual Processing**

Perception is how you interpret the world around you and make sense of it in your brain. You do so via stimuli that affect your different senses — sight, hearing, touch, smell, and taste. How you combine these senses also makes a difference. For example, in one study, consumers were blindfolded and asked to drink a new brand of clear beer. Most of them said the product tasted like regular beer. However, when the blindfolds came off and they drank the beer, many of them described it as “watery” tasting (Ries, 2009). This suggests that consumers’ visual interpretation alone can influence their overall attitude towards a product or brand.



The perceptual process begins when our sensory receptors (eyes, ears, tongue, nose, and skin) come in contact with sensory stimuli (sights, sounds, tastes, odours, and textiles) around us. Through our sensory system, we are exposed to an infinite amount of stimuli, some of which we pay attention to, and some we tune out completely. Those that receive our attention we evaluate and interpret their meaning based on our methods of cognitive and behavioural processing. After time, some of these stimuli become adapted and we seize to remark on their significance.

**Importance of Perception**

Perception is the process of selecting, organizing, and interpreting information. This process includes the perception of select stimuli that pass through our perceptual filters, are organized into our existing structures and patterns, and are then interpreted based on previous experiences. Although perception is a largely cognitive and psychological process, how we perceive the people and objects around us affects our communication. We respond differently to an object or person that we perceive favorably than we do to something we find unfavorable. But how do we filter through the mass amounts of incoming information, organize it, and make meaning from what makes it through our perceptual filters and into our social realities?

Ultimately, the perceptual process develops a consumer’s perception of a brand and formulates the brand’s position vis-à-vis the competition on what marketers call a positioning strategy.

If consumers were to only rely on sensation, it is unlikely they would be able to draw any distinction between similar products. Peanut butter, cola, ice cream…each of these product categories have competitors vying to differentiate their products from one another. If you were to organize blind taste-tests with your friends where they could only rely on the sensation of taste, they may not be able to distinguish between them. So while sensation is what we experience when our sensory receptors are engaged, it is perception, that ultimately influences our consumer decisions and forms the basis of our preferences.

For marketers, having your brand stand out in a crowded and noisy marketplace is critical to success: playing to consumers’ senses is “next level” marketing as these rich experiences can code a brand into the consumer’s memory. Capturing the consumer’s attention through stunning visual appeals, catchy sounds, tasty samples, delicious aromas and hands-on experiences (also known as Guerilla Marketing) have completely over-taken the passive advertisements and billboards of the past. When done successfully, sensory marketing transitions a brand from “barely being noticed” to earning a top position in the consumer’s mind.

Perceptual processing refers to the cognitive and neural processes involved in the interpretation and organization of sensory information from the environment. There are several models of perceptual processing that attempt to explain how the brain processes and makes sense of incoming sensory stimuli. One influential and widely accepted model is the "Perceptual Processing Model," which can be broken down into several stages:

Sensory Input:

The process begins with sensory input from the external environment, such as visual information from the eyes, auditory information from the ears, or tactile information from the skin.

Feature Detection:

In this stage, basic features of the sensory input are detected. For example, in the visual system, simple features like edges, colors, and orientations are identified. Different sensory modalities will have their own specific feature detectors.

Pattern Recognition:

Once basic features are detected, the brain works to recognize more complex patterns and objects. This involves combining and integrating the detected features to identify meaningful objects or events in the environment. For example, recognizing a face or a familiar object involves pattern recognition.

Top-Down Processing:

Top-down processing refers to the influence of higher-level cognitive processes on perception. Expectations, knowledge, and prior experiences can shape how we perceive sensory information. For instance, when looking at a blurry image, our brain might use context and knowledge to "fill in the gaps" and recognize the object.

Perceptual Organization:

Perceptual organization refers to the process of grouping and organizing sensory information into coherent perceptual units. Gestalt principles, such as proximity, similarity, continuity, and closure, play a role in this stage.

Perceptual Constancy:

Perceptual constancy refers to the brain's ability to perceive objects as stable and consistent despite changes in their appearance due to variations in lighting, viewing angle, or distance. For example, we can recognize an object as the same even if we see it from different angles.

Perceptual Inference:

Perceptual inference involves making educated guesses or inferences about sensory information based on context and prior knowledge. The brain may fill in missing details or resolve ambiguous sensory input to arrive at a coherent percept.

Action and Response:

The final stage of perceptual processing involves translating the perceived information into appropriate actions and responses. This could include motor actions, emotional reactions, or further cognitive processing.

It's important to note that perceptual processing is a highly complex and interactive process that involves interactions between different brain regions and networks. Different sensory modalities, such as vision, audition, and touch, may have slightly different processing pathways, but they all contribute to our overall perception of the world around us.

While the Perceptual Processing Model provides a general framework for understanding how perception works, it's essential to recognize that research in this field is ongoing, and the brain's mechanisms for perceptual processing are still not fully understood.

**UNIT – III**

**VISUALIZATION TECHNIQUES: SPATIAL DATA**

**3.1 Visualization Techniques**

Visualization techniques refer to various methods and tools used to represent data, information, or abstract concepts visually, often with the goal of gaining insights, communicating complex ideas, or making information more accessible and understandable. Visualization is commonly used across various fields, including data science, business, education, and science, to name a few. Here are some popular visualization techniques:

**Bar Charts and Histograms:**

Bar charts and histograms are used to display categorical and numerical data, respectively. They present data in rectangular bars of different lengths, with the length or height proportional to the values being represented.

**Line Charts:**

Line charts display data as a series of data points connected by straight lines. They are commonly used to show trends or patterns over time or across different categories.

**Pie Charts:**

Pie charts represent data as a circular graph, with each slice representing a proportion of the whole. They are useful for showing how parts relate to the whole, such as percentages of a total.

**Scatter Plots:**

Scatter plots are used to visualize the relationship between two continuous variables. Data points are plotted as individual dots on the graph, and the distribution and correlation between the variables can be observed.

**Heatmaps:**

Heatmaps use color-coding to represent data values in a two-dimensional matrix. They are commonly used to show the intensity or density of data points in a particular area.

**Choropleth Maps:**

Choropleth maps represent data by coloring regions or areas based on the values of the data. They are often used to visualize geographical data, such as population density or election results.

**Tree Maps:**

Tree maps display hierarchical data as nested rectangles, with the size of each rectangle representing the relative proportion of the data it represents.

**Word Clouds:**

Word clouds display textual data, with the size of each word indicating its frequency or importance in the text.

**Network Graphs:**

Network graphs represent relationships between entities as nodes and edges. They are useful for visualizing complex networks, such as social networks or interconnected systems.

**3D Visualizations:**

3D visualizations add an extra dimension to the representation, allowing for more immersive and interactive views of data or structures.

**Infographics:**

Infographics are visual representations that combine various visualization techniques, text, and images to convey information in a concise and engaging manner.

**Interactive Visualizations:**

Interactive visualizations enable users to manipulate and explore data dynamically, allowing for a more personalized and in-depth understanding of the information.

The choice of visualization technique depends on the nature of the data, the intended message, and the audience. Effective visualization can facilitate better understanding, reveal patterns or trends, and support decision-making processes in various domains.

Spatial data, also known as geospatial data, refers to information that has a geographic or spatial component, which means it is associated with specific locations on the Earth's surface or within a defined coordinate system. Spatial data provides valuable context and insights, as it allows for the visualization, analysis, and understanding of information in relation to its location.

**3.2 Spatial data**

There are two primary types of spatial data:

**Vector Data:**

Vector data represents spatial features as discrete points, lines, or polygons. Each feature is defined by its geometric coordinates (latitude and longitude or X, Y, Z coordinates) and attributes (additional information associated with the feature). Examples of vector data include points representing cities, lines representing roads, and polygons representing countries or administrative boundaries.

**Raster Data:**

Raster data represents spatial information as a grid of cells or pixels. Each cell in the grid has a value, which can represent attributes such as elevation, temperature, land cover, or any other continuous data. Raster data is commonly used in remote sensing, satellite imagery, and geographic information systems (GIS) for tasks like land cover classification, terrain analysis, and environmental monitoring.

Spatial data is used in various fields and applications, including:

**Geographic Information Systems (GIS):**

GIS is a system designed to capture, store, manipulate, analyze, and present spatial and geographic data. It enables users to create maps, perform spatial analysis, and make informed decisions based on geographic relationships.

**Environmental Analysis:**

Spatial data is used to study and analyze various environmental factors, such as habitat distribution, pollution levels, climate patterns, and changes in land use.

**Urban Planning and Transportation:**

Spatial data helps urban planners and policymakers in managing infrastructure, transportation networks, and urban development. It assists in optimizing routes, analyzing traffic patterns, and locating facilities.

**Natural Resource Management:**

Spatial data is essential for managing natural resources, such as forestry, agriculture, water resources, and wildlife habitats. It aids in monitoring and conservation efforts.

**Emergency Management and Disaster Response:**

Spatial data is crucial in planning for and responding to natural disasters, such as hurricanes, earthquakes, and wildfires. It helps in resource allocation, risk assessment, and evacuation planning.

**Business and Marketing:**

Spatial data is used for location-based marketing, site selection for businesses, and understanding customer demographics and preferences.

Spatial data can be collected from various sources, including satellite imagery, aerial surveys, GPS devices, sensor networks, and crowd-sourced data. The integration and analysis of spatial data play a pivotal role in understanding the world around us and making informed decisions for a wide range of applications.

**3.3 One-Dimensional Data, Two- Dimensional Data**

One-dimensional data and two-dimensional data are types of data that differ in terms of the number of dimensions they represent.

**One-Dimensional Data:**

One-dimensional data, also known as univariate data, consists of a single variable or attribute. It is essentially a list of values that can be represented along a single axis or line. Common examples of one-dimensional data include a list of temperatures recorded each day, the ages of individuals in a population, or the heights of a group of students.

Visualization techniques for one-dimensional data include histograms, bar charts, line charts, and frequency distributions. These visualizations help to summarize the distribution of the data and identify patterns or trends.

**Two-Dimensional Data:**

Two-dimensional data, also known as bivariate data, involves two variables or attributes that are observed simultaneously for each data point. The data is represented as a collection of pairs of values, and each pair corresponds to a unique combination of the two variables. Examples of two-dimensional data include the relationship between a person's weight and height, the temperature and humidity recorded at different times of the day, or the test scores of students in two different subjects.

Visualization techniques for two-dimensional data include scatter plots, bubble charts, contour plots, and heatmaps. These visualizations help to explore the relationship between the two variables and identify correlations or patterns.

It's important to note that the concepts of one-dimensional and two-dimensional data can be extended to higher dimensions, where data involves more than two variables. For example, three-dimensional data involves three variables, and the data points can be represented in a 3D space. However, as the number of dimensions increases, visualization becomes more challenging, and specialized techniques such as parallel coordinates or multidimensional scaling may be used.

In summary, one-dimensional data represents a single variable, and visualization is often done along a single axis, while two-dimensional data involves two variables, and visualization is done in two-dimensional space to explore their relationship. Understanding the dimensionality of data is essential for selecting appropriate visualization techniques and conducting meaningful data analysis.

**3.4 Three Dimensional Data, Dynamic Data**

Three-dimensional data and dynamic data are additional concepts related to the representation and characteristics of data.

**Three-Dimensional Data:**

Three-dimensional data, also known as trivariate data, involves three variables or attributes. In this case, each data point is represented by a combination of three values, and the data can be visualized in a three-dimensional space. Examples of three-dimensional data include the relationship between temperature, humidity, and air pressure at different locations in a region or the three-dimensional coordinates of points in a 3D environment.

Visualization techniques for three-dimensional data include 3D scatter plots, 3D surface plots, and volumetric rendering. These visualizations help to explore the relationships between the three variables and provide a more comprehensive understanding of the data.

**Dynamic Data:**

Dynamic data, also known as time-series data or temporal data, refers to data that changes over time. It involves multiple observations taken at different points in time for one or more variables. Dynamic data can be one-dimensional or multi-dimensional, depending on the number of variables being observed over time.

Visualization techniques for dynamic data include line charts, area charts, animated plots, and heatmaps with time as an additional dimension. These visualizations allow for the exploration of trends, patterns, and changes in the data over time, providing insights into temporal relationships and behaviors.

Dynamic data is commonly encountered in various fields, such as finance (stock prices over time), weather forecasting (temperature and precipitation patterns), health monitoring (patient vitals over time), and social media (trending topics and user interactions).

In some cases, dynamic data can also be combined with spatial data, resulting in spatio-temporal data. Spatio-temporal data involve observations that are both located in space and vary over time. This type of data is particularly relevant in fields such as environmental monitoring (tracking the movement of air pollutants) and transportation analysis (studying traffic patterns in a city over time).

Effective visualization of dynamic data requires techniques that convey changes over time in an understandable and intuitive manner. Interactive visualizations that allow users to control and manipulate the temporal aspect of the data can enhance the exploration and analysis of dynamic datasets.

**3.5 Combining Techniques**

Combining visualization techniques can be a powerful approach to gain deeper insights, convey complex information, and provide a more comprehensive understanding of the data. By integrating multiple types of visualizations, data analysts and decision-makers can explore data from different angles and uncover hidden patterns or relationships. Here are some ways to combine visualization techniques effectively:

**Small Multiples:**

Small multiples involve creating a series of similar visualizations, each depicting a subset of the data. For example, you can display multiple line charts, each representing a different region's temperature trend over time. This technique allows for easy comparison and identification of patterns across different subsets of data.

**Linked Visualizations:**

Linked visualizations involve connecting two or more visualizations, such that interactions in one visualization affect the others. For instance, you can create a scatter plot and a histogram with linked brushing, so selecting points in the scatter plot updates the corresponding data in the histogram. This facilitates exploration and correlation analysis between different variables.

**Combining Geospatial and Temporal Data:**

For spatio-temporal data, combining maps with time series plots or animations can reveal how patterns change over time and their geographic distribution. This can be particularly valuable in understanding phenomena like weather patterns, disease outbreaks, or migration trends.

**Integrated Dashboards:**

Dashboards provide a way to present multiple visualizations and insights in a single interface. By integrating various visualization techniques, such as line charts, bar charts, and maps, into a cohesive dashboard, users can get a holistic view of the data and explore relationships between different aspects.

**Visual Hierarchies:**

Creating visual hierarchies involves presenting information at different levels of detail. For example, you can display a geographic map at a high level and, upon zooming in, reveal additional details, such as regional data or individual data points. This approach allows users to explore the data at different levels of granularity.

**Visual Annotations:**

Adding visual annotations to a plot or chart can enhance its interpretability. For instance, using color-coded annotations on a scatter plot can help highlight specific data clusters or outliers, making it easier to identify important patterns.

**Animated Transitions:**

Animations can be used to show changes over time or transitions between different states in the data. Animating data points or elements in a plot can help users understand trends and temporal patterns more intuitively.

Combining visualization techniques requires careful consideration of the data and the insights you want to convey. The goal is to present information in a clear, concise, and engaging manner, enabling users to make data-driven decisions effectively. Interactive elements in the visualizations can further empower users to explore the data on their terms and gain deeper insights.

**3.6 Geospatial Data: Visualizing Spatial Data**

Geospatial data, also known as geographic or spatial data, refers to information that has a specific location on the Earth's surface or within a defined coordinate system. Geospatial data is characterized by its association with geographical coordinates, such as latitude and longitude or projected coordinates (X, Y).

Visualizing geospatial data is essential to gain insights into patterns, relationships, and trends that are location-based. There are several effective visualization techniques to represent geospatial data:

**Choropleth Maps:**

Choropleth maps use different colors or shading to represent data values for specific geographic regions or areas. The color intensity indicates the magnitude of the data variable, making it easy to identify spatial patterns and variations.

**Heatmaps:**

Heatmaps represent the density or intensity of data points in a geographical area using colors. They are especially useful for visualizing spatial distributions and identifying hotspots.

**Symbol Maps:**

Symbol maps use symbols, such as circles or icons, to represent individual data points at their specific geographic locations. The size or color of the symbols can be used to convey additional data attributes.

**Dot Density Maps:**

Dot density maps use dots to represent individual data points, with each dot representing a specific number of occurrences. This technique helps visualize the distribution of data across space.

**Flow Maps:**

Flow maps show the movement of data, such as people, goods, or information, between different locations. Flow lines or arrows indicate the direction and volume of movement.

**3D Maps:**

3D maps add a third dimension to spatial visualization, enabling a more immersive view of geographic features or terrain.

**Animated Maps:**

Animated maps show changes in geospatial data over time, allowing for the exploration of temporal patterns and trends.

**Story Maps:**

Story maps combine geographic data with narrative elements, text, images, and multimedia to tell a compelling story related to a specific location or theme.

**Web Maps:**

Web-based mapping platforms, such as Google Maps, Leaflet, or Mapbox, allow for the interactive visualization of geospatial data on web browsers, enabling users to explore and interact with the data dynamically.

**Interactive Maps:**

Interactive maps enable users to customize the visualization, filter data, zoom in and out, and access additional information through pop-ups or tooltips.

It's important to select the appropriate visualization technique based on the type of geospatial data, the analytical goals, and the intended audience. Effective visualization of geospatial data can facilitate decision-making processes, aid in problem-solving, and enhance our understanding of the geographical aspects of the data.

**3.7 Visualization of Point Data, Visualization of Line Data**

* 1. **Visualization of Point Data:**

When visualizing point data, the goal is to represent individual data points at their specific geographic locations. There are various techniques to visualize point data effectively:

**Scatter Plot Map:**

A scatter plot map displays each data point as a separate marker at its corresponding geographic coordinates. The size, color, or shape of the markers can be used to represent additional data attributes associated with each point.

**Proportional Symbol Map:**

In a proportional symbol map, the size of the symbols (e.g., circles or icons) corresponds to the magnitude of the data values at each location. Larger symbols represent higher values, making it easy to discern spatial patterns.

**Clustered Point Map:**

When dealing with a large number of data points that may overlap, using a clustered point map can help aggregate points that are close to each other. Clusters of points are represented by a single symbol, and users can interact with the map to zoom in and see individual points.

**Graduated Symbol Map:**

A graduated symbol map categorizes the data into classes or groups and represents each group with a different symbol size. It is useful when there is a range of values to visualize.

**Dot Density Map:**

A dot density map uses dots to represent individual data occurrences. Each dot represents a specified number of data points, and the overall distribution helps identify patterns and concentrations.

* 1. **Visualization of Line Data:**

When visualizing line data, the focus is on representing the continuity and flow of data along a series of connected points. Here are some effective techniques for visualizing line data:

**Line Chart:**

A line chart displays a series of data points connected by straight lines. It is suitable for visualizing trends, patterns, and changes in data over time or across a sequence.

**Flow Map:**

A flow map represents the movement of data, such as people, goods, or information, between different locations using flow lines or arrows. The thickness or intensity of the lines can indicate the volume or strength of the flow.

**Path Map:**

A path map visualizes the trajectories or paths of moving objects, such as vehicles or animals, on a map. It is effective in showing patterns of movement and routes taken.

**Spider Map:**

A spider map uses lines radiating from a central point to connect multiple locations with the central point. It is useful for visualizing connections between multiple places and a central hub.

**Sankey Diagram:**

A Sankey diagram visualizes the flow of data or resources through a system, representing the magnitude of the flow with the width of the lines.

Selecting the appropriate visualization technique depends on the specific characteristics of the data, the analytical objectives, and the audience's needs. Effective visualization can help reveal insights, relationships, and patterns in both point data and line data.

**3.8 Visualization of Area Data - Other Issues in Geospatial Data Visualization**

**Visualization of Area Data:**

When dealing with area data in geospatial visualization, the goal is to represent data that covers specific geographical regions or areas. Here are some effective techniques for visualizing area data:

Choropleth Maps:

Choropleth maps use color or shading to represent data values for specific geographic regions. Each region is filled with a color or shade that corresponds to the value of the data variable in that area. Choropleth maps are commonly used to show variations in data across regions, such as population density, unemployment rates, or election results.

Heatmaps:

Heatmaps can also be used to visualize area data, where colors indicate the intensity or density of data points within each region. Heatmaps are particularly useful for showing the concentration of data occurrences, such as crime hotspots or disease outbreaks.

Filled Contour Maps:

Filled contour maps represent data as continuous contours with color-filled areas, where each contour line represents a specific value. This technique is often used to visualize data that varies smoothly over space, such as temperature or elevation data.

Bivariate Choropleth Maps:

Bivariate choropleth maps visualize two variables simultaneously using different colors for each variable. This allows for the exploration of correlations or patterns between two variables in specific geographic regions.

Cartograms:

Cartograms are specialized maps that distort the size or shape of regions based on the data being visualized. It allows for a more accurate representation of the data's relative magnitude rather than the actual geographic size of the regions.

**Other Issues in Geospatial Data Visualization:**

Data Quality and Accuracy:

Geospatial data visualization relies heavily on the accuracy and quality of the underlying data. Inaccurate or incomplete data can lead to misleading or erroneous visualizations. Data validation and cleaning are crucial steps before creating visualizations.

Data Scale and Projection:

Choosing an appropriate map scale and projection is essential to accurately represent geographic features and maintain spatial relationships. Different map projections can distort the shape, area, or distance on the map, which can impact the interpretation of the data.

Handling Large Datasets:

Geospatial datasets can become quite large, especially in applications like remote sensing or tracking systems. Handling and visualizing big geospatial data require efficient algorithms and appropriate data reduction techniques.

Interactive Visualization:

Interactive geospatial visualization allows users to explore and interact with the data dynamically. Adding interactive elements like zoom, pan, and filtering enables users to focus on specific regions of interest and gain deeper insights.

Data Privacy and Security:

Geospatial data often includes sensitive information about individuals or locations. Ensuring data privacy and implementing appropriate security measures when visualizing geospatial data is crucial to protect personal and sensitive information.

Context and Interpretation:

Geospatial visualizations should provide sufficient context and be accompanied by informative labels, legends, and scale bars to aid in interpretation. Properly annotating the visualizations helps users understand the data and draw meaningful conclusions.

Effective geospatial data visualization can provide valuable insights into spatial patterns, trends, and relationships, aiding decision-making and problem-solving in various fields such as urban planning, environmental monitoring, and public health.