Data visualization is great for communicating with different types of people from various organizations and industries.

As the world becomes more data-driven, it is essential for us to be able to tell our stories with data and understand others too.

Data visualization helps us see and understand data, better interact with our customers, transform spreadsheets into stories and show that reports can be less intimidating.

In fact, what is data visualization?

Data visualization is the process of representing data in a graphical format using numbers, words, and images. It is a powerful tool that can help you understand data and make better business decisions.

There are many different types of data visualizations, but they all have one thing in common: they make data easier to understand. The most common type of data visualization is a line graph, which shows how a value changes over time. Other popular types of data visualizations include bar charts, pie charts, and scatter plots.

Data visualization is an important tool for anyone who works with data. Whether you're a business analyst or a scientist, data visualization can help you see trends, patterns, and relationships that you might not be able to see otherwise. And once you've found something interesting, you can use data visualization to communicate your findings to others.

The importance of visualization in today's business landscape and its usability

As the business landscape becomes more and more data-driven, organizations are realizing the importance of data visualization.

By using visual representations of data, businesses can gain insights that would otherwise be hidden in plain text. Data visualization also makes large data sets more manageable and easier to understand.

There are many ways to visualize data, and the right approach depends on the type of data being analyzed.

The already mentioned bar charts, line graphs and scatter plots are, as said, common methods to visualize and analyze data, along with heat maps.

Regardless of the technique used, data visualization can help businesses make better decisions, identify patterns and trends and improve communication.

A brief history of data visualization

Data visualization enables decision-makers to see relationships, motifs, and tendencies in data.

– The history of data visualization dates to the early 1800s, when Scottish engineer William Playfair invented the line chart and bar chart. Since then, data visualization has evolved considerably

– In the early 1900s, Florence Nightingale used graphical representations of data to improve sanitary conditions in hospitals

– In the mid-1900s, Hans Rosling used data visualizations to raise awareness about global health issues

In recent years, data visualization has become more popular with the advent of big data and powerful computing tools. Today, there are many different types of data visualizations, from simple charts and graphs to complex interactive dashboards.

Data visualizations are used by businesses, governments, and individuals to translate data into insights and make informed decisions.

Data visualization in today's business world

Data visualization is an important tool for businesses and organizations, as it helps in interpreting complex data sets. It allows businesses to easily gather data from various sources and put it to work with the purpose of supporting the decision-making process.

There are different types of data visualizations, including charts, graphs, maps, infographics, and dashboards. Each type of visualization has its own strengths and weaknesses, so businesses should choose the right type of visualization according to their needs.

Data visualization is increasingly being used in today's business world. Organizations are using data visualizations to make better decisions, communicate information more effectively and understand their customers better.

Data visualization and big data: what is big data visualization?

When it comes to data visualization and big data, there are a few things that you need to know.

First, representing data in a graphical format can be done using various techniques that we have already mentioned above.

Second, **big data** is a term used to describe a large volume of data that can be difficult to manage and analyze. Big data visualization is the process of representing big data in a graphical format to make it easier to understand and work with.

Using big data visualization offers multiple benefits, including the ability to see patterns and trends that would be difficult to spot otherwise. Big data visualization can also help you make better decisions by providing you with a clear picture of what is going on.

Finally, big data visualization can help you communicate complex information more effectively.

Data visualization types: the best-known visual representations of data & data visualization techniques

There are various types of data visualization, each with its own advantages and disadvantages. The most common and well-known visual representations of data are charts and graphs, which are excellent for conveying simple trends or relationships between variables.

However, they can be difficult to interpret if the analyzed data is complex or if there are many data points. In addition, charts and graphs can be misleading if they are not properly designed or used in conjunction with other types of data visualization.

Other common data visualization techniques include maps, timelines, and scatter plots. These techniques are often used to visualize more complex data sets and can be more effective than charts and graphs in certain situations.

Like all tools, data visualization should be used in a way that is appropriate for the specific task at hand. Choosing the right type of data visualization is an important part of effectively communicating information.

How does data visualization work?

Data visualization is the process that transforms raw data into an intelligible, coherent, and easy-to-understand pictorial format. It translates complex numerical information in a way that is easy to understand, interpret and put into storytelling.

Along with its capacity to reveal trends, patterns, and relationships in data, it can also be used to identify outliers and anomalies.

Data visualization techniques

There are countless ways to visualize data, but some techniques are more effective than others. When choosing a data visualization technique, it’s important to consider the type of data you’re working with and the message you want to communicate.

When working with large and complex data sets, it’s often helpful to use multiple data visualization techniques to get a better understanding of the data. For example, you could use a bar chart to compare the overall performance of different groups, but then use a scatter plot to drill down and see how individual values relate to each other.

The best data visualization techniques are those that make the data easy to understand and communicate the desired message effectively.

Data visualization analysis techniques

There are several different data visualization analysis techniques that can be used to gain insights from data. Some of the most common techniques include:

➔ **Data exploration** involves looking at the data to get a better understanding of what it contains. This can be done using various techniques such as visual inspection, summary statistics and data mining

➔ **Data cleaning** is an important step in any data analysis process. It involves identifying and correcting errors in the data, as well as dealing with missing values

➔ **Data transformation** is an additional step in analyzing data, often necessary in order to make it more amenable to analysis. It involves converting the data into a format that is more suitable for the chosen analysis technique

➔ **Statistical analysis** is a broad category of techniques that can be used to summarize the data and/ or uncover relationships between variables. Common statistical analyses include regression, correlation, and cluster analysis

➔ **Visualization** is arguably the most important step in any data analysis process. Data visualization allows us to see patterns in the data that would not be apparent from looking at raw numbers alone

➔ **Communication** is the final step in the data analysis process. It involves conveying the results of the analysis to stakeholders in a clear and concise manner

Data visualization analysis techniques divide into two main categories:

Univariate Analysis Techniques for Data Visualization

There are several univariate analysis techniques that can be used for data visualization. The most common ones are histograms, bar charts, and line graphs.

Histograms are used to visualize the distribution of data. They show how often certain values occur in a dataset. Bar charts are used to compare different values. Line graphs are used to visualize trends over time.

Bivariate Analysis Techniques for Data Visualization

When it comes to data visualization, there are a variety of bivariate analysis techniques that can be used to create impactful and informative visualizations.

Some common techniques include scatter plots, line graphs, and bar charts. Scatter plots are a great way to visualize relationships between two variables.

They can be used to show how one variable changes in relation to another, or to identify clusters and trends in the data. Line graphs are another popular technique for visualizing relationships between variables.

They can be used to show trends over time or to compare multiple variables against each other. Bar charts are a versatile tool for visualizing data. They can be used to compare proportions, or to show changes over time.

No matter which technique you choose, the important thing is that your visualization is clear, accurate and easy to understand. With so many options available, it can be helpful to experiment with different techniques until you find the one that works best for you and your data set.

Benefits and disadvantages of data visualization

There are both benefits and disadvantages to using data visualization as a means of understanding data.

Advantages:

– Data visualization can help you identify patterns and trends that would otherwise not be accessible

– It can also be used to communicate information in a more effective and efficient way than traditional methods such as text-based reports or spreadsheets.

However, there are also some potential drawbacks to using data visualization.

Disadvantages:

– It can be difficult to ensure that the visuals accurately represent the underlying data

– People can interpret visualizations differently, which can lead to misunderstanding or even conflict.

– Data visualizations can be time-consuming and expensive, when not adjusted to your specific needs.

That’s why, at btProvider, our mission always starts with your necessities and resources, before tackling the solutions.

We’re here to support you every step of the way and provide you with the knowledge and expertise, to make the right decisions and shoot for the right data visualization tools and subscriptions.

Good data visualization examples

There are many ways to visualize data, and the best visualization for a particular dataset depends on the nature of the data and the question you want to answer with your visualization. That said, there are some general principles that successful data visualizations follow:

➔ Good data visualization should be easy to understand. It should use simple shapes and colours that are easy to distinguish. The visualization should also use labels and annotations to make it clear what each element represents.

➔ Good data visualization should be visually pleasing. This doesn't mean that it needs to be flashy or have fancy animations, but it should be well-designed and organized in a way that is easy on the eyes.

➔ Good data visualization should tell a story. The best visualizations will help the viewer understand not only the data itself but also the implications of the data. A good visualization will make complex concepts understandable and provide insights that would be difficult to glean from raw data alone.

With these principles in mind, let's look at some specific examples of good data visualizations.

One great example of simple yet effective data visualization is this Index Mundi map of global internet usage.

This map uses colour intensity codes to show which countries have high internet usage (darker shades) and low internet usage (lighter shades). We can also see which countries have seen the biggest growth in internet usage over the past few years. This map is easy to understand and visually pleasing, making it a great example of good data visualization.

Another great example of data visualization is this line chart from NASA showing the global temperature trend from 1881 to the present day.

This chart clearly shows that the Earth's average temperature has been rising steadily over the past years and makes it easy to see how unusual recent years have been in terms of global temperatures. This chart is again easy to understand and visually appealing. It tells a clear story about the data, making it an excellent example of good data visualization.

Data visualization tools: best software instruments that translate unappealing data into catchy, understandable graphs

There are many different data visualization tools available to help you create visual representations of your data. Some of the most popular tools include:

➔ **Tableau** is a powerful data visualization tool that allows you to create interactive, visually appealing charts and graphs

➔ **Excel** is a spreadsheet program that also has some data visualization capabilities. You can use Excel to create basic charts and graphs, as well as more complex visualizations such as heat maps and bubble charts

➔ **Google Sheets** is a free online spreadsheet application with some basic data visualization features. You can use it to create line graphs, bar charts, and pie charts

➔ **R** is a programming language that is often used for statistical analysis and data science. It also has excellent data visualization capabilities, allowing you to create sophisticated graphics

➔ **Python** is another programming language that has powerful data visualization libraries. You can use Python to create a variety of different kinds of graphs and visualizations

➔ **D3.js** is a JavaScript library that can be used to create interactive data visualizations in web browsers

Each of these tools has its own strengths and weaknesses, so it's important to choose the right one for the job at hand.

For instance, bar charts are great for comparing data points side-by-side but can be difficult to read if there are too many data points.

Line graphs, on the other hand, are easy to read and follow, but can be difficult to compare multiple data sets. The best way to choose the right data visualization tool is to experiment with different ones and see which one works best for your needs. Don't be afraid to try new things – you might be surprised at what you come up with!

Open-source visualization tools

Open-source data visualization tools are tools that allow users to create visualizations of data using a variety of methods. These tools can be used to create static or interactive visualizations and can be customized to fit the needs of the user. Some popular open-source data visualization tools include D3.js, Highcharts, and Leaflet. Since we've already spoken about D3.js, let us tell you about the other two:

Highcharts is a JavaScript charting library that allows users to create a variety of charts, including bar charts, line charts, and pie charts. Highcharts also allow users to create interactive visualizations, such as dashboards and maps

Leaflet is an open-source JavaScript library for creating interactive maps. Leaflet allows users to create markers, polygons, and other shapes on their map. Leaflet also allows users to add data layers to their maps, such as weather data or traffic data.

Data visualization best practices

There is no one-size-fits-all answer to the question of what makes for effective data visualization. However, there are some best practices that graphic designers and data visualization experts typically recommend:

– When deciding how to visualize your data, consider both the message you are trying to communicate and your audience. The type of data you are working with will also play a role in determining the best way to visualize it. For example, numerical data is often best represented using charts or graphs, while categorical data may be better suited to a pictorial representation.

– Once you have decided on the type of visualization that will work best for your data, keep it simple. Data visualizations should be easy for viewers to immediately understand. Avoid clutter by only including the most essential elements.

– Use colors effectively: Colors can help direct viewers' attention and highlight important information. Use them judiciously, however, as too many colors can be confusing.

– Choose an appropriate scale: The scale of your visualization should be appropriate for the size of your audience and the amount of detail you need to include.

– Think about layout: The layout of your visualization should guide viewers through the information in a logical way.

– Test and revise: Always test your visualization with a small group of people before finalizing it. Be prepared to make revisions based on feedback from testers.

Data visualization FAQs:

As it is such a complex, interesting and continuously evolving topic, there are always more questions to be asked about data visualization. Here are some of them, as well as their answers:

How does visualizing data improve decision-making?

When it comes to making decisions, seeing is believing.

Data visualization provides a way to see the data, understand it and draw conclusions from it. By visualizing data, decision-makers can identify patterns, trends, and relationships that may not be apparent from raw data sets.

Data visualization also allows decision-makers to spot outliers and exceptions that could impact the final decision. For example, if a data set includes a few outliers, those outliers could skew the results of any analysis.

By visualizing the data, decision-makers can see the outliers and adjust accordingly. Data visualization is an important tool for any decision-maker. It allows them to see the data in a new light and make better-informed decisions.

What is no-code data visualization?

No-code data visualization is a type of data visualization that does not require any coding skills to create. This means that anyone, regardless of their technical skills, can create visually appealing and informative data visualizations. There are several no-code data visualization tools available, such as Google Sheets, **Tableau Public** or Crystal Reports.

These tools allow users to create complex visuals without needing to write any code. No-code data visualization can be an extremely powerful tool for businesses and organizations of all sizes. It can help them to communicate complex information quickly and easily to a wide audience.

What makes data visualization effective?

There are many factors that make data visualization effective.

➔ The ability to see trends and patterns at a glance is one of the most important aspects of data visualization.

➔ When data is presented in a visual format, it can be easier to see relationships and identify outliers.

➔ Data visualization can also help communicate complex ideas in a simple and clear manner.

➔ A good data visualization should be easy to understand and interpret, even for those who are not experts in the field.

➔ The use of colour, layout and other design elements can make a big difference in the effectiveness of data visualization.

➔ Finally, data visualizations should be interactive. Users should be able to explore the data and find the answers they are looking for. Interactive data visualizations allow users to ask their own questions and discover new insights about the data.

Why use data visualization?

In a world where we are constantly inundated with data, it can be hard to make sense of it all.

Data visualization is a way of representing data in a visual way that makes it easier to understand.

There are many benefits to using data visualization, such as:

– Improved understanding of complex data sets

– The ability to identify patterns and trends

– The ability to make better decisions

– Communication of ideas and information

How to evaluate and compare data visualization tools?

When it comes to data visualization, there are a lot of different tools out there. So, how do you know which one is right for you? Here are a few things to keep in mind when evaluating and comparing data visualization tools:

➔ **Ease of use**: Can you easily create the visuals you want with the tool? Is it user-friendly?

➔ **Flexibility**: Can the tool be customized to fit your specific needs?

➔ **Visualization options**: Does the tool offer a variety of ways to visualize data?

➔ **Cost**: Is the tool free or does it come with a subscription fee? Can the subscription fee vary depending on your needs?

Keep these factors in mind when evaluating data visualization tools to find the best one for your needs.

Key takeaways and conclusions

Data visualization is the process of representing data in a graphical or pictorial format. It can be used to communicate complex ideas and relationships between data sets and to reveal patterns and trends that would otherwise be hidden.

There are many different types of data visualizations, each with its own strengths and weaknesses. Some common examples include line graphs, bar charts, scatter plots and pie charts.

When choosing a data visualization tool, it is important to consider the type of data you want to represent, the level of detail you need to show, and the audience you are trying to reach.

Some general tips for creating effective data visualizations include using colours wisely, avoiding clutter, and simplifying your message.

Data visualization can be an extremely powerful way to communicate information. When used correctly, it can help people understand complex ideas quickly and make better decisions. btProvider is the only player in the Romanian Market that provides businesses and organizations with both software infrastructure and consulting services for data visualization. Whatever your data is meant to tell you, we can figure it out together.

The visualization of information is a widely used tool to improve comprehension and, ultimately, decision-making in strategic management decisions as well as in a diverse array of other

domains. Across social science research, many findings have supported this rationale. However, empirical results vary significantly in terms of the variables and mechanisms studied as well as their resulting conclusion. Despite the ubiquity of information visualization with modern software, there is little effort to create a comprehensive understanding of the powers and limitations of its use. The purpose of this article is therefore to review, systematize, and integrate extant research on the effects of information visualization on decision-making and to provide a future research agenda with a particular focus on the context of strategic management decisions. The study shows that information visualization can improve decision quality as well as speed, with more mixed effects on other variables, for instance, decision confidence. Several moderators such as user and task characteristics have been investigated as part of this interaction, along with cognitive aspects as mediating processes. The article presents integrative insights based on research spanning multiple domains across the social and information sciences and provides impulses for prospective applications in the realm of managerial decision-making.

1 Introduction

A visualization is defined as a visual representation of information or concepts designed to effectively communicate the content or message (Padilla et al. 2018) and improve understanding in the audience (Alhadad 2018). This representation can manifest in a range of imagery, from quantitative graphs (Tang et al. 2014) to qualitative diagrams (Yildiz and Boehme 2017), to abstract visual metaphors (Eppler and Aeschimann 2009) or artistic imagery. Visualization design may also intend to promote a specific behavior in the audience (Correll and Gleicher 2014). The visualization of information is associated with effective communication in terms of clarity (Suwa and Tversky 2002), speed (Perdana et al. 2018), and the understanding of complex concepts (Wang et al. 2017). Research shows, for example, that visualized risk data require less cognitive effort in interpretation than textual alternatives and are therefore comprehended more easily (Smerecnik et al. 2010), and complex sentiment data visualized in a scatterplot improve the accuracy in law enforcement decisions compared to raw data (Cassenti et al. 2019).

Visual experiences are the dominant sensory input for cognitive reasoning in everyday life, business, and science (Gooding 2006). As Davis (1986) points out, image creation and perception are part of the “unique and quintessential competencies of homo sapiens sapiens”. Hence, the visualization of information is an integral research subject in the domains of cognitive psychology, education (Alfred and Kraemer 2017), management (Tang et al. 2014) including financial reporting, strategic management, and controlling, marketing (Hutchinson et al. 2010), as well as information science (Correll and Gleicher 2014).

Management researchers study visualizations from a business perspective. First, the field of financial reporting considers the effect of financial graphs on investor perception (Beattie and Jones 2008; Pennington and Tuttle 2009). Second, the potential consequences of visualizations on decision-making are examined in the area of managerial decision support, with a focus on judgments based on quantitative data such as financial decisions (Tang et al. 2014) and performance controlling (Ballard 2020). Finally, a small number of works investigate more complex decision-making based on qualitative, multivariate, and relational information (Platts and Tan 2004). Altogether visualizations fulfill a variety of functions, from focusing attention to sharing thoughts to identifying data structures, trends, and patterns (Platts and Tan 2004).

The vast majority of existing research in visualization, however, arises from the two domains of information science and cognitive psychology. Information science research on how to design visualizations for effective user cognition stretches back almost one century (Washburne 1927). While early research focuses on comparing tables and simple graphs, newer research on human–computer interfaces covers advanced data visualizations facilitated by computing power (Conati et al. 2014). For example, interactive visualization software enables users to manipulate data directly. While promising in terms of analytic capability, the potential for biases and overconfidence is suggested as a downside (Ajayi 2014). Equally, cognitive psychology research notes that visual information may be superior over verbal alternatives in certain cognitive tasks since they can be encoded in their original form, where spatial and relational data is preserved. Thereby, visual input is inherently richer than verbal and symbolic information, which is automatically reductionistic (Meyer 1991), but more suited for discrete information retrieval due to its simplicity (Vessey and Galletta 1991). However, the processes behind visual cognition remain largely unclear (Vila and Gomez 2016).

Despite the ubiquity of visualizations in research and practice, there is no comprehensive understanding of the potential and limits of information visualization for decision-making. Although at times converging, insights from research of different areas are seldom synthesized (Padilla et al. 2018), and there has been no effort for a systematic review or overarching framework (Zabukovec and Jaklič 2015). However, a synthesis of existing research is essential and timely due to three reasons. First, information visualization is ubiquitous both in the scientific and business community, yet there are conflicting findings on its powers and limits in support of judgment and decision-making. Second, cognitive psychology research provides several promising suggestions to explain observable effects of visualizations, yet these are rarely integrated into research in other domains, including strategic decision-making. Third, the barriers to using information visualization software have fallen to a minimum, making it available to a wide range of producers and users. This raises the issue of the validity of positive effects for various task and user configurations. The goal of this paper is therefore to provide an overview of the fragmented existing research on visualizations across the social and information sciences and generate insights and a timely research agenda for its applicability to strategic management decisions.

My study advances visualization research on three paths. First, I establish a framework to summarize the numerous effects and variable interactions surrounding the use of visualizations. Second, I conduct a systematic literature review across the social and information sciences and summarize and discuss this plethora of findings along with the aforementioned structure. Third, I utilize this work as a basis for identifying and debating gaps in existing research and resulting potential avenues for future research, with a focus on the area of strategic management decisions.

The structure of the article is as follows. The next chapter briefly describes the research field, followed by the methodology of my literature search. Next, I analyze the results of my search and discuss common insights. In the ensuing chapter, I develop an agenda for management research by building on particularly relevant ideas with conflicting or incomplete evidence. Finally, I conclude my review and discuss contributions and implications for practice.

2 Definition of the research field

2.1 Definition of key terms

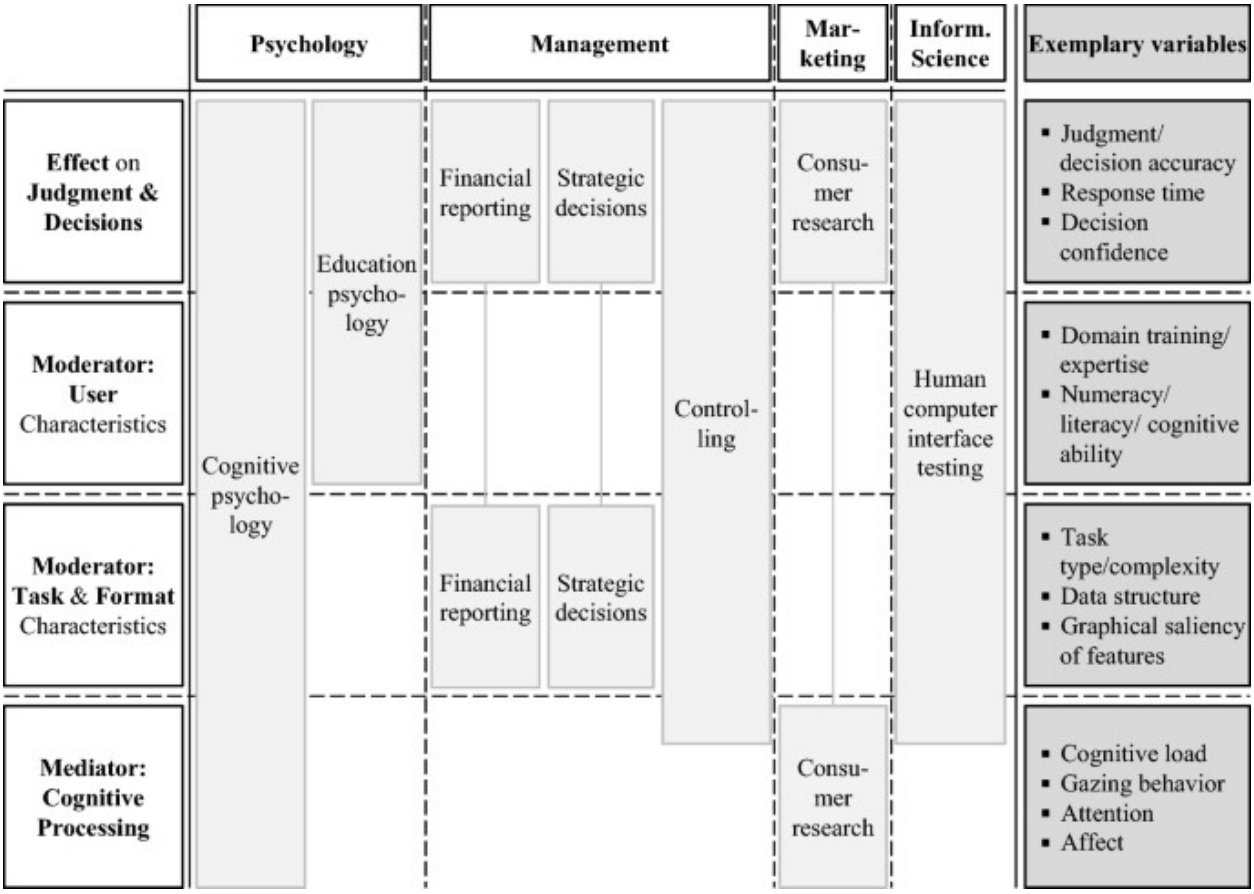
Information visualizations support the exploration, judgment, and communication of ideas and messages (Yildiz and Boehme 2017). The term "graph" is often used as a synonym for information visualization in general (Meyer 1991) as well as describing quantitative data presentation specifically (Washburne 1927). As my review exhibits, these graphs constitute the prevalent form of information visualization. Common quantitative visualizations are line and bar charts, often showcasing a development over time and regularly used in financial reporting (Cardoso et al. 2018) and controlling (Hutchinson et al. 2010). In scientific literature, probabilistic charts such as scatterplots, boxplots, and probability distribution charts (Allen et al. 2014) frequently depict risk and uncertainty. More specialized charts include decision trees to depict conditional logic (Subramanian et al. 1992), radar charts to display complex multivariate information (Peebles 2008), or cluster charts and perceptual maps for marketing decision support (Cornelius et al. 2010).

Despite the breadth of existing visualization research, its application to strategic decisions is narrow and there is an abundance of research limited to elementary tasks and choices. To provide a clear distinction, I focus my search on decisions, judgments, and inferential reasoning as more advanced forms of cognitive processing. Decision-making can be broadly defined as choosing between several alternative courses of action (Padilla et al. 2018). On the other hand, reasoning and judgment refer to the evaluation of a set of alternatives (Reani et al. 2019), without actions necessarily being attached as for decision-making. Such efforts are cognitively demanding and complex when compared to more elementary tasks, such as a choice between options (Tuttle and Kershaw 1998), and include the rigorous evaluation of alternatives across a range of attributes, which is characteristic for strategic decisions (Bajracharya et al. 2014). For this reason, I include studies that examine the influence of visualizations on some form of decision or judgment outcome. Mason and Mitroff (1981) highlight that strategic decisions, in management and elsewhere, involve complex and ambiguous information environments. Information visualization may relate to decision quality in this context since one critical factor in the effectiveness of strategic decisions is the objective and comprehensive acquisition and analysis of relevant information to define and evaluate alternatives (Dean and Sharfman 1996).

2.2 Perspectives in literature

Visualization research exists within a range of domains in the social and information sciences, which reflects the diversity of the empirical application. I identify psychology (cognitive and educational), management (financial reporting, strategic management decisions, and controlling), marketing, and information science as the primary areas of research. This heterogeneity in terms of application area provides the first dimension in my literature review. Second, I classify existing studies along the type of variable interaction they primarily investigate. Based on the framework first introduced by DeSanctis (1984), I hereby differentiate four categories: Works principally focused on (1) the effects of visualizations on comprehension and decisions as dependent variables provide the basis of all research. This relationship is then investigated through: (2) User characteristics as moderators; (3) task and format characteristics as moderators; and (4) cognitive processing as mediator. An overview of this classification, including the prevalence of extant findings across domains, is given in Fig. 1.

Fig. 1



Visualization research structured by domain and variables primarily investigated

Full size image

First, the investigation of visualization *effects on decisions and judgments* is established across all research areas mentioned, and primarily studies outcome variables such as decision accuracy (Sen and Boe 1991), speed (Falschlunger et al. 2015a), and confidence (Correll and Gleicher 2014). While these studies contribute examples for graphs influencing observable decision effectiveness and efficiency across a range of contexts, they do not investigate moderating or mediating factors.

Second, psychology research pushes this investigation further towards including *moderating effects of user characteristics*, such as domain expertise and training (Hegarty 2013), and measures of cognitive ability such as numeracy (Honda et al. 2015) or literacy (Okan et al. 2018a). The relevance of these moderating factors is validated both in studies focusing on cognition as well as experiments in educational research, for example by providing evidence that the quality of a judgment made based on a graph may depend more on the user than the format itself (Mayer and Gallini 1990).

Similarly, human–computer interface research spearheads further insights into *moderating factors of task and format* characteristics, such as task type (Porat et al. 2009), task complexity (Meyer et al. 1997), data structure (Meyer et al. 1999), and the graphical saliency of features (Fabrikant et al. 2010) through rigorous user testing. At the same time, Vessey (1991) developed the theory of cognitive fit as a concept bridging cognitive and information systems research, stating that positive effects of graphs depend on a fit between task type and format type, differentiating between symbolic and spatial archetypes.

Finally, cognitive psychology research aims at explaining the observable effects of visualization in terms of *mediating cognitive mechanisms*. Here, cognitive load theory provides the foundation, stating that an individual's working memory capacity is limited, and performance in a task or judgment depends on the cognitive load they experience while assessing information. According to this logic, cognitive load that is too high damages performance (Chandler and Sweller 1991). Reducing cognitive load by providing visualizations in complex environments is therefore often stated as a key goal of graph design (Smerecnik et al. 2010).

Importantly, the boundaries between these variable categories are fluid. Many studies investigate more than one relationship and the inclusion of moderating variables has become common. Various application areas covering these interdependencies attest to the heterogeneous nature of visualization research. However, previous reviews highlight that insights are seldom shared across fields and call for the integration of findings into new studies (Padilla et al. 2018). In particular, strategic management research does not yet follow such a holistic approach.

3 Method of literature search

3.1 Search design

The methodological basis of this paper is a systematic literature search as a means to collect and evaluate the existing findings in a systematic, transparent, and reproducible way on the specified topic (Fisch and Block 2018) in order to produce a more complete and objective knowledge presentation than in traditional reviews (Clark et al. 2021). I conduct a keyword search on the online search engines EBSCOhost and ProQuest, limited to English-language works that have been peer-reviewed, in order to ensure the quality of the sources. Gusenbauer and Haddaway (2020) identify both search engines as principal academic search systems as they fulfill all essential performance requirements for systematic reviews. On EBSCOhost, I use the databases *Business Source Premier*, *Education Research Complete*, *EconLit*, *APA PsycInfo*, *APA PsycArticles*, and *OpenDissertations* to search for *empirical* works; on ProQuest, I use the databases *British Periodicals*, *International Bibliography of the Social Sciences (IBSS)*, *Periodicals Archive Online*, and *Periodicals Index Online* with a filter on *articles* to cover the social sciences comprehensively. The keyword used is the concatenated term "(visualization OR graph OR chart) AND (decision OR judgment OR reasoning)", searched for in abstracts.[Footnote 1] The terms were chosen as "visualization" is commonly used as a category name for visualized information (Brodlie et al. 2012), and the "graph" is the focus of traditional visualization research (Vessey 1991). The term "chart" is a synonym for both quantitative and qualitative graphs which has seen increasing use particularly in the 2000s (Semmler and Brewer 2002). The terms "judgment OR decision OR reasoning" were added to ensure that studies examining

observable outcomes of visualization use, as opposed to cognitive processes such as comprehension only, were highlighted. After a review of the evolution of visualization research over time, I focus my search to articles published from the year 1990 in order to capture the recent advancements covering modern modes of information visualization.[Footnote 2] This search results in 1658 articles combined, after removing duplicates 1505 articles remain.

Next, I review all article abstracts based on the three content criteria defined in the following. I include all articles rooted in the (1) *social sciences or information sciences*, where the focus of the study lies on (2) *how a visualization* per se *or a variation within related visualizations affects a user's or audience's decision or judgment in a given task*, and the topic is studied through (3) *original empirical works.* Most articles are excluded in this process and 116 studies remain due to the prevalence of graphs as auxiliary means, not the subject of research, in various domains, particularly in medical research. I repeat this exclusion process by reading the full texts of all articles and narrow down the selection further to 81 papers.

Building on this systematic search, I conducted a supplementary search through citation and reference tracking, as well as supplementary search engines, such as JSTOR (Gusenbauer and Haddaway 2020).[Footnote 3] This includes gray literature such as conference proceedings or dissertations, which lie outside of traditional academic publishing. In addition, I limit the inclusion of gray literature to studies by researchers included in my systematic search and completed within the last 10 years in order to gather a comprehensive and up-to-date overview of the findings of working groups particularly relevant to visualization research. Thereby I identify 52 additional articles, resulting in a total of 133 articles included.

3.2 Limitations of search

Due to the plethora of existing literature mentioning the topic of visualization in various contexts and degrees of quality, I subject my search to well-defined limitations. First, I only include peer-reviewed articles in my systematic search. These are studies that have been thoroughly validated and represent the major theories within a field (Podsakoff et al. 2005). However, I incorporate gray literature of comparable quality as part of my additional exploratory search.

Second, I limit the search to information and social sciences to deliberately omit results from the broad areas of medicine and natural sciences. In these, various specific concepts are visualized as a means within research, yet not investigating the visualization itself. For the same reason, I only apply the search terms to article abstracts, since the terms “graph” and “chart” in particular will result in a high number of results when searched for in the full text, due to the common use of graphs in presenting concepts and results.

Third, I only include original empirical work in order to enable the synthesis and critical validation of empirical findings across research areas. At the same time, I acknowledge the existence of several highly relevant theoretical works, which inform my search design and structure while being excluded from the systematic literature search and analysis.

4 Results

4.1 Overview of results

I identify a total of 133 articles, published between 1990 and 2020. Interest in visualization research gained initial momentum in the early 1990s (Fig. 2). More recently, the number of studies rises starting around 2008, with the continued publication of five to ten papers per year since and a visible peak in interest around 2014/15. A significant share of recent works stems from the information science literature, and the wealth of publications around 2014 coincides with the advent of mainstream interest in big data (Arunachalam et al. 2018), which is closely linked to information visualization for subsequent analysis and decision-making (Keahey 2013). In addition, a cluster of publications by one group of authors (Falschlunger et al. 2014, 2015a, c, b) in the financial reporting domain enhances the observed peak in publications, which is therefore not indicative of a larger trend. Instead, the continued wealth of publications in the last decade shows the contemporary relevance of and interest in visualization research.

Effects of visualizations on decisions and judgments

4.2.1 Judgment/decision accuracy

The most common dependent variable investigated in visualization research is the *accuracy* of the subjects on a given comprehension, judgment, or decision task. Most studies are in psychology research, with positive effects dominating. In cognitive psychology, experiments show that well-designed visualizations can improve problem comprehension (Chandler and Sweller 1991; Huang and Eades 2005; Nadav-Greenberg et al. 2008; Okan et al. 2018b). For example, Dong and Hayes (2012) show in their experiment with 22 practitioners that a decision support system visualizing uncertainty improves the identification and understanding of ambiguous decision situations. Likewise, visualizations improve decision (Pfaff et al. 2013) and judgment accuracy (Semmler and Brewer 2002; Tak et al. 2015; Wu et al. 2017) and improve the quality of inferences made from data (Sato et al. 2019). Findings in educational psychology support this claim. In teaching, visual materials improve understanding and retention (Dori and Belcher 2005; Brusilovsky et al. 2010; Binder et al. 2015; Chen et al. 2018) in students, and support the judgment accuracy of educators when analyzing learning progress quantitatively (Lefebre et al. 2008; Van Norman et al. 2013; Géryk 2017; Nelson et al. 2017). Furthermore, Yoon's longitudinal classroom intervention (2011) using social network graphs enables students to make more reflected and information-driven strategic decisions. However, other studies arrive at more mixed or opposing findings. In their experiment, Rebotier et al. (2003) find that visual cues do not improve judgment accuracy over verbal cues in imagery processing. Other experiments even demonstrate verbal information to be superior over graphs in comprehension (Parrott et al. 2005) as well as judgment accuracy (Sanfey and Hastie 1998). Some graphs appear unsuitable for specific content, such as bar graphs depicting probabilities (Newman and Scholl 2012) and bubble charts encoding information in circle area size (Raidvee et al. 2020). In addition, more complex charts like boxplots, histograms (Lem et al. 2013), and tree charts (Bruckmaier et al. 2019) appear less effective for the accurate interpretation of statistical data in some experiments, presumably as they elicit errors and confusion in insufficiently trained students.

Studies in management and business research arrive at further, more pessimistic results. While Dull and Tegarden (1999) find in their experiment with students that three-dimensional visuals can improve the prediction accuracy in financial reporting contexts, and Yildiz and Boehme (2017) observe in their practitioner survey that a graphical model of a corporate security decision problem improves risk perception when compared to a textual description, most other studies present a less positive picture. Several studies do not find graphs superior over tables in financial judgments (Chan 2001; Tang et al. 2014; Volkov and Laing 2012), and in consumer research (Artacho-Ramírez et al. 2008). In financial reporting, a dedicated school of research investigates the effect of distorted graphs lowering financial judgment accuracy (Arunachalam et al. 2002; Beattie and Jones 2002a, b; Amer 2005; Xu 2005; Pennington and Tuttle 2009; Falschlunger et al. 2014), irrespective of whether the distortion is intended by the designer. Chandar et al. (2012) elaborate on the positive effect of the introduction of graphs and statistics in performance management for AT&T in the 1920s, but more recent case study examples are rare.

By contrast, several experimental studies from human–computer interaction research largely contribute evidence for a positive effect. Targeted visual designs lead to higher judgment accuracy in specific tasks (Subramanian et al. 1992; Butavicius and Lee 2007; Van der Linden et al. 2014; Perdana et al. 2018) and improve decision-making (Peng et al. 2019). For example, probabilistic gradient plots and violin plots enable higher accuracy in statistical inference judgments in the online experiment by Correll and Gleicher (2014) than traditional bar charts. However, experiments by Sen and Boe (1991) and Hutchinson et al. (2010) equally lack a significant effect on data-based decision-making quality. Amer and Ravindran (2010) find a potential for visual illusions degrading judgment accuracy similar to results from financial reporting, and McBride and Caldara (2013) find that visuals lower accuracy in law enforcement judgments when compared to raw data presentation (Table 2).

esponse time

The next most common outcome variable investigated in visualization research is *response time*, often referred to as efficiency. Across the board, experimenters observe that information visualization lowers response time in various judgment and decision tasks. In psychology, this includes decision-making in complex information environments (Sun et al. 2016; Géryk 2017). The opposite effect emerges from only one study, where Pfaff et al. (2013) find that a decision support system visualizing complex uncertainty information requires a longer time to use than one omitting this graphical information. In management research, Falschlunger et al. (2015a) find that visually optimized financial reports can speed up judgment both for students and practitioners. Studies originating in information science validate this picture, observing that well-designed visualizations reduce response time in quantitative

The earliest form of data visualization can be traced back the Egyptians in the pre-17th century, largely used to assist in navigation. As time progressed, people leveraged data visualizations for broader applications, such as in economic, social, health disciplines. Perhaps most notably, Edward Tufte published The Visual Display of Quantitative Information (link resides outside IBM), which illustrated that individuals could utilize data visualization to present data in a more effective manner. His book continues to stand the test of time, especially as companies turn to dashboards to report their performance metrics in real-time. Dashboards are effective data

visualization tools for tracking and visualizing data from multiple data sources, providing visibility into the effects of specific behaviors by a team or an adjacent one on performance. Dashboards include common visualization techniques, such as:

- **Tables:** This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.

- **Pie charts and stacked bar charts:** These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.

- **Line charts and area charts:** These visuals show change in one or more quantities by plotting a series of data points over time and are frequently used within predictive analytics. Line graphs utilize lines to demonstrate these changes while area charts connect data points with line segments, stacking variables on top of one another and using color to distinguish between variables.

- **Histograms:** This graph plots a distribution of numbers using a bar chart (with no spaces between the bars), representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.

- **Scatter plots:** These visuals are beneficial in reveling the relationship between two variables, and they are commonly used within regression data analysis. However, these can sometimes be confused with bubble charts, which are used to visualize three variables via the x-axis, the y-axis, and the size of the bubble.

- **Heat maps:** These graphical representation displays are helpful in visualizing behavioral data by location. This can be a location on a map, or even a webpage.

- **Tree maps,** which display hierarchical data as a set of nested shapes, typically rectangles. Treemaps are great for comparing the proportions between categories via their area size.

Access to data visualization tools has never been easier. Open source libraries, such as D3.js, provide a way for analysts to present data in an interactive way, allowing them to engage a broader audience with new data. Some of the most popular open source visualization libraries include:

- **D3.js:** It is a front-end JavaScript library for producing dynamic, interactive data visualizations in web browsers. D3.js (link resides outside IBM) uses HTML, CSS, and SVG to create visual representations of data that can be viewed on any browser. It also provides features for interactions and animations.

- **ECharts:** A powerful charting and visualization library that offers an easy way to add intuitive, interactive, and highly customizable charts to products, research papers,

presentations, etc. Echarts (link resides outside IBM) is based in JavaScript and ZRender, a lightweight canvas library.

- **Vega:** Vega (link resides outside IBM) defines itself as "visualization grammar," providing support to customize visualizations across large datasets which are accessible from the web.

- **deck.gl:** It is part of Uber's open source visualization framework suite. deck.gl (link resides outside IBM) is a framework, which is used for exploratory data analysis on big data. It helps build high-performance GPU-powered visualization on the web.
- With so many data visualization tools readily available, there has also been a rise in ineffective information visualization. Visual communication should be simple and deliberate to ensure that your data visualization helps your target audience arrive at your intended insight or conclusion. The following best practices can help ensure your data visualization is useful and clear:
- Set the context: It's important to provide general background information to ground the audience around why this particular data point is important. For example, if e-mail open rates were underperforming, we may want to illustrate how a company's open rate compares to the overall industry, demonstrating that the company has a problem within this marketing channel. To drive an action, the audience needs to understand how current performance compares to something tangible, like a goal, benchmark, or other key performance indicators (KPIs).
- **Know your audience(s):** Think about who your visualization is designed for and then make sure your data visualization fits their needs. What is that person trying to accomplish? What kind of questions do they care about? Does your visualization address their concerns? You'll want the data that you provide to motivate people to act within their scope of their role. If you're unsure if the visualization is clear, present it to one or two people within your target audience to get feedback, allowing you to make additional edits prior to a large presentation.
- **Choose an effective visual:** Specific visuals are designed for specific types of datasets. For instance, scatter plots display the relationship between two variables well, while line graphs display time series data well. Ensure that the visual actually assists the audience in understanding your main takeaway. Misalignment of charts and data can result in the opposite, confusing your audience further versus providing clarity.
- **Keep it simple:** Data visualization tools can make it easy to add all sorts of information to your visual. However, just because you can, it doesn't mean that you should! In data visualization, you want to be very deliberate about the additional information that you add to focus user attention. For example, do you need data labels on every bar in your bar chart? Perhaps you only need one or two to help illustrate your point. Do you need a variety of colors to communicate your idea? Are you using colors that are accessible to a wide range of audiences (e.g. accounting for color blind audiences)? Design your data visualization for maximum impact by eliminating information that may distract your target audience.
- Effectively designed data visualizations allow viewers to use their powerful visual systems to understand patterns in data across science, education, health, and public policy. But ineffectively designed visualizations can cause confusion, misunderstanding, or even distrust—especially

among viewers with low graphical literacy. We review research-backed guidelines for creating effective and intuitive visualizations oriented toward communicating data to students, coworkers, and the general public. We describe how the visual system can quickly extract broad statistics from a display, whereas poorly designed displays can lead to misperceptions and illusions. Extracting global statistics is fast, but comparing between subsets of values is slow. Effective graphics avoid taxing working memory, guide attention, and respect familiar conventions. Data visualizations can play a critical role in teaching and communication, provided that designers tailor those visualizations to their audience.

- This report presents research-backed guidelines for creating powerful and intuitive visualizations oriented toward communicating data to students, coworkers, and the general public. We begin by reviewing guidelines for helping viewers extract data from visualizations in precise and unbiased ways, avoiding a set of known illusions and distortions. We then describe when visual processing of visualizations is powerful (processing broad statistics) versus where it slows to a crawl (making individual comparisons), and we provide a tool kit for avoiding that slowdown. We review guidelines for ensuring that a viewer properly maps visualized values to the right concepts in the world (e.g., viewers can extract the size of an error bar on a graph, but do they understand what it means?), allowing viewers to use visualizations as effective tools for reasoning. We then review guidelines for conveying uncertainty and risk (e.g., how could a physician express survival odds for a treatment to a patient?). Finally, we summarize a set of guidelines for creating visualizations that communicate clearly and suggest resources for readers interested in learning more.
- Data visualizations range from simple graphs in elementary school classrooms, to depictions of uncertainty in election forecasts in news media, to complex data displays used by scientists and analysts. When designed effectively, these displays leverage the human visual system's massive processing power, allowing rapid foraging through patterns in data and intuitive communication of those patterns to other viewers. But when designed ineffectively, these displays leave critical patterns opaque or leave viewers confused about how to navigate unfamiliar displays.
- We review methods, empirical findings, theories, and prescriptions across the fields of visual perception, graph comprehension, information visualization, data-based reasoning, uncertainty representation, and health risk communication. These research communities study similar questions and use complementary expertise and styles of inquiry, yet they too rarely connect. We ignore artificial boundaries among these research fields, and instead integrate across them.

The Importance of Visualization Design and Literacy

- Thinking and communicating with data visualizations is critical for an educated public (Börner et al., 2019). Science education standards require students to use visualizations to understand relationships, to reason about scientific models, and to communicate data to others (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010; National Research Council, 2013). Evidence-based public policy prescriptions about climate change, vaccines, and policing are argued to be most effectively built (Kohlhammer et al., 2012) and communicated to the public (Otten et al., 2015) with visualizations. Journalists at *The New*

York Times Upshot, FiveThirtyEight, *The Economist*, and *The Washington Post* use visualizations to communicate data and evidence about statistics and policy. Data visualizations are ubiquitous in the workplace—in data-analysis software, in data-overview dashboards, and in millions of slide presentations created each day (Berinato, 2016; Parker, 2001). Physicians rely on them to show data about the risks of medical procedures, and meteorologists use them to show the uncertainty in a hurricane's potential path (Ancker et al., 2006; Ruginski et al., 2016).

- In each of these domains, low graphical literacy and ineffective design lead many viewers to struggle to understand these otherwise powerful thinking tools. Many students can find textbook visualizations too challenging to understand or integrate with nearby text (Nistal et al., 2009; Shah & Hoeffner, 2002). Public policy visualizations can be counterintuitively designed, leading many viewers to draw a conclusion opposite the one suggested by the depicted data (Engel, 2014). Dozens of best-selling guides have decried the state of visualizations in the workplace and offered prescriptions for more powerful, clear, and persuasive graphs (see the Recommended Practitioner Books section at the end of this article and Ajani et al., 2021, for a more exhaustive list). Medical-risk visualizations can lead patients to fundamentally misunderstand the base rates or risk factors for diseases or medical procedures (Ancker et al., 2006). When a prediction has a high level of uncertainty that is not intuitively conveyed, the public can lose trust in scientists and analysts. For example, when a hurricane's path deviates somewhat from the most likely trajectory, or when a politician with a 20% predicted chance to win an election prevails, these outcomes may be consistent with the uncertainty inherent to the predictions. But if the forecaster does not effectively visually communicate that uncertainty, their reputation can suffer when their prediction is "wrong" (Padilla et al., 2021).

Who Studies the Design and Comprehension of Visualizations?

- Research on the design and pedagogy of data visualizations takes place in several communities. A psychologist focusing on perception might study the mapping between a color value in a heat map and the abstract magnitude that an observer extracts from it (Stevens, 1957). A cognitive psychologist might explore how working memory limits the complexity of the statistical relationships that a viewer might extract (Halford et al., 2007; Padilla et al., 2018). An education researcher might try to remove roadblocks for students struggling to translate visual depictions to their underlying concepts (Börner & Polley, 2014; Shah & Hoeffner, 2002) or seek multimedia design principles for designing effective graphics and integrating them with text (e.g., Mayer & Fiorella, in press). Researchers in public policy communication or political science might study why viewers find some visualizations to be more trustworthy or persuasive than others (Nyhan & Reifler, 2019). Health communication researchers evaluate how to effectively communicate the risk of a medical procedure to patients with low *numeracy* (i.e., ability to work with numbers and mathematics; Ancker et al., 2006). Specialists in statistical cognition and communication seek ways to communicate uncertainty across election outcomes, bus arrival times, and hurricane paths (Hullman, 2019). Finally, a research community housed in computer and information sciences studies data visualization at multiple levels, from data types and algorithms to the

creation of user task taxonomies, to design prescriptions for visually powerful displays and fluid interaction (Munzner, 2014).

- In this article, we also draw advice from communities of practitioners who might not engage in empirical research but use extensive in-context experience to generate prescriptions for powerful and intuitive visualizations. At the end of this review, we include a list of recommended visualization-design guidebooks. Although many of these guides are oriented toward business analysts, their prescriptions extrapolate directly to science, education, and public policy visualizations. We also discuss design techniques used by a new wave of journalists focused on communicating data analysis to the general public.

The Structure of Our Review

- This review focuses on how to effectively design visualizations that communicate data to students and the general public. We review evidence-based prescriptions for designing visualizations that help people understand and reason about the patterns, models, and uncertainties carried by a data set. Another important topic, which we do not cover systematically here, is how to measure visualization literacy and the effectiveness of teaching techniques that improve it (see Börner et al., 2019; Lee et al., 2016). We also restrict our scope to quantitative visualizations, omitting discussion of qualitative visualizations of text data, diagrams, and processes (see Hegarty, 2011; Henderson & Segal, 2013, for review). We focus on research and prescriptions that are most relevant for communication to nonspecialist audiences, instead of the design of powerful tools for data analysis within expert communities.
- We first illustrate why visualizations can be such powerful tools for thinking about data. Because the human visual system is highly developed for rapid parallel extraction of behavior-relevant features and relationships, visualizations allow us to process some types of patterns across an entire two-dimensional array of values at once. We describe the limited set of visual channels that can effectively depict magnitudes to a viewer, such as the *position* of a value in a dot plot, the *size* of a circle hovering over a map, or the *color intensity* of an activation pattern in a functional MRI (fMRI) image.
- We then discuss design guidelines for ensuring that the human eye accurately decodes those depicted values. We review evidence for a ranking of some visual channels (e.g., position) as more precise than others (e.g., color intensity) for at least one common task but also discuss how new work has begun to dismantle that ranking for a broader array of tasks. We list a set of common errors and illusions that cause viewers to extract the underlying values from visual channels incorrectly—for example, *y*-axis manipulations that exaggerate differences among values, confusion about whether circles depict values with their size or diameter (which can change the extracted value by an order of magnitude), a common illusion produced by line graphs, and other illusions and categorical distortions that can arise when depicting value with color intensity. We also include a brief review of accessibility considerations for viewers with color blindness. Finally, we discuss best practices for distinguishing between groups of data (say, two groups of points on a scatterplot) by marking them with different shapes or colors.

- Next, we discuss an important dissociation in visual processing power: Whereas computing statistics across an image is broad and instantaneous, making comparisons among subsets of values is slow and limited to two or three comparisons per second. We review the types of grouping cues that loosely control what information is compared by a typical viewer and further techniques for precisely guiding a viewer to the right comparison. We discuss the importance of respecting a viewer's limited working memory, including avoiding legends and animated displays that can engage but also confuse. Finally, we review evaluations of whether visualizations should have rich and memorable designs, as opposed to a minimalist and clean aesthetic.
- The next section introduces visualization *schemas*, or knowledge structures that include default expectations, rules, and associations that a viewer uses to extract conceptual information from a data visualization. We illustrate the importance of schemas by introducing the reader to a small set of new visualization designs that will likely be unfamiliar. We then provide examples of common schema elements that are known to more graphically literate audiences (but not always respected by designers), such as the assumption that larger values are plotted upward. We shift to a brief review of human reasoning about visualizations, including formal models that draw links from visual depictions, to numeric values, to their underlying concepts and the designer's intended message. We then explore two case studies: reasoning about graphs illustrating scientific concepts and reasoning about graphs of mathematical functions.
- The subsequent sections review research on visualizing uncertainty or risk. Communication failures can start with a lack of understanding of critical statistical concepts, even among scientists. We give examples of how viewers tend to misread error bars as depicting the edges of a range of data instead of correctly understanding them as parameters of a distribution. Probability information expressed as risk is critical for people such as patients considering a medical procedure and potential evacuees who may be in the path of a hurricane, but depictions of risk are frequently misunderstood. We review guidelines for showing uncertainty or risk more intuitively, including depicting samples of discrete outcomes, showing probability density functions, and depicting data with arrays of icons.

The Power of Visualization

Visualizations let viewers see beyond summary statistics

Visualizations allow powerful processing of an entire two-dimensional rectangle of information at once, in stark contrast to the limitation of reading handfuls of symbolic numbers per second. As a demonstration, Figure 1 (top left) contains four sets of 11 pairs of values. Take a moment to compare those columns, and notice that reading symbolically represented numbers takes time. As you seek patterns within each set, or make comparisons among the four sets, progressively processing more pairs of values becomes increasingly difficult. Worse, these tasks quickly exhaust memory capacity, such that new numbers or patterns tend to displace ones that were previously seen. These limitations on symbolic processing of numbers lead viewers to rely instead on summary statistics that compress data sets into a single group of numbers. For the four sets of numbers in Figure 1, those statistics on the bottom row—means, standard deviations, and correlation coefficients—are identical, which might lead you to believe

that the numbers contributing to the statistics are similar (Anscombe, 1973). However, because statistics summarize larger sets of numbers by abstracting over them and making assumptions about the patterns that they might contain, many sets of numbers can generate the same statistics. For these four sets of numbers, relying on statistics turns out to be dangerous.

Visual channels translate numbers into images

Visualizations rely on several visual channels to transform numbers into images that the visual system can efficiently process (Bertin, 1983; Mackinlay, 1986; see Munzner, 2014, for a more complete list). Knowing these channels allows a designer to consider which might be best suited for a given data set and context—particularly given that each is associated with differential levels of precision and potential illusions, as described in the following sections. The first column of Figure 2 depicts five of the more frequently used channels. Dot plots and scatterplots, such as those in Figure 1, represent values as *position*. Bar graphs represent values not only as positions (of the tips of the bars) but also as one-dimensional *lengths* (and, some argue, even two-dimensional *areas*; Yuan et al., 2019). If two bars do not rest on the same baseline, such as segments within the same bar in a stacked bar graph, the comparison relies only on length or area. Next, two circles code numbers exclusively as two-dimensional areas (typically circles), a technique often used to overlay values across maps. *Angle* typically emerges when points are connected to form a line graph, organically allowing an encoding of the difference between adjacent points (a bigger difference creates a steeper slope and, typically, a longer line). Outside of pie charts, angle is less frequently used to depict numbers directly—perhaps on local areas of a map to represent wind directions. *Numerosity* is omitted from the figure, but it often implicitly shows higher-level attributes of data. For example, you can immediately estimate the number of points in a scatterplot, segments in a stacked bar chart, or icons in an infographic. *Intensity* is an umbrella term (often also called *lightness* or *value*) for either luminance contrast or color saturation, as used in a heat map or fMRI activation map. *Motion* is also not included in the figure, but animating a scatterplot to show values changing over time can encode the rate and direction of change in the speed and direction of the dots' motion.

How to Design a Perceptually Accurate Visualization

Understand how to leverage visual channels

Visual channels are ranked by their perceptual accuracy

These channels differ in how precisely they convey numeric values to a viewer, and knowing the ranking of these channels allows a designer to prioritize what information to show most precisely. The leftmost column of Figure 2 presents five of the channels that can depict metric data to the human visual system. This list is ordered by the typical precision with which a viewer can verbally state the ratios between the two values shown; more precise ways of communicating numbers are at the top and less precise ways are at the bottom (Cleveland & McGill, 1984, 1985; Heer & Bostock, 2010). It should be clear from the figure that the 1:7 ratio can be relatively precisely extracted for position, but that task is a bit tougher for area, and far more difficult for intensity, at the bottom of the list.

Because position is the clear winner for precision, visualization designers often prioritize the vertical and horizontal dimensions of two-dimensional space when depicting or organizing quantitative data. Faced with a single column of numbers in a spreadsheet, a visualization designer might depict those data vertically with position (in a bar or line graph) and rely on horizontal position to organize the values into categories, as in a typical bar chart. If faced with two columns of numbers, a designer might simply create two of those same types of graphs or organize each set of numbers along both the vertical and horizontal axes of position, as in a scatterplot.

The advantage of position over length for precisely depicting ratios between numbers is demonstrated in the second column of Figure 2, which shows a horizontally oriented stacked bar graph in its second row of examples. Because the black segments of the bars are aligned on a common axis at their left, their right tips provide a precise position code, allowing the viewer to see the delicate 0.9:1 ratio between the bars. However, the next set of medium-gray segments are tougher to distinguish because the positions of their right tips are no longer useful, so the viewer must rely on length—a lower-precision channel—to extract the same ratio.

Mapping visual ratios back to numbers can cause perceptual errors

Using visualizations can unlock powerful data-pattern processing. However, a designer must be aware of several perceptual illusions that can lead viewers to map visual depictions back to their original numeric values incorrectly (Huff, 1954; Tufte, 1983). If two plotted values have a 1:7 ratio, then the visualization should cause a typical viewer to see that 1:7 ratio veridically. Even for a precise visual channel such as position, this requirement can be tougher than anticipated. For example, see the dot plot and bar graph at the top of the second column of Figure 2. The dot plot uses position as its visual channel, and the bar graph depicts the same data with both position and length. The second value appears to be roughly double the first value. Look more closely at the *y*-axis: The second value is only about 1% bigger than the first; the difference appears greater because the axis baseline does not start at zero. In theory, the data are transparently depicted—but in reality, such graphs are frequently misinterpreted.

Figure 3 illustrates some real-world examples of this problem. The line graphs in the upper left, adapted from Darrell Huff's classic 1954 book *How to Lie With Statistics*, show how a line graph's scale can be stretched to make a trend appear steeper (Huff, 1954). In March 2014, a version of the bar graph at the upper right appeared on *Fox News*, a network with an avowedly opposite political orientation to Barack Obama, the U.S. president at the time. Around 6 million U.S. citizens had signed up for a new health-care program sponsored by the president, and the government specified a goal of 7 million sign-ups by March 31. Although the numbers presented are honest (a 6:7 ratio), the visualization's truncation of the *y*-axis tells a different story (a 1:3 ratio) to the viewer's visual system, suggesting a failure of the president's plan.

Why is data visualization important?

Data visualization is important for a variety of reasons, as it offers numerous benefits in various fields and industries. Here are some key reasons why data visualization is important:

- **Clarity and understanding**: Data visualization simplifies complex data, making it easier to understand. Visual representations such as charts and graphs enable individuals to quickly grasp the meaning of data, even if they aren't data experts.
- **Data-driven decision making**: In business and other fields, data visualization is essential for informed decision-making. It helps leaders and analysts make strategic choices based on data insights.
- **Time-saving**: Reading and analyzing raw data can be time-consuming. Data visualization accelerates this process, enabling quicker data analysis and decision-making.
- **Error detection**: Visual representations make it easier to detect errors or inconsistencies in data. When data is visualized, outliers or inaccuracies are often more apparent.
- **Exploration and interactivity**: Interactive data visualizations allow users to explore data and dive deeper into specific aspects. This fosters a more profound understanding and discovery of insights.
- **Forecasting and planning**: Visualizations can help with forecasting future trends and planning strategies. For example, financial analysts use visualizations to predict market trends, while urban planners use them to design efficient city layouts.
- **Real-time monitoring**: In today's fast-paced world, real-time data visualization is vital. It allows for continuous monitoring of critical metrics and rapid responses to changing situations.
- **Enhanced storytelling**: Data visualizations can turn data into compelling stories. They help individuals communicate their data-driven narratives effectively.

In summary, data visualization is a powerful tool for simplifying data. It plays a pivotal role in data analysis, decision-making, and effective communication in various domains. Its importance continues to grow in our data-driven world, and it is an essential skill for professionals and businesses aiming to leverage the potential of data.

What are the different types of data visualization?

Data visualization comes in a wide variety of forms, each tailored to specific data types, objectives, and audiences. Here are some of the common types of data visualizations:

1. **Bar charts**: Bar charts represent data as rectangular bars of varying lengths, making them suitable for comparing discrete categories or showing changes over time. They can be vertical (column charts) or horizontal (bar charts).
2. **Line charts**: Line charts display data as a series of data points connected by lines, making them ideal for showing trends over time. They're often used in time-series analysis.
3. **Pie charts**: Pie charts represent parts of a whole, with each "slice" of the pie representing a percentage or proportion of the total. They're useful for displaying how a category contributes to a total.
4. **Scatter plots**: Scatter plots use a grid to display individual data points with two numeric variables, making it easy to identify relationships and correlations between them.
5. **Heatmaps**: Heatmaps use colors to represent data values, often in a two-dimensional matrix. They're ideal for visualizing data density and patterns in large datasets.
6. **Histograms**: Histograms display the distribution of a single numeric variable by grouping data into bins or intervals. They're valuable for understanding the frequency of data values.
7. **Area charts**: Area charts are similar to line charts but use filled areas between the lines and the baseline to represent quantities over time. They're suitable for visualizing cumulative data.

8. **Bubble charts**: Bubble charts extend scatter plots by introducing a third dimension using varying sizes of bubbles. They're useful for comparing three variables simultaneously.
9. **Choropleth maps**: Choropleth maps use shading or color-coding of geographic regions to represent data values. They're often used to show regional variations in data, such as population density or economic indicators.
10. **Tree maps**: Tree maps divide data into hierarchical rectangles, making it possible to show the proportion of each category within larger categories. They're commonly used for visualizing organizational structures and hierarchical data.
11. **Sankey diagrams**: Sankey diagrams depict the flow of resources, energy, or data between various entities. They're particularly useful for illustrating processes and resource allocation.
12. **Word clouds**: Word clouds display text data with words sized proportionally to their frequency. They're excellent for visualizing word frequencies in text documents or sentiment analysis.
13. **Radar charts**: Radar charts use a circular layout with spokes to represent data values for multiple categories. They're useful for comparing entities across multiple attributes.
14. **Gantt charts**: Gantt charts display tasks or activities along a timeline, allowing for project management and scheduling visualization.
15. **Box plots**: Box plots provide a summary of a data distribution, showing the median, quartiles, and potential outliers, making them ideal for understanding data variability.
16. **Violin plots**: Violin plots combine a box plot with a kernel density plot, providing a more detailed view of data distribution than a box plot alone.

These are just a selection of the many types of data visualizations available. The visualization that you would want to choose depends on the nature of the data and the insights you want to convey. Selecting the right type of visualization is critical to effectively communicate your data's story.

What are the limitations of data visualization?

Data visualization, while a powerful tool, has its limitations that need to be acknowledged and carefully considered. One of the primary concerns is over-simplification. Visualizations often condense complex data into easily digestible forms, but this can lead to the loss of finer details and nuances in the data, potentially affecting the accuracy of the insights. Additionally, misrepresentation is a risk when using data visualization. Poorly designed visuals, whether intentionally or unintentionally, can skew the perception of data.

Another challenge is the limited context provided by visualizations. Without a complete understanding of the data sources, viewers may misinterpret the visualized data. The quality of this data is also a critical concern. The accuracy and reliability of data are fundamental, and visualizations can't rectify issues with poor-quality or incomplete data. Visualizations can become overwhelmingly complex, especially when dealing with large datasets.

A common pitfall is the overemphasis on aesthetics. While aesthetics can enhance engagement, they should not overshadow the core message and data accuracy. Proper design practices can help maintain this balance. Finally, data visualization primarily highlights correlations but doesn't necessarily unveil causation. Discerning cause-and-effect relationships often requires additional analysis and context.

Understanding these limitations is essential for using data visualization effectively and ensuring that the insights derived from visualizations are reliable and accurate.

Are there any best practices for data visualization?

There are several best practices for data visualization that can help you create clear, effective, and informative visualizations. Here are some key best practices:

- **Knowing your audience**: Understand who will be viewing your visualizations and what they are looking to gain from the data. Tailor your visualizations to meet their needs and knowledge levels.
- **Choose the right chart type**: Select the appropriate chart or graph type that best represents your data and the insights you want to convey. Each chart type is best suited for specific data and messaging.
- **Use consistent design**: Maintain consistency in color schemes, fonts, and styles throughout your visualizations to create a cohesive and professional look.
- **Label clearly**: Ensure that your charts and graphs have clear, descriptive labels for axes, data points, and any relevant components. Use a legible font size.
- **Avoid 3D effects**: Minimize or eliminate 3D effects and unnecessary embellishments. They can distort perception and make your visualizations harder to read and understand.
- **Show data proportions**: Use proper scaling to accurately represent data. Starting axes at zero is crucial to prevent misinterpretation.
- **Minimize chart junk**: Eliminate unnecessary chart elements that do not contribute to understanding. Clutter detracts from the clarity of your visualizations.
- **Provide context**: Include informative titles, captions, and contextual explanations to help viewers understand the significance of the data.
- **Use visual hierarchy**: Emphasize critical data points and labels through hierarchy. Titles and major data points should stand out, while less important elements can be de-emphasized.
- **Interactive elements**: If applicable, incorporate interactivity that allows viewers to explore the data in more detail. Interactive elements can enhance engagement and understanding.
- **Test for accessibility**: Ensure that your visualizations are accessible to individuals with disabilities. Use alt text for images and consider color contrast for readability.
- **Source and data attribution**: Clearly cite the sources of your data and provide references to maintain transparency and credibility.
- **Keep it updated**: If your data is subject to change or is time-sensitive, regularly update your visualizations to ensure they remain relevant.
- **Seek feedback**: Share your visualizations with colleagues or peers to gather feedback and refine your work. Constructive criticism can help you improve.

By following these best practices, you can create data visualizations that effectively convey your message, engage your audience, and ensure accurate interpretation of the data.

What is Data Manipulation?

Data manipulation is the process of organizing data to make it more understandable. Any type of data may be sorted alphabetically for easy comprehension. Unorganized employee information might make discovering a specific person in a company challenging.

All employee information might be sorted alphabetically, making it easy to access individual employee information. This lets website owners track traffic and popular sites, and web server logs often utilize it.

Accounting users utilize this technique to arrange data to determine production costs, future tax responsibilities, pricing trends, etc. It helps stock market forecasters estimate future stock performance. It may also be used to show information more realistically on websites, software code, or data formatting.

It is possible for computer programs, web pages, or data formatting determined by the user to manipulate data and present it to the user in a more understandable manner.

How to Use Data Manipulation Effectively

The financial data manipulation process involves cleaning, transforming, and analyzing numerical information related to an individual's or organization's finances to gain insights and make informed decisions. To perform data manipulation effectively, follow these key steps:

- **Understand Your Data:** Begin by thoroughly understanding your dataset, including its structure, variables, and any limitations or biases.
- **Data Cleaning:** Clean and preprocess the data to remove inconsistencies, missing values, and outliers. This ensures the existing data is reliable and ready for analysis.
- **Data Transformation:** Transform the data as needed, such as encoding categorical variables, normalizing numeric values, and creating new features to extract meaningful information.
- **Filtering and Selection:** Choose relevant subsets of the data processing for specific analyses. This can involve selecting specific rows, columns, or periods.
- **Aggregation and Summarization:** Aggregate and summarize data to extract insights. Common techniques include group-by operations, aggregating statistics, and creating summary tables or visualizations.
- **Feature Engineering:** Create new features or variables that may enhance the predictive power of your models. Feature engineering involves a deep understanding of the domain.
- **Data Visualization:** Visualize your data to identify trends, patterns, and outliers. Visualizations help in communicating findings and making informed decisions.
- **Modeling:** If your goal is predictive modeling, select appropriate algorithms and train models using the manipulated data. Ensure you use suitable evaluation metrics to assess model performance.
- **Iteration:** Data manipulation is often an iterative process. Analyze results, refine your data manipulations, and repeat the process as needed to achieve your goals.
- Documentation: Document all your purpose of data manipulation steps, which are critical for reproducibility and collaboration.
- **Testing and Validation:** Regularly test and validate your data manipulations and analyses to ensure consistent and reliable results.

- **Ethical Considerations:** When handling and manipulating data, consider privacy, biases, and ethical concerns.

Techniques for Data Manipulation

Data analysis might be difficult if you don't know how to manipulate data. You may use all these methods to better understand your data or its workings, from learning about various kinds of visualization to looking for outliers. Make things easy on yourself and others by using these simple tips.

Multi-step approaches to effective data manipulation may be quite successful. If you're looking to manipulate data, here are some standard techniques:

Introduction

Information has been the key to a better organization and new developments. The more information we have, the more optimally we can organize ourselves to deliver the best outcomes. That is why data collection is an important part for every organization. We can also use this data for the prediction of current trends of certain parameters and future events. As we are becoming more and more aware of this, we have started producing and collecting more data about almost everything by introducing technological developments in this direction. Today, we are facing a situation wherein we are flooded with tons of data from every aspect of our life such as social activities, science, work, health, etc. In a way, we can compare the present situation to a data deluge. The technological advances have helped us in generating more and more data, even to a level where it has become unmanageable with currently available technologies. This has led to the creation of the term 'big data' to describe data that is large and unmanageable. In order to meet our present and future social needs, we need to develop new strategies to organize this data and derive meaningful information. One such special social need is healthcare. Like every other industry, healthcare organizations are producing data at a tremendous rate that presents many advantages and challenges at the same time. In this review, we discuss about the basics of big data including its management, analysis and future prospects especially in healthcare sector.

The data overload

Every day, people working with various organizations around the world are generating a massive amount of data. The term "digital universe" quantitatively defines such massive amounts of data created, replicated, and consumed in a single year. International Data Corporation (IDC) estimated the approximate size of the digital universe in 2005 to be 130 exabytes (EB). The digital universe in 2017 expanded to about 16,000 EB or 16 zettabytes (ZB). IDC predicted that the digital universe would expand to 40,000 EB by the year 2020. To imagine this size, we would have to assign about 5200 gigabytes (GB) of data to all individuals. This exemplifies the phenomenal speed at which the digital universe is expanding. The internet giants, like Google and Facebook, have been collecting and storing massive amounts of data. For instance, depending on our preferences, Google may store a variety of information including user location, advertisement preferences, list of applications used, internet browsing history, contacts,

bookmarks, emails, and other necessary information associated with the user. Similarly, Facebook stores and analyzes more than about 30 petabytes (PB) of user-generated data. Such large amounts of data constitute '*big data*'. Over the past decade, big data has been successfully used by the IT industry to generate critical information that can generate significant revenue.

These observations have become so conspicuous that has eventually led to the birth of a new field of science termed '*Data Science*'. Data science deals with various aspects including data management and analysis, to extract deeper insights for improving the functionality or services of a system (for example, healthcare and transport system). Additionally, with the availability of some of the most creative and meaningful ways to visualize big data post-analysis, it has become easier to understand the functioning of any complex system. As a large section of society is becoming aware of, and involved in generating big data, it has become necessary to define what big data is. Therefore, in this review, we attempt to provide details on the impact of big data in the transformation of global healthcare sector and its impact on our daily lives.

Defining big data

As the name suggests, 'big data' represents large amounts of data that is unmanageable using traditional software or internet-based platforms. It surpasses the traditionally used amount of storage, processing and analytical power. Even though a number of definitions for big data exist, the most popular and well-accepted definition was given by Douglas Laney. Laney observed that (big) data was growing in three different dimensions namely, volume, velocity and variety (known as the 3 Vs) [1]. The 'big' part of big data is indicative of its large volume. In addition to volume, the big data description also includes velocity and variety. Velocity indicates the speed or rate of data collection and making it accessible for further analysis; while, variety remarks on the different types of organized and unorganized data that any firm or system can collect, such as transaction-level data, video, audio, text or log files. These three Vs have become the standard definition of big data. Although, other people have added several other Vs to this definition [2], the most accepted 4th V remains 'veracity'.

The term "*big data*" has become extremely popular across the globe in recent years. Almost every sector of research, whether it relates to industry or academics, is generating and analyzing *big data* for various purposes. The most challenging task regarding this huge heap of data that can be organized and unorganized, is its management. Given the fact that big data is unmanageable using the traditional software, we need technically advanced applications and software that can utilize fast and cost-efficient high-end computational power for such tasks. Implementation of artificial intelligence (AI) algorithms and novel fusion algorithms would be necessary to make sense from this large amount of data. Indeed, it would be a great feat to achieve automated decision-making by the implementation of machine learning (ML) methods like neural networks and other AI techniques. However, in absence of appropriate software and hardware support, big data can be quite hazy. We need to develop better techniques to handle this 'endless sea' of data and smart web applications for efficient analysis to gain workable insights. With proper storage and analytical tools in hand, the information and insights derived from big data can make the critical social infrastructure components and services (like healthcare, safety or transportation) more aware, interactive and efficient [3]. In addition,

visualization of big data in a user-friendly manner will be a critical factor for societal development.

Healthcare as a big-data repository

Healthcare is a multi-dimensional system established with the sole aim for the prevention, diagnosis, and treatment of health-related issues or impairments in human beings. The major components of a healthcare system are the health professionals (physicians or nurses), health facilities (clinics, hospitals for delivering medicines and other diagnosis or treatment technologies), and a financing institution supporting the former two. The health professionals belong to various health sectors like dentistry, medicine, midwifery, nursing, psychology, physiotherapy, and many others. Healthcare is required at several levels depending on the urgency of situation. Professionals serve it as the first point of consultation (for primary care), acute care requiring skilled professionals (secondary care), advanced medical investigation and treatment (tertiary care) and highly uncommon diagnostic or surgical procedures (quaternary care). At all these levels, the health professionals are responsible for different kinds of information such as patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data. Previously, the common practice to store such medical records for a patient was in the form of either handwritten notes or typed reports [4]. Even the results from a medical examination were stored in a paper file system. In fact, this practice is really old, with the oldest case reports existing on a papyrus text from Egypt that dates back to 1600 BC [5]. In Stanley Reiser's words, the clinical case records freeze the episode of illness as a story in which patient, family and the doctor are a part of the plot" [6].

With the advent of computer systems and its potential, the digitization of all clinical exams and medical records in the healthcare systems has become a standard and widely adopted practice nowadays. In 2003, a division of the National Academies of Sciences, Engineering, and Medicine known as Institute of Medicine chose the term "*electronic health records*" to represent records maintained for improving the health care sector towards the benefit of patients and clinicians. Electronic health records (EHR) as defined by Murphy, Hanken and Waters are computerized medical records for patients any information relating to the past, present or future physical/mental health or condition of an individual which resides in electronic system(s) used to capture, transmit, receive, store, retrieve, link and manipulate multimedia data for the primary purpose of providing healthcare and health-related services" [7].

Electronic health records

It is important to note that the National Institutes of Health (NIH) recently announced the "All of Us" initiative (https://allofus.nih.gov/) that aims to collect one million or more patients' data such as EHR, including medical imaging, socio-behavioral, and environmental data over the next few years. EHRs have introduced many advantages for handling modern healthcare related data. Below, we describe some of the characteristic advantages of using EHRs. The first advantage of EHRs is that healthcare professionals have an improved access to the entire medical history of a patient. The information includes medical diagnoses, prescriptions, data related to known allergies, demographics, clinical narratives, and the results obtained from various laboratory

tests. The recognition and treatment of medical conditions thus is time efficient due to a reduction in the lag time of previous test results. With time we have observed a significant decrease in the redundant and additional examinations, lost orders and ambiguities caused by illegible handwriting, and an improved care coordination between multiple healthcare providers. Overcoming such logistical errors has led to reduction in the number of drug allergies by reducing errors in medication dose and frequency. Healthcare professionals have also found access over web based and electronic platforms to improve their medical practices significantly using automatic reminders and prompts regarding vaccinations, abnormal laboratory results, cancer screening, and other periodic checkups. There would be a greater continuity of care and timely interventions by facilitating communication among multiple healthcare providers and patients. They can be associated to electronic authorization and immediate insurance approvals due to less paperwork. EHRs enable faster data retrieval and facilitate reporting of key healthcare quality indicators to the organizations, and also improve public health surveillance by immediate reporting of disease outbreaks. EHRs also provide relevant data regarding the quality of care for the beneficiaries of employee health insurance programs and can help control the increasing costs of health insurance benefits. Finally, EHRs can reduce or absolutely eliminate delays and confusion in the billing and claims management area. The EHRs and internet together help provide access to millions of health-related medical information critical for patient life.

Digitization of healthcare and big data

Similar to EHR, an electronic medical record (EMR) stores the standard medical and clinical data gathered from the patients. EHRs, EMRs, personal health record (PHR), medical practice management software (MPM), and many other healthcare data components collectively have the potential to improve the quality, service efficiency, and costs of healthcare along with the reduction of medical errors. The big data in healthcare includes the healthcare payer-provider data (such as EMRs, pharmacy prescription, and insurance records) along with the genomics-driven experiments (such as genotyping, gene expression data) and other data acquired from the smart web of internet of things (IoT) (Fig. 1). The adoption of EHRs was slow at the beginning of the 21st century however it has grown substantially after 2009 [7, 8]. The management and usage of such healthcare data has been increasingly dependent on information technology. The development and usage of wellness monitoring devices and related software that can generate alerts and share the health related data of a patient with the respective health care providers has gained momentum, especially in establishing a real-time biomedical and health monitoring system. These devices are generating a huge amount of data that can be analyzed to provide real-time clinical or medical care [9]. The use of big data from healthcare shows promise for improving health outcomes and controlling costs.

Big data in biomedical research

A biological system, such as a human cell, exhibits molecular and physical events of complex interplay. In order to understand interdependencies of various components and events of such a complex system, a biomedical or biological experiment usually gathers data on a smaller and/or simpler component. Consequently, it requires multiple simplified experiments to generate a wide map of a given biological phenomenon of interest. This indicates that more the data we have, the better we understand the biological processes. With this idea, modern techniques have evolved at

a great pace. For instance, one can imagine the amount of data generated since the integration of efficient technologies like next-generation sequencing (NGS) and Genome wide association studies (GWAS) to decode human genetics. NGS-based data provides information at depths that were previously inaccessible and takes the experimental scenario to a completely new dimension. It has increased the resolution at which we observe or record biological events associated with specific diseases in a real time manner. The idea that large amounts of data can provide us a good amount of information that often remains unidentified or hidden in smaller experimental methods has ushered-in the '*-omics*' era. The '*omics*' discipline has witnessed significant progress as instead of studying a single '*gene*' scientists can now study the whole '*genome*' of an organism in '*genomics*' studies within a given amount of time. Similarly, instead of studying the expression or '*transcription*' of single gene, we can now study the expression of all the genes or the entire '*transcriptome*' of an organism under '*transcriptomics*' studies. Each of these individual experiments generate a large amount of data with more depth of information than ever before. Yet, this depth and resolution might be insufficient to provide all the details required to explain a particular mechanism or event. Therefore, one usually finds oneself analyzing a large amount of data obtained from multiple experiments to gain novel insights. This fact is supported by a continuous rise in the number of publications regarding big data in healthcare (Fig. 2). Analysis of such big data from medical and healthcare systems can be of immense help in providing novel strategies for healthcare. The latest technological developments in data generation, collection and analysis, have raised expectations towards a revolution in the field of personalized medicine in near future.

Big data from omics studies

NGS has greatly simplified the sequencing and decreased the costs for generating whole genome sequence data. The cost of complete genome sequencing has fallen from millions to a couple of thousand dollars [10]. NGS technology has resulted in an increased volume of biomedical data that comes from genomic and transcriptomic studies. According to an estimate, the number of human genomes sequenced by 2025 could be between 100 million to 2 billion [11]. Combining the genomic and transcriptomic data with proteomic and metabolomic data can greatly enhance our knowledge about the individual profile of a patient—an approach often ascribed as "individual, personalized or precision health care". Systematic and integrative analysis of omics data in conjugation with healthcare analytics can help design better treatment strategies towards precision and personalized medicine (Fig. 3). The genomics-driven experiments e.g., genotyping, gene expression, and NGS-based studies are the major source of big data in biomedical healthcare along with EMRs, pharmacy prescription information, and insurance records. Healthcare requires a strong integration of such biomedical data from various sources to provide better treatments and patient care. These prospects are so exciting that even though genomic data from patients would have many variables to be accounted, yet commercial organizations are already using human genome data to help the providers in making personalized medical decisions. This might turn out to be a game-changer in future medicine and health.

Big data from omics

The big data from "omics" studies is a new kind of challenge for the bioinformaticians. Robust algorithms are required to analyze such complex data from biological systems. The ultimate goal

is to convert this huge data into an informative knowledge base. The application of bioinformatics approaches to transform the biomedical and genomics data into predictive and preventive health is known as translational bioinformatics. It is at the forefront of data-driven healthcare. Various kinds of quantitative data in healthcare, for example from laboratory measurements, medication data and genomic profiles, can be combined and used to identify new meta-data that can help precision therapies [25]. This is why emerging new technologies are required to help in analyzing this digital wealth. In fact, highly ambitious multimillion-dollar projects like "*Big Data Research and Development Initiative*" have been launched that aim to enhance the quality of big data tools and techniques for a better organization, efficient access and smart analysis of big data. There are many advantages anticipated from the processing of '*omics'* data from large-scale Human Genome Project and other population sequencing projects. In the population sequencing projects like 1000 genomes, the researchers will have access to a marvelous amount of raw data. Similarly, Human Genome Project based *Encyclopedia of DNA Elements* (ENCODE) project aimed to determine all functional elements in the human genome using bioinformatics approaches. Here, we list some of the widely used bioinformatics-based tools for big data analytics on omics data.

1. 1.

   SparkSeq is an efficient and cloud-ready platform based on Apache Spark framework and Hadoop library that is used for analyses of genomic data for interactive genomic data analysis with nucleotide precision

2. 2.

   SAMQA identifies errors and ensures the quality of large-scale genomic data. This tool was originally built for the National Institutes of Health Cancer Genome Atlas project to identify and report errors including sequence alignment/map [SAM] format error and empty reads.

3. 3.

   ART can simulate profiles of read errors and read lengths for data obtained using high throughput sequencing platforms including SOLiD and Illumina platforms.

4. 4.

   DistMap is another toolkit used for distributed short-read mapping based on Hadoop cluster that aims to cover a wider range of sequencing applications. For instance, one of its applications namely the BWA mapper can perform 500 million read pairs in about 6 h, approximately 13 times faster than a conventional single-node mapper.

5. 5.

   SeqWare is a query engine based on Apache HBase database system that enables access for large-scale whole-genome datasets by integrating genome browsers and tools.

6. 6.

   CloudBurst is a parallel computing model utilized in genome mapping experiments to improve the scalability of reading large sequencing data.

7. 7.

   Hydra uses the Hadoop-distributed computing framework for processing large peptide and spectra databases for proteomics datasets. This specific tool is capable of performing 27 billion peptide scorings in less than 60 min on a Hadoop cluster.

8. 8.

   BlueSNP is an R package based on Hadoop platform used for genome-wide association studies (GWAS) analysis, primarily aiming on the statistical readouts to obtain significant associations between genotype–phenotype datasets. The efficiency of this tool is estimated to analyze 1000 phenotypes on 10^6 SNPs in 10^4 individuals in a duration of half-an-hour.

9. 9.

   Myrna the cloud-based pipeline, provides information on the expression level differences of genes, including read alignments, data normalization, and statistical modeling.

The past few years have witnessed a tremendous increase in disease specific datasets from omics platforms. For example, the *ArrayExpress Archive of Functional Genomics* data repository contains information from approximately 30,000 experiments and more than one million functional assays. The growing amount of data demands for better and efficient bioinformatics driven packages to analyze and interpret the information obtained. This has also led to the birth of specific tools to analyze such massive amounts of data. Below, we mention some of the most popular commercial platforms for big data analytics.

Commercial platforms for healthcare data analytics

In order to tackle big data challenges and perform smoother analytics, various companies have implemented AI to analyze published results, textual data, and image data to obtain meaningful outcomes. IBM Corporation is one of the biggest and experienced players in this sector to provide healthcare analytics services commercially. IBM's Watson Health is an AI platform to share and analyze health data among hospitals, providers and researchers. Similarly, Flatiron Health provides technology-oriented services in healthcare analytics specially focused in cancer research. Other big companies such as Oracle Corporation and Google Inc. are also focusing to develop cloud-based storage and distributed computing power platforms. Interestingly, in the recent few years, several companies and start-ups have also emerged to provide health care-based analytics and solutions. Some of the vendors in healthcare sector are provided in Table 2. Below we discuss a few of these commercial solutions.

In order to analyze the diversified medical data, healthcare domain, describes analytics in four categories: descriptive, diagnostic, predictive, and prescriptive analytics. Descriptive analytics refers for describing the current medical situations and commenting on that whereas diagnostic analysis explains reasons and factors behind occurrence of certain events, for example, choosing treatment option for a patient based on clustering and decision trees. Predictive analytics focuses on predictive ability of the future outcomes by determining trends and probabilities. These methods are mainly built up of machine leaning techniques and are helpful in the context of understanding complications that a patient can develop. Prescriptive analytics is to perform analysis to propose an action towards optimal decision making. For example, decision of avoiding a given treatment to the patient based on observed side effects and predicted complications. In order to improve performance of the current medical systems integration of big data into healthcare analytics can be a major factor; however, sophisticated strategies need to be developed. An architecture of best practices of different analytics in healthcare domain is required for integrating big data technologies to improve the outcomes. However, there are many challenges associated with the implementation of such strategies.

Challenges associated with healthcare big data

Methods for big data management and analysis are being continuously developed especially for real-time data streaming, capture, aggregation, analytics (using ML and predictive), and visualization solutions that can help integrate a better utilization of EMRs with the healthcare. For example, the EHR adoption rate of federally tested and certified EHR programs in the healthcare sector in the U.S.A. is nearly complete [7]. However, the availability of hundreds of EHR products certified by the government, each with different clinical terminologies, technical specifications, and functional capabilities has led to difficulties in the interoperability and sharing of data. Nonetheless, we can safely say that the healthcare industry has entered into a ‘post-EMR’ deployment phase. Now, the main objective is to gain actionable insights from these vast amounts of data collected as EMRs. Here, we discuss some of these challenges in brief.

Storage

Storing large volume of data is one of the primary challenges, but many organizations are comfortable with data storage on their own premises. It has several advantages like control over security, access, and up-time. However, an on-site server network can be expensive to scale and difficult to maintain. It appears that with decreasing costs and increasing reliability, the cloud-based storage using IT infrastructure is a better option which most of the healthcare organizations have opted for. Organizations must choose cloud-partners that understand the importance of healthcare-specific compliance and security issues. Additionally, cloud storage offers lower up-front costs, nimble disaster recovery, and easier expansion. Organizations can also have a hybrid approach to their data storage programs, which may be the most flexible and workable approach for providers with varying data access and storage needs.

Cleaning

The data needs to cleansed or scrubbed to ensure the accuracy, correctness, consistency, relevancy, and purity after acquisition. This cleaning process can be manual or automatized

using logic rules to ensure high levels of accuracy and integrity. More sophisticated and precise tools use machine-learning techniques to reduce time and expenses and to stop foul data from derailing big data projects.

Unified format

Patients produce a huge volume of data that is not easy to capture with traditional EHR format, as it is knotty and not easily manageable. It is too difficult to handle big data especially when it comes without a perfect data organization to the healthcare providers. A need to codify all the clinically relevant information surfaced for the purpose of claims, billing purposes, and clinical analytics. Therefore, medical coding systems like Current Procedural Terminology (CPT) and International Classification of Diseases (ICD) code sets were developed to represent the core clinical concepts. However, these code sets have their own limitations.

Accuracy

Some studies have observed that the reporting of patient data into EMRs or EHRs is not entirely accurate yet [26,27,28,29], probably because of poor EHR utility, complex workflows, and a broken understanding of why big data is all-important to capture well. All these factors can contribute to the quality issues for big data all along its lifecycle. The EHRs intend to improve the quality and communication of data in clinical workflows though reports indicate discrepancies in these contexts. The documentation quality might improve by using self-report questionnaires from patients for their symptoms.

Image pre-processing

Studies have observed various physical factors that can lead to altered data quality and misinterpretations from existing medical records [30]. Medical images often suffer technical barriers that involve multiple types of noise and artifacts. Improper handling of medical images can also cause tampering of images for instance might lead to delineation of anatomical structures such as veins which is non-correlative with real case scenario. Reduction of noise, clearing artifacts, adjusting contrast of acquired images and image quality adjustment post mishandling are some of the measures that can be implemented to benefit the purpose.

Security

There have been many security breaches, hackings, phishing attacks, and ransomware episodes that data security is a priority for healthcare organizations. After noticing an array of vulnerabilities, a list of technical safeguards was developed for the protected health information (PHI). These rules, termed as HIPAA Security Rules, help guide organizations with storing, transmission, authentication protocols, and controls over access, integrity, and auditing. Common security measures like using up-to-date anti-virus software, firewalls, encrypting sensitive data, and multi-factor authentication can save a lot of trouble.

Meta-data

To have a successful data governance plan, it would be mandatory to have complete, accurate, and up-to-date metadata regarding all the stored data. The metadata would be composed of information like time of creation, purpose and person responsible for the data, previous usage (by who, why, how, and when) for researchers and data analysts. This would allow analysts to replicate previous queries and help later scientific studies and accurate benchmarking. This increases the usefulness of data and prevents creation of "data dumpsters" of low or no use.

Querying

Metadata would make it easier for organizations to query their data and get some answers. However, in absence of proper interoperability between datasets the query tools may not access an entire repository of data. Also, different components of a dataset should be well interconnected or linked and easily accessible otherwise a complete portrait of an individual patient's health may not be generated. Medical coding systems like ICD-10, SNOMED-CT, or LOINC must be implemented to reduce free-form concepts into a shared ontology. If the accuracy, completeness, and standardization of the data are not in question, then Structured Query Language (SQL) can be used to query large datasets and relational databases.

Visualization

A clean and engaging visualization of data with charts, heat maps, and histograms to illustrate contrasting figures and correct labeling of information to reduce potential confusion, can make it much easier for us to absorb information and use it appropriately. Other examples include bar charts, pie charts, and scatterplots with their own specific ways to convey the data.

Data sharing

Patients may or may not receive their care at multiple locations. In the former case, sharing data with other healthcare organizations would be essential. During such sharing, if the data is not interoperable then data movement between disparate organizations could be severely curtailed. This could be due to technical and organizational barriers. This may leave clinicians without key information for making decisions regarding follow-ups and treatment strategies for patients. Solutions like Fast Healthcare Interoperability Resource (FHIR) and public APIs, CommonWell (a not-for-profit trade association) and Carequality (a consensus-built, common interoperability framework) are making data interoperability and sharing easy and secure. The biggest roadblock for data sharing is the treatment of data as a commodity that can provide a competitive advantage. Therefore, sometimes both providers and vendors intentionally interfere with the flow of information to block the information flow between different EHR systems [31].

The healthcare providers will need to overcome every challenge on this list and more to develop a big data exchange ecosystem that provides trustworthy, timely, and meaningful information by connecting all members of the care continuum. Time, commitment, funding, and communication would be required before these challenges are overcome.

Big data analytics for cutting costs

To develop a healthcare system based on big data that can exchange big data and provides us with trustworthy, timely, and meaningful information, we need to overcome every challenge mentioned above. Overcoming these challenges would require investment in terms of time, funding, and commitment. However, like other technological advances, the success of these ambitious steps would apparently ease the present burdens on healthcare especially in terms of costs. It is believed that the implementation of big data analytics by healthcare organizations might lead to a saving of over 25% in annual costs in the coming years. Better diagnosis and disease predictions by big data analytics can enable cost reduction by decreasing the hospital readmission rate. The healthcare firms do not understand the variables responsible for readmissions well enough. It would be easier for healthcare organizations to improve their protocols for dealing with patients and prevent readmission by determining these relationships well. Big data analytics can also help in optimizing staffing, forecasting operating room demands, streamlining patient care, and improving the pharmaceutical supply chain. All of these factors will lead to an ultimate reduction in the healthcare costs by the organizations.

Quantum mechanics and big data analysis

Big data sets can be staggering in size. Therefore, its analysis remains daunting even with the most powerful modern computers. For most of the analysis, the bottleneck lies in the computer's ability to access its memory and not in the processor [32, 33]. The capacity, bandwidth or latency requirements of memory hierarchy outweigh the computational requirements so much that supercomputers are increasingly used for big data analysis [34, 35]. An additional solution is the application of quantum approach for big data analysis.

Quantum computing and its advantages

The common digital computing uses binary digits to code for the data whereas quantum computation uses quantum bits or *qubits* [36]. A *qubit* is a quantum version of the classical binary bits that can represent a zero, a one, or any linear combination of states (called *superpositions*) of those two qubit states [37]. Therefore, qubits allow computer bits to operate in three states compared to two states in the classical computation. This allows quantum computers to work thousands of times faster than regular computers. For example, a conventional analysis of a dataset with n points would require 2^n processing units whereas it would require just n quantum bits using a quantum computer. Quantum computers use quantum mechanical phenomena like superposition and quantum entanglement to perform computations [38, 39].

Quantum algorithms can speed-up the big data analysis exponentially [40]. Some complex problems, believed to be unsolvable using conventional computing, can be solved by quantum approaches. For example, the current encryption techniques such as RSA, public-key (PK) and Data Encryption Standard (DES) which are thought to be impassable now would be irrelevant in future because quantum computers will quickly get through them [41]. Quantum approaches can dramatically reduce the information required for big data analysis. For example, quantum theory can maximize the distinguishability between a multilayer network using a minimum number of layers [42]. In addition, quantum approaches require a relatively small dataset to obtain a

maximally sensitive data analysis compared to the conventional (machine-learning) techniques. Therefore, quantum approaches can drastically reduce the amount of computational power required to analyze big data. Even though, quantum computing is still in its infancy and presents many open challenges, it is being implemented for healthcare data.

Applications in big data analysis

Quantum computing is picking up and seems to be a potential solution for big data analysis. For example, identification of rare events, such as the production of Higgs bosons at the Large Hadron Collider (LHC) can now be performed using quantum approaches [43]. At LHC, huge amounts of collision data (1PB/s) is generated that needs to be filtered and analyzed. One such approach, the quantum annealing for ML (QAML) that implements a combination of ML and quantum computing with a programmable quantum annealer, helps reduce human intervention and increase the accuracy of assessing particle-collision data. In another example, the quantum support vector machine was implemented for both training and classification stages to classify new data [44]. Such quantum approaches could find applications in many areas of science [43]. Indeed, recurrent quantum neural network (RQNN) was implemented to increase signal separability in electroencephalogram (EEG) signals [45]. Similarly, quantum annealing was applied to intensity modulated radiotherapy (IMRT) beamlet intensity optimization [46]. Similarly, there exist more applications of quantum approaches regarding healthcare e.g. quantum sensors and quantum microscopes [47].

Conclusions and future prospects

Nowadays, various biomedical and healthcare tools such as genomics, mobile biometric sensors, and smartphone apps generate a big amount of data. Therefore, it is mandatory for us to know about and assess that can be achieved using this data. For example, the analysis of such data can provide further insights in terms of procedural, technical, medical and other types of improvements in healthcare. After a review of these healthcare procedures, it appears that the full potential of patient-specific medical specialty or personalized medicine is under way. The collective big data analysis of EHRs, EMRs and other medical data is continuously helping build a better prognostic framework. The companies providing service for healthcare analytics and clinical transformation are indeed contributing towards better and effective outcome. Common goals of these companies include reducing cost of analytics, developing effective Clinical Decision Support (CDS) systems, providing platforms for better treatment strategies, and identifying and preventing fraud associated with big data. Though, almost all of them face challenges on federal issues like how private data is handled, shared and kept safe. The combined pool of data from healthcare organizations and biomedical researchers have resulted in a better outlook, determination, and treatment of various diseases. This has also helped in building a better and healthier personalized healthcare framework. Modern healthcare fraternity has realized the potential of big data and therefore, have implemented big data analytics in healthcare and clinical practices. Supercomputers to quantum computers are helping in extracting meaningful information from big data in dramatically reduced time periods. With high hopes of extracting new and actionable knowledge that can improve the present status of healthcare services, researchers are plunging into biomedical big data despite the infrastructure challenges.

Clinical trials, analysis of pharmacy and insurance claims together, discovery of biomarkers is a part of a novel and creative way to analyze healthcare big data.

Big data analytics leverage the gap within structured and unstructured data sources. The shift to an integrated data environment is a well-known hurdle to overcome. Interesting enough, the principle of big data heavily relies on the idea of the more the information, the more insights one can gain from this information and can make predictions for future events. It is rightfully projected by various reliable consulting firms and health care companies that the big data healthcare market is poised to grow at an exponential rate. However, in a short span we have witnessed a spectrum of analytics currently in use that have shown significant impacts on the decision making and performance of healthcare industry. The exponential growth of medical data from various domains has forced computational experts to design innovative strategies to analyze and interpret such enormous amount of data within a given timeframe. The integration of computational systems for signal processing from both research and practicing medical professionals has witnessed growth. Thus, developing a detailed model of a human body by combining physiological data and "-omics" techniques can be the next big target. This unique idea can enhance our knowledge of disease conditions and possibly help in the development of novel diagnostic tools. The continuous rise in available genomic data including inherent hidden errors from experiment and analytical practices need further attention. However, there are opportunities in each step of this extensive process to introduce systemic improvements within the healthcare research.

High volume of medical data collected across heterogeneous platforms has put a challenge to data scientists for careful integration and implementation. It is therefore suggested that revolution in healthcare is further needed to group together bioinformatics, health informatics and analytics to promote personalized and more effective treatments. Furthermore, new strategies and technologies should be developed to understand the nature (structured, semi-structured, unstructured), complexity (dimensions and attributes) and volume of the data to derive meaningful information. The greatest asset of big data lies in its limitless possibilities. The birth and integration of big data within the past few years has brought substantial advancements in the health care sector ranging from medical data management to drug discovery programs for complex human diseases including cancer and neurodegenerative disorders. To quote a simple example supporting the stated idea, since the late 2000′s the healthcare market has witnessed advancements in the EHR system in the context of data collection, management and usability. We believe that big data will add-on and bolster the existing pipeline of healthcare advances instead of replacing skilled manpower, subject knowledge experts and intellectuals, a notion argued by many. One can clearly see the transitions of health care market from a wider volume base to personalized or individual specific domain. Therefore, it is essential for technologists and professionals to understand this evolving situation. In the coming year it can be projected that big data analytics will march towards a predictive system. This would mean prediction of futuristic outcomes in an individual's health state based on current or existing data (such as EHR-based and Omics-based). Similarly, it can also be presumed that structured information obtained from a certain geography might lead to generation of population health information. Taken together, big data will facilitate healthcare by introducing prediction of epidemics (in relation to population health), providing early warnings of disease conditions, and helping in the discovery of novel biomarkers and intelligent therapeutic intervention strategies for an improved quality of life.