Computer Aided Decision Systems Industrial Practices using Big Analytics
Professor Deepu Philip
Department of Industrial & Management Engineering
Indian Institute of Technology Kanpur
Professor Amandeep Singh
Imagineering Laboratory
Indian Institute of Technology Kanpur
Lecture 51
Data Visualization (Part 1 of 2)

Welcome to the last week of the course Computer Aided Decision Support Systems. This will be the last one or two lectures, where we will discuss Data Visualization. We have recently discussed HTML and PHP. How do we connect that when we connect that we try to get the User Interface, where the User Interface, we need to present the data, we need to present the function and the form of the data types that you have?

(Refer Slide Time: 00:45)

Contents

▸ Defining Data Visualisation
▸ The upsides and downsides
▸ The significance in today's world, Big Data
▸ Data visualisation framework
▸ Data visualisation types
▸ Data visualisation examples

▸ 2

Now, the Data Visualization will be covered in this lecture, where we will be covering the definition of it. Then, there are some Advantages and Disadvantages of Data Visualization, obviously, it is used for multiple merits, but there could be certain demerits. Some points could be missed that also we will discuss the Significance of Data Visualization in industry 4.0 or in today's world that is Big Data, how this is being used to present the structure or use some form of it. Then, Framework of Data Visualization then, Visualization Types and some examples.
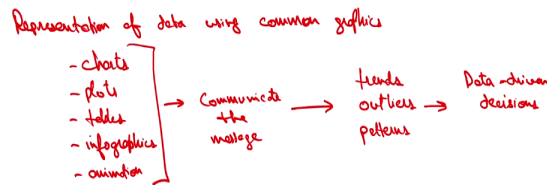
Let me start with the definition, defining Data Visualization. But definition Data Visualization is only representation data using common graphics. These graphics would be charts, plots, tables, infographics or anything like even animations.

By definition, Data Visualization helps us to have a representation of the data in various forms in various graphics, so that we are able to communicate the message. This message could be the relationships between data, the data driven insights, so that one is able to understand or one can relate with the trends, outliers, patterns. This helps us to make data driven decisions.

Making an interface, developing your program, developing your website, developing your PHP and HTML programs is okay. But finally, how to present that, what kind of colors would you choose? What is your audience, trying to understand the audience?

I will talk about major tips when we try to have a good data representation, the form of the data, the function of the data, the form should not kill the function, the function should not be so represented that the form is not even interactive, we do not even have a statically looking failure. So, it is important to have a good understanding of what your data is, what are you trying to represent? Who is the audience? All these things when you are trying to generate an idea even when we are talking about Data Visualization.

Design Thinking sessions also help us to understand it. When we try to have a concept of design thinking, we try to first empathize with what the user would like to have? What the user would

like to see? What user would like to visualize? Then, we try to define what our problem statement is, Then, we try to bring up some ideas over it. So, this is the Data Visualization Framework. Also, we will talk about in the coming slides.

So, Data Visualization is a very important step in the data science process, which allows the teams or the individuals to convey the data more effectively to colleagues and decision makers, so that the teams can organize the reporting systems, specifically defined template views to monitor performance. We cannot miss that the Data Visualization is not only limited to the performance dashboard.

For example, text mining, as an analyst may use the word cloud to capture key concepts and trends or hidden relationships with unstructured data. Alternatively, these might also use the graph structure to have a relationship between the entities of the knowledge graph itself. So, there are certain ways Data Visualization is used.

(Refer Slide Time: 05:00)

Yes, there are certain Upsides and Downsides, if I talk about the advantages or demerits.

☐ Clarity- Data Visualization helps us to have great clarity on what we are trying to present to our audience. That means, one graph of the present 1000 words. Telling just in theory or writing just in a report that okay, the GDP has risen by 5 percent in the last five years, or the average income has risen by around 2 percent. In the last one month or presenting them in a graph, putting in a pie chart India various states how the GDP has risen. So, one

graph represents 1000 words telling the whole story and putting them in one graph. That is why, we need Data Visualization.

So, our eyes just flow like water, it tries to follow equilibrium, it tries to follow the lowest path. So, Data Visualization is designed, so that the colors are also chosen. The best part or the most prominent word that we would like to show is highlighted in a way that might be put in a red color that might be put in a yellow color, so that the person sees or that might be put in a bold letters or bigger sizes. So, that one is able to see what is the overall rest of the visual presentation that we are trying to bring in.

So, we can quickly identify red from blue squares from circles, so, that is why our culture is visual, which includes everything from the arts, the reports, advertisements, journals, animations. So, Data Visualization is a form of visual art that grabs our interest and keeps our eyes on the message that the person who has designed it is trying to present. So, when we try to look at a chart, we very quickly try to see the trends and outliers in it.

- ☐ Precision- In Data Visualization, what specifically is the purpose? So, we can have a target value that could be bold or a contrasting color or maybe highlighted using certain other ways.
- ☐ Efficiency- So, Data Visualization helps us to have an efficiency or an ease of sharing the information.
- ☐ Maximize ideas, minimize ink- So, we interactive explore the opportunities, we visualize the patterns in relationships, we also maximize the ideas with the minimum inks. I think I would say 1000 words = one chart.

With the upsides, certain downsides are there that I can say the demerits. Sometimes in Big Data, we have a lot of information and what we are trying to target or the primary goal is missed. That is why, I showed you the framework or the process for the Big Data presentation, how we present the data.

Now, sometimes the basic message is missed, we try to sometimes go into the form of the data function that is missed, there is always a battle between form and function.

For instance, you have an engineer or a data scientist who is trying to write a code who understands what is the function of the data or we have the data administrator who has a goal to

be put in. The data scientists sometimes have a different viewpoint on what colors could be chosen. So, the data administrator has to take the call that the function we cannot miss, this is a more important function.

For instance, if I need to say so, IIT Kanpur has risen in SQS ranking in the last 10 years or so. I would only say a rank chart or I can say maybe a histogram showing the ranks between different years of IIT Kanpur and other institutions in India. This is my function and form both coming up.

Sometimes, I just might like to say okay, the form is more important. The color is more important and the basic function is sometimes missed. For instance, the rank of IIT Kanpur has risen in the last 10 years.

And, if I tried to say, if I try to compare with the other IITs or other institutions in India or if I tried to compare with the world top rankings IIT Kanpur would still see its ranking down there maybe at 48 or 100 or so, wherever the rank is in the last years. So, the form of the data and the function of the data function is to present that our data or IIT Kanpur has risen or enhanced its ranking in the last few years. Accordingly, it has to be put, I will talk about this more, when I will talk about the cheating case.

- ☐ So, Data Visualization sometimes gives information that is either incorrect or biased. For example, when viewing a visualization with many different data points, it is easy to have an interpretation or an inept accurate assumption or sometimes the visualization is just designed even wrong, so that it is biased or it is confusing the person.
- ☐ The correlation does not indicate cause. So, the causation is always missed in correlation, correlation is already told you in comparison to the boys in the institution the girls have scored more marks, but what is the reason of that for that we need to do further study, for that we need to have certain regression or analysis or so based upon the certain parameters, what were the parameters. The girls might be spending more hours in study, the boys might be spending more hours just roaming around or the girls might have higher ranks in the admission itself than those who are there. There might be certain parameters which could be used or which could be set and regression analysis certain, other kinds of analysis could be taken to have exactly what is the cause of this.
- ☐ Sometimes, as I said, the core statements can be lost in translation with translations when we talk about the Data Visualization framework. We try to map that data with the basic

data and try to have a visual graph of it. The basic or the core statements are sometimes missed. So, these could be certain demerits of the Data Visualization even.

(Refer Slide Time: 12:14)

Now, let me talk about the Significance of Data Visualization in the world nowadays. So, Big Data, it is important that data should make a good sense, it should be more understandable. So, the importance of Data Visualization is very simple, it helps people see, interact with and better understand the data. So, whether simple or complex Data Visualization can bring everyone on the same page regardless of their level of expertise.

Audiences though versatile, everyone has to be brought on the same page. Because these days the audience that you get, we do not know is the school's great coder or even a person who is there as a CEO might not be able to understand what is the basic code with which this website is designed or so. So, different audiences are different parts of the audience, who your audience is, and what kind of interaction you are trying to do. Are you only going to present the data? Are you going to interact with them? Or are you going to tell a story that also I will discuss what kinds of types of Data Interaction or Data Visualizations that you have. So, this is important.

And also, it is hard to think of a professional industry that does not benefit from the data nowadays. So, with the concept of the citizen data scientist is on rise, where skill sets are changing to accommodate the data driven world. So, it is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when the data informs

who, what, when, where, and how the specific field or the specific topic that you are trying to discuss is presenting.

To Data Visualization if I say, in Big Data, it sits between data analysis and you need to have storytelling so that we have Data Visualization. Now, this is the age of Big Data where we are trying to talk about the trillions of rows of the data that is being generated each day.

So, in this case, Data Visualization is not only to dress up a graph to be ready to look up the information or the infographic is there, the effective Data Visualization is a delicate balance between form and function, I will have to put that here once again form versus function. A very simple graph could be too boring if the form is too simple. The case of form so that no one notices it, but it can still make powerful information that could be extracted from it simply is boring.

For instance, if just in a black color form itself only a line graph shows from the number of students who have reason to take NPTEL courses, it could be just given okay from 500 it has risen to 50,000 in last 10 years, this could be very simple, but this 500 to 50,000 is 100 times rise, it is the powerful message that it is a 100 time rise of the number of students who are taking the courses in specific fields maybe mechanical engineering or computer science engineering. So, this could be very simple, but it could give a powerful message.

On the other hand, form versus function, if I say the form could be a very stunning visualization with lots of colors, I would say a complex visualization, but in this case, as I said in the previous slide as well, the core statement might be missing.

For instance, if I just say, in mechanical engineering, these many students have taken the courses in computer science, these many in electrical, these many people have cleared the courses a lot of information is given is just one in A4 size poster. So, it has a lot of information, but finally, what it is trying to tell is, that the number of people is risen by 100 times in who are taking the courses in NPTEL.

So, sometimes the complex visualization could miss the core statement or it could not present the message very powerfully. So, the data and visuals need to work together. There is always a balance between them. And, this art of combining these two forms and functions is actually the Data Visualization itself.

So, in Big Data, because we have a lot of data, we have a lot of ways that data is coming, what specifically is the function but specifically is the message we are trying to give from there itself the data has to be driven.

(Refer Slide Time: 18:13)

Data visualisation framework
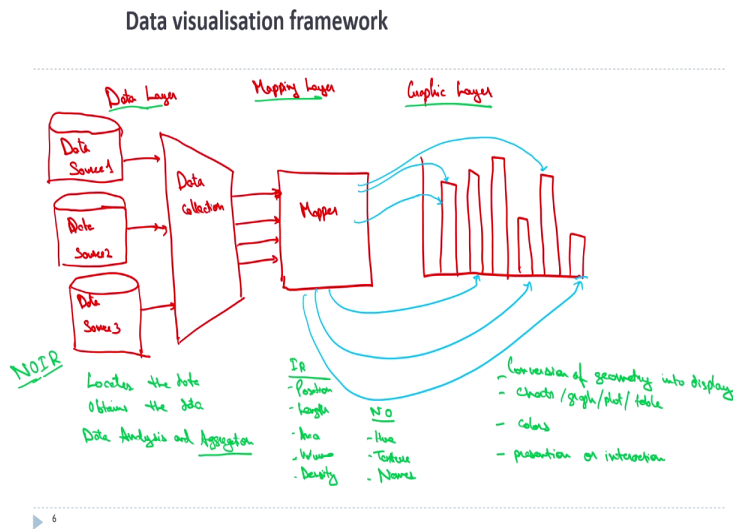
To understand this, let us try to see what the Framework of Data Visualization is. When I talk about the word framework, it means how do we design our Data Visualization, it could be simply three layers: we have a Data layer, we have a Mapping layer and a Graphic layer. We are actually going to choose the kind of plot to the king of the chart that we have and what kind of data inputs that we have.

See here generally we have different data sources. Maybe, data source one data source two and data source three, which I am putting it as my stored data. So here, the data is collected from the data sources. We call it data mining. We call it data collection. We call it data MapReduce who just connected here and then we try to map the layer.

Here, the MapReduce function comes into play more prominently. We have a mapper. But different data points come here. From here itself, we now go to the graphics. Let me say, I am putting a Histogram here and from the mapper different kinds of data points are taken from the mapper itself different data layers could be taken I might say, this layer is going here. So, it is going here. This one is also coming from here. How does this work? Let us try to talk about this thing.

The data layer has a purpose of locating and obtaining the data. Now, the importing data in the proper format is also important here, I talked about the different kinds of the data if you remember NOIR (Nominal Ordinal Interval and Ratio) scale data, this is just a categorical data, quantitative data depends upon what kind of data we have the relating data or for proper correspondence is also important here in the data layer itself.

Then, data analysis and aggregation is taken here. Because Big Data, multiple points, millions of or trillions of the datasets could be there. Whether the data is in a discrete form or continuous form, it is ordered, unordered. Do we need to transfer the categorical data into quantified form or is the quantified data solely to be presented in the lower high or yes or no. So, these translations could happen. This is known as Data Aggregation where we will try to get the final data because we are not talking about the Data Analytics right now, we will be talking about the Data Visualization in which the aggregate data is only to be presented finally, for the proper visual views that we are trying to get.

Now, this aggregate might be we might like to have a Pie chart, we might like to have Line diagram, we might like to have a Scatter plot or maybe more complex plots such as the Friend plot, the Eye plot, which could be taken depends upon the understanding of the person who is trying to read the data. If the person is simple, just a data administrator or maybe the business industrially, he would better understand the basic graphics: the Bar graph, the Line graph, Pie charts, this could be easily understood by them.

If the person is a student of statistics, the person should be could be given the data, different kinds of graphs such as, Box and Whisker plot could be given, the Bullet graph could be given, then, in Fan plot could be given, so that the person understand what is the data what is the overall information the data is trying to extract. Now, here, Data Aggregation is important. And, the kind of data that we have is important. This data layer tries to give us the information that we are going to feed into the mapping layer.

Next comes the second layer, which is a Mapping layer, which means we associate the appropriate geometry with corresponding data channels. So, it could be the data as I said, in a quantitative form, or it could be in a categorical form in a quantitative form, it could be maybe the position, length, area, volume, density, I am just taking some examples of the I and R. In the case of nominal ordinal, it could be just maybe the few textures, the names, etcetera. These are

connected to the data analysis algorithms so that we try to get the corresponding data channels which could be connected to our next layer that is the Graphic layer.

Now in the Graphic layer, what we have is the final conversion of geometry into when you say geometry kind of the data points that we have taken aggregate from the data layer and in the mapping layer, we have tried to map it to the king of the results that we are trying to present here.

So, here is the conversion of geometry into display which means we try to select the charts or graphs or plot or table and we try to select the colors and we try to see whether it has to be just presentation or interaction.

(Refer Slide Time: 26:04)

So, when I talk about Presentation or Interaction or Interactive Storytelling, this comes in our styles of Data Visualization. Data Visualization styles, they are majorly three styles, we try to have Interactive Visualization, Presentation Visualization and it could even be Interactive Storytelling.

When I say the word Interactive, it means there is an interaction between the audience and the person who is going to present the data or the website that you are trying to present. So, it has to have the user input as well.

1) Interactive Visualization- Here we have user input in between which means, it could be intended for a single investigator, or it could be for the collaborators.

For example, if am trying define proposal that I have to be for a ministry, I will take my presentation I will take to take different kinds of graphs, or different kinds of the options available, I would say this is option one, if they have a query, I will come to an option three. If we have a query, I can say okay, I have another option that is option five, it is also available. So, interaction also happens and the data is presented so that we are ready with the graphs and we are ready with different kinds of website presentations, we are able to connect with that. And, we are trying to have an interface or the input from the user as well. The user input is always there. So here for example, if I have also developed a prototype, we are already for an already taken project.

Prototype is to be presented before the final review committee. The Prototype has to have an interaction. The person would like to use the Prototype, how do we present that? This is my android interaction; this is my phone application. If you click this button this graph will open if you try to click if you try to input these values maybe for example, if it is a fluid mechanics experiment, if I say okay, if this is the fluid flow between 5 milliliters per seconds to 22 milliliters per second, this is the opening of the wall that you need to keep. So, these kinds of interactions, this could also help us to have an Interactive Visualization where we need to design it accordingly. So, the pre-renders based upon inputs are there. So, Tableau Public is also one of the examples that use Interactive Data Visualization. This is an example.

2) Presentation Visualization- It is only used for communication; this means I am not trying to interact with them. I am only giving the information. The example could be a poster that I am trying to present in a hardcopy or the poster displayed. The output number of students has risen from 500 to 50,000 in the last few years, this is just a poster, there is no interaction or input from them, so this is just a communication that means no user input is there. So, it does not support any user input. So, it is highly polished prior which means we cannot interact with them, whatever the information is taken by them or it is understood by them has to be final.

So, we have to be very careful in designing it and if it is polished prior it is well tested or well thought of prior itself that how it would present. Example could be anything, take example, I would say rank of the institute if I compare just simply presentation.

3) Interactive Storytelling- When we are trying to narrate something that we have an interactive webpage. So, in this case, the presentations are always there, but these are done through interactive webpages. Here, the story is if you do this, this happens. So, if another loop is well used here in this kind of data Presentation or Data Visualization, where we try to narrate something, we narrate using data and we tell a story. The purpose here is to invite the viewers to explore further.

For instance, if I am trying to present my proposal before the ministry, it is a technical proposal on maybe fluid mechanics itself. I will say fluid mechanics has certain experiments. These experiments are very important for the students of the first year of civil engineering, mechanical engineering, or production engineering. To understand these words, civil engineering, mechanical engineering, production engineering could be a link, where if they click, they try to understand what is civil engineering or mechanical engineering. In further itself, if I say there is a certain experiment such as Bernoulli's theorem or it could be some turbine, maybe Pelton turbine.

So, this Bernoulli' Pelton, if they click and a picture pops up which tries to tell them okay, what is this theory, what is the kind, what is the Pelton turbine? What are the other kinds of turbines which are there?

So, this interacts or invites the viewers to click more and more to have an understanding of that. Most of the web pages that you see, for instance, if you just see Wikipedia itself, the words which are there, I always try to have a link over it, which are complex words or which are words not very common. Also, the names of the person in a specific Wikipedia page, if you are trying to read the name of the person who is a prominent figure, the name of a person will also hyperlink. If you try to link that, if you try to click that, it will open the page of the person, where you can see the information about that person. So, this is known as Interactive Storytelling.

These are the ways or the styles, we try to present our visualization that could be certain tools, depending upon what kind of the business information that you have, it could be used for business intelligence, it is only to be used by the document presentation.

It depends upon what kinds of the tools that you have, what kinds of the final output or final presentation that you have, or final targets that you have. Data Visualization can help you to approach all of this.

Now, there are certain tips when we are trying to talk about Data Visualization, specifically, what we are trying to see or what we are trying to present and whom we are going to present these are very important. So, that means we need to understand four major points:

1) What is your story?
2) What is the kind of data?
3) Which chart fits best?
4) Who is your audience?

1) When I say, what is your story it means, what your data is trying to tell even. So, all the data that you are trying to present in the Data Visualization that you are trying to present here, is going to tell a story. Is the story easy to understand? For example, could your sales be up or down? Is the admission in the different institutes or in the engineering institutes in Uttar Pradesh on rise or decrease? Are the expenses within control? So, what is the reaction towards the new products which were launched last year?

So, what is a story? Whether it has to be Interactive Presentation, or Storytelling. This only is taken once you are very sure that what your story is, what is your final output that you are trying to target or what do you want the other person to understand? Once your target is fixed, what you are trying to present, what finally you are going to get out of it.

2) Then comes, what kind of data do you have or what do you want to explain through this data? So, is it qualitative? That is, we have become better than before or we have added new features in our product. Or is it even quantitative, that means you have a specific objective, objectivity is to be targeted. It could be your sales and earnings. It could be the

population statistics; it could be the number of products that were sold in the last year. So, what is the data that you are trying to present here?

Now, if you are trying to present data, depending upon the audience that you have, what kinds of charts that fits your data is very important. So, nearly all the organizations have quantitative information, which they collect or they put it in the database system itself. Data Visualization now allows the persons or organizations to present this data in a simple visual way. So that it makes sense at a glance. Quantitative data is ideal for visualization, because when you put numbers together, we can compare them and we can have a good plot out of them. It summarizes the essential characteristics that allow you to present disregard the exceptions.

Sometimes, the data is only qualitative techniques. That means, you can only present yes or no but these yes or no how many numbers of yes or no? How many times how many people have made yes or no percentages? These could be quantified further as well.

3) Which chart fits your data is very important. Now, depending upon the kind of the audience, these two are connected. One thing is, if my data is qualitative, it is only yes or no. I can only have a very limited kinds of charts, I only have maybe a number of people who have said yes or no, that could be put in a Pie chart itself, for example. Or if it is the quantitative data with different detailed information, we have the height and the weight of the students. The number of students who are there in a specific class, this is one data height, weight and also age of the students in a class or the month in which they were born, then height, weight, the month in which they are born and their eating habits. For example, the person who is obese would have a lesser height to weight ratio. So, this kind of quantitative information is important.

So, which chart fits best depends upon the kind of data and also the kind of the audience as I just mentioned before, these are the four tips or four points that you need to keep in mind.

If an audience is the data science intellect, those who are maybe able to understand the complex charts itself you can present accordingly, but if the audience here are the maybe the directors of the certain companies or you are certain investors those who have only money or you have people in ministry, those who only have to give the funds, then data has to be in a simpler form, anyway data has to be in a simple form. So, it depends upon the kind of audience that you have:

the professionals, the investors, the directors would have different views, the data scientists, the students, the learners, the organization, middle or low-level managers would have different viewpoints.

So, these four points are always to be kept in mind. With this I would like to have a break with my Data Visualization lecture. I would like to show certain examples and what are the different categories of the charts, plots, graphs that we have, that I will discuss in the next lecture. Thank you.