This is an edited post-print of a book chapter from the book:
http://www.appleacademicpress.com/handbook-of-research-for-big-data-concepts-and-techniques/9781771889803
It is recommended that interested researchers should acquire the book. Copyright agreement has been adhered to in the preparation of this post-print.

Citation (APA):
Conner, C., Samuel, J., Garvey, M., Samuel, Y., and Kretinin, A., (2020) *Conceptual Frameworks for Big-Data Visualization: Discussion on Models, Methods and Artificial Intelligence for Graphical Representations of Data*. Handbook of Research for Big Data: Concepts and Techniques, Apple Academic Press, USA. Website: http://www.appleacademicpress.com/handbook-of-research-for-big-data-concepts-and-techniques/9781771889803

Note: The purpose of this post-print is to invite suggestions for future research, and extensions of the topic of Big Data Visualization into a research guide / manual.
Correspondence: jim@aiknowledgecenter.com

Conceptual Frameworks for Big-Data Visualization: Discussion on Models, Methods and Artificial Intelligence for Graphical Representations of Data.

Cherilyn Conner
Jim Samuel, Ph.D.
Myles Garvey, Ph.D.
Yana Samuel
Andrey Kretinin, Ph.D.

Introduction

Consider this: Based on IBM research, around 2.5 quintillion bytes of data are being generated every twenty-four hours, and the volumes of data are accelerating! It is anticipated that humanity will have a whopping 49 zettabytes of data by 2022 – and just to contextualize this, we had less than one-quarter of a zettabyte of web-based data prior to 2012. As corporations deal with petabytes of data, managers are faced with huge information overload and experience the paradox of having very high quantities of data with relatively few insights. These phenomena have been broadly identified as "big-data", popularly defined in terms of veracity, volume, velocity, and variety to represent dynamically complex and often apparently chaotic data flows. While there is no dearth of taxonomies, theoretical frameworks, statistical tools, technological provisioning, computational optimizations, mathematical wizardry and extreme rigor of effort in tackling big data artifacts, the question remains: *How can we meaningfully and effectively exploit big data so as to extract insights in ways that are friendly to human sensory abilities and cognitive capabilities?*

The answer to the above question is layered – it would take a portfolio of strategies adapted to the specifics of various scenarios to provide meaningful insights. While acknowledging the

critical and irreplaceable role of data science and information systems theory, statistical methods, advanced technologies, algorithms, artificial (deep and machine) learning, mathematics and general effort, one of the most critical elements that provides very effective and efficient results is data visualization. Data visualization has been defined in a variety of ways and is generally said to be constituted by *"...a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance ..."* (quote from Stephen Few). It has been shown in past research that the type of information presented to human users influences the performance of the users subject to varying information conditions: *"...shows the impact of behavioral aspects, including the effects of a strong belief in certain information categories ... Such beliefs can affect the trading performance of market participants who believe that they have superior information.*" (Samuel et al., 2017). The way in which information is presented then becomes critical as it is not just the data, but the manner in which the insights are evoked that impacts performance by human intelligence.

Data visualization is a very powerful mechanism as it combines the representation of data, and very often the insights contained therein with human sensory friendly external stimuli, which provides easy to interpret internal value to human cognition. The present material aims to outline data visualization from multiple perspectives to provide the reader with a set of perspectives uniquely articulated and juxtaposed to evince insights about extracting insights from big data through data visualization. This chapter covers historical perspectives and a unique bundle of selected methods and cases, including research related to data visualization, to articulate, in conclusion a useful reference framework and concludes with thoughts on the potential for artificial intelligence to influence and manage data visualization for optimal results.

Brief Historic Overview

Data visualization involves the creation of graphical representations such as graphs, maps, charts, or dashboards to represent raw figures of data. Humans typically struggle more with understanding data that is represented in numbers than they do with pictorially represented data, therefore the use of visualization can help improve understanding. Friendly (2008) discusses that the creation of statistical thinking also creates further visual representations. Some examples of this include diagrams for mathematical proofs, nomograms for calculations, and graphical representations for properties of empirical numbers.

> *"The birth of statistical thinking was also accompanied by a rise in visual thinking ... Most recently, advances in statistical computation and graphic display have provided tools for visualization of data unthinkable only a half century ago." (Friendly, 2008)*

Some of the main uses of data visualization are to: compare data, explore compositions and relationships in data, track data over time, analyze distributions, evaluate performance, and make sense of geographical data. There exists an ever growing number of representations of data, some of the most frequent types of visualization are: "bar charts, tree maps, bubble charts, step charts, area charts, heat maps, pie charts, histograms, word clouds, bubble clouds, cartograms, dot

distribution maps, polar areas, spider charts, box and whisker plots, matrices, proportional symbol maps, timelines, time series, scatter plots, Gantt charts, stream graphs, tree visualizations, dendrograms, spark lines, node-link diagrams, alluvial diagrams and radial trees" (Conner et al., 2019). The following outlines a brief history of the evolution of data visualization.

Early maps and diagrams

Data visualizationbegan as early as 6200 BC with the use of geometric diagrams which were used to create maps for navigation and exploration. These maps later lead to spherical earth projections using latitude and longitude as well as star charts. Later visualization was used to create diagrams of electoral systems and knowledge. Further evolution lead to graphs of distance and speed as well as the use of rectangular coordinates for analyzing coordinates. The 16th century brought about advances in triangulation, trigonometric tables, the use of fixed positions to survey land, and cylindrical projections used to portray the globe on maps.

Advances in measurement and theory led to the creation of the pantograph to enlarge or reduce an image, the first printed astronomical pictures, the creation of the coordinate system, and weather maps. The first visual representation of statistical data was created in this time as well as the first published texts on probability theory; other statistical advances include the foundation of demographic statistics, graphs for continuous distributions and bivariate plots. The 18th century led to further advances in graphical forms as well as the beginning of color printing. Advances in statistics led to the first test of statistical significance, normal distributions, curve-fitting, beta density, and population statistics. New graphical representations formed include: contour maps, line graphs, polar coordinates, geological maps, topographical maps, and statistical maps (Friendly, 2008).

Visualization revolution

By the 19th century visualization began to grow rapidly with the late 19th century known as the "Age of Enthusiasm"; this period has also been seen as the "Golden Age" for visualization (Friendly, 2008). Statistical advances include: the first international statistics conference, statistical diagrams used on maps and in lawsuits, the use of least squares, cumulative frequency curves, Gompertz curves, logistic curves, and regression. Various new graphical depictions arose such as; pie charts, circle graphs, choropleth maps, polar-area charts and coxcombs, dot maps, flow maps, contour maps for a 3D table and later for population, ethnographic maps, a semilogarithmic grid for percentages, bilateral histograms and frequency polygons, scatterplots, correlation diagrams, and anamorphic maps. The early 20th century did not have as much advancements in visualization, however, many of the previous advancements grew in popularity, became published in textbooks, and were used in more areas including government and science. Some of the advancements that did come about in this period are: butterfly diagrams, the Lorenz curve, the Hertzsprung-Russell Diagram, Gantt charts, Path diagrams, multiple factor analysis, and Ideographs. Towards the middle of the 20th century the advancements in data visualization began to speed up again with the use and creation of: circular glyphs, high-level computing languages

such as Fortran, exploratory data analysis, triangular glyphs, plots of multivariate data using Fourier series, and interactive graphs in statistics (Friendly, 2008).

Modern visualization

In the late 20th century; through the use of large-scale software engineering, extensions of statistical models, and increased processing speed and capacity; larger data problems were able to be addressed and new representations for data were formed. Some notable advancements were a scatterplot matrix, Cartesian rectangles, mosaic displays of frequencies, Sieve diagrams, parallel coordinate plots, nested dimensions, Treemaps, sparklines, and grammatical rules for graphics (Friendly, 2008). Chen et al. (2009) note that the likely path of visualization is moving away from an offline process and towards something that is first interactive, then information assisted, and finally knowledge assisted. They believe that most development is currently towards information assisted visualization and that a transition towards knowledge assisted visualization is inevitable.

Research in information visualization has led to new types of charts and techniques used in analytics. Gray, Teahan, and Perkins (2017) introduce a number of advanced techniques that are now being used for visualization such as the following.

- Badges / Glyphs: uses icons to define a situation and repeats the same glyph icon for multiple occurrences of an input. Each glyph represents a different state of the object with respect to time.
- Headline Figures: uses summary statistics to convey messages; typical representations are percentages or integers.
- Scatterplots: scatterplots are widely recognized visually although many do not recognize their name; they are used to compare two objects or can include additional dimensions through the use of color, size, or symbols.
- Heat Maps: can be 2D or 3D to compare relationships between categorical variables through the use of colors and opacity.
- Sunburst Charts: combines a radial design similar to a pie chart with layers to show contributions, different sizes are used to represent proportions.
- Bubble Charts: show relationships between variables by utilizing size and color to encode information.
- Radar / Spider Plots: can be used for multivariate comparisons with multiple series of data.

Interactive and 3D modeling

The current state of visualization involves the visual interaction of data and the statistical models built from data. Such interaction transcends traditional data visualization (Endert, et al., 2011). The modern ability to analyze, explore, and communicate data can primarily be attributed to advances in technology such as virtual, augmented, and mixed reality, as well as tangible surfaces. Although research is currently being conducted that extends to 3D applications, published studies have yet to use this as their core topic. Rather than focus on practical applications of 3D interfaces, researchers within this literature stream have focused on the development of lower-level technologies for 3D interfaces. Bach et al. (2017) suggest a new topic of research known as immersive analytics that can be used to "explore the applicability and development of

emerging user-interface technologies for creating more engaging experiences and seamless workflows for data analysis applications." Although 3D visualizations have typically been used for physical sciences, engineering, and design; there is a growing need to also include 3D visualization for statistical and abstract data which has typically been represented in 2D (Bach et al., 2017).

There exist situations that warrant the use of 3D visualizations. Kumar and Benbasat (2018) conducted a study on 2D and 3D graphs to determine when one should be used over the other to improve comprehension. When a graph only contains two components it should be represented using a 2D plane, however, when there are three components either a 2D or a 3D graph can be used. When the data has at least one nominal or ordinal variable all three can be represented using a 2D graph without creating any redundancy, for example, two lines to represent gender. However, when all of the data is measured using a continuous scale, the only way to represent it in 2 Dimensions is by using a pair of graphs, otherwise it would have to be represented in 3D. Kumar and Benbasat (2018) found that 3D graphs outperformed 2D graphs on all of their experimental conditions, including the elementary tasks. Part of the reason for this is that visual cluttering affect 2D graphs more then 3D ones, this is particularly evident when the complexity of the data is higher. It can also take an observer longer to understand what shapes are representing which components of a third component when it is represented in a legend on a 2D graph. Therefore, if a third component is going to be represented in a legend, it should be categorical. When an analyst is looking to explore three components of information, they should first start with a 3D graph and then use a 2D graph only when it is specifically desired. (Kumar & Benbasat, 2018).

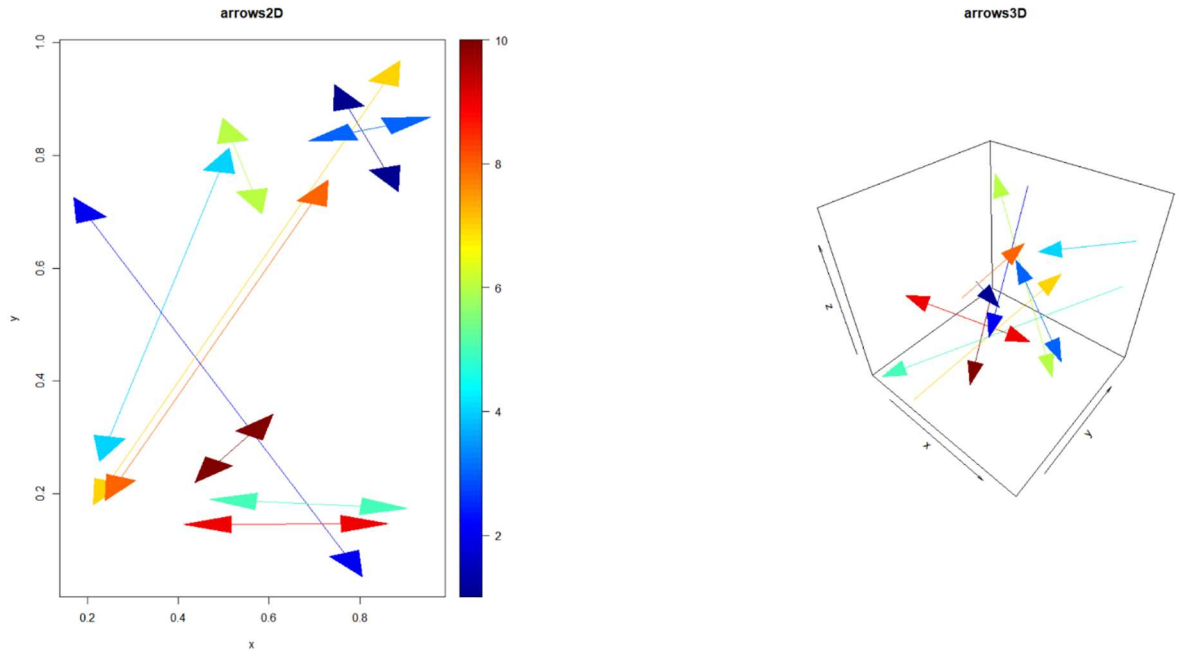

Figure 1: Paths in random uniform distributions shown in 2D and 3D space

3D modeling has also been extended into the physical world by utilizing 3D printing technology. Research in 3D modeling has found that it can be useful for modeling physical objects such as with modeling a heart based on patent data, as well as for buildings and cities. MIT has a platform which they call LuminoCity, which contains 3D printed data of the MIT campus; the systems also combines 2D screens with the 3D models as well as 3D glasses for additional visualization. Weber and Gadepally (2014) utilized the LuminoCity platform to visually analyze Twitter data from the area. This technology can be used to analyze patterns on campus as well as to look into traffic patterns (Weber & Gadepally, 2014)

Methods, Models and Cases

Visualization driven analytics has been defined as "*the science of analytical reasoning assisted by interactive visual interfaces*" with a goal of extracting information, performing analyses, and validating hypotheses through exploration (Endert et al., 2011). Although there are also automatic methods for visual analysis; the complexity of problems requires human engagement in the analysis process from an early stage. Currently the volume of information is expanding at a pace overwhelming current analytical and data consumption capabilities. Therefore, the purpose of research on visualization driven analytics is to use the abundance of data to create opportunities for decision makers to analyze the data which will allow them to make effective actions. Although techniques like classification, SVM and clustering have developed mathematically, rather than from visualization, they paved the way for an important shift from confirmatory data analysis towards that of exploratory analysis. Visualization can make analysis more effective since it can scale to larger or more challenging problems and it makes analysis more efficient since the visual representations can be used to communicated outcomes clearly to a varied audience. (Keim, Mansmann, & Thomas, 2009).

> *"Visual Analytics has the capability to transform many of our daily work processes and make them both more effective and efficient." (Keim et al., 2009)*

Visual analytics has two main focuses: analytical problems and general application areas. Analytical problems typically have an inherent logic; this allows an assessment to be made in a rational way. The problem with analytical problems is that some of them require so much computational or human resources that they are not able to be solved using current methods. General application problems can contain analytical problems within them; however, general applications are typically not as complex. In both of these classes there are three main methodologies: automatic analysis, visualization, and visual analytics. Automatic analysis is typically used when there is a way to measure the quality of the solutions to a problem, this method tends to fail when there is a local optima that is not related to the overall best solution. Visualization uses human knowledge, creativity, and intuition to solve problems; however, this fails when the data is large. Visual analytics allows for a combination of both automatic analysis and visualization; it can use algorithms and computing power while also allowing for human integration of knowledge and intuition (Keim et al., 2009).

Currently visualization analytics interpretation is primarily performed by users through subjective comprehension, analysis, and ad hoc logic. However, this approach is becoming

increasingly difficult for the generation of useful visualizations as the size of data grows. It can also be hard to ascertain if the data have been fairly analyzed or understood. This is propagated by the absence of relevant and easy to use mechanims to gauge the quality of graphical visualizations (Wang & Shen, 2011). Visual models have two main draw backs; one is that by allowing users to interact, the "*users are expected to be experts in the underlying model*", the other is that as data sets increase in size "*adjusting parameters creates an issue of scalability*" (Endert et al., 2011). Information theory has begun to be used to address the problems with visualization. Each step in the visualization process encodes information with the goal of preserving as much information as possible to create an output. Parameters must be carefully chosen to represent inevitable information loss. The entropy of a variable is used to quantify the uncertainty of the variable in a distribution; it represents the amount of data "*required to describe the variable*", therefore, "*the higher the entropy the more information it contains*" (Wang & Shen, 2011).

> *"Without a systematic and quantitative way to guide the user through the visual analysis process, visualization could soon lose its value to be a viable approach for large-scale scientific data analysis." (Wang & Shen, 2011)*

Visualization can be seen a search process where a user experiments with controls until they are satisfied with the outcome, there are two main techniques to assist with this. One technique is that of information assisted or supported visualization. Such visualizations not only tend to portray information about the data provided, but can also include properties of the visualization schema or qualifications of the results. A user can then use this knowledge to reduce their search for "*optimal control parameters*". As the quantity of data and associated sophistication grows, the use of such data, including dynamic input, to promote graphical representations will no longer be optional but be treated as being necessary. The second technique is knowledge assisted visualization, which relies more on prior knowledge on the part of the user; where the user assigns colors or meaning to objects based on their domain knowledge. Lack of knowledge on the users' part can create an obstacle in implementing visualization techniques; however, it does promote the sharing of domain knowledge among users. There are also some systems where general or domain knowledge has already been incorporated. When using knowledge assisted visualization, rule-based reasoning can be used to establish control parameters or case-based reasoning can be used to obtain knowledge through successes and failures. (Chen et al., 2009).

The terms "*data, information and knowledge*" are frequently used when talking about visualization; however, they are used to talk about different levels of abstraction, understanding, or truthfulness (Chen et al., 2009). Some cases consider visualization as being "*concerned with exploring data and information*"; or that the purpose of the information visualization is to gain unique understanding concerning the "*information space*". Other cases have a viewpoint where information visualization is for data mining and knowledge discovery. Others view the three terms to indicate different data types; these imply that all three terms could probably serve as the raw material, the substance or the produce of representation processes. The definitions for each term are not consistent across different disciplines; Chen et al. (2009) presented a clarification of each of these terms when used for visualization processes where; data refers to symbols, information is data that is processed and provides evidence about "*who, what, where, and when",* and *"knowledge is the application of data which provides answers to how*".

A problem with the visualization of information is that it is not based on a clearly defined theory, this causes the tool that is used to be difficult to validate and how valuable a method is cannot be predicted until it is implemented (Purchase, Andrienko, Jankun-Kelly, & Ward, 1970). A theory for information visualization can be used to predict the success of a new visualization method. Purchase at al. (1970) suggest that multiple theories might be needed for different levels; they note that cognitive and perceptual theories and statistical methods are already used. They associate analysis of theoretical abstractions of data visualization with the following notions: understanding of graphical representations' physical structure, exploration or manipulation of the external representation, and explorations or "*manipulation of the internal data model*". There are also limits on information visualization due to display restrictions that include pixel amount, color options, and refresh rates. Although there are many issues with measuring information in a visualization that are also ways to improve the information such as: using a novel layout, including links with the visualization, using redundant mappings, or animations (Purchase et al., 1970).

> *"Information theory has been applied to solve many tasks in imaging such as image enhancement, registration, and segmentation… In computer graphics, information theory has been utilized to effectively solve a number of problems including scene complexity analysis, pixel super sampling, viewpoint selection, light source placement, ambient occlusion, mesh simplification, and image aesthetics measure." (Weng & Shen, 2011)*

Principal Component Analysis (PCA) is a common deterministic method that is used to represent data in a condensed dimensional structure; such that the representation reflects a high-dimensional data set towards the directions possessing the highest variances. When there are two directions chosen the PCA produces summaries that of data that are easily visualized; however, the problem with PCA is that important structures may not correlate with variance. The spatialization may then mask information that could be useful. The probabilistic version of PCA known as PPCA incorporates a statistical approach to modeling. PPCA is useful because it approximates lower dimensional summaries of higher dimensional data. In user guided PPCA after an initial display is obtained the user adjusts the locations of observation indicating that they view the objects as more similar if dragged together or more different if dragged apart (Ender et al., 2011).

Another deterministic method; Multi-Dimensional Scaling (MDS); maps higher dimensional data to a lower dimensional model structure by conserving pairwise distances between observed values. In user guided MDS the user may interact and rearrange some of the observations in the visualization if they do not agree with the display. When the relative position of two points is adjusted the weights that are simultaneously consistent with the quantitative model as well as the user's adjustment are calculated. These weights are calculated by conceptually optimizing fixed adjusted points while locating optimal weights which are aligned with the visualization. Generative Topographic Mapping (GTM) is another method that uses a "*nonlinear variable modeling approach for high-dimensional data*" which "*is considered to be a probabilistic alternative to other models*". When adjusting parameters, the areas of interest respond while the areas that are distant do not change (Endert et al., 2011).

There a number of different tools and programs for data analysis and visualization; many tools focus on one specific type of visualization or analysis and are often easy to learn; however, they can be limiting in their functionality and application. Many analysts choose instead to write their own analysis and visualizations using a programing language; the most commonly used languages for data analysis are R and Python, along with Tableau, SPSS, SAS, STATA, MATLAB and such other tools. These languages are both open source and they have a large community of users who create packages and libraries to assist in all types of analysis as well as a number of online resources for users to learn and debug. Stander and Dalle Valle (2017) created a course at the university of Plymouth to introduce students to working with big data and social media data using R through RStudio and to show them a way to make professional and reproducible reports using RMarkdown. A well-known R package called "ggplot2" was utilized to create visualizations and to introduce students to different modeling techniques and students also learned a number of different types of visual and numeric methods for data analysis. Overall the course received positive feedback and student performance was able to prove the course to be successful (Stander and Dalla Valle, 2017).

Aesthetics

Figure 2: Heat map to show correlations between all continuous variables in a data set where dark red shows a strong negative correlation, dark blue shows a strong positive correlation, and white shows no correlation

Aesthetics can be used in data visualization and beyond to stimulate desire, to cause a first impression to be positively influenced, to overwhelm a viewer, or to encouraged continued usage. Aesthetic usage in data visualization has primarily been focused on graph drawings with an emphasis on minimizing bends and edge crossings while maximizing angles, orthogonality, and symmetry. Aesthetics are typically used in the domain of data visualization for their potential to promote task effectiveness; through a reduction of completion time and error rate, it can improve the efficiency and effectiveness of task performance. Although it can increase effectiveness; the use of aesthetics is not as frequently used in data visualization and is often added on at the end of development processes.

> *"Colour is an important and frequently-used feature. Examples include colour temperature gradients on maps and charts, colour-coded vector fields in flow visualization, or colour icons displayed by real-time simulation systems." (Healey, 1996)*

Cawthon and Vande Moere (2007) investigate the effectiveness of aesthetics by looking at correlations between task abandonment, response time, and perceived aesthetic. Through an online survey they found that a Botanical Viewer, Polar View, and SunBurst were the most visually pleasing visualization techniques with BeamTrees and TreeMaps being the least pleasing. The displays of Botanical Viewer, StepTree and BeamTrees have the lowest accuracy percentages; the highest percentage of accuracy was found in the SunBurst display. In terms of efficiency; Botanical Viewer and StepTree had the slowest time; SpaceTree provided the fastest response time followed by Windows Explorer. This study was able to show that displays that focus on beauty can also be effective displays and that displays with lower aesthetic scores have a higher task abandonment (Cawthon & Vande Moere, 2007).

Healey (1996) explores the use of color in data representations, however. it is important to consider what colors are most effective. They explore how elements can be identified through color, what factors will make the color of a target element easy to find, and how many colors can be used at the same time without reducing accuracy. Other studies have used color to show correlations in a data set with five dimensions; while others have found that no more than five to seven colors should be used at one time. Other research has looked into user perception of features such as hue, luminance, and height. If target variables can be quickly and accurately found due to their color, they can then be used in further analysis. Healey (1996) found that effective color should be chosen based on a combination of color distance, linear separation, and color category altogether; although effective colors can be found by using just color category. They also found users have very little difficulty with identifying targets when there are either three of five colors present; however, once they increased the color amount to seven and nine the users began to have more difficulty (Healey 1996).

Strobel, Grund and Linder (2018) conducted a study to explore the "*effect of seductive details on graph reading*" and processing time since such enamor affective details can adversely

influence with logical graph interpretation. This is so because of the diverse forces that tend to vie for the viewer's attention. The distraction hypothesis implies that seductive details can interrupt the transition between ideas since it diverts the focus of human intelligence away from logical and rational informational substance towards relatively irrelevant substance. Seductive details cause users to build models around irrelevant information rather than using the important information in the text and can also cause higher processing time. When the seductive details are present in a graph, they can cause distractions in regard to content as well as in regard to spatial aspect by adding extraneous visual objects. Strobel et al. (2018) tested to determine if seductive details affect processing via distraction, which would cause longer processing time and increased time spent looking at the seductive detail, or via disruption, which would cause the user to have to revisit areas that they looked at before looking at the seductive detail. Results showed that there was no significant difference in error rates between experimental and control conditions, however, processing time was significantly longer for both the seductive text and seductive picture conditions. Through eye tracking they found no evidence of the distraction hypothesis and found that the increased processing time was attributed to fixation on the seductive details.

Models

A model can be created off of a number of patterns where each one represents a part of the data; information visualization can be viewed as tools to help a user view the patterns available to build a model. Different patterns may be viewed as useful depending upon the data that is being analyzed. It is believed that if information visualization tools provide information about the types of patterns they fit, then a user will be able to choose an appropriate tool (Purchase et al., 1970).

> *"[Models] posit formal logical abstractions about objects, events and processes and linking them with causal or operational mechanisms in a greater unit" (Liu & Stasko, 2010).*

A common model for expressing relationships and connections between variables is a decision tree; however not all connections can be represented in this method since they must be ordered. An alternative to this in an influence diagram where "*nodes do not have to be ordered*" and "*do not have to depend on predecessors*". Howard and Matheson (2005) describe an influence diagram as *"a formal description of the problem that can be treated by computers and a representation easily understood by people in all walks of life and degrees of technical proficiency."* They are thought to be beneficial since large sections of any populace are untrained in quantitatively legitimate representation techniques. This form of representation can specify relations, functions, and numbers in both probabilistic and deterministic cases. Influence diagrams are an important tool for any formal description of relationship and for all modeling work due to its generality.

> *"One of the most perplexing aspects of making decisions under uncertainty is the problem of representing and encoding probabilistic dependencies" (Howard & Matheson, 2005).*

When looking at an influence diagram an arrow from one aleatory variable X to another variable Y, implies that such results coming from X have potential to influence the likelihood values associated with relationships to Y. An arrow coming from a decision means that what is it

pointing to is using the knowledge of the outcome of the decision. Nodes that are not connected by an arrow are probabilistically independent. Influence diagrams can usefully display required assessments for large decision problems. It is noted that an arrow between to variables means one may depend on the other not that one must depend on the other; and although adding influence arrows will not affect any probabilities, they may prevent the recognition of independencies. An influence diagram cannot represent all possible influences and assumes that missing influences do not actually exist. For decision nodes the diagram assumes that only the information available at the time of decision making is shown by the direct predecessors of the decision. However, for chance nodes the probability can be based on all non-successors. This allows for probabilistic assessments and computations to be made easily and the freedom allows a decision maker who agrees with the information but not on the observation to make decisions.

When a decision tree is able to be made from an influence diagram it must preserve the order and it "*must not allow a chance node to be a predecessor of a decision node if it is not a predecessor*". Laws of probability can be used to eliminate the conditioning of one variable on another which can lead to the inclusion of additional influence and a probability assessment must be used to determine the arithmetical likelihood of "*each chance node*" based on respective precedents. Howard and Matheson (2005) use an example with a toxic chemical to show how an initial decision tree might be based on incomplete information that must be added in and how perfect information helps to further clarify what they should be looking at to make a decision. They conclude by commenting on the ways in which influence diagrams possess process simplification potential for decision making and probabilistic modeling (Howard & Matheson, 2005).

Uncertainty

The problem with existing models is that they are mainly deterministic and only provide a single output value without indication of the uncertainty of the value. An alternative to deterministic models are probabilistic ones which spell out the risks associated with a model; they allow a decision maker to evaluate the uncertainties of the system and see a realistic picture of the potential outcomes. These models obtain their raw input from distinct sources: direct data, professional knowledge or past models. Such likelihood models that include uncertainty in model computations are increasing in popularity since they address a goal of decision analysis "*which is to provide the decision maker with a picture of the current knowledge as well as its deficiencies*"; the uncertainty is frequently expressed as a probability distribution that indicates likelihoods for possible outcomes. It is argued that a model which provide input for decision-making should contain informational substance about probabilities. This is because an attractive model could also include high probabilities of undesired outcomes (Uusitalo, Lehikoinen, Helle, & Myrberg, 2015).

A challenge that arises when combining models is determining if the definitions of a variable are compatible, this is "*critical in defining the probability distributions or the uncertainties*"; therefore, examining definitions should be the first step in model creation. The use of multiple models that are made with independent choices for simplification, assumptions about dependencies, and parameters can also be used to asses structural uncertainty. If multiple independent models all produce similar results the uncertainty can be concluded to be small; however, depending on the scenario it may be natural for the predictions to be similar (Uusitalo et

al., 2015). The Monte Carlo method is the easiest way to conduct an uncertainty analysis because it randomly draws inputs from their distribution, the "*outputs can then be seen as a random sample of their distribution*". A sensitivity analysis can also be used; this method characterizes how outputs respond to changes in inputs and emphasizes outputs that are most sensitive. If a variable changes noticeably when an input is changed within its reasonable range the variable shows large amounts of uncertainty. (Uusitalo et al., 2015).

When working with K-L models, Burnham, Anderson, and Huyvaert (2010) note that model uncertainty can cause an analyst to be uncertain as to which model is the best; one method for addressing this problem is by quantifying the model probabilities. However, it is also possible that alternative models may reflect insights unavailable through that which is statistically identified as a better model, therefore other models should be considered when making inferences. Such predictions based on model mixes can be used as a "*weighted mean of model probabilities*". Another approach uses multimodel inference to compute measures of precision, and other methods exist for ranking predictor variables. Any construction of precision measure must include uncertainty otherwise confidence may be adversely affected. Once the data is analyzed then the quantitative evidence can be reviewed to consider judgements for understanding and interpreting evidence (Burnham et al., 2010).

When performing an assessment on hazards the presence of aleatory uncertainties that are caused by unpredictable variation and epistemic uncertainties caused by a lack of knowledge are inevitable (Kunz, Gret-Regamey, & Hurni, 2011). In some fields that deal with hazards probabilistic methods and uncertainty distributions are being used, however, hazard assessments are often displayed on maps where there is currently a "*lack of information about existing uncertainties and the map users are often not aware of the uncertainties*". Currently a consensus about how to display uncertainty is lacking; some argue that this inclusion will confuse a map or lead to misunderstanding. Others do not believe that those who make decisions are not understanding of uncertainty-statistics but rather they consider that the problem is in the communication of the uncertainty. The goal is to "*provide visualizations that incorporate uncertainty to aide in analysis and decision making*"; one way to incorporate this is through interactive cartographic information systems; these systems allow a dynamic and responsive interaction with data. Users are able to exclude layers on information to focus more on the data that interests them; this helps to reduce misunderstandings that can arise from the maps (Kunz et al., 2011).

Applications

Modeling has grown in popularity for use in looking at "*environmental problems like climate change, overfishing, erosion, reduction in biodiversity*" and many more (Uusitalo et al., 2015). Researchers are looking more to understanding the uncertainties of the models because of a need to better understand visualization of uncertainty in meaningful ways; the researchers must be able to "*maximize human benefit while also minimizing harm caused to nature and avoiding any disastrous outcomes*". Although probabilistic modeling has grown in popularity, its use in ecological application remains low (Uusitalo et al., 2015). When looking at flood related data Ramos, Bartholmes, and Thielen-del Pozo (2007) found that EPS-based forecasts are able to help

with decision-making. The most useful visualization was the box counting approach in temporal diagrams; these provide essential information and are easy to understand. They also note that combining probabilistic results with deterministic forecasts helped with understanding a situation as well as to obtain an overview of the situation (Ramos et al., 2007).

Modeling is an important aspect of understanding phenomena and designing clinical systems when looking at physiological processes and pathologies (Konukoglu, et al., 2011). Modeling has begun to include model personalization where a model can adapt to a patient based on their data. Konukoglu et al. (2011) look at the estimation of parameters when working with this data which can be a challenging task due to: 1) sparsity of data, 2) uncertainty of data, 3) complexity of the model, and 4) assumptions of the model. They note that *"methods for estimating patient-specific parameters should take these challenges into account in providing not only point estimates for the parameters but also ranges of possible parameters and confidence margins"* (Konukoglu, et al., 2011). A probabilistic model can address the problem of parameter estimation by creating a joint probability distribution; however, a large number of sample simulations are required to obtain an accurate estimate, which can be computationally unrealistic. Uncertainty quantification has been able to reduce this computational cost by using "*polynomial chaos expansions of variables*" and appropriate modeling.

Konukoglu et al. (2011) propose a Bayesian method for estimation when there are a large number of parameters by using Eikonal-Diffusion (ED) methods in cardiac electrophysiology application. They focused on the ED model since it allows for condensed construction while also including nonlinearities; these models can also be solved rapidly. The probabilistic approach considers each parameter as a random variable and uses the distributions to represent the ranges of the parameters and their expected values. Such association among parameters and variables can be represented by a appropriate distributions which can then be decomposed into multiple solutions.

A Markov-Chain Monte-Carlo (MCMC) can be used to evaluate an equation sequentially at different points and a spectral representation can be used to speed up the probabilistic model process when "*the number of parameters are small and the equation is linear*". With this, a summation can be used to estimate the solution to a model equation; this is faster than solving the equation since functions are only added. These experiments were able to show the capabilities of the probabilistic method, however, they note that the results of the model are influenced by the data noise model, the model parameter variability, and the model error. Konukoglu et al. (2011) believe that their probabilistic model provides a personalized model with a confidence measure which can help when weighing the predictions of the model and can guide data acquisition. They also believe that this model can help the medical field to move away from invasive methods towards non-invasive ones. Konukoglu et al. (2011) conclude that these methods are vital for making stochastic personalization a possibility. They are already able to provide results for clinical data.

Types of textual analysis

The domain of data visualization has historically been primarily focused on numerical data since traditional statistics and data analysis focus on numerical data; however, with the prevalence of big data many of the existing data visualization methods are being adopted as there are now additional data types besides numerical. Currently there exists a demand for textual analytics since there has been a significant increase in the gathering, storing, manipulation, representation and analysis of textual data also known as "character" or "string" data. Social media has fueled the growth of textual analytics (Stieglitz, et., al., 2018). Social media activity has also grown due to a significant increase in access to and use of mobile devices (Jimenez-Marquez, Gonzalez- Carrasco, Lopez-Cuadrado, & Ruiz-Mezuca, 2019). Textual data has the potential to represent unique user information in the form of direct statements and latent implications which cannot be gathered from numerical data and other forms of data; it also lends itself to various forms of creative analysis. Organizations are focusing more on the gathering and analysis of textual information in order to improve their understanding of user sentiment, behavior, trust, loyalty and various other critical decision-influencing variables (Conner et al., 2019).

> *"The words selected by managers to describe their operations and the language used by media to report on firms and markets have been shown to be correlated with future stock returns, earnings, and even future fraudulent activities of management. Clearly, stock market investors incorporate more than just quantitative data in their valuations, but as the accounting and finance disciplines embrace this new technology, we must proceed carefully to assure that what we purport to measure is in fact so." (Loughran & McDonald, 2016)*

Although textual analysis has existed for centuries, advances in technology and search engines have caused textual analysis to spread to most if not all subjects. Textual analytics has begun to be applied to accounting and finance, however, there still exist vague taxonomies in this field. Loughran and McDonald (2016) note that literature needs to focus more on hypotheses that link to economic theory than with just applying previously used theories. They believe that there must be exposition and transparency when converting qualitative data into quantitative measures to avoid the introduction of inaccuracy and that models that are created must be replicable (Loughran & McDonald, 2016). Currently there is a shortage of meaningful and interpretable plot types and visualization methods for textual analytics. Existing visualization of textual data can be represented in four main categories: quantities visualization, trend visualization, sense visualization, and context visualization.

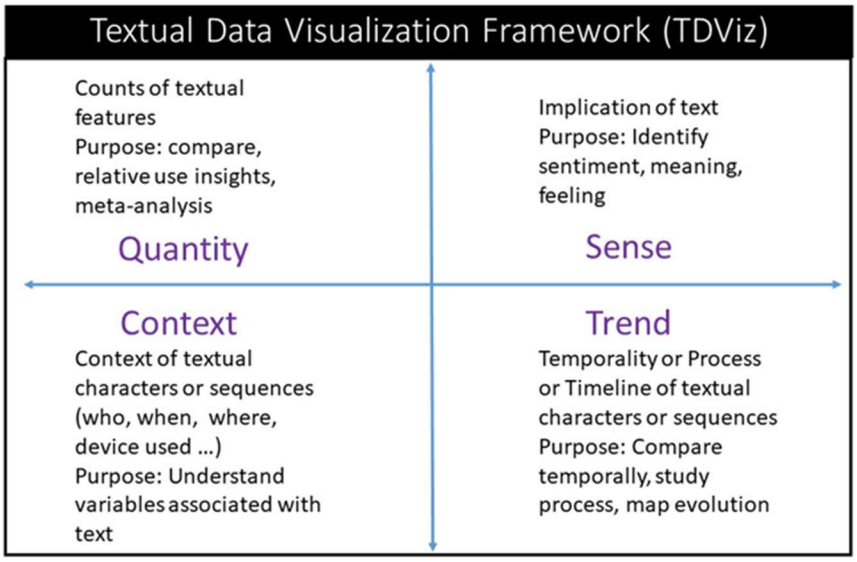

Figure 3: Textual Data Visualization Framework from Conner et al. (2019, used with permission)

Quantities Visualization

Quantities visualization consists of visual graphical representations that incorporates counts of textual characters, features, or sequences. Frequently used examples of quantity visualization are word clouds, word trees, count frequencies, table, pie charts, and phrase nets. Kabir, Karim, Newaz, and Hossain (2018) use word clouds, which display a pictorial representation of word frequency, as an initial analysis phase before using other methods. Word clouds, which are also known as tag clouds, are useful for discovering the most frequently used words since the size and sometimes color of a word is determined by the word's frequency; therefore, larger words are used more frequently. However, the use of a word cloud removes all context from the words that are being analyzed. For example, Kabir et al. (2018) used a word cloud to display the topics that a Bangladeshi politician tweeted about most, however, they were not able to tell if the topics had negative or positive connotations attached to them. Conner et al. (2019) used a word cloud (figure 4) to represent word count on textual analysis research, they found that the most frequently discussed topic regarding data visualization and textual analysis is "Sentiment".

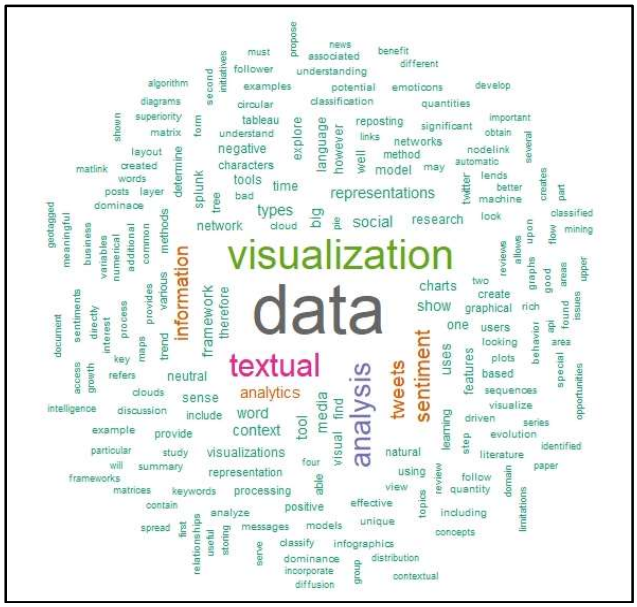

Figure 4: A word cloud created by Conner et al. (2019) to show the frequency of words used in a paper about textual analytics and data visualization (used with permission)

Chen, Lin and Yuan (2017) explore the use of word clouds as well as a number of different textual visualization techniques. One tool that can be used to help users understand concepts and opinions is SentenTree, which visualizes frequent key word patterns. Another tool they explore visualizes multiple topics with parallel rivers that use fluctuation to show the number of tweets in a topic (Chen et al., 2017).

Quantity visualization has also been explored with respect to electronic dominance, where users aim to gain relative superiority in social media and other electronic communications. Dominance can be determined using text analysis to look at the use of upper-case letters, repeated letters, emoticons, and special characters. Through an analysis of tweets about certain stocks Kretinin, Samuel, and Kashyap (2018) were able to show dominant behavior in users through repetition of tweets by the same user, use of upper case, use of longer tweets, emphatic language, and counter posts. Previous research on dominance by Samuel, Holowczak, Benbunan-Fich, and Levine (2014) used manual and automatic content analysis to analyzed chat transcriptions. This research showed the potential of automatic textual analysis in identifying dominance in communication through a computer-mediated group; this provides feedback about the nature of group discussions as well as real time dominance detection (Samuel et al., 2014).

Trend Visualization

Trend visualization applies a timeline, process, or temporality to textual characters, features, or sequences; it can be accomplished through time series, process, or evolution. Trends can be use within finance and news to identify intervals and features for relationship analysis. Timeline visualization typically uses icons for events that occur at one point in type and horizontal bars for events that take place over intervals, they also often have short descriptions of the event

as well as colors or shapes to differentiate between different relationships (Nguyen, Xu, Walker, & Wong, 2016). Chen et al. (2017) utilized a "*tool called gestaltmatrix which uses a glyph matrix*" to visualize the evolutions of user relationships. They also used a gestalt-based glyph to show the relational data in chronological order as well as GraphFlow which uses static flow visualization to show structural changes in in reposting networks overtime. One of the earliest tools for visualizing social media content, known as VisualBackChannel, uses a time stamped display of keywords as subjects as well as a circle shaped view for representing participants. Another tool called ThemeCrowds uses hierarchical visualization to create a chronological layout of keywords. RoseRiver also uses hierarchies; it finds suitable ones for different times using a tree-cut algorithm (Chen et al., 2017).

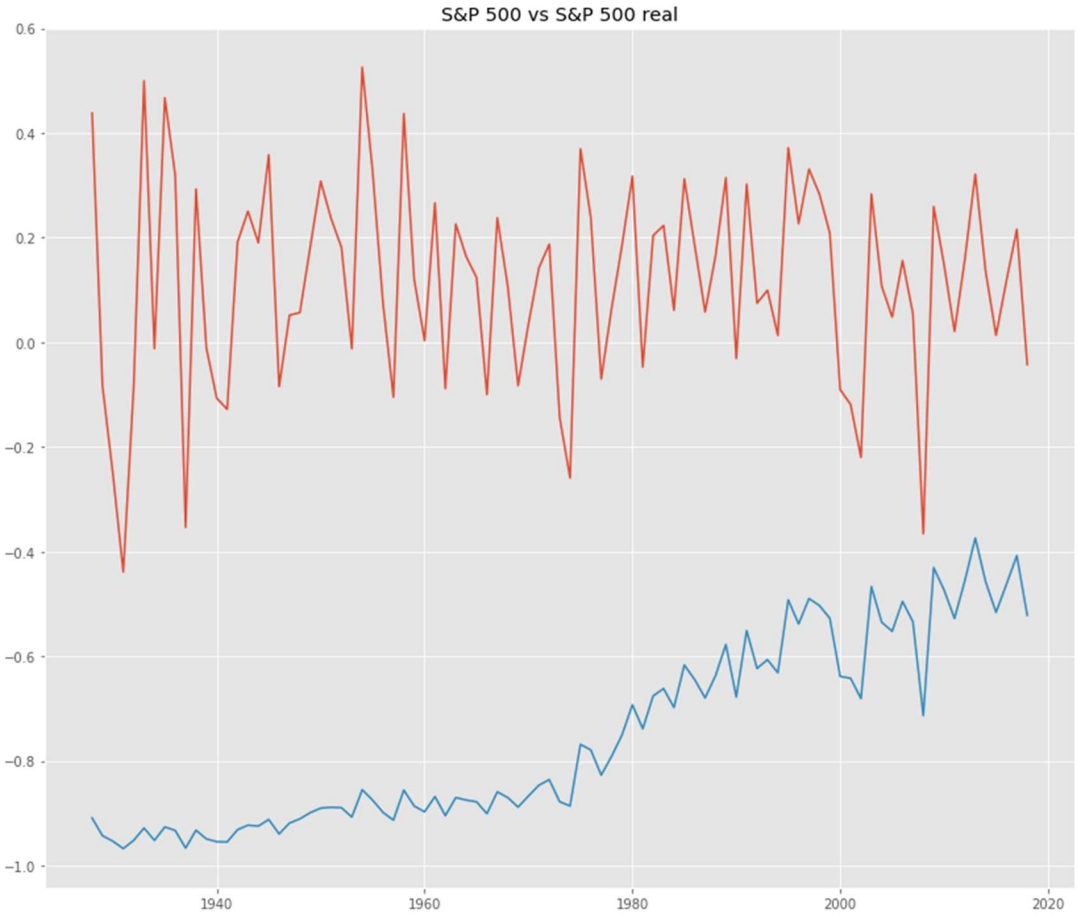

Figure 5: a view of the percentage of change in the S&P 500 with dividends in red and the percentage of change adjusted for inflation in blue

Most timelines visualizations are viewed horizontally with the earliest time on the left and the most recent time on the right, A popular example of this, LifeLines, has been used to visualize medical records. LifeLines can also be used to aggregate date for visualization using ThemeRiver or Streamgraph which uses different rivers for each theme and different widths to represent lengths of time. Other methods that have been used for visualizing set relationships with fixed data are Bubble Sets, LineSets, and KelpFusion; however, these methods are not ideal for readability. In order to address this issue Nguyen et al. (2016) create a new type of visualization called TimeSets

for visualizing set relations on a timeline by using two Gestalt principles of proximity and uniform connectedness for grouping. This method is seen to clearly show events and relationships overtime, to adjust detail levels of different events for appropriate display, and to use gradient colors for events that are shared by topics. This method allows for three levels of labeling: complete labels, trimmed labels, and aggregated labels that combine multiple events into one label. The use of aggregation allows for TimeSets to scale in order to handle large amounts of data, however there is no visual differentiation to determine how many events are included in the aggregation. Through a case study comparing TimeSets and KelpFusion they were able to show that users were slightly more accurate when using TimeSets and that most participants preferred TimeSets (Nguyen et al., 2016).

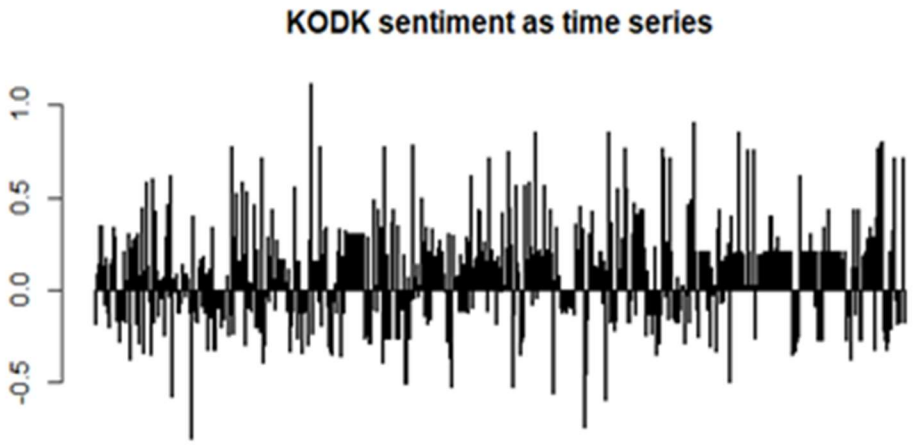

Figure 6

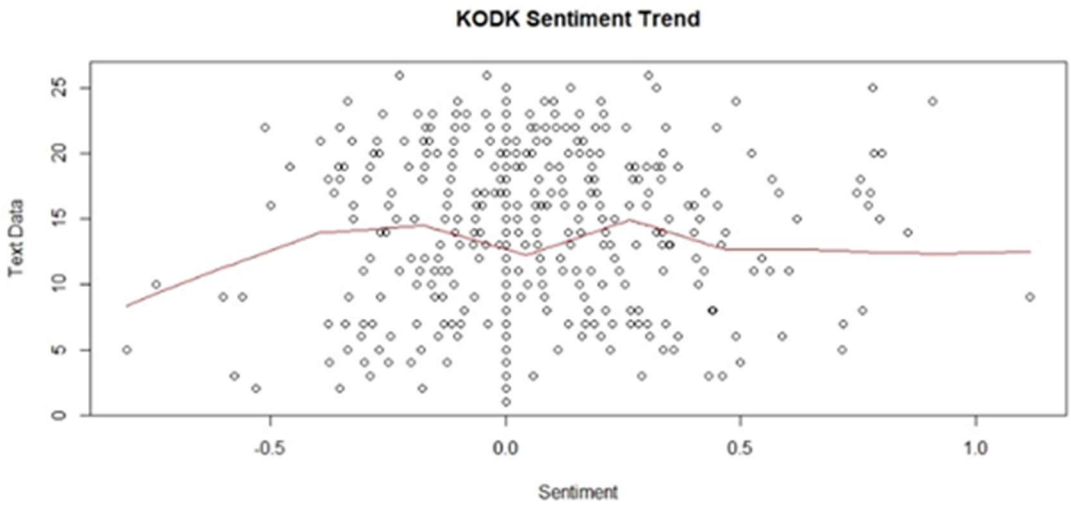

Figure 7

Figure 6 & 7: Kodak Stock Trend Analysis Visualization

Time series visualization research looks to use higher abstraction levels to convey meaningful information and to show as many variables as possible. SemanticTimeZoom is a tool that aims to portray as much quantitative and qualitative data at one time. Aigner, Rind, and Hoffmann (2012) conducted a study to validate the effectiveness and efficiency of SemanticTimeZoom since it addresses the main research challenges of "*conveying information at higher abstraction levels and showing as many variables as possible*". Other methods of visualization that were looked at were either hard to read, only able to show one qualitative abstraction, or only able to show simple abstractions. A study was conducted to compare SemanticTimeZoom to KNAVE-II, which can also show multiple abstractions of data but cannot show as many as SemanticTimeZoom. They found that in the first round of questions that SemanticTimeZoom had significantly better completion times than KNAVE-II, however, this difference was not found in the second round of questions. In both rounds there were significantly shorter times for SemanticTimeZoom in terms of task completion time and error rate. Of the 20 people who participated in the study 19 of them preferred SemanticTimeZoom; overall the study was able to show that SemanticTimeZoom can perform on the same level of KNAVE-II and that it can excel for more complicated tasks (Aigner et al., 2012)

Sense Visualization

Sense visualization uses graphical representation of analysis to explore the meaning of textual characters, features, and sequences. Natural Language Processing and a subsection of it known as sentiment analysis are typically used to analyze meaning; with research focusing on data as broad as entire documents and as narrow as single words or phrases (Kabir et al., 2018). Semantic analytics, which uses the semantics of information not just as syntax or statistical patterns is also being used to improve information systems. Since an "*understanding of semantic information can lead to actionable and timely decision making*", therefore relevant and useful visual representation tools must be used; some models that have been used to represent semantic analytics include: OntoLift, Cluster Maps, WordNet, TAP, SWETO, GlycO, TouchGraph, OWLVisz, and WEBCOM (Deligiannidis, Sheth, & Aleman-Meza, 2006).

Sentiment analysis can be used to organize the opinions portrayed in a text into groups; the most common of which are "Positive", "Negative", and less frequently "Neutral". Sentiment can be used to extract features from a text and studies have shown that there are numerous approaches for sentiment analysis (Ashraf, Verma, and Kavita, 2016). Sentiment analysis can be accomplished through a classifier-based approach which uses machine learning techniques to apply text classification as well through a lexicon-based approach which uses dictionaries of words with predetermined sentiment. It has been found that classifier-based methods work best when they are trained on a specific domain since some features of the domain may not hold the same meaning in other domains. There are various different lexicon libraries which represent words, word senses, or phrases in either a fixed category of positive or negative, in a grade scale such as strongly positive, mildly positive, neutral, mildly negative, or strongly negative, or by a value representation of strength such as an interval from -1 to +1 (Ahire, 2015).

Sentiment has been explored in various studies through a variety of different methods. Ortigosa, Martin, and Carro (2014) used a method for sentiment analysis that begins when a user

writes a message and is used to determine the polarity of the sentiment used in a post, it can also be used to explore the emotions of a user. Another method created by Pak and Paroubek (2010) spontaneously obtains data for analysis; the data is then used to apply a sentiment classifier which creates document level classification of positive, negative, or neutral. Agarwal et al. (2011) experimented with both a unigram model that utilized feature modeling as well as a tree 30 kernel-based model to first classify sentiment as either positive or negative and later to include a neutral classification. It was found that the tree kernel- based model outperformed other models and should be used for classification. Polarity ratios and n-gram graphs have also been used for content-based information by Aisopos, Papadakis, Tserpes, & Varvarigou, (2012). These models were able to provide the effectiveness that they desired while being language neutral as well as tolerant to noise; they did, however, not perform as well on large data due to extended computational time.

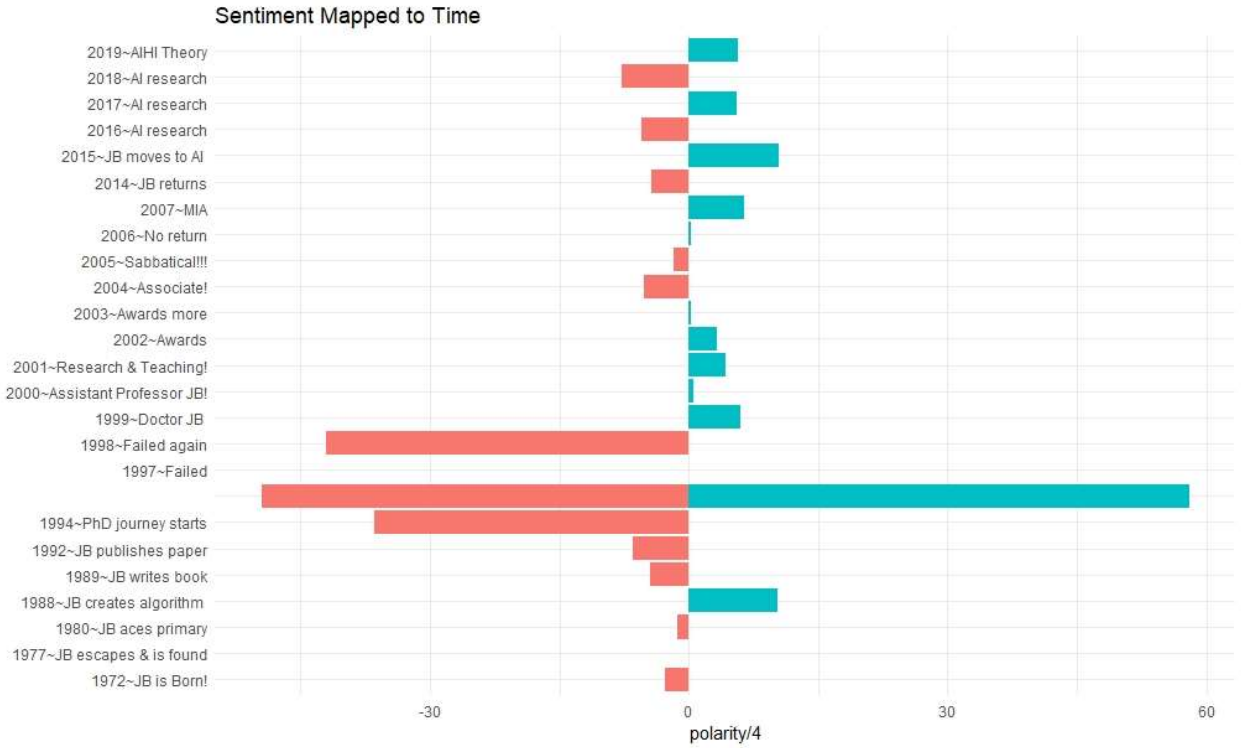

Figure 8: Blog word frequency-based sentiment related to life events in simulation.

Sentiment analysis has been used widely and can be used creatively to assign sentiment scores based on textual analysis, such as in life events simulation (Figure 8). Sentiment has also been used by Balahur et al. (2013) to distinguish good and bad news from good and bad sentiment and to look at contradictory views on newspaper articles. Business intelligence can also be viewed through sentiment through an analysis of messages on social media sites (Horakova, 2015). Sentiment can be used to determine if a tweet was sent out by a human or a bot by using word occurrence and emoticon usage to categorize users (Adarash and Kumar, 2015) Machine learning algorithms have been used to utilize capitalization, internalization, emoticons, and negation handling neutralization in tweet sentiment determination (Ashraf et al., 2016).

Research by Kabir et al. (2018) focused on the communication of feelings and sentiments in social media, they specifically chose social media since the language used is often casual and

since message lengths can be restricted; this area of sentiment is also not as well researched as those that have been developed on non-microblogging information therefore the methods that are not created specifically for social media may not be as effective. They utilized R to create a model and classify tweets from a famous Bangladeshi businessman and politician as positive, negative, and unbiased. Their findings were able to show that the user mostly posted neutral tweets, however the number of positive tweets exceeded negative ones. An important finding that must be noted is that when used on social media sentiment analysis cannot effectively detect sarcasm and will often classify it as negative sentiment (Kabir et al., 2018). Kretinin et al. (2018) also utilized sentiment to analyze a collection of tweets associated with several stocks. They classified tweets as positive, negative or neutral and focused only on how the positive and negative associated tweets related to stock price. They were able to show correlation between sentiment trends found and price movement for all of the companies that they analyzed (Figure 9).

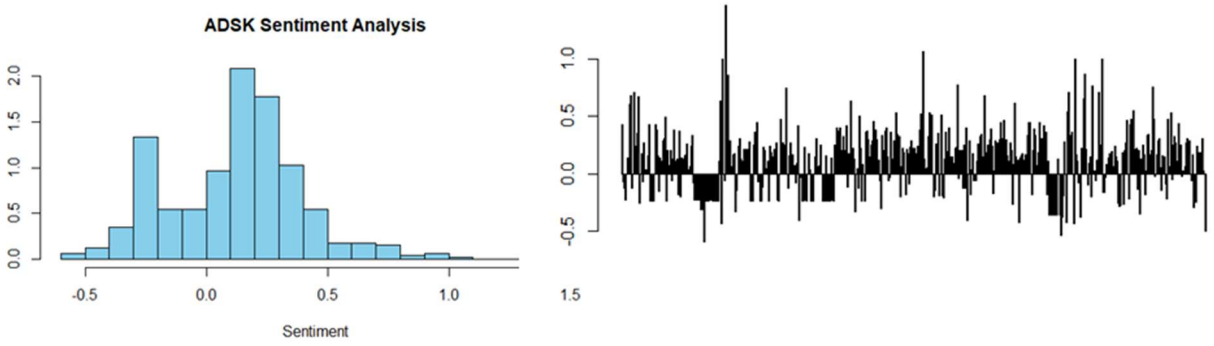

Figure 9: Autodesk Stock Sentiment Analysis Visualization

Deligiannidis et al. (2006) utilize virtual reality to create an interactive 3D visualization tool for semantic analysis called "Semantic Analytics Visualization" or SAV. This tool "*first visualizes ontologies, metadata, and heterogeneous information*"; it allows for interaction through the use of a virtual laser pointer for node selection, finally it presents results of semantic analytics techniques. This system is designed to be viewed by a group of people with one of them controlling the interaction. It is believed that a major advantage of SAV is that is can be used for analysis on the semantic web and that users can navigate, select, and query information as well as select documents. Users can also naturally interact with the environment even when there are hundreds of documents by using semi-transparency. Through studies it has been found that users have a high interest in this system and a system like this is important for interactive analysis (Deligiannidis et al., 2006).

Context Visualization

Context visualization uses graphical representations that apply meaning to the characters or sequences being analyzed; it can be used to investigate the "who, when, and where" of a social media post or to analyze the type device or connection being used. Social media users are a part of two types of networks; the first network, a follower network, is comprised all of the other users that a person follows (Chen et al., 2017). This follower network is generally based upon

relationships and interests, there are a number of tools that can be used for context visualization and tracking a user's follower network; some of these tools are as follows:

- OntoVis: uses structural and semantic abstraction to create simplifications of a network and for relationship analysis
- NodeXL: tool kit for network overview as well as discovery and exploration
- MatrixExplorer: offers node-link diagrams and matrix representation for network exploration
- MatLink: uses a hybrid representation with links on a matrix to show common neighbors and cliques
- NodeTrix: A combination of MatrixExplorer and MatLink where node links are used for global structure and a matrix is used for community structures
- iO-LAP: analyzes network data by looking at people, relation, content, and time
- GraphDic: provides easy comparisons for user attributes such as age, gender, and location
- DemographicVis: visual analytic system that supports interactive analysis of demographic groups

The other network, a reposting network, is determined by what a person posts and reposts, this reposting creates a diffusion of messages and it can also be visualized through a variety of tools such as the following.

- Google + Ripples: combines a node-link and a circular map to show information flow
- WeiboEvents: allows a user to explore the diffusion of information through either a tree layout, circular layouts, or a sail layout
- FluxFlow: provides "*an interactive system to analyze information*" spread in social media
- D-Map: provides "*a clear and intuitive visual summary of information diffusion among social communities*"
- Whisper: one of the earliest tools for analyzing the diffusion process of special temporal information; utilizes a sunflower visual metaphor

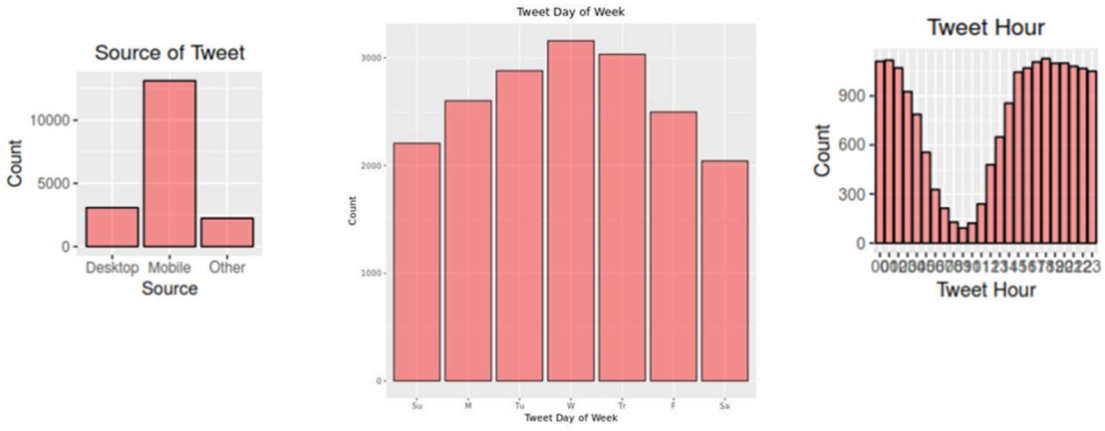

Figure 10: Viral Tweets Contextual Analysis Visualization

Context visualization and network analysis can benefit a number of industries; one suggested benefit of circular visual designs for land usage data and geo-tagged messages is in urban planning. Geo-tagged data can also be used to find patterns in points of interest (Chen et al.,

2017). Context visualization can easily be applied to Twitter data as there are more than twenty contextual variables that can be associated with a single tweet; this allows for the opportunity to gain contextual insights about the textual data (Samuel, Garvey & Kashyap, 2019).

Artificial Intelligence for Data Visualization – a Conceptual Framework

Artificial intelligence can be viewed from multiple perspectives, and though various schools of thoughts have diverse and often conflicting positions on a common definition for artificial intelligence, it is generally understood that artificial intelligence refers to technologies which match or surpass human capabilities in reasonable ways. We are seeing a huge increase in the use of artificial intelligence for a variety of purposes including, but not limited to, natural language processing, health care, financial markets, electronic training, hand writing recognition, human-computer interaction, image processing and computer vision, self-driving cars, autonomous weapons, intelligent information management and a host of other artificial intelligence applications. Artificial intelligence can be used for effective data visualization and has already been deployed in various forms of intelligent systems in gaming and in industry, such as in smart manufacturing: "*... real-time data can be visualized online via users' smart terminals. Through visualization, the results of data processing are made more accessible, straight-forward, and user-friendly*" (Tao et al., 2018). Artificial intelligence based research, modeling, automated data creation and collection, data representation and data visualization are being used across domains such as in hierarchical biology data (Kuznetsova et al., 2018); potential for visualizing library data effectively (Jiang & Carter, 2018); investigative journalism (Stray, 2019); hotel reviews and responses (Ku et al., 2019); study of unstable protocells (Points et al., 2018); business intelligence (Fombellida et al., 2018); and market segmentation (Kamthania et al., 2018), as examples of the broad examples of artificial intelligence methods and algorithm based modeling and data visualizations. There are a number of important dimensions that need consideration - artificial intelligence capabilities for visualizing structured and unstructured data, real time and non-temporal data, quantitative and qualitative data, textual analytics and language based modeling, adaptability to mathematical and statistical categories, and the psychology of the human sense of aesthetics. It would be outside of the scope of this chapter to identify and discuss such a broad set of issues in depth to evince the important underlying principles. However, we seek to emphasize the scope of possibilities as we provide an illustration of how artificial intelligence can potentially be used for automation of data visualization.

The present section aims to illustrate the scope for the use of artificial intelligence tools, methods and concepts for the purpose of data visualization. It is fairly obvious that artificial intelligence modules such as autonomous functioning of intelligent systems can be used to create real time data visualization streams with very little to no input from human intelligence. It is also evident that given the advanced capabilities of intelligent information systems, artificial intelligences can be programmed to map data visualizations to appropriate categories of data, and such intelligences can also be artificially trained to adapt to changing scenarios to capture and reflect potentially complex but useful changes to the implications of incoming data. Given a fair understanding of such basic concepts in artificial intelligence such as self-learning, autonomy and integration of complex information and intelligence modules into integrated or augmented

intelligences, additional ideas for artificial intelligence-based data visualization are summarized using the illustration below.

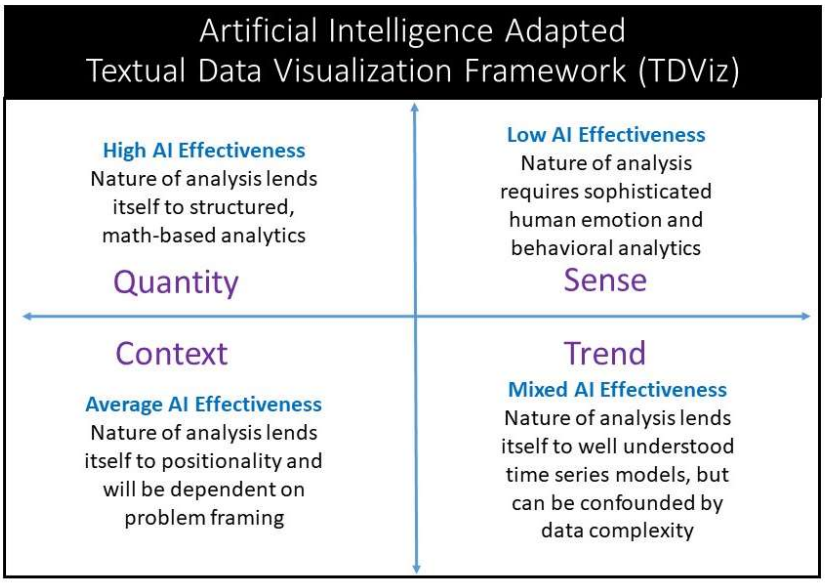

Figure 11: Artificial Intelligence Based Textual Data Visualization Framework Analysis

We adapt the textual data visualization framework (Conner et al., 2019, with permission), to analyze and demonstrate potential for alignment with artificial intelligence driven data visualization. The northwest quadrant (upper left) is reflective of high artificial intelligence effectiveness for data visualization. This is because the data and the methods anticipated to be used under this category are both standardized and fairly predictable in nature. Such a data scenario coupled with quantitative data visualization objectives would lend itself to mostly deterministic and robust data visualizations which are, by nature, programmable into artificial intelligence. Artificial intelligence can therefore be anticipated to perform well in this category of data visualizations. The southwest quadrant (lower left) is reflective of average artificial intelligence effectiveness for data visualization. This is primarily so because the data, the visualization objectives and the methods anticipated to be used under this category are standardized but not always predictable in nature, as it is expected to reflect contextuality. Such complexities coupled with quantitative data visualization objectives would lend itself to partially deterministic data visualizations which can be programmed into artificial intelligence with a reasonable, though not very high, level of success. The southeast quadrant (lower right) is reflective of mixed artificial intelligence effectiveness for data visualization. This is anticipated to be so because the visualization objectives, underlying informational stimuli and the methods expected to be used under this category are complex, and by nature have limited predictability. Since data are expected to reflect temporality, it can be expected to capture surprises which are both seasonal as well as non-seasonal. Such irregularities would data visualization programming challenging from an artificial intelligence perspective as it could be successful at times and fail at other times – hence leading to a mixed level of success for artificial intelligences success for data visualization. The northeast quadrant (upper right) is reflective of low artificial intelligence effectiveness for data visualization. This quadrant represents goals to identify and visualize human sentiment and

feelings. This being a nascent area for conclusive association of internal feelings and emotions with external stimuli or expressions, from a computational analysis, implies a significant gap in the body of knowledge. For example, even with human intelligence, multiple sentiment analysis frameworks could classify the same corpus of sentences differently. This would make it very challenging for artificial intelligence to produce high quality data visualizations which accurately reflect the underlying true sentiment from given external stimuli such as social media textual data.

Furthermore, it must be noted that with the increase in capabilities of artificial intelligence and continuing developments of learning technologies, coupled with an ability to process images and graphics, we can anticipate that artificial intelligence will rapidly improve in their data visualizations generation capabilities. With sufficient input on human behavior, human preferences for graphics and psychological factors data (or programmable functions), artificial intelligence can be expected to match human intelligence abilities to conceptualize and create new data visualizations designs and patterns.

Concluding notes.

Data visualization is a rich interdisciplinary domain with principles that can be objective across disciplines and methods that can be subjective to the specifics of the topic under consideration, and the nature of the data. In spite of much research and many developments in the practice of data visualization, there still remains much work to be done – this gap is primarily driven by the sophisticated constitution of the human mind and the general evolving understanding of the dynamic nature of information consumptions patterns of human intelligence. To the extent human intelligence can be deciphered and understood, data visualization can make predictable progress. It is also here within the scope of what is well understood about human psychology that we can aim to program artificial intelligences with data visualization capabilities, such that these artificial intelligences can function with a fair degree of autonomy creating appropriate visual illustration in real time. However, for the parts of the human mind which psychological sciences and neurosciences are still struggling to grasp, there are no easy ways to anticipate artificially intelligent data visualization systems, with satisfactory effectiveness.

The growing popularity of artificial intelligence models and methods, such as machine learning and deep learning, can be expected to have a powerful impact on the domain of data visualization: as we train models for effectiveness in data visualization parameters, mapped to measures of sensory and cognitive satisfaction, it is possible to anticipate artificial intelligence to develop data visualization capabilities that surpass human intelligences' capabilities for data visualization. However, human intelligence possesses a powerful fluidity of logic, often visible in expressions of creativity, which may not be surpassed soon by artificial intelligence. Therefore, the ideal way ahead for excellence in data visualization does not lie in human intelligence alone, nor in artificial intelligence by itself, but in an integrated functioning for human and artificial intelligences. Such integrated mechanisms for leveraging artificial intelligence are gaining popularity as mankind recognizes the tremendous possibilities that lie ahead. In conclusion, it must be emphasized that data visualization is a highly dynamic domain, and many powerful technologies driven changes can be anticipated in the future, leading to newer forms of data

visualization based on augmented intelligences and human desire for aesthetic satisfaction and excellence.

References

Adarash, M., & Ravikumar, P. (2015). Survey: Twitter data analysis using opinion mining. *International Journal of Computer Applications*, 128(5).

Agarwal, A., et al. (2011). Sentiment analysis of twitter data. *Association for Computational Linguistics*.

Ahire, S., (2015). A Survey of Sentiment Lexicons.

Aigner, W., Rind, A., & Hoffmann, S. (2012) Comparative Evaluation of an Interactive Time-Series Visualization that Combines Quantitative data with Qualitative Abstractions. *Eurographics Conference on Visualization (EuroVis) 2012*, 31(3), 995-1004.

Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012). Content vs. context for sentiment analysis: a comparative analysis over microblogs. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media - HT 12*.

Ashraf, S. S., Verma, S., & Kavita. (2016). A Survey on Sentiment Analysis Techniques on Social Media Data. *International Journal of Recent Research Aspects*, 3(3), 65-68.

Bach, B., et al. (2017). Immersive Analytics : Exploring Future Visualization and Interaction Technologies for Data Analytics. *Workshop at IEEE VIS.*

Balahur, A., et al. (2013). Sentiment analysis in the news. *arXiv*.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2010). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol*.

Cawthon, N. & Vande Moere, A., (2007). The Effect of Aesthetic on the Usability of Data Visualization. *11th International Conference Information Visualization*.

Chen, M., et al. (2009). Data, Information and Knowldege in Visualization. *IEEE Computer Graphics and Applications*.

Chen, S., Lin, L., & Yuan, X. (2017). Social Media Visual Analytics. *Computer Graphics Forum*, 36, 563-587.

Conner, C., Samuel, J., Kretinin, A., Samuel, Y. & Nadeau, L. (2019). A Picture for the Words! Textual Visualization in Big Data Analytics, *46th NBEA Annual Conference Proceedings*

Deligiannidis, L., Sheth, A. P., & Aleman-Meza, B. (2006). Semantic Analytics Visualization. *Lecture Notes in Computer Science*, 3975, 48-59.

Endert, A., et al. (2011). Observation-Level Interaction with Statistical Models for Visual Analytics. *Visual Analytics Science and Technology (VAST)*.

Fombellida, J., Martín-Rubio, I., Torres-Alegre, S., & Andina, D. (2018). Tackling business intelligence with bioinspired deep learning. *Neural Computing and Applications*, 1-8.

Friendly M. (2008) Milestones in the history of thematic cartography, statistical graphics, and data visualization.

Gray, C., Teahan, W. J., & Perkins, D., (2017) Understanding our Analytics: A visualization Survey. *Journal of Learning Analytics*.

Healey, C. G., (1996) Choosing Effective Colours for Data Visualization. *VIS '96 Proceedings of the 7th conference on Visualization '96*

Horakova, M. (2015). Sentiment Analysis Tool using Machine Learning. *Global Journal of Technology*. 195-204.

Howard, R. A., & Matheson, J. E. (2005). Influence Diagrams. *Decision Analysis*.

Jiang, Z., & Carter, R. (2018). Visualizing library data interactively: two demonstrations using R language. *Library Hi Tech News*, 35(5), 14-17.

Jimenez-Marquez, J. L., Gonzalez- Carrasco, I., Lopez-Cuadrado, J. L. & Ruiz-Mezuca, B. (2019). *Towards a big data framework for analyzing social media content. International Journal of Information management*, 44, 1-12.

Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds in R. *Informatica Economica*, 22, 25-38.

Kamthania, D., Pawa, A., & Madhavan, S. S. (2018). Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business. *Journal of computing and information technology*, 26(1), 57-68.

Keim, D., Mansmann, F., Thomas, J., (2009). Visual Analytics: How Much Visualization and How Much Analytics? *SIGKDD Explorations*, 5-8

Konukoglu, E., et al. (2011). Efficient Probabilistic Model Personalization intergrating uncertainty of Data and Parameters: Application to Eikonal-Diffusion Models in Cardiac Electrophysiology. *Progress in Biophysics and Molecular Biology*.

Kretinin, A., Samuel, J., & Kashyap, R. (2018). When the Going Gets Tough, The Tweets Get Going! An Exploratory Analysis of Tweets Sentiments in the Stock Market. *American Journal of Management*, 18(5), 23-36.

Ku, C. H., Chang, Y. C., Wang, Y., Chen, C. H., & Hsiao, S. H. (2019, January). Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews & Responses. *In Proceedings of the 52nd Hawaii International Conference on System Sciences.*

Kumar, N., Benbasat, I., (2004). The Effect of Relationship Encoding, Task Type, and Complexity on information Representation: An Empirical Evaluation of 2D and 3D Line Graphs. *MIS Quarterly,* 28(2), 255-281.

Kunz, M., Gret-Regamey, A., & Hurni, L. (2011). Visualization of uncertainty in natural hazards assessments using an interactive cartographic information system. *Nat Hazards*.

Kuznetsova, I., Lugmayr, A., & Holzinger, A. (2018). Visualisation Methods of Hierarchical Biological Data: A Survey and Review. *International SERIES on Information Systems and Management in Creative eMedia (CreMedia)*, (2017/2), 32-39.

Liu, Z., & Stasko, J. (2010). Theories in Information Visualization: What, Why and How. *Workshop on the Role of Theory in Information Visualization*.

Loughran, T. & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accountin Research*.

Nguyen, P.H., Xu, K., Walker, R., & Wong, B. W. (2016). TimeSets: Timeline Visualization with set relations. *Information Visualization*, 15(3), 253-269.

Ortigosa, A., Martin, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31, 527-541.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC.*

Points, L. J., Taylor, J. W., Grizou, J., Donkers, K., & Cronin, L. (2018). Artificial intelligence exploration of unstable protocells leads to predictable properties and discovery of collective behavior. Proceedings of the National Academy of Sciences, 115(5), 885-890.

Purchase, H. C., Andrienko, N., Jankun-Kelly, T., & Ward, M. (1970). Theoretical Foundations of Information Visualization. In *Information Visualization: Human-Centered Issues and Perspectives (Lecture Notes in Computer Science)* (pp. 51-70). Springer.

Ramos, M.-H., Bartholmes, J., & Thielen-del Pozo, J. (2007). Development of decision support products based on ensemble forecasts iin the European flood alert system. *Atmospheric Science Letters*.

Sahayak V., Shete, V., & Pathan, A. (2015). Sentiment Analysis on Twitter Data. *International Journal of Innovative Research in Advanced Engineering*, 2(1).

Samuel, J. (2017). Information Token Driven Machine Learning For Electronic Markets: Performance Effects In Behavioral Financial Big Data Analytics. *JISTEM-Journal of Information Systems and Technology Management*, 14(3), 371-383.

Samuel, J., Garvey, M., & Kashyap, R. (2019). That Message Went Viral?! Exploratory Analytics and Sentiment Analysis into the Propagation of Tweets. *In 2019 Annual Proceedings of Northeast Decision Sciences Institute (NEDSI)*. Philadelphia, USA.

Samuel, J., Holowczak, R., & Pelaez, A. (2017). The Effects of Technology Driven Information Categories on Performance in Electronic Trading Markets. *Journal of Information Technology Management.*

Samuel, J., Holowczak, R., Benbunan-Fich, R., & Levine, I. (2014). Automating Discovery of Dominance in Synchronous Computer-Mediated Communication. *2014 47th Hawaii International Conference on System Sciences*. 1804-1812.

Samuel, Y., George, J. & Samuel, J., (2018). Beyond STEM, How Can Women Engage Big Data, Analytics, Robotics And Artificial Intelligence? An Exploratory Analysis Of Confidence And Educational Factors In The Emerging Technology Waves Influencing The Role Of, And Impact Upon, Women. *In 2018 Annual Proceedings of Northeast Decision Sciences Institute (NEDSI) Conference, Rhode Island, USA*.

Stander, J. & Dalla Valle, J., (2017) On Enthusing Students About Big Data and Social Media Visualization and Analysis Using R, RStudio, and RMarkdown, *Journal of Statistics Education*, 25:2, 60-67.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39, 156-168.

Stray, J. (2019). Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism*, 1-22.

Strobel, B., Grund, S., Lindner, M. A., (2018). Do seductive details do their damage in the context of graph comprehension? Insights from eye movements. *Applied cognitive psychology,* 95-108.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157-169.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 48, 157-169.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate undertainty of deterministic models in decision support. *Environmental modelling and software*, 24-31.

Wang, C., & Shen, H.-W. (2011). Information Theory in Scientific Visualization. *Entropy*.

Weber, Z. J., & Gadepally, V., (2014). Using 3D Printing to Visualize Social Media Big Data. *ArXiv*.